SOAP: Vision-Centric 3D Semantic Scene Completion with Scene-Adaptive Decoder and Occluded Region-Aware View Projection

Supplementary Material

S-1. More Ablation Studies for Occluded Region-Aware Projection

We introduce the occluded region-aware view projection (OAP), which refines features in occluded regions through 3D deformable cross-attention between the initial voxel features $\tilde{\mathbf{F}}_i^{3D}$ and the historical voxel features $\mathbf{F}_i^{hist}$. To validate the effectiveness of OAP, we design several approaches for the view projection as in Figure S-1.

Method 1. The first method splits the voxel space into two regions: invisible and the other regions as in Figure S-1(a). Then, it fills the invisible regions with the historical voxel features, while the other regions are filled with the sum of the initial voxel features and the historical voxel features.

Method 2. The second method also splits voxel space into invisible and the other regions as in Figure S-1(b). Then, it fills the invisible regions with the historical voxel features, while features of the other regions are refined through 3D deformable cross attention (DCA) between the initial voxel features and the historical voxel features.

Method 3. Similar to the proposed OAP, the third method first identifies the occluded regions within the initial voxel features. It then splits the other remaining regions into invisible and visible regions as in Figure S-1(c). It reconstructs invisible voxel features $\mathbf{F}_i^{inv}$ and visible voxel features $\mathbf{F}_i^{vis}$ following Equation (4) and Equation (5), respectively, as done in OAP. In contrast, it reconstructs occluded voxel features $\mathbf{F}_i^{occ}$ by simply adding the initial voxel features and the historical voxel features.

Experimental Results. Table S-1 compares the alternative view projection approaches (Methods 1-3) with the proposed OAP. When comparing Method 1 to Method 2 and Method 3 to OAP, we observe that incorporating deformable cross-attention between the initial and historical voxel features yields improvements in both mIoU and IoU scores, surpassing the performance of simple addition operations. Also, Method 3 provides better performance than Method 1, while Method 2 degrades model performance compared to the proposed OAP. These results underscore the critical importance of identifying occluded regions during view projection for achieving accurate 3D SSC.

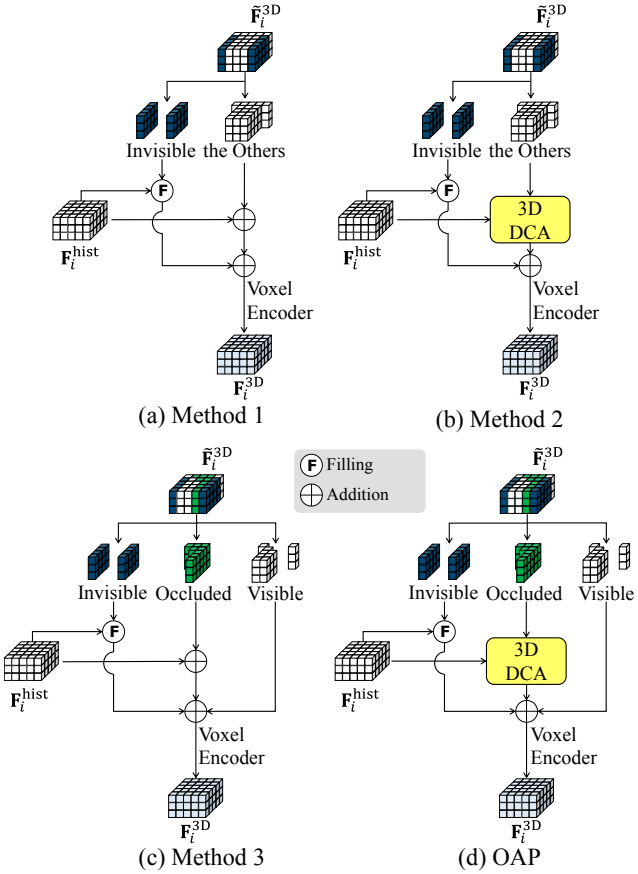

Figure S-1. Alternative approaches for occluded region-aware view projection (OAP).

Table S-1. Ablation study for occluded region-aware projection on the SemanticKITTI validation set.

