COTR: Compact Occupancy TRansformer for Vision-based 3D Occupancy Prediction

Supplementary Material

Method	Epoch	Mask	IoU	mIoU	others ⬛	barrier ◼	bicycle 🟥	bus 🟧	car 🟧	const. veh. 🟧	motorcycle 🟥	pedestrian 🟦	traffic cone ⬛	trailer 🟧	truck 🟥	drive. suf. 🟦	other flat 🟥	sidewalk 🟪	terrain 🟩	manmade 🟫	vegetation 🟩
MonoScene [3]	24	✗	-	6.1	1.75	7.23	4.26	4.93	9.38	5.67	3.98	3.01	5.90	4.45	7.17	14.91	6.32	7.92	7.43	1.01	7.65
OccFormer[41]	24	✗	30.1	20.4	6.62	32.57	13.13	20.37	37.12	5.04	14.02	21.01	16.96	9.34	20.64	40.89	27.02	27.43	18.65	18.78	16.90
BEVFormer [18]	24	✗	-	26.9	5.85	37.83	17.87	40.44	42.43	7.36	23.88	21.81	20.98	22.38	30.70	55.35	28.36	36.0	28.06	20.04	17.69
CTF-Occ [32]	24	✗	-	28.5	8.09	39.33	20.56	38.29	42.24	16.93	24.52	22.72	21.05	22.98	31.11	53.33	33.84	37.98	33.23	20.79	18.0
VoxFormer [17]	24	✔	-	40.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SurroundOcc [37]	24	✔	-	40.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
FBOcc [19]	20	✔	-	42.1	14.30	49.71	30.0	46.62	51.54	29.3	29.13	29.35	30.48	34.97	39.36	83.07	47.16	55.62	59.88	44.89	39.58
TPVFormer [10]	24	✔	66.8	34.2	7.68	44.01	17.66	40.88	46.98	15.06	20.54	24.69	24.66	24.26	29.28	79.27	40.65	48.49	49.44	32.63	29.82
+ COTR (Res-50)	24	✔	**70.6**	**39.3**	**11.66**	**45.47**	**25.34**	**41.71**	**50.77**	**27.39**	**26.30**	**27.76**	**29.71**	**33.04**	**37.76**	**80.52**	**41.67**	**50.82**	**54.54**	**44.91**	**38.27**
SurroundOcc* [37]	24	✔	65.5	34.6	9.51	38.50	22.08	39.82	47.04	20.45	22.48	23.78	23.00	27.29	34.27	78.32	36.99	46.27	49.71	35.93	32.06
+ COTR (Res-50)	24	✔	**71.0**	**39.3**	**11.37**	**45.90**	**25.97**	**43.08**	**51.56**	**25.55**	**26.21**	**26.53**	**29.48**	**33.65**	**38.87**	**80.45**	**40.34**	**50.86**	**53.88**	**45.37**	**38.94**
OccFormer[41]	24	✔	70.1	37.4	9.15	45.84	18.20	42.80	50.27	24.00	20.80	22.86	20.98	31.94	38.13	80.13	38.24	50.83	54.3	46.41	40.15
+ COTR (Res-50)	24	✔	**71.7**	**41.2**	**12.19**	**48.47**	**27.81**	**44.28**	**52.82**	**28.70**	**28.16**	**28.95**	**31.32**	**35.01**	**39.93**	**81.54**	**42.05**	**53.44**	**56.22**	**47.37**	**41.38**
BEVDet4D [8]	24	✔	73.8	39.3	9.33	47.05	19.23	41.47	52.21	27.19	21.23	23.32	21.58	35.77	38.94	82.48	40.42	53.75	57.71	49.94	45.76
+ COTR (Res-50)	24	✔	**75.0**	**44.5**	**13.29**	**52.11**	**31.95**	**46.03**	**55.63**	**32.57**	**32.78**	**30.35**	**34.09**	**37.72**	**41.84**	**84.48**	**46.19**	**57.55**	**60.67**	**51.99**	**46.33**
BEVDet4D [8]	36	✔	72.3	42.5	12.37	50.15	26.97	**51.86**	54.65	28.38	28.96	29.02	28.28	37.05	42.52	82.55	43.15	54.87	58.33	48.78	43.79
+ COTR (Swin-B)	24	✔	**74.9**	**46.2**	**14.85**	**53.25**	**35.19**	50.83	**57.25**	**35.36**	**34.06**	**33.54**	**37.14**	**38.99**	**44.97**	**84.46**	**48.73**	**57.60**	**61.08**	**51.61**	**46.72**

