# BUOL: A Bottom-Up Framework with Occupancy-aware Lifting for Panoptic 3D Scene Reconstruction From A Single Image (Supplementary Material)

Tao Chu[1,2*], Pan Zhang[2], Qiong Liu[1†], Jiaqi Wang[2]
[1] South China University of Technology [2] Shanghai AI Laboratory
{chutao, zhangpan, wangjiaqi}@pjlab.org.cn liuqiong@scut.edu.cn

## 1. Architecture

We provide the details of the 2D model and 3D model in our framework. The 2D model is shown in Figure. 1, which is composed of ResNet-50 [5] as a shared encoder, three decoders, and multiple prediction heads. We employ Panoptic-Deeplab [2] as our 2D panoptic segmentation, which contains one ASPP-decoder [1] followed by center head and offset head to predict 2D center map $c^{2d}$ and regress 2D offsets $\triangle c^{2d}$, respectively, and another ASPP-decoder with semantic head to predict semantic map $s^{2d}$. The final decoder followed by the depth head and multi-plane occupancy head is designed to predict depth $d$ and multi-plane occupancy $o^{mp}$, respectively.

With the 3D feature generated by our occupancy-aware lifting as input, a 3D refinement model is applied to predict 3D results , as shown in Figure. 2. Specifically, we convert ResNet-18 [5] and ASPP-decoder [1] to 3D to combine into our 3D encoder-decoder and reduce the channels of the 3D model due to limitations on graphics memory. Similar to the 2D model, we adapt one shared encoder and three decoders for 3D prediction. 3D semantic map $s^{3d'}$ is refined by the 3D semantic head following one decoder, and 3D offsets $\triangle c^{3d'}$ is predicted by 3D offset head following another decoder, and 3D occupancy $o^{3d}$ is obtained by the dot product of two outputs predicted by two heads with the last shared decoder. One output with $BCE$ loss is designed to obtain the course 3D occupancy, and the other with $L1$ loss regresses the Truncated Signed Distance Function (TSDF). Finally, The panoptic 3D scene reconstruction result is processed by our proposed bottom-up panoptic reconstruction.

## 2. Additional Quantitative Results

Table 1 shows the PRQ for each category on the synthetic dataset 3D-Front. Our bottom-up framework "BU" outperforms the top-down methods [3, 4, 6, 7] for almost all categories. With our occupancy-aware lifting, "BUOL" achieves further better performance for each category. Overall, our BUOL improves the baseline by +6.55% PRQ and the state-of-the-art [3] by +11.81% PRQ, respectively.

The PRQ for each category on the real-world dataset Matterport3D is shown in Table 2. Our proposed BUOL outperforms the other top-down methods [3, 4, 7] a lot. Compared with Dahnert et al. [3], our method achieves higher performance for all categories and improves the overall PRQ by 7.46%. For a fair comparison, we also compare BUOL with our strong baseline Dahnert et al. [3]+PD, and our framework achieves +4.39% PRQ better. The quantitative results show that our bottom-up framework with occupancy-aware lifting outperforms the state-of-the-art methods in both synthetic and real-world datasets.

## 3. Additional Qualitative Results

Our additional qualitative results on the synthetic dataset 3D-Front are shown in Figure. 3. Comparing the results in each row, our BUOL reconstructs instances and segments them better. For example, the chairs in rows 1, 2, 5, and 6 are more complete than the other models, and the shape of instances in each row is also closer to the ground truth.

Figure. 4 shows the additional qualitative results on the real-world dataset Matterport3D. Both thing and stuff categories are reconstructed and segmented better by our BUOL. Comparing the shape of the wall in each row, our method performs closer to the ground truth. And comparing the model results of the pillows in row 1, the chairs in rows 3 (upper left) and 5 (left), and the table in row 3, our BUOL assigns the correct category to each instance while TD-PD assigns the wrong category. And comparing the model results of the flowers and platform in row 2, the floor and instances in row 4, and the chairs in row 5, the reconstruction results of our method are better.

Comparing all the results in detail of both synthetic and real-world datasets, both "BU" and "OL" in our Bottom-Up framework with Occupancy-aware Lifting lead to better Panoptic 3D Scene Reconstruction from a single image.

*Intern at Shanghai AI Laboratory. †Corresponding author.

