Noname manuscript No.
(will be inserted by the editor)

Integrating Function, Geometry, Appearance for Scene Parsing

Yibiao Zhao · Song-Chun Zhu

Received: date / Accepted: date

Abstract Visual scene understanding is a fundamental task in computer vision systems. Traditional appearance-based classification paradigms struggle to cope with the view and appearance variations of indoor scenes and functional objects. In this paper, we present a Stochastic Scene Grammar (SSG) model to parse indoor images. The grammar is defined on a Function-Geometry-Appearance (FGA) hierarchy based on two observations: i) Functionality is the most essential property to define an indoor object, *e.g.*"an object to sit on" defines a chair, ii) The geometry (3D shape) of an object is designed to serve its function. We formulate the nature of the object functionality and contextual relations into the Stochastic Scene Grammar model, which characterizes a joint distribution over the FGA hierarchy. This hierarchical structure includes both functional concepts (the scene category, functional groups, functional objects, functional parts) and geometric entities (3D/2D/1D shape primitives). The decomposition of the grammar is terminated on the bottom-up detection of line and region likelihood. We use a simulated annealing MCMC algorithm to find the maximum a posteriori (MAP) solution, *i.e.*a parse tree. We design four data-driven steps to accelerate the search in the FGA space: i) group the line segments into 3D primitive shapes, ii) assign functional labels to these 3D primitive shapes, iii) fill in missing objects/parts according to the functional labels, and iv) synthesize 2D label maps and verify the current parse tree by the Metropolis-Hastings acceptance probability. Experimental results on several challenging indoor datasets demonstrate the proposed approach not only significantly widens the scope of indoor scene parsing algorithms from the segmentation and the 3D recovery to functional object recognition, but also yields improved overall performance.

Yibiao Zhao
University of California, Los Angeles (UCLA), USA
E-mail: ybzhao@ucla.edu
http://www.stat.ucla.edu/~ybzhao

Song-Chun Zhu
University of California, Los Angeles (UCLA), USA
E-mail: sczhu@stat.ucla.edu
http://www.stat.ucla.edu/~sczhu

1 Introduction

A central goal of computer vision is creating computational systems whose visual recognition and scene understanding accuracy is comparable to, or better than, that of biological vision. With the emergence of new technologies, such as wearable cameras and autonomous driving, computer vision is about to play a key role in peoples daily lives. The proliferation of sensors in a wide range of contexts demands an expansion in the scope of scene understanding. A large portion of the vision literature studied scenes as a classification problem or image labeling problem based on their appearance. We argue that this is ill-posed for two reasons: i) most scenes, especially indoor living spaces are defined by their functions; and ii) a space may serve multi-functions, and thus cannot be simply classified in one category. We pose the problem as scene parsing, which integrate three aspects: function, geometry and appearance. The geometric sizes of furniture are fitted to the 3D functional relations learned from data. The actions, such as sitting and sleeping, provide the top-down constraints about 3D contexts, and thus disambiguate the appearance uncertainty. The inferred plausible actions define the functions of the space.

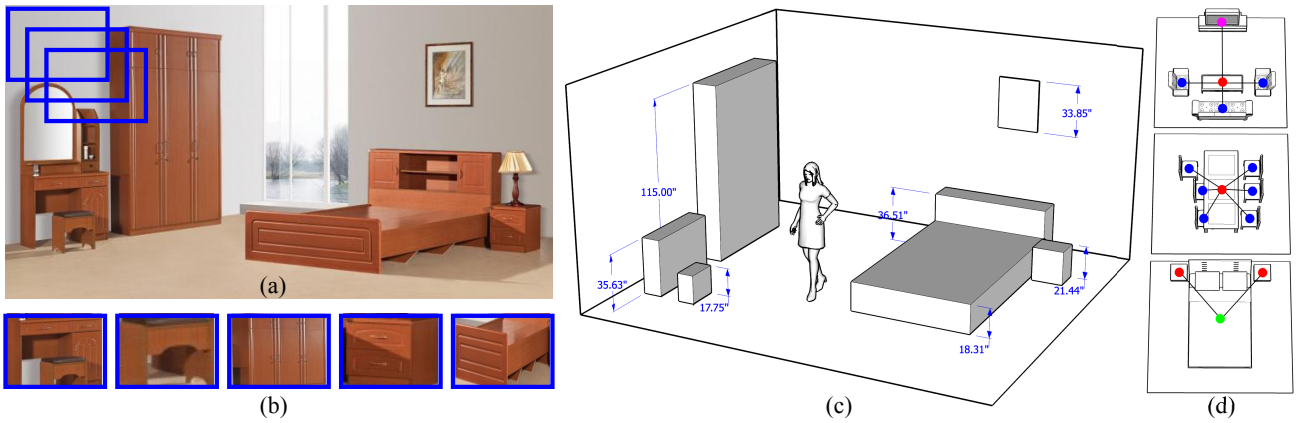


Fig. 1 Given an input image (a), traditional approaches recognize objects by analyzing the local appearance inside sliding windows (b). In this paper, we recognize the functional objects by inferring the *affordance* (c)– how likely a 3D shape is able to afford a human action, and *contextual relations* (d) – how functional objects are organized in a specific scene.

1.1 Motivation and objectives

Although object detection and labeling have made remarkable progresses in the field of computer vision, the detection of indoor objects and segmentation of indoor scenes are still challenging tasks. Over almost 50 years the recognition of explicit visual patterns, like face and handwritten digits, have matured and become ubiquitous in modern industrial and consumer products. Computers however still cannot perform many important tasks that are trivial for human vision. As demonstrated by the PASCAL VOC2012 Challenge (Everingham et al (2010)), state-of-the-art algorithms can only get 19.5% (22.6%) accuracy for the chair detection (segmentation) task. Other indoor objects, like sofas and dining tables, are also among the categories with lowest accuracies out of the twenty object categories. Both the geometric approaches in the 1980-1990s and the appearance methods dominating the last 15-years have fundamental limits. This performance gap between people and computers seemingly cannot be filled by designing better features and collecting larger dataset as people have been doing in recent years.

Image understanding is not only about the image itself but also the knowledge of the world. Humans recognize images so well because we know how the world works. Prior knowledge about the functional, physical and social mechanics of objects in the 3D world is the key for understanding. However, there is a very large gap between the observed images and the knowledge we have in our minds. In order to appropriately recognize an image, computers must have an internal abstract representation of what units of image are and how to put them together.

In Fig.1(a, b), five objects were cropped out of the image. Without the context given in image (a), even a human has difficulty identifying these objects solely based on the appearance of image patches. The classic sliding-window type of object detectors, which only observe a small isolated patch, will have difficulty distinguishing objects from one another. On the other hand, if we look at the Fig.1(c) despite the appearances of image (a), we can immediately recognize objects to sit on (chair), to sleep on (bed) and to store in (cabinet) based on their 3D shapes. For example, a cuboid of 18 inch tall could be comfortable to sit on as a chair. Moreover, the contextual relations are helpful in identifying objects with similar shapes, such as the chair on the left and the nightstand on the right. Although they are in similar shape, the nightstand is more likely to be placed beside the bed. The bed and the nightstand offer a joint functional group to serve the activity of sleeping. Some typical functional groups are illustrated in Fig.1(d).

By analogy to natural language parsing, we pose the scene understanding problem as parsing an image into a hierarchical structure of functional concepts and geometric entities using the Stochastic Scene Grammar (SSG).

1.2 Related work

Scene representation. There are four major representations: (i) Feature representation: The surge of scene understanding studies start from a series of early work on designing feature representation of spatial envelope, the gist representation by Oliva and Torralba (2001), spatial pyramid matching (SPM) by S. Lazebnik and Ponce (2006) and recent reconfigurable models by S. N. Parizi and Felzenszwalb (2012) and S. Wang and Zhu (2012). (ii) Region-based representations: The method of conditional random fields by Lafferty et al (2001) are widely

used to represent semantic relations by maximizing the affinity between region labels. These studies considered some qualitative context descriptions such as {inside, below, around, above}, which are proved to be helpful to recognize outdoor objects. Choi et al (2010) studied 2D context models that guide detectors to produce a semantically coherent interpretation of a scene. They demonstrated that 2D horizontal contexts are very sensitive to camera rotations. (iii) Non-parametric representations: such as label transfer by C. Liu and Torralba (2011), SuperParsing by Tighe and Lazebnik (2013a,b) and scene collage by Isola and Liu (2013) interpret a new scene by searching nearest neighbors from images in the scene dataset, and then transfer the label maps to the target through warping or contextual inference. Interestingly, Satkin et al (2012), Satkin and Hebert (2013) recently generalize the idea of nearest-neighbor search to the 3D scenes, so that their approach can recognize objects cross viewpoints. Lim et al (2013), Del Pero et al (2013) detected indoor objects by matching with fine-grained furniture model. (iv) Block world representation: Representation of 3d blocks allows reasoning about the physical constraints within the 3D scene. Gupta et al (2010) posed the 3D objects as blocks and inferred its 3D properties such as occlusion, exclusion and stability in addition to surface orientation labels. They showed that a global 3D prior does improve 2D surface labeling. Hedau et al (2009, 2010, 2012), Wang et al (2010), Lee et al (2009, 2010), Schwing et al (2012, 2013) parameterized the geometric scene layout of the background and/or foreground blocks and trained their models by the Structured SVM (or Latent SVM). Hu (2012), Xiao et al (2012), Hejrati and Ramanan (2012), Xiang and Savarese (2012), Pepik et al (2012), Fidler et al (2012) designed several new variants of the deformable part-based models to detect 3D entities under different view points.

