

Complete 3D Scene Parsing from an RGBD Image

Chuhang Zou · Ruiqi Guo · Zhizhong Li · Derek Hoiem

Received: date / Accepted: date

Abstract One major goal of vision is to infer physical models of objects, surfaces, and their layout from sensors. In this paper, we aim to interpret indoor scenes from one RGBD image. Our representation encodes the layout of orthogonal walls and the extent of objects, modeled with CAD-like 3D shapes. We parse both the visible and occluded portions of the scene and all observable objects, producing a *complete* 3D parse. Such a scene interpretation is useful for robotics and visual reasoning, but difficult to produce due to the well-known challenge of segmentation, the high degree of occlusion, and the diversity of objects in indoor scenes. We take a data-driven approach, generating sets of potential object regions, matching to regions in training images, and transferring and aligning associated 3D models while encouraging fit to observations and spatial consistency. We use support inference to aid interpretation and propose a retrieval scheme that uses convolutional neural networks (CNNs) to classify regions and retrieve objects with similar shapes. We demonstrate the performance of our method on our newly annotated NYUD v2 dataset [39] with detailed 3D shapes.

Keywords Visual scene understanding · 3D parsing · single image reconstruction

1 Introduction

Recovering the layout and shape of surfaces and objects is a foundational problem in computer vision. Early approaches, such as reconstruction from line drawings [35], attempt to

C. Zou¹ · R. Guo² · Z. Li¹ · D. Hoiem¹

¹Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, USA
E-mail: {czou4, zli115, dhoiem}@illinois.edu

²Google Research
E-mail: guorq@google.com

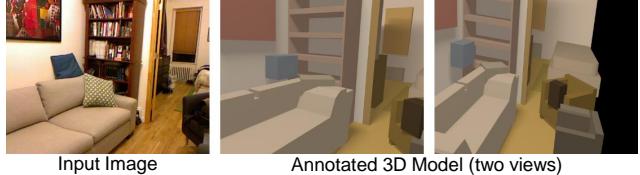


Fig. 1 Our goal is to recover a 3D model (right) from a single RGBD image (left), consisting of the position, orientation, and extent of layout surfaces and objects.

infer 3D object and surface models based on shading cues or boundary reasoning. But the complexity of natural scenes is too difficult to model with hand-coded processing and rules. More recent approaches to 3D reconstruction produce detailed literal geometric models, such as 3D point clouds or meshes, from multiple images [12], or coarse interpreted models, such as boxy objects within a box-shaped room [21], from one image.

This paper introduces an approach to recover complete 3D models of indoor objects and layout surfaces from an RGBD (RGB+Depth) image (Fig. 1). Recovering 3D models from images is highly challenging due to three ambiguities: the loss of depth information when points are projected onto an image; the loss of full 3D geometry due to occlusion; and the unknown separability of objects and surfaces. In this paper, we choose to work with RGBD images, rather than RGB images, so that we can focus on designing and inferring a useful representation, without immediately struggling with the added difficulty of interpreting geometry of visible surfaces. Even so, ambiguities due to occlusion and unknown separability of nearby objects make 3D reconstruction impossible in the abstract. If we see a book on a table, we observe only part of the book's and table's surfaces. How do we know their full shape, or even that the book is not just a bump on the table? Or how do we know that a sofa does not occlude a hole in the wall or floor? We don't. But we expect rooms to be enclosed, typically by orthogonal walls

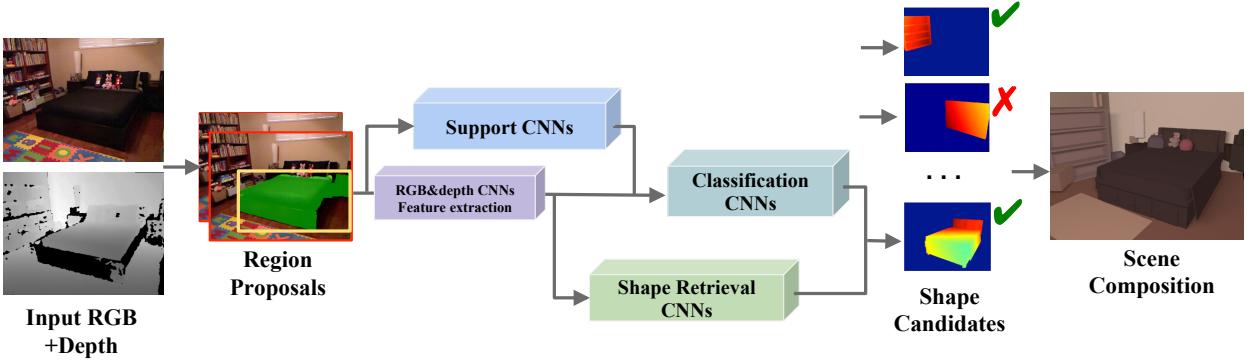


Fig. 2 Overview of our approach. Given an input RGB-D (left), we propose possible layouts and object regions. We predict each object proposal’s support height and class by our support and classification CNNs. We then retrieve a similar object shape, by using our shape retrieval CNNs to match the object region to the most similar region in the training set and transferring and aligning associated 3D models to the input depth image. The subset of proposed objects and layouts are then optimally selected based on consistency with observed depth, coverage, and constraints on occupied space. We show an example result (upper-right) and ground truth annotations (lower-right).

and horizontal surfaces, and we can guess the extent of the objects based on experience with similar objects. Our goal is to provide computers with this same interpretive ability.

Scene representation. Ideally, we want an expressive representation that supports robotics (e.g., where is it, what is it, how to move and interact with it), graphics (e.g., what would the scene look like with or without this object), and interpretation (e.g., what is the person trying to do). Importantly, we want to understand what is in the scene and what could be done, rather than only a labeling of visible surfaces. In this paper, we aim to infer a 3D geometric model that encodes the position and extent of layout surfaces, such as walls and floor, and objects such as tables, chairs, mugs, and televisions. In the long term, we hope to augment this geometric model with relations (e.g., this table supports that mug) and attributes (e.g., this is a cup that can serve as a container for small objects and can be grasped this way).

Our approach. We propose an approach to recover a 3D model of room layout and objects from an RGBD image. A major challenge is how to cope with the huge diversity of layouts and objects. Rather than restricting to a parametric model and a few detectable object classes, as in previous single-view reconstruction work, our models represent every layout surface and object with a 3D mesh that approximates the original depth image under projection. We take a data-driven approach that proposes a set of potential object regions, matches each region to a similar region in training images, and transfers and aligns the associated labeled 3D models while encouraging their agreement with observations. During the matching step, we use CNNs to retrieve objects of similar class and shape and further incorporate support estimation to aid interpretation. We hypothesize, and confirm in experiments, that support height information will help most for interpreting occluded objects because the full extent of an occluded object can be inferred from support height. The subset of proposed 3D objects and layouts that best represent the overall scene is then selected by our optimi-

mization method based on consistency with observed depth, coverage, and constraints on occupied space. The flexibility of our models is enabled through our approach (Fig. 2) to propose a large number of likely layout surfaces and objects and then compose a complete scene out of a subset of those proposals while accounting for occlusion, image appearance, depth, and layout consistency.

Detailed 3D labeling. Our approach requires a dataset with labeled 3D shape for region matching, shape retrieval, and evaluation. We make use of the NYUd v2 dataset [39] which consists of 1449 indoor scene RGBD images, with each image segmented and labeled with object instances and categories. Each segmented object also has a corresponding annotated 3D model, provided by Guo and Hoiem [15]. The 3D labeling provides ground truth 3D scene representation with layout surfaces as 3D planar regions, furniture as CAD exemplars, and other objects as coarser polygonal shapes. However, the polygonal shapes are too coarse to enable comparison of object shapes. Therefore, we extend the labeling by Guo and Hoiem with more detailed 3D annotations in the object scale. Annotations are labeled automatically and are adjusted manually as described in Sec. 3. We evaluate our method on our newly annotated groundtruth. We measure success according to accuracy of depth prediction of complete layout surfaces, voxel occupancy accuracy and semantic segmentation performance.

Our main **contributions** are:

1. We refine the NYUd v2 dataset with detailed 3D shape annotations for all the objects in each image. The labeling achieves a better representation of both ground truth object shape and whole scene depth.
2. We propose an approach to recover a 3D model of room layout and objects from an RGBD image, arguably producing the most detailed and complete 3D scene parses to date. We take a data-driven approach, generating sets of potential object regions, matching to regions in training images, and transferring and aligning associated 3D

models while encouraging fit to observations and overall consistency. We investigate impact of support estimation on scene parsing.

In contrast to multiview 3D reconstruction methods, our approach recovers complete models from a limited viewpoint, attempting to use priors and recognition to infer occluded geometry, and parses the scene into individual objects and surfaces, instead of points, voxels, or contiguous meshes. Thus, in some sense, we provide a bridge between the goals of interpretation from single-view and quantitative accuracy from multiview methods.

The overall technical contributions of this paper are substantial. Our approach tackles the challenging problem of recovering complete 3D model of indoor scene from single RGBD images. The proposed data-driven framework enables a detailed complete CAD model representation of 3D reconstruction (Sec. 3) instead of sparse voxels, point clouds or surface meshes. To enable the feasibility of the complete CAD representation, as introduced in Sec. 4, rather than performing retrieval on the entire image or superpixels, we use a more flexible region-to-region transfer to retrieve 3D shapes from a database. We then resolve conflicts in a final compositing process that accounts for fidelity to observed RGBD image. Extensive experiments demonstrate the success of our framework in handling complete 3D scene parsing quantitatively and qualitatively as shown in Sec. 5.2. Our layout estimation outperforms existing approaches. We apply ablation study to investigate the effectiveness of the design choices in each of the main step in our framework in Sec. 5.5. We demonstrate improvements due to better region proposals. Moreover, our results from automatic region proposals are nearly as good as the result from ground truth regions, demonstrating the effectiveness of our scene composition and, more generally, finding a scene hypothesis that is consistent in semantics and geometry. Our CNN-based shape retrieval approach performs quantitatively better than the CCA based retrieval method by Guo et al. [16]. Furthermore, we study in-depth to verify our hypothesis that estimating support height of objects can lead to better classification, especially for occluded objects.

This paper is an extension of our previous work [16] that predicts full 3D scene parsing from an RGBD image. Our main new contributions are the refinement of the NYUD v2 dataset with detailed 3D shape annotations, the use of CNNs to classify regions and retrieve object models with similar shapes to a region, and use of support inference to aid region classification. We also provide more detailed discussion and conduct more extensive experiments, demonstrating qualitative and quantitative improvement.

2 Related work

The most related work that recovers complete models from scene is proposed by Zhang et al. [54], where they predict 3D bounding boxes of the room and all major objects inside, together with their semantic categories from an RGB 360° full-view panorama. Different from Zhang et al.: we interpret whole-room 3D context with detailed 3D shapes and layout planes; our input is single image, which has limited field of view; and we make use of depth information, which eases the difficulty of interpreting geometry of visible surfaces. Other related topics are as follows.

