Single-Image 3D Scene Parsing Using Geometric Commonsense

Chengcheng Yu[1*], **Xiaobai Liu**[2*], **Song-Chun Zhu**[1]
[1] Department of Statistics, University of California, Los Angeles (UCLA), CA
[2] Department of Computer Science, San Diego State Univesity (SDSU), CA
chengchengyu@ucla.edu, xiaobai.liu@mail.sdsu.edu, sczhu@stat.ucla.edu

Abstract

This paper presents a unified grammatical framework capable of reconstructing a variety of scene types (e.g., urban, campus, country etc.) from a single input image. The key idea of our approach is to study a novel commonsense reasoning framework that mainly exploits two types of prior knowledge: (i) prior distributions over a single dimension of objects, e.g., that the length of a sedan is about 4.5 meters; (ii) pair-wise relationships between the dimensions of scene entities, e.g., that the length of a sedan is shorter than a bus. These unary or relative geometric knowledge, once extracted, are fairly stable across different types of natural scenes, and are informative for enhancing the understanding of various scenes in both 2D images and 3D world. Methodologically, we propose to construct a hierarchical graph representation as a unified representation of the input image and related geometric knowledge. We formulate these objectives with a unified probabilistic formula and develop a data-driven Monte Carlo method to infer the optimal solution with both bottom-to-up and top-down computations. Results with comparisons on public datasets showed that our method clearly outperforms the alternative methods.

1 Introduction

Commonsense or commonsense reasoning [Davis and Marcus, 2015; Davis *et al.*, 1993] functions in all parties of Artificial Intelligence (AI), including language, vision, planning, etc. It basically studies the consensus reality, knowledge and reasoning available to the overwhelming majority of people and attracted a lot of attention in the past literature. In the field of computer vision, however, it is still unclear how to formally describe visual commonsense knowledge, or how commonsense can be used to enhance the understanding of images or videos [Zitnick and Parikh, 2013]. This work aims to fill in this gap by studying the reasoning of geometric commonsense for 3D scene parsing. Such a parsing task aims to segment both low-level scene entities (e.g., straight edges, semantic regions) and object-level scene entities (e.g., human, vehicles) in 2D images, and estimate their geometric dimensions in the 3D world [Hoiem *et al.*, 2005; Del Pero *et al.*, 2013; Liu *et al.*, 2014; Wang *et al.*, 2015a; Mottaghi *et al.*, 2016]. Most existing 3D parsing algorithms [Hoiem *et al.*, 2008] are designed for a particular type of scene categories, e.g., urban [Liu *et al.*, 2014; Gupta *et al.*, 2010], indoor [Wang *et al.*, 2015b]. However, a practical AI system, e.g., autonomous driving, usually needs to deal with a wide variety of scene categories.

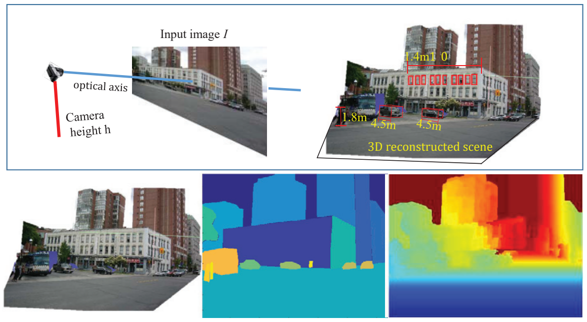


Figure 1: Single-view 3D scene reconstruction using Geometric commonsense. Top: the world is full of commonsense over geometric dimensions, e.g., that a sedan is about 4.5 meters long. Bottom: exemplar result of the proposed method, including synthesized image (left), planar segmentation (middle), and depth map (right).

Our solution to the above challenges is motivated by the fact that we human beings, unconsciously sometimes, utilize rich prior knowledge of the geometric dimensions of scene entities to understand the scene structures in images or videos [Davis *et al.*, 1993]. This knowledge can be roughly divided into two types: i) prior distributions over a single dimension of objects, e.g., the height of a female adult is about 1.75 meters, or that the length of a sedan is about 4.5 meters; ii) pair-wise comparisons between the dimensions of different scene entities at both object-level, e.g., human, windows, vehicles, etc., and part-level, e.g., straight edges, planar regions, etc. As illustrated in Figure 1, for example, the window edges on the same facade are parallel to each other and are orthogonal to the edges on the ground, a building is higher than a human, or the length of all sedans are roughly equal. These unary and pair-wise knowledge, once acquired,

[*] The first two authors contributed equally to this work. All correspondences should be addressed to Xiaobai Liu (email: xiaobai.liu@mail.sdsu.edu).

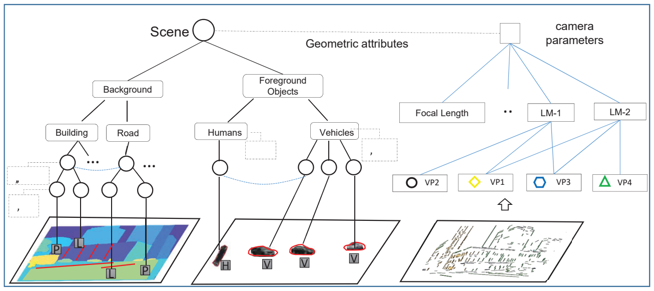


Figure 2: Attribute Parse Graph. Every graph node represents a scene entity, and is augmented with a set of geometric attributes ($\hat{l}$, line direction; d, depth;$\bar{n}$, normal orientation; w, width; h, height; f, focal length; θ, camera angle). There are five types of terminal nodes: L, line, P, planar surface; T, texture; H, human; V, vehicles.

are valid across images and scene categories, and thus form the commonsense in geometric space, called *geometric commonsense*.