	SC IoU	SSC mIoU
Method 1	45.11	18.25
Method 2	45.72	18.74
Method 3	45.83	18.94
OAP	**47.24**	**19.21**

S-2. SOAP-ResNet50

We further evaluate the performance of SOAP using ResNet50 [3] as the image encoder on the SemanticKITTI [1] and SSCBench-KITTI360 test datasets. Table S-2 and Table S-3 present quantitative comparisons of the proposed SOAP, employing either EfficientNetB7 [10] or ResNet50 [3], and other state-of-the-art methods on the SemanticKITTI and SSCBench-KITTI360 test datasets, respectively. For clarity, we denote SOAP with the ResNet50 image encoder as SOAP-ResNet50. Notably, SOAP-ResNet50 achieves the highest IoU and mIoU scores com-

Table S-2. Semantic scene completion results on SemanticKITTI test set. † represents methods that use temporal information. We categorize all methods based on their image encoders. We highlight the best results in **bold** and the second best results in underline.

Method	SC IoU	SSC mIoU	road (15.30%)	sidewalk (11.13%)	parking (1.12%)	other-ground (0.56%)	building (14.1%)	car (3.92%)	truck (0.16%)	bicycle (0.03%)	motorcycle (0.03%)	other-veh. (0.20%)	vegetation(39.3%)	trunk (0.51%)	terrain (9.17%)	person (0.07%)	bicyclist (0.07%)	motorcyclist (0.05%)	fence (3.90%)	pole (0.29%)	traf.-sign (0.08%)
ResNet50																					
VoxFormer† [7]	43.21	13.41	54.1	26.9	25.1	7.3	23.5	21.7	3.6	1.9	1.6	4.1	24.4	8.1	24.2	1.6	1.1	0.0	13.1	6.6	5.7
Symphonies [5]	42.19	15.04	58.4	29.3	26.9	11.7	24.7	23.6	3.2	3.6	2.6	**5.6**	24.2	10.0	23.1	**3.2**	1.9	**2.0**	16.1	7.7	8.0
HASSC† [11]	42.87	14.38	55.3	29.6	25.9	11.3	23.1	23.0	2.9	1.9	1.5	4.9	24.8	9.8	26.5	1.4	**3.0**	0.0	14.3	7.0	7.1
SGN† [8]	45.42	15.76	60.4	31.4	28.9	8.7	28.4	25.4	**4.5**	0.9	1.6	3.7	27.4	12.6	28.4	0.5	0.3	0.1	18.1	10.0	8.3
SOAP-ResNet50†	**47.54**	**18.72**	**63.2**	**36.6**	**35.8**	**16.3**	**30.3**	**28.4**	4.3	**4.6**	**2.9**	4.9	**31.6**	**15.0**	**33.2**	2.3	0.9	0.1	**20.6**	**11.5**	**13.3**
EfficientNetB7																					
TPVFormer [4]	34.25	11.26	55.1	27.2	27.4	6.5	14.8	19.2	3.7	1.0	0.5	2.3	13.9	2.6	20.4	1.1	2.4	0.3	11.0	2.9	1.5
NDC-Scene [13]	33.87	11.55	56.2	28.7	28.0	5.6	15.8	19.7	1.8	1.1	1.1	4.9	14.3	2.6	20.6	0.7	1.7	0.4	11.2	3.2	1.7
SurroundOcc [12]	34.72	11.86	56.9	28.3	30.2	6.8	15.2	20.6	1.4	1.6	1.2	4.4	14.9	3.4	19.3	1.4	2.0	0.1	11.3	3.9	2.4
OccFormer [15]	34.53	12.32	55.9	30.3	31.5	6.5	15.7	21.6	1.2	1.5	1.7	3.2	16.8	3.9	21.3	2.2	1.1	0.2	11.9	3.8	3.7
LowRankOcc [16]	38.47	13.56	52.8	27.2	25.1	8.8	22.1	20.9	2.9	3.3	2.7	4.4	22.9	8.9	20.8	2.4	1.7	**2.3**	14.4	7.0	7.0
HTCL† [6]	44.23	17.09	**64.4**	34.8	33.8	12.4	25.9	27.3	**5.7**	1.8	2.2	5.4	25.3	10.8	31.2	1.1	**3.1**	0.9	21.1	9.0	8.3
SOAP†	**46.09**	**19.09**	63.6	**36.2**	**36.8**	**17.2**	**28.7**	**28.9**	3.3	**5.0**	**5.3**	**7.0**	**29.8**	**15.1**	**32.3**	**3.7**	2.3	0.2	**22.5**	**11.7**	**13.1**

Table S-3. Semantic scene completion results on the SSCBench test set. † represents methods that use temporal information. We categorize all methods based on their image encoders. We highlight the best results in **bold** and the second best results in underline.

