A Stochastic Image Grammar for Fine-Grained 3D Scene Reconstruction *

Xiaobai Liu[1], **Yadong Mu**[2], **Liang Lin**[3]
[1]Department of Computer Science, San Diego State University, San Diego, 92182, CA, USA
[2] Institute of Computer Science and Technology, Peking University, Beijing, 100871, China
[3] School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, 510006, China
xiaobai.liu@mail.sdsu.edu, muyadong@gmail.com, linliang@ieee.org

Abstract

This paper presents a stochastic grammar for fine-grained 3D scene reconstruction from a single image. At the heart of our approach is a small number of grammar rules that can describe the most common geometric structures, e.g., two straights lines being co-linear or orthogonal, or that a line lying on a planar region etc. With these grammar rules, we re-frame single-view 3D reconstruction problem as jointly solving two coupled sub-tasks: i) segmenting of image entities, e.g. planar regions, straight edge segments, and ii) optimizing pixel-wise 3D scene model through the application of grammar rules over image entities. To reconstruct a new image, we design an efficient hybrid Monte Carlo (HMC) algorithm to simulate Markov Chain walking towards a posterior distribution. Our algorithm utilizes two iterative dynamics: i) *Hamiltonian Dynamics* that makes proposals along the gradient direction to search the continuous pixel-wise 3D scene model; and ii) *Cluster Dynamics*, that flip the colors of clusters of pixels to form planar region partition. Following the Metropolis-hasting principle, these dynamics not only make distant proposals but also guarantee detail-balance and fast convergence. Results with comparisons on public image dataset show that our method clearly outperforms the alternate state-of-the-art single-view reconstruction methods.

1 Introduction

Reconstructing 3D scene model from a single image has abstracted a lot of interest because of its wide applications in robotics, intelligent transportation, and surveillance etc. Despite impressive results achieved, existing 3D modeling methods are likely to miss details of the scene, e.g. rectangles of windows in facades, zebra crossing on roads, or T-junctions corners of tables. In most of urban street images, these details are directly reflected by the geometric relationships between image entities, e.g. that a straight line being parallel or orthogonal to other lines, or that a line be lying on a planar surface, or that two planar regions being orthogonal with each other, etc. However, it remains unknown how to explore these geometric constraints efficiently. There are two particular challenges: i) the segmentation of image entities, e.g. edges, planar regions etc., has illness nature ; ii) with perspective effect, apparent structures (e.g. right angle corners) do not necessarily reflect real structures in 3D world.

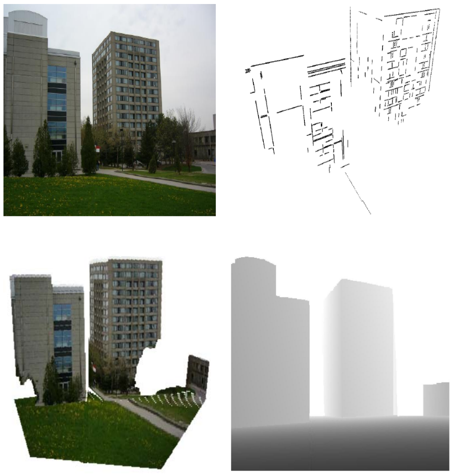

Figure 1: Single-view 3D Scene reconstruction. (a) input image; (b) novel view of 3D lines; c) novel view of of the input image ; (d) the recovered depth map.

In this work, we introduce a stochastic grammar model to address the above issues. Figure 1 illustrates an exemplar results. Our grammar includes a set of grammar rules and a probability model. Each grammar rule describes a particular geometric relationship between image entities, e.g. co-line for pixels, orthogonality for straight sedges, co-planar for straight edges and planar regions, supporting for planar regions etc. These relationships, once discovered, directly provide information of the fine-grained scene structure and thus should be persisted while optimizing a continuous 3D scene model.

To reconstruct an input image, we fit and evaluate a variety of combinations of grammars rules over the input image. In order to efficiently exploit this combinational solution space, we develop a hybrid Monte Carlo (HMC) method to simulate a Markov chain for sampling the posterior probability. Dif-

*This work was supported by the start-up funds of the San Diego State University (SDSU).

ferent from the conventional sampling methods [Liu *et al.*, 2014], we design two dynamics to make distant proposals in both continuous space and discrete spaces in order to enhance convergence speed. i) **Hamiltonian dynamics**, that make proposals in the deepest descent direction, in order to search for the continuous 3D scene model. ii) **Cluster dynamics**, that flip the partitions of a cluster of pixels, instead of single one. These dynamics are iterated until convergence. Note that our method is different from existing grammar models that only optimize discrete labeling problems [Liu *et al.*, 2014] [Hoiem *et al.*, 2005].

Contributions The two major contributions of this work include: i) we define a set of grammar rules to describe the geometric constraints between image entities and present an stochastic optimization method to automatically determine the valid constraints and recover fine-grained 3D scene model for a single image; ii) we introduce an iterative hybrid Monte Carlo method that is capable of making distant proposals in both continuous and discrete spaces. We apply the proposed method over both public image datasets and a newly created dataset. Results with comparisons show that our method clearly outperforms the state-of-the-art methods.

