Bi-SSC: Geometric-Semantic Bidirectional Fusion for Camera-based 3D Semantic Scene Completion (SupplementaryMaterial)

Yujie Xue*, Ruihui Li*, Fan Wu†, Zhuo Tang, Kenli Li, Mingxing Duan†
College of Computer Science and Electronic Engineering, Hunan University
{xueyj, liruihui, wufan, ztang, lkl, duanmingxing}@hnu.edu.cn

A. Quantitative Comparison

In the section, we present quantitative results of our method with state-of-the-art camera-based methods such as MonoScene [2], VoxFormer [8], TPVFormer [5], and OccFormer [13] on the validation sets of SemanticKITTI [1].

Semantic scene completion. As illustrated in Table 1, Bi-SSC surpasses VoxFormer-T in semantic scene understanding, exhibiting a significant performance gap. Furthermore, when compared to VoxFormer-T, which relies on historical observation data, Bi-SSC leverages binocular depth information to achieve a substantial relative mIoU improvement of 22.77%. It is worth noting that in 3D scenes, the precise comprehension of object semantics is pivotal for an accurate representation of the real world, as misunderstand could lead to erroneous decisions. Consequently, in practical camera-based applications, our approach is preferable to others.

Scene completion. The holistic scene analysis reveals that Bi-SSC also excels in scene completion, as shown in Table 1. Its IoU outperforms both monocular and binocular methods. Compared to the state-of-the-art VoxFormer-T, Bi-SSC exhibits a remarkable 0.73 increase in IoU, with an even wider performance margin compared to single-purpose methods. Notably, the values of IoU and mIoU are intertwined, and inaccurate scene completion can adversely impact the semantic understanding of the entire scene. In contrast, our approach demonstrates exceptional performance in both geometric and semantic aspects.

B. Qualitative Comparison

Figure 1 presents more visualizations. It is important to note that as we do not have access to the specifics of the test set, we can only visualize the results obtained from the validation set. Our approach demonstrates superior performance compared to other camera-based methods, particularly in heavily obscured or highly cluttered scenes. VoxFormer-T [8] exhibits a significant number of missing objects beyond the image line of sight. Notably, as depicted in the second and fourth lines of Figure 1, an observation is made that the road cannot be fully reconstructed on the left and right sides, while ours can be semantical scene completed better. Concurrently, our method excels in capturing intricate 3D details on dense objects, illustrated in the last line of Figure 1, where the outlines of vehicles and tree trunks are more accurately segmented.

C. More Ablation Study

Figure 2 shows the visualized results of using the SSF we designed versus using the attention mechanism baseline. Compared to the attention baseline, our SSF can be seen as an effective way to mitigate occluded areas in the SSC. As can be seen more clearly from the figure, SSF is good at recovering object structure and reasoning about the intersection between adjacent semantic classes. For example, in the first row of Figure 2, the road surface is more complete, and the spatiality deduced by SSF is much better when the tree and the car block each other.

D. Comparison Against 2.5/3D-input Baselines

We compared Bi-SSC with some original 2.5D/3D input baselines. In SemanticKITTI (hidden test set), the methods in Table 2 use pseudo-3D information inferred from RGB as input, and it can be observed that our model performs better on the camera-based SSC task. In Table 3, the methods use real 3D as input. Despite using image inputs with a horizontal field of view (FOV) much smaller than the LiDAR (82° vs. 180°), our completion capability is approaching recent S3CNet models, and our segmentation ability is gradually catching up to the 2.5D/3D input baselines. It can be seen that our work is gradually closing the gap between 2D and 3D. This observation is promising and encouraging for SSC, since Bi-SSC only needs cheap cameras for its infer process.

* Equal contributed. † Corresponding authors.

Method	Input	IoU	car(3.92%)	bicycle(0.03%)	motocycle(0.03%)	truck(0.16%)	other-vehicle(0.20%)	person(0.07%)	bicylist(0.07%)	motorcyclist(0.05%)	road(15.30%)	parking(1.12%)	sidewalk(11.13%)	other-grnd(0.56%)	building(14.10%)	fence(3.90%)	vegetation(39.3%)	trunk(0.51%)	terrain(9.17%)	pole(0.29%)	traf.-sign(0.08%)	mIoU
MonoScene [2]	Mono	37.12	23.55	0.20	0.77	7.83	3.59	1.79	1.03	0.00	57.47	15.72	27.05	0.87	14.24	6.39	18.12	2.57	30.76	4.11	2.48	11.50
TPVFormer [5]	Mono	35.61	23.81	0.36	0.05	8.08	4.35	0.51	0.89	0.00	56.50	**20.60**	25.87	0.85	13.88	5.94	16.92	2.26	30.38	3.14	1.52	11.36
OccFormer [13]	Mono	36.50	25.09	0.81	1.19	**25.53**	8.52	**2.78**	**2.82**	0.00	58.85	19.61	26.88	0.31	14.40	5.61	19.63	3.93	32.62	4.26	2.86	13.46
VoxFormer-T [8]	Stereo	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
Bi-SSC (Ours)	Stereo	**44.88**	**32.32**	1.27	**3.23**	18.9	**11.85**	2.29	1.76	**0.00**	**61.95**	20.0	**30.29**	**1.16**	**24.61**	**10.49**	25.91	**9.02**	**37.37**	**12.17**	**6.9**	**16.39**

