Monocular Depth Estimation

CS231N Project

Piyush Gulve

Gaurav Gupta

Introduction

- Depth information in computer vision has applications in various fields, including SLAM, AR and VR applications, object detection, semantic segmentation, etc.
- Traditional methods for depth estimation such as interpreting depth include Structure from Motion (SFM) use geometric properties of image such as using a series of 2D Images or Stereo Vision Matching recovers 3D structures from observing the scene from two viewpoints, simulating how human eyes have two cameras.
    - Other techniques, such as Feature-based mapping methods, rely on the assumption that similarities between regions in the RGB images also imply similar depth cues Sensor-based methods utilize depth sensors, like RGB-D cameras and Lidar.
- Many depth hints such as perspective and object sized can be exploited from monocular images and CNN is an ideal tool to utilize these information.

Related Work

Convolutional Neural Networks have been used for Depth Estimation over the past 10 years.

- Using Fully Connected CNN, including optimizations such as residual learning, Continuous Random Fields, Attention Guided networks.
- Multi-view stereo approaches use CNN to depth estimation from Stereo Cameras or Multiple frames of image taken from videos of a scene.
- Transfer Learning & Encoder - Decoder networks techniques have been shown to improve the the quality of the depth produced.
    - We explore couple of these in our project.

Problem Statement

- Input: single RGB image
- Output: depth image
- Evaluation: both in quantitative metric comparison and qualitative visualization

Dataset

- NYU Depth V2
- Composed of video sequences from indoor scenes
- Recorded by both the RGB and Depth cameras from the Microsoft Kinect
- 120k Images with resolution of 640x480
- We used 32K images for training with 70-30 split for training and validation
- Input images were resized to 320x240 to match test data resolution

U Net Architecture

- 4 Downsampling Blocks
- Each downsampling block consists of 3x3 conv (unpadded), ReLu, 2X2 maxpool
- 4 Up Sampling Blocks

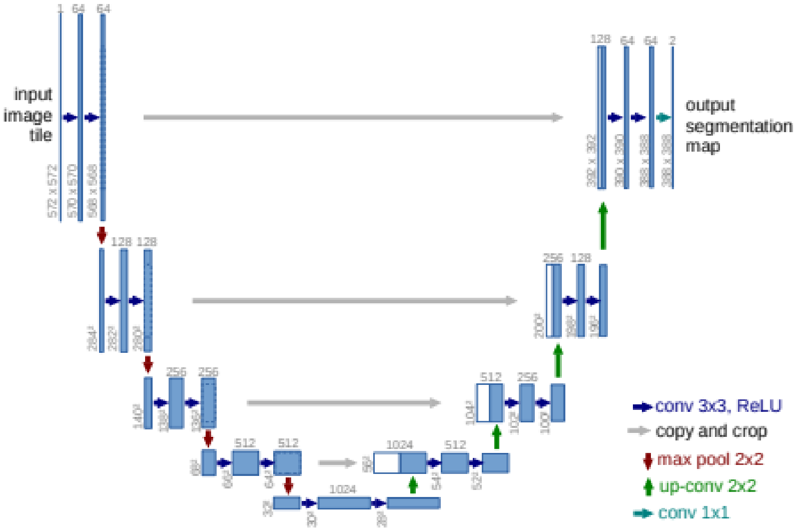


Figure U-Net Architecture [8]

Model 1 (Unet based CNN)