Method	SC IoU	SSC mIoU	car	bicycle	motorcycle	truck	other-veh.	person	road	parking	sidewalk	other-ground	building	fence	vegetation	terrain	pole	traf.-sign	other-struct.	other-object
ResNet50																				
VoxFormer† [7]	38.76	11.91	17.8	1.2	0.9	4.6	2.1	1.6	47.0	9.7	27.2	2.9	31.2	5.0	29.0	14.7	6.5	6.9	3.8	2.4
Symphonies [5]	44.12	18.58	**30.0**	1.9	**5.9**	**25.1**	**12.1**	**8.2**	54.9	13.8	32.8	**6.9**	35.1	8.6	38.3	11.5	14.0	9.6	**14.4**	**11.3**
SGN† [8]	47.06	18.25	29.0	**3.4**	2.9	10.9	5.2	3.0	58.1	15.0	36.4	4.4	42.0	7.7	38.2	23.2	16.7	16.4	9.9	5.8
SOAP-ResNet50†	**48.48**	**20.17**	30.0	3.3	4.4	7.8	6.0	5.9	**60.7**	**17.5**	**40.1**	6.3	**45.5**	**10.9**	**40.9**	**24.9**	**17.2**	**19.4**	12.9	9.7
EfficientNetB7																				
TPVFormer [4]	40.22	13.64	21.6	1.1	1.4	8.1	2.6	2.4	53.0	12.0	31.1	3.8	34.8	4.8	30.1	17.5	7.5	5.9	5.5	2.7
OccFormer [15]	40.27	13.81	22.6	0.7	0.3	9.9	3.8	2.8	54.3	13.4	31.5	3.6	36.4	4.8	31.0	19.5	7.8	8.5	7.0	4.6
SOAP†	**48.12**	**20.92**	**29.9**	**5.6**	**7.8**	**14.4**	**7.6**	**6.1**	**60.9**	**17.4**	**40.3**	**5.4**	**45.3**	**10.6**	**40.5**	**24.8**	**16.8**	**21.0**	**12.6**	**9.9**

pared to ResNet50-based methods on both SemanticKITTI and SSCBench-KITTI360. These results demonstrate that SOAP is effective, regardless of the image encoder employed.

S-3. Ablation Studies for Historical Frames

In Table S-4, we analyze the impact of the number of historical frames P on the overall model performance. We observe that SOAP produces similar performance when $P > 1$. We set the number of historical frames to $P = 4$, which provides the best trade-off between performance and efficiency.

S-4. Occluded Region Detection

The proposed SOAP utilizes the occlusion map $\mathbf{O}_i$ to distinguish the occluded regions from the voxel space. Therefore, the accuracy of the occlusion map is critical to overall SSC performance. To evaluate the occlusion detection capability of occlusion map $\mathbf{O}_i$, we establish ground-truth occlusion labels $\hat{\mathbf{O}}_i$, where occluded regions are defined as those lo-

Table S-4. Ablation study on the number of frames P in the historical voxel generation.

P	SC IoU	SSC mIoU
1	46.39	18.80
2	47.15	19.14
3	47.11	19.17
4 (current setting)	47.24	**19.21**
5	**47.29**	**19.21**

Table S-5. Occlusion detection accuracy of $\mathbf{O}_i^{3D}$ on the SemanticKITTI validation set.

