International Journal of Computer Vision manuscript No.
(will be inserted by the editor)

Towards Scene Understanding with Detailed 3D Object Representations

M. Zeeshan Zia · Michael Stark · Konrad Schindler

Received: date / Accepted: date

Abstract Current approaches to semantic image and scene understanding typically employ rather simple object representations such as 2D or 3D bounding boxes. While such coarse models are robust and allow for reliable object detection, they discard much of the information about objects' 3D shape and pose, and thus do not lend themselves well to higher-level reasoning. Here, we propose to base scene understanding on a high-resolution object representation. An object class – in our case cars — is modeled as a deformable 3D wireframe, which enables fine-grained modeling at the level of individual vertices and faces. We augment that model to explicitly include vertex-level occlusion, and embed all instances in a common coordinate frame, in order to infer and exploit object-object interactions. Specifically, from a single view we jointly estimate the shapes and poses of multiple objects in a common 3D frame. A ground plane in that frame is estimated by consensus among different objects, which significantly stabilizes monocular 3D pose estimation. The fine-grained model, in conjunction with the explicit 3D scene model, further allows one to infer part-level occlusions between the modeled objects, as well as occlusions by other, unmodeled scene elements. To demonstrate the benefits of such detailed object class models in the context of scene understanding we systematically evaluate our approach on the challenging KITTI street scene dataset. The experiments show that the model's ability to utilize image evidence at the level of individual parts improves monocular 3D pose estimation w.r.t. both location and (continuous) viewpoint.

M. Zeeshan Zia
ETH Zurich and Imperial College London
E-mail: zeeshan.zia@imperial.ac.uk

Michael Stark
Max Planck Institute for Informatics, Saarbrücken, Germany
E-mail: stark@mpi-inf.mpg.de

Konrad Schindler
Swiss Federal Institute of Technology (ETH), Zurich
E-mail: konrad.schindler@geod.baug.ethz.ch

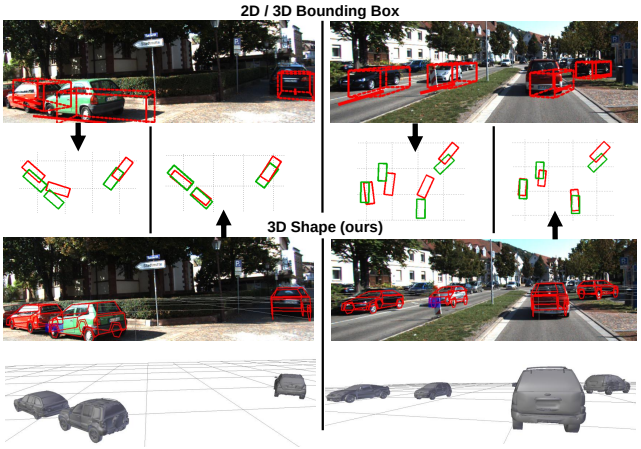


Fig. 1: *Top:* Coarse 3D object bounding boxes derived from 2D bounding box detections (not shown). *Bottom:* our fine-grained 3D shape model fits improve 3D localization (see bird's eye views).

1 Introduction

The last ten years have witnessed great progress in automatic visual recognition and image understanding, driven by advances in local appearance descriptors, the adoption of discriminative classifiers, and more efficient techniques for probabilistic inference. In several different application domains we now have semantic vision sub-systems that work on real-world images. Such powerful tools have sparked a renewed interest in the grand challenge of visual 3D scene understanding. Meanwhile, individual object detection performance has reached a plateau after a decade of steady

gains (Everingham et al. 2010), further emphasizing the need for contextual reasoning.

A number of geometrically rather coarse scene-level reasoning systems have been proposed over the past few years (Hoiem et al. 2008; Wang et al. 2010; Hedau et al. 2010; Gupta et al. 2010; Silberman et al. 2012), which apart from adding more holistic scene understanding also improve object recognition. The addition of context and the step to reasoning in 3D (albeit coarsely) makes it possible for different vision sub-systems to interact and improve each other's estimates, such that the sum is greater than the parts.

Very recently, researchers have started to go one step further and increase the level-of-detail of such integrated models, in order to make better use of the image evidence. Such models learn not only 2D object appearance but also detailed 3D shape (Xiang and Savarese 2012; Hejrati and Ramanan 2012; Zia et al. 2013). The added detail in the representation, typically in the form of wireframe meshes learned from 3D CAD models, makes it possible to also reason at higher resolution: beyond measuring image evidence at the level of individual vertices/parts one can also handle relations between parts, e.g. shape deformation and part-level occlusion (Zia et al. 2013). Initial results are encouraging. It appears that the more detailed scene interpretation can be obtained at a minimal penalty in terms of robustness (detection rate), so that researchers are beginning to employ richer object models to different scene understanding tasks (Choi et al. 2013; Del Pero et al. 2013; Zhao and Zhu 2013; Xiang and Savarese 2013; Zia et al. 2014).

Here we describe one such novel system for scene understanding based on monocular images. Our focus lies on exploring the potential of jointly reasoning about multiple objects in a common 3D frame, and the benefits of part-level occlusion estimates afforded by the detailed representation. We have shown in previous work (Zia et al. 2013) how a detailed 3D object model enables a richer pseudo-3D $(x, y, scale)$ interpretation of simple scenes dominated by a single, unoccluded object—including fine-grained categorization, model-based segmentation, and monocular reconstruction of a ground plane. Here, we lift that system to true 3D, i.e. CAD models are scaled to their true dimensions in world units and placed in a common, metric 3D coordinate frame. This allows one to reason about geometric constraints between multiple objects as well as mutual occlusions, at the level of individual wireframe vertices.

Contributions. We make the following contributions.

First, we propose a viewpoint-invariant method for 3D reconstruction (shape and pose estimation) of severely occluded objects in single-view images. To obtain a complete framework for detection and reconstruction, the novel method is bootstrapped with a variant of the poselets framework (Bourdev and Malik 2009) adapted to the needs of our 3D object model.

Second, we reconstruct scenes consisting of multiple such objects, each with their individual shape and pose, in a single inference framework, including geometric constraints between them in the form of a common ground plane. Notably, reconstructing the fine detail of each object also improves the 3D pose estimates (location as well as viewpoint) for entire objects over a 3D bounding box baseline (Fig. 1).

Third, we leverage the rich detail of the 3D representation for occlusion reasoning at the individual vertex level, combining (deterministic) occlusion by other detected objects with a (probabilistic) generative model of further, unknown occluders. Again, integrated scene understanding yields improved 3D localization compared to independently estimating occlusions for each individual object.

And *fourth*, we present a systematic experimental study on the challenging KITTI street scene dataset (Geiger et al. 2012). While our fine-grained 3D scene representation can not yet compete with technically mature 2D bounding box detectors in terms of recall, it offers superior 3D pose estimation, correctly localizing $> 43\%$ of the detected cars up to 1 m and $> 55\%$ up to 1.5 m, even when they are heavily occluded.

Parts of this work appear in two preliminary conference papers (Zia et al. 2013, 2014). The present paper describes our approach in more detail, extends the experimental analysis, and describes the two contributions (extension of the basic model to occlusions, respectively scene constraints) in a unified manner.

The remainder of this paper is structured as follows. Sec. 2 reviews related work. Sec. 3 introduces our 3D geometric object class model, extended in Sec. 4 to entire scenes. Sec. 5 gives experimental results, and Sec. 6 concludes the paper.

2 Related Work

Detailed 3D object representations. Since the early days of computer vision research, detailed and complex models of object geometry were developed to solve object recognition in general settings, taking into account viewpoint, occlusion, and intra-class variation. Notable examples include the works of Kanade (1980) and Malik (1987), who lift line drawings of 3D objects by classifying the lines and their intersections to common occurring configurations; and the classic works of Brooks (1981) and Pentland (1986), who represent complex objects by combinations of atomic shapes, generalized cones and super-quadrics. Matching CAD-like models to image edges also made it possible to address partially occluded objects (Lowe 1987) and intra-class variation (Sullivan et al. 1995).

Unfortunately, such systems could not robustly handle real world imagery, and largely failed outside controlled lab environments. In the decade that followed researchers moved

to simpler models, sacrificing geometric fidelity to robustify the matching of the models to image evidence—eventually reaching a point where the best-performing image understanding methods were on one hand bag-of-features models without any geometric layout, and on the other hand object templates without any flexibility (largely thanks to advances in local region descriptors and statistical learning).

