# On the Role of Representations for Reasoning in Large-Scale Urban Scenes

Randi Cabezas[†,1] Maroš Bláha[†,2] Sue Zheng[1] Guy Rosman[1]
Konrad Schindler[2] John W. Fisher III[1]
[1] Massachusetts Institute of Technology [2] ETH Zurich

## Abstract

*The advent of widely available photo collections covering broad geographic areas has spurred significant advances in large-scale urban scene modeling. While much emphasis has been placed on reconstruction and visualization, the utility of such models extends well beyond. Specifically, these models should support a wide variety of reasoning tasks (or queries), and thus enable advanced scene study. Driven by this interest, we analyze 3D representations for their utility to perform queries. Since representations as well as queries are highly heterogeneous, we build on a categorization that serves as a coupling interface between both domains. Equipped with our taxonomy and the notion of uncertainty in the representation, we quantify the utility of representations for solving three archetypal reasoning tasks in terms of accuracy, uncertainty and computational complexity. We provide an empirical analysis of these intertwined realms on challenging real and synthetic urban scenes and show how uncertainty propagates from representations to query answers.*

## 1. Introduction

Many applications, such as 3D scene processing, augmented reality and *autonomous agents* can benefit from having *sufficiently* detailed models of large-scale urban areas, but only if they can reason within those models. Currently, reasoning tasks are mostly treated as separate and independent of one another. Moreover, different representations are often constructed for the purpose of solving separate reasoning tasks [22, 36, 37, 70, 74, 76]. While the use of separate representations for each task may be optimal (in some sense), it is often prohibitive for autonomous agents due to constraints on sensing, computation, and storage. Such resource constraints motivate the use of a reduced set of representations for solving multiple reasoning tasks.

3D scenes can be represented in various ways including depth-maps, voxels and more. Widely utilized in computer vision and related fields, intermediate representations bridge the gap between sensing and reasoning. However, different representations make fundamentally different assumptions about what is important in the world. These assumptions determine which information is maintained or discarded from the representation. In this regard, a representation serves as an approximate *sufficient statistic* of the *data* for a given reasoning task. To quantify this notion we expand a taxonomy over the space of representations, demonstrating its utility for coupling reasoning and representation. We argue that how a representation quantifies uncertainty is crucial to understanding how uncertainty propagates from imperfect sensing to reasoning tasks.

Reasoning is the process by which we infer what is of interest about a scene from what is represented. We are

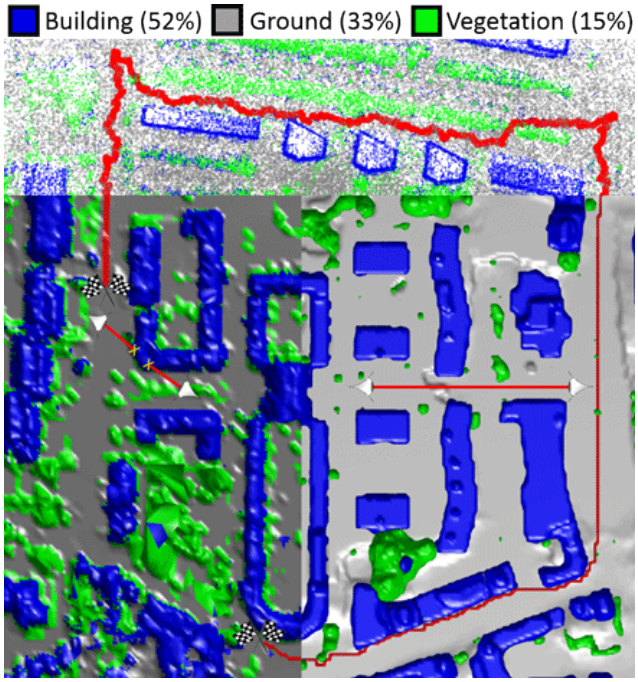



Figure 1: Exemplar for three archetypal reasoning tasks on three different representations: point cloud *(top)*, mesh *(left)* and voxels *(right)*. Queries: (1) path planning; (2) ratio of the semantic categories; (3) clear-line-of-sight.

[†] shared first authorship

interested in *abductive reasoning*, *i.e.*, obtaining quantifiable statements about the world from a set of observations. To that end, we focus on the class of reasoning tasks that can be reduced to *functions* operating on a given representation, referring to such task as *queries*. If a representation is not suitable for this purpose, one may need to simplify the meaning of the original query and/or augment the representation itself. We generally do the latter. This highlights one aspect of the interaction between a query and the representation. We show that the proposed taxonomy facilitates this process. Additionally, we consider the means by which a representation propagates sensing uncertainty to query uncertainty. For concreteness, we focus on three archetypal reasoning tasks (Fig. 1). These serve as representative examples from a wide range of choices and help illustrate tradeoffs between different representations.

Our contributions are: (1) expand a taxonomy of 3D scene representations for the goal of performing abductive reasoning combined with empirical demonstrations; (2) use archetypal queries to compare different representations in terms of accuracy, uncertainty and computational complexity; (3) highlight the importance of propagating sensing uncertainty to query uncertainty via the 3D representation.

## 2. Related Work

Reasoning about a scene encompasses tasks across vision, AI and numerous other fields. The tasks are too many to list; thus, we focus on a relevant subset. We also review relevant methods for 3D reconstruction and representation.

**Reasoning.** Many systems have been proposed to do scene reasoning over the years, with varied definitions of scene (*e.g.*, images, reconstructions) and reasoning (*e.g.*, segmentation, classification, higher-level tasks). A classical example of an early reasoning system was presented in [20]; it proposes a knowledge-based approach relying on features to perform various image processing tasks. The MESSIE system is closer in spirit to our approach [29]. It relies on processing via "specialists" (task-specific algorithms) to perform detection and segmentation. In various extensions, the system was augmented with spatial constraints [28], uncertainty handling [13], and 3D support [56, 57, 58]. Recent works explore higher complexity in 2D reasoning tasks, *e.g.*, finer segmentation or categorization [39]; expanded set of actions and events of interest [61]; or answering free-form questions [1]. In the 3D processing realm, reasoning has expanded even more rapidly, with one of the main goals being interaction between humans (or agents) and the environment [31, 42, 76]. Scene reasoning has also been studied in the AI field to create cognitive models [3] and for situational awareness [21]. These works differ from ours as these communities focus on *human* reasoning, *i.e.*, cognitive processing. In addition, the GIS community often performs reasoning based on geospatial data [53].

