Not All Voxels Are Equal: Hardness-Aware Semantic Scene Completion with Self-Distillation

Supplementary Material

In this supplemental document, we further provide the following descriptions and experiments:

- Section A: More implementation details;
- Section B: Additional experiments;
- Section C: Further discussions.

A. More Implementation Details

VoxFormer-based Implementation. Following the original settings in VoxFormer [10], we adopt monocular images from the left camera and estimated coarse geometry from stereo images as inputs. The commonly used ResNet50 backbone [5] is employed to extract image features. The 2D-to-3D transformation module consists of three deformable cross-attention layers, while the 3D backbone comprises two deformable self-attention layers. The resolution ($X' \times Y' \times Z'$) of 3D fine-grained feature ($\mathcal{F}^{\text{3D}}_{\text{fine}}$) is set to $128 \times 128 \times 16$ and the feature dimension D is set to 128. We train both **HASSC**-VoxFormer-S and **HASSC**-VoxFormer-T for 24 epochs with a learning rate of 2×10^{-4} and a batch size of one sample per GPU.

StereoScene-based Implementation. Our **HASSC**-StereoScene employs stereo images from both the left and right cameras as its inputs. We adopt EfficientNet-B7 [15] as the image encoder for fair comparison with StereoScene [7]. The 2D-to-3D transformation consists of BEV constructor and stereo constructor with an auxiliary sparse depth supervision from LiDAR point clouds like BEVDepth [9]. Following MonoScene [2], an encoder-decoder 3D U-Net is used as 3D backbone to process voxel volume features. The resolution of 3D feature map is set to $128 \times 128 \times 16$. Both StereoScene† and **HASSC**-StereoScene in our main paper are trained with 30 epochs, where the learning rate is set to 1×10^{-4} and the batch size is set to 1 per GPU.

B. Additional Experiments

B.1. Quantitative Comparison

Comparison with LiDAR-based Methods. We provide a comparison with existing LiDAR-based methods at different ranges. As shown in Tab. A1, **HASSC**-VoxFormer-T exhibits superior performance, even surpassing several LiDAR-based methods including LMSCNet [13] and SSCNet [14] (24.10%mIoU *v.s.* 22.37%mIoU / 20.02%mIoU) at short range. This outcome underscores the effectiveness of our approach in improving model accuracy.

Methods	Modality	IoU (%)↑			mIoU (%)↑		
		S	M	L	S	M	L
JS3CNet* [17]	LiDAR	63.47	63.40	53.09	**30.55**	**28.12**	**22.67**
LMSCNet* [13]	LiDAR	**74.88**	**69.45**	**55.22**	22.37	21.50	17.19
SSCNet* [14]	LiDAR	64.37	61.02	50.22	20.02	19.68	16.35
HASSC-VoxFormer-T	Camera	**66.05**	**58.01**	**44.58**	**24.10**	**20.27**	**14.74**
VoxFormer-T [10]	Camera	65.38	57.69	44.15	21.55	18.42	13.35
TPVFormer [6]	Camera	54.75	46.03	35.62	17.15	15.27	11.30
MonoScene* [2]	Camera	38.42	38.55	36.80	12.25	12.22	11.30

Table A1. Quantitative comparison with the existing LiDAR-based semantic scene completion methods. $*$ denotes that the results are reported in VoxFormer [10].

α	β	IoU (%)↑	mIoU (%)↑
0.1	1.0	**44.66**	14.40
0.2	1.0	44.58	**14.74**
0.4	1.0	44.46	14.34
0.2	0.8	44.64	14.71
0.2	1.2	44.48	14.47

Table A2. Ablation study on the coefficients (α, β) of linear transformation from $\mathcal{A}$ to $\mathcal{H}^{\text{local}}$.

t	ω	IoU (%)↑	mIoU (%)↑
1	0.75	44.52	14.25
3	0.75	44.58	**14.74**
5	0.75	**44.61**	14.57
10	0.75	44.33	14.04
3	0.5	44.46	14.61
3	1.0	44.51	14.69
1	1.0	44.53	14.50

Table A3. Ablation study on the hyper-parameters (t, ω) for hard voxel selection strategy.