Table 4. **3D Occupancy prediction performance on the Occ3D-nuScenes dataset.** We present the IoU (geometry) and mean IoU (semantic) over categories and the IoUs (semantic) for different classes.

6. Further Implementation Details

In this section, we further elaborate on the implementation details of our COTR.

Geometry-aware Occupancy Encoder. As we mentioned in Sec. 3.3, the Explicit View Transformation generates a occupancy feature $O_{\mathrm{E}} \in \mathbb{R}^{32\times200\times200\times16}$. Then, we use a 3D-ResNet according to [8] to generate multi-scale occupancy features $O^0_{\mathrm{E}} \in \mathbb{R}^{32\times200\times200\times16}, O^1_{\mathrm{E}} \in \mathbb{R}^{64\times100\times100\times8}, O^2_{\mathrm{E}} \in \mathbb{R}^{128\times50\times50\times4}$. Next, employ trilinear interpolation to sample multi-scale OCC features into the same size of $50 \times 50 \times 16$, followed by a concatenation and convolutional layer to construct the compact OCC representation $O_c \in \mathbb{R}^{256\times50\times50\times16}$. Finally, the compact OCC representation is fed into Implicit View-Transformation for further update. Since the compact OCC feature O_c has been already initialized by EVT, we only use 1 Transformer layer in IVT. With the final prediction resolution being $200 \times 200 \times 16$, we use deconvolution layers to upsample the compact OCC feature to $O \in \mathbb{R}^{256\times200\times200\times16}$, which was only utilized for mask prediction in Semantic-aware Group Decoder. In order to counteract the loss of geometric details throughout the process of downsampling, we construct a U-net[29] architecture by concatenated multi-scale features $\{O^i_{\mathrm{E}}\}^3_{i=0}$ to the upsampled features.

Loss Function. During training, we used a total of 3 different loss functions:

$$\mathcal{L} = \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{mask-cls}}\mathcal{L}_{\text{mask-cls}}, \quad (4)$$

where $\mathcal{L}_{\text{depth}}$ is the depth estimation loss in the Image Featurizers following BEVDepth [16], $\mathcal{L}_{\text{seg}}$ is a simple cross-entropy segmentation loss between a coarse prediction from O and the ground truth label, and the $\lambda_{\text{mask-cls}}$ loss combines a cross entropy classification loss and a binary mask loss for each predicted mask segment following MaskFormer [6]. We set the hyper-parameters to $\lambda_{\text{depth}} = \lambda_{\text{mask-cls}} = 1$ and $\lambda_{\text{seg}} = 10$.

7. Further Experiments

Per-class comparison with SOTA. We report more quantitative details in Table 4 about our experimental results for better comparison with other competitors. Besides TPVFormer [10] and BEVDet4D [8], we also integrate COTR

Rep.	Resolution	IoU (%)	mIoU (%)	FLOPs (G)
BEV	200 × 200	70.48	37.51	65.74
TPV	200 × (200, 16 × 2)	70.41	37.21	291.62
Voxel	200 × 200 × 16	-	-	402.98
	100 × 100 × 8	70.83	37.36	101.73
	50 × 50 × 4	70.87	37.61	**61.27**
	50 × 50 × 8	70.71	37.45	67.11
	50 × 50 × 16	**70.89**	**37.97**	78.47

Table 5. **Ablation study for different occupancy representation resolution.** All models are trained without Semantic-aware Group Decoder and long-term temporal information. We report the FLOPs of the Implicit View Transformation module.