Table 1. **Performance on the SemanticKITTI [1] validation set.** We report the performance on semantic scene completion (SSC-mIoU) and scene completion (SC-IoU) for our method and others. The best performing methods are marked in **bold**.

Method	Input	IoU(↑)	mIoU(↑)
LMSCNet [9]	$\hat{x}^{occ}$	31.38	7.07
3DSKetch [3]	$x^{rgb}, \hat{x}^{TSDF}$	26.85	6.23
JS3CNet [12]	$\hat{x}^{pts}$	34.00	8.97
AICNet [7]	$x^{rgb}, \hat{x}^{depth}$	23.93	7.09
Bi-SSC (Ours)	x^{rgb}	**45.10**	**16.73**

Table 2. **RGB-inferred** for SeamnticKITTI dataset (hidden test set).

Method	Input	IoU(↑)	mIoU(↑)
SSCNet [10]	x^{TSDF}	29.80	9.50
LMSCNet [9]	x^{occ}	**56.70**	17.60
JS3CNet [12]	x^{pts}	56.60	23.80
S3CNet [4]	x^{occ}	45.60	**29.50**
Bi-SSC (Ours)	x^{rgb}	45.10	16.73

Table 3. **Real 3D input** inference in SemanticKITTI dataset (hidden test set).

E. Quantitative analysis at different ranges.

Our superiority in far-range regions. As shown in Table 4, in the critical far region of the visual field blindness, our method shows significant improvement over other camera-based methods. Bi-SSC can obtain mIoU scores of 17.98 and 14.13 in the ranges of 12.8-25.6 meters and 25.6-51.2 meters, which are 35.9% and 161.6% higher than the state-of-the-art VoxFormer-T, respectively. In terms of scene completion, the same better performance is achieved compared to other methods. Such an improvement mainly comes from the full exploitation of visual information.

Method	IoU(↑)			mIoU(↑)		
	12.8-25.6m	25.6-51.2m	0-51.2m	12.8-25.6m	25.6-51.2m	0-51.2m
OccFormer [13]	35.66	34.18	36.50	13.39	12.76	13.46
StereoScene [6]	41.07	37.97	43.85	14.93	13.71	15.43
VoxFormer-T [8]	**54.00**	24.87	44.15	13.23	5.4	13.35
Bi-SSC (Ours)	48.58	**40.05**	**44.88**	**17.98**	**14.13**	**16.39**

Table 4. Quantitative comparison at different ranges.

F. Mutual Interactive Aggregation Module Details

Inspired by StereoScene [6], by utilizing the acquired stereo features F_s and refined features F_{refine}, the *Mutual Interactive Aggregation* (MIA) module aims to mutually reinforce and integrate their individual potentials, thereby obtaining the resulting new features.

Specifically, MIA is designed to guide interactively through cross-attention mechanism, ensuring the acquisition of reliable predictions. Following the standard protocol [11], the refined features transformer into the query $Q_{refine} \in \mathbb{R}^{C\times H\times W}$, key $K_{refine} \in \mathbb{R}^{C\times H\times W}$, and value $V_{refine} \in \mathbb{R}^{C\times H\times W}$. Likewise, the stereo features are transformed into Q_s, K_s, and V_s representations and geometric features also are transformed into Q_g, K_g, and V_g representations. These transformations enable the utilization of cross-attention operations, which can be mathematically expressed as:

$$CrossAtt_s(Q_s, K_{refine}, V_{refine}) = softmax(K^T_{refine} Q_s) \quad (1)$$

Similarly, geometric features and refinement features are calculated in this way:

$$CrossAtt_g(Q_g, K_{refine}, V_{refine}) = softmax(K^T_{refine} Q_g) \quad (2)$$

Through two interactions to encourage reliable geometric information transmission.

