

QueryOcc: Query-based Self-Supervision for 3D Semantic Occupancy

Supplementary Material

Overview of Supplementary Material

This document provides additional experimental details, analysis, and qualitative results that complement the main paper. It includes:

- Dataset and class details used for supervision and evaluation (Sec. 6).
- Implementation and training specifics for the proposed method and rendering supervision scheme (Sec. 7, Sec. 8).
- Complete per-class metrics for RayIoU and additional experiments (Sec. 9).
- Qualitative examples, including long-range predictions and behavior in contracted regions (Sec. 10).
- Discussion of the discrepancy between continuous predictions and voxelized ground truth (Sec. 11).

6. Dataset details

We follow prior work [4, 19] and focus on 15 semantic classes from the Panoptic nuScenes dataset [12]. These classes are grouped into *static* and *dynamic* categories according to their motion characteristics, as summarized in Tab. 7. As Tab. 9 shows, the dataset is a highly class-imbalanced and that the class distributions between lidar labels and pseudo-labels from images varies. For instance, the lidar point clouds on average consist of 38% drivable surface, whereas the image-plane pseudo labels are 22% in the same category.

	Class	VFM Prompt Vocabulary
Static classes	■ barrier	barricade, barrier
	■ traffic cone	traffic-cone
	■ sidewalk	sidewalk, walkway
	■ drive. surf.	highway, street
	■ terrain	grass, sand, gravel, terrain
	■ vegetation	bush, plants, tree
	■ manmade	building, wall, fence, pole, sign, light, bridge, billboard
Dynamic classes	■ bicycle	bicycle
	■ motorcycle	motorcycle, scooter
	■ cons. veh.	excavator, crane
	■ pedestrian	person, pedestrian
	■ trailer	trailer
	■ truck	lorry, truck, tractor
	■ bus	bus
	■ car	car, vehicle, sedan, van, jeep

Table 7. Semantic classes used for training and evaluation, grouped into static and dynamic categories and their corresponding prompt vocabularies used for pseudo-label generation.

7. Experiment Details

Image encoder: For the image encoder, we adopt ConvNeXt-Base [22] pretrained with DINOv3 [30]. As Tab. 8 shows, QueryOcc achieves strong performance across a broad range of backbones, including lightweight ResNets, pretrained on ImageNet [28], to larger ConvNeXt variants, indicating that the framework is not tied to a specific encoder architecture. ConvNeXt-Base offers the highest semantic and geometric accuracy while still maintaining real-time inference. The results also highlight a predictable trade-off: larger encoders improve semantic occupancy but reduce throughput.

Lift-contract-splat: The point encoder F_θ uses Fourier features with 16 log-linear frequency bands in the range [1, 10], followed by a lightweight MLP that mixes the spatial and temporal inputs. For lift-contract-splat, we set $\alpha = 0.3$, $d_{\text{near}} = 40$ m, and $d_{\text{far}} = 100$ m, and include an 'infinity bin' at 180 m to cover the maximum sensor range. We supervise the categorical depth predictions, with weight λ_{depth} in the total loss, using cross-entropy. The VFM depth predictions and projected lidar points are used for the pseudo point cloud and real lidar point cloud trainings respectively. The BEV grid maintains full resolution within ± 40 m in x and y , and contracts points outside this region using $\beta = 0.8$. Effective cell side length within the full resolution region is 0.16m.

Core model: The network uses a hidden dimension of 160 throughout. The BEV encoder consists of 10 residual blocks and four multi-scale deformable-attention layers with pre-normalization. The entire model is implemented in pure PyTorch without custom CUDA kernels.

Training: Training runs for 300k steps using AdamW (weight decay 10^{-4}), batch size 4, a peak learning rate of 10^{-6} for the image encoder and 10^{-4} for the remaining modules, with cosine decay and 6k warmup steps. VFM feature distillation uses an L_1 loss. The losses are weighted: $\lambda_{\text{occ}} = 1.0$, $\lambda_{\text{sem}} = 0.5$, $\lambda_{\text{vfm}} = 0.5$, and $\lambda_{\text{depth}} = 0.5$. We use log-frequency class weighting (using the frequencies in Tab. 9 for each supervision source respectively) without any further tuning. For the combined pseudo point cloud and real lidar point cloud trainings, we use the lidar frequencies. We follow the same prompts, listed in Tab. 7, for generating pseudo-labels from GroundedSAM [27] as GaussianFlowOcc [4] for fair comparison. We supervise 3 frames forward and 3 frames backward per sample and subsample 30k 3D points per frame for efficiency. All training and inference benchmarks are obtained on A100 GPUs.

Method	RayIoU↑			IoU↑			FPS↑
	Sem.	Dyn.	Occ.	Sem.	Dyn.	Occ.	
GaussianFlowOcc	18.7	–	–	17.1	10.1	46.9	10.2
GaussianFlowOcc*	18.2	17.2	36.0	16.1	9.9	40.2	10.2
ResNet18	20.4	17.4	42.0	18.3	9.7	51.6	14.1
ResNet34	20.8	17.8	43.1	19.0	9.9	53.3	13.8
ResNet50	22.0	19.3	43.7	19.6	11.1	52.0	13.2
ResNet101	22.5	20.4	43.9	20.5	12.3	53.9	12.5
ConvNeXt-Tiny	22.2	19.7	43.9	20.0	11.3	54.0	12.9
ConvNeXt-Small	22.8	20.6	44.3	20.5	12.3	54.1	12.0
ConvNeXt-Base	23.6	21.7	45.2	21.3	13.2	55.0	11.6

Table 8. Comparison of image encoders in our QueryOcc framework. All configurations use identical training hyperparameters and supervision. We note that QueryOcc shows state-of-the-art performance using a wide range of image encoders. ConvNeXt-Base performs best, while retaining high inference speed. * denotes reproduction for both RayIoU and IoU.

8. Rendering Supervision Details

To compare image-space 2D supervision and 4D query-based supervision in the same framework, we implement a rendering-based baseline following neural rendering formulations adapted to occupancy prediction [16, 35]. The goal is to test 2D supervision pipeline within the same architecture and using identical supervision sources.