Component			Computational Cost			
GOE	TD	CFSG	Params (M)		FLOPs (G)	
			34.97	-	541.21	-
✔			35.84	+0.87	573.82	+32.61
	✔		36.21	+1.24	628.75	+87.54
	✔	✔	36.21	+1.24	628.75	+87.54

Table 6. **Ablation study on each component's computational cost.** All models are trained without long-term temporal information.

into SurroundOcc [37] and OccFormer [41]. COTR yields significant performance improvements in both geometric completion and semantic segmentation, surpassing SurroundOcc and OccFormer by 5.5%, 1.6% in IoU and 4.7%, 3.8% in mIoU. Notably, a conspicuous amelioration primarily resides within small objects and rare objects, demonstrating that our approach can indeed apprehend finer geometric details, and substantially enhance semantic discernibility.

Ablation study for different OCC resolution. Table 5 compared different resolutions for OCC representations in our experiments. It is abundantly clear that the high-resolution (200 × 200 × 16) OCC representation incurs a massive computational overhead, with the FLOPs approximately 5× that of the compact (50 × 50 × 16) OCC representation, respectively. Additionally, it appears that preserving the height information is beneficial for the task of occupancy prediction. Overall, our compact OCC representation strikes a balance between performance and computational overhead.

Ablation study on computational cost. As shown in Table 6, our proposed COTR is an efficient approach in which each component does not add a significant amount of computational costs. It's worth noting that, since we only used the Coarse-to-Fine Semantic Grouping (CFSG) strategy during training and kept only one group during inference, CFSG doesn't introduce any extra overhead. More metrics are reported in Table 7. All baselines are concurrently equipped with EVT and IVT for a fare comparison.

method	Params (M)	FLOPs (G)	LT	Latency (s)
TPVFormer	54.05	972.75	✗	0.59
+ COTR (R50)	53.30	784.85	✗	0.43
BEVDet4D †	35.67	924.07	✗	0.67
+ COTR (R50)	38.87	747.26	✗	0.42
BEVDet4D	35.67	1049.52	✔	1.89
+ COTR (R50)	38.87	747.26	✔	1.43
BEVDet4D	121.28	4195.12	✗	1.67
+ COTR (SwinB)	104.99	3761.06	✗	1.43

Table 7. **Ablation study on efficiency.** All models are tested on a RTX A6000 GPU. † input size of 256 × 704, others same as Tab. 1.

8. Visualization

In this section, we provide more visualization results of our method.

Visual Ablations on Occluded Scenes. To validate the robustness of our method in handling occluded scenes, we present additional visual results. As shown in the first scene in Fig. 7, our method without using long-term temporal information, successfully detects small objects (such as pedestrians and bicycles) within a limited occlusion range. However, in the second scene, when a significant portion of the vehicle is obscured, the model struggles to correctly identify occluded objects due to the constrained camera perspective.

Visual Ablations on Low-light Scenes. To validate the robustness of our method in handling low-light environments, we present additional visual results. As illustrated in the first scenario of Fig. 8, our model is capable of successfully detecting unknown objects in the dark. However, the second scenario shows that while our model can detect small objects in the dark from a distance, it fails to predict successfully when part of the camera is nearly completely obscured by darkness. This limitation is primarily due to the camera's perception capabilities and other modalities such as LiDAR or Radar might be required to aid successful detection.