2 Related Works

Our work is closely related to *three* research streams in computer vision and machine learning.

Single-View 3D modeling has been extensively studied with a variety of techniques, including generative model [Han and Zhu, 2003] , context reasoning [Hoiem *et al.*, 2005], conditional random field [Heitz *et al.*, 2008], physics reasoning [Gupta *et al.*, 2010] , attributed grammar [Liu *et al.*, 2014], etc. Most of these methods were built on the classification of 2D segmentation, which did not directly solve 3D models or depth values. Other methods [Mobahi *et al.*, 2012] [Schwing and Urtasun, 2012] [Pero *et al.*, 2011] [Pero *et al.*, 2012] [Pero *et al.*, 2013] tried to recover global 3D scene without an explicit representation of scene structures. In this work we directly optimize continuous pixel-wise 3D coordinates by exploring the various geometric constraints between image entities (i.e. edge, planar regions).

Joint Recognition and Reconstruction has been investigated for a varitety of tasks, including scene labeling and reconstruction [Haene *et al.*, 2013] [Liu *et al.*, 2014], reconstruction of panorama images [Cabral and Furukawa, 2014], object recognition and modeling [Hejrati and Ramanan, 2014], layout partition and object modeling [Schwing *et al.*, 2013], joint Object Labeling and Structure-from-Motion [Xiao *et al.*, 2013] and joint tracking and mapping [Kundu *et al.*, 2014] [Zhang *et al.*, 2013]. Our method follows the same methodology to introduce a joint formula for segmenting planar regions and reconstructing the whole scene. We additionally impose the regularizations of straight lines to guide the reconstruction process.

Scene grammar has been applied for a number of image parsing problems in computer vision tasks. Koutsourakis et al. [Koutsourakis *et al.*, 2009] proposed a shape grammar to explain building facades with levels of details, their model was focused on rectified facade images not 3D geometry. Han and Zhu [Han and Zhu, 2009], Liu et al. [Liu *et al.*, 2014], Zhao and Zhu [Zhao and Zhu, 2011] and Pero et al. [Pero *et al.*, 2013] specified generative scene grammar models to model the compositional of Manhattan structures in images. Furukawa et al. [Furukawa *et al.*, 2009] studied the reconstruction of Manhattan scenes from stereo inputs. In this work, we extend these grammar models to describe the geometric constraints between image entities, from lines to planar regions to blocks, which enables detail-preserving 3D scene reconstruction.

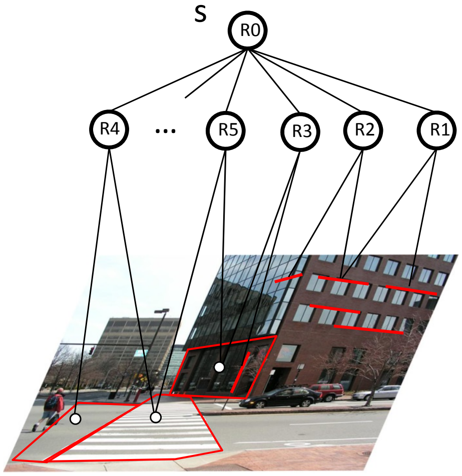

Figure 2: Illustration of parse graph. The root node S simply decomposes into a set of nodes with grammar rules R_1 to R_5. Each graph node imposes at least one constraint equation over the desired solution space.

3 Stochastic Scene Grammar for 3D Modeling

In this section, we introduce a stochastic scene grammar for single-view reconstruction problem.

3.1 Scene model

We consider urban street images in this work and use the world coordinates for the desired 3D model. These scenes are typical local Manhattan world [Liu *et al.*, 2014] where there is a family of parallel lines pointing into the sky and two or more parallel families being parallel to the groundplane. Each parallel family merges at a vanishing point in imaging plane. Once vanishing points detected, we utilize the method [Cipolla *et al.*, 1999] to compute the rotation matrix R and the intrinsic matrix K. For every image pixel (x, y), its 3D position be determined as $d_i\bar{X}$ where $\bar{X} = (KR)^{-1}(x, y, 1)$ is a 3D ray and d_i is the depth to solve.

3.2 Scene Grammar

A context-free grammar is specified by a 5-tuple $G = (V_T, V_N, R, S, P)$ where V_N is a finite set of non-terminal nodes, V_T a finite set of terminal nodes, $S \in V_N$ a start symbol, R is a set or grammar rules, and P is the probabilistic distribution for the grammar.

We apply grammar rules to generate hierarchical representations of the input image, i.e. parse graph. Figure 2 demonstrates a parse graph which includes a root node and a set of graph nodes. A parse graph is a valid interpretation of the input 2D image in the 3D space. A grammar generates a large set of valid parse graphs for one given image of the scene. We allow children nodes be shared by two or more nodes since an image entity, e.g. a straight edge, might bear relationships with two or more other entities.