For each camera pixel, we cast a ray using known intrinsics and extrinsics, sample points along the ray with exponentially increasing depth spacing, and query the model’s predicted occupancy \hat{o} and feature vector \hat{v} at each location. Rendered features are obtained via alpha compositing:

$$\hat{f}_{\text{rendered}} = \sum_i T_i o_i f_i, \quad T_i = \prod_{j < i} (1 - o_j), \quad (7)$$

where \hat{o} acts as opacity and T_i is accumulated transmittance. A lightweight 6-layer MLP (hidden dimension 64) predicts per-pixel semantic logits from the rendered features. We supervise rendered depth, semantics, and VFM features using the same supervision sources as in the main QueryOcc model: Metric3D depth, Grounded-SAM semantics, and DinoV3 features.

To mitigate depth ambiguity, we apply a single round of importance sampling around the predicted depth, concentrating samples near likely surface intersections.

9. Additional Results

This section describes additional experiments and results.

9.1. Class-wise metrics

We observe similar trends as for RayIoU in the IoU metrics shown in Tab. 10. Drivable surface, sidewalk, and terrain achieve particularly high scores, suggesting that the model captures the ground-plane structure reliably. QueryOcc also

shows clear improvements on thin or small objects such as pedestrian and traffic cone, which can risk disappearing or blurring in BEV grids. The contracted BEV representation and query-based decoder appear to preserve local spatial detail even under noisy pseudo-labels. Rare classes such as bicycle, trailer, and construction vehicle remain difficult for all methods due to severe class imbalance and possibly label noise. QueryOcc nevertheless avoids the collapse observed in several baselines and maintains competitive performance across these categories.

9.2. Input image resolution

We evaluate the effect of input resolution supervising both with and without vision foundation model (VFM) features in Tab. 11. Performance increases steadily with higher resolutions for both supervision settings, indicating that the framework can make use of additional image detail when available. As expected, higher resolutions reduce throughput, reflecting the quadratic cost of encoding and lifting denser feature maps. The consistent gap between the *main* and *w. VFM* settings show that feature-distillation supervision provides complementary cues independent of resolution.

9.3. Pseudo-label source

We compare different semantic pseudo-label sources in Tab. 12. GroundedSAMv1 [27] is used in the main experiments to match prior work, but GroundedSAMv2 [27] provides slightly higher-quality masks and improves semantic metrics. The overall performance gap is small, which indicates that the framework is not overly sensitive to the specific pseudo-label source.

9.4. Pseudo-label noise

We evaluate robustness to noisy supervision in Sec. 9.4 by injecting controlled perturbations into geometry and semantics. Depth noise is modeled as additive Gaussian noise on Metric3D’s predicted inverse depth, while semantic noise is simulated by randomly flipping a fraction of pixel labels. Performance degrades gracefully as noise increases, demonstrating that QueryOcc is robust to this type of supervision error. We expect that as vision models continue to improve in accuracy and robustness, QueryOcc will naturally benefit from these advances.

9.5. Supervision signals

We compare two supervision strategies: direct 4D query-based supervision and an image-space rendering alternative that uses the same framework and supervision sources but applies losses via alpha compositing (Sec. 8). Table 14 lists the full metrics. Rendering supervision achieves reasonable performance (slightly better than similar prior work [16, 35, 41]) but performs worse than query-based supervision across both semantic and geometric metrics. Adding feature distillation

Source	barrier	bicycle	bus	car	cons. veh.	drive. surf.	manmade	motorcycle	pedestrian	sidewalk	terrain	traffic cone	trailer	truck	vegetation
Lidar	1.11	0.02	0.55	4.65	0.19	37.93	21.30	0.05	0.28	8.37	8.24	0.09	0.66	1.93	14.63
Pseudo	2.12	0.03	0.71	6.89	0.13	21.88	35.31	0.04	0.35	6.04	7.31	0.11	0.08	1.86	17.15

Table 9. Per-class point cloud frequency (%) for lidar point cloud and pseudo point cloud supervision sources.

Model	Mean	barrier	bicycle	bus	car	cons. veh.	drive. surf.	manmade	motorcycle	pedestrian	sidewalk	terrain	traffic cone	trailer	truck	vegetation
SelfOcc [16]	10.9	0.9	1.0	15.5	15.7	0.0	48.6	16.4	0.8	5.0	13.6	17.5	0.0	0.0	19.7	8.7
LangOcc [5]	11.6	1.9	5.2	12.9	18.2	0.2	20.8	23.6	8.9	12.8	3.4	12.3	13.3	1.7	11.7	27.6
GaussTR [19] -FeatUp	13.8	4.2	8.0	24.5	26.2	9.3	23.0	17.3	8.2	13.7	4.5	8.9	10.2	0.4	25.3	23.2
GaussTR [19] -T2D	14.2	9.4	12.7	35.1	30.0	8.5	18.9	19.4	14.8	17.1	5.3	3.2	3.0	4.3	19.2	12.4
GaussianFlowOcc [4]	18.2	13.1	9.6	36.9	30.7	5.1	39.8	22.9	10.5	15.7	13.3	15.2	13.3	1.3	27.7	19.6
QueryOcc	23.6	13.6	6.2	49.1	33.9	6.3	53.1	32.2	9.4	26.1	19.2	17.5	19.3	2.5	39.9	25.7
QueryOcc+	25.8	15.7	10.9	52.5	36.2	5.9	55.2	35.5	13.6	29.5	22.1	20.1	20.0	0.6	41.0	28.5

Table 10. Per-class **RayIoU**↑ performance on the Occ3D-nuScenes dataset. QueryOcc performs better than prior work on all classes. QueryOcc+ use high-res images, camera and lidar for supervision, and VFM feature distillation supervision pushing the limits for our framework and performs best of all methods. All baselines are reproduced results, and colors indicate **First**, **Second**, **Third** in ranking.

