Patches, Planes and Probabilities: A Non-local Prior for Volumetric 3D Reconstruction

Ali Osman Ulusoy Michael J. Black Andreas Geiger
Max Planck Institute for Intelligent Systems, Tübingen, Germany
{osman.ulusoy,michael.black,andreas.geiger}@tue.mpg.de

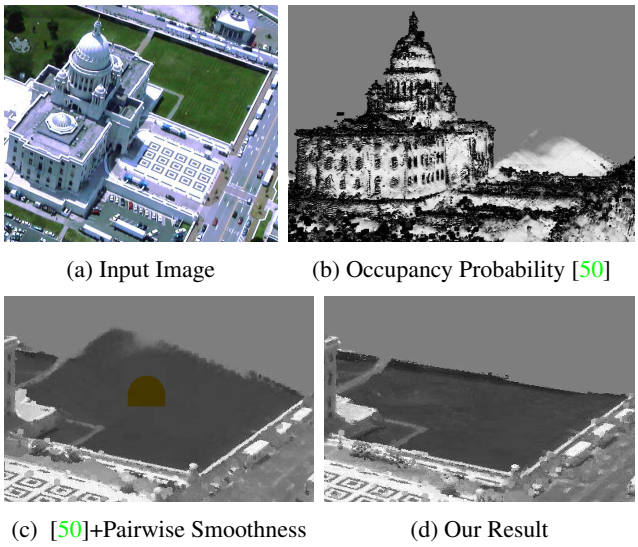
Abstract

In this paper, we propose a non-local structured prior for volumetric multi-view 3D reconstruction. Towards this goal, we present a novel Markov random field model based on ray potentials in which assumptions about large 3D surface patches such as planarity or Manhattan world constraints can be efficiently encoded as probabilistic priors. We further derive an inference algorithm that reasons jointly about voxels, pixels and image segments, and estimates marginal distributions of appearance, occupancy, depth, normals and planarity. Key to tractable inference is a novel hybrid representation that spans both voxel and pixel space and that integrates non-local information from 2D image segmentations in a principled way. We compare our non-local prior to commonly employed local smoothness assumptions and a variety of state-of-the-art volumetric reconstruction baselines on challenging outdoor scenes with textureless and reflective surfaces. Our experiments indicate that regularizing over larger distances has the potential to resolve ambiguities where local regularizers fail.

1. Introduction

Dense 3D reconstruction from multiple RGB images is a long-standing problem in computer vision with numerous practical applications. Unfortunately, it is also a highly ill-posed problem. Ambiguities arise in textureless areas or when photo-consistency assumptions are violated, e.g., at reflecting surfaces. For instance, consider the grass region in Fig. 1a. The surface contains little texture, thus multiple reconstructions satisfy the input images equally well.

Most previous work on multi-view stereo does not address such ambiguities and outputs a 3D model with no uncertainty information. In contrast, probabilistic approaches model and expose the uncertainty in the reconstruction [1, 5, 7, 8, 37, 50, 55]. Fig. 1b shows the result of a recent probabilistic method [50] that is able to expose the ambiguity caused by the textureless region. Ulusoy et al. [50] formulate 3D reconstruction as inference in a Markov random field defined over the 3D voxel grid. Image evidence (input pixels) is modeled using ray potentials that accurately incorporate visibility and free-space constraints.



(a) Input Image (b) Occupancy Probability [50]
(c) [50]+Pairwise Smoothness (d) Our Result

Figure 1: **Motivation:** (a) The grass surface contains little texture, leading to reconstruction ambiguity. (b) Voxel occupancy probabilities reveal this ambiguity where lighter colors encode higher uncertainty [50]. (c) Pairwise smoothness priors cannot resolve the ambiguity and lead to a biased 3D model. (d) Our planarity prior regularizes over large distances and helps reconstruct the correct surface.

While their method exposes reconstruction ambiguity, it is not able to resolve this ambiguity to recover the correct surface because their model does not incorporate any prior information; it models only image evidence. Luckily, the 3D world we live in is not completely random but exhibits geometric structure. Previous works impose smoothness constraints via pairwise potentials that encourage adjacent voxels to take on the same occupancy state [18, 31] or condition surface orientations on semantic information [40].

While these priors reduces surface noise to some extent, they impose regularization only locally and are therefore not sufficient to resolve large ambiguous regions as shown in Fig. 1c.

In this paper, we propose a novel prior formulation for volumetric 3D reconstruction that encourages piece-wise planarity. We are inspired by the planar nature of many elements in man-made environments, i.e., 3D range images of generic scenes can be approximated by piecewise smooth regions with discontinuities at object boundaries [23]. Fig. 1d shows that our prior is able to disambiguate large textureless regions to recover the correct surface.

Implementing a non-local prior in 3D is challenging. Even in 2D, high-order spatial priors are expensive to represent and optimize [27, 33, 39]. Representing the planarity prior directly in voxel space is complicated by the large variety of planes each single subvolume may contain, resulting in numerous high-order cliques in the MRF.

Inspired by the success of non-local segmentation [25, 29] and stereo matching [19, 32, 53] techniques, we encourage planarity within coherent image segments in all viewpoints. Our MRF reasons jointly about the occupancy and intensity of each voxel, the depth values observed at each pixel, and the planarity and plane parameters in each image segment. This hybrid 2D/3D representation with auxiliary variables allows the inference algorithm to propagate view-based planarity assumptions into 3D voxel space in a principled way and implicitly defines smoothness constraints over very large neighborhoods in 3D.

The proposed MRF model is flexible in how plane priors can be integrated. In this work, we investigate a Manhattan world prior that encourages planes to align with the three dominant orthogonal directions. Existing works on planar multi-view stereo or Manhattan world representations treat these as hard constraints [13, 14, 32]. In contrast, we take a probabilistic approach where deviations from the model are allowed as necessary (e.g., the sphere-shaped dome in Fig. 1). Besides specifying the model, we develop a message-passing algorithm for inferring approximate marginal distributions at every voxel, pixel and segment. Our experiments demonstrate that the proposed method improves upon state-of-the-art volumetric reconstruction techniques, in particular for challenging outdoor scenes with large ambiguous (e.g., textureless) areas. Our code and supplementary material are available at http://ps.is.tue.mpg.de/research_projects/volumetric-reconstruction.

2. Related Work

We first review the most relevant work on probabilistic volumetric reconstruction and then discuss approaches that exploit primitives for scene modeling. For a more complete review, we refer the reader to [15, 44].

