Holistic 3D Scene Understanding from a Single Geo-tagged Image

Shenlong Wang, Sanja Fidler and Raquel Urtasun
Department of Computer Science
University of Toronto
{slwang, fidler, urtasun}@cs.toronto.edu

Abstract

In this paper we are interested in exploiting geographic priors to help outdoor scene understanding. Towards this goal we propose a holistic approach that reasons jointly about 3D object detection, pose estimation, semantic segmentation as well as depth reconstruction from a single image. Our approach takes advantage of large-scale crowd-sourced maps to generate dense geographic, geometric and semantic priors by rendering the 3D world. We demonstrate the effectiveness of our holistic model on the challenging KITTI dataset [13], and show significant improvements over the baselines in all metrics and tasks.

1. Introduction

Inferring 3D semantic and geometric information from a single monocular image has been one of the holy grails of computer vision since the beginning [15]. In the nineties, most approaches to recover geometry were so-called Shape-from-X, where local surface orientations were inferred by exploiting texture [2], shading [33], defocus [23], and contours/silhouettes [5]. Later, learning based methods took over [36, 20, 29] inferring depth from local patches. The Manhattan world assumption was also exploited to estimate the layout of indoor scenes [18, 38], producing impressive reconstructions in this relatively limited setting.

A variety of approaches have been proposed to infer 3D objects from a single monocular image. The most common approach has been to employ a bank of 2D detectors, each trained for a specific viewpoint [37, 42, 8, 14]. Methods enabling continuous pose representation also emerged, typically yielding 3D bounding boxes around objects [19, 9, 30] or even polygonal 3D shapes [44]. Since 3D inference typically relies on lifting the 2D detections to the metric world by imposing real-physical object dimensions, having accurate localization is of crucial importance.

Holistic approaches aim at reasoning jointly about multiple related tasks. 3D reconstruction is ill-posed, however, knowledge about the objects and the scene semantics can be used to help resolve some of the inherent ambiguities. Furthermore, knowing the 3D structure of the world should simplify recognition. Recently, holistic approaches that jointly reason about both reconstruction and recognition tasks have been proposed [27, 19, 39, 16], resulting in impressive performance gains over techniques that tackle a single task.

Here we argue that there is much more prior information that one could use and is freely available. We live in an era where technology and social networks are part of our everyday's life. A single monocular image is thus no longer our only source of information, a whole cyber world sits behind it. In this paper we make use of geotagged images, and propose priors derived from map data which contains information about the scene, such as the geolocation and rough shape of roads, buildings and trees. Towards this goal, we make use of OpenStreetMaps (OSM) which is freely available. OSM is an open-source crowd-sourcing project which collects GPS trajectories of the users from their daily driving routines. Since its beginning in 2004, OSM now has more than one million contributors from around the globe. The road coverage is 50% of all the roads in the world. With a simple download click, the full world sits in our laptop.

In this paper we propose a holistic conditional random field (CRF) that reasons jointly about 3D object detection, pose estimation, semantic segmentation as well as depth reconstruction from a single image. Our approach takes advantage of large-scale crowd-sourced maps to generate dense geographic, geometric and semantic priors by rendering the 3D world. We demonstrate the effectiveness of our approach on the challenging KITTI dataset [13], and show significant improvements over the baselines in all tasks.

2. Related Work

Contextual models in computer vision: Contextual models have been used to improve scene understanding tasks, including object detection [46, 28, 32], semantic segmentation [26, 28, 27] and 3D reconstruction [27, 18]. Geometric context [21, 6] aims at capturing coarse 3D geometric structure of the scene in the form of surface orientations

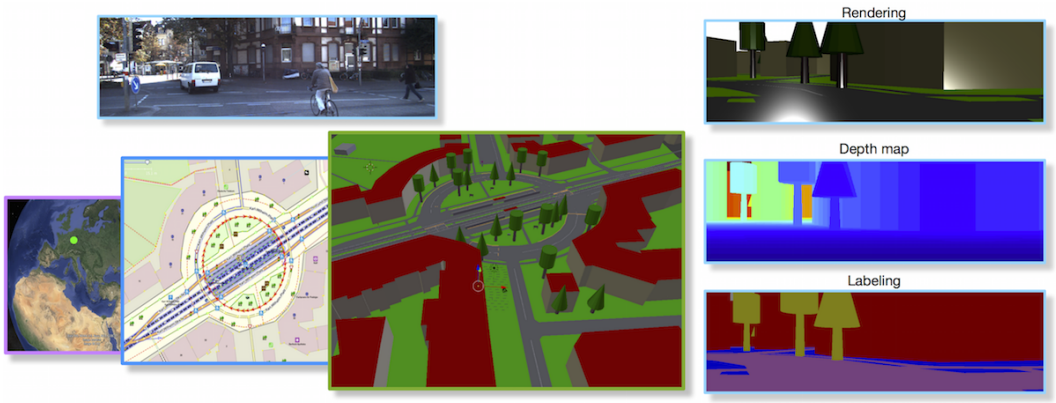

Figure 1. Maps can bring rich 3D information. Given the geolocation, a local 3D world (middle) can be built according to OpenStreetMaps (bottom left) [1]. Given the camera pose we can render a local 3D world (right) to create strong geometric and semantic priors. (*e.g.* labels).

[21], layout structure [18], relative size and ground-plane [46]. It has been intensively used for various computer vision tasks, such as object localization [9], support inference [40] and occlusion reasoning [46]. Co-occurrence statistics [26, 45, 12], canonical size, shape and appearance of certain objects [46, 22, 9] in the real-world can be exploited to efficiently help both detection and segmentation. Photogrammetric context [22, 27, 18], such as camera intrinsics and extrinsic parameters, significantly helps object location and relative scale estimation. Geolocalization can be used to provide side information about the scene type, elevation and semantics [32, 4, 12, 17]. However, most previous works exploit geographic context in a data-driven manner, which requires building a gigantic and expensive geo-referenced image database. In contrast, in this paper we make use of freely available maps to provide semantic, geographic and geometric context.