Terminal Nodes V_T We partition the input image into a set of suerpixels [Ren and Malik, 2003] and detect straight edge segments [Gioi *et al.*, 2008] to obtain terminal nodes. Each superpixel is the projection of a 3D planar surface and each edge segment is the projection of a straight line in 3D. There are around 200-300 superpixels and about 500-700 straight edge segments for each image. To reconstruct a superpixel or an straight edge V_i, we need to determine its geometric attribute, e.g., 3D position, normal orientation.

Nonterminal Nodes V_N are produced by merging terminal nodes with grammar rules. Each node in V_N indicates a combination of terminal nodes or non-terminal nodes. We impose six grammar rules, R_0 through R_6, each for a specific relationship between image entities. i) *R0*, that generates the input image into a set of grammar nodes; ii) *R1*, that merges two co-linear edge segments; ii) *R2*, that merges two orthogonal edge segments; iii) *R3*, that merges one edge segment and a planar that are co-planar; iv) *R4*, that merges two neighboring planar regions that are projections of the same 3D plane; and v) *R5*, that merges two neighboring planar regions that are orthogonal and intersected with each other. Note that a) the grammar rule R_4 can be recursively applied to get graph nodes that are used for children nodes of R_0, R_3, R_5 and R_4 itself; b) different graph nodes may share the same children nodes, which basically allows multiple interpretations of a single image entity. Among these rules, R_0 is used to generate the input image into a set of grammar modes.

In contrast to the previous methods [Liu *et al.*, 2014], in this work, the parse graph is not deep and the search space in 3D construction is relatively smaller. The sharing between nonterminals also make it a redundant representation that is potentially more robust against noises.

3.3 Probabilistic Formulation

Given an input image, our goals include i) partitioning it into image entities, i.e. planar regions and straight edges; ii) reconstruct each image entity in 3D. We achieve such goals by constructing an optimal parse graph and estimate the attributes of every graph node. To do this, we introduce a unified probabilistic framework. Let I denote the input image, our objective is to compute an optimal solution representation $W = (G, \mathcal{X}(G), K)$ where K is the number of planar regions, $\mathcal{X}(G)$ organizes all attributes of graph nodes.

The optimal solution W^* can be obtained by maximizing a posterior probability (MAP):

$$p(W|I) \propto \exp\{-K - E(I, V_N)\} \tag{1}$$

where the first term of K is used to encourage compact planar partition. The energy term $E(I, V_N)$ is defined over the nonterminals,

$$E(I, V_N) = \sum_{V \in V_N} \lambda^k E(I, V|R_k) \tag{2}$$

where λ^k is a constant related to the grammar rule k, $E(I, V|R_k)$ is conditioned on the grammar rule R_k. Note that R_0 is a lose grammar rule and does not affect the energy. In the rest of this section, we denote a terminal node of edge segment as a, a non-terminal of planar region as B.

Grammar rule R_1: $V \rightarrow (a_i, a_j)$ involves two children edge segments, a_i and a_j, that are projections of the same straight line in 3D. R_1 requires that a_i and a_j are spatially adjacent in 2D image. The attributes of an edge segment is defined as $\mathcal{X}(a_i) = (d_i, \bar{n}_i)$ where d_i denote the depth of the central point of the edge, $\bar{n}_i$ is the edge direction.

The energy term $E(I, V|R_1)$ is defined over the mutual consensus between the attributes of a_i and b_j. Let λ_{ij} denote a constant such that the following linear equation holds:

$$d_i \bar{X}_i + \lambda_{ij} \bar{n}_i = d_j \bar{X}_j \tag{3}$$

where $\bar{X}_i$ and $\bar{X}_j$ are 3D rays that are known (suppose we have calibrated the camera). We define the related energy as $\Phi(d_i, \bar{n}_i; d_j)$ as the following least square form:

$$\Phi(d_i, \bar{n}_i; d_j) = \min_{\lambda_{ij}} \|d_i \bar{X}_i + \lambda_{ij} \bar{n}_i - d_j \bar{X}_j\|^2 \tag{4}$$

Accordingly, we define $\Phi(d_j, \bar{n}_j; d_i)$ as well. Thus, we have $E(I, V|R_1)$ defined as follows:

$$\begin{aligned} E(I, V|R_1) = \Phi(d_i, \bar{n}_i; d_j) + \Phi(d_j, \bar{n}_j; d_i) \\ + \|\bar{n}_i - \bar{n}_j\|^2. \end{aligned} \tag{5}$$

where the last term is used to enforce the co-linear constraint. Eq. (5) is a convex smoothing function of four continuous variables to solve : d_i,d_j,$\bar{n}_i$, $\bar{n}_j$. In this work, we assume that the number of unknown variables are far less than the number of nonterminal nodes, which is reasonable because we allow sharing of children nodes between nonterminal nodes.

