Semantic Multi-view Stereo: Jointly Estimating Objects and Voxels

Ali Osman Ulusoy[1,2] Michael J. Black[1] Andreas Geiger[2,3]
[1]Perceiving Systems Department, MPI for Intelligent Systems Tübingen
[2]Autonomous Vision Group, MPI for Intelligent Systems Tübingen
[3]Computer Vision and Geometry Group, ETH Zürich
{osman.ulusoy,michael.black,andreas.geiger}@tue.mpg.de

Abstract

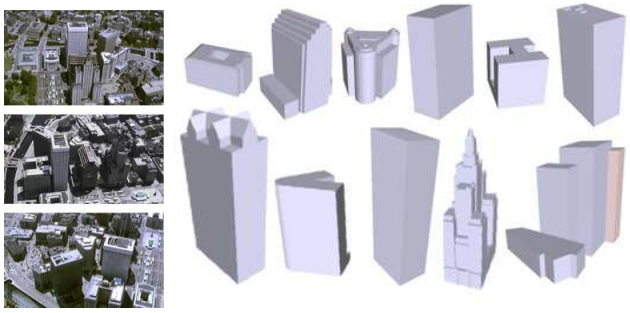
Dense 3D reconstruction from RGB images is a highly ill-posed problem due to occlusions, textureless or reflective surfaces, as well as other challenges. We propose object-level shape priors to address these ambiguities. Towards this goal, we formulate a probabilistic model that integrates multi-view image evidence with 3D shape information from multiple objects. Inference in this model yields a dense 3D reconstruction of the scene as well as the existence and precise 3D pose of the objects in it. Our approach is able to recover fine details not captured in the input shapes while defaulting to the input models in occluded regions where image evidence is weak. Due to its probabilistic nature, the approach is able to cope with the approximate geometry of the 3D models as well as input shapes that are not present in the scene. We evaluate the approach quantitatively on several challenging indoor and outdoor datasets.

1. Introduction

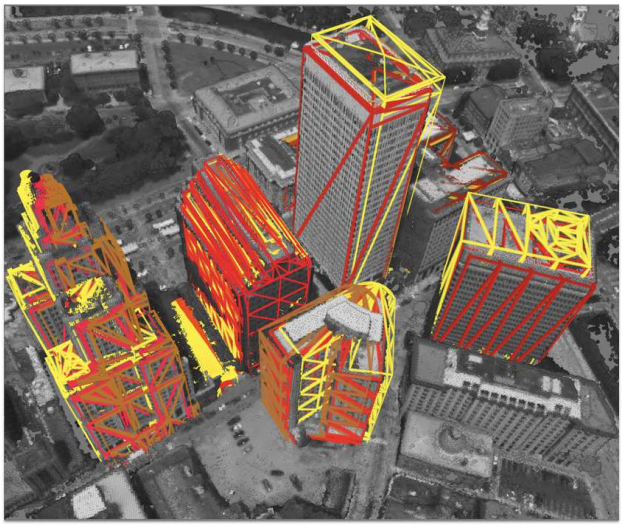
Dense 3D reconstruction from RGB images is a highly ill-posed problem. Occlusions and textureless or reflective surfaces cause fundamental ambiguities in 3D reconstruction [4, 34]. In this work, we address these ambiguities by leveraging semantic information. In particular, we propose object-level shape priors for 3D reconstruction. Our approach takes as input RGB images and a set of plausible 3D shape models, and solves for the existence and pose of each object while reconstructing a dense 3D model of the entire scene. See Fig. 1 for an illustration.

The proposed object-level shape priors yields two key benefits. First, the output of our approach is a dense reconstruction of the entire scene as well as a structural representation of the objects in it. This output yields not only an accurate mapping of the environment but also a semantic understanding in terms of the objects.

Second, the proposed prior allows for powerful regularization that can resolve large ambiguities common in 3D reconstruction. For instance, our shape prior can help reconstruct the back-side of an object even though it is occluded



(a) Images (b) 3D Shape Models



(c) Reconstruction

Figure 1: Given input images (a) and a set of object shape models (b), our approach jointly reconstructs a dense 3D model of the entire scene and solves for the existence and pose of each object model. In (c), we visualize the output of our method as a point cloud sampling of the dense reconstruction and object poses (yellow= unlikely, red=likely).

in the images. Existing works that consider low-level geometric priors such as spatial smoothness [6, 22, 38], piecewise planarity [16, 33] or Manhattan-world constraints [12, 31] cannot complete large occluded regions, especially for objects with complex geometry. Nonetheless, these priors offer complementary regularization that can be combined with our object shape prior to further improve reconstruction accuracy as we demonstrate in our experiments.

Our approach requires finding a set of 3D shape models as input. This retrieval task depends on the available semantic information. For indoor scenes, recent convolutional neural networks (CNN) together with large annotated 3D databases such as ModelNet [37] and ShapeNet [7] produce compelling results for object class detection and rough pose estimation [8, 14, 23, 32]. For outdoor reconstruction, the GPS signal can be used to collect geolocated 3D models from online collections such as 3D Warehouse[1].