Holistic scene understanding: It is well-accepted that reasoning about semantics can help reconstruction, and vice versa. Holistic models aim at solving many of these tasks jointly. For instance, [26, 45, 10, 28] conduct object detection and image labeling jointly. [27, 31] improve both depth reconstruction and semantic labeling with a holistic model. [40, 46] simultaneously detect objects and reason about the interactions between objects in the scene.

Crowd-sourcing maps: Maps obtained via crowd-sourcing can provide us with useful yet inexpensive information about the context of the scene. Maps have already been intensively used in autonomous driving scenarios for navigation and localization [4, 11]. Matzen and Snavely [32] explore geographic information for recognition tasks. In particular, they re-score box proposal for 2D vehicle detection and 3D pose estimation. In contrast, in this paper we jointly reason about depth, pixel-wise labelings as well as 3D object detection. Moreover, instead of using map meta-data directly, we render a 3D scene to encode semantic and geometric prior for our holistic model.

3. Building 3D Scene Priors From OSM

We make use of OpenStreetMaps, a freely available map dataset to extract geographic information useful for reconstruction and recognition tasks. OSM is a polygon based map representation in the world geodetic system (WGS), with rich labels such as building, road and tree. We refer the reader to the left bottom subfigure in Fig. 1 for an illustration of the data.

Given a geotagged image as well as the camera parameters, we extract a large local region of the map around the area of interest. Based on this 2D cartographic information and limited 3D information like elevation, a visual 3D world can then be easily built from OSM by extending the objects along the vertical direction, as shown in Fig. 1. In this paper, our 2D-to-3D transformation is based on OSM2World, which we modified to model buildings and trees. Moreover, we develop an OpenGL-based renderer to visualize the local world using generic textures, semantic labeling, depth and normal maps. This renderer will be used to create our priors for our holistic model. Note that the priors will be inaccurate due to the error in the geolocalization, camera pose as well as the map itself, *e.g.*, most trees are missing or misplaced. Furthermore it only contains static objects and thus will be inaccurate in places occupied by e.g., cars, pedestrians.

4. Holistic Scene Model

In this paper we are interested in utilizing geographic priors to help outdoor scene understanding. In particular, we focus on the tasks of 3D object detection, semantic segmentation as well as depth reconstruction from a single image. Towards this goal, we build 3D scene priors from freely available maps and frame the problem as one of inference in a holistic conditional random field (CRF) that reasons

jointly about all tasks and integrates semantics, geometry as well as geographic information.

4.1. Energy Definition

Given a single geo-localized image $\mathbf{x}$, we are interested in simultaneously assigning semantic labels to pixels, densely reconstructing the scene as well as detecting objects and localizing them in the 3D world. We parameterize the segmentation task with a random variable per pixel, $s_p \in \{1, ..., C\}$, encoding its semantic class. Dense depth reconstruction is parameterized with a continuous variable per pixel, $d_p \in [0, 80]$, encoding the distance in the 3D world (in meters). We parameterize each detection in 3D with four random variables, $y_i = \{x_i, z_i, \theta_i, b_i\}$, encoding the (x, z) position in the ground plane, the object pose θ_i as well a binary variable $b_i \in \{0, 1\}$ encoding whether the detection is a true positive.

Let $\mathbf{s} = (s_1, \cdots, s_N)$, $\mathbf{d} = (d_1, \cdots, d_N)$, $\mathbf{y} = (y_1, \cdots, y_M)$ be the set of all segmentation, depth estimation and detection variables, with N the size of the image and M the set of candidate detections. We define the energy of the CRF by integrating geographic context, appearance features and geometric properties:

$$\begin{aligned} E(\mathbf{y}, \mathbf{s}, \mathbf{d}) =& E_{\text{obj}}(\mathbf{y}) + E_{\text{seg}}(\mathbf{s}) + E_{\text{dep}}(\mathbf{d}) \\ &+ E_{\text{so}}(\mathbf{s}, \mathbf{y}) + E_{\text{do}}(\mathbf{d}, \mathbf{y}) + E_{\text{ds}}(\mathbf{d}, \mathbf{s}), \end{aligned} \tag{1}$$

where $E_{\text{obj}}, E_{\text{seg}}, E_{\text{dep}}$ are the energies that depend on a single task and $E_{\text{so}}, E_{\text{do}}, E_{\text{ds}}$ are the energies connecting different tasks. We now describe the potentials briefly. We refer the reader to the supplementary material for an in-depth explanation of all potentials.

3D Object Detection and Pose Estimation: The 3D object detection and pose estimation energy is defined as a sum of energies encoding appearance, geographic priors, occlusion and inter-penetration of objects in 3D:

$$\begin{aligned} E_{\text{obj}}(\mathbf{y}) = w_{\text{im}} \sum_i \phi_{\text{im}}(\mathbf{y}_i) + \mathbf{w}_{\text{geo}}^T \sum_i \phi_{\text{geo}}(\mathbf{y}_i) \\ + w_{\text{over}} \sum_{\beta \in \mathcal{E}} \phi_{\text{overlap}}(\mathbf{y}_\beta) + w_{\text{occ}} \sum_{\alpha \in \mathcal{H}} \phi_{\text{occlusion}}(\mathbf{b}_\alpha), \end{aligned} \tag{2}$$

where $\phi_{\text{im}}(\mathbf{y}_i)$ is defined as the detection score, $\phi_{\text{geo}}(\mathbf{y}_i)$ encodes geographic information in the form of the percentage of the object $\mathbf{y}$ that is visible when placed in the 3D scene and rendered back in the 2D image. Additionally, we use the distance (angular and positional) to the nearest road as well as the overlap between objects and buildings in bird's eye view. $\phi_{\text{overlap}}(\mathbf{y}_\beta)$ is a pairwise potential that penalizes object hypothesis that overlap in bird's eye view. Motivated by [9], we make use of CAD models to represent 3D cars. Towards this goal, we solve for the best transformation that aligns each CAD model with the 3D bounding box in terms of vertex error in 3D. This can be done via Procrustes analysis. We then select the CAD model that when projected to the image, best fits the 2D box. Finally, $\phi_{\text{occlusion}}(\mathbf{y}_\alpha)$ is a high-order potential that pushes object hypothesis that are occluded to be false positives, *i.e.* $b_i = 0$. This is simply done by re-rendering the CAD model in the image and calculating the occlusions generated by other cars that are in front. Fig. 2 depicts the high-order potentials as well as a real-world example for object reasoning.

