Learning Spatial-Temporal Consistency for 3D Semantic Scene Completion Supplementary Material

A. Overview

In the supplementary material, we mainly provide implementation details and more experiment results. And we further analyze the limitations of our method and outline several promising directions for future work.

B. Implementation Details

Metrics. Following standard practice in Semantic Scene Completion [1, 9, 17], we evaluate geometric accuracy using the intersection over union (IoU) and measure semantic quality with the mean IoU (mIoU). IoU reflects how well the occupied voxel structure is reconstructed, while mIoU aggregates class-wise to quantify semantic consistency. The mIoU is calculated by:

$$mIoU = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TN_c + FP_c + FN_c} \tag{1}$$

where TP_c, TN_c, FP_c, and FN_c are the true positives, ture nagatives, false positives and false negatives predictions for class c. Together, these two metrics provide a balanced assessment of both shape completion and semantic prediction, revealing how improvements in geometry and semantics jointly influence overall performance.

Coarse Occupancy Generation Module. As depicted in figure 1, the depth map predicted by MobileStereoNet is the geometric source of the depth probability. We first estimate continuous per-pixel depth, which is then discretized into depth bins and smoothed to form a soft depth distribution. This probabilistic representation enables robust depth-order and occlusion reasoning along camera rays, which cannot be achieved using raw depth values alone.

2D-to-3D Projection Module. Our 2D-to-3D projection module lifts image features into a voxel grid while injecting depth-aware geometric cues. Given temporally aggregated feature F_{agg} and the current frame depth map, the module outputs a fused voxel representation V_{agg}. Specifically, The F_{agg} is first reshaped to the target voxel V_{agg} resolution using a lightweight convolution and upsampling block, ensuring spatial alignment between image-plane features and the voxel grid. The current frame depth map are encoded through a compact convolutional module that produces depth-wise weights, each corresponding to one slice of the voxel depth dimension. These weights act as geometry-aware guidance during lifting. The adapted F_{agg} are expanded along the depth axis to size $B \times 192 \times 128 \times 128 \times 16$. Depth features are broadcast and used as multiplicative weights, allowing each depth plane to modulate image features. Finally, a 3D convolution refines the fused volume, enforcing local spatial consistency and producing a coherent voxel representation.

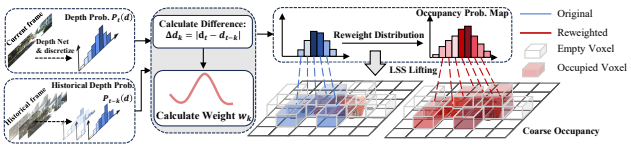

Figure 1. The illustration of Coarse Occupancy Generation.

C. Additional Quantitative Results

C.1. Results on SemanticKITTI Validation Set

Table 1 summarizes the quantitative results on the SemanticKITTI Validation Set. Our method achieves the highest overall performance, reaching 49.10 IoU and 19.10 mIoU, clearly outperforming all existing baselines. Compared with strong counterparts such as CGFormer and HTCL-S, our model delivers substantial gains in both IoU and mIoU, validating the effectiveness of our Spatial-Temporal Consistency design. Beyond the overall metrics, our approach ranks first or second across most semantic categories, showing particularly strong performance on static classes such as road and building. This improvement stems from our coarse multi-view occupancy reasoning, which compensates for missing structural cues in these regions. The consistent advantages across diverse categories highlight the capability of our framework to capture fine-grained geometry while maintaining robust scene-level understanding.

C.2. Results with Monocular Depth

Table 2 summarizes the comparison against leading approaches under both stereo and mono depth settings. The

