Single RGB Image Depth Estimation in Indoor and Outdoor Scenes

Yuanfang Wang (yolandaw), Yinghao Xu (ericx), Yuan Gao (julianyg)

I. Introduction

Depth estimation is a useful technique for multiple applications such as obstacle detection and scene reconstruction. With the research focus on Convolutional Neural Networks (CNN), depth estimation has seen rapid development recently. We propose a CNN based framework to incorporate LiDAR data for depth estimation. With , we want to use CNN to train an image-LiDAR model, that takes in an RGB image and outputs depth map obtained from its LiDAR map. We can perform depth estimation as well as prediction certainty evaluation on the outputs. By performing test-time dropout, we convert the depth prediction into a probabilistic distribution, which is useful to understand the performance of the network, as well as assisting the deployment of prediction outcomes.

II. Framework

A. General pipeline

1) Ground truth: The ground truth data generation pipeline is composed by three concatenated phases. Our goal is to generate images with true RGB-depth labels. The image data and depth data are extracted from the KITTI data set. The 2D RGB images are then projected with LiDAR depth data, which are calibrated and stored separately. Due to the sparsity and incompleteness of projected LiDAR depth data points, we perform interpolation to generate full depth maps, and map them onto the 2D images.

2) Neural Networks: The goal of CNN training is to generate depth estimations based on input 2D RGB images. The original framework has a retrain module to train the model on RGB data and depth labels to fit the model better on LiDAR depth data, and a modified prediction module to generate depth map based on input 2D RGB images.

B. Working Progress

We have completed the implementation of data pre-processing pipeline. The RGB image data are combined with calibrated LiDAR depth maps, and interpolated afterwards. We have also run demo scripts on model testing, and the results on naive models (not retrained by depth labels) are not acceptable, which is expected. Currently we have two choices of CNN models for training: the VGG-based network from ICCV 2015[1], and the global coarse-scale to local fine-scale network from NIPS 2014[2]. As mentioned in the working progress section, both CNN models require retraining on RGB-depth data; however, the author has not provided training scripts, and it is not realistic to finish writing and debugging our own training scripts within two weeks. Thus we are considering another possibility: to evaluate certainty of predicted depth maps. We are planning to append dropout layers to the trained neural network, and run depth prediction. The outcomes will form a random distribution, and by analyzing the features of the distribution, we can calculate the confidence of our depth prediction.

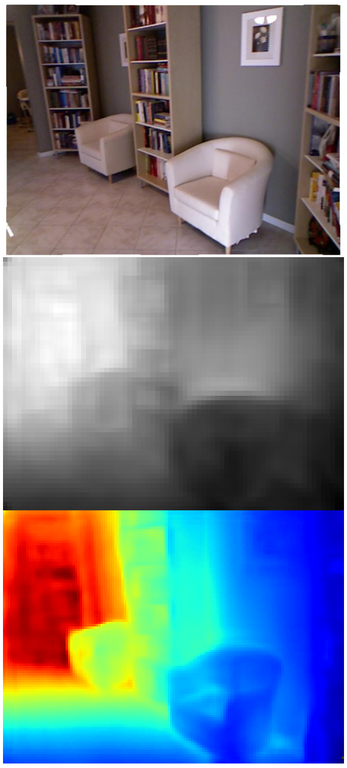

Fig. 1. The top, middle and bottom images are input RGB, depth prediction from NIPS 2014 network and from ICCV 2015 network respectively in the indoor NYU Depth v2 scene

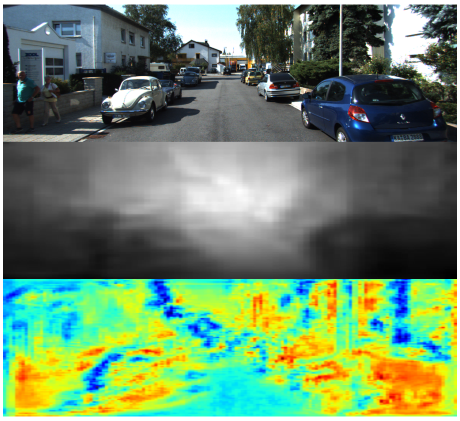

Fig. 2. The top, middle and bottom images are input RGB, depth prediction from NIPS 2014 network and from ICCV 2015 network respectively in the outdoor KITTI scene

III. Data Set

A. Ground truth

We want to test our algorithm on NYU Depth v2[3] and KITTI[4] data sets. Both the data sets contain continuous image sequences with depth information in different scenes. The NYU-Depth V2 data set is comprised of video sequences from 464 indoor scenes as recorded by both the RGB and Depth cameras from the Microsoft Kinect. We can feed our neural network with single RGB image from NYU Depth v2 to get the depth prediction and prediction certainty distribution for each pixel and compare with the depth ground truth observed by Kinect to analyze the performance of our result.