Grammar rule R_2: $V \rightarrow (a_i, a_j)$ is used to associate two children edge segments in images. This rule requires that two children edges are the projections of two straight lines in 3D that are orthogonal and intersected with each other. In local Manhattan world [Liu *et al.*, 2014], it follows that two children lines should share the same Z-component in the world coordinate, denoted as $[d_i \bar{X}_i]_Z$. We define $E(I, V|R_2)$ as follows:

$$E(I, V|R_2) = ([d_i \bar{X}_i]_Z - [d_j \bar{X}_j]_Z)^2 + \bar{n}_i^T \bar{n}_j \tag{6}$$

where the second term is used to enforce orthogonality constraint.

Grammar rule R_3: $V \rightarrow (a_i, B_j)$ is used to associate an edge segment a_i with a planar region B_j. The attributes of B_j are defined as $\mathcal{X}(B_j) = (d_j, \bar{n}_j, l_j)$ where $\bar{n}_j$ is the normal orientation, $d_j \bar{X}_j$ is the 3D position of the planar center. Thus, we require that the line with a_i lies on the plane B_j and define $E(I, V|R_3)$ using the following objective,

$$E(I, V|R_3) = \left[\bar{n}_j^T (d_j \bar{X}_j - d_i \bar{X}_i)\right]^2 \tag{7}$$

which encourages the line segment a_i to be lying on the planar B_j.

Grammar rule R_4: $V \rightarrow (B_i, B_j)$ states that two children planar surfaces share the same normal orientation, and thus should belong to the same planar surface. The surfaces B_i and B_j should be spatially adjacent in both image and 3D space. In practice, in order to address occlusions and noises, we allow the grouping of disjoint regions in image by this rule if they have high affinity in appearance. The node V is a composition of its children planar surfaces.

The energy function $E(I, V_N|R_4)$ is defined over both geometric attributes and appearance of graph nodes:

$$E(I, V_N|R_4) = \sum_{V \in V_N} \mathcal{E}^{geo}(I, V) + \mathcal{E}^{app}(I, V_N) \tag{8}$$

The geometric energy $\mathcal{E}^{geo}(I, V)$ is defined to encourage that: the central point $d_i \bar{X}_i$ of the planar B_i should lie on the plane B_j and vice versa. Thus, $\mathcal{E}^{geo}(I, V)$ is defined as:

$$\begin{aligned} \mathcal{E}^{geo}(I, V) = & \left[\bar{n}_j^T (d_j \bar{X}_j - d_i \bar{X}_i)\right]^2 \\ & + \left[\bar{n}_i^T (d_i \bar{X}_i - d_j \bar{X}_j)\right]^2 \end{aligned} \tag{9}$$

The appearance energy $\mathcal{E}^{app}(I, V_N)$ is defined over the planar partition by all the non-terminal nodes. We use the typical Ising/Potts model in statistical mechanics. Let $< s, t >$ denote a pair of adjacent superpixels, l_s and l_t their planar index or colors. We have

$$\mathcal{E}^{app}(I, V_N) = -\beta \sum_{<s,t>} f_{st} \mathbf{1}(l_s = l_t) \tag{10}$$

where $\beta > 0$ is an constant, f_{st} indicates the appearance similarity between B_i and B_j, $\mathbf{1}(l_s = l_t)$ returns 1 if superpixels s and t have the same label; otherwise, returns 0.

Grammar rule R_5: $V \to (B_i, B_j)$ is used to group two adjacent planar regions that have different yet orthogonal normal orientations. The parent node V indicates a composite structure, e.g. building blocks, or a facade standing on groundplane. Hence, we define $E(I, V|R_5)$ to be the following:

$$E(I, V|R_5) = \bar{n}_i^T \bar{n}_j \tag{11}$$

which enforces the orthogonality between B_i and B_j.

4 Inference via Hybrid Monte Carlo Method with Hamiltonian Dynamics

Our inference aims to construct an optimal parse graph by applying the grammar rules and solving the optimal attributes for each graph node, which are however intractable. We develop a Hybrid Monte Carlo method (HMC) to sample the posterior distribution in Eq. (1). It starts with an initial graph that includes a root node and a set of terminal nodes, i.e. superpixels or edge segments. Then we design a set of dynamics to reconfigure the parse graph and simulate a Markov chain in the solution space. The dynamics are either jump moves, e.g. creating new graph nodes, or diffusion moves, e.g. estimating 3D positions or normal orientation for a planar region. These stochastic dynamics are paired with each to make the solution changes reversible in order to guarantee convergence to $p(W|I)$.

Formally, a dynamic is proposed to drive the solution status from W to W', which is accepted with the probability in the Metropolis-hasting form:

$$\min(1, \frac{p(W'|I)Q(W \to W')}{p(W|I)Q(W' \to W)}) \tag{12}$$

where $Q(\cdot)$ is the proposal probability. We use three dynamics, including two jump moves in the discrete space and one diffusion move in the continuous space.

Dynamic I: Birth/Death of Non-terminal Nodes. This pair of jumps are used to create or delete a nonterminal node and thus transition the current solution into a new solution.

