SA-ConvONet: Sign-Agnostic Optimization of Convolutional Occupancy Networks

–Supplementary Material–

Jiapeng Tang[1,4], Jiabao Lei[1], Dan Xu[2], Feiying Ma[4], Kui Jia[1,5,6], and Lei Zhang[3,4]

[1]School of Electronic and Information Engineering, South China University of Technology
[2]Department of Computer Science and Engineering, HKUST, HK
[3]Department of Computing, The Hong Kong Polytechnic University, HK
[4]DAMO Academy, Alibaba Group
[5]Pazhou Lab, Guangzhou, China
[6]Peng Cheng Laboratory, Shenzhen, China

In this supplementary material, we provide more details about our network architecture in Section A. Then we present ablation studies to validate the effectiveness of each design in our approach in Section B. In the next, we demonstrate the generalization capabilities of our approach to novel categories that are different from the training category ("chair") in Section C. Finally, we show more qualitative comparison with other competitive methods on the real-world 3D scene datasets in Section D.

A. Network Architectures

PointNet: The detailed network architecture of PointNet used in the paper is depicted in Figure 1. Firstly, we map the coordinates of $\mathcal{P}$ into the feature space using a fully-connected (FC) layer and a ResNet-FC [7] block. Then, instead of using a global pooling operation to obtain a global feature like [4], we perform the grid-pooling operation [10] to locally fuse the extracted features. Specifically, we perform an average-pooling operation for the features that are within the same voxel cell from a volumetric grid with the size of 64^3. Next, we concatenate the locally pooled features with the features before pooling, and then feed the formed features into the subsequent ResNet-FC block. Overall, we use 5 ResNet blocks with intermediate grid-pooling layers to obtain the point-wise features $\mathbf{F}_0$.

3D U-Net: The network architecture of 3D-UNet is illustrated in Figure 2. The 3D U-Net [13] is used to aggregate both local and global information of the volumetric feature $\mathbf{V}_0$ that is transformed from $\mathbf{F}_0$. The dimensions of input and output features are both set to 64. To ensure that the receptive field is equal to or larger than the size of the input feature volume, the depth of the 3D U-Net is set to 4.

Occupancy Decoder: As shown in Figure 3, the occupancy decoder consists of 5 stacked ResNet-FC blocks with skip connections. And the hidden feature dimension is set to 32.

B. Ablation studies

In this section, we conduct additional ablation studies by alternatively removing one of the modules of the proposed approach to verify the effectiveness of them.

Effect of pre-training (*i.e.* w/o pre-training) Based on our approach, an alternative solution to provide initialization of the signed field for the proposed sign-agnostic optimization is to adopt the geometric initialization as in SAL [1], which initializes the implicit decoder to approximate the signed distance field of the unit sphere. The visualization comparisons are shown in Figure 4. Without the pre-trained shape prior, the sign agnostic optimization fails to reconstruct reasonable geometries.

Datasets	Methods	CD ↓	NC ↑	FS (τ) ↑	FS (2τ) ↑
ShapaNet -chair [3]	opt. enc.	**0.516**	93.42	97.15	**99.40**
	Ours	0.522	**93.51**	**97.16**	99.37
Synethetic Room [12]	opt. enc.	0.516	89.75	93.43	98.53
	Ours	**0.495**	**90.04**	**93.85**	**98.82**
ScanNet [5]	opt. enc.	0.741	86.24	81.49	95.56
	Ours	**0.728**	**86.40**	**82.08**	**95.86**

Table 1: **Additional ablation studies** on three datasets.

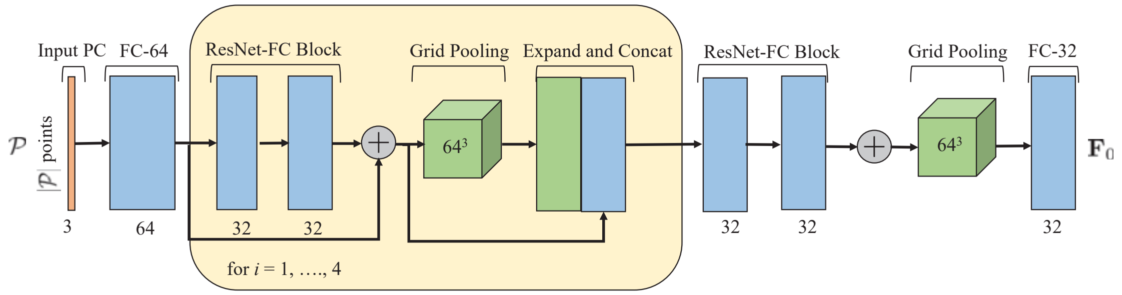


Figure 1: **ResNet [7] variants of PointNet[4].** It utilizes a stack of five ResNet-FC blocks with skip connections and grid-pooling layers to extract point-wise features $\mathbf{F_0}$ from the observed surface point cloud $\mathcal{P}$.