This paper presents a stochastic attribute scene grammar for representing both visual content and geometric commonsense constraints for parsing a single image in 3D world. Our grammar model recursively decomposes a scene into a small number of scene primitives (e.g., straight lines, planar regions, vehicles) arranged by a set of spatial relations. This results in a hierarchical representation, called parse graph, which has a root node for the whole scene and a terminal node for each scene primitive. Each graph node is described with a set of geometric attributes, e.g., 3D position, normal direction etc. We augment the attribute variables using a set of common sense knowledge that is either pre-defined or mined from online databases. These commonsense facts are defined over geometric dimensions of various categories (e.g., human height), and are widely used by human beings to understand visual inputs. Since these geometric commonsense knowledge facts are generic enough, our method is capable of reconstructing a wide variety of scene types, e.g., urban, suburban, campus etc. In contrast, most existing methods for single-view reconstruction were developed for a particular scene type (as reviewed in Section 2).

We formulate the above objectives in Bayesian framework in order to keep uncertainties during inference, which is critical to avoid pre-mature decision making. For inference, we develop an iterative data-driven Monte Carlo Markov Chain (DDMCMC) method [Tu and Zhu, 2002] that searches the optimal parse graph with both bottom-up and top-down computations. On the one hand, we partition images into compositional elements, extract visual features for elements (e.g., color) and measure their pair-wise similarities, and group elements in a bottom-up fashion to create upper-level graph nodes. On the other hand, we can decompose a graph node into multiple children nodes or propagate the attributes of a parent node to all its offspring in a top-down fashion. Both bottom-up and top-down computations are intelligently scheduled by the Metropolis Hasting Principle [Tu and Zhu, 2002] to guarantee convergence toward the posterior probability. To evaluate the proposed method, we collect an image dataset that covers five different categories: country, road, suburban, campus and urban, and manually annotate their semantic and geometric labels in 3D. This dataset is different from the previous datasets [Liu *et al.*, 2014; Hoiem *et al.*, 2008] which mostly include one or two types of scenes. Results with comparisons demonstrated that the proposed method clearly outperforms the alternative methods in the recent literature [Liu *et al.*, 2014;].

2 Relationships to Previous Works

This work is closely related to four research streams in Computer Vision and Artificial Intelligence (AI).

Commonsense Reasoning is one of the long-standing tasks in AI [Davis and Marcus, 2015; Davis *et al.*, 1993] [Davis *et al.*, 1993] and has recently attracted a lot of attentions in the field of computer vision. Such commonsense knowledge were used as context information to enhance visual recognition [Fouhey *et al.*, 2014; Grabner *et al.*, 2011], scene understanding [Wang *et al.*, 2013], activity recognition [Wang *et al.*, 2007; Wyatt *et al.*, 2005; del Rincón *et al.*, 2013] and affordance prediction [Zhu *et al.*, 2014; 2015; Wang *et al.*, 2007]. However, there is still no formal representation of visual commonsense in the past literature, which restricts the generalization capability of the developed techniques. Moreover, these works share the same insight that visual commonsense only functions in a high-level semantic understanding of images, rather than low-level pixel-wise understanding, which are not necessarily bold. In contrast, this paper studies geometric commonsense extracted from both low-level and high-level scene entities and defines a unified presentation for describing such knowledges.

3D scene reconstruction methods are mostly referred to Shape-from-X, where X stands shading [Zhang *et al.*, 1999], contour [Toppe *et al.*, 2013], focus/defocus [Nayar and Nakagawa, 1990], texture [Criminisi and Zisserman, 2000], motion [Dellaert *et al.*, 2000], photometric stereo [Lucas *et al.*, 1981] etc. Single-view reconstruction has also been studied [Hoiem *et al.*, 2005; Heitz *et al.*, 2009; Gupta *et al.*, 2010]. Most recently, Liu et al. [Liu *et al.*, 2014;] proposed an explicit image representation and presented a joint solution of 2D recognition and 3D reconstruction. These efforts utilized various modeling assumptions to guide the reconstruction process. For example, the shape-from-texture methods [Davis and Marcus, 2015] assume that the scene comprises of homogeneous texture, and the shape-from-contour methods assume that contours are known to be projections of curves on a planar surface. These methods would fail to work while dealing with complex scenes for which none of a single modeling assumption is valid. For example, the methods by Liu et al. [Liu *et al.*, 2014;] can only be used to reconstruct Manhattan or near Manhattan type scenarios in urban scenes. In this work, we introduce a concept-of-proof framework for geometric commonsense reasoning and demonstrate how this knowledge can be used to enable the 3D reconstruction of a wide variety of scene types.

Stochastic Image Grammar has been applied for a number of image parsing problems in computer vision. Koutsourakis et al. [Koutsourakis *et al.*, 2009] proposed a shape grammar to explain building facades with levels of details.