Model	Resolution	RayIoU↑			IoU↑			FPS↑
		Sem.	Dyn.	Occ.	Sem.	Dyn.	Occ.	
Main	256×256	21.6	18.9	43.4	19.1	10.2	54.1	14.1
	256×704	23.6	21.7	45.2	21.3	13.2	55.0	11.3
	504×896	23.9	22.0	45.6	21.9	14.0	55.1	6.4
	900×1600	24.7	23.1	45.9	22.9	15.3	55.5	2.4
w. VFM	256×256	21.9	19.7	43.5	19.6	10.8	54.5	13.4
	256×704	23.9	22.3	45.0	21.5	13.5	55.5	10.9
	504×896	25.0	23.5	46.2	22.7	15.0	56.0	6.2
	900×1600	25.8	24.2	46.4	23.1	16.4	56.3	2.4

Table 11. Effect of input resolution under the *main* settings with semantic and occupancy supervision as well as *with VFM* feature supervision. As the input resolution increases the performance goes up, but the inference speed goes down. The VFM model is slightly slower due to the higher output dimension in the decoder head.

Pseudo Labels	RayIoU↑			IoU↑		
	Sem.	Dyn.	Occ.	Sem.	Dyn.	Occ.
GroundedSAM v1	23.6	21.7	45.2	21.3	13.2	55.0
GroundedSAM v2	24.0	22.1	45.5	21.7	13.8	55.2

Table 12. Effect of different pseudo-label sources. Using GroundedSAMv2 gives a performance boost on the semantic metrics.

improves the query-based variant slightly, while combining rendering and query losses provides no additional benefit. This suggests that once explicit 4D supervision is available, image-space constraints become largely redundant.

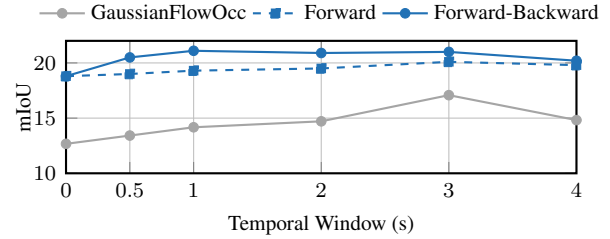


Figure 8. Effect of temporal window. Supervising across multiple timesteps improves geometric priors. Forward and backward supervision perform better than just forward.

9.6. Supervision window

Training supervision is derived from multiple adjacent time steps. Figure 8 shows that including nearby frames up to three seconds improves performance but degrades beyond that, suggesting that large temporal offsets introduce excessive difficulty for supervision. Unlike GaussianFlowOcc, our model remains stable even for narrow windows, indicating it can extract reliable geometric cues from limited temporal context. We apply supervision symmetrically in both forward and backward directions rather than predicting only the future. This bidirectional setup effectively doubles the available training signals and, as shown in Fig. 8, improves geometric reasoning around the reference frame. Together, these experiments confirm that temporal supervision provides stronger geometric priors and improves semantic occupancy for the input frame.

Depth	-	-	-	-	1e-3	5e-3	1e-2	1e-3	5e-3	1e-2
Semantics	-	5%	10%	20%	-	-	-	5%	10%	20%
Sem. \uparrow	23.6	23.2	22.8	22.4	23.2	22.4	20.8	22.9	22.0	19.2
Dyn \uparrow	21.7	20.1	19.7	19.0	20.8	20.3	19.4	20.4	19.5	16.3
Occ. \uparrow	45.2	45.7	45.3	45.1	45.2	43.1	41.5	44.5	43.4	40.8

Table 13. RayIoU for noisy depth and/or semantic psuedo-labels.

9.7. Point cloud sampling strategies

Table 15 compares alternative strategies for selecting which points to use for pseudo and lidar supervision. For lidar, moderate up weighting of dynamic points (2x–5x) improves dynamic-class performance by mitigating the inherent sparsity of moving objects. Higher weight to dynamic points beyond this offers no further benefit and can introduce instability. Sampling points uniformly from a voxel-grid with cell length 0.4m to match Occ3D slightly reduces performance. In contrast, pseudo point clouds are already dense also for dynamic objects, making their performance largely insensitive to the sampling scheme; uniform sampling performs on par with dynamic up sampling. For simplicity and reproducibility, we therefore adopt uniform downsampling.

9.8. Cell size variations

To isolate the effect of BEV resolution, Fig. 9 evaluates semantic and occupancy RayIoU across cell sizes. Performance improves as resolution increases, showing that coarse BEV grids can be a bottleneck. We hypothesize that large cells may smear thin structures, inflate surfaces, and degrade the decoder’s ability to localize geometry from queries. However, beyond roughly 0.16m, the improvements plateau and increasing resolution further yields negligible performance changes while growing memory and latency. This plateau justifies our design choice using 0.16m resolution near the ego vehicle, but contraction is required to maintain this resolution without exploding memory using long range.

9.9. Compression range sensitivity

Tab. 16 shows low sensitivity to d_{near} and K_{hr} , indicating that contracted BEV features preserve sufficient information to maintain performance.

Supervision Type				RayIoU↑			IoU↑			
Query-based				Rend	Sem.	Dyn.	Occ.	Sem.	Dyn.	Occ.
Occ.	Sem.	Feat.								
			✓	15.0	9.8	41.7	13.8	9.8	53.8	
✓	✓			23.6	21.7	45.2	21.3	13.2	55.0	
✓	✓	✓		23.9	22.3	45.0	21.5	13.5	55.5	
✓	✓	✓	✓	23.3	21.3	45.6	21.2	13.1	55.1	

Table 14. Ablation of supervision signals. Image-space rendering supervision performs worse than 4D query-based supervision in both semantic and geometric accuracy within our framework.

Sampling	Pseudo			Lidar		
	Sem.	Dyn.	Occ.	Sem.	Dyn.	Occ.
Uniform	23.6	21.7	45.2	22.9	19.4	48.3
Dyn. 2x	23.9	22.1	45.1	23.4	20.6	48.1
Dyn. 5x	23.3	21.2	44.5	24.3	22.1	47.8
Dyn. 10x	23.6	22.0	45.3	23.8	21.6	47.0
Voxel-uni	22.6	20.2	44.5	21.6	18.8	47.7

Table 15. RayIoU \uparrow for query sampling ablation under pseudo and lidar supervision. It is more important to upsample dynamic points when using lidar than pseudo-point clouds.

