Depth Estimation from a Monocular Image

Ninad Khargonkar, Sreenivas Venkobarao
University of Massachusetts, Amherst
{nkhargonkar, sreenivasv}@cs.umass.edu

Abstract

Monocular depth estimation is a vital sub-task for scene understanding, 3D reconstruction and is shown to help downstream tasks such as semantic segmentation. In the absence of depth maps from expensive RGB-D cameras, single image depth estimation becomes a crucial problem to solve. In this project, we explored methods for estimating the depth of a scene from a single image. We experiment with supervised and unsupervised methods for estimating the depth, and specifically, we investigated the effectiveness of different loss functions and analyzed the results on NYU Depth and Sintel datasets.

1. Introduction

Understanding the depth of objects in a scene is a key problem in Computer Vision, as it will improve performance on down stream tasks such as segmentation, and image classification. Conventional methods for depth estimation involves the use of special 3D sensors such as Microsoft Kinect, or stereo-camera images. However, the vast majority of images and videos available on social media and the Web are not captured with 3D sensors, or stereo-cameras. Therefore, estimating the depth for scenes and objects in this scenario becomes an important problem to solve.
Estimating depth from a single image is an ill-posed problem, since the representation of a 3D world in a 2D image creates a fundamental ambiguity. The same image could have been produced by several scenes. However, by exploiting cues about the relative shape and size of objects, a reliable estimate can be made of the depth of a scene.

In this project, we explored methods to predict depth from a single image which were learned end-to-end using Convolutional Neural Networks. We investigate existing methods, and experiment with modifying the loss functions used for training CNNs for depth estimation.
Section 2 of this report details prior work on single image depth estimation. Section 3 details our approach to the task, and section 4 briefly describes our dataset. The experiments and outcomes are described in section 5.

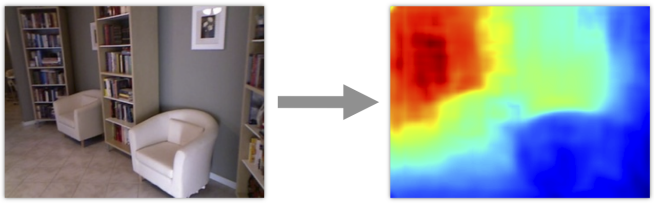


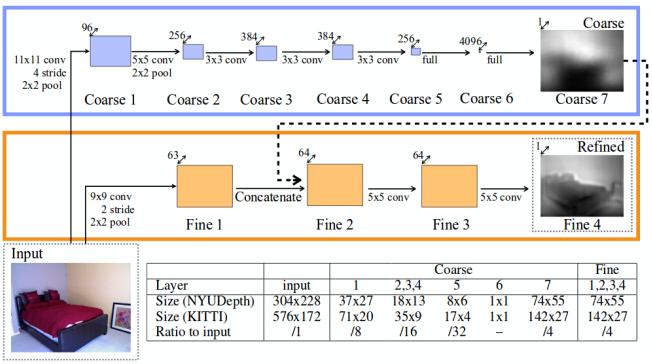
Figure 1. Example image for our task (from [11]). Given an RGB image, estimate the depth of the scene

2. Literature Review

Early work in this field involved using Markov Random Fields [12] with hand-crafted features extracted for multi-scale cues. [6] took a classification approach to model depth estimation and 3D reconstruction. They extract super-pixels from the image, and classify those super-pixels into four classes. [8] also models depth estimation as a classification problem and demonstrated that predicting a canonical depth for pixels removes the problem of perspective, and corrects for bias in the training data. They further show improvement in estimation by jointly predicting semantic labels and canonical depth for each pixel. However, they rely on hand crafted features.

Laina *et. al* [9] and Li *et. al* [10], have demonstrated the effectiveness of using fully convolutional residual networks. Both these papers modify the loss function and show competitive performance. Laina *et. al* use a version of Huber loss in place of L2 loss for regression, while Li *et. al* use classification loss by quantizing the ground truth depths in the log space.

Eigen *et. al* [2], [3] demonstrate that depth estimation can be improved using scale invariant loss to account for the fundamental ambiguity in estimating depth from 2D images. They use a multi-scale architecture and refine the estimated depth at successive scales.



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

Figure 2. Two Scale architecture - image taken from [3]

Garg et al [4] came up with a novel way to predict depth without using any ground truth depth information by leveraging the fact that once disparity for an image in a stereo pair is obtained we can trivially get the depth. Godard et al. [5] in their work improved upon the loss function and also enforced left-right consistency in the prediction of disparity maps.

