Semantically-Aware Aerial Reconstruction from Multi-Modal Data

Randi Cabezas Julian Straub John W. Fisher III
Massachusetts Institute of Technology
{rcabezas,straub,fisher}@csail.mit.edu

Abstract

We consider a methodology for integrating multiple sensors along with semantic information to enhance scene representations. We propose a probabilistic generative model for inferring semantically-informed aerial reconstructions from multi-modal data within a consistent mathematical framework. The approach, called Semantically-Aware Aerial Reconstruction (SAAR), not only exploits inferred scene geometry, appearance, and semantic observations to obtain a meaningful categorization of the data, but also extends previously proposed methods by imposing structure on the prior over geometry, appearance, and semantic labels. This leads to more accurate reconstructions and the ability to fill in missing contextual labels via joint sensor and semantic information. We introduce a new multi-modal synthetic dataset in order to provide quantitative performance analysis. Additionally, we apply the model to real-world data and exploit OpenStreetMap as a source of semantic observations. We show quantitative improvements in reconstruction accuracy of large-scale urban scenes from the combination of LiDAR, aerial photography, and semantic data. Furthermore, we demonstrate the model's ability to fill in for missing sensed data, leading to more interpretable reconstructions.

1. Introduction

Humans integrate various sensory, semantic, and contextual cues to construct internal representations of the world for robustly reasoning within their environment. This ability is little diminished in the face of sparse and noisy observations. Furthermore, humans easily extrapolate the structure of the surrounding large-scale environment from their local surroundings using cues and prior experiences of similar scenes. Such inferential feats are supported by semantic understanding and categorization of scene elements. Motivated by such abilities we develop an approach for scene reconstruction that integrates sensor data along with georeferenced semantic labels. With some exceptions, existing approaches focus solely on obtaining a semantic labeling from sensor data. We construct a probabilistic generative model that allows multi-modal data fusion, integrates semantic observations, and captures the notion that different scene components exhibit different properties via a structured prior over the different modalities. The approach, Semantically-Aware Aerial Reconstruction (SAAR), infers semantically-meaningful categories of data which are used in the reconstruction as category-specific priors, resulting in improved reconstruction accuracy.

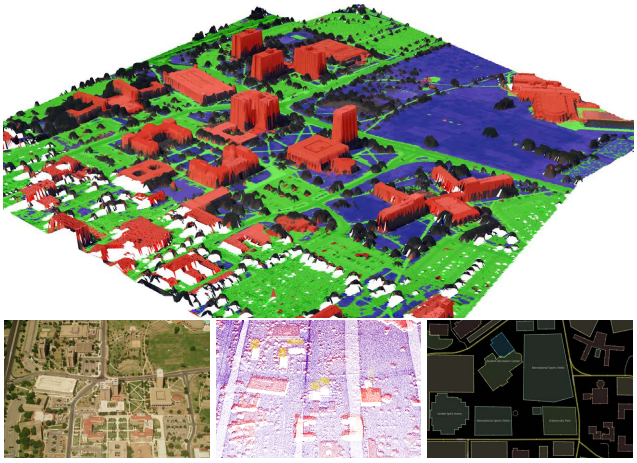

Figure 1: *Top:* Lubbock scene inferred geometry and labels. *Bottom:* Aerial image, LiDAR and OSM observations.

SAAR builds on the work of Cabezas *et al*. [4], by adding semantic observations from OpenStreetMap (OSM) [12] to the sensor data already used there: aerial oblique imagery, and Light Detection and Ranging (LiDAR). Each of these sources provides complementary features for reasoning about urban scenes. We further expand on [4] by introducing a novel structured prior, to regularize reconstructions. In contrast to prior semantic reconstruction work, our goal is to (1) exploit the learned categories to improve the reconstruction and (2) fill in missing sensor data.

Previous work in the area of semantic labeling can be categorized based on the domain of the labels: image-space or 3D-space. In the first category, Sudderth *et al*. [31] and Wang *et al*. [35] focus on the extraction of pixel-wise labels

from single images with no additional information. Along similar lines, Cao *et al*. [5] and Choi *et al*. [6] expand on prior work by introducing spatial connectivity between the labels. They show that spatial connectivity greatly improves performance. Similarly, the problem of matching noisy text tags to images and, if possible, identifying image regions that give rise to the textual description has been extensively studied [8, 19, 20, 22, 24, 25, 38]. These approaches typically learn multi-modal (text and appearance) representations and attempt to predict one modality when the other is missing. All of the aforementioned works are formulated in the image-domain and differ from the proposed approach in both formulation and goal.

Prior work in 3D scene labeling can be categorized by choice of primitive used to produce the labeling: point, voxel, mesh, or object box (while not strictly a primitive, it is included for completeness). Point-based semantic labeling has received the most attention in recent years [1, 3, 10, 14, 18, 37]. The goal is to label 3D points, either using human annotations and propagating them through the point cloud, or by following procedures similar to image-based methods and projecting the results into 3D using standard multi-view techniques. Voxel-based methods, *e.g*., [13, 16, 17, 28, 34], formulate the labeling problem as an energy minimization problem in either 2D (followed by projection to 3D), or directly in 3D. Most of these methods exploit spatial connectivity using Markov Random Fields or Conditional Random Fields (CRFs). Like some voxel-based methods, mesh-based approaches [2, 33] rely on energy minimization in a CRF; however, unlike voxel-based approaches, they tend to have a richer set of discriminative features such as texture, curvature and various mesh properties. Approaches that label scene objects (typically in the form of a bounding box) provide a higher degree of abstraction than primitive-based methods. Ren *et al*. [27] and Lin *et al*. [21] rely on a set of low-level features to find and classify scene regions.

The proposed work falls in the mesh-based category. Like other methods in this category, it relies on high-level features, including primitive appearance, geometry via location and orientation, and, if available, semantic observations, such as OSM or Geiger mode LiDAR. This data is used in a probabilistic model to learn scene categories and category-specific structured priors that can be used to regularize scene elements during reconstruction.

The contributions of this work are: (1) a novel probabilistic model that couples semantic labels and scene reconstruction, (2) mathematically consistent methods for obtaining labels and reconstructions from multi-modal data in 3D, (3) demonstration of the utility of semantic observations and structured prior distributions to improve the accuracy of scene reconstruction, and (4) the introduction of a new photo-realistic multi-modal synthetic urban city dataset.