Volumetric Reconstruction: Following early work [1, 7, 28, 43], Pollard and Mundy [37] propose a volumetric reconstruction method that updates the occupancy and color of each voxel sequentially for each image. GPU implementations of this framework show impressive results [9, 49]. However, their framework lacks a global probabilistic formulation leading to evidence overcounting [38, 50]. To address this, a number of recent approaches have phrased 3D volumetric reconstruction as MRF inference, exploiting the special characteristics of high-order ray potentials to accurately model the image formation process [18, 31, 40, 50].

Reconstruction with Primitives: Several methods exploit planar patches to represent piece-wise planar [3, 6, 10, 12, 34, 41, 45, 46, 56] or Manhattan world [13, 14, 42, 47] scenes. While the approaches produce impressive results, they enforce planarity as a hard constraint and thus only apply to piece-wise planar scenes. In contrast, our spatial prior can be viewed as a soft constraint on planarity because it allows deviations from planarity where it does not hold.

In a similar spirit, Häne et al. [22] propose a model for piecewise planar depth map fusion. Their method takes as input depth maps and a dictionary of patches and integrate these patches as soft constraints in a total variation framework that leads to improved results wrt. classical TV priors. In contrast to the proposed hybrid pixel/voxel approach, their method is restricted to a 2.5D image representation and handles only very small patches ($3-5$ pixels) while we regularize over much larger regions (up to 10k pixels).

Gallup et al. [17] sample planes from initial depth maps and exploit a semantic classifier to classify the image into planar and non-planar regions. Inference is then performed via graph cuts, segmenting the image into regions explained by planes and non-planar regions. Similar to [22], their method uses a 2.5D image representation and requires depth maps as input, while our approach integrates all constraints into a single joint volumetric reconstruction and directly takes RGB images as input.

Lafarge et al. [30] propose a method that simultaneously optimizes 3D primitives and a mesh using an objective that combines photo-consistency terms, mesh smoothness and priors on pairwise primitive arrangements. While their method demonstrate impressive results, it is limited by the topology and shape of the mesh initialization. Additionally, their method outputs a deterministic 3D model while our approach yields a probabilistic 3D interpretation.

More recently, semantic and shape information has been leveraged as prior knowledge for stereo matching [19] and multi-view reconstruction [17, 20, 21, 35, 57], e.g., by constraining the set of plausible geometries [2, 11, 19] or by modeling class specific normal distributions [20, 21]. While our focus in this work is on planarity and Manhattan world priors, semantic information can be easily integrated into our framework.

3. Probabilistic Model

This section introduces our hybrid model for probabilistic volumetric 3D reconstruction with segment-based priors. As input we assume a set of images and camera poses which we obtain using structure-from-motion [51, 52]. As our work extends [50], we use their notation whenever possible. To make this paper self-contained, we briefly repeat the image formation process described in [50] in Section 3.2. We then specify our model in Section 3.3. Details about our inference algorithm will be given in Section 4.

3.1. Notation

The 3D space is decomposed into a grid of voxels. Each voxel is assigned a unique index from the index set $\mathcal{X}$. We associate each voxel $i \in \mathcal{X}$ with two random variables: a binary occupancy variable $o_i \in \{0, 1\}$ which signals if the voxel is occupied ($o_i = 1$) or free ($o_i = 0$), and an appearance variable $a_i \in \mathbb{R}$ describing the voxel intensity (or more generally, color).

Let $\mathcal{R}$ denote the set of viewing rays of all cameras. Note that we model one viewing ray per pixel, thus $\mathcal{R}$ also corresponds to the total set of pixels. For a single ray $r \in \mathcal{R}$, let $\mathbf{o}_r = \{o_1^r, \dots, o_{N_r}^r\}$ and $\mathbf{a}_r = \{a_1^r, \dots, a_{N_r}^r\}$ denote the ordered sets of occupancy and appearance variables associated with voxels intersecting ray r. The ordering is defined by the distance to the respective camera. We further associate each pixel/ray $r \in \mathcal{R}$ with an auxiliary depth variable d_r, discretized according to the depth of each voxel along the ray r.

Each input image is segmented using the superpixelization algorithm of [54], yielding a set $\mathcal{S}$ that comprises all segments from all input images. We associate each segment $s \in \mathcal{S}$ with two random variables: a binary planarity variable $p_s \in \{0, 1\}$ indicating whether the segment is planar ($p_s = 1$) or not ($p_s = 0$), and a variable $\mathbf{n}_s \in \mathbb{R}^3$ specifying 3D plane parameters (i.e., $\mathbf{x}^\intercal \mathbf{n}_s = 1$ if $\mathbf{x} \in \mathbb{R}^3$ on plane $\mathbf{n}_s$) for this segment. We abbreviate the total set of occupancy and appearance variables in the voxel grid with $\mathbf{o} = \{o_i | i \in \mathcal{X}\}$ and $\mathbf{a} = \{a_i | i \in \mathcal{X}\}$. We further summarize all depth, planarity and plane normal variables by $\mathbf{d} = \{d_r | r \in \mathcal{R}\}$, $\mathbf{p} = \{p_s | s \in \mathcal{S}\}$ and $\mathbf{n} = \{\mathbf{n}_s | s \in \mathcal{S}\}$, respectively.

3.2. Image Formation

An image is formed by assigning each pixel the appearance of the first occupied voxel along ray r [50]:

$$I_r = \sum_{i=1}^{N_r} o_i^r \prod_{j<i} (1 - o_j^r)\, a_i^r + \epsilon \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma)$ is a noise term, I_r denotes the intensity (or color) at the pixel corresponding to ray r. Note that the term $o_i^r \prod_{j<i} (1 - o_j^r)$ evaluates to 1 for the first occupied voxel along the ray and to 0 for all other voxels.

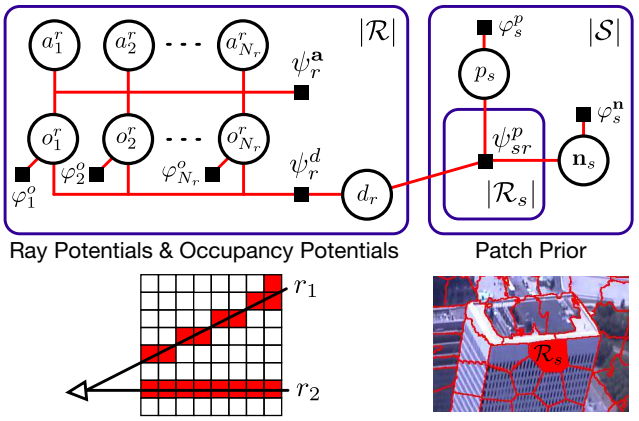


Figure 2: **Factor Graph.** Our graphical model in plate notation. $\mathcal{R}$ comprises the set of all pixels/rays in all images and $\mathcal{S}$ comprises all segments. $\mathcal{R}_s$ is the set of pixels inside segment s. Ray depth variables d_r connect voxels with segments in our hybrid 3D/2D representation. Note that ψ_{sr}^p connects only to those d_r's for wich $r \in \mathcal{R}_s$.

