PROGRESSIVE GAUSSIAN TRANSFORMER WITH ANISOTROPY-AWARE SAMPLING FOR OPEN VOCABULARY OCCUPANCY PREDICTION

SUPPLEMENTARY MATERIAL

This supplementary material offers more detailed descriptions to ensure reproducibility, along with extensive evaluations and diverse qualitative results, which collectively highlight the effectiveness, robustness, and efficiency of our proposed method, **PG-Occ**.

▷ **appendix A**: Video demonstrations comparing the open-vocabulary occupancy inference results of PG-Occ, the previous state-of-the-art method, and the Ground Truth on the Occ3D-nuScenes validation and test sets.

▷ **appendix B**: Additional experimental results, including extended quantitative comparisons and additional ablation studies.

▷ **appendix C**: More qualitative visualization results of our PG-Occ.

▷ **appendix D**: Additional details on implementation.

A VIDEO DEMONSTRATION

We provide video demonstrations of our open-vocabulary occupancy inference results on the Occ3D-nuScenes (Tian et al., 2024) validation and test sets. For better visualization quality, please refer to our project page (https://yanchi-3dv.github.io/PG-Occ/).

B ADDITIONAL EXPERIMENT RESULTS

B.1 ADDITIONAL ABLATION STUDY

Impact of the Number of Base Gaussian Queries. We evaluate the effect of varying the number of base Gaussian queries while keeping the number of extended Gaussian queries fixed at 1000. As shown in Table 6, increasing the number of base layer queries from 1000 to 4000 consistently improves the mIoU from 13.17 to 15.15, indicating enhanced perception accuracy. However, further increasing the queries to 8000 results in a slight drop in mIoU. This decline is due to the excessive number of Gaussian queries overwhelming the self-attention mechanism, thereby weakening the model's ability to capture the critical spatial interactions between queries, similar to the observations reported in Jiang et al. (2024).

Table 6: Ablation study on the number of initial queries in the base layer. The best and the second-best performances of each metric are highlighted with **bold** and underlined in the table.

Queries	mIoU	RayIoU	mAP (v)
1000	13.17	10.33	17.25
2000	14.54	12.84	18.98
4000	**15.15**	**13.92**	**21.20**
8000	<u>14.99</u>	<u>13.52</u>	<u>21.07</u>

Ablation Study of Feed-forward Densification Module Threshold. We investigate the impact of the threshold in the feed-forward densification module, which determines the minimum distance from points outside to the center of an occupancy cell, on both computational cost and prediction accuracy. Gaussian points are selected according to different thresholds, and the subsequent farthest point sampling (FPS) time, mIoU are measured. The results are summarized in Table 7. As shown in the table, decreasing the threshold selects more points, increasing the subsequent FPS time from 30 ms to 50 ms, which results in higher computational overhead. Meanwhile, a threshold of 0.2 achieves the highest mIoU of 15.15, slightly outperforming thresholds 0.0 and 0.4 (15.11 and 15.13, respectively). This indicates that the chosen threshold

Table 7: Impact of the feed-forward densification on farthest point sampling (FPS) time and occupancy prediction.

Threshold	Total FPS Time (ms)	mIoU
0.0	50	15.11
0.2	34	15.15
0.4	30	15.13

Table 9: Ablation on the loss function $\mathcal{L}$. The best performances of each metric are highlighted with **bold** in the table.

$\mathcal{L}_{L1}$	$\mathcal{L}_{SILog}$	$\mathcal{L}_{temp}$	$\mathcal{L}_{mse}$	$\mathcal{L}_{cos}$	mIoU	RayIoU	mAP (v)
✓	✓	✓	✓		13.51	12.30	18.13
✓	✓	✓		✓	15.12	13.81	20.52
✓	✓		✓	✓	14.47	13.01	19.14
✓		✓	✓	✓	15.10	13.89	20.41
	✓	✓	✓	✓	14.69	13.23	19.95
✓	✓	✓	✓	✓	**15.15**	**13.92**	**21.20**

of 0.2 effectively balances point selection for Gaussian densification, maintaining high prediction accuracy while controlling computational cost.

Ablation Study of the Number of Sampling Offsets. We evaluate the impact of the number of sampling offsets per Gaussian on occupancy prediction. As shown in Table 8, increasing the number of offsets from 8 to 32 consistently improves performance, with mIoU rising from 15.05 to 15.46. This demonstrates that a larger number of sampling points allows the network to capture more fine-grained scene details, enhancing occupancy prediction accuracy. However, the increase in offsets also leads to longer training times, growing from 8 hours for 8 offsets to 11.2 hours for 32 offsets on an 8×A800 GPU setup. These results highlight a trade-off between accuracy and computational cost, indicating that 16 offsets provide a balanced choice, achieving strong performance with moderate training time.