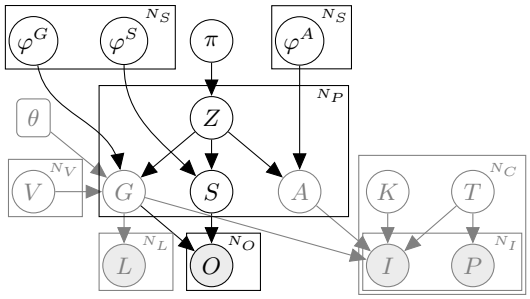

Figure 2: Graphical representation of the SAAR model. The parts taken over from [4] are depicted in gray whereas the proposed structured prior is shown in black.

2. The Probabilistic SAAR Model

The proposed SAAR model couples a latent structured prior model with a probabilistic semantic 3D world representation. The 3D representation is typically based on a collection of independent primitives (*e.g*., points, voxels, or triangles as in our case) with a series of attributes that describe the various scene aspects (*e.g*., geometry and appearance). SAAR draws from the model proposed in [4] where, in addition to the latent geometry and appearance per primitive (*i.e*., triangle), we also model a semantic label. Furthermore, we introduce a structured prior via a mixture model over the latent semantic labels, appearance, and geometry to replace the uninformative uniform priors used in [4]. We will show that posterior inference under such a prior captures meaningful scene-specific global structure, which can be leveraged to regularize the 3D reconstructions. These extensions lead to powerful regularizers for 3D reconstruction without the need of carefully hand-crafted scene priors.

From the generative viewpoint, the SAAR model describes how the combination of latent geometry $\mathbf{G}$, appearance $\mathbf{A}$, and semantic label $\mathbf{S}$ give rise to LiDAR measurements $\mathbf{L}$, observed OSM labels $\mathbf{O}$, and camera images $\mathbf{I}$ at observed GPS positions $\mathbf{P}$. The geometry is represented via vertex locations $\mathbf{V}$ and connectivity matrix θ. The images are assumed to be generated from a set of fly-by cameras with poses $\mathbf{T}$ (extrinsic) and calibration $\mathbf{K}$ (intrinsic). For convenience we let $\mathbf{W} \triangleq \{\mathbf{G}, \mathbf{A}, \mathbf{S}, \mathbf{Z}\}$, where $\mathbf{Z}$ is the primitive's categorical assignment. The probabilistic graphical representation of this model is visualized in Fig. 2 (hyperparameters are omitted for clarity). Model parameters are summarized in Tab. 1. The joint distribution for the probabilistic SAAR model is:

$$
\begin{aligned}
p(\mathbf{L}, \mathbf{I}, \mathbf{P}, \mathbf{O}, \mathbf{T}, \mathbf{K}, \mathbf{V}, \mathbf{W}, \varphi, \pi; \theta) &= p(\mathbf{W}, \varphi, \pi | \mathbf{V}; \theta) \\
&\times \prod_{v=1}^{N_V} p(V_v) \prod_{l=1}^{N_L} p(L_l | \mathbf{G}) \prod_{o=1}^{N_O} p(O_o | \mathbf{S}, \mathbf{G}) \\
&\times \prod_{c=1}^{N_C} p(T^c) p(K^c) \prod_{n=1}^{N_I^c} p(I_n^c | \mathbf{G}, \mathbf{A}, K^c, T^c) p(P_n^c | T^c) \,,
\end{aligned}
\tag{1}
$$

Variables	Description
$N_P,\ N_V,\ N_L,\ N_O,\ N_C,\ \{N_I^c\}_{c=1}^{N_C}$	Number of primitives, vertices, LiDAR and OSM points; cameras and images.
$N_S,\ N_D,\ N_A,\ N_B$	Number of semantic and OSM categories; appearance and image pixels.
$L_l \in \mathbb{R}^3$	LiDAR observation.
$O_o = (P_o \in \mathbb{R}^3, C_o \in \{1, ..., N_d\})$	OSM observation (location, categories).
$V_m \in \mathbb{R}^3, \quad G_m \in \mathbb{N}^{1\times 3},\ \theta \in \mathbb{R}^{N_p \times 3}$	Vertex location, Geometric primitive (triangle), Connectivity matrix.
$A_m = \{a_p \,\|a_p \in \mathbb{R}^3\}_{p=1}^{N_A^2}$	Primitive appearance (texture) and corresponding RGB pixel.
$S_m = \{C_o \| \forall o \text{ assigned to } G_m\}$	OSM category distribution (all observations assigned to primitive m).
$Z_m \in \{1, ..., N_s\}$	Primitive category.
$T^c \in \mathrm{SE}(3), \quad K^c \in (0, 180]$	Extrinsic trajectory (position and orientation) and Intrinsic parameter (FOV).
$I_n^c = \{p_j \,\|p_j \in \mathbb{R}^3\}_{j=1}^{N_B}$	n^{th} image taken with camera c modeled as a collection of RGB pixels.
$P_n^c \in \mathbb{R}^3$	n^{th} GPS observation of camera c.
$\pi,\ \varphi = \{\varphi^G,\ \varphi^A,\ \varphi^S\}$	Cluster proportions; geometry, appearance and semantic parameters (see text).

Table 1: List of variables used in the model.

where the structured prior over primitive geometry, appearance, and semantic label factors as:

$$p(\mathbf{W}, \varphi, \pi | \mathbf{V}; \theta) = p(\pi) \prod_{k=1}^{N_S} p(\varphi_k^G) p(\varphi_k^A) p(\varphi_k^S) \prod_{m=1}^{N_P} \big[p(Z_m|\pi) \times p(G_m|\varphi^G, Z_m, \mathbf{V}; \theta) p(S_m|\varphi^S, Z_m) p(A_m|\varphi^A, Z_m) \big] . \quad (2)$$

We now describe the terms in Eq. (1). The image likelihood, $p(I_n^c|\mathbf{G}, \mathbf{A}, K^c, T^c) = \prod_{k=1}^{N_B} \mathcal{N}(i_k; a_{m(k)}, r_{m(k)}^2)$, is modeled as a Gaussian distribution with mean corresponding to the latent appearance pixel of the associated primitive and the variance corresponding to the inverse of the dot product between the camera's viewing direction and the surface normal of the visible primitive. The LiDAR likelihood is given by $p(L_l|\mathbf{G}) = \mathcal{N}(d^2(L_l, G_{m(l)}); 0, \sigma^2)$, where $d(L_l, G_m)$ is the distance between the LiDAR point and primitive. The GPS observations follow a Gaussian distribution centered at the camera trajectory, T^c. All priors, with the exception of the structured prior, are uniform. For further details see [4].

