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

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

Abstract

Understanding and modeling the 3D scene from a single image is a practical problem. A recent advance proposes a panoptic 3D scene reconstruction task that performs both 3D reconstruction and 3D panoptic segmentation from a single image. Although having made substantial progress, recent works only focus on top-down approaches that fill 2D instances into 3D voxels according to estimated depth, which hinders their performance by two ambiguities. (1) ***instance-channel ambiguity****: The variable ids of instances in each scene lead to ambiguity during filling voxel channels with 2D information, confusing the following 3D refinement. (2)* ***voxel-reconstruction ambiguity****: 2D-to-3D lifting with estimated single view depth only propagates 2D information onto the surface of 3D regions, leading to ambiguity during the reconstruction of regions behind the frontal view surface. In this paper, we propose* ***BUOL****, a* ***B****ottom-****U****p framework with* ***O****ccupancy-aware* ***L****ifting to address the two issues for panoptic 3D scene reconstruction from a single image. For* ***instance-channel ambiguity****, a bottom-up framework lifts 2D information to 3D voxels based on deterministic semantic assignments rather than arbitrary instance id assignments. The 3D voxels are then refined and grouped into 3D instances according to the predicted 2D instance centers. For* ***voxel-reconstruction ambiguity****, the estimated multi-plane occupancy is leveraged together with depth to fill the whole regions of things and stuff. Our method shows a tremendous performance advantage over state-of-the-art methods on synthetic dataset 3D-Front and real-world dataset Matterport3D. Code and models will be released.*

1. Introduction

Joint learning of 3D reconstruction and perception is a challenging and practical problem for various applications. Existing works focus on combining 3D reconstruction with semantic segmentation [26, 27] or instance segmentation [11, 23, 28]. Recently, a pioneer work [6] unifies the tasks of 3D reconstruction, 3D semantic segmentation, and 3D instance segmentation into panoptic 3D scene re-

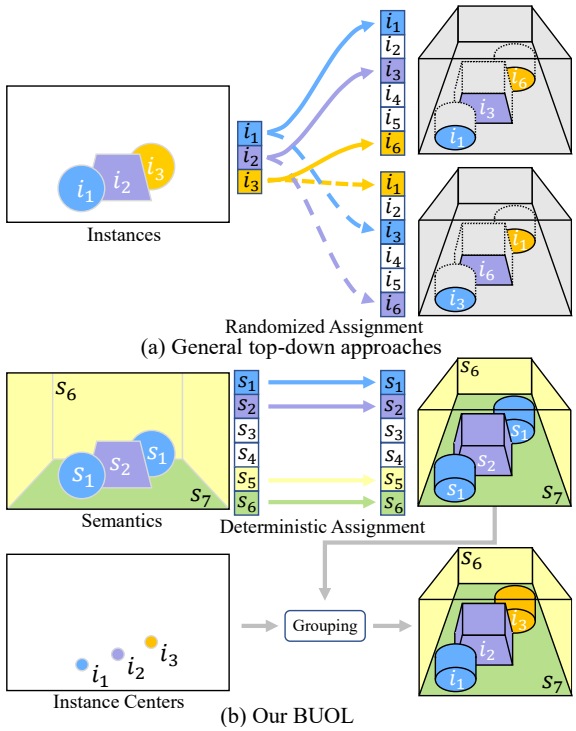

Figure 1. **Comparison of the feature lifting from 2D to 3D. (a) General Top-down approaches:** Feature lifting by depth with the two randomized instance assignments in the top-down framework. The predicted 2D instance masks $\{i_1, i_2, i_3\}$ are lifted to only the surface of 3D instances at variable channels, such as $\{i_1, i_3, i_6\}$ or $\{i_3, i_6, i_1\}$, which results in instance-channel ambiguity and voxel-reconstruction ambiguity. **(b) Our BUOL:** Occupancy-aware lifting with the deterministic semantic assignment in the bottom-up framework. The predicted 2D semantic category maps $\{s_1, s_2, s_6, s_7\}$ are lifted to the whole regions of things (s_1, s_2) and stuff (s_6, s_7), and the voxels are finally grouped into 3D instances $\{i_1, i_2, i_3\}$ by corresponding 2D instance centers.

*Intern at Shanghai AI Laboratory. †Corresponding author.

construction from a single RGB image, which assigns a category label (i.e. a thing category with easily distinguishable edges, such as tables, or a stuff category with indistinguishable edges, such as wall) [22] and an instance id (if the voxel belongs to a thing category) to each voxel in the 3D volume of the camera frustum.

Dahnert et al. [6] achieve this goal in a top-down pipeline that lifts 2D instance masks to channels of 3D voxels and predicts the panoptic 3D scene reconstruction in the following 3D refinement stage. Their method first estimates 2D instance masks and the depth map. The 2D instance masks are then lifted to fill voxel channels on the front-view surface of 3D objects using the depth map. Finally, a 3D model is adopted to refine the lifted 3D surface masks and attain panoptic 3D scene reconstruction results of all voxels.

After revisiting the top-down panoptic 3D scene reconstruction framework, we find two crucial limitations which hinder its performance, as shown in Figure 1(a). First, **instance-channel ambiguity**: the number of instances varies in different scenes. Thus lifting 2D instance masks to fill voxel channels can not be achieved by a deterministic instance-channel mapping function. Dahnert et al. [6] propose to utilize a randomized assignment that randomly assigns instance ids to the different channels of voxel features. For example, two possible random assignments are shown in Figure. 1(a), where solid and dashed arrow lines with the same color indicate a 2D mask is assigned to different voxel feature channels. This operator leads to instance-channel ambiguity, where an instance id may be assigned to an arbitrary channel, confusing the 3D refinement model. In addition, we experimentally discuss the impact of different instance assignments (*e.g.*, random or sorted by category) on performance in Section 4. Second, **voxel reconstruction ambiguity**: 2D-to-3D lifting with depth from a single view can only propagate 2D information onto the frontal surface in the camera frustum, causing ambiguity during the reconstruction of regions behind the frontal surface. As shown by dashed black lines in the right of Figure 1(a), the 2D information is only propagated to the frontal surface of initialized 3D instance masks, which is challenging for 3D refinement model to reconstruct the object regions behind the frontal surface accurately.