Category	Variable	Sub-Type	MEAN	STD
Human	Height	Female	1.6475	0.1934
		Male	1.7741	0.2217
Vehicle	Length	Sedan	4.6585	0.2395
Window	(H, W)	-	(1.5234, 1.4647)	(0.3471, 0.8462)
Door	(H, W)	-	(2.0163, 0.8514)	(0.0689, 0.0726)

Table 1: Geometric commonsense of type-I: distribution over absolute dimensions. L: length; W: width; H: height.

Researchers have [Han and Zhu, 2009; Zhao and Zhu, 2011; Del Pero *et al.*, 2011] specified generative scene grammar 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. Liu et al. [Liu *et al.*, 2014] proposed an attribute grammar for 3D scene modeling to enable compact representation of images. With only a few grammar rules, the grammar model can explain most urban images and achieved state-of-the-art performance on a few public image benchmarks. In this work, we will extend the attribute grammar to model geometric commonsense knowledge of scene entities for the reconstruction of various scene categories.

3 Our Approach

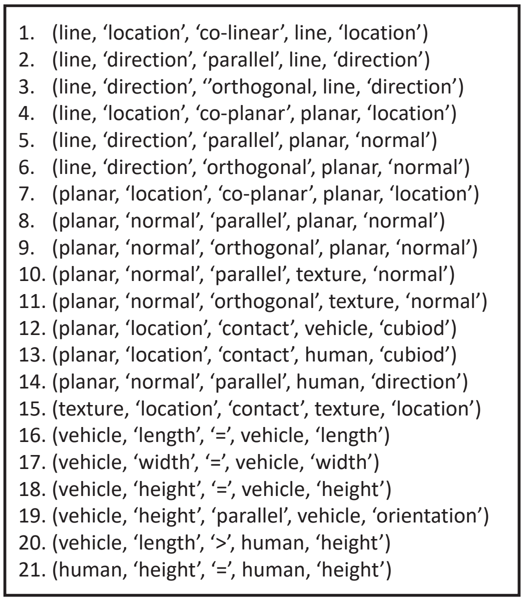
The objective of this work is to parse a single image into a set of scene entities, e.g., straight lines, surfaces, objects, etc., and reason their geometric attributes in 3D, e.g., 3D positions, normal direction etc. We consider a wide variety of scene types, e.g., urban, garden, or road.

3.1 Stochastic Scene Grammar

An attribute scene grammar is specified by a 5-tuple: $G = (V_N, V_T, S, R, P)$ where V_N and V_T are the sets of non-terminal nodes and terminal nodes respectively, S is the initial node for the scene, R is a set of production rules for spatial relationships, P is the probability for the grammar. Our grammar model results in a hierarchical graph representation, i.e. **Parse Graph**, to represent the semantic content of an image. Every graph node represents a scene primitive, including straight lines, planar regions, homogeneous or inhomogeneous texture regions, human, and vehicles, and their composites. The first three image primitives are conventionally considered as background while the other two entities are foreground objects. Figure 2 illustrates a typical parse graph used in this paper. Single-view 3D scene parsing is equivalent to creating a plausible parse graph from the input image. Note that in this work we focus on parsing outdoor images, but the proposed technique can be easily extended to deal with indoor images as well.

For every graph node, we introduce a set of geometric **Attributes** to describe their dimensions in 3D, as summarized below.

- Attributes of a straight line include its position and orientation direction in 3D. In addition, we divide all straight lines into two categories: parallel lines and non-parallel lines [Liu *et al.*, 2014]. Multiple orthogonal families of parallel lines forms one of the Manhattan frames [Liu *et al.*, 2014].



1. (line, ‘location’, ‘co-linear’, line, ‘location’)
2. (line, ‘direction’, ‘parallel’, line, ‘direction’)
3. (line, ‘direction’, ‘’orthogonal, line, ‘direction’)
4. (line, ‘location’, ‘co-planar’, planar, ‘location’)
5. (line, ‘direction’, ‘parallel’, planar, ‘normal’)
6. (line, ‘direction’, ‘orthogonal’, planar, ‘normal’)
7. (planar, ‘location’, ‘co-planar’, planar, ‘location’)
8. (planar, ‘normal’, ‘parallel’, planar, ‘normal’)
9. (planar, ‘normal’, ‘orthogonal’, planar, ‘normal’)
10. (planar, ‘normal’, ‘parallel’, texture, ‘normal’)
11. (planar, ‘normal’, ‘orthogonal’, texture, ‘normal’)
12. (planar, ‘location’, ‘contact’, vehicle, ‘cubiod’)
13. (planar, ‘location’, ‘contact’, human, ‘cubiod’)
14. (planar, ‘normal’, ‘parallel’, human, ‘direction’)
15. (texture, ‘location’, ‘contact’, texture, ‘location’)
16. (vehicle, ‘length’, ‘=’, vehicle, ‘length’)
17. (vehicle, ‘width’, ‘=’, vehicle, ‘width’)
18. (vehicle, ‘height’, ‘=’, vehicle, ‘height’)
19. (vehicle, ‘height’, ‘parallel’, vehicle, ‘orientation’)
20. (vehicle, ‘length’, ‘>’, human, ‘height’)
21. (human, ‘height’, ‘=’, human, ‘height’)

Figure 3: Geometric commonsense of Type-II: a total of 21 pair-wise relationships between scene elements.