Semantic observations $\mathbf{O}$ are modeled in SAAR as a collection of independent 3D points, where each point has a location and semantic label, $O_o \triangleq (P_o, C_o)$. The location and label are assumed to be independent, thus the likelihood is $p(O_o|\mathbf{S}, \mathbf{G}) = p(P_o|\mathbf{G}) p(C_o|\mathbf{S})$. The location model is the same as that of the LiDAR measurements; *i.e.*, the likelihood depends on the distance between the point and the generating primitive. The class likelihood is modeled as a Categorical distribution (Cat): $p(C_o|\mathbf{S}) = \mathrm{Cat}(C_o; S_{m(o)})$.

Equation (2) shows the mixture model structure of the prior on geometry, appearance, and semantic label, shown in black in Fig. 2. Note that each primitive m is assigned to a single mixture component via Z_m. Each mixture component defines a distribution over the product space of geometry, appearance and semantic labels. Following the standard approach in Bayesian mixture modeling, we assume a Dirichlet distribution (Dir) prior with hyperparameter α on the mixture weights π. The class labels Z_m are distributed according to a categorical distribution parametrized by π: $p(Z_m|\pi) = \mathrm{Cat}(Z_m; \pi)$.

Conditioned on the category assignment via label Z_m, the geometry, appearance, and semantic label of primitive m are modeled as generated independently from the associated component distribution in the respective space. For this process, we adopt a pixel-centric perspective inside each triangle. Specifically, we model the 3D location, surface normal, RGB color and semantic label of each primitive's appearance pixels as independently and identically distributed according to the corresponding mixture component distribution. We note that this modeling is performed using the latent primitive attributes. Under this model, we can collect the sufficient statistics over the different modalities and use them for both likelihood evaluations and posterior inference. In the following, we introduce the distributions for appearance, semantic labels and geometry.

Appearance: For each mixture component, we model the appearance as a three-dimensional Gaussian in the RGB color space. Hence, given appearance parameter $\varphi_k^A \triangleq (\mu_k^A, \Sigma_k^A)$ we have

$$p(A_m|\varphi_{Z_m}^A) = \prod_{i=1}^{N_m^A} \mathcal{N}(A_{m,i};\ \mu_{Z_m}^A, \Sigma_{Z_m}^A) , \quad (3)$$

where $A_{m,i}$ is the RGB color of pixel i in primitive m and the product is over all pixels in the primitive. The Gaussian parameters of φ^A are distributed according to the Normal Inverse Wishart (NIW) conjugate prior distribution [11].

Semantic labels: Labels S_m assigned to primitive m are modeled as following a categorical distribution with a Dirichlet prior:

$$p(S_m|\varphi_{Z_m}^S) = \prod_{i=1}^{N_m^A} \mathrm{Cat}(S_{m,i};\ \varphi_{Z_m}^S) . \quad (4)$$

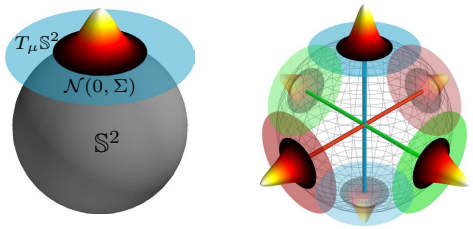

Figure 3: *Left:* Tangent space Gaussian (TG) around mean μ with covariance Σ. *Right:* Manhattan Frame (MF).

Geometry: The geometry of primitive m is modeled via the primitive's appearance pixel locations in 3D, $G^X_{m,i}$, as well as its orientations, $G^Q_{m,i}$, for pixel i. The pixel locations $G^X_{m,i}$ are assumed to be Gaussian distributed in 3D with NIW distribution priors on the Gaussian parameters $\{\mu^G_k, \Sigma^G_k\}$, along with likelihood

$$p(G^X_m|\varphi^G_{Z_m}) = \prod_{i=1}^{N^A_m} \mathcal{N}(G^X_{m,i};\, \mu^G_{Z_m}, \Sigma^G_{Z_m})\,. \quad (5)$$

As we will see, the structured prior uses the location information as a weak coupling between neighboring triangles. The orientation $G^Q_{m,i}$ of the primitive's pixel i is described via its surface normal represented as a unit-length direction vector in 3D. The space of such vectors is the unit sphere in 3D, $\mathbb{S}^2$. Accordingly, we model primitive orientations by placing zero-mean Gaussian distributions in tangent spaces to $\mathbb{S}^2$ as explored in [29, 30]. We denote this distribution, visualized in Fig. 3, the tangent space Gaussian (TG). Under the TG model, surface normals $G^Q_m \in \mathbb{S}^2$ associated with cluster $Z \triangleq Z_m$ have the following distribution

$$p(G^Q_m|\mu^Q_Z, \Sigma^Q_Z) = \prod_{i=1}^{N^A_m} \mathcal{N}(\mathrm{Log}_{\mu^Q_Z}(G^Q_{m,i}); 0, \Sigma^Q_Z) \quad (6)$$

where the Riemannian logarithm map $\mathrm{Log}_{\mu^Q_Z}(G^Q_{m,i})$ maps $G^Q_{m,i}$ into the tangent space $T_{\mu^Q_Z}S$ around the mean of the TG $\mu^Q_Z \in \mathbb{S}^2$. The TG model uses an Inverse Wishart prior [11] in the tangent plane for the covariance and a uniform prior on the sphere for the mean. We explore two different models: (1) an unconstrained model [29] with a single TG per cluster and (2) the Manhattan Frame (MF) model [30]. The MF captures the block structure of man-made environments in the space of surface normals through six TG clusters that are constrained to orthogonal and opposing locations on the sphere (Fig. 3).

3. Inference

In this section we discuss sampling-based inference for the SAAR model. We begin by outlining inference for the structured prior portion of the model, followed by an inference scheme for the full SAAR model.