3.3. Markov Random Field

We formulate volumetric 3D reconstruction as inference in a Markov random field. We specify the joint distribution over $\mathbf{o}$, $\mathbf{a}$, $\mathbf{d}$, $\mathbf{p}$ and $\mathbf{n}$ as

$$\begin{aligned} p(\mathbf{o}, \mathbf{a}, \mathbf{d}, \mathbf{p}, \mathbf{n}) = \\ \frac{1}{Z} \prod_{i \in \mathcal{X}} \varphi_i^o(o_i) \prod_{r \in \mathcal{R}} \psi_r^{\mathbf{a}}(\mathbf{o}_r, \mathbf{a}_r)\, \psi_r^d(\mathbf{o}_r, d_r) \\ \times \prod_{s \in \mathcal{S}} \varphi_s^p(p_s)\, \varphi_s^{\mathbf{n}}(\mathbf{n}_s) \prod_{r \in \mathcal{R}_s} \psi_{sr}^d(d_r, p_s, \mathbf{n}_s) \end{aligned} \tag{2}$$

where Z denotes the partition function, $\mathcal{R}_s$ contains all rays/pixels associated with segment s, and φ and ψ denote unary and high-order potentials, respectively. The corresponding factor graph is illustrated in Fig. 2 (top).

Voxel Occupancy Prior: We model our prior belief about the state of the occupancy variables using a Bernoulli distribution

$$\varphi_i^o(o_i) = \gamma^{o_i} (1 - \gamma)^{1 - o_i} \tag{3}$$

where γ is the prior probability that voxel i is occupied.

Appearance Ray Potential: Our ray potentials model the image generation process as specified by Eq. 1, i.e., they encourage the appearance of the first occupied voxel along ray r to agree with the image observation I_r at pixel r:

$$\psi_r^{\mathbf{a}}(\mathbf{o}_r, \mathbf{a}_r) = \sum_{i=1}^{N_r} o_i^r \prod_{j<i} (1 - o_j^r)\, \nu_r(a_i^r) \tag{4}$$

Here, $\nu_r(a)$ denotes the probability of observing intensity a at ray r which we model as $\nu_r(a) = \mathcal{N}(a|I_r, \sigma)$.

Depth Ray Potential: The depth ray potential models the constraint that the depth d_r at pixel r must equal the depth of the first occupied voxel along the ray,

$$\psi_r^d(\mathbf{o}_r, d_r) = \begin{cases} 1 & \text{if } d_r = \sum_{i=1}^{N_r} o_i^r \prod_{j<i}(1-o_j^r)\, d_{ri} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where d_{ri} denotes the depth of voxel i along ray r. Note that the potential $\psi_r^d(\mathbf{o}_r, d_r)$ is 1 if and only if d_r is equal to the depth of the first occupied voxel along the ray. Otherwise, the potential evaluates to 0, indicating an invalid state.

Planarity Prior: We now describe our planarity prior, which favors piece-wise planar depth maps in each of the input views. As illustrated in Fig. 2 (right), the prior comprises the three components below.

Planarity Potential: We encourage segment planarity via

$$\varphi_s^p(p_s) = \exp(\lambda_s |\mathcal{R}_s|\, p_s) \tag{6}$$

where $p_s \in \{0, 1\}$ is the planarity indicator variable of segment s. Note that each segment comprises many pixels. Thus, we weight this planarity potential by the segment area $|\mathcal{R}_s|$. The weight λ_s determines the strength of the prior, i.e., how much planarity should be enforced.

Plane Normal Potential: Our prior belief about the orientation of planar patches is modeled as a mixture distribution on the plane normal $\mathbf{n}_s$

$$\varphi_s^{\mathbf{n}}(\mathbf{n}_s) = \sum_{k=1}^{K} w_k\, \mathcal{M}\left(\frac{\mathbf{n}_s}{\|\mathbf{n}_s\|} \,\middle|\, \boldsymbol{\mu}_k, \kappa_k\right) \tag{7}$$

where $\mathcal{M}(\cdot \mid \boldsymbol{\mu}, \kappa)$ denotes the von Mises-Fisher distribution with parameters $\boldsymbol{\mu}$ (mean direction) and κ (concentration parameter). While any kind of surface orientation information can be incorporated into this potential (e.g., semantic information), in Section 6, we will investigate a rather generic Manhattan world prior ($K = 3, \kappa > 0$).

Plane Depth Potential: To ensure that the plane parameters of a planar patch agree with the depth variables of the corresponding image pixels, we define

$$\psi_{sr}^p(d_r, p_s, \mathbf{n}_s) = \begin{cases} \exp\left(-\lambda_p\, \eta(d_r - D_r(\mathbf{n}_s))\right) & \text{if } p_s = 1 \\ 1 & \text{otherw.} \end{cases} \tag{8}$$

where λ_p is a consistency weight, $\eta(\cdot)$ denotes a penalty function and $D_r(\mathbf{n}_s)$ returns the depth of plane $\mathbf{n}_s$ along ray r. The intuition behind this potential is simple: if $p_s = 0$, then all possible depth values d_r are equally likely for each $r \in \mathcal{R}_s$. In contrast, when $p_s = 1$, we favor depth values d_r that are close to the plane $\mathbf{n}_s$. To account for segmentation errors, which can lead to outliers within a segment, we model this factor using a robust Lorentzian penalty $\eta(\cdot)$.

4. Inference

We are interested in estimating the marginal distributions of $\mathbf{o}$, $\mathbf{a}$, $\mathbf{d}$, $\mathbf{p}$ and $\mathbf{n}$ in the proposed MRF. Given these marginals, we can easily calculate several quantities of interest, e.g., the probability of voxel occupancy and intensity, the probability of a segment s being planar and the expected plane normal, as well as the probability distribution of depth along each ray. Importantly, these depth distributions enable the computation of depth maps that are optimal in terms of Bayes decision theory [50].

Unfortunately, inference in our graphical model is challenging due to the large number of variables ($\mathbf{o}$, $\mathbf{a}$, $\mathbf{d}$), the high-order potentials for modeling visibility constraints ($\psi_r^{\mathbf{a}}$, ψ_r^d) and the mixed discrete ($\mathbf{o}, \mathbf{d}, \mathbf{p}$) and continuous ($\mathbf{a}, \mathbf{n}$) state spaces of the variables. Furthermore, our factor graph in Fig. 2 contains a large number of loops due to intersecting viewing rays $\mathcal{R}$. Thus, exact inference is intractable. In this section, we show how an approximation to the desired marginals can be obtained using message passing. In particular, we derive an algorithm based on sum-product particle belief propagation [24] in factor graphs [26].

