SDGOCC: Semantic and Depth-Guided Bird's-Eye View Transformation for 3D Multimodal Occupancy Prediction

Supplementary Material

A. Point-to-Pixel Correspondence

We can obtain the index between points and image pixels using the given extrinsic T_{ex} matrix between sensors and the camera's intrinsic K matrix. Each point includes 3D coordinates (x, y, z) and the obtained 2D projection pixel coordinate ${}^{C}p_i = \begin{bmatrix} u_i & v_i \end{bmatrix} \in \mathbb{R}^{H_C \times W_C}$. To apply the coordinate transformation, we add a fourth column to turn it into a 4D vector, making the equation homogeneous. The projection process is described as follows:

$$\begin{aligned} Z_i \, {}^{C}p_i &= K T_{ex} \, {}^{L}p_i \\ &= K[R \mid t] \begin{bmatrix} x_i & y_i & z_i & 1 \end{bmatrix}^T, \end{aligned} \tag{9}$$

$$T_{ex} = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix}, \tag{10}$$

where ${}^{L}p_i$ and ${}^{C}p_i$ are the corresponding 3D point coordinates and image pixel coordinates, and R and t represent the rotation matrix and the translation vector in the transformation matrix T_{ex}, Z_i is the depth of the 3D point.

B. Further Implementation Details

In this section, we further elaborate on the implementation details of our SDG-OCC.

Evaluation Metrics. To evaluate the performance of our method, we use the Mean Intersection over Union (mIoU) over all classes for evaluation, which quantifies the overlap between actual and predicted values relative to their combined set. It is calculated by

$$\mathrm{mIoU} = \frac{1}{N} \sum_{c=1}^{N} \mathrm{IoU}_c = \frac{1}{N} \sum_{c=1}^{N} \frac{TP_c}{TP_c + FP_c + FN_c}, \tag{11}$$

where TP_c, FN_c, FP_c denote true positives, false negatives, and false positives, respectively. After averaging the IoU values per class, the primary evaluation metric for semantic segmentation, mIoU, is derived.

Training Details. The image size is adjusted to 256×704 with normalization, padding and random flipping for image augmentation. To enhance performance, an exponential moving average (EMA) hook is adopted. Test-time augmentation techniques are not employed for BEV augmentation.

Loss Function. During training, we use a total of five different loss functions:

$$\begin{aligned} \mathcal{L} = \; & \lambda_{\text{depth}} \mathcal{L}_{\text{depth}} + \lambda_{\text{seg}} \mathcal{L}_{\text{seg}} + \lambda_{\text{pts}} \mathcal{L}_{\text{pts}} \\ & + \lambda_{\text{mask-occ}} \mathcal{L}_{\text{mask-occ}} + \lambda_{\text{kl}} \mathcal{L}_{\text{distill}} \end{aligned}, \tag{12}$$

Fusion Method	IoU(%)	mIoU(%)	FLOPs(G)	Params(M)
SDG-FB	95.26	51.32	1159	92.1
SDG-FA	95.35	51.66	1162	92.9
SDG-KL	95.26	50.16	1067	57.9

Table 6. The ablation study on the fusion strategies.SDG-FB represents fusion after the BEV encoder, SDG-FA represents fusion before the BEV encoder,SDG-KL represents fusion and distillation.

Method	IoU(%)	mIoU(%)	FLOPs(G)	Params(M)	Time(ms)
img†	94.62	48.51	1067	57.9	83
img+lidar	95.35	51.66	1162	92.9	133
img†+lidar	95.48	51.71	2251	100.48	217

Table 7. The ablation study on the temporal fusion. † represents the temporal fusion module for multi-frame images.

