HEIGHT-GUIDED PROJECTION REPARAMETERIZATION FOR CAMERA-LIDAR OCCUPANCY

Anonymous authors
Paper under double-blind review

ABSTRACT

3D occupancy prediction aims to infer dense, voxel-wise scene semantics from sensor observations, where the 2D-to-3D view transformation serves as a crucial step in bridging image features and volumetric representations. Most previous methods rely on a fixed projection space, where 3D reference points are uniformly sampled along pillars. However, such sampling struggles to capture the sparsity and height variations of real-world scenes, leading to ambiguous correspondences and unreliable feature aggregation. To address these challenges, we propose **HiPR**, a camera-LiDAR occupancy framework with **H**eight-Gu**i**ded **P**rojection **R**eparameterization. HiPR first encodes LiDAR into a BEV height map to capture the maximum height of the point cloud. HiPR then adjusts the sampling range of each pillar using the height prior, enabling adaptive reparameterization of the projection space. As a result, the projected points are redistributed into geometrically meaningful regions rather than fixed ranges. Meanwhile, we mask out the invalid parts of the height map to avoid misleading the feature aggregation. In addition, to alleviate the training instability caused by noisy LiDAR-derived heights, we introduce a training-time Progressive Height Conditioning strategy, which gradually transitions the conditioning signal from ground-truth heights to LiDAR heights. Extensive experiments demonstrate that HiPR consistently outperforms existing state-of-the-art methods while maintaining real-time inference. The code and pretrained models can be found at `https://anonymous.4open.science/r/HiPR_anonymous-C777`.

1 INTRODUCTION

3D occupancy prediction Ma et al. (2024a); Wang et al. (2023); Wei et al. (2023); Wu et al. (2025c); Zhu et al. (2026) aims to estimate voxel-wise scene semantics from sensor observations and is crucial for holistic scene understanding and motion planning in autonomous driving Dang et al. (2026); Du et al. (2025); Liu et al. (2025); Yang et al. (2025a); Zheng et al. (2024); Chen et al. (2026b). Camera-based occupancy prediction Cao & Vu (2026); Chen et al. (2025b;c); Leng et al. (2025); Li et al. (2026); Chen et al. (2026a) has attracted increasing attention, as cameras provide rich semantic cues at low cost and with wide coverage. To further improve performance, recent works Ming et al. (2024); Wei et al. (2025); Wolters et al. (2025); Yang et al. (2025b); Zheng et al. (2026) increasingly explore multi-modal frameworks that integrate complementary signals from LiDAR or radar.

Despite these advances, accurately lifting 2D image features into 3D space remains a critical challenge. In most previous methods, image features are associated with BEV queries through backward projection Li et al. (2023c; 2024); Liao et al. (2025), where a set of 3D reference points is projected onto image planes for feature aggregation. However, these points are predefined and uniformly sampled along globally shared pillars. Such a fixed projection space struggles to capture the sparsity and height variations of real-world scenes. As illustrated in Fig. 1(a), a car and a tall tree occupy vastly different height ranges. When using a fixed projection space (Fig. 1(b)), 3D reference points sampled above the car inevitably misalign with the actual object, leading to the erroneous aggregation of background features (e.g., the sky) instead of the target semantics. Furthermore, even for entirely empty regions devoid of objects, this fixed paradigm still redundantly assigns dense 3D reference points, forcing the network to aggregate meaningless image features. Consequently, this fixed projection space introduces ambiguous correspondences and unreliable feature aggregation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

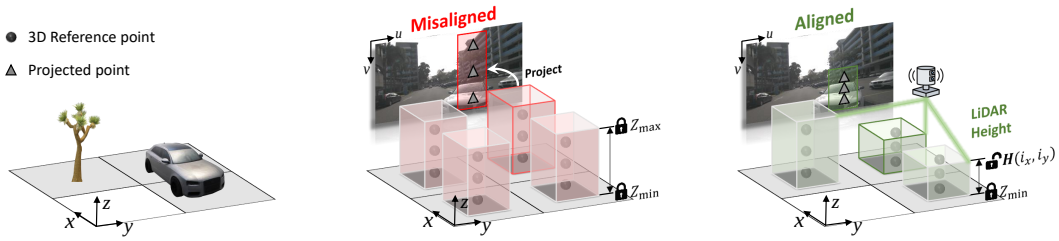


Figure 1: (a) Scene structure with varying heights. (b) Previous methods rely on a predefined fixed projection space, causing semantic misalignment on the 2D image plane. (c) Our method utilizes the LiDAR height prior to adaptively bound the sampling space of each pillar, ensuring alignment between projected reference points and scene structures for reliable feature aggregation.

To address these challenges, we propose HiPR, a camera-LiDAR occupancy framework with Height-Guided Projection Reparameterization. Specifically, as shown in Fig. 1(c), we first encode the LiDAR point cloud into a BEV height map, which captures the maximum height of the point cloud. Based on this height prior, we adjust the sampling range of each pillar. As a result, the 3D reference points are redistributed into geometrically meaningful regions rather than fixed height ranges. This enables the projected points to better align with the underlying scene structures in the image space, leading to more reliable feature aggregation. In addition, for BEV locations without LiDAR heights, we directly mask out the corresponding regions to prevent unreliable feature aggregation.

Moreover, to alleviate the training instability caused by noisy LiDAR-derived heights, we propose a training-time Progressive Height Conditioning strategy, which gradually transitions the conditioning signal from dense ground-truth heights to LiDAR heights during training.

Our main contributions are summarized as follows:

- We introduce a novel camera-LiDAR occupancy paradigm, which, for the first time, leverages the LiDAR height prior to guide the 2D-to-3D view transformation process.
- We propose HiPR, which (i) reparameterizes the projection space using a LiDAR-derived height map to align with the scene structure, and (ii) introduces a training-only Progressive Height Conditioning strategy to mitigate the impact of noisy LiDAR heights.
- Extensive experiments demonstrate that HiPR achieves state-of-the-art performance. Moreover, a lightweight variant outperforms existing real-time approaches, yielding a favorable trade-off between accuracy and efficiency. Code and weights are released.

2 Related Work

Multi-modal 3D Occupancy Prediction. Multi-modal 3D occupancy perception Duan et al. (2025); Ma et al. (2024b); Wang et al. (2024a; 2023); Wei et al. (2025); Wolters et al. (2025); Yang et al. (2025b); Zheng et al. (2026) leverages complementary information from multiple sensors to overcome the inherent limitations of single-modality perception. A common paradigm is to first lift image features into 3D space and then fuse them with point cloud features. For instance, OccFusion Ming et al. (2024) concatenates 3D features from different modalities along the feature dimension, followed by multi-scale feature extraction. However, the sensor sparsity and modality-specific visibility limit the effectiveness of naive feature concatenation. To address this issue, Co-Occ Pan et al. (2024a) enhances LiDAR features using neighboring camera features, and further introduces an implicit volumetric rendering–based regularization to supervise the fused features. In contrast, DAOcc Yang et al. (2025b) introduces 3D object detection as an auxiliary training task, providing additional supervision and improving performance on foreground categories. Beyond fusion strategies based on simple feature concatenation, recent works Song et al. (2025); Wei et al. (2025); Zheng et al. (2026) have explored more adaptive fusion mechanisms. For example, REOcc Song et al. (2025) employs cross-attention to fuse camera and radar BEV features. While effective, attention-based fusion introduces considerable computational overhead. To improve efficiency, Occ-Mamba Li et al. (2025) uses feature reordering to achieve global modeling with linear

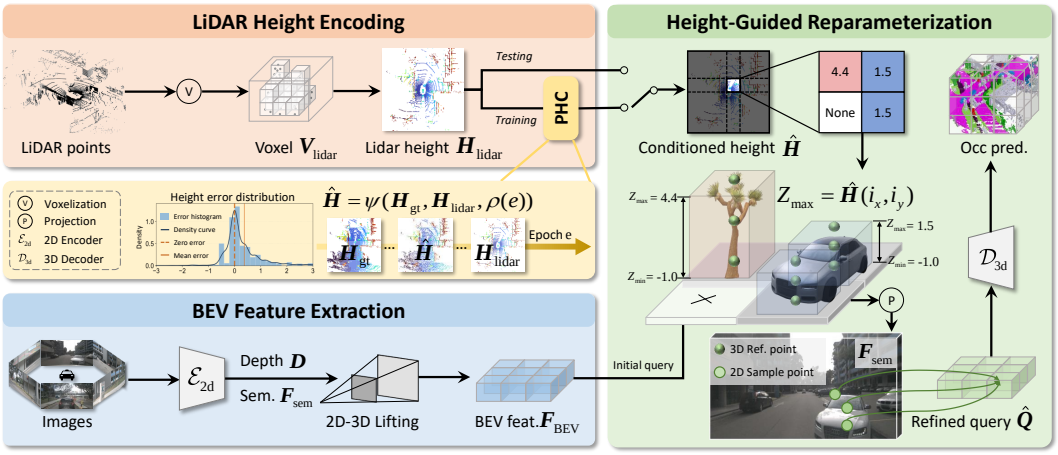


Figure 2: Overview of HiPR. Images are first transformed into a coarse BEV feature serving as the initial BEV query. Meanwhile, the LiDAR point cloud is encoded into a BEV height map and conditioned by PHC during training. Guided by the height map, the initial BEV query is refined through Height-Guided Reparameterization and finally decoded for 3D occupancy prediction.

complexity. In addition to BEV-level fusion, incorporating 2D perspective-view (PV) features has been shown to further refine 3D representations. For example, HyDRa Wolters et al. (2025) enhances PV features using BEV point-cloud features through cross-attention mechanisms. Unlike these fusion paradigms, we directly exploit LiDAR in the 2D-to-3D view transformation process.