Incorporating object shape models as priors for 3D reconstruction is challenging. Retrieved objects might not be present in the scene. Further, existing 3D models are often simplified and thus only coarse approximations of the true object shapes. Besides, even though current shape databases contain thousands of examples, no single object shape in the database might exactly match the observation. Finally, while object detectors or the GPS signal might provide a rough initialization, fine-grained object pose information is often not available.

To address these challenges, we integrate 3D object shapes with image observations in a *probabilistic* fashion. We build upon the probabilistic 3D reconstruction framework of Ulusoy et al. [34]. Their formulation accurately models 3D reconstruction from images using a Markov random field (MRF) with ray-potentials, but does not consider scene priors. In this work, we integrate object shape priors into their framework. Inference in our MRF produces probabilistic estimates of the existence and precise 3D pose of each object, as well as the dense voxel occupancy and color. Given enough image evidence, our algorithm is able to reconstruct geometric details that are not present in the input models. In case of insufficient image information, e.g., in heavily occluded regions, our approach defaults to the input model geometry under the most likely 3D pose. Finally, our approach is robust against geometric inaccuracies of the input models as well as objects that are not present in the scene. We compare our approach with state of the art 3D reconstruction methods using three aerial datasets with LIDAR ground-truth and a realistic synthetic indoor dataset.

2. Related Work

In this section, we first review existing approaches to volumetric 3D reconstruction. We then discuss methods that leverage object shape models for reconstruction.

Volumetric Reconstruction from Images: While there is a large body of literature on volumetric fusion from range images [10, 25], in this paper we focus on reconstruction directly from RGB images. Despite the increasing availability of 3D sensors, the vast majority of cameras in the world lack depth sensing capability. Consequently, image-based reconstruction is more general. Kutulakos and Seitz established the foundations of volumetric reconstruction based on photo-consistency [20]. Early probabilistic extensions of their approach include [1, 5, 27]. Unfortunately, these methods lack a global probabilistic model, which makes it difficult to interpret their probabilistic output. More recent approaches [13, 22, 30, 33, 34] phrase volumetric reconstruction as inference in an MRF where voxels along each pixel's line of sight are connected via high-order ray potentials. This approach makes precise what is optimized and further allows incorporating scene priors in a principled way.

All these approaches, except Ulusoy et al. [34], incorporate priors such as local (pairwise) smoothness [13, 22, 30] or piecewise planarity [33]. In particular, Savinov et al. exploit scene semantics and propose class-specific pairwise priors [30]. While their approach utilizes a *local* prior for all shapes of an *object class* such as building and vegetation, we exploit 3D shapes of *object instances* as a more expressive *non-local* prior.

Object Shape Priors for 3D Reconstruction: Many existing works demonstrate the usefulness of object shape priors for reconstruction. Güney et al. utilize a set of car shapes to improve stereo estimation in urban environments [15]. Salas-Moreno et al. use 3D models of furniture to improve camera tracking accuracy in indoor scenes [29]. In this work, we consider camera poses as input and focus on how object shape priors can benefit dense 3D reconstruction.

For 3D reconstruction, Pauly et al. match a database of object shapes against an incomplete point cloud from a 3D scanner and then align the best fitting shape to reconstruct occluded regions [26]. Bao et al. densify multi-view stereo point clouds by fitting 3D shape models [3]. Dame et al. use a low-dimensional shape space as a prior to improve reconstruction accuracy and completeness [11]. Zhou et al. detect objects with similar shapes in the scene and use these detections to jointly estimate a low-dimensional shape space of these objects, regularizing the reconstruction [39].

The aforementioned works consider a 3D reconstruction as input and regularize this reconstruction using shape priors. Instead, our approach takes as input RGB images and integrates image-based 3D reconstruction with detection and pose estimation of objects in the scene. This joint formulation yields two benefits over previous works. First, our approach combines images and object shapes in a principled probabilistic fashion. This allows reconstructing de-

[1] https://3dwarehouse.sketchup.com/

tails missing in the input shapes, detecting input objects that are not present in the scene and yields robustness to inaccurate shape models. Second, by accurately modeling visibility using ray-potentials, our approach yields improvements not only where the shape prior is available, but notably also in other parts of the scene. We demonstrate examples of this behavior in the experiments section.

3. Probabilistic Model

This section introduces our probabilistic model for image-based 3D reconstruction using object shape priors. As input we assume a set of images and camera poses, which we obtain using structure-from-motion [35, 36]. We further assume a set of approximate shape models of objects, which may or may not be present in the scene. Depending on the scene and available semantic information, a shape model database can be retrieved in a variety of ways. While we do not focus on the retrieval task in this paper, our experiments demonstrate examples from aerial and indoor scenes. Note that we do not assume all input objects to be present in the scene; our inference algorithm automatically estimates which object models are present. While our approach can take into account object pose information if provided, we do not assume this input. Probabilistic object pose estimates are computed as part of the inference.

As our work extends Ulusoy et al.'s probabilistic model for 3D reconstruction [34], we use their notation whenever possible. We introduce the variables of our model in Section 3.1 and specify the model in Section 3.2. Our inference algorithm is presented in Section 4.

3.1. Variables

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 empty ($o_i = 0$), and an appearance variable $a_i \in \mathbb{R}$ describing the voxel intensity.