Object function and affordance. The concept of affordance was proposed by the perceptual psychologist Gibson (1977), which refers to the perceived fundamental properties of the thing that determine how the thing could possibly be used. A pioneer work in computer vision by Stark and Bowyer (1991) proposed the use of a functional properties to recognize 3D objects. They parse an objects into a 3D geometric description, and recognize the object by searching potential functional elements. More recently, approaches have been proposed to detect objects based on human interaction. The human activity is annotated by extracting human motion from rgbd video data and used to indirectly identify objects Wei et al (2013). In this works, it is assumed that interactions are observed during training and testing. Bar-aviv and Rivlin (2006), Grabner et al (2011) detected chairs by the hallucination of embodied agents in the 3D CAD data and depth data respectively. Gupta et al (2011) proposed an algorithm to infer the human workable space by adapting human poses to the scene. Lin et al (2013), Choi et al (2013), Zhao and Zhu (2013) recently proposed holistic approaches to exploits 2D segmentation, 3D geometry, as well as contextual relations between scenes and objects for parsing rgbd and 2D images.

Single-view 3D reconstruction. Automatic 3D reconstruction from a single image is an ill posed problem. In order to recover a meaningful 3D, one have to make assumptions about the scene and use the prior knowledge to regularize the solution. These assumptions include: (i) Sketch smoothness assumption: Han and Zhu (2004) first tackled the problem by assuming the local sketch smoothness and global scene alignment. (ii) Piece-wise smoothness assumption: Saxena et al (2009) presented a fully supervised method to learn a mapping between informative features and depth values under a conditional random field framework. Payet and Todorovic (2011) proposed a joint model to recognize objects and estimate scene shape simultaneously. (iii) Surface assumption: Hoiem et al (2009) recognized the geometric surface orientation and fit ground-line that separate the floor and objects in order to pop-up the vertical surface. Delage et al (2007) proposed a dynamic Bayesian network model to infer the floor structure for autonomous 3D reconstruction from a single indoor image. Mobahi et al (2011) extracted low rank textures of repeated patterns to construct surfaces like building facades. (iv) Manhattan world representation: Recent studies on indoor scene parsing, including Hedau et al (2009, 2010, 2012), Wang et al (2010), Lee et al (2009, 2010), Schwing et al (2012, 2013), Zhao and Zhu (2011, 2013) and Del Pero et al (2011, 2012, 2013) adopted the Manhattan world representation extensively. This assumption stated that man-made scenes were built on a cartesian grid which led to regularities in the image edge gradient statistics. This enables us, from a single image, to determine the orientation of the viewer relative to the scene and also to recover scene structures which are aligned with the grid.

Stochastic image grammar. This stream of research started from "syntactic pattern recognition" by K. S. Fu and his school in the late 1970s to early 1980s. Fu (1982) depicted an ambitious program of block world scene understanding using grammars. This stream was disrupted in the 1980s and suffered from the lack of an image vocabulary that is realistic enough to express real-world objects and scenes, and reliably detectable from images. Tu et al (2005) raised the notion of image parsing to the decomposition of an image

into a hierarchical "parse graph" by a data-driven data-driven Monte Carlo sampling strategy. Zhu and Mumford (2007) proposed an AND/OR graph model to represent the compositional structures in vision. Han and Zhu (2009) applied grammar rules, in a greedy manner, to detect rectangular structures in man-made scenes. Porway and Zhu (2010) proposed an cluster sampling algorithm to parse aerial images by allowing for Markov chain jumping between competing solutions. An earlier version of our work appeared at Zhao and Zhu (2011, 2013). In this paper, we will explore the Stochastic Scene Grammar model indepth with more insights on the compositionality of functional concepts and geometric entities and computing strategy.

1.3 Overview of our approach

By analogy to natural language parsing, we pose the scene understanding problem as parsing an image into a hierarchical structure of functional concepts and geometric entities using the Stochastic Scene Grammar (SSG).

In this paper, we parse an image into a hierarchical structure, namely *a parse tree*, using the Stochastic Scene Grammar (SSG) defined on a Function-Geometry-Appearance (FGA) hierarchy. Therefore, this paper has two major contributions to the scene parsing problems:

(I) A Stochastic Scene Grammar (SSG) .

The grammar is introduced to represent a hierarchical structure of functional concepts and geometric entities. The grammar starts from a root node (the scene) and ends in a set of terminal nodes (lines/regions) as shown in Fig.2. In between, we model all intermediate functional concepts and geometric entities by three types of production rules and two types of contextual relations, as illustrated in Fig.3.

Production rules: *AND*, *OR*, and *SET*. (i) The AND rule encodes how sub-parts are composed into a larger structure. For example, three hinged faces form a 3D box, four linked line segments form a rectangle, a background and inside objects form a scene in Fig.3(i); (ii) The SET rule represents an ensemble of entities, e.g. a set of 3D boxes or a set of 2D regions as in Fig.3(ii); (iii) The OR rule represents a switch between different sub-types, e.g. a 3D foreground and 3D background have several switchable types in Fig.3(iii).

Contextual relations: *Cooperative* "+" and *Competitive* "-". (i) If the visual entities satisfy a cooperative "+" relation, they tend to bind together, e.g. hinged faces of a foreground box showed in Fig.3(a). (ii) If entities satisfy a competitive "-" relation, they compete with each other for presence, e.g. two exclusive foreground boxes competing for a same space in Fig.3(b).

(II) A Function-Geometry-Appearance (FGA) hierarchy .

On top of the Grammatical representation, our model is further developed based on observations of the FGA hierarchy as shown in Fig.2.

Function: An indoor scene is designed to serve a handful of human activities inside. The indoor objects (furniture) in the scenes are designed to support human actions, *e.g.*bed to sleep on, chair to sit on *etc.*

In the functional space, we model the probabilistic derivation of functional concepts including scene categories (bedroom), functional groups (sleeping area), functional objects (bed and nightstand), and functional parts (the mattress and the headboard of a bed).

Geometry: The 3D size (dimension) can be sufficient to evaluate how likely an object is able to afford a human action, known as the *affordance* Gibson (1977). Fortunately, most furniture has regular structures, *i.e.*rectangular shapes, therefore the detection of these objects is tractable by inferring their geometric affordance. For objects like sofas and beds, we use a more fine-grained geometric model with compositional parts, *i.e.*a group of cuboids. For example, the bed with a headboard is a better explanation of the image in terms of segmentation accuracy as shown at the bottom of Fig.2.

In the geometric space, each 3D shape is directly linked to a concept in the functional space. The contextual relations are utilized when multiple objects are assigned to the same functional group, *e.g.*a bed and a nightstand for sleeping. The distribution of the 3D geometry is learned from a large set of 3D models as shown in Fig.4.

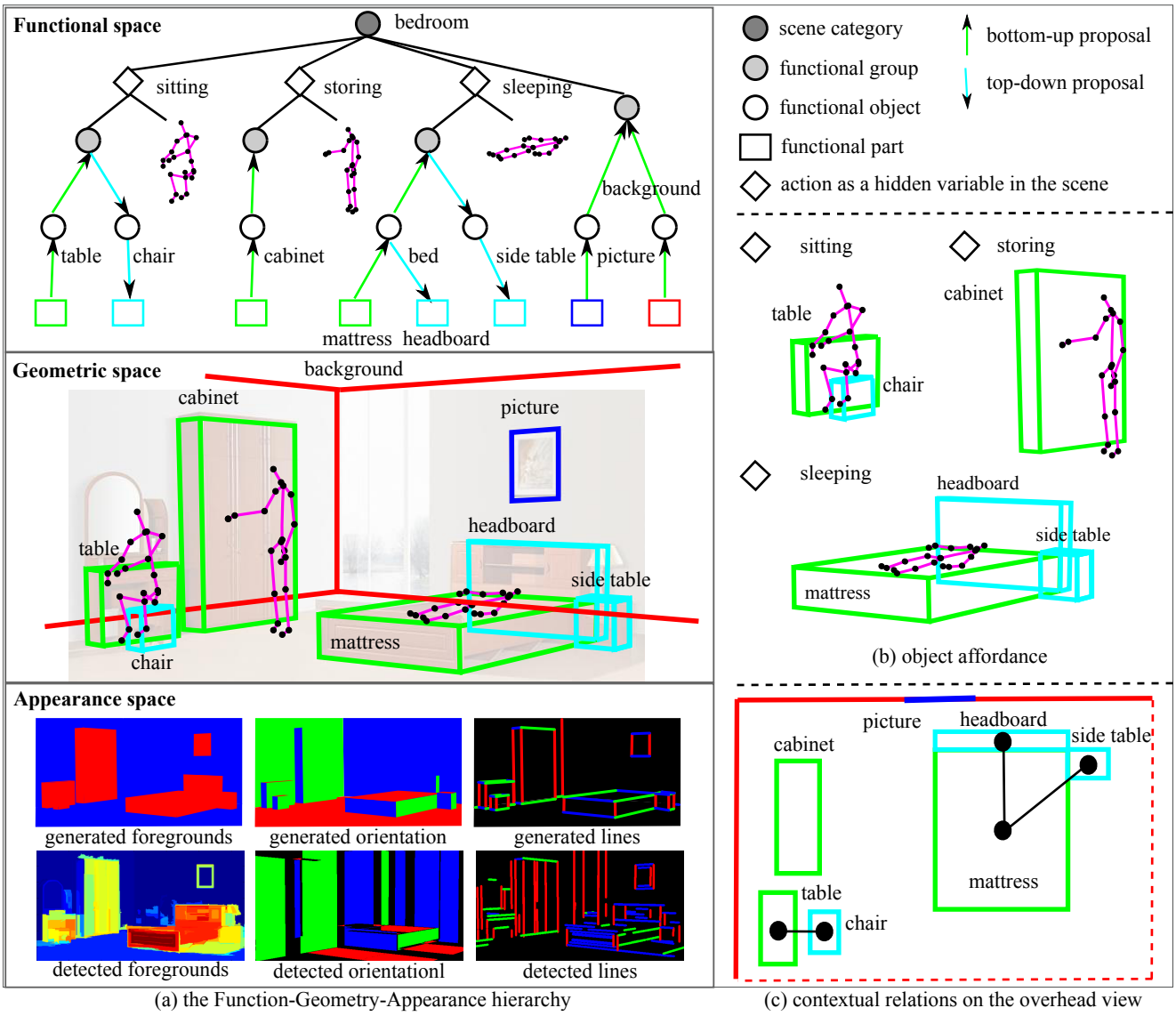
Appearance: The appearance of the furniture has large variations due to differing material properties, lighting conditions, and viewpoints. In order to land our model on the input image, we use a straight-line detection, a surface orientation estimation and a coarse foreground detection as the local evidence to support the geometry model above as shown in Fig.2.

