Are Cars Just 3D Boxes? – Jointly Estimating the 3D Shape of Multiple Objects

M. Zeeshan Zia[1,3], Michael Stark[2], and Konrad Schindler[1]

[1] Photogrammetry and Remote Sensing, ETH Zürich, Switzerland
[2] Stanford University and Max Planck Institute for Informatics
[3] Robot Vision, Imperial College London

Abstract

Current systems for scene understanding typically represent objects as 2D or 3D bounding boxes. While these representations have proven robust in a variety of applications, they provide only coarse approximations to the true 2D and 3D extent of objects. As a result, object-object interactions, such as occlusions or ground-plane contact, can be represented only superficially. In this paper, we approach the problem of scene understanding from the perspective of 3D shape modeling, and design a 3D scene representation that reasons jointly about the 3D shape of multiple objects. This representation allows to express 3D geometry and occlusion on the fine detail level of individual vertices of 3D wireframe models, and makes it possible to treat dependencies between objects, such as occlusion reasoning, in a deterministic way. In our experiments, we demonstrate the benefit of jointly estimating the 3D shape of multiple objects in a scene over working with coarse boxes, on the recently proposed KITTI dataset *of realistic street scenes.*

1. Introduction

In recent literature, there has been a strong movement away from considering objects in isolation, towards reasoning jointly about entire scenes, aiding recognition in tasks like scene understanding [20, 34, 19, 18] and object tracking [10, 4] by exploiting contextual relationships between objects and other scene entities, such as ground planes [20, 10, 36] and vertical structures [20]. At the same time, and inspired by the aspirations of the early days of computer vision [26, 6, 2, 25], it has been realized that capturing the 3D geometry of objects and scenes more accurately can lead to more accurate estimates of object location and pose [39, 29, 37, 13, 28]. The success of these models is due, at least in part, to their ability to establish correspondences across different viewpoints, and thus gain statistical strength by sharing information among them. The degree of detail of these deformable object class models ranges from about a dozen deformable parts [29] to over thirty surface vertices in a 3D wireframe [39].

Curiously, the enhanced level of geometric detail has hardly found its way into scene-level reasoning. Objects are still typically represented as 2D or 3D bounding boxes [20, 10, 27, 36] – to some degree because off-the-shelf object detectors [8, 12] output only bounding boxes.

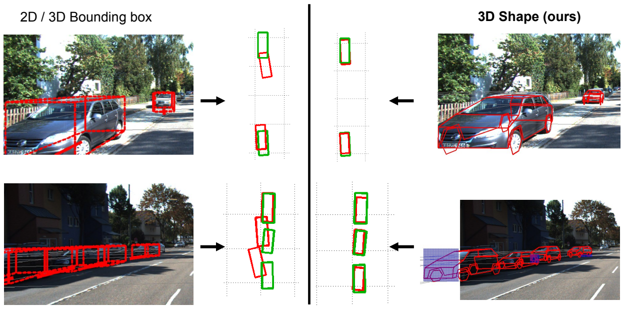


Figure 1. *Left:* Coarse 3D object bounding boxes derived from 2D bounding box detections (not shown). *Right:* our fine-grained 3D shape model fits improve 3D localization (see bird's eye views).

Such a coarse representation faces two main challenges: *(i)* there is only an implicit connection between a 2D bounding box detection and the underlying 3D geometry (by virtue of the training examples that give rise to the detection), limiting its ability to directly localize objects in 3D. And *(ii)*, coarse boxes are bound to over-estimate spatial extent in both image and 3D space, limiting their ability to describe and leverage interactions between different scene entities. While attempts have been made to mitigate *(i)* by learning parametric models that explicitly relate 2D detections and 3D object bounding boxes [3, 16] and *(ii)* by imposing soft overlap penalties [27, 36], the underlying box- or blob-like object geometry still constitutes a principal limitation of today's scene understanding approaches.

In this paper, we approach the scene understanding problem from a different angle: instead of building a scene model around an off-the-shelf 2D bounding box detector, we start directly from a fine-grained 3D object class model [39], and extend it to jointly represent scenes containing multiple objects (so far, we have focused on the car class in street scenes, see Fig. 1). The resulting scene interpretation encompasses the detailed 3D shapes of all objects in the scene, establishing an explicit connection between 2D image evidence and 3D geometry through a wireframe model (addressing challenge *(i)*) and allowing to reason about object-object interactions on the level of individual object vertices and faces (addressing challenge *(ii)*).

Our paper makes the following contributions. *First*, to our knowledge, our work is the first attempt to explore the joint estimation of multiple objects within a scene at high geometric resolution (individual parts/vertices of a wireframe model), including object-object interactions. Notably, this fine detail improves performance in both 3D object localization and viewpoint estimation over a 3D bounding box baseline (Fig. 1). *Second*, we leverage the rich detail of the 3D scene representation for occlusion modeling, combining deterministic reasoning about occlusions by detected objects with a generative probabilistic model of further, unknown occluders. This again yields improvements in 3D localization compared to the independent estimation of occlusions for each individual object. And *third*, we present a systematic experimental study on the challenging KITTI street scene dataset [15]. While our detailed 3D scene representation can not yet compete with technically mature 2D bounding box detectors in terms of recall, it is able to localize 44% of the detected highly occluded cars in our test set with an accuracy of 1 meter.

2. Related Work

Recent work involving scene-level reasoning can roughly be categorized into two major directions, based on the nature of the underlying scene representation.

Qualitative representations. The first direction is characterized by models of a qualitative nature [20, 18, 31]. Here, the focus lies on providing rough estimates of surface layout [20] and locations and relationships of the major building blocks of a scene, in the form of geometric [18] or mechanical support [31]. These representations deliver semantically rich scene descriptions, but the underlying geometry typically consists either of coarse blocks [18], or of unstructured and rather brittle superpixels [20, 31].