- Attributes of a planar or texture region include its geometric properties, i.e., position, normal direction and size in 3D, and semantic labels, e.g., ground, building, grass, road, etc.
- Attributes of a human include one’s geometric properties, i.e., positions and human height in 3D, and fine-grained semantic labels, i.e., genders, children/adult, races.
- Attributes of vehicles include positions and dimensions (length, width, height) in 3D, and catalog information, i.e., categories (sedan, car, bus).

Figure 2 visualizes an attribute graph where each node is augmented with a set of geometric attributes.

Geometric Commonsense are defined over the geometric attributes of graph nodes. There are two types of geometric commonsense.

- Type-I: prior distributions over dimensions of image entities. This type of knowledge includes, for instance, the average height of adult, the width of door, or the length of sedans. Table 1 summarizes some of these statistics (mean and standard deviation) collected by the U.S. Department of Health and Human Services [1]. These statistics can be used to regularize the creation of the desired parse graph.
- Type-II: pair-wise relationship between scene entities. This type of knowledge is defined over the comparisons of dimensions of two scene entities that are either from the same or different categories. This leads to a set of commonsense equations, each of which is defined as a 5-tuple: (entity-1, attribute-1, operator, entity-2, attribute-2), where ”operator” represents, for example, ”parallel”,

[1] U.S. Department of Health and Human Services. Anthropometric Reference data for children and adults: united states, 2007-2010. Vital and health statistics, series 11, number 252, October, 2012

”orthogonal”, ”equal”, or ”>”. Figure 3 lists the 21 pairwise relationships used in this paper.

3.2 Bayesian Formula

Given an input image, our goal is to create an attribute parse graph that is a valid interpretation of the input image. Let G denote the parse graph, we solve the optimal parse graph by maximizing a posterior (MAP): $G^* = \arg\max_G P(G|I)$ where I is the input image.

According to Bayes' rule, we rewrite the posterior distribution $P(G|I)$ as

$$P(G|I) \propto P(G)P(I|G) \tag{1}$$

where $P(G)$ is the prior model and $P(I|G)$ is the likelihood model.

The prior model is defined over the validness of commonsense constraints. Let $A.cset$ denote the set of nodes which are linked to the node A, $X(A)$ be the attributes of A. For a graph node, let i index the type-I commonsense distributions, j index the type-II commonsense equations. Then, we can define the prior model as,

$$P(G) = \frac{1}{Z}\exp\{-E(G)\} \tag{2}$$

where Z is the normalization constant. The energy function is defined as:

$$E(G) = \sum_{A\in G}\sum_{i}\mathbf{h}_i(X(A)) + \sum_{B\in A.cset}\sum_{j}\mathbf{g}_j[X(A), X(B)] \tag{3}$$

where $\mathbf{h}_i()$ denotes the normal distribution over the i-th attribute (with mean and variance in Table-1). The model $\mathbf{g}_j(X(A), X(B))$ is defined as

$$\mathbf{g}_j(X(A), X(B)) = C \cdot \mathbf{1}(X(A), X(B), j) \tag{4}$$

where C is a constant, and $\mathbf{1}(X(A), X(B), j)$ returns 1 if the j-th commonsense equation between A and B holds; otherwise 0. Note that given a parse graph G and its attributes, it is straightforward to calculate $\mathbf{h}_i()$ and check $\mathbf{g}_j()$ by definitions.

The data likelihood model $P(I|G)$ is defined over the terminal nodes of the parse graph G. In this work, there are five types of terminal nodes: line (L), planar (P), texture (T), human (H) and vehicles (V). For a graph node A, let $A.type \in \{L, P, T, H, V\}$ return the type of terminal nodes. We specify a likelihood model:

$$P(I|G) = \prod_{A\in V_T, k=A.type}\frac{Pg^k(A|I)}{Pb^k(A|I)}\prod_{B\in I-V_T}Pb^k(B|I) \tag{5}$$

where $Pg^k()$ and $Pb^k()$ denote the foreground distributions and background distributions. The likelihood ratio $\frac{Pg^k(A|I)}{Pb^k(A|I)}$ is defined for each of the five terminal node types. In implementation, we approximate the ratio using detection scores of terminal nodes. If A is a straight line, for example, we define $\frac{Pg^k(A|I)}{Pb^k(A|I)}$ to be the edge confidences by [Li *et al.*, 2012]; if A represents a human or a vehicle, we use the output confidences by the object detection method [Felzenszwalb *et al.*, 2010]. In addition, our method allows regions not covered by any terminal nodes, and aims to explain these regions using the background distributions. In particular, we define $Pb^k(B|I)$ using the confidences of classifying B as the k-th terminal.