Semantic Labeling: This energy is defined as the sum of segmentation unaries, geographic priors and smoothness:

$$\begin{aligned} E_{\text{seg}}(\mathbf{s}) =& w_{\text{in}} \sum_p \phi^s_{\text{im}}(s_p) + w^s_{\text{geo}} \sum_p \phi^s_{\text{geo}}(s_p) \\ &+ w^s_{\text{sm}} \sum_p \sum_{q \in \mathcal{N}_p} \phi_{\text{smooth}}(s_p, s_q), \end{aligned} \tag{3}$$

where ϕ^s_{im} is the score from a segmentation algorithm [34], and $\phi^s_{\text{geo}}(s_p) = 1$ if the segmentation agrees with our semantic rendering from OSM. The smoothness term $\phi_{\text{smooth}}(s_p, s_q) = \mu(s_p, s_q) \sum_m w_m k_m(\mathbf{f}_p, \mathbf{f}_q)$, with $k_m(\cdot, \cdot)$ an RBF kernel on the input features from two pixels and $\mu(\cdot, \cdot)$ a Potts function. We use a densely connected neighborhood in order to be robust to misalignment between image and geographic renderings [25].

Depth Estimation: This energy encodes geographic priors as well as smoothness:

$$E_{\text{dep}}(\mathbf{d}) = w^d_{\text{geo}} \sum_p \phi^d_{\text{geo}}(d_p) + w^d_{\text{sm}} \sum_p \sum_{q \in \mathcal{N}_p} \phi^d_{\text{smooth}}(d_p, d_q), \tag{4}$$

where the unary potential $\phi^d_{\text{geo}}(d_p)$ is the ℓ_2 distance between the hypothesis and the rendered depth in log scale. The pairwise potential is a weighted smoothness term between neighboring pixels, i.e., $\phi_{\text{smooth}}(d_p, d_q) = \mu(d_p, d_q) \sum_m w_m k_m(\mathbf{f}_p, \mathbf{f}_q)$, where $k_m(\cdot, \cdot)$ is an RBF kernel computed on input features from two pixels (coordinate and color), and $\mu(\cdot, \cdot)$ is the ℓ_2 norm.

Segmentation-Object: Given a 3D object estimation, as shown in Fig. 2, we employ [9] to match a set of CAD models with our 3D boxes. We then render those to create pixel-wise labelings of potentially dynamic objects. Then

$$E_{\text{so}}(\mathbf{s}, \mathbf{y}) = w_c \sum_{p \in \mathcal{P}(\mathbf{y}_i)} \delta(s_p = car, l(\mathbf{y}_i, p)), \tag{5}$$

with $l(y_i, p)) = 1$ if the rendered CAD model occupies pixel p, and $\delta(a, b) = 1$ if $a = b$ and 0 otherwise. Note that $\mathcal{P}(\mathbf{y}_i)$ defines the set of pixels where the CAD model reprojects.

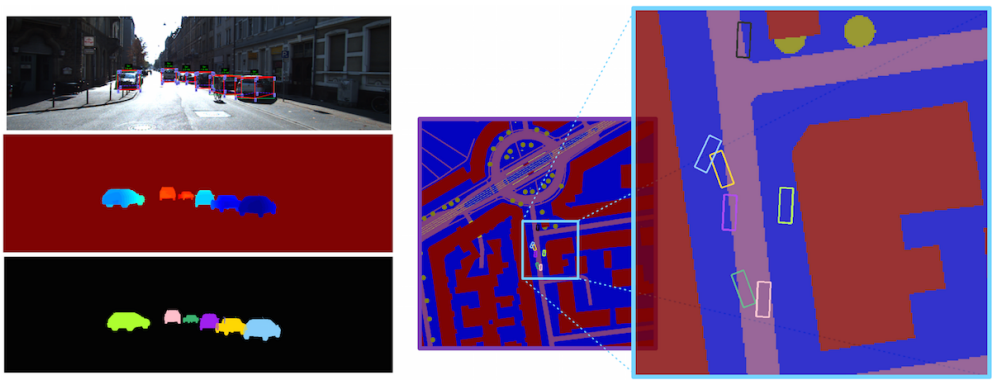

Figure 2. 3D Object localization and pose reasoning. We first obtain an initial 3D bounding box estimation from each 2D detection [8] (Left top). Based on the best matched CAD models, we generate car segmentation (Left bottom) and depth (Left middle). Given the camera pose and location, we can place the car bounding boxes in the 3D world [9] (Right top, bird-view). As we see in the right figure, the initial 3D boxes are noisy and some are even physically impossible (*e.g.* mutual overlaps, overlapping with buildings *etc.*). This is not the case with our approach, which also encourages cars to be aligned with the road direction.

Depth-Object: Similarly, we use the CAD model to generate an estimate of the depth for each pixel and define:

$$E_{\text{do}}(\mathbf{d}, \mathbf{y}) = w_{do} \sum_{p \in \mathcal{P}(\mathbf{y}_i)} \| \log(d_p) - \log(\hat{d}_p(\mathbf{y}_i)) \|_2^2, \quad (6)$$

where $\mathcal{P}(\mathbf{y}_i)$ defines the set of pixels where the CAD model reprojects.

