Mono3D++: Monocular 3D Vehicle Detection with Two-Scale 3D Hypotheses and Task Priors

Tong He, Stefano Soatto
UCLA Vision Lab
University of California, Los Angeles
{simpleig, soatto}@cs.ucla.edu

Abstract

We present a method to infer 3D pose and shape of vehicles from a single image. To tackle this ill-posed problem, we ensure two-scale projection consistency between the generated 3D hypotheses and their 2D pseudo-measurements. Specifically, we use a morphable wireframe model to generate a fine-scaled and accurate representation of vehicle shape and pose. To reduce its sensitivity to 2D landmarks, we jointly model the 3D bounding box as a coarse representation and improve the robustness. We also integrate three task priors, including unsupervised monocular depth, a ground plane constraint as well as vehicle shape priors, with forward projection errors into an overall energy function. Our system is benchmarked on the KITTI dataset and outperforms state-of-the-art methods for monocular 3D vehicle detection.

1 Introduction

Objects are regions of three-dimensional (3D) space that can move independently as a whole and have both geometric and semantic attributes (shapes, identities, affordances, etc.) in the context of a task. In this paper, we focus on vehicle objects in driving scenarios, which are one of the most important targets in autonomous driving tasks. Given an image, we wish to produce a posterior probability of vehicle attributes in 3D, or at least some point-estimates from it.

Inferring 3D vehicles from a single image is an ill-posed problem since object attributes exist in 3D but single images can only provide partial and pseudo measurements in 2D. Therefore, we propose to solve this task by tackling two issues: (i) how to ensure 3D-2D consistency between the generated 3D vehicle hypotheses and their corresponding 2D pseudo-measurements, which requires strong 3D hypotheses generators as well as robust forward innovation errors; (ii) how to refine 3D hypotheses with task priors that can be integrated into an easy-to-optimize loss function.

For the first problem, we use a joint modeling method that leverages two different 3D hypotheses generation schemes: one serves as a coarse representation of vehicle shape and pose while the other is fine-scaled. We design end-to-end trained deep networks to generate 3D hypotheses for each vehicle instance in the form of both 3D bounding box and morphable wireframe model (a.k.a linear shape model).

Figure 1: Representative 3D detection results, shown as projections on the input images. 3D morphable shape models of each vehicle are colored in green. Dashed green lines are occluded edges. For the 14 vertices of each morphable shape model, visible vertices are colored in red and occluded ones in yellow. 2D bounding boxes are colored in blue.

Shape and pose parameters will be adjusted according to the 2D pseudo-measurements via an optimization approach. A wireframe model can determine shape and pose more precisely than a 3D bounding box, but it is very sensitive to the 2D landmark measurements which can be easily affected by issues like partial occlusions, shadow, low resolution, etc. Therefore, we jointly model the 3D bounding box projection constraint to improve its robustness. Though separately studied before in various shape and pose estimation tasks (e.g. cars, hands, human bodies), these two different hypotheses generators are actually complementary and have never been jointly considered in the published literature. We conduct ablation studies to demonstrate benefits brought by jointly modeling the coarse and the fine-scaled 3D object pose and shape representations.

For the second problem, we consider three constraints on vehicles. We argue that cars should stay on the ground plane, should look like a car, and should be at a reasonable distance from the observation camera. The first argument serves as a supporting plane constraint for vehicles. The second argument is a prior term for vehicle shapes. The last argument indicates that vehicle translation in camera coordinates should
be constrained by monocular range map of the current driving scene. These constraints are jointly modeled with the aforementioned 3D-2D consistency terms in order to further improve vehicle shape and pose estimation. Again, ablation studies show that our modeling of such priors is helpful.

In summary, in this paper we propose an approach for vehicle 3D shape and pose estimation from a single image that leverages the coarse and the fine-scaled 3D hypotheses, as well as multiple task priors, as shown in Fig. 2, for joint inference. Our contributions are:

- We propose and empirically validate the joint modeling of vehicles’ coarse and fine-scaled shape and pose representations via two complementary 3D hypotheses generators, namely 3D bounding box and morphable wireframe model.
- In our method, we model multiple priors of the 3D vehicle detection task into easy-to-optimize loss functions. Such priors include unsupervised monocular depth, a ground plane constraint and vehicle shape priors.
- We build an overall energy function that integrates the proposed two improvements which significantly outperforms all recent state-of-the-art monocular 3D vehicle detectors on the KITTI dataset.