Quantitative representations. Works of the second direction are more quantitative in nature. They are often inspired by navigation-type applications like autonomous driving [10, 3, 16, 36] or robotics [27], where precise localization of road surface, other road users *etc*. is important. As a result, object geometry is represented and estimated with higher accuracy than for the qualitative reasoning of the first direction. Nevertheless it is typically limited to 3D boxes, often even with constant dimensions or fixed aspect ratio per basic-level [10, 27, 36] or fine-grained object class [32]. Detailed 3D object representations are lately revisited in [39, 40], however this work only models object instances independently, in 2D image space; whereas we extend that to reconstruct true 3D scene layout, and incorporate two aspects of joint object modeling: common ground plane and deterministic occlusion reasoning.

Recently, more fine-grained, geometric representations have been successfully applied to the understanding of indoor scenes [7, 9, 38], where a box-like room layout is populated with furniture objects. These approaches model objects as (arrangements of) boxes that are axis-aligned to the room-box [9, 38], and directly align model with image-edges. In contrast, our model is flexible in object azimuth angle and leverages individual part detectors as well as powerful poselet-style part configurations as image evidence.

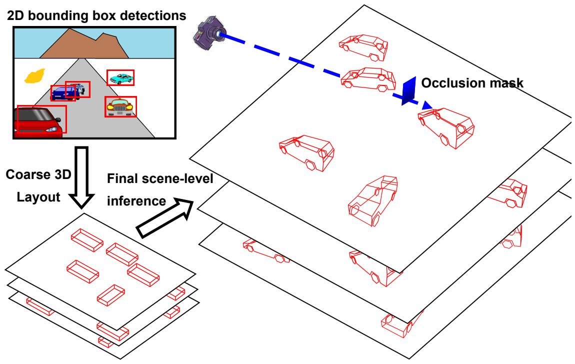


Figure 2. Scene particles (coarse 3D geometry and fine-grained shape, Sect. 3.3). Deterministic occlusion mask computation (Sect. 3.1) by ray casting and intersection (blue).

Occlusion modeling. Lately, partial occlusion has received increased attention in object class detection. Occlusions are considered on various levels of granularity, e.g. on the level of HOG [8] cells [35, 33, 14], object parts [17, 28, 40], or silhouettes [21]. With the exception of [30], which proposes to train separate detectors for multiple objects occurring with similar occlusion patterns, common to these approaches is that their occlusion modeling takes into account only a single object, namely the one that is occluded, but does not consider the remainder of the scene. As a result, these approaches have to rely on (the absence of) low-level image cues to predict occlusions, which are often noisy and unreliable. In contrast, we leverage a joint inference procedure over all objects in the scene, that takes into account interactions between different objects, such that finding object-object occlusion patterns becomes a deterministic procedure (for the known class(es)).

3. 3D Scene Model

We first describe our 3D scene model (Sec. 3.1), consisting of a common ground plane, a set of 3D deformable objects, and an explicit occlusion mask for each object. What distinguishes the model from previous work [10, 27, 36] is a much more expressive solution space that allows one to reason about the locations, shapes and interactions of objects, at the level of individual vertices and faces. We then express the likelihood of a particular scene hypothesis in that space as a combination of per-object likelihoods, computed with an existing per-object model (Sec. 3.2), and last describe our sample-based inference procedure (Sec. 3.3).

3.1. Hypothesis Space

A scene in our model consists of *(i)* a single 3D ground plane; *(ii)* a variable number of instances of a part-based, deformable object class model, which all stand on that ground plane; and *(iii)* the occlusion states of all parts in each instance. Like many other outdoor scene models [10, 27, 36] we assume known camera intrinsics for a given test image, such that one can reason in 3D metric space rather than in the 2D image plane.

Object shape. Individual object instances are represented by adapting the pseudo-3D shape model proposed in [40] for metric 3D space. An object instance is modeled as a deformable 3D wireframe $\mathbf{h}^\beta$. An active shape model governs the shape variation, *i.e*. the locations of the wireframe vertices ("parts") are determined by a small number of coefficients $\mathbf{s}$ for the strongest principal components of the deformation. The faces spanned by the vertices determine the visible (and occluding) object surfaces. At test time, shape inference amounts to estimating the coefficients $\mathbf{s}$. As in [39], we learn both shape and appearance models from 3D CAD data of the object class of interest. Note however that in our case CAD models are scaled to their real-world dimensions, for reasoning in a metric scene space.

Common ground plane. All objects are assumed to rest on a common ground plane, as often done for street scenes. While this may seem a heavy restriction, it roughly holds in most cases and drastically reduces the search space for possible object locations (2 degrees of freedom for translation and 1 for rotation, instead of 3+3). Moreover, the consensus for a common ground plane stabilizes 3D object localization, as shown in our experiments (Sec. 4.3). The ground plane itself has 2 degrees of freedom, $\boldsymbol{\theta}_{gp} = (\theta_{pitch}, \theta_{roll})$, pitch and roll angles relative to the camera coordinate frame. The height q_y of the camera above ground is assumed known and fixed.

Explicit occlusion model. Each 3D wireframe vertex has an associated binary variable that flags it as occluded or visible. In this manner one can uniformly treat occlusions caused by *(i)* other parts of the object (self occlusion) *(ii)* another object in the same scene (deterministic occlusion), or *(iii)* unmodeled occluders or missing image evidence.

In terms of parameterization, we follow [23, 40], and represent the joint occlusion states of multiple vertices by a discrete set of 2D *occlusion masks* (visualized in blue in Figs. 1, 2, 4, 5), which constitutes a non-parametric approximation of the prior distribution over possible occlusion patterns. The set of masks is denoted $\{a_i\}$ (with a_0 being the empty mask which leaves the object fully visible), and encompasses likely, re-occurring occlusion patterns such as an object being occluded from one side (Fig. 3(b), best viewed magnified), truncated by the image border (Fig. 4(d), bottom), or occluded in the middle by a post or tree (Fig. 3(c)). The occlusion state of part j is given by an indicator function $o_j(\boldsymbol{\theta}_{gp}, q_{az}, \mathbf{s}, a)$, with $\boldsymbol{\theta}_{gp}$ the ground plane parameters, q_{az} the object azimuth, $\mathbf{s}$ the object shape, and a the occlusion mask.

