Unfolding an Indoor Origami World

David F. Fouhey, Abhinav Gupta, and Martial Hebert

The Robotics Institute, Carnegie Mellon University

Abstract. In this work, we present a method for single-view reasoning about 3D surfaces and their relationships. We propose the use of mid-level constraints for 3D scene understanding in the form of convex and concave edges and introduce a generic framework capable of incorporating these and other constraints. Our method takes a variety of cues and uses them to infer a consistent interpretation of the scene. We demonstrate improvements over the state-of-the art and produce interpretations of the scene that link large planar surfaces.

1 Introduction

Over the last few years, advances in single-image 3D scene understanding have been driven by two threads of research. The first thread asks the basic representation question: What are the right primitives to extract local likelihoods of surface orientation? From geometric context [12] to recent papers on data-driven 3D primitives [6], most approaches in this thread have focused on using large amounts of labeled data to train appearance-based models for orientation likelihoods. While there have been enormous performance gains, these approaches are fundamentally limited by their local nature. The second thread that has pushed the envelope of 3D understanding focuses on reasoning. These approaches stitch together local likelihoods to create a global understanding of the scene. Some include conditional random field (CRF)-based smoothness reasoning [26], cuboidal room layout [10] and volumetric representation of objects [8, 22], 3D objects [35], and groups of objects [1, 40, 41].

Most efforts in reasoning have used either local domain-agnostic constraints or global domain-specific constraints. For instance, CRF-based approaches include the constraint that regions with similar appearance should have similar orientation. These end up, however, enforcing little more than smoothness. This has led to high-level top-down constraints given by domain-specific knowledge. For example, most indoor approaches assume a Manhattan world in which the surface normals lie on three principal directions [3]. Second only to the Manhattan-world constraint is the cuboidal room constraint, in which the camera is assumed to be inside a cube and inference becomes predicting the cube's extent [10]. While this has been enormously influential, the camera-inside-a-box representation leaves the interior and most interesting parts of the scene, for instance furniture, uninterpreted. Recent work has aimed at overcoming this by finding volumetric primitives inside scenes, conventionally cuboids [8, 22, 36, 31], and in simple scenes such as the UIUC dataset of [10], cuboid representations have

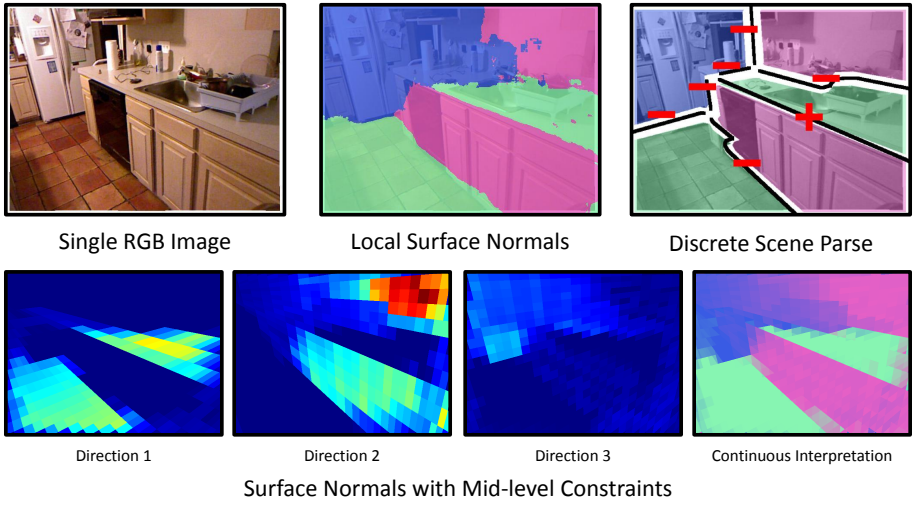


Fig. 1. We propose the use of mid-level constraints from the line-labeling era and a parametrization of indoor layout to "unfold" a 3D interpretation of the scene in the form of large planar surfaces and the edges that join them. In contrast to local per-pixel normals, we return a discrete parse of the scene in terms of surfaces and the edges between them in the style of Kanade's Origami World as well updated continuous evidence integrating these constraints. **Normal legend: blue → X; green → Y; red → Z. Edge Legend: convex +; concave −. Figures best viewed in color.**

increased the robustness of 3D scene understanding. Nonetheless, cuboid-based object reasoning is fundamentally limited by its input, local likelihoods, and it is not clear that it generalizes well to highly cluttered scenes.

In this paper, we propose an alternate idea: while there have been great efforts and progress in both low and high-level reasoning, one missing piece is mid-level constraints. Reasoning is not a one-shot process, and it requires constraints at different levels of granularity. For cluttered and realistic scenes, before we can go to cuboids, we need to a way to piece together local evidence into large planar surfaces and join them with edges. This work aims to address this problem.

These mid-level constraints linking together planes via convex and concave edges have been extensively studied in the past. There is a vast line-labeling literature (e.g., classic works [2, 14, 34]); among these works, we are principally inspired by Kanade's landmark Origami World paper [18], which reasoned directly about surfaces as first-class objects and the edges between them. As systems, line-labeling efforts failed due to weak low-level cues and a lack of probabilistic reasoning techniques; however, they hold a great deal of valuable insight.