2 Related work

To produce rich descriptions of vehicles in 3D space, methods leveraging various data are proposed, such as video, RGB-D, or RGB, among which RGB methods are most related to our work. For completeness, we also briefly review other methods.

**Video methods:** In (Song and Chandraker 2015; Murthy, Sharma, and Krishna 2017; Chhaya et al. 2016; Dong, Fei, and Soatto 2017; Fei and Soatto 2018), temporal information is explored for (moving) 3D objects localization by a recursive Bayesian inference scheme or optimizing loss functions that are based on non-rigid structure-from-motion methods (Torresani, Hertzmann, and Bregler 2003).

**RGB-D methods:** MV3D (Chen et al. 2017) encodes lidar point clouds into multi-view feature maps, which are fused with images, and uses 2D convolutions for 3D localization. In contrast, F-PointNet (Qi et al. 2018) directly processes lidar point clouds in 3D space using two variants of PointNet (Qi et al. 2017) for 3D object segmentation and amodal detection. Other methods that also use lidar point clouds include (Ren and Sudderth 2018; Xu, Anguelov, and Jain 2018). 3DOP (Chen et al. 2015) exploits stereo point clouds and evaluates 3D proposals via depth-based potentials.

**RGB methods:** Mono3D (Chen et al. 2016) scores 3D bounding boxes generated from monocular images, using a ground plane prior and 2D cues such as segmentation masks. Deep3DBox (Mousavian et al. 2017) recovers 3D pose by minimizing the reprojection error between the 3D box and the detected 2D bounding box of the vehicle. Task priors are not jointly modeled with 3D-2D innovation terms. 3DVP (Xiang et al. 2015) proposes 3D voxel with occlusion patterns and uses a set of ACF detector for 2D detection and 3D pose estimation. Its follow-up work, SubCNN (Xiang et al. 2017), uses deep networks to replace the ACF detectors for view-dependent subcategory classification. Active shape models are explored in (Zia et al. 2011; Zia, Stark, and Schindler 2013; 2014a; 2014b) for vehicle modeling. CAD models are rendered in (Mottaghi, Xiang, and Savarese 2015; Choy et al. 2015) for 3D detection by hypotheses sampling/test approaches using image features, such as HOG (Dalal and Triggs 2005). In the state-of-the-art DeepMANTA (Chabot et al. 2017), vehicle pose is adjusted by 3D-2D landmark matching. These approaches only model either the coarse or the fine-scaled 3D shape and pose representation of a vehicle, thus have limitations in accuracy and robustness. Moreover, task priors, such as monocular depth of the current driving scene, vehicle shape priors, supporting plane constraints, etc., are only partly considered and not jointly optimized with forward projection errors.

3 Method

We wish to infer the posterior distribution of object pose $g \in SE(3)$ and shape $S \subseteq \mathbb{R}^3$, given an image $I$, $P(g, S | I)$, where $SE(3)$ denotes the Euclidean group of rigid motions that characterize the position and orientation of the vehicle relative to the camera, and shape $S$ is characterized parametrically for instance using a point cloud or a linear morphable model. In the Supplementary Material we describe all the modeling assumptions needed to make in order to arrive at a tractable approximation of the posterior, maximizing which is equivalent to minimizing the weighted sum:

$$E(g, S) = E_{2D3D} + \lambda_1 E_{LP} + \lambda_2 E_{MD} + \lambda_3 E_{GP} + \lambda_4 E_S$$

(1)

in which the first two terms indicate forward projection errors of the coarse and the fine-scaled 3D hypotheses. The last

Figure 2: Our system takes a single image as input, and generates vehicles’ 3D shape and pose estimation in camera coordinates.
three terms, respectively, represent constraints enforced via unsupervised monocular depth, a ground plane assumption, and vehicle shape priors. In the next few sections we formalize and introduce details of our inference scheme, as reflected in the joint energy function (1), and how we generate the 3D hypotheses as well as the 2D pseudo-measurements via deep networks to facilitate its optimization.

