Supplementary Materials for HTCL

Bohan Li[1,2], Jiajun Deng[3], Wenyao Zhang[1,2],
Zhujin Liang[4], Dalong Du[4], Xin Jin[2] ⋆, and Wenjun Zeng[1,2]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China
[3] The University of Adelaide, Adelaide, Australia
[4] PhiGent Robotics, Beijing, China
{bohan_li, wy_zhang}@sjtu.edu.cn, jiajun.deng@adelaide.edu.au,
{zhujin.liang, dalong.du}@phigent.ai,
{jinxin, wenjunzengvp}@eitech.edu.cn

1 Additional Results

1.1 Qualitative Results

We report additional qualitative results in Figure 1 and Figure 2. As illustrated in Figure 1, we provide more qualitative comparison results between our proposed method and VoxFormer on the SemanticKITTI [1] validation set. Compared with VoxFormer, our method can predict more accurate scene layouts (e.g., crossroads in the first and second rows) and moving objects (e.g., trucks in the third row). Moreover, our method hallucinates more complete and proper out-FOV (field of view) scenes(e.g., shadow areas in the second row). As illustrated in Figure 2, we also report more visualization results on the OpenOccupancy [11] validation set. Compared to the ground truth with sparse annotations, our proposed method can generate more fine-grained realistic predictions (e.g., dense road predictions in the first and second rows).

1.2 Quantitative Results

As shown in Table 1, We conduct additional quantitative experiments on the SemanticKITTI validation dataset with other camera-based SSC methods [3, 8]. Compared to other baselines, our method achieves significant improvements in mIoU, demonstrating our method's effectiveness for semantic scene completion. Specifically, our method shows obvious superiority in capturing better moving objects (e.g., cars, bicycles, trucks) and scene layouts (e.g., roads, sidewalks).

2 More Visualization on Cross-frame Pattern Affinity

We provide more visualization results in Figure 3. Compared with the original cosine similarity, our proposed Cross-frame Pattern Affinity (CPA) effectively illustrates the contextual correspondence within the temporal content.

⋆ Corresponding author

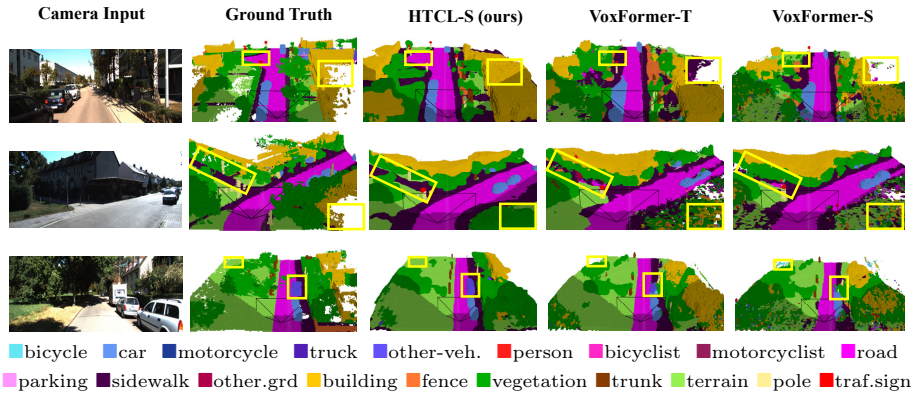

Fig. 1: Qualitative results on the SemanticKITTI validation set.

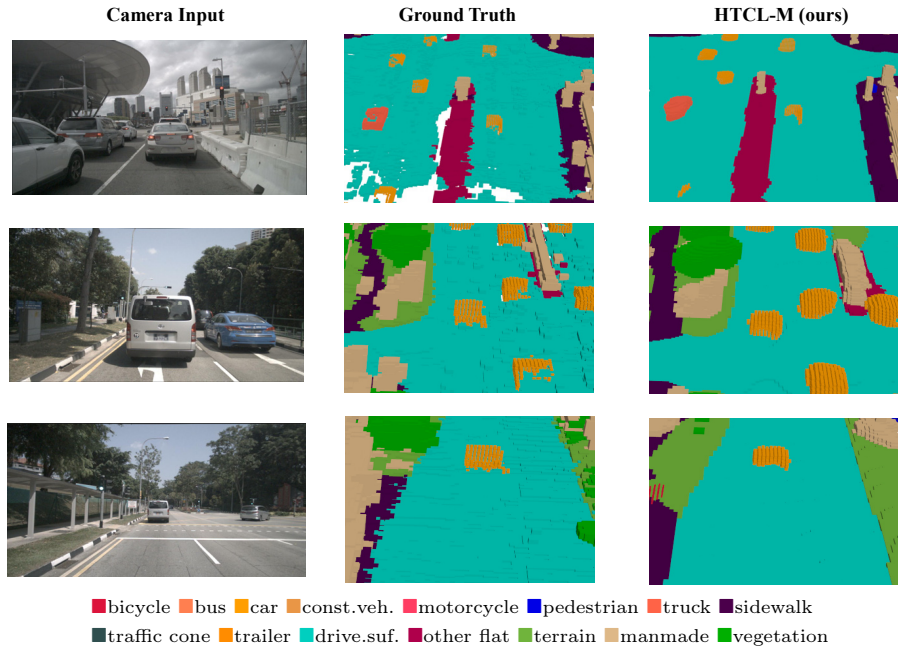

Fig. 2: Qualitative results on the OpenOccupancy validation set.