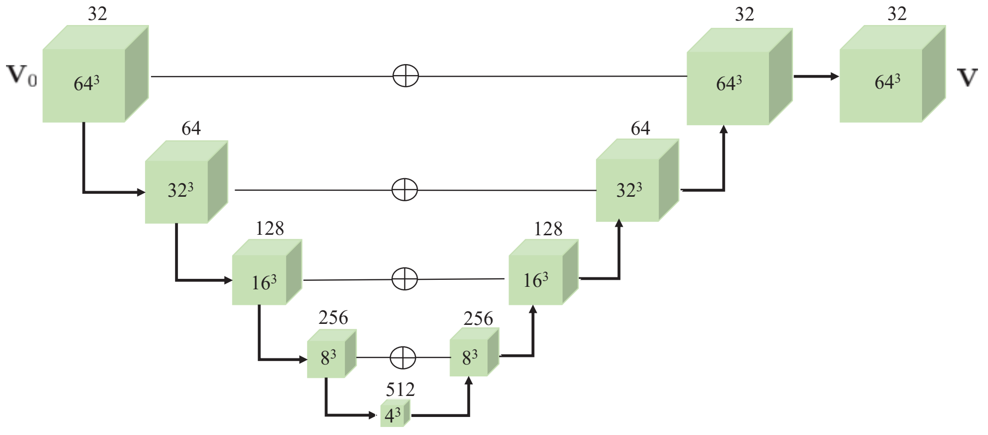


Figure 2: **3D U-Net.** To effectively fuse the global and local information of input shape, we transform $\mathbf{V}_0$ (produced from $\mathbf{F}_0$) to $\mathbf{V}$ using a 3D U-Net, which consists of a series of 3D down- and up-sampling convolutions with skip connections.

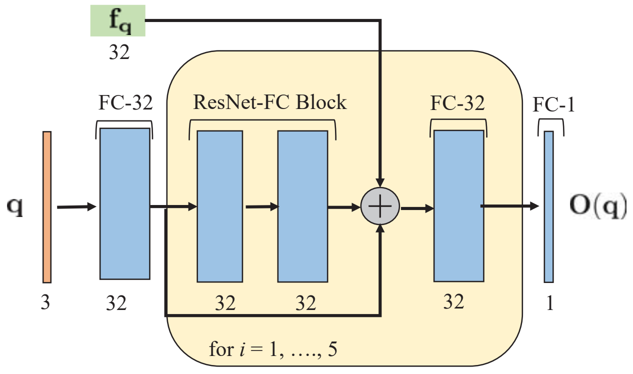


Figure 3: **Occupancy Decoder.** It contains five ResNet-FC blocks with skip connections. Given a point $\mathbf{q}$ randomly sampled in the 3D space, we query a feature vector $\mathbf{f_q}$ from the feature volume $\mathbf{V}$ according to the location of $\mathbf{q}$. Then we pass $\mathbf{q}$ and $\mathbf{f_q}$ into the occupancy decoder to predict the occupancy probability of $\mathbf{q}$ (*i.e.* $\mathbf{O_q}$).

