3D Scene Understanding by Voxel-CRF

Byung-soo Kim
University of Michigan
bsookim@umich.edu

Pushmeet Kohli
Microsoft Research Cambridge
pkohli@microsoft.com

Silvio Savarese
Stanford University
ssilvio@stanford.edu

Abstract

Scene understanding is an important yet very challenging problem in computer vision. In the past few years, researchers have taken advantage of the recent diffusion of depth-RGB (RGB-D) cameras to help simplify the problem of inferring scene semantics. However, while the added 3D geometry is certainly useful to segment out objects with different depth values, it also adds complications in that the 3D geometry is often incorrect because of noisy depth measurements and the actual 3D extent of the objects is usually unknown because of occlusions. In this paper we propose a new method that allows us to jointly refine the 3D reconstruction of the scene (raw depth values) while accurately segmenting out the objects or scene elements from the 3D reconstruction. This is achieved by introducing a new model which we called Voxel-CRF. The Voxel-CRF model is based on the idea of constructing a conditional random field over a 3D volume of interest which captures the semantic and 3D geometric relationships among different elements (voxels) of the scene. Such model allows to jointly estimate (1) a dense voxel-based 3D reconstruction and (2) the semantic labels associated with each voxel even in presence of partial occlusions using an approximate yet efficient inference strategy. We evaluated our method on the challenging NYU Depth dataset (Version 1 and 2). Experimental results show that our method achieves competitive accuracy in inferring scene semantics and visually appealing results in improving the quality of the 3D reconstruction. We also demonstrate an interesting application of object removal and scene completion from RGB-D images.

1. Introduction

Understanding the geometric and semantic structure of a scene (scene understanding) is a critical problem in various research fields including computer vision, robotics, and augmented reality. For instance, consider a robot in the indoor scene shown in the Fig. 1. In order to safely navigate through the environment, the robot must perceive the free space of the scene accurately (geometric structure). Moreover, in order for the robot to effectively interact with the environment (*e.g.,* to place a bottle on a table), it must recognize the objects in the scene (semantic structure).

Several methods have been proposed to solve the problem of scene understanding using a single RGB (2D) image. For instance, in [1, 2, 3, 4], the problem is formulated in terms of the estimation of a consistent set of semantic labels of local image regions (patches or pixels) assuming a *flat* image world. Although the results were promising, such methods do not provide information about the geometric structure of the scene. Recently, attempts have been made to jointly estimate the 3D and semantic properties of a scene using a single image or multiple images [5, 6, 7, 8, 9]. The efficacy of such methods in perceiving the scene geometry, however, is limited due to the inherent geometric ambiguity in a single image. To overcome the ambiguity, researchers have considered using depth and RGB image data for scene understanding [10, 11, 12]. Instead of labeling local 2D image regions, these methods provide semantic description of 3D elements (point clouds) acquired by a RGB-D camera [13]. However, they rely on the assumption that the 3D structure from the RGB-D device is accurate. This is not always the case due to photometric interference, discretization error, etc (see Fig. 2 for typical noisy reconstruction).

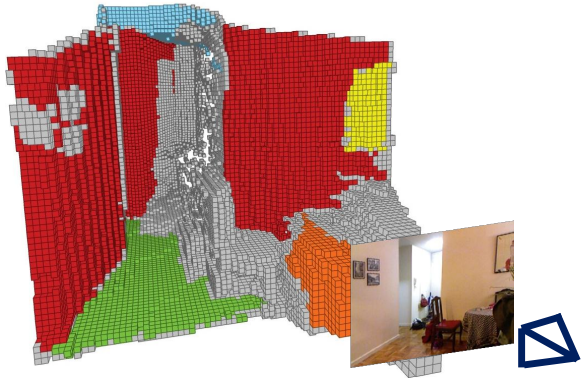
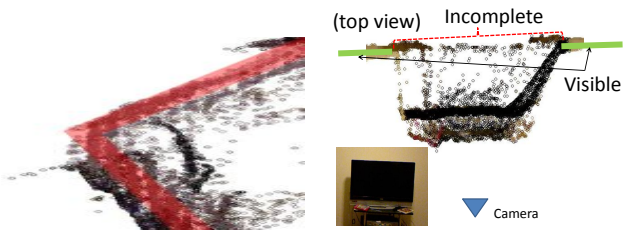


Figure 1: Given a single depth-RGB image, our proposed Voxel-CRF (V-CRF) model jointly estimates (1) a dense voxel-based 3D reconstruction of the scene and (2) the semantic labels associated with each voxel. In the figure, red corresponds to 'wall', green to 'floor', orange to 'table' and yellow to 'picture'.

In this work, we propose a method to jointly estimate the semantic and geometric structure of a scene given a single RGB-D image. Unlike [10, 11, 12] where the true geometric structure is assumed to be given and fixed, we represent a scene with a set of small cubic volume (*voxel*) in the space of interest. We jointly estimate both the semantic labeling



(a) (b)

Figure 2: (a) Reconstructed point cloud taken from the corner of the room. Ground truth '*wall*' is highlighted with a red mask. Reconstructing reliable 3D geometry from noisy point cloud is a challenging task. (b) Point clouds do not completely describe the 3D scene. For example, the wall behind the *tv* cannot be reconstructed from depth map.

and 3D geometry of voxels of the scene given a noisy set of inputs. This allows us to *i)* correct noisy geometric estimation in input data and *ii)* provide the interpretation of non-visible geometric elements (such as the wall occluded by the table in Fig. 1). Our method is based on a voxel conditional random field model which we have called Voxel-CRF (V-CRF). In our V-CRF model, each node represents a voxel in the space of interest. A voxel may or may not include one or multiple points acquired by the RGB-D sensor. The state of each voxel is summarized by two variables, *occupancy* and *semantic label*. An auxiliary variable *visibility* is introduced to help relate voxels and 2D RGB or depth observation (Sec. 3). Semantic and geometric interpretation of a scene is achieved by finding the configuration of all variables that best explains the given observation.