Height Modeling. Motivated by the observation that objects are located at distinct height ranges, BEV-SAN Chi et al. (2023) constructs a height histogram from LiDAR points to determine the boundaries of local BEV slices and then fuses these slices with global BEV features to improve detection accuracy. Beyond slice fusion, subsequent methods incorporate height cues directly into the view transformation process. OC-BEV Qi et al. (2024) augments the uniform vertical sampling in BEVFormer Li et al. (2024) with a scene-level local height prior, thereby focusing sampling on height intervals where objects are most likely to appear. Extending this idea to finer granularity, HV-BEV Wu et al. (2025a) predicts a discrete height distribution for each BEV grid and selects the top-K values for feature sampling. Height cues are also important for 3D occupancy prediction, where vertical ambiguities can be more severe. For instance, DHD Wu et al. (2025b) decouples the projection space into multiple subspaces and employs the predicted height map to select features, alleviating height ambiguity. DA-Occ Zhou et al. (2025) instead leverages the height distribution to aggregate frustum features and further rearranges voxel features to apply direction-aware convolutions. Different from these explicit height-supervision paradigms, HBEVOcc Lyu et al. (2026) targets the loss of vertical information caused by folding the height dimension into BEV channels and proposes height-aware attention to implicitly capture latent vertical cues within BEV features. In contrast, we explicitly utilize a height map derived from LiDAR to reparameterize the projection space in view transformation, enabling better spatial alignment with the varying scene structure.

3 Method

3.1 Overview

As illustrated in Fig. 2, our framework consists of three stages. First, multi-view images are processed by the BEV Feature Extraction module, where a 2D encoder generates the depth distribution $\boldsymbol{D}$ and semantic features $\boldsymbol{F}_{\text{sem}}$. These features are lifted into the BEV space through an LSS-based view transformation, yielding a coarse BEV feature $\boldsymbol{F}_{\text{BEV}}$ that serves as the initial BEV query. In parallel, the LiDAR point cloud is encoded by the LiDAR Height Encoding module into a BEV height map $\boldsymbol{H}_{\text{lidar}}$, which is further conditioned by Progressive Height Conditioning during training to obtain the conditioned height map $\hat{\boldsymbol{H}}$. Next, the Height-Guided Reparameterization module leverages $\hat{\boldsymbol{H}}$ to adaptively reparameterize each BEV pillar and mask out invalid BEV grids, refining the initial BEV query into $\hat{\boldsymbol{Q}}$. Finally, $\hat{\boldsymbol{Q}}$ is fed into a 3D decoder for 3D occupancy prediction.

3.2 BEV Feature Extraction

Given multi-view input images, we first extract image features using a 2D encoder (e.g., ResNet He et al. (2016)). For each camera view, the encoder predicts a depth distribution $\boldsymbol{D}$ and a semantic feature map $\boldsymbol{F}_{\text{sem}}$ from dense image features. The depth distribution represents the probability of each pixel belonging to predefined depth bins, while $\boldsymbol{F}_{\text{sem}}$ captures appearance and contextual information for scene understanding. Following the LSS-based view transformation Li et al. (2023a;b); Philion & Fidler (2020), we lift $\boldsymbol{F}_{\text{sem}}$ from the 2D image plane into the BEV space using $\boldsymbol{D}$ and the camera intrinsic and extrinsic parameters. This process yields a coarse BEV feature, denoted as $\boldsymbol{F}_{\text{BEV}}$, which serves as the initial BEV query for the Height-Guided Reparameterization module.

3.3 LiDAR Height Encoding

Height Map Generation. To efficiently encode sparse LiDAR observations, we project the input point cloud into a BEV height map. Given a point cloud $\mathcal{P} = \{(x_k, y_k, z_k)\}_{k=1}^{K}$ in the ego-vehicle coordinate system, we first define a 3D region of interest bounded within $[x_{\min}, x_{\max})$, $[y_{\min}, y_{\max})$, and $[z_{\min}, z_{\max})$. The space is then discretized into a voxel grid with horizontal resolution Δ_{xy} and vertical resolution Δ_z. Each point $p_k \in \mathcal{P}$ is assigned to a voxel indexed by (i_x, i_y, i_z):

$$(i_x, i_y, i_z) = \left(\left\lfloor \frac{x_k - x_{\min}}{\Delta_{xy}} \right\rfloor, \left\lfloor \frac{y_k - y_{\min}}{\Delta_{xy}} \right\rfloor, \left\lfloor \frac{z_k - z_{\min}}{\Delta_z} \right\rfloor \right). \tag{1}$$

Based on the voxelized point cloud, we construct a binary occupancy grid $\boldsymbol{V}_{\text{lidar}} \in \{0,1\}^{X \times Y \times Z}$, where $\boldsymbol{V}_{\text{lidar}}(i_x, i_y, i_z) = 1$ if the corresponding voxel contains at least one LiDAR point. To obtain a 2D height representation, we collapse the 3D grid along the vertical axis by selecting the highest occupied voxel index of each non-empty BEV cell (i_x, i_y):

$$i_z^*(i_x, i_y) = \max \{ i_z \mid \boldsymbol{V}_{\text{lidar}}(i_x, i_y, i_z) = 1 \}. \tag{2}$$

We then map the discrete index back to metric height $\boldsymbol{H}_{\text{lidar}} \in \mathbb{R}^{X \times Y}$ as:

$$\boldsymbol{H}_{\text{lidar}}(i_x, i_y) = z_{\min} + (i_z^*(i_x, i_y) + 1)\, \Delta_z, \tag{3}$$

where the upper boundary of the highest occupied voxel is taken as the representative height. For empty pillars $\sum_{i_z} \boldsymbol{V}_{\text{lidar}}(i_x, i_y, i_z) = 0$, we assign a default invalid value to $\boldsymbol{H}_{\text{lidar}}(i_x, i_y)$.

Progressive Height Conditioning. Directly using the LiDAR-derived height map $\boldsymbol{H}_{\text{lidar}}$ for Height-Guided Reparameterization may destabilize training, since $\boldsymbol{H}_{\text{lidar}}$ is inherently noisy, as evidenced by the error distribution in Fig. 2. As the height map directly determines the sampling locations of the subsequent module, such noise can lead to inaccurate sampling.

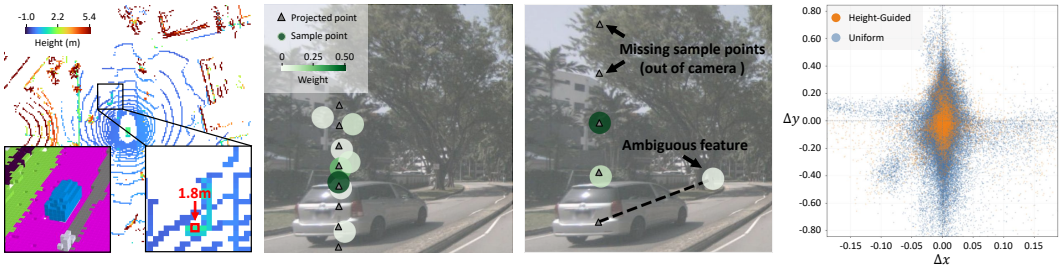
To mitigate this issue, we introduce a training-only Progressive Height Conditioning (PHC) strategy, which gradually shifts the conditioning signal from a dense ground-truth height map $\boldsymbol{H}_{\text{gt}}$ to the noisy LiDAR height map $\boldsymbol{H}_{\text{lidar}}$. Here, $\boldsymbol{H}_{\text{gt}}$ is derived from the occupancy ground truth using the same height extraction rule as in Eq. equation 2 and equation 3, ensuring that $\boldsymbol{H}_{\text{gt}}$ and $\boldsymbol{H}_{\text{lidar}}$ are geometrically aligned. During training, the final conditioning map $\hat{\boldsymbol{H}}$ is generated by:

$$\hat{\boldsymbol{H}} = \psi(\boldsymbol{H}_{\text{lidar}}, \boldsymbol{H}_{\text{gt}}, \rho(e)), \tag{4}$$

where $\psi(\cdot)$ denotes a stochastic mixing function that independently replaces each valid BEV grid in $\boldsymbol{H}_{\text{lidar}}$ with the corresponding value from $\boldsymbol{H}_{\text{gt}}$ with probability $\rho(e)$, while keeping the remaining grids unchanged. The mixing ratio $\rho(e)$ follows a cosine annealing schedule over E epochs:

$$\rho(e) = \frac{1}{2} \left(1 + \cos \left(\frac{\pi e}{E} \right) \right). \tag{5}$$

As a result, training starts with strong guidance from $\boldsymbol{H}_{\text{gt}}$ when $\rho(e) = 1$, and gradually transitions to the LiDAR-based setting used at inference as $\rho(e)$ decreases to zero. During inference, PHC is disabled and the conditioning map is simply given by $\hat{\boldsymbol{H}} = \boldsymbol{H}_{\text{lidar}}$. In this way, PHC stabilizes early training and improves the robustness of height-guided sampling against noisy LiDAR observations.



