Analysis of Deep Learning Algorithms for Single-View 3D Estimation

Hengchao Xiang, Kedarnath Vilankar, James Elder

Centre for Vision Research, York University, Toronto, Canada

INTRODUCTION

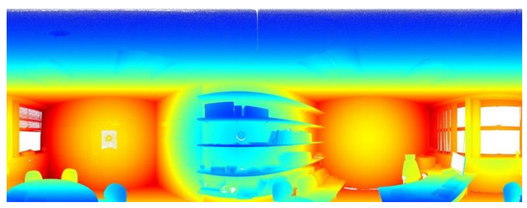
- Understanding 3D scene layout is important for many practical applications, such as 3D modelling.
- Most of the earlier work[1][2] relies on methods and technologies which are not reliable(e.g., depth from stereo and motion cues) and are expensive(e.g., LiDAR).
- Humans are good at extracting 3D scene layout from single view images.
- However, Singe-view 3D reconstruction is an ill-posed problem. Multiple single view images can have same 3D scene layout.
- Recently, Deep Convolutional Neural Networks(DCNNs)[3][4][5] have been used to predict the depth map from a single image.
- We will test and compare three deep networks on Southampton-York Natural Scenes (SYNS) dataset[6].

SYNS DATASET

Deep Convolutional Neural Networks are trained and tested on biased datasets that lack scene categories such as natural scenes, outdoor built scenes and indoor scenes. We tested these networks on our systematically sampled SYNS dataset which includes a broad scene categories.

We used 72 randomly sampled scenes (60 outdoor, 12 indoor) for the testing of the networks. Each scene is comprised of co-registered spherical high resolution HDR image(5400 x 2700 pixel) and LiDAR range data.

Spherical HDR image

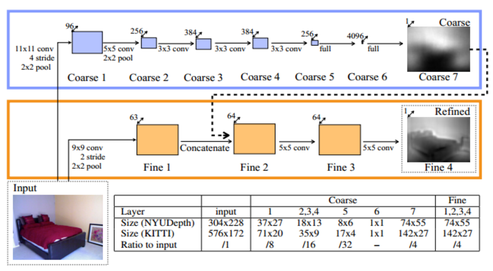


LiDAR range data

DEPTH ESTIMATION ALGORITHMS

This project evaluates the performance of three different state-of-the-art pre-trained deep networks(Eigen et al.[3], Laina et al.[4] and Fu et al.[5]) for single-view 3D estimation on SYNS dataset.

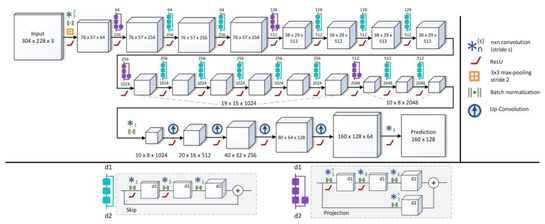
i. Eigen et al. regress dense depth maps on a single image using a two-scale architecture.



Layer	input	Coarse 1	2,3,4	5	6	7	Fine 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

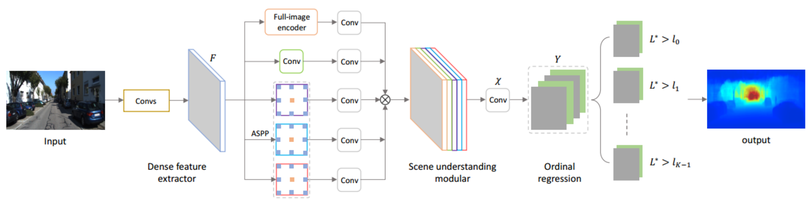
Network architecture of Eigen et al. Figure extracted from [3]

ii. Laina et al. uses a fully convolutional architecture, encompassing residual learning.



Network architecture of Laina et al. Figure extracted from [4]

iii. Fu et al. uses a log depth discretization strategy and recast network learning as an ordinal regression problem.