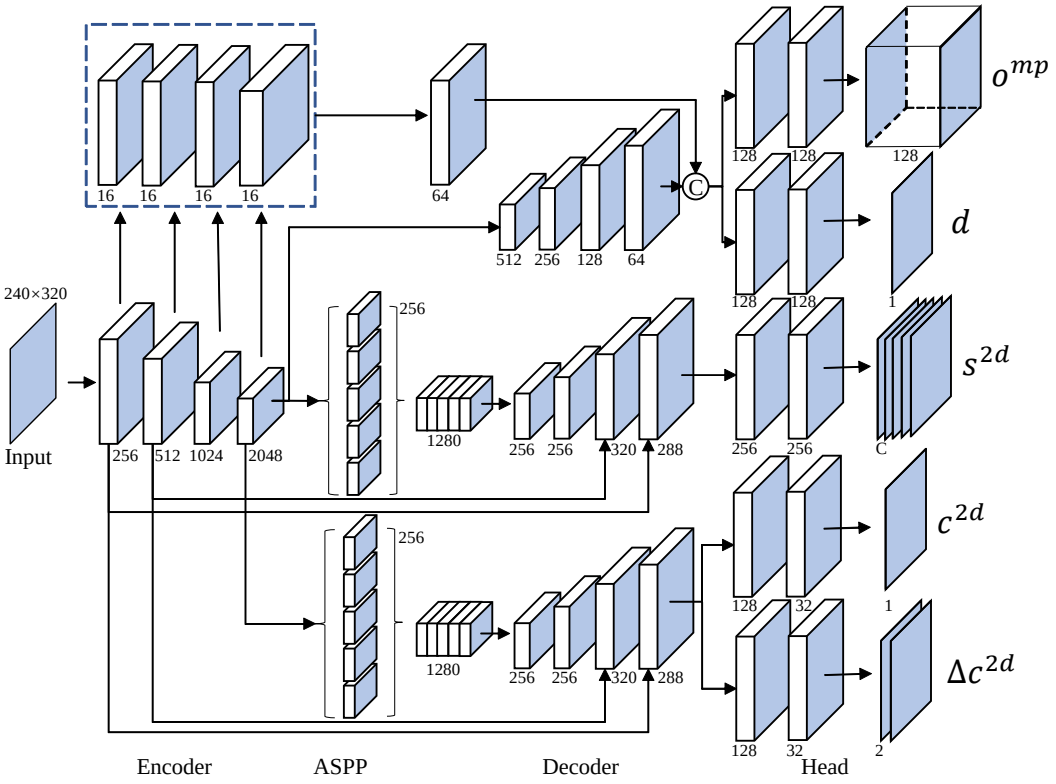


Figure 1. **The 2D model in detail of our BUOL.**

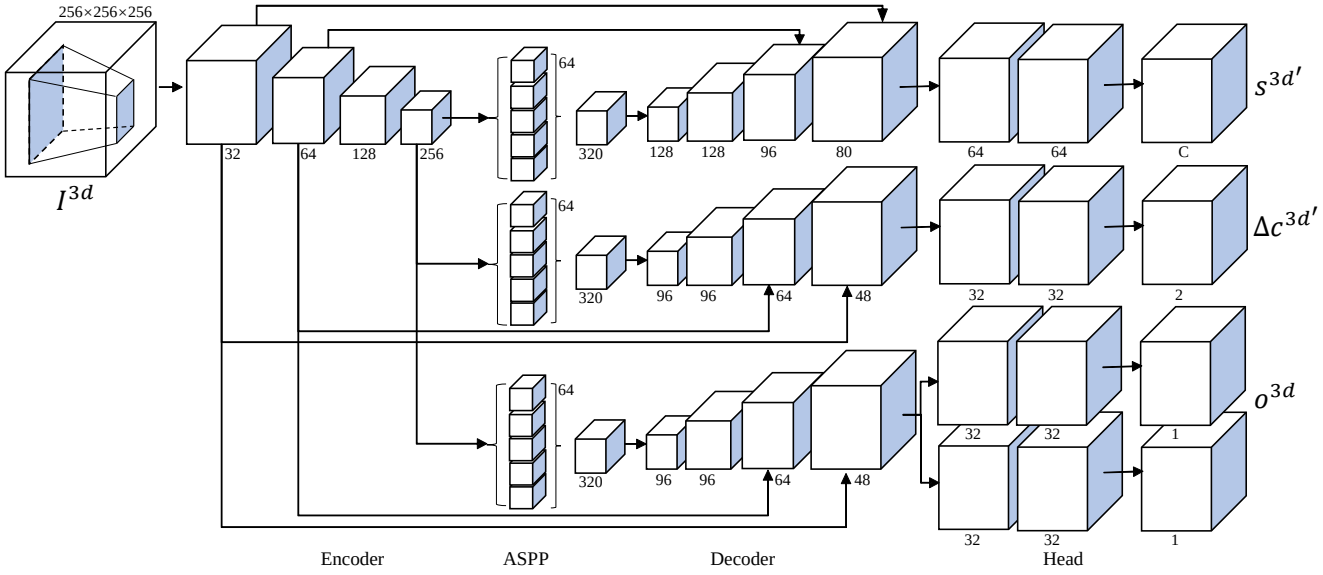


Figure 2. **The 3D model in detail of our BUOL.**

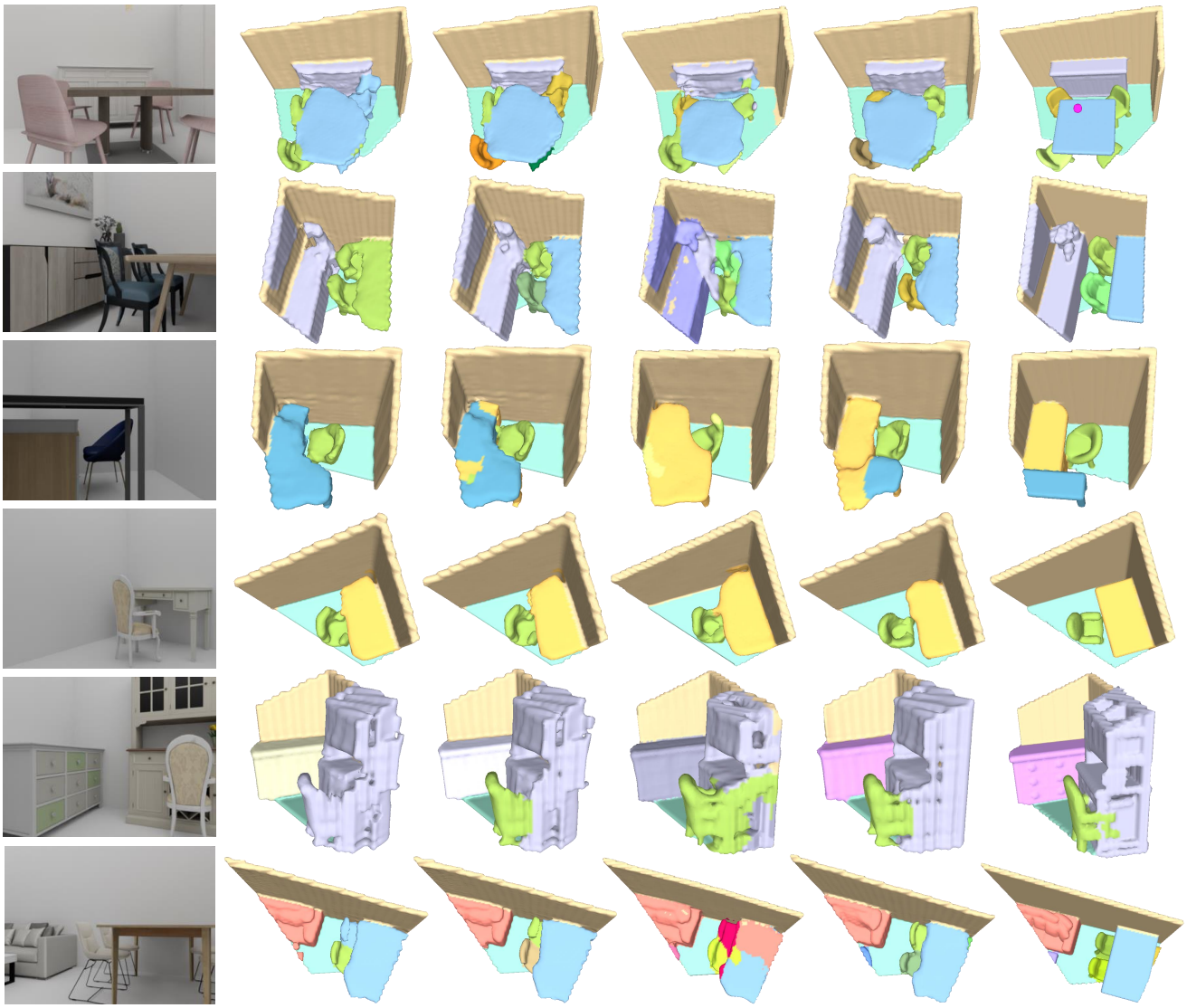


Figure 3. **Qualitative comparisons on 3D-Front.** The BUOL and BU denote our Bottom-Up framework w/ and w/o our Occupancy-aware lifting, respectively, and BU-3D denotes the bottom-up framework with instance grouping by 3D centers, and the TD-PD denotes Dahnert et al. [3]*+PD. And GT is the ground truth.