3.3 Bottom-up and Top-Down Inference

Given an input image, our inference aims to create a plausible attribute parse graph so that all attributes satisfy the geometric commonsense constraints, including both type-I and type-II knowledge. This is an intractable problem since the optimal parse graph is defined in a joint space: discrete labels (e.g., segmentation) and continuous attributes (e.g., location or orientation). To search the optimal graph, we develop a data-driven Markov Chain Monte Carlo (DDMCMC) method [Tu and Zhu, 2002] which starts with an initial graph and then reconfigures the graph with a set of dynamics to simulate a Markov Chain in the joint solution space. Two dynamics are paired with each other to guarantee the convergence to the target distribution $P(G|I)$. Our algorithm follows the Metropolis-Hastings strategy [Tu and Zhu, 2002]. Given the current graph G, we apply a dynamic to get a new graph G', and accept it with the following probability,

$$\alpha(G \to G') = \min(1, \frac{P(G'|I)Q(G \to G')}{P(G|I)Q(G' \to G)}) \tag{6}$$

where $Q()$ is the proposal probability.

The initial graph includes a root node and a set of terminal nodes. Then we introduce four dynamics that are performed in either bottom-up or top-down fashion.

Dynamics 1-2: birth/death of nonterminal nodes are used to create or delete a nonterminal node and thus transition the current parse graph to a new one. To create a new graph node, we need to select two or more candidate graph nodes to group with three criteria: i) being spatially adjacent; ii) belonging to the same semantic category; or iii) conveying type-II commonsense knowledge. These would lead to a set of candidate nodes, and we select one of them as a new graph node. The proposal probability is defined over the detection scores of these graph nodes. To delete a graph node, we need to randomly select a graph node and then remove it to reconfigure the graph. The proposal probability for this dynamic is set to be uniform, i.e., all proposal candidates have the equal chance to be selected.

Dynamics 3: changing attribute is used to modify the attributes of graph nodes. We will randomly select a graph node as well as an attribute, and assign a different value to this attribute, e.g., normal orientation for planar regions. The changes, as introduced, will be accepted with a probability. We set the proposal probability for this dynamic to be uniform.

Dynamics 4: attribute propagation is a top-down process that assigns the attributes of parent nodes to children nodes, and used to guarantee the consistency in the hierarchy. To do so, we will randomly select a parent node, and propagate all its attributes to the offspring nodes. Note that the previous works [Liu *et al.*, 2014] only have bottom-up computations which might get stuck during inference because its time con-

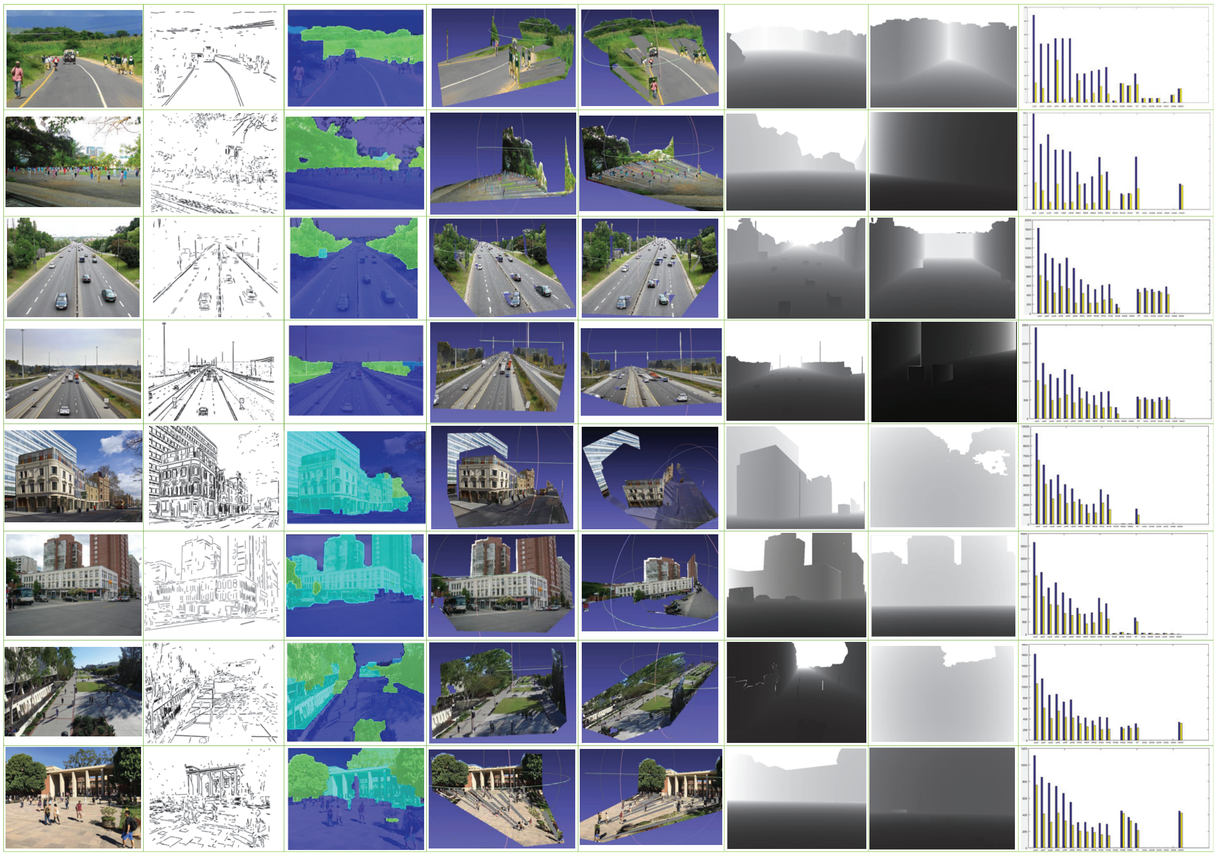


Figure 4: Exemplar results of the proposed method.

suming to flip the attributes of a subtree only using bottom-to-up dynamics.

