DepthSSC: Monocular 3D Semantic Scene Completion via Depth-Spatial Alignment and Voxel Adaptation – Supplementary Material –

Jiawei Yao[1] Jusheng Zhang[2] Xiaochao Pan[3] Tong Wu[1] Canran Xiao[4*]
[1] University of Washington [2] Sun Yat-sen University
[3] Taiyuan University of Technology [4] Central South University
jwyao@uw.edu, xiaocanran@csu.edu.cn

A. Experiment setup

Dataset We evaluate DepthSSC on the SemanticKITTI [1] dataset, renowned for its dense semantic annotations of urban driving sequences from the KITTI Odometry Benchmark. The dataset voxelizes point clouds into a 51.2m×51.2m×64m scene, represented by 256×256×32 voxel grids, and includes 20 semantic classes, including the "empty" category. SemanticKITTI includes RGB images (1220×370) and LiDAR sweeps as inputs. The dataset is divided into 10 sequences for training, 1 for validation, and 11 for testing. We also evaluate DepthSSC on the SSCBench-KITTI360 [13],which consist of 20 and 19 classes. This dataset consist of voxel grids with semantic labels covering an area of $51.2\,\text{m} \times 51.2\,\text{m} \times 6.4\,\text{m}$, with each voxel having a size of $0.2\,\text{m}$, creating a grid resolution of $256 \times 256 \times 32$.

Metric For our experimentation with the DepthSSC model, we exclusively utilize RGB images from a monocular vision setup. These images, being a primary source of input for our model, facilitate the understanding of scene structures and semantics. Our primary evaluation metrics remain focused on the intersection over union (IoU) for the occupied voxel grids. Additionally, we also adopt the mean IoU (mIoU) metric for voxel-wise semantic evaluations.

Baselines We compare our proposed DepthSSC with existing SSC baselines (JS3CNet [19], AICNet [12] and LMSCNet [22]). We also compare DepthSSC with MonoScene [3], TPVFormer [12], VoxFormer [14], NDC-Scene [23], Symphonies [9] and HASSC [21], which are best RGB-only SSC methods. Note that for the methods with more than RGB inputs, we follow [3] to adapt their results to RGB only inputs.

*Corresponding author.

Moreover, in Table 1 and Table 2 of the main paper, the notations Occ, Depth and Pts denote the occupancy grid, depth map and point cloud, which are the 3D input required by the SSC baselines. For a fair comparison, all the three inputs are converted from the depth map predicted by a pretrained depth predictor [2].For implementation details of DepthSSC, please refer to the supplementary materials.

B. Implementation details

B.1. Architectures

We adopt ResNet-50 [5] as the backbone for 2D feature extraction. The backbone consists of four stages, and we utilize features from the third stage (out_indices=(2,)) for further processing. The network is partially frozen (frozen_stages=1) and employs batch normalization (BN) for stable training. We employ a Feature Pyramid Network (FPN) [16] to process the extracted 2D features. The FPN takes the 1024-dimensional features from the ResNet backbone and transforms them into a 128-dimensional feature space. The FPN starts from level 0, adds extra convolutions on output, and produces feature maps with a spatial resolution of $H/4 \times W/4 \times 128$.

In the proposed method, voxel queries are 3D grid-shaped learnable parameters that map 2D features to the 3D volume. The voxel queries $Q \in \mathbb{R}^{64\times64\times16\times128}$ are generated at a lower resolution to reduce computational load. From these voxel queries, a subset $Q_p = \text{Reshape}(Q[M_{\text{out}}])$ is selected based on predicted occupancy from depth information, resulting in $Q_p \in \mathbb{R}^{1024\times128}$, where M_{out} is the corrected occupancy map. To handle multi-modal data, we incorporate a cross-transformer and a self-transformer. The cross-transformer utilizes a PerceptionTransformer architecture with three encoder layers, each based on VoxFormerEncoder, which processes input using deformable cross-attention mechanisms [10] to integrate multi-view image features into a unified 3D space. This encoder attends

to 8 points per pillar and employs Multi-Scale Deformable Attention [26] for effective feature fusion. Each layer in the cross-transformer has an embedding dimension of 128 and a feedforward dimension of 256, with a dropout rate of 0.1. The self-transformer follows a similar Perception-Transformer3D architecture with two encoder layers, using deformable self-attention to refine voxel features within the 3D space. The self-transformer has an embedding dimension of 128 and attends to 8 points within each voxel.

