A Continuous Occlusion Model for Road Scene Understanding

Vikas Dhiman[†] Quoc-Huy Tran[*] Jason J. Corso[†] Manmohan Chandraker[*]

[†] University of Michigan, Ann Arbor, MI, USA

[*] NEC Laboratories America, Inc. Cupertino, CA, USA

Abstract

We present a physically interpretable, continuous three-dimensional (3D) model for handling occlusions with applications to road scene understanding. We probabilistically assign each point in space to an object with a theoretical modeling of the reflection and transmission probabilities for the corresponding camera ray. Our modeling is unified in handling occlusions across a variety of scenarios, such as associating structure from motion (SFM) point tracks with potentially occluding objects or modeling object detection scores in applications such as 3D localization. For point track association, our model uniformly handles static and dynamic objects, which is an advantage over motion segmentation approaches traditionally used in multibody SFM. Detailed experiments on the KITTI raw dataset show the superiority of the proposed method over both state-of-the-art motion segmentation and a baseline that heuristically uses detection bounding boxes for resolving occlusions. We also demonstrate how our continuous occlusion model may be applied to the task of 3D localization in road scenes.

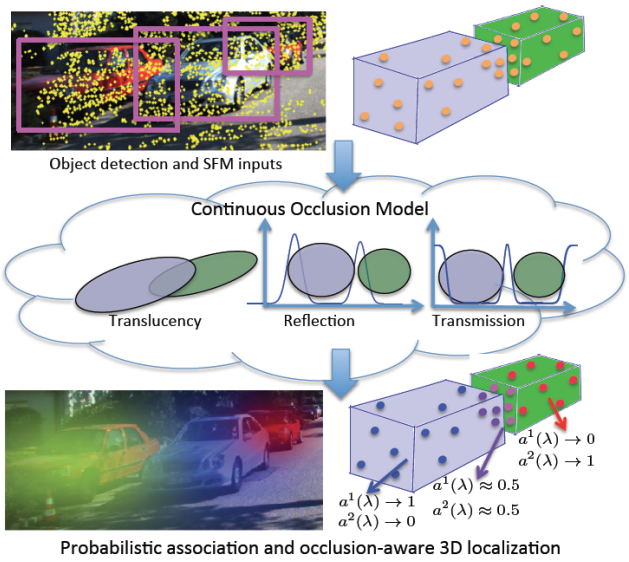

Figure 1: We propose an occlusion model in 3D that is physically-inspired and continuous. Given object detection and SFM point tracks, our unified model probabilistically assigns point tracks to objects and reasons about object detection scores and bounding boxes. It uniformly handles static and dynamic objects, thus, outperforms motion segmentation for association problems. We also demonstrate occlusion-aware 3D localization in road scenes.

1. Introduction

As a two-dimensional (2D) projection of the three-dimensional (3D) world, image formation is associated with a loss of information. This is especially significant when objects in 3D space occlude each other with respect to the camera viewpoint. In recent years, we have seen remarkable progress in various aspects of scene understanding, such as structure from motion (SFM) and object detection. However, occlusions still present a challenge, with the difficulty of physically modeling them being a major bottleneck.

Our main contribution is a novel theoretical model for occlusion handling that is continuous and fully 3D. Our model is motivated by insights from computer graphics, whereby we represent objects as translucent 3D ellipsoids. In Section 3, we develop novel continuous models for representing *transmission* and *reflection* probabilities for each ray emanating from the camera. This allows assigning probabilities for each point in space belonging to an object, which can explicitly explain image observations and reason about occlusions. This is in contrast to prior works that consider occlusions in 2D, or through discrete occluder patterns or models that are not physically interpretable [18, 19, 30, 33, 34, 35].

A key advantage afforded by our occlusion model is unification. While previous approaches to handling occlusions are application-dependent, ours is physically-inspired, thus, flexible enough to be used in a variety of scenarios. In this paper, we show that our theory can be used for uniformly modeling the association of SFM point tracks with static or dynamic objects (Section 4.1), as well as modeling object detection scores in applications like 3D localization (Section 4.2). We demonstrate the application of our formulations for road scenes from the KITTI raw dataset [7].

In particular, assigning 2D point tracks to independent,

but potentially occluding, objects is a fundamental challenge in computer vision problems such as multibody SFM [17]. Recent works use motion segmentation [1, 21] as a precursor to localizing objects, which often suffices for moving objects [25] and has also been considered for multibody SFM [12]. However, motion-based segmentation is not always applicable in road scenes, due to static parked cars, or dynamic cars that move with similar velocities. Occlusions make the problem more severe since point tracks get clustered together for static objects and may frequently appear to change association among dynamic objects in 2D. Indeed, we show in Section 5 that our point track association outperforms state-of-the-art motion segmentation methods, as well as a baseline that uses detection bounding boxes but does not consider occlusions.

Another potential application of our proposed model is towards 3D localization in road scenes. Prior works such as [24] combine information from point tracks and detection bounding boxes, but do not consider occlusions for either. In contrast, our unified occlusion model allows a probabilistic soft assignment of point tracks to objects, as well as an occlusion-aware interpretation of object detection outputs. Our model is continuous, so it remains amenable to the use of continuous optimization tools.

To summarize, our main contributions are:

- A novel theoretical model for handling occlusions that is continuous and formulated in 3D.
- Unified occlusion handling for point tracks in SFM and bounding boxes and detection scores in object detection.
- Application of our model to association of point tracks with both static and moving objects, improving over motion segmentation and occlusion-unaware baselines.
- Application of our unified formulation to 3D localization of traffic participants in road scenes.

2. Related Work

Occlusion handling in detection Several works in object detection consider occlusion by training a detector on visible parts of the object [6]. Occlusion reasoning based on 2D image silhouettes is used to improve detection performance in [10]. On the other hand, our occlusion reasoning is based on 3D entities. In recent years, object detectors have also considered occlusion reasoning using 3D cues, often learned from a dataset of CAD models [18, 19, 30]. By necessity, such frameworks are often a discrete representation of occlusion behavior, for example, in the form of a collection of occlusion masks derived from object configurations discretized over viewpoint. In contrast to these works, our occlusion modeling is also fully 3D, but allows for a continuous representation. Further, to derive 3D information, we do not use CAD models, rather we derive a probabilistic formulation based on physical insights.