However, over the past years researchers have gradually started to re-introduce more and more geometric structure in object class models and improve their performance (*e.g.* Leibe et al. 2006; Felzenszwalb et al. 2010). At present we witness a trend to take the idea even further and revive highly detailed deformable wireframe models (Zia et al. 2009; Li et al. 2011; Zia et al. 2013; Xiang and Savarese 2012; Hejrati and Ramanan 2012). In this line of work, object class models are learnt from either 3D CAD data (Zia et al. 2009, 2013) or images (Li et al. 2011). Alternatively, objects are represented as collections of planar segments (also learnt from CAD models, Xiang and Savarese 2012) and lifted to 3D with non-rigid structure-from-motion. In this paper, we will demonstrate that such fine-grained modelling also better supports scene-level reasoning.

Occlusion modeling. While several authors have investigated the problem of occlusion in recent years, little work on occlusions exists for detailed part-based 3D models, notable exceptions being (Li et al. 2011; Hejrati and Ramanan 2012).

Most efforts concentrate on 2D bounding box detectors in the spirit of HOG (Dalal and Triggs 2005). Fransens et al. (2006) model occlusions with a binary visibility map over a fixed object window and infer the map with expectation-maximization. In a similar fashion, sub-blocks that make up the window descriptor are sometimes classified into occluded and non-occluded ones (Wang et al. 2009; Gao et al. 2011; Kwak et al. 2011). Vedaldi and Zisserman (2009) use a structured output model to explicitly account for truncation at image borders and predict a truncation mask at both training and test time. If available, motion (Enzweiler et al. 2010) and/or depth (Meger et al. 2011) can serve as additional cues to determine occlusion, since discontinuities in the depth and motion fields are more reliable indicators of occlusion boundaries than texture edges.

Even though quite some effort has gone into occlusion invariance for global object templates, it is not surprising that part-based models have been found to be better suited for the task. In fact even fixed windows are typically divided into regular grid cells that one could regard as "parts" (Wang et al. 2009; Gao et al. 2011; Kwak et al. 2011). More flexible models include dedicated DPMs for commonly occuring object-object occlusion cases (Tang et al. 2012) and variants of the extended DPM formulation (Girshick et al. 2011), in which an occluder is inferred from the absence of part evidence. Another strategy is to learn a very large number of partial configurations ("poselets") through clustering (Bourdev and Malik 2009), which will naturally also include frequent occlusion patterns. The most obvious manner to handle occlusion in a proper part-based model is to explicitly estimate the oclusion states of the individual parts, either via RANSAC-style sampling to find unoccluded ones (Li et al. 2011), or via local mixtures (Hejrati and Ramanan 2012). Here we also store a binary occlusion flag per part, but explicitly enumerate allowable occlusion patterns and restrict the inference to that set.

Qualitative scene representations. Beyond detailed geometric models of individual objects, early computer vision research also attempted to model entire scenes in 3D with considerable detail. In fact the first PhD thesis in computer vision (Roberts 1963) modeled scenes comprising of polyhedral objects, considering self-occlusions as well as combining multiple simple shapes to obtain complex objects. Koller et al. (1993) used simplified 3D models of multiple vehicles to track them in road scenes, whereas Haag and Nagel (1999) included scene elements such as trees and buildings, in the form of polyhedral models, to estimate their shadows falling on the road, as well as vehicle motion and illumination.

Recent work has revisited these ideas at the level of plane- and box-type models. E.g., Wang et al. (2010) estimate the geometric layout of walls in an indoor setting, segmenting out the clutter. Similarly, Hedau et al. (2010) estimate the layout of a room and reason about the locations of the bed as a box in the room. For indoor settings it has even been attempted to recover physical support relations, based on RGB-D data (Silberman et al. 2012). For fairly generic outdoor scenes, physical support, volumetric constraints and occlusions have been included, too, still using boxes as object models (Gupta et al. 2010). Also for outdoor images, Liu et al. (2014) partition single views into a set of oriented surfaces, driven by grammar rules for neighboring segments. It has also been observed that object detections carry information about 3D surface orientations, such that they can be jointly estimated even from a single image (Hoiem et al. 2008). Moreover, recent work suggests that object detection can be improved if one includes the density of common poses between neighboring object instances (Oramas et al. 2013).

All the works indicate that even coarse 3D reasoning allows one to better guess the (pseudo-)3D layout of a scene, while at the same time improving 2D recognition. Together with the above-mentioned strength of fine-grained shape models when it comes to occlusion and viewpoint, this is in our view a compelling reason to add 3D contextual constraints also to those fine-grained models.

Quantitative scene representations. A different type of methods also includes scene-level reasoning, but is tailored to specific applications and is more quantitative in nature. Most

works in this direction target autonomous navigation, hence precise localization of reachable spaces and obstacles is important. Recent works for the autonomous driving scenario include: (Ess et al. 2009), in which multi-pedestrian tracking is done in 3D based on stereo video, and (Geiger et al. 2011; Wojek et al. 2013), both aiming for advanced scene understanding including multi-class object detection, 3D interaction modeling, as well as semantic labeling of the image content, from monocular input. Viewpoint estimates from semantic recognition can also be combined with interest point detection to improve camera pose and scene geometry even across wide baselines (Bao and Savarese 2011).

For indoor settings, a few recent papers also employ detailed object representations to support scene understanding (Del Pero et al. 2013), try to exploit frequently co-occurring object poses (Choi et al. 2013), and even supplement geometry and appearance constraints with affordances to better infer scene layout (Zhao and Zhu 2013).

3 3D Object Model

We commence by introducing the fine-grained 3D object model that lies at the core of our approach. Its extension to entire multi-object scenes will be discussed in Sec. 4. By modeling an object class at the fine level of detail of individual wireframe vertices the object model provides the basis for reasoning about object extent and occlusion relations with high fidelity. To that end, we lift the pseudo-3D object model that we developed in Zia et al. (2013) to metric 3D space, and combine it with the explicit representation of likely occlusion patterns from Zia et al. (2013). Our object representation then comprises a model of global object geometry (Sec. 3.1), local part appearance (Sec. 3.2), and an explicit representation of occlusion patterns (Sec. 3.3). Additionally, the object representation also includes a grouping of local parts into semi-local part configurations (Sec. 3.4), which will be used to initialize the model during inference (Sec. 4.3). We depict the 3D object representation in Fig. 2.

3.1 Global Object Geometry

We represent an object class as a deformable 3D wireframe, as in the classical "active shape model" formulation (Cootes et al. 1995). The vertices of the wireframe are defined manually, and wireframe exemplars are collected by annotating a set of 3D CAD models (i.e., selecting corresponding vertices from their triangle meshes). Principal Component Analysis (PCA) is applied to obtain the mean configuration of vertices in 3D as well as the principal modes of their relative displacement. The final geometric object model then consists of the mean wireframe $\boldsymbol{\mu}$ plus the m principal component directions $\mathbf{p}_j$ and corresponding standard deviations

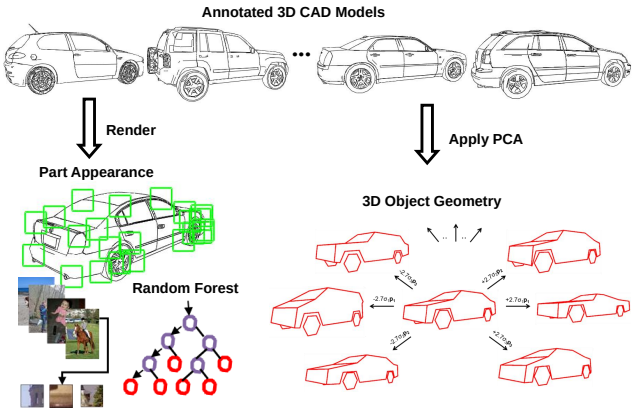


Fig. 2: 3D Object Model.

σ_j, where $1 \leq j \leq m$. Any 3D wireframe $\mathbf{X}$ can thus be represented, up to some residual $\boldsymbol{\epsilon}$, as a linear combination of r principal components with geometry parameters $\mathbf{s}$, where s_k is the weight of the k^{th} principal component:

$$\mathbf{X}(\mathbf{s}) = \boldsymbol{\mu} + \sum_{k=1}^{r} s_k \sigma_k \mathbf{p}_k + \boldsymbol{\epsilon} \tag{1}$$

Unlike the earlier Zia et al. (2013), the 3D CAD models are scaled according to their real world metric dimensions. [1]The resulting metric PCA model hence encodes physically meaningful scale information in world units, that allow one to assign absolute 3D positions to object hypotheses (given known camera intrinsics).