In this paper, we propose **BUOL**, a **B**ottom-**U**p framework with **O**ccupancy-aware **L**ifting to address the above two ambiguities for panoptic 3D scene reconstruction from a single image. For instance-channel ambiguity, our bottom-up framework lifts 2D semantics to 3D semantic voxels, as shown in Figure. 1(b). Compared to the top-down methods shown in Figure. 1(a), instance-channel ambiguity is tackled by a simple deterministic assignment mapping from semantic category ids to voxel channels. The voxels are then grouped into 3D instances according to the predicted 2D instance centers. For voxel-reconstruction ambiguity, as shown in Figure. 1(b), the estimated multi-plane occupancy is leveraged together with depth by our occupancy-aware lifting mechanism to fill regions inside the things and stuff besides front-view surfaces for accurate 3D refinement.

Specifically, our framework comprises a 2D priors stage, a 2D-to-3D lifting stage, and a 3D refinement stage. In the 2D priors stage, the 2D model predicts 2D semantic map, 2D instance centers, depth map, and multi-plane occupancy. The multi-plane occupancy presents whether the plane at different depths is occupied by 3D things or stuff. In the 2D-to-3D lifting stage, leveraging estimated multi-plane occupancy and depth map, we lift 2D semantics into deterministic channels of 3D voxel features inside the things and stuff besides the front-view surfaces. In the 3D refinement stage, we predict dense 3D occupancy in each voxel for reconstruction. Meanwhile, the 3D semantic segmentation is predicted for both the thing and stuff categories. The 3D offsets towards the 2D instance centers are also estimated to identify voxels belonging to 3D objects. The ground truth annotations of 3D panoptic reconstruction, *i.e.*, 3D instance/semantic segmentation masks and dense 3D occupancy, can be readily converted to 2D instance center, 2D semantic segmentation, depth map, multi-plane occupancy, and 3D offsets for our 2D and 3D supervised learning. During inference, we assign instance ids to 3D voxels occupied by thing objects based on 2D instance centers and 3D offsets, attaining final panoptic 3D scene reconstruction results.

Extensive experiments show that the proposed bottom-up framework with occupancy-aware lifting outperforms prior competitive approaches. On the pre-processed 3D-Front [10] and Matterport3D [2], our method achieves +11.81% and +7.46% PRQ (panoptic reconstruction quality) over the state-of-the-art method [6], respectively.

2. Related Work

3D reconstruction. Single-view 3D reconstruction learns 3D geometry from a single-view image. Pixel2Mesh attempts to progressively deform an initialized ellipsoid mesh for a single object, while DISN predicts the underlying signed distance fields to generate the single 3D mesh. UCLID-Net [12] back-projects 2D features by the regressed depth map to object-aligned 3D feature grids, and CoReNet [30] is proposed to lift 2D features to 3D volume by ray-traced skip connections.

To reconstruct the object or scene in more detail, some works adopt multi-view images as input. Pix2Vox [33] is proposed to select high-quality reconstructions for each part in 3D volumes generated by different view images. TransformerFusion [1] also selectively stores features extracted from multi-view images.

3D segmentation. Some 3D segmentation methods directly

use a basic geometry as input. For 3D semantic segmentation, 3DMV [7] combines the features extracted from 3D geometry with lifted multi-view image features to predict per-voxel semantics. ScanComplete [8] is proposed to predict complete 3D geometry with per-voxel semantics by devising 3D geometry with filter kernels invariant to the overall scene size.

For 3D instance segmentation, there exist some top-down and bottom-up methods as follows. Some methods [16, 17] based on box proposals predicted by 3D-RPN pay more attention to the fusion of 3D geometry features and lifted image features. SGPN [32] predicts point grouping proposals for point cloud instance segmentation. RfD-Net [29] focuses on predicting instance mesh of the high objectness proposal predicted by point cloud proposal network. Instead of directly regressing bounding box, GSPN [34] generates proposals by reconstructing shapes from noisy observations to provide location of instances.

Most bottom-up methods adopt center as the goal of instance grouping. PointGroup [20] and TGNN [19] learn to extract per-point features and predict offsets to shift each point toward its object center. Lahoud et al. [25] propose to generate instance labels by learning a metric that groups parts of the same object instance and estimates the direction toward the instance's center of mass. There also exist other bottom-up methods. OccuSeg [13] predicts the number of occupied voxels for each instance to guide the clustering stage of 3D instance segmentation. HAIS [4] introduces point aggregation for preliminarily clustering points to sets and set aggregation for generating complete instances.

3D segmentation with reconstruction. For 3D semantic segmentation with reconstruction, Atlas [27] is proposed to directly regress a truncated signed distance function (TSDF) from a set of posed RGB images for jointly predicting the 3D semantic segmentation of the scene. AIC-Net [26] is proposed to apply anisotropic convolution to the 3D features lifted from 2D features by the corresponding depth to adapt to the dimensional anisotropy property voxel-wisely.

As far as we know, the instance segmentation with reconstruction works follow the top-down pipeline. Mesh R-CNN [11] augments Mask R-CNN [14] with a mesh prediction branch to refine the meshes converted by predicted coarse voxel representations. Mask2CAD [23] and Patch2CAD [24] leverage the CAD model to match each detected object and its patches, respectively. Total3DUnderstanding [28] is proposed to prune mesh edges with a density-aware topology modifier to approximate the target shape.