Ablation Study of Loss. We systematically evaluate the impact of different loss function combinations, including $\mathcal{L}_{L1}$, $\mathcal{L}_{SILog}$, $\mathcal{L}_{temp}$, $\mathcal{L}_{mse}$, and $\mathcal{L}_{cos}$. The results are summarized in Table 9. We observe that using all five loss functions consistently yields the best performance, achieving a mIoU of 15.15%, a RayIoU of 13.92%, and an mAP of 21.20%. Omitting any individual loss results in a slight drop across these metrics, indicating that each component contributes to both geometric accuracy and feature alignment. These findings confirm that the combination of complementary losses enables PG-Occ to more effectively capture fine-grained scene details and improve overall 3D occupancy prediction.

Table 8: Impact of the number of sampling offsets per Gaussian on performance and training time.

Sampling Points	mIoU	Training Time
8	15.05	8 hours
16	15.15	9 hours
32	15.46	11.2 hours

Ablation Study of Photometric Supervision. Our occupancy prediction aims to recover both geometric and semantic components. Due to the challenges of large-scale outdoor scenes and limited view supervision, photometric information often fails to provide effective geometric supervision. Moreover, color features do not reliably correspond to semantic categories in such scenes, so we exclude them during training. We performed an ablation study by adding a photometric prediction head to regress 3D color values for supervision. The quantitative results are summarized in Table 10.

Table 10: Ablation study on photometric supervision. The best performances of each metric are highlighted with **bold**.

Color Supervision	mIoU	RayIoU
w/o color supervision	**15.15**	**13.92**
w/ color supervision	14.96	13.89

Ablation Study of Pose Noise. We investigate the robustness of PG-Occ to pose noise during temporal-spatial feature fusion by adding Gaussian perturbations with different standard deviations to the historical ego poses during inference, as summarized in Table 11. The results show that a slight amount of pose noise leads to a minor improvement in mIoU, while larger noise levels cause a small decrease followed by stabilization, demonstrating that PG-Occ is robust to pose errors. We attribute this robustness to two factors: first, the nuScenes dataset poses are not perfectly accurate and contain small misalignments, which naturally provide tolerance to minor noise; second, significant pose inaccuracies result in feature sampling failures on the camera plane, preventing unreliable features from degrading system performance and thereby maintaining relatively high perception accuracy.

Table 11: Impact of Gaussian pose noise on occupancy prediction.

Standard Deviation	0	0.01	0.1	0.5
mIoU	15.15	15.19	15.12	15.12

B.2 Layer-wise Time Consumption

The Table 12 reports the inference time of each Gaussian transformer layer in milliseconds. The base layer, using the fewest Gaussians, achieves the fastest speed at 27.4 ms. As more Gaussians are added in the First and Second Progressive Layers, the inference time correspondingly increases to 58.3 ms and 60.6 ms, reflecting the higher computational cost of processing denser representations.

Table 12: Inference time of different layers.

Component	Time (ms)
Base Layer	27.4
First Progressive Layer	58.3
Second Progressive Layer	60.6

B.3 Robustness Evaluation with Different Pretrained Depth Models

In the main paper, we adopt Metric3D V2 (Hu et al., 2024) as our default depth estimator. To further examine the robustness of PG-Occ under different sources of depth supervision, we additionally train and evaluate our model using UniDepth V2 (Piccinelli et al., 2025) pseudo-depth. The Table 13 reports both the depth estimation errors and open-vocabulary semantic occupancy performance on the nuScenes validation set.

Similar to the setting with Metric3D V2 pseudo-depth, using UniDepth V2 also enables PG-Occ to recover depth estimates that outperform the pseudo-labels. For example, while UniDepth V2 provides an Abs Rel of 0.158, PG-Occ improves it to 0.137, showing that PG-Occ can consistently refine imperfect depth supervision regardless of the depth model used.

In terms of the open-vocabulary occupancy prediction task, the performance remains stable across depth models. When switching from Metric3D V2 to UniDepth V2 pseudo-depth, the mIoU only changes slightly from 15.15 to 15.08, confirming that PG-Occ is largely insensitive to the specific choice of depth estimator.

These findings highlight two key properties of PG-Occ. **(i)** PG-Occ does not depend on any specific depth architecture; it only requires coarse geometric cues for initialization and supervision, making it naturally compatible with diverse metric depth models. **(ii)** PG-Occ consistently refines these cues and maintains robust occupancy performance even as the upstream depth model changes.