Algorithm 1 Structured Prior Inference

1: Initialize φ^G, φ^A , φ^S and π from priors.
2: **for** $i \in \{1, \ldots, N_{\text{iter}}\}$ **do**
3: Sample $\mathbf{Z}$ according to Eq. (7).
4: Sample φ^G, φ^A φ^S and π according to Eq. (8).
5: **end for**

3.1. Structured Prior Inference

As stated in the previous section, the structured prior for geometry, appearance, and semantic labels is equivalent to a mixture model over the aforementioned modes. This motivates the use of a Gibbs sampler that iterates between sampling labels Z_m for each primitive and sampling mixture component parameters for posterior inference. We present the necessary posterior distributions in the following; for a detailed derivation see Sup. Mat.

Label posterior: The posterior distribution for a label Z_m of primitive m given the mixture parameters is

$$\begin{aligned} &p(Z_m = k|\mathbf{Z}_{\setminus m}, \mathbf{G}, \mathbf{A}, \mathbf{S}, \pi, \varphi^G, \varphi^A, \varphi^S) \\ &\quad \propto p(A_m, G_m, S_m, \varphi^G, \varphi^A, \varphi^S|Z_m)p(Z_m = k|\pi) \\ &\quad = p(G_m, \varphi^G|Z_m)p(A_m, \varphi^A|Z_m)p(S_m, \varphi^S|Z_m)\pi_k \\ &\quad \propto p(G_m|\varphi^G_k)p(A_m|\varphi^A_k)p(S_m|\varphi^S_k)\pi_k\,. \end{aligned} \quad (7)$$

For each primitive m we can evaluate Eq. (7) under all clusters k using the likelihoods described in the previous section to obtain a discrete probability distribution. After normalization, we sample the indicator Z_m.

Parameter posteriors: The posterior distributions over the parameters of the mixture model factors into the different modes as

$$\begin{aligned} &p(\pi, \varphi^G, \varphi^A, \varphi^S|\mathbf{Z}, \mathbf{G}, \mathbf{A}, \mathbf{S}) \\ &\quad \propto p(\pi|\mathbf{Z})p(\varphi^A|\mathbf{A}, \mathbf{Z})p(\varphi^G|\mathbf{G}, \mathbf{Z})p(\varphi^S|\mathbf{S}, \mathbf{Z}) \\ &\quad \propto p(\pi|\mathbf{Z}) \prod_{k=1}^{N_S} p(\varphi^A_k|A_{\mathcal{I}_k})p(\varphi^G_k|G_{\mathcal{I}_k})p(\varphi^S_k|S_{\mathcal{I}_k}) \end{aligned} \quad (8)$$

where we use the indicator set $\mathcal{I}_k = \{m : z_m = k\}$ to collect all primitives that are assigned to cluster k. Due to conjugacy of the priors on π, φ^G, φ^A, and φ^S, the posterior parameter distributions take the same form as the prior distributions. This allows efficient sampling of posterior parameters after updating the sufficient statistics [11]. Posterior sampling for the TG and MF distribution in φ^G is carried out as described in [29] and [30] respectively. The Gibbs sampler for the structured prior is outlined in Alg. 1.

3.2. SAAR Model Inference

Inference on the structured prior portion of the model is coupled with the inference scheme proposed in [4] to sample from the posterior of the SAAR model. Conceptually, the SAAR algorithm interleaves the geometry, appearance,

Algorithm 2 Full SAAR Model Inference

1: Initialize world primitives and camera pose
2: Sample assignment of LiDAR and OSM data.
3: Initialize φ^G, φ^A, φ^S, π and $\mathbf{Z}$
4: **for** $i \in \{1, \ldots, N_{\text{up}}\}$ **do**
5: Estimate appearance $\mathbf{A}$.
6: Optimize over camera pose $\mathbf{T}$ and $\mathbf{K}$.
7: Run Structured Prior inference procedure (Alg. 1).
8: Estimate appearance $\mathbf{A}$.
9: Optimize over world primitive geometry, $\mathbf{V}$.
10: Sample assignment of LiDAR and OSM data.
11: **end for**

and camera pose sampling with posterior sampling of the structured prior as outlined in Alg. 2. Specifically, the updates to the appearance and vertex locations now take the structured priors into account. The appearance updates have the same form as in [4]:

$$p(\mathbf{A}|\mathbf{I}, \mathbf{G}, \mathbf{K}, \mathbf{T}, \mathbf{Z}, \varphi^A) \tag{9}$$
$$\propto \prod_{c=1}^{N_C} \prod_{n=1}^{N_I^c} \prod_{k=1}^{N_B} p(I_k^{n,c}|\mathbf{G}, \mathbf{A}, K^c, T^c) \prod_{m=1}^{N_P} \prod_{i=1}^{N_A} p(A_{m,i}|\varphi^A, Z_m)\,,$$

where we now utilize the inferred appearance parameters of the assigned cluster. Since the form of the equation does not change, we can still solve it in closed-form. Similarly, the vertex terms are augmented with the semantic label

$$p(\mathbf{V}, \mathbf{G}|\mathbf{I}, \mathbf{L}, \mathbf{O}, \mathbf{Z}, \mathbf{S}, \varphi^G; \theta) \propto \prod_{o=1}^{N_o} p(O_o|\mathbf{G}, \mathbf{S}) \prod_{l=1}^{N_L} p(L_l|\mathbf{G})$$
$$\times \prod_{c=1}^{N_C} \prod_{n=1}^{N_I^c} p(I_n^c|\mathbf{G}, \mathbf{A}, K^c, T^c) \prod_{m=1}^{N_P} p(G_m|\varphi^G, \mathbf{V}; \mathbf{Z}; \theta)\,. \tag{10}$$

We optimize over Eq. (10) using a downhill simplex optimization method [23] to update the latent vertex locations to their maximum a posteriori configuration as in [4].

4. Results

To facilitate quantitative experiments, we created a multi-modal photo-realistic urban city synthetic dataset, Synthetic City (SynthCity). In the following, we briefly describe SynthCity and present several experiments to validate the propose model. Experiments include an evaluation of model performance in terms of reconstruction accuracy, and an ablation study to identify the best modalities for scene categorization. Furthermore, we demonstrate the utility of the model by showing its ability to operate in the presence of noisy data and to estimate missing sensor data. Finally, we present results on real-world scenes where we qualitatively show the method's improvements over the baseline of [4]. Throughout this section we'll use the shorthand notation TG-N (or MF-N) to refer to a TG (or MF) orientation model with N semantic categories, *i.e.*, $N_S = N$.