3.1 Notation

We assume we are given a color image \( I : D \subset \mathbb{R}^2 \to \mathbb{S}^2 \) sampled as a positive-valued matrix. An axis-aligned subset \( b \subset D \) is called “2D bounding box”, and represented by the location of its center \( (t_x, t_y) \in \mathbb{R}^2 \), and scales \( (e^w, e^h) \in \mathbb{R}_+^2 \), all in pixel units and represented in exponential coordinates \((w, h)\) to ensure positivity. We assume the camera is calibrated so these can be converted to Euclidean coordinates. Equivalently, a 2D bounding box can be represented by an element of the scaled translation subgroup of the affine group in 2D:

\[
g_b = \begin{bmatrix} e^w & t_x \\ e^h & t_y \\ 1 \end{bmatrix} \in \mathbb{A}(2). \tag{2}
\]

In space, we call a gravity-aligned parallelepiped \( B \subset \mathbb{R}^3 \) a 3D bounding box, resting on the ground plane, whose shape is represented by three scales \( \sigma = (e^L, e^H, e^W) \in \mathbb{R}_+^3 \), again in exponential coordinates to ensure positivity, and whose pose is represented by its orientation \( \theta \in [0, 2\pi) \) and position on the ground plane \( T \in \mathbb{R}^3 \) relative to the camera reference frame. A 3D bounding box can also be represented as an element of the scaled translation subgroup of the affine group in 3D:

\[
g_B = \begin{bmatrix} R(\theta), e_d \end{bmatrix} \begin{bmatrix} e^L & T_X \\ e^H & T_Y \\ e^W & T_Z \\ 1 \end{bmatrix} \in \mathbb{A}(3) \tag{3}
\]

where \( R(\theta) \) is a rotation around \( Y \) by \( \theta \), so \( (R(\theta), T) \in \mathbb{SE}(3) \), and \( e^T_d = [0, 0, 0, 1] \). Here \( T_Y = 0 \) is the camera height from the ground plane. Assuming the ground plane is represented by its (scaled) normal vector \( N \in \mathbb{R}^3 \), we describe it as the locus \( \{ T \in \mathbb{R}^3 \mid N^T T = 1 \} \) (the ground plane cannot go through the optical center). Therefore, the vector \( T \) is subject to the constraint \( N^T T = 1 \).

We call \( S \subset \mathbb{R}^3 \) a shape, represented by a set of \( K \) points\(^1\) \( P_k \in \mathbb{R}^3 \) in a normalized reference frame. Note that \( p \in D \subset \mathbb{R}^2 \) are corresponding landmark points within the 2D bounding box. Equivalently, \( S \in \mathbb{R}^{3 \times K} / \mathbb{A}(3) \) is in the affine shape space (Kendall 1984) of \( K \) points, where the quotient is restricted to transformations of the form (3). \( Z : D \subset \mathbb{R}^2 \to \mathbb{R}_+^3 \) is a depth map, that associates to each pixel \((x, y)\) in \( D \) a positive scalar \( Z(x, y) \).

3.2 Inference scheme

As shown in Fig. 2, the inference criterion we use combines a generative component, whereby we jointly optimize the innovation (forward prediction error) between the projection of the 3D hypotheses and the image pseudo-measurements, monocular depth map constraints, geometric constraints (ground plane), in addition to penalizing large deformations of the shape prior. In the Supplementary Material we derive an approximation of the posterior of 3D pose and shape \( P(g_B, S|I) \), maximizing which is equivalent to minimizing the negative log here:

\[
-\log[P(g_B|g_B)P(p|g_B, S)P(Z_b|T_Z)P(T)P(S)] \tag{4}
\]

which is in accordance with (1). The first term is the 2D bounding box compatibility with the projected 3D bounding box. It is a coarse statistic of the geometric innovation.

\[
E_{2D3D} = \|g_b - \pi(g_B)\| \tag{5}
\]

where the subscript suggests 2D/3D consistency, and \( \pi \) denotes the central perspective map, assuming a calibrated camera. The second term is the fine-scaled geometric innovation, \( i.e., \), the distance between the predicted position of projected wireframe model vertices, and their 2D pseudo-measurements by a landmark detection network.