The configuration of variables in the V-CRF model needs to be consistent with certain important geometric and semantic rules that ensure stable and more accurate 3D reconstruction and classification of the elements in the scene. This includes relationships such as '*supported by*' or '*attached to*' (Sec. 4.2). Geometric and semantic relationships based on higher-level elements such as certain groups of voxels which belong to the same plane (or object) are encoded using interactions between groups of voxels. These relationships are especially useful for consistent labeling of voxels in an occluded space (Sec. 4.3). The parameters associated with the above-mentioned interaction functions are learned from training data.

Given our V-CRF model, we solve the scene understanding problem by minimizing the energy function associated with the V-CRF. Instead of assuming that the true 3D geometry is given, we jointly estimate the geometric and semantic structure of the scene by finding the best configuration of all occupancy and semantic label variables of all voxels in the space. Our inference algorithm iterates between 1) deciding voxels to be associated with observations and 2) reasoning about the geometric and semantic description of voxels. In each iteration, we obtain an approximate solution using graph-cuts based inference [14].

In summary, the contributions of this paper are 5 folds. 1) We propose a new voxel based model for 3D scene understanding with RGB-D data that jointly infers the geometric and semantic structure of the scene (Sec. 3). 2) We improve structure estimation given noisy and incomplete 3D reconstruction provided by RGB-D sensors. 3) Geometric and semantic rules are proposed and modeled in the V-CRF model (Sec. 3&4). 4) An efficient iterative method is proposed for performing inference in the V-CRF model (Sec. 5). 5) We demonstrate (through qualitative and quantitative results and comparisons on benchmarks) that V-CRF produces accurate geometric and semantic scene understanding results (Sec. 6). Some applications enabled by the V-CRF model are discussed in Sec. 6.4.

2. Related Work

Our model is related to [5, 15] in that blocks are used to represent 3D space. On the other hand, unlike [5, 15], our blocks are defined at a fine resolution that enables us to understand scenes (such as cluttered indoor environments) in more detail. The methods proposed in [16, 17, 7] are also relevant to our work. These methods analyze geometric properties of the underlying scene and infer free space. However, our model can produce more fine grained labelling of geometric and semantic structure which is important for cluttered scenes. The approaches for scene understanding described in [10, 11, 12, 18] are based on RGB-D data. Similar to our method, these methods assign semantic labels to image elements such as 3D points or superpixels. However, image elements in these works are defined only over visible elements. Also, it is assumed that image elements are already correctly localized in 3D space. In contrast, our model can reason about the labelling of both visible and occluded image elements.

Our work is also closely related to [19, 20] in the use of a random field model for joint semantic labeling and geometric interpretation. [19] encouraged consistent semantic and geometric labelling of pixels by penalizing sudden changes in depth or semantic labeling results. [20] showed that the joint geometric-semantic labelling model helps in geometry estimation. Similar to our occlusion reasoning, they showed that the depth of fully occluded regions can be inferred by having stereo images. However, they did not consider a complete reconstruction of the scene. The problem of labelling occluded regions is also discussed in [21], where relative relationships between objects and background are used to infer labels of the occluded region. However, the lack of a voxel representation restricts [21] to reconstruction of the foreground and background layers. In contrast, in theory, our model can reconstruct any number of layers in the scene.

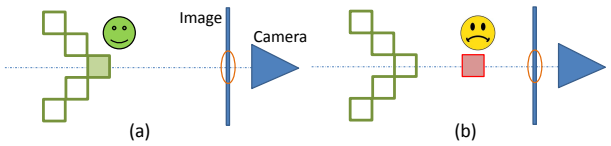


Figure 3: Ambiguity of assigning image observations to the voxels in a view ray. Five voxels with green outline are the ground truth voxels in a correct place. (a) For the successful cases, the voxel can be reconstructed from a depth data. (b) Unfortunately, due to noisy depth data, incorrect voxels are reconstructed in many cases.

3. Voxel-CRF

We now describe the Voxel-CRF (V-CRF) model. We represent the semantic and geometric structure of the scene with a 3D lattice where each cell of the lattice is a *voxel*. V-CRF is defined over a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{C})$, where $\mathcal{V}$ are vertices representing voxels, edges $\mathcal{E}$ connect horizontally or vertically adjacent pairs of vertices, cliques $\mathcal{C}$ are groups of voxels which are related, *e.g.,* voxels on the same 3D plane, or voxels that are believed to belong to an object (through an object detection bounding box). The state of each voxel is described with a structured label $\ell_i = (o_i, s_i)$ and the visibility v_i. The first variable o_i represents voxel occupancy; *i.e.,* it indicates whether voxel i is *empty* ($o_i = 0$) or *occupied* ($o_i = 1$). The second variable s_i indicates the index of semantic class the voxel belong to; *i.e.,* $s_i \in \{1, \cdots, |S|\}$ if the voxel is occupied ($o_i = 1$), or $s_i = \emptyset$ if $o_i = 0$, where $|S|$ is the number of semantic classes (*e.g.,* table, wall, ...). Estimation of the structured label $L = \{\ell_i\}$ over the V-CRF model produces a geometric and semantic interpretation of the scene.