**3D Reconstructions.** A reconstruction is the process by which a 3D representation of the world is computed. Contemporary advances in multi-view reconstruction - including increases in model size [66], accuracy at millimeter scale [69], robust [25, 63] and efficient processing algorithms [54, 73] - have pushed the limits of urban models. Furthermore, research has recently focused on incorporating attributes into traditionally pure geometric models. In this context, pioneering *semantic reconstruction* methods [6, 12, 32, 40, 44, 45, 68] emerged with the goal of augmenting the representation with the type of scene elements. These works show how to learn categories and demonstrate that such knowledge improves geometric fidelity.

**3D Representations.** The impact of a representation on reasoning tasks has been studied in the AI field. In [16] *Davis et al*. axiomatically outlined the key properties of a *knowledge representation* as follows: (1) it is a *surrogate* for the world, *i.e.*, a substitute that cannot contain all information; (2) it biases our view of the world; (3) it encompasses a set of supported inferences about the world; (4) it organizes information for efficient computation of inferences. In [46], *Lafarge* independently proposed a list of criteria for characterizing 3D reconstructions congruent with the above properties. It includes geometric accuracy – corresponding to (1) from above; representation complexity – (4); regularity of output – (2); visual appearance – (2); and level of automation – (4). The only gap in this list is point (3), which we will cover in this work by qualifying the feasibility of tasks for specific representations. Additionally, we clarify point (2) by providing a criteria for describing how 3D representations bias our world view.

## 3. Representations

The set of commonly used 3D representations (*e.g.*, voxels, depthmaps, etc.) make fundamentally different assumptions about the world. These assumptions determine which characteristics of the world are maintained or discarded from the representation. These choices impact the degree to which a representation *approximates* a sufficient statistic, *i.e.*, how closely the distribution of the query conditioned on the representation approximates the distribution of the query conditioned on the world. Formally, let $Q$, $\mathbf{X}$ and $\mathbf{W}$ be random variables denoting the query, the representation and the world respectively; then $\mathbf{X}$ is an approximate sufficient statistic if $p(Q|\mathbf{X}, \mathbf{W}) \approx p(Q|\mathbf{X})$. The approximation arises from missing subtleties of the world in the representation due to, *e.g.*, sensing limitations or computation limits when constructing or using the representation.

Often, a user of a representation may not know the quality of the estimated representation parameters which depends on data and reconstruction methods. Thus, a repre-

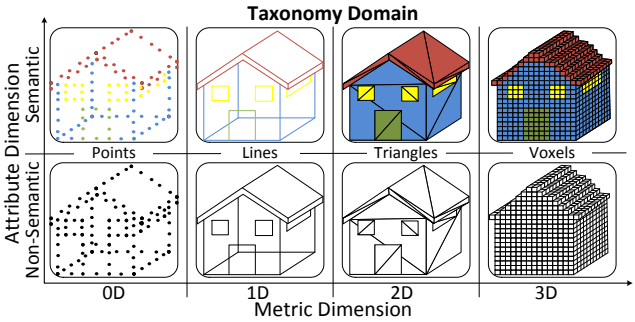



Figure 2: Representation taxonomy combining metric and attribute (*i.e.*, semantic categorization) components.

sentations should also quantify parameter uncertainty. That is, a representation should report not only what it "knows", but how well it knows it. To that end, we treat representations as a collection of $n$ random variables, jointly denoted by $\mathbf{X} = \{X_i\}_{i=1}^{n}$, which have a distribution, $p_X(\mathbf{X})$, over possible parameter values.

### 3.1. Representation Taxonomy

We rely on a taxonomy of various 3D representations to better understand and quantify their effect on a reasoning task. The taxonomy contains two complementary characteristics: *metric* and *attribute*. The metric characteristic can be broken down into the inherent dimension of the representation element. Expanding on prior work, [59, §2.5], a valid breakdown can be 0D, 1D, 2D, or 3D. Points, lines, polygons, and polyhedra are the canonical examples for each dimension. The highest metric property that can be measured (*i.e.*, mensuration [64]): visual, length, area, or volume, is determined by the inherent dimension of the element. We label points as visual employing their geometric definition as a demarcation of a location since they are devoid of any metric information. Note that while different choices of connectivity between elements yield separate representations (*e.g.*, polygonal soup vs. polygonal mesh), these still fall within the same category of the taxonomy. Furthermore, we observe that connectivity must occur using elements of one dimension less (*e.g.*, lines connect polygons).

Representations can have a variety of attribute characteristics such as appearance, material properties or semantic categorization. Each characteristic provides an additional axis for differentiation. When combined, the metric and attribute components yield a taxonomy of the representations; see Fig. 2 for the case of semantic labels in the attribute direction. We will use the notation "nD-X" (*e.g.*, 0D-semantic) to refer to a specific taxonomy group.

A few observations can be made from the taxonomy outlined above. Higher-dimension representations retain aspects of the data that are discarded by lower-dimension representations. This implies that we can discard the additional information to convert a higher-dimension representation to a lower-dimension one (*e.g.*, marching cubes [48] for transforming voxels into a mesh). On the other hand, converting a lower-dimensional representation to higher-dimension often entails incorporating scene aspects in a manner that is independent of the measurement process, as in the methodology of [24] (*i.e.*, by employing external assumptions). The nature of the assumptions depends on the specific conversion while the accuracy of the resultant representation depends on how well the assumptions match the world (*e.g.*, the accuracy of a 2D mesh obtained from a set of 0D points is highly dependent on the density of the original points relative to the complexity of the underlying surface).

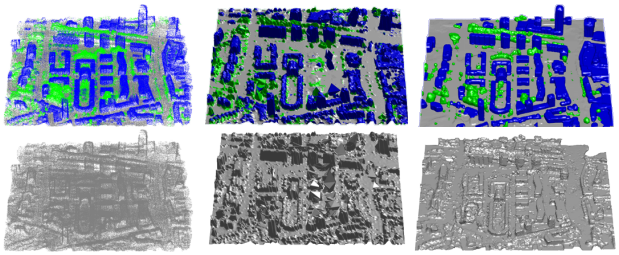
### 3.2. Exemplar Representations

Motivated by the taxonomy of §3.1, we consider three representations: point cloud (PC), triangular mesh (mesh), and voxel. These are canonical examples of 0D, 2D and 3D metric representations. We omit 1D representations as they are rarely used for large-scale reconstructions. For each metric class we consider attribute class semantic and non-semantic resulting in six different representation (Fig. 3a).