Since all object hypotheses reside in the same 3D coordinate system, mutual occlusions can be derived deterministically from their depth ordering (Fig. 2, top): we cast rays from the camera center to each wireframe vertex of all other objects, and record intersections with faces of any other object as an appropriate occlusion mask. Each object instance $\mathbf{h}^\beta = (\mathbf{q}, \mathbf{s}, a)$ comprises 2D translation and azimuth $\mathbf{q} = (q_x, q_z, q_{az})$ on the ground plane, shape parameters $\mathbf{s}$, and occlusion mask a. Accordingly, we write $\Gamma\big(\{\mathbf{h^1}, \mathbf{h^2}, \ldots, \mathbf{h^n}\}\backslash\mathbf{h}^\beta, \mathbf{h}^\beta, \boldsymbol{\theta}_{gp}\big)$, i.e. the operator Γ returns the index of the occlusion mask for $\mathbf{h}^\beta$ as a function of the other objects in a given scene estimate.

3.2. Probabilistic Formulation

All evidence in our model comes from object part detection, and the prior for allowable occlusions is given by per-object occlusion masks (Sect. 3.1).

Object likelihood. The likelihood of an object being present at a particular location in the scene is measured by responses of a bank of (viewpoint-independent) sliding-window part detectors, evaluated at projected image coordinates of the corresponding 3D wireframe vertices. Specifically, we use a multi-class random forest trained on dense shape-context descriptors [1]. The likelihood $\mathcal{L}(\mathbf{h}^\beta, \boldsymbol{\theta}_{gp})$ for an object $\mathbf{h}^\beta$ is the sum over the responses (log-likelihoods) of all visible parts, or a constant likelihood for occluded parts [23, 40]:

$$\mathcal{L}(\mathbf{h}^\beta, \boldsymbol{\theta}_{gp}) = \max_{\varsigma} \left[\frac{\sum_{j=1}^{m} \big(\mathcal{L}_v + \mathcal{L}_o\big)}{\sum_{j=1}^{m} o_j(\boldsymbol{\theta}_{gp}, q_{az}, \mathbf{s}, a_0)} \right]. \quad (1)$$

The denominator normalizes for the varying number of self-occluded parts at different viewpoints. $\mathcal{L}_v$ is the evidence $S_j(\varsigma, \mathbf{x}_j)$ for part j if it is visible, found by looking up the detection score at image location $\mathbf{x}_j$ and scale ς, normalized with the background score $S_b(\varsigma, \mathbf{x}_j)$. $\mathcal{L}_o$ assigns a fixed likelihood c to an occluded part:

$$\mathcal{L}_v = o_j(\boldsymbol{\theta}_{gp}, q_{az}, \mathbf{s}, a) \log \frac{S_j(\varsigma, \mathbf{x}_j)}{S_b(\varsigma, \mathbf{x}_j)}, \quad (2)$$

$$\mathcal{L}_o = \big(o_j(\boldsymbol{\theta}_{gp}, q_{az}, \mathbf{s}, a_0) - o_j(\boldsymbol{\theta}_{gp}, q_{az}, \mathbf{s}, a)\big) c\,. \quad (3)$$

Scene-level likelihood. To score an entire scene we combine object hypotheses and ground plane into a scene hypothesis $\psi = \{q_y, \boldsymbol{\theta}_{gp}, \mathbf{h^1}, ..., \mathbf{h^n}\}$. The likelihood of a complete scene is then the sum over all object likelihoods, such that the objective for scene interpretation becomes:

$$\hat{\psi} = \arg\max_{\psi} \left[\sum_{\beta=1}^{n} \mathcal{L}(\mathbf{h}^\beta, \boldsymbol{\theta}_{gp}) \right]. \quad (4)$$

Note, the domain $Dom\big(\mathcal{L}\big(\mathbf{h}^\beta, \boldsymbol{\theta}_{gp}\big)\big)$ must be limited such that the occluder mask a^β of an object hypothesis $\mathbf{h}^\beta$ is dependent on relative poses of all the objects in the scene: an object hypothesis $\mathbf{h}^\beta$ can only be assigned occlusion masks a_i which respect object-object occlusions—*i.e*. at least all the vertices covered by $\Gamma\big(\{\mathbf{h^1}, \mathbf{h^2}, \ldots, \mathbf{h^n}\}\backslash\mathbf{h}^\beta, \mathbf{h}^\beta, \boldsymbol{\theta}_{gp}\big)\big)$ must be covered, even if a different mask would give a higher objective value. Also note that the ground plane in our current implementation is a hard constraint—objects off the ground are impossible in our parameterization.

3.3. Inference

The objective (4) is high-dimensional, non-convex, and not smooth (due to the binary occlusion states). Note that deterministic occlusion reasoning potentially introduces dependencies between all pairs of objects, and the common ground plane effectively ties all other variables to the ground plane parameters $\boldsymbol{\theta}_{gp}$. In order to still do approximate inference and reach strong local maxima of the likelihood function, we have designed an inference scheme that proceeds in stages, lifting an initial 2D guess (*Initialization*) about object locations to a coarse 3D model (*Coarse 3D geometry*), and refining that coarse model into a final collection of consistent 3D shapes (*Final scene-level inference*).

Initialization. Like many other complex systems with non-convex objective functions, ours relies on a good initialization that harvests as much information as possible from the available image evidence. We obtain this initialization with a large number of dedicated, viewpoint-dependent detectors for re-occurring configurations of wireframe vertices, similar in spirit to poselets [5, 40]. We train a bank of single-component DPM [12] detectors, with separate detectors for full objects as well as different degrees of partial occlusion (up to $\approx 80\%$) in order to ensure high recall and a large number of object hypotheses to chose from. Each of these detectors predicts both a (full-object) 2D bounding box and a viewpoint, discretized to 8 azimuth directions. We fix the number of object instances detected at this stage (using thresholding on the detection scores), and do not change this number during inference.