IoU	Recall	Precision
75.62	75.90	99.52

cated behind ground-truth depth bins derived from the projected LiDAR point clouds. Table S-5 reports the occlusion recall, precision, and IoU between the predicted occlusion map $\varphi(\mathbf{O}_i)$ and the ground-truth occlusion labels $\varphi(\hat{\mathbf{O}}_i)$ within the voxel space. Remarkably, the occlusion map $\mathbf{O}_i$ achieves a precision of 99.52, which indicates that the most identified occluded regions align with the ground-truth occluded regions. This highlights OAP's ability to refine voxel features, resulting in superior 3D SSC performance. Additionally, the occlusion map $\mathbf{O}_i$ achieves reliable recall and IoU performance.

S-5. Results Using Monocular Depth

We also evaluate the performance of SOAP using the monocular depth model [2], which is employed in other methods [5, 7, 14]. Table S-6 presents the results on the SemanticKITTI validation set. The scores of other methods [5, 7, 14–16] are obtained from their respective papers. We observe that SOAP with monocular depth still outperforms other methods with significant margins, demonstrating its superior adaptability.

S-6. Scalability of OAP

Incorporation OAP into other models: To verify the scalability of the occluded region-aware view projection (OAP), we integrate the proposed OAP into OccFormer and SGN. Table S-7 below reports that OAP significantly enhances the performance, even when SGN inherently uses temporal information.

S-7. Ablation Study for Scene-Adaptive Decoder

Novelty-Based Token Selection. In this study, we introduce the novelty-based token selection, which leverages token scores as $\mathbf{S} = \mathbf{S}^{\text{cls}} + \alpha\mathbf{S}^{\text{nov}}$, where $\mathbf{S}^{\text{nov}}$ is novelty scores and α is its balancing parameter. Table S-8 compares the performance of SOAP with respect to the novelty score $\mathbf{S}^{\text{nov}}$ and its corresponding balancing parameter α. The result without $\mathbf{S}^{\text{nov}}$ means the token scores are computed solely by the classification scores $\mathbf{S} = \mathbf{S}^{\text{cls}}$. The model exhibits similar performance across different values of α, with the best performance observed at $\alpha = 0.2$. Also, token selection without novelty scores results in slight reduction in the performance.

Table S-6. Comparison on the stereo and monocular settings. We replace the stereo depth estimator [9] with the monocular depth model [2] for the monocular setting.

Mehtod	Stereo		Mono	
	IoU	mIoU	IoU	mIoU
VoxFormer [7]	44.15	13.35	38.08	11.27
Symphonies [5]	41.92	14.89	38.37	12.20
SGN [8]	46.21	15.32	41.87	12.91
OccFormer [15]	-	-	36.50	13.46
LowRankOcc [16]	-	-	37.85	14.21
SOAP	**47.24**	**19.21**	**43.64**	**16.62**

Table S-7. The scalability of the proposed OAP.

Method	OccFormer	+ OAP	SGN	+ OAP
IoU	36.50	**46.60**	46.21	**47.07**
mIoU	13.46	**17.18**	15.31	**17.04**

tSNE Visualization of the Semantic Repository. Figure S-2 illustrates the t-SNE visualization of tokens in the semantic repository $\mathbf{R}$ across update iterations, under (a) $\mathbf{S} = \mathbf{S}^{\text{cls}}$ and (b) $\mathbf{S} = \mathbf{S}^{\text{cls}} + \alpha\mathbf{S}^{\text{nov}}$. Different colors represent different semantic classes. The feature space of $\mathbf{R}$ progressively exhibits clear separation among different semantic classes, indicating that the semantic repository contains distinct features for each semantic class. However, we also observe that updating with $\mathbf{S} = \mathbf{S}^{\text{cls}}$ tends to select tokens with similar features for each semantic class, leading to reduced semantic diversity. In contrast, updating with the novelty scores $\mathbf{S} = \mathbf{S}^{\text{cls}} + \alpha\mathbf{S}^{\text{nov}}$ preserves the semantic diversity of the semantic repository $\mathbf{R}$.

Number of Representative Tokens. We also analyze the impacts of the number of representative tokens K on the overall model performance. As shown in Table S-9, SOAP demonstrates stable performance even with smaller K. Even though both IoU and mIoU are improved with the increase of the K, we select $K = 10$ as our current setting, balancing the trade-off between performance and efficiency.