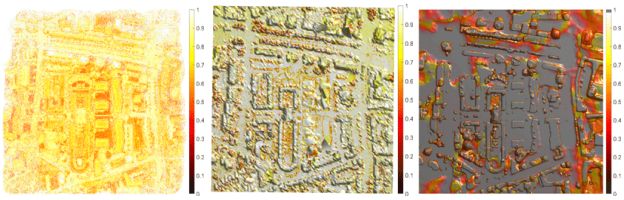
We now turn to the question of how to build the representations. As stated earlier, representations are models of the world and reconstruction algorithms are the *inference procedures* by which the models are estimated. Here we focus on a single prototypical reconstruction algorithm for each representation with the assumption that these are illustrative of all others. In practice, this assumption is limiting as there are many algorithms that can be used to infer a representation. Each method relying on potentially different assumptions, *e.g.*, input or pre-processing, use of regularizers, optimization, *etc*. This wide array of choices inevitably impacts the quality of the representation, however, if algorithms are consistent in their representation of uncertainty the above assumption approximately holds. We emphasize that any reconstruction method, *e.g.*, [4, 10, 17, 27, 63, 66, 67, 69], can be used in place of the ones described here.

To build the models in this work, we rely solely on aerial imagery. We have chosen a coarse compilation of semantic labels comprising mutually exclusive classes: building, ground, and vegetation (voxel representations explicitly model the additional class free-space). An important consideration is how each model quantifies uncertainty, Fig. 3b.

We build the PC using VisualSFM [72] and densify it with plane sweep. To obtain the semantics for this representation, we utilize a supervised multi-class version of Adaboost [8] with image and geometry features. Location uncertainty in the PC is modeled by an isotropic Gaussian distribution (a simplified version of the uncertainty regions of reconstructed points in MVS from well-distributed cameras [18, 33]). Semantic uncertainty arises from the classifier's per-class probability distribution. The non-semantic mesh representation is based on [11], while the semantic



(a) Semantic and non-semantic models. *L-R*: PC, mesh, and voxels.



(b) Semantic uncertainty. Color shows the most likely class probability; higher values imply more certainty. *L-R*: PC, mesh, voxels.

Figure 3: Enschede scene models and semantic uncertainty.

mesh is based on [12]. Both models capture geometric uncertainty as ambiguity in the latent positions of the mesh vertices. The semantic mesh learns a mixture of semantic labels for each triangle. The non-semantic voxel model is obtained using [75]; this leads to a partitioning of the space into two classes (free-space or solid) with equally-sized voxels. Geometry uncertainty is implicit in each voxel's Bernoulli distribution (we sample this distribution instead of thresholding at 0.5 as in the original work). The semantic voxel representation is generated using [32]; conceptually, it extends the non-semantic voxel model by subdividing the solid space into specific semantic categories. For simplicity, we assume that all elements in the representations are independent though this is not the case for shared vertices in the mesh nor for attributes in the voxel representation.

# 4. Reasoning

Reasoning is the process by which we infer what is of interest about a scene from what is represented. As stated earlier, a query encapsulates reasoning as a computable operation over a representation. As such, the plain statement of a query must be interpreted in the context of what is computable over the representation. Here we focus on deterministic functions, $f(\cdot)$ with parameters $\theta$ that operate on the representation $\mathbf{X}$; the query result $Q$ is given by $Q = f(\mathbf{X}; \theta)$. Consistent with [16], queries and representations are intertwined; the latter is necessary to formulate the computation of the former. For example the query "What is the percentage of vegetation in the scene?" intends to quantify the vegetation in the scene. We can establish that a representation needs semantics to identify vegetation and, at minimum, 0D-metric information to quantify the amount although any other metric (*e.g.*, area or volume) could also

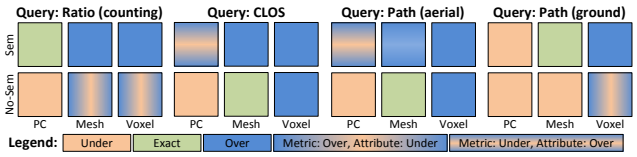



Figure 4: Query details in representation's taxonomy.

be used provided it is computable over the elements of the representation. We term queries that have exactly one mapping for performing the task on a representation as *exactly-specified* (*e.g.*, the earlier example in a 0D-semantic representation). Queries with multiple mappings are termed *over-specified*, *e.g.*, 2D-semantic representations can use counts or area. Lastly, queries that require information not included in the representation are termed *under-specified*, *e.g.* "What is the volume percentage of vegetation in the scene?" in a 0D-semantic representation. This last example necessitates computing volume which requires 3D-metric knowledge or lifting lower metric dimensions via some assumptions for the calculation.

This classification quickly informs us of the relation between reasoning and representations and the potential ambiguities or need for additional information. Given that higher-dimension representations contain aspects beyond and including those of lower-dimension representations inevitably leads one to conclude that they can address a richer set of queries. We will refer to the representations as under-/exactly-/over-determined for a specific query.

## 4.1. Exemplar Queries

For concreteness, we focus on three archetypal reasoning tasks: clear-line-of-sight (CLOS), scene element category analysis, and path planning. CLOS is a computation between two points, an element of more complex reasoning tasks such as surveillance and electromagnetic wave propagation, relying on only the geometry of a local region. Scene category analysis is a crucial element of urban planning analysis and path planning is an element of navigation and mobility analysis. Both queries rely on global scene information. These queries are representative of a large subset of reasoning tasks in urban reconstructions [51]. Through them we will explore tradeoffs between different representations and queries. We now describe each query and discuss representation specific details (*c.f.* Fig. 4). Importantly we focus on simple query formulations to stress the difference in properties across representations.

**Clear Line of Sight.** We state the clear-line-of-sight (CLOS) query as "Is there a clear line of sight between points A and B?" In its simplest form, it is a geometric computation over a surface (more complicated methods can be extended to perform this task while relaxing the surface requirement, *e.g.*, [38, 49]). The query formulation requires surface-level non-semantic information, mak-

ing 2D-non-semantic exactly-determined. Representations with additional information (*e.g.*, semantic or higher metric-dimension) are over-determined, while those with fewer are under-determined. Consequently, we augment the PC representation by attributing a notional volume (consistent with the footprint of a pixel) to each point, resulting in a solid spherical element. Although other assumptions could be used (*e.g.*, oriented patches as in [25]), we use spheres for computational simplicity.