Occlusion handling in tracking Occlusions have also been handled in tracking-by-detection frameworks by considering occluder patterns in the image [13, 29]. A notable exception is the work of Milan et al. [15] that explicitly models occlusions in the continuous domain to determine a visibility ratio for each object in multi-target tracking. However, the occlusion model in [15] is essentially the overlap of image projections of a Gaussian representation of the object. Our occlusion modeling, on the other hand, is fully 3D, based on physical modeling of object-ray intersections and much more general in determining the probability of a point in space as belonging to an object. While our model can also be used to determine a visibility ratio similar to [15], it has far more general applications and can be quantitatively evaluated, as shown by our experiments on point track associations.

Motion segmentation and multibody SFM An application for our occlusion modeling is to determine point track associations in scenes with multiple objects. For moving objects, this is within the purview of motion segmentation, which has been approached through algebraic factorization methods [4, 27, 26], statistical methods [11, 9, 20] and clustering methods [31, 8]. Some recent efforts include robust algebraic segmentation with hybrid perspective constraints [21] and spectral clustering with point track spatial affinities [1]. Unlike our work, such methods cannot handle static objects, or dynamic objects with little relative motion. Closer to our application, motion segmentation is also used within multibody SFM frameworks [12, 16, 17]. In contrast to these works, our formulation does not distinguish between moving and static objects and also explicitly reasons about occlusions due to 3D object geometries for associating point tracks to individual objects.

3D localization One of the vital goals of 3D scene understanding is to localize 3D objects in complex scenes. Monocular frameworks like ours have also reasoned about occlusions, for instance, partial object detectors are considered in [28]. A detailed part-based representation of objects based on annotated CAD models is used for monocular scene understanding in [33, 34, 35], which also allows reasoning about mutual occlusions between objects. In contrast to these works, our monocular framework uses a physical modeling of occlusion in continuous space and derives unified representations for SFM points and object detection bounding boxes. This makes our model more general, extensible and amenable for continuous optimization.

3. Continuous Occlusion Model

A common parametric modeling for objects, especially traffic participants in road scene understanding, is as opaque

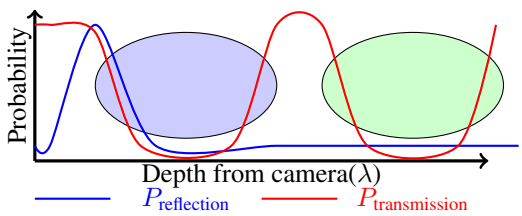

Figure 2: We represent objects as translucent ellipsoids, which leads to the formulation of transmission and reflection probabilities. This figure shows the reflection probability for the first object (in violet), which has high values around the camera-facing side of the object. Also, note that the transmission probability is inversely proportional to occupancy.

cuboids.[1] However, such models introduce discontinuities in the problem formulation and do not adequately account for uncertainties in pose and dimensions. With this motivation, we introduce our representation of 3D objects and our modeling of object-object relationships, which lead to a continuous occlusion model that correctly accounts for uncertainties in position and dimensions. We refer the reader to Figure 2 for an illustration of the proposed concepts.

Occupancy model for traffic participants Intuitively, we consider traffic participants to be regions of 3D space with a high probability of occupancy. We model the uncertainty in occupancy as a translucency function, with regions more likely to be occupied by an object considered more opaque and regions more likely to be free space considered more transparent. Based on this intuition, we model objects as translucent 3D ellipsoids whose opacity is maximum at the center and falls off towards the edges. In particular, we model the occupancy at 3D location $\mathbf{x}$ corresponding to an object O_i centered at $\mathbf{p}_i$ as

$$f^i_{\text{occ}}(\mathbf{x}) = \mathcal{L}(\mathbf{x}; \mathbf{p}_i, \mathbf{\Sigma}_i), \tag{1}$$

where $\mathcal{L}(\cdot)$ is the logistic function given by

$$\mathcal{L}(\mathbf{x}; \mathbf{p}, \mathbf{\Sigma}) = \frac{1}{1 + e^{-k(1-d(\mathbf{x},\mathbf{p}))}}, \tag{2}$$

with $d(\mathbf{x}, \mathbf{p}) = (\mathbf{x}-\mathbf{p})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{p})$ being the Mahalanobis distance. We set $k = 10\ln(49)$ as the value that allows the logistic function $\mathcal{L}$ to drop to 0.98 at a distance $d = 0.9$ from the object center. The spread of the ellipsoid, determined by $\mathbf{\Sigma}_i$, depends on the dimensions of the object. Please refer to the supplementary material for the computation of $\mathbf{\Sigma}_i$ from object dimensions.

Image formation Given the above occupancy representation of the scene, a point on an object is observed in the camera when precisely two conditions are satisfied. First, the backprojected ray from the observed image pixel is transmitted through free space until it reaches the object. Second, the ray encounters an opaque enough object surface and is reflected. More formally, the probability of observation of a point $\mathbf{x}_j$ on object O_i is given by

$$P^{ij}_{observation} = P^{ij}_{reflection} P^{j}_{transmission}. \tag{3}$$

The reflection probability ensures the presence of an object to constitute the observation, while the transmission probability allows us to model occlusions. The forms of these two functions are described next.

Reflection probability Consider a 3D point $\mathbf{x}_j$ observed in the image at pixel $\mathbf{u}_j$. Let $\mathbf{K}$ be the intrinsic calibration matrix for the camera and $\hat{\mathbf{r}}_j = \frac{\mathbf{K}^{-1}\mathbf{u}_j}{\|\mathbf{K}^{-1}\mathbf{u}_j\|}$ be the unit vector along the backprojected ray from the camera center passing through $\mathbf{u}_j$. Then, the probability of reflection at depth λ along the ray $\hat{\mathbf{r}}_j$, by an object O_i, is determined by the gradient of the object's occupancy function f^i_{occ} as

$$P^{ij}_{\text{reflection}}(\lambda) = \frac{1}{Z}(\max\{0, \nabla f^i_{occ}(\mathbf{x}_j)^\top \hat{\mathbf{r}}_j\})^2. \tag{4}$$

The gradient $\nabla f^i_{occ}(\mathbf{x}_j)$ encourages the reflection probability to be high near object boundaries, the $\max$ ensures that negative probability due to the gradient in the direction opposite to the ray is clipped off and squaring allows the function to be smooth near zero. Here, Z denotes the normalization factor. We note that in the extreme case of an object being fully opaque (that is, $\nabla f^i_{occ}(\mathbf{x}_j) \in \{0, 1\}$), the above model reverts to a (squared) Lambertian reflection. Figure 2 shows an example of the reflection probability.