3.2 Local Part Appearance

We establish the connection between the 3D geometric object model (Sec. 3.1) and an image by means of a set of *parts*, one for each wireframe vertex. For each part, a multi-view appearance model is learned, by generating from training patches with non-photorealistic rendering of 3D CAD models from a large number of different viewpoints (Stark et al. 2010), and training a sliding-window detector on these patches.

Specifically, we encode patches around the projected locations of the annotated parts ($\approx 10\%$ in size of the full object width) as dense shape context features (Belongie et al. 2000). We learn a multi-class Random Forest classifier where each class represents the multi-view appearance of a particular part. We also dedicate a class trained on background patches, combining random real image patches with rendered non-part patches to avoid classifier bias. Using synthetic renderings for training allows us to densely sample the

[1] While in the earlier work they were scaled to the same size, so as to keep the deformations from the mean shape small.

relevant portion of the viewing sphere with minimal annotation effort (one time labeling of part locations on 3D CAD models, i.e. no added effort in creating the shape model).

3.3 Explicit Occluder Representation

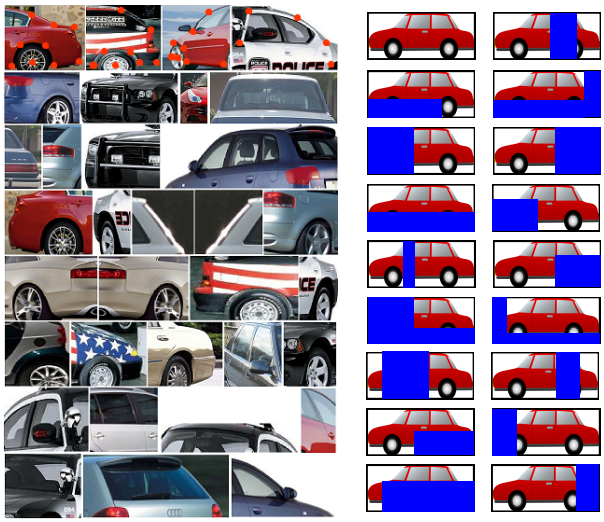
The 3D wireframe model allows one to represent partial occlusion at the level of individual parts: each part has an associated binary variable that stores whether the part is visible or occluded. Note that, in theory, this results in a exponential number of possible combinations of occluded and unoccluded parts, hindering efficient inference over occlusion states. We therefore take advantage of the fact that partial occlusion is not entirely random, but tends to follow reoccurring patterns that render certain joint occlusion states of multiple parts more likely than others (Pepik et al. 2013): the joint occlusion state depends on the shape of the occluding physical object(s).

Here we approximate the shapes of (hypothetical) occluders as a finite set of occlusion masks, following (Kwak et al. 2011; Zia et al. 2013). This set of masks constitutes a (hard) non-parameteric prior over possible occlusion patterns. The set is denoted by $\{a_i\}$, and for convenience we denote the empty mask which leaves the object fully visible by a_0. We sample the set of occlusion masks regularly from a generative model, by sliding multiple boxes across the mask in small spatial increments (the parameters of those boxes are determined empirically). Figure 3(b) shows a few out of the total 288 masks in our set, with the blue region representing the occluded portion of the object (car). The collection is able to capture different modes of occlusion, for example truncation by the image border (Fig. 8(d), first row), occlusion in the middle by a post or tree (Fig. 8(d), 2nd row), or occlusion of only the lower parts from one side (Fig. 8(d), third row).

Note that the occlusion mask representation is independent of the cause of occlusion, and allows to uniformly treat occlusions that arise from *(i)* self occlusion (a part is occluded by a wireframe face of the same object), *(ii)* occlusion by another object that is part of the same scene hypothesis (a part is occluded by a wireframe face of another object), *(iii)* occlusion by an unknown source (a part is occluded by an object that is not part of the same scene hypothesis, or image evidence is missing).

3.4 Semi-Local Part Configurations

In the context of people detection and pose estimation, it has been realized that individual body parts are hard to accurately localize, because they are small and often not discriminative enough in isolation (Bourdev and Malik 2009). Instead, it has proved beneficial to train detectors that span



(a) (b)

Fig. 3: (a) Individual training examples for a few part $configurations$ (top row shows labeled part locations), (b) example occlusion masks.

multiple parts appearing in certain poses (termed "poselets"), seen from a certain viewpoint, and selecting the ones that exhibit high discriminative power against background on a validation set (alternately, the scheme of Maji and Malik (2009) could also be used). In line with these findings, we introduce the notion of part *configurations*, i.e. semi-local arrangements of a number of parts, seen from a specific viewpoint, that are adjacent (in terms of wireframe topology). Some examples are depicted in Fig. 3(a)). These configurations provide more reliable evidence for each of the constituent parts than individual detectors. We use detectors for different configurations to find primising 2D bounding boxes and viewpoint estimates, as initializations for fitting the fine-grained 3D object models.

Specifically, we list all the possible configurations of 3-4 adjacent visible parts that are not smaller than $\approx 20\%$ of the full object (for the eight coarse viewpoints). Some configurations cover the full car, whereas others only span a part of it (down to $\approx 20\%$ of the full object). However we found the detection performance to be rather consistent even if other heuristics were used for part configuration generation. We then train a bank of single component DPM detectors, one for each configuration, in order to ensure high recall and a large number of object hypotheses to choose from. At test time, activations of these detectors are merged together through agglomerative clustering to form full object hypothesis, in the spirit of the poselet framework (Bourdev and Malik 2009). For training, we utilize a set of images labeled at the level of individual parts, and with viewpoint labels from a small discrete set (in our experiments 8 equally spaced viewpoints). All the objects in these images are fully

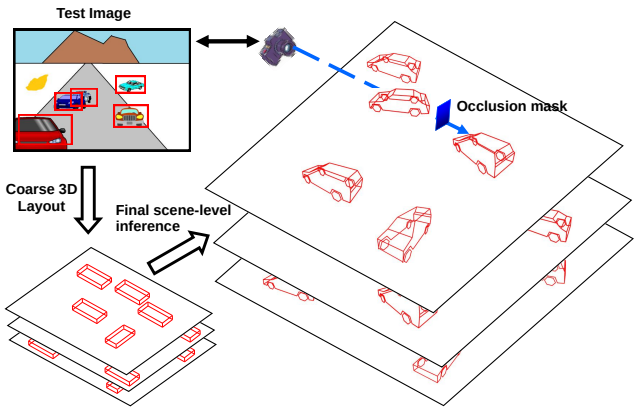


Fig. 4: 3D Scene Model.

visible. Thus, we can store the relative scale and bounding box center offsets, w.r.t. the full object bounding box, for the part-configuration examples. When detecting potentially occluded objects in a test image, the activations of all configuration detectors predict a full object bounding box and a (discrete) pose.

Next we recursively merge nearby (in $x, y, scale$) activations that have the same viewpoint. Merging is accomplished by averaging the predicted full object bounding box corners, and assigning it the highest of the detection scores. After this agglomerative clustering has terminated all clusters above a fixed detection score are picked as legitimate objects. Thus we obtain full object bounding box predictions (even for partially visible objects), along with an approximate viewpoint.

4 3D Scene Model

We proceed by extending the single object model of Sec. 3 to entire scenes, where we can jointly reason about multiple objects and their geometric relations, placing them on a common ground plane and taking into account mutual occlusions. As we will show in the experiments (Sec. 5), this joint modeling can lead to significant improvements in terms of 3D object localization and pose estimation compared to separately modeling individual objects. It is enabled by a joint scene hypothesis space (Sec. 4.1), governed by a probabilistic formulation that scores hypotheses according to their likelihood (Sec. 4.2), and an efficient approximate inference procedure for finding plausible scenes (Sec. 4.3). The scene model is schematically depicted in Fig. 4.

4.1 Hypothesis Space

Our 3D scene model comprises a common ground plane and a set of 3D deformable wireframes with corresponding occlusion masks (Sec. 3). Note that this hypothesis space is more expressive than the 2.5 D representations used by previous work (Ess et al. 2009; Meger et al. 2011; Wojek et al. 2013), as it allows reasoning about locations, shapes, and interactions of objects, at the level of individual 3D wireframe vertices and faces.

Common ground plane. In the full system, we constrain all the object instances to lie on a common ground plane, as often done for street scenes. This assumption usually holds 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. We parametrize the ground plane with the pitch and roll angles relative to the camera frame, $\boldsymbol{\theta}_{gp} = (\theta_{pitch}, \theta_{roll})$. The height q_y of the camera above ground is assumed known and fixed.

