AdaSFormer: Adaptive Serialized Transformers for Monocular Semantic Scene Completion from Indoor Environments

Supplementary Material

The supplementary material consists of two parts: 1) network details and 2) additional visualization results.

In the *Network Details* section, we provide additional information about the overall network architecture. In the *Additional Visualization Results* section, we present additional qualitative results on the NYUv2 [6] and Occ-ScanNet [10] datasets, further demonstrating the superiority of our method.

1. Network Details

The framework mainly consists of four components: 2D feature extraction, 2D-to-3D projection, 3D encoder, and 3D decoder.

2D Feature Extraction. 2D Features are first extracted from RGB images using a pre-trained EfficientNet [7] backbone. The depth of each pixel is then predicted with an off-the-shelf method [9] and employed to project the 2D features into 3D space.

2D-3D Projection. In the 2D-to-3D projection, we utilize surface projection [4, 8], which maps 2D features to their corresponding 3D locations using camera intrinsics, extrinsics, and estimated depth. This approach significantly reduces projection errors compared to sight projection [1] and LSS [5]. Additionally, in contrast to query proposal-based projection [3], it offers a reduction in computational cost by lowering the computational burden of the projection operation. Moreover, by projecting only onto the observed surface, this approach greatly reduces the computational cost of the subsequent Transformer-based network.

3D Network. We stack a series of ASFormer Blocks to construct the 3D encoder. Each block is a hybrid structure that integrates attention and convolution operations. The attention module provides a global receptive field, while the convolution module propagates features by incorporating local neighborhood information.

Each block consists of four key components: Adaptive Serialized Attention (ASA), Center-Relative Positional Encoding (CRPE), Convolution-Modulated Layer Normalization (CMLR), and the DDR Block [2]. These components collaboratively enhance receptive-field adaptability, spatial reasoning, cross-module feature alignment, and neighborhood feature completion. It is worth noting that the Center-Relative Positional Encoding (CRPE) is applied only to the first layer of the first block, since it is mainly used to model the richness of the input information.

The 3D decoder is composed of a series of deconvolutional layers. We do not adopt a transformer-based architecture, as the number of non-empty voxels grows substantially at this stage, which makes convolutional decoding more efficient than transformer-based decoding and reduces the computational overhead. Moreover, during the encoder stage, the transformer has already captured sufficient global information.

This design of the ASFormer Block offers three main advantages:

1) Complementary functionality: The Adaptive Serialized Attention (ASA) is employed to capture global contextual features, while the convolutional layers propagate features to spatial neighboring regions, effectively propagating local geometric information.

2) Bridging heterogeneous features: Since Transformers and convolutional networks extract fundamentally different types of features, directly interleaving these modules can result in learning difficulties and limited performance gains. To mitigate this, we propose Convolution-Modulated Layer Normalization, which modulates the Transformer features using convolutional information, ensuring stable and effective feature fusion.

3) Efficiency considerations: By integrating convolution, the Transformer avoids computing attention over all voxels, thereby reducing computational cost.

2. Additional Visualization Results

In this section, we present additional visualization results on the NYUv2 and Occ-ScanNet datasets. As shown in Fig. 1 and Fig. 2, our method outperforms ISO [10] by a substantial margin, owing to the enlarged receptive field. Specifically, as shown in Fig. 1 (fourth row), our method captures the relative spatial and semantic relationships between the table and chairs more effectively, leading to more accurate and complete completion results.

Similarly, as illustrated in Fig. 2 (third row), our method better models the spatial and semantic relationships between various objects and the surrounding walls, thereby achieving more precise and comprehensive scene completion. This more comprehensive scene awareness not only deepens the model's understanding of structural cues but also leads to more reliable predictions in occluded regions.