To create a non-terminal node, we first randomly select one of the five grammar rules R_1, ..., or R_5, and create a list of candidates that are plausible according to the predefined constraints. Taking R_1 as example, two children edge segments should i) have the same orientation; ii) be spatially connected. Each candidate in this list is represented by its potential. Let B_i^k denote the i^{th} candidate for the grammar rule R_k, $|B_i^k|$ the size of the region associated with B_i^k, its energy is $E(I, B_i^k|R_k)$. The proposal probabilities for selecting B_i^k is calculated based on the weighted list,

$$Q(W \to W') = 1 - \frac{|B_i^k| E(I, B_i^k|R_k)}{\sum_j |B_j^k| E(I, B_j^k|R_k)} \tag{13}$$

Similarly, we obtain another set of candidate nodes to delete based on their energies, and the proposal probabilities for deleting the node D_i^k is calculated as

$$Q(W \to W') = \frac{|D_i^k| E(I, D_i^k|R_k)}{\sum_j |D_j^k| E(I, D_j^k|R_k)} \tag{14}$$

Dynamic II: Hamiltonian Dynamic We use the Hamiltonian mechanics [Audin and Babbitt, 2008] [Almeida, 1992] to make proposals for the continuous variables in W, i.e. d_i and $\bar{n}_i$. Hamiltonian method uses physical system dynamics rather than probability distribution to make proposals. Let a vector $\theta = (\{d_i, \bar{n}_i\})$ organize all the desired continuous attributes. $E(\theta)$ is the energy defined in Eq. (2). Consider θ as a position at the energy landscape, h as the momentum at time t. Sampling θ is equal to moving θ through the energy landscape with a varying moment h. The energy $H(\theta, h)$ at a time-step, known as Hamiltonian, is a combination of the energy $E(\theta)$ and the kinetic energy $K(h)$, i.e. $H(\theta, h) = E(\theta) + K(h)$. We set $K(h) = h^T h/2$ as conventional. The partial derivatives of the Hamiltonian determine how θ and h change over time, according to the Hamilton's equations:

$$\frac{\partial \theta}{\partial t} = \frac{\partial K(h)}{\partial h} = h, \quad \frac{\partial h}{\partial t} = -\frac{\partial E(\theta)}{\partial \theta} \tag{15}$$

Starting with initial state at time t=0, we can iteratively compute the states of θ and the moment h at each time, following the Euler's method [Audin and Babbitt, 2008]:

$$\theta^{t+1} = \theta^t - \alpha h^t \quad h^{t+1} = h^t - \alpha \frac{E(\theta)}{\partial \theta} \tag{16}$$

where the subscript t denotes the state at time t. The proposal probability for Hamiltonian Dynamic is defined over the energy changes after L times updates. Let (θ, h) denote the initial state, (θ^*, h^*) the updated states after L times, we have

$$Q(W \to W') = \frac{1}{\mathcal{Z}} \exp\{H(\theta, h) - H(\theta^*, h^*)\} \tag{17}$$

where $\mathcal{Z}$ is a normalization constant. We set $\mathcal{Z}$ such that the sum of probabilistic of all the proposals is unit one.

Algorithm 1 Inference .

```
**Input:** a single image;
**Output:** parse graph and 3D scene model;
Initialization: detecting straight edge segments; partitioning superpixels; initialize the parse graph $G$;
Iterate until convergence,
    - Randomly select one of the two MCMC dynamics;
    - Make proposals accordingly to reconfigure the current parse graph;
    - Accept the change with a probability
```

Hamiltonian dynamics can make proposals that are far from the current solution state and can still be accepted with high probability, i.e., distant proposals, because it exploits the steepest descent direction and current moment, rather than probabilistic distribution in other MCMC methods [Liu *et al.*, 2014]. In addition, Hamiltonian dynamics has been approved to be reversible and ergodic [Almeida, 1992].

Dynamic III: Merge/split planar regions This pair of jumps is used to regroup superpixels around the boundaries of planar regions. We obtain the list of candidate proposals for the merge/split dynamics as follows. Firstly, we select the superpixels that locate on the boundaries of two different regions and use them as graph nodes. These superpixels are usually with big ambiguities. Secondly, we link every two adjacent nodes to form an adjacent graph, and measure the edge weight using the appearance similarities of neighbor superpixels. Thirdly, we sample the edge status of 'on' or 'off' based on their edge weights to obtain connected components (CCP). We select one of the CCPs and change its semantic label to get a new solution state W'. This procedure is similar to that used by Barbu et al. [Barbu and Zhu, 2007] for graph labeling problem. Let CCP_i^k denote the i^{th} CCP, $g(\text{CCP}_i^k|W)$ denote the confidence of its label in the solution W, the proposal probability for selecting the i^{th} candidate is defined as follows:

$$Q(W \to W') = \frac{g(\text{CCP}_i^k|W')/g(\text{CCP}_i^k|W)}{\sum_j g(\text{CCP}_j^k|W')/c(\text{CCP}_j^k|W)} \tag{18}$$

The confidence $g(\text{CCP}_i^k|W)$ is defined as the appearance similarity between the selected CCP and the other nodes with the same color [Barbu and Zhu, 2007].