Table 13: Robustness evaluation across different depth models. The best performances of each metric are highlighted with **bold**.

Method	Abs Rel	Sq Rel	RMSE	RMSE log	mIoU
Metric3D V2 (Hu et al., 2024)	0.170	4.016	6.453	0.291	—
UniDepth V2 (Piccinelli et al., 2025)	0.158	2.232	5.491	0.259	—
PG-Occ (Metric3D V2)	0.139	1.159	5.466	0.269	**15.15**
PG-Occ (UniDepth V2)	**0.131**	**1.129**	**5.049**	**0.248**	15.08

B.4 Effectiveness in Multimodal Settings with LiDAR and Cameras

While our approach targets image-based occupancy prediction, which is an important direction in the field, it is also effective in multimodal settings combining LiDAR and cameras. To validate this capability, we performed additional experiments on the nuScenes dataset, substituting pseudo-depth inputs with ground-truth sparse LiDAR point clouds. The results demonstrate that PG-Occ effectively leverages multimodal inputs, maintaining robust semantic occupancy prediction even for challenging scenes.

As summarized in 14, even a direct, naive replacement—without any tailored optimization—significantly boosts 3D occupancy metrics: mIoU rises from 15.15 (Depth) to 18.98 (Sparse LiDAR), RayIoU from 13.92 to 15.22, and mAP(v) from 21.20 to 29.53. This validates the pipeline's capability to generalize beyond pseudo-depth and effectively absorb multimodal spatial cues. Notably, our approach yields robust improvements with

Table 14: 3D occupancy performance with pseudo-depth vs. LiDAR inputs.

Method	mIoU	RayIoU	mAP (v)
PG-Occ (Depth)	15.15	13.92	21.20
PG-Occ (LiDAR)	18.98	15.22	29.53

basic point cloud substitution, indicating that further gains remain achievable via more sophisticated fusion techniques.

C Additional Visualization Results

C.1 PG-Occ Capabilities

We present more qualitative results in Fig. 9 demonstrating that, using single-frame multi-view inputs and feed-forward inference, PG-Occ accurately estimates scene depth and generates open-vocabulary feature renderings capturing semantics beyond fixed categories. It supports zero-shot semantic 3D occupancy prediction and enables flexible open-vocabulary text queries for object retrieval and localization.

C.2 BEV Visualization

In this subsection, we present BEV (bird's-eye view) occupancy visualizations produced by PG-Occ. This perspective provides a comprehensive overview of the scene layout, allowing clear observation of spatial relationships among various objects. As illustrated in Fig. 10, our method accurately reconstructs both large and small scene elements, including vehicles, pedestrians, and barriers, while maintaining sharp and consistent occupancy boundaries. We select a variety of diverse scenes to demonstrate the robustness and generalization capability of our approach, highlighting its ability to handle complex environments effectively.

C.3 Ego-centric perspective occupancy visualization with previous SOTA method

In this subsection, we visualize occupancy from the vehicle's perspective and compare our results with the previous state-of-the-art method, GaussTR (Jiang et al., 2024). This comparison aims to highlight the strengths and improvements of our approach in estimating occupancy within the scene. As illustrated in Fig. 11, our method demonstrates superior detection results for small objects compared to GaussTR (Jiang et al., 2024), particularly for car, bicycle, bus, truck, and barrier. Interestingly, our approach is capable of detecting elements that are not well annotated in the Ground Truth, such as the pedestrians and bicycles shown in the second visualization of the figure.

C.4 Third perspective occupancy visualization

In this subsection, we present the visualization of our method from two different third-person perspectives. As illustrated in Fig. 12, we compare the zero-shot semantic occupancy estimations generated by our approach with the Ground Truth. The visualizations illustrate the effectiveness of our method in accurately capturing the spatial occupancy of various objects within the scene. The results underscore our model's ability to perform zero-shot semantic occupancy estimation, enabling it to infer the occupancy of objects it has not encountered during training. However, it is important to note that due to occlusion issues present in the scene, our self-supervised method may face challenges in making accurate predictions in areas lacking visual observations. Nevertheless, it can still yield reasonable inferences to a certain extent.

