SPSG: Self-Supervised Photometric Scene Generation from RGB-D Scans –Supplementary Material–

Angela Dai[1] Yawar Siddiqui[1] Justus Thies[1] Julien Valentin[2] Matthias Nießner[1]
[1]Technical University of Munich [2]Google

A. Network Architecture

We detail our network architecture specifications in Figure 1. Convolution parameters are given as (nf_in, nf_out, kernel_size, stride, padding). Each convolution (except those producing final outputs for geometry and color) is followed by a Leaky ReLU and batch normalization.

B. Additional Results

B.1. Additional Ablation Studies

We additionally evaluate the effect of the CIELAB color space that our approach uses for color generation, in comparison to RGB space. Table 1 quantitatively evaluates the color generation, showing that CIELAB space is more effective, and Figure 2 shows that using CIELAB space allows our approach to capture a greater diversity of colors in our output predictions.

We also evaluate the geometric reconstruction when trained with a 3D ℓ_1 loss only in Table 2; here, Baseline-3D can improve on an ℓ_1 loss only, and ours leverages the advantages of a view-guided synthesis for the best reconstruction performance.

4

B.2. Runtime Performance

Since our network architecture is composed of 3D convolutions, we can generate an output prediction in a single forward pass for an input scan, with runtime performance dependent on the 3D volume of the test scene as $\mathcal{O}(\text{dimx}\times\text{dimy}\times\text{dimz})$. A small scene of size $1.5\times3.0\times2.6$ meters ($72 \times 152 \times 128$ voxels), inference time is 0.33 seconds; a medium scene of size $2.8 \times 3.9 \times 2.6$ meters ($140 \times 196 \times 128$ voxels) takes 0.86 seconds, and a large scene of size $6.0\times6.6\times2.6$ meters ($300\times328\times128$ voxels) takes 2.4 seconds.

Memory footprint During training, our approach operates with a memory footprint of 5.5GB with a batch size of 2. At test time, the memory footprint of the small, medium, and large scenes previously mentioned is 0.7GB, 1.7GB, and 6GB respectively. Very large test scenes can be realized by running our method by chunks of the receptive field size; for instance, our largest test scene spans $34.5 \times 49.2 \times 2.6$ meters ($1727\times2461\times128$ voxels), with a memory footprint of 1.4GB in this fashion.

B.3. Qualitative Results

We provide additional qualitative results of colored reconstruction of Matterport3D [1] scans and ShapeNet [2] chairs in Figures 3 and 4, respectively. As can be seen, our method consistently generates sharper results compared to the baseline methods. In Figure 3, the comparison to [4] is shown. Since the approach does not complete geometry, we provide our predicted geometry as input. In contrast to our method, it is not properly estimating color tones like for the green chair in the bottom row of the figure. Figure 4 shows more examples for our experiments on the ShapeNet dataset in comparison to Im2Avatar [6], PIFu [5] and Texture Fields [4].

Additionally, we show qualitative comparisons of the geometric completion of our approach on Matterport3D [1] scans in Figure 5, in comparison to SG-NN [3]. Both methods were trained on our Matterport3D chunks data, where inputs were composed of 30% of frames available, and targets of 50% of frames available; test target scenes are visualized with all available frames. The direct 3D supervision guiding SG-NN contains fused errors from small camera estimation misalignments and depth noise (small shifts in the target TSDF), resulting in a tendency to produce a few visible seams in the resulting reconstructions. In contrast, our view-guided synthesis helps to avoid these artifacts, producing more compelling scene geometry.