In the absence of uncertainty, the computation is straightforward: check if any representation element intersects the line segment connecting A and B, if there are none then there is a clear line of sight. The problem of segment intersection with triangles, voxels and spheres is well known and mature algorithms exist. We use a variant of [50] for triangle intersections, a 3D implementation of Bresenham's algorithm [9] for voxels and follow [34] for spheres. If we assume segment endpoints are chosen uniformly at random, the computational costs associated with these algorithms grow as $\mathcal{O}(n)$, $\mathcal{O}(n^{\frac{1}{3}})$, and $\mathcal{O}(n)$ respectively, where $n$ is the number of elements. Importantly, $n$ is not the same across representations and may grow at different rates for increasing scene complexity.

**Category Analysis.** The category analysis query computes the *relative* quantity of different categories in the scene, *e.g.*, "What is the percentage of category A in the scene?" This particular query illustrates how representations influence the interpretation of a reasoning task since the plain statement is open to wide interpretation. Each yielding multiple formulations for each representation (*e.g.*, counting, computing area or volume). For simplicity, we use the counting metric for all representations; this metric has computation complexity $\mathcal{O}(n)$ and has the benefit of being a linear function, *e.g.* $Q = \frac{1}{n}\sum_i \mathbb{1}[x_i = c]$, where $c$ is the class of interest. In this case, the 0D-semantic representation is exactly-determined for this query. All other semantic representations are over-determined while the non-semantic ones are under-determined and require a proxy for the semantic categories. We use simple heuristics to approximate some classes; *e.g.*, ground and buildings are labeled by thresholding vertical position and surface normal directions.

We note that in our evaluation we explore variations in metric-choice (*i.e.*, computing area or volume) to highlight the distinctions between query formulation over different representations. These choices bias results in a manner specific to the representation and to the underlying assumptions used to lift the representation to compute the query.

**Path Planning.** This query is stated as "Can an agent get from point A to B?", where agents are ground-bound (*e.g.*, pedestrian) or aerial (*e.g.*, UAVs). This query is interesting in that the formulation depends on properties of the both the representation *and* the agent. The query identifies if a viable path exists, as such any path planning algorithm may be used in the query formulation. Both ground and aerial agents require at least 2D-metric information. In addition, ground-bound agents require semantic information to determine the feasibility of an element (*e.g.*, sidewalk or road), as such non-semantic representations require augmentation.

For the case where the path traverses elements of the representation, *i.e.*, all voxel paths and ground-based paths in mesh and point cloud, the computation additionally requires connectivity to determine feasibility. The notion of neighbors is well-defined in mesh and voxel representations. Connectivity is assumed in the point cloud by defining a set of closest $N$ points as neighbors. Given connectivity, the computation can be made via shortest path algorithms [19, 41, 60]. In our implementation we use [19], which has computational complexity of $\mathcal{O}(n\log(n))$, where $n$ is the number of elements in the representation.

The case of aerial paths in the mesh and point cloud representations is more involved as the traversable element is not directly represented. This merits some form of configuration-space planning [65]; we choose Rapidly-exploring Random Trees (RRTs) [47]) as an example of such approaches. For this method the mesh must be augmented with bounds encoding the extent of the representation while the point cloud must be augmented with the concept of obstacles, *i.e.*, we again attribute the points with volume to form non-traversable space. The convergence rate of the RRT algorithm to a valid path is difficult to characterize as it depends on the structure of the obstacles [14]. As such, we limit the number of iterations to $T_{max}$ and arrive at a computational cost of $\mathcal{O}(nT_{max}^2)$.

### 4.2. Query Computation and Uncertainty

We now outline the query computation over uncertain representations. Recall that we treat representations as a collection of random variables, jointly denoted by $\mathbf{X}$, with probability density function $p_X(\mathbf{X})$. Queries are deterministic functions, $f(\cdot)$, that operate on the representations with parameters $\theta$; a query answer $Q$ is then expressed as $Q = f(\mathbf{X};\theta)$. We are interested in both query result and its associated uncertainty (which is a direct consequence of the uncertainty in the representation). We can obtain this distribution by performing a change of variables and integrating out nuisance variables [52, Eq. 8-8]:

$$p_Q(Q) = \int p_x(f^{-1}(\mathbf{Y}), \mathbf{X}_{\setminus 1}) \left| \frac{\partial f^{-1}}{\partial Q}(\mathbf{Y}) \right| d\mathbf{X}_{\setminus 1} \,, \quad (1)$$

where $f^{-1}(\cdot)$ is the inverse of $f(\cdot)$, $\mathbf{Y} = \{Q, \mathbf{X}_{\setminus 1}\}$ and $\mathbf{X}_{\setminus i}$ is used to remove element $i$ from the set $\mathbf{X}$. However, Eq. (1) is often intractable or, as $f(\cdot)$ may not have a unique inverse. Entropy is one measure by which we may quantify the uncertainty of the query result. We can bound the entropy of $Q$ in relation to the entropy of $\mathbf{X}$:

$$H(Q) \leq H(\mathbf{X}) - H(\mathbf{X}_{\setminus 1}|Q) + \mathbb{E}\left[\left|\frac{\partial f}{\partial X_1}(\mathbf{X})\right|\right] , \quad (2)$$

where equality holds iff $f(\cdot)$ is one-to-one [52, Eq. 15-113]. However, Eq. (2) has the same challenges as Eq. (1), with the added drawback of being a very loose bound.