D Additional Implementation Details

D.1 Voxelization

As described in section 3.4, after obtaining the progressive 3D Gaussian representation that models the scene based on the current camera input. We convert the 3D feature Gaussian blobs output to a semantic occupancy field. To begin with, we take n arbitrary text prompts c_{text} and encode them using the CLIP text encoder to obtain their corresponding feature embeddings f_{text}. And then compute the similarity between these text embeddings and the text-aligned features f_i of Gaussian i, subsequently. The text probability for each 3D feature Gaussian blob G under c_{text} can be calculated

Table 15: Text prompts used for zero-shot semantic occupancy estimation on the Occ3D-nuScenes dataset (Tian et al., 2024). '-' indicates that no prompts were made for this class.

nuScenes Class	**Prompts**
others	-
barrier	barrier
bicycle	bicycle
bus	bus
car	car
construction vehicle	construction vehicle
motorcycle	motorcycle
pedestrian	person
traffic cone	cone
trailer	trailer
truck	truck
driv. surface	road
other flat	-
sidewalk	sidewalk
terrain	terrain, grass
manmade	building, wall, fence, pole, sign
vegetation	vegetation
empty	sky

as follow equation:

$$p_i = \sigma(f_i \cdot f_{text}^T) \tag{14}$$

where p_i represents the text probability of the i-th Gaussian blob under c_{text}, and σ denotes the softmax operation.

After that, we define a voxel grid within the region of interest (ROI) occupancy range and then calculate the influence of each Gaussian on each voxel, accumulating the results. This process is affected by the anisotropy parameter s, r of the Gaussians, their opacity o, and assigned text probability p. The formulation for this voxelization can be written as:

$$\mathcal{V}_o = \sum_{i=1}^{N} G_i(x; \mu_i, s_i, r_i, o_i) = \sum_{i=1}^{N} \exp(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i))o_i, \tag{15}$$

$$\mathcal{V}_p = \sum_{i=1}^{N} G_i(x; \mu_i, s_i, r_i, o_i) = \sum_{i=1}^{N} \exp(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i))p_i, \tag{16}$$

where $\mathcal{V}_o$, $\mathcal{V}_p$ denote the final occupancy probability and semantic 3D occupancy field, x denotes the voxel grid position of occupancy, Σ is the Gaussian covariance matrix of each Gaussian, revived from its scale s_i and rotation quaternion r_i.

In the evaluation of the nuScenes retrieval dataset experiment in Section 4.2, since the ground truth consists of text annotations for sparse LiDAR points $\mathcal{P}$, we treat each LiDAR point p as the center x of a voxel. This allows us to obtain the corresponding final occupancy probability and text feature, as shown in the following formula.

$$\mathcal{P}_o = \sum_{i=1}^{N} G_i(p; \mu_i, s_i, r_i, o_i) = \sum_{i=1}^{N} \exp(-\frac{1}{2}(p-\mu_i)^T \Sigma_i^{-1}(p-\mu_i))o_i, \tag{17}$$

$$\mathcal{P}_f = \sum_{i=1}^{N} G_i(p; \mu_i, s_i, r_i, o_i) = \sum_{i=1}^{N} \exp(-\frac{1}{2}(p-\mu_i)^T \Sigma_i^{-1}(p-\mu_i))f_i, \tag{18}$$

where $\mathcal{P}_o$, $\mathcal{P}_f$ denote the final occupancy probability and the corresponding text feature of the LiDAR point cloud $\mathcal{P}$.

D.2 TEXT PROMPT

Due to the imprecise semantics in the Occ3D-nuScenes (Tian et al., 2024) dataset, we made some minor adjustments to the prompts used in PG-Occ, as shown in Table 15. Specifically, we do not detect the categories 'others' or 'other flat,' as they can lead to ambiguities. Note that further fine-tuning of these ambiguous prompts could enhance performance.

For the retrieval task in Section 4.2, we directly use the prompt provided by the dataset.

D.3 ADDITIONAL MODEL AND TRAINING DETAILS

D.3.1 SUPERVISION STRATEGY

Metric3D V2 (Hu et al., 2024) and MaskCLIP (Zhou et al., 2022) are utilized for depth and feature supervision. The loss weight parameters are set as follows: $\lambda_{SILog} = 0.15$, $\lambda_{temp} = 10$, and $\lambda_{mse} = 10$. The learning rate is initialized at 2e-4 with a weight decay of 0.01, using the AdamW optimizer.

D.3.2 MODEL ARCHITECTURE AND TRAINING SETUP

We adopt ResNet-50 (He et al., 2016) as the image feature backbone, utilizing the previous seven frames to capture spatio-temporal information. PG-Occ is initialized with 4,000 Gaussian queries in the base layer and progressively adds 1,000 queries per layer, resulting in one base and two progressive layers with an embedding dimension of 256. All training experiments are conducted on 8 A800 GPUs for 8 epochs, while inference is performed on a single A800 GPU. To improve computational efficiency, we use a resolution of 180×320 for depth and feature rasterization, as well as for Gaussian point initialization.