We associate one viewing ray r for each pixel in the input images. Let $\mathcal{R}$ denote the set of viewing rays of all cameras. 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 sets of occupancy and appearance variables associated with voxels intersecting ray r, ordered by the distance to the respective camera.

An image is formed by assigning each pixel the appearance of the first occupied voxel along the pixel's ray r [34]:

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

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

We now introduce the variables related to the object shape models. Let $\mathcal{S}$ denote the set of input object shapes. We associate each shape model $s \in \mathcal{S}$ with a binary random variable $b_s \in \{0, 1\}$, which denotes whether the model is present in the scene ($b_s = 1$) or not ($b_s = 0$). We represent the pose of each shape model using a continuous variable $\mathbf{p}_s \in \Omega$ which comprises 3D translation, rotation and scaling on a continuous but bounded domain Ω.

We abbreviate the total set of occupancy and appearance variables in the voxel grid with $\mathbf{o} = \{o_i | i \in \mathcal{X}\}$ and $\mathbf{a} = \{a_i | i \in \mathcal{X}\}$ and summarize the set of shape model variables using $\mathbf{b} = \{b_s | s \in \mathcal{S}\}$ and $\mathbf{p} = \{\mathbf{p}_s | s \in \mathcal{S}\}$.

3.2. Markov Random Field

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

$$p(\mathbf{o}, \mathbf{a}, \mathbf{b}, \mathbf{p}) = \frac{1}{Z} \prod_{i \in \mathcal{X}} \varphi_i^o(o_i) \prod_{r \in \mathcal{R}} \psi_r(\mathbf{o}_r, \mathbf{a}_r) \tag{2}$$
$$\times \prod_{s \in \mathcal{S}} \left[\varphi_s^b(b_s)\ \varphi_s^p(\mathbf{p}_s) \prod_{q \in \mathcal{Q}_s(\mathbf{p}_s)} \kappa_q(\mathbf{o}_q, b_s, \mathbf{p}_s) \right]$$

where Z denotes the partition function, φ are unary potentials, and ψ and κ are high-order potentials.

Voxel Occupancy Prior: We model the 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: The ray potentials penalize deviations from the image formation model as specified in Eq. 1. They encourage the appearance of the first occupied voxel along ray r to agree with the image observation I_r at pixel r:

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

where N_r is the number of voxels along ray r. Here, $\nu_r(a)$ denotes the probability of observing intensity a at ray r. We follow [34] and model this term using a Gaussian distribution $\nu_r(a) = \mathcal{N}(a | I_r, \sigma)$.

The preceding two potentials (Eq. 3+4) were introduced in [34] and model 3D reconstruction from images. The following potentials formulate the proposed object shape prior.

Raylet Potential: Transitions between empty and occupied voxels in the volume imply surfaces. If a shape model

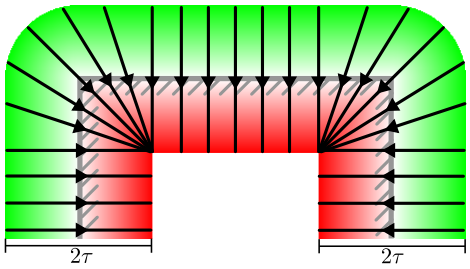


Figure 2: Raylets (black) are located at the surface of each shape model (gray) and oriented oriented with the gradient of the truncated signed distance function (green to red).

is present in the scene, i.e. $b_s = 1$, its surface should agree with voxel transitions in the volume. In particular, voxels directly in front of the model surface should be empty and voxels at the surface should be occupied. We introduce high-order *raylet* potentials to implement this behavior as a soft constraint. First, we define raylets as short ray segments centered at the surface of each model and aligned with the negative gradient of the truncated signed distance function (TSDF) as illustrated in Fig. 2. Each raylet is truncated at distance $\pm\tau$ from the surface. The raylet potential connects all voxels intersecting the raylet and prefers the first occupied voxel along the raylet to coincide with the shape model surface. Since the raylet travels from the outside towards the inside of the surface, it encourages voxels outside the surface to be empty and voxels at the surface to be occupied. The voxels inside the surface are not affected. The finite extent of the raylet ensures that the shape model affects only voxels in its immediate surrounding, hence minimizing interference with other surfaces in the scene.

We denote the set of all raylets defined by shape s in a canonical pose as $q \in \mathcal{Q}_s$. The raylets $\mathcal{Q}_s$ transform according to the model pose $\mathbf{p}_s$, which we denote as $\mathcal{Q}_s(\mathbf{p}_s)$. As with the camera viewing rays, each raylet $q \in \mathcal{Q}$ intersects an ordered set of voxels $\mathbf{o_q} = \{o_1^q, \ldots, o_{N_q}^q\}$.