3. Approach

We explored two recent methods of depth estimation using convolutional neural networks. The first method uses a multi-scale approach - i.e., the network learns to estimate depth at successively finer scales. [2], [3]. The second method demonstrates that using a reconstruction objective on stereo-pair images can induce a learning of disparity maps. [5] Using this disparity map, the depth is predicted.

3.1. Supervised Method Using Ground Truth Depth Maps

The multi-scale architecture we used is shown in figure 2. This architecture generates a coarse depth map estimate at a highly downsampled resolution using AlexNet [7] architecture. This coarse estimate is fine tuned by a second stack of convolution layers. The authors introduce a scale invariant loss term, and model the problem as a regression task. We investigated the performance of this model on the Sintel dataset, by transfer learning, fine tuning and training from scratch.

An improved version of this multiscale architecture was demonstrated in [2]. The authors demonstrated that the same architecture could be applied to related tasks (depth estimation, normal prediction, semantic segmentation) by merely tweaking the loss function. They introduce an additional term besides the scale invariant metric to account for the horizontal and vertical gradients. This encourages consistency in the local structure.

3.2. Unsupervised Method using Stereo pair images

Acquiring ground truth depth data can be an expensive process and often the data is sparse and/or noisy in nature. Another possible bottleneck is that common depth data sets often target a specific type of scene and hence do not have information for a wide variety of scenes. Many of the problems with monocular depth estimation are somewhat relaxed when we have access to a stereo image pair of the same scene.

When a rectified pair of images is available (from a stereo camera pair for example), obtaining the depth information for the scene is relatively simpler task than the monocular setting. Through use of geometry (triangle similarity), the depth can be predicted once the disparity map is calculated and relation is a simple inverse mapping with a constant factor

$$depth = \frac{Bf}{disparity} \tag{1}$$

Here, B is the stereo camera baseline and f is the focal length. The diagram for obtaining the relation can be seen in Fig 3. The $disparity$ of a pixel is the displacement of it across the left and right images and due to the alignment assumption, the displacement will be in the horizontal direction (along the epipolar line) and hence its calculation is then relatively straightforward as seen in Fig. 4.

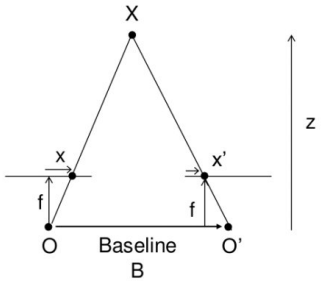


Figure 3. Geometric relations for disparity to depth

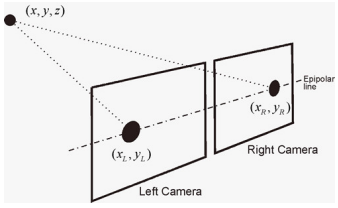


Figure 4. Epipolar line: The displacement for a pixel will be along the horizontal direction

Thus in a stereo-pair image based method, we do not require any ground truth depth information for the learning process, instead we require an rectified stereo image data set from which we can generate the depth. Godard *et. al*

[5] show that a network can be trained to predict a disparity map given the left image of a stereo pair and therefore it only requires stereo pairs during training and it primarily learns using an image reconstruction loss. Once a disparity map for the left image is predicted, the image reconstruction loss is calculated for the right image in the stereo pair by estimating the right image from the left image using simple bilinear sampling over the disparity value for each pixel. In addition to the image reconstruction loss, some other constraints are also enforced which help in improving the quality of the results.

3.3. Network details

The neural network is fully convolutional in nature with an encoder - decoder scheme to get the disparity map out of an input image. The encoder setup is the usual convolution and pooling layers while the decoder setup uses upconvolutions to get disparity maps at 4 scales along with skip connections from the encoder part for providing information to the predictions at the higher resolutions scales. Also, the disparity outputs double in size in each successive scale.

