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

Tong He,[1,2] **Stefano Soatto**[1]
[1]University of California, Los Angeles, [2]Megvii (Face++)
{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 optimize 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 representation of vehicle shape and pose. To reduce its sensitivity to 2D landmarks, we jointly model the 3D bounding box as a coarse representation which improves 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.

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. 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 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 scoring mechanisms; (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). 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. We conduct ablation studies to demonstrate benefits brought by jointly modeling the coarse and the fine-scaled 3D object pose and shape representations.

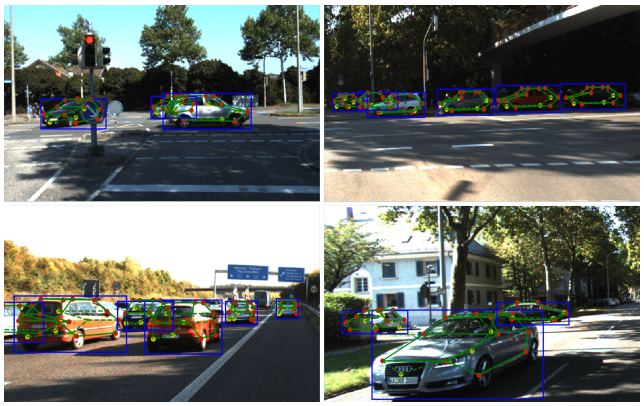

Figure 1: Representative 3D detection results, shown as projections on the input images. 3D morphable shape models of each vehicle are colored in green. Dashed green lines are occluded edges. For the 14 vertices of each morphable shape model, visible vertices are colored in red and occluded ones in yellow.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

For the second problem, we consider three constraints on vehicles. 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 a monocular range map of the current driving scene. These constraints are jointly modeled with 3D-2D consistency terms in order to further improve vehicle shape and pose estimation.

In summary, in this paper we propose an approach for vehicle 3D shape and pose estimation from a single image that

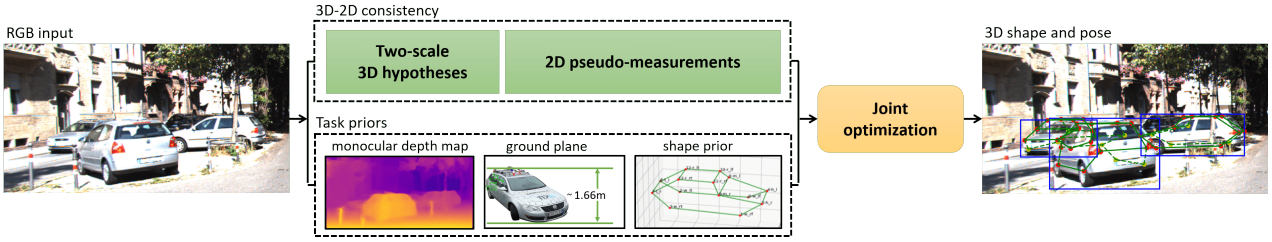

Figure 2: Our system takes a single image as input, and generates vehicles' 3D shape and pose estimation in camera coordinates.

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 improves the 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.

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-point 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 $S \subset \mathbb{R}^3$, 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 [1] we describe all the modeling assumptions needed to arrive at a tractable approximation of the posterior, maximizing which is equivalent to minimizing the weighted sum:

$$E(g, S) = E_{\text{2D3D}} + \lambda_1 E_{\text{LP}} + \lambda_2 E_{\text{MD}} + \lambda_3 E_{\text{GP}} + \lambda_4 E_{\text{S}} \quad (1)$$

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

[1] https://tonghehehe.com/det3d

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:

$$g_B = [R(\theta),\ e_4] \begin{bmatrix} e^L & & & T_X \\ & e^H & & T_Y \\ & & e^W & T_Z \\ & & & 1 \end{bmatrix} \in \mathbb{A}(3) \tag{3}$$

where $R(\theta)$ is a rotation around Y by θ, so $(R(\theta), T) \in SE(3)$, and $e_4^T = [0,\ 0,\ 0,\ 1]$. 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[2] $P_k \in \mathbb{R}^3$ 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). $Z : D \subset \mathbb{R}^2 \to \mathbb{R}_+$ 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:

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

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.

