Supplemental Material for ScanComplete: Large-Scale Scene Completion and Semantic Segmentation for 3D Scans

Angela Dai[1,3,5] Daniel Ritchie[2] Martin Bokeloh[3] Scott Reed[4] Jürgen Sturm[3] Matthias Nießner[5]
[1]Stanford University [2]Brown University [3]Google [4]DeepMind [5]Technical University of Munich

In this supplemental document, we provide additional details for our ScanComplete submission. First, we show a qualitative evaluation on real-world RGB-D data; see Sec. 1. Second, we evaluate our semantics predictions on real-world benchmarks; see Sec. 2. Further, we provide details on the comparisons to Dai et al. [3] in Sec. 3 and visualize the subvolume blocks used for the training of our spatially-invariant network in Sec. 4. In Sec. 5, we compare the timings of our network against previous approaches showing that we not only outperform them in terms of accuracy and qualitative results, but also have a significant runtime advantage due to our architecture design. Finally, we show additional results on synthetic data for completion and semantics in Sec. 6.

1. Qualitative Evaluation Real Data

In Fig. 3 and Fig. 4, we use our network which is trained only on the synthetic SUNCG set, and use it infer missing geometry in real-world RGB-D scans; in addition, we infer per-voxel semantics. We show results on several scenes on the publicly-available ScanNet [2] dataset; the figure visualizes real input, completion (synthetically-trained), semantics (synthetically-trained), and semantics (synthetically pre-trained and fine-tuned on the ScanNet annotations).

2. Quantitative Evaluation on Real Data

For evaluation of semantic predictions on real-world scans, we provide a comprehensive comparison on the ScanNet [2] and Matterport3D [1] datasets, which both have ground truth per-voxel annotations. The results are shown in Tab. 1. We show results for our approach that is only trained on the synthetic SUNCG data; in addition, we fine-tune our semantics-only network on the respective real data. Unfortunately, fine-tuning on real data is challenging when using a distance field representation given that the ground truth data is incomplete. However, we can use pseudo-ground truth when leaving out frames and corresponding it to a more (but still not entirely) complete reconstruction when using an occupancy grid representation. This strategy works on the Matterport3D dataset, as we have relatively complete scans to begin with; however, it is not applicably to the more incomplete ScanNet data.

3. Comparison Encoder-Predictor Network

In Fig. 1, we visualize the problems of existing completion approach by Dai et al. [3]. They propose a 3D encoder-predictor network (3D-EPN), which takes as input a partial scan of an object and predicts the completed counterpart. Their main disadvantage is that block predictions operate independently; hence, they do not consider information of neighboring blocks, which causes seams on the block boundaries. Even though the quantitative error metrics are not too bad for the baseline approach, the visual inspection reveals that the boundary artifacts introduced at these seams are problematic.

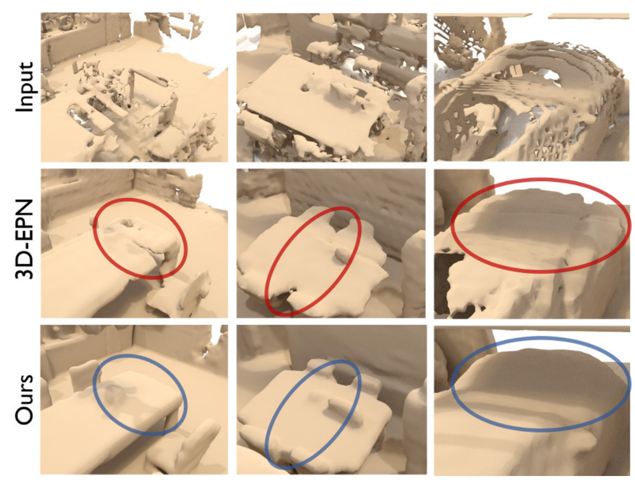

Figure 1. Applying the 3D-EPN approach [3] to a scene by iteratively, independently predicting fixed-size subvolumes results in seams due to inconsistent predictions. Our approach, taking the entire partial scan as input, effectively alleviates these artifacts.