Coarse 3D geometry. Since we reason in a fixed, camera-centered 3D coordinate frame, the initial detections can be lifted to 3D space, by casting rays through 2D bounding box centers and instantiating objects on these rays, such that their reprojections are consistent with 2D boxes, and reside on a common ground plane. We perform grid search over ground plane parameters, object locations and viewpoints, jointly for all initial detections: a 3D bounding box with the mean dimensions of our 3D shape model is projected into the image for a small range of viewpoints (close to the initialization viewpoint) and over a range of ground planes. We keep track of all 3D hypotheses whose projected centroids and scales are (within some tolerance) consistent with the 2D bounding boxes. We point out that we lift to actual metric 3D scene coordinates, unlike some other works, *e.g.* [40], that work in $(x, y, scale)$-space.

The resulting 3D scene hypotheses (ground plane + compatible pose of each detected car) in turn are the starting point for the final inference stage. Rather than committing to a single best scene hypothesis, we keep a number of promising hypotheses and maintain that set of "scene particles" for further inference, in the spirit of [22]. All occlusion masks are initialized to a_0.

Final scene-level inference. The final inference procedure is based on block coordinate descent to decouple the shape and viewpoint variables from the occlusion masks, combined with ideas from smoothing-based optimization [24]. As the set of scene particles already covers the sensible range of ground planes, the ground plane parameters of the individual particles are kept fixed and not further optimized. This stabilizes the optimization. Each particle is iteratively refined in two steps: first, the shape and viewpoint parameters of all objects are updated, by testing many random perturbations around the current values and keeping the best one. The random perturbations follow a normal distribution that is adapted in a data-driven fashion ("smoothing-based optimization" [24]). Then, object occlusions are recomputed and occlusions by unmodeled objects are updated, by exhaustive search over the set of possible masks. For each scene particle these two update steps are iterated, and the particle with the highest objective value ψ forms our MAP estimate. Alg. 1 summarizes the optimization scheme.

```
Given: Scene particle $\psi'$: initial objects $\mathbf{h}^\beta = (\mathbf{q}^\beta, \mathbf{s}^\beta, a^\beta)$,
$\beta = 1 \dots n$; fixed $\boldsymbol{\theta}_{gp}$; $a^\beta = a_0$ (all objects fully visible)
for fixed number of iterations do
    1. for $\beta = 1 \dots n$ do
        draw samples $\{\mathbf{q}_j^\beta, \mathbf{s}_j^\beta\}^{j=1..m}$ from a Gaussian
            $\mathcal{N}(\mathbf{q}^\beta, \mathbf{s}^\beta; \Sigma^\beta)$ centered at current values;
        update $\mathbf{h}^\beta = \text{argmax}_j \ \mathcal{L}(\mathbf{h}^\beta(\mathbf{q}_j^\beta, \mathbf{s}_j^\beta, a^\beta), \boldsymbol{\theta}_{gp})$
    end
    2. for $\beta = 1 \dots n$ do
        update occlusion mask (exhaustive search)
        $a^\beta = \text{argmax}_j \ \mathcal{L}(\mathbf{h}^\beta(\mathbf{q}^\beta, \mathbf{s}^\beta, a_j), \boldsymbol{\theta}_{gp})$
    end
    3. Recompute sampling variance $\Sigma^\beta$ of Gaussians [24]
end
```

Algorithm 1: Inference run for each scene particle.

4. Experiments

In the following, we evaluate the ability of our fine-grained 3D scene representation to recover 3D object location and pose in challenging street scenes from a public data set [15], given a single image of the scene of interest and known camera intrinsics. Note that this task is a lot more demanding than 2D image space localization, since it involves monocular estimation of the depth w.r.t. the camera as well as continuous metric viewpoint estimation. The evaluation is divided into three parts: first, we verify that the first stage of our pipeline, object pre-detection, is on par with the state-of-the-art in terms of 2D bounding box localization (Sect. 4.2). Second, we evaluate how accurately different variants of our model can localize objects in 3D and estimate their viewpoints, outperforming corresponding 3D bounding box-based baselines by significant margins (Sect. 4.3). And third, we explore the relation between 3D performance and 2D image space prediction of the individual object parts in our 3D model (Sect. 4.4).

4.1. Dataset

We use the KITTI *3D object detection and orientation estimation* benchmark dataset [15] as a testbed for our approach, since it provides challenging images of realistic street scenes with varying levels of occlusion and clutter, but nevertheless controlled enough conditions for thorough evaluation. It consists of around 7500 training and 7500 test images of street scenes captured from a moving vehicle and comes with labeled 2D and 3D object bounding boxes and viewpoints (generated with the help of a laser scanner).

Test set. To investigate the effect of fine-grained 3D scene modeling, we chose a subset of the original training set, for which we manually annotate part positions and part occlusions for all cars in the images. We need images with multiple cars that are large enough to identify their parts. Given the large annotation effort, we select every 5th image from the training set with at least two cars with height greater than 75 pixels, resulting in 260 images with 982 cars, of which 672 are partially occluded. This ensures that, while being biased towards more complex scenes, we still sample a representative portion of the dataset.[1]

Training set. To train the DPM detectors for initialization (*c.f*. Sect. 3.3), we utilize a labeled dataset of 1000 car images downloaded from the internet, and 150 images from the KITTI dataset (none of which are part of the test set).

4.2. Object Pre-Detection

As a sanity check, we first verify that our 2D pre-detection matches the state-of-the-art. To that end we evaluate a standard 2D bounding box detection task according to PASCAL VOC criteria [11] (50% intersection-over-union between predicted and ground truth bounding boxes). As normally done we restrict the evaluation to objects of a certain minimum size and visibility. Specifically, we only consider cars $>$ 50 pixels in height which are at least 20% visible. The minimum size is slightly stricter than the 40 pixels that [15] use for the dataset (since we need to ensure enough support for the part detectors), whereas the occlusion threshold is much more lenient than their 80% (since we are specifically interested in occluded objects).

