Understanding Indoor Scenes using 3D Geometric Phrases

Wongun Choi[1], Yu-Wei Chao[1], Caroline Pantofaru[2], and Silvio Savarese[1]

[1]University of Michigan, Ann Arbor, MI, USA
[2]Google, Mountain View, CA, USA*
{wgchoi, ywchao, silvio}@umich.edu, cpantofaru@google.com

Abstract

Visual scene understanding is a difficult problem interleaving object detection, geometric reasoning and scene classification. We present a hierarchical scene model for learning and reasoning about complex indoor scenes which is computationally tractable, can be learned from a reasonable amount of training data, and avoids oversimplification. At the core of this approach is the 3D Geometric Phrase Model which captures the semantic and geometric relationships between objects which frequently co-occur in the same 3D spatial configuration. Experiments show that this model effectively explains scene semantics, geometry and object groupings from a single image, while also improving individual object detections.

1. Introduction

Consider the scene in Fig. 1.(a). A scene classifier will tell you, with some uncertainty, that this is a dining room [21, 23, 15, 7]. A layout estimator [12, 16, 27, 2] will tell you, with different uncertainty, how to fit a box to the room. An object detector [17, 4, 8, 29] will tell you, with large uncertainty, that there is a dining table and four chairs. Each algorithm provides important but uncertain and incomplete information. This is because the scene is cluttered with objects which tend to occlude each other: the dining table occludes the chairs, the chairs occlude the dining table; all of these occlude the room layout components (i.e. the walls).

It is clear that truly understanding a scene involves integrating information at multiple levels as well as studying the interactions between scene elements. A scene-object interaction describes the way a scene type (e.g. a dining room or a bedroom) influences objects' presence, and vice versa. An object-layout interaction describes the way the layout (e.g. the 3D configuration of walls, floor and observer's pose) biases the placement of objects in the image, and vice versa. An object-object interaction describes the way objects and their pose affect each other (e.g. a dining table suggests that a set of chairs are to be found around it). Combining predictions at multiple levels into a global estimate can improve each individual prediction. As part of a larger system, understanding a scene semantically and functionally will allow us to make predictions about the presence and locations of unseen objects within the space.

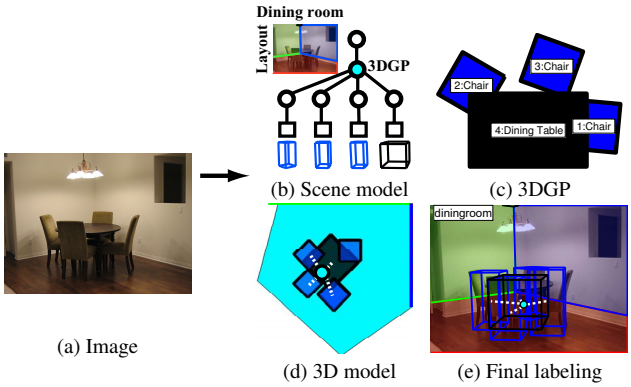


Figure 1. Our unified model combines object detection, layout estimation and scene classification. A single input image (a) is described by a scene model (b), with the scene type and layout at the root, and objects as leaves. The middle nodes are latent *3D Geometric Phrases*, such as (c), describing the 3D relationships among objects (d). Scene understanding means finding the correct parse graph, producing a final labeling (e) of the objects in 3D (bounding cubes), the object groups (dashed white lines), the room layout, and the scene type.

We propose a method that can automatically learn the interactions among scene elements and apply them to the holistic understanding of indoor scenes. This scene interpretation is performed within a hierarchical interaction model and derived from a single image. The model fuses together object detection, layout estimation and scene classification to obtain a unified estimate of the scene composition. The problem is formulated as image parsing in which a parse graph must be constructed for an image as in Fig. 1.(b). At the root of the parse graph is the scene type and layout while the leaves are the individual detections of objects. In between is the core of the system, our novel *3D Geometric Phrases* (3DGP) (Fig. 1.(c)).

A 3DGP encodes geometric and semantic relationships

*This work was done while C. Pantofaru was at Willow Garage, Inc.

between groups of objects which frequently co-occur in spatially consistent configurations. As opposed to previous approaches such as [5, 24], the 3DGP is defined using 3D spatial information, making the model rotation and viewpoint invariant. Grouping objects together provides contextual support to boost weak object detections, such as the chair that is occluded by the dining table.

Training this model involves both discovering a set of 3DGPs and estimating the parameters of the model. We present a new learning scheme which discovers 3DGPs in an unsupervised manner, avoiding expensive and ambiguous manual annotation. This allows us to extract a few useful sets of GPs among exponentially many possible configurations. Once a set of 3DGPs is selected, the model parameters can be learned in a max-margin framework. Given the interdependency between the 3DGPs and the model parameters, the learning process is performed iteratively (Sec. 5).

To explain a new image, a parse graph must estimate the scene semantics, layout, objects and 3DGPs, making the space of possible graphs quite large and of variable dimension. To efficiently search this space during inference, we present a novel combination of bottom-up clustering with top-down Reversible Jump Markov Chain Monte Carlo (RJ-MCMC) sampling (Sec. 4).

As a result of the rich contextual relationships captured by our model, it can provide scene interpretations from a single image in which i) objects and space interact in a physically valid way, ii) objects occur in an appropriate scene type, iii) the object set is self-consistent and iv) configurations of objects are automatically discovered (Fig. 1.(d,e)). We quantitatively evaluate our model on a novel challenging dataset, the *indoor-scene-object* dataset. Experiments show our hierarchical scene model constructed upon 3DGPs improves object detection, layout estimation and semantic classification accuracy in challenging scenarios which include occlusions, clutter and intra-class variation.