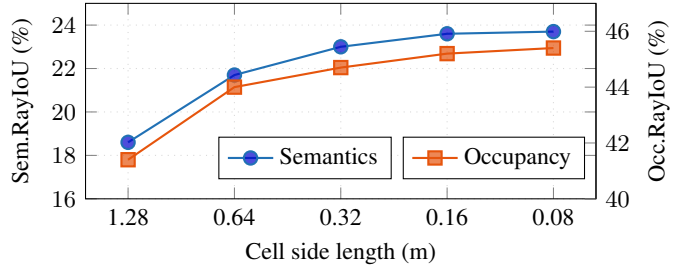


Figure 9. Effect of BEV cell size on RayIoU \uparrow metrics. Smaller cells leads to a notable improvements in performance, underscoring the need for an efficient, compact, and fine-grained BEV representation. In our framework the performance plateaus when using cell sizes smaller than 0.16.

$d_{\text{near}}, K_{\text{hr}}$	40m	30m	20m	10m
Sem. \uparrow	23.6	23.7	23.5	23.4
Dyn. \uparrow	21.7	21.4	21.2	21.1
Occ. \uparrow	45.2	45.7	45.8	45.4

Table 16. RayIoU exhibits low sensitivity to parameter choice

9.10. Increasing dataset size

We further study scaling by adding Argoverse 2 and the NAVSIM split of nuPlan [8, 10, 38] to the training data mix. To reduce preprocessing costs, we assume semantic pseudo-labels are available only for nuScenes, while the additional datasets provide lidar occupancy and VFM feature supervision. As shown in Table 17, adding either AV2 or nuPlan improves mean RayIoU by approximately +2 points over the nuScenes-only baseline, demonstrating that QueryOcc can leverage additional data even under weaker and heterogeneous supervision. Notably, nuPlan yields larger gains than AV2, likely due to its scale and closer distribution to nuScenes. However, combining all three datasets does not yield further improvements, indicating diminishing returns from additional data. Together with Table 6, this suggests

Datasets			RayIoU \uparrow		
nuSc	AV2	nuPl	Sem.	Dyn.	Occ.
✓			22.9	19.4	48.3
✓	✓		24.3	20.8	50.5
✓		✓	25.3	22.3	50.3
✓	✓	✓	25.0	22.3	50.1

Table 17. Effect of combining additional datasets with nuScenes.

that performance is ultimately limited by supervision quality: while additional data provides initial gains, weaker VFM-based supervision constrains further improvement compared to pseudo-label supervision.

9.11. Lidar segmentation

We leverage the lidar segmentation annotations introduced in Panoptic nuScenes [12] to analyze the semantic quality of both the pseudo-labels, and the predictions of QueryOcc trained using lidar point cloud supervision. These annotations provide per-point ground truth labels, enabling fine-grained evaluation beyond voxel-based metrics. For evaluation, we query QueryOcc at the 3D locations of the lidar points and compare the predicted semantic classes to the ground truth labels. We apply the same projection-based procedure used to construct pseudo-labels to obtain comparable predictions from the pseudo-labels for the full point cloud. Tab. 18 shows that QueryOcc consistently outperforms the pseudo-labels across all reported metrics. A more detailed view in Tab. 19 reveals that per-class intersection-over-union improves for all categories except *trailer*, indicating that the model does not simply replicate the supervision signal but systematically enhances it. This trend is further supported by the confusion matrices in Figure 10, where QueryOcc exhibits reduced inter-class confusion compared to the pseudo-labels, particularly among major semantic groups. However, ambiguities between semantically similar classes, such as *truck* and *trailer*, persist for both. This indicates that QueryOcc is able to denoise and refine the supervision signal, despite being trained on imperfect pseudo-labels.

Method	IoU	Rec.	F1	Acc.
Pseudo-labels	35.5	45.0	48.6	45.0
QueryOcc	49.8	64.5	62.0	64.5

Table 18. QueryOcc performs better on lidar segmentation metrics than the pseudo-ground truth.

10. Qualitative examples

We present qualitative examples from QueryOcc in Figs. 11 to 13 and separate videos. Overall, these examples show that QueryOcc produces sharp geometry, maintains fine-grained detail, and infers plausible structures behind occlusions.

Scene A (Fig. 11): Vehicle surfaces appear sharper and less boxy than the voxelized ground-truth occupancy, which is visibly inflated due to the 0.4 m discretization (black

boxes). Continuous-field self-supervised methods (including QueryOcc and *e.g.* GaussianFlowOcc), are typically not constrained by this voxel resolution and naturally produce higher-frequency details than the grid can represent. QueryOcc also recovers small, narrow structures such as road signs (white box), showing that the contracted BEV retains sufficient local detail for fine-scale prediction.

Scene B (Fig. 12): The motorcyclist on the right is correctly detected (black box), while nearby thin poles (white box) are missed. This reflects class imbalance and the difficulty of recovering extremely narrow structures. We note that the occupancy yields some structure for the bottom of the poles, but due to the class-uncertainty no semantic prediction is produced. Larger background regions such as terrain, sidewalk, vegetation are estimated well.

Scene C (Fig. 13): Pedestrians (black boxes) are accurately reconstructed, including the partially occluded walker with an umbrella in the left rear camera. The method naturally misses a pedestrian fully occluded by a bus (white box), but still infers plausible surface structure, *i.e.* drivable surface and sidewalk, behind the occlusion. The predicted occupancy also suggests a possible car, which likely reflects dataset bias toward parked vehicles in similar configurations. We hypothesize that temporal supervision across adjacent frames, in which the occluding objects move and expose new areas, encourages the model to infer plausible structures behind occlusions.

Long-range Scene D (Fig. 14): We extend the visualization beyond the ± 40 m Occ3D evaluation range to ± 60 m to assess how the model behaves in the contracted far-field region (outside the black dashed box). Notably, the contracted region still provides sufficient spatial signal for the decoder to reconstruct consistent surfaces, showing that the representation retains usable geometric information even after contraction.

Long-range Scene E (Fig. 15): This example illustrates the model’s ability to recover structure in a more complex far-field scenario. Although the road bends outside the high-resolution area, the decoder still reconstructs its curvature from contracted BEV features, indicating that the contraction preserves directional and structural cues at long ranges. The predictions remain geometrically stable up to 60m, suggesting that the model generalizes beyond the evaluation boundary and can leverage contracted features effectively.