Transmission probability Since we are modeling occupancy as transparency, we derive inspiration from optics for the modeling of translucent objects. A model for transmission of light across a distance α, through a medium of density ρ and opacity β is given by the Beer-Lambert law as

$$I(\alpha) = I_0 e^{-\beta\rho\alpha}. \tag{5}$$

In our formulation of scene occupancy, both opacity and density at a scene point $\mathbf{x}_j$ are encapsulated within the total occupancy function summed over all objects, $f_{occ}(\mathbf{x}_j) = \sum_i f^i_{occ}(\mathbf{x}_j)$. Further, the domain of our occupancy function $f_{occ}(\mathbf{x}_j)$ is $[0, 1]$ instead of $[0, \infty)$ for opacity β. Thus, we replace $e^{-\beta\rho}$ by the transparency function $1 - f_{occ}(\mathbf{x}_j)$ and consequently, the transmission probability over a small distance $d\lambda$ is given by

$$P^{j}_{transmission}(\lambda + d\lambda) = P^{j}_{transmission}(\lambda)(1 - f_{occ}(\mathbf{x}_j))^{d\lambda}. \tag{6}$$

Thus, for an image point $\mathbf{u}_j$ to correspond to a 3D point $\mathbf{x}_j$ at depth λ along the backprojected ray $\hat{\mathbf{r}}_j$, the ray must be

[1] Notable exceptions exist, such as [34, 35], but we note that such models are expensive, application-specific and still discontinuous.

transmitted through space with the probability

$$P^j_{transmission}(\lambda) = \prod_c^{\lambda} (1 - f_{occ}(\lambda \hat{\mathbf{r}}_j))^{d\lambda}. \quad (7)$$

Here, $\prod_c^{\lambda}$ represents a *product integral* from c to λ, where c is the position of camera screen, considered here to be equivalent to the focal length of the camera .[2]

In practice, the integral for transmission probability (7) is difficult to compute even numerically. So we choose a parameterization in the form of a product of sigmoid functions, which is a reasonable approximation to the behaviour of the transmission probability, as follows:

$$P^j_{transmission}(\lambda) = \prod_i (1 - \mathcal{L}_u(\mathbf{u}; \boldsymbol{\mu}_i, \boldsymbol{\Gamma}_i)\mathcal{L}_\lambda(\lambda; \nu_i)), \quad (8)$$

where $\mathcal{L}_u(.)$ is sigmoid in the image domain, with $\boldsymbol{\mu}_i$ and $\boldsymbol{\Gamma}_i$ representing the elliptical projection of object O_i in the image and $\mathcal{L}_\lambda(.)$ is sigmoid in the depth domain, with ν_i being the mean depth of object O_i. That is,

$$\mathcal{L}_u(\mathbf{u}; \boldsymbol{\mu}_i, \boldsymbol{\Gamma}_i) = \frac{1}{1 + e^{-k_u(1-(\mathbf{u}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Gamma}_i^{-1}(\mathbf{u}-\boldsymbol{\mu}_i))}}, \quad (9)$$

$$\mathcal{L}_\lambda(\lambda; \nu_i) = \frac{1}{1 + e^{-k_d(\lambda - \nu_i)}}. \quad (10)$$

In Figure 3, we compare the exact and approximate formulations of transmission probability given by (7) and (8), respectively. Note that the choice of mean depth of the object causes some deviation from the exact transmission probability. However, the shift of transmission probability anywhere through the object is still a reasonable approximation as occluded points can only lie outside the object. On the other hand, it yields significant computational savings since ray intersections with an ellipsoid are expensive to evaluate densely.

Thus, we have modeled the transmission probability to effectively capture the effect of occlusion due to all traffic participants in a scene that lie along a particular ray. We reiterate that our reflection and transmission probabilities are continuous functions, which allows us to keep the problem formulation in the continuous domain.

4. Unified Occlusion Models

In this section, we highlight the versatility of our occlusion modeling by demonstrating its unified application to two different problems: associating point tracks with objects and 3D object localization using objects and point tracks. Table 1 summarizes inputs and outputs for these problems.

[2] A product integral is a simple integral in the log domain

$$\prod_c^{\lambda} (1 - f_{occ}(\lambda \hat{\mathbf{r}}_j))^{d\lambda} = e^{\int_c^{\lambda} \ln(1 - f_{occ}(\lambda \hat{\mathbf{r}}_j)) d\lambda}.$$

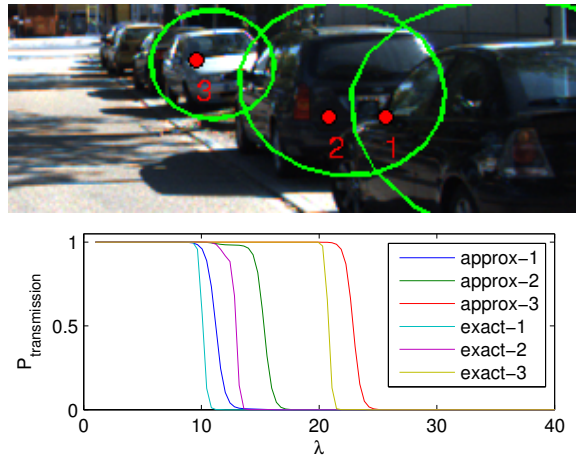

Figure 3: Comparisons between the approximate and exact formulations of $P^j_{\text{transmission}}(\lambda)$. The drop in the approximate version is delayed because we assume drop at the object center rather than the camera-facing face of the object.