The variable v_i encodes the visibility of a voxel i where $v_i = 1$ and $v_i = 0$ indicate whether the voxel is *visible* or *non-visible*, respectively. Any given ray from the camera touches a single visible (occupied) voxel. Due to the high amount of noise in the RGB-D sensor, it is difficult to unambiguously assign 2D observations (image appearance and texture) to voxels in 3D space (see Fig. 3 for an illustration). The visibility variables v_i allow us to reason about this ambiguity. Provided that we know which single voxel is visible on the viewing-ray, we can assign the 2D image observation to the corresponding voxel. Since *visibility* is a function of *occupancy*, and vice versa, we infer the optimal configuration of the two in an iterative procedure.

V-CRF can be considered as a generalization of existing CRF-based models for scene understanding in 3D [10, 11, 12], where $\{o_i\}$ and $\{v_i\}$ are assumed to be given, and semantic labels $\{s_i\}$ are inferred only for visible and occupied scene elements. In contrast, V-CRF model is more flexible by having o_i and v_i as random variables, and this enables richer scene interpretation by *i)* estimating occluded regions, *e.g.,* (o_i, s_i)=(*occupied*,*table*), v_i= *occluded*, and *ii)* correcting noisy depth data.

4. Energy Function

Given a graph $\mathcal{G}$, we aim to find $V^* = \{v_i^*\}$ and $L^* = \{\ell_i^*\}$ that minimize the energy function $E(V, L, O)$, where $O = \{C, I, D\}$, C is the known camera parameters, I is the observation from a RGB image, and D is the observation from a depth map. The energy function can be written as a sum of potential functions defined over individual, pairs, and group of voxels as: $E(V, L, O) =$

$$\sum_i \phi_u(v_i, \ell_i, O) + \sum_{i,j} \phi_p(v_i, \ell_i, v_j, \ell_j, O) + \sum_c \phi_c(V_c, L_c, O) \tag{1}$$

where i and j are indices of voxels and c is the index of higher-order cliques in a graph. The first term models the observation cost for individual voxels, while the second and third terms model semantic and geometric consistency among pairs and groups of voxels, respectively.

4.1. Observation for Individual Voxels

The term ϕ_u represents the cost of the assignment (v_i, ℓ_i) for a voxel i. We model the term for two different cases, when voxel i is occupied ($o_i = 1$) and when it is empty ($o_i = 0$).

$$\phi_u(v_i, \ell_i, O) = \begin{cases} k_1 & \text{if } o_i = 1, s_i \neq \emptyset \\ k_2 & \text{if } o_i = 0, s_i = \emptyset \end{cases} \tag{2}$$

where k_1 and k_2 are defined as $k_1 =$:

$$w_1^u v_i \log P(s_i|O) - w_2^u \log f_s(d_i - d^m_{r(i)}) - w_3^u \log \frac{|\mathrm{P}_i|}{|\mathrm{P}_i^{max}|} \tag{3}$$

$$\text{and } k_2 = -w_4^u \log(1 - f_s(d_i - d^m_{r(i)})) - w_5^u \log(1 - \frac{|\mathrm{P}_i|}{|\mathrm{P}_i^{max}|}). \tag{4}$$

When the voxel i is occupied ($o_i = 1$), it is composed of three terms. The first term incorporates the observations $P(s_i|O)$ from an image if it is visible ($v_i = 1$), to estimate a structured label of the voxel. The second term models the uncertainty in the depth value from the RGBD image through a normal distribution $f_s \sim \mathcal{N}(0, \sigma^2)$. Larger the disparity between depth according to the data map value $d_r^m(i)$, which is value associated with a ray $r(i)$ for a voxel i, and the voxel i's depth d_i, more likely it is to be labeled as an empty state. The third term models the occupancy based on density of 3D points in a voxel i. Note that there can be more than one image pixel corresponding to a voxel. We measure the ratio $|\mathrm{P}_i|/|\mathrm{P}_i^{max}|$, which is the ratio between the number of detected points in 3D cubical volume associated with a voxel i over the maximum number of 3D points in a voxel i, *i.e.,* the number of rays penetrating through a voxel i. If there is an object at voxel i, and the surface is perpendicular to the camera ray, the number of points is the largest. If this ratio is small (*i.e.* few points), the energy function encourages $o_i = 0$.

In the case the voxel i is empty ($o_i = 0$), the energy models the sensitivity of the sensor (first term) and the den-

sity of point clouds (second term). Different terms are balanced with weights $w^u_{\{\cdot\}}$, which are learned from the training dataset as discussed in Sec. 5.

4.2. Relating Pairs of Voxels

The pairwise energy terms penalize labellings of pairs of voxels that are geometrically or semantically inconsistent. Two different types of neighborhoods are considered to define pairwise relationships between voxels: *i)* adjacent voxels in 3D lattice structure, and *ii)* adjacent voxels in its 2D projection. The pairwise costs depend on visibility, spatial relationship, and appearance similarity of a pair of voxels. Appearance similarity between a pair of voxels (*e.g.,* color) is represented by c_{ij} which is a discretized color difference between voxels i and j, similar to [22]. If voxel j is empty or occluded, we use c_i, *i.e.* in this case the cost is the function of the color of the visible voxel i. The pairwise cost on the labelling of voxels also depends on their visibility and is defined as:

$$\phi_p(v_i = 1, \ell_i, v_j = 1, \ell_j) = w_1^{pw}(s_{ij}, c_{ij})T[\ell_i \neq \ell_j] \quad (5)$$

$$\phi_p(v_i = 0, \ell_i, v_j = 1, \ell_j) = w_2^{pw}(s_{ij}, c_j)T[\ell_i \neq \ell_j] \quad (6)$$

$$\phi_p(v_i = 1, \ell_i, v_j = 0, \ell_j) = w_3^{pw}(s_{ij}, c_i)T[\ell_i \neq \ell_j] \quad (7)$$