4. Training Block Pairs

In Fig. 2, we visualize the subvolumes used for training our fully-convolutional network on the three hierarchy

ScanNet												
	bed	ceil.	chair	floor	furn.	obj.	sofa	table	tv	wall	wind.	avg
ScanNet [2]	**60.6**	47.7	**76.9**	**90.8**	**61.6**	**28.2**	75.8	67.7	6.3	81.9	25.1	56.6
Ours (SUNCG)	42.6	69.5	53.1	70.9	23.7	20.0	**76.3**	63.4	**29.1**	57.0	26.9	48.4
Ours (ft. ScanNet; sem-only)	52.8	**85.4**	60.3	90.2	51.6	15.7	72.5	**71.4**	21.3	**88.8**	**36.1**	**58.7**

Matterport3D												
	bed	ceil.	chair	floor	furn.	obj.	sofa	table	tv	wall	wind.	avg
Matterport3D [1]	**62.8**	0.1	20.2	92.4	**64.3**	17.0	27.7	10.7	**5.5**	76.4	15.0	35.7
Ours (Matterport3D; sem-only)	38.4	93.2	**62.4**	94.2	33.6	**54.6**	15.6	**40.2**	0.7	51.8	**38.0**	47.5
Ours (Matterport3D)	41.8	**93.5**	58.0	**95.8**	38.3	31.6	**33.1**	37.1	0.01	**84.5**	17.7	**48.3**

Table 1. Semantic labeling accuracy on real-world RGB-D. Per-voxel class accuracies on Matterport3D [1] and ScanNet [2] test scenes. We can see a significant improvement on the average class accuracy on the Matterport3D dataset.

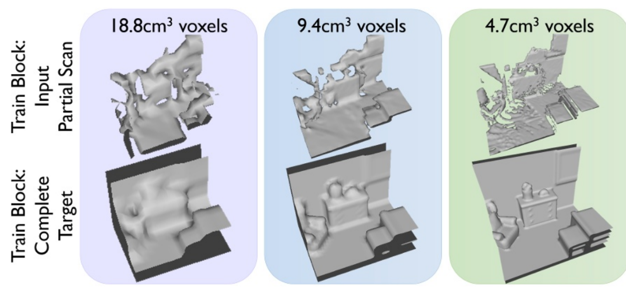

Figure 2. Subvolume train-test pairs of our three hierarchy levels.

levels of our network. By randomly selecting a large variety of these subvolumes as ground truth pairs for training, we are able train our network such that it generalizes to varying spatial extents at test time. Note again the fully-convolutional nature of our architecture, which allow the precessing of arbitrarily-sized 3D environments in a single test pass.

5. Timings

We evaluate the run-time performance of our method in Tab. 2 using an Nvidia GTX 1080 GPU. We compare against the baseline 3D-EPN completion approach [3], as well as the ScanNet semantic voxel prediction method [2]. The advantage of our approach is that our fully-convolutional architecture can process and entire scene at once. Since we are using three hierarchy levels and an auto-regressive model with eight voxel groups, our method requires to run a total of 3×8 forward passes; however, note again that each of these passes is run over entire scenes. In comparison, the ScanNet voxel labeling method is run on a per-voxel column basis. That is, the $x-y$-resolution of the voxel grid determines the number of forward passes, which makes its runtime significantly slower than our approach even though the network architecture is less powerful (e.g., it cannot address completion in the first place).

The original 3D-EPN completion method [3] operates on a 32^3 voxel grid to predict the completion of a single model. We adapted this approach in to run on full scenes; for efficiency reasons we change the voxel resolution to $32 \times 32 \times 64$ to cover the full height in a single pass. This modified version is run on each block independently, and requires the same number of forward passes than voxel blocks. In theory, the total could be similar to one pass on a single hierarchy level; however, the separation of forward passes across several smaller kernel calls – rather than fewer big ones – is significantly less efficient on GPUs (in particular on current deep learning frameworks).

6. Additional Results on Completion and Semantics on SUNCG

Fig. 5 shows additional qualitative results for both completion and semantic predictions on the SUNCG dataset [4]. We show entire scenes as well as close ups spanning a variety of challenging scenarios.