11. Continuous Predictions vs. Voxelized Ground Truth

The qualitative examples reveal a systematic discrepancy between our predictions and the voxelized ground truth used in Occ3D. The Occ3D ground truth is constructed by fusing multiple lidar sweeps into a fixed 0.4m grid, which inflates geometry, thickens object boundaries, and removes fine-grained structure.

Source	Mean	barrier	bicycle	bus	car	cons. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	sidewalk	terrain	manmade	vegetation
Pseudo-labels	35.5	19.1	9.5	48.5	55.6	9.2	23.2	28.5	15.6	4.1	32.8	79.3	43.6	46.3	62.5	54.2
QueryOcc	49.8	27.6	16.4	78.3	80.7	15.1	46.0	55.9	32.3	0.3	62.0	87.9	44.0	53.6	74.5	73.2

Table 19. Per-class IoU (%) against lidar segmentation ground truth.

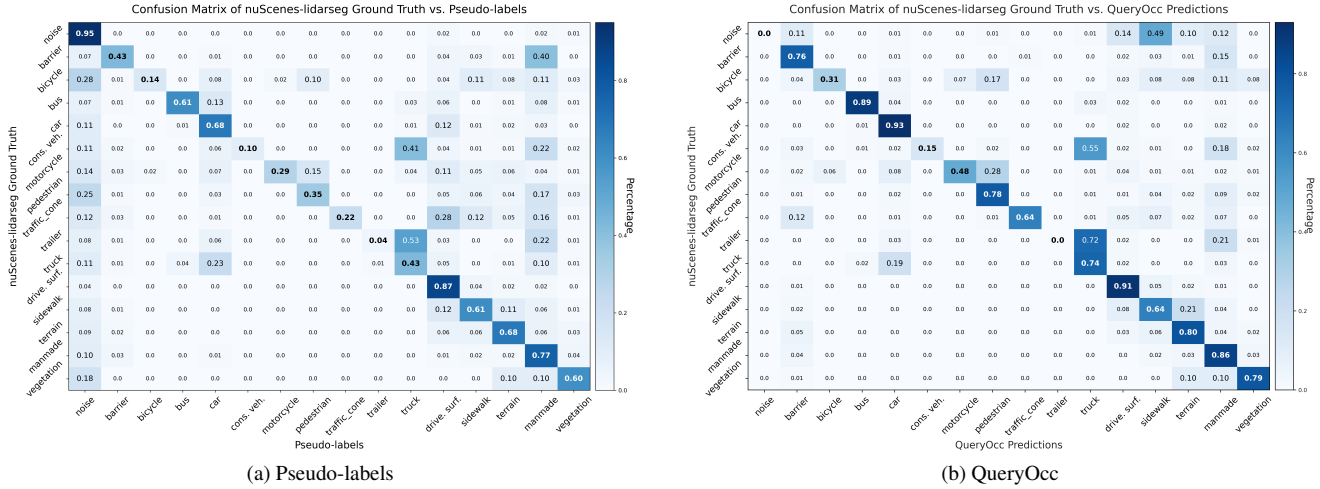


Figure 10. Confusion matrices for the relevant classes on the nuScenes lidar segmentation dataset [12].

Continuous-field methods, including ours and rendering-based approaches such as GaussianFlowOcc, learn a smooth occupancy function in \mathbb{R}^3 and are not restricted by voxel resolution. They can therefore produce sharper transitions and thin structures that the discretized labels cannot represent. These representational differences can make predictions appear visually higher-resolution than the voxel grid, particularly around vehicles, poles, trees, and other narrow structures, without leading to increased metric scores. This mismatch also motivates the use of RayIoU, which is less sensitive to voxel inflation and better aligned with the underlying geometry.

Overall, the discrepancy is inherent to voxel-based supervision and evaluation. Continuous 4D methods generate finer spatial detail than the evaluation grid can express, which can put them at a disadvantage under voxel-based metrics even when the underlying geometry is consistent.