More Ablation Studies. Firstly, we conduct ablation experiments on the coefficients (α, β) of linear transformation from $\mathcal{A}$ to $\mathcal{H}^{\text{local}}$. The results, as shown in Table A2, indicate that the value of α has a greater impact on the completion accuracy than β. We set $\alpha = 0.2$ and $\beta = 1.0$ empirically.

Then, the ablations on the hyper-parameters (t, ω) for hard voxel selection are provided in Tab. A3. The settings of t and ω are crucial for selecting hard voxels to ensure diversity, which can prevent the model from over-fitting in local areas as described in Sec. 3.2 of main paper. As illustrated in Tab. A3, appropriately expanding the scope of hard voxel selection is beneficial to improving model accuracy.

Comparison on Detailed Semantic Categories on Official Benchmark. The detailed semantic categories comparison on the *test set* of SemanticKITTI [1] is illustrated in Tab. A4. **HASSC**-VoxFormer-T outperforms all the camera-based methods. Notably, **HASSC**-VoxFormer-S shows obvi-

Methods	SC IoU (%)	SSC car (3.92%)	bicycle (0.03%)	motorcycle (0.03%)	truck (0.16%)	other-veh. (0.20%)	person (0.07%)	bicyclist (0.07%)	motorcyclist (0.05%)	road (15.30%)	parking (1.12%)	sidewalk (11.13%)	other-grnd (0.56%)	building (14.1%)	fence (3.90%)	vegetation (39.3%)	trunk (0.51%)	terrain (9.17%)	pole (0.29%)	traf.-sign (0.08%)	mIoU (%)
LMSCNet* [3DV20] [13]	31.38	14.30	0.00	0.00	0.30	0.00	0.00	0.00	0.00	46.70	13.50	19.50	3.10	10.30	5.40	10.80	0.00	10.40	0.00	0.00	7.07
3DSketch* [CVPR20] [3]	26.85	17.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	37.70	0.00	19.80	0.00	12.10	3.40	12.10	0.00	16.10	0.00	0.00	6.23
AICNet* [CVPR20] [8]	23.93	15.30	0.00	0.00	0.70	0.00	0.00	0.00	0.00	39.30	19.80	18.30	1.60	9.60	5.00	9.60	1.90	13.50	0.10	0.00	7.09
JS3C-Net* [AAAI21] [17]	34.00	20.10	0.00	0.00	0.80	4.10	0.00	0.20	0.20	47.30	19.90	21.70	2.80	12.70	8.70	14.20	3.10	12.40	1.90	0.30	8.97
MonoScene [CVPR22] [2]	34.16	18.80	0.50	0.70	3.30	4.40	1.00	1.40	0.40	54.70	24.80	27.10	5.70	14.40	11.10	14.90	2.40	19.50	3.30	2.10	11.08
TPVFormer [CVPR23] [6]	34.25	19.20	1.00	0.50	3.70	2.30	1.10	2.40	0.30	55.10	27.40	27.20	6.50	14.80	11.00	13.90	2.60	20.40	2.90	1.50	11.26
OccFormer [ICCV23] [19]	34.53	21.60	1.50	1.70	1.20	3.20	2.20	1.10	0.20	55.90	**31.50**	**30.30**	6.50	15.70	11.90	16.80	3.90	21.30	3.80	3.70	12.32
NDC-Scene [ICCV23] [18]	36.19	19.13	**1.93**	**2.07**	**4.77**	**6.69**	**3.44**	2.77	**1.64**	**58.12**	25.31	28.05	6.53	14.90	12.85	17.94	3.49	25.01	4.43	2.96	12.58
VoxFormer-S [CVPR23] [10]	42.95	20.80	1.00	0.70	3.50	3.70	1.40	2.60	0.20	53.90	21.10	25.30	5.60	19.80	11.10	22.40	7.50	21.30	5.10	4.90	12.20
HASSC-VoxFormer-S	**43.40**	22.80	1.60	1.00	4.70	3.90	1.60	**4.00**	0.30	54.60	23.80	27.70	6.20	21.10	13.10	23.80	8.50	23.30	5.80	5.50	13.34
VoxFormer-T [CVPR23] [10]	43.21	21.70	1.90	1.60	3.60	4.10	1.60	1.10	0.00	54.10	25.10	26.90	7.30	**23.50**	13.10	24.40	8.10	24.20	6.60	5.70	13.41
HASSC-VoxFormer-T	42.87	**23.00**	1.90	1.50	2.90	4.90	1.40	3.00	0.00	55.30	25.90	29.60	**11.30**	23.10	**14.30**	**24.80**	**9.80**	**26.50**	**7.00**	**7.10**	**14.38**

Table A4. Performance comparisons of detailed semantic categories with the state-of-the-art camera-based methods on the ***hidden test set*** of SemanticKITTI [1]. $*$ denotes that the method is converted to the camera-based model by MonoScene [2]. The top two with the highest accuracy are highlighted in **bold** and underlined, respectively.