(a) Height map & Voxel GT (b) Height-Guided (c) Uniform (d) Offsets comparison

Figure 3: Comparison of height-guided and uniform sampling. (a) Height map and voxel ground truth. (b) Height-guided sampling concentrates projected reference points and sampled features around the object region. (c) Uniform sampling introduces invalid projections, including out-of-camera samples and ambiguous feature correspondences. (d) Learned cross-attention offsets show that Height-Guided Reparameterization yields a more compact distribution than uniform sampling.

3.4 Height-Guided Reparameterization

Revisiting Standard Backward Projection. In standard BEVFormer-style backward projection Li et al. (2023c; 2024); Liao et al. (2025), each BEV query aggregates image features from a set of 3D reference points sampled uniformly along a globally fixed pillar:

$$\hat{\boldsymbol{Q}}(i_x, i_y) = \sum_{j=1}^{N_z} \phi\left(\boldsymbol{Q}(i_x, i_y), \mathcal{R}\left(i_x, i_y, z_j\right), \boldsymbol{F}_{\text{sem}}\right), \tag{6}$$

where $\boldsymbol{Q}(i_x, i_y)$ and $\hat{\boldsymbol{Q}}(i_x, i_y)$ denote the input and updated BEV queries at location (i_x, i_y), $\phi(\cdot)$ represents the deformable cross-attention operator Li et al. (2024), $\mathcal{R}\left(i_x, i_y, z_j\right)$ stands for the 3D reference point at height z_j, $\boldsymbol{F}_{\text{sem}}$ refers to the multi-view image features, and N_z is the number of vertical sampling points. The vertical sampling locations Li et al. (2023c; 2024); Liao et al. (2025) are uniformly defined as:

$$z_j = z_{\min} + \alpha_j (z_{\max} - z_{\min}), \qquad \alpha_j = \frac{j-1}{N_z - 1}. \tag{7}$$

Since the vertical sampling locations $\{z_j\}_{j=1}^{N_z}$ are identical for all BEV grids, this formulation is inherently agnostic to scene geometry. As a result, reference points are often wasted in empty spaces, and even for object regions, the uniform sampling causes their 2D projections to misalign with the actual targets, leading to ambiguous correspondences and unreliable feature aggregation.

Reparameterization. To deal with this problem, we propose utilizing the conditioned height map $\hat{\boldsymbol{H}}(i_x, i_y)$ to dynamically reparameterize the projection space of each BEV query.

Specifically, we define a height-validity mask:

$$\boldsymbol{M}(i_x, i_y) = \mathbb{I}\big(\hat{\boldsymbol{H}}(i_x, i_y) \text{ is valid}\big), \tag{8}$$

and construct height-guided sampling locations as:

$$\tilde{z}_j(i_x, i_y) = z_{\min} + \alpha_j \big(\hat{\boldsymbol{H}}(i_x, i_y) - z_{\min}\big), \qquad \alpha_j = \frac{j-1}{N_z - 1}. \tag{9}$$

Accordingly, the query update process is reformulated as:

$$\hat{\boldsymbol{Q}}(i_x, i_y) = \begin{cases} \displaystyle\sum_{j=1}^{N_z} \phi(\boldsymbol{Q}(i_x, i_y), \mathcal{R}(i_x, i_y, \tilde{z}_j(i_x, i_y)), \boldsymbol{F}_{\text{sem}}), & \boldsymbol{M}(i_x, i_y) = 1, \\ \boldsymbol{Q}(i_x, i_y), & \boldsymbol{M}(i_x, i_y) = 0. \end{cases} \tag{10}$$

In this way, the conditioned height map $\hat{\boldsymbol{H}}$ defines the upper height bound of each pillar and restricts feature aggregation to valid BEV regions. The effectiveness of this reparameterization is illustrated in Fig. 3. Consider a target vehicle with a physical height of 1.8 m, as indicated by the LiDAR height

Table 1: Performance on Occ3D. The camera visible mask is used during training. Inputs include Camera (C), Radar (R) and LiDAR (L). R denotes ResNet, and Swin denotes Swin Transformer.

Method	Input	Backbone	mIoU	others	barrier	bicycle	bus	car	cons. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation
BEVDetOcc Huang et al. (2021)	C	Swin-B	42.0	12.2	49.6	25.1	52.0	54.5	27.9	28.0	28.9	27.2	36.4	42.2	82.3	43.3	54.6	57.9	48.6	43.6
FB-Occ Li et al. (2023c)	C	R50	39.8	13.8	44.5	27.1	46.2	49.7	24.6	27.4	28.5	28.2	33.7	36.5	81.7	44.1	52.6	56.9	42.6	38.1
FlashOcc Yu et al. (2023)	C	R50	32.0	6.2	39.6	11.3	36.3	44.0	16.3	14.7	16.9	15.8	28.6	30.9	78.2	37.5	47.4	51.4	36.8	31.4
COTR Ma et al. (2024a)	C	R50	44.5	13.3	52.1	32.0	46.0	55.6	32.6	32.8	30.4	34.1	37.7	41.8	84.5	46.2	57.6	60.7	52.0	46.3
ProtoOcc Kim et al. (2025)	C	R50	39.6	12.4	45.9	26.3	44.4	51.8	26.8	27.6	28.0	27.5	32.8	36.9	81.8	45.7	53.1	56.5	42.2	36.6
STCOcc Liao et al. (2025)	C	R50	45.0	15.2	52.3	32.2	50.5	56.5	31.7	33.9	33.4	33.8	38.9	44.9	83.9	47.4	57.1	60.1	50.6	42.7
RadOcc Zhang et al. (2024)	C+L	Swin-B	49.4	10.9	58.2	25.0	57.9	62.9	34.0	33.5	50.1	32.1	48.9	52.1	82.9	42.7	55.3	58.3	68.6	66.0
OccFusion Ming et al. (2024)	C+L+R	R101	46.7	12.4	50.3	31.5	57.6	58.8	34.0	41.0	47.2	29.7	42.0	48.0	78.4	35.7	47.3	52.7	63.5	63.3
HyDRa Wolters et al. (2025)	C+R	R50	44.4	15.1	51.1	32.7	52.3	56.3	29.4	35.9	35.1	33.7	39.1	44.1	80.4	45.1	52.0	55.3	52.1	44.4
EFFOcc Shi et al. (2025)	C+L	R50	52.8	12.1	59.7	33.4	61.8	65.0	35.5	46.0	57.1	41.0	47.9	54.6	82.8	44.0	56.4	60.2	**71.1**	**69.6**
SDG-Fusion Duan et al. (2025)	C+L	R50	51.7	13.2	57.8	24.3	60.3	64.3	36.2	39.4	52.4	35.8	**50.9**	53.7	84.6	47.5	58.0	61.6	70.7	67.7
SDG-KL Duan et al. (2025)	C+L	R50	50.2	12.3	57.1	23.7	58.8	62.7	34.6	36.2	50.1	32.1	49.9	51.2	84.1	46.1	57.2	61.5	69.6	65.8
ALOcc-2D-mini Chen et al. (2025a)	C+L	R50	50.0	15.7	54.6	36.6	55.7	60.6	34.8	41.0	44.9	39.3	44.5	51.1	83.6	48.5	57.3	60.2	62.7	58.2
ALOcc-2D Chen et al. (2025a)	C+L	R50	53.5	16.5	57.8	41.6	57.9	63.8	37.6	45.0	52.1	45.8	49.6	54.4	**85.3**	50.5	59.7	62.3	67.1	62.0
DAOcc Yang et al. (2025b)	C+L	R50	54.3	13.0	60.7	39.8	**64.0**	**66.5**	36.3	49.0	**60.1**	44.3	50.7	55.9	82.9	44.6	56.8	60.6	70.1	68.3
HiPR-mini	C+L	R50	53.1	17.0	58.8	38.8	58.0	62.8	37.7	45.4	55.7	43.1	49.3	53.5	83.9	49.3	57.9	60.7	66.9	63.8
HiPR	C+L	R50	**55.5**	**17.6**	**61.0**	**42.0**	60.9	65.3	**40.0**	**49.3**	55.6	**48.5**	50.7	**56.2**	**85.3**	**51.2**	**60.0**	**62.4**	70.3	66.8

map in Fig. 3(a). Under the standard uniform sampling paradigm (Fig. 3(c)), the model samples reference points up to the maximum predefined ceiling (e.g., $z_{\max} = 5.4$ m). This lack of geometric awareness results in numerous invalid projections that either fall out of the camera field of view or sample ambiguous background features (e.g., trees). In contrast, by dynamically truncating the sampling range to the actual object height ($[z_{\min}, 1.8]$ m), our height-guided method encourages the sampling points to concentrate on the semantically meaningful vehicle body (Fig. 3(b)). Furthermore, as shown in Fig. 3(d), the learned spatial offsets $(\Delta x, \Delta y)$ of our method exhibit a compact, zero-centered distribution. This suggests that the initial 3D projections are well aligned with informative regions and require only minor adjustments. In contrast, uniform sampling produces a more scattered distribution, indicating that the model relies on larger offsets to capture valid features.

