Pano2Scene: 3D Indoor Semantic Scene Reconstruction from a Single Panorama Image

Wei Zeng[1]
w.zeng@uva.nl
Sezer Karaoglu[1,2]
s.karaoglu@3duniversum.com
Theo Gevers[1,2]
th.gevers@uva.nl

[1] Computer Vision Laboratory
University of Amsterdam
The Netherlands
[2] 3DUniversum
Science Park 400
The Netherlands

Abstract

3D indoor semantic scene reconstruction from 2D images is challenging as it requires both scene understanding and object reconstruction. Compared to perspective images, panoramas provide larger field of view and carry more scene information. In this paper, to reconstruct the 3D indoor semantic scene from a single panorama image, we propose a pipeline that jointly learns to predict the 3D scene layout, complete the object shapes and reconstruct the full scene point cloud. Experiments on the Stanford 2D-3D dataset demonstrate the generality and suitability of the proposed method.

1 Introduction

3D indoor semantic scene reconstruction from 2D images is important for different computer vision applications such as robot-environment interaction and interior design. At the same time, panorama images are currently enjoying a surge in popularity and witnessing increased adoption in robotic applications and marketing productions. In this paper, we focus on the reconstruction of a full 3D indoor semantic scene point cloud from a single panorama image.

Most of the previous works on semantic scene reconstruction are dealing with *perspective* images. Tulsiani *et al*. [24] propose a voxel-based representation to reconstruct the 3D structure of the scene, but the resolution is limited and the computational cost for scene-level voxel reconstruction yields considerable amount of overhead during training and inference. Izadinia *et al*. [16] reconstruct a scene by retrieving similar meshes from a large database of furniture CAD models. However, the method requires many iterations of model rendering and the accuracy is highly dependent on the similarity of the CAD models in the database. Recently, [21] proposes a method to jointly reconstruct the room layout, object bounding boxes and meshes from a single perspective image. However, the method requires dense and clean meshes for proper object mesh reconstruction, which is tedious and labor-consuming for real scenes. Furthermore, previous methods can only partially reconstruct indoor scenes as the perspective images have limiting effects on the field of view. In contrast to previous

methods, our approach reconstructs the full 3D indoor semantic scene point cloud from a single 2D panorama image. The proposed pipeline jointly learns to predict the 3D scene (room) layout, complete object (furniture) shapes and reconstruct the full scene point cloud.

Previous methods on panorama layout estimation predict the 2D layout edges and corners in the input panorama and by post-processing them to match the (Manhattan) 3D layout [19, 22, 33]. However, object clutter poses a challenge to properly extract the occluded edges and corners. In addition, constraints are imposed in the optimization process to compute the 2D to 3D conversion. In contrast to existing methods, our method directly estimates the layout depth map. The predicted layout depth map can serve as a coarse 3D layout by converting it to a 3D point cloud of the scene layout. Hence, the proposed method does not require extra constraints for the 3D optimization.

As to object shape completion, most previous work focus on object-level completion [10, 23, 28, 29], in which the input are normally clean partial point clouds. In contrast, our method aims to full scene-level object completion. Due to the accuracy limitation of scene-level depth estimation and instance segmentation, the partial object point clouds inferred from the predicted depth and instance masks are typically noisy and deformed. The proposed method projects the noisy global feature vectors onto the manifold of the clean ones to overcome the noise and deformations in the predicted partial point clouds.

To impose the global constraints for the scene-level reconstruction and enforce consistency between the reconstructed scene point cloud and the panorama input, it is critical to jointly train our pipeline end-to-end. The proposed method equirectangularly projects the inferred complete object point clouds back on the 2D panorama to minimize the losses of the projected object masks and depth with respect to the ground truth. Experimental results indicate that joint training further advances the reconstruction accuracy. To the best of our knowledge, our approach is the first to reconstruct the full 3D indoor semantic scene point cloud from a single panorama image.

In summary, our contributions are as follows:

- A unified semantic scene reconstruction pipeline is proposed to reconstruct the full 3D indoor semantic scene point cloud from a single panorama image.
- To recover the 3D layout of the scene, the method estimates the layout depth map and reconstruct the parameterized 3D layout. Our method obtains state-of-the-art performance for 3D layout estimation from a single panorama image.
- To generate the full point cloud of objects in the scene, the method completes the object point cloud from the visible partial point cloud via global feature vector mapping to obtain robustness to noise and deformations in the predicted partial point cloud.
- To enforce consistency between the reconstructed scene point cloud and the panorama input, the method projects the inferred complete object point cloud back on the 2D panorama and jointly trains the full pipeline end-to-end.