	Symbol	Description
Input	$\mathbf{u}_j(t)$	2D feature track j at time t
	$\mathbf{d}^i(t)$	2D detection bounding box of object O_i at time t
Initialization with [23]	$\mathbf{p}^c(t)$	Position of camera at time t
	$\boldsymbol{\omega}^c(t)$	Orientation of camera at time t
	$\mathbf{p}^i_0(t)$	Initial position of object O_i at time t
	$\boldsymbol{\omega}^i_0(t)$	Initial orientation of object O_i at time t
	$\mathbf{B}^i_0$	Initial 3D dimensions of object O_i
Output	P^{ij}_{assoc}	Probability of assigning feature track j to object O_i
	$\mathbf{p}^i(t)$	Position of object O_i at time t
	$\boldsymbol{\omega}^i(t)$	Orientation of object O_i at time t
	$\mathbf{B}^i$	3D dimensions of object O_i

Table 1: Notation of inputs and outputs for object-point association and 3D object localization. Note that object dimensions are independent of time.

4.1. Object-Point Association

Given 2D image points $\{\mathbf{u}_j\}$ that are tracked between consecutive frames and a set of objects $\{O_i\}$ appearing in the frames, we aim to associate $\mathbf{u}_j$ with O_i. Based on our continuous occlusion model in Section 3, the association probability $a^{ij}(\lambda)$ between point track $\mathbf{u}_j$ and object O_i at depth λ can be defined as

$$a^{ij}(\lambda) = P^{ij}_{\text{reflection}}(\lambda) P^j_{\text{transmission}}(\lambda), \quad (11)$$

where $P^{ij}_{\text{reflection}}(\lambda)$ and $P^j_{\text{transmission}}(\lambda)$ are from (4) and (8) respectively. Note that the fraction $a^{ij}(\lambda)$, although called association probability, does not capture the entire information that we have available for computing the association of point tracks to objects.

Rather, to compute the association probability between point track $\mathbf{u}_j$ and object O_i, we should also use the reprojection error. When the association of point track $\mathbf{u}_j$ and object O_i is correct and the point of reflection is at depth λ, the corresponding reprojection error $E^{ij}_{\text{reproj}}(\lambda)$ must be zero, otherwise the error becomes a measure of distance from the

true solution. The error $E^{ij}_{\text{reproj}}(\lambda)$ can be used for associating point tracks and objects by converting it to the probability domain as

$$P^{ij}_{\text{reproj}}(\lambda) = \frac{1}{Z'} \exp(-E^{ij}_{\text{reproj}}(\lambda)), \tag{12}$$

where Z' is the normalization coefficient.

Using both of the evidence terms in (11) and (12), we can define the new association probability P^{ij}_{assoc}, as follows:

$$P^{ij}_{\text{assoc}} = \frac{1}{Z''} \int_0^{\infty} a^{ij}(\lambda) \exp(-E^{ij}_{\text{reproj}}(\lambda)) d\lambda, \tag{13}$$

where Z'' is the new normalization coefficient.

Once we have computed the association probability P^{ij}_{assoc} for every pair of point tracks and objects, we can assign each point track to the object with the highest association probability. The point tracks having very small association probabilities are assigned to the background.

In contrast to the principled approach above, a heuristic baseline may simply assign a point track to the detection bounding box enclosing it (and the background if outside all bounding boxes). For regions where bounding boxes overlap, it may assign point tracks to the object that has the smallest mean depth among the competing bounding boxes. As we demonstrate in our experiments, such heuristics are sub-optimal compared to using (13) from our occlusion model.

4.2. 3D Object Localization

In this section, we exploit our continuous occlusion model for another application, namely, 3D object localization in road scenes, which further demonstrates its versatility. Given a set of 2D tracked feature points $\{\mathbf{u}_j(t)\}$ and 2D detection bounding boxes $\{\mathbf{d}^i(t)\}$ at frame t, the goal is to localize 3D traffic participants. In particular, for each traffic participant, we wish to estimate its position $\mathbf{p}^i(t)$ and orientation $\boldsymbol{\omega}^i(t)$ on the ground plane and its 3D dimensions $\mathbf{B}^i(t)$. Please refer to Table 1 for a summary of inputs and outputs.

We construct a graphical model for representing relationships among objects, as well as between objects and point tracks. Figure 4 illustrates an example of the graph and energies. The negative log likelihood is decomposed as follows:

$$\begin{aligned} -\log P(\{\mathbf{p}^i(t), \boldsymbol{\omega}^i(t), \mathbf{B}^i(t)\} | \{\mathbf{u}_j(t)\}, \{\mathbf{d}^i(t)\}) = \\ -\tilde{Z} + \sum_{t=s_i}^{e_i} \lambda_{\text{track}} \mathcal{E}^{ijt}_{\text{track}} + \\ \sum_{t=s_i}^{e_i} \sum_{i=1}^{N} \left(\lambda_{\text{detect}} \mathcal{E}^{it}_{\text{detect}} + \lambda_{\text{dyn}} \mathcal{E}^{it}_{\text{dyn}} + \lambda_{\text{size}} \mathcal{E}^{it}_{\text{size}} \right) \quad , \end{aligned}$$

where $\mathcal{E}^{ijt}_{\text{track}}$ and $\mathcal{E}^{it}_{\text{detect}}$ reason about image observations such as point tracks and bounding boxes, while $\mathcal{E}^{it}_{\text{dyn}}$ and $\mathcal{E}^{it}_{\text{size}}$ impose smoothness constraints and size priors respectively.

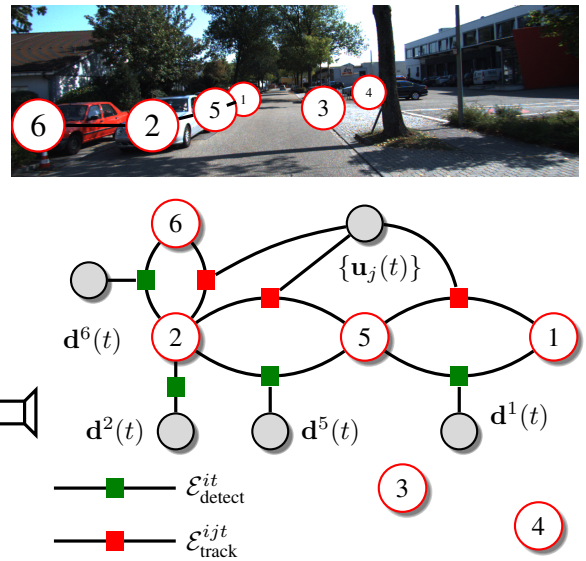

Figure 4: (Top) A sample road scene with occlusions, where the unknowns of each object are modeled as random variables. (Bottom) The graphical model corresponding to the above frame. In particular, the numbered nodes denote the unknown state variables of each object (position, orientation, and dimensions), the shaded nodes are observed variables (detection bounding boxes and point tracks), and the colored squares represent various energies that capture object-object interactions.