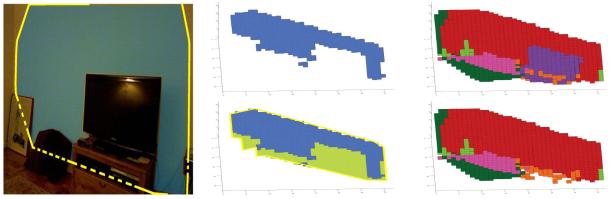
$$\phi_p(v_i = 0, \ell_i, v_j = 0, \ell_j) = w_4^{pw}, \quad (8)$$

where $T[\cdot]$ is the indicator function, s_{ij} is a spatial relationship between voxels i and j. i and j are chosen differently for 2D and 3D cases as discussed below. These functions penalize if ℓ_i and ℓ_j are inconsistent. The exact penalty for inconsistent assignments depends on the relative spatial location s_{ij} and colors c_{ij} of the voxel pairs. $w^{pw}_{\{\cdot\}}(s_{ij}, c_{ij})$ are weights that are learned from the training data.

Adjacent pairs in 3D. For all adjacent pairs of voxels, we specify their spatial relationship s_{ij}, where $s_{ij} \in \{vertical, horizontal\}$. The color difference between i and j is also used to modulate the cost $w^{pw}_{\{\cdot\}}(\cdot)$, where we cluster color difference between two voxels as in [22], c_{ij} is the index of a closest cluster.

Adjacent pairs in 2D. On top of adjacent voxels in 3D, the adjacency between a pair of voxels in the projected 2D images is formulated as pairwise costs. For example, occlusion boundaries are useful cues to distinguish voxels that belong to different objects; if two voxels are across a detected occlusion boundary (when projected in the view of the camera), they are likely to have different semantic labels. On the other hand, if two voxels across the boundary are still close in 3D, they are likely to have a same semantic label. The relationship of voxels are automatically indexed as follows. First, we extract pairs of 2D pixels from 2D RGB images which are on the opposite side of the occlusion boundaries. The pair of 2D pixels are then projected into 3D voxels. From the training data, we collect the relative surface feature between voxels i and j[1] and cluster them to represent different types of corners, depending on their geometric properties in 3D. Finally, the spatial index s_{ij} indicates a cluster ID. We learn different weights for different cluster automatically from the training data.

[1]The surface feature for adjacent regions i and j is composed of surface norm, color, and height.



(a) (b) (c)

Figure 4: (Best visible in a high resolution) (a) A detected plane using [23] is highlighted with the blue mask. Its convex hull is drawn with the yellow polygon and it includes both visible and occluded region of a planar surface. (b) A group of voxels associated with the detected planar surface (top) and a group of voxels associated with the convex hull (bottom). The voxels in the convex hull not only enforce consistency for visible voxels, but also for occluded voxels. (c) V-CRF result: our model not only allows the labeling of visible voxels for TV (top), but also the labeling of the occluded region corresponding to the 'wall'. For visibility, we removed the voxels corresponding to the TV. (bottom).

4.3. Relating Groups of Voxels

We now introduce higher-order potentials that encode the relationship among more than two voxels. The potentials enforce semantic and geometric consistency among voxels in a clique $c \in \mathcal{V}_C$ of voxels that can be quite far from each other. The relationships for a group of voxels can be represented using the Robust Pott's model [1]. Different types of 3D priors can be used, *e.g.,* surface detection, object detection, or room layout estimation; however, in this work, we consider two types of voxel groups $\mathcal{V}_C$, 1) 3D surfaces that are detected using a Hough voting based method [23] and 2) categorical object detections [24] as follows[2].

3D Surfaces. The first type is the group of voxels that belong to a 3D surface (wall, tables etc). From the depth data and its projected point clouds, we can identify 3D surfaces [23] and these are useful to understand images for two reasons. First, a surface is likely to belong to an object or a facet of the indoor room layout, and there is consistency among labels of voxels for a detected plane. Second, the part of the plane occluded by other objects can be inferred by extending the plane to include the convex hull[3] of the detected surface (See Fig. 4). According to the law of closure of Gestalt theory, both visible and invisible regions inside this convex hull are likely to belong to the same object.

Object Detections. Object detection methods provide a cue to define groups of voxels (bounding box) that take the same label, as used for 2D scene understanding in [25, 4], where we grouped a set of visible voxels which fall inside

[2]Room layout estimation is not used due to heavy clutter in the evaluated dataset.

[3]3D plane with smallest perimeter containing all the points associated with a detected surface.

in the object bounding box. We use off-the-shelf detectors, *e.g.,* proposed in [24], to find 2D object bounding boxes and then find the corresponding voxels in 3D to form a clique.

4.4. Relating Voxels in a Camera Ray

V-CRF model enforces that there is only one visible voxel for each ray from a camera. This is enforced by the following energy term.

$$\phi_c(V_{c_r}, L_{c_r}, O) = \begin{cases} 0 & \text{if } \sum_{i \in c} v_i = 1 \\ \infty & \text{otherwise} \end{cases} \quad (9)$$

where c_r is indices of voxels in a single ray.

5. Inference and Learning

In this section, we discuss our inference and learning procedures. We propose an inference method where structured labels $\{\ell_i\}$ and visibility labels $\{v_i\}$ are iteratively updated (Sec. 5.1). The parameters of the model are learned using Structural SVM framework [26] (Sec. 5.2).