But in the outdoor scene, because the large range of environment depth, from zero to one hundred meters or more (while for the indoor environments, the depth is normally within 5 meters), and the strong sunlight, Kinect sensor fails to observe the ground truth of depth information. So the KITTI data set used LiDAR sensor instead. This data set contains several outdoor scenes captured by data collection car mounted with forward camera system and 360 degree 64 lines LiDAR sensor Velodyne. The depth information captured by Velodyne are sparse and non-equally distributed 3D point clouds, different from dense and equally distributed 2D depth "images" got from indoor sensor Kinect.

In order to evaluate our performances in both indoor and outdoor scene with ground truth captured by different sensing devices, we projected the LiDAR data from KITTI data set onto corresponding 2D images and then interpolated the sparse points to generate a 2D dense and equally distributed depth "image" like what we can directly get from the Kinect sensor.

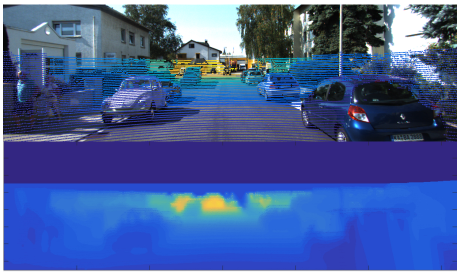

Fig. 3. The top and bottom images are the projected sparse points and the interpolated "depth image" respectively of the KITTI data set

B. KITTI data pre-processing

For each frame of the synchronized LiDAR point cloud and image from KITTI data set. We first project the point cloud onto 2D image using the calibration parameters and then interpolate the sparse 2D points with the value of depth.

1) From 3D point cloud to 2D: In the camera model of pinhole projection, the position relationship between an 3D point and its correspondent 2D image point has the following steps.
First, transform the 3D point from world coordinate $[X, Y, Z, 1]^T$ to camera coordinate $[X_c, Y_c, Z_c, 1]^T$ by

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} \sim \begin{bmatrix} \mathbf{R} & \mathbf{T} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

where $\mathbf{R}$ and $\mathbf{T}$ are the rotation and translation matrix from the LiDAR sensor to the camera sensor of correspondent frame from the dataset. Second, transform the point $[X_c, Y_c, Z_c, 1]^T$ in camera coordinate to 2D image plane $[x, y, 1]^T$ by

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix}$$

The third step is from the camera image plane to the pixel coordinate $[u, v, 1]^T$,

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \sim \mathbf{K} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

where $\mathbf{K}$ is the camera calibration matrix provided by the KITTI data set.

2) Depth interpolation: The KITTI training data set contains the RGB image frames of various street scenes and corresponding depth data points generated by a LiDAR range sensor. The transformation of the LiDAR point-cloud in to the image plane results in a sparse distribution of depth points. Besides, the depth points are sometimes incomplete due to the complexity in the scenes.

Therefore, in order to train the network effectively, we need to generate a dense depth map from sparse, incomplete depth data. The method we employ is interpolation. We first fill the empty pixels in the depth image with 0's. Next, a 10-by-10 window centered around each zero pixel is used to calculate the weighted average of non-zero pixels within the window, the result of which is then assigned to the zero pixels. The weights of non-zero pixels are determined by the relative distance from the center pixel. In this way, the resulting dense depth images not only retain the true depth information from LIDAR point cloud, but in the mean time also have high resolution that can be exploited in the training framework.

IV. Experiment Plans

We want to evaluate our result of depth prediction and prediction certainty distribution of singe RGB image on both the indoor NYU Depth v2 and the outdoor KITTI data set. As mentioned previously, the network from ICCV 2015 only trained on indoor scene. It has poor depth prediction in the outdoor environment. As a result, we chose to use the network from NIPS 2014 to generate the depth estimation and prediction certainty distribution of both indoor and outdoor scene.

Another issue of depth prediction is that the scale retrieval of estimated depth. For the indoor scene, the depth range is from zero to ten meters. In the outdoor scene, the depth range is generally from zero to sixty meters. To maximize our prediction in different scene with different range and accuracy, in the training process, we normalized the range to 0 to 255. As a result, in the evaluation process, we need first retrieve the scale between our estimation and the ground truth by minimizing the L2 error, then recalculate the error using the retrieved scale.

It is expected that when the error of our depth prediction is large, the correspondent prediction certainty distribution will have large variance.

References

[1] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," *arXiv:1411.4734*, 2015.

[2] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," 2014.

[3] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[4] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.