Room layout is often modeled as a 3D box (cuboid) [21, 11, 38, 53, 54]. A box provides a good approximation to many rooms, and it has few parameters so that accurate estimation is possible from single RGB images [21, 11, 38, 6] or panoramas [54]. Even when a depth image is available, a box layout is often used (e.g.,[53]) due to the difficulty of parameterizing and fitting more complex models. Others, such as [7, 26, 32], estimate a more detailed Manhattan-structured layout of perpendicular walls based on visible floor-wall-ceiling boundaries. Methods also exist to recover axis-aligned, piecewise-planar models of interiors from large collections of images [12] or laser scans [49], benefiting from more complete scene information with fewer occlusions. We model the walls, floor, and ceiling of a room with a collection of axis-aligned planes with cutouts for windows, doors, and gaps, as shown in Figure 5. Thus, we achieve similar model complexity to other methods that require more complete 3D measurements.

Our use of **region transfer** is inspired by the SuperParsing method of Tighe and Lazebnik [44], which transfers pixel labels from training images based on retrieval. Similar ideas have also been used in other modalities: Karsch et al. [24] transfer depth, Guo and Hoiem [14] transfer polygons of background regions, Yamaguchi et al. [51] transfer clothing items. Exemplar-based 3D modeling is also employed by Satkin and Hebert [37] to transfer 3D geometry and object labels from entire scenes. Rather than retrieving based on entire images [27, 55] (which is too constraining) or superpixels (which may not correspond to entire objects), we take an approach of proposing a bag of object-like regions and resolving conflicts in a final compositing process that accounts for fidelity to observed depth points, coverage, and consistency. In this way, our approach also relates to work on segmentation [39, 17, 9] and parsing [34, 2] of RGBD images and generation of bags of object-like regions [3, 10, 33]. The difference between our method and the above region transfer work is that: (1) we transfer 3D object models, rather than semantic labels; (2) we make use of larger, possibly overlapping object proposals, versus smaller, disjoint superpixels; and (3) we learn a distance function to improve the retrieval.

3D objects are also often modeled as 3D boxes when estimating from RGB images [22, 50, 56, 30, 54] or RGB-D images [30]. However cuboids do not provide good shape approximations to chairs, tables, sofas, and many other common objects. Another approach is to fit CAD-like models to depicted objects. From RGB images, Lim et al. [28, 29] find furniture instances, and Aubry et al. [1] recognize chairs using HOG-based part detectors. In RGB-D images, Song and Xiao [40] search for chairs, beds, toilets, sofas, and tables by sliding 3D windows and enumerating all possible poses. Song et al. then extend the method to detect a larger variety of classes [41] with CNN-based approach and interprets semantic voxel representations from single depth map [43]. Gupta et al. [19] fit shape models of 6 classes with poses to improve object detection. Xiang et al. [48] align 3D shapes from 100 categories to 2D images, contributing a large scale database for 3D object recognition. Different from the above methods, our approach finds a detailed shape for any object and layout in the scene from a single RGBD image. Our approach is efficient and light-weight compared to ScanNet [5], which utilizes RGB-D video dataset to annotate instance-level semantic segmentations, since our model operates on single images instead of videos. Moreover, our detailed shape prediction can be utilized as an refining tool for existing dataset with 3D bounding box ground truth.

Support height estimation. Guo and Hoiem [15] localize the height and full extent of support surfaces from one RGBD image. In addition, object height priors have shown to be crucial geometric cues for better object detection in both 2D [23, 46] and 3D [30, 40, 19]. Deng et al. [8] apply height above ground to distinguish objects. We propose to use objects' support height to aid region class interpretation, which helps by distinguishing objects that appear at different height levels: e.g. chairs should be on the floor and alarm clocks should be on the table, and by inferring the full extent of an occluded object.

3 Detailed 3D annotations for indoor scenes

We conduct our experiments on the NYUDv2 dataset [39], which provides complete 3D labeling of both objects and layouts of 1449 RGB-D indoor images. Each object and layout has a 2D segment labeling and a corresponding annotated 3D model, provided by Guo and Hoiem [15]. The 3D annotations use 30 models to represent 6 categories of furniture that are most common and use extruded polygons to label all other objects. These models provide a good approximation of object extent but are often poor representations of object shape, as shown in Fig 3. Therefore, we extend the NYUD v2 dataset by replacing the extruded polygons with CAD models collected from ShapeNet [4] and ModelNet [47]. To align name-space between datasets, we manually map all model class labels to the 633-class 3D object

labels in NYUD v2 dataset. The shape retrieval and alignment process is performed automatically and then adjusted manually, as follows.

Coarse alignment. For each ground truth 2D region r_i in the NYUD v2 dataset, we retrieve model set $M = \{M_i\}$ from our collected models that have the same class label as r_i . We also include the region's original coarse 3D annotation by Guo and Hoiem [15] in the model set M , so that we can preserve the original labeling if no provided CAD models are better fit in depth. We initialize each M_i 's 3D location as the world coordinate center of the 3D annotation labeled by Guo and Hoiem. We resize M_i to have the same height as the 3D annotation.

Fine alignment. Next, we align each retrieved 3D object model M_i to fit the available depth map of the corresponding 2D region r_i in the target scene. The initial alignment is often not in the correct scale and orientation; e.g., a region of a left-facing chair often resembles a right-facing chair and needs to be rotated. We found that using Iterative Closest Point to solve for all parameters did not yield good results. Instead, we enumerate 16 equally-spaced orientations from -180 to 180 from top-down view and allows 2 minor scale revision ratio as $\{1.0, 0.9\}$. We perform ICP to solve for translation initialized using scale and rotation, and pick the best ICP result based on the following cost function:

$$\begin{aligned} \text{FittingCost}(M_i, T_i) = & C_{depth} \sum_{j \in r_i \cap s(M_i, T_i)} |\mathcal{I}_d(j) - \hat{d}(j; M_i, T_i)| \\ & + \sum_{j \in r_i \cap \neg s(M_i, T_i)} C_{missing} \\ & + C_{occ} \sum_{j \in \neg r_i \cap s(M_i, T_i)} \max(\hat{d}(j; M_i, T_i) - \mathcal{I}_d(j), 0) \end{aligned} \quad (1)$$

where T_i represents scale, rotation, and translation, $s(\cdot)$ is the mask of the rendered aligned object, $\mathcal{I}_d(j)$ denotes the observed depth at pixel j and $\hat{d}(j)$ means the rendered depth at j . The first term encourages depth similarity to the ground truth RGBD region. The second penalizes pixels in the proposed region that are not rendered. We loosely allow the rendered object depth in the scene to be further away than the sensor depth. This is because depth map only gives the depth of the closest object, not all objects, due to possible occlusion. We do not want to penalize predicting a larger depth than is observed in the scene. The third term penalizes pixels in the rendered model that are closer than the observed depth image (so the model does not stick out into space known to be empty).

Based on the fitting cost of Eq. 1, our algorithm picks the model M_i with the best translation, orientation, and scale T_i . The fitting scales T_i along different dimensions are the same. This helps simplify the search procedure for the alignment. We found that the large variety of CAD models in

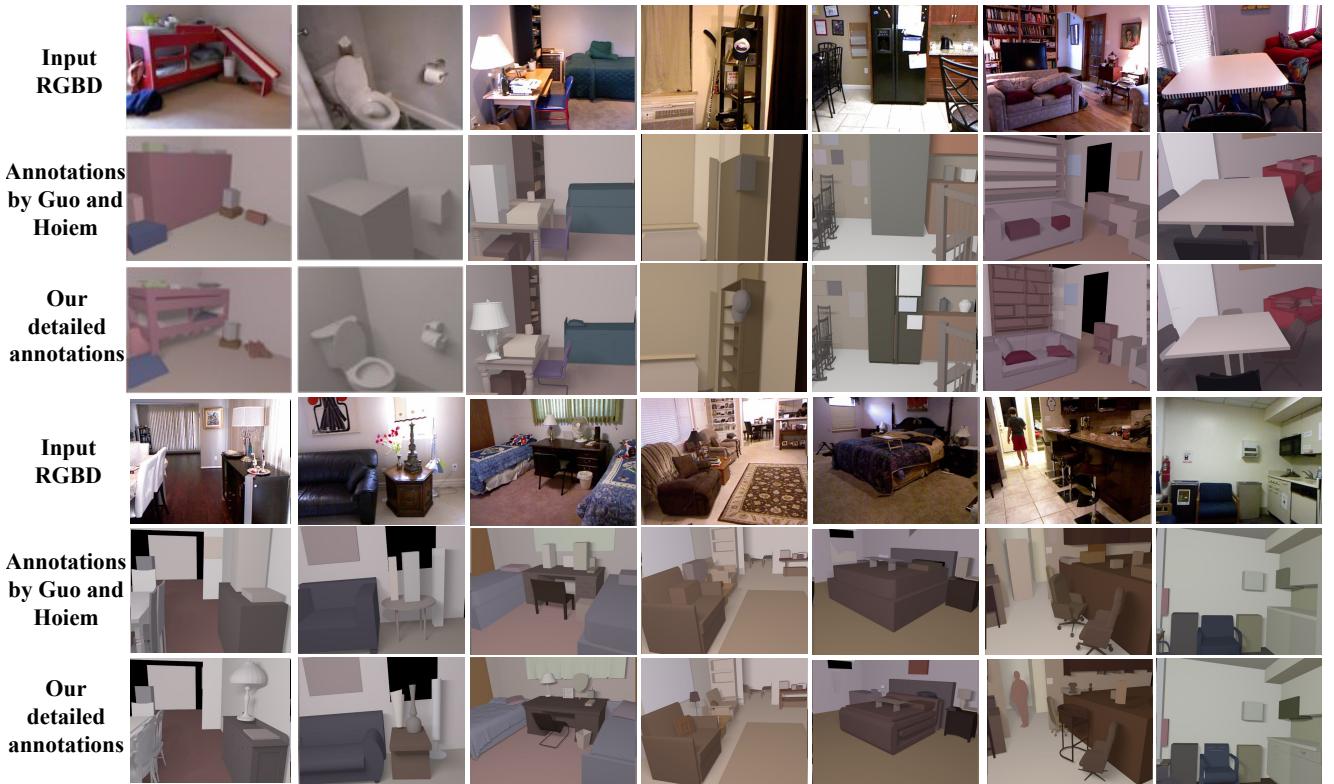


Fig. 3 Samples of our detail 3D annotation in NYUd v2 dataset. We show from the left column to the right with different levels (larger to smaller) of relative depth error improvements compared with Guo and Hoiem[15]. Our annotations are much detailed in object shape scale.