We design a four-step inference algorithm that enables a MCMC chain to travel up and down through the FGA hierarchy:

i). A bottom-up appearance-geometry (AG) step groups noisy line segments in the A space into 3D primitive shapes, *i.e.*cuboids and rectangles, into the G space;

ii). A bottom-up geometry-function (GF) step assigns functional labels in the F space to detected 3D primitive shapes, *e.g.*to sleep on;

iii). A top-down function-geometry (FG) step further fills in the missing objects and the missing parts in the G space according to the assigned functional la-



(a) the Function-Geometry-Appearance hierarchy (b) object affordance (c) contextual relations on the overhead view

Fig. 2 Integrating function, geometry and appearance for scene parsing. The functional concepts impose the object affordance and contextual relations to the geometric entities. The final parsing result is evaluated on top of the synthesis of appearance likelihood maps.

bels, *e.g.*a missing nightstand of a sleeping group or a missing headboard of a bed;

iv). A top-down geometry-appearance (GA) step synthesizes 2D label maps in the A space, and makes an accept/reject decision of a current proposal by the Metropolis-Hastings acceptance probability.

2 Stochastic Scene Grammar

The Stochastic Scene Grammar (SSG) is defined as a four-tuple $G = (S, V, R, P)$, where S is a start symbol at the root (scene); $V = V^N \cup V^T$, V^N is a finite set of non-terminal nodes (structures or sub-structures), V^T is a finite set of terminal nodes (line segments); $R = \{r : \alpha \to \beta\}$ is a set of production rules, each of which represents a generating process from a parent node α to its child nodes $\beta = Ch_\alpha$. $P(r) = P(\beta|\alpha)$ is an expansion probability for each production rule $(r : \alpha \to \beta)$. A set of all valid configurations C derived from production rules is called a *language*:

$$L(G) = \{C : S \overset{\{r_i\}}{\to} C, \{r_i\} \subset R, C \subset V^T\}. \quad (1)$$

2.1 Production rules

We define three types of stochastic production rules R^{AND},R^{OR},R^{SET} to represent the structural *regularity*

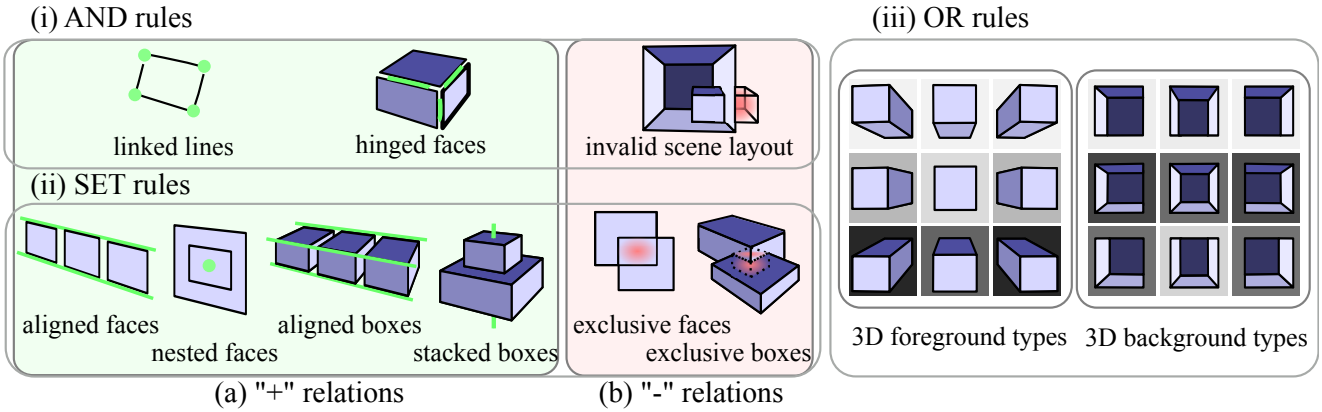


Fig. 3 Three types of production rules: AND (i), SET (ii) OR (iii), and two types of contextual relations: cooperative "+" relations (a), competitive "-" relations (b).

and *flexibility* of visual entities. The regularity is enforced by the AND rule and the flexibility is expressed by the OR rule. The SET rule is a mixture of OR and AND rules.

(i) An AND rule ($r^{AND} : A \rightarrow a \cdot b \cdot c$) represents the *decomposition* of a parent node A into three sub-parts a, b, and c. The probability $P(a, b, c|A)$ measures the compatibility (contextual relations) among sub-structures a, b, c. As seen Fig.3(i), the grammar outputs a high probability if the three faces of a 3D box are well hinged, and a low probability if the foreground box lays out of the background.

(ii) An OR rule ($r^{OR} : A \rightarrow a \mid b$) represents the *switching* between two sub-types a and b of a parent node A. The probability $P(a|A)$ indicates the preference for one subtype over others. For the 3D foreground in Fig.3(iii), the three sub-types in the third row represent objects below the horizon. These objects appear with high probabilities. Similarly, for the 3D background in Fig.3(iii), the camera rarely faces the ceiling or the ground, hence, the three sub-types in the middle row have higher probabilities (darker color means higher probability). Moreover, OR rules also model the discrete size of entities, which is useful to rule out the extremely large or small entities.

(iii) An SET rule ($r^{SET} : A \rightarrow \{a\}_k, k \geq 0$) represents an *ensemble* of k visual entities. The SET rule is equivalent to a mixture of OR and AND rules ($r^{SET} : A \rightarrow \emptyset \mid a \mid a \cdot a \mid a \cdot a \cdot a \mid \cdots$). It first chooses a set size k by ORing, and forms an ensemble of k entities by ANDing. It is worth noting that the OR rule essentially changes the graph topology of the output parse tree by changing its node size k.

As a result, the AND, OR, SET rules generate various functional concepts and geometric entities which satisfy contextual relations as seen in Fig.3

2.2 Contextual relations

There are two kinds of contextual relations, *Cooperative* "+" relations and *Competitive* "-" relations, which are involved in the AND and SET rules.

(i) The cooperative "+" relations specify the *concurrent* patterns in a scene, e.g. hinged faces, nested rectangle, aligned windows in Fig.3(a). The visual entities satisfying a cooperative "+" relation tend to bind together. The cooperative "+" relation is introduced by either functional context in Sect.3.3 or geometric decomposition in Sect.3.4.

(i) The competitive "-" relations specify the *exclusive* patterns in a scene. If entities satisfy competitive "-" relations, they compete with each other for presence. As shown in Fig.3(b), if a 3D box is not contained by its background, or two 2D/3D objects are exclusive with one another, these cases will rarely be in a solution simultaneously. The "-" relations is introduced by physical constraints in Sect.3.3.

If several visual entities satisfy a cooperative "+" relation, they tend to bind together as a *tight structures*. We group visual entities into these tight structures as much as possible in the early stage of inference according to the geometric decomposition (Sect.4). If the entities do not violate any competitive "-" relation, they may be loosely combined as a *loose structures*, whose combinations are sampled in a later stage of inference (Sect.4). The high-level functional concept will also impose "+" relations in the later stage of inference. If an object is assigned with functional label, then the algorithm will be able to sample its parts or nearby objects according to the 3D contextual relations as explained in Sect.3.3.

With the three production rules and two contextual relations, SSG is able to handle an enormous number

of scene configurations and large geometric variations, which are the major difficulties in our task.

2.3 Bayesian formulation of the grammar

We define a posterior distribution for a solution (a parse tree) pt conditioned on an input image I. This distribution is specified in terms of the statistics defined over the derivation of production rules.

$$\begin{aligned} P(pt|I) &\propto P(pt)P(I|pt) \\ &= P(S)\prod_{v\in V^N} P(Ch_v|v)\prod_{v\in V^T} P(I|v) \end{aligned} \tag{2}$$

where I is the input image, pt is the parse tree. The probability derivation represents a generating process of the production rules $\{r : v \to Ch_v\}$ from the start symbol S to the nonterminal nodes $v \in V^N$, and to the children of non-terminal nodes Ch_v. The generating process stops at the terminal nodes $v \in V^T$ and generates the image I.

We use a probabilistic graphical model of an AND/OR graph [12, 17] to formulate our grammar. The graph structure $G = (V, E)$ consists of a set of nodes V and a set of edges E. The edges define a parent-child conditional dependency for each production rule. The posterior distribution of a parse graph pt is given by a family of Gibbs distributions:

$$P(pt|I;\lambda) = \frac{1}{Z(I;\lambda)}\exp\{-E(pt|I)\}, \tag{3}$$

where $Z(I;\lambda) = \sum_{pt\in\Omega}\exp\{-E(pt|I)\}$ is a partition function summation over the solution space Ω.

