Title
Analysis by Synthesis: 3D Image Parsing Using Spatial Grammar and Markov Chain Monte Carlo

Author
Qi, Siyuan

Publication Date
2015

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

Analysis by Synthesis:

3D Image Parsing Using Spatial Grammar and Markov Chain Monte Carlo

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Computer Science

by

Siyuan Qi

2015

Abstract of the Thesis

Analysis by Synthesis:

3D Image Parsing Using Spatial Grammar and Markov Chain Monte Carlo

by

Siyuan Qi

Master of Science in Computer Science

University of California, Los Angeles, 2015

Professor Song-Chun Zhu, Chair

Scene understanding is a fundamental problem in computer vision research. We address this problem in an *"analysis by synthesis"* fashion - explain observed data (an 2D image) according to a set of spatial grammar (describes the underlying functional arrangement and 3D geometric structure of a scene) that generate it. The inference process is carried out in a Bayesian framework. The posterior probability includes a prior probability reflecting the knowledge of indoor 3D scene structure encoded by grammar, and a likelihood that evaluates the accuracy of the re-projected image and the physical plausibility. The most reasonable explanation of the image is given by a parse tree that maximizes the posterior probability, and it is found by reversible-jump Markov Chain Monte Carlo sampling.

The thesis of Siyuan Qi is approved.

Demetri Terzopoulos

Yingnian Wu

Song-Chun Zhu, Committee Chair

University of California, Los Angeles

2015

Table of Contents

List of Figures

Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Prof. Song-Chun Zhu for the continuous support of my study and research, for his motivation, enthusiasm, and immense knowledge. I would never have been able to finish the thesis without the guidance and help from my thesis committee members.

I would like to thank my friends in the lab and in particular, I want to extend my gratefulness to Yibiao Zhao, who led me into the project and provided insightful comments.

I would also like to thank my family for the support they provided me through my entire life. Without their love and encouragement, I would not have finished this thesis.

CHAPTER 1

Introduction

A central goal of computer vision is to create a computational system that recognizes and understands the world at a level that is comparable to the biological vision system. A large portion of vision literature studied the scene understanding problem as a classification problem. We argue that discriminative approaches that classifies each pixel to a semantic label, are insufficient to understand the geometric and functional relations between objects and the scene. Hence we propose a generative method that explains a scene according to a spatial grammar we defined. An explanation of a scene is given by a parse tree, which integrates high level functional category of objects, layout structures, 3D geometric parameters of object, and camera parameters. A solution of the scene parsing problem is a parse tree that maximizes the posterior probability, which combines a prior probability reflecting high-level knowledge of functional arrangement and 3D scene structure encoded by a stochastic grammar, and a likelihood evaluated by calculating the difference between synthesized image and the observed image. The inference process is accomplished by reversible-jump Markov Chain Monte Carlo sampling. Hence the final result of the proposed approach expresses the high level structure of the scene and explains the scene by understanding.

1.0.1 Motivation

Traditional computer vision tasks such as object detection and labeling have made remarkable progress over the past years. However, recognizing and understanding indoor scenes in the way human perceive the world is still a challenging task. Most computer

vision algorithms are designed to analyze low-level information and features, but they are unable to understand the images based on commonsense knowledge of the 3D world. In order to construct an abstract knowledge representation of the real world, we propose the scene grammar and pose the scene understanding problem as parsing an image, in analogy to natural language parsing problems.

1.0.2 Literature review

Scene representation There are mainly four scene representations: (i) Feature-based representation: A series of early work on exploring feature representation of scenes leads the research of scene understanding, including spatial envelope (gist representation) by Oliva and Torralba [OT01], spatial pyramid matching by S.Lazebnik and Ponce [LSP06] reconfigurable models by S.N.Parizi and Felzenszwalb [POF12], and S. Wang and Zhu [WWZ12]. (ii) Region-based representations: The method of conditional random felds by Lafferty et al [LMP01], and 2D context models by Choi et al [CLT10]. (iii) Non-parametric representation: label transfer by C. Liu and Torralba [LYT11], SuperParsing by Tighe and Lazebnik [TL13b, TL13a], scene collage by Isola and Liu [IL13]. Satkin et al [SLH12], Satkin and Hebert [SH13] applies nearest-neighbor search to 3D scenes. Lim et al [LPT13], Del Pero et al [DBK13] detected indoor objects by matching with fine-grained furniture model. (iv) Block world representation: Block world representation: Representation of 3d blocks enables reasoning about the physical constraints within the 3D scene. Gupta et al (2010) represents the 3D objects as blocks and infers 3D properties of objects such as occlusion, exclusion and stability in addition to labeling surface orientation. They showed that a global 3D prior does improve 2D surface labeling. Hedau et al [HHF09, HHF10, HHF12], Wang et al [WGK10b], Lee et al [LHK09, LGH10], Schwing et al [SHP12, SFP13] parameterized the geometric scene layout of the background and/or foreground blocks. Their models are trained by the Structured SVM (or Latent SVM). Hu [Hu12], Xiao et al [XRT12], Hejrati and Ramanan [HR12], Xiang and Savarese [XS12], Pepik et al [PGS12], Fidler et al [FDU12] designed several variants of the deformable

part-based models to detect 3D entities under different view points.

Object function and affordance. The concept of affordance, which was first proposed by the perceptual psychologist Gibson [Gib77], refers to the perceived fundamental properties of an object that determine how the object could possibly be used. More recently, approaches have been proposed to detect objects based on human interactions. In Wei et al [WZZ13], human activities are annotated by extracting human motion from rgbd video data and used to indirectly identify objects. In Bar-aviv and Rivlin [BR06], Grabner et al [GGG11], they hallucinate embodied agents in the 3D CAD data and depth data respectively, and detected chairs accordingly. Gupta et al [GSE11] proposed an approach to infer the human workable space by adapting human poses to the scene. Lin et al [LFU13], Choi et al [CCP13], Zhao and Zhu [ZZ13] raised holistic approaches to exploits 2D segmentation, 3D geometry, as well as contextual relations between scenes and objects for parsing rgbd and 2D images. Recently, Zhu and Zhao [ZZZ15] proposed a method to analyze the potential tool-use of arbitrary objects by physical simulations.

Single-view 3D reconstruction. In order to achieve a meaningful 3D reconstruction from a single image, assumptions about the scene have to be made and prior knowledge is necessary to regularize the solution. These assumptions include: (i) Sketch smoothness assumption: Han and Zhu [HZ04] first approached the problem by assuming the local sketch smoothness and global scene alignment. (ii) Piece-wise smoothness assumption: Saxena et al [SSN09] presented a fully supervised method to learn a mapping between informative features and depth values under a conditional random field framework. Payet and Todorovic [PT11] proposed a joint model to recognize objects and estimate scene shape simultaneously. (iii) Surface assumption: Hoiem et al [HEH09] 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 [DLN07] proposed a dynamic Bayesian network model to infer the floor structure for autonomous 3D reconstruction from a single indoor image. Mobahi et al [MZY11] 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 [HHF09, HHF10, HHF12], Wang et al [WGK10a], Lee et al [LHK09, LGH10], Schwing et al [SHP12, SFP13], Zhao and Zhu [ZZ11, ZZ13] and Del Pero et al [DGB11, DBF12, DBK13] 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.

Stochastic image grammar. Fu [Fu82] depicted an program of block world scene understanding using grammars. Tu et al [TCY05] introduced the decomposing an image into a hierarchical "parse graph" by a data-driven data-driven Monte Carlo sampling strategy. Zhu and Mumford [ZM07] proposed an AND/OR graph model to represent the compositional structures in vision. Han and Zhu [HZ09] applied grammar rules, in a greedy manner, to detect rectangular structures in man-made scenes. Porway and Zhu [PZ10] proposed an cluster sampling algorithm to parse aerial images. An earlier version of the work presented in this thesis appeared at Zhao and Zhu [ZZ11, ZZ13].

CHAPTER 2

Model

2.1 A 3D representation of visual scenes

Motivated by the goal of understanding the scene as human do, we model the a scene as a functional arrangement of 3D objects by an observation: *objects, especially man-made, are defined by their affordance and actions that they are involved, and scenes are defined by activities and actions that they can provide space for.* Therefore, an indoor scene can be decomposed into different functional groups, and the groups can be further decomposed and the scene forms a hierarchical structure, e.g. a bedroom can be decomposed into a "bed set" and a "dressing set", and a "bed set" can be further decomposed into a "bed", two "nightstands" and an "ottoman", the "dressing set" can be decomposed into a "dressing table" and a "chair". We represent this grammatical semantic meaning of a scene as a tree structure as shown in Figure 2.1. In addition, the geometric information is encoded into this tree structure, as the geometric relations between a parent and a child node are represented by a geometric transformation. Hence this generative model combines the grammar model and the geometric model, and it is able to reconstruct a scene according to the tree.