Inspired by these pioneering efforts, we introduce mid-level constraints based on convex and concave edges and show how these edges help link multiple surfaces in a scene. Our contributions include: (a) a generic framework and novel parametrization of superpixels that helps to incorporate likelihoods and con-

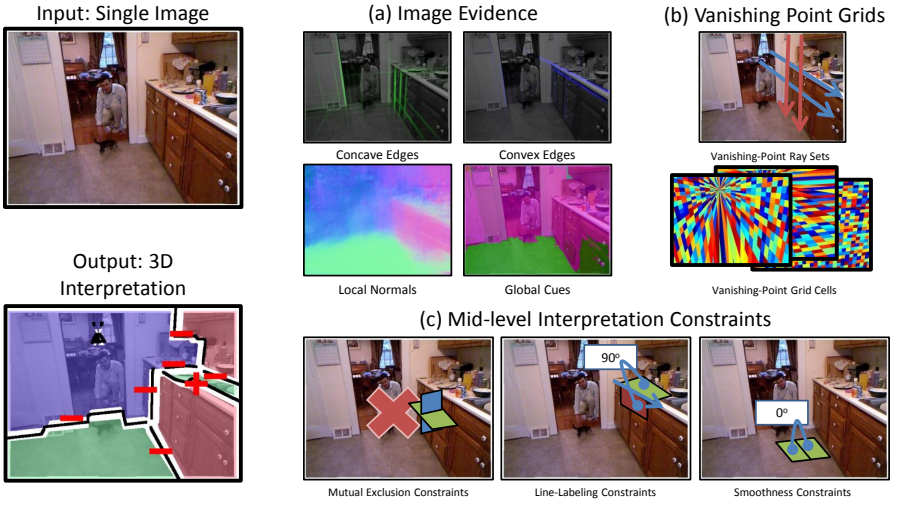


Fig. 2. Overview of the proposed approach. (Left) We take a single image of an indoor scene and produce a 3D interpretation in terms of surfaces and the edges that join them. (Right) We accumulate evidence from inferred surface normal discontinuities in the scene (convex blue, concave green), local surface normals, and room layout. (b) We formulate the problem as assembling a coherent interpretation from a collection of vanishing-point aligned grids. (c) This interpretation must respect observed edges joining segments as well as constraints such as mutual exclusion and smoothness.

straints at all levels of reasoning; (b) the introduction of mid-level constraints for 3D scene understanding as well as methods for finding evidence for them; (c) a richer mid-level interpretation of scenes compared to local normal likelihoods that can act as a stepping-stone for subsequent high-level volumetric reasoning.

2 Related Work

Determining the 3D layout of a scene has been a core computer vision problem since its inception, beginning with Robert's ambitious 1965 "Blocks World" thesis [28]. Early work such as [2, 14] often assumed a simple model in which the perceptual grouping problem was solved and there were no nuisance factors. Thus, the 3D layout problem could be posed as constraint satisfaction over the visible lines. These methods, however, failed to pan out in natural images because the actual image formation process is much more noisy than was assumed.

After many decades without success, general 3D layout inference in relatively unconstrained images began making remarkable progress [5, 11, 12, 29] in the mid-2000s, powered by the availability of training data. This sparked a renaissance during which progress started being made on a variety of long-standing 3D understanding problems. In the indoor world, a great deal of effort went into developing constrained models for the prediction of room layout [10] as well as

features [6, 23, 27] and effective methods for inference [4, 22, 31, 32]. While these high-level constraints have been enormously successful in constrained domains (e.g., less cluttered scenes with visible floors such as the datasets of [10, 38]), they have not been successfully demonstrated on highly cluttered scenes such as the NYU v2 Depth Dataset [33]. Indeed, on these scenes, it turns out to be difficult *even with depth* to find a variety of simple primitives such as cuboids [16, 17, 39], support surfaces [7], or segmentations [9]. We believe that one missing ingredient is effective mid-level constraints, and in this work propose such constraints in the form of line-labels. We emphasize that our goal is to complement prior high-level constraints: we envision a system in which all of these cues cooperate to produce an understanding of the scene at a number of levels of grouping in a hierarchical but feedback manner. This work acts as a stepping stone towards this vision, and introduces a framework for layout estimation that we demonstrate can easily integrate constraints from a variety of levels.

Other work in single-image layout prediction has drawn inspiration from classic line-labeling approaches, but has focused on the task of finding occlusion boundaries [13, 15, 24]. While important, occlusion boundaries only provide a 2.1D sketch (e.g., like [25]) and provide no information in a single image about the surface orientation without their complementary normal discontinuity labels. In this work, we focus on this other class of labels, namely convex and concave edges. These have been applied in the context of stereo [37], but have largely been ignored in the single-image layout community, apart from work on shape recovery using hand-marked folds [19].

3 Overview

In this work, our goal is: given a single image, group pixels into planes, and infer the orientations of these planes and the convex/concave nature of edges. Similar to previous indoor scene understanding approaches, we assume a Manhattan world, which restricts the orientation of planes to three principal directions. Generally, this constraint is implicitly encoded by first grouping pixels into regions via appearance and then solving the surface normal problem as a 3-way classification problem. Our key idea is to reformulate the problem and solve the grouping and classification problem jointly by using top-down superpixels. Given the estimated vanishing points, we determine three possible grids of superpixels aligned with these vanishing points and the problem of classification becomes finding the "active" grid cell at every pixel.