2 Related Work

Layout Prediction: Traditional methods treat this task as an optimization problem. Delage *et al*. [5] propose a dynamic Bayesian network model to recover the 3D model of the indoor scene. Hedau *et al*. [12] model the room with a parametric 3D box by iteratively localizing

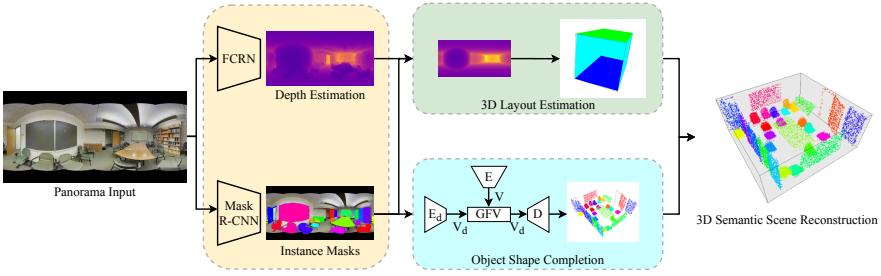
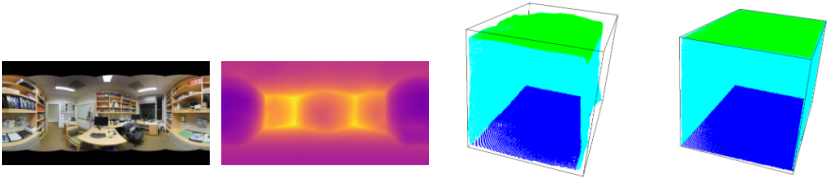


Figure 1: Overview of our pipeline. The whole process consist of three modules: 1) depth and instance segmentation; 2) 3D layout estimation; 3) object shape completion. The output is the reconstructed 3D indoor semantic scene point cloud of the panorama input.

clutter and refitting the box. Recently, neural network-based methods took stride in tackling this problem. Zou *et al*. [33] predict the layout boundary and corner map directly from the input panorama. Yang *et al*. [27] leverage both the equirectangular panorama-view and the perspective ceiling-view to learn different cues of the room layout. Sun *et al*. [22] encode the room layout as three 1D vectors and propose to recover the 3D room layouts from 1D predictions. Other works aims to leverage depth information for room reconstruction [20, 31, 32], but they all deal with perspective images and use ground truth depth as input. In contrast to previous methods, in this paper, we predict the layout depth map as the intermediate representation to recover the 3D layout of the input panorama.

Point Cloud Completion: Fan *et al*. [8] propose an architecture consisting of an encoder which encodes the input into an embedding, and a decoder which generates the point cloud from the embedding. Yang *et al*. [28] generate a point cloud structured as a manifold through a series of deformation (folding) operations on the Euclidean plane. Yuan *et al*. [29] combine the fully-connected decoder and the folding decoder to generate point clouds in two stages. Groueix *et al*. [10] design a decoder that learns a manifold by computing a mapping from the Euclidean plane to the ground-truth point cloud. Tchapmi *et al*. [23] propose a decoder that generates a structured point cloud without assuming any specific structure or topology on the underlying point set. [30] decomposes the single-view point cloud generation into depth estimation and point cloud completion. The proposed method extends this concept to scene-level reconstruction. The global feature vector mapping is used to obtain robustness to noise and deformations in the inferred partial point cloud.

Semantic Scene Reconstruction: Indoor scene reconstruction from images is an essential task in computer vision and graphics. A number of existing approaches [4, 14] estimate the object poses together with the room layout [5, 12]. However, these methods focus on the prediction of the 3D bounding box for each object. Other methods [15, 16] use 3D model retrieval modules to improve the shape quality in scene reconstruction. [9, 21, 24] only need a single image as input to reconstruct multiple object shapes in a scene. [9] produces object meshes, but ignores scene context and suffers from the artifacts of mesh generation on cubified voxels. [24] is designed for voxel reconstruction with limited resolution. [21] requires clean and dense meshes to supervise object mesh reconstruction. Different from all existing methods, our proposed method combines depth estimation, instance segmentation, 3D layout estimation and object shape completion through a joint learning (end-to-end) pipeline to reconstruct the 3D indoor semantic scene point cloud from a single panorama image.