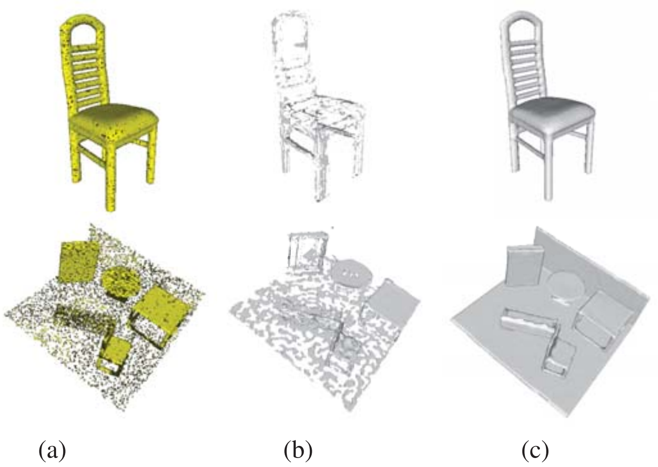


Figure 4: **Additional Qualitative Ablation Studies:** (a) input point clouds, (b) without the pre-training of convolutional occupancy networks, and (c) Ours.

Effect of only optimizing the encoder (*i.e.* opt. enc.) In all experiments, we choose to optimize the whole network parameters with the unsigned binary cross-entropy loss during inference. An alternative solution is to only optimize the encoder (*i.e.* PointNet and 3D U-Net) while freezing the occupancy decoder. The comparisons shown in Table 1 clearly demonstrate that jointly optimizing the whole network can achieve better generality to unseen shapes.

Ablation studies on the iteration number of the test-time optimization. Fig. 5 and 6 show the quantitative and qualitative results w.r.t. the number of iterations, respectively. Notably, the 'Iter 0' represents the result before optimization. We can observe that after about 600 iterations of the test-time optimization, the results become stable.

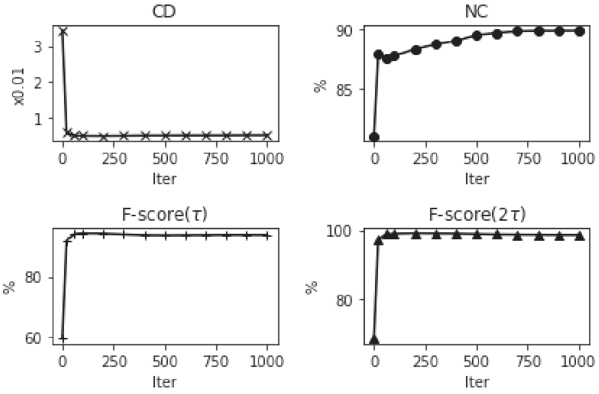
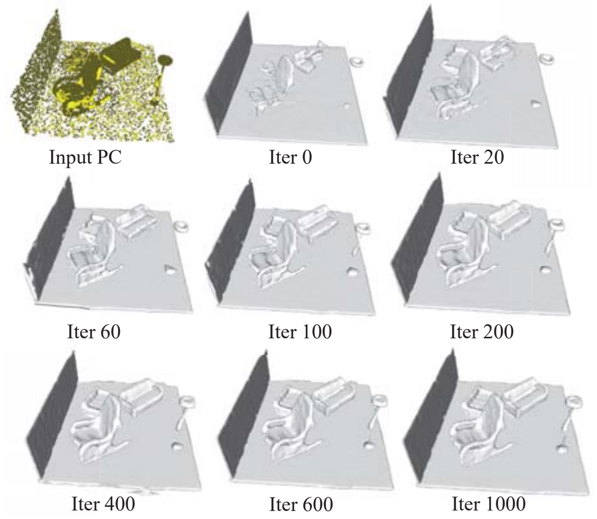


Figure 5: Quantitative results obtained at different iterations during the test-time optimization. Experiments are conducted on the synthetic room dataset with the input of 30,000 points.


Input PC Iter 0 Iter 20

Iter 60 Iter 100 Iter 200

Iter 400 Iter 600 Iter 1000

Figure 6: Examples of qualitative results of a synthetic room obtained at different iterations during the test-time optimization.

Ablation studies on the sparsity level of the input. Quantitative results for different sparsity levels of the input are shown in Table 2 and 3. We can observe that the results of different evaluation metrics only show a slightly small variance, which clearly demonstrates the robustness of our method against the input sparsity.

$\|\mathcal{P}\|$	CD ↓	NC ↑	FS (τ) ↑	FS (2τ) ↑
5,000	0.529	89.71	95.51	99.00
10,000	0.524	92.37	96.85	99.20
20,000	0.522	93.29	97.11	99.06
30,000	0.522	93.51	97.16	99.37
40,000	0.502	93.57	97.11	99.35
50,000	0.502	93.61	97.04	99.29

Table 2: Quantitative results at different sparsity levels of the input point cloud $\mathcal{P}$ on the ShapeNet 'chair' category.