3.5 Loss Function

Following Chen et al. (2025a); Cheng et al. (2022); Liu et al. (2024), we train the query-based decoder with a classification loss, a mask loss, and a dice loss. The overall objective is defined as:

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}, \tag{11}$$

where $\mathcal{L}_{\text{cls}}$ denotes the cross-entropy loss for query classification, $\mathcal{L}_{\text{mask}}$ denotes the cross-entropy loss for occupancy masks, and $\mathcal{L}_{\text{dice}}$ denotes the dice loss for mask overlap optimization. In our implementation, the loss weights are set to $\lambda_{\text{cls}} = 2.0$, $\lambda_{\text{mask}} = 5.0$, and $\lambda_{\text{dice}} = 5.0$.

4 Experiments

Datasets and Metrics. We evaluate our method on the nuScenes dataset Caesar et al. (2020), which contains 700 training scenes and 150 validation scenes with annotations at 2 Hz. Based upon nuScenes, Occ3D Tian et al. (2023) and SurroundOcc Wei et al. (2023) further provide voxel-wise occupancy annotations. For Occ3D, the voxel region spans $[-40, 40]$ m along the X and Y axes and $[-1, 5.4]$ m along the Z axis, with a voxel size of $0.4 \times 0.4 \times 0.4$ m. It contains 18 semantic categories, including 17 semantic classes and 1 empty class. For SurroundOcc, the voxel region is $[-50, 50]$ m along the X and Y axes and $[-5, 3]$ m along the Z axis, with a voxel size of $0.5 \times 0.5 \times 0.5$ m. Compared with Occ3D, SurroundOcc excludes the "others" category. Following prior work Yang et al. (2025b), we use mIoU and RayIoU as evaluation metrics.

Implementation Details. Built upon the ALOcc series Chen et al. (2025a), we develop two variants: a lightweight **HiPR-mini** for real-time performance, and a standard **HiPR** that incorporates stereo depth estimation and larger channel dimensions. Following DAOcc Yang et al. (2025b), we aggregate LiDAR sweeps from 10 adjacent timestamps as input. We optimize all models using AdamW Loshchilov & Hutter (2017) with an initial learning rate of 2×10^{-4} for 24 epochs and a

Table 2: Comparison of RayIoU on Occ3D. The camera visible mask is not used during the training.

Method	Input	Backbone	RayIoU	$\text{RayIoU}_{1m, 2m, 4m}$		
BEVFormer Li et al. (2024)	C	R101	32.4	26.1	32.9	38.0
FB-Occ Li et al. (2023c)	C	R50	39.0	33.0	40.0	44.0
RenderOcc Pan et al. (2024b)	C	Swin-B	19.5	13.4	19.6	25.5
SparseOcc Liu et al. (2024)	C	R50	36.1	30.2	36.8	41.2
FlashOcc Yu et al. (2023)	C	R50	38.5	32.8	39.3	43.4
OPUS Wang et al. (2024b)	C	R50	41.2	34.7	42.1	46.7
STCOcc Liao et al. (2025)	C	R50	42.1	36.9	42.8	46.7
ALOcc-2D-mini Chen et al. (2025a)	C	R50	39.3	32.9	40.1	44.8
ALOcc-2D Chen et al. (2025a)	C	R50	43.0	37.1	43.8	48.2
DAOcc Yang et al. (2025b)	C+L	R50	48.4	44.5	48.9	51.9
HiPR-mini	C+L	R50	48.4	44.2	49.0	52.0
HiPR	C+L	R50	**50.0**	**45.9**	**50.5**	**53.4**

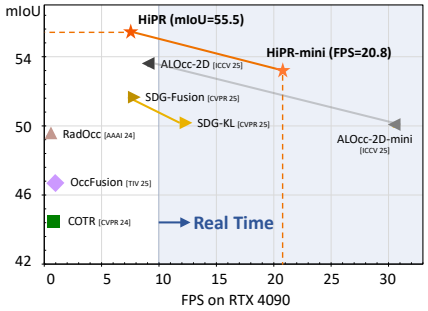


Figure 4: FPS-mIoU comparison.

Table 3: Performance on SurroundOcc. Unlike Occ3D, the evaluation is performed over all voxels.

Method	Input	Backbone	mIoU	barrier	bicycle	bus	car	cons. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation
BEVFormer Li et al. (2024)	C	R101	16.8	14.2	6.6	23.5	28.3	8.7	10.8	6.6	4.1	11.2	17.8	37.3	18.0	22.9	22.2	13.8	22.2
TPVFormer Huang et al. (2023)	C	R101	17.1	16.0	5.3	23.9	27.3	9.8	8.7	7.1	5.2	11.0	19.2	38.9	21.3	24.3	23.2	11.7	20.8
SurroundOcc Wei et al. (2023)	C	R101	20.3	20.6	11.7	28.1	30.9	10.7	15.1	14.1	12.1	14.4	22.3	37.3	23.7	24.5	22.8	14.9	21.9
GaussianFormer Huang et al. (2024)	C	R101	17.3	19.5	11.3	26.1	29.8	10.5	13.8	12.6	8.7	12.7	21.6	39.6	23.3	24.5	23.0	9.6	19.1
GaussianFormer2 Huang et al. (2025)	C	R101	20.8	21.4	13.4	28.5	30.8	10.9	15.8	13.6	10.5	14.0	22.9	40.6	24.4	26.1	24.3	13.8	22.0
M-CONet Wang et al. (2023)	C+L	R101	24.7	24.8	13.0	31.6	34.8	14.6	18.0	20.0	14.7	20.0	26.6	39.2	22.8	26.1	26.0	26.0	37.1
Co-Occ Pan et al. (2024a)	C+L	R101	27.1	28.1	16.1	34.0	37.2	17.0	21.6	20.8	15.9	21.9	28.7	42.3	25.4	29.1	28.6	28.2	38.0
OccFusion Ming et al. (2024)	C+L	R101	27.6	25.2	19.9	34.8	36.2	20.0	23.1	25.3	17.5	**22.7**	30.1	39.5	23.3	25.7	27.6	29.5	40.6
OccFusion Ming et al. (2024)	C+L+R	R101	28.3	28.3	21.0	**35.1**	36.8	20.3	26.2	25.9	19.2	21.3	30.6	40.1	23.8	25.6	27.6	29.8	40.8
OccCylindrical Ming et al. (2025)	C+L	R50	28.7	26.2	22.1	31.5	36.8	18.0	27.8	**29.9**	**23.9**	20.6	28.3	43.0	23.1	28.0	27.8	**30.8**	**41.0**
HiPR-mini	C+L	R50	28.6	27.9	21.7	33.2	36.8	22.8	27.7	27.0	19.9	20.2	30.0	41.4	26.8	28.7	27.7	28.3	37.5
HiPR	C+L	R50	**30.4**	**29.9**	**23.1**	34.8	**38.4**	**23.4**	**29.7**	28.8	21.9	22.4	**32.0**	**43.8**	**28.6**	**30.6**	**29.9**	30.6	39.4

total batch size of 16. Our implementation is built on MMDetection3D Chen et al. (2019). Experiments are conducted on NVIDIA GeForce RTX 4090 and L40 GPUs.

4.1 Comparison with the State-of-the-Art

Comparison on Occ3D. As shown in Tab. 1 and Tab. 2, we evaluate HiPR on Occ3D under both training settings: with and without the camera mask. Overall, our method achieves state-of-the-art performance in both scenarios. Under the *with camera mask* setting, we report mIoU. HiPR outperforms the second-best DAOcc Yang et al. (2025b) by 1.2 mIoU. Compared to OccFusion Ming et al. (2024), which leverages additional radar input and a larger backbone (ResNet-101 He et al. (2016)), HiPR achieves a substantial gain of 8.8 mIoU. Under the *without camera mask* setting, we report RayIoU Liu et al. (2024). HiPR significantly outperforms camera-based methods such as STCOcc Liao et al. (2025) by a large margin of 7.9 RayIoU. Compared to the multi-modal method DAOcc, the lightweight variant HiPR-mini achieves comparable performance, while the full HiPR further improves upon it by 1.6 RayIoU. Fig. 5 shows qualitative comparisons between HiPR and DAOcc. DAOcc misclassifies large objects and fails to reconstruct distant structures accurately, while HiPR yields more accurate semantics and more complete reconstructions.

Comparison on SurroundOcc. Tab. 3 presents a quantitative comparison on the SurroundOcc Wei et al. (2023) dataset. Overall, HiPR consistently outperforms existing methods. In particular, it surpasses the second-best method, OccCylindrical Ming et al. (2025), by 1.7 mIoU. Meanwhile, despite using only a lightweight ResNet-50 He et al. (2016) backbone, HiPR still outperforms competitive approaches such as OccFusion Ming et al. (2024), Co-Occ Pan et al. (2024a), and GaussianFormer Huang et al. (2024). Furthermore, HiPR achieves clear improvements on challenging categories such as bicycle and motorcycle, which are typically sparse and difficult to model, thereby demonstrating its ability to better capture fine-grained structures.