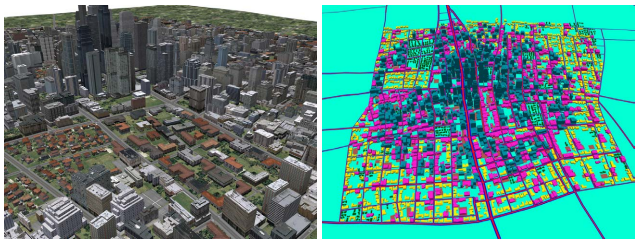

Figure 4: Sample view of SynthCity dataset (City3) along with ground-truth mesh colored by semantic categories.

4.1. SynthCity Dataset

The lack of ground-truthed aerial datasets motivates the creation of the SynthCity dataset. We used Esri's CityEngine [9] to create five randomly-generated realistic cities (ToyCity, ToyCity2, City1-1, City3, and City4), each with eight different types of scene elements: open space, streets, sidewalks, parking lots, residential buildings, office buildings, high-rise buildings, and vegetation. Custom build rules allowed pseudo-random variation of geometry and appearance, resulting in a collection of elements that are not only photo-realistic, but also closely match the layout of real-world cities (Fig. 4). LiDAR measurements were obtained by simulating real LiDAR collections [15]. Please refer to the Sup. Mat. for further details.

4.2. Improved Reconstructions - SynthCity

We exploit the resulting categorization and priors to improve reconstruction accuracy via label-dependent updates to geometry and appearance. The goal is to compare the effect of the structured prior and semantic information on reconstruction accuracy. Fig. 5 compares various configurations of SAAR (using all available data: appearance, location, orientation and semantic observations) with the non-structured prior model [4], and with the LiDAR-only method of Zhou *et al.* [39] on the SynthCity dataset. The distance metric used is the mean error between the estimated and true mesh as computed by Metro [7]. From the figure we can see that in all cases SAAR produces better reconstructions than both baseline approaches. We emphasize that the only difference between SAAR and [4] is the use of the structured prior and semantic observations. See Sup. Mat. for additional comparisons.

4.3. Clustering Results

We used SynthCity to obtain quantitative accuracy metrics for the SAAR model. This allows us to study which modalities lead to improved semantic labeling via reduction of reconstruction error. We considered various feature combinations using appearance, orientation, location, and semantic observations under both the TG and the MF models. We quantified the utility of each of the modalities by

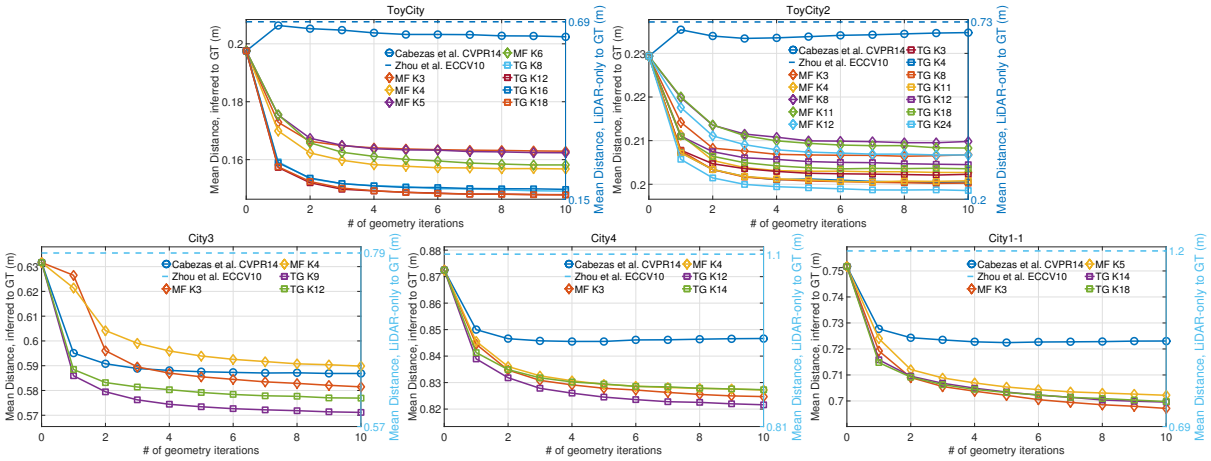

Figure 5: Mean geometry error for SynthCity reconstructions under [4] and SAAR (left y-axis); and [39] (right y-axis).

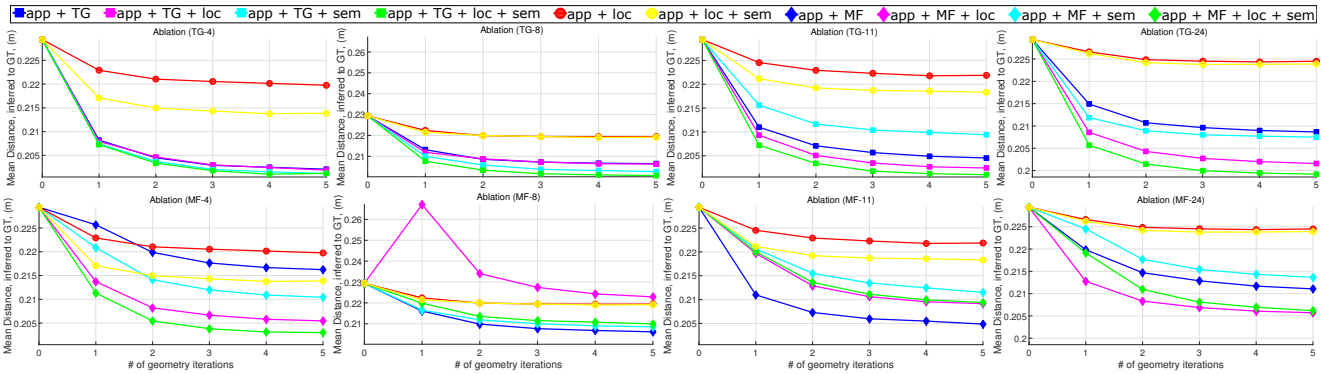

Figure 6: Ablation study SynthCity Toy2. *Columns:* Number of clusters (N_s): 4, 8, 11, 24. *Rows:* TG and MF models (square and diamond markers) respectively; color indicates combination of modalities: appearance, location, orientation and semantic (see legend). Reconstruction error decreases as modalities are added, independent of model or parameters used.

measuring the error of the estimated geometry against the ground-truth geometry. Fig. 6 shows that generally the reconstruction accuracy improves as more features are added. For example, consider TG-24 (top-right plot): as we add the location feature (magenta line) to the appearance and orientation features (blue line) we improve reconstruction accuracy; including the semantic feature (green line) provides further improvements. This behavior is seen across all TG models and most of the MF models. We hypothesize that as the number of clusters grow, the MF model gets stuck in local optima and thus its performance suffers. It is important to note that in the absence of semantic information (magenta lines) the structured prior model still performs well. Fig. 7 shows qualitative comparison of the labeling.

