Real-time Monocular Depth Estimation with Extremely Light-weight Neural Network

Paper ID: 1954

Mian-Jhong Chiu
Institute of Multimedia Engineering
National Chiao Tung University
Hsinchu County, Taiwan
0210028@gmail.com

Wei-Chen Chiu
Department of Computer Science
National Chiao Tung University
Hsinchu County, Taiwan
walon@cs.nctu.edu.tw

Hua-Tsung Chen
Department of Computer Science
National Chiao Tung University
Hsinchu County, Taiwan
huatsung@cs.nctu.edu.tw

Jen-Hui Chuang
Department of Computer Science
National Chiao Tung University
Hsinchu County, Taiwan
jchuang@cs.nctu.edu.tw

Abstract

Using a single image to perform depth estimation has become one of the main focuses in resent research works. However, prior works usually rely on highly complicated computation and power-consuming GPU to achieve such task; therefore, we focus on developing a real-time light-weight system for depth prediction in this paper. We propose a supervised learning based CNN with detachable decoders that produce depth predictions with different scales. We also formulate a novel log depth loss function that computes the depth difference in log space. To train our model efficiently, we generate depth map and semantic segmentation with complex teacher models. Via a series of ablation studies and experiments, it is validated that our model can efficiently performs real-time depth prediction with only 0.32M parameters, with the best trained model outperforms previous works on KITTI dataset for various evaluation matrices.

Approach

Data Pre-processing

During the training process, a multi-task learning method is conducted, wherein the multi-task loss function is evaluated to update the model parameters. Therefore, two types of data are needed for the training process: (i) semantic segmentation and (ii) depth map as ground truths.

For (i), we generate such data using a teacher model - DeepLabV3 [19], a CNN model which generates the semantic segmentation with pixel-wise label. As for (ii), we employ Pyramid Stereo Matching Network (PSMNet) [20] as a teacher model. However, the above PSMNet depth map is generated via stereo matching, which is certainly not the measured ground truth; therefore, we subtract PSMNet depth map from the sparse depth map to compute mean error per distance for the training set. To compensate such error, a look-up table is built and used to adjust the depth value of each pixel of the PSMNet depth map to a more accurate depth value.

Light-weight Neural Network Design

Figure 1 shows the proposed neural network architecture design, wherein each layer is denoted with its layer name, number of output channels, and strides, while the multiplication on top of a layer gives the repetition of that layer. The light blue region represents the encoder, which down-samples input RGB image and eventually outputs a set of low-resolution and high-expressive feature maps. The decoder is represented by four light orange regions D1~D4, each composed of a PixelShuffle layer and 2~4 bottlenecks, which eventually generates depth map and semantic segmentation map with multiple resolutions. For each decoder block, a feed-forward feature map from previous layer is concatenated with a feature map passed from some encoder layers via skip connections and fed into a pixel shuffle layer for up-sampling. After that, the feature maps are fed into 2~4 bottlenecks to increase the feature expressiveness before further passed to two separate output layers, one for depth prediction and the other for semantic segmentation.

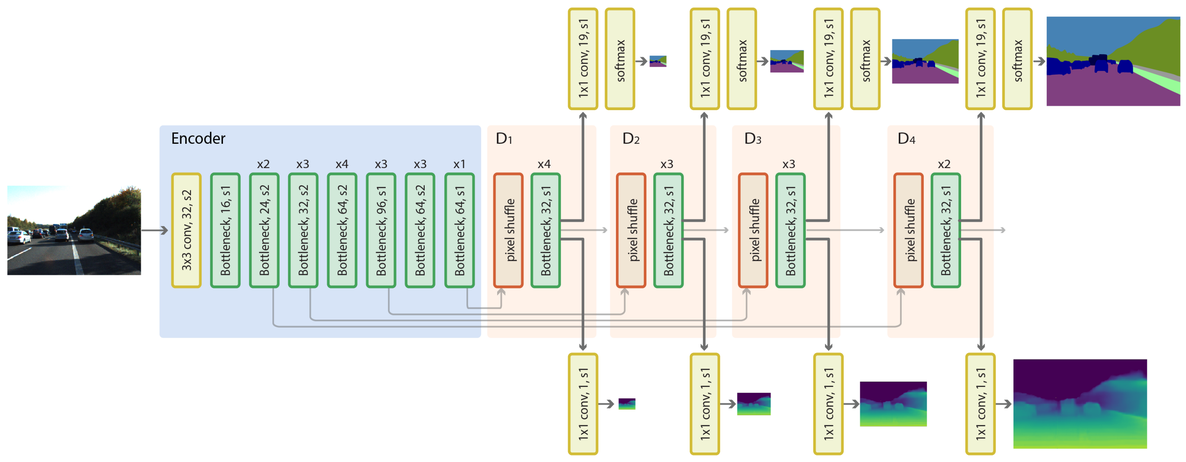

Figure 1. The proposed CNN architecture. Each layer is specified by layer name, number of channels, and strike size. The multiplication (×) on top of a layer gives the repetition of that layer.