Speed–Accuracy Trade-off. Fig. 4 presents the speed–accuracy trade-off of HiPR on Occ3D under the setting trained without camera visible mask. Following prior works He et al. (2025); Hou et al. (2024); Kim et al. (2025), methods running above 10 FPS are regarded as real-time. Overall, HiPR

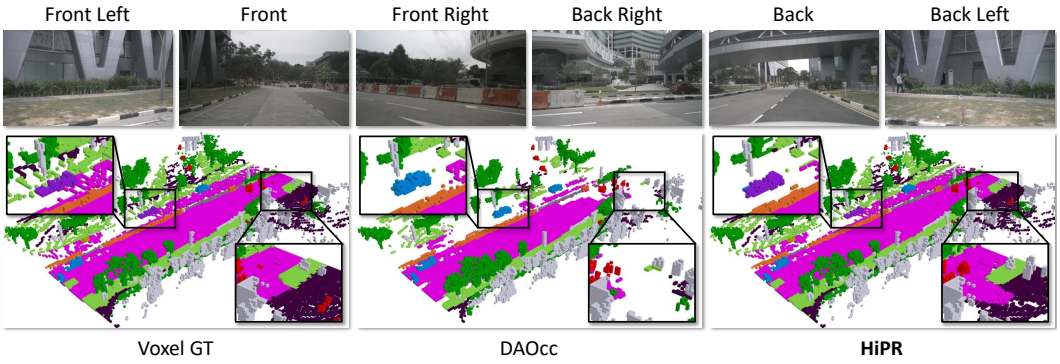


Figure 5: Visual comparisons of DAOcc and our HiPR on the validation set of Occ3D.

Table 4: Ablation study of HiPR on the Occ3D dataset. HGR refers to Height-Guided Reparameterization and PHC refers to Progressive Height Conditioning.

Model	HGR		PHC	mIoU
	Height-Guided Sampling	Height-Validity Mask		
Baseline				50.01
i	✓			51.04
ii	✓	✓		52.81
iii	✓	✓	✓	53.10

Table 5: Ablation study of different sampling strategies in HGR module.

Strategy	mIoU		
	Overall	Road	Object
Full-range	50.57	61.95	47.48
Upper bound (3 m)	50.63	62.42	47.32
Upper bound (2 m)	50.81	62.58	47.57
Road-biased	51.03	62.65	47.87
Height-guided	**52.81**	**62.85**	**49.95**

attains state-of-the-art accuracy, while HiPR-mini enables real-time inference with near state-of-the-art performance, outperforming existing feature-fusion-based multi-modal methods Duan et al. (2025); Ming et al. (2024); Zhang et al. (2024). Specifically, compared with the real-time method SDG-KL Duan et al. (2025), HiPR-mini is 7.2 FPS faster and improves mIoU by 2.9 points, demonstrating its superior speed–accuracy trade-off for efficient 3D occupancy prediction.

4.2 Ablation Study

All experiments in this section are trained with the camera visible mask on the Occ3D dataset.

HiPR Designs. Tab. 4 presents the ablation study of the proposed Height-Guided Reparameterization (HGR) and Progressive Height Conditioning (PHC). The baseline model is ALOcc-2D-mini. Starting from this baseline, introducing height-guided sampling improves the mIoU from 50.01 to 51.04. In this setting, valid pillars are reparameterized according to the height map, while invalid regions still use uniformly sampled pillars over the maximum height range. Further incorporating the height-validity mask brings a larger improvement, boosting the mIoU to 52.81. This demonstrates that restricting feature aggregation to valid BEV regions is crucial for suppressing invalid projections and improving geometric alignment. As shown in Fig. 6, the BEV features are more focused and semantically consistent in occupied regions, while noisy responses in empty areas are effectively suppressed. Finally, PHC improves the performance by 0.29 mIoU, indicating that progressively transitioning stabilizes training and improves the robustness of height-guided sampling.

Sampling Strategies in HGR. Tab. 5 compares different sampling strategies in HGR. Starting from HiPR-mini, we keep all other components fixed and vary only the sampling strategy. Uniform sampling over the full vertical range achieves 50.57 mIoU. Restricting the sampling range with fixed upper bounds of 3 m and 2 m yields only marginal gains of 0.06 and 0.24 mIoU, respectively, with most of the improvement coming from road-related classes. Since driving scenes are strongly biased toward near-ground structures, we further evaluate a road-biased strategy, which allocates half of the reference points over the full vertical range and the other half to the lower region $[-1, 2]$ m. This strategy improves the overall mIoU to 51.03, but still provides limited gains on object classes. In contrast, our height-guided strategy adapts the sampling range to each pillar and achieves 52.81

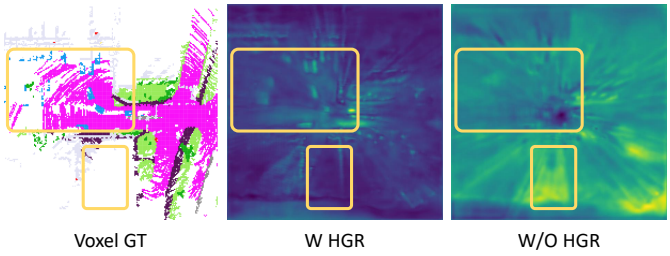


Figure 6: Comparison of BEV features with and without HGR.

Table 6: Ablation study of conditioning strategy in PHC module.

Setting	mIoU
LiDAR Only	52.81
GT Only	31.83
Linear Interpolation	52.83
Replace (**Ours**)	53.10

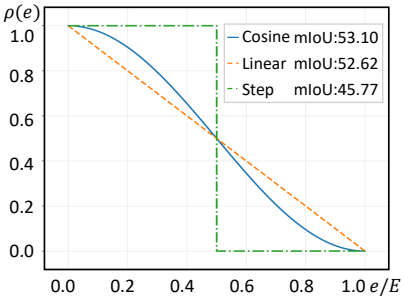


Figure 7: PHC schedules ablation.

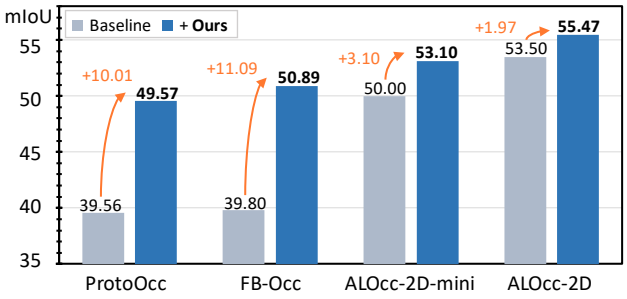


Figure 8: Generalization across different frameworks.

mIoU, outperforming the full-range baseline by 2.24 mIoU. Notably, the improvement is much larger on object classes (+2.47 mIoU) than on road-related classes (+0.90 mIoU), demonstrating the importance of pillar-specific adaptation for structures with diverse vertical extents.

Conditioning Strategy in PHC. As shown in Tab. 6, we compare different conditioning strategies in PHC. Using only LiDAR-derived heights as the baseline achieves 52.81 mIoU. In contrast, using ground-truth heights throughout training severely degrades performance to 31.83 mIoU, revealing a large gap between the clean training signal and noisy LiDAR at inference. We further compare two mixing strategies. Linear interpolation blends ground-truth and LiDAR-derived height maps but brings only limited improvement, while our hard replacement strategy achieves the best performance of 53.10 mIoU. Moreover, we ablate different PHC schedules in Fig. 7. The step schedule causes an abrupt transition and degrades performance by 7.33 mIoU, whereas the cosine schedule gradually shifts from ground-truth to LiDAR-derived heights and achieves the best result.

4.3 Generalization Capability

We further evaluate the generalization of our design across different occupancy frameworks, including ProtoOcc Kim et al. (2025), FB-Occ Li et al. (2023c), and the ALOcc-2D series Chen et al. (2025a). For ProtoOcc and ALOcc, we insert HGR before the occupancy head, while for FB-Occ, we replace the standard backward projection with HGR. As shown in Fig. 8, our method consistently improves all baselines. Specifically, it brings gains of 10.01 and 11.09 mIoU on vision-based ProtoOcc and FB-Occ, respectively, and improves the LiDAR-assisted ALOcc variants by 3.10 and 1.97 mIoU. These consistent gains demonstrate that HGR and PHC generalize well across different frameworks and can serve as effective plug-and-play modules for 3D occupancy prediction.

5 Conclusion

In this paper, we present HiPR, a novel camera–LiDAR framework that reparameterizes the projection space using the LiDAR height prior. By encoding LiDAR observations into a BEV height map, HiPR introduces height-guided sampling and height-validity mask to adaptively adjust sampling range and grid validity, thereby improving geometric alignment in the projection space. In addition, we leverage ground-truth heights during training to stabilize optimization. Extensive experiments demonstrate the superiority of HiPR in accuracy, efficiency, and generalization.

AI USE STATEMENT

In this work, we used generative AI tools for translation and language polishing. We have not used generative AI tools for research idea generation or method design. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our research focuses on 3D occupancy prediction, a key perception problem in autonomous driving. All training and evaluation are conducted using publicly available datasets. This research does not involve human subjects and has been conducted in accordance with standard research ethics and responsible research practices.

REPRODUCIBILITY STATEMENT

To support reproducibility and facilitate further research by the community, we provide detailed descriptions of the model architecture, training settings, and evaluation protocols in this paper. The source code and pretrained models are available at `https://anonymous.4open.science/r/HiPR_anonymous-C777`.