Inferring the active grid cells using image evidence is a severely underconstrained problem, like most single image 3D tasks. Therefore, we include a variety of constraints based on (a) mutual exclusion; (b) appearance and smoothness; (c) convex/concave edge likelihoods; and (d) global room layout. Some are enforced as a unary, while others, such as (c), are binary in nature. Our objective is therefore a quadratic with mutual exclusion constraints. Additionally, the superpixel variables must be integer if one wants a single interpretation of each pixel (1 corresponding to active, 0 to non-active). The resulting problem is thus NP-

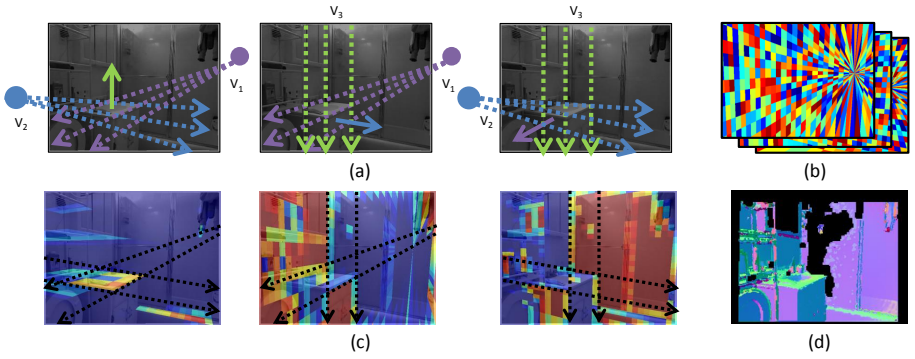


Fig. 3. Parametrization of the method. (a) We sweep rays (dotted lines) from each vanishing point, defining a pencil of lines. The intersection of two pencils of lines defines a superpixel with normal (solid line) perpendicular to the normals of the generating vanishing points. (b) We represent layout by activations of these superpixel grids. (c) We show the likelihoods on each grid cell for the ground truth surface normals (d).

hard in general. We propose to optimize both the integral and relaxed problems: integral solutions are themselves rich inferences about the scene, but we believe the relaxed solutions can act as inputs to higher-level reasoning processes.

We formalize our parametrization of the problem in Section 4 and discuss how we combine all the available evidence into one model in Section 5. Finally, we introduce an approach to finding surface normal discontinuities in Section 6.

4 Parametrization

The first step in our approach is estimating the vanishing points and creating grid cells in 3 principal directions. These act as superpixels defined by geometry rather than appearance. These grids are generated by sweeping rays from pairs of vanishing points, as shown in Fig. 3. The orientation of cells in the grids is defined by the normal orthogonal to the two generating vanishing points. Thus, a cell not only defines a grouping but also an orientation. Therefore any interpretation of the scene in terms of Manhattan-world surface normals that respects this grid can be represented as a binary vector $\mathbf{x}$ encoding which grid cells are active. To illustrate this, we show the likelihoods for the ground truth over grid cells in Fig. 3 (c). This formulation generalizes many previous parametrization of the 3D layout problem, for instance the parametrization proposed in [10, 31]. As we demonstrate with our potentials, our parametrization enables the easy arbitration between beliefs about layout encoded at every pixel such as [6, 20] and beliefs encoded parametrically, such as room layouts or cuboids [10, 22, 31]. Note that our grids overlap, but only one grid cell can be active at each pixel location; we enforce this with a mutual exclusion constraint.

5 Unfolding an interpretation of the scene

We now present how we combine our image evidence to find an interpretation of the scene. We first explain how we obtain surface normal likelihoods and use them as unaries to provide evidence for grid cells in Sec. 5.1. We then explain how we can enforce pairwise constraints on these grid cells given edge evidence in Sec. 5.2. Finally, we introduce a binary quadratic program that arbitrates between these cues and constraints to produce a final interpretation of the scene in Sec. 5.3.

5.1 Unary Potentials

The first kind of cue for finding whether grid cell i is active ($x_i = 1$) or not ($x_i = 0$) is local evidence at the grid cell location. In this work, we use two complementary cues based on techniques for inferring evidence of surface normals and transform them into potentials that capture how much we should prefer x_i to take the value 1. Recall that every grid cell represents not only a grouping but also an orientation, and therefore one can easily convert between likelihoods for orientation at a location and likelihoods of each grid cell being activated.

Local evidence: A wide variety of approaches have been proposed for estimating local surface normal likelihoods from image evidence. We adopt the top-performing approach for doing this, Data-driven 3D Primitives [6] (3DP), which builds a bank of detectors that are associated with local surface configurations. At test time, the method convolves the bank with the image at multiple scales and transfers the associated local surface configuration to the test image wherever each detector has a high response. We soft-assign each pixel to each grid, producing a probability map of each orientation over the image. The local evidence potential of a grid cell i $\phi_{\text{local}}(i)$ is the probability of its orientation averaged over its support.

Global room-fitting evidence: Global room fitting constraints have been important and successful in the single image 3D understanding community. These seek to model the room as a vanishing-point-aligned box. We run the room-fitting method of Hedau et al. [10], which produces a ranked collection of 3D cuboid room hypotheses, where each wall corresponds to one of our grids' directions. At every pixel, we build a histogram of each direction, weighted by the inverse of the rank of the room and suppressing pixels predicted as clutter. The room-fitting evidence potential of a grid cell i $\phi_{\text{room}}(i)$ is the frequency of its orientation averaged over its support.

5.2 Binary Potentials

The second kind of cue for whether a cell i is active or not comes from considering it in conjunction with its neighbors. We use binary potentials characterizing preferences for pairs of grid cells. These operate within the same grid (i.e., on cells of the same orientation) and across grids (i.e., on cells with different orientations). These allow us to probabilistically integrate mid-level constraints via

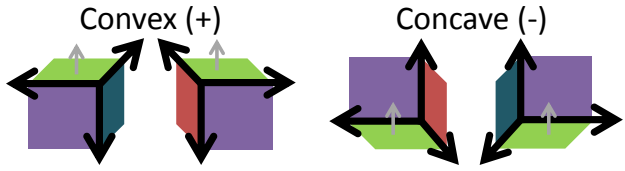


Fig. 4. A subset of valid arrangements in the image plane of surfaces (colors) and convex and concave edges (black) in our scene interpretation method with a gray surface normal arrow for disambiguation.

convex and concave edges. In this section, we describe our potential for achieving this as well as a standard appearance smoothness potential.