Methods	SC IoU (%)	SSC car	bicycle	motorcycle	truck	other-veh.	person	bicyclist	motorcyclist	road	parking	sidewalk	other-grnd	building	fence	vegetation	trunk	terrain	pole	traf.-sign	mIoU (%)
VoxFormer-T	44.15	26.54	**1.28**	0.56	7.26	7.81	1.93	1.97	0.00	53.57	19.69	26.52	0.42	19.54	7.31	26.10	6.10	33.06	9.15	4.94	13.35
HASSC-VoxFormer-T*	44.12	26.63	0.26	0.57	6.63	8.45	**2.67**	2.47	0.00	56.57	**20.81**	29.04	0.77	19.89	7.87	**27.08**	7.51	**34.49**	**9.42**	**5.55**	14.03
HASSC-VoxFormer-T	**44.58**	**27.33**	1.07	**1.14**	**17.06**	**8.83**	2.25	**4.09**	0.00	**57.23**	19.89	**29.08**	**1.26**	**20.19**	**7.95**	27.01	**7.71**	33.95	9.20	4.81	**14.74**

Table A5. Performance comparisons of detailed semantic categories for ablation on the ***validation set*** of SemanticKITTI. $*$ denotes no self-distillation during training.

Methods	SC IoU (%)	SSC person	rider	car	trunk	plants	traf.-sign	pole	building	fence	bike	ground	mIoU (%)
SSCNet† [CVPR17] [14]	40.82	10.13	0.10	1.68	2.72	31.80	1.36	6.24	32.95	4.41	15.23	25.84	12.04
HASSC-SSCNet	**42.66**	8.86	**0.17**	1.01	**3.82**	**34.22**	0.63	6.12	**35.03**	**5.57**	**17.04**	25.84	**12.57**
LMSCNet† [3DV20] [13]	51.97	5.60	0.00	2.03	2.66	38.00	2.21	5.91	40.05	13.52	28.14	43.93	16.55
HASSC-LMSCNet	**52.38**	**7.53**	**0.36**	0.66	**3.07**	**39.77**	**2.83**	3.37	**40.47**	**15.29**	**32.20**	**46.57**	**17.47**

Table A6. Quantitative comparisons with the LiDAR-based baseline models on the ***validation set*** of SemanticPOSS [12]. We only use the hard voxel mining (HVM) head **without** self-training strategy. † denotes the results are reproduced from the original implementation. The improved results compared to the corresponding baselines are marked in **blue**.

ous improvements in all semantic categories comparing to VoxFormer-S. We also present class-wise mIoU results on the *validation set* for ablation. As shown in Tab. A5, the hard voxel mining yields obvious improvements on *road* (+3.00% mIoU) and *sidewalk* (+2.52% mIoU) categories, which possess substantial volumes and regular geometries. The self-distillation combined with HVM head further improves the accuracy on *truck*/*car*/*bicyclist* benefiting from the consistency in geometry.

More Experiments on Other Dataset. To examine the scalability of our proposed hard voxel mining (HVM) head, we conduct experiments on another challenging dataset SemanticPOSS [12]. Since SemanticPOSS is collected in campus scenarios, there are many moving objects like *person* and *rider* that are hard to identify. SemanticPOSS contains six sequences and only provides LiDAR point clouds

