# Supplementary Material for 3D Sketch-aware Semantic Scene Completion via Semi-supervised Structure Prior

Xiaokang Chen[1*] Kwan-Yee Lin[2] Chen Qian[2] Gang Zeng[1†] Hongsheng Li[3]
[1]Key Laboratory of Machine Perception (MOE), School of EECS, Peking University
[2]SenseTime Research [3]The Chinese University of Hong Kong

## 1. Introduction

This supplementary material presents: (1) details of three datasets; (2) the detail description of training/inference strategies and network architecture for 3D Sketch Hallucination Module proposed in our approach; (3) additional ablation studies and analysis of our approach; (4) comparison with state-of-the-art methods on SUNCG dataset; (5) visualization of SSC results of the proposed method and SSCNet [6] on NYUCAD dataset.

## 2. Dataset Details

**NYU** consists of 1449 indoor scenes that are captured via a Kinect sensor, which makes it a challenging dataset. There are 795 for training and 654 for test. We follow [6] and use the 3D annotated labels provided by [4] for semantic scene completion task. To address the misalignment of some label volumes and their corresponding depth maps, **NYUCAD** uses the depth maps generated from the projections of the 3D annotations. **SUNCG** is a synthetic dataset that consists of 45622 indoor scenes. The depth images and semantic scene volumes are acquired by setting different camera orientations. The training set contains about 150K depth images and the corresponding test set consists of totally 470 pairs sampled from 170 non-overlap scenes.

## 3. Implementation Details

In this section, more details of the *3D Sketch Hallucination Module* proposed in our approach are provided. The architecture is shown in Figure 1.

**Training.** During training, $\hat{G}_{\text{raw}}$ and $G_{\text{gt}}$ are concatenated and fed into several convolutions. It will output the mean and diagonal covariance for the posterior $q(\hat{z}|G_{gt}, \hat{G}_{raw})$ and the output size is $\frac{1}{4}$ of the input size. Then we use the reparameterization trick to construct a latent code which approximates a gaussian distribution. The dimension of the latent code is 16 and the resolution of it is $(15 \times 9 \times 15)$, so it is light-weight. Note that the blocks marked with dotted

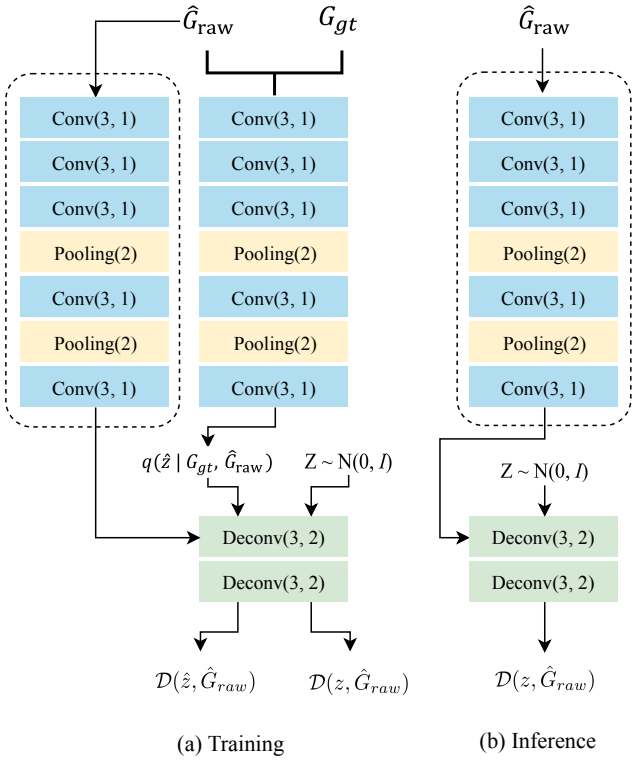


Figure 1. **Details of 3D Sketch Hallucination Module in the main paper**. The convolution parameters are shown as (kernel size, dilation). The pooling parameters are shown as (downsample rate). The Deconvolution parameters are shown as (kernel size, upsample rate).