2.2 Stochastic scene grammar model

The stochastic scene grammar is applied to represent the stochastic distribution of such a structure, namely a *parse tree*. The grammar starts from a root node and ends in a set of terminal nodes. The production rules characterize the compositionality of functional

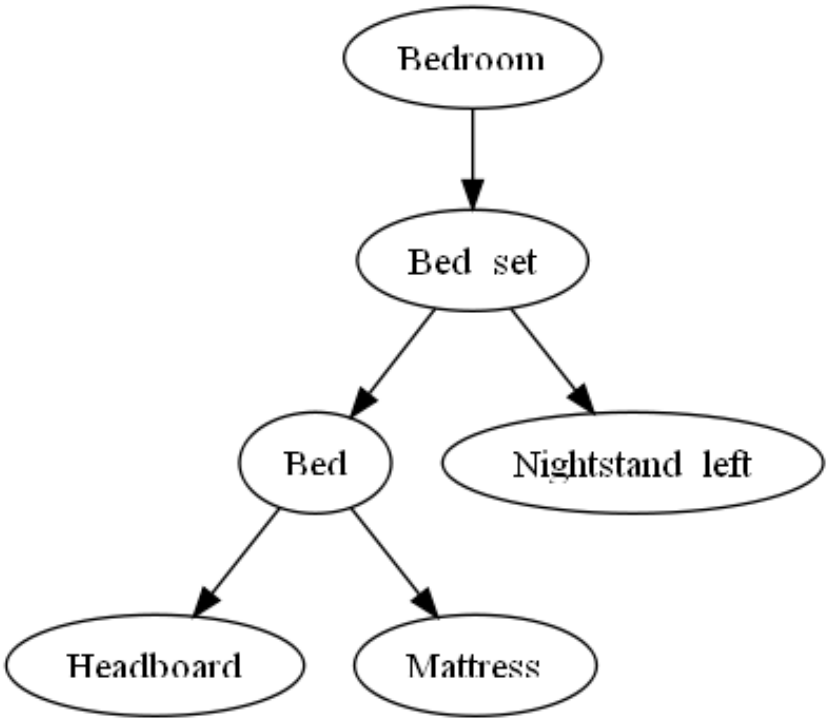

Figure 2.1: An example of a parse tree of a bedroom.

entities on the tree, and a geometric transformation is calculated between each parent-child pair.

The stochastic scene grammar is defined as a four-tuple $G = (S, V, R, P)$, where S is a start symbol; V is a set of nodes which includes the non-terminal nodes V^{NT} and terminal nodes V^T: $V = V^{NT} \cup V^T$; $R = \{r : \alpha \to \beta\}$ is a set of production rules that represent the top-down sampling process from a parent node α to its children node β; $P : p(r) = p(\beta|\alpha)$ is the probability for each production rule.

The posterior probability for a parse tree pt conditioned on an input image I is formulated as

$$p(pt|I) \propto p(pt)p(I|pt) = p_{phy}(pt)p_{grm}(pt)p(I|pt), \tag{2.1}$$

where the prior probability $p(pt)$ is composed of the physical probability $p_{phy}(pt)$ and the grammar probability $p_{grm}(pt)$. $p_{phy}(pt)$ represents the probability that the parse tree is

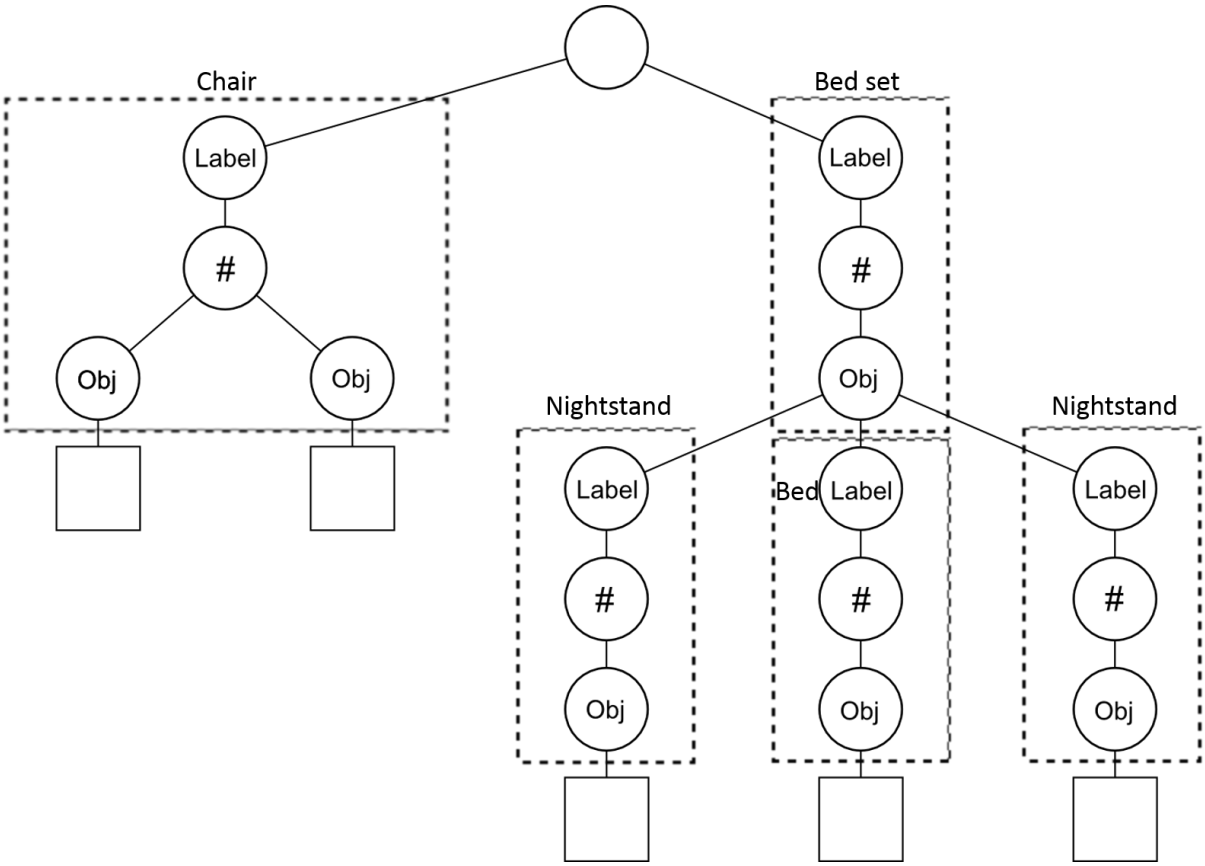

Figure 2.2: An example of a parse tree of a bedroom.

physically plausible, and $p_{grm}(pt)$ represents the probability of generating the parse tree from the root node. $p(I|pt)$ is the likelihood of the image given the parse tree, which is calculated by evaluating the difference between the re-projected and the original image.

2.3 Geometric model

For each object we define an anchor point around which the object can rotate itself, and the geometric transformation between the object and its parent that specifies the position of the child's anchor point under the parent's coordinate system, the child's rotation angle.

By default, we set the anchor point to be the center of the object. we set the default

base orientation for an object so that its back is against the wall of parent that is closest in distance to its sampled position in parent, and a small noise of orientation is sampled and added to the base orientation. Take into consider the fact that most objects lay flat on other objects, we only model the relative orientation/rotation the horizontal plane. Hence for each node, we have 7 parameters: size (l, w, h) (3 DoF), relative position (x, y, z) (3 DoF) and relative orientation θ (1 DoF). We use Gaussian mixture model as their prior distributions: $p(\theta) = \sum_{i=1}^{K} N(\mu_i, \Sigma_i)$. Notice that since the position and orientation parameters are describing the relative geometric relation between a node and its parent, the sampling process is independent from parent.

The absolute position of a point in the child cuboid can be recursively calculated by applying the relative transformations from child to parent. Denote the child as c, parent as p, and the anchor points for child and parent as ac and ap. The coordinate of a point in the child coordinate system X_c can be transformed to a coordinate in the parent coordinate system X_p equation 2.2. Hence the position of a point in the world coordinate system can be calculated recursively.

$$\begin{aligned} X_p &= H_{c\to p} X_c \\ &= H_{ap\to p} H_{c\to ac} X_c \\ &= \begin{bmatrix} cos(\theta) & sin(\theta) & x_{ap} \\ -sin(\theta) & sin(\theta) & y_{ap} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & x_{ac} \\ 0 & 1 & y_{ac} \\ 0 & 0 & 1 \end{bmatrix} X_c \end{aligned} \tag{2.2}$$

The process of generating a child node from a parent node is illustrated in Figure 2.3. (1) The size of the child node is first sampled from a prior distribution according to the label, e.g., bed. (2)-(3) Anchor points are sampled for the child node and the parent node respectively. (4) The child cuboid is placed so that its anchor point matches the parent's anchor point. (5) Find the default orientation of the child node by placing the "back" surface to be the closest surface to a wall [YYT11]. (6) Rotate the child node cuboid by the sampled angle.

Hence, after specifying the probability distributions for each node in a set of grammar, it is straightforward to top-down sample a whole parse tree from the root node. Figure 2.4 shows four examples of random samples of a bedroom.

