Project:Monocular Depth Estimation

李鹏飞
黄奇浩
李欣怡

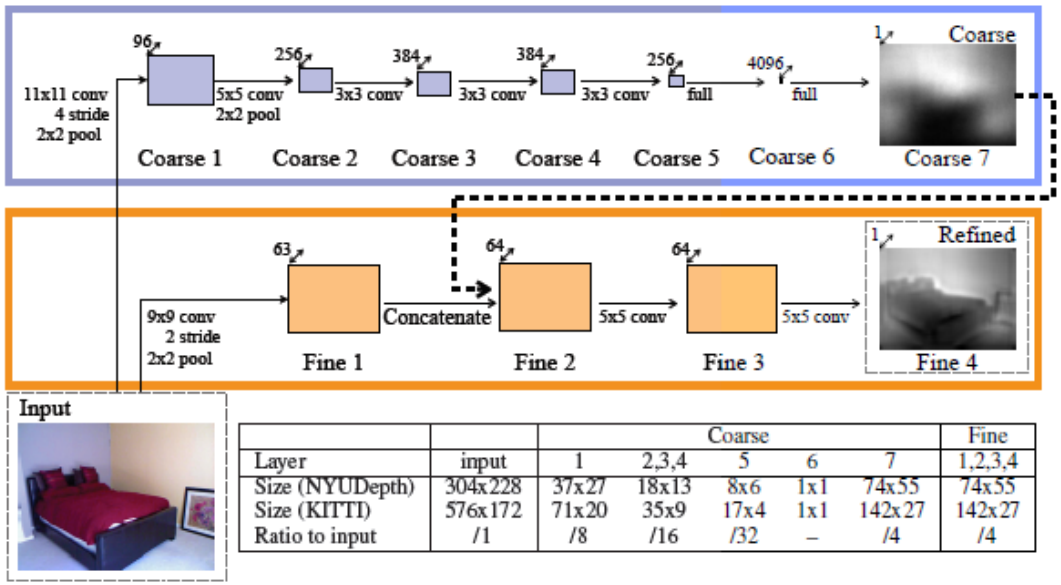
26^{th} Jun 2018

Introduction

- Monocular Depth Estimation
- We analyzed three methods of getting monocular depth images using depth estimation.
  - CNN: Eigen D
  - FCNR: Laina I
  - Unsupervised learning: Garg R,Godard C
- We implement CNN and FCNR using NYUv2 dataset by using pytorch. We will introduce three methods, our project details and test result.

CNN

- Multi-scale network for estimation.
- Coarse estimation for global map and refined by estimation of local maps.
- We *implemented* this method and more detailed discussion about our projects in coming chapter.



		Coarse					Fine
Layer	input	1	2,3,4	5	6	7	1,2,3,4
Size (NYUDepth)	304x228	37x27	18x13	8x6	1x1	74x55	74x55
Size (KITTI)	576x172	71x20	35x9	17x4	1x1	142x27	142x27
Ratio to input	/1	/8	/16	/32	–	/4	/4

Figure 1: Demonstration of CNN procedure

FCNR:ResNet + FCN

- A fully convolutional architecture,encompassing residual learning to model the ambiguous mapping between rgb and depth maps.
- A single end-to-end training architecture without replying on post-processing techniques.
- Novel way of learning feature map up-sampling, Huber loss for optimization and fewer parameters and training data for the proposed model.