Methods	IoU	road (15.30%)	sidewalk (11.13%)	parking (1.12%)	other-grnd (0.56%)	building (14.10%)	car (3.92%)	truck (0.16%)	bicycle (0.03%)	motocycle (0.03%)	other-vehicle (0.20%)	vegetation (39.3%)	trunk (0.51%)	terrain (9.17%)	person (0.07%)	bicylist (0.07%)	motorcyclist (0.05%)	fence (3.90%)	pole (0.29%)	traf.-sign (0.08%)	mIoU
LMSCNet[11]†	28.61	40.68	18.22	4.38	0.00	10.31	18.33	0.00	0.00	0.00	0.00	13.66	0.02	20.54	0.00	0.00	0.00	1.21	0.00	0.00	6.70
AICNet[8]†	29.59	43.55	20.55	11.97	0.07	12.94	14.71	4.53	0.00	0.00	0.00	15.37	2.90	28.71	0.00	0.00	0.00	2.52	0.06	0.00	8.31
JS3C-Net[15]†	38.98	50.49	23.74	11.94	0.07	15.03	24.65	4.41	0.00	0.00	6.15	18.11	4.33	26.86	0.67	0.27	0.00	3.94	3.77	1.45	10.31
MonoScene[1]	37.12	57.47	27.05	15.72	0.87	14.24	23.55	7.83	0.20	0.77	3.59	18.12	2.57	30.76	1.79	1.03	0.00	6.39	4.11	2.48	11.50
TPVFormer[3]	35.61	56.50	25.87	20.60	0.85	13.88	23.81	8.08	0.36	0.05	4.35	16.92	2.26	30.38	0.51	0.89	0.00	5.94	3.14	1.52	11.36
OccFormer[17]	36.50	58.85	26.88	19.61	0.31	14.40	25.09	**25.53**	0.81	1.19	8.52	19.63	3.93	32.62	2.78	2.82	0.00	5.61	4.26	2.86	13.46
VoxFormer-T[9]	44.15	53.57	26.52	19.69	0.42	19.54	26.54	7.26	1.28	0.56	7.81	26.10	6.10	33.06	1.93	1.97	0.00	7.31	9.15	4.94	13.35
HASSC-T[13]	44.58	57.23	29.08	19.89	1.26	20.19	27.33	17.06	1.07	1.14	8.83	27.01	7.71	33.95	2.25	**4.09**	0.00	7.95	9.20	4.81	14.74
Symphonies[4]	41.92	56.37	27.58	15.28	0.95	21.64	28.68	20.44	2.54	2.82	**13.89**	25.72	6.60	30.87	**3.52**	2.24	0.00	8.40	9.57	5.76	14.89
H2GFormer-T[14]	44.69	57.00	29.37	21.74	0.34	20.51	14.29	6.80	0.95	0.91	9.32	27.44	7.80	36.26	1.15	0.10	0.00	7.98	9.88	5.81	14.29
BRGScene[6]	43.34	61.90	31.20	**30.70**	**10.70**	24.20	22.80	2.80	3.40	2.40	6.10	23.80	8.40	27.00	2.90	2.20	**0.50**	**16.50**	7.00	7.20	15.36
CGFormer[16]	45.99	65.51	32.31	20.82	0.16	23.52	34.32	19.44	**4.61**	2.71	7.67	26.93	8.83	39.54	2.38	4.08	0.00	9.20	10.67	7.84	16.87
HTCL-S[7]	45.51	63.70	32.48	23.27	0.14	24.13	34.30	20.72	3.99	**2.80**	11.99	26.96	8.79	37.73	2.56	2.70	0.00	11.22	11.49	6.95	17.13
Ours	**49.10**	**67.57**	**35.69**	24.28	1.42	**28.76**	**34.85**	25.23	2.09	1.47	11.88	**32.99**	**13.8**	**43.49**	3.39	2.74	0.00	9.15	**14.01**	**10.06**	**19.10**

Table 1. Quantitative results on the SemanticKITTI validation set. † represents the results obtained when these methods use RGB inputs, which are implemented and reported in MonoScene [1]. The best results are in **Bold**.

reported numbers of previous methods are taken from their original papers. Our method consistently achieves the highest IoU and mIoU in all settings. In the stereo-depth regime, our approach improves upon the strongest single frame method CGFormer [16] by +2.89 IoU and +2.23 mIoU, indicating a more reliable geometric reconstruction. When settting to monocular depth known to be more challenging due to scale ambiguity, our model still surpasses all competitors with substantial margins. These improvements confirm that our method generalizes well across varying depth qualities and remains highly competitive in both settings.