Panoptic 3D Scene Reconstruction from a single image is first proposed by Dahnert et al. [6], and they deliver a state-of-the-art top-down strategy with Mask R-CNN [14] as 2D instance segmentation and random assignment for instance lifting. Our BUOL is the first bottom-up method for panoptic/instance segmentation with reconstruction from a single image.

3. Methodology

In this section, we propose a bottom-up panoptic 3D scene reconstruction method with occupancy-aware lifting. Given a single 2D image, we aim to learn corresponding 3D occupancy and 3D panoptic segmentation. To achieve this goal, as shown in Figure. 2, we first extract the 2D priors, which includes 2D semantics, 2D instance centers, scene depth, and multi-plane occupancy. Then, an efficient occupancy-aware feature lifting block is designed to lift the 2D priors to 3D features, thus giving a good initialization for the following learning. Finally, a bottom-up panoptic 3D scene reconstruction model is utilized to learn the 3D occupancy and 3D panoptic segmentation, where a 3D refinement model maps the lifted 3D features to 3D occupancy, 3D semantics, and 3D offsets, and an instance grouping block is designed for 3D panoptic segmentation. In addition, the ground truth of 2D priors and 3D offsets adopted by our method can be easily obtained by ground truth annotations of 3D panoptic reconstruction (*i.e.* 3D semantic map, instance masks, and occupancy).

3.1. 2D Priors Learning

Given a 2D image $x \in \mathbb{R}^{H\times W\times 3}$, where H and W is image height and width, panoptic 3D scene reconstruction aims to map it to semantic labels $\hat{s}^{3d}$ and instance ids $\hat{i}^{3d}$. It's hard to directly learn 3D knowledge from a single 2D image, so we apply a 2D model F_θ to learn rich 2D priors:

$$s^{2d}, d, c^{2d}, o^{mp} = F_\theta(x), \tag{1}$$

where $s^{2d} \in [0,1]^{H\times W\times C}$ is 2D semantics with C categories. $d \in \mathbb{R}^{H\times W}$ is the depth map. $c^{2d} \in \mathbb{R}^{N\times 3}$ is predicted locations of N instance centers ($\mathbb{R}^{N\times 2}$) with corresponding category labels ($\{0,1,...C-1\}^{N\times 1}$). $o^{mp} \in [0,1]^{H\times W\times M}$ is the estimated multi-plane occupancy which presents whether the M planes at different depths are occupied by 3D things or stuff, and the default M is set as 128.

3.2. Occupancy-aware Feature Lifting

After obtaining the learned 2D priors, we need to lift them to 3D features for the following training. Here, an occupancy-aware feature lifting block is designed for this goal, as shown in Figure. 3. First, we lift the 2D semantics s^{2d} to coarse 3D semantics I_s^{3d} in the whole region of things and stuff rather than only on the front-view surface adopted by previous work [6],

$$I_s^{3d}(u,v,z) = \begin{cases} s^{2d}(K_{cam}^{-1}[u,v,1]), & \text{if } z \geq d(u,v) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

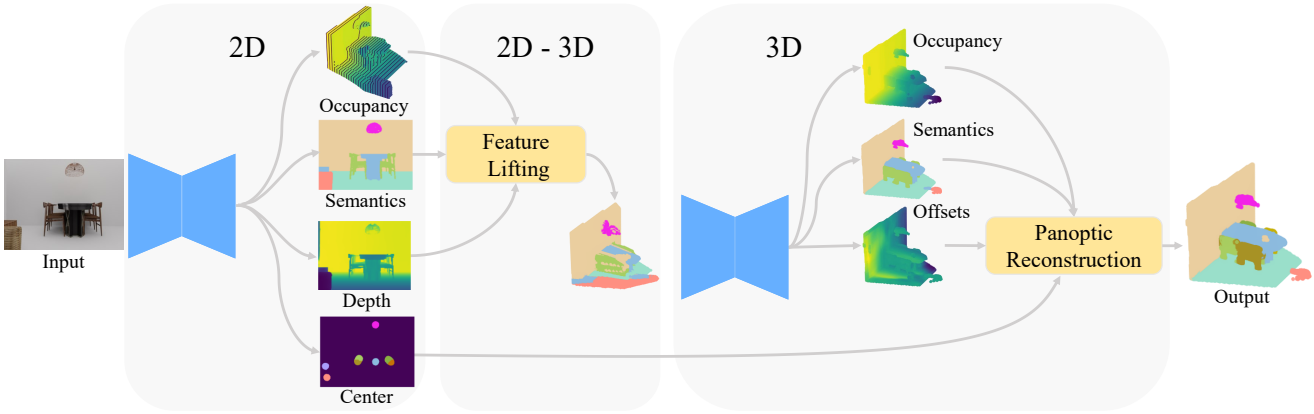

Figure 2. **The illustration of our framework.** Given a single image, we first predict 2D priors by 2D model, then lift 2D priors to 3D voxels by our occupancy-aware lifting, and finally predict 3D results using the 3D model and obtain panoptic 3D scene reconstruction results in a bottom-up manner.

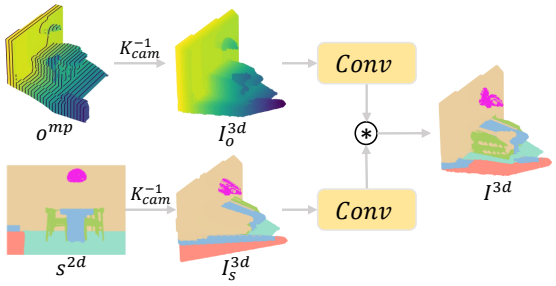

Figure 3. **Occupancy-aware Lifting.** We lift multi-plane occupancy and 2D semantics predicted by the 2D model to 3D features. $*$ is Hadamard product.

where K_{cam} is the camera intrinsic matrix, $d(u,v)$ is depth at location (u,v). The region $z < d(u,v)$ is free space, where is set 0 to ignore.

