Learning 3D Scene Semantics and Structure from a Single Depth Image

Bo Yang [1]*, Zihang Lai [1]*, Xiaoxuan Lu [1], Shuyu Lin [1],
Hongkai Wen [2], Andrew Markham [1], Niki Trigoni [1]
[1]University of Oxford
[2]University of Warwick
firstname.lastname@cs.ox.ac.uk

Abstract

In this paper, we aim to understand the semantics and 3D structure of a scene from a single depth image. Recent deep neural networks based methods aim to simultaneously learn object class labels and infer the 3D shape of a scene represented by a large voxel grid. However, individual objects within the scene are usually only represented by a few voxels leading to a loss of geometric detail. In addition, significant computational and memory resources are required to process the large scale voxel grid of a whole scene. To address this, we propose an efficient and holistic pipeline, ***3R-Depth****, to simultaneously learn the semantics and structure of a scene from a single depth image. Our key idea is to deeply fuse an efficient 3D shape estimator with existing recognition (*e.g.*, ResNets) and segmentation (*e.g.*, Mask R-CNN) techniques. Object level semantics and latent feature maps are extracted and then fed to a shape estimator to extract the 3D shape. Extensive experiments are conducted on large-scale synthesized indoor scene datasets, quantitatively and qualitatively demonstrating the merits and superior performance of 3R-Depth.*

1. Introduction

To enable an intelligent machine to navigate within and interact with the world, it is essential to understand the 3D structure and semantic meanings of its surrounding environment. With the widespread availability of off-the-shelf RGB-D sensors such as Microsoft Kinect and Google Tango, high-quality depth images of the environment can be acquired easily. A fundamental and open question is how to learn both the 3D geometry and semantic annotation of the entire scene.

Classic approaches address the above question in two separate pipelines. (1) Early methods in [6][17] only consider semantic segmentation for visible surfaces, ignoring the 3D geometry of the environment. (2) The approaches in [3][15] simply recover the 3D structure without extracting the semantic meanings. Basically, both pipelines rely on hand-crafted image feature extraction and matching. Furthermore, classic techniques for 3D structure recovery usually require multiple images scanned from different viewing angles, which is inefficient and even infeasible in many real-world scenarios.

With the advancement of deep neural nets, recent works such as SSCNet [19] and ScanComplete [2] are among the first work to simultaneously learn semantic labels and recover the 3D geometry for a scene.

SSCNet [19] takes a single depth view as the input, and predicts a completed voxel grid of the scene, with each voxel labeled with a semantic class. Although achieving impressive results, it has two drawbacks. (1) Since the input partial scene is represented by a $240 \times 144 \times 240$ voxel grid, and the output complete scene is a small $60 \times 36 \times 60$ voxel grid, many individual objects, *e.g.* chairs or tables, only consist of a few voxels. As a result, fine geometric details are unlikely to be recovered in the scene. (2) Since the majority of the scene tends to be unoccupied, most of the input and output voxel grid are '0'. Therefore, it is a waste of computation and memory to learn the whole sparse voxel grid.

Dai *et al*. introduce ScanComplete [2] to simultaneously complete 3D structure and infer per-voxel semantic labels for a large-scale scene. A sequence of depth images along a trajectory are firstly fused and voxelized, generating a large voxel grid, *e.g.*, $1480 \times 1230 \times 64$, to represent the partial scene. To curtail the high computation and memory costs incurred by the partial large voxel grid, ScanComplete uniformly samples subvolumes, *e.g.*, $32 \times 32 \times 32$ subgrids, from the large voxel grid, and then select meaningful subvolumes, *e.g.*, containing chairs, tables, to train a network. While this divide-and-conquer strategy is promising to deal with large-scale semantic scene completion, it is limited by the following reasons. (1) Since the input large voxel grid is manually fused from a sequence of depth images, a large amount of pre-processing work is required. In addition, the created voxel grid would inevitably consume large memory. (2) Before training the network, each sampled subvolume is manually filtered by checking whether it contains interesting information or not. (3) Similar to SSCNet, geometric details of individual objects are unlikely to be recovered, as each object may consist of few voxels.