2. Related Work

Image understanding has been explored on many levels, including object detection, scene classification and geometry estimation.

The performance of generic object recognition has improved recently thanks to the introduction of more powerful feature representations [20, 4]. Felzenszwalb *et al*. proposed a deformable part model (DPM) composed of multiple HoG components [8] which showed promising performance for single objects. To improve detection robustness, the interactions between objects can be modeled. Category-specific 2D spatial interactions have been modeled via contextual features by Desai *et al*. [5], whereas Sadeghi *et al*. [24] modeled groups of objects as *visual phrases* in 2D image space that were determined by a domain expert. Li *et al*. [18] identified a set of useful *visual phrases* from a training set using only 2D spatial consistency. Improving upon these, Desai *et al*. [5] proposed a method that can encode detailed pose relationships between co-appearing objects in 2D image space. In contrast to these approaches, our 3DGPs are capable of encoding both 3D geometric and contextual interactions among objects and can be automatically learned from training data.

Researchers have also looked at the geometric configuration of a scene. Geiger *et al*. [10] related traffic patterns and vanishing points in 3D. To obtain physically consistent representations, Gupta *et al*. [11] incorporated the concept of physical gravity and reasoned about object supports. Bao *et al*. [2, 1] utilized geometric relationship to help object detection and scene structure estimation. Several methods attempted to specifically solve indoor layout estimation [12, 13, 27, 30, 22, 26, 25]. Hedau *et al*. proposed a formulation using a cubic room representation [12] and showed that layout estimation can improve object detection [13]. This initial attempt demonstrated promising results, however experiments were limited to a single object type (bed) and a single room type (bedroom). Other methods [16, 30] proposed to improve layout estimation by analyzing the consistency between layout and the geometric properties of objects without accounting for the specific categorical nature of such objects. Fouhey *et al*. [9] incorporated human pose estimation into indoor scene layout understanding. However, [9] does not capture relationships between objects or between an object and the scene type.

A body of work has focused on classifying images into semantic scene categories [7, 21, 23, 15]. Li *et al*. [19] proposed an approach called *object bank* to model the correlation between objects and scene by encoding object detection responses as features in a SPM and predicting the scene type. They did not, however, explicitly reason about the relationship between the scene and its constituent objects, nor the geometric correlation among objects. Recently, Pandey *et al*. [21] used a latent DPM model to capture the spatial configuration of objects in a scene type. This spatial representation is 2D image-based, which makes it sensitive to viewpoint variations. In our approach, we instead define the spatial relationships among objects in 3D, making them invariant to viewpoint and scale transformation. Finally, the latent DPM model assumes that the number of objects per scene is fixed, whereas our scene model allows an arbitrary number of 3DGPs per scene.

3. Scene Model using 3D Geometric Phrases

The high-level goal of our system is to take a single image of an indoor scene and classify its scene semantics (such as room type), spatial layout, constituent objects and object relationships in a unified manner. We begin by describing the unified scene model which facilitates this process.

Image parsing is formulated as an energy maximization

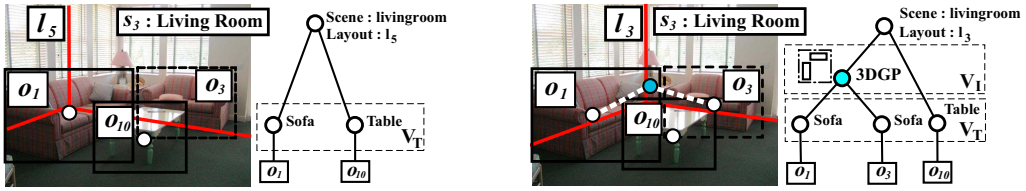


Figure 2. Two possible parse graph hypotheses for an image - on the left an incomplete interpretation (where no 3DGP is used) and on the right a complete interpretation (where a 3DGP is used). The root node S describes the scene type s_1, s_3 (bedroom or livingroom) and layout hypothesis l_3, l_5 (red lines), while other white and skyblue round nodes represent objects and 3DGPs, respectively. The square nodes $(o_1, ..., o_{10})$ are detection hypotheses obtained by object detectors such as [8] (black boxes). Weak detection hypotheses (dashed boxes) may not be properly identified in isolation (left). A 3DGP, such that indicated by the skyblue node, can help transfer contextual information from the left sofa (strong detections denoted by solid boxes) to the right sofa.

problem (Sec. 3.1), which attempts to identify the parse graph that best fits the image observations. At the core of this formulation is our novel *3D Geometric Phrase* (3DGP), which is the key ingredient in parse graph construction (Sec. 3.2). The 3DGP model facilitates the transfer of contextual information from a strong object hypothesis to a weaker one when the configuration of the two objects agrees with a learned geometric phrase (Fig. 2 right).

Our scene model $\mathcal{M} = (\Pi, \theta)$ contains two elements; the 3DGPs $\Pi = \{\pi_1, ..., \pi_N\}$ and the associated parameters θ. A single 3DGP π_i defines a group of object types (e.g. sofa, chair, table, etc.) and their 3D spatial configuration, as in Fig. 1(d). Unlike [30], which requires a training set of hand crafted composition rules and learns only the rule parameters, our method automatically learns the set of 3DGPs from training data via our novel training algorithm (Sec. 5). The model parameter θ includes the observation weights α, β, γ, the semantic and geometric context model weights η, ν, the pair-wise interaction model μ, and the parameters λ associated with the 3DGP (see eq. 1).