The semantic observations in SAAR can be used to attribute meaning to the learned clusters. One possible method of achieving this is by analyzing the learned semantic component distributions φ^S. The collection of high probability semantic observations under each cluster forms the inferred meaning of the cluster under the model. Fig. 8 shows the semantic component distribution, MF-3 and MF-4, for the ToyCity2 scene. By looking at the learned semantic distribution of MF-3, we can see that the blue cluster has high probability observations of types: sidewalk, roads and parking lots, thus justifying the attachment of the interpretation of the cluster as "ground". The labeling produced by MF-4 follows a similar pattern as MF-3. Moreover, the effect of adding one more cluster component can be clearly seen in the figure: the "ground" cluster in MF-3 (blue) is divided into two clusters in MF-4 (blue and black). Each of the new clusters now has a more distinct meaning: the blue cluster is solely "green space" while the black cluster is the "road network". We note that this behavior arises naturally from the model and no special conditions were used to produce this result.

4.4. Real-World Scenes

SAAR was also tested on real-world scenes. Lacking ground truth, we provide qualitative comparisons. The results of this section are based on the CLIF 2007 dataset [32]

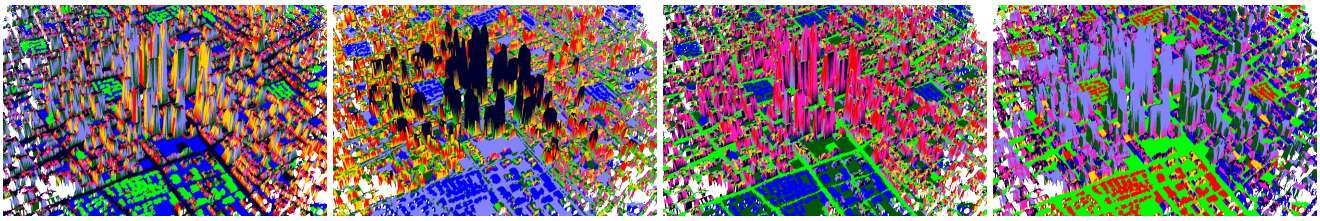

Figure 7: Learned categories for SynthCity City3 using TG-8 model. Colors represent cluster assignment. *Left-to-Right:* appearance+TG, appearance+TG+location, appearance+TG+semantic, appearance+TG+location+semantic.

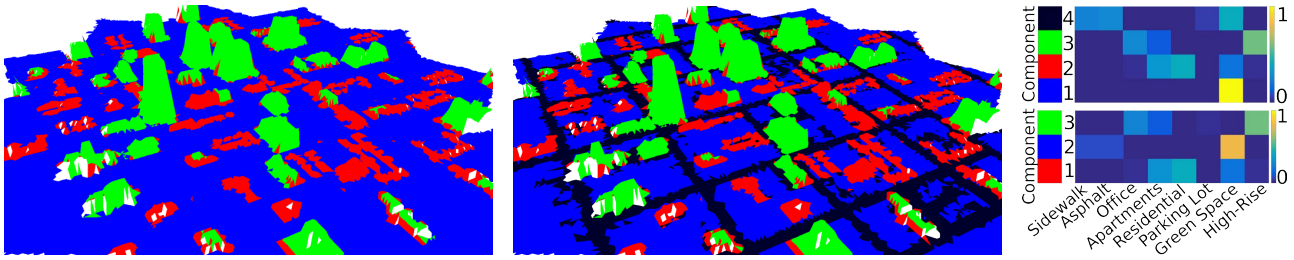

Figure 8: ToyCity2 clusters. *L-R:* view of MF-3 and MF-4 (colors indicate cluster assignment); semantic mixture components φ^S. Increasing N_s by one causes the ground cluster (blue MF-3) to split into ground and roads (blue and black MF-4).

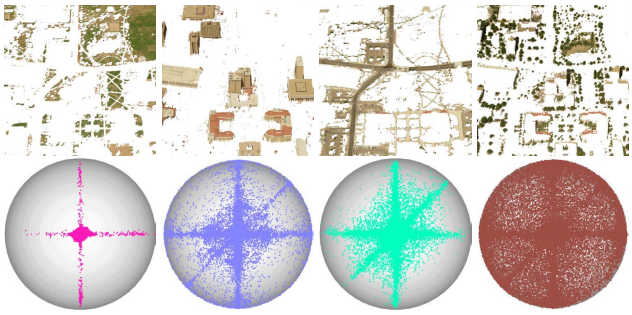

Figure 9: SAAR MF-4 model. *Top:* Image pixels corresponding to learned clusters (ground, buildings, roads, trees). *Bottom:* cluster orientations (spheres oriented so that scene up direction points out of the page).

and the Lubbock dataset. For each of the datasets, semantic information was obtained from OSM [36]. Both datasets contain semantic categories: road, building, parking lot, water, recreational and ground; additionally CLIF contains rails, while Lubbock contains grass (see Sup. Mat.). Reconstructions for these datasets are shown in Figs. 10 and 11. Qualitatively these reconstructions have more detail than the ones shown in [4]. Note that unlike prior work, SAAR clusters orientations in the correct space, *i.e.*, the 3D unit-sphere. This clustering is visualized in Fig. 9 for the Lubbock dataset using MF-4 model. As the figure shows, orientations are indicative of scene elements; *e.g.*, "trees", do not have any preferred orientation. On the other hand, "ground" and "building" clusters have very compact distributions centered around the scene's up direction.