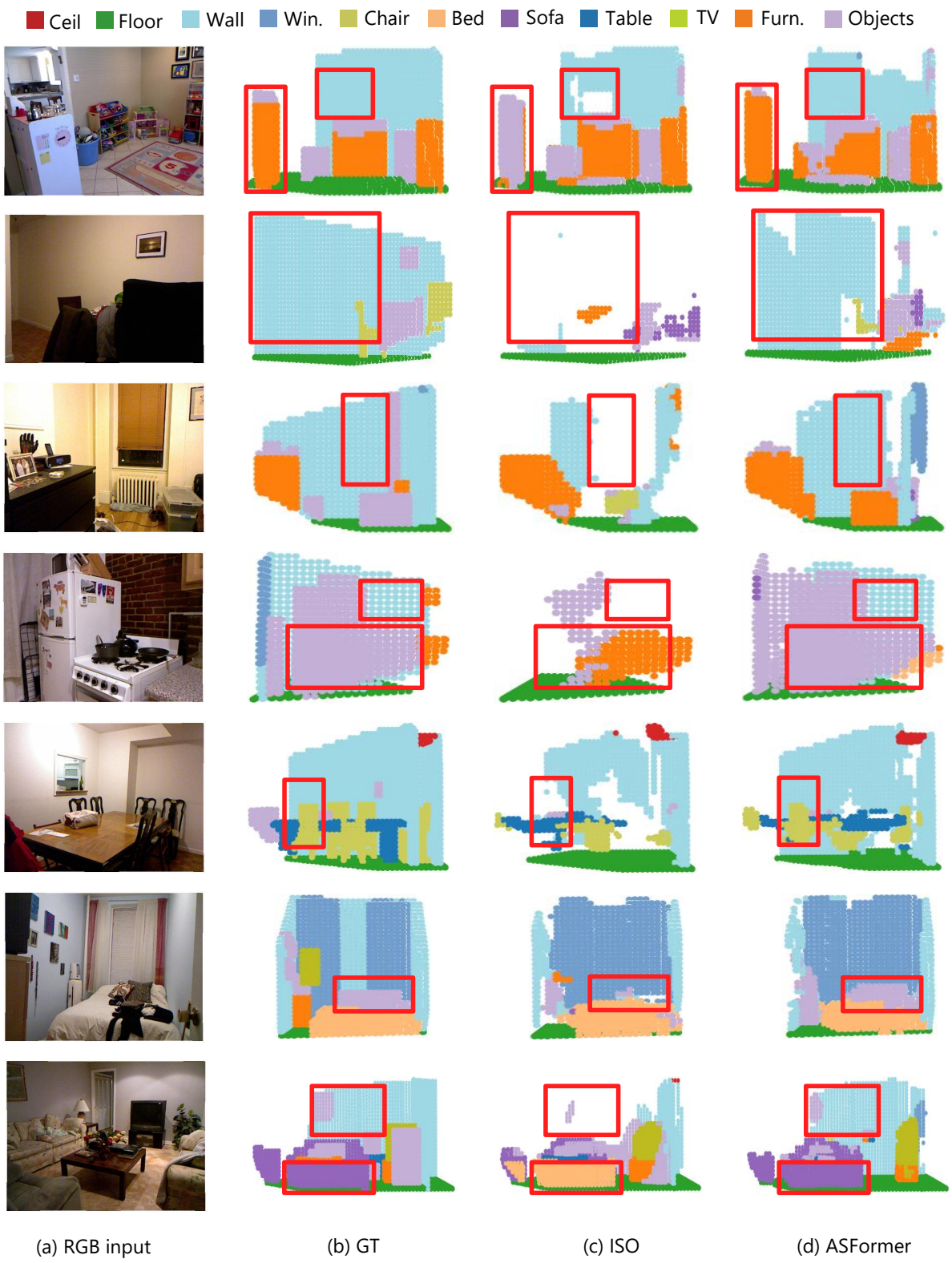

Figure 1. **Qualitative Comparisons** of semantic scene completion results on the NYUv2 testset [6] with different methods. Figures from left to right: (a) The single RGB input; (b) The ground truth; (c) ISO [10]; and (d) Our proposed method.