(a) Panorama Input (b) Estimated Layout Depth (c) 3D Layout before Optimization (d) Final 3D Layout

Figure 2: Demonstration of the 3D layout estimation module. This module estimates the (b) layout depth of the (a) panorama input, then recovers the parameterized 3D scene layout. (c) is the 3D layout point cloud directly recovered from the estimated layout depth map. The (d) final 3D layout constrain the layout shape to Manhattan world assumption.

3 Method

The overview of the proposed pipeline is illustrated in Figure 1. Our pipeline consists of three modules: (1) the depth and instance segmentation module predicts the depth map and instance segmentation masks from the panorama input; (2) the 3D layout estimation module recovers the 3D scene layout from the estimated layout depth map; (3) the object shape completion module infers the complete object point cloud from the visible partial point cloud. The proposed pipeline reconstructs the full scene point cloud by embedding the outputs of all modules together by joint training and inference. The details of each module are discussed in this section.

3.1 Depth and Instance Segmentation

The first module of the proposed pipeline consists of depth estimation as well as instance segmentation from a single panorama image.

Depth Estimation: For this task, different CNN models for depth estimation exist [5, 6, 7, 18, 25, 32]. In our work, the fully convolutional ResNet-50 architecture proposed by Laina *et al*. [18] is used. However, the equirectangular panorama may suffer from horizontal distortions. To reduce this distortion effect, the encoder uses a modified input block. As shown by [32], the input block uses rectangle filters and varies the resolution to account for different distortion levels. One more up-projection layer is also added to the original FCRN network architecture so that the output depth preserves the input resolution.

Instance Segmentation: In our method, Mask R-CNN [11] is used to segment images at the instance-level to obtain the object labels and corresponding masks. The ResNet-101 backbone is used and initialized with pre-trained weights on the MSCOCO dataset [1].

3.2 3D Layout Estimation

To obtain the global geometric structure of the scene, the proposed approach predicts the 3D scene layout. Instead of predicting 2D representations (e.g. edge and corner maps), our method directly predicts the layout depth map from the input panorama. The input of this proposed module is the concatenation of the predicted depth map and the instance segmentation masks from previous module. A ResNet-18 is used to build our encoder for the layout depth estimation network. The decoder restores the original input resolution by

means of up-sampling operators followed by 3×3 convolutions. The skip connections are also added to prevent shifting of the prediction results during the up-sampling step. The output is the estimated layout depth map with the same resolution as the input panorama.

To constrain the layout so that the floor and ceiling are planar and walls are perpendicular to each other (Manhattan world assumption), the proposed method recovers the parameterized 3D layout through optimization in 3D space, as shown in Figure 2. Using the scene layout point cloud converted from the predicted layout depth map, the floor/ceiling plan map is obtained by projecting the point cloud to the XZ plane. Similar to [27], a regression analysis is applied on the edges of the floor plan map and cluster them into sets of horizontal and vertical lines in 3D space. Then, the floor plan is recovered by using the straight, axis-aligned, wall-floor boundaries. The room height is efficiently computed by voting for the ceiling-floor distance along the Y axis. As a result, the recovered parameterized 3D layout meets the Manhattan world assumption.

3.3 Object Shape Completion

After the depth map is estimated, the point cloud of the visible scene is calculated based on the camera model. With the predicted instance masks, the visible partial point cloud of each object is inferred. Limited by the accuracy of the scene-level depth estimation and instance segmentation, the inferred partial point clouds are typically noisy and deformed.

Object Point Cloud Completion: The full point cloud is inferred by learning a mapping from the space of partial observations to the space of complete shapes. To this end, an encoder-decoder network architecture similar to PCN [29] is used. The aim of the encoder $\mathbf{E}$ is to concisely represent the geometric information of the partial point cloud by a global feature vector $\mathbf{v}$. The decoder $\mathbf{D}$, taking the global feature vector $\mathbf{v}$ as input, first produces a sparse point cloud by a fully-connected decoder [8]. Then, a detailed point cloud is obtained by a folding-based decoder [28]. In order to learn a prior over the complete 3D point cloud, we train the encoder-decoder network ($\mathbf{E}$, $\mathbf{D}$) taking the partial point cloud inferred from the ground truth depth and instance masks, i.e. the clean partial point cloud, as input.