Then, we resort to multi-plane occupancy o^{mp} learned in the 2D stage to remove the meaningless region of the coarse 3D semantics I_s^{3d} and obtain the lifted 3D features. Formally, the lifted 3D features are calculated as the product of I_s^{3d} and coarse 3D occupancy I_o^{3d},

$$I^{3d} = Conv(I_s^{3d}) * Conv(I_o^{3d}), \text{where}$$
$$I_o^{3d}(u,v,z) = \begin{cases} o^{mp}(K_{cam}^{-1}[u,v,z]), & \text{if } z \geq d(u,v) \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $Conv$ is a Conv-BN-ReLU block. $*$ is Hadamard product.

As shown in Figure. 3, our multi-plane occupancy can give supplementary shape cues for the occluded region, thus the lifted features are capable to serve as a good 3D initialization for the following 3D refinement, greatly reducing the pressure of the 3D refinement model.

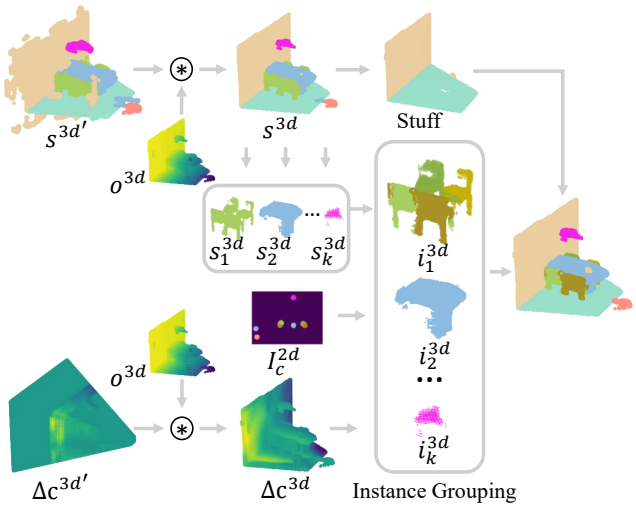

Figure 4. **Panoptic Reconstruction.** The predicted 3D semantics and 3D offsets are first refined by 3D occupancy, and then the reconstructed 3D results are combined with 2D instance centers for 3D instance grouping, and finally, 3D instances and stuff are combined to obtain panoptic 3D scene reconstruction. $*$ is Hadamard product.

3.3. Bottom-up Panoptic Reconstruction

Usually, the lifted 3D features are coarse and cannot be used for panoptic reconstruction directly. To refine the coarse features, a powerful 3D encoder-decoder model G_ϕ is used to predict 3D occupancy, 3D semantic map, and 3D offsets:

$$s^{3d'}, \triangle c^{3d'}, o^{3d} = G_\phi(I^3), \quad (4)$$

where $s^{3d'}, \triangle c^{3d'}, o^{3d}$ is refined 3D semantic map, 3D offsets and 3D occupancy, respectively.

The panoptic reconstruction utilizes the refined 3D results for 3D reconstruction and 3D panoptic segmentation,

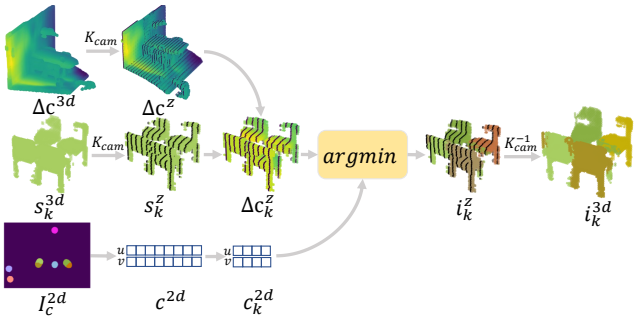

Figure 5. **Instance Grouping.** We convert both 2D instance centers and 3D offsets of each category at multi-plane to group 3D instances.

as shown in Figure. 4. For 3D reconstruction, guided by the 3D occupancy o^{3d}, we can obtain reconstructed semantics by $s^{3d} = s^{3d'} * o^{3d}$ and reconstructed offsets by $\triangle c^{3d} = \triangle c^{3d'} * o^{3d}$, where $*$ is Hadamard product. For 3D panoptic segmentation, we need to assign the instance ids to the voxels of the *things*. To achieve this, we propose grouping instances with the estimated 2D instance centers, 3D offsets, and 3D semantics.

The proposed instance grouping block is shown in Figure. 5. We first convert 3D offsets $\triangle c^{3d}$ to multi-plane by $\triangle c^z = \triangle c^{3d}(K_{cam}[u, v, z])$, where $z \in \{0, 1, ..., M-1\}$ corresponds to different depths. Then multi-plane semantics of category k can also be calculated by $s_k^z = s_k^{3d}(K_{cam}[u, v, z])$. And the 3D offsets of category k can be calculated by $\triangle c_k^z = \triangle c^z * s_k^z$.

Meanwhile, we can get 2D instance centers c^{2d} from 2D center map I_c^{2d}, and then the instance centers of category k, c_k^{2d}, can be indexed from c^{2d}. Finally, 2D instance centers and 3D offsets of category k are combined to group 3D instance at multi-plane:

$$i_k^z(u, v) = argmin_{k_j} \|c_{k_j}^{2d} - (u + \triangle c_k^z(u, v)_u, v + \triangle c_k^z(u, v)_v)\|, \quad (5)$$

where $c_{k_j}^{2d} \in \mathbb{R}^2$ is the jth 2D instance center of category k. i_k^z is the predicted instance id at depth z. The 3D instance id of category k at location (u, v, z) can be calculated by $i_k^{3d}(u, v, z) = i_k^z(u, v)(K_{cam}^{-1}[u, v, z])$.

