Monocular Semantic Scene Completion via Masked Recurrent Networks

Supplementary Material

In this supplementary material, we first detail the architecture of the 3D network used in the initial MSSC. Next, we provide a qualitative comparison of the NYUv2 and SemanticKITTI datasets [1]. We then include an ablation study on mask updating.

1. Details on the 3D network in Initial SSC

Our method first performs an initial SSC prediction and then carries out the recurrent refinement. In this section, we detail the 3D network architecture in the initial SSC stage as shown in Fig. 1.

The projected 3D features are first passed through an encoder, which includes two AIC [5] blocks and a channel-wise attention (CA) module [3, 4]. Each block is composed of four AIC modules. An AIC module could model various objects or stuff with severe variations in shapes and layouts by an anisotropic receptive field. The channel-wise attention module is placed between the two AIC blocks to reweight and capture the channel-wise dependencies. After encoding, two deconvolution layers are applied to upsample the features to the original resolution of the input. Next, the features are fed into the SSC head to obtain the semantic scene completion results.

2. Qualitative Comparisons

Fig. 2 and Fig. 3 show the qualitative comparison on the NYUv2 and SemanticKITTI datasets. Our method outperforms MonoScene in the occluded regions and more effectively recovers fine-grained details. We choose MonoScene [2] as a baseline for comparison because it targets the same type of scenarios as our method, providing a unified approach for both indoor and outdoor environments.

From Fig. 2 and Fig. 3, it is evident that our method demonstrates superior capability in recovering various object categories, showcasing a stronger ability for information restoration. MonoScene struggles to accurately complete the semantic details of objects, often resulting in incomplete or imprecise reconstructions. In contrast, our method effectively captures finer details and restores more comprehensive semantic information, leading to a more accurate and visually coherent scene representation.

3. Different Design Choices of Mask Updating

In this section, we evaluate the effectiveness of different design choices for mask updating as shown in Table 1.

Mask Updating Module. The mask updating module is proposed to sequentially update the mask M. We experiment with omitting the mask updating module. We can observe that it boosts 0.62% mIoU by adding the mask updating module.

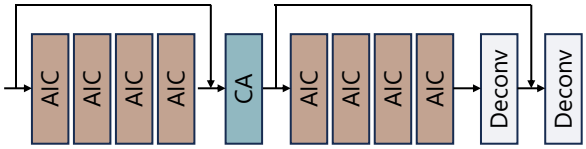

Figure 1. Details on the 3D network in initial SSC.

Table 1. Ablation study on the different design choices of the Mask Updating module in MonoMRN.

Design Choices	SC-IoU(%)	SSC-mIoU(%)
With Mask Updating Module	52.33	30.11
W/O Mask Updating Module	**53.16**	**30.73**
With Mask Initialization	51.26	29.62
W/O Mask Initialization	**53.16**	**30.73**
With Mask Loss	51.66	29.67
W/O Mask Loss	**53.16**	**30.73**

Mask Initialization. In the early stages of training, the mask predictions from the mask updating module are of low quality and fluctuate dramatically, causing the model to focus on inaccurate occupied regions. We tested a version that only used the mask updating module. Mask initialization could obtain 1.11% performance gain.

Mask Loss. We introduce mask loss to provide supervision of the occupied regions. We can observe that mask loss enhances the performance by 1.06% mIoU.