Global Feature Vector Mapping: If the inferred noisy and deformed partial point cloud is passed through the encoder $\mathbf{E}$, a "noisy" global feature vector is obtained, i.e. one that does not lie on the manifold of representations learnt by the above encoder-decoder network ($\mathbf{E}$, $\mathbf{D}$). Hence, the task of completing the point cloud is reduced to projecting the noisy global feature vector onto the manifold of clean ones. The cleaned global feature vector can then be passed through the decoder $\mathbf{D}$ to obtain a complete point cloud. Taking the estimated partial object point cloud as input, another encoder $\mathbf{E}_d$ is trained. As shown in Figure 1, the global feature vector $\mathbf{v}_d$ from $\mathbf{E}_d$ is mapping to the clean global feature vector $\mathbf{v}$ from $\mathbf{E}$. Then the global feature vector $\mathbf{v}_d$ is passed through the pre-trained decoder $\mathbf{D}$ to output the completed point cloud. The parameters of $\mathbf{D}$ are not updated during this step. Through the global feature vector mapping in latent space, the network becomes robust to noise and deformations in the predicted partial point cloud.

3.4 Joint Learning for Semantic Scene Reconstruction

In this section, the learning targets are discussed with the corresponding loss functions, and we describe our joint loss for end-to-end training.

Individual Losses: For depth estimation, the reverse Huber (berHu) loss is used:

$$L_{dep} = \begin{cases} e_i & e_i \leq c \\ \frac{e_i^2+c^2}{2c} & else \end{cases} \tag{1}$$

where $e_i = \|d_i - g_i\|_1$, d_i and g_i denote the predicted and ground truth depth respectively. We follow [18] to set $c = \frac{1}{5} max_i(\|d_i - g_i\|_1)$.

For instance segmentation, the same losses as Mask R-CNN [11] are used:

$$L_{seg} = l_{cls} + l_{box} + l_{mask} \tag{2}$$

where l_{cls}, l_{box} and l_{mask} denote the classification loss, the bounding-box loss and the mask loss, respectively.

For 3D layout estimation, to account for both pixel-wise accuracy and spatially coherent results of the layout depth estimation, the depth gradient and normals are incorporated with the reverse Huber loss, as done by [13]:

$$L_{layout\_dep} = l_{depth} + l_{gradient} + l_{normal} \tag{3}$$

where l_{depth}, $l_{gradient}$ and l_{normal} denote the reverse Huber loss, the depth gradient loss and the normal loss, respectively.

For point cloud completion, the Earth Mover's Distance (EMD) is used to measure the distance between the predicted point cloud P_p and the ground truth point cloud P_{gt}. The EMD requires $P_p, P_{gt} \subseteq R^3$ to have equal size $s = |P_p| = |P_{gt}|$, defined by:

$$L_{EMD} = \frac{1}{|s|} \min_{\phi: P_p \to P_{gt}} \sum_{x \in P_p} \|x - \phi(x)\|_2^2 \tag{4}$$

where $\phi : P_p \to P_{gt}$ is a bijection. As to the latent space mapping for the global feature vectors, the pipeline minimizes the L2 distance between the global feature vector v_d from the estimated partial object point cloud and the one v from the ground truth.

$$L_{gfv} = \|v_d - v\|_2^2 \tag{5}$$

Joint Losses: To enforce consistency between the reconstructed scene point cloud and the panorama input, we define: (1) projected mask loss L_{proj_m}, as the average binary cross-entropy loss between the projected object masks and the ground truth masks; (2) projected depth loss L_{proj_d}, as the L2 loss between the projected object depth and the ground truth object depth. As done by [17], we only use the non-zero pixels in the projected depth map and search their neighbors to reduce the influence of projection errors. With these two proposed joint losses, the poses and scales of the reconstructed object point cloud are constrained and consistent with the panorama input image.