where $\mathcal{L}_{\text{depth}}$ and $\mathcal{L}_{\text{seg}}$ represent the losses for depth estimation and image semantic segmentation in the view transformation, both using cross-entropy loss. $\mathcal{L}_{\text{pts}}$ denotes the voxel supervision loss based on the target scale in LiDAR segmentation, integrating both Lovasz loss and cross-entropy loss. $\mathcal{L}_{\text{mask-occ}}$ is the loss between the predicted occupancy grid and the ground truth (GT), employing masked cross-entropy loss. $\mathcal{L}_{\text{distill}}$ is the loss function for knowledge distillation, representing the weight ratio of unidirectional knowledge transfer from multimodal to unimodal representations, where $\alpha = \beta$=1.0. And we set the hyperparameters to $\lambda_{\text{depth}} = 0.05$ and $\lambda_{\text{seg}} = 0.5$ for the depth and segmentation losses, respectively, and additionally include weights λ_{pts}=$\lambda_{\text{mask-occ}}$=$\lambda_{\text{kl}}$=1.0 to balance their contributions in the overall loss function.

C. Further Experiments

In this section, we provide more experiments of our proposed method.

Feature Fusion Strategies. The feature fusion strategies of the feature fusion has a significant impact on the final performance. We conducte experiments on the image BEV features obtained from the initial view transformation and those processed by the BEV encoder. According to the results in Tab. 6, applying the encoder before fusion and distillation yields better performance.

Effectiveness of Temporal Fusion. Temporal augmentation is a crucial technique for enhancing 3D perception performance. While it introduces multi-frame information, it causes only a small increase in memory overhead yet brings substantial performance improvements. To validate the effectiveness of temporal fusion, we conducte ex-

translations (m)	rotations (°)	IoU↑	mIoU↑
0.1	1	-1.4%	-0.12%

Table 8. Random perturbation on the matrix

method	IoU	mIoU
Baseline+FOAD	94.49 %	43.51%
Baseline+Fusion	94.74%	44.92%

Table 9. Ablation experiment on the neighborhood attention

periments using common temporal fusion methods from BevStereo, with results shown in Tab. 7. Integrating temporal information into camera-only models yields a noticeable performance improvement with minimal computational overhead and a small compromise in processing speed. However, when multimodal data is further integrated with temporal information, the improvement becomes negligible, indicating a potential intrinsic coupling between the temporal dynamics of multi-frame image sequences and LiDAR information.

Sensitivity to intrinsic and extrinsic. LSS methods are sensitive to both intrinsic and extrinsic factors due to the unidirectional mapping between image pixels and 3D points. Any misalignment in the intrinsic parameters of the camera or the extrinsic parameters of the 3D scene can lead to significant errors. In contrast, our approach leverages copoints from LiDAR as priors and employs bidirectional projection between 3D points and image pixels. The perturbations in the transformation matrix have minimal impact on the bidirectional projection process, resulting in a more stable and robust mapping. Random perturbations, including translations and rotations (see Tab. 8), demonstrate that our method maintains minimal performance degradation, highlighting its robustness.

Ablation Study on the Fusion-Distillation Module. We conducted ablation experiments on the baseline, separately for the performance-based fusion method and the speed-based distillation method. The results are shown in Tab. 9, where it can be observed that the fusion-based method performs slightly better than the fusion-distillation method.

D. Visualization

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

Visualization on Occluded Scenes. To validate the robustness of our method in handling occluded scenes, we present additional visual results on both OCC3D-nuScenes and Openocc-nuScenes dataset. As shown in the first scene in Fig. 8, our method is able to detect pedestrians that are partially occluded. As shown in the second scene, our approach performs well in identifying vehicles that are largely occluded by other vehicles.

Visualization 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 Fig. 9, on the OCC3D-nuScenes, our model demonstrates the capability to recognize small objects, such as pedestrians and motorcycles, in low-light scenes, and it can effectively detect motorcycles at a considerable distance.