Line-labeling: The presence of a convex or concave edge tells us not only that a discontinuity may exist, but also what sorts of labels can occur on either side. For instance, in Manhattan-world scenes, a convex edge at the top of a counter tells us there is a horizontal surface above and a vertical surface below. Because this edge constrains the labels of adjoining surfaces, it is more powerful than a simple smoothness term, which would only permit a labeling change at the edge.

We therefore include a potential that combines inferred convex and concave edges with a dictionary of surface configurations to reward interpretations of the scene in terms of grid cell activations that match our assumptions and available image evidence. We present a basic method for obtaining evidence of convexity and concavity in Section 6, but our potential is agnostic to the source of evidence.

We build an explicit enumeration of arrangements in the image plane that satisfy observing a scene formed by continuous Manhattan-world aligned polyhedra, e.g., a concave edge joining two vertical surfaces with the rightwards facing surface on the left. One half of the preferred arrangements is displayed in Fig. 4; the other half is an analogous y-inverted set. Some scenes may not satisfy our assumptions about the world and our image evidence may be wrong, and we therefore do not make hard decisions as in past line-labeling work [2, 14, 18], but instead form a potential encouraging interpretations that agree with our beliefs. Specifically, given two grid cells with different orientations, we can determine what edge we expect to see in our dictionary, and reward the mutual activation of the two grid cells if we see that edge. We use the potential $\psi_{\text{line}}(i,j) = \exp(-\beta_{line} e_{i,j}^2)$ where $e_{i,j}$ is the inferred probability of that edge from image evidence (i.e., mean image evidence over the edge joining two superpixels). We compute this potential over adjacent pairs of grid cells (i.e., sharing a vanishing point ray) but with different orientations. We compute this separately for convex and concave edges, letting the learning procedure decide their weight.

Smoothness: Adjacent and similar looking parts of the scene should generally have similar labels. As is common in the segmentation literature, we use a Potts-like model: we compute color histograms over LAB space (10 bins per dimension) for grid cells i and j, yielding histograms h_i and h_j; the potential is $\psi_{\text{smooth}}(i,j) = \exp(-d(h_i,h_j)^2)$, where d is the χ^2 distance. We compute the

potential over adjacent grid cells with the same orientation, rewarding similarly colored regions for having similar orientation.

5.3 Inference

We need to resolve possibly conflicting potentials and infer the best interpretation of the scene given the available evidence. Mathematically, we formulate this as an optimization over a vector $\mathbf{x} \in \{0,1\}^n$, where each x_i represents whether grid cell i is active and where $\mathbf{x}$ contains the grid cells from all grids.

Our unary potentials $\{\mathbf{u}_i\}$ and binary potentials $\{\mathbf{B}_j\}$ are collated as a vector $\mathbf{c} = \sum_k \lambda_k \mathbf{u}_k$ and matrix $\mathbf{H} = \sum_l \alpha_l \mathbf{B}_l$ respectively, where c_i and $H_{i,j}$ respectively represent the costs of turning grid cell i on and the cost of turning both grid cell i and j on. Since two active overlapping cells imply that their pixels have two interpretations, we add a mutual-exclusion constraint. This is enforced on cells i and j that are on different grids and have sufficient overlap ($|\cap|/|\cup| \geq 0.2$ in all experiments). This can be formulated as a linear constraint $x_i + x_j \leq 1$. Finally, since our output is in the image plane and our cells are not all the same size, we weight the unary potentials by their support size and binaries by the minimum size of the cells involved.

Our final optimization is a binary quadratic program,

$$\arg\max_{\mathbf{x}\in\{0,1\}^n} \mathbf{c}^T\mathbf{x} + \mathbf{x}^T\mathbf{H}\mathbf{x} \quad \text{s.t.} \quad \mathbf{C}\mathbf{x} \leq \mathbf{1}, \tag{1}$$

where $\mathbf{C}$ stacks the mutual-exclusion linear constraints. Inference of this class of problems is NP-hard; we obtain a solution with the Gurobi solver, which first solves a continuous relaxation of Eqn. 1 and then performs a branch-and-bound search to produce an integral solution. The relaxed solution also acts as an updated belief about the scene, and may serve as a cue for the next layer of scene reasoning. We learn trade-off parameters $\{\lambda_k\}$, $\{\alpha_l\}$ by grid-search for each of the five potentials on a reduced set of images in the training set.

6 Finding Convex and Concave Edges

Our method needs a source of convex and concave edges in a scene. In this section, we describe a simple method for obtaining them. We produce surface normal discontinuity maps from depth data and adapt the 3D Primitives approach [6] to transfer oriented surface normal discontinuities.

We begin by generating surface normal discontinuity labels for our method. Similar to [9], we sweep a half-disc at 8 orientations at 7 scales over the image to get cross-disc normal angles at every pixel and orientation. These are noisy at each scale, but the stable ones (i.e., low variance over all scales), tend to be high quality. Example normal discontinuity labels are shown in Fig. 5.