* This work was done during an internship at SenseTime Research.
† Gang Zeng is the corresponding author.

| #Index | #Stage | Structure Prior | CVAE | SC-IoU(%) | SSC-mIoU(%) |
|---|---|---|---|---|---|
| (a) | 1 | ✗ | ✗ | 65.9 | 37.9 |
| (b) | 2 | ✗ | ✗ | 67.4 | 38.8 |
| (c) | 2 | ✓ | ✗ | 70.7 | 40.1 |
| (d) | 2 | ✓ | ✓ | **71.3** | **41.1** |

Table 1. **Ablation studies on different modules**. We perform this ablation study on NYU dataset.

| Input | Shape | Semantic Labels | Sketch | SC-IoU(%) | SSC-mIoU(%) |
|---|---|---|---|---|---|
| TSDF+RGB | ✓ | | | 69.4 | 40.2 |
| TSDF+RGB | | ✓ | | **71.5** | 40.3 |
| TSDF+RGB | | | ✓ | 71.3 | **41.1** |

Table 2. **Ablation studies on different representations of structure prior**. We perform this ablation study on NYU dataset.

lines are used to downsample the input $\hat{G}_{\text{raw}}$, so it has the same resolution with the latent code. Then the latent code and the downsampled $\hat{G}_{\text{raw}}$ will be concatenated and fed into deconvolutions to obtain $\mathcal{D}(\hat{z}, \hat{G}_{raw}))$. As mentioned in the main paper, we also train a GSNN. We sample four latent codes with the same size of the one mentioned before from $p(z) \sim \mathcal{N}(0, I)$, and conduct the same operation as before to obtain $\mathcal{D}(z, \hat{G}_{raw}))$. The four results will be averaged.

**Inference.** During inference, we directly sample four latent codes from $p(z) \sim \mathcal{N}(0, I)$, and concatenate the latent code with the downsampled $\hat{G}_{\text{raw}}$ to obtain $\mathcal{D}(z, \hat{G}_{raw}))$. The four results will be averaged.

## 4. More Quantitative Results

Besides the results analyzed in the main paper, we demonstrate and discuss more ablation studies of our model on more experimental settings in this section.

**Different Modules in the Framework.** In the main paper, we show the effectiveness of different modules in the framework on NYUCAD dataset. Here, we conduct the same ablation study on NYU dataset. Figure 2 shows the detailed network architectures of corresponding methods in Table 1. As can be seen from Table 1, under the low-resolution input, simple multi-modality strategy (Index (b)) has little strength compared with habitual SSC framework (Index (a)). In contrast, the proposed 3D sketch-aware feature embedding strategy (Index (c)) helps boosts the performances with $4.8\%$ SC-IoU and $2.2\%$ SSC-mIoU improvements. Since it explicitly encodes sketch information of the scene, which is resolution-insensitive and compact. Such a depth feature embedding could encourage the subsequent network to infer the invisible areas of the scene with well structure-preserving details. The further improvements provided by Index (d) demonstrate the effectiveness of the 3D sketch hallucination module, which could help generate more accurate and realistic results.

**Different Representations of Structure Prior.** To assess the effectiveness of the 3D sketch as a structure prior to the subsequent SSC network, we compare it with other available priors to guide the feature embedding of depth information on NYU dataset. The results are shown in Table 2. As mentioned in the main paper, 'Shape' refers to the binary description of the scene, and 'Semantic Labels' refers to the semantic description of the scene. They could be regarded as the coarse predictions of scene completion and scene segmentation. Then the features embedded from the supervision of the 'Shape' and 'Semantic Labels' are fed into the subsequent SSC network to infer the final results. We observe that the proposed 3D sketch achieves the best overall performance considering SC-IoU and SSC-mIoU together. This ablation study provides additional evidence that the proposed 3D sketch is an effective representation of depth information and better facilitates the subsequent network learning the concept of objects' structure when compared with other possible structure priors.