Object instances. Each object in the scene is an instance of the 3D wireframe model described in Sec. 3.1. An individual instance $\mathbf{h}^\beta = (\mathbf{q}, \mathbf{s}, a)$ comprises 2D translation and azimuth $\mathbf{q} = (q_x, q_z, q_{az})$ relative to the ground plane, shape parameters $\mathbf{s}$, and an occlusion mask a.

Explicit occlusion model. As detailed in Sec. 3.3, we represent occlusions on an object instance by selecting an occluder mask out of a pre-defined set $\{a_i\}$, which in turn determines the binary occlusion state of all parts. That is, 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. 4): 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. 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.

4.2 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 and relative object positions (Sec. 4.1).

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 (Sec. 3.2), evaluated at projected image coordinates of the corresponding 3D wireframe vertices.[2] The likelihood $\mathcal{L}(\mathbf{h}^\beta, \boldsymbol{\theta}_{gp})$ for an object $\mathbf{h}^\beta$ standing on the ground

[2] In practice this amounts to a look-up in the precomputed response maps.

plane $\boldsymbol{\theta}_{gp}$ is the sum over the responses of all visible parts, with a constant likelihood for occluded parts (m is the total number of parts, a_0 is the 'full visibility' occluder mask):

$$\mathcal{L}(\mathbf{h}^{\beta}, \boldsymbol{\theta}_{gp}) = \max_{\boldsymbol{\varsigma}} \left[\frac{\sum_{j=1}^{m} \big(\mathcal{L}_v + \mathcal{L}_o\big)}{\sum_{j=1}^{m} o_j(\boldsymbol{\theta}_{gp}, q_{az}, \mathbf{s}, a_0)} \right]. \quad (2)$$

The denominator normalizes for the varying number of self-occluded parts at different viewpoints. $\mathcal{L}_v$ is the evidence (pseudo log-likelihood) $S_j(\boldsymbol{\varsigma}, \mathbf{x}_j)$ for part j if it is visible, found by looking up the detection score at image location $\mathbf{x}_j$ and scale $\boldsymbol{\varsigma}$, normalized with the background score $S_b(\boldsymbol{\varsigma}, \mathbf{x}_j)$ as in (Villamizar et al. 2011). $\mathcal{L}_o$ assigns a fixed likelihood c, estimated by cross-validation on a held-out dataset:

$$\mathcal{L}_v = o_j(\boldsymbol{\theta}_{gp}, q_{az}, \mathbf{s}, a) \log \frac{S_j(\boldsymbol{\varsigma}, \mathbf{x}_j)}{S_b(\boldsymbol{\varsigma}, \mathbf{x}_j)} \ , \quad (3)$$

$$\mathcal{L}_o = \big(o_j(\boldsymbol{\theta}_{gp}, q_{az}, \mathbf{s}, a_0) - o_j(\boldsymbol{\theta}_{gp}, q_{az}, \mathbf{s}, a)\big) c \ . \quad (4)$$

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:

$$\hat{\psi} = \arg\max_{\psi} \left[\sum_{\beta=1}^{n} \mathcal{L}(\mathbf{h}^{\beta}, \boldsymbol{\theta}_{gp}) \right] . \quad (5)$$

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 (except for experiments in which we "turn off" the ground plane for comparison).

4.3 Inference

The objective function in Eqn. 5 is high-dimensional, highly 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, Occlusion Reasoning*).

Initialization. We initialize the inference from coarse 2D bounding box pre-detections and corresponding discrete viewpoint estimates (Sec. 3.4), keeping all pre-detections above a confidence threshold. Note that this implicitly determines the maximum number of objects that will be considered in the scene hypothesis under consideration.

Coarse 3D geometry. Since we reason in a fixed, camera-centered 3D coordinate frame, the initial detections can be directly 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 the 2D boxes and discrete viewpoint estimates, and reside on a common ground plane. In order to avoid discretization artifacts, we then refine the lifted object boxes by imputing the mean object shape and performing a grid search over ground plane parameters and object translation and rotation (azimuth). In this step, rather than commiting to a single scene-level hypothesis, we retain many candidate hypotheses (*scene particles*) that are consistent with the 2D bounding boxes and viewpoints of the pre-detections within some tolerance.

Occlusion reasoning. We combine two different methods to select an appropriate occlusion mask for a given object, *(i)* deterministic occlusion reasoning, and *(ii)* occlusion reasoning based on (the absence of) part evidence.

(i) Since by construction we recover the 3D locations and shapes of multiple objects in a common frame, we can calculate whether a certain object instance is occluded by any other modeled object instance in our scene. This is calculated efficiently by casting rays to all (not self-occluded) vertices of the object instance, and checking if a ray intersects any other object in its path before reaching the vertex. This deterministically tells us which parts of the object instance are occluded by another modeled object in the scene, allowing us to choose an occluder mask that best represents the occlusion (overlaps the occluded parts). To select the best mask we search through the entire set of occluders to maximize the number of parts with the correct occlusion label, with greater weight on the occluded parts (in the experiments, twice as much as for visible parts).

(ii) For parts not under deterministic occlusion, we look for missing image evidence (low part detection scores for multiple adjacent parts), guided by the set of occluder masks. Specifically, for a particular wireframe hypothesis, we search through the set of occluder masks to maximize the summed part detection scores (obtained from the Random Forest classifier, Sec. 3.2), replacing the scores for parts behind the occluder by a constant (low) score c. Especially in this step, leveraging local context in the form of occlusion masks stabilizes individual part-level occlusion estimates, which by themselves are rather unreliable because of the noisy evidence.

Given: Scene particle ψ': 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 $\mathbf{h}^\beta = \text{argmax}_j \ \mathcal{L}\big(\mathbf{h}^\beta(\mathbf{q}_j^\beta, \mathbf{s}_j^\beta, a^\beta), \boldsymbol{\theta}_{gp}\big)$
  end
  2. for $\beta = 1 \dots n$ **do**
    update occlusion mask (exhaustive search)
    $a^\beta = \text{argmax}_j \ \ \mathcal{L}\big(\mathbf{h}^\beta(\mathbf{q}^\beta, \mathbf{s}^\beta, a_j), \boldsymbol{\theta}_{gp}\big)$
  end
  3. Recompute sampling variance Σ^β of Gaussians (Leordeanu and Hebert 2008)
end

Algorithm 1: Inference run for each scene particle.

Final scene-level inference. Finally, we search a good local optimum of the scene objective function (Eqn. 5) using an iterative stochastic optimization scheme shown in Algorithm 1. Each particle is iteratively refined in two steps: first, the shape and viewpoint parameters of all objects are updated. Then, object occlusions are recomputed and occlusions by unmodeled objects are updated, by exhaustive search over the set of possible masks.

The update of the continuous shape and viewpoint follows the smoothing-based optimization of Leordeanu and Hebert (2008). In a nutshell, new values for the shape and viewpoint parameters are found by testing many random perturbations around the current values. The trick is that the random perturbations follow a normal distribution that is adapted in a data-driven fashion: in regions where the objective function is unspecific and wiggly the variance is increased to suppress weak local minima; near distinct peaks the variance is reduced to home in on the nearby stronger optimum. For details we refer to the original publication.

For each scene particle the two update steps – shape and viewpoint sampling for all cars with fixed occlusion masks, and exhaustive occlusion update for fixed shapes and viewpoints – are iterated, and the particle with the highest objective value ψ forms our MAP estimate. As the space of ground planes is already well-covered by the set of multiple scene particles (in our experiments 250), we keep the ground plane parameters of each particle constant. This stabilizes the optimization. Moreover, we limit ourselves to a fixed number of objects from the pre-detection stage. The scheme could be extended to allow adding and deleting object hypotheses, by normalizing the scene-level likelihood with the number of object instances under consideration.

5 Experiments

In this section, we extensively analyze the performance of our fine-grained 3D scene model, focusing on its ability to derive 3D estimates from a single input image (with known camera intrinsics). To that end, we evaluate object localization in 3D metric space (Sec. 5.4.1) as well as 3D pose estimation (Sec. 5.4.2) on the challenging KITTI dataset (Geiger et al. 2012) of street scenes. In addition, we analyze the performance of our model w.r.t. part-level occlusion prediction and part localization in the 2D image plane (Sec. 5.5). In all experiments, we compare the performance of our full model with stripped-down variants as well as appropriate baselines, to highlight the contributions of different system components to overall performance.