References

[1] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 1, 2

[2] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 1, 2, 3, 4

[3] A. Dai, C. R. Qi, and M. Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 1, 2, 3

[4] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. *Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 4

	#Convs	Scene Size (voxels)			
		$82 \times 64 \times 64$	$100 \times 64 \times 114$	$162 \times 64 \times 164$	$204 \times 64 \times 222$
3D-EPN [3]	8 + 2fc	20.4	40.4	79.6	100.5
ScanNet [2]	9 + 2fc	5.9	19.8	32.5	67.2
Ours (base level)	32	0.4	0.4	0.6	0.9
Ours (mid level)	42	0.7	1.3	2.2	4.7
Ours (high level)	42	3.1	7.8	14.8	31.6
Ours (total)	-	4.2	9.5	17.6	37.3

Table 2. Time (seconds) to evaluate test scenes of various sizes measured on a GTX 1080.

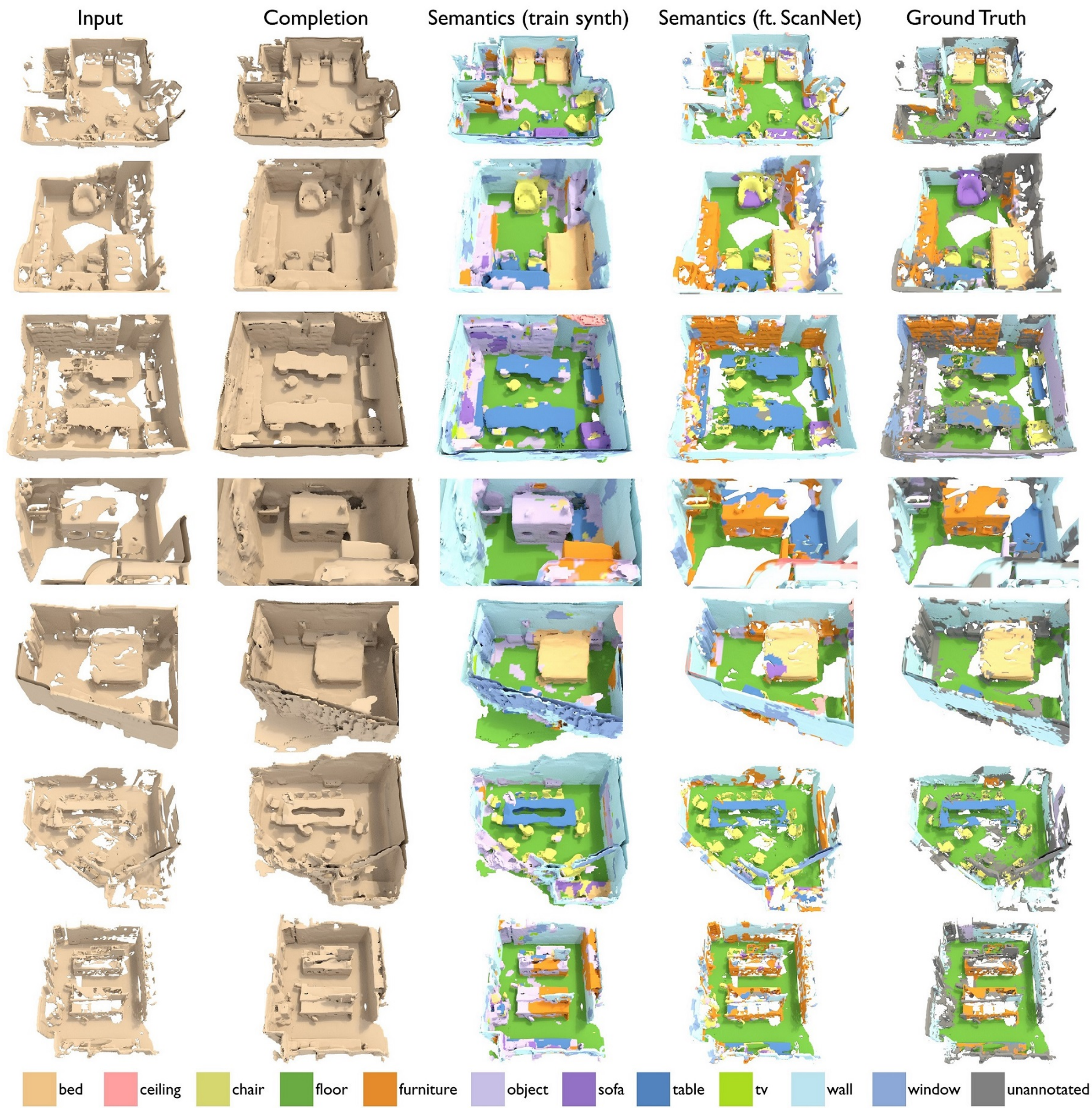

Figure 3. Additional results on ScanNet for our completion and semantic voxel labeling predictions.