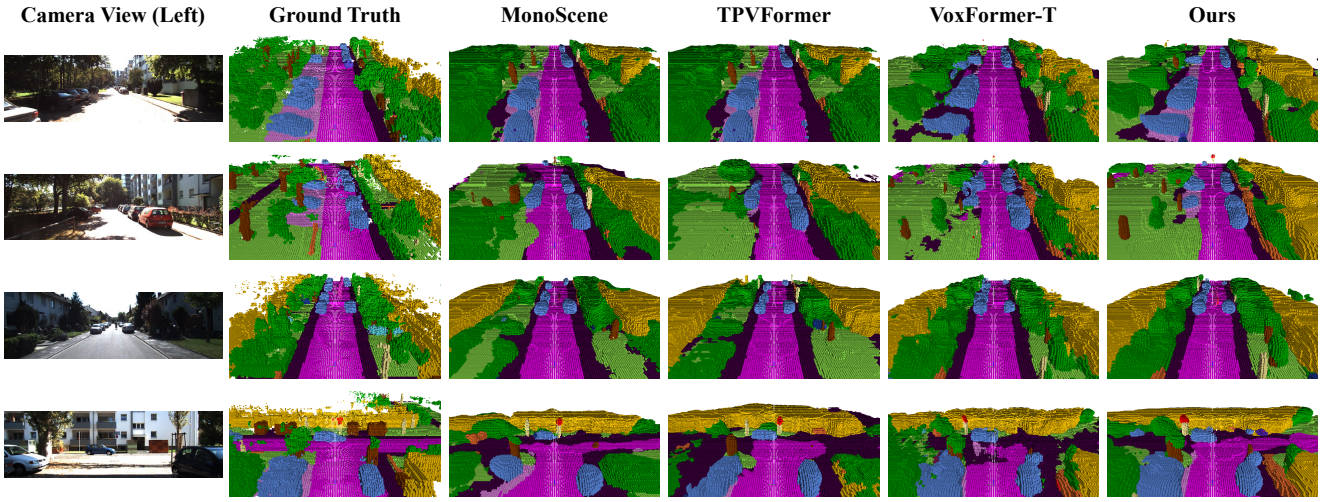

Figure A1. Additional visual results of our method (**HASSC**-VoxFormer-T) and the state-of-the-art camera-based methods on the ***validation set*** of SemanticKITTI.

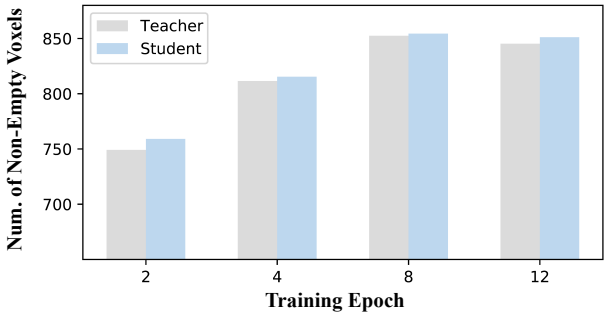

Figure A2. The number of the non-empty voxels in selected hard voxels changes during training on both student and teacher branches.

as input. We follow the original setting and split the (00-01, 03-05) / 02 as training/validation set. LiDAR-based methods including SSCNet [14] and LMSCNet [13] are utilized as baseline models. We only replace the vanilla completion head with HVM head and retrain these models. Note that the self-training strategy is not used here. As shown in Tab. A6, our HVM head achieves stable improvements with various models including SSCNet (+0.53%mIoU, +1.84%IoU) and LMSCNet (+0.92%mIoU, +0.41%IoU).

B.2. Qualitative Comparison

Visualization of Selected Hard Voxels. We visualize the number of the non-empty voxels in selected hard voxels on both student and teacher branches. As illustrated in Fig. A2, the number of the non-empty voxels continues to increase during training, which indicates that our dynamic hard voxel selection strategy favors more difficult non-empty voxels.

Comparison on Validation Set. Here, we provide more qualitative comparisons on the *validation set* of SemanticKITTI. As shown in Fig. A1, our method (**HASSC**-VoxFormer-T) obtains stable completion results in complex scenarios compared to other camera-based methods.

Comparison on Test Set. Additionally, several qualitative comparisons on the *test set* are presented in Fig. A4. Our method consistently performs well on the more challenging *test set* with 11 sequences.

C. Further Discussions

C.1. Explanatory Experiments

Failure Cases. There are mainly two factors affecting performance of our method: 1) some categories exhibit a long-tail distribution with infrequent occurrence, like *bicyclist* (0.07%), *traffic sign* (0.08%) and *trunk* (0.51%); 2) certain classes, like *vegetation*, often have sever occlusion, which brings in difficulties in geometry estimation. We provide several failure cases caused by the false geometric estimation in the Fig. A3.