Combining the stuff from 3D semantics, and the 3D instances grouped by our instance grouping block, we finally predict the panoptic 3D scene reconstruction results from a single image.

3.4. Loss for BUOL

The total loss for the proposed BUOL contains 2D loss and 3D loss. The 2D priors training loss is defined as follows:

$$\mathcal{L}^{2d} = w_p^{2d}\mathcal{L}_p^{2d} + w_d^{2d}\mathcal{L}_d^{2d} + w_o^{mp}\mathcal{L}_o^{mp} \quad (6)$$

where weights w_p^{2d},w_d^{2d} and w_o^{mp} are used to balance the objective. The panoptic segmentation loss is

$$\mathcal{L}_p^{2d} = w_s^{2d}CE(s^{2d}, \hat{s}^{2d}) + w_c^{2d}L1(I_c^{2d}, \hat{I}_c^{2d}) \quad (7)$$

which is composed of semantic map *cross entropy* loss and instance center regression $L1$-norm loss. The ground truth center map $\hat{I}_c^{2d}$ are defined as 2D Gaussian-encoded heatmaps centered in instance mass, and the ground truth of 2D instances and 2D semantics are rendered by 3D instances and 3D semantics, respectively. The depth estimation loss $\mathcal{L}_d^{2d}$ follows [18] to penalize the difference between the estimated depth d and the ground truth depth $\hat{d}$ which is generated by the 3D geometry. The multi-plane occupancy loss $\mathcal{L}_o^{mp}$ is defined as:

$$\mathcal{L}_o^{mp} = BCE(o^{mp}, \hat{o}^{mp}), \quad (8)$$

where the $\hat{o}^{mp}$ is obtained by sampling the 3D ground truth occupancy $\hat{o}^{3d}$ at multi-plane, *i.e.* $\hat{o}^{mp} = \hat{o}^{3d}(K_{cam}[u, v, z])$.

The 3D loss of BUOL is composed of 3D occupancy loss, 3D semantic loss, and 3D offset loss, defined as follows:

$$\begin{aligned}\mathcal{L}^{3d} = w_o^{3d}\mathcal{L}_o^{3d}(o^{3d}, \hat{o}^{3d}) + w_s^{3d}CE(s^{3d'}, \hat{s}^{3d}) \\ + w_{\triangle c}^{3d}L1(\triangle c^{3d'}, \triangle \hat{c}^{3d})\end{aligned} \quad (9)$$

where $w_o^{3d}, w_s^{3d}, w_{\triangle c}^{3d}$ are weighting coefficients. The 3D occupancy loss $\mathcal{L}_o^{3d}$ is composed of a binary classification loss BCE and a regression loss $L1$, and the details can be referred to supplemental materials. The ground truth $\triangle \hat{c}^{3d}$ for each voxel is offset between its 2D instance center and location in its nearest depth plane, which can be generated by 3D ground truth instances.

To stabilize the training, we first train 2D model F_θ with $\mathcal{L}^{2d}$. After converging, the 3D loss $\mathcal{L}^{3d}$ is applied to train 3D model G_ϕ.

4. Experiments

In this section, we conduct experiments on the preprocessed synthetic dataset 3D-Front [10] and real-world dataset Matterport3D [2]. We compare our method with state-of-the-art panoptic 3D scene reconstruction methods and provide an ablation study to highlight the effectiveness of each component.

4.1. Experiment Setup

Datasets. 3D-Front [10] is a synthetic indoor dataset with 18,797 room scenes and 11 categories (9 for things, and 2 for stuff) in 6,801 mid-size apartments. To generate data for panoptic 3D scene reconstruction, we follow Dahnert et al. [6], and first randomly sample rooms and camera locations, then use BlenderProc [9] to render RGB images along with depth, semantic map, and instance mask and finally use signed distance function (SDF) to get 3D ground truth. It contains 96,252/11,204/26,933 train/val/test images corresponding to 4,389/489/1,206 scenes, respectively.

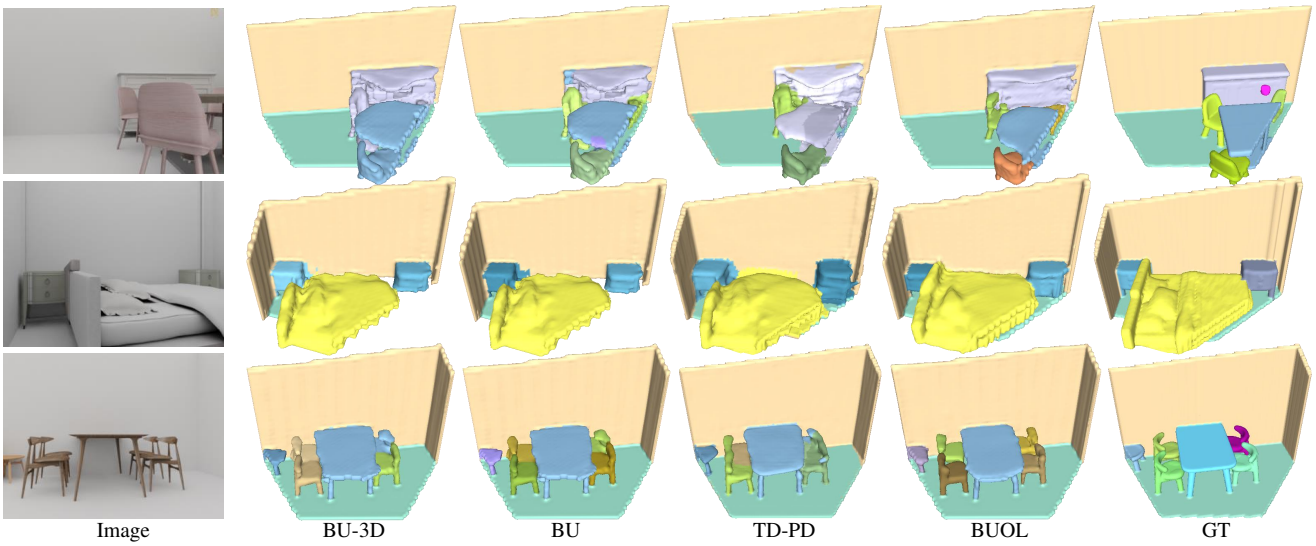

Figure 6. **Qualitative comparisons against competing methods on 3D-Front.** The BUOL and BU denote our Bottom-Up framework w/ and w/o our Occupancy-aware lifting, respectively, and BU-3D denotes the bottom-up framework with instance grouping by 3D centers, and the TD-PD denotes Dahnert et al. [6]*+PD. And GT is the ground truth.