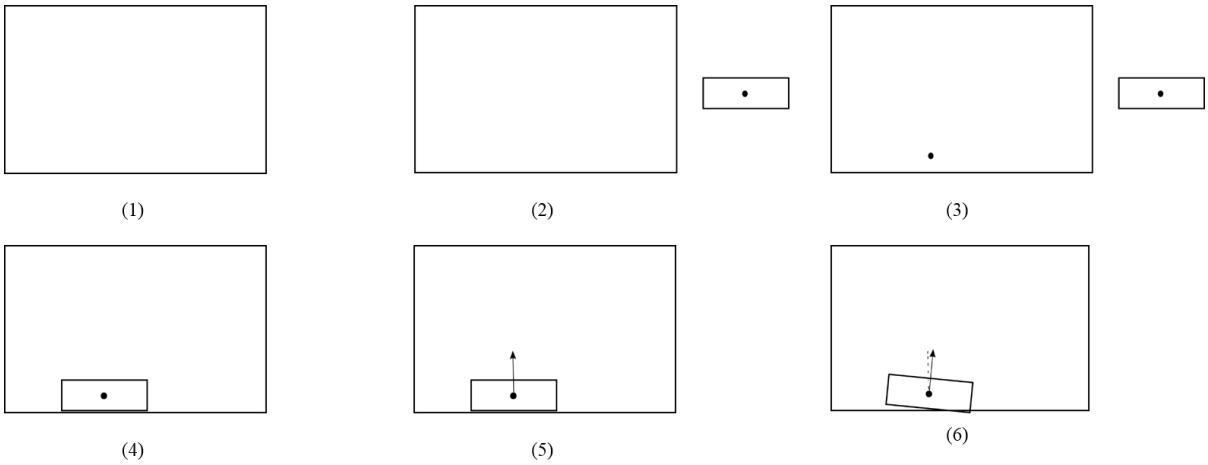

Figure 2.3: Sampling process of a child node from a parent node

2.4 Probability formulation

2.4.1 Prior probability

The prior probability $p(pt)$ can be decomposed into a physical probability $p_{phy}(pt)$ and a grammar probability $p_{grm}(pt)$. The physical probability $p_{phy}(pt)$ is defined as a Boltzmann distribution: $\frac{1}{Z}\exp(-E_{phy}(pt))$, where the energy is proportional to the volume of intersection of cuboids. This encourages less collisions in the physical world. The grammar probability is the probability of top-down sampling a parse tree from the root node according to the grammar:

$$\begin{aligned} p_{grm}(pt) &= p(root) \prod_{r:\alpha\to\beta} p(\beta|\alpha) \\ &= p(root) \prod_{r:\alpha\to\beta} [p(\#\beta|\alpha) \prod_{\beta} p(att_\beta)] \end{aligned} \tag{2.3}$$

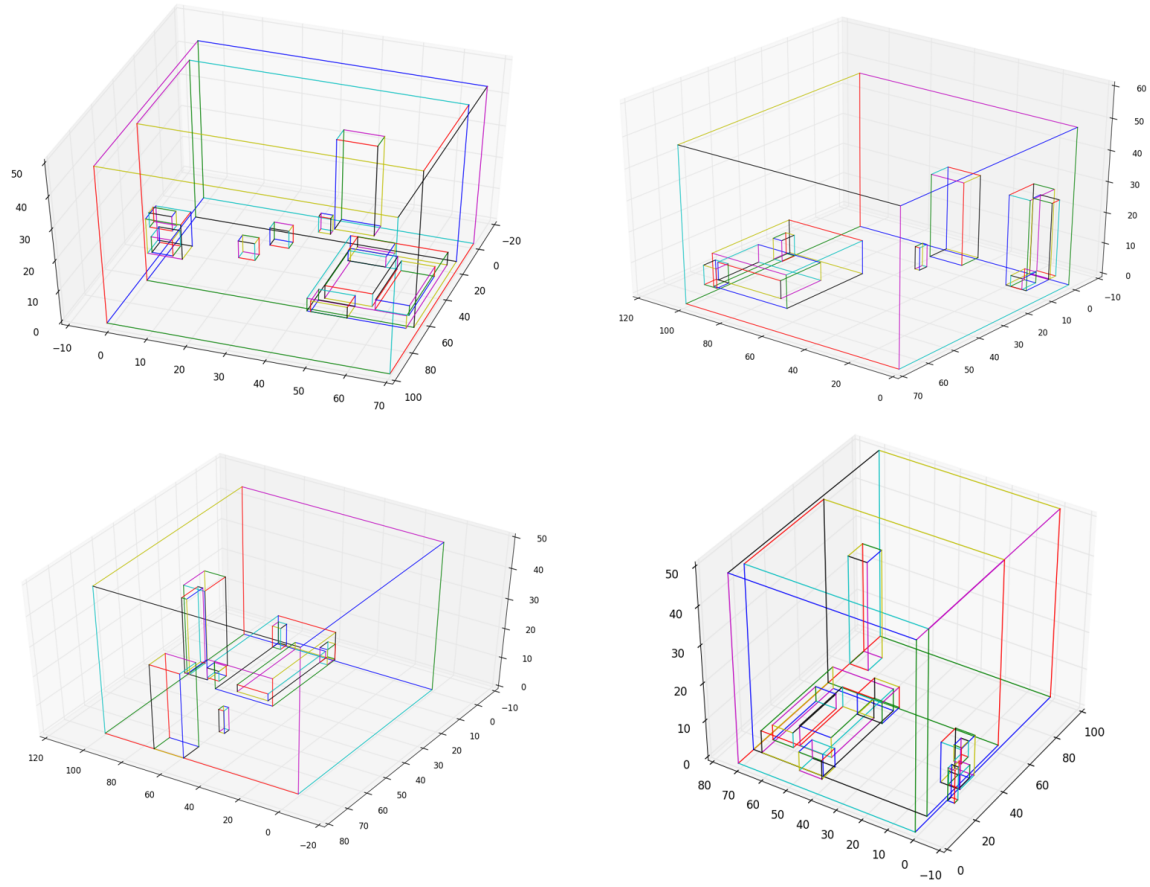

Figure 2.4: Top-down samples of a bedroom

where $p(\#\beta|\alpha)$ is the probability of number of children nodes given parent, and $p(att_\beta)$ is the probability of the attributes (size, position and orientation) of a particular node.

2.4.2 Likelihood

We evaluate the likelihood of the sampled parse tree $p(I|pt)$ by comparing the re-projected image with the approximated ground truth labeling of the original image.

Given the complexity in graphics rendering, such as environmental lighting, object material and textures, it is not practical to infer all the scene parameters. So we propose to render two abstract images - the object label map and orientation map for each image,

which encode semantics and geometry of the scene, and are invariant to lighting factors. The likelihood is formulated as the probability of I given the current parse tree:

$$p(I|pt) = \int_{I_s} p(I|I_s)p(I_s|pt) = \int_{I_s} p(I|I_s)1(I_s = f(pt, c)) = p(I|I_s = f(pt, c)) \tag{2.4}$$

where f is the graphics engine that takes the parse tree pt and camera parameters c and outputs a synthesized abstract images or label maps I_s. The $p(I|I_s)$ is defined as a Boltzmann distribution: $\frac{1}{Z}\exp(-E(I|I_s))$, where the energy is determined by the number of different pixels between the pre-computed labeling maps from recognition and the reprojected labeling maps of the parse tree.

CHAPTER 3

Inference

The algorithm aims to find a parse tree that can best explain the scene. The entire inference process includes a pre-processing step and a sampling step using Monte Carlo Markov Chain (MCMC). The pre-processing step takes the images as input, and estimates the camera parameters, generates proposals for cuboids (object detection), and label maps for likelihood evaluation. The parser then takes the result from the pre-processing step as input, and finally outputs the optimal parse tree and the reconstructed 3D scene. In the parsing step, we apply top-down and bottom-up inference [HZ09, WZ11] during the reversible jump Monte Carlo Markov Chain sampling process. The overall structure is shown in Figure 3.1.

Figure 3.1: A high-level abstraction of the algorithm

3.1 Preprocessing and cuboid proposals

The preprocessing process generates proposals for the next-stage sampling and computes labeling maps for verification of samples during the sampling process. We detect all possible 3D cuboids, 2D labeling of the image and estimate camera parameters from

vanishing points.

First we estimate the 3D geometry of the image. The recovery of 3D geometric measure includes utilizing line segment clustering and vanishing point estimation to determine the intrinsic and extrinsic parameters of the camera, and grouping the line segments to reconstruct 3D cuboid proposals. Each cuboid proposal is assigned with a weight indicating the confidence based on the clustering of line segments. After normalizing the weights, an initial probability distribution for all cuboids detected in the image is built.

Second, we compute the orientation map and segmentation map for the image These label maps serve as evidences to evaluate the likelihood in order to calculate the posterior probability, so that we can decide whether the sample should be accepted or rejected during the MCMC sampling process. An example of the label maps is shown in Figure 3.3.

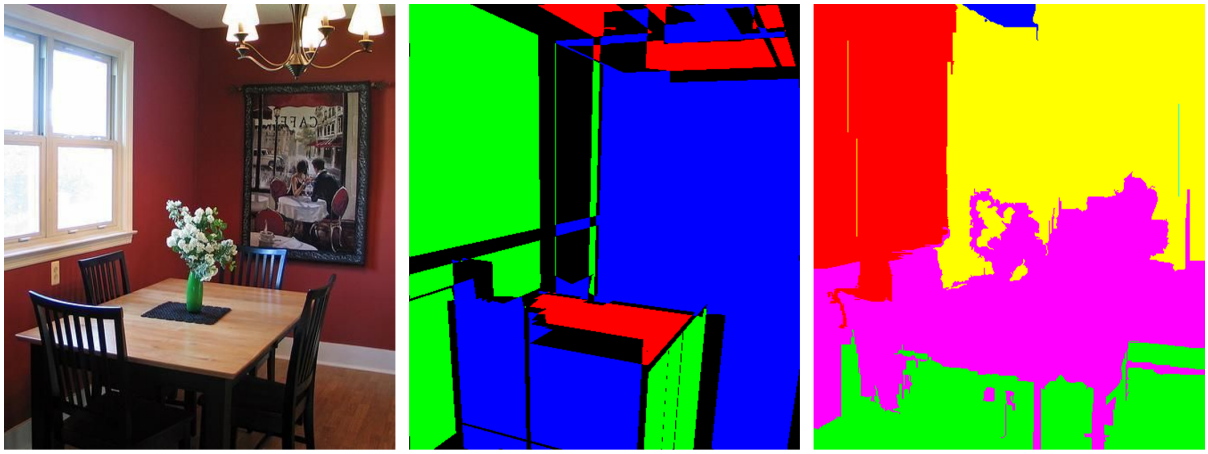

Figure 3.3: Example: orientation map and segmentation map computed from an image of a living room.