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 International Conference on Computer Vision*, 2019. 1

[2] Anh-Quan Cao et al. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 3

[3] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 1

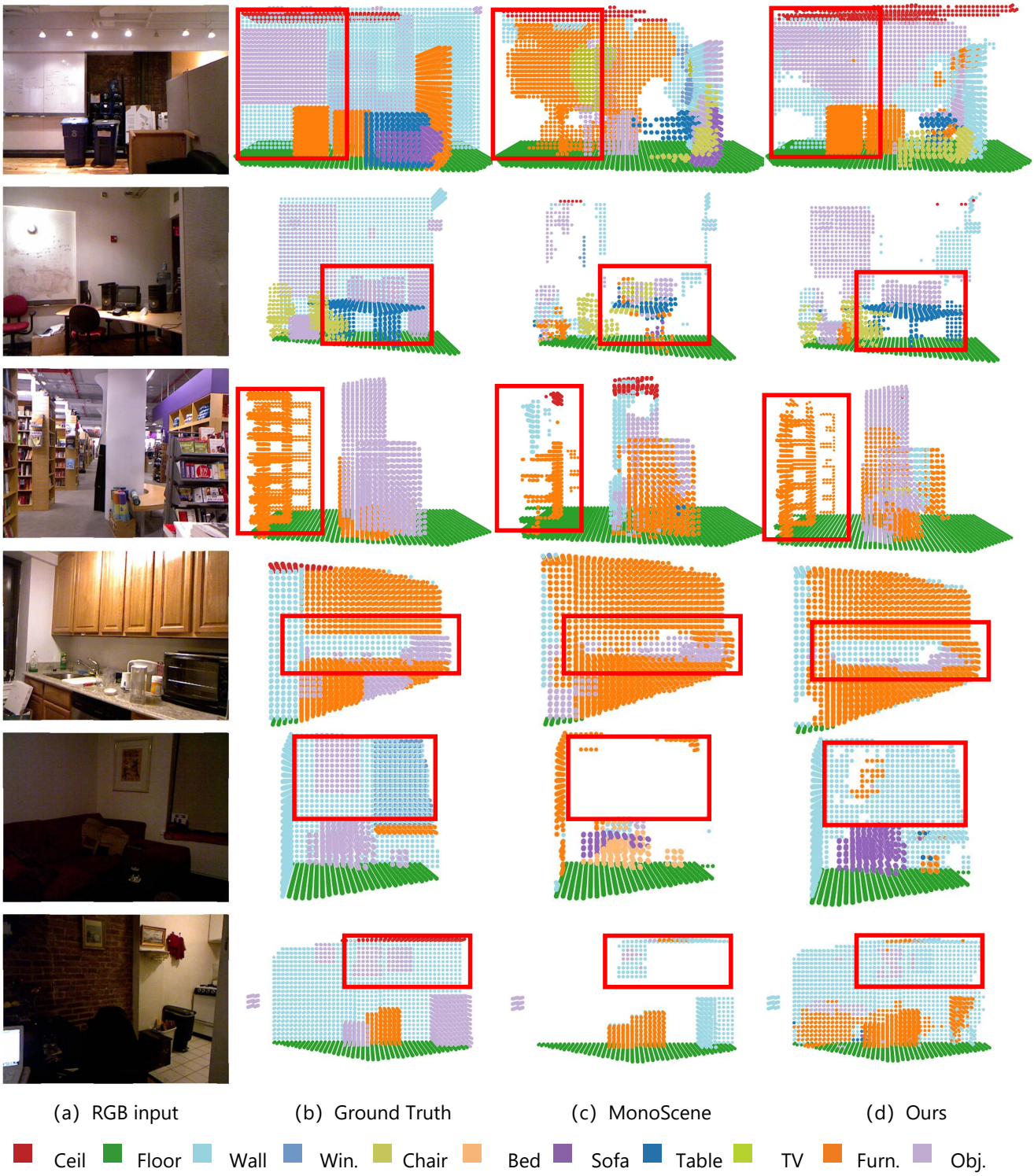

Figure 2. Qualitative comparison on NYUv2. The leftmost column presents the input RGB images, while the subsequent columns sequentially show the results of Ground Truth, MonoScene [2], and our method.