5.1 Dataset

In order to evaluate our approach for 3D layout estimation from a single view, we require a dataset with 3D annotations. We thus turn to the KITTI *3D object detection and orientation estimation* benchmark dataset (Geiger et al. 2012) 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 evaluations. It consists of around $7,500$ training and $7,500$ 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. Since annotations are only made publicly available on the training set of KITTI, we utilize a portion of this training set for our evaluation. We choose only images with multiple cars that are large enough to identify parts, and manually annotate all cars in this subset with 2D part locations and part-level occlusion labels. Specifically, we pick every 5th image from the training set with at least two cars with height greater than 75 pixels. This gives us 260 test images with 982 cars in total, of which 672 are partially occluded, and 476 are severely occluded. Our selection shall ensure that while being biased towards more complex scenes, we still sample a representative portion of the dataset.

Training set. We use two different kinds of data for training our model, *(i)* synthetic data in the form of rendered CAD models, and *(ii)* real-world training data. *(i)* We utilize 38 commercially available 3D CAD models of cars for learning the object wireframe model as well as for learning viewpoint-invariant part appearances, (c.f. Zia et al. 2013). Specifically, we render the 3D CAD models from 72 different azimuth angles (5° steps) and 2 elevation angles (7.5° and 15° above the ground), densely covering the relevant part of the viewing sphere, using the non-photorealistic style of Stark et al. (2010). Rendered part patches serve as positive part examples, randomly sampled image patches as well as non-part samples from the renderings serve as negative background examples to train the multi-class Random Forest classifier. The classifier distinguishes 37 classes (36 parts

and 1 background class), using 30 trees with a maximum depth of 13. The total number of training patches is $162,000$, split into $92,000$ part and $70,000$ background patches. *(ii)* We train 118 part configuration detectors (single component DPMs) labeled with discrete viewpoint, 2D part locations and part-level occlusion labels on a set of $1,000$ car images downloaded from the internet and 150 images from the KITTI dataset (none of which are part of the test set). In order to model the occlusions, we semi-automatically define a set of 288 occluder masks, the same as in Zia et al. (2013).

5.2 Object Pre-Detection

As a sanity check, we first verify that our 2D pre-detection (Sec. 3.4) matches the state-of-the-art. To that end we evaluate a standard 2D bounding box detection task according to the PASCAL VOC criterion ($> 50\%$ 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 Geiger et al. (2012) 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 (Felzenszwalb et al. 2010), both trained on the same training set (Sec. 5.1). Precision-recall curves are shown in Fig. 6. We observe that our detector bank (green curve, 57.8% AP) in fact performs slightly better than the original DPM (red curve, 57.3% AP). In addition, it delivers coarse viewpoint estimates and rough part locations that we can leverage for initializing our scene-level inference (Sec. 4.3). The pre-detection takes about 2 minutes per test image on a single core (evaluation of 118 single component DPMs and clustering of their votes).

5.3 Model Variants and Baselines

We compare the performance of our full system with a number of stripped down variants in order to quantify the benefit that we get from each individual component. We consider the following variants:

(i) FG: the basic version of our fine-grained 3D object model, without ground plane, searched occluder or deterministic occlusion reasoning; this amounts to independent modeling of the objects in a common, metric 3D scene coordinate system. *(ii)* FG+SO: same as (i) but with searched occluder to represent occlusions caused by unmodeled scene elements. *(iii)* FG+DO: same as (i) but with deterministic occlusion reasoning between multiple objects. *(iv)* FG+GP: same as (i), but with common ground plane. *(v)* FG+GP+DO+SO: same as (i), but with all three components, common ground plane, searched occluder, and deterministic occlusion turned on. *(vi)* the earlier pseudo-3D shape model (Zia et al. 2013), with probabilistic occlusion reasoning; this uses essentially the same object model as (ii), but learns it from examples scaled to the *same* size rather than the *true* size, and fits the model in 2D $(x, y, scale)$-space rather explicitly recovering a 3D scene interpretation.

We also compare our representation to two different baselines, *(vii)* COARSE: a scene model consisting of 3D bounding boxes rather than detailed cars, corresponding to the coarse 3D geometry stage of our pipeline (Sec. 4.3); and *(viii)* COARSE+GP: like *(vii)* but with a common ground plane for the bounding boxes. Specifically, during the coarse grid search we choose the 3D bounding box hypothesis whose 2D projection is closest to the corresponding pre-detection 2D bounding box.

	full dataset		occ >0 parts		occ >3 parts	
	<1m	<1.5m	<1m	<1.5m	<1m	<1.5m
Fig. 5 plot	*(a)*	*(b)*			*(c)*	*(d)*
(i) FG	23%	35%	22%	31%	23%	32%
(ii) FG+SO	26%	37%	23%	33%	27%	36%
(iii) FG+DO	25%	37%	26%	35%	27%	38%
(iv) FG+GP	40%	53%	40%	52%	38%	49%
(v) FG+GP+DO+SO	**44%**	**56%**	**44%**	**55%**	**43%**	**60%**
(vi) Zia et al. (2013)	—	—	—	—	—	—
(vii) COARSE	21%	37%	21%	40%	20%	42%
(viii) COARSE+GP	35%	54%	28%	48%	27%	47%

Table 1: 3D localization accuracy: percentage of cars correctly localized within 1 and 1.5 meters of ground truth.

5.4 3D Evaluation

Having verified that our pre-detection stage is competitive and provides reasonable object candidates in the 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. Our current unoptimized implementation takes around 5 minutes to evaluate the local part detectors in a sliding-window fashion at multiple scales over the whole image, and further 20 minutes per test image for the inference, on a single core. This is similar to recent deformable face model fitting work, e.g. Schönborn et al. (2013). However, both the sliding-window part detector and the sample-based inference naturally lend themselves to massive parallization. In fact the part detector only

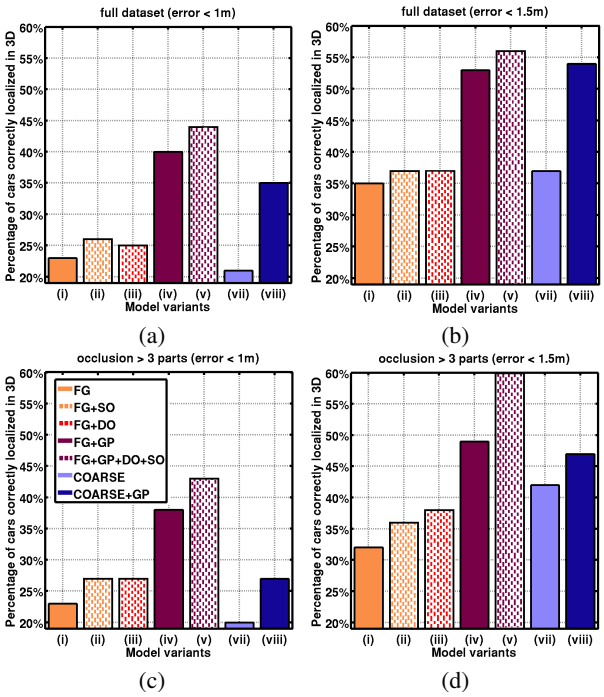


Fig. 5: 3D localization accuracy: percentage of cars correctly localized within 1 (a,c) and 1.5 (b,d) meters of ground truth, on all (a,b) and occluded (c,d) cars.

needs to be evaluated within the pre-detection bounding boxes, which we do not exploit at present. Moreover, we set the number of iterations conservatively, in most cases the results already converge far earlier.

5.4.1 3D Object Localization

Protocol. We measure 3D localization performance by the fraction of detected object centroids that are correctly localized up to deviations of 1, and 1.5 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 estimates lie within 45° of the ground truth, thus excluding failures of pre-detection. We perform the analysis for three settings (Tab. 1): *(i)* over our full testset (517 of 982 TPs); *(ii)* only over those cars that are partially occluded, *i.e.* 1 or more of the parts that are not self-occluded by the object are not visible (234 of 672 TPs); and *(iii)* only those cars that are severely occluded, *i.e.* 4 or more parts are not visible (113 of 476 TPs). Fig. 5 visualizes selected columns of Tab. 1 as bar plots to facilitate the comparison.

Results. In Tab. 1 and Fig 5, we first observe that our full system (FG+GP+DO+SO, dotted dark red) is the top performer for all three occlusion settings and both localization error thresholds, localizing objects with 1 m accuracy in $43-44\%$ of the cases and with 1.5 m accuracy in 55–60% of the cases. Fig. 8 visualizes some examples of our full system FG+GP+DO+SO vs. the stronger baseline COARSE+GP.

Second, the basic fine-grained model FG (orange) outperforms COARSE (light blue) by 1–3 percent points (pp) corresponding to a relative improvement of 4–13% at 1 m accuracy. The gains increase by a large margin when adding a ground plane: FG+GP (dark red) outperforms COARSE+GP (dark blue) by 5–12 pp (13–43%) 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 1.5 m are less clear-cut. It appears that from badly localized initializations just inside the 1.5 m radius, the final inference sometimes drifts into incorrect local minima outside of 1.5 m.