While Eq. (1) and (2) are useful in that they directly relate representation uncertainty to query uncertainty, we cannot compute them for all queries. Alternatively, we may estimate the quantities using Monte-Carlo sampling techniques [55]. The approach relies on sampling $M$ representation instances $\tilde{\mathbf{x}} = \{\mathbf{x}^m\}_{m=1}^M$ from $p_X(\mathbf{X})$ where the method of sampling is specific to the representation, *e.g.* sampling locations for the point cloud representation. Evaluation of the query function on these samples yields samples of the query result; *i.e.*, $\tilde{\mathbf{q}} = \{q^m\}_{m=1}^M$, where $q^m = f(\mathbf{x}^m; \theta)$. We can then use the query answer samples to estimate the distribution (*e.g.*, using kernel density estimators [5]), entropy (*e.g.*, using [7] or [2]) or compute statistics of interest such as mean and variance:

$$\mu_Q \approx \frac{1}{M}\sum_{m=1}^{m} q^m \quad \text{and} \quad \sigma_Q^2 \approx \frac{1}{M^2}\sum_{m=1}^{M}[q^m - \mu_Q]^2 \ . \quad (3)$$

**Reducing Query Uncertainty.** An important property of the query computation presented here is that it directly relates uncertainty in representation to uncertainty in the query. In principle, this link can be exploited to reduce query uncertainty, *i.e.*, for discrete queries, the following relationship holds $H(Q) = H(X_i|\mathbf{X}_{\setminus i}) - H(X_i|\mathbf{X}_{\setminus i}, Q) + H(\mathbf{X}_{\setminus i}) - H(\mathbf{X}_{\setminus i}|Q)$ where $X_i$ is any single representation variable which the query depends upon. The first two terms in the RHS capture the *influence* of $X_i$ on the query uncertainty and is commonly known as the mutual information $I(X_i; Q|\mathbf{X}_{\setminus i}) \triangleq H(X_i|\mathbf{X}_{\setminus i}) - H(X_i|\mathbf{X}_{\setminus i}, Q)$. The mutual information expression provides a mechanism for identifying the representation element driving the uncertainty in the query answer. Once key representation elements have been identified, we can use information planning techniques [30, 71] to select additional measurements to further refine that representation element and consequently reduce uncertainty in the query answer.

**Special Cases.** A few special cases appear in our analysis. First, if the query only depends on a subset of representation elements, Eq. (2) simplifies to depend only on that subset. This result has the ability to simplify local queries and reduce the computational complexity associated with them. Second, linear queries, which take the form $Q = \sum_i^n \alpha_i X_i$, allow us to compute desired quantities in closed-form, *e.g.*, using [35, p. 134] we can compute the mean and covariance for the query answer distribution. If random variables $X_i$ are also independent (as in the case of this work) Eq. (1) simplifies to a convolution. This convolution has well known forms for most common distributions [35, ch. 3]. Of interest to us is the case when $X_i \sim$ Bernoulli$(p_i)$, (as in scene category analysis) then $Q$ follows a Poisson binomial distribution, $Q \sim$ PoissonBinomial$(p_1, \ldots, p_n)$ with mean $\mu_Q = \sum_{i=1}^n \alpha_i p_i$ and variance $\sigma_Q^2 = \sum_{i=1}^n \alpha_i^2 (1 - p_i) p_i$ [23].

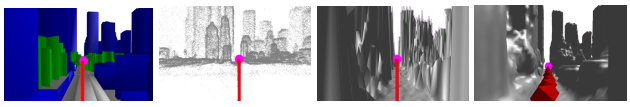


Figure 5: SynthCity CLOS. *L-R:* ground-truth, non-semantic PC, mesh and voxel (10 samples each). All clear.

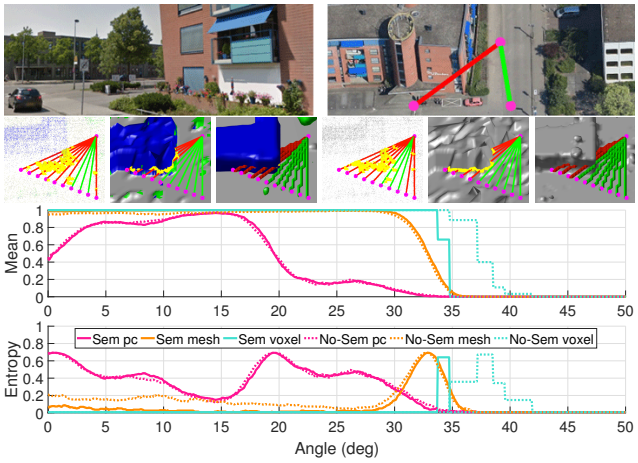



Figure 6: Enschede CLOS corner, target moves from visible to occluded. *Top:* views from start location and above. *Center:* representations with rays colored by mean answer and intersections in yellow. *Bottom:* mean and entropy curves.

## 5. Experiments

We present a set of experiments designed to assess the performance of the representations with regards to the queries outlined above. As testbed, we use representations of two scenes, a synthetic example, SynthCity [12] and a real data example of Enschede, Netherlands [62].

### 5.1. Clear Line of Sight

We begin CLOS with a simple example in SynthCity: a clear line along a road. The non-semantic models, Fig. 5, are able to obtain the correct answer with high certainty under a low number of samples. We expand on this by considering the scenario where a target moves from visible to occluded while the starting location remains fixed, Fig. 6. The figure shows that representations agree and are highly certain of their answers (*i.e.*, low entropy) when the target is clearly occluded or clearly visible. However, when the target approaches the estimated building boundary the representations disagree in their query answer and certainty levels. In this example, the PC has a large uncertainty region, leading to a low mean answer; mesh and voxel models have a much more compact uncertainty region and agree on where the target becomes occluded. The results of compar-

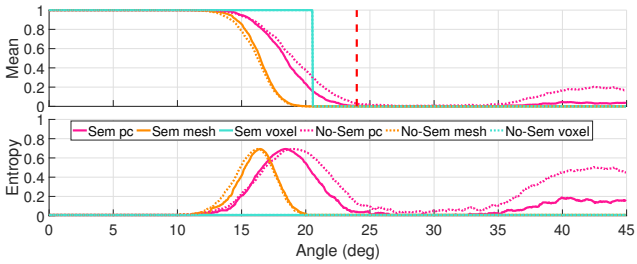



Figure 7: SynthCity CLOS corner mean and entropy curves as target moves from visible to occluded. Vertical dashed line denotes ground-truth building boundary.

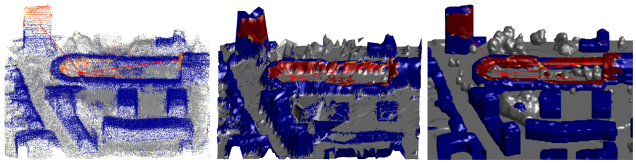


Figure 8: Which building points can see the target (red triangle)? Color shows mean answer, blue - not visible, red - fully visible. *L-R:* Enschede semantic PC, mesh and voxel.