To overcome the limitations of the prior art for se-

*equal contribution

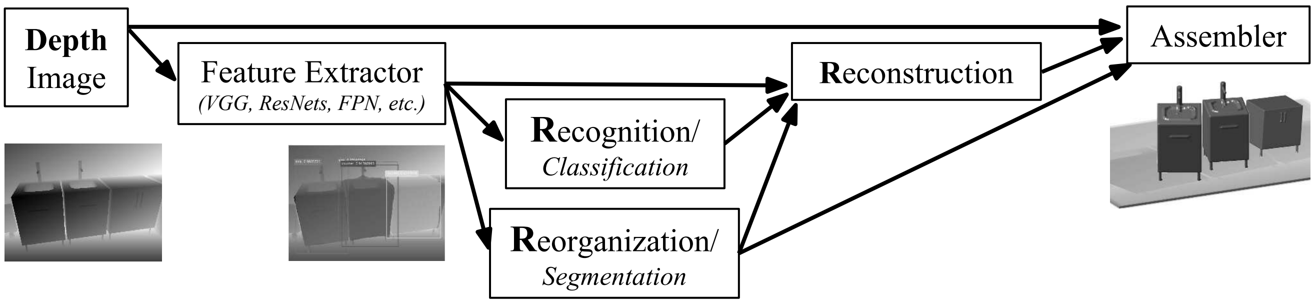


Figure 1: Flow chart of 3R-Depth.

mantic scene completion, we introduce a novel, efficient and holistic pipeline, named **3R-Depth**, to simultaneously learn the semantics and structure of a scene. As shown in Figure 1, 3R-Depth is an end-to-end trainable framework which simultaneously completes three tasks: **R**ecognition, **R**eorganization/Segmentation, and **R**econstruction [14] from a single **Depth** image. In particular, 3R-Depth takes a raw depth image as the input and then simultaneously segments and classifies each object in the scene, after which the segmented objects, which are also recognized, are reconstructed with full and fixed-size 3D shapes. In this way, scale issues are sidestepped.

By drawing on the powerful recognition techniques such as ResNets [9], the state of the art segmentation approaches such as Mask-RCNN [7], and the reconstruction methods such as 3D-RecGAN [24], our 3R-Depth is designed with the following features and advantages over existing approaches:

- 3R-Depth only takes a single raw depth image as the input, which does not require pre-processing and is also memory and computation efficient.
- 3R-Depth firstly learns accurate semantics from the depth image, and then estimates a fixed and high resolution 3D shape, *i.e.*, a $64 \times 64 \times 64$ voxel grid, for each object in the scene. Geometric details of individual objects can be well-recovered, irrespective of different sizes of objects in the raw depth input.
- The reconstructed 3D object shapes can be easily assembled to form a 3D scene, according to the corresponding depth values, thus recovering the 3D scene structure with semantic labels.

2. Related Work

(1) Recognition and Segmentation. Recent deep neural networks have led to a series of breakthroughs in image recognition, from the early AlexNet [11] and VGGNet [18], to the recent ResNet [8] and DenseNet [10]. These deep models greatly benefit many related vision tasks including object detection and segmentation. Both Fast/Faster R-CNN [4][16] and Fully Convolutional Network (FCN) [13] emerge as powerful baseline systems for object detection and semantic segmentation accordingly. Building on R-CNN [5] and Fast R-CNN [4], Faster R-CNN [16] applies attention mechanism with a Region Proposal Network (RPN) and then achieves leading performance in object detection. The most recent Mask R-CNN [7] proposes an RoIAlign operation together with an independent subnetwork to predict a binary mask for each RoI (Region of Interest) on Faster R-CNN, achieving the state of the art performance in instance segmentation task. So far, Mask R-CNN is able to simultaneously predict an accurate object class label and a pixel-level mask for each instance in the input RGB image. This powerful framework naturally becomes a fundamental component for large scene understanding including semantics extraction and 3D structure recovering.