3 Extensive Experiments on BEV Detection

To further demonstrate the potential of our method, we present preliminary experimental evaluations on the Bird-Eye-View (BEV) detection [6, 7, 9, 10, 13] in the nuScenes [2] validation dataset. Specifically, we utilize *BEVDet* [5] as the

Table 1: Quantitative results on the SemanticKITTI validation set with the state-of-the-art camera-based SSC methods. The "S-T", "S" and "M" denote temporal stereo images, single-frame stereo images, and single-frame monocular images, respectively.

Methods	HTCL-S (ours)	VoxFormer-T	VoxFormer-S	MonoScene
Input	S-T	S-T	S	M
IoU	**45.51**	44.15	44.02	37.12
mIoU	**17.13**	13.35	12.35	11.50
car	**34.30**	26.54	25.79	23.55
bicycle	**3.99**	1.28	0.59	0.20
motorcycle	**2.80**	0.56	0.51	0.77
truck	**20.72**	8.10	7.26	7.83
other-veh.	**11.99**	7.81	3.77	3.59
person	**2.56**	1.93	1.78	1.79
bicyclist	2.30	1.97	**3.32**	1.03
motorcyclist	0.00	0.00	0.00	0.00
road	**63.70**	53.57	54.76	57.47
parking	**23.27**	19.69	15.50	15.72
sidewalk	**32.48**	26.52	26.35	27.05
other.grd	0.14	0.42	0.70	**0.87**
building	**24.13**	19.54	17.65	14.24
fence	**11.22**	7.31	7.64	6.39
vegetation	**26.96**	26.10	24.39	18.12
trunk	**8.79**	6.10	5.08	2.57
terrain	**37.73**	33.06	29.96	30.76
pole	**11.49**	9.15	7.11	4.11
traf.sign	**6.95**	4.94	4.18	2.48

baseline configuration and substitute the original model of *BEVDet* with our proposed *HTCL-M*, while maintaining the same detection head. The evaluation results are reported in Table 2 and visually depicted in Figure 4. As shown in Table 2, our HTCL-M outperforms BEVDet4D-Base with a relative improvement of 20.19% mAP and 6.34% NDS, respectively. These results illustrate the effectiveness of our proposed methodology, indicating its potential applicability to a broader spectrum of downstream tasks.

Table 2: Quantitative results of BEV Detection on the nuScenes validation set. We conduct preliminary experiments by employing the detection head.

Methods	**Resolution**	**mAP ↑**	**NDS ↑**
BEVDet-Base	1600× 640	0.397	0.477
BEVDet4D-Base	1600× 640	0.426	0.552
PETR-R101	1408× 512	0.357	0.421
BEVDepth-R101	512× 1408	0.412	0.535
HTCL-M (ours)	1600× 640	**0.512**	**0.587**

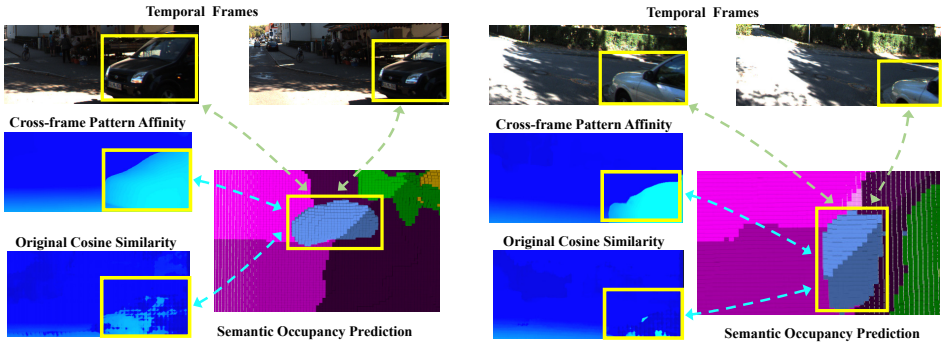

Fig. 3: Visualization of the heat maps from our proposed Cross-frame Pattern Affinity (CPA) and the original cosine similarity.

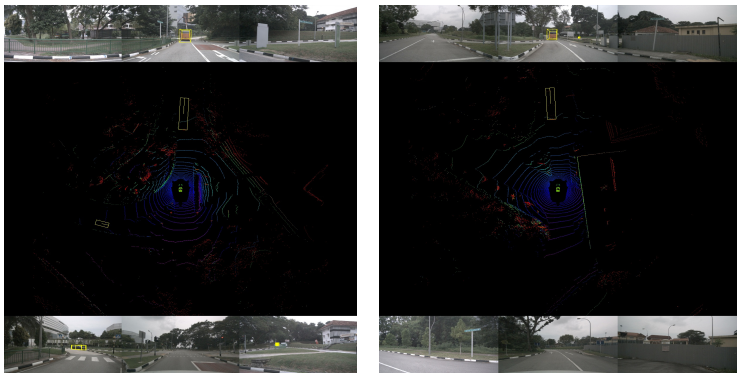

Fig. 4: Visualization results of BEV detection on the nuScenes validation set.