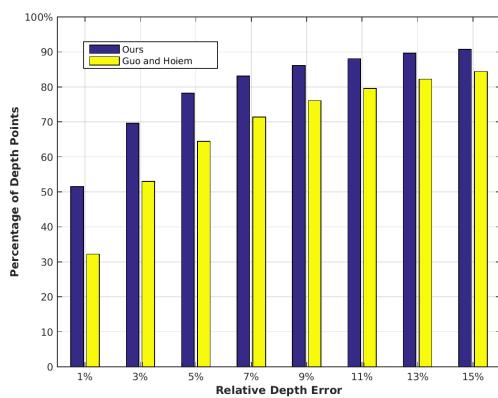


Fig. 4 Cumulative relative depth error of our detailed 3D annotations and the 3D annotations by Guo and Hoiem [15] in NYUd v2 dataset.

the ShapeNet dataset already provide enough shape candidates to select from, in regardless of the scales along the three dimensions of each shape. We set the term weights C_{depth} , $C_{missing}$, C_{occ} as 1.0, 0.9, 0.5 using grid search in the validation set. The search criteria is to find the set of term weights that minimize both the rendered depth difference to the ground truth depth and the rendered 2D segmentation difference to the ground truth 2D region annotation. For original 3D polygonal labeling, we fix M_i and use equation 1 to search for the best fitting scale and translation that minimize the cost. For efficiency, we first obtain the top 5 models

based on the fitting cost, each maximized only over the 16 initial orientations before ICP. For each of these models, we then solve for the best translation T_i for each scale and rotation based on Eq. 1 and finally select the aligned model with the lowest fitting cost.

Post-processing. Automatic fitting may fail due to high occlusion or missing depth values. We manually conduct a post-processing check and refine bad-fitting models. Using a GUI, an annotator checks the automatically produced shape for each region. If the result is not satisfactory, the annotator compares to other top model fits, and if none of those are good matches, then the fitting optimization based on Eq. 1 is applied to the original polygonal 3D labeling. This helps to ensure that our detailed shape annotations are a strict improvement over the original course annotations. In total, we fit 3792 different CAD models to the object regions and through manual checking observe failures in only 10% of cases for automatic fitting. These failures are manually corrected.

Validation. Figure 4 reports the cumulative relative error of the rendered depth of our detailed 3D annotations compared with ground truth depth in NYUd v2 dataset. The relative error r_D is computed as:

$$r_D = \frac{1}{|S_I|} \sum_{I \in S_I} \sum_{p \in I} \frac{|d_p - \hat{d}_p|}{d_p} \quad (2)$$

where $S_I = \{I_1, I_2, \dots, I_N\}$ is all the RGBD images in the dataset; p represents a pixel in each image I ; d_p is the ground truth depth of pixel p from sensor; and \hat{d}_p is the rendered depth of the 3D label annotation at pixel p . For comparison, we report the r_D of the 3D annotations by Guo and Hoiem [15]. Our annotations have more points with low relative depth error, and achieve a better modeling of depth for each image.

4 Approach

Given an RGBD image as input, we aim to find a set of layout and object models that fit RGB and depth observations and provide a likely explanation for the unobserved portion of the scene. A major challenge is how to cope with the huge diversity of layouts and objects. Existing approaches [41, 19, 40] focuses on recovering a few detectable objects and are not feasible as an expressive scene representation to support whole scene interpretation. Song et al. [43] propose a parametric model that employs a neural network to perform semantic completion from single depth images. However, a post-processing step is still needed for the approach by Song et al. to match candidate 3D shapes that fit both visible semantic information and scene regularity, e.g. no 3D overlap between shapes. Different from the above methods, we represent all layouts and objects in the scene as CAD models and propose a framework to parse both the visible and occluded portions of the layout and all observable objects, producing a complete 3D parse. Our resulting 3D parse is consistent with the observation and reasonable in 3D. We formulate this as:

$$\{\mathbf{M}, \theta\} = \arg \min (\text{AppearanceCost}(\mathcal{I}_{RGBD}, \mathbf{M}, \theta) + \text{DepthCost}(\mathcal{I}_D, \mathbf{M}, \theta) + \text{ModelCost}(\mathbf{M}, \theta)). \quad (3)$$

\mathcal{I}_{RGBD} is the RGB-D image; \mathcal{I}_D is the depth image alone; \mathbf{M} is a set of candidate 3D layout surfaces and object models; and θ is the set of parameters for each surface/object model, including translation, rotation, scaling, and whether each candidate model is included. AppearanceCost encourages that object models should match underlying region appearance, rendered objects should cover pixels that look like objects (versus layout surfaces), and different objects should have evidence from different pixels. DepthCost encourages similarity between the rendered scene and observed depth image. ModelCost penalizes intersection of 3D object models.

We propose to tackle this complex optimization problem in stages (Figure 2): (1) propose candidate layout surfaces and objects; (2) retrieve 3D models of the proposed candidate and improve the transformation of each surface/object model to the depth image; (3) choose a subset of models that best explains the scene. Layout elements (wall, floor,

and ceiling surfaces) are proposed by scanning for planes that match observed depth points and pixel labels and then finding boundaries and holes (Sec. 4.2). Parsing 3D objects is particularly difficult. We propose an exemplar-based approach, matching regions in the input RGBD image to regions in the training set, and transferring and aligning corresponding 3D models (Sec. 4.3). We then choose a subset of objects and layout surfaces that minimizes the depth, appearance, and model costs using a specialized search (Sec. 4.4). Despite the challenges of matching an optimization, we experimentally find that our approach produces results from automatic regions that are nearly as good as those produced from ground truth regions, even though using ground truth regions greatly simplifies the matching and selection process.

4.1 RGBD image pre-processing

We perform experiments on our newly re-annotated NYU v2 dataset, using the standard training/test split. The dataset consists of 1449 RGB-D images, with each image segmented and labeled into object instances and categories. Each segmented object has a corresponding annotated detailed 3D model labeled through the process in Sec. 3. Given a test image, we use the code from [39] to obtain an oversegmentation with boundary likelihoods and the probability that each pixel j corresponds to an object $P_{object}(j; \mathcal{I}_{RGBD})$ (instead of wall, floor, or ceiling). We use the code from [44] to find the major orthogonal scene orientations, which is used to align the scene and obtain height value for each pixel.

4.2 Layout proposals

We generate layout proposals representing the full extent of possible layout surfaces, such as walls, floor, and ceiling. The layouts of these surfaces can be complex. For example, the scene in Figure 5 has several ceiling-to-floor walls, one with a cutout for shelving, and a thin strip of wall below the ceiling on the left. The cabinet on the left could easily be mistaken for a wall surface. Our approach is to propose a set of planes in the dominant room directions. These planes are labeled into “floor”, “ceiling”, “left wall”, “right wall”, or “front wall” based on their position and orientation. Then, the extent of the surface is determined based on observed depth points.

To find layout plane proposals, we aggregate appearance, depth, and location features computed from [39] and train a separate linear SVM classifier for each of the five layout categories to detect planes. Appearance feature is the color histogram under YCbCr space. We define

$$p(p_i; P) = \mathcal{N}(dist(p_i, P), \sigma_p) \mathcal{N}(dist(n_i, P), \sigma_n) \quad (4)$$

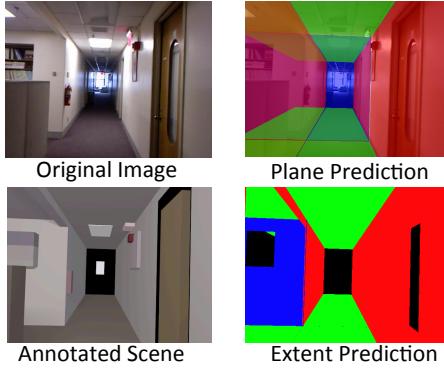


Fig. 5 Layout proposal. We detect axis-aligned horizontal and vertical planes in the input depth image and estimate the extent of each surface. Human-annotated layout is on the lower-left.

as the probability that a point with position p_i and normal n_i belongs to plane P , where the distances are point-to-plane and angular distance, and $\sigma_p = 0.025$ and $\sigma_n = 0.0799$ are based on Kinect measurement error. Each pixel also has a probability of belonging to floor, wall, ceiling, or object, using code from [39]. The plane detection features are $f_1 = \sum_i p(p_i; P)$; $f_2 \dots f_5$, the sum in f_1 weighted by each of the four label probabilities; f_6 , the number of points behind the plane by at least 3% of plane depth; $f_7 \dots f_{11} = (f_1 \dots f_5)/f_6$; and f_{12} , a plane position prior estimated from training data. Non-maximum suppression is used to remove weaker detections within 0.15m of a stronger detected plane. Remaining planes are kept as proposals if their classification score is above a threshold, typically resulting in 4-8 layout surfaces.

The maximum extent of a proposed plane is determined by its intersection with other planes: for example, the floor cuts off wall planes at the base. Furthermore, hole cut-outs (e.g. fireplace in a wall) are made by finding connected components of pixels with depth 5% behind the plane, projecting those points onto the plane, fitting a bounding box to them, and removing the bounding box from the plane surface. Intuitively, observed points behind the plane are evidence of an opening, and the use of a bounding box enforces a more regular surface that hypotheses openings behind partly occluded areas. Note that the fitted bounding box for hole cut-out on walls could be small compared to the ground truth due to occlusion. We observe that our simple approach produces excellent results in nearly all cases. Our method produces lower pixel labeling error for layout estimation than other approaches (see experiments in Sec. 5.2).

4.3 Object candidates

To produce object candidate regions, we use the method for RGBD images by Gupta et al. [18] and extract top ranked 2000 region proposals for each image. Our experiments show that this region retrieval is more effective than the method based on Prims algorithm [33] used in our previous work [16].

Likely object categories and 3D shapes are then assigned to each candidate region.

4.3.1 CNN-based Shape Retrieval

We train and use CNN networks to predict the object category and support height of each region, as shown in Fig. 6 and Fig. 7. The support height is used as a feature for the object classification. We also train and use a CNN with a Siamese network design to find the most similar 3D shape of a training object, based on region and depth features.

Support height prediction. We predict support height for each object with the aim of better predicting class and position. As shown in Fig. 6, our support height prediction network jointly predicts the candidate support height probability and the support type: whether the object is supported from below or behind. We first find candidate support heights using the method of Guo and Hoiem [15] and use the network to estimate the probability of each candidate being the correct height. The support type prediction and the candidate support height probability prediction share the same convolution structure and the first two fully connected layers. Our network input consists two features: crops of 1) the depth maps and height maps of the region proposal and 2) the depth maps and height maps of the region that extends from the bottom of the region proposal to the estimated floor position. The former feature helps infer support type through its shape and the latter helps infer the support height through the vertical change in texture and depth. We found that having those two inputs together performs better for both support height and support type predictions. To create the input feature vector, we subtract the candidate support height from the height cropped images, re-size all four crops to 32×32 patches, and concatenate channel-wise.

In the test set, we identify the closest candidate support height with 92% accuracy, with an average distance error of 0.18m. As a feature for classification, we use the support height relative to the camera height, which leads to slightly better performance than using support height relative to the estimated ground. This is because dataset images are taken from consistent heights but estimated ground height may be mistaken.