Method	PRQ	RSQ	RRQ	PRQ_{th}	RSQ_{th}	RRQ_{th}	PRQ_{st}	RSQ_{st}	RRQ_{st}
SSCNet [31]+IC	11.50	32.90	33.00	8.03	32.07	24.69	26.95	36.75	70.25
Mesh R-CNN [11]	-	-	-	20.90	38.00	53.20	-	-	-
Total3D [28]	15.08	36.63	40.15	13.77	34.88	38.89	20.94	44.49	45.85
Dahnert et al. [6]*	42.20	55.59	73.19	36.51	51.47	69.21	67.78	74.15	91.09
Dahnert et al. [6]*+PD	47.46	60.48	76.09	42.25	56.90	72.45	70.94	76.59	92.45
Our BUOL	**54.01**	**63.81**	**82.99**	**49.73**	**60.57**	**80.67**	**73.30**	**78.37**	**93.42**

Table 1. Comparison to the state-of-the-art on 3D-Front. "*" denotes the trained model with the official codebase released by the authors.

Matterport3D [2] is a real-world indoor dataset that contains 90 building-scale scenes. For panoptic 3D scene reconstruction, Matterport3D is pre-processed in the same way as 3D-Front to generate the ground truth of 34,737/4,898/8,631 train/val/test images corresponding to 61/11/18 scenes. It contains the same 11 categories as 3D-Front and another stuff category "ceiling".

Metrics. We adopt panoptic reconstruction quality PRQ, reconstructed segmentation quality RSQ, and reconstructed recognition quality RRQ [6] as our metrics. In addition, PRQ_{th} and PRQ_{st} denote PRQ of things and stuff, respectively. PRQ is calculated by the average measure across C categories, with PRQ_k for category k defined as:

$$\begin{aligned} PRQ_k &= RSQ_k * RRQ_k \\ &= \frac{\sum_{(i,\hat{i})\in TP_k} IoU(i,\hat{i})}{|TP_k|} * \frac{2|TP_k|}{2|TP_k| + |FP_k| + |FN_k|} \\ &= \frac{\sum_{(i,\hat{i})\in TP_k} 2IoU(i,\hat{i})}{2|TP_k| + |FP_k| + |FN_k|} \end{aligned} \tag{10}$$

where TP_k, FP_k, and FN_k denote true positives, false positives, and false negatives for category k, respectively, and intersection over union (IoU) is the metric between predicted mask i and ground truth mask $\hat{i}$. The predicted segments are matched with ground truth if the voxelized IoU is no less than 25%. Following Dahnert et al. [6], we set the evaluate resolution for panoptic 3D scene reconstruction to 3cm for synthetic data and 6cm for real-world data.

Implementation. We adopt ResNet-50 [15] as our shared 2D backbone of 2D Panoptic-Deeplab [5], and use three branches to learn rich 2D priors. One decoder with the semantic head is used for semantic segmentation, and one decoder followed by the center head is utilized for instance center estimation. Another decoder with a depth head and multi-plane occupancy head is designed for geometry priors. For the 3D model, we convert 2D ResNet-18 [15] and ASPP-decoder [3] to 3D models as our 3D encoder-decoder, and design 3D occupancy head, 3D semantic head, and 3D offset head for panoptic 3D scene reconstruction. For the two datasets, we apply Adam [21] solver with the initial learning rate 1e-4 combined with polynomial learning rate decay scheduler for 2D learning, and the initial learning rate

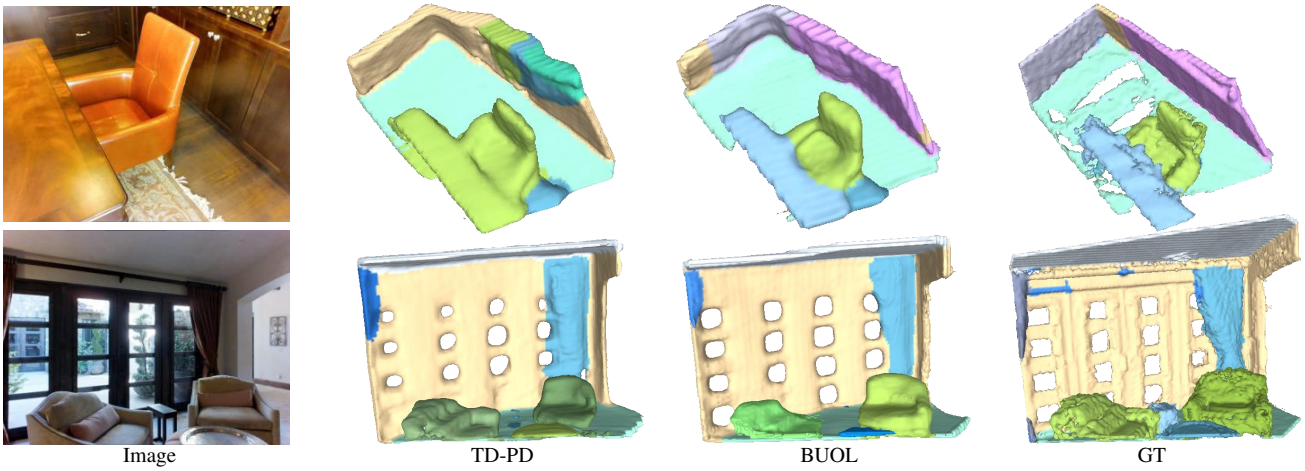

Figure 7. **Qualitative comparisons against competing methods on Matterport3D.** The BUOL denotes our Bottom-Up framework with Occupancy-aware lifting, and the "TD-PD" denotes Dahnert et al. [6]*+PD. And GT is the ground truth.