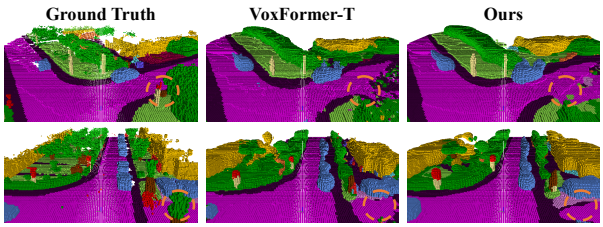

Figure A3. Visualization of some typical failure cases.

About the Formulation of Hardness. We have explored alternative designs of hardness, such as utilizing the highest confidence ("A") or loss magnitude ("B") as indicators for global hardness, and quadratic function encoding for local hardness ("C"). The Tab. A7 indicates that Eq.1 and Eq.3 of main paper can better integrate global information with local geometric priors.

Method	Design A	Design B	Design C	Ours
mIoU (%)↑	14.21	14.01	14.31	**14.74**

Table A7. Comparison of different design options.

About the Impact of Empty Voxels. Empty voxels, typically more numerous, are simpler to classify. We leverage prior ratio information of empty/non-empty voxels with various hardness levels for hard voxel weighting and model training to enhance voxel-wise SSC. As shown in the Tab. A8, **HASSC** effectively accelerates network convergence and attains promising accuracy.

Epoch	2	4	6	8	12	Final (best)
VoxFormer-T	10.57	11.91	11.95	12.47	13.06	13.33
HASSC-VoxFormer-T	**10.95**	**12.37**	**13.66**	**14.20**	**14.74**	**14.74**

Table A8. Illustration of accuracy variations during training.

C.2. Limitation and Future Work

Although having achieved promising improvements over existing methods especially at short/middle range with our proposed scheme, there is still a large performance gap between camera-based methods and LiDAR-based methods [4, 16] in the full range. Besides, over-fitting on hard voxels is indeed a potential problem. Our method is also constrained by the inaccurate geometry estimation and the long-tail distribution.

In the future, we will leverage neural radiance fields (NeRFs) [11] to extract the geometric and semantic clues contained in the image sequences to further improve the performance of *vision-centric* methods.

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

[2] Anh-Quan Cao and Raoul de Charette. MonoScene: Monocular 3D semantic scene completion. In *CVPR*, pages 3991–4001, 2022. 1, 2

[3] Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *CVPR*, pages 4193–4202, 2020. 2

[4] Ran Cheng, Christopher Agia, Yuan Ren, Xinhai Li, and Liu Bingbing. S3CNet: A sparse semantic scene completion network for LiDAR point cloud. In *CoRL*, pages 2148–2161, 2021. 4

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

[6] 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. 1, 2

[7] Bohan Li, Yasheng Sun, Xin Jin, Wenjun Zeng, Zheng Zhu, Xiaoefeng Wang, Yunpeng Zhang, James Okae, Hang Xiao, and Dalong Du. Stereoscene: Bev-assisted stereo matching empowers 3d semantic scene completion. *arXiv preprint arXiv:2303.13959*, 2023. 1

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

[9] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *AAAI*, pages 1477–1485, 2023. 1

[10] 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. 1, 2

[11] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 4

[12] Yancheng Pan, Biao Gao, Jilin Mei, Sibo Geng, Chengkun Li, and Huijing Zhao. Semanticposs: A point cloud dataset with large quantity of dynamic instances. In *IV*, pages 687–693, 2020. 2

[13] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. LMSCNet: Lightweight multiscale 3D semantic completion. In *3DV*, pages 111–119, 2020. 1, 2, 3

[14] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, pages 1746–1754, 2017. 1, 2, 3

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

[16] Zhaoyang Xia, Youquan Liu, Xin Li, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, and Yu Qiao. SCPNet: Semantic scene completion on point cloud. In *CVPR*, pages 17642–17651, 2023. 4

[17] 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 *AAAI*, pages 3101–3109, 2021. 1, 2