We formulate the raylet potential as

$$\kappa_q(\mathbf{o}_q, b_s, \mathbf{p}_s) = \begin{cases} \sum_{i=1}^{N_q} o_i^q \prod_{j<i} (1 - o_j^q)\, \eta_i^q(\mathbf{p}_s) & \text{if } b_s = 1 \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

where $\eta_i^q(\mathbf{p}_s)$ is the probability of voxel i explaining the shape model with pose $\mathbf{p}_s$. This probability is measured using the distance between voxel i and the object surface. Denoting the unsigned distance between voxel i along raylet q to the model surface as $d_i^q(\mathbf{p}_s)$, we define $\eta(\cdot)$ as

$$\eta_i^q(\mathbf{p}_s) = \exp\left(\lambda_p \max\left(0, 1 - \frac{d_i^q(\mathbf{p}_s)}{\tau}\right)\right) \tag{6}$$

where $\lambda_p > 0$ is a hyperparameter of our model. For voxels close to the surface η evaluates high, whereas for voxels further away from the surface η is small.

While the raylet potential attains its highest value when the voxel geometry matches the surface prior, i.e. all voxels in front of the surface are empty and the voxel at the surface is occupied, it allows for deviations, which helps cope with inaccuracies in the input models. Finally, if the model is not present in the scene, i.e., $b_s = 0$, the raylet potential does not influence the voxel geometry and it is equal to 1. Since $\eta \geq 1$, the potential favors solutions where surfaces in the reconstruction are explained by plausible shape models.

Object Presence Prior: We model the prior belief about the presence of each shape model in the scene using

$$\varphi_s^b(b_s) = \exp(-\lambda_b \left|\mathcal{Q}_s\right| b_s). \tag{7}$$

where we choose $\lambda_b > 0$ to favor simple explanations of the scene with few object models. Note that we scale the potential by the number of raylets $\left|\mathcal{Q}_s\right|$ to achieve invariance to raylet sampling.

Object Pose Prior: If available, prior knowledge about the object pose can be integrated via the pose prior $\varphi_s^p(\mathbf{p}_s)$. In this work, we make no assumptions about the object pose and therefore use a uniform prior $\varphi_s^p(\mathbf{p}_s) \propto 1$.

4. Inference

In this section, we briefly present our inference algorithm based on belief propagation. Additional details and detailed derivations of the message equations can be found in the supplementary document.

In this work, we are interested in estimating a probabilistic 3D reconstruction rather than the most likely one. Our inference technique estimates the marginal distributions of occupancy and appearance at each voxel, as well as the existence and pose parameters of each shape model in the database. The marginal distributions enable analysis of uncertainty in the 3D reconstruction and are thus useful for subsequent algorithms that utilize the resulting 3D models.

Inference in our MRF is challenging due to the high order ray and raylet potentials (Eq. 4+5), the mixed discrete $(\mathbf{o}, \mathbf{b})$ and continuous $(\mathbf{a}, \mathbf{p})$ state spaces of the variables, and the large number of variables (millions of voxels) and factors (hundreds of millions of ray and raylet potentials). Moreover, our MRF contains a large number of loops due to intersecting viewing rays $\mathcal{R}$ and raylets $\mathcal{Q}_s$, rendering exact inference intractable. We thus present an approximate inference algorithm. Our approach is based on sum-product particle belief propagation [18] and addresses the aforementioned challenges.

While naïve belief propagation on high-order ray potentials is intractable, Ulusoy et al. [34] demonstrate that the algebraic structure of the ray potentials allows the complexity to be reduced from exponential to linear time. The raylet

potentials proposed in this paper possess a similar structure, which we exploit to achieve efficient (linear time) message passing. For details, we refer to the supp. document.

Additionally, the continuous variables $(\mathbf{a}, \mathbf{p})$ in our model complicate belief propagation. In particular, the integrals that arise in the message equations do not admit closed-form solutions. For the pose variables $\mathbf{p}$, we follow a particle based strategy [18] and maintain a sample distribution $\{\mathbf{p}_s^{(1)}, \ldots, \mathbf{p}_s^{(K)}\}$ to approximate the continuous state space of $\mathbf{p}$. This discretization allows Monte Carlo estimates of the integral equations (see supp. document). We discuss our sampling strategy in Section 5.

For the voxel appearance variables $\mathbf{a}$, the messages to the variable can be computed analytically and represented as a constant plus weighted Gaussian distributions. The variable-to-factor messages however cannot be computed analytically. We follow [34] and approximate these messages using Mixture-of-Gaussians (MoG) distributions.

5. Implementation

Due to the highly loopy nature of our MRF, the quality of inference depends on the message passing schedule. Empirically we found the following strategy to perform well. First, we pass messages among the ray potentials, ignoring the raylet potentials, i.e., the shape prior. This corresponds to the method of [34] and yields an initial 3D reconstruction. We then incorporate the raylet potentials into the inference, which regularizes the reconstruction according to the 3D shape models. We interleave message passing for the ray and raylet potentials until convergence. As object surfaces are regularized, the ray potentials exploit the refined free-space and visibility constraints to improve the reconstruction in other parts of the scene as well. We show examples of this behavior in Section 6.