Algorithm 1 summarizes the proposed inference algorithm. It starts with an initial parse graph and uses a set of reversible dynamics to reconfigure the parse graph until convergence. Different from previous sampling-based methods [Liu *et al.*, 2014] [Tu and Zhu, 2002], the proposed algorithm can make distant proposals in both continuous and discrete space and thus can converge fast to the target distribution.

5 Experiment

Datasets We use the three datasets [Liu *et al.*, 2014]: CMU dataset, LMW-A, and LMW-B. The CMU dataset was originally collected by Hoiem et al. [Hoiem *et al.*, 2008]. It includes a subset of 100 images provided by Liu et al. [Gupta *et al.*, 2010]. We used 50 images for training and the rest for testing as [Gupta *et al.*, 2010]. *LMW-A* consists of 50 images from the collections in [Hoiem *et al.*, 2008]. There are 4.6 VPs per image on average. *LMW-B* consists of 50 images from the dataset of EurasianCities in [E.Tretyak *et al.*, 2012] with 4.2 VPs per image on average. We further collect 950 images from different sources, i.e. *LMW-C*, and manually annotate VPs, region labels and surface normal orientations . These images are selected from the PASCAL VOC [Everingham *et al.*, 2015] and Labelme projects [Russell *et al.*, 2007]. There are 3.5 VPs per image on average.

Baseline We compare our method with three popular single-view 3D reconstruction methods: i) the geometric parsing method Hoiem et al. [Hoiem *et al.*, 2005]; ii) the method by Gupta et al. [Gupta *et al.*, 2010], and iii) the recently proposed attributed grammar method by Liu et al. [Liu *et al.*, 2014]. We use the implementation parameters in their respective papers. To evaluate the effects of individual grammar rules, we implement three variants of the proposed method in order . i) *Ours-I*, that explores geometric relationships between lines/edges with three grammar rules: $R1$ (co-linear), $R2$ (orthogonality), and $R3$ (attachment). ii) *Ours-II*, that explores geometric relationships between planars/regions with the grammar rules: $R4$ (co-planar) and $R5$ (supporting). iii) *Ours-III*, that uses all five grammar rules.

Implementation To calibrate camera, we assume the camera optical center coincides with the image center. Thus, we select the parameter configuration that achieves the maximum log-probability. Similar simulation based maximum likelihood estimation (MLE) method has been used in previous works [Tu and Zhu, 2002] [Zhao and Zhu, 2011]. We train our models on the subset of CMU dataset, and use other three datasets for testing. We train our models on the subset of CMU dataset, and use other three datasets for testing. To make proposals for R_1 through R_5, we set the spatial distance between two children nodes (i.e. edge or planar region) to be less than 40 pixels, and the orientation angles, if applicable, to be less than 10 degrees. We extract the appearance features in [Hoiem *et al.*, 2005] from image regions. All images without groundtruth annotations in our dataset are used for the self-taught learning. The maximal iterations of HMC algorithm is fixed to 2000. For each image, the average processing time is 50 seconds on a Workstation(i7@3.6GHZ with 16GB memory).

Qualitative Evaluation Fig. 3 visualizes exemplar results by the proposed method Ours-III. We plot input images in the 1^{st} column. In the 2^{nd} column, we show the synthesized edge maps from novel viewpoints, obtained by applying the dynamic-I only (i.e., optimizing Eq. 2). We use the Matlab optimization toolbox to solve the quadratic programming problem. In the 3^{rd} column, we plot the synthesized images from novel viewpoints. In the next two columns, we plot the depth maps recovered by our method and the method by Hoiem et al. [Hoiem *et al.*, 2008], respectively. The obtained depth maps are much better than those by [Hoiem *et al.*, 2008]. Note that the method in [Hoiem *et al.*, 2008] [Gupta *et al.*, 2010] [Liu *et al.*, 2014] needs a post-processing step to approximate the depth map. In contrast, our method directly optimizes 3D depth values while respecting different types of geometric constraints.