4.1. Message Passing

Let $\mu_{f\to x}$ denote the message sent from factor f to variable x, and let $\mu_{x\to f}$ denote the corresponding variable-to-factor message. The messages from the unary factors to the variables, as well as the variable-to-factor messages are straightforward and we omit them here to save space. In the following, we present the message equations for the appearance ray potential $\psi_r^{\mathbf{a}}$, the depth ray potential ψ_r^d and the plane depth potential ψ_{sr}^p. The supplementary document contains detailed derivations of all the equations. Below, we assume that all incoming messages to a factor are normalized such that they sum/integrate to 1.

Plane Depth Messages: The continuous message from the plane depth potential ψ_{sr}^p to the plane parameters $\mathbf{n}_s$ is given by

$$\begin{aligned} \mu_{\psi_{sr}^p \to \mathbf{n}_s}(\mathbf{n}_s) &= \sum_{p_s} \sum_{d_r} \psi_{sr}^p(d_r, p_s, \mathbf{n}_s)\mu(p_s)\mu(d_r) \qquad (9) \\ &= \mu(p_s = 1) \sum_{d_r} \psi_{sr}^p(d_r, p_s = 1, \mathbf{n}_s)\mu(d_r) + \mu(p_s = 0) \end{aligned}$$

where we have abbreviated the incoming messages using $\mu(p_s) = \mu_{p_s \to \psi_r^d}(p_s)$ and $\mu(d_r) = \mu_{d_r \to \psi_r^d}(d_r)$. Note that the message $\mu_{\psi_{sr}^p \to \mathbf{n}_s}(\mathbf{n}_s)$ becomes uniform if there is strong evidence of non-planarity, i.e., $\mu(p_s = 0) = 1$, from the other pixels in the segment, e.g., in case of a highly

curved surface. Otherwise, i.e. $\mu(p_s = 1) = 1$, the message evaluates high for planes $\mathbf{n}_s$ that agree with the depth distribution $\mu(d_r)$.

The message to the binary planarity variable p_s reads as

$$\mu_{\psi^p_{sr}\to p_s}(p_s = 1) = \int_{\mathbf{n}_s} \sum_{d_r} \psi^p_{sr}(d_r, p_s = 1, \mathbf{n}_s)\mu(\mathbf{n}_s)\mu(d_r)$$
$$\mu_{\psi^p_{sr}\to p_s}(p_s = 0) = 1 \tag{10}$$

with $\mu(\mathbf{n}_s) = \mu_{p_s\to\psi^d_r}(\mathbf{n}_s)$ and $\mu(d_r) = \mu_{d_r\to\psi^d_r}(d_r)$. The planarity message is high if the depths of likely planes (where $\mu(\mathbf{n}_s)$ is high) coincide with likely depths (high value of $\mu(d_r)$). Otherwise, the depth distribution at the pixel cannot be explained with the incoming plane distribution, and therefore, the planarity message is low.

Finally, the message to the depth variable d_r is given by

$$\mu_{\psi^p_{sr}\to d_r}(d_r) = \sum_{p_s} \int_{\mathbf{n}_s} \psi^p_{sr}(d_r, p_s, \mathbf{n}_s)\mu(p_s)\mu(\mathbf{n}_s) \tag{11}$$
$$= \mu(p_s = 1) \int_{\mathbf{n}_s} \psi^p_{sr}(d_r, p_s = 1, \mathbf{n}_s)\mu(\mathbf{n}_s) + \mu(p_s = 0)$$

If there is strong evidence of non-planarity from the other pixels in the segment, i.e. $\mu(p_s = 0) = 1$, then the message to the depth variable becomes uniform, i.e. the depth variables are not affected. Otherwise, the message is high for values d_r that match the depth of likely planes, i.e., $d_r \approx D_r(\mathbf{n}_s)$ where $\mu(\mathbf{n}_s)$ is high.

Appearance Ray Messages: In a naïve application of belief propagation, computing the factor-to-variable messages requires exponential time in the number of variables involved in the potential. Since each viewing ray intersects hundreds of voxels, the appearance ray potentials typically involve hundreds of variables, making this computation intractable. However, the special structure of the ray potentials allows for reducing the computation time from exponential to linear in the number of variables [31,50]. Exploiting this property, Ulusoy et al. [50] derived the sum-product message equations for this factor, which we include below for completeness:

$$\mu_{\psi^{\mathbf{a}}_r\to o^r_i}(o^r_i = 1) = \sum_{j<i} \mu(o^r_j = 1) \prod_{k<j} \mu(o^r_k = 0)\,\rho_{rj} + \prod_{k<i} \mu(o^r_k = 0)\,\rho_{ri} \tag{12}$$

$$\mu_{\psi^{\mathbf{a}}_r\to o^r_i}(o^r_i = 0) = \sum_{j<i} \mu(o^r_j = 1) \prod_{k<j} \mu(o^r_k = 0)\,\rho_{rj} + \sum_{j>i} \mu(o^r_j = 1) \prod_{\substack{k<j \\ k\neq i}} \mu(o^r_k = 0)\,\rho_{rj} \tag{13}$$

The incoming occupancy messages are abbreviated by $\mu(o^r_i) = \mu_{o^r_i\to\psi^{\mathbf{a}}_r}(o^r_i)$ and ρ_{rj} is a photo-consistency measure at the jth voxel along ray r, see [50] for details. This message has an intuitive interpretation: it increases the occupancy probability of voxels that are photo-consistent and visible. The probability of voxels between the camera and the likely surface location are decreased. For occluded voxels the message is uniform.

Ulusoy et al. also showed that the messages to the continuous appearance variables, i.e. $\mu_{\psi^{\mathbf{a}}_r\to a_i}$, can be computed analytically and that they can be compactly represented as a constant plus a weighted Gaussian distribution. The variable-to-factor messages $\mu_{a_i\to\psi^{\mathbf{a}}_r}$ cannot be computed analytically. We follow [50] and approximate them using Mixture-of-Gaussians (MoG) distributions. We refer the reader to [50] (Section 4.1) for all necessary details.

Depth Ray Messages: The message from ψ^d_r to d_r is readily given by

$$\mu_{\psi^d_r\to d_r}(d_r = d_{ri}) \propto \mu(o^r_i) \prod_{j<i} \mu(o^r_j = 0) \tag{14}$$

where the incoming occupancy messages are again abbreviated by $\mu(o^r_i) = \mu_{o^r_i\to\psi^d_r}(o^r_i)$.