(2) 3D Object Reconstruction. 3D object shapes can be recovered from either a single depth/RGB image or multiple images. Recent deep learning approaches achieve compelling results in single depth view reconstruction. 3D ShapeNets [22] is the first work that uses neural networks to infer 3D shapes from a single depth view. Firman *et al.* [3] propose a random decision forest to estimate unknown voxels. Varley *et al.* [21] propose a neural network to recover the complete 3D shape from a single depth view. 3D-EPN [1] firstly predicts a 32^3 object shape and then synthesizes a higher resolution shape from a large shape database. 3D-PRNN [25] infers a few number of shape primitives using RNNs. Yang *et al.* [24][23] incorporate adversarial learning for 3D shape estimation from a single depth view. Although existing work can achieve encouraging results, they only predict the shape of a single clean object without considering the semantics and 3D structure of a large scale scene.

(3) Semantic Scene Estimation. SSCNet [19] is among the first work to simultaneously predict semantics and 3D shapes using deep neural nets. The recent ScanComplete [2] takes divide-and-conquer strategy to complete large scale volumetric 3D scene. Tulsiani *et al.* [20] recently propose a network to predict both object shape together with orientation and the scene layout from an RGB image, but their network does not simultaneously predict object semantic labels. Besides, in their network, the object bounding box is separately extracted using existing algorithms.

3. 3R-Depth

Our 3R-Depth is a general framework towards 3D scene understanding. In this section, 3R-Depth is instantiated by integrating the state of the art techniques as plug-ins for each component. In this way, it is trivial to drop in alternative or improved components.

3.1. Feature Extraction

Given a single depth image as the input, this module aims to extract the latent features from the scene for subsequent recognition, segmentation and reconstruction. In our 3R-Depth, the ResNet50 architecture [9] is applied, although other existing architecture such as VGG and FPN are also applicable. In particular, the input depth image has the resolution of 640×480, which is the same as the depth images generated by Microsoft Kinect V2, while the output is a tensor of latent feature maps.

3.2. Classification and Segmentation

Given the learnt latent feature maps from section 3.1, the two tasks of classification and segmentation are simultaneously conducted using a separate subnetwork. A large number of candidate bounding boxes are firstly proposed, after which the size and location of each bounding box are optimized or filtered given the supervision of ground truth labels through the separate network. Instance-level segmentation is further learnt using another mask-branch network given ground truth supervision. In our 3R-Depth, the existing Mask-RCNN architecture [7] for instance-level classification and segmentation is applied, although the recent PANet [12] is also applicable. The output of the recognition and segmentation modules are a series of bounding boxes and binary masks, which are associated with class labels, on the top of input feature maps. Each of the labeled bounding box and mask corresponds to a specific object in the input depth image.

3.3. Reconstruction

Given the learnt latent feature maps from section 3.1, and the estimated bounding boxes and masks from section 3.2, this module aims to reconstruct the 3D shape for each segmented instance. We firstly multiply the latent feature maps by each of the binary masks, which results in the feature maps for each object, and then the learnt class label for each instance from section 3.2 is concatenated with the masked feature maps. The resulted object-level feature maps are directly fed into the reconstruction module to infer the corresponding 3D shapes. In our 3R-Depth, the 3D decoder of 3D-RecGAN [24] is leveraged to estimate 3D shapes with a resolution of $64 \times 64 \times 64$ voxel grids.

3.4. Assembler

Given the learnt 3D object shapes from section 3.3 and the estimated bounding boxes and masks from section 3.2, the corresponding depth values of each object are directly retrieved from the input depth image. The reconstructed object shapes are rescaled and assembled back to the scene, recovering the 3D scene semantics and structure accordingly.