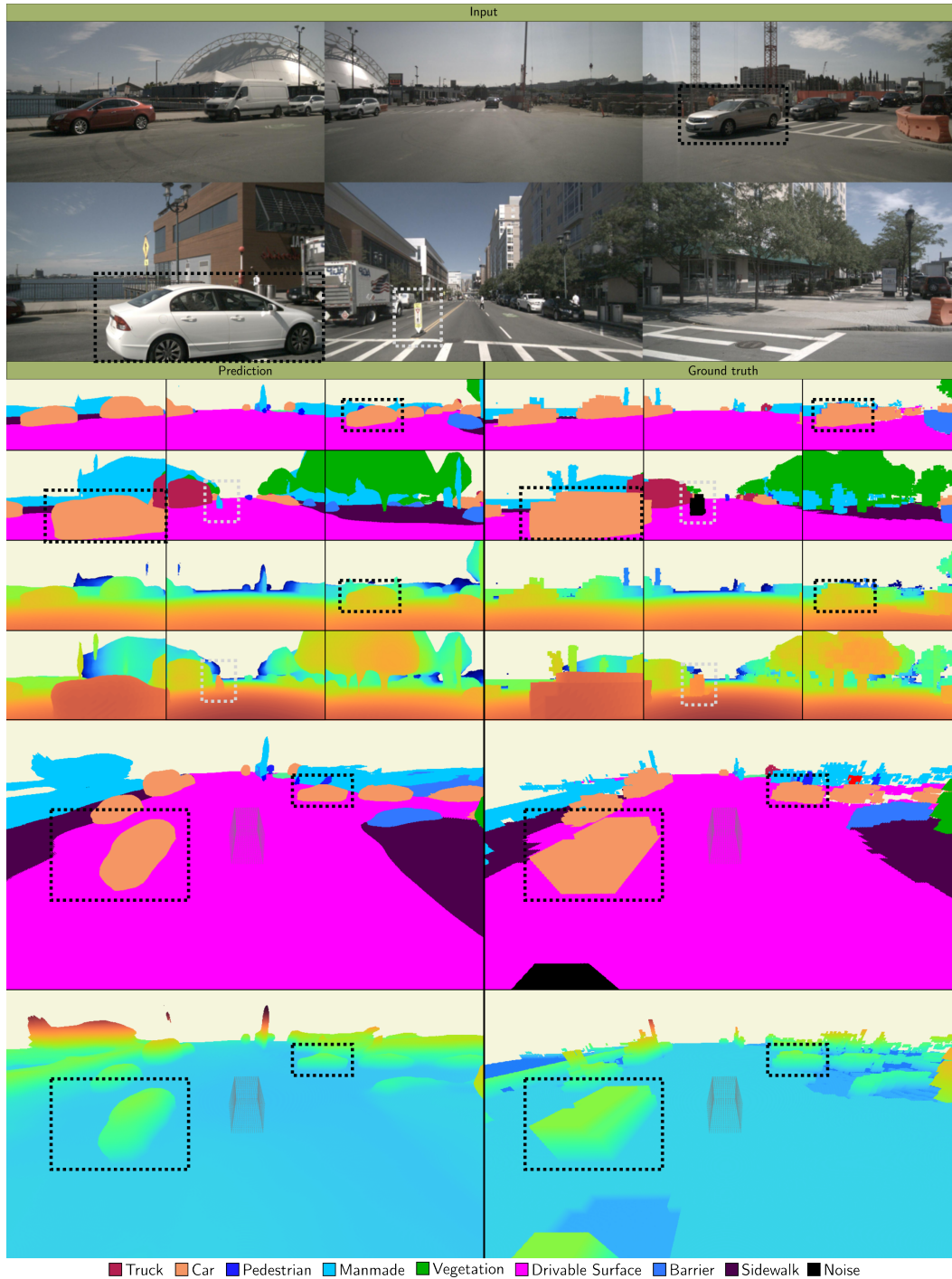


Figure 11. **Scene A:** Qualitative example from QueryOcc. The model reconstructs vehicle geometry with high fidelity (black boxes) and recovers small narrow structures such as road signs (white box). Semantic and occupancy predictions remain spatially consistent across camera views and exhibit smoother geometry than the coarse voxel grid from the Occ3D ground truth.

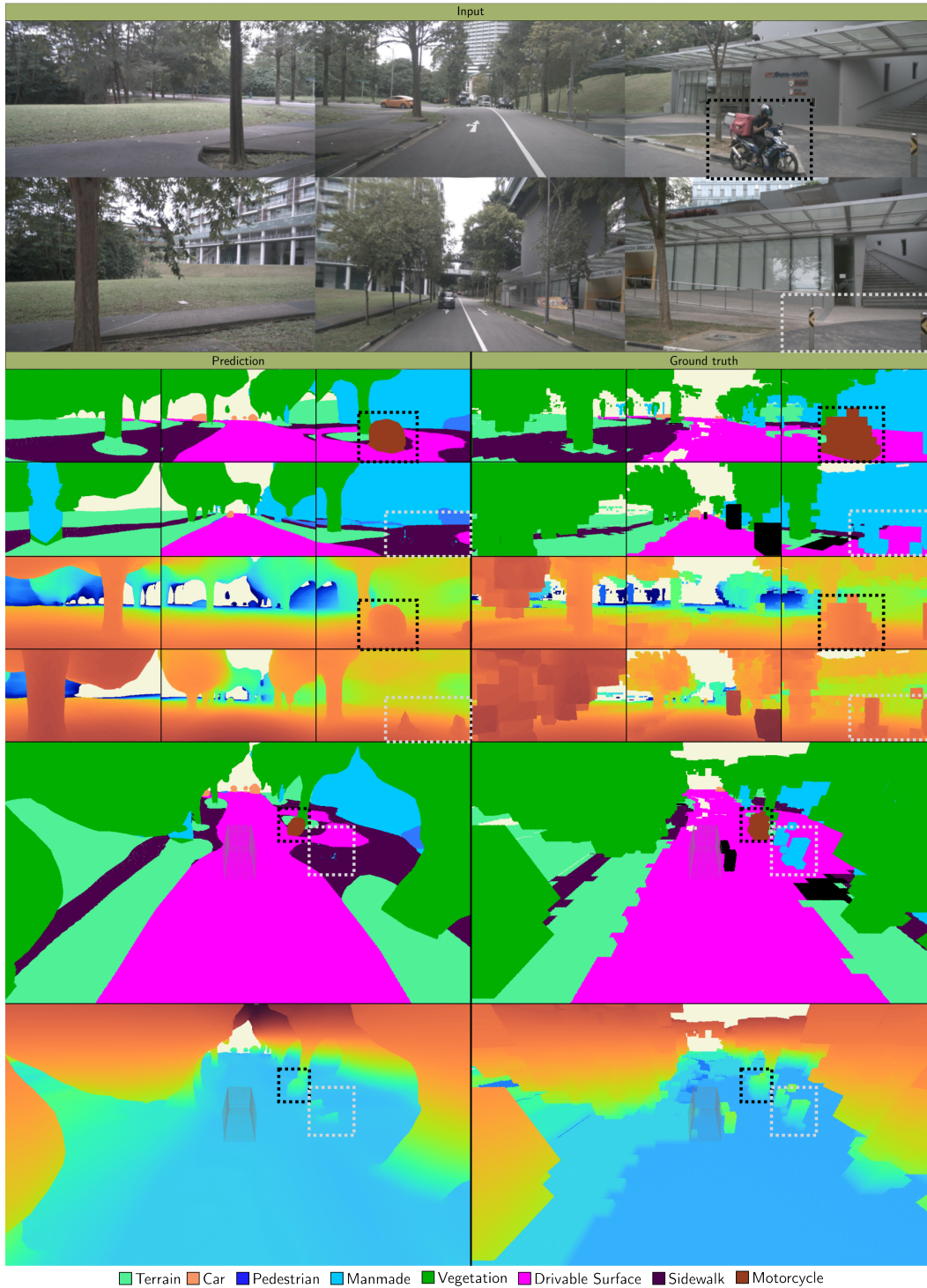


Figure 12. **Scene B:** The motorcyclist is correctly identified (black box), while thin manmade poles are missed semantically despite partial occupancy signals (white box), illustrating the difficulty of extremely narrow structures. Broader scene elements such as sidewalk, terrain, and vegetation are predicted coherently.