[18] 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 9455–9465, 2023. 2

[19] Yunpeng Zhang, Zheng Zhu, and Dalong Du. OccFormer: Dual-path transformer for vision-based 3D semantic occupancy prediction. In *ICCV*, 2023. 2

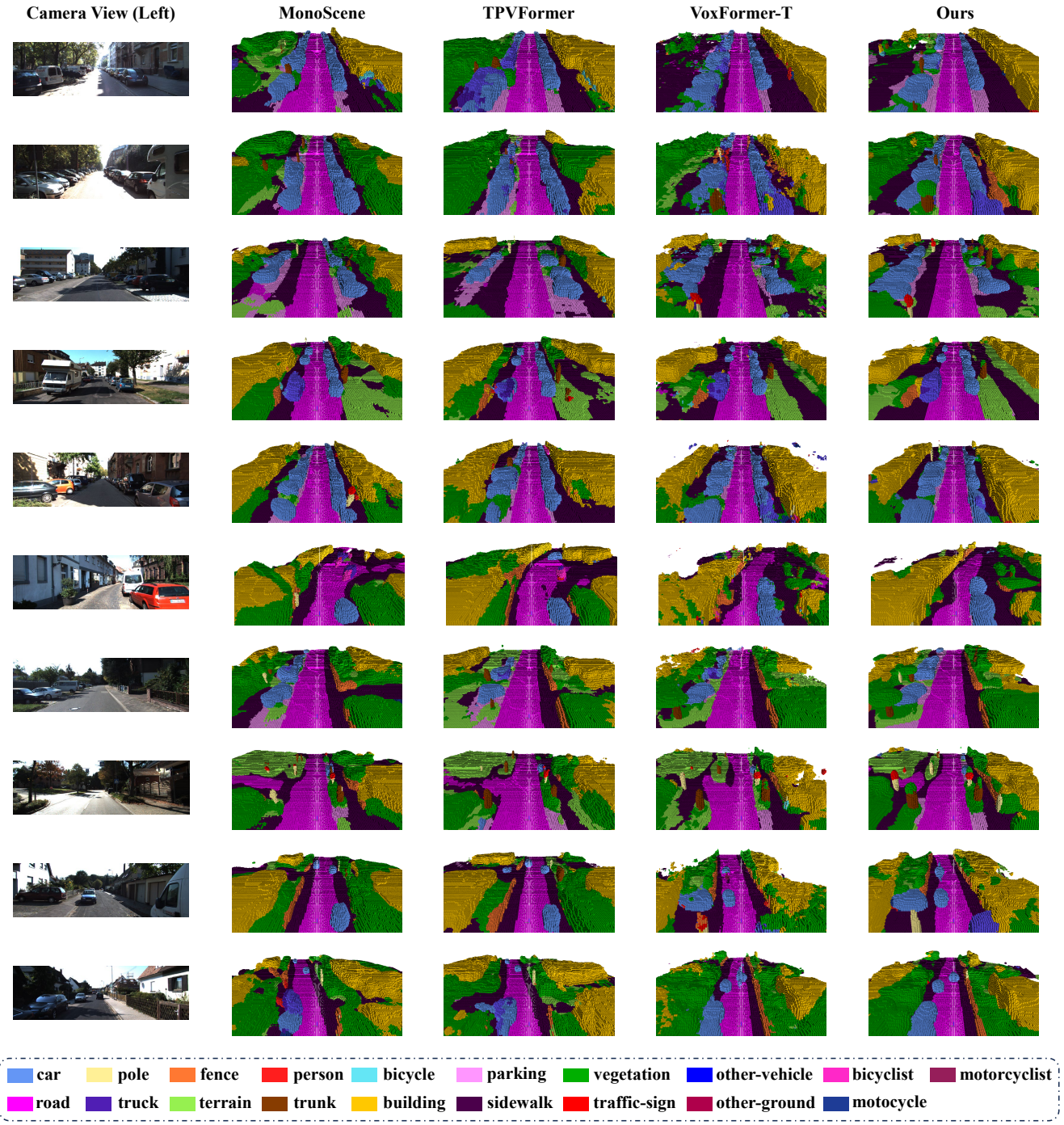

Figure A4. Visual results of our method (**HASSC**-VoxFormer-T) and other camera-based methods on the ***hidden test set*** of SemanticKITTI.