We define a parse graph $G = \{S, \mathbb{V}\}$ as a collection of nodes describing geometric and semantic properties of the scene. $S = (C, H)$ is the root node containing the scene semantic class variable C and layout of the room H, and $\mathbb{V} = \{V_1, ..., V_n\}$ represents the set of non-root nodes. An individual V_i specifies an object detection hypothesis or a 3DGP hypothesis, as shown in Fig. 2. We represent an image observation $I = \{O_s, O_l, O_o\}$ as a set of hypotheses with associated confidence values as follows. $O_o = \{o_1, ..., o_n\}$ are object detection hypotheses, $O_l = \{l_1, ..., l_m\}$ are layout hypotheses and $O_s = \{s_1, ..., s_k\}$ are scene types (Sec. 3.3).

Given an image I and scene model $\mathcal{M}$, our goal is to identify the parse graph $G = \{S, \mathbb{V}\}$ that best fits the image. A graph is selected by i) choosing a scene type among the hypotheses O_s, ii) choosing the scene layout from the layout hypotheses O_l, iii) selecting positive detections (shown as o_1, o_3, and o_{10} in Fig. 2) among the detection hypotheses O_o, and iv) selecting compatible 3DGPs (Sec. 4).

3.1. Energy Model

Image parsing is formulated as an energy maximization problem. Let $\mathbb{V}_T$ be the set of nodes associated with a set of detection hypotheses (objects) and $\mathbb{V}_I$ be the set of nodes corresponding to 3DGP hypotheses, with $\mathbb{V} = \mathbb{V}_T \cup \mathbb{V}_I$. Then, the energy of parse graph G given an image I is:

$$
\begin{aligned}
E_{\Pi,\theta}(G, I) = & \underbrace{\alpha^\top \phi(C, O_s)}_{\text{scene observation}} + \underbrace{\beta^\top \phi(H, O_l)}_{\text{layout observation}} + \underbrace{\sum_{V \in \mathbb{V}_T} \gamma^\top \phi(V, O_o)}_{\text{object observation}} \\
& + \underbrace{\sum_{V \in \mathbb{V}_T} \eta^\top \psi(V, C)}_{\text{object-scene}} + \underbrace{\sum_{V \in \mathbb{V}_T} \nu^\top \psi(V, H)}_{\text{object-layout}} \\
& + \underbrace{\sum_{V, W \in \mathbb{V}_T} \mu^\top \varphi(V, W)}_{\text{object overlap}} + \underbrace{\sum_{V \in \mathbb{V}_I} \lambda^\top \varphi(V, Ch(V))}_{\text{3DGP}} \qquad (1)
\end{aligned}
$$

where $\phi(\cdot)$ are unary observation features for semantic scene type, layout estimation and object detection hypotheses, $\psi(\cdot)$ are contextual features that encode the compatibility between semantic scene type and objects, and the geometric context between layout and objects, and $\varphi(\cdot)$ are the interaction features that describe the pairwise interaction between two objects and the compatibility of a 3DGP hypothesis. $Ch(V)$ is the set of child nodes of V.

Observation Features: The observation features ϕ and corresponding model parameters α, β, γ capture the compatibility of a scene type, layout and object hypothesis with the image, respectively. For instance, one can use the spatial pyramid matching (SPM) classifier [15] to estimate the scene type, the indoor layout estimator [12] for determining layout and Deformable Part Model (DPM) [8] for detecting objects. In practice, rather than learning the parameters for the feature vectors of the observation model, we use the confidence values given by SPM [15] for scene classification, from [12] for layout estimation, and from the DPM [8] for object detection. To allow bias between different types of objects, a constant 1 is appended to the detection confidence, making the feature two-dimensional as in [5] [1].

Geometric and Semantic Context Features: The geometric and semantic context features ψ encode the compatibility between object and scene layout, and object and scene

[1] This representation ensures that all observation features associated with a detection have values distributed from negative to positive, make graphs with different numbers of objects are comparable.

type. As discussed in Sec. 3.3, a scene layout hypothesis l_i is expressed using a 3D box representation and an object detection hypothesis p_i is expressed using a 3D cuboid representation. The compatibility between an object and the scene layout ($\nu^\top \psi(V, H)$) is computed by measuring to what degree an object penetrates into a wall. For each wall, we measure the object-wall penetration by identifying which (if any) of the object cuboid bottom corners intersects with the wall and computing the (discretized) distance to the wall surface. The distance is 0 if none of the corners penetrate a wall. The object-scene type compatibility, $\eta^\top \psi(V, C)$, is defined by the object and scene-type co-occurrence probability.

Interaction Features: The interaction features φ are composed of an object overlap feature $\mu^\top \varphi(V, W)$ and a 3DGP feature $\lambda^\top \varphi(V, Ch(V))$. We encode the overlap feature $\varphi(V, W)$ as the amount of object overlap. In the 2D image plane, the overlap feature is $A(V \cap W)/A(V) + A(V \cap W)/A(W)$ where $A(\cdot)$ is the area function. This feature enables the model to learn inhibitory overlapping constraints similar to traditional non-maximum suppression [4].

3.2. The 3D Geometric Phrase Model

The 3DGP feature allows the model to favor a group of objects that are commonly seen in a specific 3D spatial configuration, e.g. a coffee table in front of a sofa. The preference for these configurations is encoded in the 3DGP model by a deformation cost and view-dependent biases (eq. 2).