Results. We compare our bank of single component DPM detectors to the original deformable part model (DPM [12]), both trained on the same training set (Sec. 4.1). Precision-recall curves are shown in Fig. 3(b). We observe that our detector bank (green curve, 57.8% AP) in fact performs slightly better than original DPM [12] (red curve, 57.3% AP). In addition it delivers coarse viewpoint estimates and rough part locations that we can leverage (Sec. 3.3).

4.3. 3D Evaluation

Having verified that our pre-detection stage is competitive and provides reasonable object candidates in the 2D image plane, we now move on to the more challenging task of estimating the 3D location and pose of objects from monocular images with known camera intrinsics. As we will show, the fine-grained representation leads to significant performance improvements over a standard baseline that considers only 3D bounding boxes, on both tasks.

(a)

total: 982	inliers after	3D localization		VP estimation err.	
detected: **517**	inference	<1m	<2m	<10°	median
(i) FG+SO	94%	26%	47%	**66%**	6°
(ii) FG+SO+DO	93%	25%	47%	65%	**5°**
(iii) FG+SO+GP	94%	40%	65%	**66%**	**5°**
(iv) FG+SO+DO+GP	**96%**	**44%**	**67%**	65%	6°
(v) Zia et. al [40]	**96%**	—	—	—	—
(vi) COARSE	—	21%	45%	38%	13°
(vii) COARSE+GP	—	35%	66%	51%	10°

(b)

total: 672	inliers after	3D localization		VP estimation err.	
detected: **234**	inference	<1m	<2m	<10°	median
(i) FG+SO	94%	23%	44%	62%	6°
(ii) FG+SO+DO	93%	24%	44%	62%	6°
(iii) FG+SO+GP	94%	39%	62%	62%	**5°**
(iv) FG+SO+DO+GP	**96%**	**44%**	**63%**	**65%**	**5°**
(v) Zia et. al [40]	**96%**	—	—	—	—
(vi) COARSE	—	21%	49%	41%	13°
(vii) COARSE+GP	—	28%	60%	51%	10°

Table 1. 3D localization & viewpoint estimation accuracy: (a) full dataset, (b) occluded cars only. Best values per column in bold.

4.3.1 3D Object Localization

In order to isolate the contributions of individual components of the scene model, we evaluate and compare the following methods in all following experiments (Tab. 1): *(i)* FG+SO: the basic version of our fine-grained 3D object model with search for occluders, however without ground plane and deterministic occlusion reasoning; this amounts to independent modeling of the objects in a common, metric 3D scene coordinate system. *(ii)* FG+SO+DO: same as (i) but with deterministic occlusion reasoning between objects. *(iii)* FG+SO+GP: same as (i) but with common ground plane. *(iv)* FG+SO+DO+GP: same as (i), but with both deterministic occlusion reasoning and ground plane. *(v)* the pseudo-3D shape model of [40], with probabilistic occlusion reasoning; this uses essentially the same object model as (i), but fits it in 2D $(x, y, scale)$-space rather explicitly recovering a 3D scene interpretation. We also compare our representation to two different baselines, *(vi)* COARSE: a scene model consisting of 3D bounding boxes rather than detailed cars, corresponding to a fine grid search over pose parameters using the mean car shape, that project to the 2D bounding box from initialization (Sec. 3.3); and *(vii)* COARSE+GP: like *(vi)* but with a common ground plane for the bounding boxes.

Protocol. We measure 3D localization performance by the fraction of detected objects that are correctly localized on the ground plane up to deviations of 1, and 2 meters. These thresholds may seem rather strict for the viewing geometry of KITTI, but in our view larger tolerances make little sense for cars with dimensions $\approx 4.0 \times 1.6$ meters.

In line with existing studies on pose estimation, we base the analysis on true positive (TP) initializations that meet the PASCAL VOC criterion for 2D bounding box overlap and whose coarse viewpoint estimate lies within 45° of the ground truth, thus excluding failures of pre-detection. We

[1] The data, ground truth labels, and code are made available at http://www.igp.ethz.ch/photogrammetry/downloads.

perform the analysis for two settings (Tab. 1): *(a)* over our full testset (517 of 982 TPs), and *(b)* only over those cars that are partially occluded (234 of 672 TPs).

Results. In Tab. 1, we first observe that in terms of localization our full system (FG+SO+DO+GP) is the top performer in both settings and with both thresholds, localizing objects with 1 m accuracy in 44% of the cases and with 2 m accuracy in 63–67% of the cases.

Second, the basic fine-grained model FG+SO outperforms COARSE by 2–5 percent points (pp) corresponding to a relative improvement of 6–22% at 1 m accuracy, and on the full dataset also at 2 m accuracy (albeit by a more moderate 5%). The same applies when adding ground plane: FG+SO+GP outperforms COARSE+GP by 5–11 pp (14–38%) at 1 m accuracy. In other words, cars are not 3D boxes. Modeling their detailed shape and pose yields better scene descriptions, with and without ground plane constraint. The results at 2 m are less clear-cut. It appears that from badly localized initializations just inside the 2 m radius, the final inference sometimes drifts into incorrect local minima outside of 2 m.

Third, FG+SO+DO+GP brings further improvement. At 1 m it outperforms the next best coarse model COARSE+GP by a remarkable 9–16pp (25–55%). Notably, the gain is largest at high accuracy and on occluded objects. Fig. 4 confirms these results qualitatively: the bird's eye views (cols. c & e) show clearly improved agreement between the models estimates (red) and the ground truth (green).