By following a similar argument to the derivation of the appearance ray messages (see supplementary document), the message to the occupancy variable o^r_i is given by

$$\mu_{\psi^d_r\to o^r_i}(o^r_i = 1) = \sum_{j<i} \mu(o^r_j = 1) \prod_{k<j} \mu(o^r_k = 0)\,\mu(d_{rj}) + \prod_{k<i} \mu(o^r_k = 0)\,\mu(d_{ri}) \tag{15}$$

$$\mu_{\psi^d_r\to o^r_i}(o^r_i = 0) = \sum_{j>i} \mu(o^r_j = 1) \prod_{\substack{k<j \\ k\neq i}} \mu(o^r_k = 0)\,\mu(d_{rj}) + \sum_{j<i} \mu(o^r_j = 1) \prod_{k<j} \mu(o^r_k = 0)\,\mu(d_{rj}) \tag{16}$$

where the incoming depth messages are abbreviated as $\mu(d_{ri}) = \mu_{d_r\to\psi^d_r}(d_r = d_{ri})$. Note that the messages Eq. 15+16 are very similar to those of the appearance ray factor in Eq. 12+13. The difference is that the photo-consistency measure ρ in Eq. 12+13 is replaced with the incoming depth message $\mu(d)$ which carries information from our planarity prior. The messages in Eq. 15+16 intuitively increase the occupancy probability of voxels at likely depths, i.e., where $\mu(d_{rj})$ is high. The probability of voxels between the camera and the likely depth are decreased whereas the message to the occluded voxels are uniform.

4.2. Particle Belief Propagation

Unfortunately, the continuous plane parameter variables $\mathbf{n}$ complicate the message computations. In particular, Eq. 9 is of continuous form and therefore difficult to represent. Further, the integrals that arise in Eq. 10+11 cannot be calculated in closed form.

To tackle these challenges, we exploit particle belief propagation [24], which has been adopted by many works with great success [4, 19, 36, 53]. The main idea is to discretize the continuous space with a finite set of particles and use this discretization to approximate the integral equations with a Monte Carlo estimate.

For each segment, we draw K particles, $\{\mathbf{n}_s^{(k)}\}_{k=1}^K$, from a proposal distribution $W_s(\mathbf{n})$. Using these particles, we approximate the integral in Eq. 10 via importance sampling

$$\mu_{\psi_{sr}^p \to p_s}(p_s = 1) \approx \frac{1}{K} \sum_{k=1}^{K} \sum_{d_r} \psi_{sr}^p(d_r, 1, \mathbf{n}_s) \frac{\mu(\mathbf{n}_s^{(k)})}{W_s(\mathbf{n}_s^{(k)})} \mu(d_r) \tag{17}$$

where $W_s(\cdot)$ denotes an appropriate proposal distribution. The approximation of the depth variable message in Eq. 11 is similar and can be found in the supplementary material.

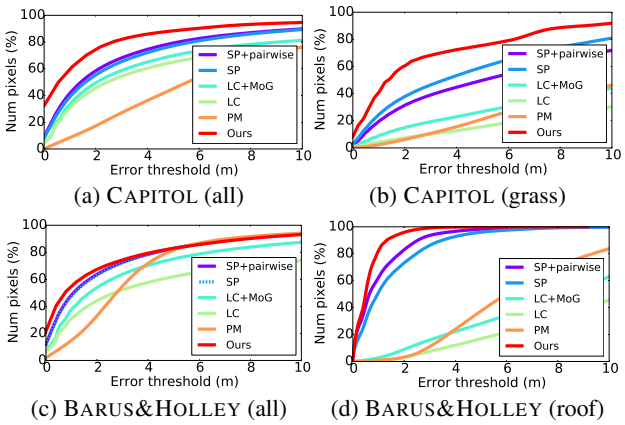
Note that the quality of these approximations depend on how well the set of particles explores the continuous space. For instance, consider a perfectly planar patch. If none of the particles come close to the correct plane, the planarity message in Eq. 10 will be underestimated, hence incorrectly lowering the planarity belief of the segment.

To avoid this situation, in this work, we take advantage of a data-driven strategy [48] that generates samples from an initial 3D reconstruction of the scene. More specifically, we first run inference for our model without the planarity patch prior and compute the most likely depth at each pixel. We then repeatedly draw three pixels from the segment, prioritizing pixels with low depth certainty (negative entropy) and fit a plane to them. Promising planes, i.e., planes that explain most of the depth values inside the segment, are added to the particle set. In addition to this particle set, we also generate particles by conditioning on the plane prior implied by the normal potentials. We draw plane normals from the prior in Eq. 7 and then for each sample, optimize the depth of the plane to best match the depth estimates within the segment. Compared to a naïve approach such as uniformly sampling the space of plane parameters, our data-driven strategy avoids unlikely particles. Further details of our particle sampling strategy are presented in the supplementary document.

Given the set of plane particles $\{\mathbf{n}_s^{(k)}\}_{k=1}^K$, we build a kernel density estimate in order to obtain an approximation to the proposal distribution $W_s(\mathbf{n})$

$$W_s(\mathbf{n}) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{N}(\mathbf{n}; \boldsymbol{\mu} = \mathbf{n}_s^{(k)}, \boldsymbol{\Sigma} = \sigma_{\text{kde}}\, \mathbf{I}) \tag{18}$$

where σ_{kde} denotes the kernel bandwidth, which we empirically set to a fixed value for all our experiments.



(a) CAPITOL (all) (b) CAPITOL (grass)

(c) BARUS&HOLLEY (all) (d) BARUS&HOLLEY (roof)

Figure 3: Reconstruction accuracy plots for the CAPITOL and BARUS&HOLLEY datasets.

5. Implementation

This section provides details of our implementation. We initialize all ray messages uniformly and first pass messages without the planarity prior, iterating over each image at least once. This yields an initial 3D model. The beliefs in the occupancy variables are then propagated to the per-pixel depth variables via Eq. 11. The median of these depth distributions yields a depth map which we use to segment the images using a depth-aware superpixelization algorithm [54].