4. Evaluation

4.1. Data Synthesize

To the best of our knowledge, there is no existing dataset that suits our 3R-Depth for evaluation. Therefore, we synthesize our own dataset based on the large-scale SUNCG indoor scene repository [19]. We render approximately 1.2 million views from randomly selected 25 thousand scenes with the provided toolbox. Similar to [19], we exclude bad viewpoints from rendered views. Specifically, the rendered images are filtered by the following criteria: (1) object area should be larger than 10% of the image, and (2) no larger than 20% of any object is occluded. Finally, 57 thousand valid views are generated in total, with a 8:2 split for training and testing. We select 15 common object categories as classes of interest.

4.2. Metrics

Segmentation performance is evaluated by mean average precision (AP), averaged for Intersection of Union, IoU $\in [0.5 : 0.05 : 0.95]$ (COCO's standard metric) [7]. 3D reconstruction is evaluated by the mean IoU between predicted 3D voxel grids and their ground truth. The IoU for an individual voxel grid is formally defined in [24].

4.3. Results

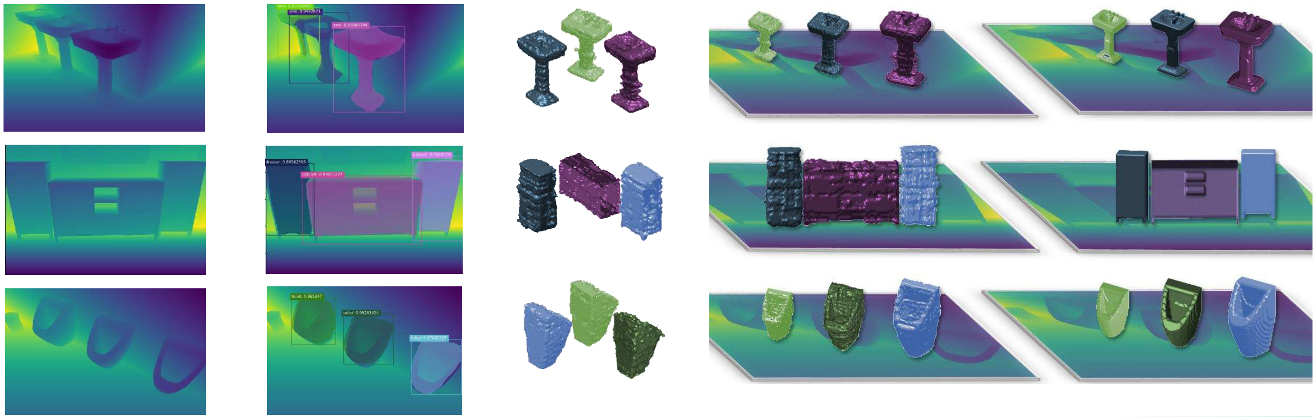
Table 1 shows per category mask prediction AP scores and reconstruction IoU scores, while Figure 2 shows the qualitative results. As to mask prediction, our integration of Mask-RCNN achieves superior accuracy on depth images, which is consistent with its outstanding performance for mask prediction on RGB images in its original paper. As to individual 3D object reconstruction, the 3D shape estimator also achieves satisfactory performance. However, we observe that the 3D shape estimator performs better on regular categories, *e.g.*, toilet, bathtub, than irregular ones.

5. Conclusion

In this paper, an efficient and holistic pipeline is proposed for 3D scene understanding from a single depth image. In this pipeline, instance level semantics are accurately extracted through the integration of ResNets and Mask-RCNN, while the high resolution 3D instance shape is inferred with an efficient 3D decoder which is deeply fused with recognition and segmentation nets. After all individual objects have been classified, segmented and reconstructed, they are assembled together according to the available depth values. Our approach is extensively evaluated on the large-

Table 1: Per-category mask AP and reconstruction IoU on SUNCG dataset.

class	toilet	chair	table	sofa	bed	shelves	night stand	lamp	desk	cabinet	sink	bathtub	bookshelf	dresser	counter	mean
maskAP	0.479	0.585	0.589	0.875	0.808	0.526	0.620	0.674	0.698	0.737	0.832	0.934	0.376	0.881	0.683	0.686
IoU(64^3)	0.811	0.580	0.294	0.678	0.677	0.668	0.665	0.478	0.550	0.708	0.867	0.860	0.579	0.673	0.744	0.656



input depth image

predicted class labels and masks for object instance

predicted 3D shapes for objects

assembled structure for the 3D scene

ground truth structure for the 3D scene

Figure 2: Qualitative results of our 3R-Depth.

scale SUNCG dataset and is able to recover high-quality 3D scene semantics and structures.