And fourth, adding the ground plane always boosts the performance for both the FG+SO and COARSE models, strongly supporting the case for joint 3D scene reasoning: at 1 m accuracy the gains are 14 pp (FG+SO *vs.* FG+SO+GP), 19 pp (FG+SO+DO *vs.* FG+SO+DO+GP), and 14 pp (COARSE vs. COARSE+GP) on the full data set (for qualitative results see Fig. 5). For the fine-grained models this trend persists under occlusion (in fact the gains are slightly larger for occluded cars), whereas the coarse model benefits a lot less from the ground plane with occlusion: 16 pp (FG+SO *vs.* FG+SO+GP), 20 pp (FG+SO+DO *vs.* FG+SO+DO+GP), but only 7 pp (COARSE vs. COARSE+GP).

Finally, deterministic occlusion reasoning when coupled with ground plane considerably improves performance (FG+SO+GP *vs.* FG+SO+DO+GP): at 1 m accuracy the gains are 4–5pp (9–12%). Not surprisingly, deterministic occlusion reasoning only helps when the location estimates are already reasonably good (*i.e.* when ground plane assumption is used); otherwise part occlusions are already captured well by the latent occlusion variables of the basic model. This explains the negligible difference in localization accuracy for FG+SO+DO as compared to FG+SO.

Note that in some images the base detector finds only one car. In that case the scene model cannot bring any improvement, but also does not deteriorate the result.

We obtain even richer 3D "reconstructions" by replacing wireframes with nearest database 3D CAD models Fig. 3(b-c), accurately recognizing hatchbacks (b1, b4, c2), sedans (b2, c1) and approximating the van (b3) by a station wagon.

4.3.2 Viewpoint Estimation

Beyond 3D location, 3D scene interpretation also requires the viewpoint of every object, or equivalently its orientation in metric 3D space. Many object classes are elongated, thus their orientation is valuable at different levels, ranging from low-level tasks such as detecting occlusions and collisions to high-level ones like enforcing long-range regularities (*e.g.* cars parked at the roadside are usually parallel).

Protocol. We measure viewpoint estimation accuracy in two ways: as the percentage of detected objects for which the angular error is below 10°, and as the medianangular error between estimated and ground truth azimuth angle, averaged over detected objects.

Results. In Tab. 1, we first observe that the full system FG+SO+DO+GP outperforms the best coarse model COARSE+GP by significant margins of 14 pp on both the full dataset and on occluded objects, decreasing the median error by 4–5°. The qualitative results in Fig. 4 again confirm this. Second, all FG+SO models as well as [40] deliver quite reliable viewpoint estimates with only minor differences in performance ($\leq$ 3 pp, or 1°). And third, the ground plane helps considerably for the COARSE models, improving by 10–13 pp, decreasing the median error by 3°.

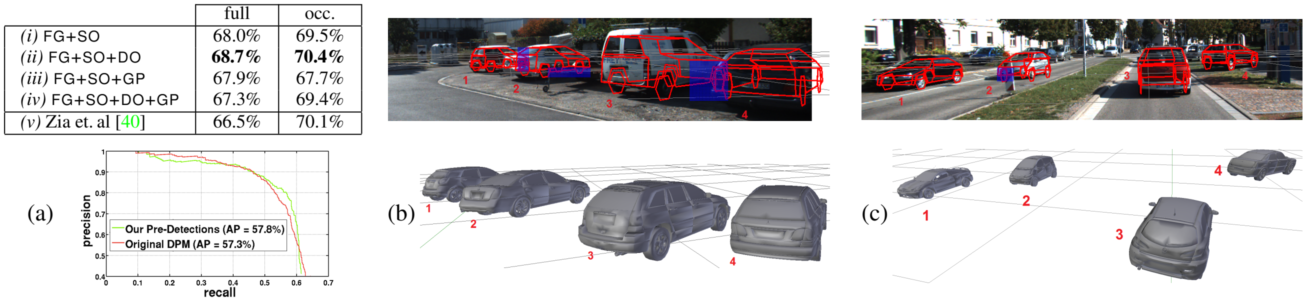
Estimates of 3D orientation are not provided by Zia et. al [40], however the viewpoint estimates in 2D image space (apparent azimuth of the object as seen in the image) are given. In comparison our full system FG+SO+DO+GP improves viewpoint estimates by 6–13 pp while decreasing the median error by 2–3° advocating detailed 3D reasoning even for 2D viewpoint estimation.

4.4. 2D Evaluation

While the objective of this work is to enable accurate localization and pose estimation in 3D (Sec. 4.3), we also present an analysis of 2D performance (part localization and occlusion prediction in the image plane), since such 2D measures are sometimes used in the context of monocular 3D modeling. An interesting finding is that significantly better 3D localization (Sec. 4.3.1) due to scene-level constraints does not translate to better localization of reprojected parts in the image plane. Rather, the correlation is weak and if anything slightly negative. In other words, whenever possible *3D reasoning should be evaluated in 3D space*, rather than in 2D projection.

To quantify the localization accuracy of object parts in the 2D image plane we count how many of the 36 parts that make up our deformable car model match manual annotations in the images. Effectively, we thus evaluate goodness-of-fit of the estimated deformable model's reprojection.[2]

[2] Note, there is no 3D counterpart to this part-level evaluation, since we see no way to obtain sufficiently accurate 3D part annotations.



	full	occ.
(i) FG+SO	68.0%	69.5%
(ii) FG+SO+DO	**68.7%**	**70.4%**
(iii) FG+SO+GP	67.9%	67.7%
(iv) FG+SO+DO+GP	67.3%	69.4%
(v) Zia et. al [40]	66.5%	70.1%

Figure 3. (a) Part localization accuracy (top), 2D pre-detection (bottom). (b-c) Example detections and corresponding 3D reconstructions.

Protocol. We follow the evaluation protocol commonly applied for human body pose estimation and report the average percentage of correctly localized parts (PCP), using a relative threshold adjusted to the size of the reprojected car (20 pixels for a car of size 500 × 170 pixels, *i.e.* ≈ 4% of the total length, as in [40]).