4 Additional Ablation Studies

We conduct additional ablation studies on the Multi-group Context Generation and the Multi-level Deformable Block, as presented in Table 3. As introduced in the main paper, we employ multiple groups of contextual features to facilitate diverse independent similarity learning. The results in Table 3 demonstrate that leveraging 3 contextual groups yields a significant performance improvement, while employing more groups (5 groups) leads to a relatively slight improvement. Similarly, the enhancement of utilizing more feature levels (5 levels) in the Multi-level Deformable Block is also relatively minor. Therefore, considering the time consumption and parameter efficiency, we adopt 3 contextual groups in the Multi-group Context Generation and 3 feature levels in the Multi-level Deformable Block as the default settings.

Table 3: Ablation studies on the Multi-group Context Generation and the Multi-level Deformable Block.

Multi-group Context Generation			Multi-level Deformable Block			mIoU (%)	Time (s)
1 Group	3 Groups	5 Groups	1 Level	3 Levels	5 Levels		
✓				✓		15.26	0.283
		✓		✓		17.21	0.312
	✓		✓			16.51	0.286
	✓				✓	17.18	0.309
	✓			✓		17.13	0.297

5 Loss Function

PoseNet implementation We implement the PoseNet following previous video depth estimation methods [4,12]. To reduce the learning burden, we pre-train the PoseNet and freeze it for temporal semantic occupancy learning. The PoseNet is trained without ground truth in a self-supervised manner. At the pre-training stage, the training objective of the PoseNet is:

$$\mathcal{L}_{pose} = \min_{n} \mathrm{PE}\left(I_t, I_{t+n\to t}\right) + \lambda * \mathcal{L}_{smooth}. \tag{1}$$

where PE is a combination of SSIM and L1 losses between reference image I_t and source image I_{t+n}. L_{smooth} is the smoothness loss for pixel-level regularization from [4, 12]. λ is the balance coefficient. We will add the details in Section 7 (Network Training) of the supplementary material.

Network Training. We follow the basic learning objective of MonoScene [3] for semantic scene completion. Standard semantic loss $\mathcal{L}_{\text{sem}}$ and geometry loss $\mathcal{L}_{\text{geo}}$ are leveraged for semantic and geometry supervision, while an extra class weighting loss $\mathcal{L}_{ce}$ is also added. To further enforce the ensembled volume, we adopt a binary cross entropy loss $\mathcal{L}_{depth}$ to encourage the sparse depth distribution. The overall learning objective of this framework is formulated as:

$$\mathcal{L} = \mathcal{L}_{depth} + \lambda_{ce}\mathcal{L}_{ce}. \tag{2}$$

where several λs are balancing coefficients.

6 Limitation and Potential Negative Impact

The running speed of our model could be further enhanced as more lightweight networks are more practical for real-world applications. We leave this to our future work. While the promising semantic scene completion results of the proposed method could promote the development of autonomous driving, the legal challenges, as well as the privacy and data security risks of autonomous driving remain subjects of debate.

References

1. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: ICCV (2019)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
3. Cao, A.Q., de Charette, R.: Monoscene: Monocular 3d semantic scene completion. In: CVPR (2022)
4. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: CVPR (2020)
5. Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790 (2021)
6. Kopf, J., Rong, X., Huang, J.B.: Robust consistent video depth estimation. In: CVPR (2021)
7. Li, S., Luo, Y., Zhu, Y., Zhao, X., Li, Y., Shan, Y.: Enforcing temporal consistency in video depth estimation. In: ICCV (2021)
8. Li, Y., Yu, Z., Choy, C., Xiao, C., Alvarez, J.M., Fidler, S., Feng, C., Anandkumar, A.: Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In: CVPR (2023)
9. Lin, X., Lin, T., Pei, Z., Huang, L., Su, Z.: Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. arXiv preprint arXiv:2211.10581 (2022)
10. Liu, Y., Wang, T., Zhang, X., Sun, J.: Petr: Position embedding transformation for multi-view 3d object detection. In: ECCV. Springer (2022)
11. Wang, X., Zhu, Z., Xu, W., Zhang, Y., Wei, Y., Chi, X., Ye, Y., Du, D., Lu, J., Wang, X.: Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. ICCV (2023)
12. Watson, J., Mac Aodha, O., Prisacariu, V., Brostow, G., Firman, M.: The temporal opportunist: Self-supervised multi-frame monocular depth. In: CVPR (2021)
13. Zhang, H., Shen, C., Li, Y., Cao, Y., Liu, Y., Yan, Y.: Exploiting temporal consistency for real-time video depth estimation. In: ICCV (2019)