\[
E_{LP} = \sum_{k=1}^{K} \|p_k - \pi(g_B P_k)\|^2_2 \tag{6}
\]

where LP stands for landmark projection. To approximate the third term, we produce a dense range map of the current driving scene via an unsupervised monocular depth map estimation network.

\[
E_{MD} = \|T_Z - Z_b\| \tag{7}
\]

where MD means monocular depth, and \( Z_b \in \mathbb{R} \) is the average depth of an image crop specified by \( (I, g_b) : D \subset \mathbb{R}^2 \to \mathbb{S}^2 \). The fourth term assumes a geometric constraint by the ground plane, characterized by the normal vector \( N \).

\[
E_{GP} = \|N^T T - 1\| \tag{8}
\]

in which GP indicates ground plane. As generic regularizers, we also assume small deformations (shape coefficients \( \alpha_n \) close to their mean) of the shape model.

\[
E_S = \sum_{n=1}^{N} \|\alpha_n - \frac{1}{N} \sum_{n} \alpha_n\|^2_2 \tag{9}
\]

The overall loss function is the weighted sum, with multipliers \( \lambda \):

\[
E = E_{2D3D}(g_b, g_B) + \lambda_1 E_{LP}(p, g_B, P) + \lambda_2 E_{MD}(Z_b, T) + \lambda_3 E_{GP}(T) + \lambda_4 E_S(\alpha) \tag{10}
\]

3.3 Mono3D++ network

Our inference scheme leverages independence assumptions to factor the posterior probability into components, resulting in the compound loss described above. To initialize the minimization of (10), we separate it into modules that are implemented as deep networks, which is shown in Fig. 3. In the next paragraphs we describe each component in more details.

\(^1\)We overload the notation and use \( P \) for for points in space and probabilities. Which is which should be clear from the context.
2D Bounding box. We use a one-stage detection architecture similar to SSD (Liu et al. 2016) to estimate a posterior over (a regular subsampling of) the set of bounding boxes, \( P(l, g_b | I) \). Before setting putative bounding boxes on latent feature maps at different scales, we fuse feature maps from shallow and deep layers. It is shown in (Ren et al. 2017) that this step can improve both the semantic meaning and the resolution of the feature maps for better object detection.

\[
H_{P,Q}(l_j, \hat{b}_i) + \lambda d(g_b, \hat{g}_b) \chi(l_j) \tag{11}
\]

where \( j = j(i) = \arg \max_j \text{IoU}(b_j, \hat{b}_i) \). Here \( H \) denotes the cross-entropy between the true class posterior \( P(l | I) \) and the model realized by our network \( Q(l | I) \), and \( d \) is a distance in \( \mathbb{H}(2) \) that sums the Euclidean distance of the translational terms and the exponential coordinates of the scale terms, \( (w, h) \). Here \( i \) is the index of each putative bounding box (a.k.a. “anchor box”) which is sampled with negative/positive sample ratio of 3:1 from a larger group of regularly sampled anchor boxes on multiple latent feature maps. The index \( j \) to be chosen to match \( i \) consists of a data association problem (Bowman et al. 2017). We apply an indicator function \( \chi(\cdot) \) before the bounding box coordinate regression term so that this loss is only optimized for positive anchor boxes.

2D Landmark. We employ a stacked hourglass network (Newell, Yang, and Deng 2016), with skip-connections, to approximate the posterior of individual land-

\[
\frac{1}{K} \sum_{k=1}^{K} ||\hat{w}_{k,i} - w_{k,i}||^2 \tag{12}
\]

where \( \hat{w}_{k,i} \in \mathbb{R}^{64 \times 64} \) is the predicted heat map for the \( k \)th landmark and \( w_{k,i} \in \mathbb{R}^{64 \times 64} \) is a 2D Gaussian with standard deviation of 1 pixel centered at the \( k \)th landmark ground truth location. When a landmark is occluded or out of view, its ground truth heat map is set to all zeros. In inference, the MAP estimate \( p_k \in \mathbb{R}^2 \) is offset by a quarter of a pixel in the direction of its next highest neighbor before transforming back to the image coordinate space. Each vehicle is modeled by 14 landmarks.