Here, λ_{track}, λ_{detect}, λ_{dyn}, λ_{size} are energy weights, N is the number of objects in the sequence, s_i and t_i are respectively the starting and ending frames of object O_i, and $\tilde{Z}$ is the normalization coefficient. Next, we present our unified continuous occlusion modeling for both point track and bounding box energies. Due to space constraints, we present the details of other energies in the supplementary material.

Continuous point track energy with occlusion Let $\Omega^i(t)$ be the pose of object O_i at time t in world coordinates, which is computed using the camera pose at time t and the relative pose of object O_i with respect to the camera at time t. We denote $\pi_{\Omega^i(t)}(.)$ and $\pi^{-1}_{\Omega^i(t-1)}(.)$ as the forward and inverse projection functions that project a 3D point to the 2D image and vice versa. Then, the reprojection error for 2D point $\mathbf{u}_j(t)$ with hypothesized depth λ, is given by

$$E^{ij}_{\text{reproj}}(\lambda) = \left\| \mathbf{u}_j(t) - \pi_{\Omega^i(t)} \left(\pi^{-1}_{\Omega^i(t-1)} (\mathbf{u}_j(t-1), \lambda) \right) \right\|^2. \tag{14}$$

Note that the inverse projection $\pi^{-1}_{\Omega^i(t)}(.)$ depends on both the 2D point $\mathbf{u}_j(t)$ and the unknown depth λ. Also note that the inverse projection relies on the object pose at time $t-1$ while the forward projection relies on the object pose at time t, which can be different.

For an object O_i, let $\{\Omega(t)\}_i$ be the poses of all occluding objects at time t (inclusive of object O_i) and $\{\mathbf{B}\}_i$ be

their corresponding 3D dimensions. Then, we model the continuous point track energy with explicit occlusion reasoning as the expected reprojection error over the association probability

$$\mathcal{E}^{ijt}_{\text{track}}(\{\Omega(t)\}_i, \{\Omega(t-1)\}_i, \{\mathbf{B}\}_i) = \sum_{i=1}^{N}\sum_{j=1}^{M}\int_0^{\infty} a^{ij}(\lambda)E^{ij}_{\text{reproj}}(\lambda)d\lambda, \quad (15)$$

where N and M are, respectively, the number of objects and points and $a^{ij}(\lambda)$ is the association probability of point $\mathbf{u}_j(t)$ with object O_i at depth λ, given by (11).

Continuous bounding box energy with occlusion Object detection is usually followed by non-maximal suppression that results in discarding similar bounding boxes. When we are jointly optimizing detections with other cues, it is usually not desirable to use a single bounding box. To retain the entire detection output while maintaining the continuous form of our energies, we approximate the distribution of detection scores with a multi-modal sum of Gaussian-like logistic functions. In particular, let 2D bounding box $\mathbf{d}^i(t)$ be parameterized as a 4D vector $[x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}}]^\top$. We fit a parametric function to the detection scores, of the form

$$S(\mathbf{d}^i(t)) = \sum_k A_k \exp\left(-\boldsymbol{\epsilon}^i_k(t)^\top \boldsymbol{\Lambda}_k^{-1}\boldsymbol{\epsilon}^i_k(t)\right), \quad (16)$$

where A_k is an amplitude and $\boldsymbol{\epsilon}^i_k(t) = \mathbf{d}^i(t) - \boldsymbol{\mu}_k$, with $\boldsymbol{\mu}_k$ the mean and $\boldsymbol{\Lambda}_k$ the covariance. We use a non-linear solver to minimize the above, with initialization from non-maximal suppressed outputs. The optimization is constrained by the symmetry and positive definiteness of $\boldsymbol{\Lambda}_k$, $x_{\text{max}} \geq x_{\text{min}}$ and $y_{\text{max}} \geq y_{\text{min}}$.

Detection scores with occlusion reasoning With our model of $P^j_{\text{transmission}}(\lambda)$ described in Section 3, we compute the probability of a point $\mathbf{u}$ in the image to be occluded, assuming the point is on object O_i with mean depth ν_i, as

$$\Theta_i(\mathbf{u}, \nu_i) = 1 - P_{\textit{transmission}}(\nu_i, \mathbf{u}). \quad (17)$$

If a portion of the proposed detection bounding box is known to be occluded, one would like to decrease the confidence in the detection score about the localization of that end of the object. Assuming that occlusions are more likely on the boundaries of the detection bounding box, we wish to decrease the confidence on the mean detection boundaries around the occluded boundaries. To re-model detection scores scaled by continuous occlusion, we sample $\Theta_i(\mathbf{u}, \nu_i)$ at the hypothesized detection boundaries from the Gaussian mixture model (GMM) $S(.)$ in (16) and augment the detection boundary covariance matrix by $\mathcal{P}_i = \rho_i\rho_i^\top$, where $\rho_i = \Theta_i(\mathbf{u}, \nu_i)$. The new covariance matrix for the detection scores is given by $\boldsymbol{\Lambda}'_k = \mathcal{P}_i + \boldsymbol{\Lambda}_k$ for all k. The detection score GMM $S'(.)$ with explicit occlusion reasoning is given by replacing the covariance matrix, as follows:

$$S'(\mathbf{d}^i(t)) = \sum_k A_k \exp\left(-\boldsymbol{\epsilon}^i_k(t)^\top \boldsymbol{\Lambda}_k^{'-1}\boldsymbol{\epsilon}^i_k(t)\right). \quad (18)$$

The energy of detection scores is simply taken to be the inverse of the above detection score, that is,

$$\mathcal{E}^{it}_{\text{detect}}(\{\Omega^i(t)\}_i, \{\mathbf{B}^i\}_i) = \frac{1}{S'(\mathbf{d}^i(t))}. \quad (19)$$

Inference on graphical model We apply the Metropolis-Hastings method [14] to perform inference on the graphical model. Since we optimize over continuous variables, we use the Gaussian distribution as the proposal function. We choose this over alternatives such as block-coordinate descent since they are slower in our experiments.