| Method | Cabinet | Bed | Chair | Sofa | Table | Desk | Dresser | Lamp | Other | Wall | Floor | PRQ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SSCNet [7]+IC | 7.80 | 16.60 | 7.90 | 13.30 | 12.10 | 5.50 | 0.50 | 0.70 | 7.90 | 15.20 | 38.70 | 11.50 |
| Mesh R-CNN [4] | 29.70 | 13.30 | 24.10 | 24.40 | 28.50 | 23.50 | 14.40 | 1.40 | 28.70 | - | - | - |
| Total3D [6] | 17.25 | 4.56 | 18.76 | 14.07 | 19.40 | 16.79 | 7.04 | 8.13 | 17.97 | 8.27 | 33.61 | 15.08 |
| Dahnert et al. [3]* | 43.59 | 49.86 | 27.43 | 42.07 | 40.18 | 34.05 | 40.43 | 8.77 | 42.21 | 57.63 | 77.93 | 42.20 |
| Dahnert et al. [3]*+PD | 47.67 | 59.00 | 36.59 | 51.96 | 42.72 | 38.70 | 47.18 | 12.47 | 43.93 | 62.22 | 79.66 | 47.46 |
| Our BU | 54.81 | 55.77 | 43.10 | 54.81 | 50.60 | 49.02 | 49.23 | 14.74 | 49.15 | 62.73 | 74.37 | 50.76 |
| Our BUOL | **56.06** | **64.13** | **46.46** | **56.61** | **52.72** | **52.08** | **50.97** | **17.44** | **51.08** | **64.63** | **81.97** | **54.01** |

Table 1. The PRQ for each category on 3D-Front. "*" denotes the trained model with the official codebase released by the authors.


Figure 4. **Qualitative comparisons on Matterport3D.** The BUOL denotes our Bottom-Up framework with Occupancy-aware lifting, and the "TD-PD" denotes Dahnert et al. [3]*+PD. And GT is the ground truth.

| Method | Cabinet | Bed | Chair | Sofa | Table | Desk | Dresser | Lamp | Other | Wall | Floor | Ceiling | PRQ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SSCNet [7]+IC | 0.07 | 0.11 | 0.61 | 0.07 | 0.53 | 0.00 | 0.00 | 0.00 | 0.19 | 0.34 | 3.96 | 0.00 | 0.49 |
| Mesh R-CNN [4] | 3.10 | 10.00 | 14.80 | 12.00 | 7.90 | 0.00 | 0.00 | 2.80 | **6.00** | - | - | - | - |
| Dahnert et al. [3] | 12.33 | 10.24 | 9.75 | 14.40 | 8.07 | 0.00 | 0.00 | 0.00 | 2.26 | 10.92 | 16.54 | 4.88 | 7.01 |
| Dahnert et al. [3]*+PD | 9.73 | 20.13 | 11.95 | 12.19 | 4.87 | 0.00 | 0.00 | 2.68 | 4.42 | 16.72 | 31.53 | 6.73 | 10.08 |
| Our BUOL | **13.24** | **27.67** | **16.26** | **17.88** | **11.68** | **1.21** | **1.52** | **3.58** | 5.73 | **19.97** | **38.26** | **16.59** | **14.47** |

Table 2. The PRQ for each category on Matterport3D. "*" denotes the trained model with the official codebase released by the authors.

## References


[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[2] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12475–12485, 2020. 1

[3] Manuel Dahnert, Ji Hou, Matthias Nießner, and Angela Dai. Panoptic 3d scene reconstruction from a single rgb image. *Advances in Neural Information Processing Systems*, 34:8282–8293, 2021. 1, 3, 4

[4] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9785–9795, 2019. 1, 3, 4

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[6] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020. 1, 3

[7] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017. 1, 3, 4