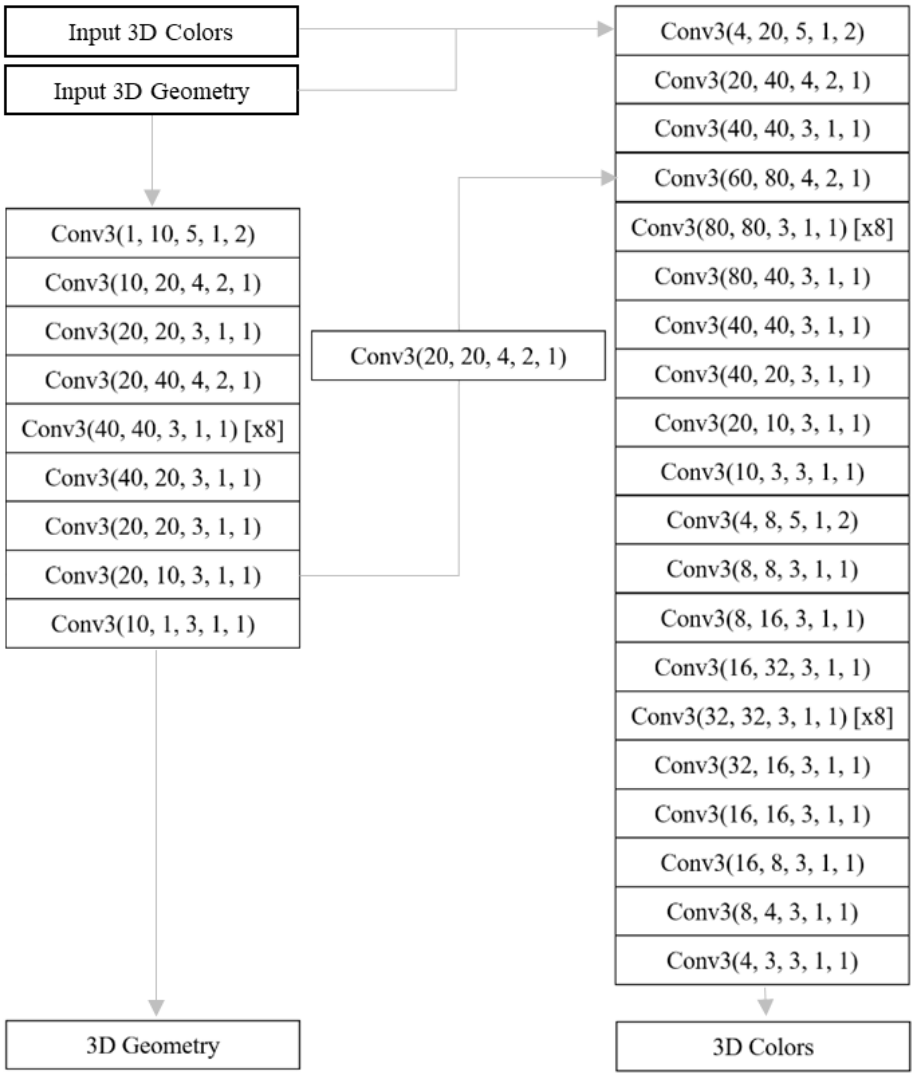

Figure 1: Network architecture specification. Given an incomplete RGB-D scan, we take its 3D geometry and color as input, and leverage a fully-convolutional neural network to predict the complete 3D model represented volumetrically for both geometry and color.

Method	SSIM (↑)	Feature-ℓ_1 (↓)	FID (↓)
Using RGB	0.702	0.222	58.8
Ours	**0.709**	**0.219**	**56.03**

Table 1: Comparison of our approach using CIELAB color space to using RGB on Matterport3D [1] scans. CIELAB produces more effective color generation.

References

[1] Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*, pages 667–676, 2017. 1, 2, 3, 4, 6

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 5

[3] Angela Dai, Christian Diller, and Matthias Nießner. Sg-nn:

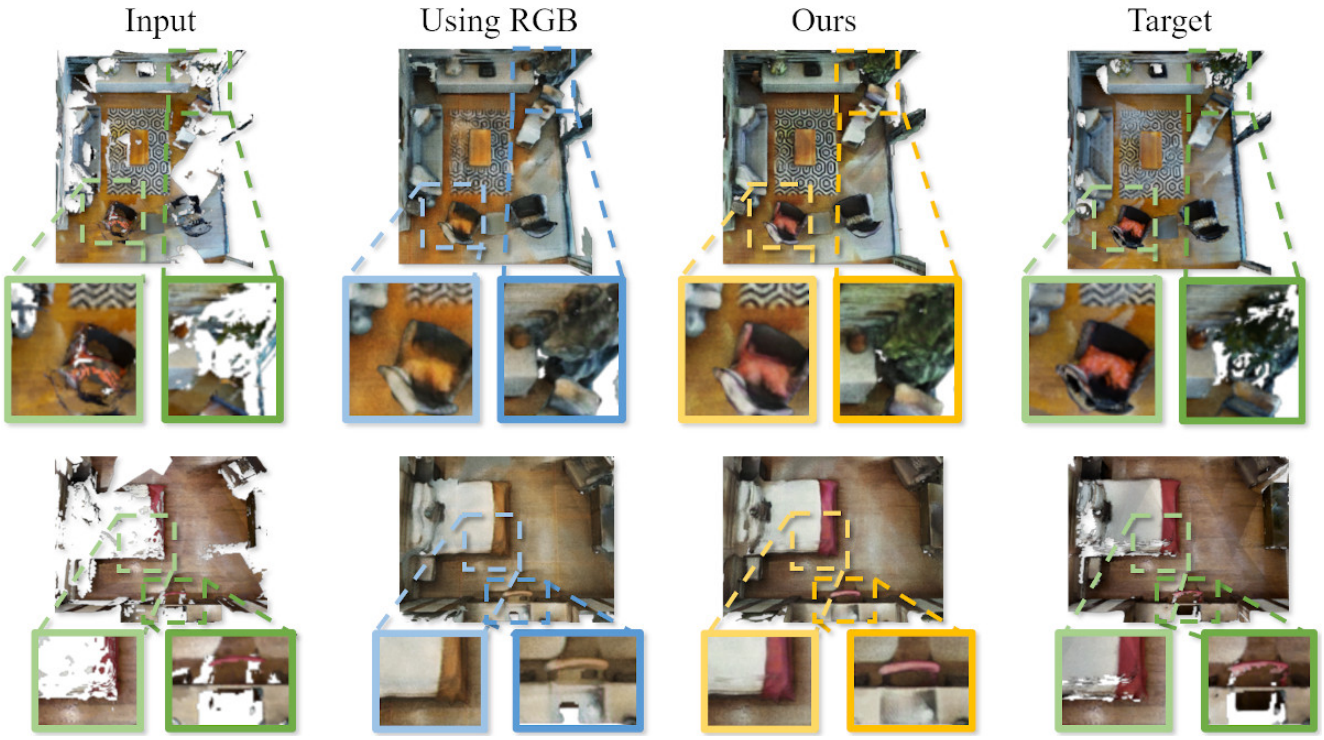

Figure 2: Qualitative comparison of our approach using CIELAB color space vs RGB color space on Matterport3D [1] scans. Using CIELAB space allows us to capture more diversity in output color generation.

	Matterport3D		
Method	IoU (↑)	Recall (↑)	Chamfer Dist. (↓)
3D ℓ_1 loss only	0.31	0.58	0.02
Baseline-3D	0.33	0.58	0.04
Ours	**0.39**	**0.64**	**0.01**

Table 2: Additional ablations on geometric reconstruction from Matterport3D [1] scans.

Sparse generative neural networks for self-supervised scene completion of rgb-d scans. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2020. 1

[4] Michael Oechsle, Lars Mescheder, Michael Niemeyer, Thilo Strauss, and Andreas Geiger. Texture fields: Learning texture representations in function space. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4531–4540, 2019. 1, 5

[5] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2304–2314, 2019. 1, 5

[6] Yongbin Sun, Ziwei Liu, Yue Wang, and Sanjay E Sarma. Im2avatar: Colorful 3d reconstruction from a single image. *arXiv preprint arXiv:1804.06375*, 2018. 1, 5