Given a 3DGP node V, the spatial deformation (dx_i, dz_i) of a constituent object is a function of the difference between the object instance location o_i and the learned expected location c_i with respect to the centroid of the 3DGP (the mean location of all constituent objects m_V). Similarly, the angular deformation da_i is computed as the difference between the object instance orientation a_i and the learned expected orientation α_i with respect to the orientation of the 3DGP (the direction from the first to the second object, a_V). Additionally, 8 view-point dependent biases for each 3DGP encode the amount of occlusion expected from different view-points. Given a 3DGP node V and the associated model π_k, the potential function can be written as follows:

$$\lambda_k^\top \varphi_k(V, Ch(V)) = \sum_{p \in \mathcal{P}} b_k^p \, \mathbb{I}(a_V = p) - \sum_{i \in Ch(V)} d_k^{i\top} \varphi_k^d(dx_i, dz_i, da_i) \tag{2}$$

where $\lambda_k = \{b_k, d_k\}$, $\mathcal{P}$ is the space of discretized orientations of the 3DGP and $\varphi_d(dx_i, dz_i, da_i) = \{dx_i^2, dz_i^2, da_i^2\}$. The parameters d_k^i for the deformation cost φ_k^i penalize configurations in which an object is too far from the anchor. The view-dependent bias b_k^p "rewards" spatial configurations and occlusions that are consistent with the camera location. The amount of occlusion and overlap among objects in a 3DGP depends on the view point; the view-dependent bias encodes occlusion and overlap reasoning. Notice that the spatial relationships among objects in a 3DGP encodes their relative positions in 3D space, so the 3DGP model is rotation and view-point invariant. Previous work which encoded the 2D spatial relationships between objects [24, 18, 5] required large numbers of training images to capture the appearance of co-occuring objects. On the other hand, our 3DGP requires only a few training examples since it has only a few model parameters thanks to the invariance property.[2]

3.3. Objects in 3D Space

We propose to represent objects in 3D space instead of 2D image space. The advantages of encoding objects in 3D are numerous. In 3D, we can encode geometric relationships between objects in a natural way (e.g. 3D euclidean distance) as well as encode constraints between objects and the space (e.g. objects cannot penetrate walls or floors). To keep our model tractable, we represent an object by its 3D bounding cuboid, which requires only 7 parameters (3 centroid coordinates, 3 dimension sizes and 1 orientation.) Each object class is associated to a different prototypical bounding cuboid which we call the cuboid model (which was acquired from the commercial website www.ikea.com similarly to [22].) Unlike [13], we do not assume that objects' faces are parallel to the wall orientation, making our model more general.

Similarly to [12, 16, 27], we represent the indoor space by the 3D layout of 5 orthogonal faces (floor, ceiling, left, center, and right wall), as in Fig. 1(e). Given an image, the intrinsic camera parameters and rotation with respect to the room space (K, R) are estimated using the three orthogonal vanishing points [12]. For each set of layout faces, we obtain the corresponding 3D layout by back-projecting the intersecting corners of walls.

An object's cuboid can be estimated from a single image given a set of known object cuboid models and an object detector that estimates the 2D bounding box and pose (Sec. 6). From the cuboid model of the identified object, we can uniquely identify the 3D cuboid centroid O that best fits the 2D bounding box detection o and pose p by solving for

$$\hat{O} = \underset{O}{\operatorname{argmin}} \, ||o - P(O, p, K, R)||_2^2 \tag{3}$$

where $P(\cdot)$ is a projection function that projects 3D cuboid O and generates a bounding box in the image plane. The above optimization is quickly solved with a simplex search method [14]. In order to obtain robust 3D localization of each object and disambiguate the size of the room space given a layout hypothesis, we estimate the camera height (ground plane location) by assuming all objects are lying on a common ground plane. More details are discussed in the supplementary material.

[2] Although the view-dependent biases are not view-point invariant, there are still only a few parameters (8 views per 3DGP).

4. Inference

In our formulation, performing inference is equivalent to finding the best parse graph specifying the scene type C, layout estimation H, positive object hypotheses $V \in \mathbb{V}_T$ and 3DGP hypotheses $V \in \mathbb{V}_I$.

$$\hat{G} = \underset{G}{\operatorname{argmax}} E_{\Pi,\theta}(G, I) \tag{4}$$

Finding the optimal configuration that maximizes the energy function requires exponential time. To make this problem tractable, we introduce a novel bottom-up and top-down compositional inference scheme. Inference is performed for each scene type separately, so scene type is considered given in the remainder of this section.

Bottom-up: During bottom-up clustering, the algorithm finds all candidate 3DGP nodes $\mathbb{V}_{cand} = \mathbb{V}_T \cup \mathbb{V}_I$ given detection hypothesis O_o (Fig. 3 top). The procedure starts by assigning one node V_t to each detection hypothesis o_t, creating a set of candidate terminal nodes (leaves) $\mathbb{V}_T = \{\mathbb{V}_T^1, ..., \mathbb{V}_T^{K_o}\}$, where K_o is the number of object categories. By searching over all combinations of objects in $\mathbb{V}_T$, a set of 3DGP nodes, $\mathbb{V}_I = \{\mathbb{V}_I^1, ..., \mathbb{V}_I^{K_{GP}}\}$, is formed, where K_{GP} denotes the cardinality of the learned 3DGP model Π given by the training procedure (Sec. 5). A 3DGP node V_i is considered valid if it matches the spatial configuration of a learned 3DGP model π_k. Regularization is performed by measuring the energy gain obtained by including V_i in the parse graph. To illustrate, suppose we have a parse graph G that contains the constituent objects of V_i but not V_i itself. If a new parse graph $G' \leftarrow G \cup V_i$ has higher energy $0 < E_{\Pi,\theta}(G', I) - E_{\Pi,\theta}(G, I) = \lambda_k^\top \varphi_k(V_i, Ch(V_i))$, then V_i is considered as a valid candidate. In other words, let π_k define the 3DGP model shown in Fig. 4(c). To find candidates $\mathbb{V}_I^k$ for π_k, we search over all possible configurations of selecting one terminal node among the sofa hypotheses $\mathbb{V}_T^{sofa}$ and one among the table hypotheses $\mathbb{V}_T^{table}$. Only candidates that satisfy the regularity criteria are accepted as valid. In practice, this bottom-up search can be performed very efficiently (less than a minute per image) since there are typically few detection hypotheses per object type.