Categorization. Our classification network gets input of the region proposal's support height and type, along with CNN features from both RGB and depth. The network predicts the probability for each class as shown in Fig. 7. To model the various classes of shapes in indoor scene, we classify the regions into the 78 most common classes, which have at least 10 training samples. Less common objects are classified as “other prop”, “other furniture” and “other structure” based on rules by Silberman et al. [39]. In addition, we identify a region proposal that is not representative for an object shape (e.g. a piece of chair leg region when the

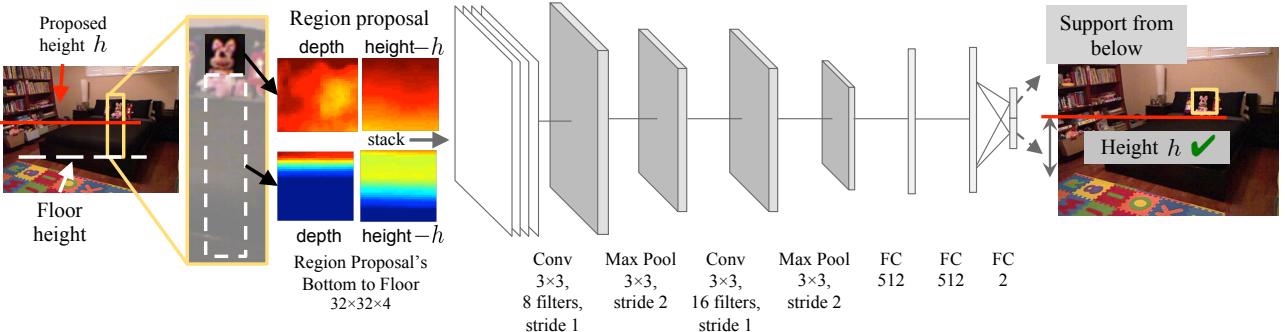


Fig. 6 The CNN for predicting a candidate object’s support height. We perform ReLU between the convolutional layer and the max pooling layer. Local response normalization is performed before the first fully connected (FC) layer. We add dropout with 0.5 before each FC layer during training.

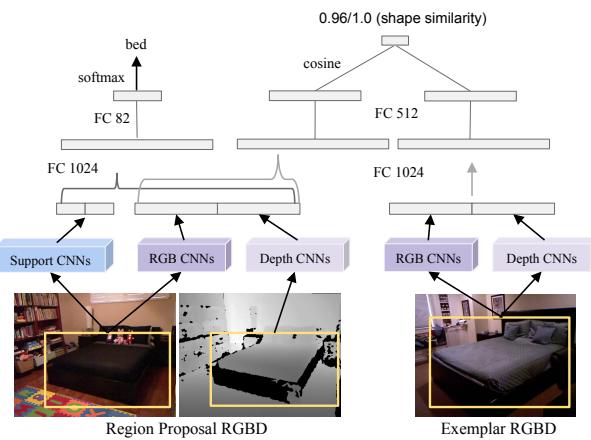


Fig. 7 The CNNs for region classification (left) and similar shape retrieval (right). We perform ReLU and dropout with 0.5 after the first FC layer for both of the networks during training.

whole chair region is visible) as a “bad region” class. This leads to our $78 + 3 + 1 = 82$ -class classifier. The input support height and type are directly predicted by our support height prediction network. To create our classification features, we copy the two predicted support values 100 times each (a useful trick to reduce sensitivity to local optima for important low-dimensional features) and concatenate them to the region proposal’s RGB and HHA features from Gupta et al. [18] in both the 2D bounding box and masked region as in [20]. Experiments show that using the predicted support type and the support height improves the classification accuracy by about 1%, with larger improvement for occluded objects.

Shape candidate retrieval. Using a Siamese network (Fig. 7), we learn a region-to-region similarity measure that predicts 3D shape similarity of the corresponding objects. The network embeds the RGB and HHA features used in our classification network into a space where cosine distance correlates to shape similarity, as in [52]. In training, we use surface-to-surface distance [36], the normalized point cloud distance between the densely sampled points on each of the 3D meshes, as the ground truth similarity. The surface-to-

surface distance is calculated from two centered complete 3D shapes complete 3D meshes, and thus no self-occlusion shall be considered during ground-truth computation. We train the network to penalize errors in shape similarity orderings. Each region pair’s shape similarity score is compared with the next pair’s among the randomly sampled batch in the current epoch and penalized only if the ordering disagrees with the ground truth similarity. We attempted sharing embedding weights with the classification network but observed a 1% drop in classification performance. We also found predicted class probability to be unhelpful for predicting shape similarity.

Candidate region selection. We apply the above retrieval scheme to each of the 2000 region proposals in each image, obtaining shape similarity rank compared with all the training samples and 81-object class and non-object class probability for each region proposal. In order to reduce the number of retrieved candidates before the scene interpretation in Sec. 4.4, we first reduce the number of region proposals using non-maximal suppression based on non-object class probability and threshold on the non-object class probability. We set the threshold to obtain 190 region proposals for each image, on average. We select the two most probable classes for each remaining region proposal and select five most similar shapes for each class, leading to ten shape candidates for each region proposal.

Then, we further refine these ten retrieved shapes. We align each shape candidate to the target scene by translating the model using the offset between the depth point mass centers of the region proposal and the retrieved region. We then perform ICP and use a similar method to that described in Section 3 to speed up the process. We use a grid of initial values for rotation and scale and pick the best one for each shape based on the fitting energy in Eq. 1. We set the term weight: $C_{missing}$ as 0.6, C_{depth} as 1.0, C_{occ} as 0.9 based on a grid search on the validation set. The overall selection costs on average 3.7 seconds for each object. We tried using the estimated support height for each region proposal for aligning the related 3D shape models but observed a worse

performance in the scene composition result. This is because a relatively small error in object’s support height estimation can cause a larger error in fitting.

Finally, we select the two most promising shape candidates based on the following energy function,

$$E_l(m_i) = w_f E_{fitting}(m_i, t_i) + w_c \log P(c_i|r_i) + w_b \log P(b_i|r_i) \quad (5)$$

$E_{fitting}$ is the fitting energy defined in Eq. 1 that we used for alignment. $P(c_i|r_i)$ and $P(b_i|r_i)$ are the softmax class probability and the non-object class probability output by our classification network for the region proposal r_i . We normalize $P(c_i|r_i)$ to sum to 1, in order not to penalize the non-object class twice in the energy function. We set the term weights $w_f = 1.0$, $w_c = -1500$, $w_b = 1300$ using a grid search. Note that $E_{fitting}$ is on the scale of the number of pixels in the region.

Training details. We first find meta-parameters on the validation set after training classifiers on the training set. Then, we retrain on both training and validation in order to report results on the test set. We train our networks with the region proposals that have the highest (and at least 0.5) 2D intersection-over-union (IoU) with each ground truth region in the train set. We train the support height prediction network with the ground truth support type for the region proposal and set the ground truth support height as the closest support height candidate that is within 0.15 meters from the related 3D annotation’s bottom. For training regions that are supported from behind, we do not penalize the support height estimation, since our support height candidates are for vertical support. For training the classification network, we also include the non-object class region proposals that have < 0.3 IoU with the ground truth regions. We randomly sample the same number of the non-object regions as the total number of the object regions during training. To avoid unbalanced weights for different classes, we sample from the dataset the same number of training regions for each class in each epoch. When training the shape similarity network, we use the ground truth detailed CAD annotation obtained from Sec. 3. When conducting the shape candidate retrieval, the pool of 3D CAD models is the detailed annotation from the training split of the NYUD v2 dataset. We translate each 3D model to origin and re-size to $200 \times 200 \times 200$ -voxel cuboid before computing the surface-to-surface distance. We use ADAM [25] to train each network with the hyper-parameter of $\beta_1 = 0.9$, $\beta_2 = 0.999$. The learning rate for each network is: support height prediction 0.0008, classification 0.003 and Siamese network 0.0001.

4.4 Scene composition

Finally, given a set of candidate objects from Sec. 4.3 and layout plane candidates from Sec. 4.2, we need to choose a subset that closely reproduces the original depth image when rendered, adhere to pixel predictions of object occupancy, and correspond to minimally overlapping 2D regions and 3D aligned models. Instead of fitting each model to the depth map separately, we integrate the object and layout proposals together, optimize on proposals and compose a complete scene. The models $M = \{M_i\}$ and their alignment $T = \{T_i\}$ are fixed. The remaining parameters are \mathbf{y} with $y_i \in \{0, 1\}$ indicating whether each layout or object model is part of the scene. We choose \mathbf{y} to minimize Equation 6:

$$\begin{aligned} & \text{selectionCost}(\mathbf{y}) = \\ & \sum_j \text{clip}\left(\left|\log_2 \frac{\hat{d}(j; \mathbf{M}, \mathbf{T}, \mathbf{y})}{\mathcal{I}_d(j)}\right| - \log_2(1.03), [0, 1]\right) + \\ & \sum_j |\text{isObject}(j; \mathbf{M}, \mathbf{T}, \mathbf{y}) - P_{\text{object}}(j; \mathcal{I}_{RGBD})| + \\ & \sum_j \max\left(\sum_i y_i r_i(j) - 1, 0\right) + \\ & \sum_{i,k>i} y_i y_k \text{overlap3d}(M_i, T_i, M_k, T_k). \end{aligned} \quad (6)$$

Here j represents the rendered pixel for each candidate shape. \hat{d} renders object models selected by \mathbf{y} . The first term minimizes error between rendered model and observed depth. We use log space so that the error matters more for close object. We subtract $\log_2(1.03)$ and clip at 0 to 1 because errors less than 3% of depth are within noise range, and we want to improve only reduction of depth if the predicted and observed depth is within a factor of 2. The second term encourages rendered object and layout pixels to match the probability of object map (P_{object}) of each image based on [39]. $r_i(j)$ is whether pixel j belongs to the model M_i rendered in 2D. The function $\text{isObject}(j) = 1$ identifies that a pixel corresponds to an object, rather than a layout model. The third term penalizes choosing object models that correspond to overlapping region proposals (each selected object should have evidence from different pixels). The second and third term, combined with region retrieval, account for the appearance cost between the scene model and observations. The fourth term penalizes 3D overlap of pairs of aligned models, to encourage scene consistency. To improve efficiency of this terms’ computation, we approximate the object occupancy by rendering the closest and furthest points of the object at each pixel. We assume that all pixels between are occupied, and penalize using 3D overlap – the length of the range of overlapping occupied depth, summed over each pixel.

The depth rendering and consistency terms of Equation 6 involve high-order dependencies, leading to a hard binary

problem. We initialize $\mathbf{y} = \mathbf{0}$ and estimate the solution in three steps. First, we perform greedy search to minimize depth error (minimize `selectionCost` while weighting first term by a factor of 10) by iteratively adding the next candidate that maximizes `selectionCost` gain. Then, this solution is used to initialize a hill-climbing search on `selectionCost` (experiments indicated insignificant benefit to weighting terms). The hill-climbing search iteratively adds ($y_i = 1$) or removes ($y_i = 0$), where i indicates the candidate model that results in the greatest cost reduction, until no change yields further improvement. Finally, for all layout proposals and a subset of object proposals that are not yet selected, our algorithm tries adding the proposed model and removing all models whose renderings overlap and keeps the change if the cost is reduced. In experiments, we found this search procedure to outperform a variety of other attempted methods including general integer programming algorithms and relaxations.