Method	Stereo Depth		Mono Depth	
	IoU(%)	mIoU(%)	IoU(%)	mIoU(%)
VoxFormer	44.15	13.35	38.08	11.27
OccFormer	-	-	36.50	13.46
Symphonize	41.92	14.89	38.37	12.20
SGN	46.21	15.32	41.87	12.91
CGFormer	45.99	16.87	41.82	14.06
Ours	49.10	19.10	44.26	16.67

Table 2. Comparison on the stereo and monocular settings on the SemanticKITTI validation set.

C.3. Ablation Study for Backbone Networks

Table 3 presents the effect of different backbone networks on semantic occupancy performance. Using ResNet50 [2], our method achieves the highest results among all ResNet50-based models, reaching 49.10 IoU and 19.10 mIoU, outperforming Symphonies [4], HASSC [13], SGN [10], and SOAP [5]. When switching to the more powerful EfficientNet-B7 [12] encoder, our approach again delivers the best performance (48.28 IoU and 19.58 mIoU), surpassing all current state-of-the-art methods. These results demonstrate that our method consistently enhances semantic occupancy prediction across diverse backbone architectures, exhibiting strong generalization and robustness regardless of the underlying image encoder.

Method	Image encoder	IoU(%)↑	mIoU(%)↑
Symphonies	ResNet50	41.92	14.89
HASSC	ResNet50	44.58	14.74
SGN	ResNet50	46.21	15.32
SOAP	ResNet50	48.12	18.80
Ours	ResNet50	**49.10**	**19.10**
OccFormer	EfficientNetB7	36.50	13.46
HTCL-S	EfficientNetB7	45.51	17.13
CGFormer	EfficientNetB7	45.99	16.87
SOAP	EfficientNetB7	47.24	19.21
Ours	EfficientNetB7	**48.28**	**19.58**

Table 3. Ablation Study for Backbone Networks on the SemanticKITTI validation set.

D. Additional Qualitative Results

D.1. More Analysis of the Occupancy Change

We further visualize the occupancy differences before and after applying our coarse occupancy module, as shown in Fig. 2. The resulting change map highlights the regions updated by our coarse occupancy generation, showing that incorporating additional coarse occupancy provides richer structural cues for semantic prediction. In addition, we compare our results with those of HTCL-S. The comparison demonstrates that introducing coarse occupancy yields noticeably more complete and structurally coherent semantic reconstructions.

D.2. More Comparison with Other Methods

In Fig. 3, we present additional qualitative comparisons against the state-of-the-art methods VoxFormer-T and HTCL-S. Our approach consistently delivers more coherent global structures and reconstructs fine-grained details. In addition, the predictions exhibit improved environmental consistency and stronger semantic awareness, leading to more reliable and visually plausible occupancy results.

E. Limitations and Future Work

Despite the notable improvements of the proposed method over existing approaches, several limitations remain. As indicated by the parameter–performance analysis, adopting a larger backbone (e.g., replacing it with a EfficientNetB7 network) can yield substantial performance gains, but at the cost of significantly increased FLOPs and parameter counts. Designing more efficient architecture strategy that scale to higher input resolutions without incurring excessive computational overhead would further enhance performance.

Finally, although our method has been evaluated on widely used benchmark datasets, real-world deployment presents additional challenges, such as sensor synchronization errors and dynamic lighting conditions. Extending our experiments to more diverse real-world datasets and improving robustness under open-set conditions—as well as implementing the method within multi-view perception pipelines—represent promising directions for future research.

References

[1] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 3991–4001, 2022. 1, 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on CVPR*, pages 770–778, 2016. 2

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

[4] 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 *Proceedings of the IEEE/CVF Conference on CVPR*, pages 20258–20267, 2024. 2

[5] Hyo-Jun Lee, Yeong Jun Koh, Hanul Kim, Hyunseop Kim, Yonguk Lee, and Jinu Lee. Soap: Vision-centric 3d semantic scene completion with scene-adaptive decoder and occluded region-aware view projection. In *Proceedings of the CVPR Conference*, pages 17145–17154, 2025. 2

[6] Bohan Li, Yasheng Sun, Zhujin Liang, Dalong Du, Zhuanghui Zhang, Xiaofeng Wang, Yunnan Wang, Xin Jin, and Wenjun Zeng. Bridging stereo geometry and bev representation with reliable mutual interaction for semantic scene completion. *arXiv preprint arXiv:2303.13959*, 2023. 2