The energy is decomposed into three potential terms:

$$\begin{aligned} E(pt|I) = &\sum_{v\in V^{OR}} E^{OR}(A_T(Ch_v)) \\ &+ \sum_{v\in V^{AND}} E^{AND}(A_G(Ch_v)) \\ &+ \sum_{\Lambda_v\in\Lambda_I, v\in V^T} E^T(I(\Lambda_v)) \end{aligned} \tag{4}$$

(i) **The energy for OR nodes** is defined over "type" attributes $A_T(Ch_v)$ of ORing child nodes. The potential captures the prior statistics on each switching branch. $(r : v \to Ch_v)$.

$$\begin{aligned} E^{OR}(A_T(v)) &= -\log P(v \to A_T(v)) \\ &= -\log\{\frac{\#(v \to A_T(v))}{\sum_{u\in Ch(v)}\#(v \to u)}\}. \end{aligned} \tag{5}$$

The switching probability of foreground objects and the background layout is shown in Fig.3(iii).

(ii) **The energy for AND nodes** is defined over "geometric" attribute $A_G(Ch_v)$ of ANDing child nodes. They are Markov Random Fields (MRFs) inside a tree-structure. We define both "+" relations and "-" relations as

$$E^{AND} = \lambda^+h^+(A_G(Ch_v)) + \lambda^-h^-(A_G(Ch_v)), \tag{6}$$

where $h(*)$ are sufficient statistics in the exponential model, λ are their parameters. For 2D faces as an example, the "+" relation specifies a quadratic distance between their connected joints

$$h^+(A_G(Ch_v)) = \sum_{a,b\in Ch_v}(X(a) - X(b))^2, \tag{7}$$

and the "-" relation specifies an overlap rate between their occupied image area

$$h^-(A_G(Ch_v)) = (\Lambda_a \cap \Lambda_b)/(\Lambda_a \cup \Lambda_b), a, b \in Ch_v. \tag{8}$$

(iii) **The energy for Terminal nodes** is defined over bottom-up image features $I(\Lambda_v)$ on the image area Λ_v. The features used in this paper include: (a) a foreground map, (b) a 3D orientation map, (c) a line segments map as shown in Fig.2. This term only captures the features from their dominant image area Λ_v, and avoids the double counting of the shared edges and the occluded regions as discussed in Sect.3.5.

3 Integrating function, geometry and appearance

We define our grammar model in the context of indoor scenes over the functional space $\mathcal{F}$, the geometric space $\mathcal{G}$ and the appearance space $\mathcal{A}$ as shown in Fig.2. The model involves the notion of the functional concept, the object affordance, the contextual relation, the decomposition of geometric entities, the quantization of appearance space, and single view 3D reconstruction.

3.1 The functional concepts

The functional space $\mathcal{F}$ contains the categorical variables of functional concepts, including the scene categories Fs, the functional groups Fg, the functional objects Fo, and the functional parts Fp. Starting from a start symbol S, we define following production rules:

S → Fs: S → [bedroom] | [living room]
Fs→Fg: [bedroom] → [sleeping][background] | ⋯
Fg→Fo: [sleeping] → [bed] | [bed][night stand] | ⋯
Fo→Fp: [bed] → [headboard][mattress] | [mattress]

The OR symbol "|" separates alternative explanations of the grammar derivation. Each alternative explanation has a branching probability $q(\alpha \to \beta)$, which is learned by simply counting the frequency of each production rules on the labels of thousands of images in

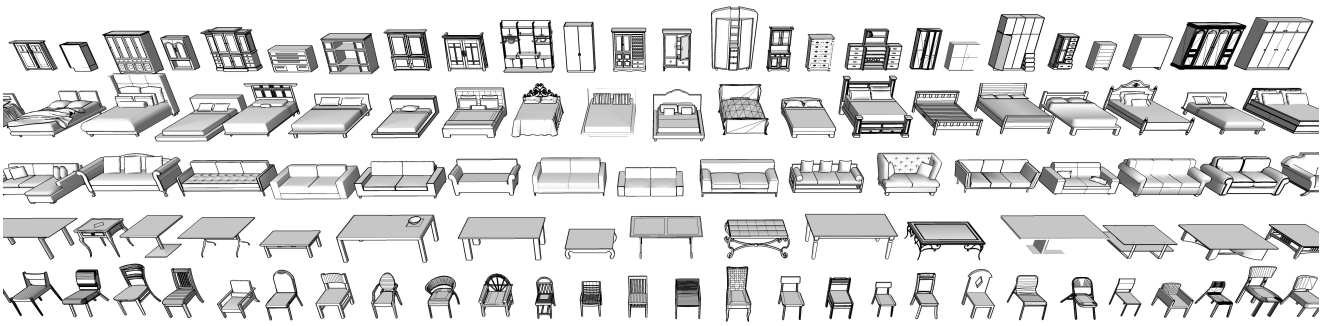


Fig. 4 A collection of indoor functional objects from the Google 3D Warehouse

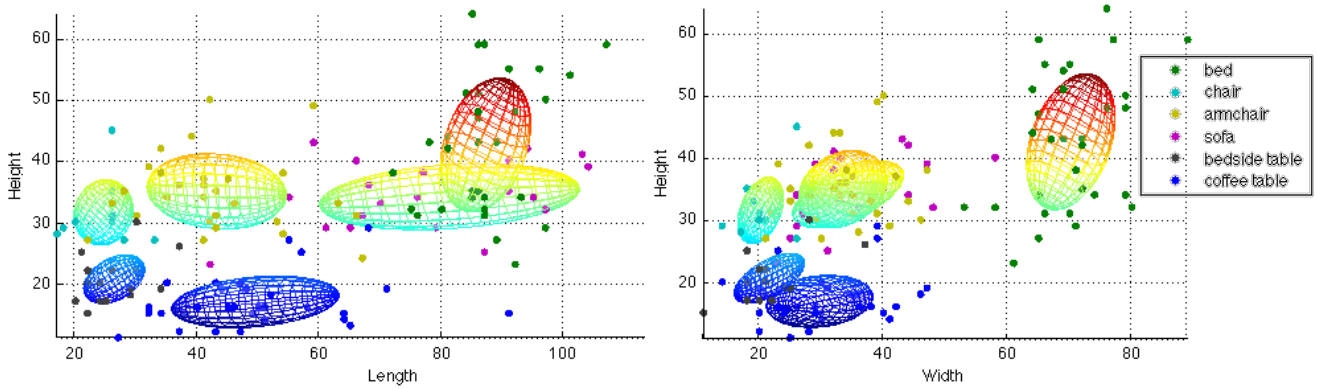


Fig. 5 The distribution of the 3D sizes of the functional objects (in unit of inch).

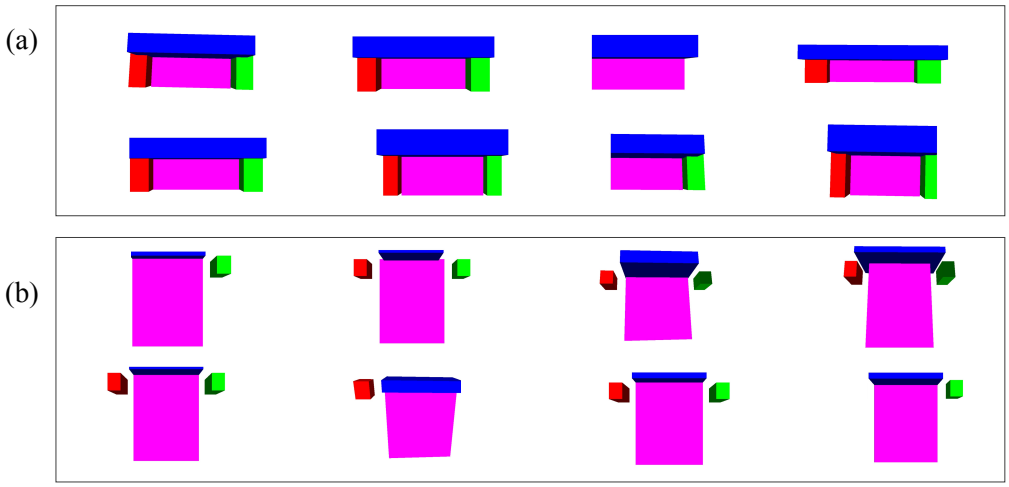


Fig. 6 Samples drawn from the distributions of 3D geometric models (a) the functional object "sofa" and (b) the functional group "sleeping".

the SUN dataset by Xiao et al (2010) under the "bedroom" category and the "living room" category. We can see that all the production rules are SET rules, each of which is a combination of an OR rule and and AND rule together.

The functional concepts impose the object affordance by modeling the distribution of 3D size (dimension) for each geometric primitive Gp as shown in Fig.5. Meanwhile, they also introduce the contextual relations among these geometric primitives by modeling the distance distribution between them.

3.2 The object affordance

Object affordance models the distribution of 3D size (dimension) for each functional part, for example, how large the bed mattress is. If we consider human actions as hidden variables in the space, then the affordance probability measures how likely the geometric shape of an object is able to afford an action. As shown in Fig.2, a cube around 1.5ft tall is comfortable to sit on despite its appearance, and a "table" of 6ft tall loses its original function – to place objects on while sitting in front of. We model the 3D sizes of the functional parts by a mixture of Gaussians. The model characterizes the Gaussian nature of the object sizes and allows for simultaneous alternatives of canonical sizes, such as king size bed, full size bed *etc.* It is also a kind of SET rule, where OR represents the mixture, and AND represents the joint distribution among several dimensions of each Gaussian. We estimated the model by EM clustering, and we manually picked a few typical primitives as the initial mean for the Gaussian, *e.g.*a coffee table, a side table and a desk from the table category.

