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

Shenlong Wang, Sanja Fidler, Raquel Urtasun
Department of Computer Science, University of Toronto.

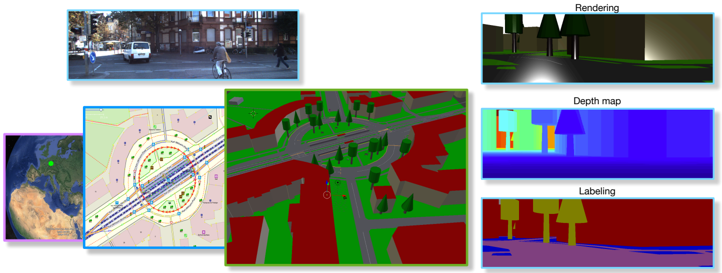

Figure 1: Maps can bring rich 3D information.

Inferring 3D semantic and geometric information from a single monocular image has been one of the holy grails of computer vision since the beginning. 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 [3], and show significant improvements over the baselines in all metrics and tasks.

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.

We make use of OpenStreetMaps [1], 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.

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, y) 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:

$$E(\mathbf{y}, \mathbf{s}, \mathbf{d}) = E_{\text{obj}}(\mathbf{y}) + E_{\text{seg}}(\mathbf{s}) + E_{\text{dep}}(\mathbf{d}) + E_{\text{so}}(\mathbf{s}, \mathbf{y}) + E_{\text{do}}(\mathbf{d}, \mathbf{y}) + E_{\text{ds}}(\mathbf{d}, \mathbf{s}) \quad (1)$$

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 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 the full paper and supplementary material for an in-depth explanation of all potentials.

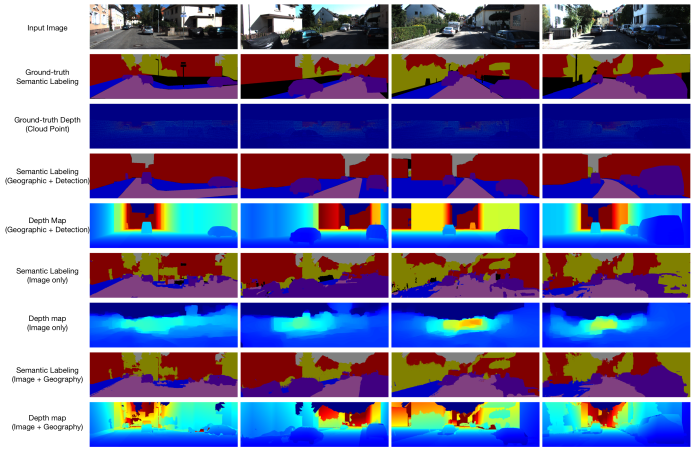

Figure 2: Overall performance for holistic tasks.

Depth	Karsch et al. [4]	Geography	Ours
$\delta < 1.25$	53.07%	61.25%	**69.44%**

Semantic labeling	Ren *et al.* [5]	Tighe *et al.* [6]	Ours
IoU	71.93%	60.67%	**74.78%**

Table 1: Monocular depth estimation and semantic labeling performance

In our experiments, we evaluate our approach on the challenging KITTI dataset [3] over three different tasks. We tested our performance quantitively on two subsets, according to the availability of the ground-truth data. Fig. 2 depicts the overall performance of our proposed method and several competing algorithms in semantic labeling and depth reconstruction. We also measure the performance quantititively and our approach overperforms all the competing algorithms [4, 5, 6].

[1] Openstreetmap. https://www.openstreetmap.org/.

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

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[4] Kevin Karsch, Ce Liu, and Sing Bing Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *PAMI*, 2014.

[5] Xiaofeng Ren, Liefeng Bo, and Dieter Fox. Rgb-(d) scene labeling: Features and algorithms. In *CVPR*, 2012.

[6] Joseph Tighe and Svetlana Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.