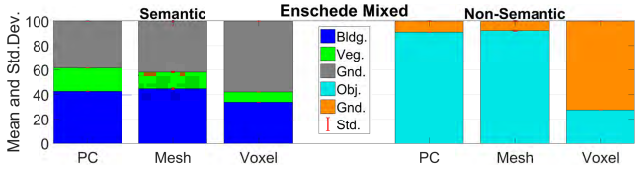



Figure 9: Matching formulation for each representation of the category analysis. Mean and standard deviations on Enschede (100 samples per representation).

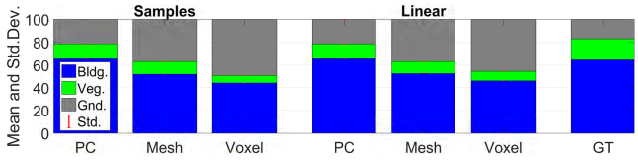



Figure 10: SynthCity category analysis query (counts), comparison of samples and exact linear computation.

ing semantic and non-semantic models are consistent. This is expected as semantics only indirectly influence CLOS – through their effect on the reconstruction process.

To identify biases in the representations' query answers, we repeat the experiment using synthetic data, Fig. 7. Overall, we see similar characteristics for areas far from the boundary. As the target nears the true boundary, the mean answer transitions prematurely for all representations. This collapsing of the representation has been shown to be a result of the use of regularizers [15, 26, 43]. As a final example, we show we can compose several CLOS queries to obtain a more complex task, *e.g.*, "Which scene elements can see point A?" Fig. 8 shows the visibility of a courtyard in Enschede for semantic models.

## 5.2. Scene Category Analysis

We begin the scene category analysis by exploring the interaction between query formulation and the capabilities of the representation, *e.g.*, we vary the query's metric quantification method from counts to area to volume.

For this task, under-determined representations require augmentation (and assumptions). The PCs are raised to a 2D-metric representation by triangulating the points in the ground then pushing the triangle vertices to their height; for 3D-metric, each point is augmented with volume. The mesh representation is lifted to 3D-metric by estimating a ground plane. In the absence of semantics, we estimate only two classes, "ground" and "object" by relying on simple heuristics (*i.e.*, height and surface orientation thresholds). The results for SynthCity can be seen in Fig. 11. For semantic representations, the best performance is achieved by the representation whose metric properties match the query formulation and progressively degrades as assumptions are incorporated. As expected, non-semantic representations perform worse than their semantic counterparts due to the query's high reliance on semantic information.

In order to remove any variability introduced by differences in reconstruction algorithms, we also discard additional information in over-determined representations and compare their performance in a lower formulation dimension. For example, the voxel bar in the area plot of Fig. 11 is obtained by doing marching cubes on the voxel representation and computing the area of the resulting mesh. By comparing the derived-mesh and the original voxel to ground-truth we see that better performance is obtained with matching original dimension (similar results hold for the mesh). This further indicates that it is the interplay between assumptions and formulation and not the difference in reconstruction methods that drives the changes in the figure.

Given the above results, Fig. 9 shows the matching formulation for each representation, *i.e.*, counting for point cloud, area for mesh and volume for voxels for Enschede. All representations agree that buildings compose roughly 40% of the scene, independent of the query formulation. Furthermore, we see that that voxels have a larger percentage for ground since they explicitly model the region below buildings. Lastly, recall that the counting query is linear; thus, means and variances can be obtained in closed-form. We compare the sample-based approximation against exact computation in Fig. 10 and observe that the sampled quantities closely track the exact computation.

## 5.3. Path Planning

Fig. 12 shows statistics and spatial distribution of the existence of four paths (three ground and one aerial) in the semantic models of the Enschede scene. The figure shows that all representations are fairly confident in the existence of a path; visual inspection of the scene confirms the path

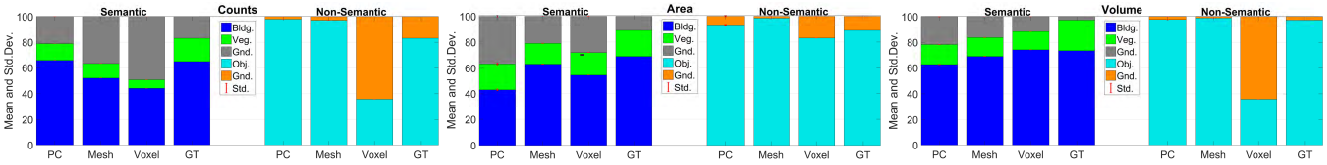
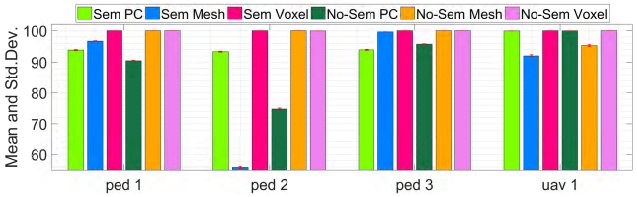
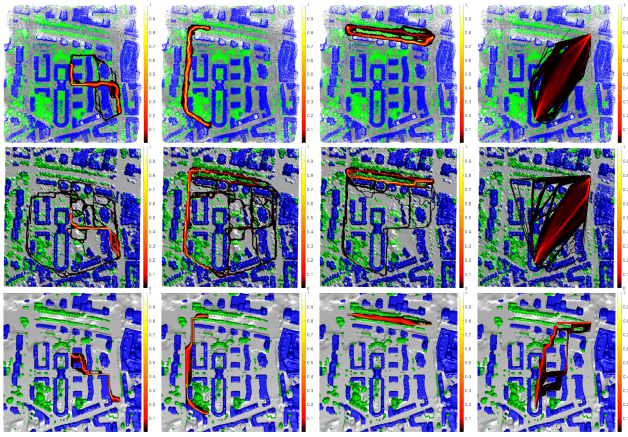



Figure 11: SynthCity estimate of categories for varied metric quantification of the scene category query (100 samples each).




(a) Path existence statistics (mean and standard deviation).



(b) Path distributions. *Columns:* 3 pedestrian and 1 UAV path (20k and 5k samples respectively). *Rows:* semantic PC, mesh and voxel.

Figure 12: Path details for four long-distance paths.