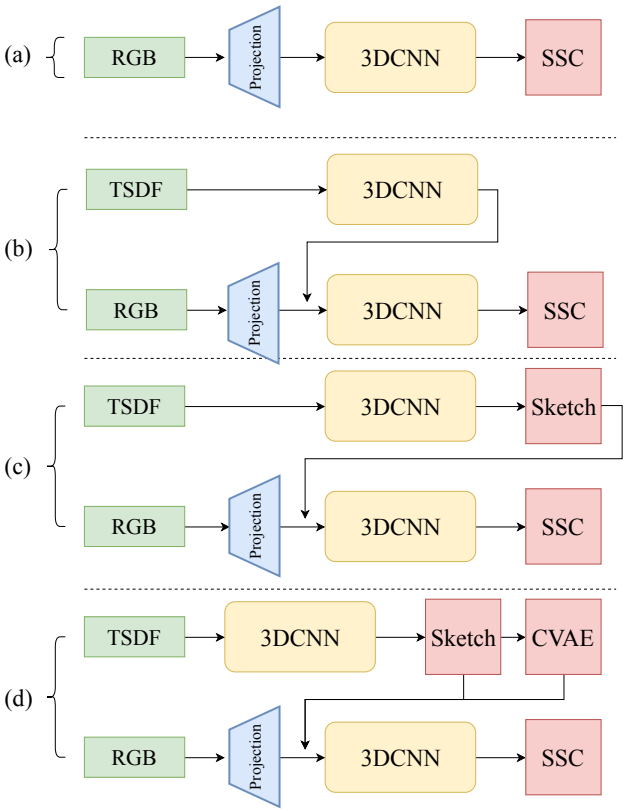


Figure 2. **The illustration of different methods mentioned in Table 1**. (a)(b)(c)(d) correspond to four rows of the Table.

**Different Modal Inputs.** We further conduct ablation studies on the effect of different modal input to our framework on NYUCAD dataset. Results are listed in Table 3. The experimental phenomenon is consistent with the same experimental setting on NYU dataset listed in the main paper. Specifically, as shown in the third row, our frame-

| Input for Stage1 | Input for Stage2 | SC-IoU(%) | SSC-mIoU(%) |
|---|---|---|---|
| RGB | RGB | 81.5 | 51.9 |
| RGB | TSDF | 82.5 | 52.5 |
| TSDF | TSDF | 83.7 | 51.3 |
| TSDF | RGB | **84.2** | **55.2** |

Table 3. **Ablation studies on different modal input**. We perform this ablation study on NYUCAD dataset.

work could achieve 83.7% SC-IoU and 51.3% SSC-mIoU with *only depth modality as input*. Comparing with CCP-Net [9] mentioned in the main paper, which is the most state-of-the-art SSC method that increases the voxel resolution of depth source to $240 \times 144 \times 240$, our model with single depth modality produces better results on SC-IoU ( 83.7% vs 82.4%) and competitive performance on SSC-IoU ( 51.3% vs 53.2%). Note that the input and output of our model are both under $60 \times 36 \times 60$ resolution. With RGB modality injected into the second stage (the fourth row), our model could achieve leading state-of-the-art results with 84.2% SC-IoU and 55.2% SSC-mIoU. The overall results listed in Table 3 show: (1) RGB modality has more strength on semantic segmentation, while depth modality has more strength on completion. (2) the proposed 3D sketch-aware feature embedding is not conflicted with the modality input of subsequent SSC network.

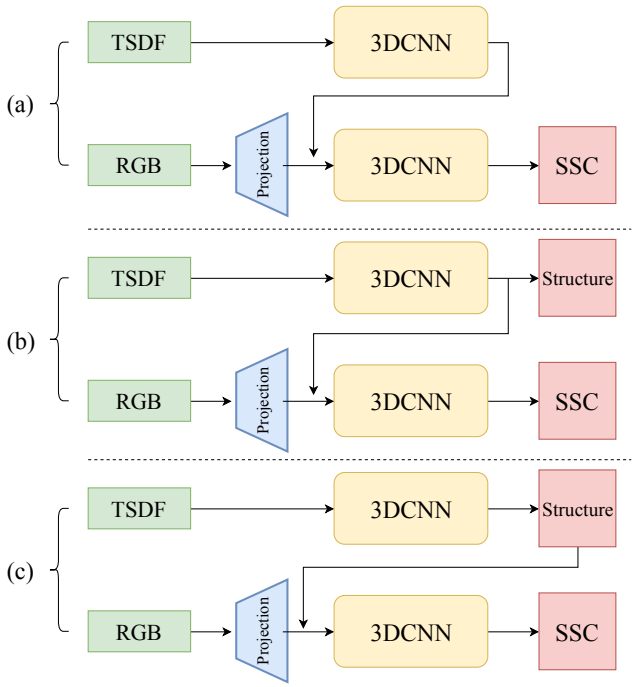


Figure 3. **Illustration of different types of embedding strategies mentioned in Table 4**. (a)Implicit embedding with no supervision. (b) Implicit embedding with supervision on structure representations. (c) Explicit embedding with supervision on structure representations.

| Supervision | Embedding | SC-IoU(%) | SSC-mIoU(%) |
|---|---|---|---|
| None | Implicit | 67.4 | 38.8 |
| Shape | Implicit | 69.4 | 39.2 |
| | Explicit | **69.4** | **40.2** |
| Semantic | Implicit | 71.0 | 39.7 |
| | Explicit | **71.5** | **40.3** |
| Sketch | Implicit | 70.6 | 40.6 |
| | Explicit | **71.3** | **41.1** |

Table 4. **Ablation studies on different types of embeddings**. We perform this ablation study on NYU dataset.