Given this data, we use a simple transfer method to infer labels for a new scene with a bank of 3D primitive detectors from [6]. Each detector is associated

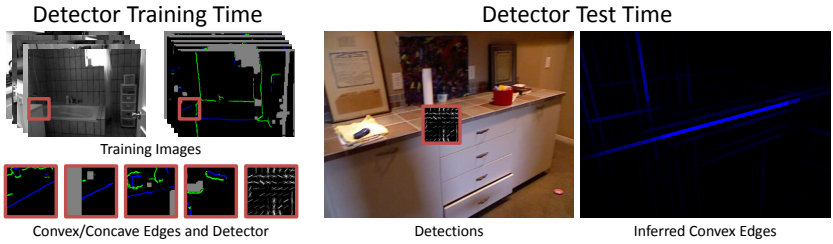


Fig. 5. An illustration of our approach for finding surface normal discontinuities. At training time, we automatically compute surface normal discontinuity maps (convex blue, concave green, missing data gray). At test time, we run a bank of 3D primitive detectors in the image; these detectors are trained to recognize a set of patches. We transfer the convex and concave edge patches associated with each detector.

with a set of bounding boxes corresponding to the locations on which the detector was trained. In the original approach, the surface normals in these boxes were transferred to new images. Instead of transferring surface normals, we transfer the normal discontinuity label, separating by type (convex/concave) and orientation (8 orientations). Thus edge probabilities only accumulate if the detections agree on both type and orientation. At every pixel, the empirical frequency of normal discontinuity labels gives a probability of each edge type at each orientation. This is complementary to the local evidence unary: at the corner of a room, for instance, while a set of detectors may not agree on the specific surface normal configuration, they might agree that there is a concave edge.

7 Experiments

Our output space is a richer interpretation of images compared to per-pixel prediction of surface normals. Evaluating this output in terms of line labelings or linkages of planes is not possible since there are no baseline approaches, ground-truth labels, or established evaluation methodologies. We therefore evaluate one of the outputs of our approach, surface normals, for which there exist approaches and methodologies. We adopt the setup used by the state-of-the-art on this task [6], and evaluate on the challenging and highly cluttered NYU v2 dataset [33].

7.1 Experimental Setup

Training and testing: Past work on single image 3D using NYU v2 [6, 21] has reported results on a variety of train/test splits, complicating inter-method comparisons. To avoid this, we report results on the training and testing splits used by Silberman et al. [33] in their support prediction problem.
Quantitative Metrics: We quantitatively evaluate the surface normal aspect of our approach. However, we strongly believe that the existing per-pixel quantitative metrics for this task are sometimes misleading. For instance, in Fig. 6,

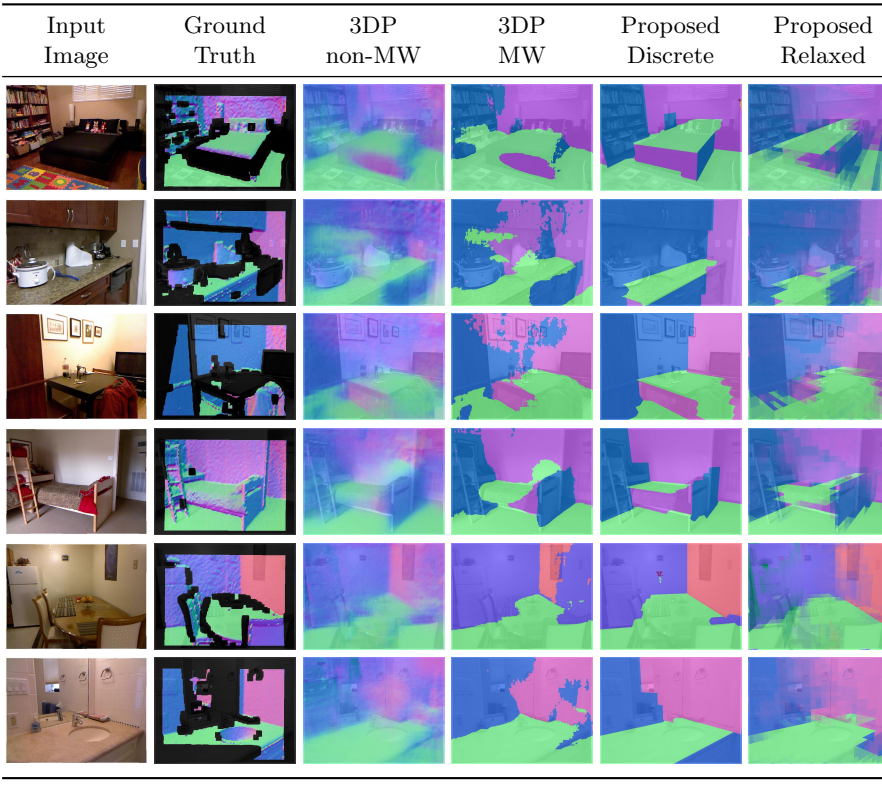


Fig. 6. Selected results on the NYU Dataset comparing our approach to the state-of-the-art, 3D Primitives. To help visualize alignment, we blend the predicted normals with the image.



Fig. 7. Surface connection graphs of scenes automatically obtained from a single image. Our method infers a parse of the scene in terms of large vanishing-point-aligned regions and the edges between them (convex: +, concave: −).

Table 1. Results on NYU v2 for per-pixel surface normal prediction. Our approach improves over Manhattan-world methods in every evaluation metric.