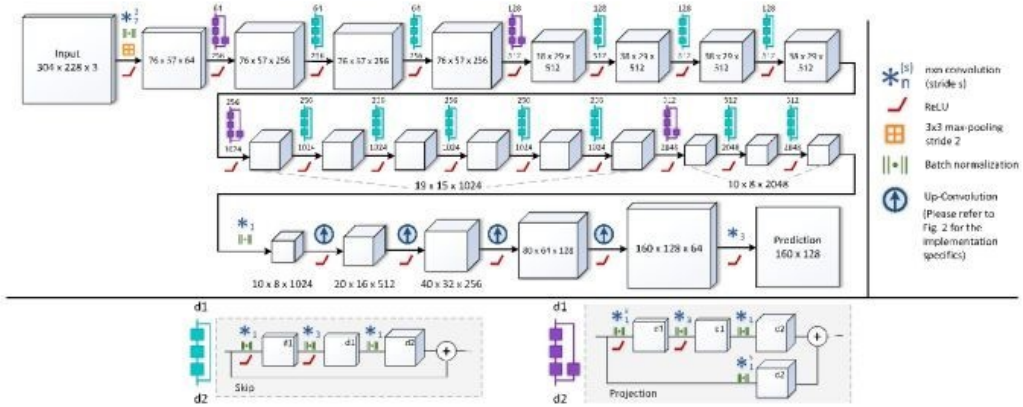


Figure 2: Demonstration of FCNR procedure

Unsupervised learning: transform the monocular depth estimation into stereo depth computing

- Convert the monocular depth estimation into stereo depth computing.
- Combine the forward mapping and inverse or backward mapping method for comparing the original depth image and result image.
- FCN, skip-connect and pre-training results for encoder.

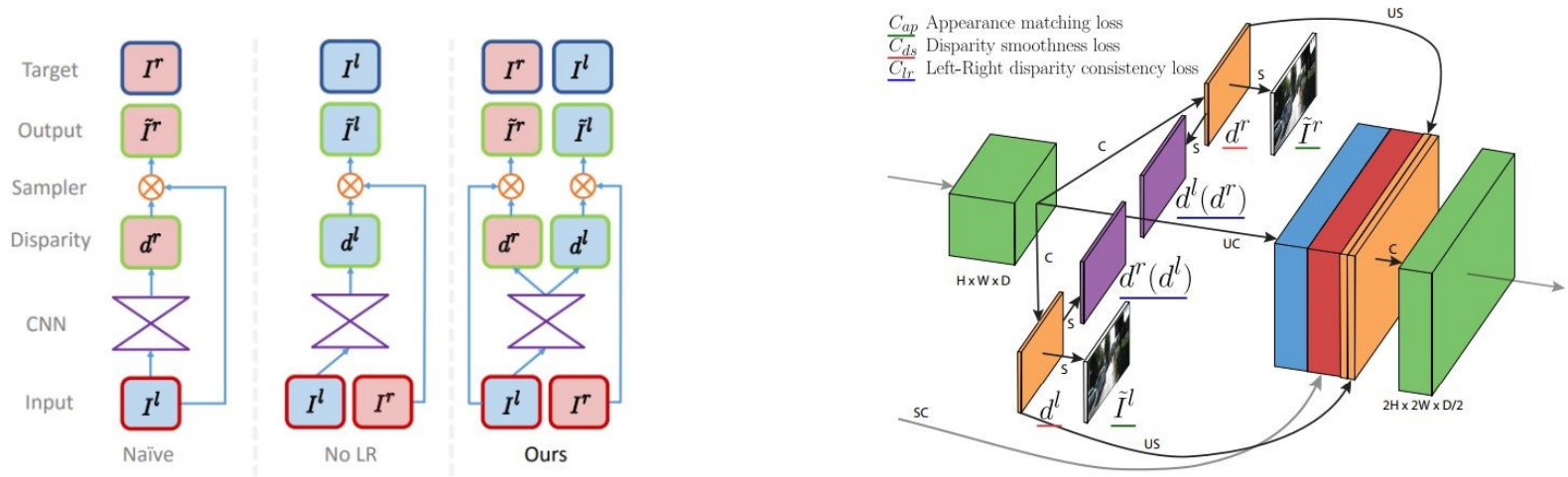


Figure 3: Demonstration of Unsupervised Learning

Multi-scale Network structure

- This model is a multi-scale deep network that first predicts a coarse global output based on the entire image area, then refines it using finer-scale local networks.
- A related work by Eigen et al. apply two convolutional networks in stages for single-image depth map prediction.
- This multi-scale network can be applied to three disparate tasks: depth, Semantic labels and surface normal.