4 Experiments

Dataset To evaluate the generalization capability of the proposed method, we collect 500 images for each of the following categories: 1) country; 2) suburban; 3) road; 4) campus; and 5) urban, resulting in a collection of 2,500 images. The urban images mostly follow the Manhattan assumption [Liu *et al.*, 2014] while the country images do not. These images are selected from existing datasets [Hoiem *et al.*, 2008; Liu *et al.*, 2014; Everingham *et al.*, 2015]. For every category, we use 100 images for training and use the rest for testing. We manually annotate semantic and geometric labels for every image. To label an image, we first divide every image into three main categories: ground, sky and vertical, and further divide the vertical category into porous, solid and oriented surfaces. The number of planar orientations is equal to the number of horizontal vanishing points, as defined in [Liu *et al.*, 2014]. For quantitative comparisons, all images are manually annotated with vanishing points (VPs), surface segmentation and surface orientation (represented by the correspondent VPs for each surface).

Baseline We compare the proposed method with three popular methods by Hoiem et al. [Hoiem *et al.*, 2008], Gupta et al. [Gupta *et al.*, 2010], and Liu. et al. [Liu *et al.*, 2014]. Among these algorithms, the method in [Hoiem *et al.*, 2008] only utilizes appearance models, the method in [Gupta *et al.*, 2010] tries to reason block structures, and the method in [Liu *et al.*, 2014] tries to explore the parallel or orthogonal relationships between planar regions, which are special cases of the proposed geometric commonsense equations. These baseline methods are restricted to their capability to deal with arious scene types. In contrast, the proposed method can adaptively find suitable commonsense knowledge and thus are applicable to all scene types.

Implementation We use the method by Ren et al. [Ren and Malik, 2003] to partition each image into 200-300 superpixels, the method by Li et al. [Li *et al.*, 2012] to detect straight lines as well as vanishing points, the method [Hoiem *et al.*, 2008] to identify planar regions and texture regions, and the method [Felzenszwalb *et al.*, 2010] to detect human and vehicles. These algorithms are called in pre-processing steps and the results are used as inputs for the proposed parsing algorithm. We set the maximal iteration of DDMCMC to be 2000. It costs about 1-2 minutes to converge on a Dell Workstation (i7-4790 CPU@3.6GHZ with 16GB memory). We implemented two variants of our method to evaluate the effectiveness of individual commonsense. i) *Our-I*, that only utilizes the type-II geometric commonsense and ii) *Our-II*, that utilizes both type-I and type-II commonsense equations.

Qualitative result Figure 4 visualizes a few exemplar results by the proposed method. In Columns 1, 2 and 3, we show the input images, edge maps, and semantic region partition, respectively. We further show two synthesized viewpoints in columns 4 and 5, and the estimated depth map in column 6. For comparisons, we plot the depth map obtained by Gupta et al. [Gupta *et al.*, 2010] in column 7. We can

	Country	Suburban	Road	Campus	Urban
Our-II	64.4%	60.6%	74.4%	71.6%	75.2%
Our-I	59.1%	58.6%	72.3%	69.4%	72.5%
Liu et al.	54.6%	57.3%	68.6%	62.3%	68.7%
Gupta et al.	53.8%	51.4%	55.8%	57.4%	62.2%
Hoiem et al.	52.3%	50.6%	59.3%	58.6%	63.3%

Table 2: Numerical comparisons on surface orientation.

	Country	Suburban	Road	Campus	Urban
Our-II	67.9%	65.7%	71.3%	76.1%	75.8%
Our-I	62.8%	64.8%	69.8%	74.3%	73.9%
Liu et al.	56.5%	63.3%	65.3%	70.6%	67.2%
Gupta et al.	52.1%	57.8%	58.1%	65.7%	61.3%
Hoiem et al.	53.2%	52.9%	54.4%	66.3%	64.6%

Table 3: Numerical comparisons on segmentation.

observe that the obtained 3D scene model include the variety of vivid details, contain windows, small size facades (e.g., in the first row and second row), and doors (in the third row) etc. The obtained depth map is more accurate than those by [Gupta *et al.*, 2010]. In contrast to [Gupta *et al.*, 2010] that needs a post-processing step to approximate the depth, our method directly optimizes the geometric attributes while respecting various commonsense constraints.

Quantitative result We compare the proposed method with three baseline methods for two parsing problems: surface orientation prediction and region segmentation. For *normal orientation estimation*, we use the metric of *accuracy*, i.e., the percentage of pixels that have the correct normal orientation label, and average accuracies over test images for every dataset. On the estimation of main geometric classes, i.e., 'ground', 'vertical', and 'sky', both our method and baseline methods can achieve high-quality results with accuracy 0.98 or more. Therefore, we focus on the vertical subclasses, like [Gupta *et al.*, 2010], and discard the superpixels belonging to ground and sky while calculating the accuracies of all methods.