Third, we further process the cuboid proposals by building a non-parametric distribution of the cuboid proposals. Originally, all cuboid proposals come from the first step and share the same initial probability distribution. Since different objects have differ-

Figure 3.2: Sampled reconstructed cuboids from line segments detected from image

ent distributions of sizes, we filter all cuboid proposals by the sizes of different objects and combine the score with the original weights as shown in Figure 3.4, to generate different cuboid distributions for specific objects. We represent each cuboid as a six dimensional variable (three dimensions of position and three dimensions of size), and the non-parametric distribution is approximated by a weighted KDE (kernel density estimation):

$$f(x) = \frac{w}{h} \sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

Notice that the landscape of a distribution becomes smoother when the bandwidth for KDE increases. The sampling process would easily get stuck in local minimums if the bandwidth is set too small. Hence it is important to apply an annealing algorithm on the bandwidth and re-estimate the cuboid distributions during the sampling process.

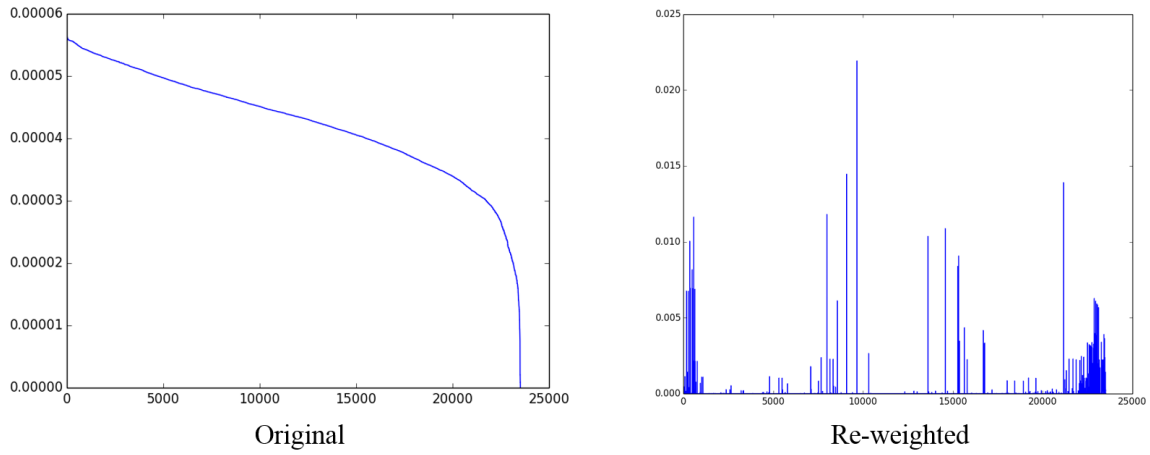

Figure 3.4: Original cuboid proposals and re-weighted cuboid proposals for "Bed"

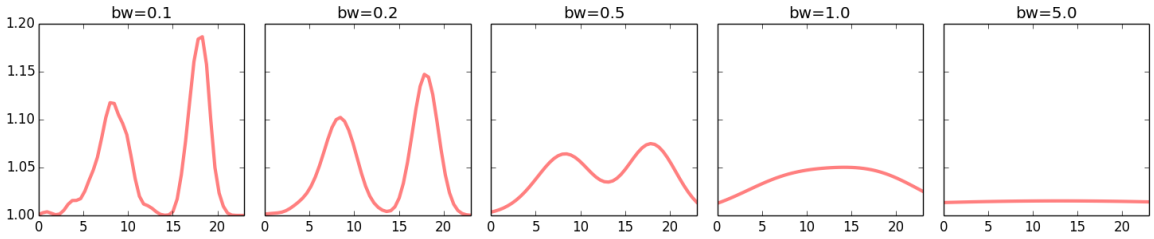

Figure 3.5: An illustration of kernel density estimation under different bandwidths

3.2 Reversible jump MCMC

To find an optimal solution in the huge solution space of parse trees, we adopt the reversible-jump Monte Carlo Markov Chain algorithm considering the fact that the dimension of the solution is not fixed.

3.2.1 Metropolis-Hastings acceptance rate

Conventionally, the true probability of a distribution in the MCMC framework is denoted as $\pi(x)$, and for the traditional MCMC moves that does not involve dimension change, the Metropolis-Hastings acceptance ratio is defined as:

$$\alpha = min(1, \frac{\pi(x^*)q(x|x^*)}{\pi(x)q(x^*|x)}) \tag{3.1}$$

In reversible jump MCMC, the following acceptance rate is used for a change from m to n dimensions:

$$\alpha = min(1, \frac{\pi(x^*)q(x|x^*)}{\pi(x)q(x^*|x)} J_{f_{m\to n}}) \tag{3.2}$$

where $J_{f_{m\to n}}$ is the Jacobian of the dimension matching function. $f_{m\to n}$ is the dimension matching function. $J_{f_{m\to n}}$ is the Jacobian of the dimension matching function:

$$J_{f_{m\to n}} = |det \frac{\partial f_{m\to n}(x_m, u_{m,n})}{\partial (x_m, u_{m,n})}| \tag{3.3}$$

It is used to map the variables at dimensionalities m and n into a space of common dimensionality. It is usually done by introducing additional $n - m$ parameters, or

projecting out the corresponding $m - n$ parameters.

Hence in our case the Jacobian matrix becomes:

$$J_{f_{pt \to pt^*}} = |det \frac{\partial f_{pt \to pt^*}(pt, \Delta pt)}{\partial (pt, \Delta pt)}| \tag{3.4}$$

Notice that each variable in Δpt is independently sampled from pt since they represent relative relations between nodes, hence the Jacobian is 1 in this case [YYW12]. Therefore the acceptance rate is formulated as:

$$\begin{aligned} \alpha &= min(1, \frac{\pi(pt^*)q(pt|pt^*)}{\pi(pt)q(pt^*|pt)}) \\ &= min(1, \frac{p(pt^*|I)q(pt|pt^*)}{p(pt|I)q(pt^*|pt)}) \\ &= min(1, \frac{p(pt^*)p(I|pt^*)q(pt|pt^*)}{p(pt)p(I|pt)q(pt^*|pt)}) \end{aligned} \tag{3.5}$$

3.2.2 Simulated annealing

We introduce a linearly decreasing temperature T to control the landscape of posterior probability π:

$$\pi(pt|I) = \frac{1}{Z} \exp(-\frac{E(pt|I)}{T}) \tag{3.6}$$

We also update the bandwidth of the non-parametric distribution of the data proposals along with temperature, to increase the dynamic of the Markov Chain, and make the process more effective.

3.3 Top-down / bottom-up proposals

We design jump and diffusion methods to ensure the ergodicity of the reversible jump Markov Chain. The following two types of dynamics are designed for jump proposals: add: sample a subtree from a non-terminal node randomly chosen from the current parse tree; delete: delete a subtree whose root is a node randomly chosen from the current parse tree.

Proposal probabilities q:

Add:

$$q(pt^*|pt) = p(v \in V^{NT})p(child|v)p(pt_{child}) \tag{3.7}$$

Delete:

$$q(pt^*|pt) = p(v \in pt) \tag{3.8}$$

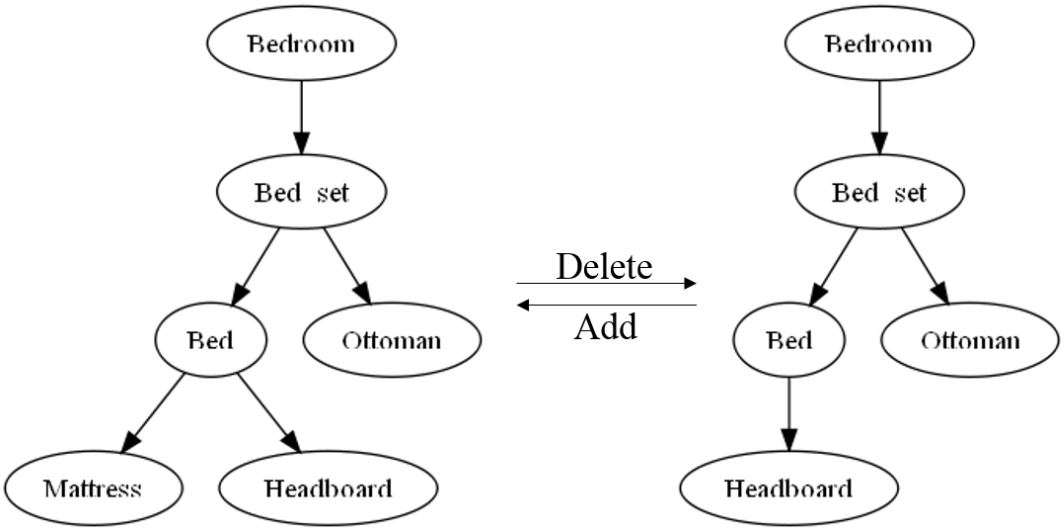

Figure 3.6: Jump: add and delete

Three dynamics are designed for the diffusion moves:

(i) α diffusion: data-driven bottom-up detection. This dynamic directly draws cuboid proposals from the non-parametric distribution built up by the line segments detected from the image. After a node is chosen from the parse tree, the six-dimension cuboid variable will be assigned to the node. Note that the non-parametric distribution gives us the cuboid coordinates in the world coordinate system, thus recalculation is needed for the geometric transformations between this node and its parent and children.

$$q(pt^*|pt) = p(v \in pt)p_{KDE}(att_v|label_v) \tag{3.9}$$

where p_{KDE} represents the non-parametric distribution of the cuboid proposals.