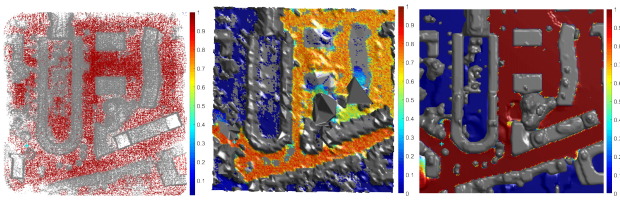


Figure 13: Mean path existence from a given starting point (cyan "+") to all ground elements. *L-R*: Enschede semantic PC, mesh and voxel (all 100 samples).

existence. Several observations can be made: first, the point cloud paths are very consistent in overall motion but have some spread. Paths in the mesh tend to have more variability, especially for the second pedestrian path where uncertainty in the mesh induces large variations in the path. Voxel paths seem to have the least variability, suggesting that the estimated semantic class likelihoods are fairly certain.

Similar to the general visibility experiment of the CLOS query, we can compose several path queries to reason about the general "reachability" of scene elements. This task can be stated as "Which locations can be reached from a specific starting point?" Fig. 13 illustrates the results for semantic models of Enschede. It is important to note that both mesh and voxel models have areas that cannot be reached as expected from the scene topology (dark blue in the figure).

## 6. Discussion

We explore the role of 3D representations in the context of abductive reasoning over large-scale urban scenes. To quantify the interaction between reasoning and representation, we focus on two concepts: a taxonomy of the space of 3D representations and the interpretation of representations as an approximate sufficient statistic for a task. These concepts serve as coupling between the domains; namely they help identify representations suitable for performing a task. We show how exact computation of task is generally infeasible and instead propose a sampling-based approach to determine the influence of uncertainty in the representation on the query answer. On the empirical front we quantify the utility of three different representations for solving three archetypal reasoning tasks. Our analyses highlight key characteristics of the representation-reasoning interaction. Namely, the scene analysis query demonstrates the link between query formulation and representation metric dimension, *i.e.* more accurate answers were obtained when these properties were aligned. This result agrees with the intuition that more assumptions lead to more inaccurate answers.

Many facets of this problem require additional analysis. A strong assumption made in this work is that the reconstruction algorithms used here are indicative of all others. This should be validated by comparing other methods. Additional representations should be considered, for simplicity our analysis skipped dynamic representations or mixtures of multiple metric dimension elements. A host of other facets can also be thoroughly investigated including: (1) robustness of representation and reconstruction algorithms, (2) scalability for varied representations and queries, (3) more queries on targeted applications, (4) effect of structure in representations. Additional material can be downloaded from http://people.csail.mit.edu/rcabezas.

**Acknowledgments.** The authors thank Julian Straub and Christopher Dean for helpful discussions. Randi Cabezas was partially supported by the ONR (N00014-17-1-2072) and John Fisher by the ONR MURI program (N00014-11-1-0688).

## References


[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

[2] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. Van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 1997.

[3] F. Bergmann and B. Fenton. Scene based reasoning. *Artificial General Intelligence*, 2015.

[4] R. Bhotika, D. J. Fleet, and K. N. Kutulakos. A probabilistic theory of occupancy and emptiness. *ECCV*, 2002.

[5] C. M. Bishop. *Pattern recognition and Machine Learning*. Springer, 2006.

[6] M. Bláha, C. Vogel, A. Richard, J. D. Wegner, T. Pock, and K. Schindler. Large-scale semantic 3D reconstruction: an adaptive multi-resolution model for multi-class volumetric labeling. In *CVPR*, 2016.

[7] J. A. Bonachela, H. Hinrichsen, and M. A. Munoz. Entropy estimates of small data sets. *Journal of Physics A: Mathematical and Theoretical*, 2008.

[8] D. Bonbouzid, R. Busa-Fekete, N. Casagrande, F.-D. Collin, and B. Kégl. MULTIBOOST: A Multiple-purpose Boosting Package. *JMLR*, 2012.

[9] J. E. Bresenham. Algorithm for computer control of a digital plotter. *IBM Systems journal*, 1965.

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

[11] R. Cabezas, O. Freifeld, G. Rosman, and J. W. Fisher III. Aerial Reconstructions via Probabilistic Data Fusion. *CVPR*, 2014.

[12] R. Cabezas, J. Straub, and J. W. Fisher III. Semantically-Aware Aerial Reconstruction from Multi-Modal Data. *ICCV*, 2015.

[13] V. Clément, G. Giraudon, S. Houzelle, and F. Sandakly. Interpretation of remotely sensed images in a context of multisensor fusion using a multispecialist architecture. *Geoscience and Remote Sensing, IEEE Transactions on*, 1993.

[14] M. Clifton, G. Paul, N. Kwok, D. Liu, and D.-L. Wang. Evaluating performance of multiple rrts. In *International Conference on Mechtronic and Embedded Systems and Applications*, 2008.

[15] D. Cremers and K. Kolev. Multiview stereo and silhouette consistency via convex functionals over convex domains. *PAMI*, 2011.

[16] R. Davis, H. Shrobe, and P. Szolovits. What is a knowledge representation? In *AI Magazine*, 1993.

[17] J. S. De Bonet and P. Viola. Poxels: Probabilistic voxelized volume reconstruction. In *ICCV*, 1999.

[18] G. Di Leo, C. Liguori, and A. Paolillo. Propagation of uncertainty through stereo triangulation. In *2010 IEEE Instrumentation&Measurement Technology Conference Proceedings*, 2010.

[19] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1959.

[20] B. A. Draper, R. T. Collins, J. Brolio, A. R. Hanson, and E. M. Riseman. The schema system. *IJCV*, 1989.

[21] M. R. Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 1995.

[22] J. Fasola, P. E. Rybski, and M. M. Veloso. Fast goal navigation with obstacle avoidance using a dynamic local visual model. In *SPAI*, 2005.

[23] M. Fernandez and S. Williams. Closed-form expression for the poisson-binomial probability density function. *IEEE Transactions on Aerospace and Electronic Systems*, 2010.

[24] F. W. Fichtner. Semantic enrichment of a point cloud based on an octree for multi-storey pathfinding. Master's thesis, Delft University of Technology, 2016.

[25] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *PAMI*, 2010.

[26] P. Gargallo, E. Prados, and P. Sturm. Minimizing the reprojection error in surface reconstruction from images. In *ICCV*, 2007.

[27] P. Gargallo, P. Sturm, and S. Pujades. An occupancy–depth generative model of multi-view images. In *ACCV*, 2007.

[28] P. Garnesson and G. Giraudon. Spatial context in an image analysis system. In *ECCV*, 1990.