**Different Types of Embedding Strategies.** At last, we verify the effect of different types of embedding strategies based on different structure priors on NYU dataset. In order to explain explicit and implicit depth embedding more clearly, we also illustrate the corresponding architectures in Figure 3. Figure 3(a) represents implicit depth embedding with no supervision. The feature comes from the TSDF branch is directly fed into the RGB branch. Figure 3(b) represents implicit depth embedding with supervision. Figure 3(c) represents the proposed method. We abstract TSDF feature into a concrete shape representation and use it as structure prior to the RGB branch. Experimental results in Table 4 show great advantage of the proposed 3D sketch-aware feature embedding strategy as an explicit and compact encoding of depth information. Note that the proposed 3D Sketch Hallucination Module (*i.e.*, CVAE part) is incorporated in both models (b) and (c). Both models (b) and (c) could obtain significant gains based on the already very strong baseline model (a). This further illustrates the effectiveness of the proposed method in another aspect.

**What is the 3D sketch beneficial for?** We try to find out what benefits does the network actually get from the 3D sketch-aware feature embedding and its hallucination schema. As mentioned in the main paper, the hallucinated 3D sketch is expected to provide structure prior to the entire scene, including both visible and invisible areas. To verify this, we conduct experiments to explore how much does the network benefit from the 3D sketch on visible and invisible areas, respectively. Results are listed in Table 5. We observe there is a relatively uniform increase in both visible (5.9%) and invisible (6.5%) areas, which illustrates the sketch not only provides structure prior in the visible areas but also infers reasonable structures in the invisible areas.

We also conduct experiments to explore the effect of the sketch prior in invisible areas on the theoretical upper limit of the proposed method. Results are listed in Table 6. Firstly, we observe that simply supply 3D sketch prior in visible areas could boost the performance (compared with the first row in Table 6 in the main paper). Besides, providing 3D sketch prior in both visible and invisible areas could significantly improve performance. We owe it to the fact

| Sketch | Visible SSC-mIoU(%) | Invisible SSC-mIoU(%) | SSC-mIoU(%) |
|---|---|---|---|
| ✗ | 58.5 | 46.0 | 48.7 |
| ✓ | **64.4** (5.9↑) | **52.5** (6.5↑) | **55.2** (6.5↑) |

Table 5. **Benefits of the sketch in visible/invisible areas**. The ablation study is evaluated on NYUCAD dataset.

| Oracle Sketch | SC-IoU(%) | SSC-mIoU(%) |
|---|---|---|
| Visible | 80.6 | 51.2 |
| Visible+Invisible | **94.2** | **65.0** |

Table 6. **Impact of visible/invisible oracle sketch**. We perform this ablation study on NYUCAD dataset.

| Region | Prec. (%) | Recall (%) | IoU (%) |
|---|---|---|---|
| Visible | 78.0 | 78.7 | 64.4 |
| Invisible | 72.1 | 64.6 | 51.7 |
| Visible+Invisible | 73.5 | 67.6 | 54.4 |

Table 7. **Results of the sketch in visible/invisible areas**. We perform this ablation study on NYUCAD dataset.

| Method | Speed(FPS) | Memory(M) | SSC-mIoU(%) |
|---|---|---|---|
| SSCNet [6] | 0.7 | 5305 | 24.7 |
| DDRNet [2] | 1.5 | 1829 | 30.4 |
| Ours | **6.7** | **1753** | **41.1** |

Table 8. **The inference speed and GPU memory usage of our method and previous methods.** All results are acquired on a GTX 1080 Ti GPU and evaluated on the NYU [5] test set.

that the complete sketch provides the complete structural information of the object in the scene, which is beneficial for the network to learn the shape and semantics of the object and recognize it.

We further use quantitative indicators to measure the quality of the learned 3D sketch. Results are listed in Table 7. We observe that the learned 3D sketch has high accuracy and recall in both visible and invisible areas, so as to supply sufficient structure prior for the subsequent SSC network in the proposed method.

**Efficiency Analysis.** Following DDRNet [2], we report the speed and GPU memory usage of the proposed method as well as the some previous methods. As shown in Table 8, the proposed method achieves much faster speed and a significant performance gain compared with DDRNet [2]. Moreover, our method requires less GPU memory cost than previous methods, which demonstrates the efficiency advantages of the proposed method.