Depth-Segmentation: We also employ a depth regressor which uses geometric and semantic features as follows:

$$E_{\text{ds}}(\mathbf{d}, \mathbf{s}) = w_{ds} \sum_{p} \| \log(d_p) - \log(\hat{d}(\mathbf{f}_p; s_p)) \|_2^2, \quad (7)$$

where $\hat{d}(\mathbf{f}_p; c) = w_c^T \mathbf{f}_p + b$ is a local linear regression model trained for different semantic labels, with $\mathbf{f}_p$ encoding the pixel coordinates, color, as well as gradient features and log-depth from our geographic context computed from rendering the 3D scene.

4.2. Inference

Inference in our model can be done by computing the minimum energy configuration

$$\min_{\mathbf{y},\mathbf{s},\mathbf{d}} E(\mathbf{y}, \mathbf{s}, \mathbf{d})$$

Note that this inference is NP-hard, and is particularly difficult as it contains a mixture of discrete and continuous variables. We perform approximate inference by running block coordinate descent. Thus we iteratively solve for each task, fixing the other ones, but taking into account the dependencies between the tasks. We refer the reader to Alg. 1 for a summary of our inference algorithm. We now describe how to do inference over each task.

Solving for 3D Detection: This involves minimizing

$$\min_{\mathbf{y}} E(\mathbf{y}, \mathbf{s}, \mathbf{d}) = \min_{\mathbf{y}} E_{\text{obj}}(\mathbf{y}) + E_{\text{so}}(\mathbf{s}^t, \mathbf{y}) + E_{\text{do}}(\mathbf{d}^t, \mathbf{y}),$$

with $\mathbf{s}^t, \mathbf{d}^t$ the current estimates of segmentation and depth. As we have a continuous-discrete MRF with relatively small number of nodes and edges, we proposed to use Gibbs sampling to do inference, where each object is updated by a Metropolis-Hastings sampler with a Gaussian proposal distribution centered at the current location and pose.

Solving for segmentation: This involves minimizing

$$\min_{\mathbf{s}} E(\mathbf{y}, \mathbf{s}, \mathbf{d}) = \min_{\mathbf{s}} E_{\text{seg}}(\mathbf{s}) + E_{\text{so}}(\mathbf{s}, \mathbf{y}^t) + E_{\text{ds}}(\mathbf{d}^t, \mathbf{s}),$$

with $\mathbf{y}^t, \mathbf{d}^t$ the current solutions for the 3D detection, localization and depth estimation tasks. This reduces to the problem of inference in a fully connected MRF with Gaussian potentials. We exploit the efficient mean-field inference algorithm of [25] to solve this problem, which iteratively approximates the original distribution with a product of independent marginal distributions by minimizing the KL-divergence between the two distributions.

Solving for depth: This involves minimizing

$$\min_{\mathbf{d}} E(\mathbf{y}, \mathbf{s}, \mathbf{d}) = \min_{\mathbf{d}} E_{\text{dep}}(\mathbf{d}) + E_{\text{ds}}(\mathbf{d}, \mathbf{s}^t) + E_{\text{do}}(\mathbf{d}, \mathbf{y}^t),$$

with fixed 3D detection and semantic segmentation $\mathbf{y}^t, \mathbf{s}^t$. This is a continuous-valued Gaussian MRF, which can be efficiently solved by Gaussian belief propagation [43].

5. Experiments

In this section, we evaluate our approach on the challenging KITTI dataset [13]. We tested our performance