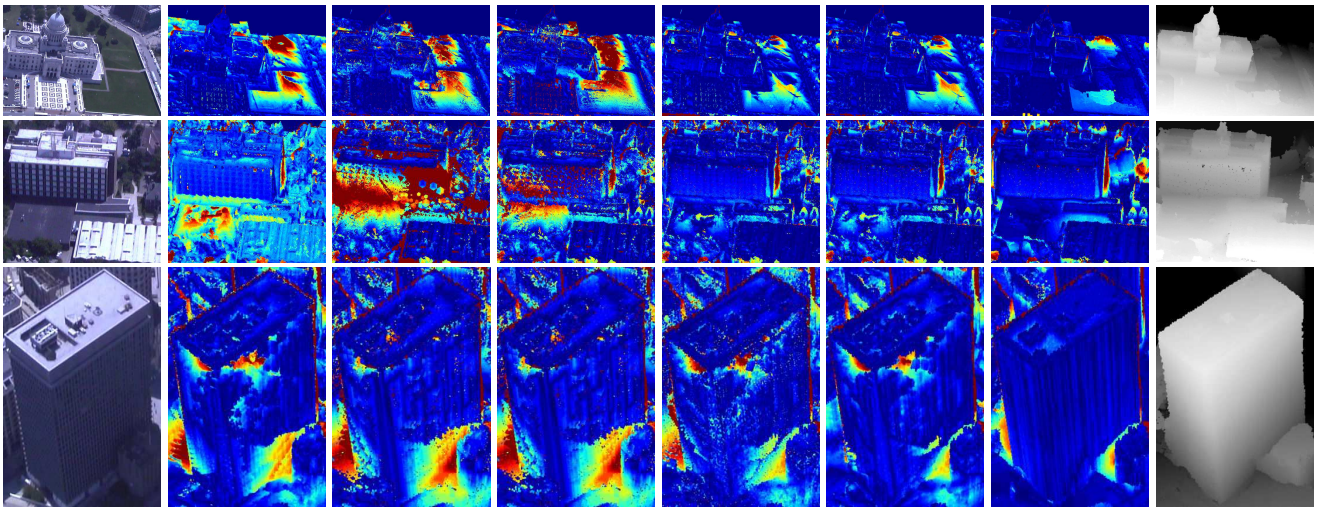
We further use these depth maps to generate plane particles $\{\mathbf{n}_s^{(k)}\}_{k=1}^K$ for each segment as discussed in Section 4.2, using $K = 64$ throughout all our experiments. Next, we compute the messages from the plane depth factor ψ_{sr}^p to the plane parameters (Eq. 9) and to the planarity variables (Eq. 10). Finally, the information aggregated from the plane depth factor and all other pixels in the image segment is passed back to the depth variables using Eq. 11. The depth variables in turn influence the occupancies along each image ray via Eq. 15+16.

We interleave this process with the message updates for the appearance ray factors and iterate over each image until convergence. All necessary algorithmic details can be found in the supplementary document. Our implementation uses GPU amenable octree data structures and GPU parallelization for message computations. Our current implementation takes roughly 20 seconds to process a 1 Megapixel image and 30 million voxels.

6. Experimental Evaluation

We evaluate our algorithm on three aerial datasets with LIDAR ground truth[1] provided by Restrepo et al. [38]. The datasets exhibit several challenges, including large fea-

[1]The original datasets are distributed with sparse LIDAR point clouds. Ulusoy et al. triangulated these point clouds to obtain a dense ground truth mesh [50]. We use their meshes for a fair comparison to the baselines.



(a) Image (b) PM (c) LC (d) LC+MoG (e) SP (f) SP+pairwise (g) Ours (h) Our Depth

Figure 4: **Qualitative Results.** This figure shows qualitative results on the CAPITOL (top), BARUS&HOLLEY (middle) and DOWNTOWN (bottom) dataset. (a) Region of interest. (b-g) Visualization of errors. Cooler colors correspond to lower error. (h) Depth map predicted by our model. We encourage the reader to zoom in for details.

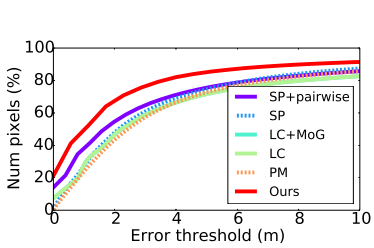
tureless regions, reflective and specular surfaces, severe occlusions and transient objects. The input images are roughly one megapixel in size and each dataset contains $\sim$ 200 images. The datasets are referred to as CAPITOL, BARUS&HOLLEY and DOWNTOWN, and example image crops are presented in Fig. 4 (first column).

Baselines: We compare our results to several state-of-the-art baselines. First, we compare to the sum-product algorithm of Ulusoy et al. [50], whose formulation is equivalent to our model without the patch prior and which we call "SP" in the following. As their method does not encode any spatial regularization, we introduce an additional baseline "SP+pairwise" with pairwise smoothness potentials that encourage adjacent voxels to take the same occupancy label [31] . We optimize the parameters of this potential using the CAPITOL dataset. Further details can be found in the supplementary document. Next, we compare to the approach of Liu and Cooper ("LC") whose max-product formulation utilizes ray potentials [31] as our method, but suffers from a systematic bias in ambiguous regions as shown in [50]. We also include comparisons to an improved version of Liu and Cooper's approach as proposed by [50] with a more robust voxel color model, which we refer to as "LC+MoG". Finally, we include a comparison to Pollard and Mundy's approach ("PM") [37], which lacks a global probabilistic formulation but nevertheless achieves very good results on several datasets [9, 49], including the Middlebury benchmark [44].

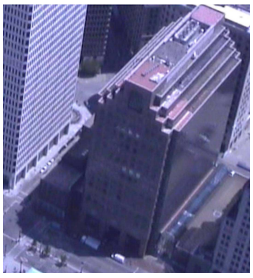
Evaluation: We evaluate reconstruction accuracy by comparing depth maps predicted by each method to ground truth. More specifically, we compute depth maps for all input viewpoints and report the sum of per-pixel absolute errors over all pixels in all views. We create ground truth depth maps by projecting the LIDAR mesh onto the view. For algorithms producing deterministic MAP outputs, i.e., "LC" and "LC+MoG", we consider the first occupied voxel along the ray as a depth prediction. For algorithms which compute occupancy marginals, we compute Bayes optimal depth estimates at each pixel under the ℓ_1 loss as described in [50] and the supplementary document.

We set the parameters of the superpixelization algorithm [54] to product roughly 500 segments. We found this segmentation granularity to yield a reasonable tradoff between over- and undersegmentation of the images. We empirically chose a single set of parameters that we use for all three datasets: $\lambda_s = 5, \lambda_d = 1, \kappa = 20$.