## 5. Comparisons with State-of-the-art Methods on SUNCG

To verify the generalization of our proposed method, we further compare our method with state-of-the-art methods on SUNCG dataset. Results are listed in Table 9. Note that not all scenes in SUNCG have corresponding RGB images, thus we only use TSDF volume as the input for both two stages. Results show that our method with low-resolution input obtains significant improvements on both the SC-IoU and SSC-mIoU metrics when comparing with previous state-of-the-art methods, which illustrates the effectiveness of the proposed modules.

| Methods | Resolution | SC-IoU(%) | SSC-mIoU(%) |
|---|---|---|---|
| SSCNet [6] | $(240, 60)$ | 73.5 | 46.4 |
| ForkNet [7] | $(80, 80)$ | 86.9 | 63.4 |
| SATNet [3] | $(60, 60)$ | 78.5 | 64.3 |
| VVNetR-120 [1] | $(120, 60)$ | 84.0 | 66.7 |
| CCPNet [9] | $(240, 60)$ | 86.5 | 69.1 |
| ESSCNet [8] | $(240, 60)$ | 84.5 | 70.5 |
| Ours | $(60, 60)$ | **88.2** | **76.5** |

Table 9. **Results on SUNCG dataset**. Bold numbers represent the best scores. *Resolution(a, b)* means the input resolution is $(a \times 0.6a \times a)$ and the output resolution is $(b \times 0.6b \times b)$.

## 6. Visualization of SSC Results on NYUCAD

In this section, we show the qualitative results of the proposed method and SSCNet [6]. Figure 4 visualizes the SSC result of the proposed method and SSCNet [6] on NYUCAD dataset. Our method achieves better intra-class consistency and inter-class distinction compared with SSCNet. In the first and the second rows of Figure 4, SSCNet fails to complete the wall because the wall has too many regions missing. In contrast, the proposed method leverages the structure prior well and successfully infer invisible areas. In the following four rows, we observe that our results have consistent semantic surface and precise boundaries under the constraint of the predicted sketch, which demonstrates the effectiveness of the structure prior.

We also show more quantitative evaluations of the effectiveness of our approach in the form of video. ***For a straightforward visual perception, please refer to the supplemental video and better turn the audio on.*** Specifically, the first scene is a classroom. In the second column, which is the prediction result from SSCNet, the ceiling is missing, and the region of cabinets and chairs are in chaos. However, our proposed method completes the ceiling well and shows great consistency on the floor. The second scene is a classroom with several desks. Our proposed method shows good intra-class consistency on the board on the wall, and it completes the ceiling well. From the above two examples, we find that the proposed method could handle a case in which large portions of geometry are missing in the partial observation (such as the ceiling). The third scene is the living room. We observe that in the second column, a large area of the wall is missing and the prediction of pillows on the sofa is not precise. However, our proposed method not only completes the wall well, but also accurately recognizes the