$\|\mathcal{P}\|$	CD ↓	NC ↑	FS (τ) ↑	FS (2τ) ↑
5,000	0.511	89.24	93.50	98.57
10,000	0.494	89.86	94.01	98.89
20,000	0.494	90.03	93.85	98.73
30,000	0.495	90.04	93.85	98.82
40,000	0.488	90.04	93.85	98.73
50,000	0.476	89.98	93.95	98.99

Table 3: Quantitative results at different sparsity levels of the input point cloud $\mathcal{P}$ on the synthetic room dataset.

C. Novel Categories Generalization

In this section, we analyze the generalization performance of our approach and the baselines on the object-level reconstruction. We directly evaluate them on novel categories such as "bench", "lamp" and "watercraft" that are different from the training "chair" category. As shown in Figure 7, our approach can preserve more geometric details such as small holes, long rods, and thin parts, while the baselines cannot. This demonstrates the superior generalization capabilities of the proposed approach to unseen categories.

D. Real-world Scenes Generalization

In this section, we first describe the implementation details of sign-agnostic optimization of convolutional occupancy networks in a sliding-window manner, and then provide more qualitative comparison on the real-world scenes datasets including ScanNet-V2 [5] and Matterport3d [2].

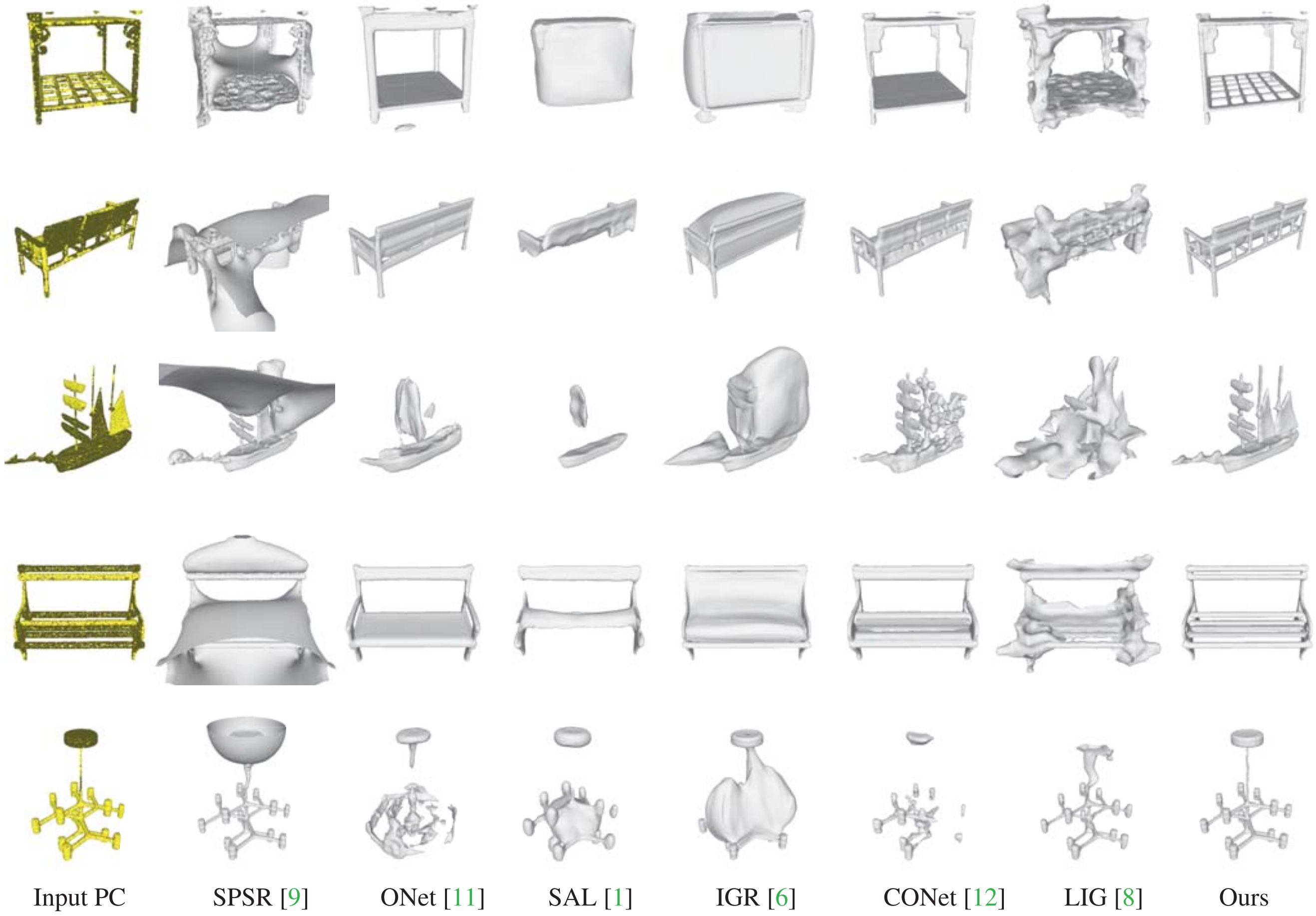


Figure 7: **Generalize to Novel Categories.** We directly evaluate our approach and baselines on unseen, novel categories including “bench”, “lamp”, and “watercraft” that are very different from the training “chair” category.

D.1. Implementation Details of Sign-Agnostic Optimization in a Sliding-Window Manner

In the experiments of object-level and synthetic scene reconstruction, we perform pre-training and sign-agnostic optimization within the unit cube. However, this strategy cannot deal with real-world scenes that are arbitrarily sized and represented in meters. Although we can resize these scenes into the unit cube, convert them into volumetric grids of size 64^3, and then process them using the 3D U-Net as described in Section A, we may not be able to recover fine-grained geometries as the low-resolution voxelization process loses much information about surface details, while the high-resolution voxelization such as $128^3, 256^3$ would suffer from the heavy computation cost and memory issues. Thanks to the translation equivalence of fully convolutional networks, we can apply the proposed model to local patches cropped from large scenes and perform implicit surface reconstruction in a sliding-window manner, which can help us preserve the input information while avoiding memory issues of 3D CNNs.