Third, modeling fine-grained occlusions either independently (FG+SO, dotted orange) or deterministically across multiple objects (FG+DO, dotted red) brings marked improvements on top of FG alone. At 1 m they outperform FG by 1–4 pp (2–15%) and by 2–4 pp (7–19%), respectively. We get similar improvements at 1.5 m, with FG+SO and FG+DO outperforming FG by 2–4 pp (4–14%), and 2–6 pp (4–19%) respectively. Not surprisingly, the performance boost is greater for the occluded cases, and both occlusion reasoning approaches are in fact beneficial for 3D reasoning. Fig. 9 visualizes some results with and without occlusion reasoning.

And last, adding the ground plane always boosts the performance for both the FG and COARSE models, strongly supporting the case for joint 3D scene reasoning: at 1 m accuracy the gains are 15–18 pp (65–81%) for FG+GP vs. FG, and 7–14 pp (30–67%) for COARSE+GP vs. COARSE. Similarly, at 1.5 m accuracy we get 17–21 pp (51–68%) for FG+GP vs. FG, and 5–17 pp (10–47%) for COARSE+GP vs. COARSE. for qualitative results see Fig. 10.

We obtain even richer 3D "reconstructions" by replacing wireframes with nearest neighbors from the database of 3D CAD models (Fig. 11), accurately recognizing hatchbacks (a, e, f, i, j, l, u), sedans (b, o) and station wagons (d, p, v, w, x), as well as approximating the van (c, no example in database) by a station wagon. Specifically, we represent the estimated wireframe as well as the annotated 3D CAD exemplars as vectors of corresponding 3D part locations, and find the nearest CAD exemplar in terms of Euclidean distance, which is then visualized. Earlier, the same method was used to perform fine-grained object categorization (Zia et al. 2013).

	full dataset				occ >0 parts				occ >3 parts			
	<5°	<10°	3D err	2D err	<5°	<10°	3D err	2D err	<5°	<10°	3D err	2D err
(i) FG	44%	**69%**	**5°**	**4°**	41%	65%	6°	4°	35%	**58%**	**7°**	5°
(ii) FG+SO	42%	66%	6°	**4°**	39%	62%	6°	4°	33%	53%	8°	5°
(iii) FG+DO	**45%**	68%	**5°**	**4°**	44%	**66%**	6°	4°	36%	56%	**7°**	**4°**
(iv) FG+GP	41%	63%	6°	**4°**	40%	62%	6°	4°	36%	52%	8°	5°
(v) FG+GP+DO+SO	44%	65%	6°	**4°**	**47%**	65%	**5°**	**3°**	**44%**	55%	8°	**4°**
(vi) Zia et al. (2013)	-	-	-	6°	-	-	-	6°	-	-	-	6°
(vii) COARSE	16%	38%	13°	9°	20%	41%	13°	6°	21%	40%	14°	9°
(viii) COARSE+GP	25%	51%	10°	6°	27%	51%	10°	5°	23%	40%	14°	7°

Table 2: 3D viewpoint estimation accuracy (percentage of objects with less than 5° and 10° error) and median angular estimation errors (3D and 2D)

5.4.2 *Viewpoint Estimation*

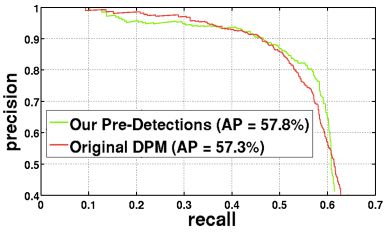
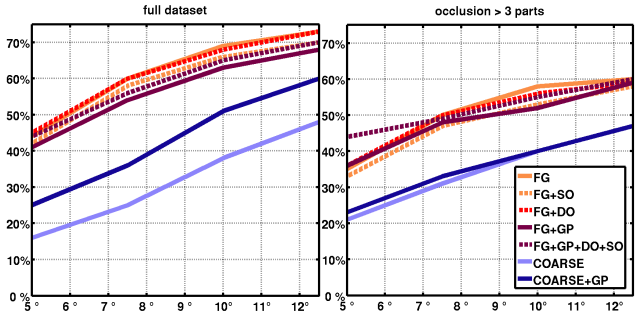


Fig. 6: Object pre-detection performance.



(a) (b)

Fig. 7: Percentage of cars with VP estimation error within $x°$.

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 can evaluate object orientation (azimuth) in 2D image space as well as in 3D scene space. 2D viewpoint is the apparent azimuth of the object as seen in the image. The actual azimuth relative to a fixed scene direction (called 3D viewpoint), is calculated from the 2D viewpoint estimate and an estimate of 3D object location. We measure viewpoint estimation accuracy in two ways: as the percentage of detected objects for which the 3D angular error is below $5°$ or $10°$, and as the median angular error between estimated and ground truth azimuth angle over detected objects, both in 3D and 2D.

Results. Table 2 shows the quantitative results, again comparing our full model and the different variants introduced in Sec. 5.3, and distinguishing between the full dataset and two subsets with different degrees of occlusion. In Fig. 7 we plot the percentage of cars whose poses are estimated correctly up to different error thresholds, using the same color coding as Fig. 5.

First, we observe that the full system FG+GP+DO+SO (dotted dark red) outperforms the best coarse model COARSE+GP (dark blue) by significant margins of 19–21 pp and 14–15 pp at $5°$ and $10°$ errors respectively, improving the median angular error by $4°$–$6°$.

Second, all FG models (shades of orange and red) deliver quite reliable viewpoint estimates with smaller differences in performance ($\leq$ 6 pp, or $1°$ median error) for $10°$ error, outperforming their respective COARSE counterparts (shades of blue) by significant margins. Observe the clear grouping of curves in Fig. 7. However, for the high accuracy regime ($\leq 5°$ error), the full system FG+GP+DO+SO (dotted dark red) delivers the best performance for both occluded subsets, beating the next best combination FG+DO (dotted red) by 3 pp and 8 pp, respectively.

Third, the ground plane helps considerably for the COARSE models (shades of blue), improving by 9 pp for $\leq 5°$ error, and 13 pp for $\leq 10°$ over the full data set. Understandably, that gain gradually dissolves with increasing occlusion.

And fourth, we observe that in terms of median 2D viewpoint estimation error, our full system FG+GP+DO+SO outperforms the pseudo-3D model of (Zia et al. 2013) by $2°$–$3°$, highlighting the benefit of reasoning in true metric 3D space.

	full dataset		occ >0 parts		occ >3 parts	
	occl. pred. acc.	#cars >70% parts	occl. pred. acc.	#cars >70% parts	occl. pred. acc.	#cars >70% parts
(i) FG	82%	69%	70%	**68%**	57%	43%
(ii) FG+SO	87%	66%	80%	63%	77%	35%
(iii) FG+DO	84%	70%	72%	67%	62%	**47%**
(iv) FG+GP	82%	68%	68%	67%	57%	46%
(v) FG+GP+DO+SO	**88%**	**71%**	82%	67%	79%	44%
(vi) Zia et al. (2013)	87%	64%	**84%**	61%	**84%**	32%
(vii) COARSE	—	—	—	—	—	—
(viii) COARSE+GP	—	—	—	—	—	—

Table 3: 2D accuracy. Part-level occlusion prediction accuracy and percentage of cars which have >70% parts accurately localized.

5.5 2D Evaluation

While the objective of this work is to enable accurate localization and pose estimation in 3D (Sec. 5.4), we also present an analysis of 2D performance (part localization and occlusion prediction in the image plane), to put the work into context. Unfortunately, a robust measure to quantify how well the wireframe model fits the image data requires accurate ground truth 2D locations of even the occluded parts, which are not available. A measure used previously in Zia et al. (2013) is 2D part localization accuracy only evaluated for the visible parts, but we now find it to be biased, because evaluating the model for just the visible parts leads to high accuracies on that measure, even if the overall fit is grossly incorrect. We thus introduce a more robust measure below.

Protocol. We follow the evaluation protocol commonly applied for human body pose estimation and evaluate the number of correctly localized parts, using a relative threshold adjusted to the size of the reprojected car (20 pixels for a car of size 500×170 pixels, *i.e.* $\approx 4\%$ of the total length (c.f. Zia et al. 2013)). We use this threshold to determine the percentage of detected cars for which 70% or more of all (not self-occluded) parts are localized correctly, evaluated only on cars for which at least 70% of the (not self-occluded) parts are visible according to ground truth. We find this measure to be more robust, since it favours sensible fits of the overall wireframe.