[7] 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 *European Conference on Computer Vision*, pages 131–148. Springer, 2024. 2

[8] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 3351–3359, 2020. 2

[9] 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 *Proceedings of the IEEE/CVF conference on CVPR*, pages 9087–9098, 2023. 1, 2

[10] 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 Transactions on Image Processing*, 2024. 2

[11] Luis Roldao, Raoul De Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pages 111–119. IEEE, 2020. 2

[12] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 2

[13] 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 *Proceedings of the IEEE/CVF Conference on CVPR*, pages 14792–14801, 2024. 2

[14] Yu Wang and Chao Tong. H2gformer: Horizontal-to-global voxel transformer for 3d semantic scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5722–5730, 2024. 2

[15] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3101–3109, 2021. 2

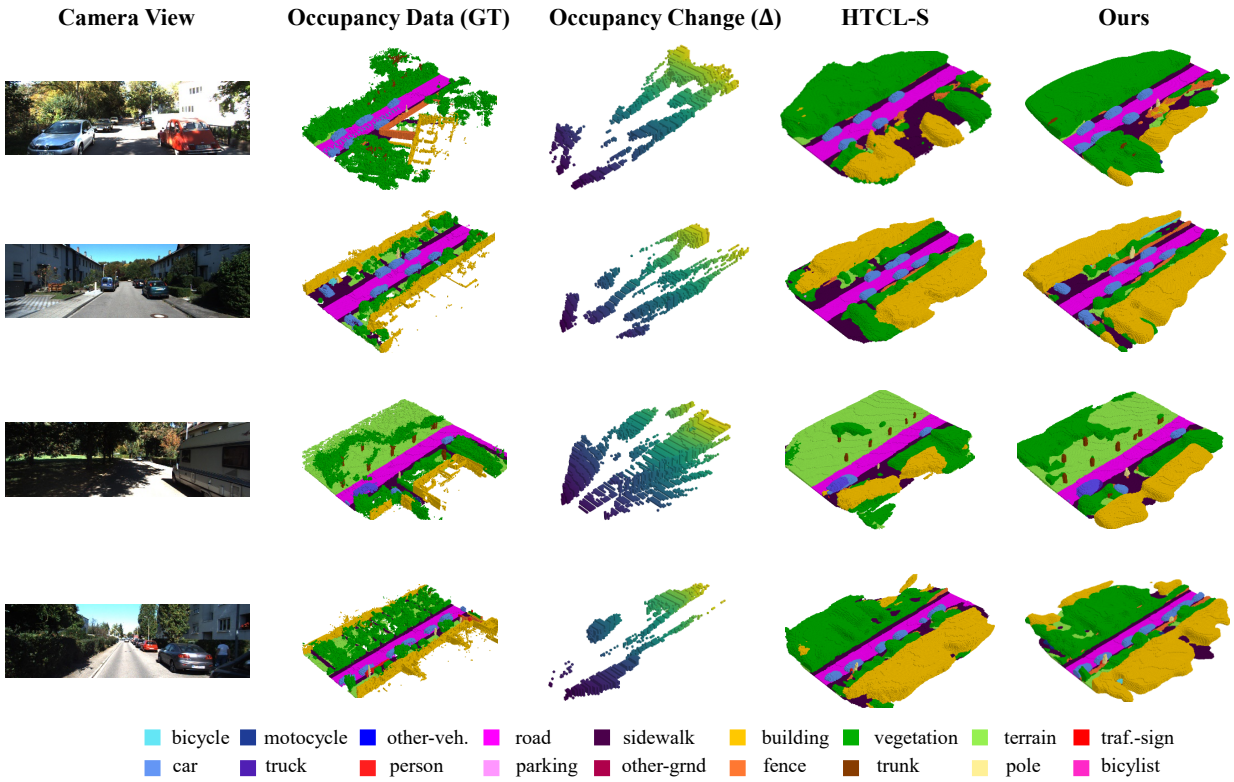

Figure 2. Comparison between the ground truth occupancy and coarse occupancy changes. The change map shows the difference before and after applying our coarse occupancy generation.