In order to learn a better affordance model, we collected a dataset of functional indoor furniture, as shown in Fig.4. The functional objects in the dataset are modeled with real-world measurements, and therefore we can generalize our model to real images by learning from this dataset. We found that the real-world 3D sizes of the objects has less variance than the projected 2D sizes. As we can see, these functional categories are quite distinguishable solely based on their sizes as shown in Fig.5. For example, the coffee tables and side tables are very short and usually lower than the sofas, and the beds generally wider than others. The object poses are aligned in the dataset. We keep four copies of the Gaussian model for four alternative orientations along x, $-x$, y and $-y$ axes to make the model rotation invariant in the testing stage.

3.3 The contextual relations

We define two types of relations: functional relations which are the *Cooperative* "+" relations and physical constraints which are the *Competitive* "-" relations.

Functional relations are defined by the distributions of the 3D relative relations among the parts of an objects Fo or the objects of an functional group Fg. The relative relations is modeled by the distribution of distances between correspondent dimensions of two entities. Fig.6 shows some typical samples drawn from our learned distribution. This term enables top-down prediction of the missing parts objects as we will discuss in Sect.4.

Physical constraints avoid invalid geometric configurations that violate physical laws: Two objects can not penetrate each other and the objects must be contained in the room. The model penalizes the penetrating area between foreground objects Λ_f and the exceeding area beyond the background room borders Λ_b as $1/z \exp\{-\lambda(\Lambda_f + \Lambda_b)\}$, where we take λ as a large number, and $\Lambda_f = \Lambda(v_i) \bigcap \Lambda(v_j)$, $\Lambda_b = \Lambda(v_i) \bigcap \overline{\Lambda(bg)}$.

3.4 The decomposition of geometric entities

As shown in Fig.8, the geometric space $\mathcal{G}$ contains the geometric primitives of 3D cuboids, 2D rectangles and 1D line segments. Each primitives can be decomposed into several lower dimensional shapes. Parsing starts from detection of line segments in the 2D image space as shown in Fig.7(a). The composition of the geometric entities is coded by a series of AND rules where the relations between children nodes are set to a constraint within a threshold. The threshold is set to 5 pixels in the image, which means we tolerate 5 pixels offset between those rigidly combined components. The OR rule also plays a role by representing alternative ways of composition under different the view points. The production rules of geometric decomposition is illustrated in Fig.3 and the result of geometric decomposition is shown in Fig.8.

3.5 The appearance space

We define the appearance model by applying the idea of *analysis-by-synthesis*. In the functional space and the geometric space, we specify how the underlying causes generate a scene image. There is still a gap between the synthesized scene and the observed image, because we can not render a real image without knowing the accurate lighting condition and material parameters. In order to fill this gap, we make use of discriminative approaches: a line segment detector by Von Gioi et al (2010), a foreground detector by Hedau et al (2009) and a surface orientation detector by Lee et al (2009) to produce a line map $Al(\mathcal{I})$, a foreground map $Af(\mathcal{I})$ and a surface orientation map $Ao(\mathcal{I})$ respectively. We evaluate our model by calculating the pixel-wise difference between the maps generated by our model and the maps from bottom-up detection as shown at the bottom of Fig.2 (a).

$$P(\mathcal{I}|\mathcal{G}) \propto \exp(\lambda[d(\mathrm{Al}(\mathcal{G}), \mathrm{Al}(\mathcal{I})) + d(\mathrm{Af}(\mathcal{G}), \mathrm{Af}(\mathcal{I})) + d(\mathrm{Ao}(\mathcal{G}), \mathrm{Ao}(\mathcal{I}))]) \quad (9)$$

As shown in the Fig.2, we decompose the scene structure according to the grammar rules, and project all the

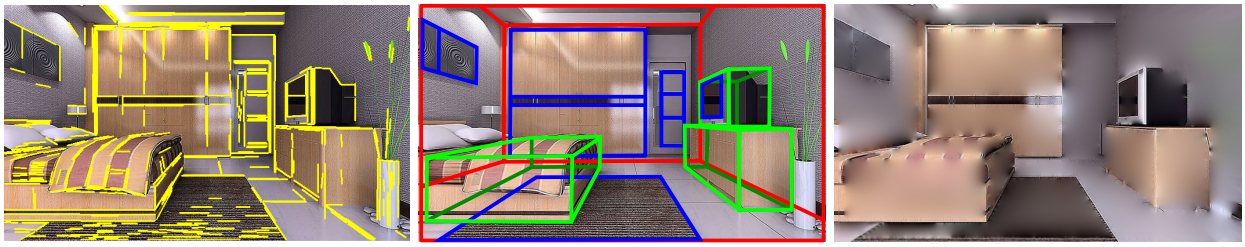


Fig. 7 Input image and output result of the geometric parsing.

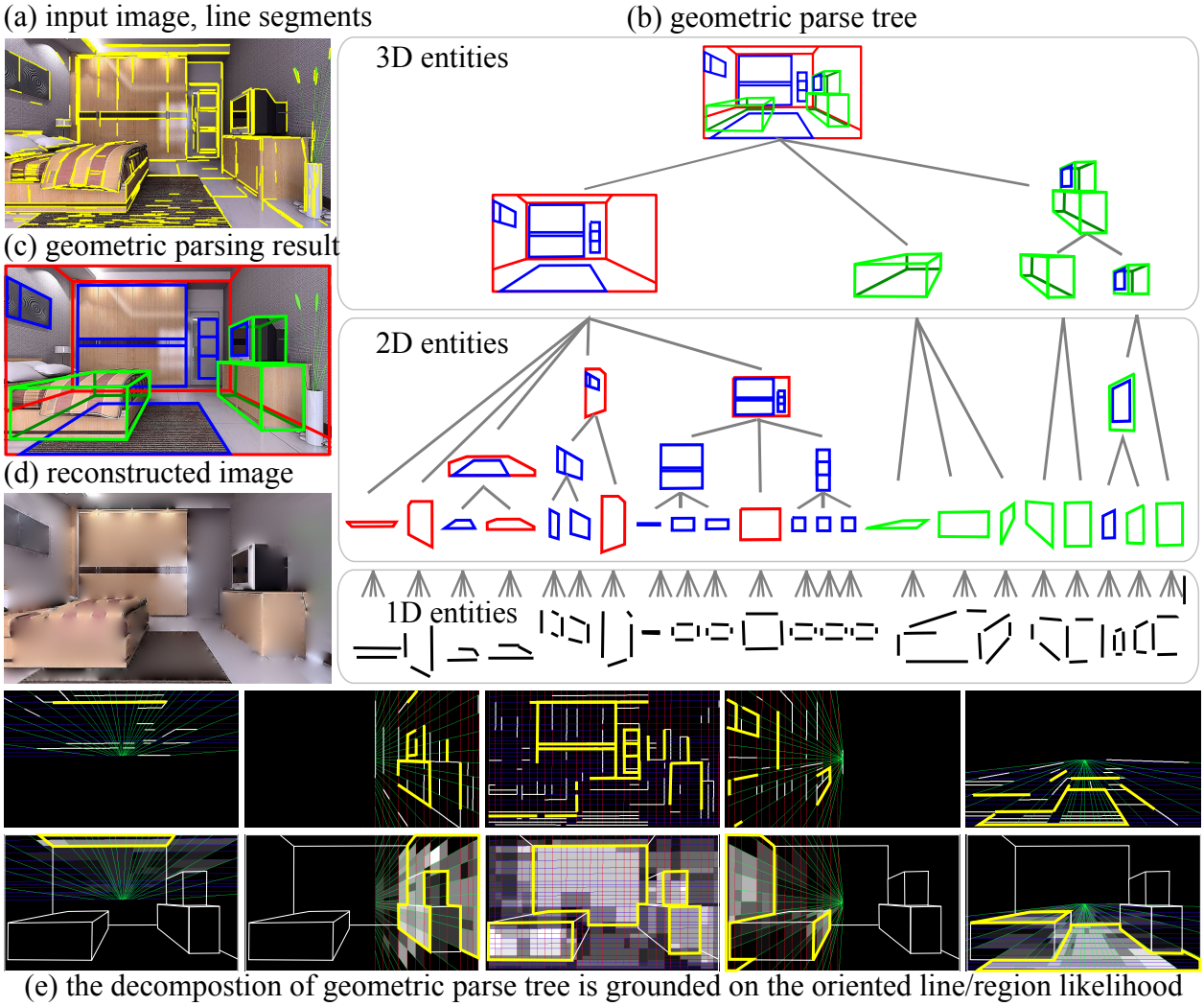


Fig. 8 The decomposition of geometric entities and the quantization of image likelihood.

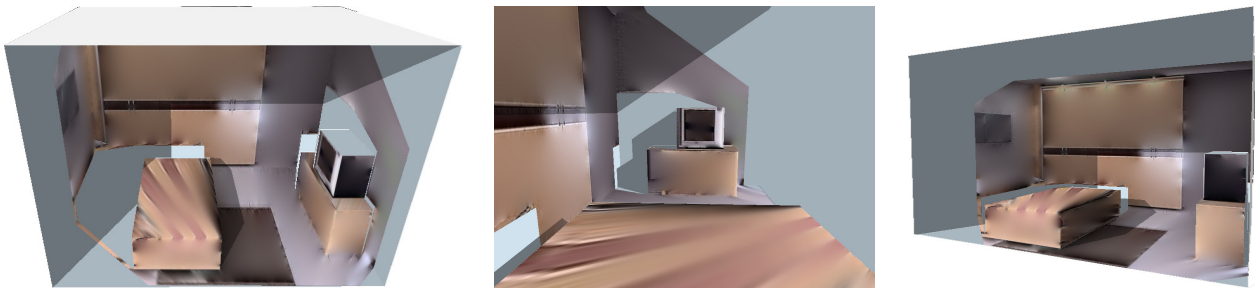


Fig. 9 3D synthesis of novel views based on the parse result.

terminal primitives to 2D surface with respect to five normal directions: facing down, facing left, facing front, facing right and facing up. The surface facing back is not visible from the camera position.