Further, we calculate the percentage of (not self-occluded) parts for which the correct occlusion label is estimated. For the model variants which do not use the occluder representation (FG and FG+GP), all candidate parts are predicted as visible.

Results. Tab. 3 shows the results for both 2D part localization and part-level occlusion estimation. We observe that our full system FG+GP+DO+SO is the highest performing variant over the full data set (88% part-level occlusion prediction accuracy and 71% cars with correct part localization). For the occluded subsets, the full system performs best among all FG models on occlusion prediction, whereas the results for part localization are less conclusive. An interesting observation is that methods that use 3D context (FG+GP+DO+SO, FG+GP, FG+DO) consistently beat (FG+SO), *i.e.* inferring occlusion is more brittle from (missing) image evidence alone than when supported by 3D scene reasoning.

Comparing the pseudo-3D baseline (Zia et al. 2013) and its proper metric 3D counterpart FG+SO, we observe that, indeed, metric 3D improves part localization by 2–3 pp (despite inferior part-level occlusion prediction). In fact, all FG variants outperform Zia et al. (2013) in part localization by significant margins, notably FG+GP+DO+SO (6–12 pp).

On average, we note that there is only a weak (although still positive) correlation between 2D part localization accuracy and 3D localization performance (Sec. 5.4). In other words, whenever possible *3D reasoning should be evaluated in 3D space*, rather than in the 2D projection.[3]

6 Conclusion

We have approached the 3D scene understanding problem from the perspective of detailed deformable shape and occlusion modeling, jointly fitting the shapes of multiple objects linked by a common scene geometry (ground plane). Our results suggest that detailed representations of object shape are beneficial for 3D scene reasoning, and fit well with scene-level constraints between objects. By itself, fitting a detailed, deformable 3D model of cars and reasoning about occlusions resulted in improvements of 16–26% in object localization accuracy (number of cars localized to within 1m in 3D), 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 30–67%. 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 104–113% improvement in 3D localization compared to just lifting 2D detections, as well as a reduction of the median 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 woud be 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

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

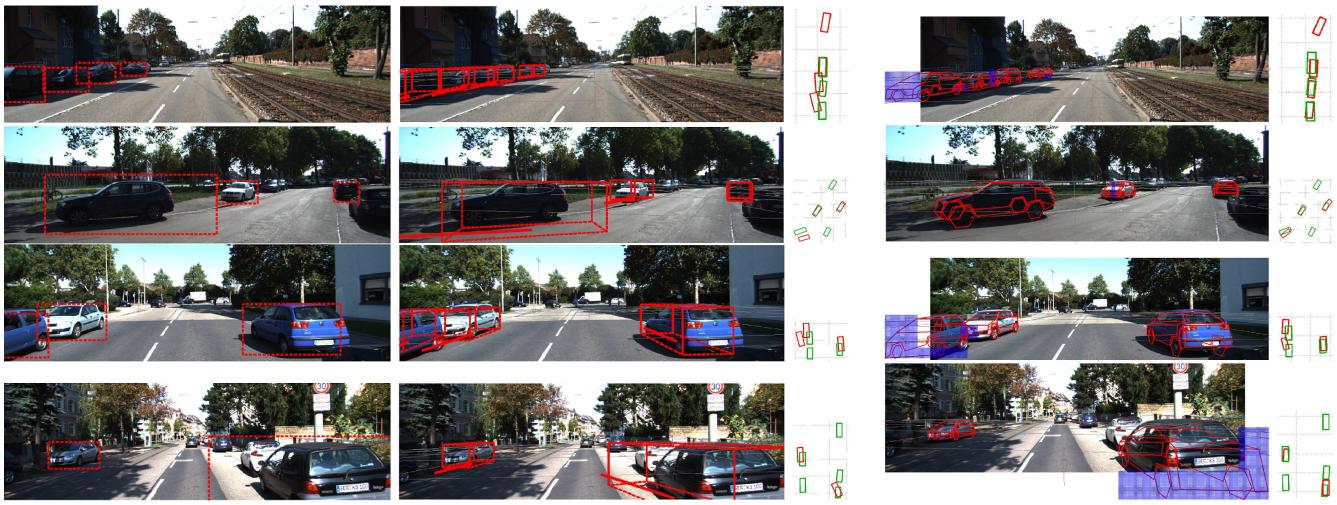


Fig. 8: COARSE+GP (a-c) vs FG+GP+DO+SO (d,e). (a) 2D bounding box detections, (b) COARSE+GP based on (a), (c) bird's eye view of (b), (d) FG+GP+DO+SO shape model fits (blue: estimated occlusion masks), (e) bird's eye view of (d). Estimates in red, ground truth in green.

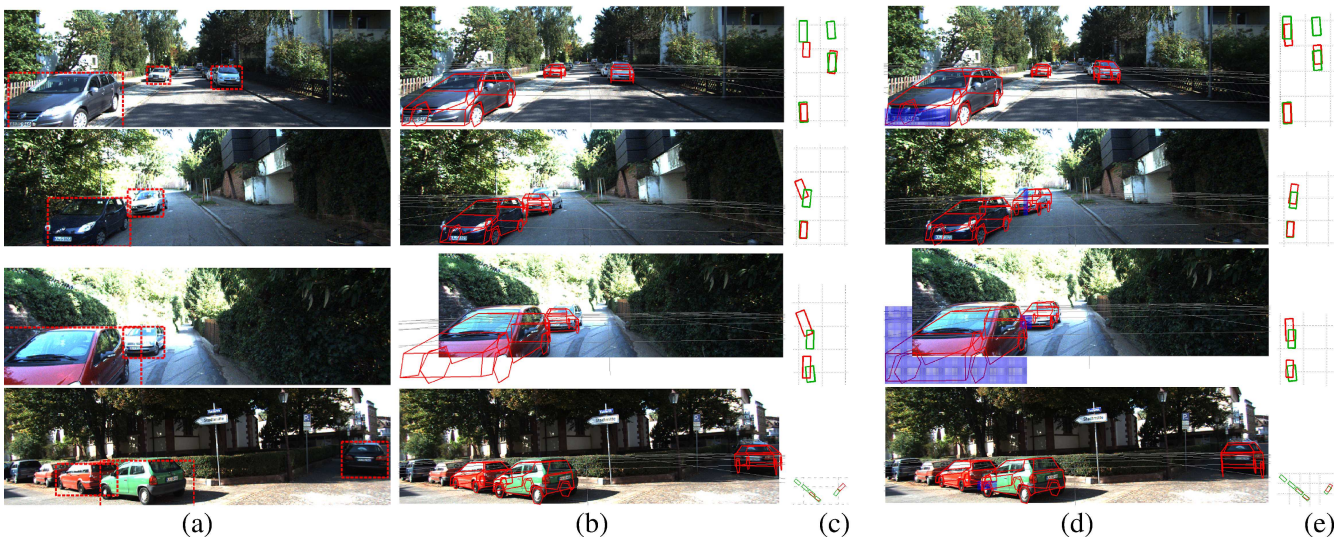


Fig. 9: FG+GP (a-c) vs FG+GP+DO+SO (d,e). (a) 2D bounding box detections, (b) FG+GP based on (a), (c) bird's eye view of (b), (d) FG+GP+DO+SO shape model fits (blue: estimated occlusion masks), (e) bird's eye view of (d). Estimates in red, ground truth in green.

2D pre-detection 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

S.Y. Bao, S. Savarese, Semantic Structure from Motion, in *CVPR*, 2011

S. Belongie, J. Malik, J. Puzicha, Shape Context: A New Descriptor for Shape Matching and Object Recognition, in *NIPS*, 2000

L. Bourdev, J. Malik, Poselets: Body part detectors trained using 3D human pose annotations, in *ICCV*, 2009

R.A. Brooks, Symbolic reasoning among 3-d models and 2-d images. Artificial Intelligence (1981)

W. Choi, Y.-W. Chao, C. Pantofaru, S. Savarese, Understanding Indoor Scenes Using 3D Geometric Phrases, in *CVPR*, 2013

T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models, their training and application. CVIU **61**(1) (1995)

N. Dalal, B. Triggs, Histograms of Oriented Gradients for Human Detection, in *CVPR*, 2005

L. Del Pero, J. Bowdish, B. Kermgard, E. Hartley, K. Barnard, Understanding Bayesian rooms using composite 3D object models, in

Fig. 10: FG (a-c) vs FG+GP (d,e). (a) 2D bounding box detections, (b) FG based on (a), (c) bird's eye view of (b), (d) FG+GP shape model fits, (e) bird's eye view of (d). Estimates in red, ground truth in green.