Particle Sampling: In the following, we describe our approach to sampling the pose parameter particles. Ideally, we would like to draw K particles $\{\mathbf{p}_s^{(1)}, \ldots, \mathbf{p}_s^{(K)}\}$ for each shape model s directly from the belief of $\mathbf{p}_s$,

$$\text{belief}(\mathbf{p}_s) = \prod_{q \in \mathcal{Q}_s} \mu_{\kappa_q \to \mathbf{p}_s}(\mathbf{p}_s) \tag{8}$$

where $\mu_{\kappa_q \to \mathbf{p}_s}$ is the message from the raylet potential κ_q to the pose variable $\mathbf{p}$. Unfortunately, directly sampling from this distribution is difficult. We therefore resort to Metropolis-Hastings (MH) sampling [17] and run a Markov Chain to obtain the desired sample set. However, a straightforward application of MCMC sampling [2] to Eq. 8 is highly inefficient as each function evaluation requires processing all voxels along each raylet of shape model s, densely querying the voxel grid. Instead, we seek a proposal distribution $\omega_s(\mathbf{p})$ that is efficient to evaluate and approximates Eq. 8 sufficiently well.

We observe that most voxels along each raylet can be ignored when computing $\mu_{\kappa_q \to \mathbf{p}_s}$. Since the raylet potential in Eq. 5 evaluates the TSDF of only the first visible voxel, voxels with small occupancy belief do not contribute significantly to the equation. Thus, we consider only the voxels with substantial occupancy belief to accelerate MCMC sampling. In particular, our approach extracts a sparse cloud of voxel centers from the volume, ignoring voxels with low occupancy belief. The proposal distribution $\omega(\mathbf{p}_s)$ is

$$-\log \omega(\mathbf{p}_s) = \sum_{\ell=1}^{L} \max\left(0, 1 - \frac{d^\ell(\mathbf{p}_s)}{\tau}\right) \tag{9}$$

where L is the number of voxels with substantial occupancy belief and $d^\ell(\mathbf{p}_s)$ denotes the distance of voxel ℓ to the model surface at pose $\mathbf{p}_s$. Our parallelized implementation requires about 1ms to evaluate a single proposal $\mathbf{p}_s$ given 100k 3D points. For each surface model s, we draw $K = 64$ samples from Eq. 9.

Runtime: Our implementation uses grid-octree data structures [24] and GPGPU parallelization for message passing. Passing all ray potential messages takes 7 seconds for a 1MP image and a scene with roughly 30 million voxels. The MCMC sampling (10K iterations) and the raylet potential message passing for a single shape model typically takes roughly 5 and 10 seconds, respectively.

6. Experimental Evaluation

We evaluate our algorithm on four challenging datasets with ground truth geometry. Sample images from each dataset are presented in Fig. 3.

The LIVINGROOM dataset contains realistic renderings of a synthetic living room. The data is part of the "augmented ICL-NUIM dataset" distributed by Choi et al. [9]. We use the "Living room 2" camera trajectory [2] and sample every tenth image, for a total of 234 images. The images are 640x480 pixels in size. Choi et al. [9] used this dataset for camera tracking and reconstruction from depth images. In our work, we assume fixed camera poses and consider reconstruction from RGB images. To simulate a realistic setting, we do not use the ground truth camera poses provided by the dataset but obtain the poses and camera calibration using structure from motion [35, 36]. This dataset is highly challenging due to the large textureless surfaces such as walls, limited viewpoints and many reflective materials.

The other three datasets were captured in urban environments from an aerial platform. The images, camera poses and LIDAR points are provided by Restrepo et al. [28]. The images are one Megapixel in size and each dataset contains ∼200 images. The original datasets are distributed with sparse LIDAR points. Ulusoy et al. triangulated these

[2] http://redwood-data.org/indoor/dataset.html/

(a) LIVINGROOM (b) CAPITOL

(c) DOWNTOWN (d) DOWNTOWN2

Figure 3: Sample images from the datasets we use.

points to obtain a dense ground truth mesh [34]. We use their meshes for a fair comparison to the baselines.

While the DOWNTOWN and DOWNTOWN2 datasets were captured roughly at the same location, illumination conditions are significantly different as seen in Fig. 3c+3d. While DOWNTOWN was recorded on a cloudy day, DOWNTOWN2 was captured on a sunny day close to sunset, causing long shadows and strong reflections.

Object Shape Proposals: Our approach requires a set of plausible object shapes. The method to retrieve these proposals is scene-dependent and in particular, depends on the available semantic information.

For the LIVINGROOM dataset, we cropped four objects from the ground-truth mesh: a chair, sofa, table and cupboard. While these models allow for evaluation given unknown object pose but "perfect" object shape, in most realistic scenarios, available shape models are often approximate. We therefore test our algorithm's robustness to approximate input object shapes using IKEA furniture models from [21] that only coarsely resemble the true object shapes.

For the aerial datasets, we use the approximate geolocation information to retrieve relevant 3D models from Trimble 3D Warehouse. All three aerial datasets were collected in downtown Providence, Rhode Island, USA. A search for the keywords "Providence, Rhode Island" on the Trimble 3D Warehouse returned several building models. We use the rough geolocation information of each model to filter out models that are not contained within the scene boundary. For the CAPITOL dataset, this resulted in a single building model which is the Rhode Island State House as shown in Fig. 7a. For the DOWNTOWN and DOWNTOWN2 datasets, we obtained eleven building models as shown in Fig. 1b. The retrieved models are geometrically inaccurate and do not match the ground truth. Moreover, five of the eleven retrieved models are located in the periphery of the scene and visible only in a few input images. Our inference typically assigns low probability to the presence of these objects. We provide detection experiments in the supp. document.