Loss Function

In this paper, a loss function formulated as a weighted sum of depth loss and segmentation loss, is applied to each decoder block jointly. By analyzing the depth value distribution of ground truth depth map on KITTI dataset, we discovered that most of ground truth pixels has small depth value while only few pixels have large depth value, which means that the model should be more focused on the depth prediction for nearby objects. Therefore, we proposed a log depth loss that focuses more on nearby pixels by applying a log transform to each pixel of depth map, i.e.,

$$G(d) = \frac{(\log d - \log m) \times M}{\log M - \log m}; m = 4, M = 80,$$

where d gives the depth value of each pixel. The lower bound m and upper bound M are set to 4 and 80, respectively according to the limitation of the LiDaR sensor.

Experimental Results

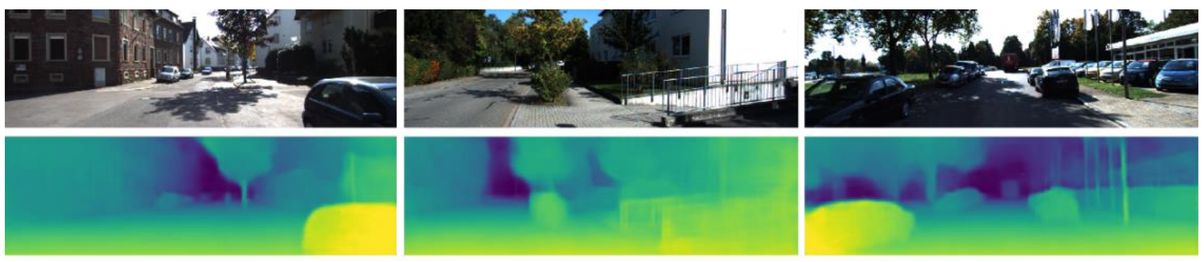

Figure 2. Some qualitative results of our multi-tasks model.
Top: input images, bottom: predicted depth map

Table 1. EVALUATION OF MODELS TRAINED WITH DIFFERENTLY PREPROCESSED TRAINING DATA.

Improvement			RMSE (meter)
Depth Teacher	Log Depth	Segment Teacher	
✓			3.945
✓	✓		3.884
✓	✓	✓	3.871

Table 2. THE COMPARISON OF PREDICTION ERROR AND TOTAL USAGE OF MODEL PARAMETERS.

Method	RMSE	Params
Elkerdawy et al.	5.891	5.9 M
Poggi et al.	6.030	1.9 M
Nekrasov et al.	3.453	2.99 M
Ours	3.871	0.32 M

Table 3. COMPARED WITH PREVIOUS WORKS ON KITTI DATASET.

Model	ARD	SRD	RMSE	RMSE log	Threshold (No cap)		
					<1.25	<1.56	<1.95
Kuznietsov et al.	0.113	0.741	4.621	0.189	0.862	0.969	0.986
Godard et al.	0.133	1.158	5.370	0.208	0.841	0.949	0.978
Eigen et al.	0.203	1.548	6.370	0.282	0.702	0.890	0.958
Liu et al.	0.201	1.584	6.471	0.273	0.680	0.898	0.967
Luo et al.	0.094	0.626	4.252	0.177	0.891	0.965	0.894
Godard et al.	0.114	0.991	5.029	0.203	0.864	0.951	0.978
Luo et al.	0.128	0.935	5.011	0.209	0.831	0.945	0.979
Yin et al.	0.149	1.060	5.567	0.226	0.796	0.935	0.975
Guo et al.	0.111	0.771	4.449	0.185	0.868	0.958	0.983
Yang et al.	0.092	0.547	3.390	0.177	0.989	0.962	0.982
Amiri et al.	0.078	0.417	3.464	0.126	0.923	0.984	0.995
Fu et al.	0.072	0.307	2.727	0.120	0.932	0.984	0.994
Ours	0.106	0.502	3.871	0.160	0.897	0.998	0.9997