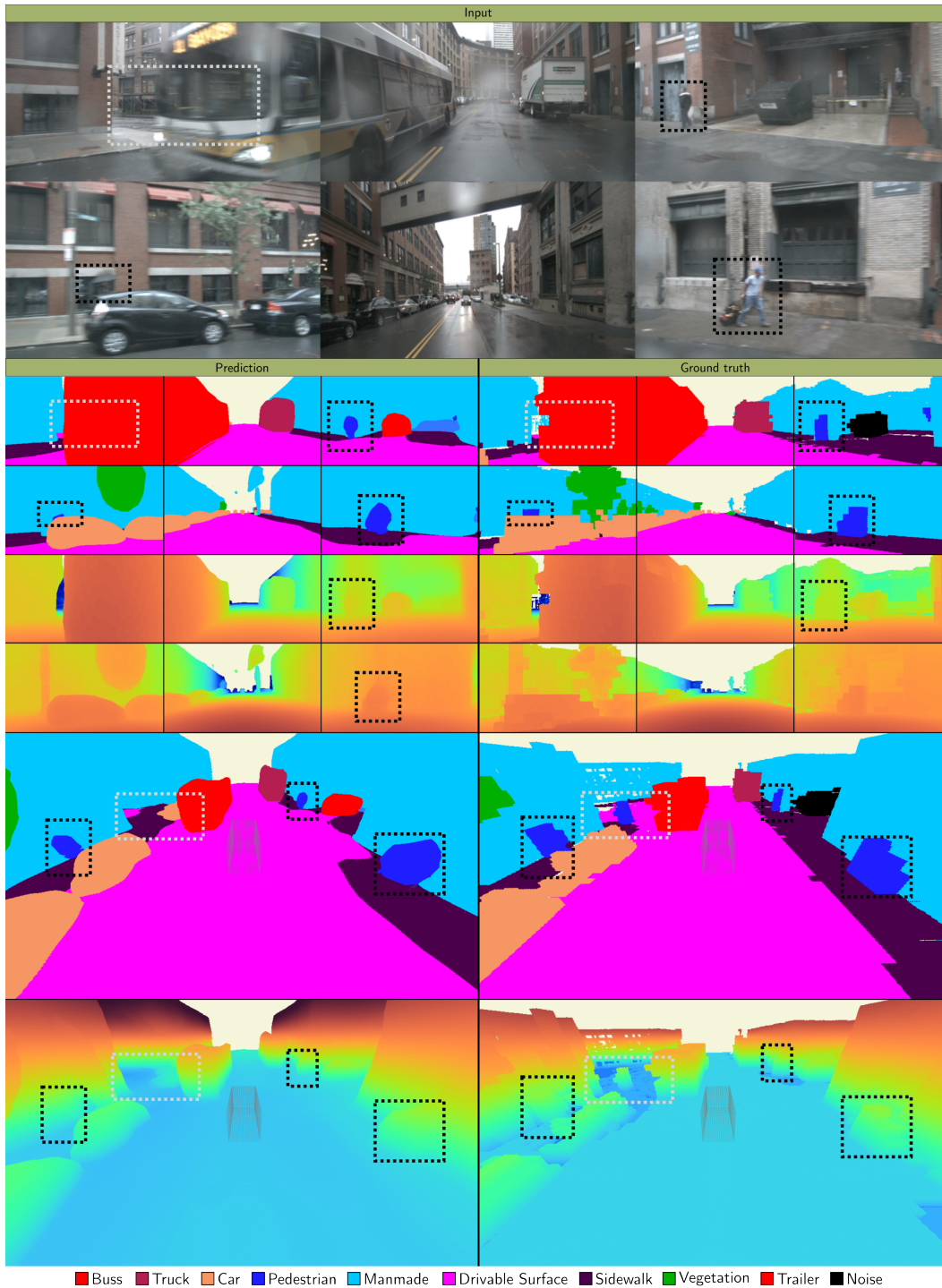


Figure 13. **Scene C:** Pedestrians (black boxes) are reconstructed with high consistency, including partially occluded instances. A fully occluded pedestrian behind a bus is missed (white box), though the model infers plausible surface continuation behind an occlusion. The predicted occupancy also reflects common priors such as possible parked vehicles in similar street configurations.