End-to-end joint training using all loss functions is defined by:

$$\begin{aligned} L = \lambda_{dep} L_{dep} + \lambda_{seg} L_{seg} + \lambda_{layout\_dep} L_{layout\_dep} \\ + \lambda_{EMD} L_{EMD} + \lambda_{gfv} L_{gfv} + \lambda_{proj\_m} L_{proj\_m} + \lambda_{proj\_d} L_{proj\_d} \end{aligned} \tag{6}$$

where λ_* are the weights used to balance contribution of each component loss.

Method	3D IoU(%)	Corner error(%)	Pixel error(%)
LayoutNet [33]	76.33	1.04	2.70
DuLa-Net [27]	79.36	0.79	2.55
HorizonNet [22]	79.79	0.71	**2.39**
wo/ depth & semantic	77.25	1.23	3.40
w/ pred. depth	81.82	0.84	3.02
w/ pred. semantic	78.26	1.25	3.31
Ours	**84.88**	**0.70**	2.40

Table 1: Quantitative results and ablation study of 3D layout estimation on the Stanford 2D-3D dataset. Our method outperforms all existing methods.

4 Experiments

In this section, the performance of our proposed pipeline is evaluated on 3D layout estimation, object point cloud completion and 3D semantic scene reconstruction.

Dataset: The dataset used for training and testing is the Stanford 2D-3D dataset [2]. The Stanford 2D-3D dataset contains 1413 *RGB* panoramic images collected from 6 large-scale indoor environments, including offices, conference rooms, and other open spaces. For each room there is real-scanned point cloud and annotated as furniture (*board*, *bookcase*, *chair*, *sofa*, *table*) or building elements (*ceiling*, *floor*, *wall*, *door*, *window*) or *clutter*.

Metrics: The 3D layout estimation is evaluated by: 1) 3D IoU: intersection over union between predicted 3D layout and the ground truth; 2) Corner error (CE): average Euclidean distance between predicted corners and ground-truth corners; 3) Pixel error (PE): pixel-wise error between predicted surface classes and the ground truth.

The object point cloud completion is evaluated by the Chamfer Distance (CD) and Earth Mover's Distance (EMD). The Chamfer Distance measures the difference between the predicted point cloud P_p and the ground truth point cloud P_{gt}, defined by:

$$L_{CD} = \frac{1}{|P_p|}\sum_{x\in P_p} \min_{y\in P_{gt}} \|x-y\|_2^2 + \frac{1}{|P_{gt}|}\sum_{y\in P_{gt}} \min_{x\in P_p} \|x-y\|_2^2 \tag{7}$$

The 3D scene reconstruction is also evaluated by the Chamfer Distance and Earth Mover's Distance.

Implementation: To initialize our networks properly, the pipeline follows a two-stage training procedure: we first train depth estimation, instance segmentation, 3D layout estimation, and object shape completion network individually. Then, we combine all the networks and jointly train the pipeline end-to-end with the loss L in Equation 6.

4.1 3D layout Estimation:

A quantitative comparison of different methods for 3D layout estimation on the Stanford 2D-3D dataset is summarized in Table 1. LayoutNet [33] predicts the layout boundary and corner maps directly from the input panorama. DuLa-Net [27] leverages both the equirectangular panorama-view and the perspective ceiling-view to learn different cues for room layout. HorizonNet [22] encodes the room layout as three 1D vectors and proposes to recover the 3D room layout from 1D predictions by a RNN. Besides, ablation studies of the proposed method are conducted as: 1) *wo/ depth&semantic*: predicting the layout depth directly from the input; 2) *w/ pred. depth*: only with the predicted depth; 3) *w/ pred. semantic*: only with the predicted semantic. The proposed method shows state-of-the-art performance and

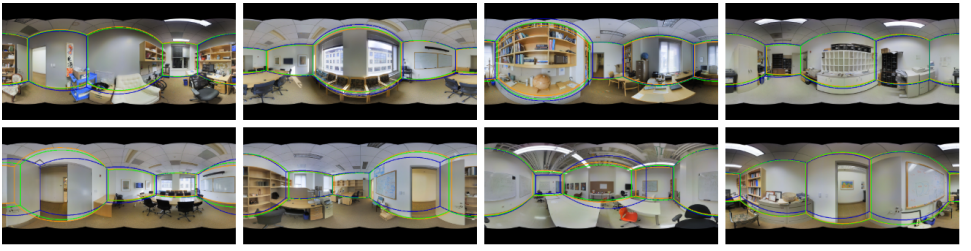


Figure 3: Qualitative comparison on 3D layout estimation. For each example, the predicted layout (HorizonNet [22]: blue, our proposed method: green) is shown together with the ground truth (orange) under an equirectangular view.