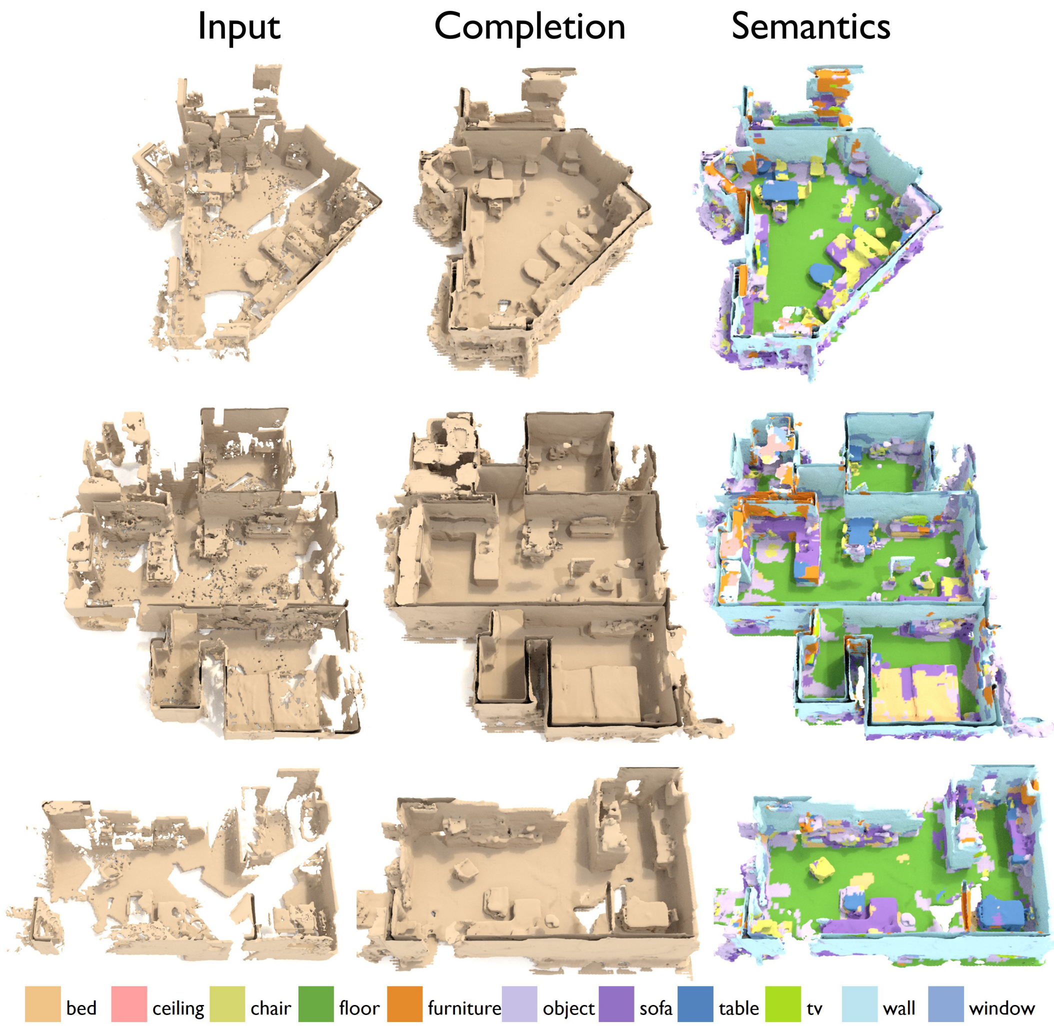

Figure 4. Additional results on Google Tango scans for our completion and semantic voxel labeling predictions.

	bed	ceil.	chair	floor	furn.	obj.	sofa	table	tv	wall	wind.	avg
ScanNet [2]	11.7	88.7	13.2	81.3	11.8	13.4	25.2	18.7	4.2	53.5	0.5	29.3
SSCNet [4]	33.1	42.4	21.4	42.0	24.7	8.6	39.3	25.2	13.3	47.7	24.1	29.3
Ours	**50.4**	**95.5**	**35.3**	**89.4**	**45.2**	**31.3**	**57.4**	**38.2**	**16.7**	**72.2**	**33.3**	**51.4**

Table 3. Semantic labeling on SUNCG scenes, measured as IOU per class over the visible surface of the partial test scans.