Coarse Object Localization: To accelerate the MCMC pose sampling process (see Section 5), we first coarsely discretize the pose space and evaluate the pose likelihood Eq. 9 at each point. We then use the modes of this distribution to initialize Markov chains that explore the pose space locally. In particular, we use the knowledge of the ground plane, that is estimated via robust plane fitting, to restrict the poses to translations on the ground plane and rotations around the up vector. Fast evaluation of Eq. 9 and the restricted pose space allows for exhaustive search in a few seconds. While this strategy worked well for the aerial scenes, we observed a few failure cases for the LIVINGROOM dataset. For these cases, a rough initial pose estimate can be obtained by semantic segmentation or object detection.

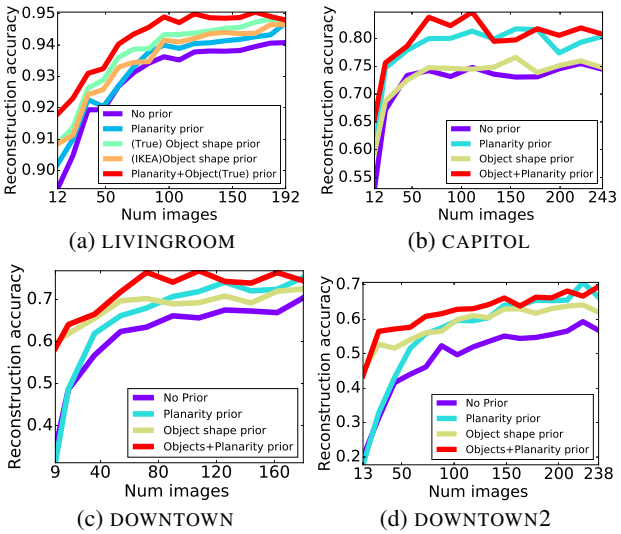
Model Parameters: We use the same set of parameters for the aerial and indoor datasets. Our supp. document provides details and experiments with varying sets of parameters.

Baselines: We compare our results to several state-of-the-art approaches. First, we compare against Ulusoy et al. [34] whose formulation is equivalent to removing the object shape prior from our model and which we refer to as "No prior" in the following. Second, we compare against [33], which integrates a planarity prior into the formulation of [34] and achieves state-of-the-art results on both the CAPITOL and the DOWNTOWN datasets. We refer to this baseline as "Planarity prior". Finally, we evaluate a combination of the planarity prior [33] with our object shape prior, which we refer to as "Object+Planarity".

Evaluation Protocol: We follow [34] and evaluate reconstruction accuracy as the absolute error in depth map prediction with respect to the depth maps that are generated by projecting the ground truth meshes into all input views. In particular, we compute the percentage of pixels falling below an error threshold while varying this threshold from 0 to 3 meters for the indoor dataset and 0 to 10 meters for the aerial datasets. See Fig. 5 in [34] for an illustration. We obtain a single accuracy value between 0 and 1 by considering the normalized area under this curve.

To compute depth maps from the probabilistic 3D models, we follow Ulusoy et al. [34], who showed that choosing the median value of each pixel's depth distribution minimizes our error metric. They further showed that sum-product belief propagation yields per-pixel depth distributions as a by-product. Note that the depth distributions in their approach rely only on the image evidence whereas the depth distributions in our approach integrate information from both the input images and object shape models.

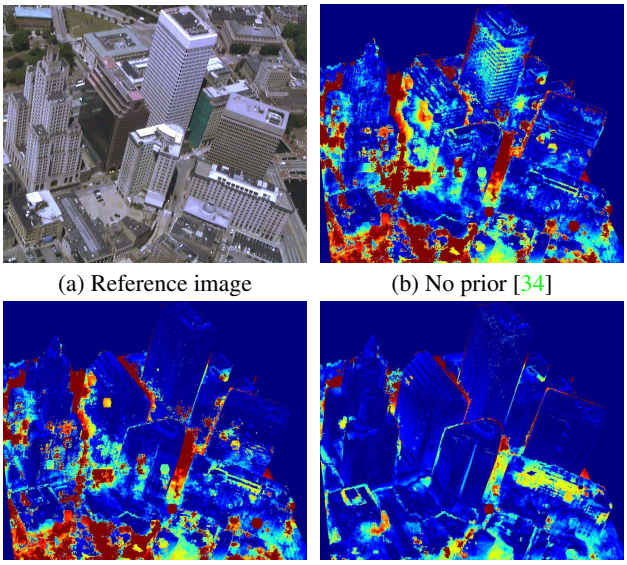
We evaluate the aforementioned baselines on all four



(a) LIVINGROOM (b) CAPITOL

(c) DOWNTOWN (d) DOWNTOWN2

Figure 4: Comparison of reconstruction accuracy of several baselines for varying number of images. Higher is better.



(a) Reference image (b) No prior [34]

(c) Shape model fitting (d) Proposed shape prior

Figure 5: Visualization of depth errors for DOWNTOWN. Cooler colors depict lower errors. See text for details.

datasets and report the results in Fig. 4. We also varied the number of images in each dataset by subsampling the input views uniformly in space. This experiment illustrates the benefit of the proposed object shape prior in particular when reconstructing from a small set of input images.