More specifically, we also pre-train our model on the synthetic indoor scene dataset [12] where the size of scenes is approximately a real-world unit of 4.4m × 4.4m × 4.4m. Similar to the setting of [12], we set the voxel size as 0.02m such that each scene is contained in a volumetric grid with size 220^3. During the network pre-training, we utilize the Res-PointNet and 3D U-Net described in Section A to learn corresponding convolutional features from each cropped subvolume. Then we predict the occupancy probabilities of query points uniformly sampled from the grid of input subvolume. Specifically, we randomly sample one point within the whole scene and use it as the center of the subvolume. The size of each cropped subvolume (*i.e.* $H \times W \times D$) is set to $25 \times 25 \times 25$. Since the receptive field of 3D U-Net is 64, we set the size of input subvolumes to $(H + 63)(W + 63) + (D + 63) = 88 \times 88 \times 88$. At each iteration, we use a batch size of 4 subvolumes.

During the test-optimization stage, we divide the large scene into overlapped subvolumes and then perform sign-agnostic optimization for each subvolume in a sliding-window manner. We determine the size of cropped subvolumes according to the size of input scenes such that they are compatible with the GPU memory. Notably, we do not need the padding operation as the cropped subvolumes overlap.

D.2. Additional Qualitative Results

We have provided more qualitative comparisons on the ScanNet [5] in Figure 8. Besides, more visualized results on the Matterport3D [2] are shown in Figure 9. From these results, we can clearly observe that our method achieves more superior performance to large scenes with multiple rooms than the existing state-of-the-arts. And in comparison with those baselines such as SPSR [9, 8] that heavily rely on accurate surface normals, our approach can avoid the degenerated results caused by inaccurate normal estimation. Besides, compared to CONet [12], our approach can reconstruct more complete geometries and preserve complicated geometric details well, which validates the effectiveness of the proposed sign-agnostic optimization during inference. Overall, our method simultaneously maximizes the scalability to large scenes, generality to unseen shapes, and applicability to real scans that lack reliable surface normals.

References

[1] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *CVPR*, 2020.

[2] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *3DV*, 2017.

[3] 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.

[4] R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017.

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.

[6] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *ICML*, 2020.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[8] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. Local implicit grid representations for 3d scenes. In *CVPR*, 2020.