5.1. Inference

We find the most probable labelling of L and V under the model by minimizing the energy function Eq. 1. Efficiency of the inference step is a key requirement for us as V-CRF is defined over a voxel space which can be much larger than the number of pixels in the image. We propose an efficient graph-cut based approximate iterative inference procedure that is described below.

In the t^{th} iteration, we estimate the value of the visibility variables V_t from L_{t-1} by finding out the first occupied voxel in each ray from a camera. Given V_t, we solve the energy minimization problem $\text{argmin}_L E(V_t, L, O)$ instead of Eq. 1, and update L_t. This procedure is illustrated in Alg. 1. Note that, by fixing V_t, the energy (Eq. 1) becomes independent of V and can be minimized using graph-cut [1, 14].

1 Initialize V_t, $t = 0$.;
2 Build a V-CRF with unary, pairwise and higher-order potential terms, by fixing V_t.;
3 (Scene understanding) Solve $L_{t+1} = \text{argmin}_L E(V_t, L, O)$ with the graph-cut method;
4 (Updating visibility) From L_{t+1}, update V_{t+1};
5 Go back to Step. 2.;

Algorithm 1: Iterative inference process for L and V.

5.2. Learning

The energy function introduced in Sec. 4 is the sum of unary, pairwise, and higher-order potentials. Since the weights $W = (w^u_{\{\cdot\}}, w^{pw}_{\{\cdot\}}, w^g_{\{\cdot\}})$ are linear in the energy function, we formulate the training problem as a structural SVM problem [26].

Specifically, given N RGB-D images $(I^n, D^n)_{n \in 1 \sim N}$ and their corresponding ground truth labels L^n, we solve the following optimization problem:

$$\min_{W, \xi \geq 0} W^T W + C \sum_n \xi^n(L) \quad (10)$$
$$s.t. \quad \xi^n(L) = max_L(\Delta(L; L^n) + E(L^n|W) - E(L|W))$$

where C controls the relative weights of the sum of the violated terms $\{\xi^n(L)\}$ with respect to the regularization term. $\Delta(L; L^n)$ is the loss function for the visible voxels according to its structured label that guarantees larger loss when L is more different from L^n. Note that the loss function can be decomposed into a sum of local losses on individual voxels, and the violated terms can be efficiently inferred by the graph-cut method. Similar to [4], stochastic subgradient decent method is be used to solve Eq. 10.

6. Experiments

We evaluate our framework on two datasets [28, 12].

6.1. Implementation Details

Appearance Term. For the appearance term $P(s_i|O)$ for visible voxels in Sec. 4.1, we incorporate responses of [2] and [27], which are state-of-the-art methods using 2D and 3D features, respectively.

3D Surface Detection. We find groups of voxels composing 3D surfaces using off-the-shelf plane detector [23], which detects a number of planes from point clouds by hough voting in a parameterized space. Different types of parameterized space can be used; in this work, we used Randomzied Hough Voting. Please see [23] for details.

Object Detection. We use pre-trained DPM detector [24] Release 4 [29] to provide detections for higher-order cliques. Among various semantic classes, we used reliable detection results from sofa, chair, and tv/monitors.

Voxel Initialization. To build V-CRF model, the 3D space of interest is divided with voxels having size of $(4cm)^3$ for testing. For training, voxels are divided into $(8cm)^3$ for efficiency. Since the difference in resolution is small we could use the relationships learned from the training set on the test set with reasonable results. Initialization is performed by assigning appearance likelihood for each point in a cloud to a voxel. Note that more than one point from a cloud can be associated with a single voxel; for simplicity, we used averaged appearance likelihood responses from multiple points for Eq. 2.

6.2. NYU DEPTH Ver. 1

We first evaluate our framework on the NYU Depth dataset Ver. 1 (NYUD-V1) [28], where pixelwise annotations are available for 13 classes. The dataset contains 2347 images from 64 different indoor environments. We used the same 10 random splits of training and testing set used in