In addition to the stereo image reconstruction loss, some other constraints like smoothness and left-right consistency are also introduced by Godard et at [5]. The smoothness term enforces the depth to be locally smooth except at places where the image gradient is high as depth discontinuity is found to occur at regions with large value for gradient. This is enforced with an exponential weight term for depth gradients in each x and y directions with image gradients in x and y directions respectively:

$$L_{ij}^{smooth} = |\partial_x d_{ij}|e^{-||\partial_x I_{ij}||} + |\partial_y d_{ij}|e^{-||\partial_y I_{ij}||} \quad (2)$$

The left-right consistency term ensures that the disparity map is also aligned with the left image and to give better results on object boundaries. The network predicts both the left and right disparities from the left image alone and then reconstruction loss takes into account both the ground truth left and right images. The broad overview is shown in Fig. 5

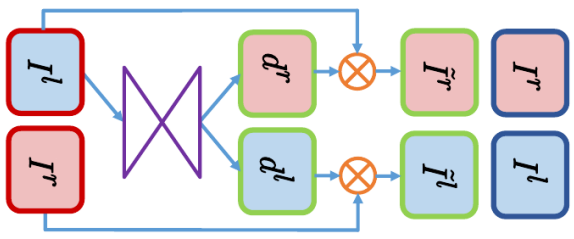


Figure 5. Left-right consistency using sampling from both left and right disparity maps

4. Dataset

The NYUv2 dataset [11] contains 400K natural images of indoor scenes, with corresponding depth maps, generated using a Kinect camera. Owing to our limited resources, we train and test on the 1449 densely labeled pairs of aligned RGB and depth images from the NYU dataset. We also experimented with training on the Sintel dataset [1]. The Sintel dataset is different from the NYU dataset in that it is largely images of outdoor scenes generated artificially.

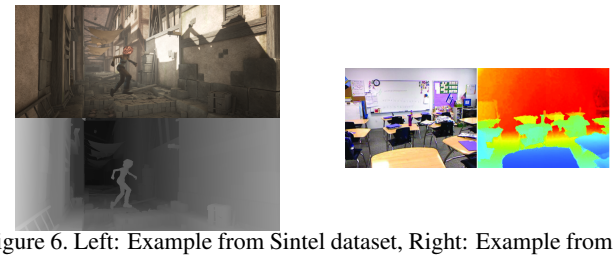


Figure 6. Left: Example from Sintel dataset, Right: Example from NYU dataset

5. Experiments on Supervised Depth Estimation

We initially experimented with the multi-scale architecture of [2], but owing to resource constraints, we shifted to their earlier work [3].

In order to test the generalization of [3] to different datasets, we used the architecture pre-trained on the NYUv2 dataset in [3] as a black-box feature extractor and trained a neural network classifier to regress onto the depth for the Sintel dataset. Specifically, we extracted the fc6 features from the coarse scale (a 4096-d vector), and the conv-3 features from the fine scale of the architecture. The conv-3 feature maps had a dimensionality of $74 \times 55 \times 64$. In order to reduce dimensionality, we used a single convolutional kernel, and got an output of size $74 \times 55 \times 1$. We flattened this into a (4070-d) vector. The two vectors were concatenated, and a network was trained to predict the depth using the scale invariant loss function defined in [3].

This approach failed to produce good results, indicating that the model was not able to extract features that were helpful for predicting the depth on the Sintel dataset. This is to be expected since the network was trained on real-world images of indoor scenes such as furniture, while the Sintel dataset was a synthetic dataset containing images of human and non-human characters.

We explored training the network architecture from scratch for Sintel. However, this approach led to overfitting since the network was too large to be trained on the limited amount of data in the Sintel dataset.

The per sample loss function we used for these experiments was the scale invariant loss defined in [3]:

Model	Coarse	Fine
Two Scale Model [3]	0.043	0.044
Model with Modified Loss	0.101	0.0835

Table 1. Performance comparison using RMSE (log. scale inv. loss) metric on NYU v2 dataset. These values are for the 1449 images from the NYUv2 dataset.

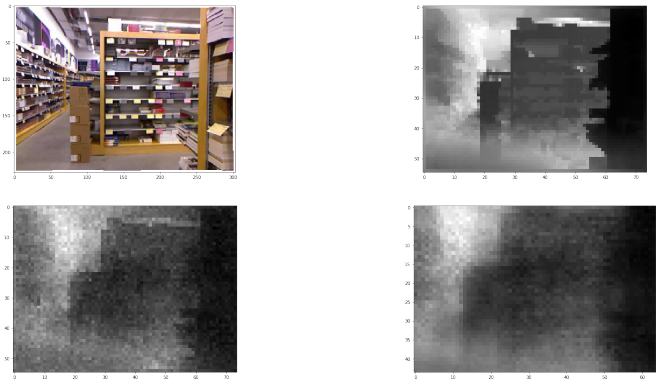


Figure 7. Top: Example Image from NYU Dataset with corresponding true depth values. Bottom: Predicted Depth at Coarse Scale and Fine Scale from model using the scale invariant loss.

$$D(y, y^*) = \frac{1}{2n^2} \sum_{i,j} ((\log y_i - \log y_j) - (\log y_i^* - \log y_j^*))^2 \tag{3}$$

Here, y is the predicted depth map, and y^* is the ground truth map. Since the true scale is ambiguous in 2D depth estimation, this loss function imposes a relative consistency constraint - the network has to produce depth estimates which are consistent within the image, even if the scale is different from the ground truth depth.

