In-depth: Depth Map Estimation From Monocular RGB Image

Xu Guo, Isha Singhal, Meijiao Png {xuguo, isha22, mpng}@stanford.edu)
Department of Computer Science, Stanford University

Project Video: https://youtu.be/Dd7YxdT1jCs

Overview

RGB-D images augment conventional RGB images with additional depth information on a per-pixel basis. This additional information can be used in various applications that include 3D reconstruction, AR/VR and image processing.

While modern consumer technology such as smartphones have enabled more people to take RGB photos, it is still difficult to obtain RGB-D images. There has been numerous efforts in industry to integrate specialized sensors into hardware to capture depth information (Google Project Tango, lenovo Phab2 pro, intel realsense). However, the efforts have not been successful because depth sensing capabilities require extra hardware, accurate calibration and extra design space. It is usually hard to justify the large BOM cost (bill of material cost), production line change, and drastic industrial design change of the phone to incorporate depth sensors.

In this paper, we evaluate deep learning approaches to construct "application ready" depth image using a single RGB camera image, which leapfrogs the need of specialized depth sensor.

Features

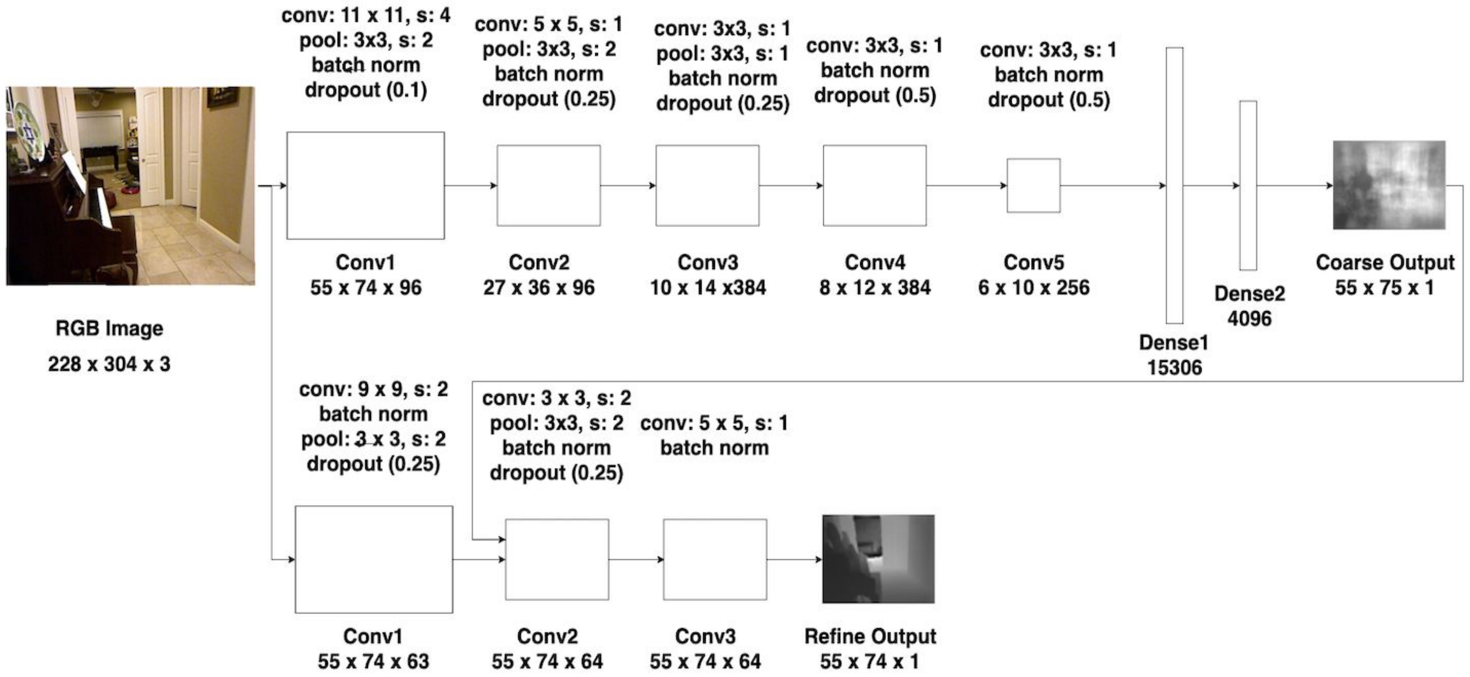
Input: RGB image with dimension: (228, 304, 3)
Output:
- Depth image with dimension (55, 74, 1)

Data Processing

Our dataset comes from NYU Depth dataset which contains 1449 pairs of RGBD images of indoor scenes recorded by Microsoft Kinect. Although the dataset contains information including segmentation, object label, etc, our training only takes two sections of the data: depth and image.

Color image data consist of 3 RGB channels with uint8 integers, depth image consists of a single channel float point data representing real world distance measured in meters. In the preprocessing step, the RGB channels data are normalized to [0.0, 1.0], while the depth value are kept as is. The normalization of depth value did not improve either the estimation accuracy or the learning velocity, thus the depth value are unprocessed to simplify the metric computation.

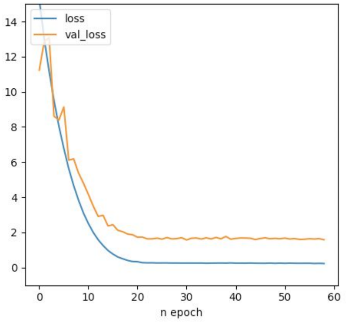
Architecture



Loss Curve

$$L(y, \hat{y}) = 1/n \sum d_i^2 + \lambda/n^2 (\sum d_i)^2, \lambda \in [0, 1]$$

$$d = log(y^i) - log(\hat{y}^i)$$



Metrics

Absolute relative difference: $\frac{1}{T} \sum_1^T |y - \hat{y}|/y$

Squared relative difference: $\frac{1}{T} \sum_1^T |y - \hat{y}|^2/y$

Root Mean Squared Error (RMSE): $\sqrt{\frac{1}{T} \sum_1^T |y - \hat{y}|^2}$

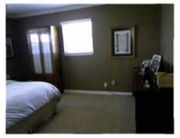
	Baseline	Augmentation	Dropout	Regularization	Dropout + Augmentation
Abs Rel Diff	0.294	**0.286**	0.272	0.318	0.309
Sqrt Rel Diff	0.391	**0.356**	0.349	0.471	0.375
RMSE	1.161	**1.076**	1.113	1.296	1.056

Table 1: Metric of our approach

	Make3D	Eigen	Karsch
Abs Rel Diff	0.408	0.215	0.350
Sqrt Rel Diff	0.581	0.212	0.223
RMSE	1.24	0.907	1.2

Table 2: Metric of other approaches

Results

Augmentation Model	Input	Ground truth	Predict
From training set			
From dev set			

References

[1] David Eigen, Christian Puhrsch, Rob Fergus Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. Curran Associates, Inc., NIPS2014-5539, 2014.
[2] https://github.com/jguoxu/cnn_depth_tensorflow
[3] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In NIPS, 2005.
[4] Daniel Stanley Tan, Chih-Yuan Yao, Conrado Ruiz, Jr., and Kai-Lung Hua Single-Image Depth Inference Using Generative Adversarial Networks. 2019 Apr 10. doi: 10.3390/s19071708
[5]https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html
[6] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, Il Hong Suh From Big to Small: Multi-Scale Local Planar Guidance for Monocular Depth Estimation
[7] K. Karsch, C. Liu, S. B. Kang, and N. England. Depth extraction from video using nonparametric sampling. In TPAMI, 2014.