REFERENCES

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pp. 11621–11631, 2020.

Anh-Quan Cao and Tuan-Hung Vu. Occany: Generalized unconstrained urban 3d occupancy. In *CVPR*, 2026.

Dubing Chen, Jin Fang, Wencheng Han, Xinjing Cheng, Junbo Yin, Chengzhong Xu, Fahad Shahbaz Khan, and Jianbing Shen. Alocc: Adaptive lifting-based 3d semantic occupancy and cost volume-based flow predictions. In *ICCV*, pp. 4156–4166, 2025a.

Dubing Chen, Huan Zheng, Jin Fang, Xingping Dong, Xianfei Li, Wenlong Liao, Tao He, Pai Peng, and Jianbing Shen. Rethinking temporal fusion with a unified gradient descent view for 3d semantic occupancy prediction. In *CVPR*, pp. 1505–1515, 2025b.

Dubing Chen, Huan Zheng, Yucheng Zhou, Xianfei Li, Wenlong Liao, Tao He, Pai Peng, and Jianbing Shen. Semantic causality-aware vision-based 3d occupancy prediction. In *ICCV*, pp. 24878–24888, 2025c.

Dubing Chen, Huan Zheng, Tianyi Yan, Yucheng Zhou, Runzhou Tao, Zhongying Qiu, Jianfei Yang, and Jianbing Shen. Fdr-occ: Factorized dense routing for full-spectrum 3d occupancy prediction. In *ECCV*, pp. 95–111. Springer, 2026a.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Sitao Chen, Zhuangwei Zhuang, Hui Luo, Qingyao Wu, and Mingkui Tan. Geometry-aware spatio-temporal context modeling for 4d occupancy forecasting. *arXiv preprint arXiv:2608.15279*, 2026b.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pp. 1290–1299, 2022.

Xiaowei Chi, Jiaming Liu, Ming Lu, Rongyu Zhang, Zhaoqing Wang, Yandong Guo, and Shanghang Zhang. Bev-san: Accurate bev 3d object detection via slice attention networks. In *CVPR*, pp. 17461–17470, 2023.

Chenxu Dang, Haiyan Liu, Jason Bao, Pei An, Xinyue Tang, An Pan, Jie Ma, Bingchuan Sun, and Yan Wang. Sparseworld: A flexible, adaptive, and efficient 4d occupancy world model powered by sparse and dynamic queries. In *AAAI*, volume 40, pp. 3497–3505, 2026.

Jiayuan Du, Yiming Zhao, Zhenglong Guo, Yong Pan, Wenbo Hou, Zhihui Hao, Kun Zhan, and Qijun Chen. Sparseworld-tc: Trajectory-conditioned sparse occupancy world model. *arXiv preprint arXiv:2511.22039*, 2025.

ZaiPeng Duan, ChenXu Dang, Xuzhong Hu, Pei An, Junfeng Ding, Jie Zhan, YunBiao Xu, and Jie Ma. Sdgocc: Semantic and depth-guided bird's-eye view transformation for 3d multimodal occupancy prediction. In *CVPR*, pp. 6751–6760, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Yulin He, Wei Chen, Siqi Wang, Tianci Xun, and Yusong Tan. Achieving speed-accuracy balance in vision-based 3d occupancy prediction via geometric-semantic disentanglement. In *AAAI*, volume 39, pp. 3455–3463, 2025.

Jiawei Hou, Xiaoyan Li, Wenhao Guan, Gang Zhang, Di Feng, Yuheng Du, Xiangyang Xue, and Jian Pu. Fastocc: Accelerating 3d occupancy prediction by fusing the 2d bird's-eye view and perspective view. In *ICRA*, pp. 16425–16431. IEEE, 2024.

Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *CVPR*, pp. 9223–9232, 2023.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. In *ECCV*, pp. 376–393. Springer, 2024.

Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. Gaussianformer-2: Probabilistic gaussian superposition for efficient 3d occupancy prediction. In *CVPR*, pp. 27477–27486, 2025.

Jungho Kim, Changwon Kang, Dongyoung Lee, Sehwan Choi, and Jun Won Choi. Protoocc: Accurate, efficient 3d occupancy prediction using dual branch encoder-prototype query decoder. In *AAAI*, volume 39, pp. 4284–4292, 2025.

Ziyang Leng, Jiawei Yang, Wenlong Yi, and Bolei Zhou. Occupancy learning with spatiotemporal memory. In *ICCV*, pp. 26569–26578, 2025.

Heng Li, Yuenan Hou, Xiaohan Xing, Yuexin Ma, Xiao Sun, and Yanyong Zhang. Occmamba: Semantic occupancy prediction with state space models. In *CVPR*, pp. 11949–11959. IEEE, 2025.

Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *AAAI*, volume 37, pp. 1486–1494, 2023a.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *AAAI*, volume 37, pp. 1477–1485, 2023b.

Zhihao Li, Shanshan Zhang, and Jian Yang. Ashsr: Enhancing query-based occupancy prediction via anti-occlusion sampling and hard sample reweighting. *Neurocomputing*, 676:133007, 2026.

Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023c.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2020–2036, 2024.

Zhimin Liao, Ping Wei, Shuaijia Chen, Haoxuan Wang, and Ziyang Ren. Stcocc: Sparse spatial-temporal cascade renovation for 3d occupancy and scene flow prediction. In *CVPR*, pp. 1516–1526, 2025.

Haisong Liu, Yang Chen, Haiguang Wang, Zetong Yang, Tianyu Li, Jia Zeng, Li Chen, Hongyang Li, and Limin Wang. Fully sparse 3d occupancy prediction. In *ECCV*, pp. 54–71. Springer, 2024.

Ruixun Liu, Lingyu Kong, Derun Li, and Hang Zhao. Occvla: Vision-language-action model with implicit 3d occupancy supervision. *arXiv preprint arXiv:2509.05578*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Chuandong Lyu, Wenkai Li, Iman Yi Liao, Fengqian Ding, Han Liu, and Hongchao Zhou. Hbevocc: Height-aware bird's-eye-view representation for 3d occupancy prediction from multi-camera images. *Sensors*, 26(3):934, 2026.

Qihang Ma, Xin Tan, Yanyun Qu, Lizhuang Ma, Zhizhong Zhang, and Yuan Xie. Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction. In *CVPR*, pp. 19936–19945, 2024a.

Yukai Ma, Jianbiao Mei, Xuemeng Yang, Licheng Wen, Weihua Xu, Jiangning Zhang, Xingxing Zuo, Botian Shi, and Yong Liu. Licrocc: Teach radar for accurate semantic occupancy prediction using lidar and camera. *IEEE Robotics and Automation Letters*, 10(1):852–859, 2024b.

Zhenxing Ming, Julie Stephany Berrio, Mao Shan, and Stewart Worrall. Occfusion: Multi-sensor fusion framework for 3d semantic occupancy prediction. *IEEE Transactions on Intelligent Vehicles*, 2024.

Zhenxing Ming, Julie Stephany Berrio, Mao Shan, Yaoqi Huang, Hongyu Lyu, Nguyen Hoang Khoi Tran, Tzu-Yun Tseng, and Stewart Worrall. Occcylindrical: Multi-modal fusion with cylindrical representation for 3d semantic occupancy prediction. *arXiv preprint arXiv:2505.03284*, 2025.

Jingyi Pan, Zipeng Wang, and Lin Wang. Co-occ: Coupling explicit feature fusion with volume rendering regularization for multi-modal 3d semantic occupancy prediction. *IEEE Robotics and Automation Letters*, 9(6):5687–5694, 2024a.

Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Hongwei Xie, Bing Wang, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *ICRA*, pp. 12404–12411. IEEE, 2024b.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, pp. 194–210. Springer, 2020.

Zhangyang Qi, Jiaqi Wang, Xiaoyang Wu, and Hengshuang Zhao. Ocbev: Object-centric bev transformer for multi-view 3d object detection. In *3DV*, pp. 1188–1197. IEEE, 2024.

Yining Shi, Kun Jiang, Jinyu Miao, Ke Wang, Kangan Qian, Yunlong Wang, Jiusi Li, Tuopu Wen, Mengmeng Yang, Yiliang Xu, et al. Effocc: Learning efficient occupancy networks from minimal labels for autonomous driving. In *IROS*, pp. 17008–17015. IEEE, 2025.

Chaehee Song, Sanmin Kim, Hyeonjun Jeong, Juyeb Shin, Joonhee Lim, and Dongsuk Kum. Reocc: Camera-radar fusion with radar feature enrichment for 3d occupancy prediction. In *IROS*, pp. 15367–15374. IEEE, 2025.

Xiaoyu Tian, Tao Jiang, Longfei Yun, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *arXiv preprint arXiv:2304.14365*, 2023.

Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. Occgen: Generative multi-modal 3d occupancy prediction for autonomous driving. In *ECCV*, pp. 95–112. Springer, 2024a.

Jiabao Wang, Zhaojiang Liu, Qiang Meng, Liujiang Yan, Ke Wang, Jie Yang, Wei Liu, Qibin Hou, and Ming-Ming Cheng. Opus: occupancy prediction using a sparse set. *NeurIPS*, 37:119861–119885, 2024b.

Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *ICCV*, pp. 17850–17859, 2023.

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

Zhiqiang Wei, Lianqing Zheng, Jianan Liu, Tao Huang, Qing-Long Han, Wenwen Zhang, and Fengdeng Zhang. Ms-occ: Multi-stage lidar-camera fusion for 3d semantic occupancy prediction. *arXiv preprint arXiv:2504.15888*, 2025.

Philipp Wolters, Johannes Gilg, Torben Teepe, Fabian Herzog, Anouar Laouichi, Martin Hofmann, and Gerhard Rigoll. Unleashing hydra: Hybrid fusion, depth consistency and radar for unified 3d perception. In *ICRA*, pp. 7467–7474. IEEE, 2025.

Di Wu, Feng Yang, Benlian Xu, Pan Liao, Wenhui Zhao, and Dingwen Zhang. Hv-bev: Decoupling horizontal and vertical feature sampling for multi-view 3d object detection. *IEEE Transactions on Intelligent Transportation Systems*, 2025a.

Yuan Wu, Zhiqiang Yan, Zhengxue Wang, Xiang Li, Le Hui, and Jian Yang. Deep height decoupling for precise vision-based 3d occupancy prediction. In *ICRA*, pp. 12647–12654. IEEE, 2025b.

Yuan Wu, Zhiqiang Yan, Yigong Zhang, Xiang Li, and Jian Yang. See through the dark: Learning illumination-affined representations for nighttime occupancy prediction. *arXiv preprint arXiv:2505.20641*, 2025c.

Yu Yang, Jianbiao Mei, Yukai Ma, Siliang Du, Wenqing Chen, Yijie Qian, Yuxiang Feng, and Yong Liu. Driving in the occupancy world: Vision-centric 4d occupancy forecasting and planning via world models for autonomous driving. In *AAAI*, volume 39, pp. 9327–9335, 2025a.

Zhen Yang, Yanpeng Dong, Jiayu Wang, Heng Wang, Lichao Ma, Zijian Cui, Qi Liu, Haoran Pei, Kexin Zhang, and Chao Zhang. Daocc: 3d object detection assisted multi-sensor fusion for 3d occupancy prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025b.

Zichen Yu, Changyong Shu, Jiajun Deng, Kangjie Lu, Zongdai Liu, Jiangyong Yu, Dawei Yang, Hui Li, and Yan Chen. Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. *arXiv preprint arXiv:2311.12058*, 2023.

Haiming Zhang, Xu Yan, Dongfeng Bai, Jiantao Gao, Pan Wang, Bingbing Liu, Shuguang Cui, and Zhen Li. Radocc: Learning cross-modality occupancy knowledge through rendering assisted distillation. In *AAAI*, volume 38, pp. 7060–7068, 2024.

Lianqing Zheng, Jianan Liu, Runwei Guan, Long Yang, Shouyi Lu, Yuanzhe Li, Xiaokai Bai, Jie Bai, Zhixiong Ma, Hui-Liang Shen, et al. Doracamom: Joint 3d detection and occupancy prediction with multi-view 4d radars and cameras for omnidirectional perception. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. In *ECCV*, pp. 55–72. Springer, 2024.

Yuchen Zhou, Yan Luo, Xiaogang Wang, Xingjian Gu, Mingzhou Lu, and Xiangbo Shu. Da-occ: Direction-aware 2d convolution for efficient and geometry-preserving 3d occupancy prediction in autonomous driving. *arXiv preprint arXiv:2507.23599*, 2025.

Xubo Zhu, Haoyang Zhang, Fei He, Rui Wu, Yanhu Shan, Wen Yang, and Huai Yu. Dr. occ: Depth- and region-guided 3d occupancy from surround-view cameras for autonomous driving. *arXiv preprint arXiv:2603.01007*, 2026.

A APPENDIX

Table 7: Comparison of computational cost. Metrics were measured on an NVIDIA 4090 GPU.

Method	mIoU	FPS	Param. (M)	Train Mem. (GB)
FB-Occ Li et al. (2023c)	39.80	14.4	68.53	6.01
+Ours	50.89	14.8	68.53	5.97
ALOcc-2D-mini Chen et al. (2025a)	50.00	30.5	36.23	2.15
+Ours	53.10	20.8	37.21	2.58
ALOcc-2D Chen et al. (2025a)	53.50	8.1	63.70	3.38
+Ours	55.47	7.5	63.96	4.58

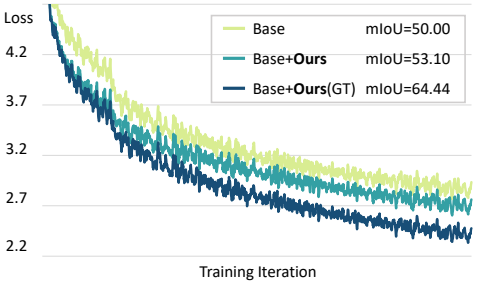


Figure 9: Upper-Bound analysis.

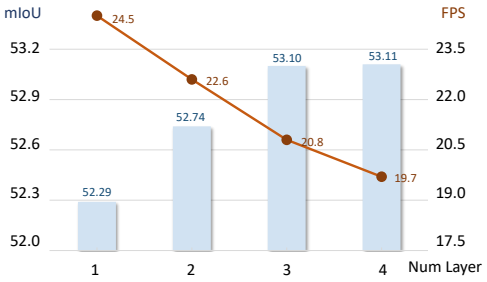


Figure 10: Ablation on HGR layers.

A.1 COMPUTATIONAL COST

Tab. 7 presents the computational cost analysis of our method on different models. Overall, our method consistently improves prediction accuracy with only limited additional computational overhead. Specifically, on FB-Occ Li et al. (2023c), our method achieves a 11.09 improvement in mIoU. Meanwhile, by filtering out BEV grids with the height validity mask, our method reduces unnecessary computation and maintains comparable efficiency. For the ALOcc series Chen et al. (2025a), our method also brings consistent performance gains. On ALOcc-2D-mini, the mIoU increases from 50.00 to 53.10, with only a moderate increase in parameters and training memory. On ALOcc-2D, our method further improves the mIoU from 53.50 to 55.47, while maintaining comparable inference speed. These results demonstrate that our method can effectively enhance prediction accuracy across different baselines without introducing prohibitive computational costs.

A.2 UPPER-BOUND ANALYSIS

To evaluate the potential of height guidance, we compare the baseline ALOcc-2D-mini, our LiDAR-based height-guided variant, and an upper-bound model that leverages ground-truth height maps. Notably, the upper-bound setting uses ground-truth height not only during training but also at inference time. As illustrated in Fig. 9, incorporating LiDAR height guidance improves the baseline by 3.10 mIoU, demonstrating the effectiveness of our reparameterization strategy. Furthermore, ground-truth height guidance accelerates convergence and yields substantially better performance, improving mIoU by 14.44. These results suggest that a more accurate height prior provides stronger geometric constraints and highlight the upper-bound potential of our height-guided design.

A.3 ANALYSIS OF HEIGHT-GUIDED REPARAMETERIZATION

Effect of Height-Validity Mask. To isolate the contribution of the height-validity mask, we apply the mask to the uniform-sampling baseline while keeping all other settings unchanged. As shown in Tab. 8, introducing the validity mask improves the mIoU from 50.57 to 51.37, indicating that suppressing feature aggregation in pillars without valid LiDAR observations effectively reduces unreliable projections. Under the same mask setting, replacing uniform sampling with height-guided sampling further improves the mIoU from 51.37 to 52.81. These controlled comparisons show that

Table 8: Ablation on the height-validity mask.

Strategy	Validity Mask	mIoU
Uniform	×	50.57
Uniform	✓	51.37
Height-guided (Ours)	✓	**52.81**

Table 9: Comparison of occupied-voxel hit ratios under different vertical sampling strategies.

Strategy	Hit Ratio (%)	mIoU
Full-range uniform	31.26	50.57
Height-guided (Ours)	**62.98**	**52.81**

Table 10: Evaluation of LiDAR-derived height. F denotes number of LiDAR frames

Metric	0–10 m		10–30 m		>30 m		Overall	
	1 F	10 F	1 F	10 F	1 F	10 F	1 F	10 F
Coverage (%)	63.29	87.25	24.71	54.11	13.69	28.49	26.37	50.53
Bias (m)	-0.20	-0.11	-0.32	-0.04	-0.01	+0.38	-0.23	+0.01
Acc@0.4m (%)	84.17	85.15	62.42	72.08	47.87	50.99	66.94	71.36

the two components are complementary: the validity mask determines whether a pillar is refined, whereas height-guided sampling determines where its reference points are placed vertically.

Reference Point Sampling Efficiency. We further examine whether height-guided reparameterization improves the utilization of the limited reference-point budget. For a controlled comparison, we use the same LiDAR-valid pillars and 8 reference points per pillar for both full-range uniform and height-guided sampling. We define the occupied-voxel hit ratio as the percentage of sampled 3D reference points that fall inside occupied ground-truth voxels. As shown in Tab. 9, height-guided sampling increases the hit ratio by 31.72 percentage points, while improving the mIoU from 50.57 to 52.81. This increase indicates that HGR redistributes reference points from geometrically irrelevant regions toward scene structures, thereby utilizing the fixed sampling budget more effectively.