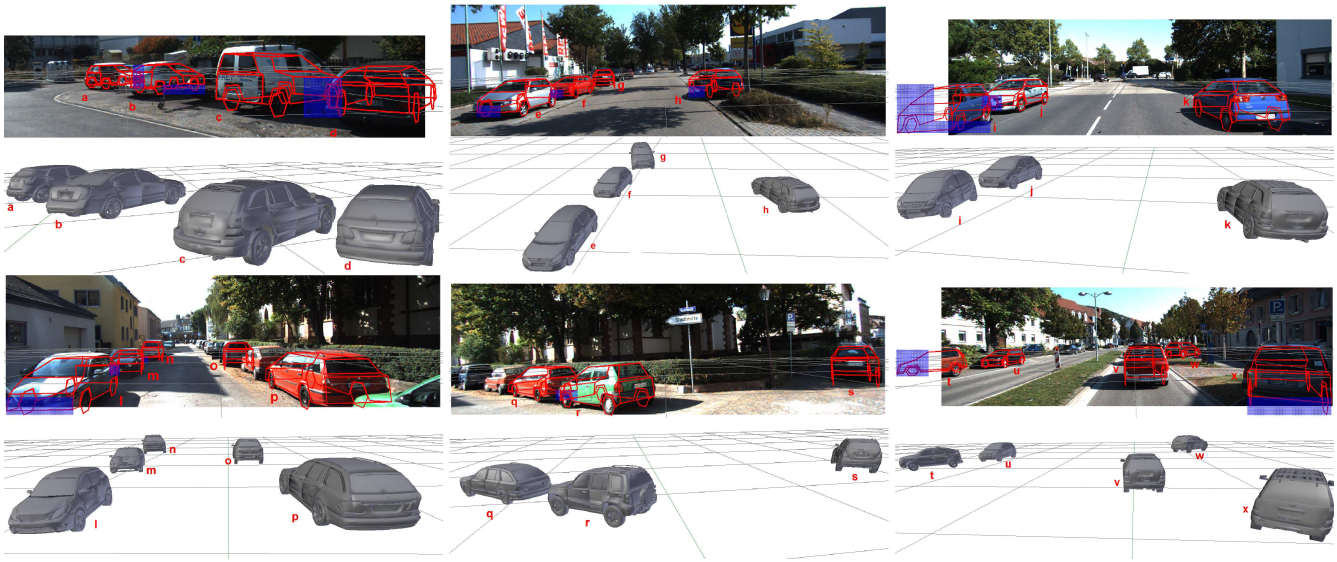


Fig. 11: Example detections and corresponding 3D reconstructions.

CVPR, 2013

M. Enzweiler, A. Eigenstetter, B. Schiele, D.M. Gavrila, Multi-Cue Pedestrian Classification with Partial Occlusion Handling, in *CVPR*, 2010

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

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

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

R. Fransens, C. Strecha, L.V. Gool, A Mean Field EM-algorithm for Coherent Occlusion Handling in MAP-Estimation, in *CVPR*, 2006

T. Gao, B. Packer, D. Koller, A Segmentation-aware Object Detection Model with Occlusion Handling, in *CVPR*, 2011

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

A. Geiger, C. Wojek, R. Urtasun, Joint 3D Estimation of Objects and Scene Layout, in *NIPS*, 2011

R.B. Girshick, P.F. Felzenszwalb, D. McAllester, Object detection with grammar models, in *NIPS*, 2011

A. Gupta, A.A. Efros, M. Hebert, Blocks World Revisited: Image Understanding Using Qualitative Geometry and Mechanics, in *ECCV*, 2010

M. Haag, H.-H. Nagel, Combination of edge element and optical flow estimates for 3d-model-based vehicle tracking in traffic image sequences. IJCV **35**(3), 295–319 (1999)

V. Hedau, D. Hoiem, D.A. Forsyth, Thinking Inside the Box: Using Appearance Models and Context Based on Room Geometry, in *ECCV*, 2010

M. Hejrati, D. Ramanan, Analyzing 3D Objects in Cluttered Images, in *NIPS*, 2012

D. Hoiem, A. Efros, M. Hebert, Putting objects in perspective. IJCV **80**(1), 3–15 (2008)
T. Kanade, A Theory of Origami World. Artificial Intelligence (1980)
D. Koller, K. Daniilidis, H.H. Nagel, Model-based object tracking in monocular image sequences of road traffic scenes. IJCV **10**(3), 257–281 (1993)
S. Kwak, W. Nam, B. Han, J.H. Han, Learning occlusion with likelihoods for visual tracking, in *ICCV*, 2011
B. Leibe, A. Leonardis, B. Schiele, An implicit shape model for combined object categorization and segmentation. Toward Category-Level Object Recognition (2006)
M. Leordeanu, M. Hebert, Smoothing-based Optimization, in *CVPR*, 2008
Y. Li, L. Gu, T. Kanade, Robustly aligning a shape model and its application to car alignment of unknown pose. PAMI **33**(9) (2011)
X. Liu, Y. Zhao, S.-C. Zhu, Single-View 3D Scene Parsing by Attributed Grammar, in *CVPR*, 2014
D. Lowe, Three-dimensional object recognition from single two-dimensional images. AI **31**(3), 355–395 (1987)
S. Maji, J. Malik, Object Detection Using a Max-Margin Hough Transform, in *CVPR*, 2009
J. Malik, Interpreting Line Drawings of Curved Objects. IJCV **1**(1), 73–103 (1987)
D. Meger, C. Wojek, B. Schiele, J.J. Little, Explicit occlusion reasoning for 3d object detection, in *BMVC*, 2011
J. Oramas, L. De Raedt, T. Tuytelaars, Allocentric Pose Estimation, in *ICCV*, 2013
A. Pentland, Perceptual organization and representation of natural form. AI (1986)
B. Pepik, M. Stark, P. Gehler, B. Schiele, Occlusion Patterns for Object Class Detection, in *CVPR*, 2013
L.G. Roberts, Machine Perception of Three-Dimensional Solids, PhD thesis, MIT, 1963
S. Schönborn, A. Forster, B. Egger, T. Vetter, A Monte Carlo Strategy to Integrate Detection and Model-Based Face Analysis, in *GCPR*, 2013
N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor Segmentation and Support Inference from RGBD Images, in *ECCV*, 2012
M. Stark, M. Goesele, B. Schiele, Back to the Future: Learning Shape Models from 3D CAD Data, in *BMVC*, 2010
G.D. Sullivan, A.D. Worrall, J.M. Ferryman, Visual Object Recognition Using Deformable Models of Vehicles, in *IEEE Workshop on Context-Based Vision*, 1995
S. Tang, M. Andriluka, B. Schiele, Detection and Tracking of Occluded People, in *BMVC*, 2012
A. Vedaldi, A. Zisserman, Structured output regression for detection with partial truncation, in *NIPS*, 2009
M. Villamizar, H. Grabner, J. Andrade-Cetto, A. Sanfeliu, L.V. Gool, F. Moreno-Noguer, Efficient 3D Object Detection using Multiple Pose-Specific Classifiers, in *BMVC*, 2011
H. Wang, S. Gould, D. Koller, Discriminative Learning with Latent Variables for Cluttered Indoor Scene Understanding, in *ECCV*, 2010
X. Wang, T. Han, S. Yan, An HOG-LBP human detector with partial occlusion handling, in *ICCV*, 2009
C. Wojek, S. Walk, S. Roth, K. Schindler, B. Schiele, Monocular visual scene understanding: understanding multi-object traffic scenes. PAMI (2013)
Y. Xiang, S. Savarese, Object Detection by 3D Aspectlets and Occlusion Reasoning, in *3dRR*, 2013
Y. Xiang, S. Savarese, Estimating the Aspect Layout of Object Categories, in *CVPR*, 2012
Y. Zhao, S.-C. Zhu, Scene Parsing by Integrating Function, Geometry and Appearance Models, in *CVPR*, 2013
M.Z. Zia, U. Klank, M. Beetz, Acquisition of a dense 3D model database for robotic vision, in *ICAR*, 2009
M.Z. Zia, M. Stark, K. Schindler, Explicit Occlusion Modeling for 3D Object Class Representations, in *CVPR*, 2013
M.Z. Zia, M. Stark, K. Schindler, Are Cars Just 3D Boxes? – Jointly Estimating the 3D Shape of Multiple Objects, in *CVPR*, 2014
M.Z. Zia, M. Stark, B. Schiele, K. Schindler, Detailed 3d representations for object recognition and modeling. PAMI **35**(11), 2608–2623 (2013)