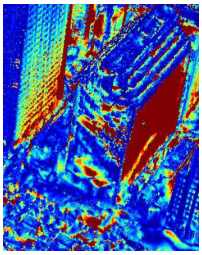
All three datasets contain many surfaces that are orthogonal to each other as can be seen in Fig. 4 (first column). This structure motivates a Manhattan world prior on the plane orientations in Eq. 7, which favors planes oriented along the X, Y or Z directions. However, this prior requires orienting the scene such that the dominant directions in the scene coincide with the X, Y and Z direction. Towards this goal, we first compute the ground plane normal (i.e., the Z direction) as the RANSAC fit to a sparse point cloud generated by running our model without the planarity prior and accumulating the 3D point clouds from each depth map. The only remaining unknown is the rotation around the Z axis, which we compute as the entropy minimizer of the point cloud projections onto canonical orthogonal X and Y planes (see [16] for details).



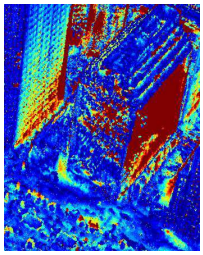
(a) DOWNTOWN (all)



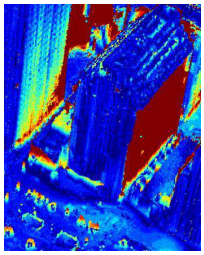
(b) Image



(c) SP



(d) SP+pairwise



(e) Ours

Figure 5: (5a) Accuracy plots for DOWNTOWN, (5b) Region of interest, (5c-5e) Visualization of errors.

Discussion of Results: Fig. 3 and Fig. 5a present cumulative accuracy plots for the three datasets and all methods. Overall our algorithm outperforms previous methods for the CAPITOL and DOWNTOWN datasets and achieves similar performance for BARUS&HOLLEY.

We visualize the error maps for all datasets and methods in Fig. 4. For the CAPITOL and BARUS&HOLLEY datasets (first two rows), the majority of errors are localized on the large featureless regions, e.g., the grass region for CAPITOL and the black rooftop in BARUS&HOLLEY. The results show that "PM" as well as algorithms that estimate a MAP solution, i.e., "LC" and "LC+MoG", yield a systematic bias in textureless regions, leading to large errors. The sum-product ("SP") result is able to reveal the ambiguity in the region (see Fig. 1b) and obtains better results. Nonetheless, "SP" cannot resolve the ambiguity without additional prior information. The model with the pairwise smoothness potentials ("SP+pairwise") yields denser and smoother results, leading to lower errors on most structures, e.g. the building facades in CAPITOL and BARUS&HOLLEY. However, results on the grass region in CAPITOL and the rooftop in BARUS&HOLLEY indicate that the pairwise model is unable to resolve ambiguities on the large textureless regions. In contrast, our algorithm achieves significantly more accurate results in these regions as seen in Fig. 4g. To investigate this more closely, we evaluate errors only on the large featureless surfaces. The results are displayed in Fig. 3b+3d and show the clear improvement of our algorithm. Fig. 1d depicts a rendering of our reconstruction where the ambiguity in the model is resolved successfully.

Besides textureless regions, our planarity prior helps improve reconstruction of reflective surfaces as well. The last row of Fig. 4a displays a building with a reflective surface from DOWNTOWN. All baselines yield large errors on the building surface since it violates the Lambertian surface assumption. Our planarity prior helps correct these errors to achieve a denser and more accurate result. Fig. 5 displays results for another reflective building. All algorithms produce large errors for the side surface of the building which has a mirror like reflectivity. However, the front side of the building contains a mixture of reflective and non-reflective materials. While all algorithms produce some correct depth values in this region, the results are in general very noisy and contain large holes. The model with pairwise smoothness potentials ("SP+pairwise") fails to bridge these large gaps in the reconstruction. In contrast, our algorithm regularizes over the entire facade and is able to reconstruct this facade surface correctly.

The results confirm that even though our prior favors piecewise planar reconstructions, non-planar structures are preserved, e.g. the building dome in Fig. 4 (first row) shows very little error. Our algorithm is robust to non-planar structures due to two facts: First, our model allows for turning off the planarity prior wherever necessary. Second, we also allow for outliers within a segment, owing to our robust penalty function $\eta(\cdot)$ in Eq. 8. The latter property is important to deal with imprecise segmentation boundaries.

For BARUS&HOLLEY, our planarity priors introduces errors around some of the shrubs and trees, which lowers the overall accuracy (Fig. 3c). However, for some of these regions, the LIDAR ground truth is not accurate either due to the tree tops which have been extruded to the ground level. In the supplementary document, we present an additional evaluation excluding such regions, as well as further examples and failure cases.

7. Conclusion

We have presented a novel non-local prior for probabilistic volumetric 3D reconstruction that encourages planarity within image segments and regularizes over large voxel neighborhoods. Our experiments show that the proposed prior is able to resolve reconstruction ambiguities of textureless and partially reflective surfaces and achieves state-of-the-art results in reconstruction accuracy for highly challenging aerial datasets. In our future work, we plan to incorporate semantic information in our model. Furthermore, our prior formulation also allows for geometries beyond planar segments. We believe that integrating more complex primitives such as spheres, cylinders, deformable shapes or even wholistic 3D scene models will be promising extensions of the presented model.

References

[1] M. Agrawal and L. S. Davis. A probabilistic framework for surface reconstruction from multiple images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2001. 1, 2

[2] S. Bao, M. Chandraker, Y. Lin, and S. Savarese. Dense object reconstruction with semantic priors. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[3] A. Bartoli. A random sampling strategy for piecewise planar scene segmentation. *Computer Vision and Image Understanding (CVIU)*, 105(1):42–59, 2007. 2

[4] F. Besse, C. Rother, A. Fitzgibbon, and J. Kautz. PMBP: PatchMatch Belief Propagation for correspondence field estimation. *International Journal of Computer Vision (IJCV)*, 110(1):2–13, 2014. 6

[5] R. Bhotika, D. J. Fleet, and K. N. Kutulakos. A probabilistic theory of occupancy and emptiness. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2002. 1

[6] A. Bodis-Szomoru, H. Riemenschneider, and L. Van Gool. Fast, approximate piecewise-planar modeling based on sparse structure-from-motion and superpixels. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[7] J. D. Bonet and P. Viola. Poxels: Probabilistic voxelized volume reconstruction. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 1999. 1, 2

[8] A. Broadhurst, T. W. Drummond, and R. Cipolla. A probabilistic framework for space carving. *ICCV*, 2001. 1

[9] F. Calakli, A. O. Ulusoy, M. I. Restrepo, G. Taubin, and J. L. Mundy. High resolution surface reconstruction from multi-view aerial imagery. In *3DIMPVT*, 2012. 2, 7