In order to properly quantize the geometric space and speed up the computation, we first group detected line segment into three main groups with respect to three vanishing points. And we further group the line segments into a series of rays pointing from the vanishing points to each line segments. We enforce the angle between two nearby rays larger than 2°, therefore line segments along same orientation will be grouped together. We will also interpolate rays between two nearby rays if the angle between them are larger than 5°. Any two groups of rays will form a oriented mesh as shown in Fig.2.(e) (Please zoom 800% in to see the detail of rays). This quantization process guarantee that each detected line will be represented by several pieces of edges on the mesh, and each pixel will fall into a cell as well. In this way, the line/region likelihood of bottom detection is stored in the quantized meshes for each surface orientation. The brighter the intensity the higher the likelihood for each cell.

In Fig.2.(e), the yellow lines on the upper penal represent the activated line segments. The line segment is activated when the geometric parsing result (c) match with the bottom-up detection result (a). The edge probability measures how many line segments are activated, which implicit encourages more line segment to be explained by final parsing result. the region with yellow boundary on the lower penal represent the activated surface region. A surface region is activated only the surface orientation is matched with geometric parsing results on (c) with a consideration of depth ordering. The depth ordering guarantee the occluded region will not affect the likelihood of parsing result. Therefore, the quantization of image likelihood not only accelerates the inference process by a lookup table of precomputation, but also avoids the double counting of the shared edges and the occluded regions.

3.6 Single View 3D Scene Reconstruction

Another important component of our model is the recovery of 3D geometric measure in the real world scale from the parsing result. It enables us to utilize the general knowledge about object affordance and contextual relations to identify functional objects and groups as discussed before.

Camera calibration: We cluster line segments to find three vanishing points whose corresponding dimensions are orthogonal to each other Hedau et al (2009). The vanishing points are then used to determine the intrinsic and extrinsic calibration parameters Criminisi et al (2000); Hartley and Zisserman (2004). We assume that the aspect ratio is 1 and there is no skew. Any pair of finite vanishing points can be used to estimate the focal length. If all three vanishing points are visible and finite in the same image, then the optical center can be estimated as the orthocenter of the triangle formed by the three vanishing points. Otherwise, we set the optical center to the center of an image. Once the focal length and optical center has been determined, the camera rotational matrix can be estimated accordingly Hartley and Zisserman (2004).

3D reconstruction. We now present how to back-project a 2D structure to the 3D space and how to derive the corresponding coordinates. Considering a 2D point p in an image, there is a collection of 3D points that can be projected to the same 2D point p. This collection of 3D points lays on a ray from the camera center $C = (Cx, Cy, Cz)^T$ to the pixel $p = (x, y, 1)^T$. The ray $P(\lambda)$ is defined by $(X, Y, Z)^T = C + \lambda R^{-1} K^{-1} p$, where λ is the positive scaling factor that indicates the position of the 3D point on the ray. Therefore, the 3D position of the pixel lies at the intersection of the ray and a plane (the object surface). We assume a camera is 4.5ft high. By knowing the distance and the normal of the floor plane, we can recover the 3D position for each pixel with the math discussed above. Any other plane contacting the floor can be inferred by its contact point with the floor. Then we can gradually recover the whole scene by repeating the process from the bottom up. If there is any object too close to the camera to see the bottom, we will put it 3 feet away from the camera.

4 Problem inference

We design a top-down/bottom-up algorithm to infer an optimal parse tree pt. The compositional structure of the continuous geometric parameters and discrete functional labels introduces a large solution space, which is infeasible to enumerate all the possible explanations. Neither the sliding windows (top-down) nor the binding (bottom-up) approaches can handle such an enormous number of configurations independently.

In this paper, we design a reversible jump MCMC algorithms to construct the parsing tree and re-configures it dynamically using a set of moves. Formally, our scene parsing algorithm simulates a Markov chain $\mathcal{MC} =< \Omega, v, \mathcal{K} >$ with kernel $\mathcal{K}$ in space Ω and with probability v for the starting state. We specify stochastic dynamics by defining the transition kernels of revisable jumps. For each Markov chain move is defined by a kernel with a transition matrix $\mathcal{K}(pt^*|pt : I)$, which represents the

probability that the Markov chain make a transition from state pt to pt^* when a move is applied.

The kernels are constructed to obey the detailed balance condition:

$$p(pt|I)\mathcal{K}(pt^*|pt:I) = p(pt^*|I)\mathcal{K}(pt|pt^*:I). \quad (10)$$

Kernels which change the graph structure are grouped into reversible pairs. For example, the kernel for node creation $\mathcal{K}_+$ is paired with the kernel for node deletion $\mathcal{K}_-$ to form a combined move of node switch. To implement the kernel, at each time step the algorithm randomly selects the choice of move and then uses kernel $\mathcal{K}(pt^*|pt:I)$ to select the transition from state pt to state pt^*. Note that the probability $\mathcal{K}(pt^*|pt:I)$ depend on the input image I. This distinguishes our algorithms as a Data-Driven MCMC from conventional MCMC computing (Tu and Zhu (2002); Tu et al (2005)).

The Kernel is designed using proposal probabilities and correspondent acceptance probability.

$$\mathcal{K}(pt^*|pt:I) = Q(pt^*|pt:I)\alpha(pt^*|pt:I) \quad (11)$$

The acceptance probability follows:

$$\alpha(pt \to pt^*) = min\{1, \frac{Q(pt|pt^*,I)}{Q(pt^*|pt,I)} \cdot \frac{P(pt^*|I)}{P(pt|I)}\} \quad (12)$$

The Metropolis-Hasting form ensures that the Markov chain search satisfies the detailed balance principle. A simulated annealing technology is also used to find the maximum of complex posteriori distribution with multiple peaks while other approaches may trap the algorithm at local optimal peaks.

Data: an input 2D image
Result: an output parse tree
Initialization;
while *the rejection time larger than K* **do**
  Choose one of the following moves randomly;
  – adding a geometric entities
  – switching the functional label of a geometric entities
  – removing a geometric entities

  if *adding/removing a non-terminal node* **then**
    Recursively adding/removing its children;
  end
  Calculate the posterior probability and validate the solution by projecting the 3D parse tree structure on the 2D image plane;
  Accept/reject the new parse tree with the acceptance probability;
end
Return the parse tree with the highest posterior;

Algorithm 1: Inference algorithm

In this paper, we handle three types of moves: adding, removing, switching. The adding move attaches a sub-tree to the current parse tree; the removing move deletes a sub-tree from the current parse tree; the switching move changes the the attribute of a tree node. If the switching move involve an adding move and an removing move of a sub-tree, then it is a combined move as we discussed above. These three types of moves are alternated throughout inference.

To simplify the problem, we detect all the possible geometric entities in a bottom-up initialization step. Therefore, all the moves are revisable jumps within the discrete state-space. The diffuse move of fine tuning the continuous geometric parameters will dramatically slow down the Markov chain search. We summarize the entire process here, and describe the details in the rest of this section.

4.1 Initialization

The algorithm starts from detecting straight line segments by Von Gioi et al (2010). Based on the Manhattan assumption, we group the line segments into N groups, each of which is correspondent to a vanishing point. We then select three dominate orthogonal vanishing point to build our coordinate system. We assume the camera parameters are reliably calibrated in this step, the calibration algorithm is discussed in Sect.3.6.

As illustrated in Fig.8, we incrementally group noisy line segment into geometric structures *i.e.*2D rectangles and 3D cuboids by filtering the entities based on geometric AND rules layer by layer. The rectangles are formed by filtering over the combinations of two pairs of parallel lines or T junctions. Similarly, the cuboids are formed by filtering over the combinations of two hinged rectangles.

We design a four-step MCMC algorithm that enables a Markov chain travel up and down through the FGA hierarchy: $\mathcal{A} \to \mathcal{G} \to \mathcal{F} \to \mathcal{G} \to \mathcal{A}$. In each iteration, the algorithm proposes a new parse tree pt^* based on the current one pt according to the proposal probability.

4.2 Bottom-up appearance-geometry (AG) step

With all the geometric entities detected and saved in a pool and all the likelihood stored in a lookup table as mentioned before, the proposed probability for a new geometric entity g^* is defined as the probability to choose an entity from the pool:

$$Q_1(g^*|\mathcal{I}_\Lambda) = \frac{P_A(\mathcal{I}_\Lambda|g^*)P(g^*)}{\int_{g\in Gp} P_A(\mathcal{I}_\Lambda|g)P(g)} \quad (13)$$

where the $P_A(\mathcal{I}_\Lambda|g)$ is defined in a similar form of likelihood in Eq.9 except that we only calculate the image likelihood within a local patch $\mathcal{I}_\Lambda$. The likelihood within a local patch can be calculated very fast with the lookup table.

The $P(g)$ characterizes the prior distribution of the geometric size, *i.e.*how likely a geometric entity g can be generated from the functional space.

$$P(g) = \int_{\mathcal{F}} P(\mathcal{F}, g) = \int_{\mathcal{F}} P(\mathcal{F})P(g|\mathcal{F}) \tag{14}$$

Given a functional label, the distribution of geometric size $P(g|\mathcal{F})$ is defined by a Gaussian model. Thus the prior distribution of geometric size $\int_{\mathcal{F}} P(\mathcal{F})P(g|\mathcal{F})$ is a mixture of a large number of Gaussians. And $P(\mathcal{F})$ is a hyperprior of mixture coefficients. It is worth noting that this proposal probability Q_1 is independent of the current parse tree pt. Therefore we can precompute the proposal probability for each possible geometric proposal, which dramatically reduces the computational cost of the chain search.