Top-down: Given all possible sets of nodes $\mathbb{V}_{cand}$, the optimal parse graph G is found via Reversible Jump Markov Chain Monte Carlo (RJ-MCMC) sampling (Fig. 3 bottom). To efficiently explore the space of parse graphs, we propose 4 reversible jump moves, *layout selection*, *add*, *delete* and *switch*. Starting from an initial parse graph G_0, the RJ-MCMC sampling draws a new parse graph by sampling a random jump move, and the new sample is either accepted or rejected following Metropolis-Hasting rule. After N iterations, the graph that maximizes the energy function $\operatorname{argmax}_G E(G, I)$ is selected as the solution. The initial parse graph is obtained by 1) selecting the layout with highest observation likelihood [12] and 2) greedily adding

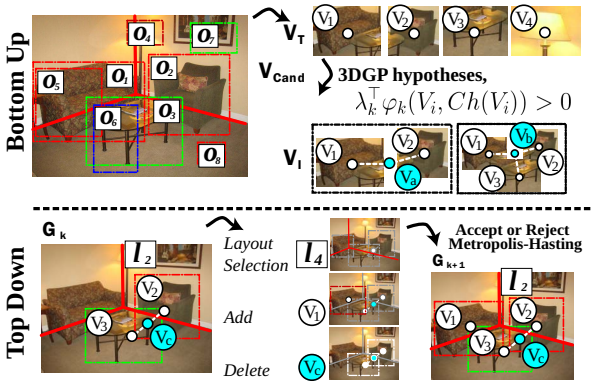


Figure 3. **Bottom-up:** Candidate objects $\mathbb{V}_T$ and 3DGP nodes $\mathbb{V}_I$ are vetted by measuring spatial regularity. Red, green and blue boxes indicate sofas, tables and chairs. Black boxes are candidate 3DGP nodes. **Top-down:** the Markov chain is defined by 3 RJ-MCMC moves on the parse graph G_k. Given G_k, a new G' is proposed via one move and acceptance to become G_{k+1} is decided using the Metropolis-Hasting rule. Moves are shown in the bottom-right subfigures. Red and white dotted boxes are new and removed hypotheses, respectively.

object hypotheses that most improve the energy, similarly to [5]. The RJ-MCMC jump moves used with a parse graph at inference step k are defined as follows.

Layout selection: This move generates a new parse graph G_{k+1} by changing the layout hypothesis. Among $|L|$ possible layout hypotheses (given by [12]), one is randomly drawn with probability $exp(l_k)/\sum_i^{|L|} exp(l_i)$, where l_k is the score of the k^{th} hypothesis.

Add: This move adds a new 3DGP or object node from $V_i \in \mathbb{V}_{cand} \setminus G_k$ into G_{k+1}. To improve the odds of picking a valid detection, a node is sampled with probability $exp(s_i)/\sum_j^{|\mathbb{V}_{cand} \setminus G_k|} exp(s_j)$, where s_i is the aggregated detection score of all children. For example, in Fig. 3(bottom), s_i of V_c is the sum of the sofa and table scores.

Delete: This move removes an existing node $V_i \in G_k$ to generate a new graph G_{k+1}. Like the *Add* move, a node is selected with probability $exp(-s_i)/\sum_j^{|G_k|} exp(-s_j)$.

5. Training

Given input data $x = (O_s, O_l, O_o)$ with labels $y = (C, H, V_T)$ per image, we have two objectives during model training: i) learn the set of 3DGP models Π and ii) learn the corresponding model weights θ. Since the model parameters and 3DGPs are interdependent (e.g. the number of model parameters increases with the number of GPs), we propose an iterative learning procedure. In the first round, a set of 3DGPs is generated by a propose-and-match scheme. Given Π, the model parameters θ are learned using a latent max-margin formulation. This formulation accommodates the uncertainty in associating an image to a parse graph G similarly to [8, 28]; i.e. given a label y, the root node and terminal nodes of G can be uniquely identified, but the

3DGP nodes in the middle are hidden.

Generating Π: This step learns a set of 3DGPs, Π, which captures object groups that commonly appear in the training set in consistent 3D spatial configurations. Given an image, we generate all possible 3DGPs from the ground truth annotations $\{y\}$. The consistency of each 3DGP π_k is evaluated by matching it with ground truth object configurations in other training images. We say that a 3DGP is matched if $\lambda_k^\top \varphi_k(V, Ch(V)) > th$ (see Sec. 4). A 3DGP model π_k is added to Π if it is matched more than K times. This scheme is both simple and effective. To avoid redundancy, agglomerative clustering is performed over the proposed 3DGP candidates. Exploring all of the training images results in an over-complete set Π that is passed to the parameter learning step.