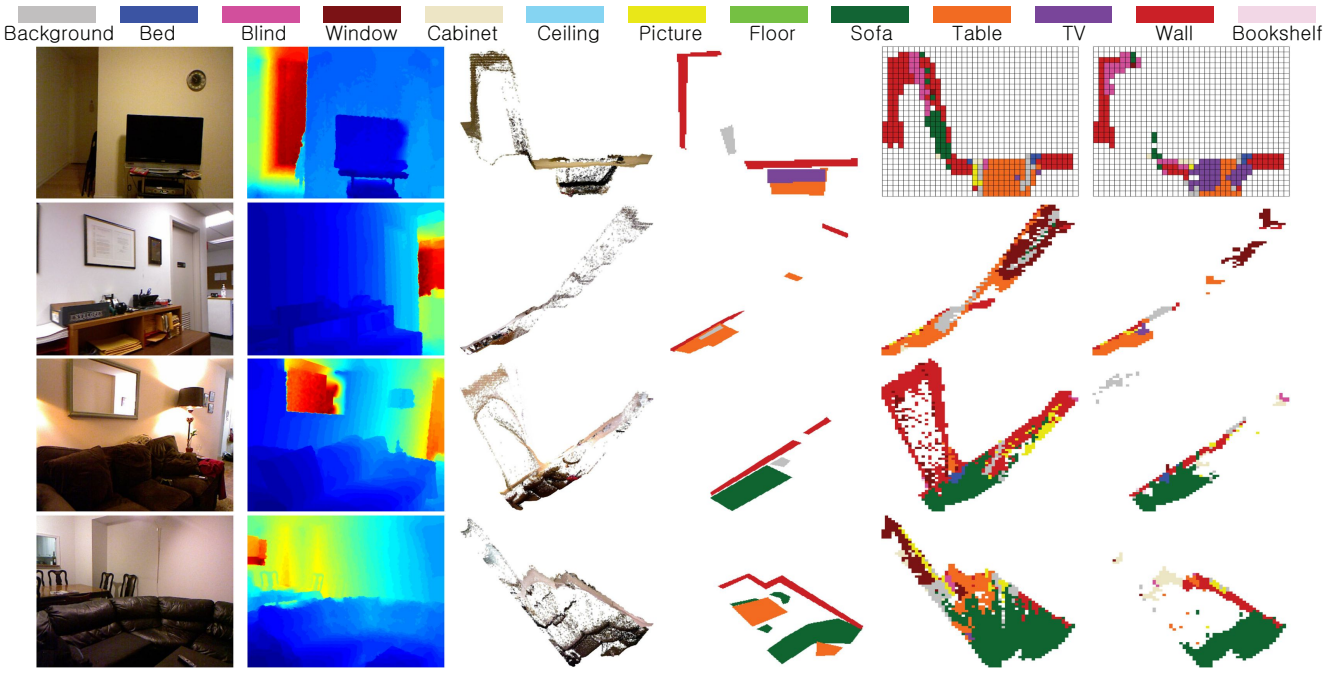


Figure 5: Four typical examples show that the 3D geometry of the scene is successfully estimated by solving V-CRF model. Given (a) a RGB image and (b) a depth map, (c) reconstructed 3D geometry (top view) suffers from noise and may not produce realistic scene understanding results. (d) Annotated top-view structured labels (occupied or not, semantic labels). (e) Results from other methods, *e.g.,* [27]. (f) V-CRF achieves labeling and reconstruction results that are closer to the ground truth than [27]. For instance, the empty space (hall) in the first image is successfully constructed with V-CRF, whereas [27] fails. Even with the error due to reflection of the mirror on the third example, V-CRF is capable of reconstructing realistic scenes along with accurate semantic labeling results. We draw a grid to visualize voxels from top view for the first example only.

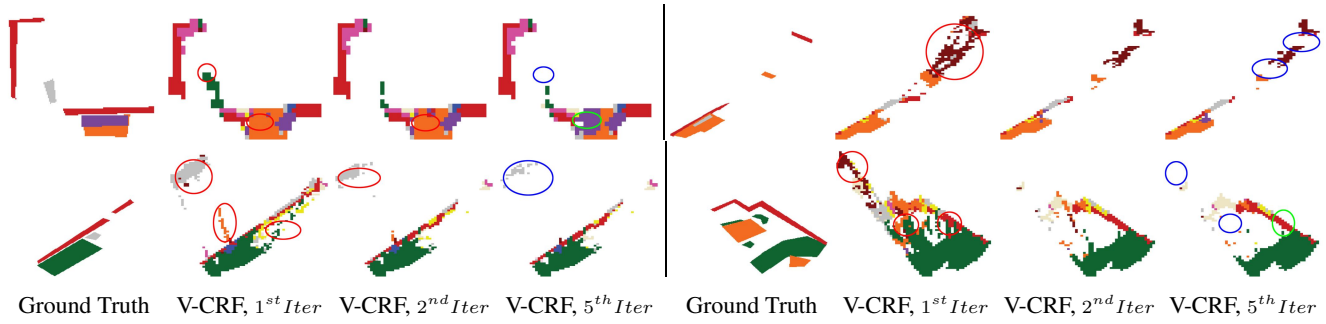


Figure 6: Examples show that the iterative inference process improves scene understanding (Sec. 5.1). We visualize joint geometric and semantic scene understanding results from its top view. (1,5th column) The annotated top-view ground truth labeling. (2,6th column) V-CRF results after 1st iteration, (3,7th column) after 2nd iteration, (4,8th column) and after 5th iteration. Clearly, as the number of iterations increases, both geometry estimation accuracy and semantic labeling accuracy are improved, as highlighted with blue circles and green circles, respectively. Red circles highlight areas that have been better reconstructed across iterations.

[27] and compared the performance against [27, 2] as well as variants of our model.

The proposed framework solves semantic and geometric scene understanding jointly. Yet, evaluating the accuracy in 3D is not an easy task because of the lack of ground truth geometry due to the noisy depth data and incomplete region of occluded part. We propose two metrics for evaluating accuracy - one based on a top view analysis and one evaluating only the visible voxels.

Metric 1: Top-view analysis. Similar to [16, 7], top-view analysis can help understand the results of the framework and perceive the free space of the scene as well as the occluded regions. While [28] only provides frontal view annotation, we annotated top-view ground truth labels as depicted in Fig. 5 (d), where free space and object occupancy as well as semantic labeling can be evaluated. We propose

	[2]	[27]	U	U+PW	U+PW+G
Geo	76.6	80.0	85.8	87.4	**87.7**
S,1^{st}	19.1	38.3	40.4	41.1	41.6
S,5^{th}	-	-	41.7	43.7	**44.6**

Table 1: Top-view analysis for NYUD-V1. Different columns are for benchmark methods [2, 27] and different components of our model (U:only unary terms, $U+PW$:unary and pairwise, and $U+PW+G$:full model). Geometric accuracies are reported in the first line. Semantic accuracies (2nd and 3rd lines) is measured after 1st and 5th iterations of inference steps. By having more components, our model gradually improves the accuracy, and iterative procedure further helps. Full model V-CRF achieves the state-of-the-art performance of 87.7% and 44.6% for geometric and semantic estimation accuracy, respectively. The typical examples can be found from Fig. 5.

a novel user interface for efficient top-view annotation [30]. Specifically, 1320 images from 54 different scenes are annotated[4], where the labeling space is {*empty*, *bed*, *blind*, *window*, *cabinet*, *picture*, *sofa*, *table*, *television*, *wall*, *bookshelf*, *other objects*}.