	Summary Stats. (Lower Better)			% Good Pixels (Higher Better)		
	Mean	Median	RMSE	11.25°	22.5°	30°
	Manhattan-world Techniques					
Proposed	**35.1**	**19.2**	**48.7**	**37.6**	**53.3**	**58.9**
Fouhey et al. [6]	36.0	20.5	49.4	35.9	52.0	57.8
Hedau et al. [10]	40.0	23.5	44.1	34.2	49.3	54.4
Lee et al. [23]	43.3	36.3	54.6	25.1	40.4	46.1
	non-Manhattan-world Techniques					
Fouhey et al. [6]	**34.2**	**30.0**	**41.4**	**18.6**	**38.6**	**49.9**
Karsch et al. [20]	40.7	37.8	46.9	8.1	25.9	38.2
Hoiem et al. [12]	36.0	33.4	41.7	11.4	31.3	44.5
Saxena et al. [30]	48.0	43.1	57.0	10.7	27.0	36.3

row 1, our method does worse than [6] on mean and median error, even though it conveys the cuboidal nature of the bed more precisely and segments it into three faces. However, in the absence of other metrics, we still evaluate performance on the metrics introduced in [6]: summary statistics (mean, median, root mean square error) and percent-good-pixels metrics (the fraction of pixels with error less than a threshold t). Note that each metric characterizes a different aspect of performance, not all of which are equally desirable.

Baselines: Our primary point of comparison is 3DP [6], which is the state-of-the-art and outperforms a diverse set of approaches. In particular, the informative comparison to make is with the Manhattan-world version of 3DP: Manhattan-world methods generally produce results that are nearly correct (correct vanishing point) or off by 90° (incorrect one), which is implicitly rewarded by some metrics (% Good Pixels) and penalized by others (mean, RMSE). This makes comparisons with methods not making the assumption difficult to interpret. Nonetheless, to give context, we also report results for the baseline approaches of [6], including but separately presenting non-Manhattan-world ones.

Implementation details: *Vanishing points:* We use the vanishing point detector introduced in [10]. *Grid cells:* The grids used in this work are formed by 32 and 64 rays from exterior and interior vanishing points. *Implausible Grid Cells:* Some grid cells near vanishing points represent implausible surfaces (e.g., an enormous plane at just the right angle); we softly suppress these. *Holes in gridding:* the grid cells will not all line up, leaving a small fraction of the scene uninterpreted. We fill these with nearest-neighbor inpainting. *Training our potentials:* Most potentials are learned from data, but we must use their test-time behavior on the train set to learn our potential trade-offs. For the room-fitting potential, we use 2× cross-validated output; for 3DP-based potentials, we suppress detections on the few images on which the detector was trained.

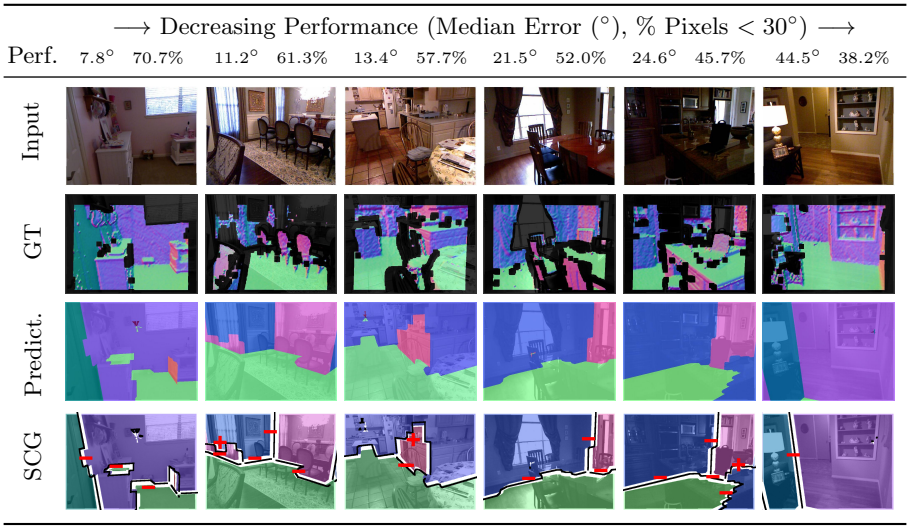


Fig. 8. Results automatically sampled across the method's performance range. Performance reported as median error and % Pixels < 30°. Results were sorted by mean rank over all criteria and six results were automatically picked to evenly divide the list.

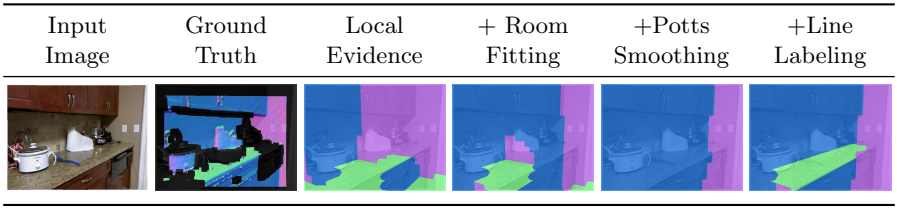


Fig. 9. Qualitative analysis of the method components. Right-to-left: We start with local, then global unaries, followed by smoothness then line-labeling binaries.

7.2 Results

Predicting Surface Normals: We show selected qualitative results in Fig. 6 that illustrate the contributions of our method, as well as an automatically selected performance range in Fig. 8. Consider Fig. 6: our method can often accurately capture object boundaries, especially in comparison with the normal-snapping approach described in [6], which produces noticeable spotting or bending artifacts. Our top-down parametrization mitigates this issue by constraining the space of interpretations, resulting in more plausible explanations. Our mid-level constraints help with the recovery of hard-to-see surfaces, such as surfaces on top of counters or beds (rows 1, 4) or sides of cabinets (row 6). These small surfaces are frequently smoothed away by the Potts model, but are recovered when our line labeling potentials are used.