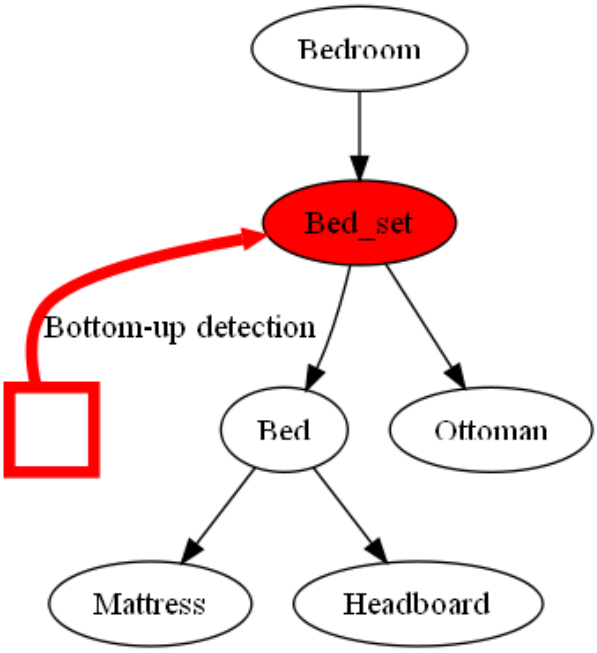

Figure 3.7: Diffusion: α channel

(ii) β diffusion: grammar-driven bottom-up prediction. This dynamic proposes cuboid for a parent node in the parse tree from the children nodes by inversely sampling a geometric transformation. After a parent node in the parse tree is chosen, we randomly chose a child node from its children, and sample a geometric transformation from the prior distribution of the child node. Therefore the attributes of the chosen node will be recalculated, and the transformations between the chosen node and other children will be recalculated as well.

$$q(pt^*|pt) = p(v \in V^{NT})p(size_v)p(child|v)p(pos_{child})p(ori_{child}) \tag{3.10}$$

where $size_v$, pos_v, and ori_v are the size, relative position and relative orientation for the node v respectively.

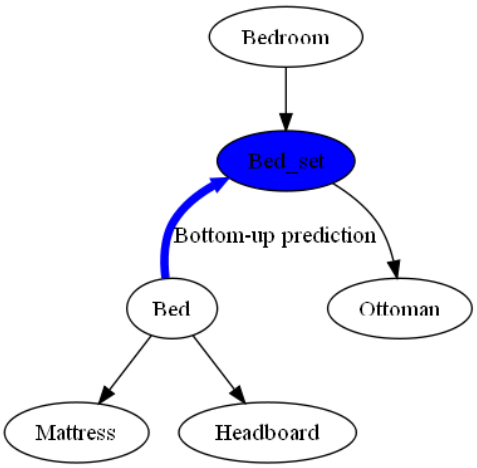

Figure 3.8: Diffusion: β channel

(iii) γ diffusion: grammar-driven top-down prediction. This dynamic proposes cuboid by top-down sampling a child node in the parse tree from its parent node based on the prior distribution, i.e., pre-defined geometric model. Sampling directly from the prior distribution is helpful in adjusting occluded objects or parts in certain contexts. This could also be used in the initialization of the parse tree (create a parse tree according to our prior knowledge before seeing the picture).

$$q(pt^*|pt) = p(v \in pt)p(att_v|par_v) \tag{3.11}$$

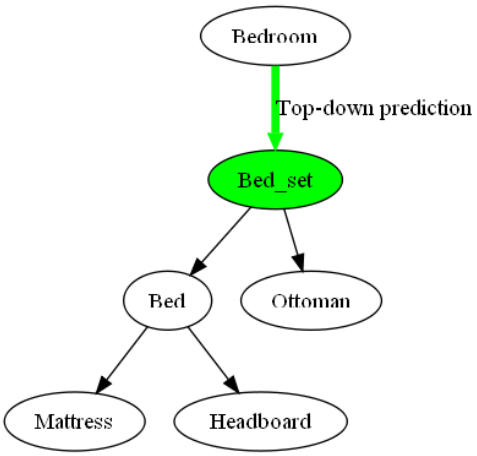

Figure 3.9: Diffusion: γ channel

To make the Markov Chain more efficient, we design two different moves for each diffusion dynamic: individual diffusion and group diffusion. Individual diffusion: only the node chosen will be moved, all the other nodes keep their absolute position in the world coordinate system. Hence the geometric transformations between the node and its children need to be recalculated to keep the absolute positions of its children. This helps when the algorithm tries to adjust certain cuboids locally without affecting other cuboids. Group diffusion: move the whole subtree of the children node. This is useful when the algorithm tries to move/diffuse a functional group , e.g., a bed set, together to a desired place. An illustration is shown in Figure 3.10.

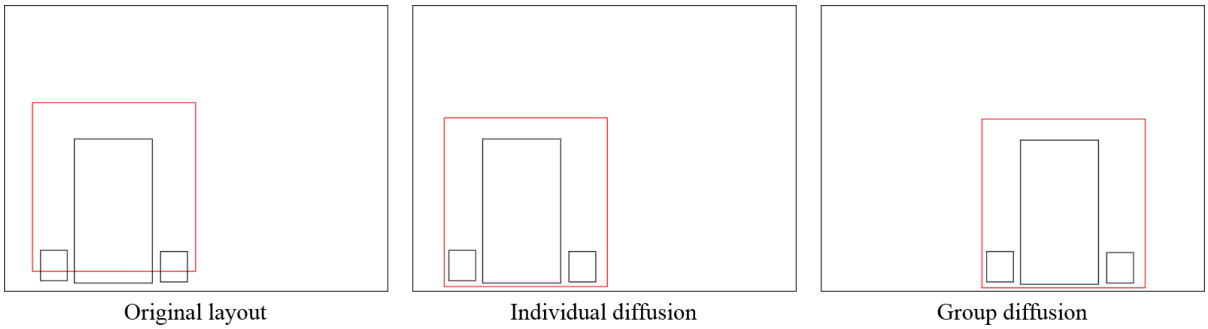

Figure 3.10: Individual diffusion and group diffusion

CHAPTER 4

Results

Figure 4.2 shows the inference process of parsing an image of a bedroom. Several steps are retrieved and shown as intermediate results in the process. The seven column shows the Markov Chain starts from random and gradually burns-in as the temperature cools down.The top three rows show parse images, rendered label maps, rendered 3D depth maps and their corresponding parse trees respectively. The red/blue/green arrows denote three different dynamics: the α bottom-up detection, β bottom-up prediction and γ top-down prediction. The output of the algorithm is a parse tree and the reconstructed 3D scene shown in the last row.

The algorithm has a strong flexibility, and can be easily adapted to different kinds of scene. We tested it on various types of indoor scenes. Figure 4.3 and Figure 4.4 shows the reconstruction result of an auditorium and a lobby respectively. We can see that the reconstruction results on rather complicated scenes such as auditoriums are also satisfying.

The algorithm is also tested on the UIUC indoor scene dataset. Some results are shown in Figure 4.5 and Figure 4.6. The first column shows the original picture, the second column to fourth column shows the reconstructed 3D structure, orientation map and label map respectively. The last column shows the energy curve during the MCMC sampling. For more results, please refer to a series of previous work [ZZ11, ZZ13].