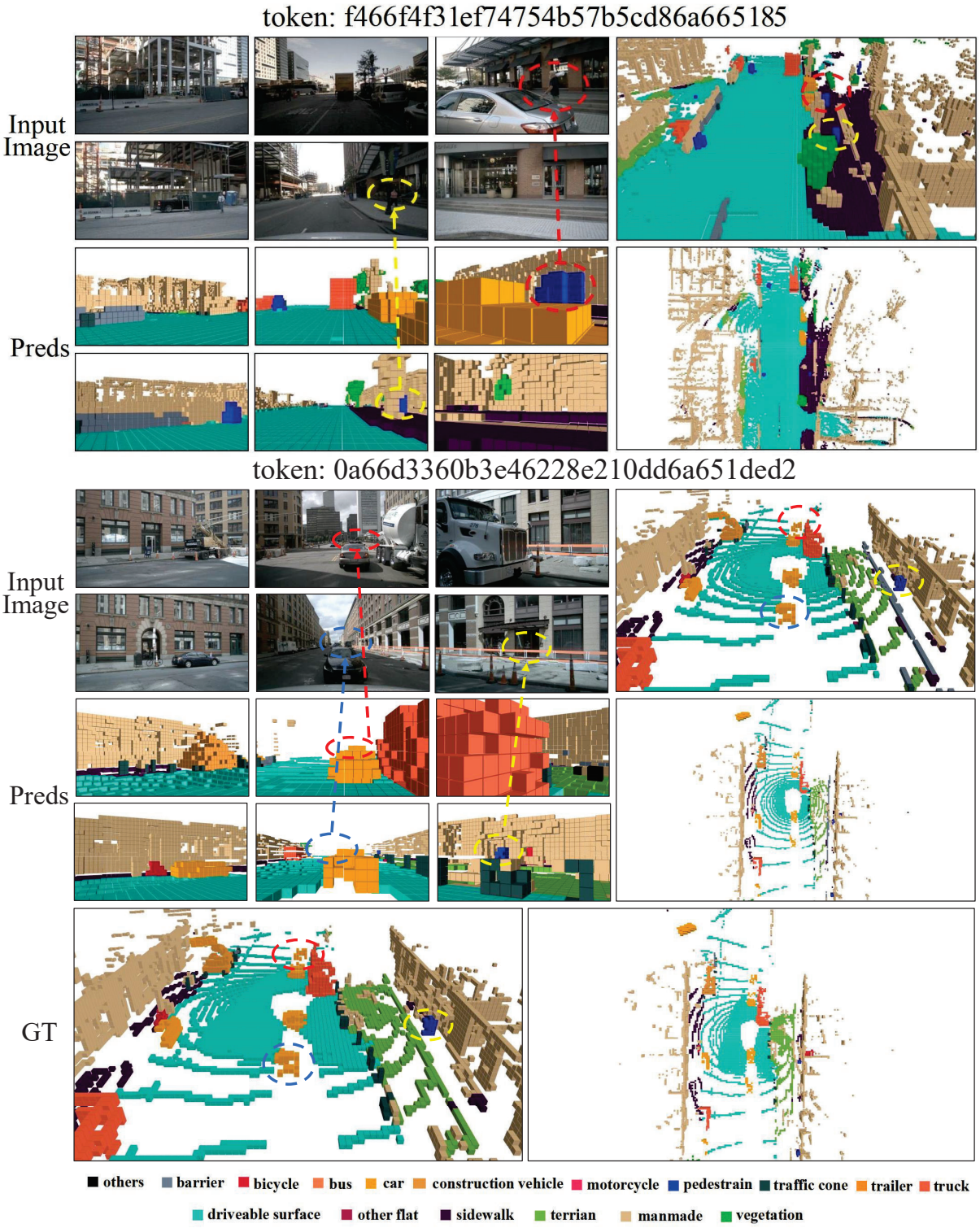

Figure 8. Visualizations for occlusion on OCC3D-nuScenes validation set. For each scene, the six images in the *"Input Images"* line left are the inputs to our model captured by font-left, front, front-right, back-right, back, and back-right cameras. The six images in the *"Preds"* line left denote our 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.