Our Spatially-Transformed Graph Fusion (ST-GF) module addresses the misalignment issue between depth maps and voxel queries. The Adaptive Spatial Adjustment Network (ASAN) predicts a 3D affine transformation matrix Θ for each voxel query $Q \in \mathbb{R}^{64\times64\times16\times128}$ and depth prediction $D \in \mathbb{R}^{64\times64\times16\times1}$. Using Θ, the grid generator maps the output space to the input space, applying trilinear interpolation to adjust voxel positions. Transformed voxels are clustered into nodes, with edges representing spatial relationships. Node features are fused via graph convolution, and the refined features are backpropagated to the original voxel space, ensuring accurate scene comprehension. The resolution-adaptive deformable attention mechanism adjusts the positions and quantities of query points in deformable self-attention based on the dynamically assigned resolution. For each voxel V_i, its position in three-dimensional space can be represented as $p = (x, y, z)$. We adjust these positions based on the resolution $R(V_i)$, allowing voxels with higher complexity to have a higher query density. The adjusted query points are calculated as $p' = p + \delta R(V_i) + \Delta p$, where $\delta = 0.1$ is a constant.

B.2. Hyperparameter for Training

We utilize 4 NVIDIA Tesla A100 GPUs to train the DepthSSC model across 30 epochs, processing a batch size of 4 images in each iteration. These RGB images are of the resolution 1220×370. During training, we incorporate a random horizontal flip for data augmentation. For optimization, we employ the AdamW optimizer, initiating with a learning rate of 1e-4 coupled with a weight decay of 1e-4. By the time we reach the 5th epoch, we decrease the learning rate by 10%. Both stage-1 and stage-2 are trained separately for 24 epochs, using a learning rate of 2×10^{-4}.

C. Additional Ablation Studies

ST-GF Ablation Experiments. The ST-GF module combines spatial transformation and graph structure features to ensure accurate alignment of spatial information between the depth map and voxel queries in 3D scene completion. Alternative alignment techniques, such as Iterative Closest Point (ICP) [24], feature-based registration [11], and regularization-based matching [17], can also be used. Regularization-based matching minimizes a distance metric between the source and target, while feature-based registration uses extracted feature points for matching. As shown in Table 1, ST-GF outperforms other alignment methods, demonstrating its effectiveness. However, any alignment method improves the performance of the original VoxFormer.

Alignment Method	Ours(SemanticKITTI) IoU ↑	mIoU ↑
ST-GF	**45.97**	**14.59**
ICP	44.28 (-1.69)	12.64(-1.95)
Feature-based Reg.	44.45(-1.52)	12.76(-1.83)
Regularization Matching	44.87(-1.10)	12.98(-1.61)

Table 1. **Ablation study** for ASAN in ST-GF module.

Distance Metric	Ours(SemanticKITTI) IoU ↑	mIoU ↑
Euclidean Distance	**45.97**	**14.59**
Cosine Similarity	45.46(-0.51)	13.25(-1.34)
Manhattan Distance	45.33(-0.64)	12.92(-1.67)

Table 2. **Ablation study** on connection strength computation using different distance metrics.

Connection Strength Ablation Experiments. In the ST-GF module, connection strength represents relationships or similarities between nodes, influencing which relationships are fused during graph convolution. We compare different distance metrics for semantic scene completion (SSC). Table 2 shows that Euclidean distance achieves the best performance, as it accurately captures the actual distance and relative positional relationships in 3D space. Cosine similarity, focusing more on direction than magnitude, is less suitable for this task. Manhattan distance considers spatial aspects but does not account for the shortest distance between two points, leading to potential information loss or inaccuracies.

Resolution-Adaptive Deformable Attention Ablation Experiments. Resolution-adaptive deformable attention addresses the varying geometric complexities in 3D data by enabling finer voxel resolutions in complex regions. To validate its effectiveness, we compared it against non-uniform voxelization [6], dynamic kernel [4], and non-local operations [18]. These methods partially address geometric complexity but have limitations. Table 4 presents the ablation results. Non-uniform voxelization can cause data discontinuities and biases. Dynamic kernel methods face alignment issues with different kernel sizes and shapes. Non-local operations, while capturing long-range dependencies, are computationally intensive for 3D data and less effective at local complexities. The results demonstrate that resolution-