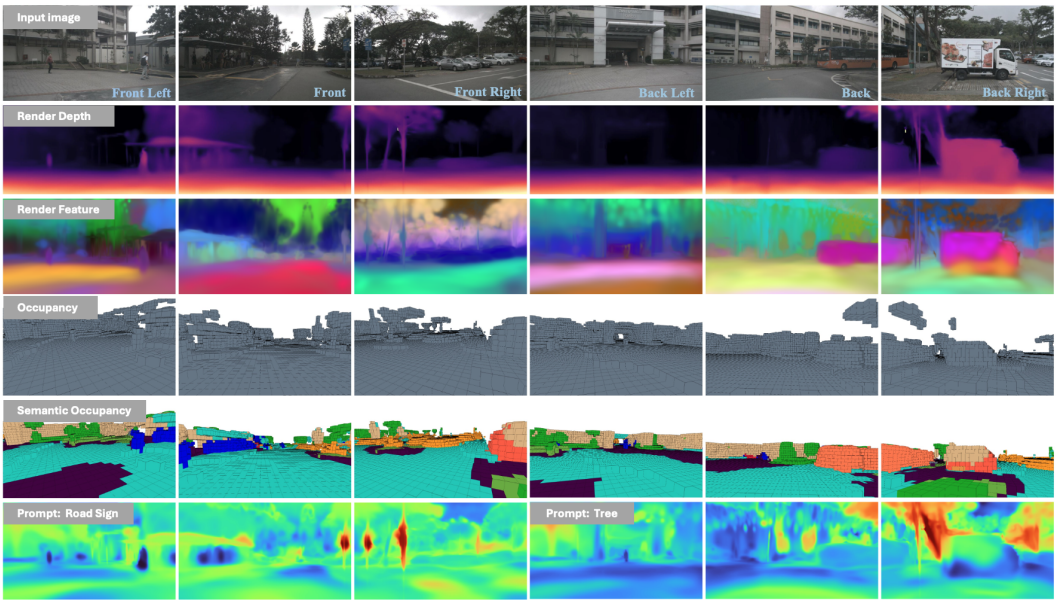

Figure 9: PG-Occ capabilities. Given only single-frame multi-view inputs and using only feed-forward passes, PG-Occ can: (1) estimate depth (row 2); (2) render open-vocabulary model features (row 3); (3) predict 3D occupancy in a zero-shot manner (rows 4); (4) predict semantic 3D occupancy in a zero-shot manner (rows 5); (5) support open-vocabulary text queries (rows 6).

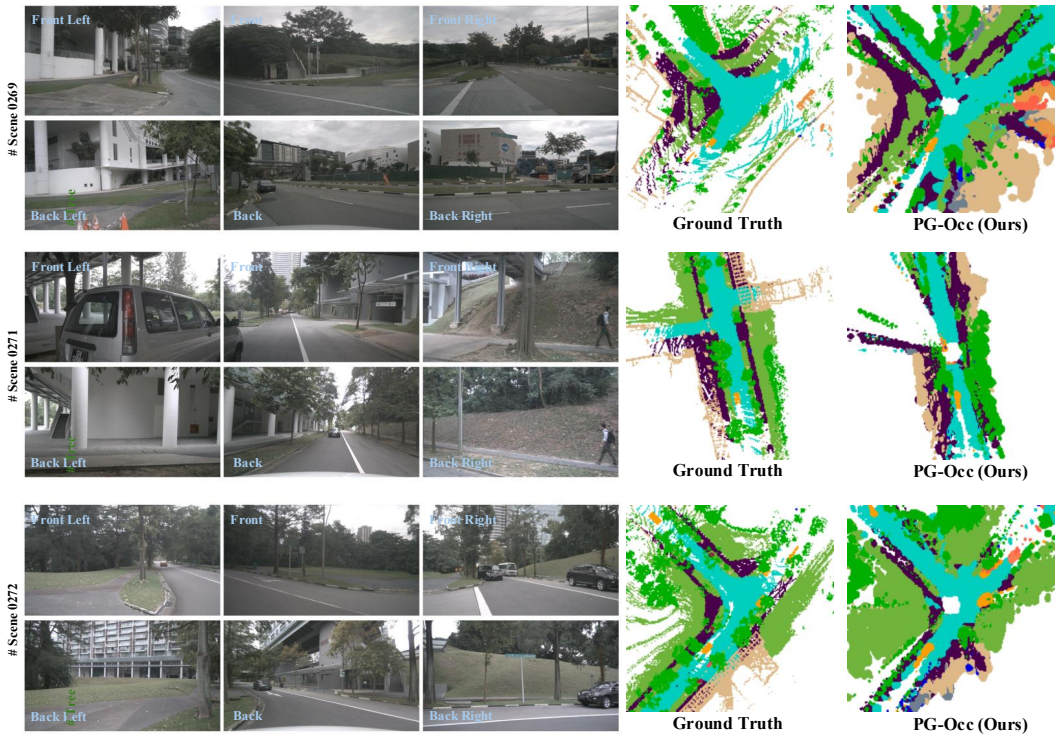

Figure 10: BEV visualization of open-vocabulary scene perception by PG-Occ. The figure illustrates predicted occupancy and semantic structures from a bird's-eye perspective, emphasizing the model's ability to capture spatial relationships and overall scene layout.