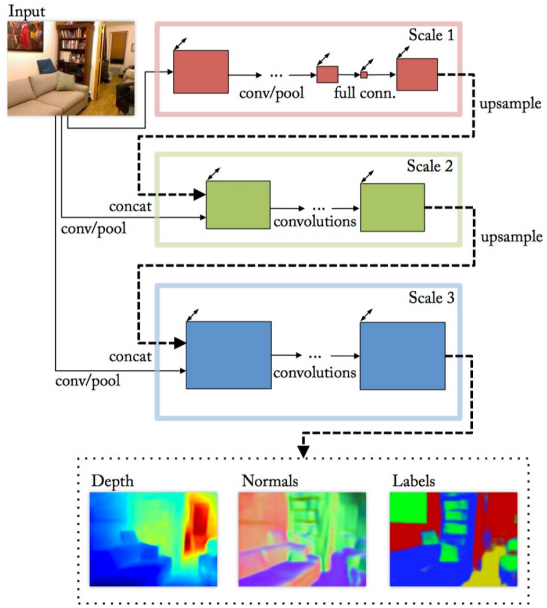


Figure 4: Demonstration of Multi-scale Network

Keypoints in Network Training

- Parameters in each layers of the multi-scale deep network should be initialized carefully we tried kaiming-normal and xaiver-uniform method to initialize parameters
- To avoid gradient explosion problem gradient clipping is necessary.
- First scale in Network should be initialized with **ImageNet** in order to obtain low-level feature maps

Figure 5: ImgaeNet Dataset

Loss functions

- In absence of camera motion to do triangulation, depth estimation from a single image of a generic scene is an ill-posed problem.
- Loss function comparing the predicted and ground-truth log depth maps D and D°

$$L_{depth}(D, D^*) = \frac{1}{n}\sum_i d_i^2 - \frac{1}{2n^2}(\sum_i d_i)^2 + \frac{1}{n}\sum[(\nabla_x d_i)^2 + (\nabla_y d_i)^2]$$

- A simple elementwise loss using a dot product

$$L_{normals}(N, N^*) = -\frac{1}{n}\sum_i N_i \cdot N_i^*$$

- Pixelwise cross-entropy loss for semantic prediction

$$L_{semantic}(C, C^*) = -\frac{1}{n}\sum_i C_i * \log(C_i)$$

Fully Convolutional Residual Networks

Advantage:

- Remove the fully-connected layer, this network operates on an input image of any size
- Retain the spatial structure without fully-connected layer, achieve better result in better result
- Deeper neural network with residual blocks

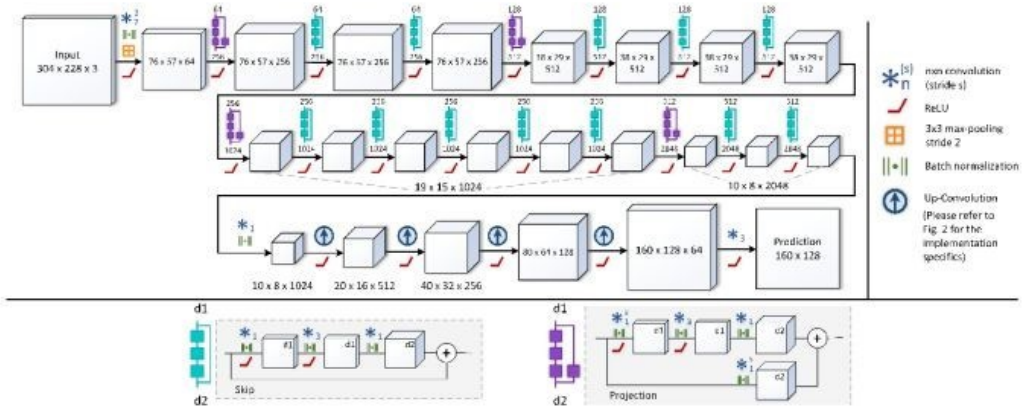
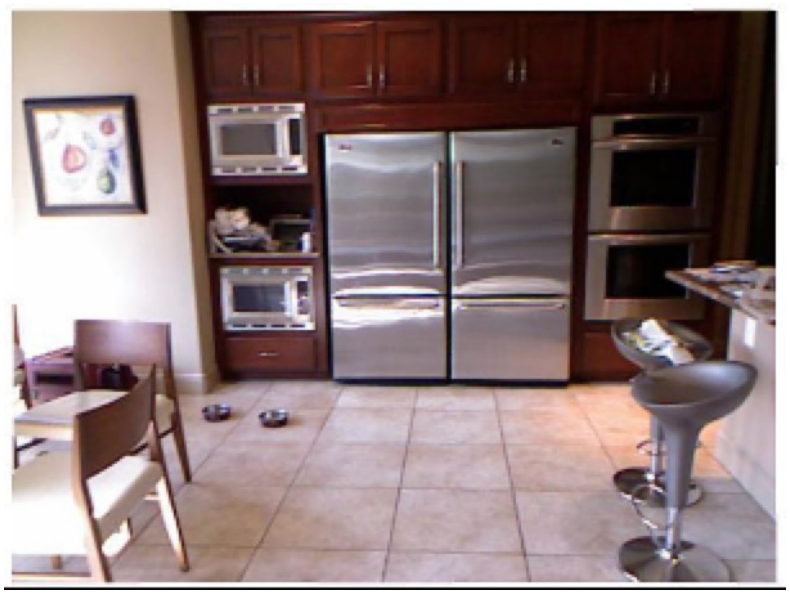