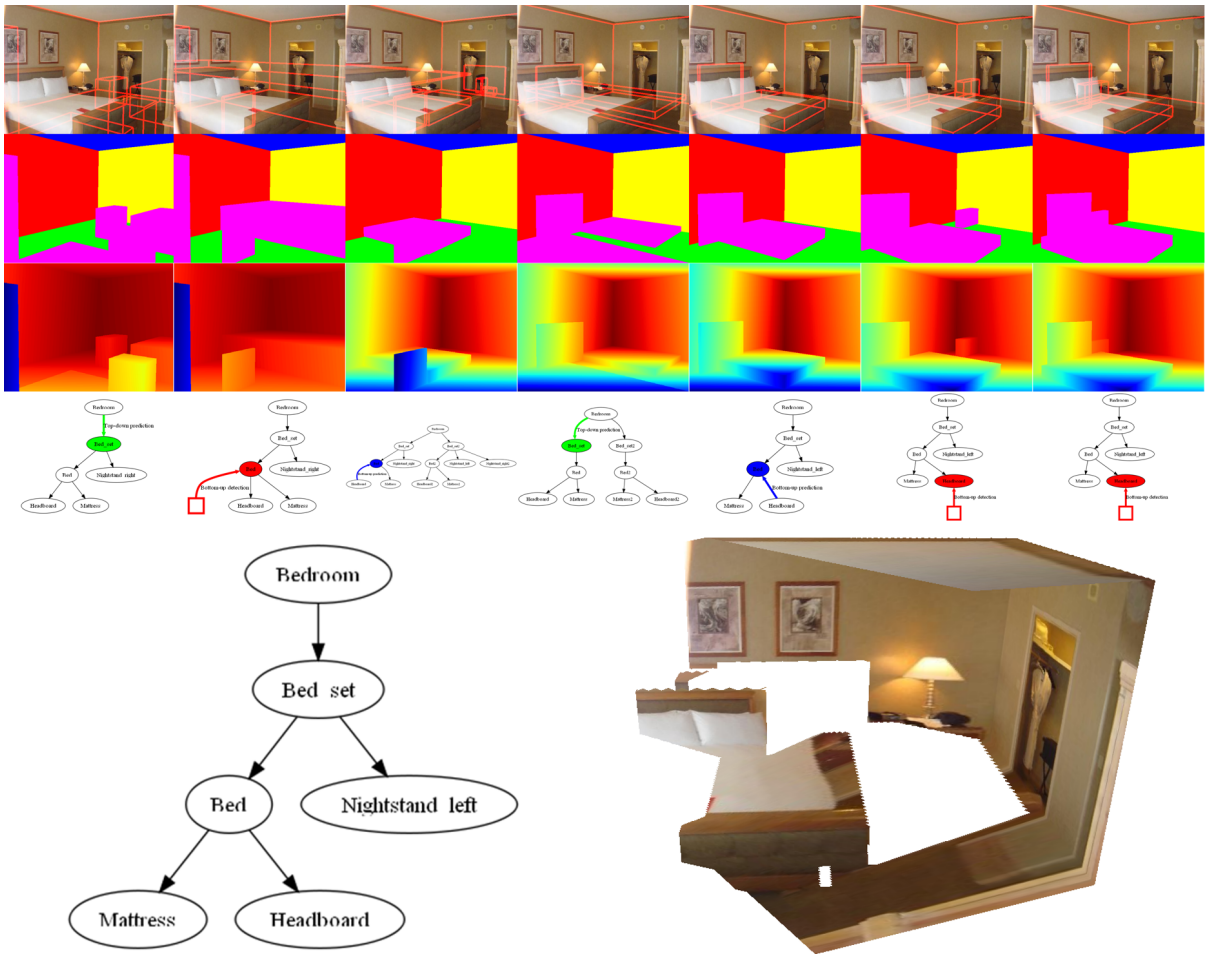

Figure 4.1: The analysis-by-synthesis MCMC sampling process.

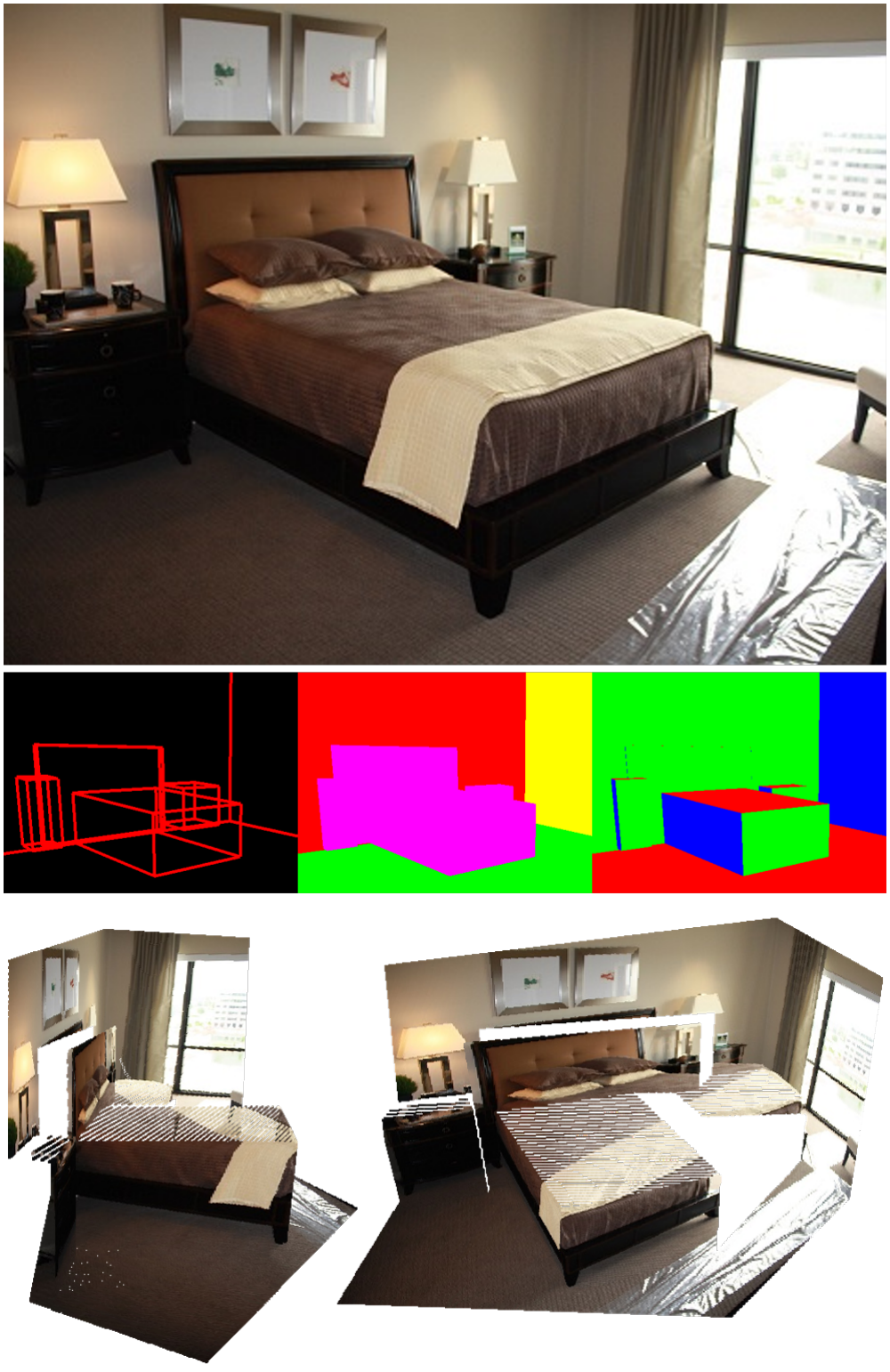

Figure 4.2: Result of parsing an image of a bedroom.

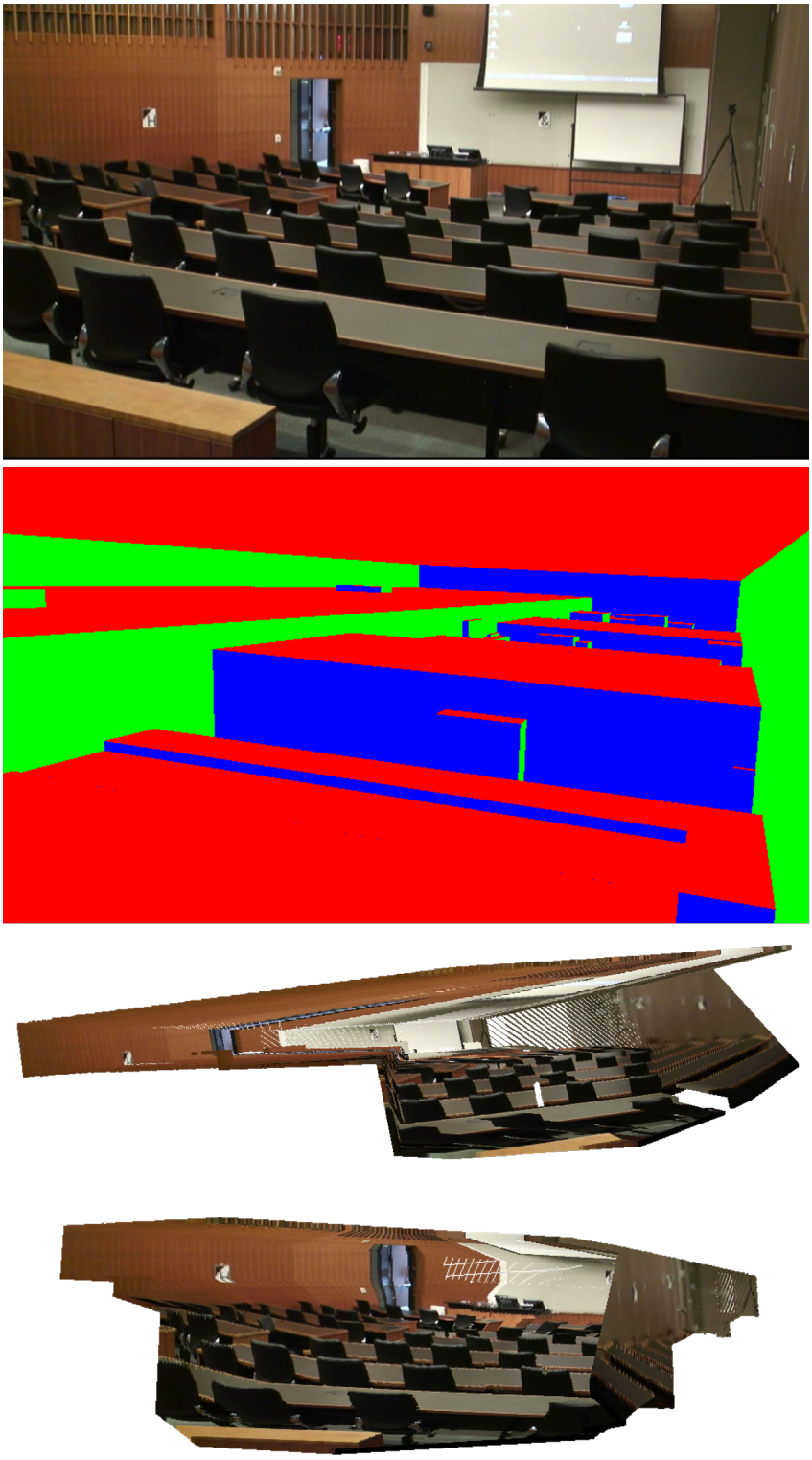

Figure 4.3: Result of parsing an image of a auditorium.

Figure 4.4: Result of parsing an image of a lobby.