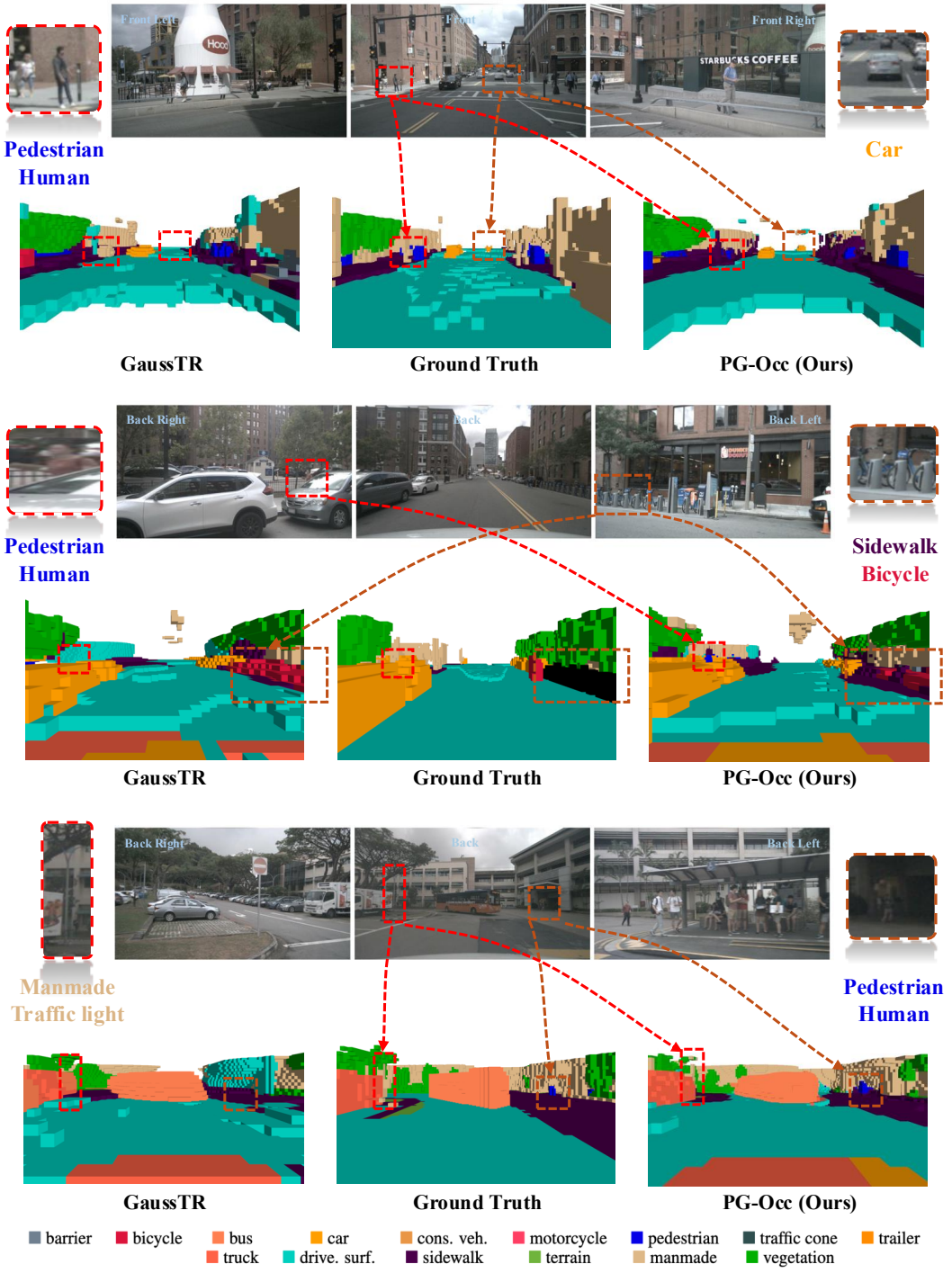

Figure 11: Qualitative comparisons of zero-shot semantic occupancy estimation from an ego-centric multi-camera perspective. Each row shows input images from multiple viewpoints (top), corresponding occupancy predictions by GaussTR (left bottom), the ground truth occupancy (middle bottom), and our PG-Occ method (right bottom). Dashed boxes and lines highlight specific objects—such as pedestrians, cars, bicycles, and traffic lights—that have been successfully detected and reconstructed. Our approach demonstrates superior detection and reconstruction of small or distant objects, better preserves spatial relationships, and provides more accurate object shapes compared with GaussTR. Colors indicate semantic categories as defined in the legend. For best inspection of fine details, we recommend viewing the color version and zooming in.