Number of Repetitions. Table S-10 analyze the impact of the repeated number of the scene-adaptive decoder. The results indicate that SOAP achieves robust stability across different configurations for the scene-adaptive decoder. We pick the repeated numbers 9 that provide the best trade-off between performance and efficiency.

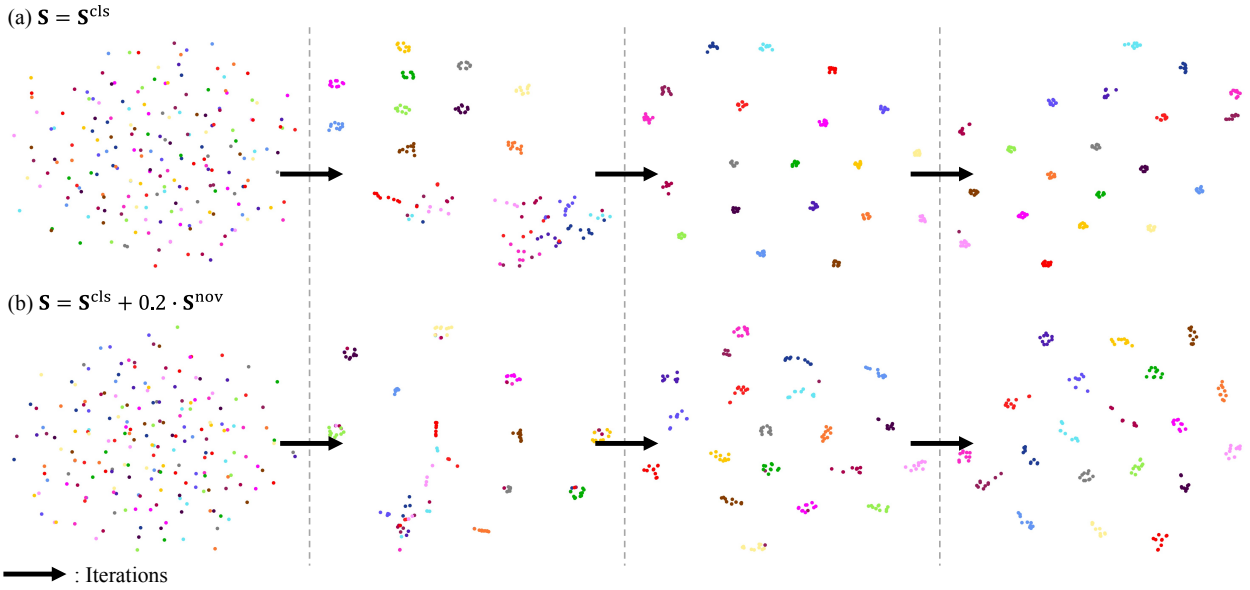

Figure S-2. t-SNE visualization of semantic tokens in the semantic repository $\mathbf{R}$ according to update processes with (a) $\mathbf{S} = \mathbf{S}^{\text{cls}}$ and (b) $\mathbf{S} = \mathbf{S}^{\text{cls}} + \alpha\mathbf{S}^{\text{nov}}$. Different colors represent different semantic classes.

Table S-8. Ablation study for token scores $\mathbf{S}$ on the SemanticKITTI validation set.

Token scores $\mathbf{S}$	α	SC IoU	SSC mIoU
$\mathbf{S}^{\text{cls}}$	-	47.11	19.07
$\mathbf{S}^{\text{cls}} + \alpha\mathbf{S}^{\text{nov}}$	0.1	47.15	19.18
$\mathbf{S}^{\text{cls}} + \alpha\mathbf{S}^{\text{nov}}$	0.2	**47.24**	**19.21**
$\mathbf{S}^{\text{cls}} + \alpha\mathbf{S}^{\text{nov}}$	0.4	47.11	19.19
$\mathbf{S}^{\text{cls}} + \alpha\mathbf{S}^{\text{nov}}$	0.6	47.20	19.13

Table S-9. Ablation study for the number of representative tokens K in the semantic repository on the SemanticKITTI validation set.

K	SC IoU	SSC mIoU
5	47.18	19.08
10 (current setting)	47.24	19.21
15	**47.26**	**19.23**
20	**47.26**	**19.23**

Table S-10. Ablation study for the repeated number of scene-adaptive decoder (SAD) on the SemanticKITTI validation set.