		Window	Door	Table	Chair	Sofa	Bookcase	Board	Reconstructed Scene
CD	PointNet-FC (baseline)	0.011	0.022	0.039	0.015	0.043	0.018	0.006	0.022
	FoldingNet [28]	0.009	0.018	0.035	0.009	0.039	0.016	0.004	0.019
	PCN [29]	0.009	0.019	0.022	0.007	0.030	0.014	0.005	0.015
	Ours (with GFV)	**0.004**	0.006	0.014	0.007	0.019	0.009	**0.002**	0.009
	Our Final (with GFV & joint training)	**0.004**	**0.004**	**0.011**	**0.004**	**0.017**	**0.008**	**0.002**	**0.007**
EMD	PointNet-FC (baseline)	0.017	0.039	0.068	0.027	0.080	0.039	0.008	0.040
	FoldingNet [28]	0.015	0.032	0.073	0.022	0.051	0.032	0.006	0.033
	PCN [29]	0.014	0.031	0.033	0.016	0.048	0.021	0.006	0.024
	Ours (with GFV)	**0.008**	0.011	0.025	0.015	0.033	0.018	**0.004**	0.016
	Our Final (with GFV & joint training)	**0.008**	**0.007**	**0.021**	**0.010**	**0.030**	**0.015**	**0.004**	**0.013**

Table 2: Point cloud reconstruction. Our proposed method performs the best for object shape reconstruction and overall scene reconstruction.

outperforms other existing methods. By leveraging the layout depth map as an intermediate representation, the proposed network abstracts the geometric structure of the scene from both a local and global perspective. This results in more geometric cues for the scene layout prediction and is less affected by occlusions.

The qualitative results for the 3D layout estimation are shown in Figure 3. The proposed method outperforms the other methods and shows robustness to occlusion. As presented by the first two examples in the second row, with more global structure information and semantic content, the detection of occluded corners are more accurate. As shown in the third example in the second row, since the proposed method explicitly incorporates depth information, the corners are located more precisely (avoiding locations in the middle of the wall which has continuous depth).

4.2 Object Point Cloud Completion:

To evaluate our proposed object shape completion module, the results are compared with the baseline PointNet-FC as well as state-of-the-art methods FoldingNet [28] and PCN [29], as shown in Table 2. The baseline PointNet-FC consists of the PointNet encoder and a fully connected decoder with 4 layers of output dimensions 256, 512, 1024, and $3 \times N$ ($N = 1024$ in our experiments). The FoldingNet proposes a folding-based decoder that deforms a canonical 2D grid onto the underlying 3D object surface of a point cloud. The PCN generates point clouds in 2 stages where the first stage is a lower resolution point cloud and the second stage is the final output. With the global feature vector mapping (GFV) to regulate the inferred partial point cloud, our proposed module outperforms other methods for all categories.

A number of qualitative results are shown in Figure 4. The ground truth of Stanford

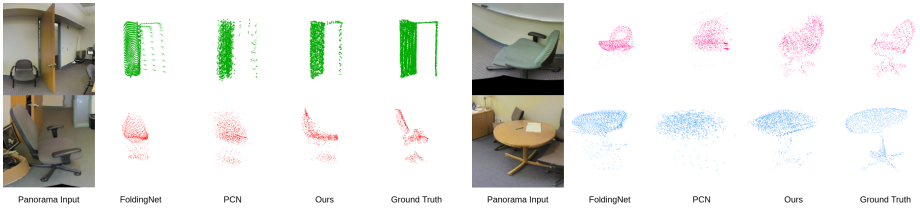


Figure 4: Qualitative comparison on 3D object point cloud completion.

2D-3D dataset, used as supervision, are real-scanned point clouds which may be incomplete. This makes the object point cloud completion more challenging. FoldingNet tends to generate continuous surfaces which could not preserve the gap between points (e.g. doors) and details (e.g. chair leg in the second example of the first row). PCN could abstract the global structures of the objects but the details are negatively affected by the inferred noisy partial point clouds (e.g. both chairs examples). Through global feature vector mapping, our proposed method generates plausible complete object point clouds.