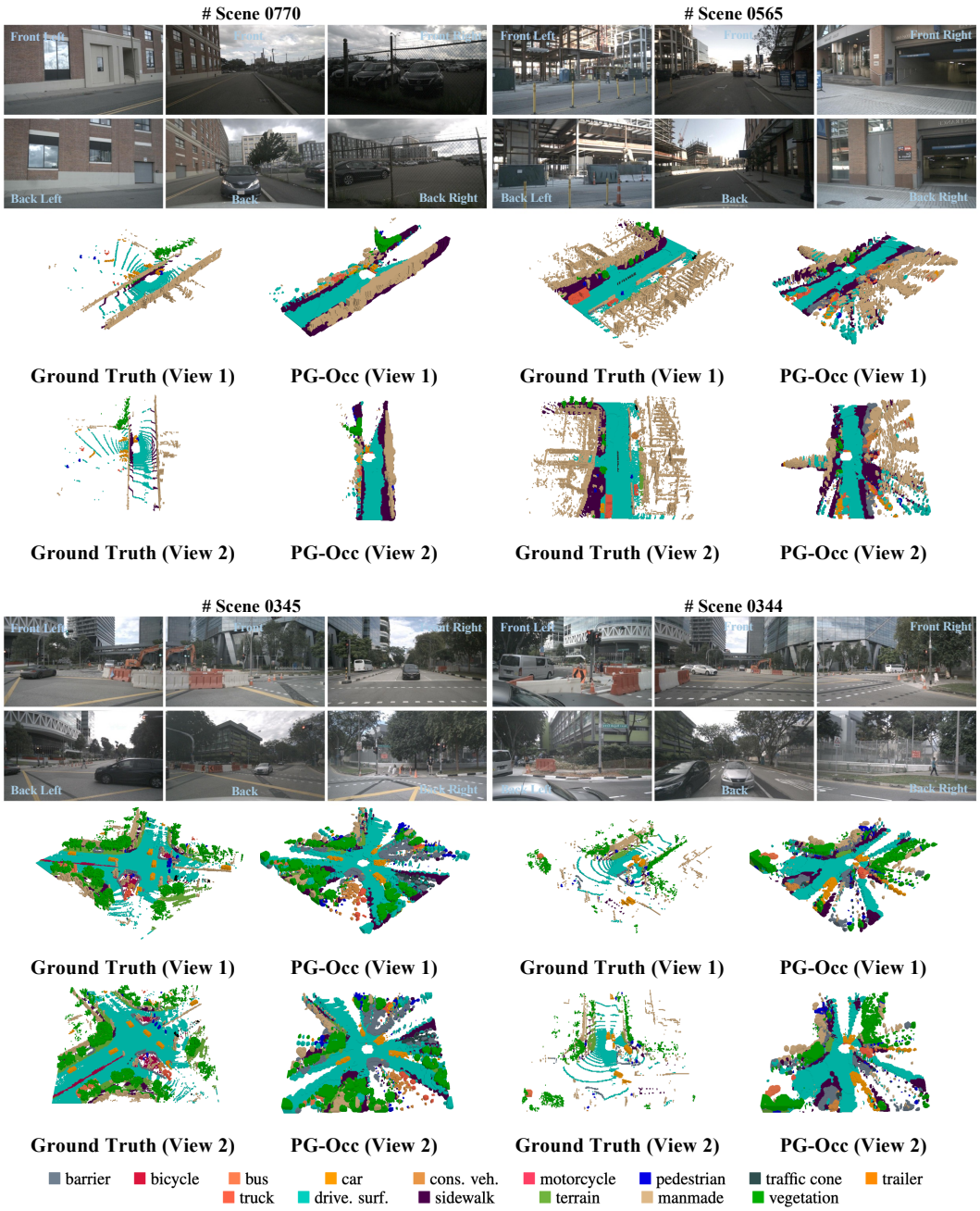

Figure 12: Qualitative zero-shot semantic occupancy results on the third perspective for two views. For each view (View 1 and View 2), we show the predictions of our method (PG-Occ) alongside the Ground Truth. The results demonstrate that PG-Occ accurately captures semantic occupancy patterns across different perspectives.