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

where the subscript suggests 2D/3D consistency, and π 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

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

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.

$$E_{\text{MD}} = \|T_Z - Z_b\| \tag{7}$$

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.

$$E_{\text{GP}} = \|N^T T - 1\| \tag{8}$$

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

$$E_{\text{S}} = \sum_{n=1}^{N} \|\alpha_n - \frac{1}{N}\sum_{n} \alpha_n\|_2^2 \tag{9}$$

The overall loss function is the weighted sum, with multipliers λ:

$$\begin{aligned} E = {} & E_{\text{2D3D}}(g_b, g_B) + \lambda_1 E_{\text{LP}}(p, g_B, P) \\ & + \lambda_2 E_{\text{MD}}(Z_b, T) + \lambda_3 E_{\text{GP}}(T) + \lambda_4 E_{\text{S}}(\alpha) \end{aligned} \tag{10}$$

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.

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

[2] We overload the notation and use P for for points in space and probabilities. Which is which should be clear from the context.

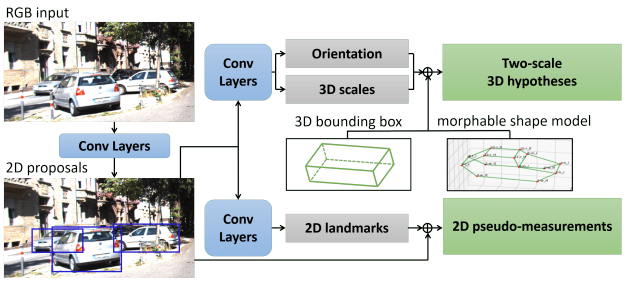

Figure 3: The two-scale 3D hypotheses consist of the rotated and scaled 3D bounding box and morphable wireframe model. The image pseudo-measurements include 2D bounding boxes and landmarks. In our inference scheme, we use the hypotheses and the pseudo-measurements to initialize the optimization of (10) and generate the final 3D pose and shape estimation of a vehicle.

this step can improve both the semantic meaning and the resolution of the feature maps for better object detection.

$$H_{P,Q}(l_j, \hat{l}_i) + \lambda d(g_{b_j}, \hat{g}_{b_i})\chi(l_j) \tag{11}$$

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 $Q(\hat{l}|I)$, and d is a distance in $\mathbb{A}(2)$ 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 landmarks within each 2D bounding box $P(p_k|I, \hat{g}_b)$. We use the mean-square error as loss.

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

where $\hat{w}_{k,i} \in \mathbb{R}^{64\times64}$ is the predicted heat map for the k_{th} landmark and $w_{k,i} \in \mathbb{R}^{64\times64}$ 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. 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 (64×64) crop of the (centered and re-scaled) image $I \circ \hat{g}_b{}^{-1}$. We design a joint loss with multi-scale supervision for training pose and 3D scales.

Method	3DVP	3DOP	Mono3D	GoogLenet DeepMANTA	VGG16 DeepMANTA	Mono3D++
Type	Mono	Stereo	Mono	Mono	Mono	Mono
Time	40 s	3 s	4.2 s	0.7 s	2 s	**0.6 s**

Table 1: Inference time comparisons against a paragon set of both monocular and stereo methods.

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

where Q is an approximation of $P(\theta|I, \hat{g}_b)$, both of which are Von Mises distributions over the discrete set $\{0, \ldots, 359^o\}$. We use a cross-entropy loss for orientation estimation. At inference time, the MAP estimate $\hat{\theta}_i \in [0, 2\pi)$ is used as a vehicle's azimuth estimation. $P(\sigma|I, \hat{g}_b, \theta)$ 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 $S_m \in \mathbb{R}^{3K\times1}$ of each vehicle instance can be factorized as the sum of the mean shape $\bar{S} \in \mathbb{R}^{3K\times1}$ of this category deformed using a linear combination of N basis shapes, $V_n \in \mathbb{R}^{3K\times1}$. For each car, the orthographic projection constraint between its 3D shape S_m and 2D landmarks $p_{k,m}$ is constructed as:

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

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

$$\zeta_{k,m} \sim N(0, \sigma^2 I_{2\times2}), \quad \alpha_{n,m} \sim N(0, 1), \quad R_m^T R_m = I_{3\times3} \tag{16}$$

in which $c_m \in \mathbb{R}^{2\times2}$ is the scaling factor of the orthography (para-perspective projection). $R_m \in \mathbb{R}^{2\times3}$ and $t_m \in \mathbb{R}^{3\times1}$ are the orthographic rotation and translation of each object instance in the camera coordinate, respectively. $P_{k,m} \in \mathbb{R}^{3\times1}$ 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, with skip-connections, to predict per-pixel disparity (Godard, Aodha, and Brostow 2017). The left view is input to the encoder, and the right view used as supervision for appearance matching at multiple output scales of the decoder. The total loss is accumulated across four output scales combining three terms. One measures the quality of image matching, one measures disparity smoothness, and the last measures left/right disparity consistency. Next we describe each term relative to the left view. Each term is replicated

	1 meter			2 meters			3 meters		
Method	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
3DVP	45.6 / -	34.3 / -	27.7 / -	65.7 / -	54.6 / -	45.6 / -	- / -	- / -	- / -
SubCNN	39.3 / -	31.0 / -	26.0 / -	70.5 / -	56.2 / -	47.0 / -	- / -	- / -	- / -
Mono3D	- / 46.0	- / 38.3	- / 34.0	- / 71.0	- / 59.9	- / 53.8	- / 80.3	- / 69.3	- / 62.7
GoogLenet DeepMANTA	70.9 / 65.7	58.1 / 53.8	49.0 / 47.2	90.1 / 89.3	77.0 / 75.9	66.1 / 67.3	- / -	- / -	- / -
VGG16 DeepMANTA	66.9 / 69.7	53.2 / 54.4	44.4 / 47.8	88.3 / 91.0	74.3 / 76.4	63.6 / 67.8	- / -	- / -	- / -
Mono3D++	**80.6 / 80.2**	**67.7 / 65.1**	**56.0 / 54.6**	**93.3 / 92.7**	**83.0 / 80.8**	**71.8 / 70.5**	**95.0 / 95.4**	**86.7 / 85.4**	**76.2 / 75.9**

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.

for the right view to enforce left-view consistency. Appearance is measured by

$$\frac{1}{D}\sum_{a,b}\beta\frac{1-SSIM(I^l_{ab},\widetilde{I}^l_{ab})}{2}+(1-\beta)||I^l_{ab}-\widetilde{I}^l_{ab}||_1 \tag{17}$$

which combines single-scale SSIM (Wang et al. 2004) and the L^1 distance between the input image I^l and its reconstruction $\widetilde{I}^l$ obtained by warping I^r using the disparity d^r with a differentiable image sampler from the spatial transformer network (Jaderberg et al. 2015). Disparity smoothness is measured by

$$\frac{1}{D}\sum_{a,b}|\partial_x d^l_{ab}|e^{-||\partial_x I^l_{ab}||}+|\partial_y d^l_{ab}|e^{-||\partial_y I^l_{ab}||} \tag{18}$$

which contains an edge-aware term depending ∂I (Heise et al. 2013). Finally, left/right consistency is measured using the L^1 norm.

$$\frac{1}{D}\sum_{a,b}\|d^l_{ab}-d^r_{ab-d^l_{ab}}\|_1. \tag{19}$$

3.4 Implementation

It takes about one week to train the 2D bounding box network, and two hours for the orientation/3D scale network on KITTI with 4 TITAN-X GPUs. The landmark detector is trained on Pascal3D. The training process for the monocular depth estimation network is unsupervised using KITTI stereo-pairs, which takes around 5 to 12 hours depending on the amount of data available. In theory, these deep networks could be unified into a single one and trained jointly, but this is beyond our scope here. Learning the morphable shape model takes about 2.5 minutes using 2D vehicle landmarks. At inference time, we use the Ceres solver (Agarwal, Mierle, and Others) to optimize the weighted loss (10). On average it converges in 1.5 milliseconds within about 15 iterations. Detailed timing comparisons are shown in Table 1.

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; 2017; 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), AP_{3D}, and bird's eye view based average localization precision, AP_{loc} (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).

		1 / 2 / 3 meters		
Method	Type	Easy	Moderate	Hard
Mono3D++	Mono	80.2 / 92.7 / 95.4	65.1 / 80.8 / 85.4	54.6 / 70.5 / 75.9
3DOP	Stereo	78.6 / 87.4 / 89.5	66.9 / 80.0 / 84.2	59.4 / 71.9 / 76.0

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.

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. AP_{3D} is computed from the IoU of 3D bounding boxes. AP_{loc} 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 AP_{3D} and AP_{loc} 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 above, thus preventing direct comparison. For object 2D orientation and bounding box evaluation, we also compare with Faster-RCNN (Ren et al. 2015) as well as Deep3DBox.