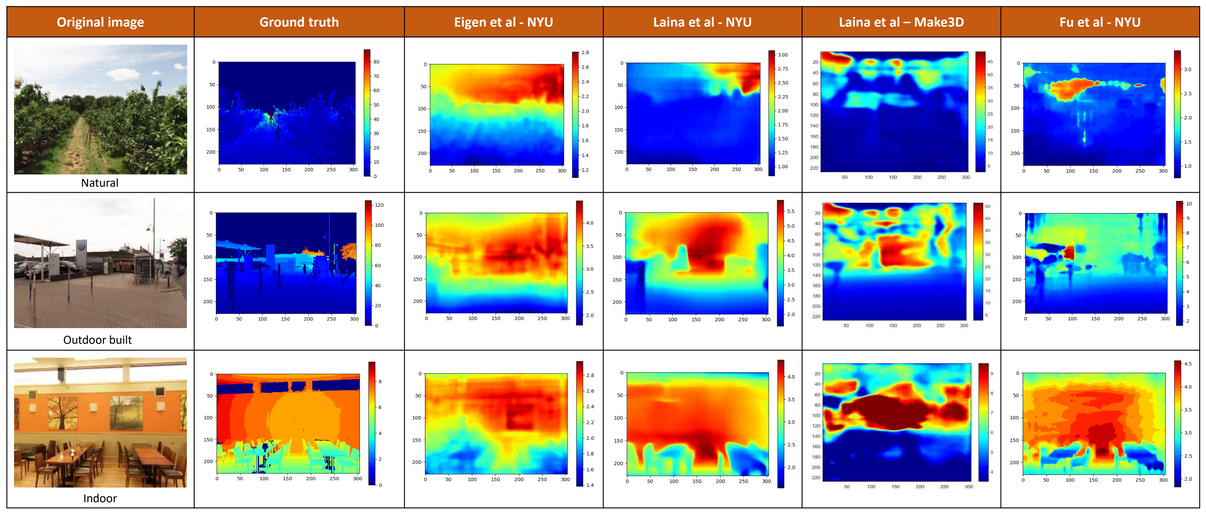
Network architecture of Fu et al. Figure extracted from [5]

The original HDR image in SYNS is high resolution and spherical. We therefore select 12 distinct patches for each scene to match the intended FOV and resolution for each method.

12 patches from one image

RESULTS

- Qualitative results showing predictions using different pre-trained networks on SYNS dataset.



- Quantitative evaluation using four different error metrics.

Evaluation metrics:

1) $rel = \frac{1}{N}\sum\frac{|D-D^*|}{D^*}$

2) $rms = \sqrt{\frac{1}{N}\sum(D - D^*)^2}$

3) $log_{10} = \frac{1}{N}\sum|log_{10}(D) - log_{10}(D^*)|$

4) $rmspe = \sqrt{\frac{1}{N}\sum\frac{(D-D^*)^2}{D^*}}$

where D is the estimated depth, D* is the ground truth, and N is total number of pixels in all images

	Natural				Built Outdoor				Indoor			
	Eigen et al	Laina et al	Fu et al	Laina et al -Make3D	Eigen et al	Laina et al	Fu et al	Laina et al -Make3D	Eigen et al	Laina et al	Fu et al	Laina et al -Make3D
rel	0.73	0.77	0.68	0.35	0.79	0.85	0.80	0.48	0.43	0.38	0.34	0.60
rms	17.42	17.64	16.89	10.19	26.02	26.53	26.11	19.32	3.91	3.72	3.35	3.63
log10	0.64	0.73	0.58	0.19	0.80	0.97	0.84	0.34	0.28	0.25	0.20	0.21
rmspe	0.75	0.78	0.70	0.48	0.81	0.86	0.81	0.56	0.48	0.44	0.40	0.98

Performance of pre-trained networks on SYNS dataset

	Eigen et al	Laina et al	Fu et al
rel	0.16	0.13	0.12
rms	0.64	0.57	0.51
log10	-	0.06	0.05

Performance on NYU dataset

CONCLUSIONS

- The results indicate that all three networks trained on NYU dataset perform poorly on SYNS dataset compared with performance on NYU dataset.
- The algorithms trained on indoor scenes(NYU dataset) perform best in predicating the depth of indoor scene categories. However, because of the biased training, they perform poorly on other scene categories.
- For the indoor scenes, the network of Fu et al. gives the best predictions.
- The pre-trained model of Laina et al. trained on Make3D dataset gives better depth predications on outdoor scenes than indoor scenes. This is because the Make3D dataset consists mostly of outdoor scenes.
- In future, we intend to train and evaluate these algorithms on the broad range of scene categories of SYNS dataset.

REFERENCES

[1] Memisevic, R., & Conrad, C. (2011). Stereopsis via deep learning. In NIPS Workshop on Deep Learning (Vol. 1, p. 2).

[2] Sinz, F. H., Candela, J. Q., Bakır, G. H., Rasmussen, C. E., & Franz, M. O. (2004, August). Learning depth from stereo. In Joint Pattern Recognition Symposium (pp. 245-252). Springer, Berlin, Heidelberg.

[3] Eigen, D., Puhrsch, C., & Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. In Advances in neural information processing systems (pp. 2366-2374).

[4] Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., & Navab, N. (2016, October). Deeper depth prediction with fully convolutional residual networks. In 3D Vision (3DV), 2016 Fourth International Conference on (pp. 239-248). IEEE.

[5] Fu, H., Gong, M., Wang, C., Batmanghelich, K., & Tao, D. (2018, June). Deep Ordinal Regression Network for Monocular Depth Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2002-2011).

[6] Adams, W. J., Elder, J. H., Graf, E. W., Leyland, J., Lugtigheid, A. J., & Muryy, A. (2016). The southampton-york natural scenes (syns) dataset: Statistics of surface attitude. Scientific reports, 6, 35805.