[10] A.-L. Chauve, P. Labatut, and J.-P. Pons. Robust piecewise-planar 3d reconstruction and completion from large-scale unstructured point data. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010. 2

[11] A. Dame, V. Prisacariu, C. Ren, and I. Reid. Dense reconstruction using 3D object shape priors. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[12] F. Fraundorfer, K. Schindler, and H. Bischof. Piecewise planar scene reconstruction from sparse correspondences. *Image and Vision Computing (IVC)*, 24(4):395–406, 2006. 2

[13] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Manhattan-world stereo. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[14] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Reconstructing building interiors from images. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2009. 2

[15] Y. Furukawa and C. Hernandez. Multi-view stereo: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 9(1-2):1–148, 2013. 2

[16] D. Gallup, J.-M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007. 7

[17] D. Gallup, J.-M. Frahm, and M. Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010. 2

[18] P. Gargallo, P. Sturm, and S. Pujades. An occupancy–depth generative model of multi-view images. In *Proc. of the Asian Conf. on Computer Vision (ACCV)*, pages 373–383, 2007. 1, 2

[19] F. Güney and A. Geiger. Displets: Resolving stereo ambiguities using object knowledge. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 6

[20] C. Haene, N. Savinov, and M. Pollefeys. Class specific 3d object shape priors using surface normals. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[21] C. Haene, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3D scene reconstruction and class segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[22] C. Haene, C. Zach, B. Zeisl, and M. Pollefeys. A patch prior for dense 3d reconstruction in man-made environments. In *Proc. of the International Conf. on 3D Digital Imaging, Modeling, Data Processing, Visualization and Transmission (THREEDIMPVT)*, 2012. 2

[23] J. Huang, A. B. Lee, and D. Mumford. Statistics of range images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2000. 2

[24] A. Ihler and D. McAllester. Particle belief propagation. *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009. 4, 6

[25] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision (IJCV)*, 82(3):302–324, 2009. 2

[26] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009. 4

[27] N. Komodakis and N. Paragios. Beyond pairwise energies: Efficient optimization for higher-order MRFs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[28] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *International Journal of Computer Vision (IJCV)*, 38(3):199–218, 2000. 2

[29] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical random fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 36(6):1056–1077, 2014. 2

[30] F. Lafarge, R. Keriven, M. Bredif, and H.-H. Vu. A hybrid multiview stereo algorithm for modeling urban scenes. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 35(1):5–17, 2013. 2

[31] S. Liu and D. Cooper. Statistical inverse ray tracing for image-based 3d modeling. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 36(10):2074–2088, 2014. 1, 2, 5, 7

[32] W. Luo and H. Maitre. Using surface model to correct and fit disparity data in stereo vision. In *Proc. of the International Conf. on Pattern Recognition (ICPR)*, 1990. 2

[33] P. Marquez-Neila, P. Kohli, C. Rother, and L. Baumela. Non-parametric higher-order random fields for image segmentation. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 2

[34] B. Micusik and J. Kosecka. Multi-view superpixel stereo in urban environments. *International Journal of Computer Vision (IJCV)*, 89(1):106–119, 2010. 2

[35] A. Owens, J. Xiao, A. Torralba, and W. T. Freeman. Shape anchors for data-driven multi-view reconstruction. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2013. 2

[36] J. Pacheco, S. Zuffi, M. J. Black, and E. Sudderth. Preserving modes and messages via diverse particle selection. In *Proc. of the International Conf. on Machine learning (ICML)*, 2014. 6

[37] T. Pollard and J. L. Mundy. Change detection in a 3-d world. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007. 1, 2, 7

[38] M. I. Restrepo. *Characterization of Probabilistic Volumetric Models for 3-d Computer Vision*. PhD thesis, Brown University, 2013. 2, 6

[39] S. Roth and M. J. Black. Fields of experts. *International Journal of Computer Vision (IJCV)*, 82(2):205–229, 2009. 2

[40] N. Savinov, L. Ladicky, C. Häne, and M. Pollefeys. Discrete optimization of ray potentials for semantic 3d reconstruction. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2

[41] A. Saxena, M. Sun, and A. Y. Ng. Make3D: learning 3D scene structure from a single still image. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 31:824–840, 2009. 2

[42] M. Schönbein and A. Geiger. Omnidirectional 3d reconstruction in augmented manhattan worlds. In *Proc. IEEE International Conf. on Intelligent Robots and Systems (IROS)*, 2014. 2

[43] S. Seitz and C. Dyer. Photorealistic scene reconstruction by voxel coloring. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1997. 2

[44] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006. 2, 7

[45] S. N. Sinha, D. Steedly, and R. Szeliski. Piecewise planar stereo for image-based rendering. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2009. 2

[46] N. Srinivasan and F. Dellaert. A rao-blackwellized mcmc algorithm for recovering piecewise planar 3d models from multiple view rgbd images. In *Proc. IEEE International Conf. on Image Processing (ICIP)*, 2014. 2

[47] J. Straub, G. Rosman, O. Freifeld, J. J. Leonard, and J. W. Fisher. A mixture of manhattan frames : Beyond the manhattan world. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[48] Z. Tu, S. C. Zhu, and H. Shum. Image segmentation by data driven markov chain monte carlo. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2001. 6

[49] A. O. Ulusoy, O. Biris, and J. L. Mundy. Dynamic probabilistic volumetric models. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2013. 2, 7

[50] A. O. Ulusoy, A. Geiger, and M. J. Black. Towards probabilistic volumetric reconstruction using ray potentials. In *Proc. of the International Conf. on 3D Vision (3DV)*, 2015. 1, 2, 3, 4, 5, 6, 7

[51] C. Wu. Towards linear-time incremental structure from motion. *Proc. of the International Conf. on 3D Vision (3DV)*, 2013. 3

[52] C. Wu, S. Agarwal, B. Curless, and S. Seitz. Multicore bundle adjustment. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011. 3

[53] K. Yamaguchi, T. Hazan, D. McAllester, and R. Urtasun. Continuous markov random fields for robust stereo estimation. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2012. 2, 6

[54] K. Yamaguchi, D. McAllester, and R. Urtasun. Robust monocular epipolar flow estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 3, 6, 7

[55] A. Yao and A. Calway. Dense 3-d structure from image sequences using probabilistic depth carving. In *Proc. of the British Machine Vision Conf. (BMVC)*, 2003. 1

[56] L. Zebedin, J. Bauer, K. Karner, and H. Bischof. Fusion of feature-and area-based information for urban buildings modeling from aerial imagery. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2008. 2

[57] C. Zhou, F. Güney, Y. Wang, and A. Geiger. Exploiting object similarity in 3d reconstruction. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2015. 2