[29] P. Garnesson, G. Giraudon, and P. Montesinos. An image analysis, application for aerial imagery interpretation. In *Pattern Recognition, 1990. Proceedings., 10th International Conference on*, 1990.

[30] C. Guestrin, A. Krause, and A. Singh. Near-optimal sensor placements in gaussian processes. *ICML*, 2005.

[31] A. Gupta, S. Satkin, A. Efros, and M. Hebert. From 3D scene geometry to human workspace. *CVPR*, 2011.

[32] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3D scene reconstruction and class segmentation. In *CVPR*, 2013.

[33] R. Hartley and A. Zisserman. Multiple View Geometry in Computer Vision. *Cambridge University Press*, 2003.

[34] M. Held. ERIT - a collection of efficient and reliable intersection tests. *Journal of Graphics Tools*, 1997.

[35] R. V. Hogg, J. W. McKean, and A. T. Craig. *Introduction to mathematical statistics*. Pearson 7th ed., 2013.

[36] W. Hu and S.-C. Zhu. Learning 3D object templates by quantizing geometry and appearance spaces. *PAMI*, 2015.

[37] S. Ikehata, H. Yang, and Y. Furukawa. Structured indoor modeling. In *ICCV*, 2015.

[38] S. Katz and A. Tal. On the Visibility of Point Clouds. *ICCV*, 2015.

[39] A. Kembhavi, T. Yeh, and L. S. Davis. Why did the person cross the road (there)? scene understanding using probabilistic logic models and common sense reasoning. *ECCV*, 2010.

[40] B.-S. Kim, P. Kohli, and S. Savarese. 3D Scene Understanding by Voxel-CRF. *ICCV*, 2013.

[41] R. Kimmel and J. A. Sethian. Fast marching methods on triangulated domains. In *Proceedings of the National Academy of Science*, 1998.

[42] H. Kjellström, D. Kragić, and M. J. Black. Tracking people interacting with objects. In *CVPR*, 2010.

[43] K. Kolev and D. Cremers. Continuous ratio optimization via convex relaxation with applications to multiview 3d reconstruction. In *CVPR*, 2009.

[44] A. Kundu, Y. Li, F. Dellaert, F. Li, and J. Rehg. Joint semantic segmentation and 3D reconstruction from monocular video. *ECCV*, 2014.

[45] Ľ. Ladický, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. Torr. Joint optimization for object class segmentation and dense stereo reconstruction. *IJCV*, 2012.

[46] F. Lafarge. Some new research directions to explore in urban reconstruction. In *Joint Urban Remote Sensing Event (JURSE)*, 2015.

[47] S. M. LaValle. Rapidly-exploring random trees a new tool for path planning. 1998.

[48] W. E. Lorensen and H. E. Cline. Marching cubes: a high resolution 3D surface construction algorithm. *Computer Graphics*, 1987.

[49] R. Mehra, P. Tripathi, A. Sheffer, and N. J. Mitra. Visibility of Noisy Point Cloud Data. *IEEE International Conference on Shape Modeling And Applications*, 2010.

[50] T. Möller and B. Trumbore. Fast, minimum storage ray/triangle intersection. In *ACM SIGGRAPH 2005 Courses*, 2005.

[51] P. Musialski, P. Wonka, D. G. Aliaga, M. Wimmer, L. Gool, and W. Purgathofer. A survey of urban reconstruction. In *Computer graphics forum*, 2013.

[52] A. Papoulis and S. U. Pillai. *Probability, random variables, and stochastic processes*. Mc-Graw Hill, 1991.

[53] A. B. Peter, A. M. Rachael, and D. L. Christopher. *Principles of Geographical Information Systems*. Oxford University Press, 2015.

[54] M. Pollefeys, D. Nistér, J. M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S. J. Kim, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénius, R. Yang, G. Welch, and H. Towles. Detailed real-time urban 3D reconstruction from video. *IJCV*, 2007.

[55] C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[56] F. Sandakly and G. Giraudon. Multispecialist system for 3D scene analysis. In *ECAI*, 1994.

[57] F. Sandakly and G. Giraudon. Scene analysis system. In *Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*, 1994.

[58] F. Sandakly and G. Giraudon. 3D scene interpretation for a mobile robot. *Robotics and autonomous systems*, 1997.

[59] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. *CVPR*, 2006.

[60] J. A. Sethian. A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences*, 1996.

[61] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S.-C. Zhu. Joint inference of groups, events and human roles in aerial videos. *CVPR*, 2015.

[62] Slagboom en Peeters Aerial Survey. http://www.slagboomenpeeters.com/3d.htm.

[63] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the World from Internet Photo Collections. *IJCV*, 2007.

[64] T. Thomas. *Principles of geometry, mensuration, trigonometry, land-surveying, and levelling*. Longman, Brown, Green&Longmans, 1848.

[65] S. Thrun, W. Burgard, and D. Fox. *Probabilistic robotics*. MIT press, 2005.

[66] E. Tola, C. Strecha, and P. Fua. Efficient Large Scale Multi-View Stereo for Ultra High Resolution Image Sets. *Machine Vision and Applications*, 2012.

[67] A. O. Ulusoy, A. Geiger, and M. J. Black. Towards probabilistic volumetric reconstruction using ray potentials. In *3DV*, 2015.

[68] 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.

[69] H. Vu, P. Labatut, J. Pons, and R. Keriven. High accuracy and visibility-consistent dense multiview stereo. *IEEE Transactions on PAMI*, 2012.

[70] S. Wang, S. Fidler, and R. Urtasun. Holistic 3D scene understanding from a single geo-tagged image. In *CVPR*, 2015.

[71] J. L. Williams, J. W. Fisher III, and A. S. Willsky. Performance guarantees for information theoretic active inference. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS)*, March 2007.

[72] C. Wu. VisualSFM: A Visual Structure from Motion System, 2011.

[73] C. Wu. Towards linear-time incremental structure from motion. In *3D Vision*, 2013.

[74] J. Xiao and Y. Furukawa. Reconstructing the worlds museums. *IJCV*, 2014.

[75] C. Zach. Fast and High Quality Fusion of Depth Maps. *3DV*, 2008.

[76] B. Zheng, Y. Zhao, J. Yu, K. Ikeuchi, and S.-C. Zhu. Scene understanding by reasoning stability and safety. *IJCV*, 2015.