[9] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM ToG*, 2013.

[10] Yiyi Liao, Simon Donne, and Andreas Geiger. Deep marching cubes: Learning explicit surface representations. In *CVPR*, 2018.

[11] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019.

[12] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *ECCV*, 2020.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer, 2015.

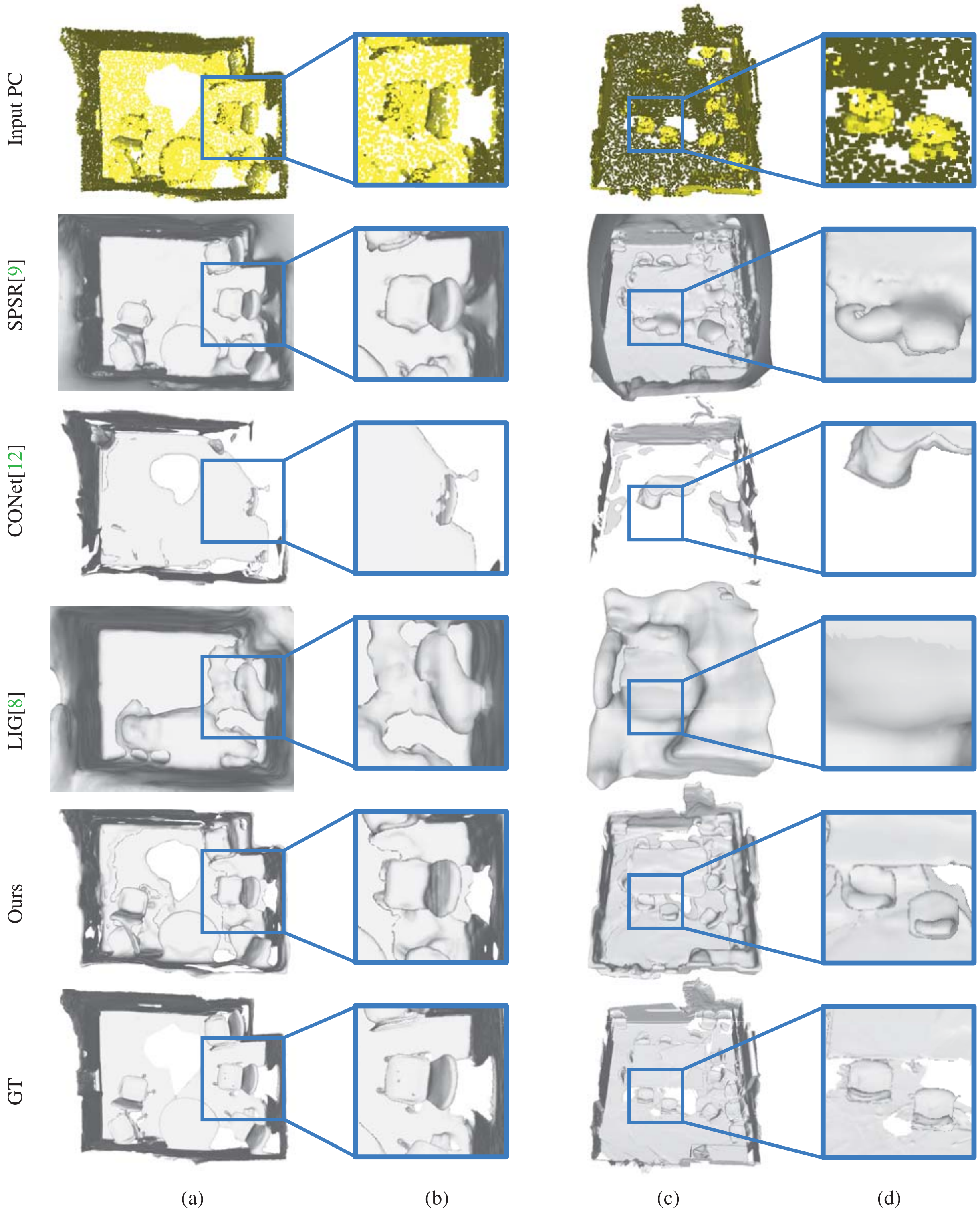


Figure 8: **Scene-level Reconstruction on ScanNet [5].** Qualitative comparisons for surface reconstruction from unorientated scans of ScanNet. All methods except SPSR are trained on the synthetic room dataset and directly evaluated on ScanNet.