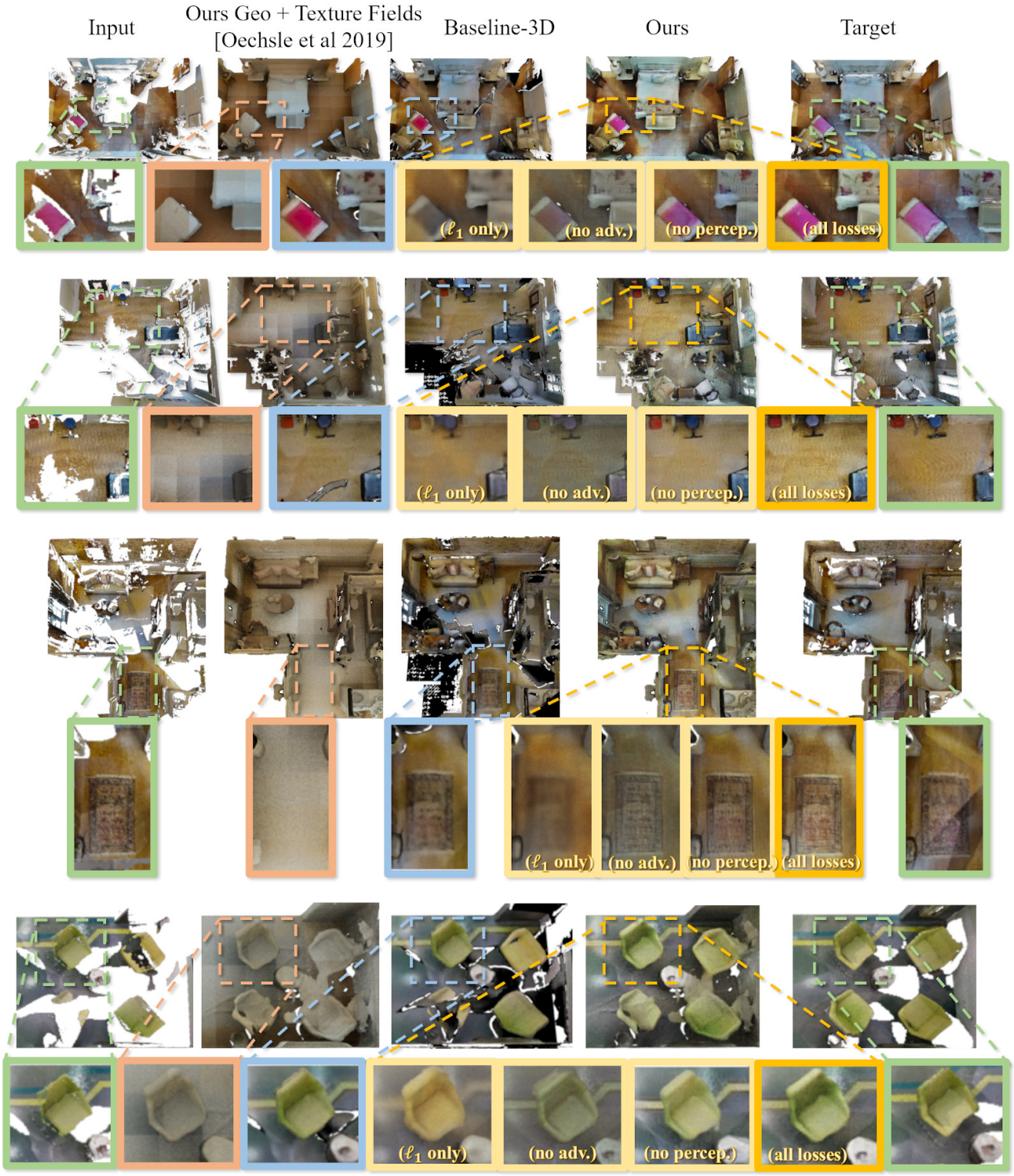

Figure 3: Additional qualitative evaluation of colored reconstruction on Matterport3D [1] scans.

Figure 4: Additional qualitative evaluation of colored reconstruction of our method against Im2Avatar [6], PIFu [5], and Texture Fields [4] (run on geometry predicted by our method) on ShapeNet [2] chairs.