5 Experiments

The Overall performance is evaluated on the final complete 3D scene prediction for objects and layouts using the ground truth detailed 3D annotation (in Sec. 5.2). We compare our current approach to our previous approach (Guo et al. [16]). For some measures, we also compute the performance of the detailed 3D ground truth annotations as an upper bound. We investigate the effectiveness of the design choices in each of the main steps: region proposal, classification and shape retrieval, and scene composition. We also investigate the effects of incorporating support inference. To avoid overfitting to the NYUv2 dataset, we train our approach on the standard training split, tune all the parameters on the validation split and report results on the test split.

Improving on Guo et al. [16]. Our current approach improves on our previous work (“**Guo et al.**”) in two aspects. First, we use region proposals from Gupta et al. [18] (“MCG”) instead of the region proposals generated by randomized Prims algorithm [33]. In Guo et al.’s method, region proposals are generated by first starting with an oversegmentation and boundary strengths by Silberman et al. [39]. A neighborhood graph is then created, with superpixels as nodes and boundary strength as the weight connecting adjacent superpixels. Finally, the randomized Prims algorithm is applied on the graph to obtain a set of region proposals. The seed region of the Prims algorithm is sampled according to the objectness of the segment, so that the segments that are confident layout are never sampled as seeds. Size constraints and merging threshold are used to produce a more diverse set of segmentations. Regions that are near-duplicates of other regions are suppressed to make the set of proposals more compact. For each image, 100 candidate

region proposals are generated. Second, our method considers both object category and shape similarity for retrieval, while Guo et al. use only object category as a proxy for object similarity. Guo et al. apply canonical-correlation analysis (CCA) to find embeddings of visual features that improve retrieval [13]. CCA finds pairs of linear projections of the two views $\alpha^T X$ and $\beta^T Z$ that are maximally correlated. Here $X = \{x_i\}$ are the feature vectors of the regions. In experiments, Guo et al. use the combination of 3D features and RGBD-SIFT, which were found to provide better performance than pre-trained CNN features. $Z = \{z_i\}$ are the label matrix, which contains the one-hot indicator vectors of the object category labels of the corresponding region. The similarity weight matrix is then computed as $W = \alpha\alpha^T$, which is used to parametrize the distance metric:

$$\text{dist}_W(x_i, x_j) = \sqrt{(x_i - x_j)^T W (x_i - x_j)} \quad (7)$$

In the experiments for Guo et al., the top 3 object models nearest to each region proposal according to dist_W are retrieved.

Detailed ground truth labeling as an upper bound.

We include the evaluation of our detailed ground truth annotation for the pixel-wise measures in Sec. 3, denoted by “**3D ground truth**”. Note that even the “3D ground truth” sometimes does not achieve high accuracy, because rendered models may not follow image boundaries and some small objects are not modeled in annotations.

Comparison with Deep Sliding Shapes [41] (“DSS”).

We compare our approach with DSS, the state-of-the-art amodal 3D object detector in RGBD images. Given single RGBD image, DSS predicts multiple (around 300 after NMS) overlapping tight 3D bounding boxes and a confidence score for each object class. Each object class is predicted independently, therefore there is no constraint on 3D overlap between them. The evaluation criteria is the mean average precision (mAP) for each class. Different from DSS, our approach aims at producing an exact reconstruction of scene composed by all coherent objects and layouts with detailed CAD representation, while penalizing on object 3D overlap to ensure consistency. We evaluate on final scene composition performance to match observed RGB and depth. Though our approach have different goal, metrics and optimization strategy compared with DSS, we can still convert the results to each other’s format and compare.

Comparison with Semantic Scene Completion [43] (“SSCNet”). We compare our approach with the scene-level approach SSCNet, which is the state-of-the-art semantic scene completion method from single depth image. SSCNet predicts occupancy in a voxel space of $60 \times 36 \times 60$ within the view frustum of the scene. Each predicted voxel is assigned to either one of the 3 layout labels (wall, ceiling, floor) or one of the 8 common furniture classes (e.g. bed,

chair, table). Both SSCNet and our approach predict semantic labeling of the observed scene. Different from SSCNet, our framework utilizes RGB information in addition to observed depth. Our approach finds a mesh-based representation for any object and layout in the scene, compared with SSCNet which produces a limited number of classes under the voxel-based representation and is not flexible in rendering the scene with a higher resolution. For comparison, we align and voxelize our predicted 3D scene model to the same configurations as SSCNet and evaluate on their metric.

5.1 Performance measures

We evaluate our 3D scene completion approach on our newly annotated NYUd v2 dataset. We use both 2D and 3D quantitative measures to evaluate different aspects of our solutions: (1) instance segmentation and semantic segmentation performance induced by rendering; (2) label/depth accuracy of predicted layout surfaces; (3) voxel accuracy of occupied space and freespace. Among these, the depth accuracy of predicted layout surfaces and voxel accuracy of object occupancy are the most direct evaluations of our 3D model accuracy. The instance segmentation and semantic segmentation indicates accuracy of object localization and classification. Where possible, we compare on these measures to prior work and sensible baselines.

For the comparison with DSS, we directly use DSS's 3D bounding box prediction on the 654 test images of the NYUd v2 dataset. We use our detailed annotated ground truth to compare on our metrics. Since DSS has no layout prediction, we only compare on the 19 classes that DSS concerns. We use the following three metrics for parsing objects: 3D voxel object occupancy, 2D semantic segmentation and instance segmentation. Note that DSS produces multiple overlapping predictions with confidence. We first filter out detections with score lower than 0.5 in each image. We then apply NMS with an IoU threshold of 0 for each class separately to obtain non-overlapping top ranked 3D boxes. DSS has no 2D segmentation mask for each detection, we thus project the 3D box on the 2D image to get the predicted object segmentation and depth. We ignore the layouts and other object classes in both our 3D/2D predictions and the ground truth. On the other hand, to compare on DSS's amodal 3D object detection metric [40], we extract a tight 3D bounding box around each of our predicted detailed shape and set the confidence of each box to 1. We also compare with the result of Sliding Shapes by Song et al. [40]. Sliding Shapes predicts detailed shapes, while the DSS approach only predicts 3D bounding boxes. Same as [40] and [41], we evaluate on the ground truth of five main furniture on the test set which is the intersection of NYUd v2 and the Sliding Shapes test set.

For the comparison with SSCNet, we evaluate on SSCNet's metric given the input of the kinect depth map and RGB observations on the test split of the NYUd v2 dataset. Since our predicted complete scene model is aligned with the camera coordinate, we first align our predictions to the world coordinate given groundtruth floor plane. We then voxelize the scene into $60 \times 36 \times 60$ voxel grids to obtain the same resolution as in SSCNet and then compare. We map our 84-class semantic labeling to the 11-class used in SSCNet. We then evaluate on two metrics [43]: (1) scene completion with precision, recall and voxel-level IoU; (2) semantic scene completion with voxel-level IoU. Though our method is not optimized for voxel-based evaluation, we can still compare and evaluate to validate our approach on the scene-level 3D parsing performance.

5.2 Evaluation of scene composition

Qualitative results. We show representative examples of predicted 3D scenes in Figures 8 and 9 of our approach, 3D ground truth, and Guo et al.'s method. Compared with the ground truth annotation, our method produces a similar layout parsing and a reasonable prediction and localization of main furniture in the scene. Compared with our previous approach of Guo et al., our method performs better, which results from the better region classification and shape estimation capability that will be analyzed in the ablation study Sec. 5.5.

To demonstrate the generalization capability to other dataset, we show qualitative results on the SUN-RGBD dataset [42] in Figure 10. Since SUN-RGBD dataset does not have detailed shape annotations to train on, we directly use our model trained on NYUd v2 dataset. Our method is able to obtain the accurate 3D parse of layouts and the main furniture like beds, shelves and desks. Errors mainly come from region proposal classification failure and missing small objects from the region proposal step, due to the domain difference between the two datasets.

84-class instance segmentation. We can infer region labels by projecting the 3D scene models to the 2D images. Though instance segmentation neglects the evaluation of 3D localization of shapes and layout, which is the main purpose of our method, it can be a reasonable evaluation of object localization and classification. We evaluate the 84-class instance segmentation (81 object classes and 3 layout classes: wall, ceiling, floor) on the inferred region labels of our scene composition result. The evaluation follows the protocol in RMRC [45] that computes the coverage of ground truth regions, weighted or unweighted by area, for each image. The average coverage across all images is reported in Table 1. We see improvements, compared to Guo et al., due to both classification method (CNN vs. CCA) and region proposal method (MCG vs. Prim's). Also, it is interesting to see that

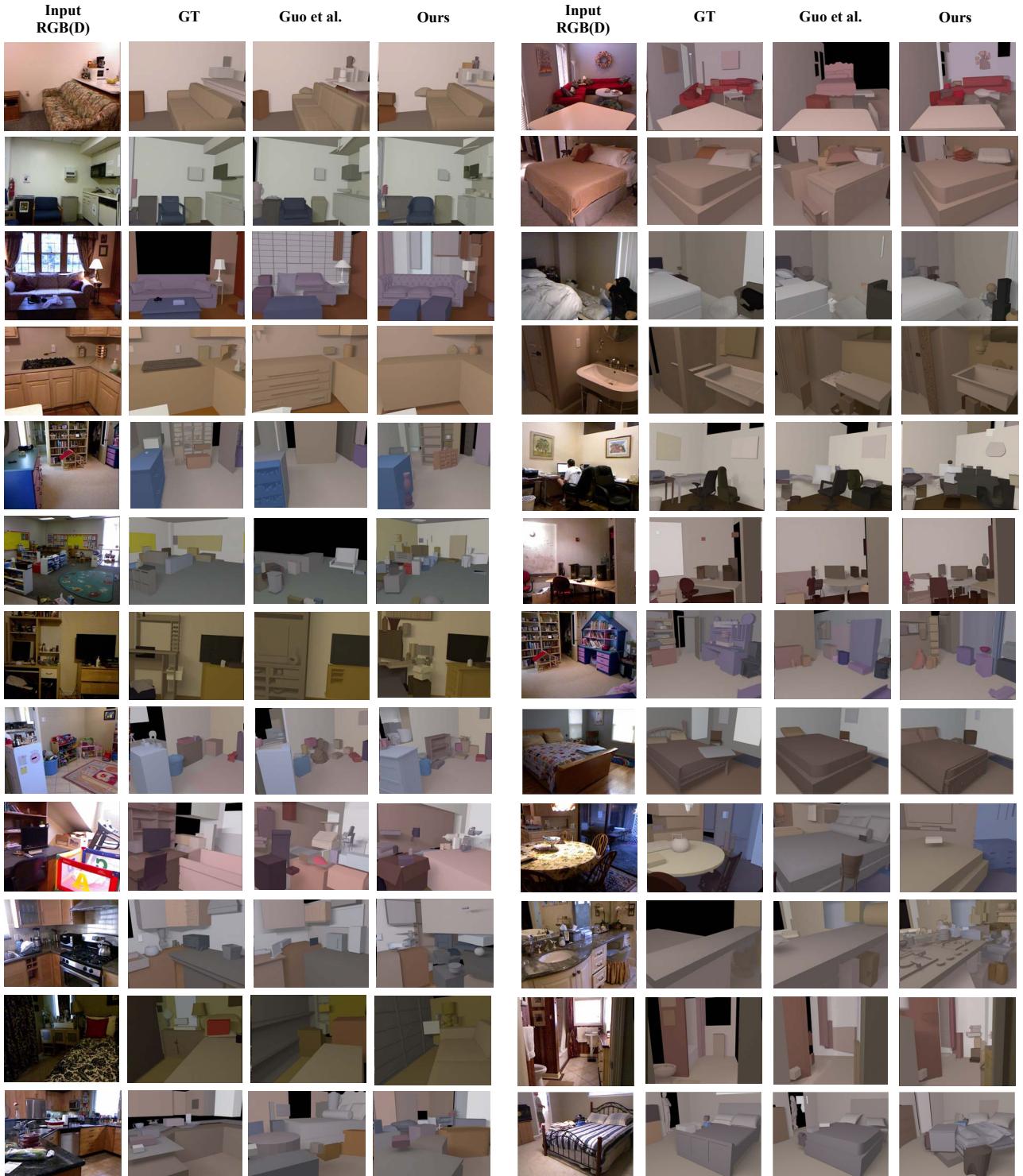


Fig. 8 Qualitative results on scene composition with automatic region proposals. We randomly sample images from the top 25% (first four rows), medium 50% (row 6-9) and worst 25% (last four rows) based on 84-class semantic segmentation accuracy.