Quantitative Results For *surface orientation* estimation,

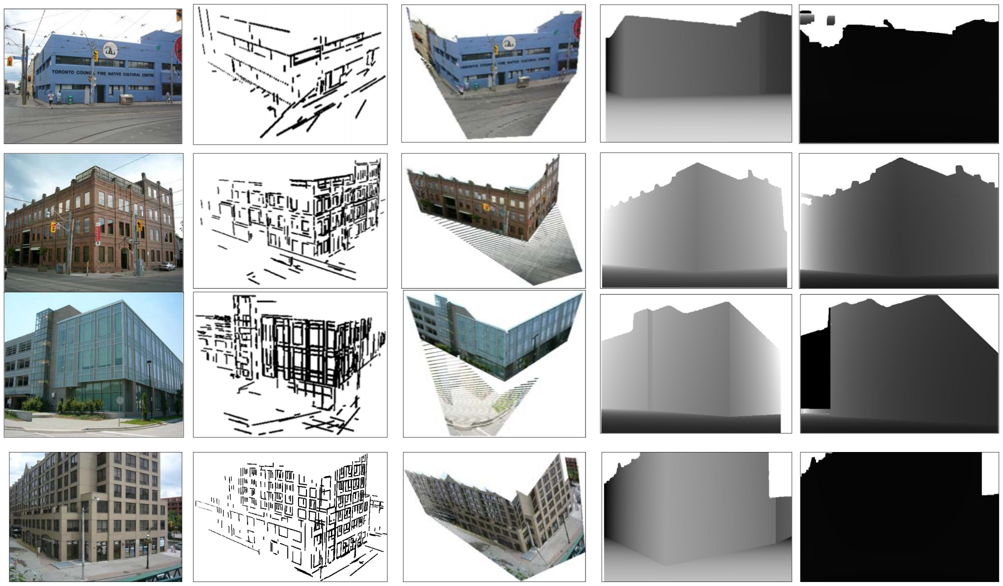

Figure 3: Results on LMW-A. Column-1: input images; Column-2: reconstructed 3D edge mode; Column-3: novel viewpoint synthesized; Column-4: depth map recovered by the proposed method Ours-III; Column-5: depth map by Hoiem et al.[16]

Table 1: Numerical comparisons on surface orientation.

	CMU [Hoiem *et al.*, 2008]	LMW-A	LMW-B	LMW-C
Ours-III	**82.31** %	**72.56** %	**70.90** %	**64.17** %
Ours-II	79.14 %	71.92 %	65.89 %	62.34 %
Ours-I	78.39 %	79.46 %	64.35 %	61.47 %
Liu et al. [Liu *et al.*, 2014]	76.34 %	67.90 %	64.30 %	62.34 %
Hoiem et al. [Hoiem *et al.*, 2008]	68.80 %	56.30 %	52.70 %	53.28 %
Gupta et al. [Gupta *et al.*, 2010]	73.72 %	62.21 %	59.21 %	58.39 %

Table 2: Numerical comparisons on region labelling

	CMU[Hoiem *et al.*, 2008]	LMW-A	LMW-B	LMW-C
Ours-III	**85.42**%	**73.51** %	**73.29** %	**79.81** %
Ours-II	82.19%	71.48 %	72.54 %	78.63 %
Ours-I	70.32%	69.72 %	71.08 %	77.28 %
Liu et al. [Liu *et al.*, 2014]	72.71%	66.45%	65.14 %	63.17 %
Hoiem et al. [Hoiem *et al.*, 2005]	65.32 %	58.37%	57.70 %	59.25 %
Gupta et al. [Gupta *et al.*, 2010]	68.85%	59.21%	60.28%	60.19%

we use the metric of *accuracy*, calculated by the percentage of pixels that have the correct label and averaged over the test images.

Table 1 reports the numerical comparisons on four datasets. Only results on verticle classes are reported. From the results, we can observe the following. Firstly, the proposed *Ours-III* clearly outperforms other baseline methods on all the four datasets. Taking the CMU dataset for instance, the method by Gupta et al. [Gupta *et al.*, 2010] has an average performance of 73.72%, whereas ours performs at 82.31%. On the other three datasets that have accurate normal orientation annotations, the improvements by our method are even more. As stated by Gupta et al. [Gupta *et al.*, 2010], it is hard to improve vertical subclass performance. Our method, however, can improve these two baselines with large margins. Secondly, *Ours-III* clearly outperforms other two variants, i.e., *Ours-I* and *Ours-II* that use less types of grammar rules. These comparisons justify the effectiveness of the proposed grammar model. Thirdly, *Ours-III* has good margins over our previous method [Liu *et al.*, 2014]. Although [Liu *et al.*, 2014] follows the same methodology, this work directly optimize pixel-wise 3D coordinates while respecting the variety of knowledge imposed with grammar rules, which leads to better performance.

Table 2 reports the *region labeling* performance on the four datasets. We use the *best spatial support* metric as [Gupta *et al.*, 2010], which first estimates the best overlap score of each ground truth labeling and then averages it over all ground-truth labeling. Our method improves the method [Gupta *et al.*, 2010] with the margins of 9.47, 16.57, 14.30, 13.10 and 19.20 percentages on the four datasets, respectively. Note that all the three variants of our methods outperform the baseline methods, which demonstrates that jointly solving reconstruction of lines and planes can bring considerable improvements in region labeling.

6 Conclusion

This paper introduced a grammar model for single-view 3D scene reconstruction. Each grammar rule describes a geometric relationship between straight lines/edges and planes/regions. We specify a probability model to deal with uncertainty reasoning, and introduce a Hybrid Monte Carlo (HMC) algorithm that can make distant proposals in both continuous and discrete spaces. Extensive evaluations on challenging images show that our method can clearly outperform the state-of-the-art methods. This paper contributes a generic framework for optimizing joint representation that comprises of both continuous and discrete variables. The developed techniques can be applied to solve existing vision tasks, e.g. joint tracking and segmentation, or motivate novel vision tasks.