Figure 6: Network architecture

Demonstration of Predicting Depth



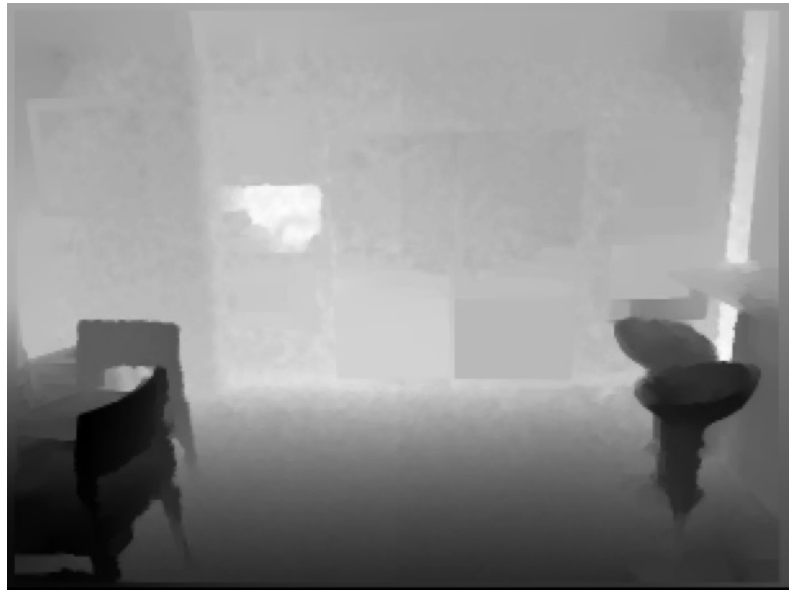


Figure 7: input RGB Depth Label Data to Network

Demonstration of Predicting Depth

- Predicting depth map shows poor representation in detail regions such as void gap in left chair.
- In the right side, two chairs' depth information merges while the original depth map shows it with difference.
- Depth in the background shows strong contour especially in the edge.