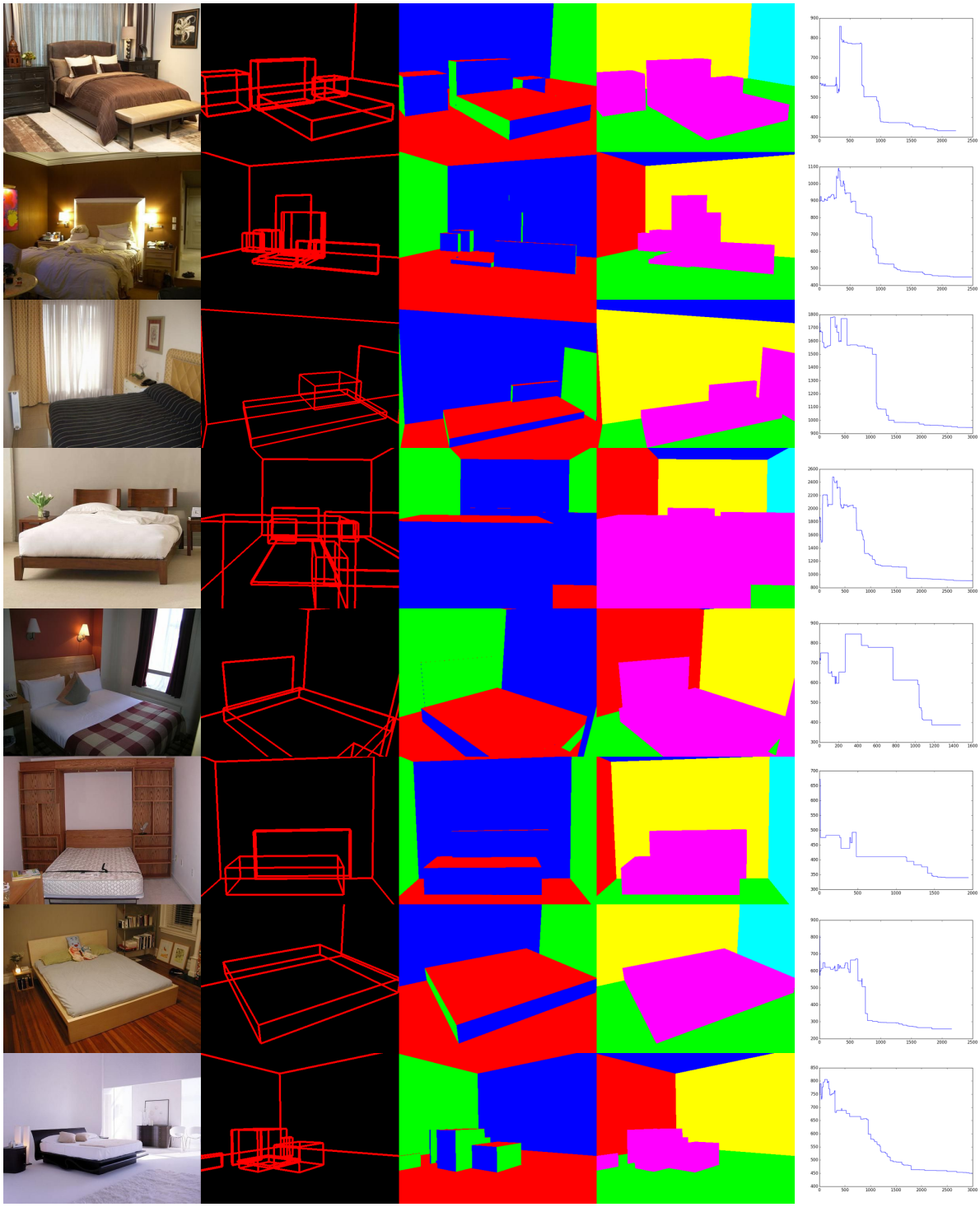

Figure 4.5: More results from the UIUC dataset.

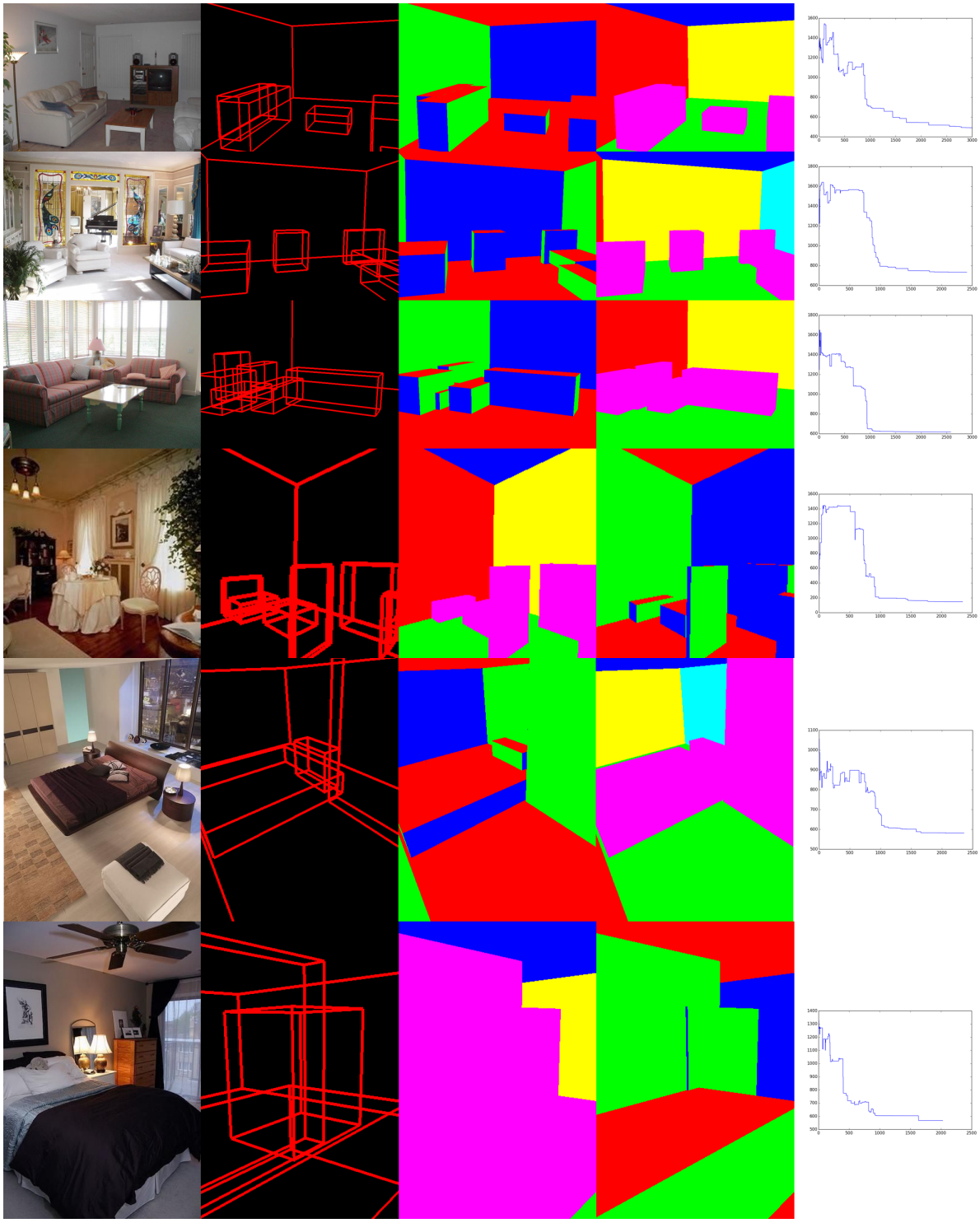

Figure 4.6: More results from the UIUC dataset.

CHAPTER 5

Discussion and Conclusion

Image understanding is not only about the image itself but also the knowledge of the world. Humans recognize images at a deep level because we know how the world works. Prior knowledge about the functional, physical mechanics of objects in the 3D world is the key for understanding. However, there is a huge 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.

There are four major representations for scene understanding: i) Non-parametric models: such as label transfer by [LYT11], SuperParsing by [TL13b] and scene collage by [IL13] 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, [SH13] recently generalize the idea of nearest-neighbor search to the 3D scenes, so that their approach can recognize objects cross viewpoints. [LPT13], [DBK13] detected indoor objects by matching with fine-grained furniture models. ii) Parametric models: Representation of parametric 3D shapes allows reasoning about the physical constraints within the 3D scene. [GEH10] posed the 3D objects as blocks and inferred their 3D properties such as occlusion, exclusion and stability in addition to surface orientation labels. [HHF12, WGK10b, LGH10, SFP13] parameterized the geometric scene layout of the background and/or foreground blocks and trained their models by the Structured SVM (or Latent SVM).

The approaches proposed in this paper and some early work [ZZ11, ZZ13] represent a

3D scene by a generative language model, a stochastic grammar, to characteristic the 3D structures of the world. We then augmented hidden commonsense knowledge, i.e. object functionality and physics, to grammar structure. The high-level knowledge projected onto a sentence/image thus forms contexts of the language/vision. With the analysis-by-synthesis approach, a computer can imagine 3D scenes by building a mental picture of the world or even action associated to the functional objects, which provide a machine the fundamental potential to reason, predict, answer questions, and hold intelligent dialog.

References

[BR06] Ezer Bar-aviv and Ehud Rivlin. "Functional 3D Object Classification Using Simulation of Embodied Agent." In *BMVC*, 2006.

[CCP13] W. Choi, Y. Chao, C. Pantofaru, and S. Savarese. "Understanding Indoor Scenes using 3D Geometric Phrases." In *CVPR*, 2013.