4.3 Bottom-up geometry-function (GF) step

This step assigns functional labels to the 3D geometric entity detected in the G space. The switching of functional labels can be happened in any layers of the functional parse tree as shown in Fig.2, which include switching of scene category, switching functional group, switching of functional object and switching of functional part label. The proposal probability of switching an functional label f^* on the functional parse tree is defined as

$$Q_2(f^*|pa, cl) = \frac{P(cl|f^*)P(f^*|pa)}{\int_f P(cl|f)P(f|pa)} \tag{15}$$

where the cl are the children of f^*, and pa is the parent of f^* on the current parse tree pt. In this way, the probability describes the compatibility of the functional label f^* with its parent pa and its children cl on the tree. With the geometry primitives fixed on the bottom, this proposal makes the chain jumping in the functional space to find a better functional explanation for these primitives. The search on the functional space is fast since the functional label space is small. With the Markov property on the tree, $Q_2(f^*|pa, cl)$ is equivalent to the marginal probability $P(f^*|pt)$.

4.4 Top-down function-geometry (FG) step

This step fills in the missing object in a functional group or the missing part in a functional object. For example, once a bed is detected, the algorithm will try to propose nightstands beside it by drawing samples from the geometric prior and the contextual relations. The problem of sampling with complex constraints was carefully studied by Yeh et al.Yeh et al (2012). Fig.6 shows some typical samples. The proposal probability $Q_3(g^*|\mathcal{F})$ of a new geometric primitive g^* is just a Gaussian distribution for geometric size described in Sect.3.2 and a Gaussian distribution of contextual relation described in Sect.3.3. The contextual relation handles relative position of the new primitive given its existing neighbor.

Here, we can see that $Q_1(\mathcal{I} \to \mathcal{G})$ proposes g^* by the bottom-up image detection, and $Q_3(\mathcal{F} \to \mathcal{G})$ proposes g^* by the top-down functional prediction. They are two kinds of approximation of the marginal distribution $P(g^*|pt)$.

On the other hand, removing an existing tree node is relatively a simple problem. We proposes to delete a geometric primitive with uniform probability. Both the add and delete operation will trigger GF step of re-assigning a functional label.

4.5 Top-down geometry-appearance (GA) step

This step projects the 3D geometric model to the 2D image plane with respect to the relative depth order and camera parameters. The projection is a deterministic step as described in Sect.3.5. It generates the image feature maps used to calculate the overall likelihood in Eq.9. The image features are shown at the bottom of Fig.2 (a) . The algorithm will calculate the acceptance probability according to the proposal probability of previous three steps and posterior probability according to the Eq.12.

5 Experiments

Our algorithm has been evaluated on the UIUC indoor dataset Hedau et al (2009), the UCB dataset Del Pero et al (2011), and the SUN dataset Xiao et al (2010). The UCB dataset contains 340 images and covers four cubic objects (bed, cabinet, table and sofa) and three planar objects (picture, window and door). The ground-truths are provided with hand labeled segments for geometric primitives. The UIUC indoor dataset contains 314 cluttered indoor images and the ground-truth is two label maps of the background layout with/without foreground objects. We picked two categories in the SUN dataset: the bedroom with 2119 images and the living room with 2385 images. This dataset contains thousands of object labels and was used to train our functional model as discussed in Sect.3.2

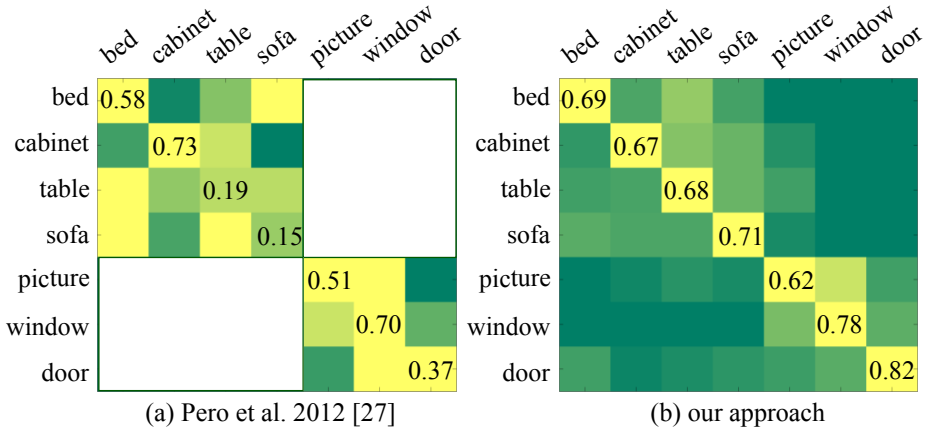


Fig. 10 The confusion matrix of functional object classification on the UCB dataset.

Quantitative evaluation:

We first compared the confusion matrix of functional object classification rates among the successfully detected objects on the UCB dataset as shown in Fig.10. The state-of-the-art work by Del Pero et al (2012) performed slightly better on the cabinet category, but our method get better performance on the table and sofa categories. This is mainly attributed to our fine-grained part model and functional groups model. It is worth noting that our method reduced the confusion between the bed and the sofa. Because we also introduced the hidden variables of scene categories, which help to distinguish between the bedroom and living room according to the objects inside.

In Table.1, we compared the precision and recall of functional object detection with Del Pero et al (2012). The result shows our top-down process did not help the detection of planner objects. But it largely improves the accuracy of cubic object detection from 30.8% to 34.8% with the recall from 24.3% to 29.7%.

In Table.2, we also test our algorithm on the UCB dataset and the UIUC dataset together with five state-of-the-art algorithms: Hedau et al (2009), Wang et al (2010), Lee et al (2010), Del Pero et al (2011) and Del Pero et al (2012). The results show the pixel-level segmentation accuracy of proposed algorithms not only significantly widens the scope of indoor scene parsing algorithm from the segmentation and 3D recovery to the functional object recognition, but also yields improved overall performance.

Qualitative evaluation:

Some experimental results on the UIUC and the SUN datasets are illustrated in Fig.11. The green cuboids are cubic objects proposed by the bottom-up AG step, and the cyan cuboids are the cubic objects proposed by the top-down FG step. The blue rectangles are the detected planar objects, and the red boxes are the background layouts. The functional labels are given to the right of each image. Our method has detected most of the indoor objects, and recovered their functional labels very well. The top-down predictions are very useful to detect highly occluded nightstands as well as the headboards of the beds. As shown in the last row, our method sometimes failed to detect certain objects. The bottom left image fails to identify the drawer in the left but a door. In the middle bottom image, the algorithm failed to accurately locate the mattress for this bed with a curtain. The last image is a kind of typical failure example due to the unusual camera position. We assumed the camera position is 4.5 feet high, while this camera position in this image is higher than our assumptions. As a result, the algorithm detected a much larger bed instead.

Table 1 The precision (and recall) of functional object detection on the UCB dataset.

UCB dataset	planar objects	cubic objects
Del Pero et al (2012)	27.7% (19.7%)	31.0% (20.1%)
Ours w/o top-down	28.1%(18.5%)	30.8% (24.3%)
Ours w/ top-down	28.1%(18.7%)	34.8% (29.7%)

Table 2 The pixel classification accuracy of background layout segmentation on the UCB dataset and the UIUC dataset.

	UCB dataset	UIUC dataset
Hedau et al (2009)	-	78.8%
Wang et al (2010)	-	79.9%
Lee et al (2010)	-	83.8%
Del Pero et al (2011)	76.0%	73.2%
Del Pero et al (2012)	81.6%	83.7%
Our approach	82.8%	85.5%

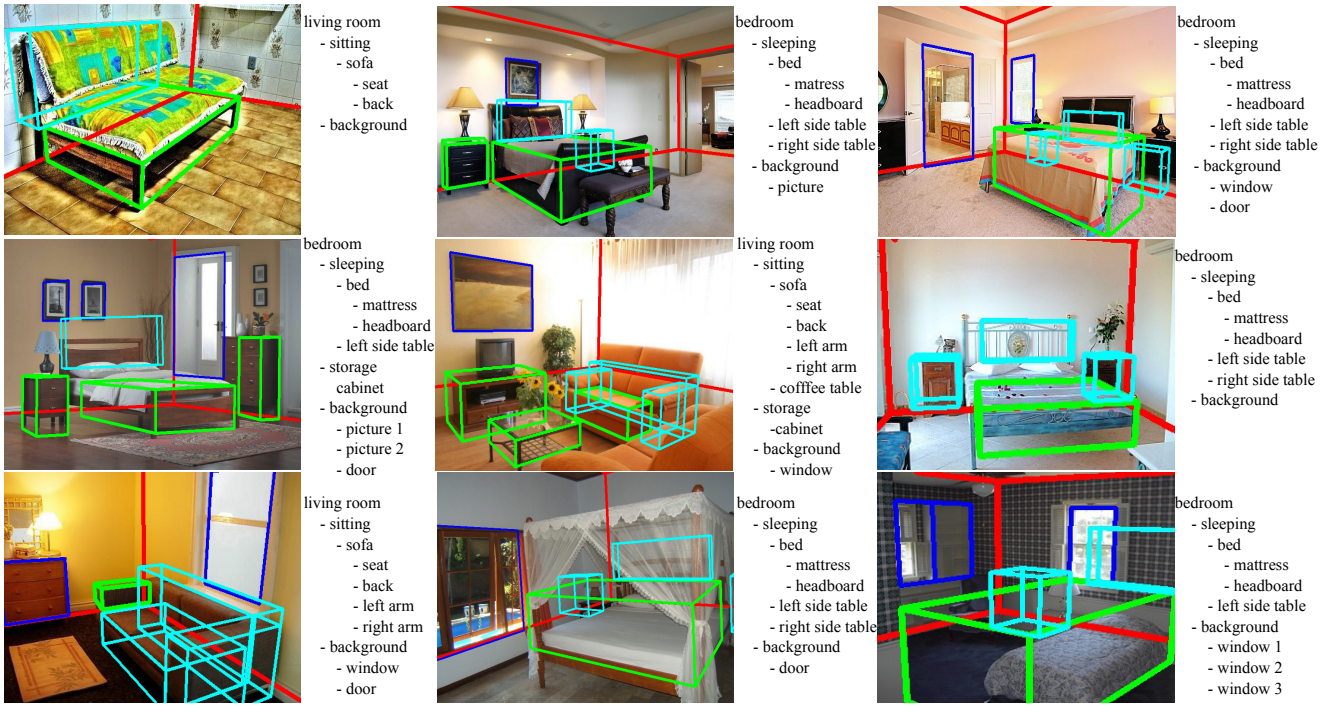


Fig. 11 Parsing results include cubic objects (green cuboids are detected by bottom-up step, and cyan cuboids are detected by top-down prediction), planar objects (blue rectangles), background layout (red box). The parse tree is shown to the right of each image.