Method	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation	mIoU
MonoScene [3]	1.75	7.23	4.26	4.93	9.38	5.67	3.98	3.01	5.90	4.45	7.17	14.91	6.32	7.92	7.43	1.01	7.65	6.06
TPVFormer [8]	7.22	38.90	13.67	**40.78**	45.90	17.23	19.99	18.85	14.30	26.69	34.17	55.65	35.47	37.55	30.70	19.40	16.78	27.83
BEVDet [7]	4.39	30.31	0.23	32.36	34.47	12.97	10.34	10.36	6.26	8.93	23.65	52.27	24.61	26.06	22.31	15.04	15.10	19.38
OccFormer [25]	5.94	30.29	12.32	34.40	39.17	14.44	16.45	17.22	9.27	13.90	26.36	50.99	30.96	34.66	22.73	6.76	6.97	21.93
BEVFormer [15]	5.85	37.83	17.87	40.44	42.43	7.36	23.88	21.81	20.98	22.38	30.70	55.35	28.36	36.0	28.06	20.04	17.69	26.88
DepthSSC(Ours)	**12.11**	**40.26**	**28.35**	37.42	**54.56**	**23.15**	**29.20**	**28.73**	**29.21**	**31.22**	**38.67**	**72.88**	**46.05**	**47.85**	**33.23**	**37.02**	24.37	**38.84**

Table 3. 3D semantic occupancy prediction performance on the validation set of Occ3D-nuScenes [20].

Region-Adaptive Method	Ours(SemanticKITTI) IoU ↑	mIoU ↑
Resolution-Adaptive Deformable Attention	**45.97**	**14.59**
Non-Uniform Voxelization	40.61(-5.36)	10.63(-3.96)
Dynamic Kernel Methods	42.58(-3.39)	11.80(-2.79)
Non-Local Operations	41.47(-4.50)	11.39(-3.20)

Table 4. **Ablation study** on resolution-adaptive deformable attention.

adaptive deformable attention outperforms these methods, capturing local geometric details more effectively.

D. Results on Occ3D

The results in Table 3 demonstrate the superiority of our DepthSSC over several state-of-the-art methods on the Occ3D-nuScenes validation set. DepthSSC achieves the highest mIoU of 38.84, outperforming existing approaches such as TPVFormer (27.83) and BEVFormer (26.88) by a significant margin.

Specifically, ST-GF allows for more accurate spatial alignment between voxel queries and depth maps, leading to enhanced object recognition in challenging categories such as vegetation (37.02) and manmade structures (33.23). GAV further refines voxel resolution dynamically, which is particularly beneficial in capturing fine details in complex categories like bicycles (28.35) and pedestrians (28.73). DepthSSC also excels in driveable surface detection (72.88), indicating that the proposed fusion and voxelization techniques are effective in both object-level and scene-level predictions. The substantial improvements across a range of categories, especially in highly dynamic or small-scale objects, validate the robustness of our approach in semantic scene completion tasks.

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] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 1

[3] 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, 3

[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] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[6] Zeyu Hu, Xuyang Bai, Jiaxiang Shang, Runze Zhang, Jiayu Dong, Xin Wang, Guangyuan Sun, Hongbo Fu, and Chiew-Lan Tai. Voxel-mesh network for geodesic-aware 3d semantic segmentation of indoor scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[7] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 3

[8] 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. 3

[9] 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 Computer Vision and Pattern Recognition*, pages 20258–20267, 2024. 1

[10] Sangwon Kim, Dasom Ahn, and Byoung Chul Ko. Cross-modal learning with 3d deformable attention for action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10265–10275, 2023. 1

[11] Kavitha Kuppala, Sandhya Banda, and Thirumala Rao Barige. An overview of deep learning methods for image registration with focus on feature-based approaches. *International Journal of Image and Data Fusion*, 11(2):113–135, 2020. 2

[12] 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. 1

[13] Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, et al. Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. *arXiv preprint arXiv:2306.09001*, 2023. 1

[14] 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

[15] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 3

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[17] Yonghuai Liu, Luigi De Dominicis, Baogang Wei, Liang Chen, and Ralph R Martin. Regularization based iterative point match weighting for accurate rigid transformation estimation. *IEEE transactions on visualization and computer graphics*, 21(9):1058–1071, 2015. 2

[18] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 120–136. Springer, 2020. 2

[19] 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. 1

[20] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[21] 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 Computer Vision and Pattern Recognition*, pages 14792–14801, 2024. 1

[22] 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*, 2021. 1

[23] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9455–9465, 2023. 1

[24] Juyong Zhang, Yuxin Yao, and Bailin Deng. Fast and robust iterative closest point. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3450–3466, 2021. 2

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

[26] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2