3D Orientation and scale hypotheses. \( P(\theta, \sigma | I, \hat{g}_b) \) is approximated by a deep network with ResNet backbone (He et al. 2016), yielding \( Q(\hat{g}_b | I, \hat{g}_b) \) where \( I_{\hat{g}_b} \) is a \((64 \times 64) \) crop of the (centered and re-scaled) image \( I \in \mathbb{R}^{64 \times 64} \). We design a joint loss with multi-scale supervision for training pose and 3D scales.

\[
H_{P,Q}(\theta_j, \hat{\theta}_i) + \lambda d(\sigma_j, \hat{\sigma}_i) \tag{13}
\]

where \( Q \) is an approximation of \( P(\theta | I, \hat{g}_b) \), both of which are Von Mises distributions over the discrete set \( \{0, \ldots, 359\} \). We use a cross-entropy loss for orientation estimation. At inference time, the MAP estimate \( \theta_i \in [0, 2\pi) \) is used as a vehicle’s azimuth estimate. \( P(\sigma | I, \hat{g}_b, \theta) \) is estimated jointly with azimuth in the same network using the \( L^1 \) distance \( d(\sigma_j, \hat{\sigma}_i) \). Empirically, we found that imposing the orientation loss on the intermediate feature map and the size loss on the last layer generates better results than minimizing both losses on the same layer.

Shape hypotheses. The 3D morphable shape model is learnt using 2D landmarks via an EM-Gaussian method (Torresani, Hertzmann, and Bregler 2003; Kar et al. 2015). The hypothesis is that 3D object shapes are confined to a low-dimensional basis of the entire possible shape space. Therefore, the normalized 3D shape \( S_m \) is a \( \mathbb{R}^3 \times 1 \) of each vehicle instance can be factorized as the sum of the mean shape \( S \in \mathbb{R}^3 \times 1 \) of this category deformed using a linear combination of \( N \) basis shapes, \( V_n \in \mathbb{R}^3 \times 1 \). For each car, the orthographic projection constraint between its 3D shape \( S_m \) and 2D landmarks \( p_{k,m} \) is constructed as:

\[
p_{k,m} = c_m R_m (P_{k,m} + t_m) + \zeta_{k,m} \tag{14}
\]

\[
S_m = \bar{S} + \sum_{n=1}^{N} \alpha_{n,m} V_n \tag{15}
\]

\[
\zeta_{k,m} \sim N(0, \sigma^2 I_2 \times 2), \quad \alpha_{n,m} \sim N(0, 1), \quad R_m^T R_m = I_3 \times 3 \tag{16}
\]

in which \( c_m \in \mathbb{R}^{2 \times 2} \) is the scaling factor of the orthography (para-perspective projection). \( R_m \in \mathbb{R}^{2 \times 3} \) and \( t_m \in \mathbb{R}^{3 \times 1} \) are the orthographic rotation and translation of each object instance in the camera coordinate, respectively. \( P_{k,m} \in \mathbb{R}^{3 \times 3} \) represents the \( k \)th 3D landmark in \( S_m \).

Monocular depth. \( P(Z | I) \) is learned from stereo disparity, converted to depth using camera calibration. Similar to vehicle landmark detection, we use an hourglass network,
Table 2: 3D localization comparisons with monocular methods on KITTI val1/val2 by ALP of 1, 2 and 3 meters thresholds for 3D box center distance.

<table>
<thead>
<tr>
<th>Method</th>
<th>Type</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>3DVP</td>
<td>Mono</td>
<td>45.6</td>
<td>34.3</td>
<td>27.7</td>
<td>65.7</td>
<td>34.6</td>
<td>45.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SubCNN</td>
<td>Mono</td>
<td>39.3</td>
<td>31.0</td>
<td>26.0</td>
<td>70.5</td>
<td>56.2</td>
<td>47.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mono3D</td>
<td>1 / 46.0</td>
<td>1 / 38.3</td>
<td>1 / 34.0</td>
<td>1 / 71.0</td>
<td>1 / 59.9</td>
<td>1 / 55.8</td>
<td>-</td>
<td>1 / 80.3</td>
<td>1 / 69.3</td>
<td>1 / 62.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GoogleNet DeepMANTA</td>
<td>70.9 / 65.7</td>
<td>58.1 / 53.8</td>
<td>49.0 / 47.2</td>
<td>90.1 / 89.3</td>
<td>71.0 / 75.9</td>
<td>66.1 / 61.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DeepMANTA</td>
<td>66.9 / 69.7</td>
<td>53.2 / 54.4</td>
<td>44.4 / 47.8</td>
<td>88.3 / 91.0</td>
<td>74.3 / 76.4</td>
<td>63.6 / 61.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mono3D++</td>
<td>80.6 / 80.2</td>
<td>67.7 / 65.1</td>
<td>56.0 / 54.6</td>
<td>93.3 / 92.7</td>
<td>83.0 / 80.8</td>
<td>71.8 / 70.5</td>
<td>95.0 / 95.4</td>
<td>86.7 / 85.4</td>
<td>76.2 / 75.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: 3D localization comparisons with the state-of-the-art stereo method by ALP under 3D box center distance thresholds of 1, 2 and 3 meters. Note that our method only uses a single image for inference, while 3DOP needs stereo-pairs.