Table 2 reports the numerical comparisons on five scene categories. From the results, we can observe the following. Firstly, the proposed method clearly outperforms the other baseline methods on all five scene categories. The improvements over country images are most significant since our method can adaptively select the informative cues over planar regions, and discard the edge/gradient cues while the other methods can not. As stated by Gupta et al. [Gupta *et al.*, 2010], it is difficult to improve vertical subclass performance. Our method, however, can improve these three baselines with large margins. Secondly, The proposed method outperforms the recent grammar based method [Liu *et al.*, 2014], which tries to create a hierarchical graph as well. As summarized, the method in [Liu *et al.*, 2014] applies Manhattan or mixture Manhattan type assumptions, which do not always hold in images. In contrast, the proposed commonsense knowledge are effective across a wide variety of images.

For *region segmentation*, 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. We discard the superpixels belonging to ground and sky while calculating the accuracies of all methods. Table 3 reports the region labeling performance on the five scene categories. Our method outperforms the method [Liu *et al.*, 2014] with the margins of 11.4, 2.4, 6.0, 5.5 and 8.6 percentages on the five categories, respectively. These comparisons show that jointly solving recognition and reconstruction has the ability to considerably improve recognition accuracies.

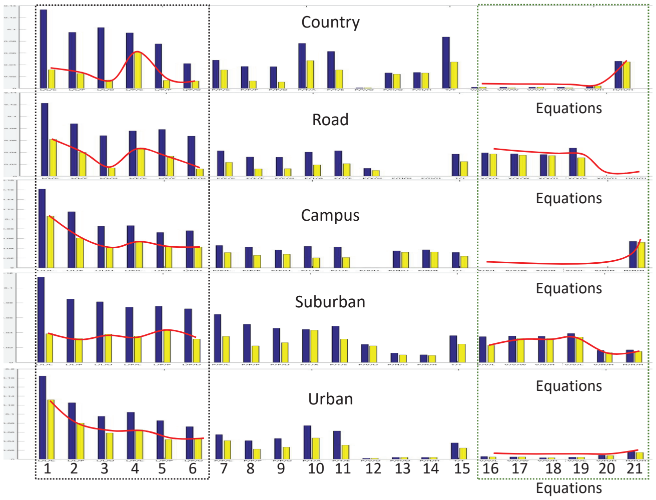


Figure 5: Average occurrence frequencies of commonsense equations on images of five scene categories.

Reasoning of valid geometric commonsense An interesting aspect of the proposed method is that we can identify the invalid commonsense equations and graph nodes that have inconsistent attributes with their children nodes. To do so, we introduce a heuristic test that comprises of two major steps. First, we solve the optimal parse graph as well as its attribute variables using the proposed DDMCMC algorithm. Second, with the optimal graph, we check if a constraint equation is satisfied using a simple threshold based method. In the 8-th column of the Figure 4, we visualize the occurrence frequencies of valid equations (yellow bars) and the total number of equations (blue bars) for every image. Figure 5 further plots the average occurrence frequencies of valid equations in individual categories. Note that the distributions of valid equations (red curves) vary significantly across categories, which shows that our method can adaptively find the suitable commonsense knowledge for various types of images.

5 Conclusion

This paper presented a probabilistic method for single-view 3D scene reconstruction using geometric commonsense and specify a generic probabilistic formula to solve multiple 3d parsing problems simultaneously. We developed a stochastic optimization algorithm to search the optimal parse graph with both bottom-to-up and top-down computations. In evaluations, we collected a new image dataset to include a variety of scene categories and annotated their 3D scene models. Results with comparisons demonstrated that our method is capable of accurately reconstructing a wide variety of scene categories. We also demonstrated that the proposed method can be used to disclose the valid commonsense knowledge used to explained an image. As the first piece of works in its catalog, our studies are able to enhance our understandings of geometric commonsense knowledge and their critical role in computer vision. The developed techniques can also be easily extended to the broader scope of commonsense reasoning.

Acknowledgements

This projct was supported by NSF #1657600, ONR MURI N00014-16-1-2007, NSF IIS #1423305, and DARPA XAI N66001-17-2-4029.

References

[Criminisi and Zisserman, 2000] Antonio Criminisi and Andrew Zisserman. Shape from texture: Homogeneity revisited. In *BMVC*, pages 1–10, 2000.

[Davis and Marcus, 2015] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103, 2015.

[Davis *et al.*, 1993] Randall Davis, Howard Shrobe, and Peter Szolovits. What is a knowledge representation? *AI magazine*, 14(1):17, 1993.

[Del Pero *et al.*, 2011] Luca Del Pero, Jinyan Guan, Ernesto Brau, Joseph Schlecht, and Kobus Barnard. Sampling bedrooms. In *CVPR*, pages 2009–2016. IEEE, 2011.

[Del Pero *et al.*, 2013] Luca Del Pero, Joshua Bowdish, Bonnie Kermgard, Emily Hartley, and Kobus Barnard. Understanding bayesian rooms using composite 3d object models. In *CVPR*, pages 153–160, 2013.

[del Rincón *et al.*, 2013] Jesús Martínez del Rincón, Maria J Santofimia, and Jean-Christophe Nebel. Common-sense reasoning for human action recognition. *Pattern Recognition Letters*, 34(15):1849–1860, 2013.