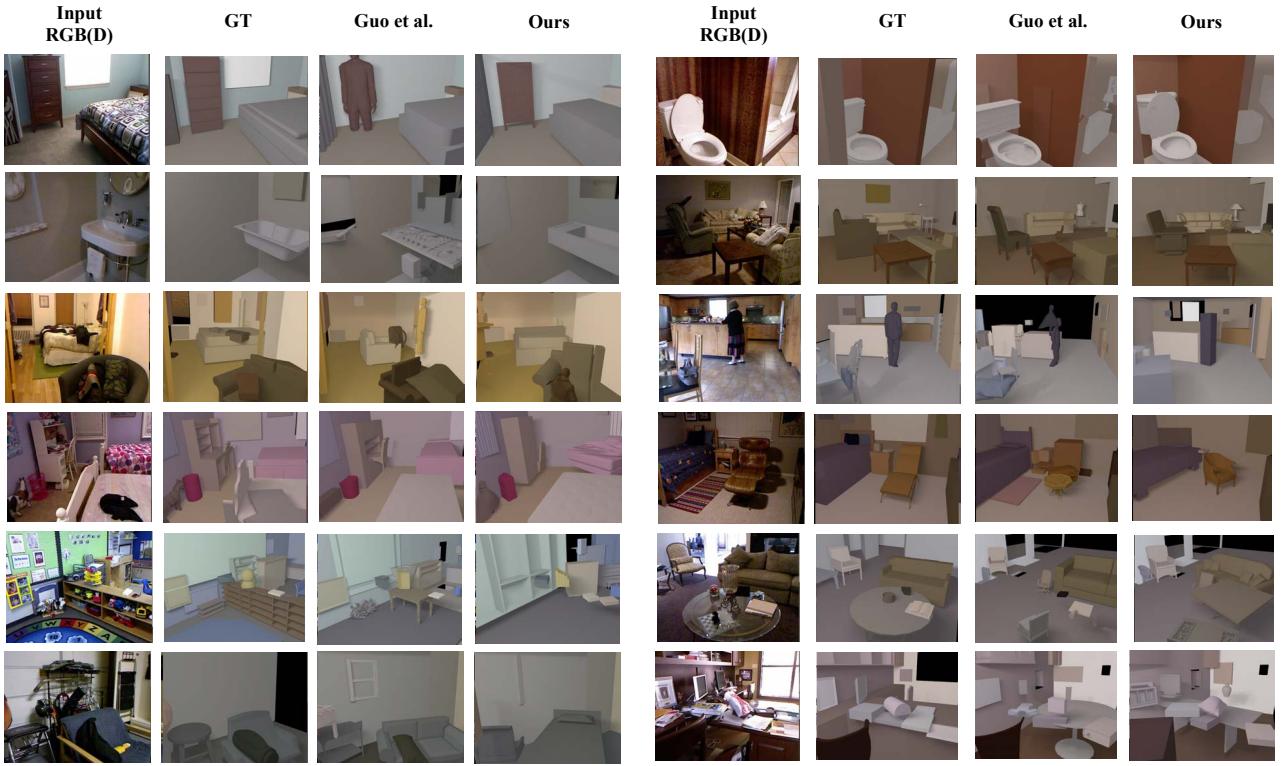


Fig. 9 Qualitative results on scene composition given ground truth 2D labeling as region proposals. We randomly sample images from the top 25% (first two rows), medium 50% (row 3-4) and worst 25% (last two rows) based on 84-class semantic segmentation accuracy.



Fig. 10 Qualitative results on SUN-RGBD dataset. Our method is able to estimate accurate 3D layout and the 3D occupancy of main furniture.

Table 1 Results of 84-class instance segmentation on both automatic region proposals and ground truth regions.

Method	Mean coverage weighted	Mean coverage unweighted
Guo et al. [16]	48.12	29.75
MCG+CCA	50.43	32.91
Ours w/o support	50.43	34.34
Ours	50.23	33.92
Ours w/ GT support	50.52	34.32
Ours w/ GT-region	52.39	35.61
Detailed ground truth labeling	65.42	49.15

using ground truth regions does not dramatically improve results, indicating that the region proposal and scene composition methods are highly effective.

Table 2 Results of 84-class semantic segmentation on both automatic region proposals and ground truth regions.

Method	Objects+Layouts		
	avg class	avg instance	avg pixel
Guo et al. [16]	6.96	30.39	43.85
MCG + CCA	8.04	32.56	46.17
Ours w/o support	10.51	32.45	45.94
Ours	10.77	32.90	46.25
Ours w/ GT support	11.28	33.71	47.46
Ours w/ GT-region	16.63	37.61	54.86
Detailed ground truth labeling	67.85	77.53	85.78

84-class semantic segmentation. Semantic segmentation is evaluated by the percent of pixels that are correctly classified, either averaged across pixels (“avg pixel”), instances (“avg instance”), or classes (“avg class”), as reported in Table 2. The average across pixels tends to be highest,

Table 3 Depth error for visible and occluded portions of layouts. Sensor error is the difference between the input depth image and annotated layout.

Method	Layout Depth Error		
	overall	visible	occluded
Sensor	0.517	0.059	0.739
Guo et al. [16]	0.166	0.074	0.204
Ours	0.150	0.074	0.181

Table 4 Pixel labeling error for layout surfaces with 5 categories (full dataset, 654 image).

Method	Layout Pixel Error		
	overall	occluded	visible
NYU Parser [39]	34.6	50.0	5.2
Guo et al. [16]	10.9	14.0	4.8
Ours	10.6	13.6	4.8

Table 5 Results of pixel labeling error for layout surfaces with 5 categories compared with Zhang et al. [53] (on intersection of test subsets in [39] and [53], 47 images). Note that [53] models layout as boxes, while our method can model more general layouts

Method	Layout Pixel Error		
	overall	occluded	visible
Zhang et al [53]	10.0	13.0	5.9
Guo et al. [16]	5.4	7.6	2.3
Ours	5.1	7.2	2.0

since the most common and largest objects are more likely to be correctly labeled. Again, our approach outperforms Guo et al. due to improved region proposals and classification. Support prediction also helps here. We also compute the 40-class semantic segmentation to compare to the state-of-the-art [31]: 18.3 average class / 33.8 average instance / 47.7 average pixel vs. their 34.0/49.5/65.4. Since our method projects 3D models of classified regions onto the images, it may not adhere to image boundaries as well as a direct semantic segmentation method and also may have more difficulty with small objects.

Layout estimation. In Table 4, we evaluate accuracy of labeling background surfaces into “left wall”, “right wall”, “front wall”, “ceiling”, and “floor”. Ground truth is obtained by rendering the 3D annotation of layout surfaces, and our prediction is obtained by rendering our predicted layout surfaces. The labels of “openings” (e.g., windows that are cut out of walls) are assigned based on the observed depth and surface normal. We compare to the RGBD region classifier of Silberman et al. [39] on the full test set. As expected, we outperform significantly on occluded surfaces (13.6% vs. 50.0% error) but also outperform on visible surfaces (4.8% vs. 5.2% error), which is due to the benefit of a structured scene model. Our method outperforms Guo et al.’s method on occluded surfaces, which means the layout estimation take benefits from the better object interpretation during the scene composition step. We also compare to Zhang et al. [53] who estimate box-like layout from RGBD images on the in-

Table 6 Results of freespace voxel estimation. Unoccupied voxel precision/recall using our method, given ground truth segmentation (GT-region) or only automatic region proposals. A baseline of freespace inferred from depth point (Sensor) is compared.

Method	Freespace	
	precision	recall
Sensor	1.000	0.7874
Guo et al. [16]	0.954	0.914
Ours	0.954	0.919
Ours w/ GT-region	0.955	0.925

Table 7 Results of predicted object occupied voxel precision/recall, compared to fitting bounding boxes to ground truth regions (Bbox).

Method	Occupancy			
	precision	recall	precision- ϵ	recall- ϵ
Bbox	0.487	0.298	0.756	0.569
Guo et al. [16]	0.504	0.380	0.751	0.646
Ours	0.478	0.397	0.741	0.710
Ours w/ GT-region	0.549	0.417	0.815	0.681

tersection of their test set with the standard test set (Table. 5). These images are easier than average, and our method outperforms substantially, cutting the error nearly in half (5.1% vs. 10.0%).

We evaluate layout depth prediction, the rendered depth of the room without foreground objects (Table 3). Error is the difference in depth from the ground truth layout annotation. On visible portions of layout surfaces, the error of our prediction is very close to that of the sensor, with the difference within the sensor noise range. On occluded surfaces, the sensor is inaccurate (because it measures the foreground depth, rather than that of the background surface), and the average depth error of our method is only 0.15 meters, which is quite good considering that the sensor noise is conservatively 0.03*depth.

Occupancy and Freespace Evaluation. We evaluate our scene prediction performance based on voxel prediction. The voxel representation has advantages of being computable from various volumetric representations, viewpoint-invariant, and usable for models constructed from multiple views (as opposed to depth- or pixel-based evaluations). The scope of the evaluation is the space surrounded by annotated layout surfaces. Voxels that are out of view or behind solid ground truth walls are not evaluated. We render objects separately and convert them into a solid voxel map. The occupied space is the union of all the voxels from all objects and layouts; free space is the complement of the set of occupied voxels.

Our voxel representation is constructed on a fine grid with $0.03m$ resolution to allow inspection of shape details of the 3D model objects we use. The voxel prediction and recall are presented in Table 6 and Table 7.

Table 8 Comparison with Deep Sliding Shapes (DSS) [41]. We evaluate voxel occupancy (our metric), semantic segmentation, and instance segmentation metric based on the predicted 19 object classes defined in DSS.