We report quantitative results in Table 1. Our method outperforms the state-of-the-art for Manhattan-world prediction in every metric. This is important

Table 2. Component-wise analysis on NYU v2 [33]: we report results with parts of the full system removed to analyze the contributions of each method to overall performance.

	Mean	Median	RMSE	11.25°	22.5°	30°
Unaries Only	36.0	19.9	49.6	36.8	52.4	58.0
Smoothness Only	35.4	19.7	48.9	37.4	53.0	58.5
Full Method	**35.1**	**19.2**	**48.7**	**37.6**	**53.3**	**58.9**

since each metric captures a different aspect of performance. It also does better than the non-Manhattan-world methods in all metrics except the ones that heavily penalize Manhattan-world techniques, mean and RMSE; nonetheless, even on mean error, it is second place overall. Although our system outperforms the state-of-the-art, we stress that per-pixel metrics must be considered carefully.

Qualitative scene parses: Our approach produces surface connection graphs in the style of Kanade's Origami World [18]. We decode plane relationships in an interpretation via our edge dictionary illustrated in Fig. 4: given two adjoining surfaces and their orientations, we decode their relation according to our scene formation assumptions (contiguous Manhattan-world polyhedra). We then automatically render qualitative parses of the scene as shown in Fig. 7. As this work does not handle occlusion, failures in decoding relationships occur at configurations that are impossible without occlusion (e.g., vertical-atop-other-vertical).

Ablative analysis: We now describe experiments done to assess the contributions of each component. We show an example in Fig. 9 that characterizes qualitatively how each part tends to change the solution: an initial shape is captured by the local evidence potential and is improved by the room fitting potential. Smoothness potentials remove noise, but also remove small surfaces like counters. The line-labeling potentials, however, can enable the better recovery of the counter. We show quantitative results in Table 2: each step contributes; our line-labeling potentials reduce the median error the most.

Confidence of predictions: Accurately predicting scene layout in every single pixel of every single image is, for now, not possible. Therefore, a crucial question is: can a method identify when it is correct? This is important, for instance, as a cue for subsequent reasoning or for human-in-the-loop systems.

We compute performance-vs-coverage curves across the dataset by sweeping a threshold over per-pixel confidence for our approach and [6] in Fig. 10. Our method's confidence at a pixel is the normalized maximum value of overlapping superpixel variables in the relaxed solution. Our method ranks its predictions well: when going from 100% to 25% coverage, the median error drops by 7.3° and % Pixels $< 11.25°$ increases by 11.7%. Thus the framework is capable of identifying which of its predictions are most likely to be correct. Additionally,

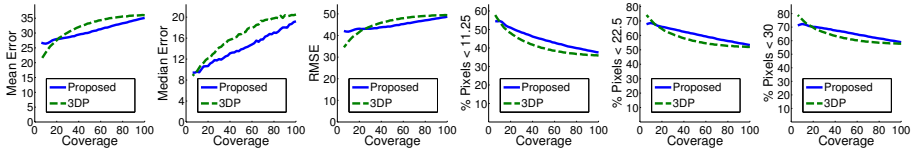


Fig. 10. Performance vs. coverage curves for the system and the next-best method, 3DP. Each plot quantifies accuracy in a metric against fraction of pixels predicted.

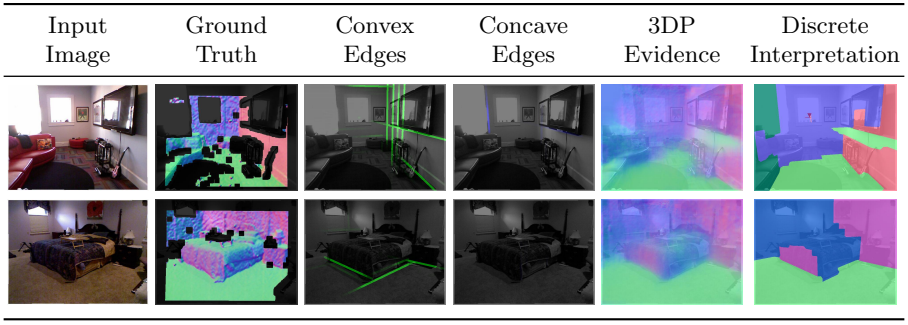


Fig. 11. Failure modes and limitations: (Top) Local evidence can be misleading. An edge is seen below the TV, and our method "folds" its interpretation accordingly. (Bottom) We model only unary and binary relationships between cells; higher order reasoning may allow the recognition of the top of the bed from its sides.

the method out-performs [6] on all metrics, averaging along the curve and at all operating points except the ultra-sparse regime.

Failure modes and limitations: We report some failure modes and limitations in Fig. 11. Our primary failure mode is noisy evidence from inputs. These tend to correspond to mistaken but confident interpretations (e.g., the fold preferred by our model in the first row). Sometimes layouts inferred by our system violate high-level constraints: in the second row, for instance, our interpretation is unlikely globally although it makes sense locally. By reasoning about the proposed pieces, we can reject it without the surface on top necessary to make it plausible. This is consistent with our vision of the 3D inference process: this paper has argued that mid-level constraints are valuable, not that they are the end of the scene interpretation story. Rather than solve all problems at once, we must pass updated evidence to subsequent reasoning.

Acknowledgments: This work was supported by an NDSEG Fellowship to DF, NSF IIS-1320083, ONR MURI N000141010934, and a gift from Bosch Research & Technology Center.