Fig. 5 shows typical examples of scene understanding from single view RGB-D images from the proposed V-CRF. Note that our model improves reconstruction errors in depth map as well as semantic understanding against a benchmark method, *e.g.,* [27]. Fig. 6 illustrates the results for different number of iterations; we observe that most of minor errors are corrected in the first iteration, whereas more severe errors are gradually improved over the iterative inference process.

Quantitative results can be found in Table. 1. The free space estimation accuracy is measured by evaluating binary classification results for occupancy (empty/non-empty) from the top-view of the image (Table. 1, 1st line 'Geo'). The occupancy map from the top-view is an important measure and relevant to a number of applications such as robotics. Compared to [27], our method achieves 7.7% overall improvement. Especially, our unary potential gives 5.8% boost over [27] (pairwise potentials and higher-order potentials further improves the accuracy). Note that our unary potential not only models appearance but also models geometric properties of the occupancy. This allows V-CRF model to achieve better performance even with the simple unary model, compared to [27].

We also observe that semantic labeling accuracy is simultaneously improved in Table. 1, the second and the third lines. Here, we analyze *i)* the effect of different energy terms and *ii)* the effect of the iterative procedure. It shows that our full model with larger number of iterations achieves the state-of-the-art average accuracy of 44.6%, which is 6.3% higher than the projected results from [27]. The typical examples can be found in Fig. 5.

Metric 2: Visible voxels. The accuracy of semantic labels for visible voxels is presented in Table. 2. For this evaluation, we used the original labeling over 2347 images with 13 classes annotations [28]. Compared to the state-of-the-art method [27], our full model achieves 4.5% improvement in average recall rate.

[4]Bedroom, kitchen, livingroom, office scenes are annotated.

	[2]	[27]	U	U+PW	U+PW+G
S,5^{th}	42.8	65.5[5]	69.5	69.9	**70.0**

Table 2: Visible voxel analysis for NYUD-V1. Semantic labeling accuracies of the visible voxel, after 5th iteration of the inference. Full V-CRF (U+PW+G) model achives the best performance compared against [2, 27] and variants of our models (U, U+PW).

	[2]	[27]	U	U+PW	U+PW+G
Geo (top)	73.2	78.2	85.0	**87.1**	**87.1**
S,5^{th} (top)	16.3	23.9	31.0	32.9	**33.6**
S,5^{th} (visible)	38.6	53.7	61.3	63.2	**63.4**

Table 3: The evaluation results on NYUD-V2. The first two lines are for top-view analysis, and the third line is the analysis for visible voxels. The accuracy is worse than that of NYUD-V1 due to diversity in the dataset. Still, our methods achieves the highest accuracy for both geometry estimation and semantic labeling tasks.

6.3. NYU DEPTH Ver. 2

The NYU Depth dataset Ver. 2 (NYUD-V2) [12] contains 1449 RGB-D images collected from 464 different indoor scenes having more diversity than NYUD-V1. We split the data into 10 random sets for training and testing and evaluate performance for top-view labeling, and for visible voxels, as in NYUD-V1. The experimental results show that the accuracy is worse than that of NYUD-V1 due to diversity of the dataset, but still full V-CRF model achieves the best performance compared against [2, 27].

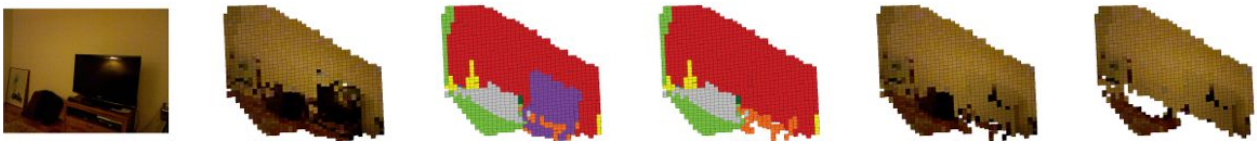
Metric 1: Top-view analysis. We annotated top-view with the same labeling space used for NYUD-V1. This consist of 762 images from 320 different indoor scenes. The first and the second rows in Table. 3 show the performance of geometry estimation and semantic labeling from the top view, respectively. Our model achieves the best performance in both semantic and geometric accuracy (9.7% and 8.9% improvement over [27]).

Metric 2: Visible voxel. The third row in Table. 3 shows semantic labeling accuracy for visible voxels. Our full model achieves 63.4% (9.7% improvement over [27]).

6.4. Augmented Reality: Object Removal.

One interesting application is an augmented reality scenario where one can remove or move around objects. This is not possible in most of conventional augmented reality methods [31] where one can put a new object in a scene

[5]This number is equivalent to 2D semantic labeling accuracy 76.1% reported in super-pixel-based evaluation [27]. 2D super-pixel-based evaluation cannot address the accuracy of 3D scene labeling and tends to penalize less for inaccurate labeling for distant 3D regions.