Results. In Fig. 3(a), we observe relatively small differences in performance across different model variants. Interestingly, there is a slight tendency for 3D models without ground plane assumption to perform better: for the full test set as well as for occluded set, FG+SO and FG+SO+DO outperform FG+SO+GP and FG+SO+DO+GP. Although the trend is weak and needs to be investigated further, it does at first glance seem surprising, as it is negatively correlated with 3D performance (Sec. 4.3). Taking into account the strongly non-linear relation between 2D and 3D errors, especially in depth direction along the camera axis, the result in fact confirms intuition: more flexible models have greater freedom to (over-)fit image evidence of individual cars. But sacrificing legitimate scene-level constraints comes at a cost, since they can no longer stabilize the more brittle monocular 3D reasoning.

We also note in passing that part occlusions are consistently predicted well by all models, with ≈ 87% correct predictions on full dataset, and ≈ 81% on occluded cars.

5. Conclusion

We have approached the 3D scene understanding problem from the perspective of deformable shape modeling, jointly fitting shapes of multiple objects linked by a common scene geometry (ground plane). Our results suggest that detailed representations of object shape are highly beneficial for 3D scene reasoning, and fit well with scene-level constraints between objects. By itself, fitting a detailed, deformable 3D model of cars resulted in improvements of 6–22% in object localization accuracy, over a baseline which just lifts objects' bounding boxes into the 3D scene. Enforcing a common ground plane for all 3D bounding boxes improved localization by 33–66%. When both aspects are combined into a joint model over multiple cars on a common ground plane, each with its own detailed 3D shape and pose, we get a striking 104–108% improvement in 3D localization compared to just lifting 2D detections, as well as a reduction of the orientation error from 13° to 5°. We also find that the increased accuracy in 3D scene coordinates is not reflected in improved 2D localization of the shape model's parts, supporting our claim that 3D scene understanding should be carried out (and evaluated) in an explicit 3D representation.

An obvious limitation of the present system, to be addressed in future work, is that it only includes a single object category, and applies to the simple (albeit important) case of scenes with a dominant ground plane. In terms of technical approach it is desirable to develop a better and more efficient inference algorithm for the joint scene model. Finally, the bottleneck where most of the recall is lost is the 2D predetection stage. Hence, either better 2D object detectors are needed, or 3D scene estimation must be extended to run directly on entire images without initialization, which will require greatly increased robustness and efficiency.

Acknowledgements. This work has been supported by the Max Planck Center for Visual Computing & Communication.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. *CVPR 2009*.

[2] A.Pentland. Perceptual organization and representation of natural form. *AI'86*.

[3] S. Y. Bao and S. Savarese. Semantic structure from motion. *CVPR11*.

[4] S. Y. Bao, Y. Xiang, and S. Savarese. Object co-detection. *ECCV 2012*.

[5] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. *ICCV 2009*.

[6] R. A. Brooks. Symbolic reasoning among 3-d models and 2-d images. *AI'81*.

[7] W. Choi, Y. W. Chao, C. Pantofaru, and S. Savarese. Understanding indoor scenes using 3d geometric phrases. In *CVPR 2013*.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR 2005*.

[9] L. Del Pero, J. Bowdish, B. Kermgard, E. Hartley, and K. Barnard. Understanding Bayesian rooms using composite 3d object models. *CVPR 2013*.

[10] A. Ess, B. Leibe, K. Schindler, and L. V. Gool. Robust multi-person tracking from a mobile platform. *PAMI*, 31(10):1831–1846, 2009.

[11] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[12] P. F. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9), 2010.

[13] S. Fidler, S. Dickinson, and R. Urtasun. 3d object detection and viewpoint estimation with a deformable 3d cuboid model. *NIPS 2012*.

[14] T. Gao, B. Packer, and D. Koller. A segmentation-aware object detection model with occlusion handling. *CVPR 2011*.

[15] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. *CVPR 2012*.

[16] A. Geiger, C. Wojek, and R. Urtasun. Joint 3d estimation of objects and scene layout. In *NIPS 2011*.

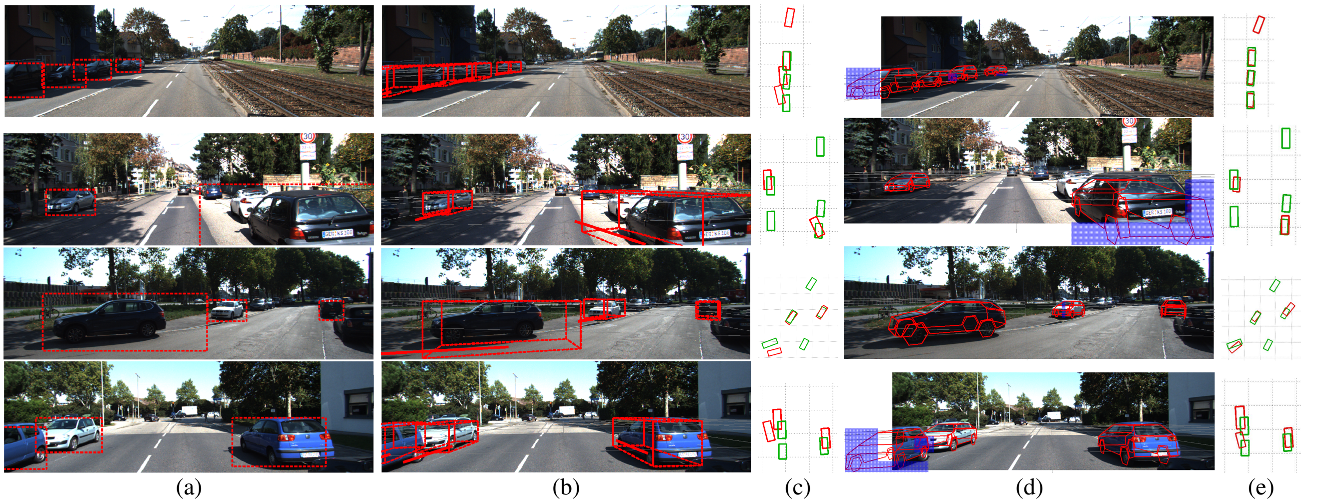


Figure 4. COARSE+GP (a-c) *vs*. FG+SO+DO+GP (d,e). (a) 2D bounding box detections, (b) COARSE+GP based on (a), (c) bird's eye view of (b). (e) FG+SO+DO+GP shape model fits (blue: estimated occlusion masks), (f) bird's eye view of (e). Estimates in red, ground truth in green.