Method	Voxel occupancy				Semantic segmentation			Instance segmentation	
	precision	recall	precision- ϵ	recall- ϵ	avg class	avg instance	avg pixel	Mean coverage weighted	Mean coverage unweighted
DSS[41]	0.318	0.585	0.572	0.667	22.92	29.52	41.45	27.36	18.21
Ours	0.381	0.275	0.642	0.490	24.72	27.57	33.64	33.75	25.57

Table 9 Comparison with Semantic Scene Completion (SSCNet) [43]. We follow the metric used by SSCNet and evaluate on the NYUd v2 test split with kinect depth map.

Method	Scene completion			Semantic scene completion											
	precision	recall	IoU	ceiling	floor	wall	window	chair	bed	sofa	table	tvs	furniture	objects	avg
SSCNet	57.0	94.5	55.1	15.1	94.7	24.4	0	12.6	32.1	35	13	7.8	27.1	10.1	24.7
Ours	69.9	63.0	49.4	19.4	68.2	21.0	16.5	10.7	43.1	22.2	0.7	4.3	23.0	15.3	22.8

Table 10 3D Amodal object detection in NYUd v2 dataset. We evaluate on the mAP (%) as defined on [40]

Method	bed	toilet	sofa	table	chair
Sliding Shapes [40]	33.5	67.3	33.8	34.5	29
DSS [41]	84.7	89.9	55.4	70.5	61.1
Ours	36.9	36.1	14.12	4.81	3.1

There is inherent annotation and sensor noise in our data, which is often much greater than $0.03m$. Objects, when they are small, of nontrivial shape, or simply far away, result in very poor voxel accuracy, even though they agree with the input image. Therefore, we perform evaluation with a tolerance, proportional to the depth of the voxel, for which we use $\epsilon = 0.05 * \text{depth}$, based on the sensor resolution of Kinect. Specifically, an occupied voxel within ϵ of a ground truth voxel is considered to be correct (for precision) and to have recalled that ground truth voxel.

We compare to two simple baselines. For free space, we evaluate the *observed* free space from depth sensor. The free space from observed depth predicts 100% of the visible free space but recalls none of the free space that is occluded. For occupied space, our baseline generates bounding boxes based on ground truth segmentations with 10% outlier rejection. We outperform this baseline, whether using ground truth segmentations or automatic region proposals to generate the model, and we perform similarly to the Guo et al. variation. Also, note that precision is higher than recall, which means it is more common to miss objects than to generate false ones.

5.3 Comparison with Deep Sliding Shape

We show in Table 8 the comparison between our approach and the DSS method. Our approach demonstrates better precision of voxel occupancy, average class semantic segmentation and both instance segmentation metrics, indicating a better shape representation based on CAD model over bounding box. DSS shows better recall in object voxel occupancy, which is benefited from their 3D local search regime for each object in the whole scene. The better recall of predicted

3D objects also lead to their better performance in the average pixel accuracy and average instance accuracy of semantic segmentation.

We compare our performance on 3D object detection metrics in Table 10. Although 3D object detection is not our direct goal, our 3D detection on bed class is better than Sliding Shapes. We observe that our lower mAP for chair, sofa and table is due to three factors: 1) the failing of distinguishing between similar object classes: e.g. we predict on desk shape which is in ground truth a table, and we confuse between sofa and chair (with arms like sofa shape); 2) 3D localization error, which is especially for toilet case; 3) our method is constrained to produce objects that do not overlap, even with classes outside the five main furniture.

5.4 Comparison with Semantic Scene Completion Network

We show in Table 9 the comparison between our approach and the SSCNet method. For scene completion, our approach outperforms SSCNet in precision, which benefits from our detailed shape modeling for both objects and layout. SSCNet shows better performance in recall, due to their voxel-level prediction from observed occupancy space. Our lower recall also leads to our slightly lower performance in voxel-level occupancy IoU than SSCNet. For semantic scene completion, our predictions on ceiling, window, bed and other object classes outperform SSCNet. Our average voxel-level IoU is less than 2% lower than SSCNet. Errors mainly come from objects like table with fewer observed depth points to fit object and shape classification confusion like tv (classified as other objects) and sofa (classified as bed or chair) class.

5.5 Ablation study

Region proposals. In Tables 1 and 2, we see the impact of changing the region proposal method from Guo et al.’s Prim’s based algorithm to Gupta et al.’s [18] Multiscale Combinatorial Grouping (MCG). The “MCG + CCA” method is

Table 11 Quantitative evaluation for our retrieval method compared with Guo et al.’s method.

Method	avg class accuracy (%)			avg 3D IoU			avg surface distance (m)		
	top 1	top 2	top 3	top 1	top 2	top 3	top 1	top 2	top 3
Guo et al. [16]	11.97	16.36	19.65	0.134	0.177	0.200	0.033	0.029	0.027
Ours	41.56	54.47	62.07	0.191	0.231	0.249	0.026	0.024	0.023

Table 12 Our 81-class classification accuracy on ground truth 2D regions in the test set. We compare two methods: our classification network with/without estimating support height. We compute the average accuracy for each class, average precision based on the predicted probability and the accuracy averaged over instances. The classification networks are trained and evaluated 10 times and the means and standard deviations (reflecting variation due to randomness in learning) are reported. 15 common object class results are also listed. Bold numbers signify better performance.

Method	avg per class	avg precision	avg over instance	picture	chair	cabinet	pillow	bottle	books	paper	table	box	window	door	sofa	bag	lamp	clothes
w/o support height	43.7±0.3	37.7±0.1	40.8±0.3	57.5	46.1	39.5	66.5	55.6	30.0	40.7	36.7	10.6	61.4	54.9	63.0	14.9	64.6	25.3
w/ support height	44.7±0.3	39.7±0.1	42.7±0.2	57.5	53.1	44.35	69.0	54.9	33.5	43.5	39.5	14.1	62.4	57.8	65.9	15.1	65.9	26.9

Table 13 Classification accuracy for ground truth regions under different occlusion ratios in test set

Occlusion Ratio	< 0.5	> 0.5
w/ support height	45.8	38.5
w/o support height	44.2	35.7

Guo et al. with only a substitution of the region proposal method and improves over “Guo et al.” in each case. We can also see that using ground truth regions improves performance over automatic region proposals, but the improvement is less than we had expected, which indicates that our region proposal and scene composition steps are effective.

Retrieval. We evaluate our candidate shape retrieval method compared with Guo et al.’s retrieval method, as shown in table 11. To isolate from the influence of errors in other steps like region proposals generation or scene composition, we assume that the ground truth regions are given (and the candidate selection in Sec. 4.3.1 for the CNN-based method is not performed). We evaluate top N retrieved class accuracy and top N retrieved shape similarity, based on the shape intersection over union (IoU) and the surface-to-surface distance [36]. In our experiment, we set $N \in \{1, 2, 3\}$. To isolate from rotation ambiguity in shape similarity measurement, we rotate each retrieved object to find the best shape similarity score with the ground truth 3D shape. Our CNN-based retrieval method outperforms the Guo et al.’s CCA method under all the evaluation criteria.

We can indirectly measure the effectiveness of our **scene composition** by comparing “Ours” to “GT-Region” (use the best fitting 3D object model and class for each ground truth 2D region). For semantic segmentation, occupancy (Table 7), and free space (Table 6, we see a relatively small improvement by using ground truth regions. Qualitatively as well, our automatic region proposal and scene composition produce similarly good results to those produced using ground truth regions. Although the region proposal method is far from perfect, our scene composition step appears to be effective at keeping the good region proposals and discarding the others.

Support height prediction. We first examine the effect of predicting support height in region categorization. In table 12, we report the region classification accuracy of our classification network given or without the support prediction on the ground truth 2D regions in the test set. Overall, incorporating support height prediction improves the classification results. For certain classes, objects often appear on the same height (e.g. chair, desk, rug) have better classification accuracy given object’s height, while objects that can appear on several heights (e.g. picture) do not have this benefit. Table 13 shows the per-instance classification accuracy under different occlusion ratios of the ground truth regions in the test set. The occlusion ratio is computed from the observed 2D region over the 2D projection of the ground truth 3D label. The improvement in the classification accuracy is larger for highly occluded area, which agrees with our hypothesis that estimating object’s support height will help classify occluded regions.

As an ablation study of the effect of incorporating support inference in the overall scene composition performance, in Table 1 and Table 2, we report our result with or without support height prediction and using the ground truth support height as the upper bound for including support reasoning. As we can see, incorporating support improves the semantic segmentation performance by about 1%. Figure 11 shows the effect of applying support reasoning on semantic segmentation accuracy of each class. Classes that are often at a certain height: toilet, printer, lamp, etc. will have benefits from support height information. Classes that might appear in several levels of height: mask, pixture, light won’t gain benefits from this. Using ground truth support height generally leads to a larger benefit than estimated support height, as expected.

6 Discussions

We discuss the performance of our framework and the contributions of each of the component. Our framework fol-

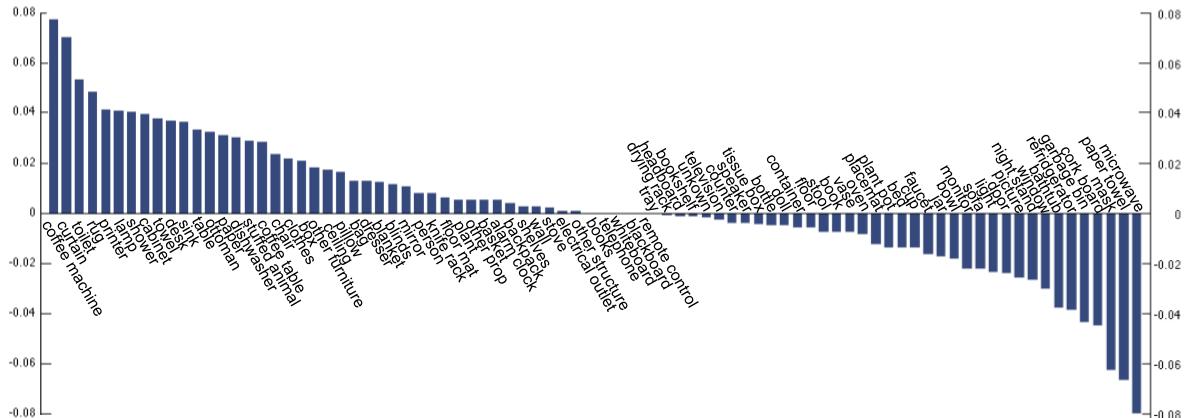


Fig. 11 Semantic segmentation accuracy of each class w/ support minus accuracy w/o support accuracy with automatic region proposals

lows three steps: region proposal, classification and shape retrieval, and scene composition. Each step solves for a loosely separate sub-problem, but are all closely related to produce a compact result for the complete 3D scene parsing task. For in-depth investigation of each of the component, as in the ablation study in Sec. 5.5, we perform experiments and analyze the effectiveness of the design choices in each step. For the region proposal step, we see improvements on applying the state-of-the-art RGBD region proposal method by Gupta et al. [18] instead of our previous Prim’s based method. We observe less difference in overall performance if we use ground truth regions in our framework, which indicates the downstream step’s robustness to errors from region proposals. For the retrieval step, our use of CNNs to classify and retrieve similar shapes outperforms our previous simpler CCA-based retrieval approach. We validate our hypothesis that estimating support height of objects can lead to better classification to improve retrieval, especially for occluded objects. The heavy occlusion characteristic of indoor scene makes the region classification harder, and accumulates errors to the retrieval step. To resolves the retrieval errors, we incorporate the scene composition step with combinatorial optimization to select a subset of shape candidates that fits the observation and scene regularity like no 3D overlap between shapes. We combine greedy search and a hill-climbing method to efficiently solve the hard optimization problem which outperforms a variety of other attempted methods including general integer programming algorithms and relaxations. Note that possible improvements for the framework exists, as we will discuss in the following paragraphs.