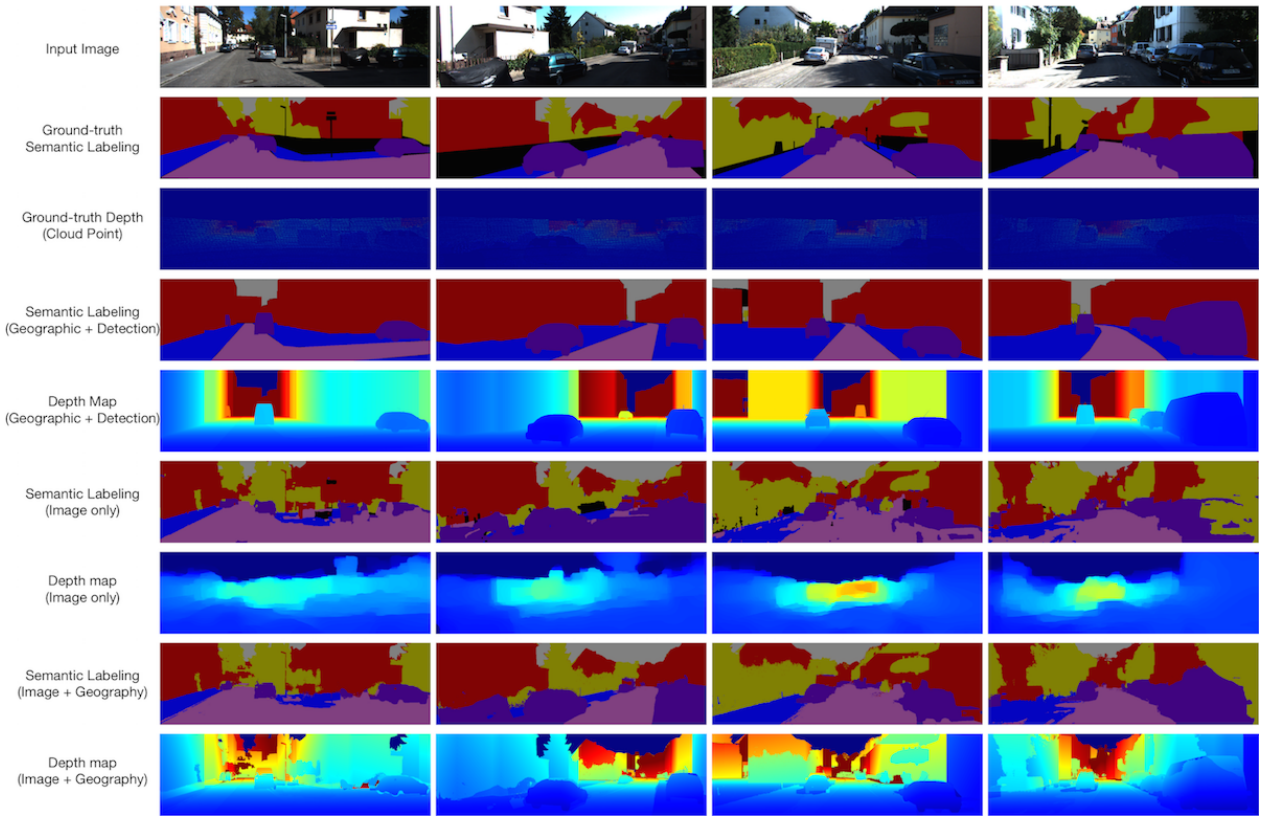

Figure 3. Overall performance for holistic tasks.

Algorithm 1 Holistic Inference via Block Coordinate Descent

Input: Image $\mathbf{x}$, geolocation and camera pose.
Get initial detections $\mathbf{y}^0$ from image [8].
Render the 3D scene from OpenStreetMaps.
Get initial $\mathbf{s}^0$ and $\mathbf{d}^0$ from 3D scene.
repeat
Solve $\mathbf{y}^{t+1} = \arg\min_{\mathbf{y}} E(\mathbf{y}, \mathbf{s}^t, \mathbf{d}^t)$
Render object depth and labeling using $\mathbf{y}^t$
Solve $\mathbf{s}^{t+1} = \arg\min_{\mathbf{s}} E(\mathbf{y}^{t+1}, \mathbf{s}, \mathbf{d}^t)$
Solve $\mathbf{d}^{t+1} = \arg\min_{\mathbf{d}} E(\mathbf{y}^{t+1}, \mathbf{s}^{t+1}, \mathbf{d})$
until convergence or reach max iteration
Output: $\mathbf{y}, \mathbf{s}, \mathbf{d}$

quantitively on two subsets, according to the availability of the ground-truth data. For depth reconstruction, 3D object detection and pose estimation we use the KITTI tracking sequences, as they are very challenging and contain many moving objects. For semantic segmentation, we use the KITTI visual odometry dataset where annotations by [35] are available. We train our model parameters by cross-validation.

5.1. Depth Reconstruction

In KITTI, the ground-truth depth is captured by a Velodyne LIDAR. To create a per-pixel depth estimation, we first project for each image the 3D point cloud onto the camera plane, with the provided calibration matrix. This results in a sparse ground truth image. Note that due to moving objects we cannot aggregate multiple Velodyne point clouds in contrast to the stereo benchmark where the scene is static and 7 frames are aggregated to produce denser depth. Nonetheless, we get 18,000 pixels labeled on average, which we employ to compute our error metric. When the projection of multiple pixels overlap, we pick the one with closest capture time. Although the ground truth is not perfect, it is sufficiently precise given the capabilities to reconstruct depth from a single image. We split the 21 sequences into training (0-9, 3049 images), validation (19-20, 1894 images) and testing (10-18, 2724 images). We employ the relative depth ratio $\delta = \max(\frac{d_{\text{gt}}}{d_{\text{out}}}, \frac{d_{\text{out}}}{d_{\text{gt}}})$ as our accuracy measure. We report the ratio of pixel with correctly estimated depth, depending on whether the relative depth ratio is smaller than three thresholds (1.25, 1.25^2 and 1.25^3). As baseline we train a support vector regressor (SVR) for each semantic label [7]. We use local features (color, pixel co-