In the above cross-attention operation, the features after cross-attention are first input into the residual CNN network

for regularization and channel number adjustment to generate a transform representing $F \in \mathbb{R}^{C\times D\times H\times W}$. Then utilize average pooling to compress the depth dimension D and spatial dimension $H \times W$, denoted as:

$$d_c = \frac{1}{D \times H \times W} \sum_{d=1}^{D} \sum_{i=1,j=1}^{H,W} F(d, i, j) \tag{3}$$

and the channel dependency was captured by the excitation module, which is updated as:

$$\hat{d}_c = \sigma(C_2\delta(C_1 d_c)) \tag{4}$$

where the C_1 and C_2 represent 1×1×1 convolutions with dimensionality-reduction. The σ indicates sigmoid gate and the δ denotes standard GELU. By using point-wise convolution with the change feature F, we get the output. Formally,

$$F_{refine} = \mathbb{M}(\hat{d}_c \odot F) \tag{5}$$

Where $\mathbb{M}$ consists of point-wise convolution of GELU activation and group normalization.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019. 1, 2

[2] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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 clouds. In *Conference on Robot Learning*, pages 2148–2161. PMLR, 2021. 2

[5] 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 Computer Vision and Pattern Recognition*, pages 9223–9232, 2023. 1, 2

[6] 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. 2

[7] 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 Computer Vision and Pattern Recognition*, pages 3351–3359, 2020. 2

[8] 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 Computer Vision and Pattern Recognition*, pages 9087–9098, 2023. 1, 2