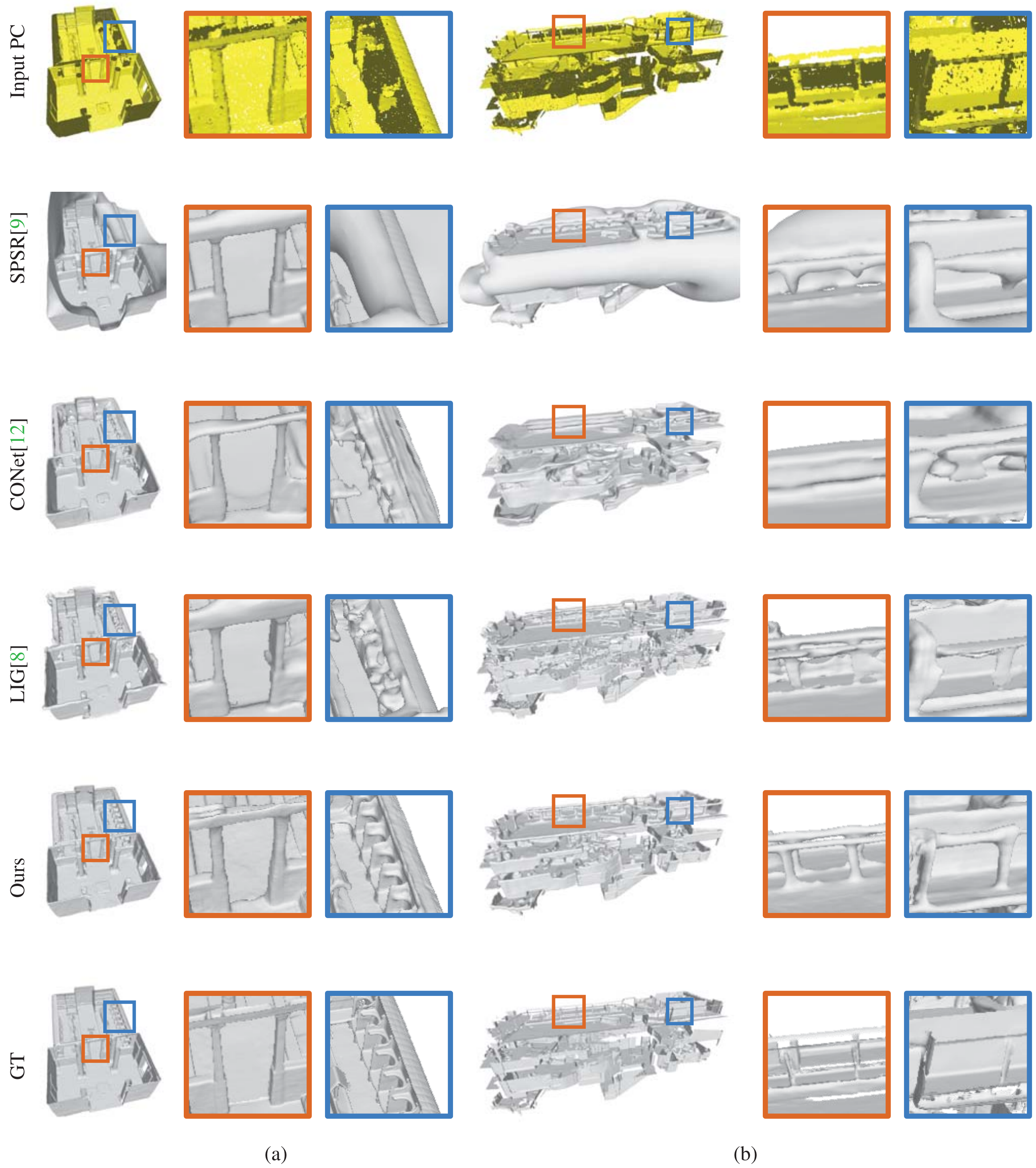


Figure 9: **Scene-level Reconstruction on Matterport 3D [2].** Qualitative comparisons for surface reconstruction from un-orientated scans of Matterport3D. All methods except SPSR are trained on the synthetic room dataset and directly evaluated on Matterport 3D.