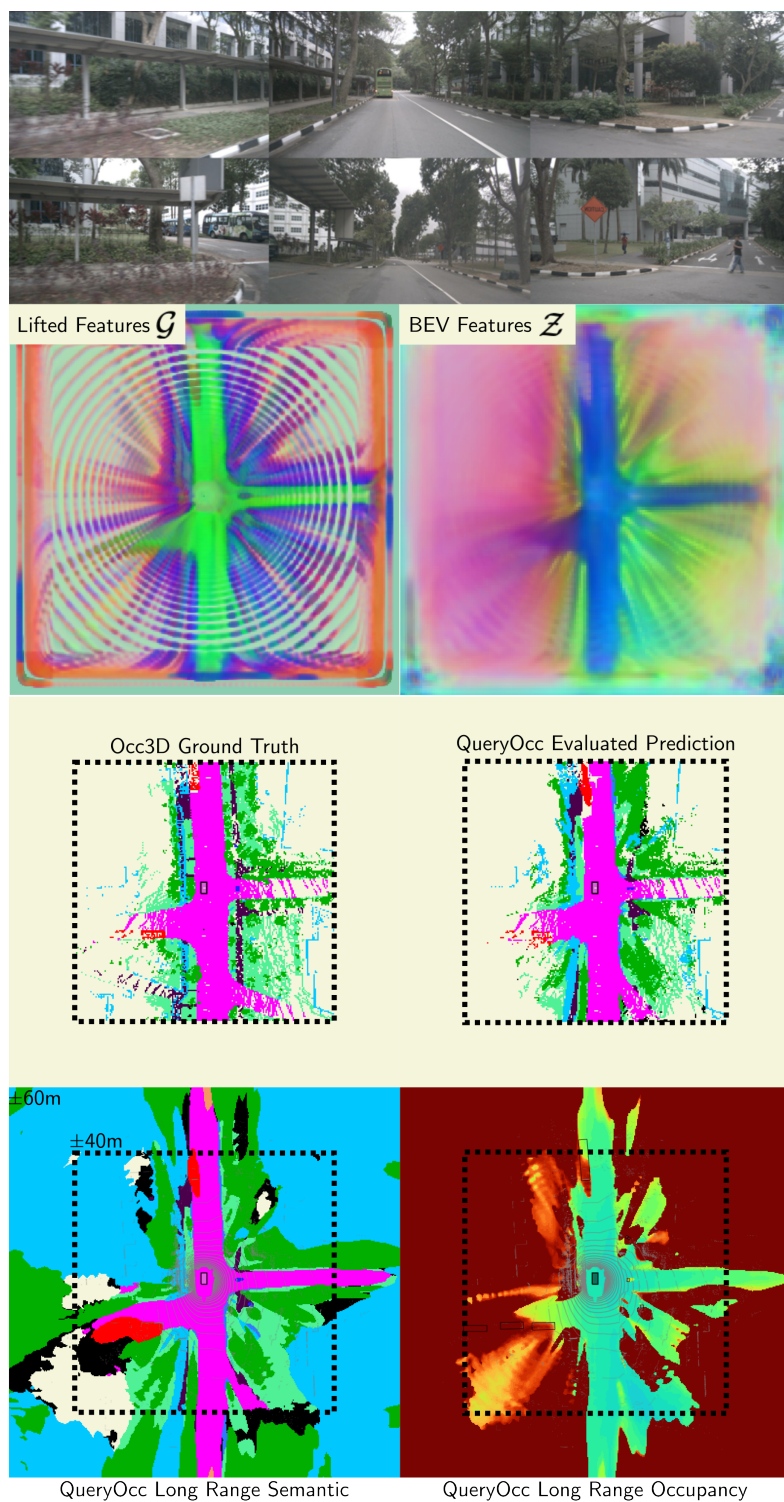


Figure 14. **Scene D:** Predictions beyond the Occ3D evaluation range. We visualize lifted features, BEV features, predictions, and ground truth, extending the view from the standard ± 40 m to ± 60 m. The decoder reconstructs plausible geometry both inside the high-resolution region and in the contracted far-field area (black dashed box). Road layout and free-space structure remain consistent even well outside the evaluation boundary.

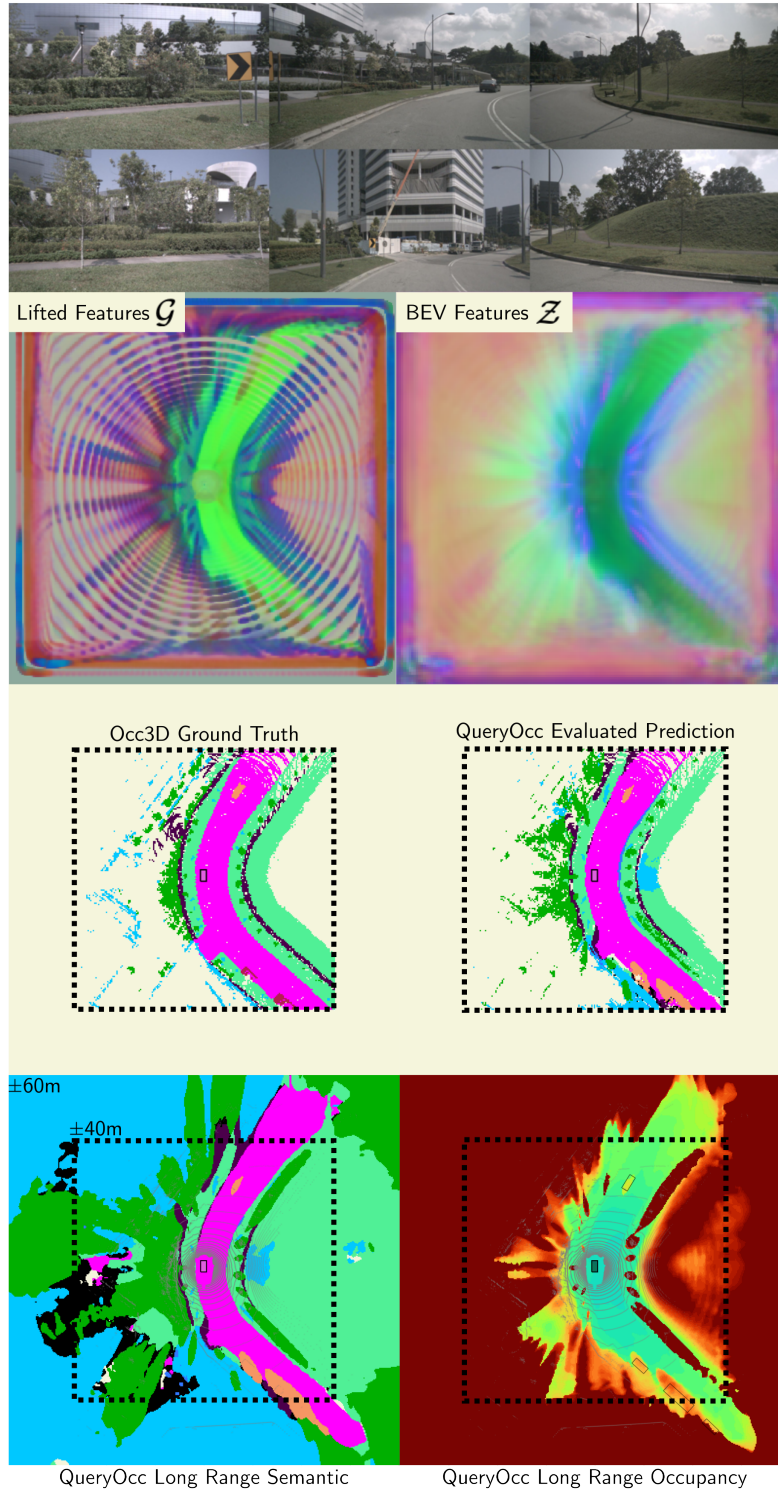


Figure 15. **Scene E:** Geometric consistency in the contracted far-field. We again extend visualization to 60m range. The decoder recovers the bending road and surrounding free-space structure from the contracted region outside the high-resolution area (black dashed box), demonstrating that the representation preserves spatial cues even at long distances.