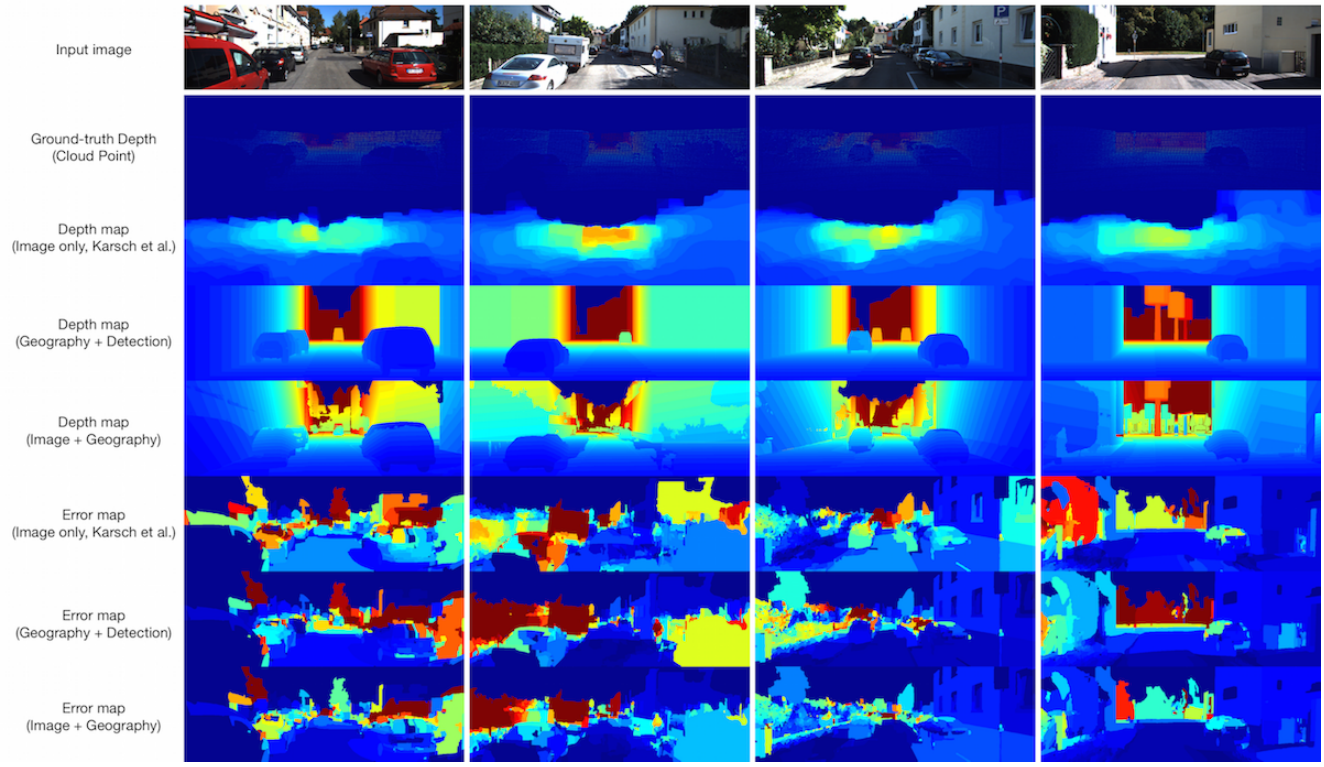

Figure 4. Depth reconstruction from rendered 3D map. From top to bottom: input image, ground-truth depth, image based depth estimation [24], geographic based depth map, image+geographic depth estimation, error map for image only depth, error map for geographic only depth and error map for image+geographic depth.

ordinates, histogram of gradients, local binary descriptors) as input to the regressor. We also runs the SIFTFlow based depth transfer algorithm, which is the current state-of-the-art single image depth estimation method [24].

As shown in Tab. 1 our geographic prior is fairly accurate but generates mistakes due to incorrect building heights or missing vegetations in OSM, which occur very frequently. In contrast, our holistic approach leverages semantic segmentation and results in a 10% improvement over the prior. Moreover, our experiment results also suggest that purely image-based depth reconstruction cannot generate reliable results, due to the depth ambiguity that is present when using local image features. Additionally, we computed the oracle performance that our approach could ever achieve. This is not 100% as the UCM superpixels we employed (with boundary threshold 0.04) are not perfect. For each superpixel, an average depth is estimated based on the ground truth. The performance is around $78.16\%, 93.87\%, 98.71\%$ for the three measurements respectively. Our method is fairly close to the oracle performance.

We refer the reader to Fig. 4 for a qualitative illustration of some examples. As shown in this figure, our holistic model can generate good dense depth estimates. Furthermore, semantic information from the image can help produce more accurate results with fine details and compensate errors brought by inaccurately rendered 3D scenes. We also show a failure case in the leftmost column, where depth is wrongly estimated due to missing trees in the geographic 3D scene and little geometric information from the image.

Method	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Image only	26.21%	52.34%	78.52%
Karsch *et al.* [24]	53.07%	83.91%	**93.57%**
Geographic only	61.25%	76.13%	85.49%
Proposed	**69.44%**	**85.51%**	92.59%

Table 1. Depth reconstruction performance

5.2. Semantic Labeling

We train on 100 training images from the visual odometry benchmark, and split the testing images into two parts, 23 images for validation and 23 for test. As baseline we employ the state-of-the-art hierarchical semantic labeling algorithms of [34] and [41]. We utilize six semantic classes for evaluation, namely sky, building, road, sidewalk, vegetation and car, since classes like pole and pedestrians contain

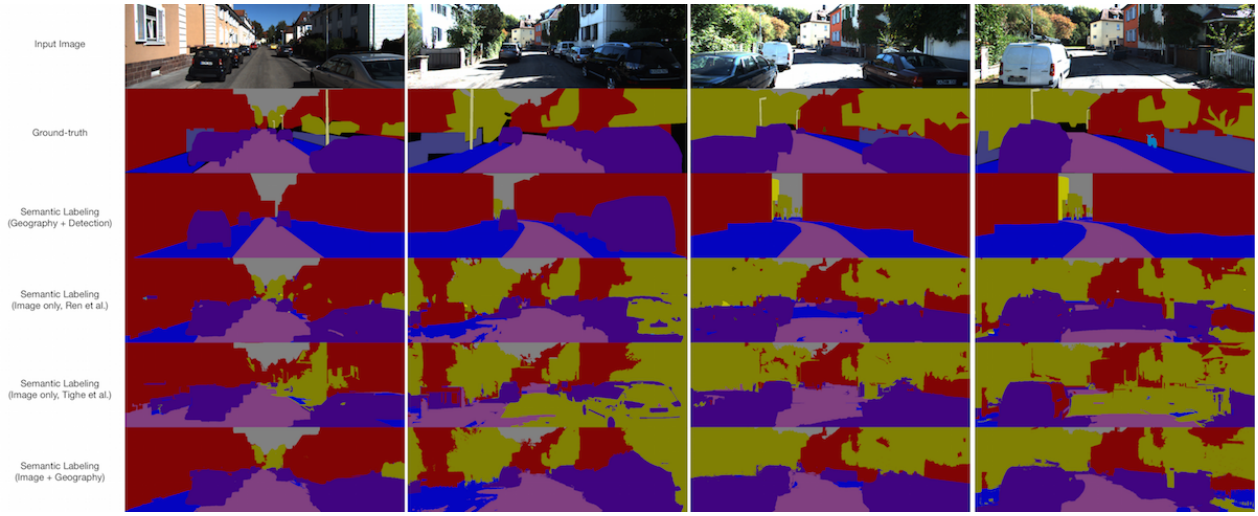

Figure 5. Semantic labeling from rendered 3D map and 3D car detection. From top to bottom: input image, ground-truth labeling, semantic labeling from geographic information, image based semantic labeling of [34], image based semantic labeling of [41], and our proposed semantic labeling using geographic and image information. Color code: sky, building, road, sidewalk, fence, vegetation, car.

only 2 examples. We use PASCAL intersection-over-union (IOU) as our measure of performance.