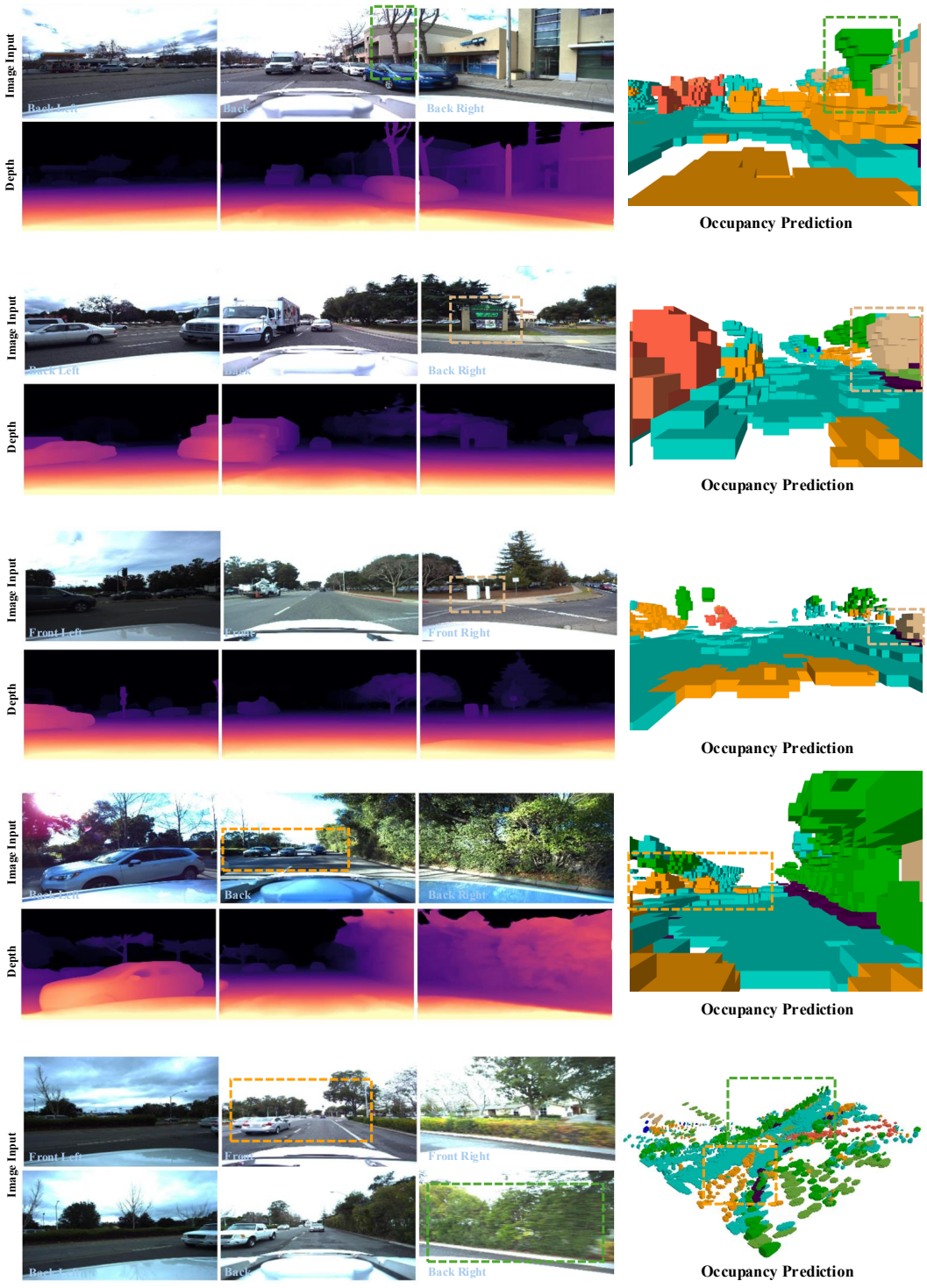

Figure 13: Zero-shot generalization on the Lyft Level-5 dataset (Christy et al., 2019). Our model is not retrained or fine-tuned on the Lyft Level-5 dataset but used directly after training on nuScenes. This scenario involves a substantial domain shift, including differences in image resolution, camera intrinsics, viewpoints, and overall scene distribution. Despite these challenges, our method maintains strong zero-shot generalization, accurately predicting occupancy and successfully recovering small or rarely seen objects in completely unseen scenes.