Repeated number of SAD	SC IoU	SSC mIoU
5	47.12	19.17
7	46.93	19.23
9 (current setting)	**47.24**	19.21
11	46.95	**19.35**

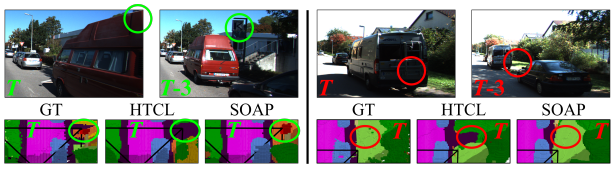

Figure S-3. Visualization for severe occlusion examples.

S-8. Qualitative Results

Figure S-5 and Figure S-6 show qualitative comparisons on the SemanticKITTI validation set and SSCBench test set, respectively. We obtain the results of SGN, HTCL, and MonoScene using source codes and network parameters provided by respective authors. The first and last columns in both Figures are the input images and the ground truth semantic labels, respectively. The second to fourth columns in Figure S-5 visualize the 3D semantic scene completion results from SGN, HTCL, and the proposed SOAP. The second to fourth columns in Figure S-6 are the results from MonoScene, SGN, and the proposed SOAP. We see that SOAP reconstructs 3D scenes more precisely than SGN, HTCL, and MonoScene.

Visualization for severe occlusion examples. Figure S-3 illustrates the qualitative results of SOAP on severe occlusion examples. As occluded regions in the current T frame become visible in a previous $T-3$ frame, SOAP generates reliable 3D SSC predictions even under challenging conditions of severe occlusion.

Regions occluded by moving objects. Figure S-3 illustrates that road regions occluded by a moving object in the

Figure S-4. t-SNE visualization of features for regions occluded by moving objects.

current frame are often visible in previous frames. The right figure presents a t-SNE visualization of features for occluded points (red) and visible points (green) across the frames, along with randomly sampled points (gray). This highlights that occluded regions caused by moving objects can often be refined using features in historical frames.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*, pages 9297–9307, 2019. 1

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, pages 4009–4018, 2021. 3

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[4] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *CVPR*, pages 9223–9232, 2023. 2

[5] Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Tianwei Lin, Wenyu Liu, and Xinggang Wang. Symphonize 3d semantic scene completion with contextual instance queries. In *CVPR*, pages 20258–20267, 2024. 2, 3

[6] Bohan Li, Jiajun Deng, Wenyao Zhang, Zhujin Liang, Dalong Du, Xin Jin, and Wenjun Zeng. Hierarchical temporal context learning for camera-based semantic scene completion. In *ECCV*, pages 131–148. Springer, 2024. 2

[7] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *CVPR*, pages 9087–9098, 2023. 2, 3

[8] Jianbiao Mei, Yu Yang, Mengmeng Wang, Junyu Zhu, Jongwon Ra, Yukai Ma, Laijian Li, and Yong Liu. Camera-based 3d semantic scene completion with sparse guidance network. *IEEE TIP*, 2024. 2, 3

[9] Faranak Shamsafar, Samuel Woerz, Rafia Rahim, and Andreas Zell. Mobilestereonet: Towards lightweight deep networks for stereo matching. In *WACV*, pages 2417–2426, 2022. 3

[10] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019. 1

[11] Song Wang, Jiawei Yu, Wentong Li, Wenyu Liu, Xiaolu Liu, Junbo Chen, and Jianke Zhu. Not all voxels are equal: Hardness-aware semantic scene completion with self-distillation. In *CVPR*, pages 14792–14801, 2024. 2

[12] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *ICCV*, pages 21729–21740, 2023. 2

[13] Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, and Hongsheng Li. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In *ICCV*, pages 9421–9431, 2023. 2

[14] Zhu Yu, Runming Zhang, Jiacheng Ying, Junchen Yu, Xiaohai Hu, Lun Luo, Siyuan Cao, and Huiliang Shen. Context and geometry aware voxel transformer for semantic scene completion. In *NeurIPS*, 2024. 3

[15] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *ICCV*, pages 9433–9443, 2023. 2, 3