4.3 3D Semantic Scene Reconstruction:

To impose the global constraints for the scene-level reconstruction and enforce consistency between the reconstructed scene point cloud and the panorama input, we combine all the modules and jointly train the pipeline end-to-end. As this is the first work, to the best of our knowledge, to reconstruct the full 3D semantic scene point cloud from a single panorama image, we report the quantitative results in Table 2 and illustrate the qualitative results of the reconstructed semantic scene point cloud in Figure 5. As shown in Table 2, with joint (end-to-end) training, the performance of the reconstruction is further improved. The examples in the first row in Figure 5 show the scenes with clear views and less occlusions. The examples in the second row exhibit the input panoramas with more clutter and large occlusions. The third row presents the results for more complicate scenes. All the results manifest that, with different complexities, our pipeline maintains visually appealing reconstructed semantic scene point cloud.

As an example, to show the generalization ability of our model, we apply our method on unseen data provided by the SUN360 dataset [26], where no ground truth depth or point clouds are available. As shown in Figure 6, although the SUN360 dataset has completely different indoor configurations, our approach still obtains plausible 3D reconstruction results.

5 Conclusion

In this paper we propose a pipeline to reconstruct the 3D indoor semantic scene point cloud from a single panorama image. The proposed pipeline joint learns to predict the 3D scene layout, complete the object shapes and reconstruct the full scene point cloud. By estimating the layout depth map, the method recovers the parameterized 3D scene layout. To generate the full object point cloud, the method completes the noisy partial point cloud via global feature vector mapping. The full pipeline is joint training end-to-end to ensure the consistency between the reconstructed scene point cloud and the panorama input. Experimental results demonstrate the the generality and suitability of the proposed pipeline.

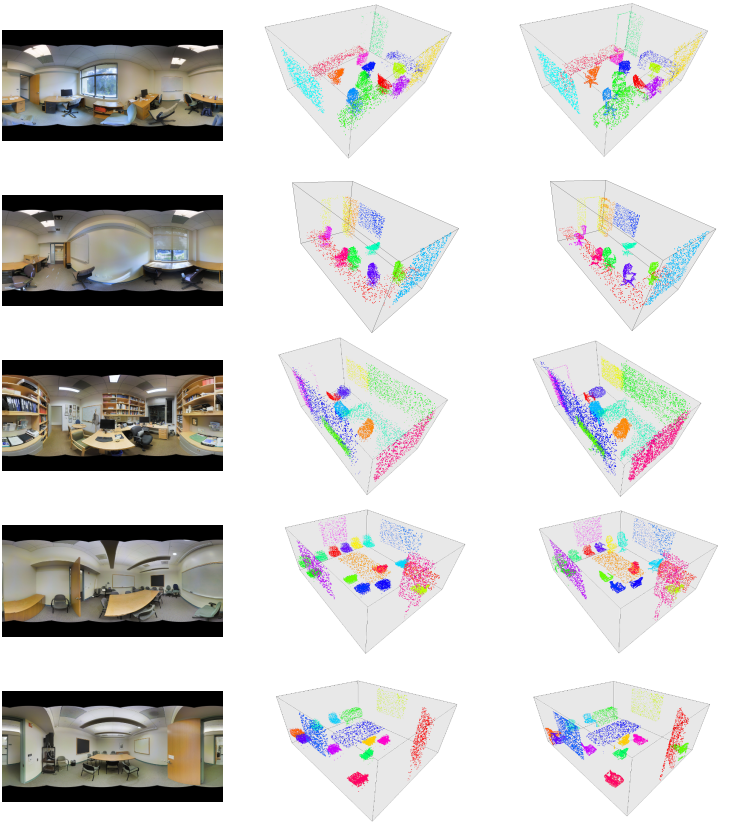


Figure 5: Qualitative results for 3D semantic scene reconstruction. Given a single panorama image, our method (end-to-end) reconstructs the 3D indoor semantic scene point cloud.