4.5. Handling Missing Data

A main advantage of a probabilistic formulation is the ability to easily handle noisy and missing data. Here we show SAAR's ability to infer cluster assignment of partially-observed data and predict a missing modality. Specifically, we learn primitive assignments and cluster parameters using the scene's visible data (visibility refers to triangles that have image evidence). The learned cluster parameters are then used to predict the assignment of non-visible scene primitives using their location, orientation and semantic observations. Once the non-visible primitives are assigned to a cluster, we can predict their appearance by sampling from the corresponding appearance component posterior. The results of applying this procedure to City3 and CLIF scenes are shown in Fig. 10. The smooth boundaries between visible and non-visible assignment indicate that the model is able to infer partial assignments well. An exception is the river in CLIF, where the visible evidence dominate the semantic evidence for visible primitives. Due to the small number of clusters used, four, the predicted appearance captures only the main color trend.

4.6. Timing

The structured prior inference (Alg. 1) is relatively fast: approximately 0.02s per plane per iteration for TG models and 0.5s for MF models. This yields an overall run-time between 5-30 min and 30-60 min for TG and MF models in SynthCity respectively (workstation: 48 cores at 2.6GHz with NVIDIA GTX 780 graphics card). Full model inference (Alg. 2) is considerably slower due to the high com-

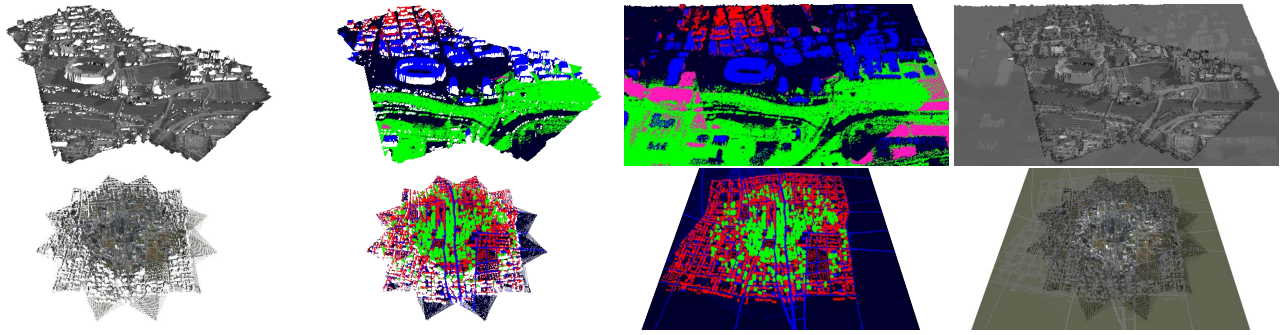

Figure 10: Missing data reasoning via SAAR (MF-4). *Top:* CLIF; *Bottom:* City3. *Left-to-right:* visible scene primitives color coded according to appearance and cluster assignment, filled in cluster assignment and appearance for all primitives.

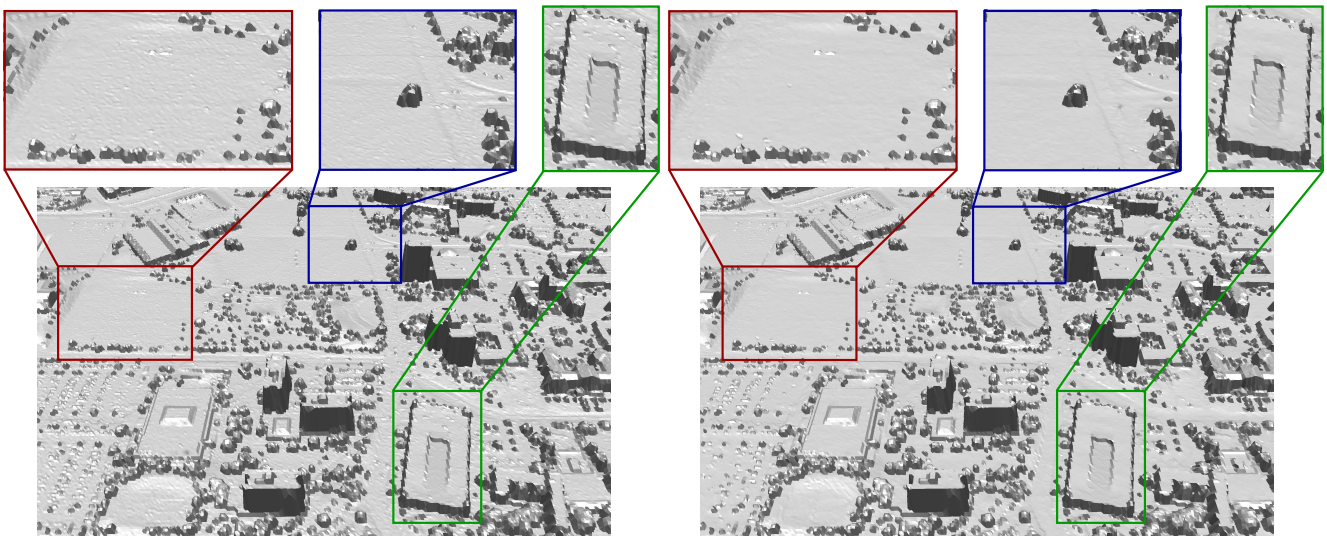

Figure 11: Lubbock reconstructions. *Left:* Cabezas *et al.* [4]; *Right:* SAAR (MF-4). Notice the regularized flat horizontal surfaces obtained with SAAR.

putational costs of the geometry updates in [4]. In our non-optimized implementation, the effect of using the structured prior, *i.e.*, evaluating Eqs. (9) and (10), is to increase runtime by 1-3 hours. The total runtime of a geometry update in SynthCity is between 2-20 hours (workstation: 24 cores at 2.3GHz with NVIDIA GTX Titan graphics card).

5. Conclusion

We propose SAAR, a probabilistic generative model for inferring semantically-consistent aerial reconstructions from multi-modal data. Using the novel SynthCity dataset, we have demonstrated that SAAR improves both reconstructions qualitatively and quantitatively by incorporating semantic data and utilizing a structured prior over geometry, appearance and semantic labels. Furthermore, by virtue of the generative model construction and robust inference algorithm, noisy or missing data does not hinder the model's ability. The latter was demonstrated on two different real-world datasets. The proposed model offers a mathematically-consistent framework for integrating both semantic and sensed data to generate richer scene reconstructions. Recent efforts in extending OpenStreetMap data into the third dimension [26] can benefit from approaches similar to the proposed one. Important extensions to the work presented here include the addition of spatial connectivity to the primitive's labels; investigations of other modalities to include for better cluster; as well as, more fine-grained categorization capabilities. All source code as well as the novel multi-modal SynthCity dataset can be downloaded from http://people.csail.mit.edu/rcabezas.