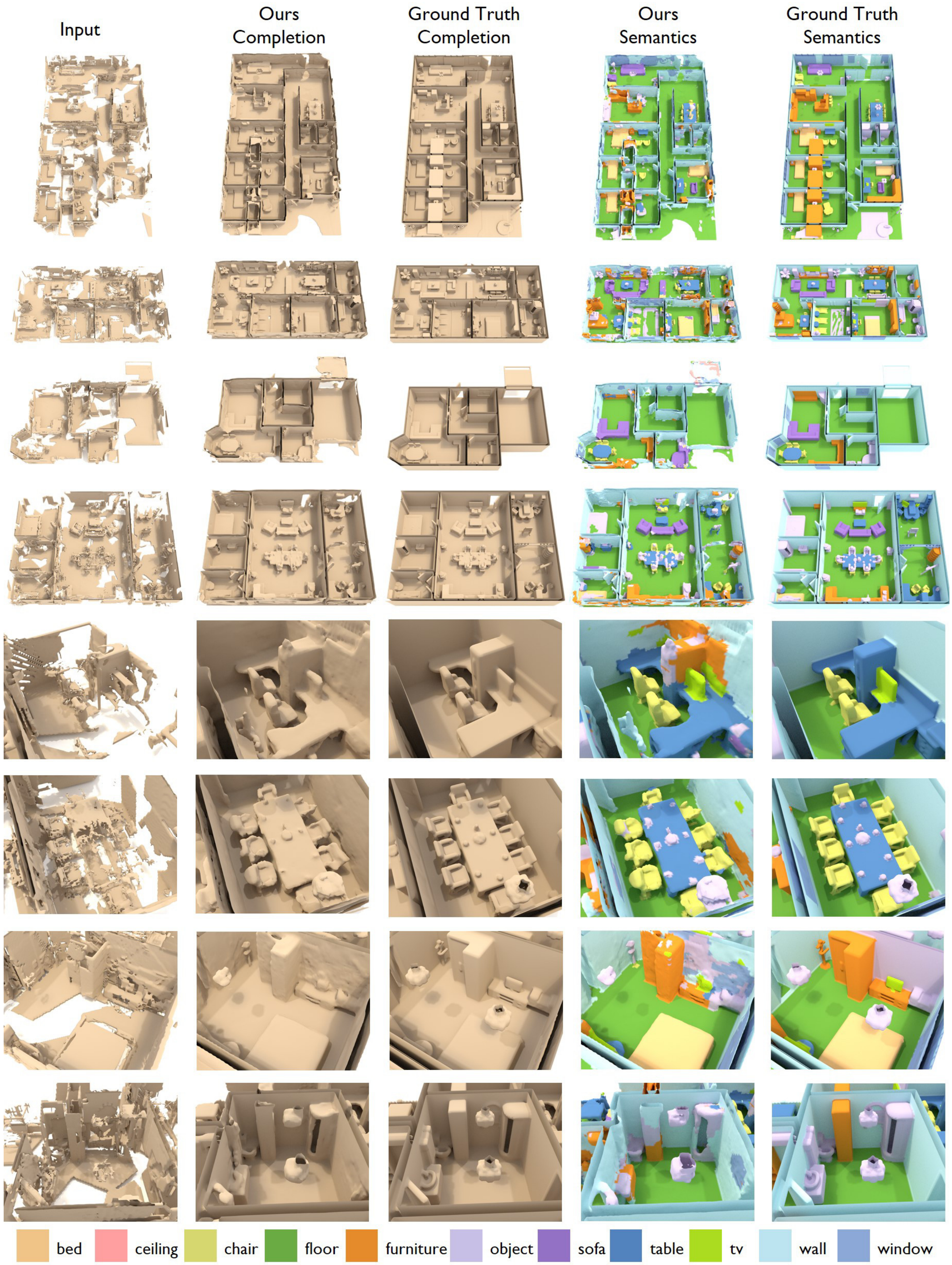

Figure 5. Additional results on SUNCG for our completion and semantic voxel labeling predictions.