We noticed that the depth maps were very noisy. So we added a variance term to the loss function. Minimizing the variance over the difference between the ground truth disparity and the predicted disparity would generate stable depth estimates.

$$D(y, y^*) = \frac{1}{2n^2} \sum_{i,j} ((\log y_i - \log y_j) - (\log y_i^* - \log y_j^*))^2 + Var(log(y) - log(y^*)) \tag{4}$$

We found that the denoising term performed better on the Sintel dataset in comparison to the NYU dataset. Qualitative results for the various approaches are shown in figures 7, 8, and 9.

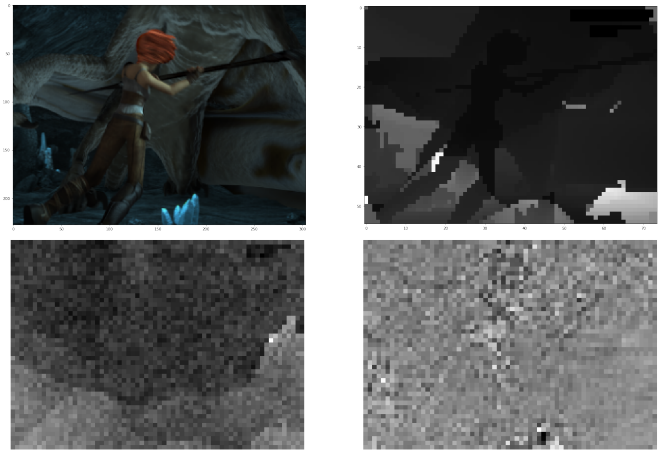


Figure 8. Top: Example Image from Sintel Dataset with corresponding true depth values. Bottom: Predicted Depth at Coarse Scale and Fine Scale using the scale invariant loss.

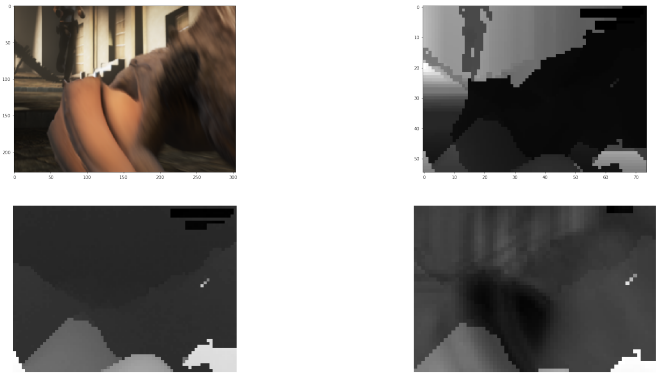


Figure 9. Top: Example Image from Sintel Dataset with corresponding true depth values. Bottom: Predicted Depth at Coarse Scale and Fine Scale from model using the modified loss function.

6. Experiments on Unsupervised Depth Estimation

6.1. Fine tuning on Sintel dataset

For the qualitative result, we take a stereo pair from the sintel data set and show the both disparity map predicted by the network using just the left image and a disparity map calculated from both the images of the stereo pair (ground truth). One can see the network learns to predict decent disparity maps after fine-tuning on the sintel dataset. The fine tuning loss plot is show in Fig. 10 and the depth visualizations are show in Fig. 11.

6.2. Uncertainty in predictions

In order to analyze the distribution of the disparity map for a single test image, we apply 50 different color augmentations to the same image and get 50 different outputs to compare against. Then we obtain the variance for the

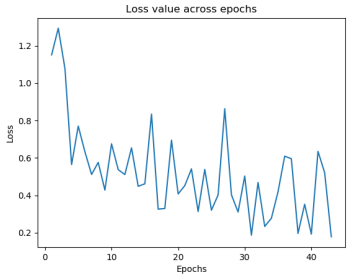


Figure 10. Fine tuning loss

Figure 11. The top row has the left and right image of the stereo pair. On the bottom row, the left image is the ground truth while the right image is the predicted

disparity prediction for each pixel and observe where the network is most uncertain in making the predictions.