Learning θ and pruning Π: Given a set of 3DGPs Π, the model parameters are learned by iterative *latent completion* and *max-margin* learning. In latent completion, the most compatible parse graph G is found for an image with ground truth labels y by finding compatible 3DGP nodes V_I. This maximizes the energy over the latent variable (the 3DGP nodes), $\hat{h}_i$, given an image and label (x_i, y_i).

$$\hat{h}_i = \underset{h}{\operatorname{argmax}}\, E_{\Pi,\theta}(x_i, y_i, h) \tag{5}$$

After latent completion, the 3DGP models which are not matched with a sufficient number (< 5) of training examples are removed, keeping the 3DGP set compact and ensuring there are sufficient positive examples for max-margin learning. Given all triplets of $(x_i, y_i, \hat{h}_i)$, we use the cutting plane method [5] to train the associated model parameter θ by solving the following optimization problem.

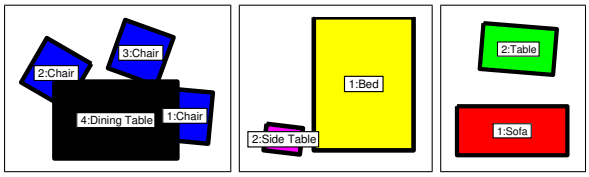
$$\min_{\theta,\xi} \frac{1}{2}\|\theta\|^2 + C\sum_i \xi^i$$
$$s.t. \quad \max_h E_{\Pi,\theta}(x_i, y, h) - E_{\Pi,\theta}(x_i, y_i, \hat{h}_i) \le \xi^i - \delta(y, y_i),\ \forall i, y \tag{6}$$

where C is a hyper parameter in an SVM and ξ^i are slack variables. The loss contains three components, $\delta(y, y_i) = \delta_s(C, C_i) + \delta_l(H, H_i) + \delta_d(V_T, V_{Ti})$. The scene classification $\delta_s(C, C_i)$ and detection $\delta_d(V_T, V_{Ti})$ losses are defined using hinge loss. We use the layout estimation loss proposed by [12] to model the layout estimation loss $\delta_l(H, H_i)$. The process of generating Π and learning the associated model parameters θ is repeated until convergence.

Using the learning set introduced in Sec. 6, the method discovers 163 3DGPs after the initial generation of Π and retains 30 after agglomerative clustering. After 4 iterations of pruning and parameter learning, our method retains 10 3DGPs. Fig. 4 shows selected examples of learned 3DGPs (the complete set is presented in supplementary material.)

6. Experimental Results

Datasets: To validate our proposed method, we collected a new dataset that we call the *indoor-scene-object* dataset,



(a) (b) (c)

Figure 4. Examples of learned 3DGPs. The object class (in color) and the position and orientation of each object is shown. Note that our learning algorithm learns spatially meaningful structures without supervision.

which we contribute to the community. The *indoor-scene-object* dataset includes 963 images. Although there exist datasets for layout estimation evaluation [12], object detection [6] and scene classification [23] in isolation, there is no dataset on which we can evaluate all the three problems simultaneously. The *indoor-scene-object* dataset includes three scene types: living room, bedroom, and dining room, with $\sim$300 images per room type. Each image contains a variable number of objects. We define 6 categories of objects that appear frequently in indoor scenes: sofa, table, chair, bed, dining table and side table. In the following experiments, the dataset is divided into a training set of 180 images per scene, and a test set of the remaining images. Ground truth for the scene types, face layouts, object locations and poses was manually annotated. We used $C = 1$ to train the system without tuning this hyper parameter.

Scene Classifier: The SPM [15] is utilized as a baseline scene classifier, trained via libSVM [3]. The baseline scene classification accuracy is presented in Table 1. The score for each scene type is the observation feature for scene type in our model ($\phi(C, O_s)$). We also train two other state-of-the art scene classifiers SDPM [21] and Object bank [19] and report the accuracy in Table. 1.

Indoor layout estimation: The indoor layout estimator as trained in [12] is used to generate layout hypotheses with confidence scores for O_l and the associated feature $\phi(H, O_l)$. As a sanity check, we also tested our trained model on the indoor UIUC dataset [12]. Our model with 3DGPs increased the original 78.8% pixel accuracy rate [12] to 80.4%. Pixel accuracy is defined as the percentage of pixels on layout faces with correct labels.

To further analyze the layout estimation, we also evaluated per-face estimation accuracy. The per-face accuracy is defined as the intersection-over-union of the estimated and ground-truth faces. Results are reported in Table. 2.

Object detection: The baseline object detector (DPM [8]) was trained using the PASCAL dataset [6] and a new dataset we call the *furniture* dataset containing 3939 images with 5426 objects. The bounding box and azimuth angle (8 view points) of each object were hand labeled. The accuracy of

	Obj. Bank [19]	SDPM [21]	SPM [15]	W/o 3DGP	3DGP
Acc.	76.9 %	86.5 %	80.5 %	85.5 %	**87.7** %

Table 1. Scene classification results using state-of-the-art methods (left-two), the baseline [15] (center) and our model variants (right-two). Our model outperforms all the other methods.