Point tracks	Ours	BBox	BM	RAS
Dynamic & occluded	**13.2**	21.3	30.9	30.1
Occluded	**15.7**	19.8	39.5	37.8
Dynamic	**6.6**	11.4	15.3	17.7
All	**8.6**	12.6	21.9	21.5

Table 2: Mean association errors on different sets of input point tracks over all sequences. Errors are in terms of average fractions of foreground points incorrectly associated to objects per sequence.

5. Experiments

In this section, we benchmark our continuous occlusion model for point-to-object association against the baseline method using detection bounding boxes and state-of-the-art methods for motion segmentation [1, 21]. We then show how the proposed model may be applied for 3D object localization in road scenes. For our experiments, we use 35 sequences of the KITTI raw dataset [7], which are recorded under a variety of driving conditions and include 10,088 frames and 634 object tracks in total.

5.1. Association Experiments

Setup We first perform the association experiment that compares the accuracy of point-to-object association using our proposed model against a heuristic baseline and state-of-the-art motion segmentation methods. The detection bounding box baseline method (BBox) is as described at the end of Section 4.1. For motion segmentation, we use robust algebraic segmentation with hybrid perspective constraints (RAS) [21] and spectral clustering with point track spatial affinities (BM) [1].

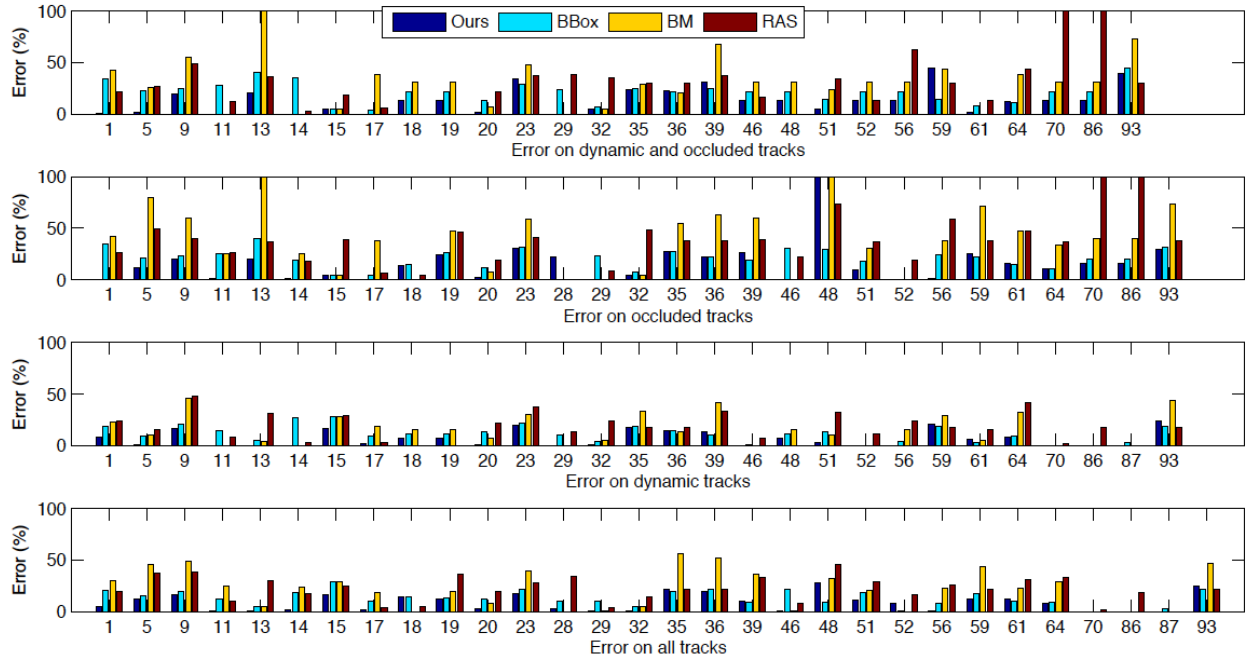

Figure 5: Association errors on different sets of input point tracks. Numbers on the x-axis represent sequence numbers in the KITTI raw dataset. Errors are in terms of average fractions of foreground points incorrectly associated to objects per sequence.