It is found that in many cases, occlusion edges induced uncertainty into the prediction. The color augmentations that were used are random gamma (between values of 0.8 and 1.2) and contrast (between values of 0.5 and 1.5). The uncertainty visualizations are shown in Fig. 12. One can see higher variance value at the occlusion edges.

Figure 12. An image with the visualization of the variance in predictions

6.3. Occlusion aware loss term

As seen in the previous sub-section, occlusion edges are somewhat problematic for the network and produce disparity estimates which have higher uncertainty. To address this we substitute an occlusion-aware term instead of the edge-aware weight term in the loss function to enforce the smoothness constraint. As seen in Fig 13, occlusion boundaries can be approximately estimated by looking at gradient of predicted depth map.

Figure 13. An image from sintel data set (market scene) with the x-gradient of depth

Instead of weighing the L_1 norm of the disparity with inverse exponential term of the image gradient, we weight it with the inverse exponential term of the normalized depth gradient (depth can be found up to a constant factor by inverting the disparity predicted). Our motivation comes from the paper by Wang et al [13], where they use thresholded normalized depth gradient as cue for the occlusion boundary. The normalization ensures robustness of the depth gradient to changes in depth across the image. Therefore the smoothness loss can now be written as: ($depth_{ij} = \frac{1}{d_{ij}}$ and d_{ij} is the predicted pixel disparity).

$$L_{ij}^{smooth} = |\partial_x d_{ij}| e^{-||\partial_x depth_{ij}||} + |\partial_y d_{ij}| e^{-||\partial_y depth_{ij}||} \tag{5}$$

The expectation from such a constraint on the smoothness loss was that it would only enforce the smoothness on non-occlusion boundary regions. Since the quality of the disparity (and hence depth) is decent, we thought that using the predicted depth itself in the loss would help in fine tuning. One inadvertent consequence is the partial reduction in smoothness along 'normal' edges in the images and although the results are not too different from earlier model, we observe some minute qualitative differences in the output as seen in Fig 14.

7. Conclusion

Although fully supervised end to end trainable models for single image depth estimation exist [3], we found that

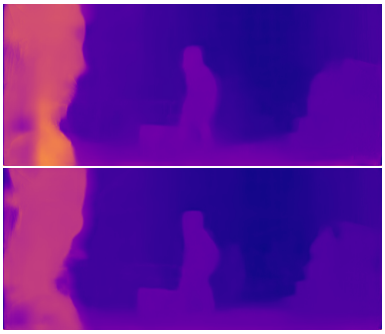


Figure 14. Top: Output of model without the depth-aware term in the smoothness constraint. Bottom: Output of the model which has the depth-aware term. On close inspection, one can see some background objects being detected in the bottom image.

it is too specific for a scene and does not generalize well across different data sets. Enforcing a denoising term in the loss function did not shown any conclusive improvement in the depth prediction and overall the supervised approach did not yield good results for us (although we learned to implement architectures from scratch in tensorflow). The unsupervised approach which used the stereo pairs during training showed somewhat promising results in transferring information across to an entirely new data set perhaps due to the fact that training on stereo reconstruction loss made the model generalize well. Some some weak supervision for handling occlusion edges by introducing an occlusion aware term improved the estimates by a small amount. As a part of the future work we would like to tweak around the reconstruction loss in the model and observe any changes in the results.

References

[1] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, October 2012.

[2] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

[3] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[4] Ravi Garg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756. Springer, 2016.

[5] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. *arXiv preprint arXiv:1609.03677*, 2016.

[6] Derek Hoiem, Alexei A Efros, and Martial Hebert. Automatic photo pop-up. *ACM transactions on graphics (TOG)*, 24(3):577–584, 2005.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[8] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–96, 2014.

[9] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016.

[10] Bo Li, Yuchao Dai, Huahui Chen, and Mingyi He. Single image depth estimation by dilated deep residual convolutional neural network and soft-weight-sum inference. *arXiv preprint arXiv:1705.00534*, 2017.

[11] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[12] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. Learning depth from single monocular images. In *Advances in neural information processing systems*, pages 1161–1168, 2006.

[13] Ting-Chun Wang, Alexei A Efros, and Ravi Ramamoorthi. Occlusion-aware depth estimation using light-field cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3487–3495, 2015.