The results indicate that both the proposed object shape prior and the planarity prior [33] improve reconstruction accuracy over the baseline with no prior [34]. For LIVINGROOM, our shape prior performs better than the planarity prior independent of the number of images used. For CAPITOL, the planarity prior achieves higher accuracy due to the flat textureless grass region where planarity is an appropriate prior [33]. For DOWNTOWN and DOWNTOWN2, the object shape prior results in significantly better performance than the planarity prior, in particular for small number of input images. Given sufficiently many images, the planarity prior achieves similar or better results. Overall, combining the planarity and object shape prior achieves the best results. We provide a more detailed analysis below.

Small number of images: Fig. 4 shows that for a small number ($\sim$10) of images, the object shape prior achieves significant improvements over the baseline without priors. In contrast, the planarity prior yields little to no improvement because it requires an adequate initial reconstruction to sample plane hypothesis from. For $\sim$10 images, the initial reconstructions are highly ambiguous, therefore impairing the planarity prior.

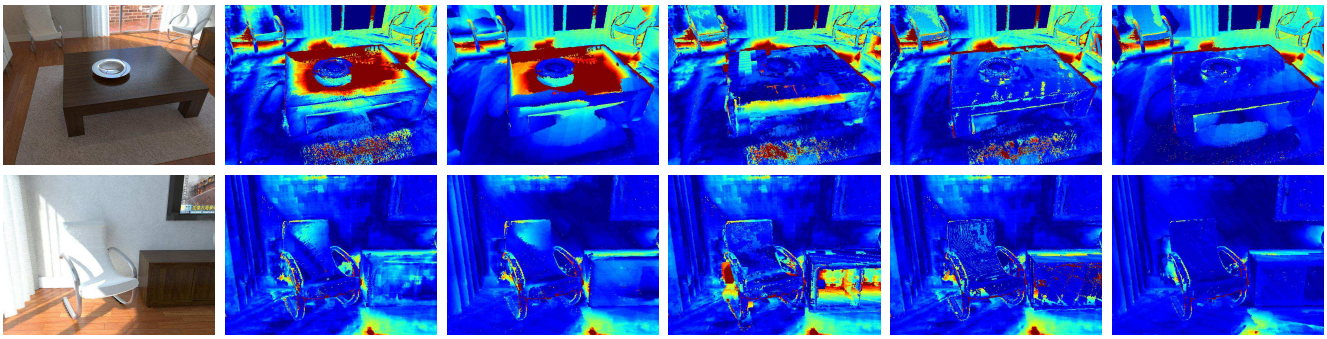
Fig. 5 visualizes the depth errors in one of the input views. Our approach (Fig. 5d) significantly improves accuracy with respect to the baseline (Fig. 5b). Note that improvements are visible everywhere in the scene and not limited at building surfaces for which shape priors apply. Our inference scheme exploits the geometric knowledge induced by the prior to refine free-space areas and visibility constraints in the *entire* scene, leading to higher accuracy also in regions for which no shape priors are available.

This improvement is made possible by our probabilistic model that accurately models visibility using ray-potentials and integrates shape priors in a principled manner. In contrast, existing methods first reconstruct a 3D model from the images and then fuse shape models into this 3D reconstruction [3, 11, 39]. Such approaches can not achieve improvements where no shape priors are available. We demonstrate the benefit over such methods by comparing to a baseline that reconstructs a 3D model using no prior [34] and then incorporates shape models using a single iteration of raylet-to-voxel message passing. As shown in Fig. 5c, the result is significantly worse compared to our approach (Fig. 5d). We provide further examples in the supplementary.

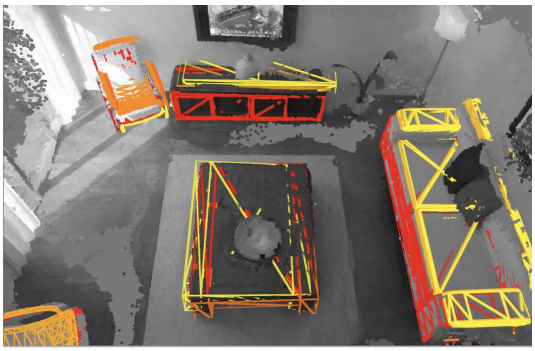
Robustness to approximate input shape: We evaluate the robustness of our approach to approximate input shapes using the LIVINGROOM dataset. The results in Fig. 4a indicate that our approach improves accuracy even when using IKEA models that are only coarse approximations to the true 3D shapes. See Fig. 6a+6b for a comparison. As expected, performance increases further when using the true 3D shapes. We provide qualitative results in Fig. 6.