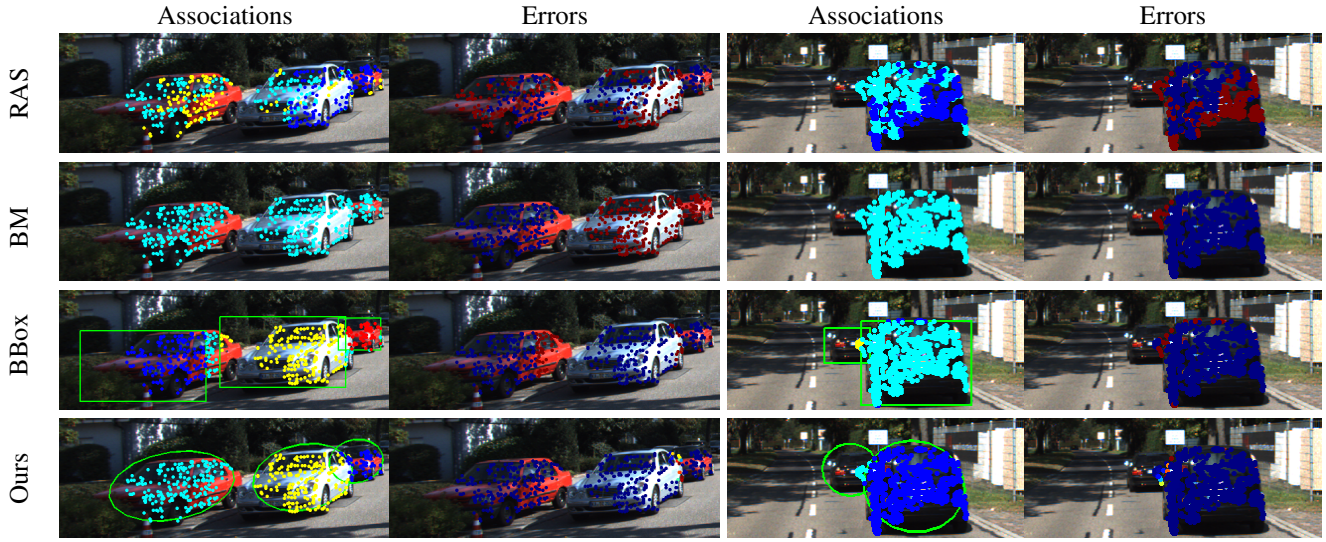

Figure 6: Qualitative results of the association experiment. The "Associations" columns show the point track assignments to appropriate objects. Each color represents a different object to which point tracks can be associated to. The "Errors" columns show the probabilistic errors in association: low error points are in blue while high error points are in red. Note that our method changes smoothly at the object boundaries with intermediate probabilities, while the baseline method has merely 0 and 1 errors.

For each sequence, the methods of [5] and [2] are used for computing detection bounding boxes and object tracklets, respectively. We then apply [32] to extract point tracks. Note that our method can handle occlusions in both static and dynamic scenes, but motion segmentation focuses on dynamic scenes. For a complete evaluation, we organize the point tracks into four sets: all point tracks, occluded point tracks, dynamic point tracks and dynamic as well as occluded point tracks. The parameters (position, orientation, and dimensions) of all objects (cars) estimated by the method of [23] are provided to our method (for computing association probability) and the baseline BBox method (for depth ordering). The number of objects is known a priori in our model (from object tracking [2]) and is also provided to other methods such as BBox and RAS.

Method	t	dim
Point cloud fitting	6.87	4.02
Initialization by [23]	5.61	3.23
$\mathcal{E}^{ijt}_{\text{trackNoOcc}} + \mathcal{E}^{it}_{\text{detectNoOcc}} + \mathcal{E}^{it}_{\text{size}} + \mathcal{E}^{it}_{\text{dyn}}$	3.95	1.72
$\mathcal{E}^{ijt}_{\text{trackNoOcc}} + \mathcal{E}^{it}_{\text{detect}} + \mathcal{E}^{it}_{\text{size}} + \mathcal{E}^{it}_{\text{dyn}}$	4.81	2.16
$\mathcal{E}^{ijt}_{\text{track}} + \mathcal{E}^{it}_{\text{detectNoOcc}} + \mathcal{E}^{it}_{\text{size}} + \mathcal{E}^{it}_{\text{dyn}}$	4.05	**1.59**
$\mathcal{E}^{ijt}_{\text{track}} + \mathcal{E}^{it}_{\text{detect}} + \mathcal{E}^{it}_{\text{size}} + \mathcal{E}^{it}_{\text{dyn}}$	**3.24**	2.16

Table 3: Localization experiment results with different combinations of energies. We report translation error (t) and dimension error (dim) in meters per car. Yaw angles for static objects are not optimized by our model. These experiments use the set of occluded tracks to demonstrate the effect of our modeling.

Results Figure 5 shows the association errors – the percentages of point tracks incorrectly assigned to objects – for all methods on the four sets of input point tracks, for each sequence. The mean results over all sequences are summarized in Table 2. From Figure 5, our method is usually the most accurate among all methods, leading to the best mean error on all sets of input point tracks in Table 2, which is followed by the bounding box baseline method. This clearly shows the advantage of our continuous occlusion model over the simple baseline method for resolving occlusions. RAS and BM often have the highest errors in Figure 5, thus, the highest mean errors on all sets of input point tracks in Table 2.

More importantly, both RAS and BM rely on motions of objects for clustering point tracks, therefore they cannot work well with static point tracks (for example, point tracks that belong to parked cars). This fact can be observed in Table 2, where there are large differences in the mean errors of both methods on data containing static point tracks (rows 2 and 4) and data consisting of dynamic point tracks only (rows 1 and 3). In contrast, our method and the baseline method are relatively independent of object motions, resulting in smaller performance gaps. Further, our method also outperforms motion segmentation on dynamic objects (row 3), which shows the effect of detection bounding boxes and by a more significant margin when occlusions are present (row 1), which shows the effect of our occlusion modeling.

Qualitative comparisons of point track associations from various methods are shown in Figure 6. We note the low errors using our occlusion model and the smooth transition of assignment across object boundaries.

5.2. Localization Experiments

We report errors in translation and dimension estimates, measured in meters per car, in Table 3. The average depth of cars in the dataset is approximately 20 meters. We compare four combinations of energies against the initialization using [23] and a simple baseline which fits a 3D cuboid on the 3D point cloud reconstructed using SFM within detection bounding boxes in consecutive frames (for unobservable dimensions, such as when only the back of a car is visible, we rely on 3D size priors). The energy $\mathcal{E}^{ijt}_{\text{trackNoOcc}}$ represents the point track energy without accounting for occlusions, that is, we model $\mathcal{E}^{ijt}_{\text{track}}$ in the absence of $a^{ij}(\lambda)$. Similarly, $\mathcal{E}^{it}_{\text{detectNoOcc}}$ is the bounding box energy without the modification of $\mathbf{\Lambda}_k$ that accounts for occlusion. We use $\lambda_{\text{track}} = 1$, $\lambda_{\text{detect}} = 1$, $\lambda_{\text{dyn}} = 10$, $\lambda_{\text{size}} = 7$. Please refer to the supplementary material for a detailed list of parameter settings.

From Table 3, the baseline method has the highest errors, which is likely due to lack of point tracks and incorrect point-to-object associations (using detection bounding boxes). Moreover, minimizing different combinations of energies yields lower errors than the initialization with [23], which shows the advantage of our energy minimization. Finally, we observe that the use of the continuous occlusion model improves the localization accuracy in terms of the translation error, which is the most significant metric affected by all cues. Occlusion modeling for detection increases dimension error since we explicitly allow greater uncertainty in occluded edges of the bounding box. Note that none of our energies optimize yaw angles for static objects, which can be handled in practice through either the detector orientation or external information such as lane geometry.