References

[Almeida, 1992] A. Almeida. Hamiltonian systems: Chaos and quantization. *Cambridge monographs on mathematical physics*, 1992.

[Audin and Babbitt, 2008] M. Audin and D. Babbitt. Hamiltonian systems and their integrability. *American Mathmatical Society*, 2008.

[Barbu and Zhu, 2007] A. Barbu and S-C. Zhu. Generalizing swendsen-wang to sampling arbitrary posterior probabilities. *TPAMI*, 2007.

[Cabral and Furukawa, 2014] R. Cabral and Y. Furukawa. Piecewise planar and compact floorplan reconstruction from images. In *CVPR*, 2014.

[Cipolla *et al.*, 1999] R. Cipolla, T. Drummond, and D. Robertson. Camera calibration from vanishing points in images of architectural scenes. In *BMVC*, 1999.

[E.Tretyak *et al.*, 2012] E.Tretyak, O. Barinova, P. Kohli, and V. Lempitsky. Geometric image parsing in man-made environments. *IJCV*, 97(3):305–321, 2012.

[Everingham *et al.*, 2015] M. Everingham, S. Eslami, L. Van Gool, C. Williams, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015.

[Furukawa *et al.*, 2009] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Manhattan-world stereo. In *CVPR*, 2009.

[Gioi *et al.*, 2008] R. Gioi, J. Jakubowicz, J. Morel, and G. Randall. Lsd: A fast line segment detector with a false detection control. *TPAMI*, 2008.

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

[Haene *et al.*, 2013] C. Haene, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3d scene reconstruction and class segmentation. In *CVPR*, 2013.

[Han and Zhu, 2003] F. Han and S-C. Zhu. Bayesian reconstruction of 3d shapes and scenes from a single image. In *Proc. of Int'l workshop on High Level Knowledge in 3D Modeling and Motion*, October 2003.

[Han and Zhu, 2009] F. Han and S-C. Zhu. Bottom-up/top-down image parsing with attribute grammar. *TPAMI*, 31(1):59–73, 2009.

[Heitz *et al.*, 2008] G. Heitz, S. Gould, A. Saxena, and D. Koller. Cascaded classification models: Combining models for holistic scene understanding. In *NIPS*, 2008.

[Hejrati and Ramanan, 2014] M. Hejrati and D. Ramanan. Analysis by synthesis: Object recognition by object reconstruction. In *CVPR*, 2014.

[Hoiem *et al.*, 2005] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. *ICCV*, 2005.

[Hoiem *et al.*, 2008] D. Hoiem, A. Efros, and M. Hebert. Closing the loop on scene interpretation. In *CVPR*, 2008.

[Koutsourakis *et al.*, 2009] P. Koutsourakis, L. Simon, O. Teboul, G. Tziritas, and N. Paragios. Single view reconstruction using shape grammars for urban environments. In *ICCV*, pages 1795–1802, 2009.

[Kundu *et al.*, 2014] A. Kundu, Y. Li, F. Daellert, F. Li, and J. Rehg. Joint semantic segmentation and 3d reconstruction from monocular video. In *ECCV*, 2014.

[Liu *et al.*, 2014] X. Liu, Y. Zhao, and S-C. Zhu. Single-view 3d scene parsing by attributed grammar. In *CVPR*, 2014.

[Mobahi *et al.*, 2012] H. Mobahi, Z. Zhou, A. Yang, and Y. Ma. Holistic 3d reconstruction of urban structures from low-rank textures. In *ACCV*, 2012.

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

[Pero *et al.*, 2012] L. Pero, J. Guan, E. Hartley, B. Kermgard, and K. Barnard. Bayesian geometric modeling of indoor scenes. In *CVPR*, 2012.

[Pero *et al.*, 2013] L. Pero, J. Guan, E. Hartley, B. Kermgard, and K. Barnard. Understanding bayesian rooms using composite 3d object models. In *CVPR*, 2013.

[Ren and Malik, 2003] X. Ren and J. Malik. Learning a classification model for segmentation. In *ICCV*, 2003.

[Russell *et al.*, 2007] B. Russell, A. Torralba, K. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 2007.

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

[Schwing *et al.*, 2013] A. Schwing, S. Fidler, M. Pollefeys, and R. Urtasun. Box in the box: Joint 3d layout and object reasoning from single images. In *ICCV*, 2013.

[Tu and Zhu, 2002] Z. Tu and S-C. Zhu. Image segmentation by data-driven markov chain monte carlo. *TPAMI*, 24(5):657–673, 2002.

[Xiao *et al.*, 2013] J. Xiao, A. Owens, and A. Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *ICCV*, 2013.

[Zhang *et al.*, 2013] H. Zhang, A. Geiger, and R. Urtasun. Understanding high-level semantics by modeling traffic patterns. In *ICCV*, 2013.

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