4 Experiments

We evaluate our method on the KITTI object detection benchmark. This dataset contains 7,481 training images and 7,518 test images. To facilitate comparison with competing approaches, we isolate a validation set from the training set according to the same protocol of (Xiang et al. 2015; 2016; Chabot et al. 2017) called (train1, val1), and the same used by (Chen et al. 2015; 2016; Chabot et al. 2017) called (train2, val2). We report results using five evaluation metrics: three for 3D and two for 2D localization. For the former, we use average localization precision (ALP) (Xiang et al. 2015), average precision based on 3D intersection-over-union (IoU), AP3D, and bird’s eye view based average localization precision, APloc (Geiger, Lenz, and Urtasun 2012). Although 2D localization is not our goal, as a sanity check we also measure average precision (AP) and average orientation similarity (AOS).

ALP is based on the distance between the center of the detected 3D boxes and the annotated ground truth. A 3D detection is correct if the distance is below a threshold. AP3D is computed from the IoU of 3D bounding boxes. APloc is obtained by projecting the 3D bounding boxes to the ground plane (bird’s eye view) and computing the 2D IoU with ground truth. Note that while ALP only measures distance between centers, both AP3D and APloc jointly evaluate a vehicle’s translation, orientation and 3D scales. AOS measures 2D orientation relative to ground truth.

The paragon set for our experiments consists of 3DVP (Xiang et al. 2015), SubCNN (Xiang et al. 2017), Mono3D (Chen et al. 2016), 3DOP (Chen et al. 2015) and DeepMANTA (Chabot et al. 2017). While 3DOP is the state-of-the-art stereo method, the rest are monocular. Among the monocular methods, DeepMANTA is the current state-of-the-art. Also related to our method for monocular 3D vehicle detection is Deep3DBox (Mousavian et al. 2017), which however used different evaluation metrics from the ones...
Figure 4: Recall/3D localization precision curves for 1 meter (left), 2 meters (middle) and 3 meters (right) precision on KITTI val2. Solid lines are our results. Dashed lines are Mono3D. Dash-dot lines are 3DOP, which uses stereo-pairs at inference time.

Table 4: Comparisons on AP$_{3D}$ under different 3D IoU thresholds with both monocular and stereo methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Type</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono3D</td>
<td>Mono</td>
<td>62.9 / 25.2 / 2.5</td>
<td>48.2 / 18.2 / 2.3</td>
<td>42.7 / 15.5 / 2.3</td>
</tr>
<tr>
<td>Mono3D++</td>
<td>Mono</td>
<td>48.2 / 18.2 / 2.3</td>
<td>34.3 / 11.5 / 2.3</td>
<td>28.1 / 10.1 / 2.3</td>
</tr>
<tr>
<td>3DOP</td>
<td>Stereo</td>
<td>35.0 / 5.5 / 1.5</td>
<td>26.4 / 5.5 / 1.5</td>
<td>20.0 / 4.8 / 1.5</td>
</tr>
</tbody>
</table>

Table 5: Comparisons on AP$_{loc}$ under different 2D IoU thresholds with both monocular and stereo methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Type</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono3D</td>
<td>Mono</td>
<td>13.6 / 12.2 / 9.02</td>
<td>8.75 / 8.99 / 16.43</td>
<td>9.02 / 11.5 / 22.4</td>
</tr>
<tr>
<td>Mono3D++</td>
<td>Mono</td>
<td>68.33 / 64.5 / 50.8</td>
<td>55.00 / 50.0 / 50.5</td>
<td>50.8 / 47.6 / 35.6</td>
</tr>
<tr>
<td>3DOP</td>
<td>Stereo</td>
<td>76.1 / 71.4 / 64.5</td>
<td>64.5 / 60.2 / 53.6</td>
<td>50.5 / 48.2 / 50.1</td>
</tr>
</tbody>
</table>

Table 6: Ablation studies on val1/val2 by ALP under 3D box center distance threshold of 1 meter.