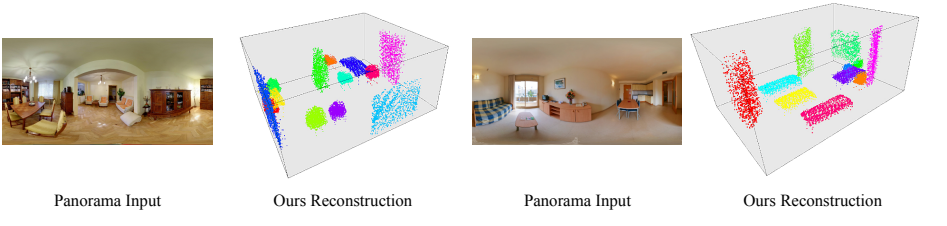


Figure 6: Qualitative results for 3D semantic scene reconstruction on unseen image samples from the SUN360 dataset.

Acknowledge

The first author (Wei Zeng) is funded by the China Scholarship Council (CSC) from the Ministry of Education of P.R. China.

References

[1] Waleed Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017.

[2] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

[3] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[4] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8648–8657, 2019.

[5] Erick Delage, Honglak Lee, and Andrew Y Ng. A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2418–2428. IEEE, 2006.

[6] Marc Eder, Pierre Moulon, and Li Guan. Pano popups: Indoor 3d reconstruction with a plane-aware network. In *2019 International Conference on 3D Vision (3DV)*, pages 76–84. IEEE, 2019.

[7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[8] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 6, 2017.

[9] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9785–9795, 2019.

[10] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 216–224, 2018.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017.

[12] Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering the spatial layout of cluttered rooms. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1849–1856. IEEE, 2009.

[13] 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.

[14] Siyuan Huang, Siyuan Qi, Yinxue Xiao, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation. In *Advances in Neural Information Processing Systems*, pages 207–218, 2018.

[15] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. Holistic 3d scene parsing and reconstruction from a single rgb image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 187–203, 2018.

[16] Hamid Izadinia, Qi Shan, and Steven M Seitz. Im2cad. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5134–5143, 2017.

[17] Li Jiang, Shaoshuai Shi, Xiaojuan Qi, and Jiaya Jia. Gal: Geometric adversarial loss for single-view 3d-object reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 802–816, 2018.

[18] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 International Conference on 3D Vision (3DV)*, pages 239–248. IEEE, 2016.

[19] Chen-Yu Lee, Vijay Badrinarayanan, Tomasz Malisiewicz, and Andrew Rabinovich. Roomnet: End-to-end room layout estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4865–4874, 2017.

[20] Chen Liu, Pushmeet Kohli, and Yasutaka Furukawa. Layered scene decomposition via the occlusion-crf. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–173, 2016.

[21] 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. *arXiv preprint arXiv:2002.12212*, 2020.

[22] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1047–1056, 2019.

[23] Lyne P Tchapmi, Vineet Kosaraju, Hamid Rezatofighi, Ian Reid, and Silvio Savarese. Topnet: Structural point cloud decoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 383–392, 2019.

[24] Shubham Tulsiani, Saurabh Gupta, David F Fouhey, Alexei A Efros, and Jitendra Malik. Factoring shape, pose, and layout from the 2d image of a 3d scene. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 302–310, 2018.

[25] Xiaolong Wang, David Fouhey, and Abhinav Gupta. Designing deep networks for surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 539–547, 2015.

[26] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2695–2702. IEEE, 2012.

[27] Shang-Ta Yang, Fu-En Wang, Chi-Han Peng, Peter Wonka, Min Sun, and Hung-Kuo Chu. Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3363–3372, 2019.

[28] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, 2018.

[29] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *2018 International Conference on 3D Vision (3DV)*, pages 728–737. IEEE, 2018.

[30] Wei Zeng, Sezer Karaoglu, and Theo Gevers. Inferring point clouds from single monocular images by depth intermediation. *arXiv preprint arXiv:1812.01402*, 2018.

[31] Jian Zhang, Chen Kan, Alexander G Schwing, and Raquel Urtasun. Estimating the 3d layout of indoor scenes and its clutter from depth sensors. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1273–1280, 2013.

[32] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 448–465, 2018.

[33] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. Layoutnet: Reconstructing the 3d room layout from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2051–2059, 2018.

[34] Chuhang Zou, Ruiqi Guo, Zhizhong Li, and Derek Hoiem. Complete 3d scene parsing from an rgbd image. *International Journal of Computer Vision*, 127(2):143–162, 2019.