Method	PRQ	RSQ	RRQ	PRQ_{th}	RSQ_{th}	RRQ_{th}	PRQ_{st}	RSQ_{st}	RRQ_{st}
SSCNet [31]+IC	0.49	21.68	1.50	0.19	22.75	0.59	1.43	20.43	4.43
Mesh R-CNN [11]	-	-	-	6.29	31.12	15.60	-	-	-
Dahnert et al. [6]	7.01	28.57	17.65	6.34	26.06	16.06	10.78	40.03	26.77
Dahnert et al. [6]*+PD	10.08	36.04	22.53	7.33	33.23	16.68	18.33	44.47	40.07
Our BUOL	**14.47**	**45.71**	**30.91**	**10.97**	**45.30**	**23.81**	**24.94**	**46.93**	**52.22**

Table 2. Comparison to the state-of-the-art on Matterport3D. "*" denotes the trained model with the official codebase released by the authors.

5e-4 decayed at 32,000th and 38,000th iteration. During training, we first train the 2D model for 50,000 iterations with batch size 32, then freeze the parameters and train the 3D model for 40,000 iterations with batch size 8. All the experiments are conducted with 4 Tesla V100 GPUs. In addition, we initialize the model with the pre-trained ResNet-50 for 3D-Front, and the pre-trained model on 3D-Front for Matterport3D which is the same as Dahnert et al. [6].

4.2. Comparison with State-of-the-art Methods

3D-Front. For synthetic dataset, we compare BUOL with the state-of-the-art method [6] and other mainstream [11, 28, 31]. The results are shown in Table 1. Our proposed BUOL and Dahnert et al. [6] both outperform other methods a lot. However, with the proposed bottom-up framework and occupancy-aware lifting, our BUOL outperforms the state-of-the-art method by a large margin, +11.81%. For a fair comparison, we replace the 2D segmentation in Dahnert et al. [6] with the same 2D model as ours, denoted as Dahnert et al. [6]+PD (also denoted as TD-PD). Comparing to this method in Table 1, our BUOL also shows an advantage of +6.55% PRQ. The qualitative comparison results in Figure. 6 also show our improvement. In the second row, our BUOL reconstructs the bed better than TD-PD with occupancy-aware lifting. In the last row, our BUOL can recognize all the chairs while TD-PD obtains the sticky chair. In the first rows, both BU and OL in BUOL achieve the better Panoptic 3D Scene Reconstruction results.

Matterport3D. We also compare BUOL with some methods [6, 11, 31] on real-world dataset. The results are shown in Table 2. Our BUOL outperforms the state-of-the-art Dahnert et al. [6] by +7.46% PRQ with the proposed bottom-up framework and occupancy-aware lifting. For fairness, we also compare BUOL with Dahnert et al. [6]+PD, and our method improves the PRQ by +4.39%. Figure. 7 provides the qualitative results. In the first row, our BUOL can segment all instances corresponding to ground truth, which contains a chair, a table and two cabinets, and TD-PD can only segment the chair. In the second row, our BUOL reconstruct the wall and segment curtains batter than TD-PD. In addition, although the highest performance, the PRQ of the Matterport3D is still much lower than that of the 3D-Front due to its noisy ground truth.

4.3. Ablation Study

In this section, we verify the effectiveness of our BUOL for panoptic 3D scene reconstruction. As shown in Table 3, for a fair comparison, TD-PD is the state-of-the-art

Method	PRQ	RSQ	RRQ	PRQ_{th}	PRQ_{st}
TD-PD	47.46	60.48	76.09	42.25	70.94
BU-3D	46.73	59.17	76.68	41.77	69.07
BU	50.76	60.66	81.94	46.80	68.55
BUOL	**54.01**	**63.81**	**82.99**	**49.73**	**73.30**

Table 3. Ablation study of the proposed method vs baselines.

Inst/Sem	Assignment	PRQ	RSQ	RRQ
Instance	random	47.46	60.48	76.09
	category	48.92	61.20	77.48
Semantics	category	50.76	60.66	81.94

Table 4. Comparison of different assignments.

top-down method [6] with the same 2D Panoptic-Deeplab as ours, which is our baseline method. BU denotes our proposed bottom-up framework. Different from BU, 2D Panoptic-Deeplab in TD-PD is used to predict instance masks instead of semantics and instance centers. BU-3D denotes the bottom-up framework which groups 3D instances by the predicted 3D centers instead of 2D centers.

Top-down vs. Bottom-up. TD-PD and BU adopt the same 2D model. The former lifts the instance masks to the 3D features, while the latter lifts the semantic map and groups 3D instances with 2D instance centers. Comparing the two settings in Table 3, BU significantly boosts the performance of RRQ by +5.85% which proves our bottom-up framework with proposed 3D instance grouping achieves more accurate 3D instance mask than direct instance mask lifting. The drop of PRQ_{st} for stuff may come from the lower capability of used 3D ResNet + ASPP, compared with other methods equipped with stronger but memory-consuming 3D UNet. Overall, the proposed bottom-up framework achieves +3.3% PRQ better than the top-down method. Figure. 6 provides qualitative comparison of BU and TD-PD. The bottom-up framework performs better than the top-down method. For example, in the last row of Figure. 6, TD-PD fails to recognize the four chairs, while BU reconstructs and segments better.