[Dellaert *et al.*, 2000] Frank Dellaert, Steven M Seitz, Charles E Thorpe, and Sebastian Thrun. Structure from motion without correspondence. In *CVPR*, volume 2, pages 557–564. IEEE, 2000.

[Everingham *et al.*, 2015] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015.

[Felzenszwalb *et al.*, 2010] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010.

[Fouhey *et al.*, 2014] David F Fouhey, Vincent Delaitre, Abhinav Gupta, Alexei A Efros, Ivan Laptev, and Josef Sivic. People watching: Human actions as a cue for single view geometry. *IJCV*, 110(3):259–274, 2014.

[Furukawa *et al.*, 2009] Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. Manhattan-world stereo. In *CVPR*, pages 1422–1429. IEEE, 2009.

[Grabner *et al.*, 2011] Helmut Grabner, Juergen Gall, and Luc Van Gool. What makes a chair a chair? In *CVPR*, pages 1529–1536. IEEE, 2011.

[Gupta *et al.*, 2010] Abhinav Gupta, Alexei Efros, and Martial Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. *ECCV*, pages 482–496, 2010.

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

[Heitz *et al.*, 2009] Geremy Heitz, Stephen Gould, Ashutosh Saxena, and Daphne Koller. Cascaded classification models: Combining models for holistic scene understanding. In *Advances in Neural Information Processing Systems*, pages 641–648, 2009.

[Hoiem *et al.*, 2005] Derek Hoiem, Alexei A Efros, and Martial Hebert. Geometric context from a single image. In *ICCV*, volume 1, pages 654–661. IEEE, 2005.

[Hoiem *et al.*, 2008] Derek Hoiem, Alexei A Efros, and Martial Hebert. Closing the loop in scene interpretation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[Koutsourakis *et al.*, 2009] Panagiotis Koutsourakis, Loic Simon, Olivier Teboul, Georgios Tziritas, and Nikos Paragios. Single view reconstruction using shape grammars for urban environments. In *ICCV*, pages 1795–1802. IEEE, 2009.

[Li *et al.*, 2012] Bo Li, Kun Peng, Xianghua Ying, and Hongbin Zha. Vanishing point detection using cascaded 1d hough transform from single images. *Pattern Recognition Letters*, 33(1):1–8, 2012.

[Liu *et al.*,] Xiaobai Liu, Yadong Mu, and Liang Lin. A stochastic image grammar for fine-grained 3d scene reconstruction.

[Liu *et al.*, 2014] Xiaobai Liu, Yibiao Zhao, and Song-Chun Zhu. Single-view 3d scene parsing by attributed grammar. In *CVPR*, pages 684–691, 2014.

[Lucas *et al.*, 1981] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981.

[Mottaghi *et al.*, 2016] Roozbeh Mottaghi, Sanja Fidler, Alan Yuille, Raquel Urtasun, and Devi Parikh. Human-machine crfs for identifying bottlenecks in scene understanding. *TPAMI*, 38(1):74–87, 2016.

[Nayar and Nakagawa, 1990] Shree K Nayar and Yasuo Nakagawa. Shape from focus: An effective approach for rough surfaces. In *ICRA*, pages 218–225. IEEE, 1990.

[Ren and Malik, 2003] Xiaofeng Ren and Jitendra Malik. Learning a classification model for segmentation. In *ICCV*, volume 1, pages 10–17, 2003.

[Toppe *et al.*, 2013] Eno Toppe, Claudia Nieuwenhuis, and Daniel Cremers. Relative volume constraints for single view 3d reconstruction. In *CVPR*, pages 177–184, 2013.

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

[Wang *et al.*, 2007] Shiaokai Wang, William Pentney, Ana-Maria Popescu, Tanzeem Choudhury, and Matthai Philipose. Common sense based joint training of human activity recognizers. In *IJCAI*, volume 7, pages 2237–2242, 2007.

[Wang *et al.*, 2013] Shuo Wang, Jungseock Joo, Yizhou Wang, and Song-Chun Zhu. Weakly supervised learning for attribute localization in outdoor scenes. In *CVPR*, pages 3111–3118, 2013.

[Wang *et al.*, 2015a] Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Holistic 3d scene understanding from a single geo-tagged image. In *CVPR*, pages 3964–3972, 2015.

[Wang *et al.*, 2015b] Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Lost shopping! monocular localization in large indoor spaces. In *ICCV*, pages 2695–2703, 2015.

[Wyatt *et al.*, 2005] Danny Wyatt, Matthai Philipose, and Tanzeem Choudhury. Unsupervised activity recognition using automatically mined common sense. In *AAAI*, volume 5, pages 21–27, 2005.

[Zhang *et al.*, 1999] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *TPAMI*, 21(8):690–706, 1999.

[Zhao and Zhu, 2011] Yibiao Zhao and Song-Chun Zhu. Image parsing with stochastic scene grammar. In *Advances in Neural Information Processing Systems*, pages 73–81, 2011.

[Zhu *et al.*, 2014] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *ECCV*, pages 408–424. Springer, 2014.

[Zhu *et al.*, 2015] Yixin Zhu, Yibiao Zhao, and Song Chun Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. In *CVPR*, pages 2855–2864, 2015.

[Zitnick and Parikh, 2013] C Lawrence Zitnick and Devi Parikh. Bringing semantics into focus using visual abstraction. In *CVPR*, pages 3009–3016, 2013.