As shown in Tab. 2, our algorithm outperforms [34] by 3%. Importantly, it improves all categories. The weak performance of the purely geographic unary is mainly due to the misalignment and inaccuracy of OSM, which suggests an important future direction to improve the geographic 3D world from the image information. Despite this low performance, our holistic model improves in all classes. As shown in Fig. 5 and Fig. 3 our geographic prior can provide additional information for semantic reasoning, *e.g.* correcting wrongly labeled road pixels due to shadows. On the other hand, with the help of 3D detection our algorithm can also generate instance segmentations that have been missed by the image-based segmentation algorithm. Moreover, the use of densely connected pairwise potentials reduces the noise while preserving the boundaries. Additionally, given our estimated depth and semantic labels, we visualize the 3D point cloud in Fig. 7, by transforming the points from the image to the world coordinates and coloring with RGB intensity as well as estimated semantic labels.

5.3. 3D Object Detection

We evaluate our 3D object detection on the tracking dataset, following the same train/validation/test splits that we use for the depth reconstruction experiment. As baseline we use the method of [9], which fits a CAD model in 3D to 2D bounding boxes. Note that those boxes are also the input to our algorithm. In particular, they are the output of the deformable part-based model [8]. We evaluate both mean average precision (mAP) as well as F1-measure for 2D object detection. We compute localization results for all true-positive DPM detections with heights larger than 30 pixels. Localization is evaluated as the percentage of detections that have an error less than a fix distance (1m and 2m respectively). We also report the median localization error. Pose estimation is evaluated as the median pose error in degrees. As shown in Tab. 3, our holistic approach significantly improves in all metrics over the baseline. Specially, as shown in Fig. 6, pose estimation errors can be significantly reduced in our model.

We conduct an additional pilot experiment on a small subset of the dataset, which contains 419 images. We asked the authors of [46] to run their method on this subset. [46] utilizes poselets [3] as the detector and can only return results on cars of size more than 50 pixels, as their approach requires sufficient image evidence to reason about parts. However, our approach can tackle the more difficult setting of dealing with small cars of half the size. Nevertheless, we report results on the subset where their approach works. On pose estimation, our method (4.09° median error) achieves 2° degree improvement over both Zia *et al.* [46] (6.28° median error) and Fidler *et al.* [9] (6.16° median error), and comparable localization performance (1.31 meters) vs (1.25 meters for Zia *et al.* and 1.33 meters for Fidler *et al.*) in terms of median error. Under this easy configuration (large bounding boxes) the initial localization error is very small, our localization potentials do not have major impact since the hypothesis do not overlap with buildings or other objects.

6. Conclusions

In this paper we have proposed a holistic approach that reasons jointly about 3D object detection, pose estimation, semantic segmentation as well as depth reconstruction from

Method	Overall	sky	building	road	sidewalk	vegetation	car
Ren *et al.* [34]	71.93%	87.35%	78.67%	72.58%	41.28%	80.93%	59.47%
Tighe *et al.* [41]	60.67%	81.41%	72.17%	52.16%	17.33%	69.91%	52.30%
Geographic only	44.21%	32.41%	59.25%	63.01%	36.41%	7.36%	35.66%
Proposed	**74.78%**	**88.62%**	**80.12%**	**80.89%**	**43.64%**	**81.59%**	**63.50%**

Table 2. Semantic labeling performance: intersection-over-union

Method	Config	mAP	F1	< 1m	< 2m	median loc	$< 10°$	median pos
Fidler *et al.* [9]	[8], height > 30	58.1%	0.67	19.96%	37.32%	2.92m	71.97%	4.85°
Proposed	[8], height > 30	**60.6%**	**0.69**	**22.96%**	**40.73%**	**2.67m**	**77.32%**	**3.85°**

Table 3. 3D object detection and pose estimation performance

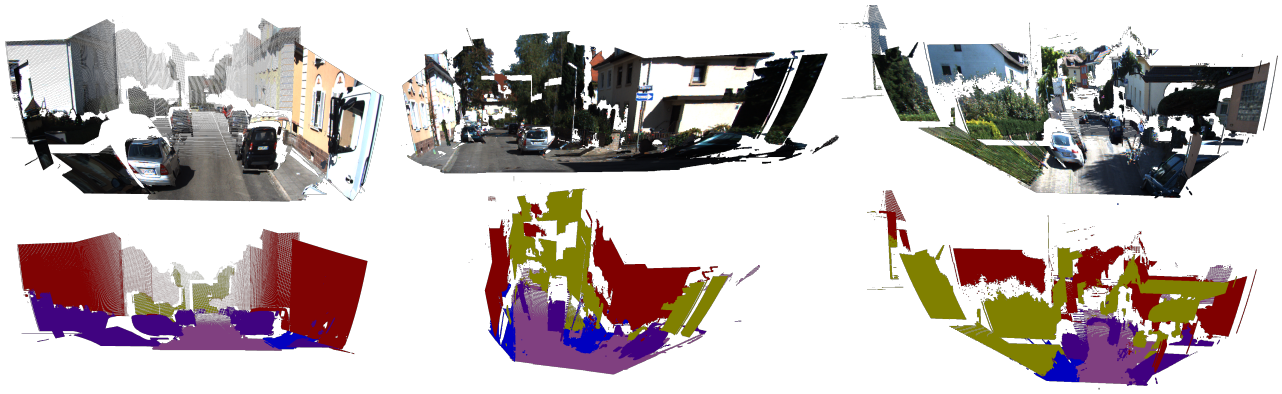

Figure 7. 3D point cloud visualization of depth reconstruction results. Top row: coloring with input image; bottom row: coloring with our semantic labeling result.