[CLT10] Myung J. Choi, Joseph J. Lim, Antonio Torralba, and Alan S. Willsky. "Exploiting Hierarchical Context on a Large Database of Object Categories." In *CVPR*, 2010.

[DBF12] L. Del Pero, J. Bowdish, D. Fried, B. Kermgard, E. Hartley, and K. Barnard. "Bayesian geometric modeling of indoor scenes." In *CVPR*, pp. 2719–2726, 2012.

[DBK13] L. Del Pero, Joshua Bowdish, Bonnie Kermgard, Emily Hartley, and Kobus Barnard. "Understanding Bayesian rooms using composite 3D object models." In *CVPR*, 2013.

[DGB11] L. Del Pero, J. Guan, E. Brau, J. Schlecht, and K. Barnard. "Sampling bedrooms." In *CVPR*, 2011.

[DLN07] E. Delage, H. Lee, and A. Ng. "Automatic single-image 3d reconstructions of indoor manhattan world scenes." *Robotics Research*, p. 305321, 2007.

[FDU12] Sanja Fidler, Sven Dickinson, and Raquel Urtasun. "3D Object Detection and Viewpoint Estimation with a Deformable 3D Cuboid Model." In *NIPS*, 2012.

[Fu82] K.S. Fu. *Syntactic Pattern Recognition and Applications*. Prentice-Hall, second edition, 1982.

[GEH10] Abhinav Gupta, Alexei A. Efros, and Martial Hebert. "Blocks World Revisited: Image Understanding Using Qualitative Geometry and Mechanics." In *European Conference on Computer Vision(ECCV)*, 2010.

[GGG11] Helmut Grabner, Juergen Gall, and Luc Van Gool. "What Makes a Chair a Chair?" In *CVPR*, 2011.

[Gib77] James J . Gibson. *The Theory of Affordances*. Lawrence Erlbaum, 1977.

[GSE11] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. "From 3D scene geometry to human workspace." In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1961–1968, Washington, DC, USA, 2011. IEEE Computer Society.

[HEH09] D. Hoiem, A. Efros, and M. Hebert. "Automatic Photo Pop-up." *TOG*, **31**(1):59–73, 2009.

[HHF09] V. Hedau, D. Hoiem, and D. Forsyth. "Recovering the spatial layout of cluttered rooms." In *ICCV*, 2009.

[HHF10] V. Hedau, D. Hoiem, and D. Forsyth. "Thinking Inside the Box: Using Appearance Models and Context Based on Room Geometry." In *ECCV*, 2010.

[HHF12] Varsha Hedau, Derek Hoiem, and David Forsyth. "Recovering Free Space of Indoor Scenes from a Single Image." In *CVPR*, 2012.

[HR12] Mohsen Hejrati and Deva Ramanan. "Analyzing 3D Objects in Cluttered Images." In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pp. 602–610. 2012.

[Hu12] Wenze Hu. "Learning 3D object templates by hierarchical quantization of geometry and appearance spaces." In *CVPR*, pp. 2336–2343, 2012.

[HZ04] Feng Han and Song-Chun Zhu. "Bayesian Reconstruction of 3D Shapes and Scenes From A Single Image." In *Proc. IEEE Workshop on Perceptual Organization in Computer Vision*, 2004.

[HZ09] F. Han and S. C. Zhu. "Bottom-Up/Top-Down Image Parsing with Attribute Grammar." *PAMI*, 2009.

[IL13] P. Isola and C. Liu. "Scene collaging: analysis and synthesis of natural images with semantic layers." In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[LFU13] Dahua Lin, Sanja Fidler, and Raquel Urtasun. "Holistic Scene Understanding for 3D Object Detection with RGBD cameras." In *ICCV*, 2013.

[LGH10] D. Lee, A. Gupta, M. Hebert, and T. Kanade. "Estimating Spatial Layout of Rooms using Volumetric Reasoning about Objects and Surfaces Advances in Neural Information Processing Systems." *Cambridge: MIT Press*, pp. 609–616, 2010.

[LHK09] D. Lee, M. Hebert, and T. Kanade. "Geometric Reasoning for Single Image Structure Recovery." In *CVPR*, 2009.

[LMP01] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. "Conditional random fields: probabilistic models for segmenting and labeling sequence data." *ICML*, pp. 282–289, 2001.

[LPT13] Joseph J. Lim, Hamed Pirsiavash, and Antonio Torralba. "Parsing IKEA Objects: Fine Pose Estimation." In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[LSP06] S. Lazebnik, C. Schmid, and J. Ponce. "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories." In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[LYT11] C. Liu, J. Yuen, and A. Torralba. "Nonparametric scene parsing via label transfer." *IEEE Trans. on Patt. Anal. Mach. Intell (TPAMI)*, 2011.

[MZY11] Hossein Mobahi, Zihan Zhou, Allen Y. Yang, and Yi Ma. "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, 2011.

[OT01] A. Oliva and A. Torralba. "Modeling the shape of the scene: a holistic representation of the spatial envelope." *International Journal of Computer Vision (IJCV)*, 2001.

[PGS12] Bojan Pepik, Peter Gehler, Michael Stark, and Bernt Schiele. "3D2PM - 3D Deformable Part Models." In *ECCV*, Firenze, Italy, 2012.

[POF12] S. N. Parizi, J. Oberlin, and P. Felzenszwalb. "Reconfigurable models for scene recognition." In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[PT11] Nadia Payet and Sinisa Todorovic. "Scene Shape from Textures of Objects." In *CVPR*, 2011.

[PZ10] J. Porway and S. C. Zhu. "Hierarchical and Contextual Model for Aerial Image Understanding." *IJCV*, **88**(2):254–283, 2010.

[SFP13] Alexander G. Schwing, Sanja Fidler, Marc Pollefeys, and Raquel Urtasun. "Box In the Box: Joint 3D Layout and Object Reasoning from Single Images." In *ICCV*, 2013.

[SH13] Scott Satkin and Martial Hebert. "3DNN: Viewpoint Invariant 3D Geometry Matching for Scene Understanding." In *ICCV*, 2013.

[SHP12] Alexander G. Schwing, Tamir Hazan, Marc Pollefeys, and Raquel Urtasun. "Efficient Structured Prediction for 3D Indoor Scene Understanding." In *CVPR*, 2012.

[SLH12] Scott Satkin, Jason Lin, and Martial Hebert. "Data-Driven Scene Understanding from 3D Models." In *BMVC*, 2012.

[SSN09] A. Saxena, M. Sun, and A. Ng. "Make3D: Learning 3D Scene Structure from a Single Still Image." *PAMI*, **31**(5):824–840, 2009.

[TCY05] Z. Tu, X. Chen, A. Yuille, and S.C. Zhu. "Image parsing: unifying segmentation, detection and recognition." *IJCV*, **63**(2):113–140, 2005.

[TL13a] J. Tighe and S. Lazebnik. "Finding things: image pars- ing with regions and per-exemplar detectors." In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[TL13b] J. Tighe and S. Lazebnik. "Superparsing: scalable non- parametric image parsing with superpixels." *International Journal of Computer Vision (IJCV)*, 2013.

[WGK10a] H. Wang, S. Gould, and D. Koller. "Discriminative Learning with Latent Variables for Cluttered Indoor Scene Understanding." In *ECCV*, 2010.

[WGK10b] Huayan Wang, Stephen Gould, and Daphne Koller. "Discriminative learning with latent variables for cluttered indoor scene understanding." In *ECCV*, pp. 497–510, 2010.

[WWZ12] S. Wang, Y. Wang, and S.C. Zhu. "Hierarchical Space Tiling in Scene Modeling." In *Asian Conf. on Computer Vision (ACCV)*, 2012.

[WZ11] T. Wu and S.C. Zhu. "A numerical study of the bottom-up and top-down inference processes in and-or graphs." *International Journal of Computer Vision (IJCV)*, 2011.

[WZZ13] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. "Modeling 4D Human-Object Interactions for Event and Object Recognition." In *ICCV*, 2013.

[XRT12] Jianxiong Xiao, Bryan Russell, and Antonio Torralba. "Localizing 3D cuboids in single-view images." In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *NIPS*, pp. 755–763. 2012.

[XS12] Yu Xiang and Silvio Savarese. "Estimating the Aspect Layout of Object Categories." In *CVPR*, 2012.

[YYT11] Lap-Fai Yu, Sai Kit Yeung, Chi-Keung Tang, Demetri Terzopoulos, Tony F. Chan, and Stanley Osher. "Make it home: automatic optimization of furniture arrangement." *ACM Trans. Graph.*, **30**(4):86, 2011.

[YYW12] Yi-Ting Yeh, Lingfeng Yang, Matthew Watson, Noah D. Goodman, and Pat Hanrahan. "Synthesizing open worlds with constraints using locally annealed reversible jump MCMC." *ACM Trans. Graph.*, **31**(4):56:1–56:11, July 2012.

[ZM07] S. C. Zhu and D. Mumford. "A stochastic grammar of images." *Foundations and Trends in Computer Graphics and Vision*, **2**(4):259–362, 2007.

[ZZ11] Yibiao Zhao and Song-Chun Zhu. "Image Parsing via Stochastic Scene Grammar." In *NIPS*, 2011.

[ZZ13] Yibiao Zhao and Song-Chun Zhu. "Scene Parsing by Integrating Function, Geometry and Appearance Models." In *CVPR*, 2013.

[ZZZ15] Yixin Zhu, Yibiao Zhao, and Song-Chun Zhu. "Understanding Tools: Task-Oriented Object Modeling, Learning and Recognition." In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.