2D instance center vs. 3D instance center. We also compare the 2D instance center with the 3D instance center for 3D instance grouping. To estimate the 3D instance center, the center head is added to the 3D refinement model, called BU-3D. Quantitative comparing BU-3D and BU in Table 3, we can find the PRQ_{st} for stuff is similar, but when grouping 3D instances with the 2D instance centers, the PRQ_{th} for thing has improved by 5.03%, which proves 3D instance grouping with 2D instance center performing better than that with the 3D instance center. We conjecture that the error introduced by the estimated depth dimension may impact the position of the 3D instance center. Meanwhile, grouping in multi-plane is easier for 3D offset learning via reducing one dimension to be predicted. Qualitative comparison BU with BU-3D is shown in Figure. 6, due to inaccurate 3D instance centers, the result of BU-3D in the last row misclassifies a chair as a part of the table, and the result in the first row does not recognize one chair.

Voxel-reconstruction ambiguity. We propose occupancy-aware lifting to provide the 3D features in full 3D space to tackle voxel-reconstruction ambiguity. Quantitative comparing BUOL with BU in Table 3, our proposed occupancy-aware lifting improves PRQ_{th} by 2.93% for thing and PRQ_{st} by 4.75% for stuff, which verifies the effectiveness of multi-plane occupancy predicted by the 2D model. It facilitates the 3D model to predict more accurate occupancy of the 3D scene. In addition, with our occupancy-aware lifting, PRQ_{st} for stuff of the 3D model with ResNet-18 + 3D ASPP outperforms the model TD-PD with 3D U-Net by 2.36% PRQ. As shown in Figure. 6, with occupancy-aware lifting, BUOL reconstructs 3D instances better than others. For example, in the second row of Figure. 6, BUOL can reconstruct the occluded region of the bed, while other settings fail to tackle this problem.

Instance-channel ambiguity. To analyze the instance-channel ambiguity in the top-down method, we conduct experiments based on TD-PD, as shown in Table 4. When lifting instance masks with random assignment, the model achieves 47.76% PRQ. However, fitting random instance-channel assignment makes the model pay less attention to scene understanding. To reduce the randomness, we try to apply instance-channel with sorted categories, which improves PRQ to 48.92%. Because an arbitrary number of instances with different categories may exist in an image, resulting in the randomness of instance number even for the same category. To further reduce the randomness, our proposed bottom-up method, also called BU, lifts semantics with deterministic assignment, and gets 50.76% PRQ, which proves that the pressure of the 3D model can be reduced with the reduction in the randomness of instance-channel assignment and the bottom-up method can address the instance-channel ambiguity.

5. Conclusion

In this paper, we propose a bottom-up framework with occupancy-aware lifting (BUOL) for panoptic 3D scene reconstruction. Our bottom-up framework lifts 2D semantics instead of 2D instances to 3D to avoid instance-channel ambiguity, and the proposed occupancy-aware lifting leverages multi-plane occupancy predicted by 2D model to avoid voxel-reconstruction ambiguity. BUOL outperforms state-of-art approaches with top-down framework for both 3D reconstruction and 3D perception in a series of experiments. We believe that BUOL will drive the area of panoptic 3D scene reconstruction from a single image forward.

References

[1] Aljaz Bozic, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular rgb scene reconstruction using transformers. *Advances in Neural Information Processing Systems*, 34:1403–1414, 2021. 2

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 2, 5, 6

[3] 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. 6

[4] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15467–15476, 2021. 3

[5] 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. 6

[6] 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, 2, 3, 5, 6, 7, 8

[7] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–468, 2018. 3

[8] Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott Reed, Jürgen Sturm, and Matthias Nießner. Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2018. 3

[9] Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Youssef Zidan, Dmitry Olefir, Mohamad Elbadrawy, Ahsan Lodhi, and Harinandan Katam. Blenderproc. *arXiv preprint arXiv:1911.01911*, 2019. 5

[10] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021. 2, 5

[11] 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, 6, 7

[12] Benoit Guillard, Edoardo Remelli, and Pascal Fua. Uclidnet: Single view reconstruction in object space. *Advances in Neural Information Processing Systems*, 33:3244–3253, 2020. 2

[13] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2940–2949, 2020. 3

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[15] 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. 6

[16] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4421–4430, 2019. 3

[17] Ji Hou, Angela Dai, and Matthias Nießner. Revealnet: Seeing behind objects in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2098–2107, 2020. 3

[18] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1043–1051. IEEE, 2019. 5

[19] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1610–1618, 2021. 3

[20] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pages 4867–4876, 2020. 3

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[22] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019. 2

[23] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. Mask2cad: 3d shape prediction by learning to segment and retrieve. In *European Conference on Computer Vision*, pages 260–277. Springer, 2020. 1, 3

[24] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. Patch2cad: Patchwise embedding learning for in-the-wild shape retrieval from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12589–12599, 2021. 3

[25] Jean Lahoud, Bernard Ghanem, Marc Pollefeys, and Martin R Oswald. 3d instance segmentation via multi-task metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9256–9266, 2019. 3

[26] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 3351–3359, 2020. 1, 3

[27] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *European conference on computer vision*, pages 414–431. Springer, 2020. 1, 3

[28] 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, 6, 7

[29] Yinyu Nie, Ji Hou, Xiaoguang Han, and Matthias Nießner. Rfd-net: Point scene understanding by semantic instance reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4608–4618, 2021. 3

[30] Stefan Popov, Pablo Bauszat, and Vittorio Ferrari. Corenet: Coherent 3d scene reconstruction from a single rgb image. In *European Conference on Computer Vision*, pages 366–383. Springer, 2020. 2

[31] 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. 6, 7

[32] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2569–2578, 2018. 3

[33] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2690–2698, 2019. 2

[34] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3947–3956, 2019. 3