3D Localization. We use ALP with distance thresholds of 1, 2 and 3 meters in Table 2 including both val1 and val2. Our method improves the state-of-the-art monocular method, DeepMANTA, by 10.5% on average. Even though our method is monocular, we compare to the stereo method 3DOP using val2 in Table 3. Surprisingly, we outperform

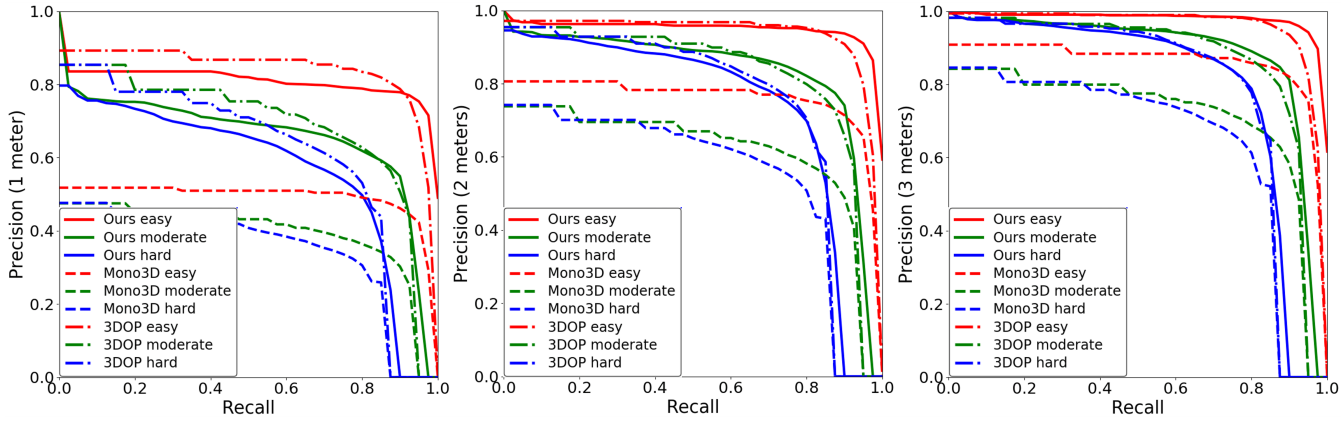

Figure 4: Recall/3D localization precision curves for 1 meter (left), 2 meters (middle) and 3 meters (right) precision on KITTI val2. Solid lines are our results. Dashed lines are Mono3D. Dash-dot lines are 3DOP, which uses stereo-pairs at inference time.

		3D IoU 0.25 / 0.50 / 0.70		
Method	Type	Easy	Moderate	Hard
Mono3D	Mono	62.9 / 25.2 / 2.5	48.2 / 18.2 / 2.3	42.7 / 15.5 / 2.3
Mono3D++	Mono	**71.9 / 42.0 / 10.6**	**59.1 / 29.8 / 7.9**	**50.5 / 24.2 / 5.7**
3DOP	Stereo	85.5 / 46.0 / 6.6	68.8 / 34.6 / 5.1	64.1 / 30.1 / 4.1

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

		2D IoU 0.50 / 0.70		
Method	Type	Easy	Moderate	Hard
Mono3D	Mono	30.5 / 5.2	22.4 / 5.2	19.2 / 4.1
Mono3D++	Mono	**46.7 / 16.7**	**34.3 / 11.5**	**28.1 / 10.1**
3DOP	Stereo	55.0 / 12.6	41.3 / 9.5	34.6 / 7.6

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

3DOP on "easy" and "moderate" cases and is comparable on "hard" case. Detailed comparisons by precision/recall curves are shown in Fig. 4. We outperform the monocular by large margins and is better than the stereo in some cases.

3D Detection and Bird's Eye View Localization. We use AP_{3D} with 3D IoU thresholds of 0.25, 0.5 and 0.7, as well as AP_{loc} with 2D IoU thresholds of 0.5 and 0.7. Table 4 shows comparisons with both monocular and stereo methods using val2. Our method surpasses all monocular ones uniformly. Though the monocular setting is more challenging than the stereo one due to the lack of depth, our method still outperforms 3DOP by about 39% to 61% on AP_{3D} with IoU threshold of 0.7. Table 5 shows comparison on AP_{loc} with both monocular and stereo using val2. Again, our results surpass monocular ones uniformly. Even if compared with the stereo method, we gain around 21% to 33% relative improvement on AP_{loc} under 0.7 IoU.