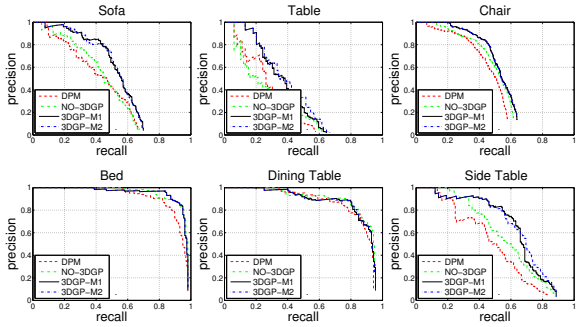


Figure 5. Precision-recall curves for DPMs [8] (red), our model without 3DGP (green) and with 3DGP using M1 (black) and M2 (blue) marginalization. Average Precision (AP) of each method is reported in Table.3.

each baseline detector is presented in Fig. 5 and Table 3. The detection bounding boxes and associated confidence scores from the baseline detectors are used to generate a discrete set of detection hypotheses O_o for our model. To measure detection accuracy, we report the precision-recall curves and average precision (AP) for each object type, with the standard intersection-union criteria for detections [6]. The marginal detection score $m(o_i)$ of a detection hypothesis is obtained by using the log-odds ratio that can be approximated by the following equation similarly to [5].

$$m(o_i) = \begin{cases} E_\Pi(\hat{G}, I) - E_\Pi(\hat{G}_{\setminus o_i}, I), \; o_i \in \hat{G} \\ E_\Pi(\hat{G}_{+o_i}, I) - E_\Pi(\hat{G}, I), \; o_i \notin \hat{G} \end{cases} \quad (7)$$

where $\hat{G}$ is the solution of our inference, $\hat{G}_{\setminus o_i}$ is the graph without o_i, and $\hat{G}_{+o_i}$ is the graph augmented with o_i. If there exists a parent 3DGP hypothesis for o_i, we remove the corresponding 3DGP as well when computing $\hat{G}_{\setminus o_i}$.

To better understand the effect of the 3DGP, we employ two different strategies for building the augmented parse graph $\hat{G}_{+o_i}$. The first scheme $M1$ builds $\hat{G}_{+o_i}$ by adding o_i as an object hypothesis. The second scheme $M2$ attempts to also add a parent 3DGP into $\hat{G}_{+o_i}$ if 1) the other constituent objects in the 3DGP (other than o_i) already exist in $\hat{G}$ and 2) the score is higher than the first scheme (adding o_i as an individual object). The first scheme ignores possible 3DGPs when evaluating object hypotheses that are not included in $\hat{G}$ due to low detection score, whereas the second scheme also incorporates 3DGP contexts while measuring the confidence of those object hypotheses.

Results: We ran experiments using the new *indoor-scene-object* dataset. To evaluate the contribution of the 3DGP to the scene model, we compared three versions algorithms: 1) the baseline methods, 2) our model without 3DGPs (including geometric and semantic context features), and 3) the full model with 3DGPs. In both 2) and 3), our model was trained on the same data and with the same setup.

Method	Pix. Acc	Floor	Center	Right	Left	Ceiling
Hedau [12]	81.4 %	73.4 %	68.4 %	71.0 %	71.9 %	56.2 %
W/O 3DGP	**82.8** %	76.9 %	**69.3** %	**71.8** %	**72.5** %	**56.3** %
3DGP	82.6 %	**77.3** %	**69.3** %	71.5 %	72.4 %	55.8 %

Table 2. Layout accuracy obtained by the baseline [12], our model without 3DGP and with 3DGP. Our model outperforms the baseline for all classes.

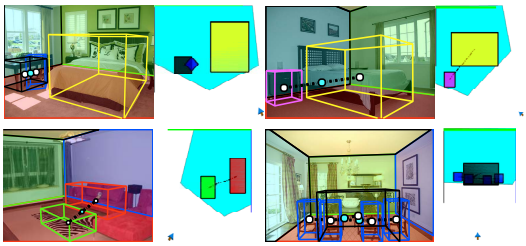


Figure 6. 2D and 3D (top-view) visualization of the results using our 3DGP model. Camera view point is shown as an arrow. This figure is best viewed in color.

As seen in the Table 3, our model (without or with 3DGPs) improves the detection accuracy significantly ($2-16\%$) for all object classes. We observe significant improvement using our model without 3DGPs for all objects except tables. By using 3DGPs in the model, we further improve the detection results, especially for side tables ($+8\%$ in AP). This improvement can be explained by noting that the 3DGP consisting of a bed and side-table boosts the detection of side-tables, which tend to be severely occluded by the bed itself (Fig. 4 (middle)). Fig. 7 provides qualitative results. Notice that M2 marginalization provides higher recall rates in lower precision areas for tables and side tables than M1 marginalization. This shows that the 3DGP can transfer contextual information from strong object detection hypotheses to weaker detection hypotheses.

The scene model (with or without 3DGPs) significantly improves scene classification accuracy over the baseline ($+7.2\%$) by encoding the semantic relationship between scene type and objects (Table. 1). The results suggest that our contextual cues play a key role in the ability to classify the scene. Our model also outperforms state-of-the-art scene classifiers [19, 21] trained on the same dataset.

Finally, we demonstrate that our model provides more accurate layout estimation (Table. 2) by enforcing that all objects lie inside of the free space (see Fig. 7). We observe that our model does equal or better than the baseline [12] in $94.1\%(396/421)$ of all test images. Although the pixel label accuracy improvement is marginal compared to the baseline method, it shows a significant improvement in the floor estimation accuracy (Table. 2). We argue that the floor is the most important layout component since its extent directly provides information about the free space in the scene; the intersection lines between floor and walls uniquely specify the 3D extent of the free space.

Method	Sofa	Table	Chair	Bed	D.Table	S.Table
DPM [8]	42.4 %	27.4 %	45.5 %	91.5 %	85.5 %	48.8 %
W/O 3DGP	44.1 %	26.8 %	49.4 %	**94.7** %	**87.8** %	57.6 %
3DGP-M1	**52.9** %	37.0 %	52.5 %	94.5 %	86.7 %	64.5 %
3DGP-M2	**52.9** %	**38.9** %	**52.6** %	94.6 %	86.7 %	**65.4** %

Table 3. Average Precision of the DPM [8], our model without 3DGP and with 3DGP. Our model significantly outperforms DPM baseline in most of the object categories.