<table>
<thead>
<tr>
<th>Method</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>v1</td>
<td>13.6 / 12.2</td>
<td>11.3 / 12.3</td>
<td>11.3 / 12.3</td>
</tr>
<tr>
<td>v2</td>
<td>68.33 / 64.5</td>
<td>55.00 / 50.5</td>
<td>50.8 / 47.6</td>
</tr>
<tr>
<td>v3</td>
<td>76.1 / 71.4</td>
<td>64.5 / 60.2</td>
<td>50.5 / 48.2</td>
</tr>
<tr>
<td>v4</td>
<td>79.45 / 71.8</td>
<td>62.76 / 59.1</td>
<td>52.79 / 50.5</td>
</tr>
</tbody>
</table>

Table 7: Ablation studies on val1/val2 by AP$_{3D}$ with 3D IoU threshold of 0.25.
Table 8: Ablation studies on val1/val2 by AP$_{loc}$ under 2D IoU threshold of 0.5.

<table>
<thead>
<tr>
<th>Method</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>v1</td>
<td>2.06 / 2.27</td>
<td>2.30 / 2.27</td>
<td>2.29 / 2.36</td>
</tr>
<tr>
<td>v2</td>
<td>37.27 / 30.18</td>
<td>27.48 / 24.82</td>
<td>23.67 / 21.49</td>
</tr>
<tr>
<td>v3</td>
<td>32.08 / 37.25</td>
<td>22.12 / 28.30</td>
<td>25.84 / 24.14</td>
</tr>
<tr>
<td>v4</td>
<td>30.50 / 46.68</td>
<td>36.85 / 34.32</td>
<td>29.05 / 28.13</td>
</tr>
</tbody>
</table>

Figure 5: Typical issues (e.g. partial occlusions, shadow, low resolution) that affect 2D vehicle landmark measurements.

Qualitative Results. Fig. 1 shows representative outputs of our method, including cars at different scales, 3D shapes, poses and occlusion patterns. Fig. 5 illustrates typical issues addressed by jointly modeling a vehicle’s coarse and fine-scaled 3D shape and pose representations. When vehicle landmarks suffer from partial occlusions, shadow or low resolution, we can still leverage 2D bounding boxes in order to enforce the two-scale geometric innovation constraints.

Generality. Although not our interested targets, chairs share similar constraints to vehicles like the two-scale 3D hypotheses innovation, a ground plane assumption, shape priors, etc. Thus to demonstrate the generality of our method, we also test on chairs and report results in the Supplementary Material.

Failure Modes. In Fig. 6 we illustrate some failure cases, which include field of view truncation, causing the bounding box projection constraint $E_{2D3D}$ in the overall energy to enforce the 3D bounding box’s 2D projection to be within the truncated 2D box measurement. Failures can also occur due to inaccurate orientation estimation, and under-representation in the training set (oversized SUV) which causes the normalized morphable wireframe model to be rescaled by incorrect 3D size estimation.

5 Conclusion

We have presented a method to infer vehicle pose and shape in 3D from a single RGB image that considers both the coarse and the fine-scaled 3D hypotheses, and multiple task priors, as reflected in an overall energy function (10) for joint optimization. Our inference scheme leverages independence assumptions to decompose the posterior probability of pose and shape given an image into a number of factors. For each term we design a loss function that is initialized by output from deep network as shown in Fig. 3. Our method shows significant improvements over state-of-the-art monocular 3D vehicle detectors under various evaluation settings.

References


Chen, X.; Kundu, K.; Zhu, Y.; Berneshawi, A.; Ma, H.; Fidler, S.; and Urtasun, R. 2015. 3d object proposals for accurate object class detection. In NIPS.