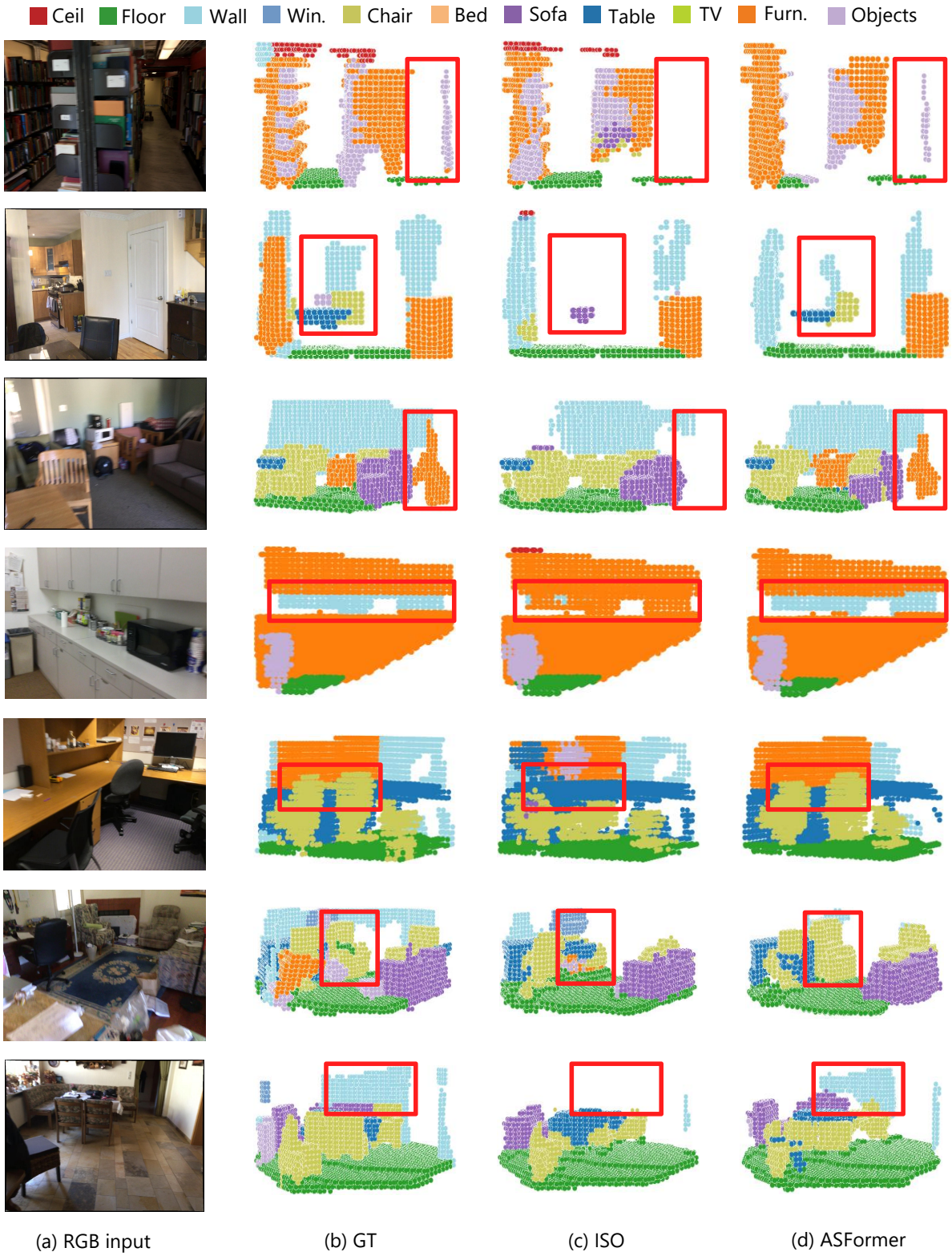

Figure 2. **Qualitative Comparisons** of semantic scene completion results on the OCC-ScanNet testset [10] with different methods. Figures from left to right: (a) The single RGB input; (b) The ground truth; (c) ISO [10]; and (d) Our proposed method.

References

[1] Anh-Quan Cao and Raoul De Charette. MonoScene: Monocular 3D semantic scene completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 1

[2] Jie Li, Yu Liu, Dong Gong, Qinfeng Shi, Xia Yuan, and Chunxia Zhao. RGBD based dimensional decomposition residual network for 3D semantic scene completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7693–7702, 2019. 1

[3] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9087–9098, 2023. 1

[4] Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and think: Disentangling semantic scene completion. In *Advances in Neural Information Processing Systems*, pages 261–272, 2018. 1

[5] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In *European conference on computer vision*, pages 194–210, 2020. 1

[6] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGB-D images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012. 1, 2

[7] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 1

[8] Xuzhi Wang, Xinran Wu, Song Wang, et al. Monocular semantic scene completion via masked recurrent networks. In *IEEE/CVF Conference on International Conference on Computer Vision*, 2025. 1

[9] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 1

[10] Hongxiao Yu, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Monocular occupancy prediction for scalable indoor scenes. In *European Conference on Computer Vision*, pages 38–54, 2024. 1, 2, 3