(a) (b) (c) (d) (e) (f)

Figure 7: (a) RGB image. (b) 3D Reconstruction with V-CRF. (c) Semantic labeling results. (d) Associated voxels for detected 'television' is removed. Note that the region behind the TV is labeled as wall by modeling energy terms for pairwise voxels and planes. (e) As an augmented reality application, TV is removed and voxels are colored with the same color as the adjacent voxels with label 'wall'. (f) All the foreground objects are removed. The occluded region behind the bag is not well reconstructed since there was no plane found behind it. More examples can be found at [30].

but cannot remove the existing objects, since it requires a model to *i)* identify semantic and geometric properties of the objects, *ii)* estimate occluded region behind the object. In contrast, V-CRF model can solve this problem. Fig. 7 shows that our model can be used to detect, say, a TV set in the scene and remove it. Note that the occluded region behind the TV is reconstructed using pairwise relationships among voxels as discussed in Sec. 4.2 and the concept of 3D surface prior as introduced in Sec. 4.3.

7. Conclusion

We have presented the V-CRF model for jointly solving the problem of semantic scene understanding and geometry estimation that incorporates 3D geometric and semantic relationships between scene elements in a coherent fashion. Our formulation generalizes many existing 3D scene understanding frameworks. Experimental results indicate that our method quantitatively and qualitatively achieves good performance on the challenging NYU Depth dataset (Version 1 and 2).

Acknowledgement

We acknowledge the support of NSF CAREER grant #1054127, NSF CPS grant #0931474 and a KLA-Tencor Fellowship. We thanks Dr. Wongun Choi for his valuable comments and constructive suggestions.

References

[1] P. Kohli, L. Ladicky, and P. H. S. Torr, "Robust higher order potentials for enforcing label consistency," in *CVPR*, 2008. 1, 4, 5

[2] L. Ladicky, C. Russell, P. Kohli, and P. Torr, "Associative hierarchical crfs for object class image segmentation," in *ICCV*, 2009. 1, 5, 6, 7

[3] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *ECCV*, 2006. 1

[4] B. Kim, M. Sun, P. Kohli, and S. Savarese, "Relating things and stuff by high-order potential modeling," in *ECCV Workshop (HiPot)*, 2012. 1, 4, 5

[5] A. Gupta, A. Efros, and M. Hebert, "Blocks world revisited: Image understanding using qualitative geometry and mechanics," *ECCV*, 2010. 1, 2

[6] D. Hoiem, A. A. Efros, and M. Hebert, "Geometric context from a single image," in *ICCV*, 2005. 1

[7] V. Hedau, D. Hoiem, and D. Forsyth, "Recovering free space of indoor scenes from a single image," in *CVPR*, 2012. 1, 2, 6

[8] W. Choi, Y.-W. Chao, C. Pantofaru, and S. Savarese, "Understanding indoor scenes using 3d geometric phrases," in *CVPR*. 1

[9] S. Y. Bao and S. Savarese, "Semantic structure from motion," in *CVPR*, 2011. 1

[10] H. S. Koppula, A. Anand, T. Joachims, and A. Saxena, "Semantic labeling of 3d point clouds for indoor scenes," in *NIPS*, 2011. 1, 2, 3

[11] D. Munoz, J. A. Bagnell, and M. Hebert, "Co-inference machines for multi-modal scene analysis," in *ECCV*, 2012. 1, 2, 3

[12] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," ECCV, 2012. 1, 2, 3, 5, 7

[13] Microsoft Kinect, http://www.xbox.com/en-US/kinect. 1

[14] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *PAMI*, 2001. 2, 5

[15] Z. Jia, A. Gallagher, A. Saxena, and T. Chen, "3d-based reasoning with blocks, support, and stability," in *CVPR*, 2013. 2

[16] S. Satkin, J. Lin, and M. Hebert, "Data-driven scene understanding from 3d models," 2012. 2, 6

[17] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert, "From 3d scene geometry to human workspace," in *CVPR*, 2011. 2

[18] S. Gupta, P. Arbelaez, and J. Malik, "Perceptual organization and recognition of indoor scenes from rgb-d images," in *CVPR*, 2013. 2

[19] L. Ladickỳ, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. Torr, "Joint optimization for object class segmentation and dense stereo reconstruction," *IJCV*, 2012. 2

[20] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. Sinha, "Object stereo: joint stereo matching and object segmentation," in *CVPR*, 2011. 2

[21] R. Guo and D. Hoiem, "Beyond the line of sight: labeling the underlying surfaces," in *ECCV*, 2012. 2

[22] J. Gonfaus, X. Boix, J. Van De Weijer, A. Bagdanov, J. Serrat, and J. Gonzalez, "Harmony potentials for joint classification and segmentation," in *CVPR*, 2010. 4

[23] D. Borrmann, J. Elseberg, K. Lingemann, and A. Nüchter, "The 3d hough transform for plane detection in point clouds: A review and a new accumulator design," *3D Research*, 2011. 4, 5

[24] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *PAMI*, 2010. 4, 5

[25] L. Ladickỳ, P. Sturgess, K. Alahari, C. Russell, and P. Torr, "What, where and how many? combining object detectors and crfs," *ECCV*, 2010. 4

[26] T. Joachims, T. Finley, and C. Yu, "Cutting-plane training of structural svms," *Machine Learning*, 2009. 5

[27] X. Ren, L. Bo, and D. Fox, "Rgb-(d) scene labeling: Features and algorithms," in *CVPR*, 2012. 5, 6, 7

[28] N. Silberman and R. Fergus, "Indoor scene segmentation using a structured light sensor," in *ICCV Workshop (3DRR)*, 2011. 5, 6, 7

[29] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Discriminatively trained deformable part models, release 4." http://people.cs.uchicago.edu/ pff/latent-release4/. 5

[30] Project page, http://cvgl.stanford.edu/projects/vcrf/. 7, 8

[31] D. Van Krevelen and R. Poelman, "A survey of augmented reality technologies, applications and limitations," *International Journal of Virtual Reality*, 2010. 7