[16] Linqing Zhao, Xiuwei Xu, Ziwei Wang, Yunpeng Zhang, Borui Zhang, Wenzhao Zheng, Dalong Du, Jie Zhou, and Jiwen Lu. Lowrankocc: Tensor decomposition and low-rank recovery for vision-based 3d semantic occupancy prediction. In *CVPR*, pages 9806–9815, 2024. 2, 3

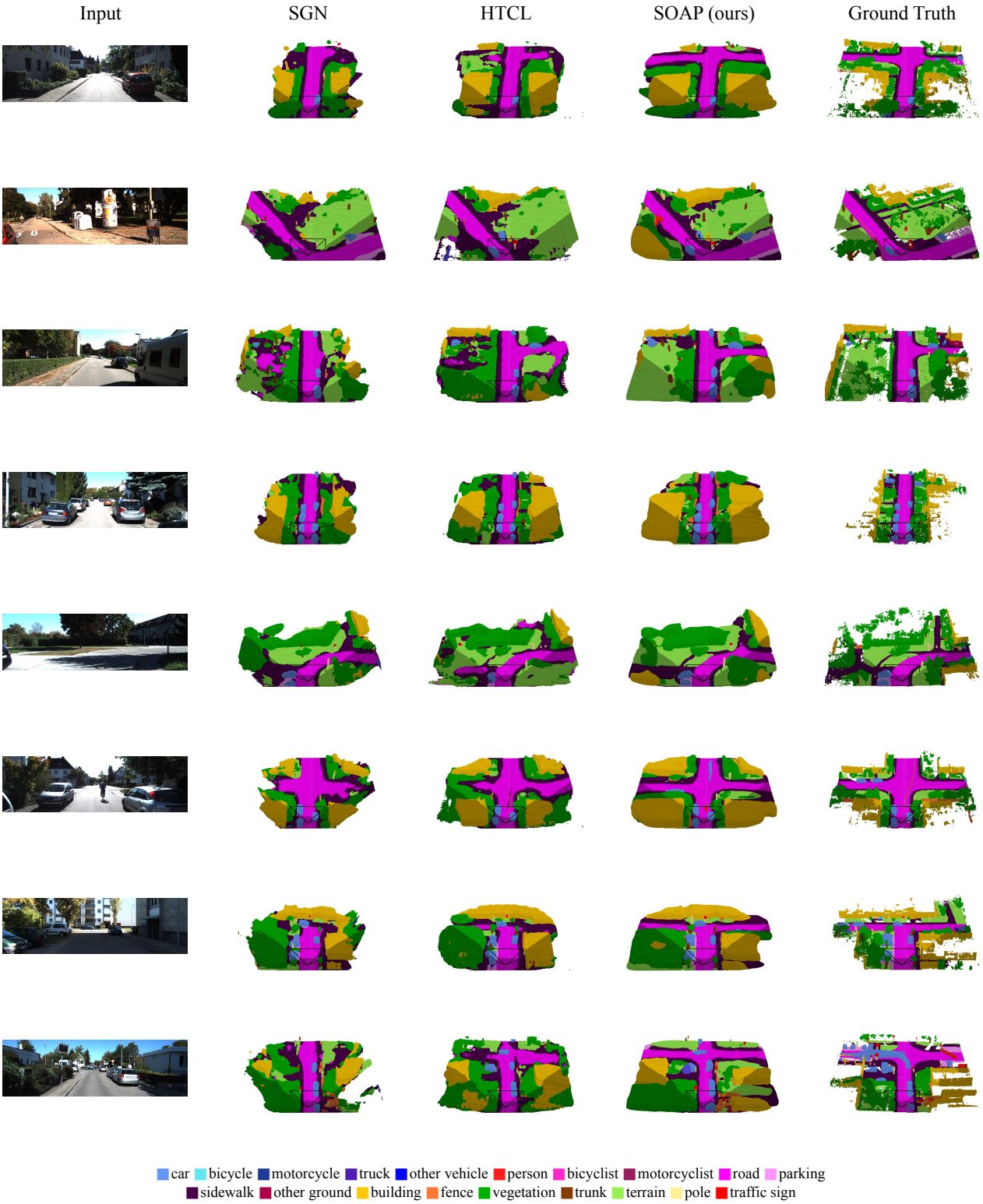

Figure S-5. Qualitative results on SemanticKITTI validation set.