6 Conclusion

This paper presents a stochastic grammar built on a function-geometry-appearance (FGA) hierarchy. Our approach parses an indoor image by inferring the object function and the 3D geometry. The functionality defines an indoor object by evaluating its "affordance". The affordance measures how much an object can support the corresponding human action, e.g. a bed is able to support the action of sleep. We found it is effective to recognize certain object functions according to its 3D geometry regardless of observing the actions.

The function helps to build an intrinsic bridge between the man-made object and the human action, which can motivate other interesting studies in the future: functional objects/areas in a scene attract human's needs and/or intentions; other risky areas (like sharp corners) apply repulsive force to human actions. As a result, a parsed scene with functional labels defines a human action space, and it also helps to predict people's behavior by making use of the function cues. On the other hand, given an observed action sequence, it is possible to accurately recognize the functional objects associated with the rational actions.

Acknowledgements This work is supported by ONR MURI grant N00014-10-1-0933 and DARPA MSEE grant FA 8650-11-1-7149.

References

Bar-aviv E, Rivlin E (2006) Functional 3d object classification using simulation of embodied agent. In: BMVC

C Liu JY, Torralba A (2011) Nonparametric scene parsing via label transfer. IEEE Trans on Patt Anal Mach Intell (TPAMI)

Choi MJ, Lim JJ, Torralba A, Willsky AS (2010) Exploiting hierarchical context on a large database of object categories. In: CVPR

Choi W, Chao Y, Pantofaru C, Savarese S (2013) Understanding indoor scenes using 3d geometric phrases. In: CVPR

Criminisi A, Reid I, Zisserman A (2000) Single view metrology. International Journal of Computer Vision (IJCV) 40(2):123–148

Del Pero L, Guan J, Brau E, Schlecht J, Barnard K (2011) Sampling bedrooms. In: CVPR

Del Pero L, Bowdish J, Fried D, Kermgard B, Hartley E, Barnard K (2012) Bayesian geometric modeling of indoor scenes. In: CVPR, pp 2719–2726

Del Pero L, Bowdish J, Kermgard B, Hartley E, Barnard K (2013) Understanding bayesian rooms using composite 3d object models. In: CVPR

Delage E, Lee H, Ng A (2007) Automatic single-image 3d reconstructions of indoor manhattan world scenes. Robotics Research p 305321

Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. IJCV 88(2):303–338

Fidler S, Dickinson S, Urtasun R (2012) 3d object detection and viewpoint estimation with a deformable 3d cuboid model. In: NIPS

Fu KS (1982) Syntactic pattern recognition and applications. Prentice-Hall

Gibson JJ (1977) The Theory of Affordances. Lawrence Erlbaum
Grabner H, Gall J, Gool LV (2011) What makes a chair a chair? In: CVPR
Gupta A, Efros AA, Hebert M (2010) Blocks world revisited: Image understanding using qualitative geometry and mechanics. In: European Conference on Computer Vision(ECCV)
Gupta A, Satkin S, Efros AA, Hebert M (2011) From 3d scene geometry to human workspace. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Washington, DC, USA, pp 1961–1968
Han F, Zhu SC (2004) Bayesian reconstruction of 3d shapes and scenes from a single image. In: Proc. IEEE Workshop on Perceptual Organization in Computer Vision
Han F, Zhu SC (2009) Bottom-up/top-down image parsing with attribute grammar. PAMI
Hartley RI, Zisserman A (2004) Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, ISBN: 0521540518
Hedau V, Hoiem D, Forsyth D (2009) Recovering the spatial layout of cluttered rooms. In: ICCV
Hedau V, Hoiem D, Forsyth D (2010) Thinking inside the box: Using appearance models and context based on room geometry. In: ECCV
Hedau V, Hoiem D, Forsyth D (2012) Recovering free space of indoor scenes from a single image. In: CVPR
Hejrati M, Ramanan D (2012) Analyzing 3d objects in cluttered images. In: Bartlett P, Pereira F, Burges C, Bottou L, Weinberger K (eds) Advances in Neural Information Processing Systems 25, pp 602–610
Hoiem D, Efros A, Hebert M (2009) Automatic photo pop-up. TOG 31(1):59–73
Hu W (2012) Learning 3d object templates by hierarchical quantization of geometry and appearance spaces. In: CVPR, pp 2336–2343
Isola P, Liu C (2013) Scene collaging: analysis and synthesis of natural images with semantic layers. In: IEEE International Conference on Computer Vision (ICCV)
Lafferty JD, McCallum A, Pereira FCN (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. ICML pp 282–289
Lee D, Hebert M, Kanade T (2009) Geometric reasoning for single image structure recovery. In: CVPR
Lee D, Gupta A, Hebert M, Kanade T (2010) Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces advances in neural information processing systems. Cambridge: MIT Press pp 609–616
Lim JJ, Pirsiavash H, Torralba A (2013) Parsing ikea objects: Fine pose estimation. In: IEEE International Conference on Computer Vision (ICCV)
Lin D, Fidler S, Urtasun R (2013) Holistic scene understanding for 3d object detection with rgbd cameras. In: ICCV
Mobahi H, Zhou Z, Yang AY, Ma Y (2011) Holistic 3d reconstruction of urban structures from low-rank textures. In: Proceedings of the International Conference on Computer Vision - 3D Representation and Recognition Workshop, pp 593–600
Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. International Journal of Computer Vision (IJCV)
Payet N, Todorovic S (2011) Scene shape from textures of objects. In: CVPR
Pepik B, Gehler P, Stark M, Schiele B (2012) 3d2pm - 3d deformable part models. In: ECCV, Firenze, Italy
Porway J, Zhu SC (2010) Hierarchical and contextual model for aerial image understanding. IJCV 88(2):254–283
S Lazebnik CS, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
S N Parizi JO, Felzenszwalb P (2012) Reconfigurable models for scene recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
S Wang YW, Zhu S (2012) Hierarchical space tiling in scene modeling. In: Asian Conf. on Computer Vision (ACCV)
Satkin S, Hebert M (2013) 3dnn: Viewpoint invariant 3d geometry matching for scene understanding. In: ICCV
Satkin S, Lin J, Hebert M (2012) Data-driven scene understanding from 3d models. In: BMVC
Saxena A, Sun M, Ng A (2009) Make3d: Learning 3d scene structure from a single still image. PAMI 31(5):824–840
Schwing AG, Hazan T, Pollefeys M, Urtasun R (2012) Efficient structured prediction for 3d indoor scene understanding. In: CVPR
Schwing AG, Fidler S, Pollefeys M, Urtasun R (2013) Box in the box: Joint 3d layout and object reasoning from single images. In: ICCV
Stark L, Bowyer K (1991) Achieving generalized object recognition through reasoning about association of function to structure. PAMI 13:10971104
Tighe J, Lazebnik S (2013a) Finding things: image parsing with regions and per-exemplar detectors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
Tighe J, Lazebnik S (2013b) Superparsing: scalable nonparametric image parsing with superpixels. International Journal of Computer Vision (IJCV)
Tu Z, Zhu SC (2002) Image segmentation by data-driven markov chain monte carlo. PAMI 24(5):657–673
Tu Z, Chen X, Yuille A, Zhu S (2005) Image parsing: unifying segmentation, detection and recognition. IJCV 63(2):113–140
Von Gioi R, Jakubowicz J, Morel JM, Randall G (2010) Lsd: A fast line segment detector with a false detection control. TPAMI 32(4):722–732
Wang H, Gould S, Koller D (2010) Discriminative learning with latent variables for cluttered indoor scene understanding. In: ECCV, pp 497–510
Wei P, Zhao Y, Zheng N, Zhu SC (2013) Modeling 4d human-object interactions for event and object recognition. In: ICCV
Xiang Y, Savarese S (2012) Estimating the aspect layout of object categories. In: CVPR
Xiao J, Hays J, Ehinger K, Oliva A, Torralba A (2010) Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR, pp 3485 –3492
Xiao J, Russell B, Torralba A (2012) Localizing 3d cuboids in single-view images. In: Bartlett P, Pereira F, Burges C, Bottou L, Weinberger K (eds) NIPS, pp 755–763
Yeh YT, Yang L, Watson M, Goodman ND, Hanrahan P (2012) Synthesizing open worlds with constraints using locally annealed reversible jump mcmc. ACM Trans Graph 31(4):56:1–56:11
Zhao Y, Zhu SC (2011) Image parsing via stochastic scene grammar. In: NIPS
Zhao Y, Zhu SC (2013) Scene parsing by integrating function, geometry and appearance models. In: CVPR
Zhu SC, Mumford D (2007) A stochastic grammar of images. Foundations and Trends in Computer Graphics and Vision 2(4):259–362