Combining image and shape evidence: Our method combines image evidence and the input shape models to produce

(a) True shape models for LIVINGROOM (b) Approximate shape models from IKEA [21]

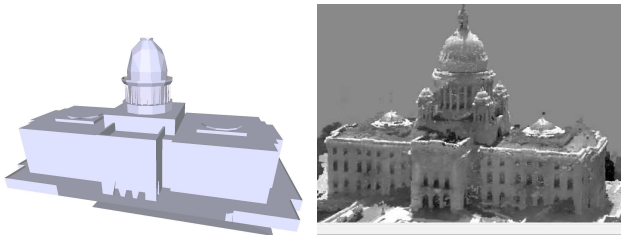


(c) Reference image (d) No prior (e) Planarity prior (f) IKEA prior (g) True prior (h) Object+Planarity



(i) Our Reconstruction

Figure 6: (d-h) Visualization of depth errors for LIVINGROOM. Cooler colors correspond to lower error. **Top-row:** The reflective table surface causes large errors for the baseline without prior (d) [34]. The planarity prior (e) [33] is unable to correct this error. Our approach significantly improves accuracy even when using approximate IKEA models (f). Using the correct table prior (g) further improves the result. Combining the true object shapes with the planarity prior (h) yields the best results. **Bottom-row:** As above, our approach improves over the baseline with no prior even when using the inaccurate IKEA chair model. However, using the IKEA cupboard (second from right in (b)) causes incorrect holes in the reconstruction (f). **Left:** (i) Point cloud of our dense 3D reconstruction overlaid with a subset of the pose samples. Objects are colored according to their belief (yellow=unlikely, red=likely).



(a) Input shape model (b) Reconstruction result

Figure 7: Our method is able to combine image evidence (see Fig. 3b) and the approximate shape models (a) to produce detailed reconstructions (b).

detailed reconstructions. Fig. 7 presents an example where our method has successfully recovered fine scale structures that are not present in the input model. Note that our reconstruction includes details such as the small towers next to the cupola and the tip of the cupola even though they are absent from the input shape model.

7. Conclusion

In this paper, we present a probabilistic approach that integrates object-level shape priors with image-based 3D reconstruction. Our experiments demonstrate that the proposed shape prior significantly improves reconstruction accuracy, in particular when the number of input images is small. To the best of our knowledge, our approach is the first to simultaneously reconstruct a dense 3D model of the entire scene and a structural representation of the objects in it. Our experiments demonstrate the benefit of this joint inference. Further, we believe such an integrated representation of 3D geometry and semantics is a step towards holistic scene understanding and will benefit applications such as augmented reality and autonomous driving.

Future directions include incorporating parametric shape models to improve generality of our prior. We also believe recent 3D pose estimation methods [14, 19, 23] can be used to improve pose proposals during inference.

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. 2

[2] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1-2):5–43, 2003. 5

[3] 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, 7

[4] 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

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

[6] 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] A. X. Chang, T. A. Funkhouser, L. J. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. Shapenet: An information-rich 3d model repository. *arXiv.org*, 1512.03012, 2015. 2

[8] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun. 3d object proposals for accurate object class detection. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 2

[9] S. Choi, Q. Zhou, and V. Koltun. Robust reconstruction of indoor scenes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 5

[10] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *ACM Trans. on Graphics (SIGGRAPH)*, 1996. 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, 7

[12] 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

[13] 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. 2

[14] S. Gupta, P. A. Arbeláez, R. B. Girshick, and J. Malik. Aligning 3D models to RGB-D images of cluttered scenes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 8

[15] 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

[16] 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

[17] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970. 5

[18] A. Ihler and D. McAllester. Particle belief propagation. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009. 4, 5

[19] W. Kehl, F. Milletari, F. Tombari, S. Ilic, and N. Navab. Deep learning of local rgb-d patches for 3d object detection and 6d pose estimation. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2016. 8

[20] 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

[21] J. J. Lim, A. Khosla, and A. Torralba. FPM: Fine pose parts-based model with 3D CAD models. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 6, 8

[22] 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. 2

[23] F. Massa, B. C. Russell, and M. Aubry. Deep exemplar 2d-3d detection by adapting from real to rendered views. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 8

[24] A. Miller, V. Jain, and J. L. Mundy. Real-time rendering and dynamic updating of 3-d volumetric data. In *Proc. of the Workshop on General Purpose Processing on Graphics Processing Units (GPGPU)*, page 8, 2011. 5

[25] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Proc. of the International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011. 2

[26] M. Pauly, N. J. Mitra, J. Giesen, M. H. Gross, and L. J. Guibas. Example-based 3d scan completion. In *Eurographics Symposium on Geometry Processing (SGP)*, 2005. 2

[27] 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. 2

[28] M. I. Restrepo. *Characterization of Probabilistic Volumetric Models for 3-d Computer Vision*. PhD thesis, Brown University, 2013. 5

[29] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. J. Kelly, and A. J. Davison. SLAM++: simultaneous localisation and mapping at the level of objects. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[30] 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. 2

[31] 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

[32] S. Song and J. Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[33] A. O. Ulusoy, M. Black, and A. Geiger. Patches, planes and probabilities: A non-local prior for volumetric 3d reconstruction. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 6, 7, 8

[34] 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, 8

[35] C. Wu. Towards linear-time incremental structure from motion. In *Proc. of the International Conf. on 3D Vision (3DV)*, 2013. 3, 5

[36] 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, 5

[37] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[38] C. Zach, T. Pock, and H. Bischof. A globally optimal algorithm for robust tv-l1 range image integration. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2007. 2

[39] 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, 7