Number of HGR Layers. We conduct this ablation by varying the number of HGR layers in HiPR-mini. As shown in Fig. 10, mIoU consistently improves as the number of layers increases, but the gain becomes marginal beyond three layers, while FPS continues to decrease. Therefore, we adopt three HGR layers to achieve a better accuracy-efficiency trade-off.

A.4 Analysis of LiDAR-Derived Height Priors

Distance-Wise Height-Prior Quality. We first analyze the quality of the LiDAR-derived height prior at different ego-centric distances. We evaluate three complementary metrics: *coverage*, *height bias*, and *Acc@0.4m*. Coverage measures the proportion of pillars with valid LiDAR height observations. Height bias is defined as $\boldsymbol{H}_{\mathrm{LiDAR}} - \boldsymbol{H}_{\mathrm{GT}}$. Acc@0.4m measures the proportion of valid height estimates whose absolute error is within 0.4 m. As shown in Tab. 10, the quality of the single-frame height prior degrades noticeably with distance. Its coverage decreases from 63.29% within 0–10 m to 24.71% at 10-30 m and only 13.69 beyond 30 m. Similarly, Acc@0.4m drops from 84.17 to 62.42% and 47.87%, respectively. These results indicate that single-frame LiDAR observations become increasingly sparse and less reliable in distant regions. To investigate whether denser observations can provide a stronger geometric prior, we additionally aggregate ten LiDAR sweeps. Multi-frame aggregation substantially improves height coverage across all distance ranges. Overall coverage increases from 26.37% to 50.53%, while Acc@0.4m improves from 66.94% to 71.36%. Meanwhile, the overall height bias changes from -0.23 m to +0.01 m, indicating that temporal aggregation largely alleviates the systematic underestimation observed in the single-frame setting.

Effect of Multi-Frame Aggregation. We further investigate whether the improved height prior translates into better occupancy prediction and where the resulting gains occur. HiPR uses ten aggregated LiDAR frames to construct the height prior, consistent with the LiDAR setting of DAOcc Yang et al. (2025b). For a controlled comparison, we construct a single-frame variant, denoted as HiPR*, by using only the current LiDAR frame while keeping the remaining model unchanged. As shown in Tab. 11, aggregating multiple LiDAR frames improves the overall mIoU from 54.70 to 55.47, confirming that a denser and more accurate height prior further benefits height-guided reparameterization. Under the same 10-frame LiDAR setting, HiPR also outperforms DAOcc by 1.14 mIoU. The distance-wise results further show that the benefit of multi-frame aggregation is concentrated mainly in the middle- and long-range regions.

This trend is consistent with the height-prior analysis in Tab. 10: single-frame LiDAR already provides relatively dense and accurate height observations within 10 m, whereas its coverage decreases sharply at larger distances. Therefore, multi-frame aggregation provides greater benefits in regions where the single-frame height prior is sparse, establishing a clear connection between height-prior quality and occupancy performance.

Table 11: Distance-stratified performance.

Distance	DAOcc	HiPR*	HiPR
0–10 m	61.59	**63.59**	63.16
10–30 m	57.47	57.27	**58.08**
>30 m	43.46	44.45	**45.44**
Overall	54.33	54.70	**55.47**

A.5 Height-Stratified Evaluation

To further investigate whether the benefit of height-guided reparameterization is limited to near-ground structures, we conduct a height-stratified evaluation. Specifically, we divide the vertical occupancy space into a lower region, $[-1.0, 2.2)$ m, and an upper region, $[2.2, 5.4]$ m. Since semantic labels in the upper region are substantially sparser and more imbalanced, we report binary occupancy IoU for the two height ranges to more directly evaluate geometric reconstruction quality. As shown in Tab. 12, fixed-range sampling strategies mainly improve occupancy reconstruction in the lower region. For example, the road-biased strategy improves the lower-region IoU from 85.07 to 86.51, whereas its improvement in the upper region is only +0.51 IoU. This behavior is expected because these fixed strategies allocate more reference points to lower heights and therefore primarily benefit road-level structures. In contrast, height-guided sampling consistently improves both vertical regions. Compared with full-range uniform sampling, it improves the overall binary occupancy IoU by 2.90 points, including gains of 2.45 IoU in the lower region and 5.67 IoU in the upper region. In particular, the substantially larger improvement in the upper region demonstrates that the benefit of HGR does not simply arise from concentrating reference points near the ground. Instead, by adapting the sampling range independently for each BEV pillar, HGR better accommodates scene structures with diverse vertical extents and allocates the limited sampling budget to geometrically relevant regions.

Table 12: Height-stratified evaluation.

Strategy	mIoU	IoU		
		Overall	Lower	Upper
Full-range	50.57	80.26	85.07	66.86
Upper bound (3 m)	50.63	80.43	86.07	67.09
Upper bound (2 m)	50.81	80.62	86.40	66.96
Road-biased	51.03	80.84	86.51	67.37
Height-guided (Ours)	**52.81**	**83.16**	**87.52**	**72.53**

A.6 Limitation

Although HiPR achieves strong performance, it relies on LiDAR-derived height priors, which can become sparse and less reliable in distant regions. In addition, modeling scene geometry with a single height representation may be insufficient for complex multi-layer structures. We leave the exploration of richer geometric representations and more robust height prior as future work.

A.7 Additional Visual Results

We provide additional visualizations of DAOcc Yang et al. (2025b) and our HiPR in Fig. 11.

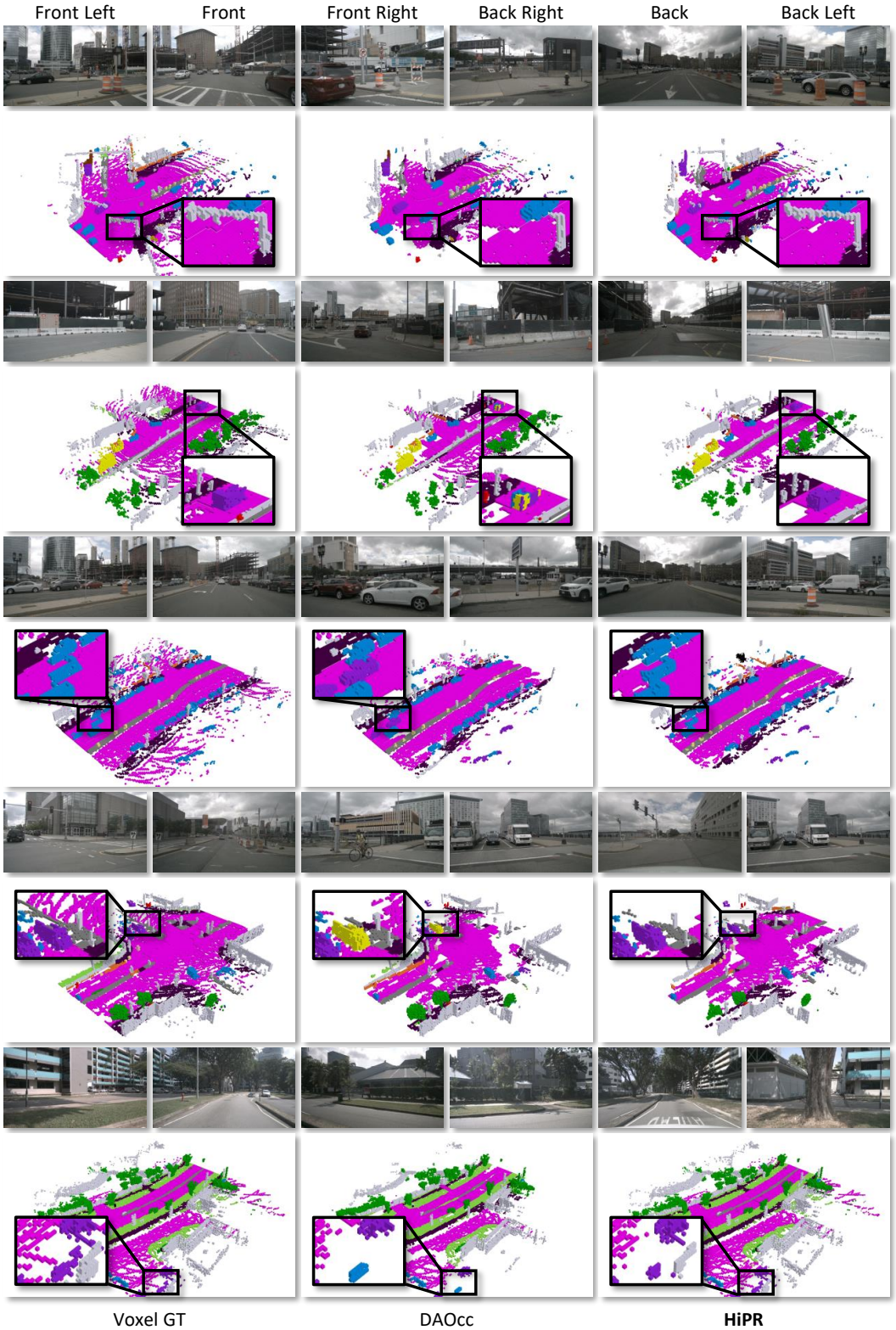


Figure 11: Qualitative comparison between DAOcc and our HiPR on the Occ3D validation set.