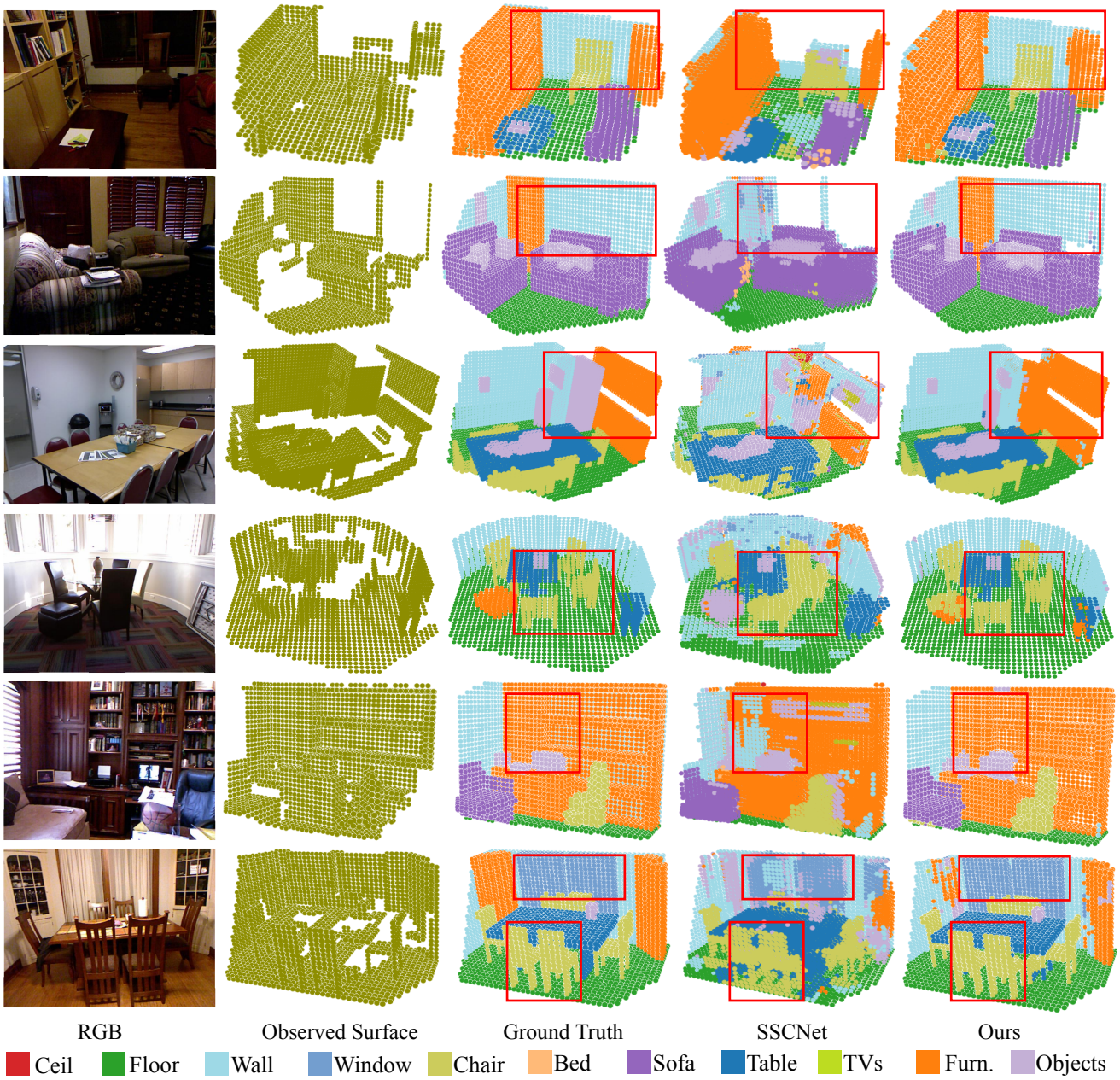


Figure 4. **Semantic Scene Completion results on NYUCAD dataset.** From left to right: (1) RGB input, (2) observed surface, (3) ground truth, (4) results of SSCNet, (5) our results. Our results achieve higher voxel-level accuracy compared with SSCNet. Better viewed in color and zoom in.

objects on the wall. The last scene is another living room. Our proposed method accurately recognizes the chairs and achieves nice intra-class consistency and inter-class distinction.

## References

[1] Yu-Xiao Guo and Xin Tong. View-volume network for semantic scene completion from a single depth image. In *IJCAI*, pages –, 2018.

[2] Jie Li, Yu Liu, Dong Gong, Qinfeng Shi, Xia Yuan, Chunxia Zhao, and Ian Reid. Rgbd based dimensional decomposition residual network for 3d semantic scene completion. In *CVPR*, pages –, 2019.

[3] Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and think: Disentangling semantic scene completion. In *NIPS*, pages 261–272, 2018.

[4] Jason Rock, Tanmay Gupta, Justin Thorsen, JunYoung Gwak, Daeyun Shin, and Derek Hoiem. Completing 3d object shape from one depth image. In *ICCV*, pages 2484–2493, 2015.

[5] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[6] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, pages 1746–1754, 2017.

[7] Yida Wang, David Joseph Tan, Nassir Navab, and Federico Tombari. Forknet: Multi-branch volumetric semantic completion from a single depth image. In *ICCV*, pages 8608–8617, 2019.

[8] Jiahui Zhang, Hao Zhao, Anbang Yao, Yurong Chen, Li Zhang, and Hongen Liao. Efficient semantic scene completion network with spatial group convolution. In *ECCV*, pages 733–749, 2018.

[9] Pingping Zhang, Wei Liu, Yinjie Lei, Huchuan Lu, and Xiaoyun Yang. Cascaded context pyramid for full-resolution 3d semantic scene completion. In *ICCV*, pages 7801–7810, 2019.