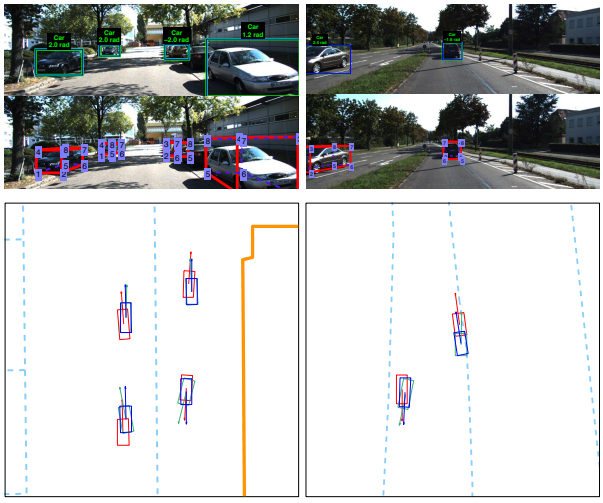

Figure 6. Bird's eye view visualization of our localization and pose estimation (GT, Fidler *et al.* [9], Ours)

a single image. Our approach is able to take advantage of large-scale crowd-sourced maps to generate dense geographic, geometric and semantic priors by rendering the 3D world. We have demonstrated the effectiveness of our holistic model on the challenging KITTI dataset [13], and showed significant improvements over the baselines in all metrics and tasks. While we employed OpenStreetMaps, we would like to emphasize that GoogleEarth and other similar resources could be used as well.

7. Acknowledgements

This work was partially supported by ONR-N00014-14-1-0232.

References

[1] Osm2world. http://osm2world.org. 2

[2] A. A. M. Loh. *The recovery of 3-D structure using visual texture patterns*. PhD thesis, University of Western Australia, 2006. 1

[3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009. 7

[4] M. A. Brubaker, A. Geiger, and R. Urtasun. Lost! leveraging the crowd for probabilistic visual self-localization. In *CVPR*, 2013. 2

[5] K. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *CVPR*, 2003. 1

[6] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, 2009. 1

[7] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. *NIPS*, 1997. 5

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2010. 1, 4, 5, 7, 8

[9] S. Fidler, S. Dickinson, and R. Urtasun. 3d object detection and viewpoint estimation with a deformable 3d cuboid model. In *NIPS*, 2012. 1, 2, 3, 4, 7, 8

[10] S. Fidler, R. Mottaghi, A. Yuille, and R. Urtasun. Bottom-up segmentation for top-down detection. In *CVPR*, 2013. 2

[11] G. Floros, B. van der Zander, and B. Leibe. Openstreetslam: Global vehicle localization using openstreetmaps. In *ICRA*, 2013. 2

[12] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3d traffic scene understanding from movable platforms. *PAMI*, 2013. 2

[13] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1, 4, 8

[14] C. Gu and X. Ren. Discriminative mixture-of-templates for viewpoint classification. In *ECCV*, pages 408–421, 2010. 1

[15] A. Guzmán. Decomposition of a visual scene into three-dimensional bodies. In *Fall Joint computer conference*. ACM, 1968. 1

[16] C. Hane, N. Savinov, and M. Pollefeys. Class specific 3d object shape priors using surface normals. In *CVPR*, 2014. 1

[17] J. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. In *CVPR*, 2008. 2

[18] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. In *ICCV*, 2009. 1, 2

[19] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *ECCV*, 2010. 1

[20] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. In *TOG*, 2005. 1

[21] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, 2005. 1, 2

[22] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *IJCV*, 2008. 2

[23] H. Jin and P. Favaro. A variational approach to shape from defocus. In *ECCV*, 2002. 1

[24] K. Karsch, C. Liu, and S. B. Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *PAMI*, 2014. 6

[25] P. Krähenbühl and V. Koltun. Parameter learning and convergent inference for dense random fields. In *ICML*, 2013. 3, 4

[26] L. Ladickỳ, C. Russell, P. Kohli, and P. H. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*. 2010. 1, 2

[27] L. Ladickỳ, J. Shi, and M. Pollefeys. Pulling things out of perspective. In *CVPR*, 2014. 1, 2

[28] L. Ladickỳ, P. Sturgess, K. Alahari, C. Russell, and P. H. Torr. What, where and how many? combining object detectors and crfs. In *ECCV*. 2010. 1, 2

[29] D. C. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. In *CVPR*, 2009. 1

[30] J. Liebelt and C. Schmid. Multi-view object class detection with a 3d geometric model. In *CVPR*, 2010. 1

[31] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *CVPR*, 2010. 2

[32] K. Matzen and N. Snavely. Nyc3dcars: A dataset of 3d vehicles in geographic context. In *ICCV*, 2013. 1, 2

[33] E. Prados and O. Faugeras. Shape from shading. Springer, 2006. 1

[34] X. Ren, L. Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In *CVPR*, 2012. 3, 6, 7, 8

[35] G. Ros, S. Ramos, M. Granados, A. Bakhtiary, D. Vazquez, and A. Lopez. Vision-based offline-online perception paradigm for autonomous driving. In *WACV*, 2015. 5

[36] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *PAMI*, 2009. 1

[37] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. In *CVPR*, 2000. 1

[38] A. Schwing, T. Hazan, and R. Urtasun. Efficient structured prediction for 3d indoor scene understanding. In *ECCV*, 2012. 1

[39] A. G. Schwing, S. Fidler, M. Pollefeys, and R. Urtasun. Box in the box: Joint 3d layout and object reasoning from single images. In *ICCV*, 2013. 1

[40] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*. 2012. 2

[41] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013. 6, 7, 8

[42] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing visual features for multiclass and multiview object detection. *PAMI*, 2007. 1

[43] Y. Weiss and W. T. Freeman. Correctness of belief propagation in gaussian graphical models of arbitrary topology. *Neural computation*, 2001. 4

[44] Y. Xiang and S. Savarese. Estimating the aspect layout of object categories. In *CVPR*, 2012. 1

[45] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012. 2

[46] M. Z. Zia, M. Stark, K. Schindler, and R. Vision. Are cars just 3d boxes?–jointly estimating the 3d shape of multiple objects. In *CVPR*, 2014. 1, 2, 7