6. Conclusions and Future Work

We have presented a theoretically novel continuous model for occlusion reasoning in 3D. A key advantage is its physical inspiration that lends flexibility towards occlusion reasoning for varied elements of scene understanding, such as point tracks, object detection bounding boxes and detection scores. We demonstrate unified modeling for different applications such as object-point track associations and 3D localization. Our occlusion model can uniformly handle static and dynamic objects, which is an advantage over motion segmentation methods for object-point association. A challenge is that inference for 3D localization is currently slow, requiring a few minutes per window of frames, which prevents exhaustive cross-validation for tuning of weights. Our future work will explore speeding up the inference, for example, by approximating the graph with a tree using the Chow-Liu method [3], which will allow belief propagation for fast inference. Another direction for future work is to replace a single ellipsoid by a set of spheres for modelling a translucent object [22], which will better capture object boundary and appearance while remaining a continuous model.

Acknowledgments This work was part of V. Dhiman's internship at NEC Labs America, in Cupertino. V. Dhiman and J. J. Corso were also supported by NSF NRI IIS 1522904.

References

[1] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, pages 282–295, 2010. 2, 6

[2] W. Choi and S. Savarese. Multi-target tracking in world coordinate with single, minimally calibrated camera. In *ECCV*, pages 553–567, 2010. 7

[3] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968. 8

[4] J. Costeira and T. Kanade. A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, 29(3):159–179, 1998. 2

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 7

[6] T. Gao, B. Packer, and D. Koller. A segmentation-aware object detection model with occlusion handling. In *CVPR*, pages 1361–1368, 2011. 2

[7] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, pages 3354–3361, 2012. 1, 6

[8] A. Goh and R. Vidal. Segmenting motions of different types by unsupervised manifold clustering. In *CVPR*, pages 1–6, 2007. 2

[9] A. Gruber and Y. Weiss. Multibody factorization with uncertainty and missing data using the em algorithm. In *CVPR*, pages 707–714, 2004. 2

[10] E. Hsiao and M. Hebert. Occlusion reasoning for object detection under arbitrary viewpoint. In *CVPR*, pages 3146–3153, 2012. 2

[11] K. Kanatani. Motion segmentation by subspace separation and model selection. In *ICCV*, pages 586–591, 2001. 2

[12] A. Kundu, K. M. Krishna, and C. V. Jawahar. Realtime multibody visual SLAM with a smoothly moving monocular camera. In *ICCV*, pages 2080–2087, 2011. 2

[13] S. Kwak, W. Nam, B. Han, and J. H. Han. Learning occlusion with likelihoods for visual tracking. In *ICCV*, pages 1551–1558, 2011. 2

[14] D. J. MacKay. Introduction to monte carlo methods. In *Learning in graphical models*, pages 175–204. Springer, 1998. 6

[15] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multitarget tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):58–72, 2014. 2

[16] R. Namdev, A. Kundu, K. Krishna, and C. Jawahar. Motion segmentation of multiple objects from a freely moving monocular camera. In *ICRA*, pages 4092–4099, 2012. 2

[17] K. E. Ozden, K. Schindler, and L. V. Gool. Multibody structure-from-motion in practice. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1134–1141, 2010. 2

[18] B. Pepik, P. Gehler, M. Stark, and B. Schiele. 3D2PM – 3D deformable part models. In *ECCV*, pages 356–370, 2012. 1, 2

[19] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Occlusion patterns for object class detection. In *CVPR*, pages 3286–3293, 2013. 1, 2

[20] S. Rao, R. Tron, R. Vidal, and Y. Ma. Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories. In *CVPR*, pages 1–8, 2008. 2

[21] S. R. Rao, A. Y. Yang, S. S. Sastry, and Y. Ma. Robust algebraic segmentation of mixed rigid-body and planar motions from two views. *International Journal of Computer Vision*, 88(3):425–446, 2010. 2, 6

[22] H. Rhodin, N. Robertini, C. Richardt, H.-P. Seidel, and C. Theobalt. A versatile scene model with differentiable visibility applied to generative pose estimation. In *ICCV*, pages 765–773, 2015. 8

[23] S. Song and M. Chandraker. Robust scale correction in monocular SFM for autonomous driving. In *CVPR*, pages 1566–1573, 2014. 4, 7, 8

[24] S. Song and M. Chandraker. Joint SFM and detection cues for monocular 3D localization in road scenes. In *CVPR*, pages 3734–3742, 2015. 2

[25] R. Tron and R. Vidal. A benchmark for the comparison of 3D motion segmentation algorithms. In *CVPR*, pages 1–8, 2007. 2

[26] R. Vidal and R. Hartley. Motion segmentation with missing data using powerfactorization and gpca. In *CVPR*, pages 310–316, 2004. 2

[27] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (gpca). In *CVPR*, pages 621–628, 2003. 2

[28] C. Wojek, S. Walk, S. Roth, K. Schindler, and B. Schiele. Monocular visual scene understanding: Understanding multi-object traffic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):882–897, 2013. 2

[29] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2):247–266, 2007. 2

[30] Y. Xiang and S. Savarese. Object detection by 3D aspectlets and occlusion reasoning. In *ICCV Workshops*, pages 530–537, 2013. 1, 2

[31] J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *ECCV*, pages 94–106, 2006. 2

[32] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime TV-L1 optical flow. In *DAGM on Pattern Recognition*, pages 214–223, 2007. 7

[33] M. Zia, M. Stark, and K. Schindler. Explicit occlusion modeling for 3D object class representations. In *CVPR*, pages 3326–3333, 2013. 1, 2

[34] M. Zia, M. Stark, and K. Schindler. Are cars just 3D boxes? - Jointly estimating the 3D shape of multiple objects. In *CVPR*, pages 3678–3685, 2014. 1, 2, 3

[35] M. Z. Zia, M. Stark, and K. Schindler. Towards scene understanding with detailed 3d object representations. *International Journal of Computer Vision*, 112(2):188–203, 2014. 1, 2, 3