Common errors made by our approach include splitting large objects into small ones, completely missing small objects, difficulty with highly occluded objects such as chairs. Based on the previous discussions, we see the major sources of error of our framework are: 1) region classification error and 2) shape fitting error. Errors mainly result from the heavy occlusion from objects in indoor scene. Fewer errors accumulate in the shape retrieval step, referring to the com-

parison between Table 12 and 11. The method does not seem to be very sensitive to the region proposal method, based on the comparison in Table 1, 2 and 7. In fact, we were surprised to find that results from automatic region proposals are often comparable to those from ground truth regions. In part, this is because the later fitting and optimization steps are effective at discarding bad regions and retaining good ones.

We see many interesting directions for future work: 1) modeling object context such as chairs tend to be near/under tables, relative pose among objects and the layouts; 2) modeling self-similarity, e.g. most chairs within one room will look similar to each other; 3) incorporating semantic constraints. For example, SSCNet by Song et al [43] employs neural network to perform semantic completion from single depth image. For further improvement, our system can incorporate the semantic completion from Song et al. to generate candidate region proposals and add semantic constraint for scene composition in Sec. 4.4.

7 Conclusions

We proposed an approach to predict a complete 3D scene model of individual objects and surfaces from a single RGBD image. Our representation encodes the layout of walls as 3D planes and all interpretable objects as 3D mesh models, parsing both the visible and occluded portion of the scene. We take a data-driven approach, generating sets of potential object regions, matching to regions in training images, and transferring and aligning associated 3D models while encouraging fit to observations and spatial consistency. We incorporate support inference to aid interpretation and propose a retrieval scheme that uses CNNs to classify regions and find objects with similar shapes. We demonstrate the performance of our method on our newly annotated NYUD v2 dataset with detailed 3D shapes. Compared to our earlier work (Guo et al. [16]), we demonstrate improvements due to better region proposals and use of CNNs to classify

and retrieve similar shapes. We also show that our results from automatic region proposals are nearly as good as from ground truth regions, demonstrating the effectiveness of our scene composition and, more generally, of finding a scene hypothesis that is consistent in semantics and geometry. We also showed that estimating support height of objects can lead to better classification, especially for occluded objects.

Acknowledgements This research is supported in part by ONR MURI grant N000141010934 and ONR MURI grant N000141612007. We thank David Forsyth for insightful comments and discussion and Saurabh Singh, Kevin Shih and Tanmay Gupta for their comments on an earlier version of the manuscript.

References

1. Aubry M, Maturana D, Efros AA, Russell BC, Sivic J (2014) Seeing 3d chairs: Exemplar part-based 2d-3d alignment using a large dataset of cad models. In: CVPR
2. Banica D, Sminchisescu C (2013) Cpmc-3d-o2p: Semantic segmentation of rgbd images using cpmc and second order pooling. CoRR
3. Carreira J, Sminchisescu C (2012) Cpmc: Automatic object segmentation using constrained parametric min-cuts. PAMI 34(7):1312–1328
4. Chang AX, Funkhouser T, Guibas L, Hanrahan P, Huang Q, Li Z, Savarese S, Savva M, Song S, Su H, Xiao J, Yi L, Yu F (2015) ShapeNet: An Information-Rich 3D Model Repository. Tech. Rep. arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago
5. Dai A, Chang AX, Savva M, Halber M, Funkhouser T, Nießner M (2017) Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proc. Computer Vision and Pattern Recognition (CVPR), IEEE
6. Dasgupta S, Fang K, Chen K, Savarese S (2016) Delay: Robust spatial layout estimation for cluttered indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 616–624
7. Delage E, Lee H, Ng AY (2006) A dynamic Bayesian network model for autonomous 3D reconstruction from a single indoor image. In: CVPR
8. Deng Z, Todorovic S, Jan Latecki L (2015) Semantic segmentation of rgbd images with mutex constraints. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1733–1741
9. Dollár P, Zitnick CL (2013) Structured forests for fast edge detection. In: ICCV
10. Endres I, Hoiem D (2010) Category independent object proposals. In: ECCV
11. Flint A, Murray DW, Reid I (2011) Manhattan scene understanding using monocular, stereo, and 3d features. In: ICCV
12. Furukawa Y, Curless B, Seitz SM, Szeliski R (2009) Reconstructing building interiors from images. In: ICCV
13. Gong Y, Ke Q, Isard M, Lazebnik S (2013) A multi-view embedding space for internet images, tags, and their semantics. IJCV
14. Guo R, Hoiem D (2012) Beyond the line of sight: Labeling the underlying surfaces. In: ECCV
15. Guo R, Hoiem D (2013) Support surface prediction in indoor scenes. In: ICCV
16. Guo R, Zou C, Hoiem D (2015) Predicting complete 3d models of indoor scenes. arXiv preprint arXiv:150402437
17. Gupta S, Arbelaez P, Malik J (2013) Perceptual organization and recognition of indoor scenes from RGB-D images. In: CVPR
18. Gupta S, Girshick R, Arbeláez P, Malik J (2014) Learning rich features from rgbd images for object detection and segmentation. In: European Conference on Computer Vision, Springer, pp 345–360
19. Gupta S, Arbeláez P, Girshick R, Malik J (2015) Aligning 3d models to rgbd images of cluttered scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4731–4740
20. Hariharan B, Arbeláez P, Girshick R, Malik J (2014) Simultaneous detection and segmentation. In: Computer vision—ECCV 2014, Springer, pp 297–312
21. Hedau V, Hoiem D, Forsyth D (2009) Recovering the spatial layout of cluttered rooms. In: ICCV
22. Hedau V, Hoiem D, Forsyth D (2010) Thinking inside the box: Using appearance models and context based on room geometry. In: ECCV
23. Hoiem D, Efros AA, Hebert M (2008) Putting objects in perspective. International Journal of Computer Vision 80(1):3–15
24. Karsch K, Liu C, Kang SB (2012) Depth extraction from video using non-parametric sampling. In: ECCV
25. Kingma D, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980
26. Lee DC, Hebert M, Kanade T (2009) Geometric reasoning for single image structure recovery. In: CVPR
27. Li Y, Su H, Qi CR, Fish N, Cohen-Or D, Guibas LJ (2015) Joint embeddings of shapes and images via cnn image purification. ACM Trans Graph
28. Lim JJ, Pirsiavash H, Torralba A (2013) Parsing IKEA objects: Fine pose estimation. In: ICCV
29. Lim JJ, Khosla A, Torralba A (2014) Fpm: Fine pose parts-based model with 3d cad models. In: European Conference on Computer Vision, Springer, pp 478–493
30. Lin D, Fidler S, Urtasun R (2013) Holistic scene understanding for 3d object detection with rgbd cameras. In:

ICCV

31. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3431–3440
32. Mallya A, Lazebnik S (2015) Learning informative edge maps for indoor scene layout prediction. In: Proceedings of the IEEE International Conference on Computer Vision, pp 936–944
33. Manen S, Guillaumin M, Van Gool L (2013) Prime object proposals with randomized prim's algorithm. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2536–2543
34. Ren X, Bo L, Fox D (2012) Rgb-(d) scene labeling: Features and algorithms. In: CVPR
35. Roberts LG (1963) Machine perception of 3-D solids. PhD thesis, Massachusetts Institute of Technology
36. Rock J, Gupta T, Thorsen J, Gwak J, Shin D, Hoiem D (2015) Completing 3d object shape from one depth image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2484–2493
37. Satkin S, Hebert M (2013) 3dnn: Viewpoint invariant 3d geometry matching for scene understanding. In: ICCV
38. Schwing AG, Urtasun R (2012) Efficient exact inference for 3d indoor scene understanding. In: ECCV
39. Silberman N, Hoiem D, Kohli P, Fergus R (2012) Indoor segmentation and support inference from rgbd images. Computer Vision–ECCV 2012 pp 746–760
40. Song S, Xiao J (2014) Sliding shapes for 3d object detection in depth images. In: European conference on computer vision, Springer, pp 634–651
41. Song S, Xiao J (2016) Deep sliding shapes for amodal 3d object detection in rgb-d images. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
42. Song S, Lichtenberg SP, Xiao J (2015) Sun rgb-d: A rgb-d scene understanding benchmark suite. In: CVPR, vol 5, p 6
43. Song S, Yu F, Zeng A, Chang AX, Savva M, Funkhouser T (2017) Semantic scene completion from a single depth image. Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition
44. Tighe J, Lazebnik S (2010) Superparsing: Scalable non-parametric image parsing with superpixels. In: ECCV
45. Urtasun R, Fergus R, Hoiem D, Torralba A, Geiger A, Lenz P, Silberman N, Xiao J, Fidler S (2013) Reconstruction meets recognition challenge
46. Walk S, Schindler K, Schiele B (2010) Disparity statistics for pedestrian detection: Combining appearance, motion and stereo. In: Computer Vision–ECCV 2010, Springer, pp 182–195
47. Wu Z, Song S, Khosla A, Yu F, Zhang L, Tang X, Xiao J (2015) 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1912–1920
48. Xiang Y, Kim W, Chen W, Ji J, Choy C, Su H, Mottaghi R, Guibas L, Savarese S (2016) Objectnet3d: A large scale database for 3d object recognition. In: European Conference Computer Vision (ECCV)
49. Xiao J, Furukawa Y (2012) Reconstructing the world's museums. In: ECCV
50. Xiao J, Russell BC, Torralba A (2012) Localizing 3d cuboids in single-view images. In: NIPS
51. Yamaguchi K, Kiapour MH, Berg TL (2013) Paper doll parsing: Retrieving similar styles to parse clothing items. In: ICCV
52. Yih Wt, Toutanova K, Platt JC, Meek C (2011) Learning discriminative projections for text similarity measures. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning, Association for Computational Linguistics, pp 247–256
53. Zhang J, Kan C, Schwing AG, Urtasun R (2013) Estimating the 3d layout of indoor scenes and its clutter from depth sensors. In: ICCV
54. Zhang Y, Song S, Tan P, Xiao J (2014) Panoccontext: A whole-room 3d context model for panoramic scene understanding. In: European Conference on Computer Vision, Springer, pp 668–686
55. Zhang Y, Bai M, Kohli P, Izadi S, Xiao J (2016) Deepcontext: Context-encoding neural pathways for 3d holistic scene understanding. In: ICCV
56. Zhao Y, Zhu SC (2013) Scene parsing by integrating function, geometry and appearance models. In: CVPR