Acknowledgments. The authors thank Sue Zheng, Christopher Dean, and Oren Freifeld for general and helpful discussions. This research was partially supported by the Office of Naval Research (ONR) MURI program (N000141110688) and by VITALITE, which receives support from Army Research Office (ARO) MURI (W911NF-11-1-0391).

References

[1] S. Y. Bao, M. Bagra, Y. W. Chao, and S. Savarese. Semantic structure from motion with points, regions, and objects. In *Computer Vision and Pattern Recognition*, 2012.

[2] S. Y. Bao, M. Chandraker, Y. Lin, and S. Savarese. Dense object reconstruction with semantic priors. In *Computer Vision and Pattern Recognition*, 2013.

[3] S. Y. Bao and S. Savarese. Semantic structure from motion. In *Computer Vision and Pattern Recognition*, 2011.

[4] R. Cabezas, O. Freifeld, G. Rosman, and J. W. Fisher III. Aerial Reconstructions via Probabilistic Data Fusion. In *Computer Vision and Pattern Recognition*, 2014.

[5] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *ICCV*, 2007.

[6] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *Computer Vision and Pattern Recognition*, 2010.

[7] P. Cignoni, C. Rocchini, and R. Scopigno. Metro: measuring error on simplified surfaces. In *Computer Graphics Forum*. Wiley Online Library, 1998.

[8] L. Du, L. Ren, D. Dunson, and L. Carin. A Bayesian model for simultaneous image clustering, annotation and object segmentation. *Advances in Neural Information Processing Systems*, 2009.

[9] Esri. CityEngine. http://www.esri.com/software/cityengine.

[10] N. Fioraio and L. Di Stefano. Joint detection, tracking and mapping by semantic bundle adjustment. In *Computer Vision and Pattern Recognition*, 2013.

[11] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2014.

[12] M. Haklay and P. Weber. Openstreetmap: User-generated street maps. In *Pervasive Computing*, 7(4):12–18, 2008.

[13] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3D scene reconstruction and class segmentation. In *Computer Vision and Pattern Recognition*, 2013.

[14] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments, 2012.

[15] M. E. Hodgson and P. Bresnahan. Accuracy of Airborne Lidar-Derived Elevation : Empirical Assessment and Error Budget. In *Photogrammetric Engineering Remote Sensing*, 2004.

[16] B.-S. Kim, P. Kohli, and S. Savarese. 3D Scene Understanding by Voxel-CRF. In *International Conference on Computer Vision*, 2013.

[17] A. Kundu, Y. Li, F. Dellaert, F. Li, and J. M. Rehg. Joint Semantic Segmentation and 3D Reconstruction from Monocular Video. In *European Conference on Computer Vision*, 2014.

[18] F. Lafarge, C. Mallet. Creating large-scale city models from 3D-point clouds: a robust approach with hybrid representation. In *IJCV*, 2012.

[19] L. J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Computer Vision and Pattern Recognition Workshops*, 2009.

[20] L. J. Li, C. Wang, Y. Lim, D. M. Blei, and L. Fei-Fei. Building and using a Semantivisual image hierarchy. In *Conference on Computer Vision and Pattern Recognition*, 2010.

[21] D. Lin, S. Fidler, and R. Urtasun. Holistic scene understanding for 3D object detection with RGBD cameras. In *International Conference on Computer Vision*, 2013.

[22] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *Computer Vision and Pattern Recognition*, 2014.

[23] J.A. Nelder, R. Mead. A simplex method for function minimization. In *The Computer Journal*, 1965.

[24] Z. Niu, G. Hua, X. Gao, and Q. Tian. Spatial-DiscLDA for visual recognition. In *Computer Vision and Pattern Recognition*, 2011.

[25] Z. Niu, G. Hua, X. Gao, and Q. Tian. Context Aware Topic Model for Scene Recognition, 2012.

[26] M. Over, A. Schilling, S. Neubauer, and A. Zipf. Generating web-based 3D city models from OpenStreetMap: The current situation in Germany. *Computers, Environment and Urban Systems*, 34(6):496–507, 2010.

[27] X. Ren, L. Bo, and D. Fox. RGB-(D) scene labeling: Features and algorithms. In *Computer Vision and Pattern Recognition*, 2012.

[28] S. Sengupta, P. Sturgess. Semantic octree: Unifying recognition, reconstruction and representation via an octree constrained higher order MRF. In *ICRA*, 2015.

[29] J. Straub, J. Chang, O. Freifeld, and J. W. Fisher III. A Dirichlet process mixture model for spherical data. In *Artificial Intelligence and Statistics*, 2015.

[30] J. Straub, G. Rosman, O. Freifeld, J. J. Leonard, and J. W. Fisher III. A mixture of Manhattan frames: Beyond the Manhattan world. In *CVPR*, 2014.

[31] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, 2005.

[32] US Air Force. Columbus Large Image Format Dataset 2007. https://www.sdms.afrl.af.mil/index.php?collection=clif2007.

[33] J. P. C. Valentin, S. Sengupta, J. Warrell, A. Shahrokni, and P. H. S. Torr. Mesh based semantic modelling for indoor and outdoor scenes. In *CVPR*, 2013.

[34] V. Vineet, O. Miksik, M. Lidegaard, M. Nießner, S. Golodetz, V. A. Prisacariu, O. Kähler, D. W. Murray, S. Izadi, P. Perez, P. H. S. Torr. Incremental Dense Semantic Stereo Fusion for Large-Scale Semantic Scene Reconstruction. In *ICRA*, 2015.

[35] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition Workshops*, 2009.

[36] O. Wiki. Openstreetmap wiki, 2014. [Online; accessed 18-April-2015].

[37] J. Xiao, A. Owens, and A. Torralba. SUN3D: A database of big spaces reconstructed using SfM and object labels. In *International Conference on Computer Vision*, 2013.

[38] O. Yakhnenko. Multi-modal hierarchical Dirichlet process model for predicting image annotation and image-object label correspondence. *SIAM Conference on Data Mining*, 2009.

[39] Q. Y. Zhou, U. Neumann. 2.5D Dual Contouring: A Robust Approach to Creating Building Models from Aerial LiDAR Point Clouds. In *European Conference on Computer Vision*, 2010.