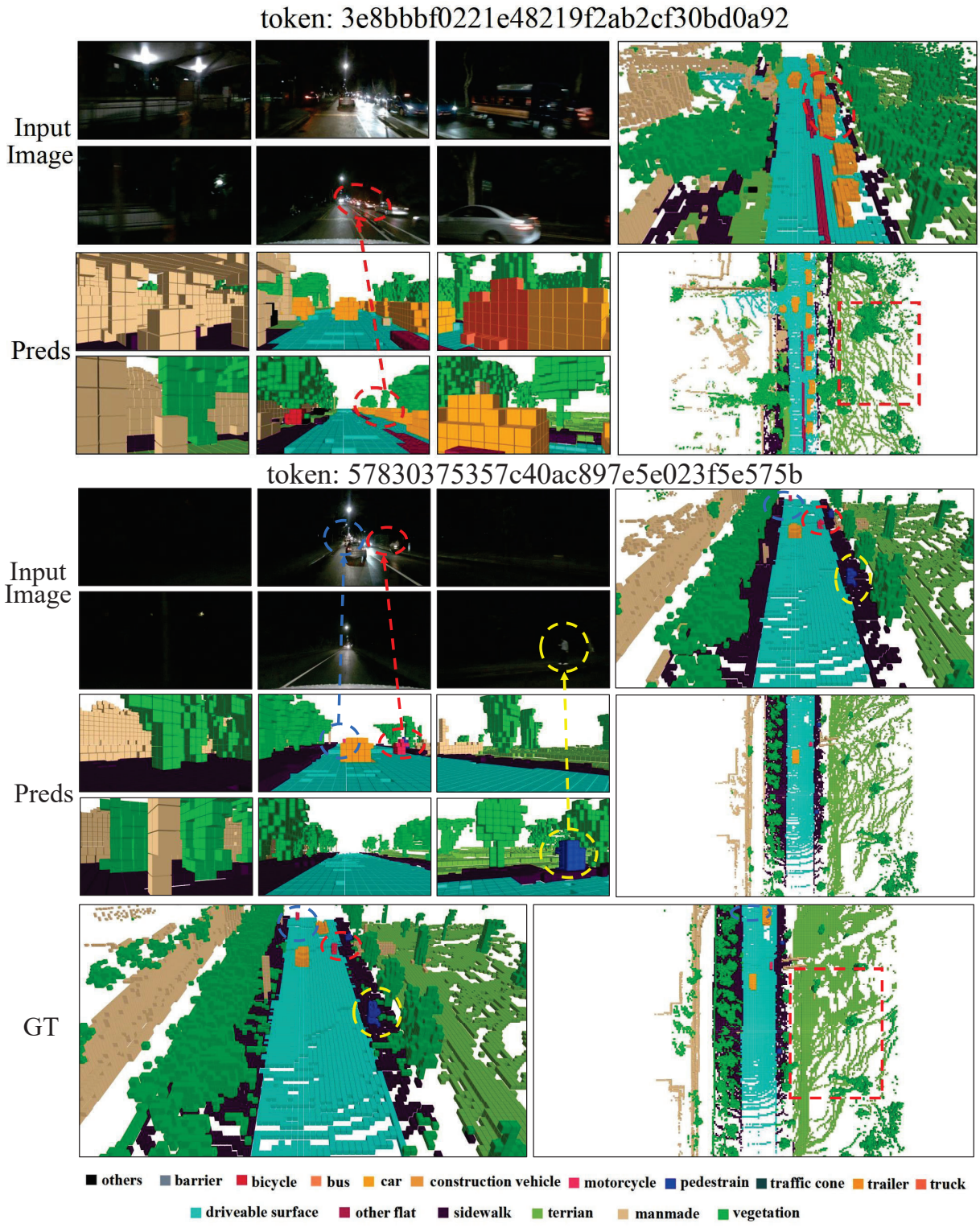

Figure 9. Visualizations for low-light environments on OCC3D-nuScenes validation set.