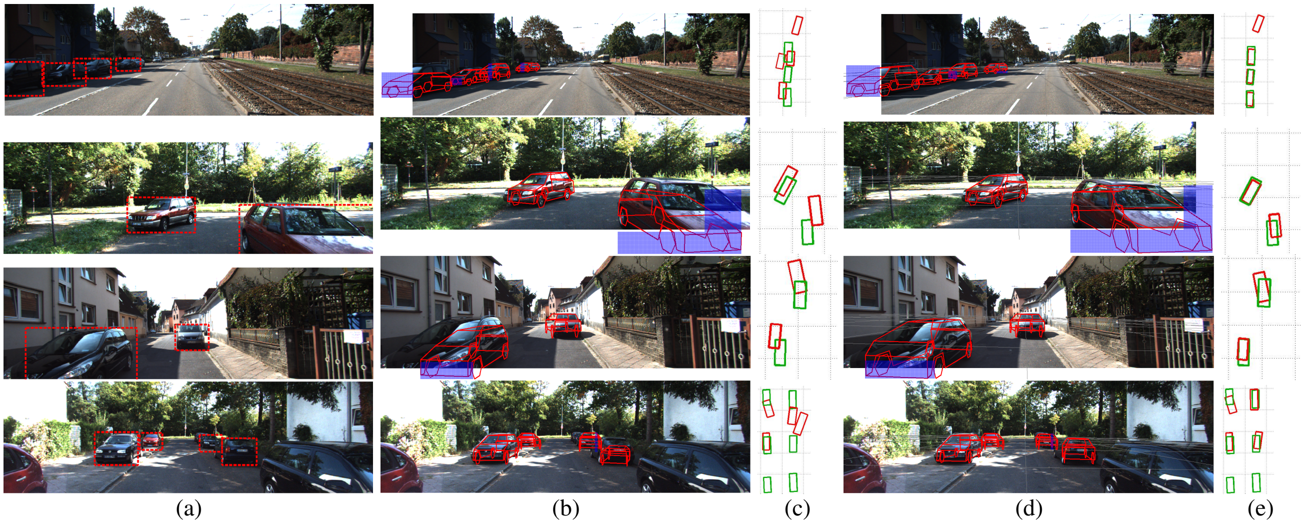


Figure 5. FG+SO+DO (a-c) *vs*. FG+SO+DO+GP (d,e). (a) 2D bounding box detections, (b) FG+SO+DO based on (a), (c) bird's eye view of (b). (d) FG+SO+DO+GP shape model fits (blue: estimated occlusion masks), (e) bird's eye view of (d). Estimates in red, ground truth in green.

[17] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. Object detection with grammar models. *NIPS 2011*.

[18] A. Gupta, A. A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. *ECCV 2010*.

[19] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. *ECCV 2010*.

[20] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. *IJCV*, 80(1):3–15, 2008.

[21] E. Hsiao and M. Hebert. Occlusion reasoning for object detection under arbitrary viewpoint. *CVPR 2012*.

[22] M. Isard and A. Blake. CONDENSATION – conditional density propagation for visual tracking. *IJCV*, 1(29), 1998.

[23] S. Kwak, W. Nam, B. Han, and J. H. Han. Learning occlusion with likelihoods for visual tracking. *ICCV 2011*.

[24] M. Leordeanu and M. Hebert. Smoothing-based optimization. *CVPR 2008*.

[25] D. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31(3), 1987.

[26] D. Marr and H. K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *PRSL B*, 200:269–294, 1978.

[27] D. Meger, C. Wojek, B. Schiele, and J. Little. Explicit occlusion reasoning for 3d object detection. *BMVC 2011*.

[28] M.Hejrati and D.Ramanan. Analyzing 3d objects in cluttered images. *NIPS12*.

[29] B. Pepik, P. Gehler, M. Stark, and B. Schiele. 3DDPM - 3d deformable part models. *ECCV 2012*.

[30] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Occlusion patterns for object class detection. In *CVPR 2013*.

[31] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. *ECCV 2012*.

[32] M. Stark, J. Krause, B. Pepik, D. Meger, J. Little, B. Schiele, and D. Koller. Fine-grained categorization for 3d scene understanding. In *BMVC 2012*.

[33] A. Vedaldi and A. Zisserman. Structured output regression for detection with partial truncation. *NIPS 2009*.

[34] H. Wang, S. Gould, and D. Koller. Discriminative learning with latent variables for cluttered indoor scene understanding. *ECCV 2010*.

[35] X. Wang, T. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. *ICCV 2009*.

[36] C. Wojek, S. Walk, S. Roth, K. Schindler, and B. Schiele. Monocular visual scene understanding: understanding multi-object traffic scenes. *PAMI*, 2013.

[37] Y. Xiang and S. Savarese. Estimating the aspect layout of object categories. *CVPR 2012*.

[38] Y. Zhao and S.-C. Zhu. Scene parsing by integrating function, geometry and appearance models. In *CVPR 2013*.

[39] M. Z. Zia, M. Stark, B. Schiele, and K. Schindler. Detailed 3d representations for object recognition and modeling. *PAMI*, 35(11):2608–2623, 2013.

[40] M. Z. Zia, M. Stark, and K. Schindler. Explicit occlusion modeling for 3d object class representations. In *CVPR 2013*.