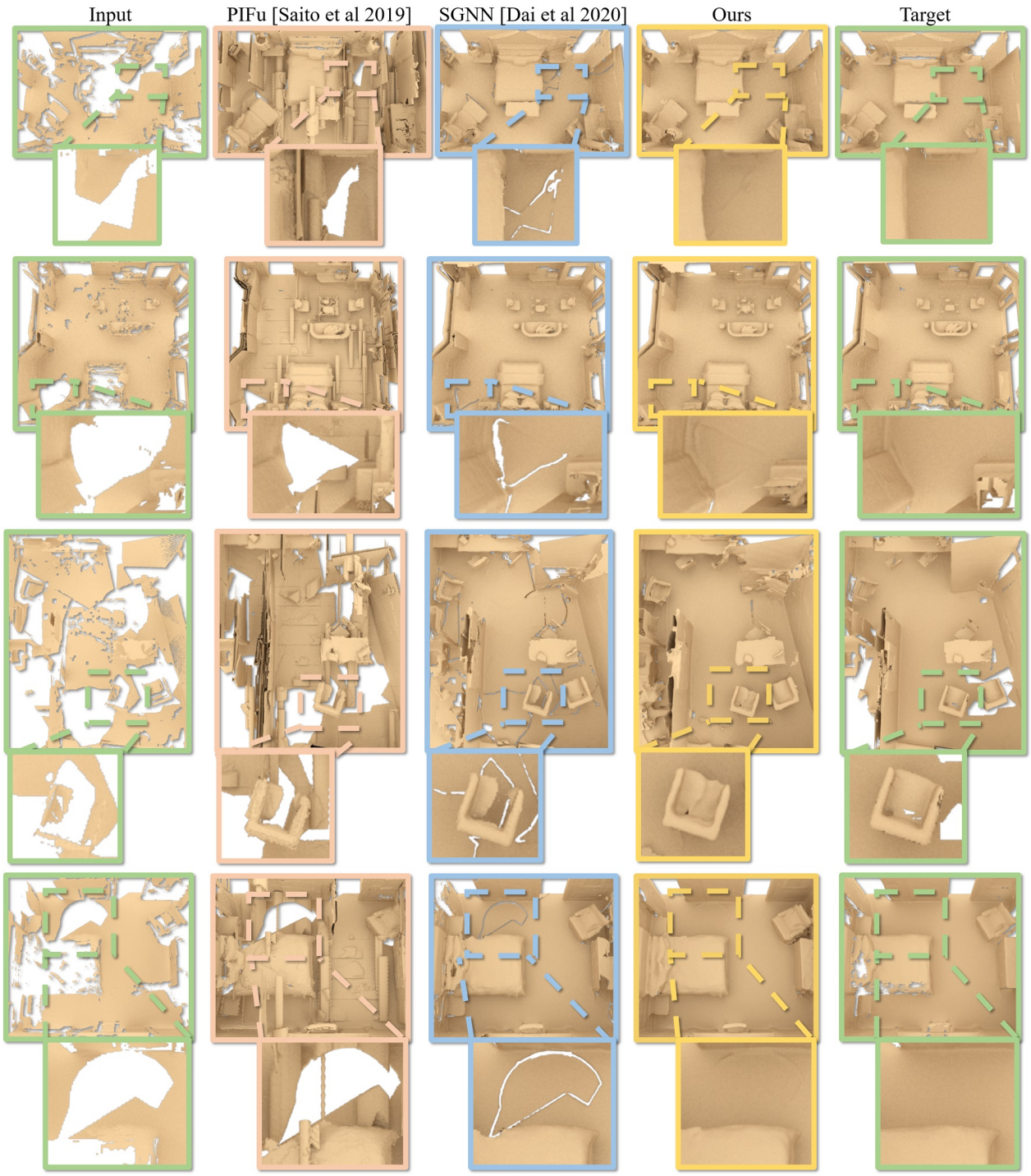

Figure 5: Qualitative comparison of geometric completion results on Matterport3D [1] scans. Our view-guided approach mitigates learning from artifacts in the fused 3D reconstruction of the scenes (e.g., small frame misalignments, which can cause seams such as in the SG-NN reconstructions), producing more accurate scene geometry.