Ablation Studies. In Table 6, 7 and 8 we use val1 and val2 with ALP, AP_{3D} and AP_{loc} to validate our joint modeling of the coarse and the fine-scaled 3D hypotheses, as well as task priors. "v1" indicates our inference scheme at initialization; "v2" only models the coarse geometric innovation and a ground plane constraint; "v3" adds the fine-scaled geometric innovation and vehicle shape priors. Best results are achieved by our overall model "v4", which further considers unsupervised monocular depth. Due to the page limit, extended comparisons over different threshold values are reported in the Supplementary Material.

2D Detection and Orientation. As a sanity check, the 2D detection AP and AOS are also evaluated with monocular and stereo methods. Our estimation is on par with the state-of-the-art results. Detailed comparisons are included in the Supplementary Material.

Method	Easy	Moderate	Hard
v1	13.6 / 16.9	12.2 / 13.3	11.3 / 12.5
v2	68.5 / 68.2	58.3 / 57.5	50.8 / 47.6
v3	76.1 / 73.2	64.5 / 60.2	53.6 / 50.0
v4	**80.6 / 80.2**	**67.7 / 65.1**	**56.0 / 54.6**

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

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 focus here, chairs share similar constraints to vehicles like the two-scale 3D hypotheses innovation, a ground plane assumption, shape priors, etc.

Method	Easy	Moderate	Hard
v1	10.50 / 11.50	8.75 / 8.99	9.02 / 16.43
v2	68.33 / 58.50	55.00 / 49.96	49.17 / 45.09
v3	71.39 / 66.59	59.06 / 54.88	50.59 / 48.26
v4	**79.45 / 71.86**	**62.76 / 59.11**	**52.79 / 50.53**

Table 7: Ablation studies on val1/val2 by AP_{3D} with 3D IoU threshold of 0.25.

Method	Easy	Moderate	Hard
v1	2.06 / 2.27	2.30 / 2.27	2.29 / 2.36
v2	37.27 / 30.18	27.48 / 24.82	23.67 / 21.49
v3	42.68 / 37.25	32.12 / 28.50	25.84 / 24.14
v4	**50.50 / 46.68**	**36.85 / 34.32**	**29.05 / 28.13**

Table 8: Ablation studies on val1/val2 by AP_{loc} under 2D IoU threshold of 0.5.

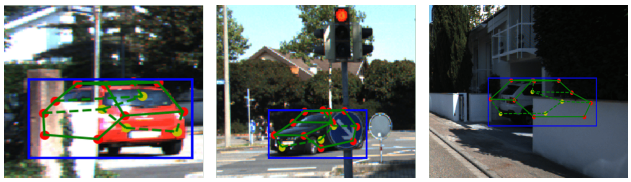

Figure 5: Typical issues (e.g. partial occlusions, shadow, low resolution) that affect 2D vehicle landmark measurements.

Thus to demonstrate the generality of our method, we also test on chairs and report results in the Supplementary Material.

Figure 6: Failure cases caused by field of view truncation, inaccurate orientation or 3D scale estimation.

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 underrepresentation 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 improves the state-of-the-art for monocular 3D vehicle detection under various evaluation settings.

Acknowledgement

Research supported by ONR N00014-17-1-2072, N00014-13-1-034 and ARO W911NF-17-1-0304.

References

Agarwal, S.; Mierle, K.; and Others. Ceres solver. http://ceres-solver.org.

Bowman, S. L.; Atanasov, N.; Daniilidis, K.; and Pappas, G. J. 2017. Probabilistic data association for semantic slam. *2017 IEEE International Conference on Robotics and Automation (ICRA)* 1722–1729.

Chabot, F.; Chaouch, M. A.; Rabarisoa, J.; Teulière, C.; and Chateau, T. 2017. Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1827–1836.

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*.

Chen, X.; Kundu, K.; Zhang, Z.; Ma, H.; Fidler, S.; and Urtasun, R. 2016. Monocular 3d object detection for autonomous driving. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2147–2156.

Chen, X.; Ma, H.; Wan, J.; Li, B.; and Xia, T. 2017. Multi-view 3d object detection network for autonomous driving. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 6526–6534.