[9] 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

[10] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017. 2

[11] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2

[12] 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

[13] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2304.05316*, 2023. 1, 2

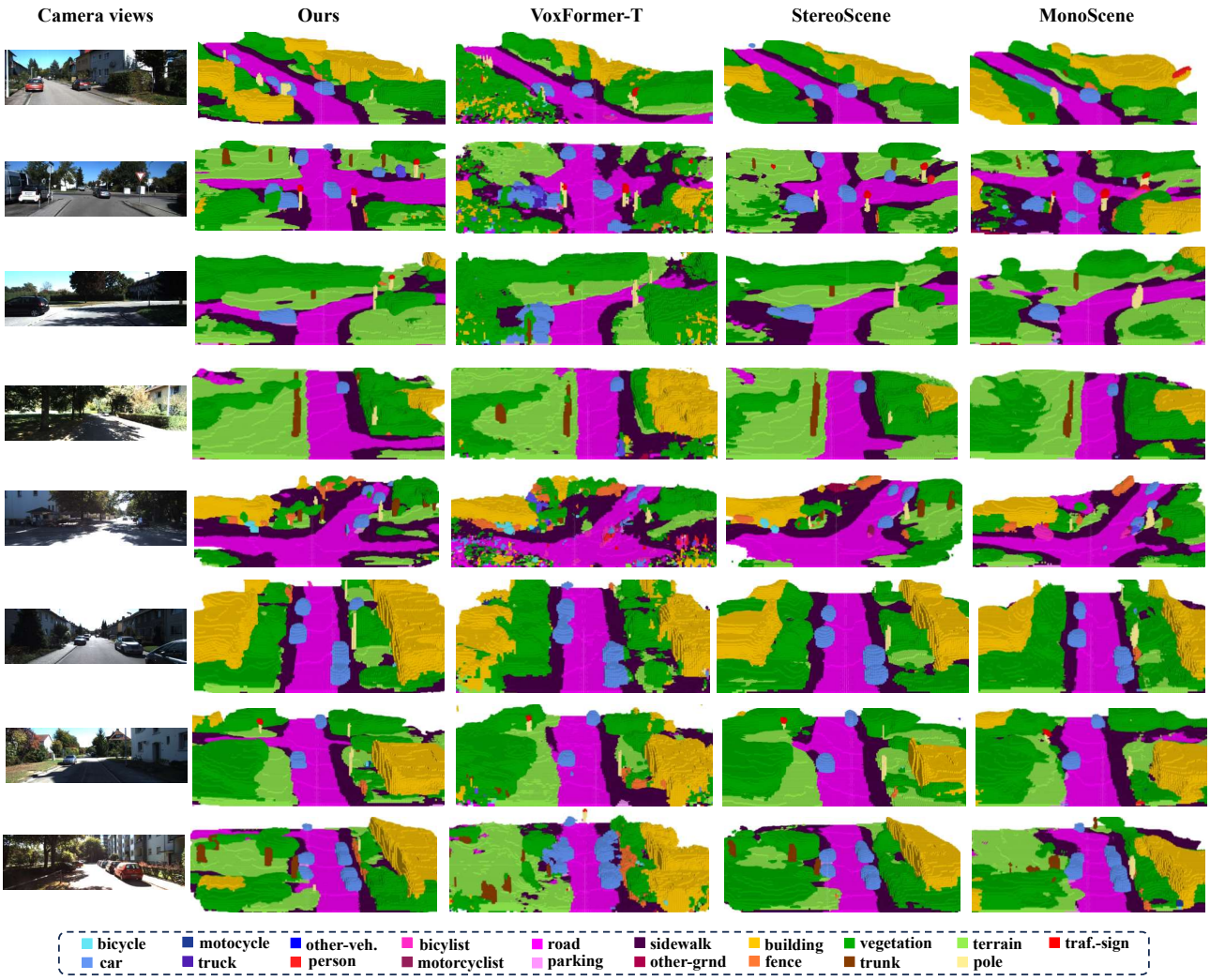

Figure 1. **Qualitative results in SemanticKITTI dataset.** Our approach better captures the layout of the scene, it reconstructs and estimates the geometry of the obscured roads and shaded areas of the car.

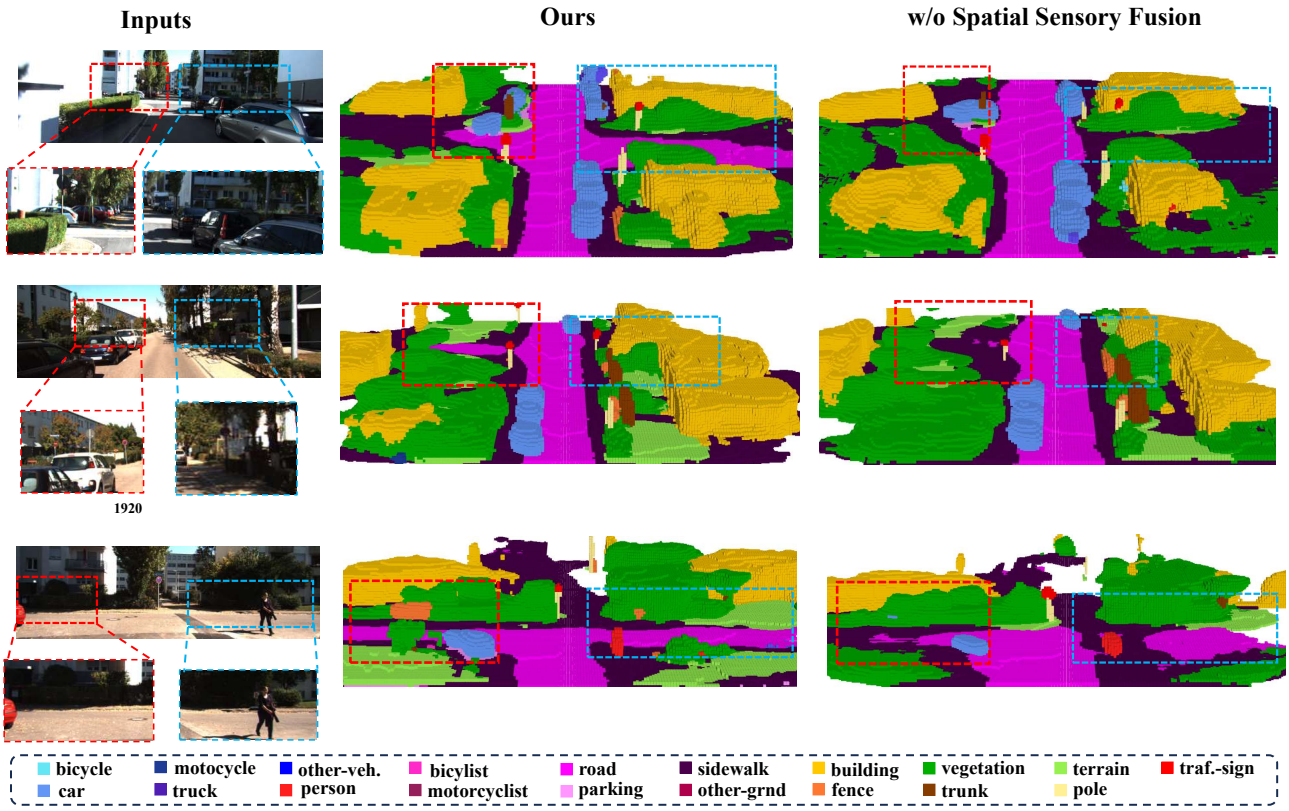

Figure 2. **Qualitative results of the Spatial Sensory Fusion (SSF) module influence.** SSF has a positive contribution to blocking and shading blurred areas. It can be clearly seen that our SSF completes the crossroads better, and the poles in the shaded areas can also be segmented.