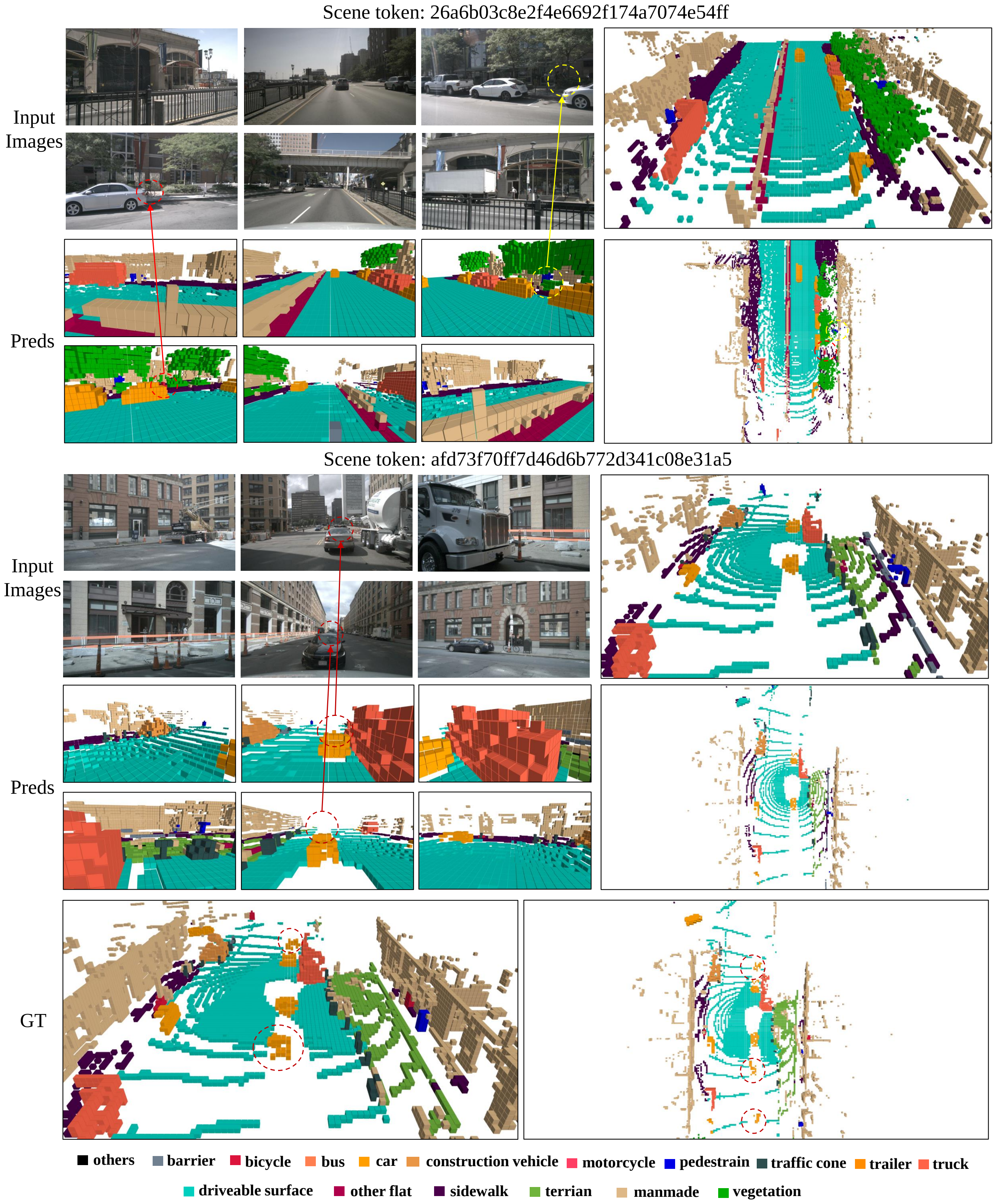

Figure 7. Visualizations for occlusion on OCC3D-nuScenes validation set. For each scene, the six images in the *"Input Images"* line left are the inputs to our model captured by font-left, front, front-right, back-right, back, and back-right cameras. The six images in the *"Preds"* line left denote our prediction results with the corresponding views as the inputs. The two images on the right provide a global view of our predictions. The two images in the *"GT"* line provide a global view of ground truth.