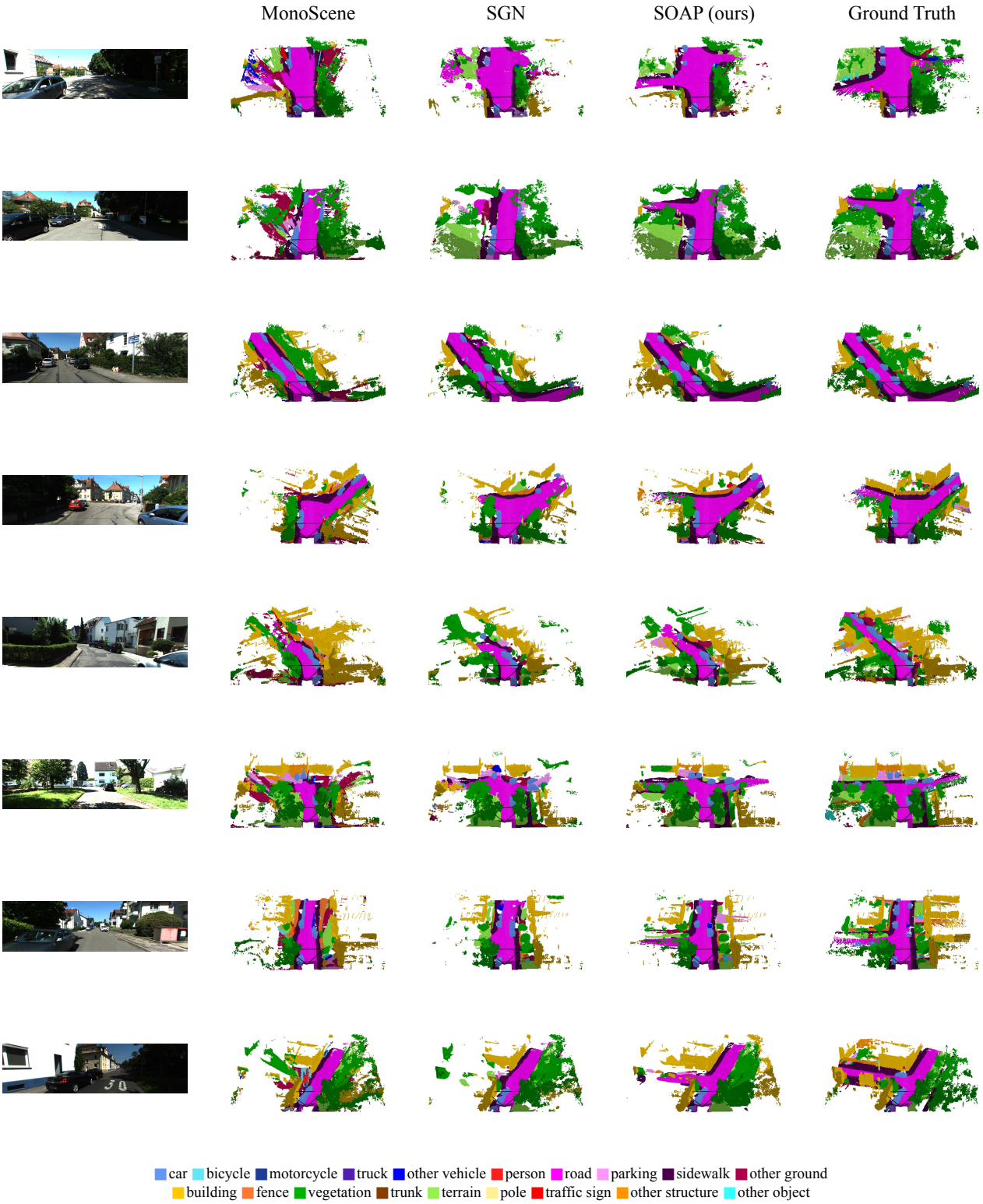

Figure S-6. Qualitative results on KITTI360-SSCBench test set.