References

1. Choi, W., Chao, Y.W., Pantofaru, C., Savarese, S.: Understanding indoor scenes using 3D geometric phrases. In: CVPR (2013)
2. Clowes, M.: On seeing things. Artificial Intelligence 2, 79–116 (1971)
3. Coughlan, J., Yuille, A.: The Manhattan world assumption: Regularities in scene statistics which enable Bayesian inference. In: NIPS (2000)
4. Del Pero, L., Bowdish, J., Fried, D., Kermgard, B., Hartley, E.L., Barnard, K.: Bayesian geometric modeling of indoor scenes. In: CVPR (2012)
5. Delage, E., Lee, H., Ng, A.Y.: A dynamic Bayesian network model for autonomous 3D reconstruction from a single indoor image. In: CVPR (2006)
6. Fouhey, D.F., Gupta, A., Hebert, M.: Data-driven 3D primitives for single image understanding. In: ICCV (2013)
7. Guo, R., Hoiem, D.: Support surface prediction in indoor scenes. In: ICCV (2013)
8. Gupta, A., Efros, A., Hebert, M.: Blocks world revisited: Image understanding using qualitative geometry and mechanics. In: ECCV (2010)
9. Gupta, S., Arbelaez, P., Malik, J.: Perceptual organization and recognition of indoor scenes from RGB-D images. In: CVPR (2013)
10. Hedau, V., Hoiem, D., Forsyth, D.: Recovering the spatial layout of cluttered rooms. In: ICCV (2009)
11. Hoiem, D., Efros, A., Hebert, M.: Automatic photo pop-up. In: SIGGRAPH (2005)
12. Hoiem, D., Efros, A., Hebert, M.: Geometric context from a single image. In: ICCV (2005)
13. Hoiem, D., Efros, A.A., Hebert, M.: Recovering occlusion boundaries from an image. IJCV 91(3), 328–346 (2011)
14. Huffman, D.: Impossible objects as nonsense sentences. Machine Intelligence 8, 475–492 (1971)
15. Jia, Z., Gallagher, A., Chang, Y.J., Chen, T.: A learning based framework for depth ordering. In: CVPR (2012)
16. Jia, Z., Gallagher, A., Saxena, A., Chen, T.: 3D-based reasoning with blocks, support, and stability. In: CVPR (2013)
17. Jiang, H., Xiao, J.: A linear approach to matching cuboids in RGBD images. In: CVPR (2013)
18. Kanade, T.: A theory of origami world. Artificial Intelligence 13(3) (1980)
19. Karsch, K., Liao, Z., Rock, J., Barron, J.T., Hoiem, D.: Boundary cues for 3D object shape recovery. In: CVPR (2013)
20. Karsch, K., Liu, C., Kang, S.B.: Depth extraction from video using non-parametric sampling. In: ECCV (2012)
21. Ladický, L., Shi, J., Pollefeys, M.: Pulling things out of perspective. In: CVPR (2014)
22. Lee, D.C., Gupta, A., Hebert, M., Kanade, T.: Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In: NIPS (2010)
23. Lee, D.C., Hebert, M., Kanade, T.: Geometric reasoning for single image structure recovery. In: CVPR (2009)
24. Liu, M., Salzmann, M., He, X.: Discrete-continuous depth estimation from a single image. In: CVPR (2014)
25. Nitzberg, M., Mumford, D.: The 2.1D sketch. In: ICCV (1990)
26. Ramalingam, S., Kohli, P., Alahari, K., Torr, P.: Exact inference in multi-label CRFs with higher order cliques. In: CVPR (2008)

27. Ramalingam, S., Pillai, J., Jain, A., Taguchi, Y.: Manhattan junction catalogue for spatial reasoning of indoor scenes. In: CVPR (2013)
28. Roberts, L.: Machine perception of 3D solids. In: PhD Thesis (1965)
29. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: NIPS (2005)
30. Saxena, A., Sun, M., Ng, A.Y.: Make3D: Learning 3D scene structure from a single still image. TPAMI 30(5), 824–840 (2008)
31. Schwing, A.G., Fidler, S., Pollefeys, M., Urtasun, R.: Box In the Box: Joint 3D Layout and Object Reasoning from Single Images. In: ICCV (2013)
32. Schwing, A.G., Urtasun, R.: Efficient Exact Inference for 3D Indoor Scene Understanding. In: ECCV (2012)
33. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: ECCV (2012)
34. Sugihara, K.: Machine Interpretation of Line Drawings. MIT Press (1986)
35. Xiang, Y., Savarese, S.: Estimating the aspect layout of object categories. In: CVPR (2012)
36. Xiao, J., Russell, B., Torralba, A.: Localizing 3D cuboids in single-view images. In: NIPS (2012)
37. Yamaguchi, K., Hazan, T., McAllester, D., Urtasun, R.: Continuous markov random fields for robust stereo estimation. In: ECCV (2012)
38. Yu, S.X., Zhang, H., Malik, J.: Inferring spatial layout from a single image via depth-ordered grouping. Workshop on Perceptual Organization (2008)
39. Zhang, J., Chen, K., Schwing, A.G., Urtasun, R.: Estimaing the 3D Layout of Indoor Scenes and its Clutter from Depth Sensors. In: ICCV (2013)
40. Zhao, Y., Zhu, S.: Image parsing via stochastic scene grammar. In: NIPS (2011)
41. Zhao, Y., Zhu, S.: Scene parsing by integrating function, geometry and appearance models. In: CVPR (2013)