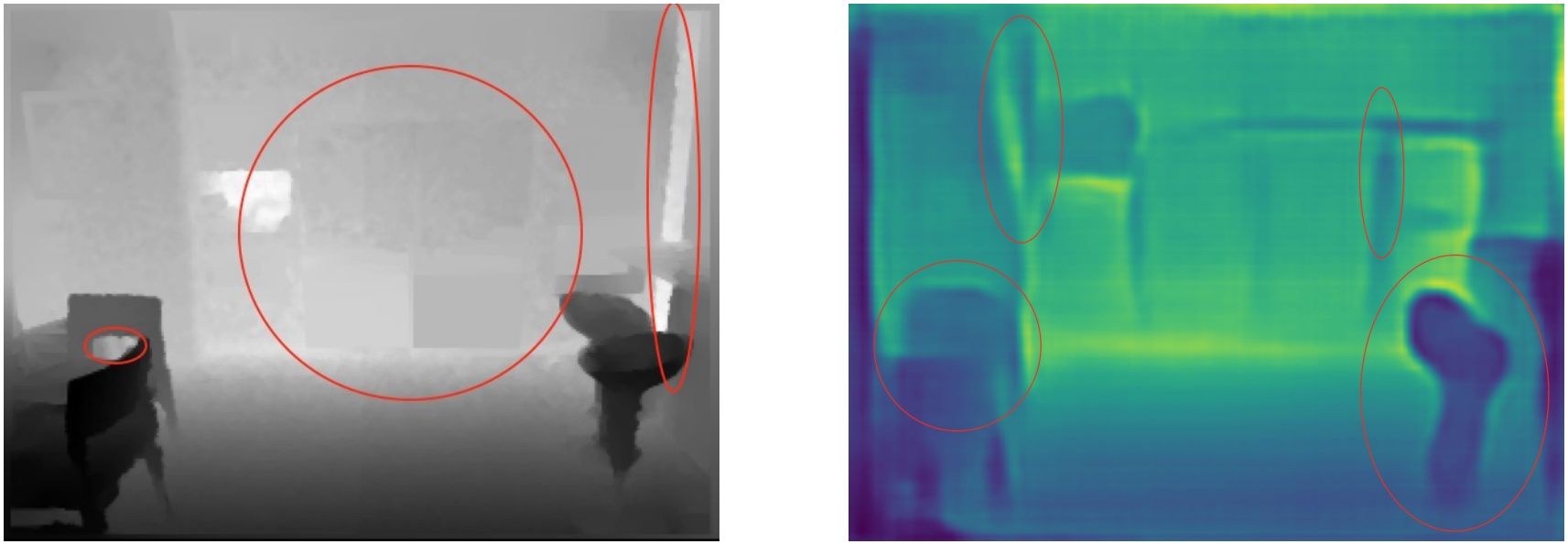
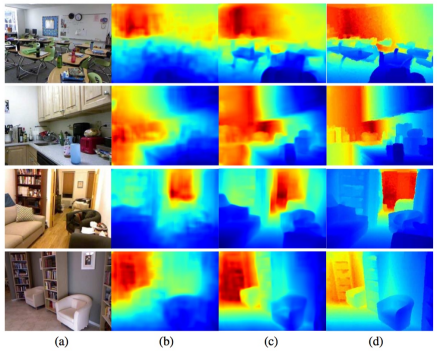


Figure 8: Depth Comparison of Original and Predicting Depth from Network

Results from Other Papers

- Multi-Scale shows better performance comparing to simple CNN arch.
- ResNet enables us to design a deeper network, FCN's features enables the network to scratch map's spatial info.
- Accuracy with delta = 1.25: Multi-Scale 0.769, FCRN 0.811



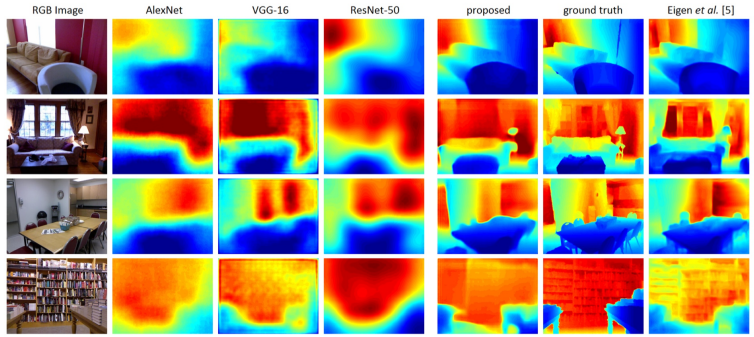


Figure 9: Depth Comparison of Original and Predicting Depth from Network

References

- Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 2650-2658.
- Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C]//Advances in neural information processing systems. 2014: 2366-2374.
- Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.
- Laina I, Rupprecht C, Belagiannis V, et al. Deeper depth prediction with fully convolutional residual networks[C]//3D Vision (3DV), 2016 Fourth International Conference on. IEEE, 2016: 239-248.

References

- Garg R, BG V K, Carneiro G, et al. Unsupervised cnn for single view depth estimation: Geometry to the rescue[C]//European Conference on Computer Vision. Springer, Cham, 2016: 740-756.
- Godard C, Mac Aodha O, Brostow G J. Unsupervised monocular depth estimation with left-right consistency[C]//CVPR. 2017, 2(6): 7.

- Thank You.