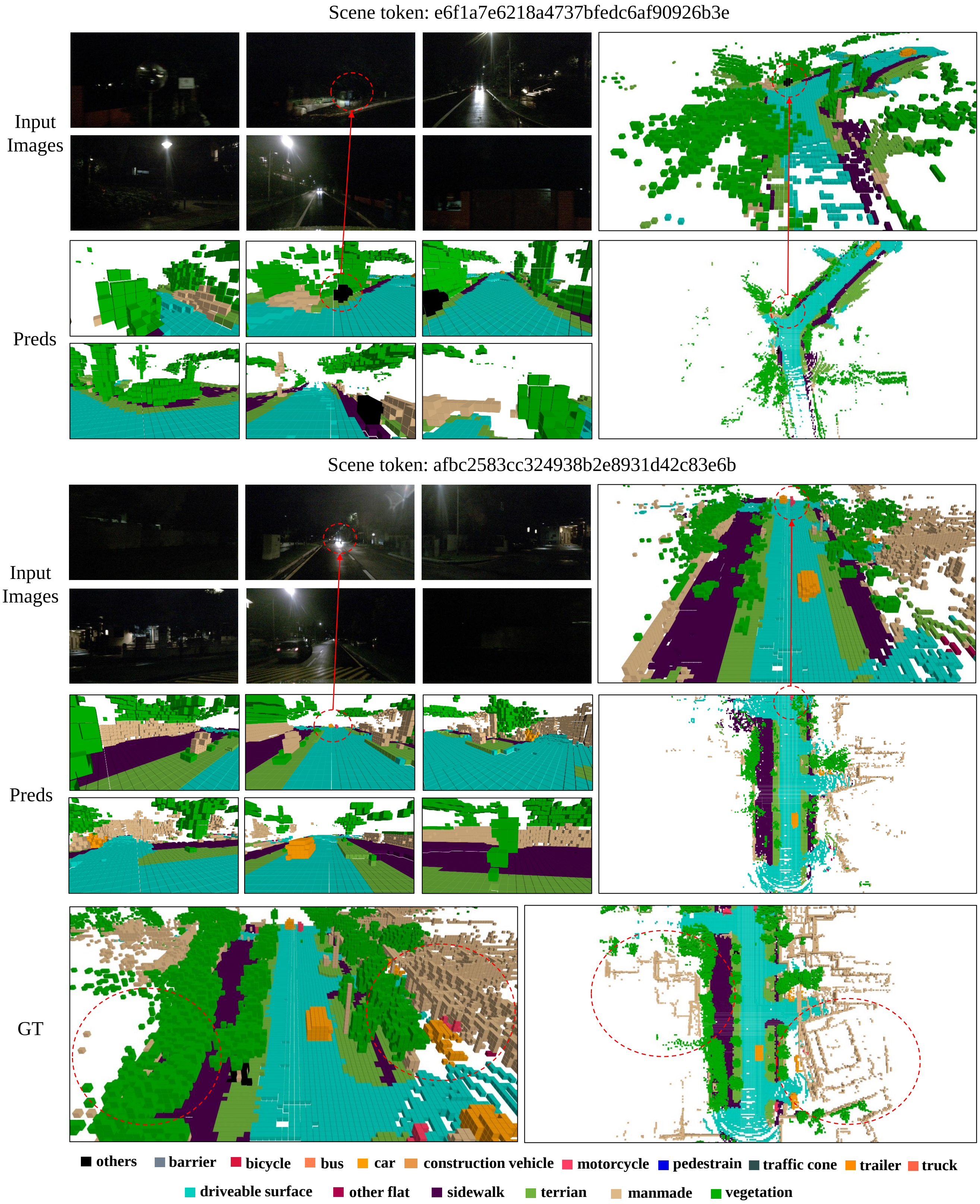

Figure 8. Visualizations for low-light environments on OCC3D-nuScenes validation set.