[16] Zhu Yu, Runmin Zhang, Jiacheng Ying, Junchen Yu, Xiaohai Hu, Lun Luo, Si-Yuan Cao, and Hui-Liang Shen. Context and geometry aware voxel transformer for semantic scene completion. *Advances in Neural Information Processing Systems*, 37:1531–1555, 2024. 2

[17] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9433–9443, 2023. 1, 2

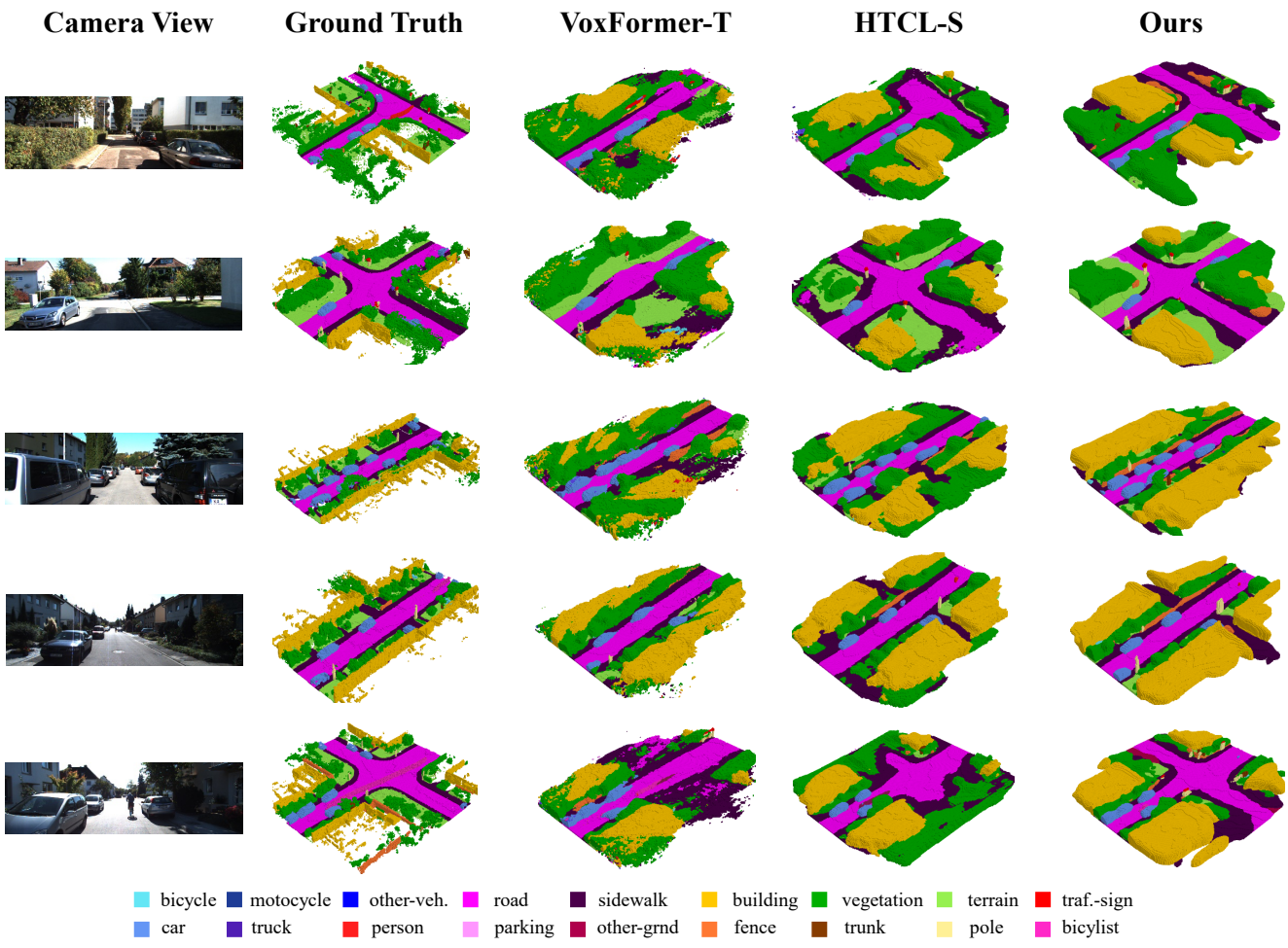

Figure 3. More qualitative comparisons on SemanticKITTI validation set.