Chhaya, F.; Reddy, N. D.; Upadhyay, S.; Chari, V.; Zia, M. Z.; and Krishna, K. M. 2016. Monocular reconstruction of vehicles: Combining slam with shape priors. *2016 IEEE International Conference on Robotics and Automation (ICRA)* 5758–5765.

Choy, C. B.; Stark, M. J.; Corbett-Davies, S.; and Savarese, S. 2015. Enriching object detection with 2d-3d registration and continuous viewpoint estimation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2512–2520.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* 1:886–893 vol. 1.

Dong, J.; Fei, X.; and Soatto, S. 2017. Visual-inertial-semantic scene representation for 3d object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 3567–3577.

Fei, X., and Soatto, S. 2018. Visual-inertial object detection and mapping. *CoRR* abs/1806.08498.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. *2012 IEEE Conference on Computer Vision and Pattern Recognition* 3354–3361.

Godard, C.; Aodha, O. M.; and Brostow, G. J. 2017. Unsupervised monocular depth estimation with left-right consistency. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 6602–6611.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778.

Heise, P.; Klose, S.; Jensen, B.; and Knoll, A. 2013. Pm-huber: Patchmatch with huber regularization for stereo

matching. *2013 IEEE International Conference on Computer Vision* 2360–2367.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; and Kavukcuoglu, K. 2015. Spatial transformer networks. In *NIPS*.

Kar, A.; Tulsiani, S.; Carreira, J.; and Malik, J. 2015. Category-specific object reconstruction from a single image. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1966–1974.

Kendall, D. G. 1984. Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society* 16(2):81–121.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S. E.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *ECCV*.

Mottaghi, R.; Xiang, Y.; and Savarese, S. 2015. A coarse-to-fine model for 3d pose estimation and sub-category recognition. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 418–426.

Mousavian, A.; Anguelov, D.; Flynn, J. J.; and Kosecka, J. 2017. 3d bounding box estimation using deep learning and geometry. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 5632–5640.

Murthy, J. K.; Sharma, S.; and Krishna, K. M. 2017. Shape priors for real-time monocular object localization in dynamic environments. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* 1768–1774.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *ECCV*.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*.

Qi, C. R.; Liu, W.; Wu, C.; Su, H.; and Guibas, L. J. 2018. Frustum pointnets for 3d object detection from rgb-d data. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE.

Ren, Z., and Sudderth, E. B. 2018. 3d object detection with latent support surfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 937–946.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39:1137–1149.

Ren, J. S. J.; Chen, X.; Liu, J.; Sun, W.; Pang, J.; Yan, Q.; Tai, Y.-W.; and Xu, L. 2017. Accurate single stage detector using recurrent rolling convolution. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 752–760.

Song, S., and Chandraker, M. K. 2015. Joint sfm and detection cues for monocular 3d localization in road scenes. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 3734–3742.

Torresani, L.; Hertzmann, A.; and Bregler, C. 2003. Learning non-rigid 3d shape from 2d motion. In *NIPS*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13:600–612.

Xiang, Y.; Choi, W.; Lin, Y.; and Savarese, S. 2015. Data-driven 3d voxel patterns for object category recognition. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1903–1911.

Xiang, Y.; Choi, W.; Lin, Y.; and Savarese, S. 2017. Subcategory-aware convolutional neural networks for object proposals and detection. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* 924–933.

Xu, D.; Anguelov, D.; and Jain, A. 2018. Pointfusion: Deep sensor fusion for 3d bounding box estimation.

Zia, M. Z.; Stark, M. J.; Schiele, B.; and Schindler, K. 2011. Revisiting 3d geometric models for accurate object shape and pose. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)* 569–576.

Zia, M. Z.; Stark, M. J.; and Schindler, K. 2013. Explicit occlusion modeling for 3d object class representations. *2013 IEEE Conference on Computer Vision and Pattern Recognition* 3326–3333.

Zia, M. Z.; Stark, M. J.; and Schindler, K. 2014a. Are cars just 3d boxes? jointly estimating the 3d shape of multiple objects. *2014 IEEE Conference on Computer Vision and Pattern Recognition* 3678–3685.

Zia, M. Z.; Stark, M. J.; and Schindler, K. 2014b. Towards scene understanding with detailed 3d object representations. *International Journal of Computer Vision* 112:188–203.