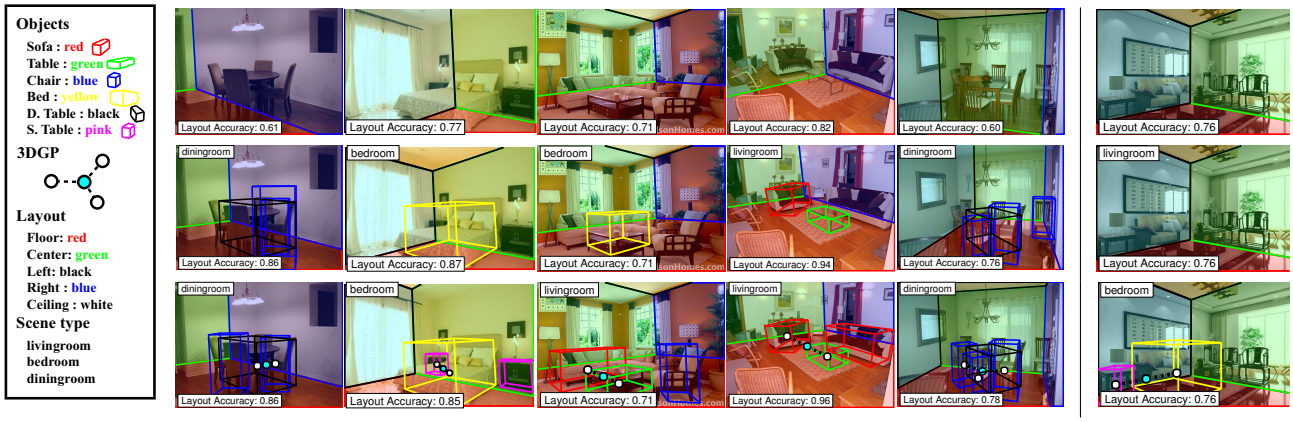


Figure 7. Example results. First row: the baseline layout estimator [12]. Second row: our model without 3DGPs. Third row: our model with 3DGPs. Layout estimation is largely improved using the object-layout interaction. Notices that the 3DGP helps to detect challenging objects (severely occluded, intra-class variation, etc.) by reasoning about object interactions. Right column: false-positive object detections caused by 3DGP-induced hallucination. See supplementary material for more examples. This figure is best shown in color.

7. Conclusion

In this paper, we proposed a novel unified framework that can reason about the semantic class of an indoor scene, its spatial layout, and the identity and layout of objects within the space. We demonstrated that our proposed object 3D Geometric Phrase is successful in identifying groups of objects that commonly co-occur in the same 3D configuration. As a result of our unified framework, we showed that our model is capable of improving the accuracy of each scene understanding component and provides a cohesive interpretation of an indoor image.

Acknowledgement: We acknowledge the support of the ONR grant N00014111038 and a gift award from HTC.

References

[1] S. Y. Bao and S. Savarese. Semantic structure from motion. In *CVPR*, 2011. 2

[2] S. Y. Bao, M. Sun, and S. Savarese. Toward coherent object detection and scene layout understanding. In *CVPR*, 2010. 1, 2

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. 6

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1, 2, 4

[5] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *IJCV*, 2011. 2, 3, 4, 5, 6, 7

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) challenge. *IJCV*, 2010. 6, 7

[7] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. *CVPR*, pages 524–531, 2005. 1, 2

[8] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9), Sept. 2010. 1, 2, 3, 5, 6, 7

[9] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic. People watching: Human actions as a cue for single-view geometry. In *ECCV*, 2012. 2

[10] A. Geiger, C. Wojek, and R. Urtasun. Joint 3d estimation of objects and scene layout. In *NIPS*, 2011. 2

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

[12] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered room. In *ICCV*, 2009. 1, 2, 3, 4, 5, 6, 7, 8

[13] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *ECCV*, 2010. 2, 4

[14] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM J. on Optimization*, 1998. 4

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 1, 2, 3, 6

[16] D. Lee, A. Gupta, M. Hebert, and T. Kanade. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *NIPS*, 2010. 1, 2, 4

[17] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Statistical Learning in Computer Vision, ECCV*, 2004. 1

[18] C. Li, D. Parikh, and T. Chen. Automatic discovery of groups of objects for scene understanding. In *CVPR*, 2012. 2, 4

[19] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, December 2010. 2, 6, 7

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, Nov. 2004. 2

[21] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011. 1, 2, 6, 7

[22] L. D. Pero, J. Bowdish, D. Fried, B. Kermgard, E. L. Hartley, and K. Barnard. Bayesian geometric modeling of indoor scenes. In *CVPR*, 2012. 2, 4

[23] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009. 1, 2, 6

[24] A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 2, 4

[25] S. Satkin, J. Lin, and M. Hebert. Data-driven scene understanding from 3d models. In *BMVC*, 2012. 2

[26] A. G. Schwing and R. Urtasun. Efficient exact inference for 3d indoor scene understanding. In *ECCV*, 2012. 2

[27] H. Wang, S. Gould, and D. Koller. Discriminative learning with latent variables for cluttered indoor scene understanding. In *ECCV*, 2010. 1, 2, 4

[28] Y. Wang and G. Mori. Hidden part models for human action recognition: Probabilistic versus max margin. *PAMI*, 2011. 5

[29] Y. Xiang and S. Savarese. Estimating the aspect layout of object categories. In *CVPR*, 2012. 1

[30] Y. Zhao and S.-C. Zhu. Image parsing via stochastic scene grammar. In *NIPS*, 2011. 2, 3