References

[1] A. Dai, C. R. Qi, and M. Nießner. Shape Completion using 3D-Encoder-Predictor CNNs and Shape Synthesis. *CVPR*, 2017. 2

[2] A. Dai, D. Ritchie, M. Bokeloh, S. Reed, J. Sturm, and M. Nießner. ScanComplete: Large-Scale Scene Completion and Semantic Segmentation for 3D Scans. *CVPR*, 2018. 1, 2

[3] M. Firman, O. M. Aodha, S. Julier, and G. J. Brostow. Structured Prediction of Unobserved Voxels From a Single Depth Image. *CVPR*, 2016. 1, 2

[4] R. Girshick. Fast R-CNN. *ICCV*, 2015. 2

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR*, 2014. 2

[6] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from RGB-D images. *CVPR*, 2013. 1

[7] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask R-CNN. *ICCV*, 2017. 2, 3

[8] K. He and J. Sun. Deep Residual Learning for Image Recognition. *CVPR*, 2015. 2

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *CVPR*, 2016. 2, 3

[10] G. Huang, Z. Liu, K. Q. Weinberger, and L. v. d. Maaten. Densely Connected Convolutional Networks. *CVPR*, 2017. 2

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012. 2

[12] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path Aggregation Network for Instance Segmentation. *CVPR*, 2018. 3

[13] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. *CVPR*, 2015. 2

[14] J. Malik, P. Arbelaez, J. Carreira, K. Fragkiadaki, R. Girshick, G. Gkioxari, S. Gupta, B. Hariharan, A. Kar, and S. Tulsiani. The three R's of computer vision: recognition, reconstruction, and reorganization. *Pattern Recognition Letters*, 72(1):4–14, 2016. 2

[15] D. T. Nguyen, B.-S. Hua, M.-K. Tran, Q.-H. Pham, and S.-K. Yeung. A Field Model for Repairing 3D Shapes. *CVPR*, 2016. 1

[16] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *NIPS*, 2015. 2

[17] X. Ren, L. Bo, and D. Fox. RGB-(D) scene labeling: features and algorithms. *CVPR*, 2012. 1

[18] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ICLR*, 2015. 2

[19] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic Scene Completion from a Single Depth Image. *CVPR*, 2017. 1, 2, 3

[20] S. Tulsiani, S. Gupta, D. Fouhey, A. A. Efros, and J. Malik. Factoring Shape, Pose, and Layout from the 2D Image of a 3D Scene. *CVPR*, 2018. 2

[21] J. Varley, C. Dechant, A. Richardson, J. Ruales, and P. Allen. Shape Completion Enabled Robotic Grasping. *IROS*, 2017. 2

[22] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3D ShapeNets: A Deep Representation for Volumetric Shapes. *CVPR*, 2015. 2

[23] B. Yang, S. Rosa, A. Markham, N. Trigoni, and H. Wen. 3D Object Dense Reconstruction from a Single Depth View. *arXiv*, 2018. 2

[24] B. Yang, H. Wen, S. Wang, R. Clark, A. Markham, and N. Trigoni. 3D Object Reconstruction from a Single Depth View with Adversarial Learning. *ICCV Workshops*, 2017. 2, 3

[25] C. Zou, E. Yumer, J. Yang, D. Ceylan, and D. Hoiem. 3D-PRNN: Generating Shape Primitives with Recurrent Neural Networks. *ICCV*, 2017. 2