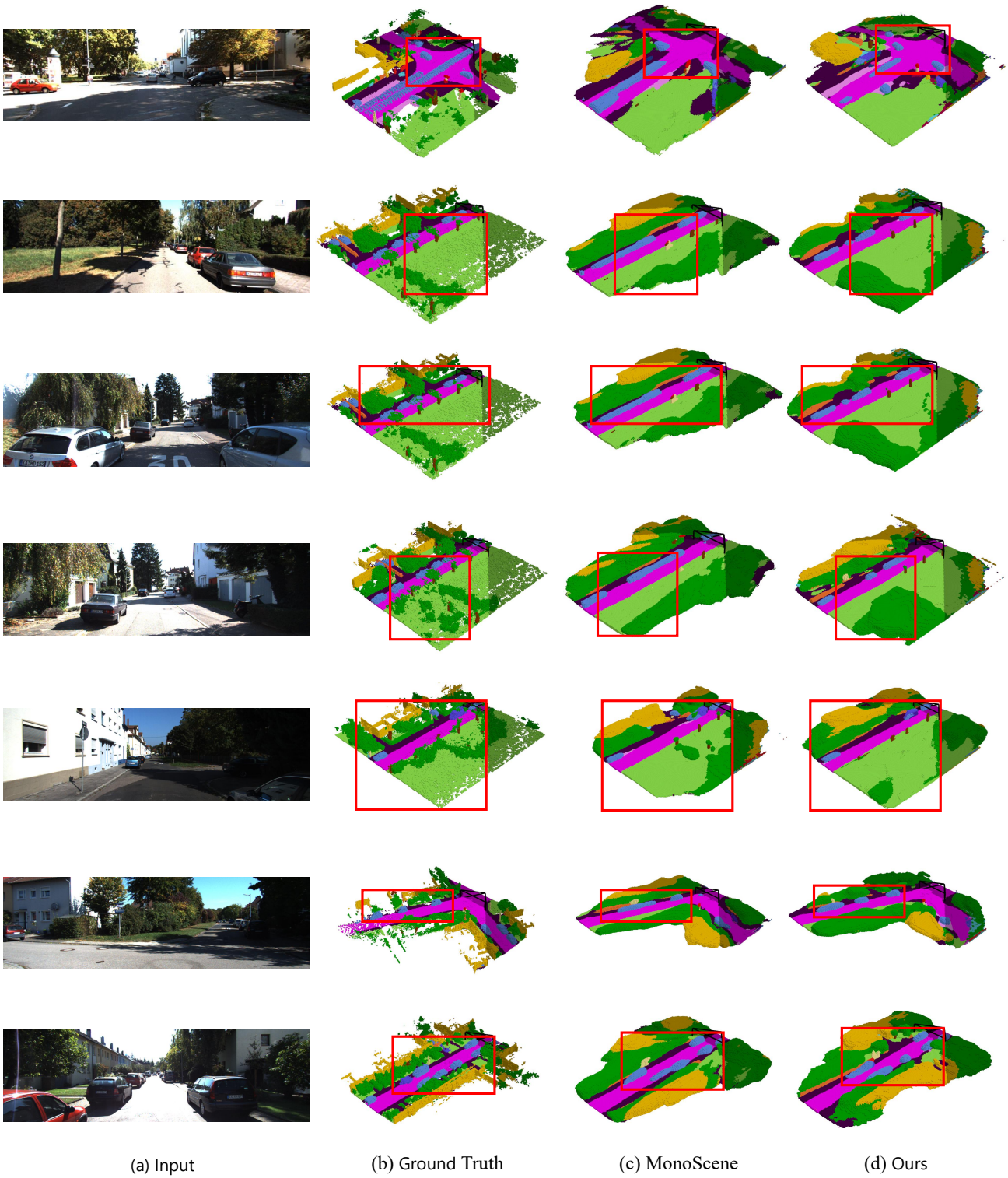

Figure 3. Qualitative comparison on SemanticKITTI validation set. The leftmost column presents the input RGB images, while the subsequent columns sequentially show the results of Ground Truth, MonoScene [2], and our method.

[4] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 1

[5] 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*, 2020. 1