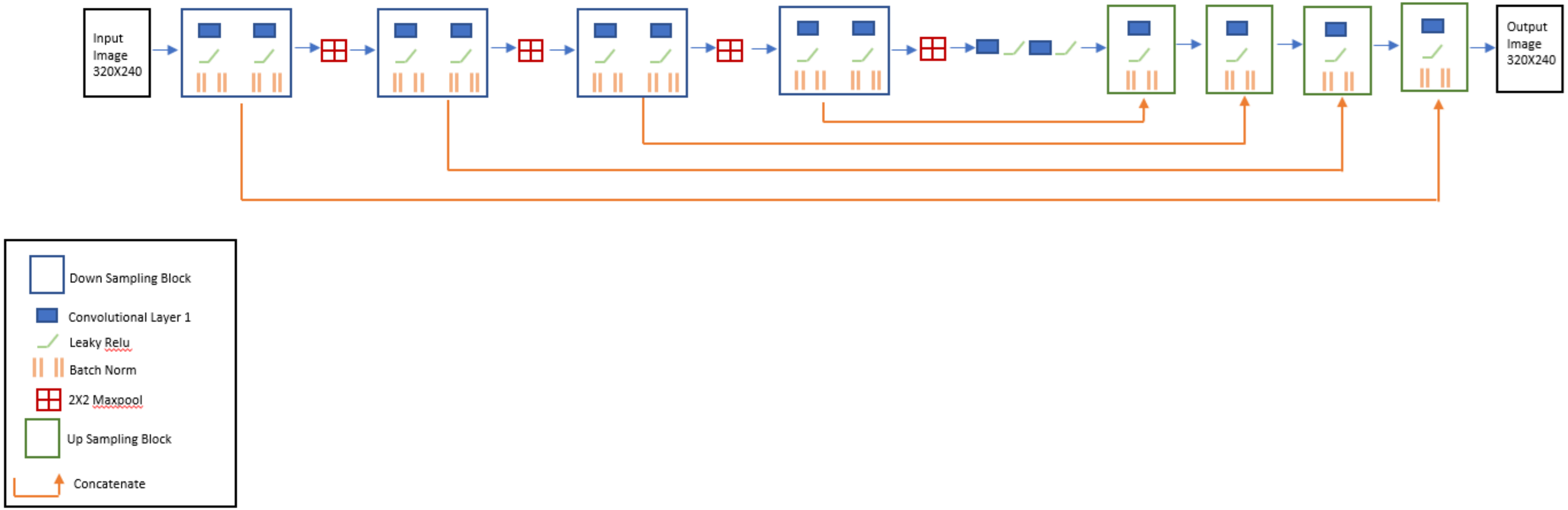
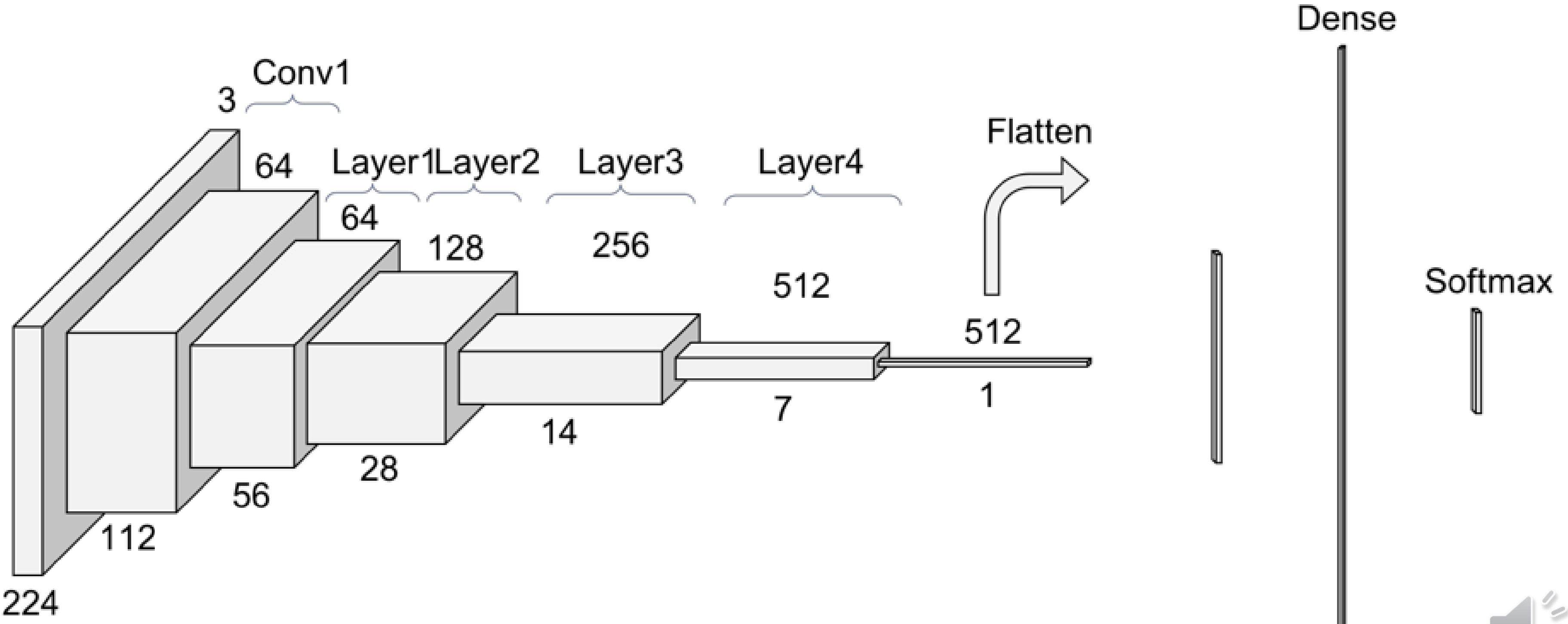


Figure Our model architecture

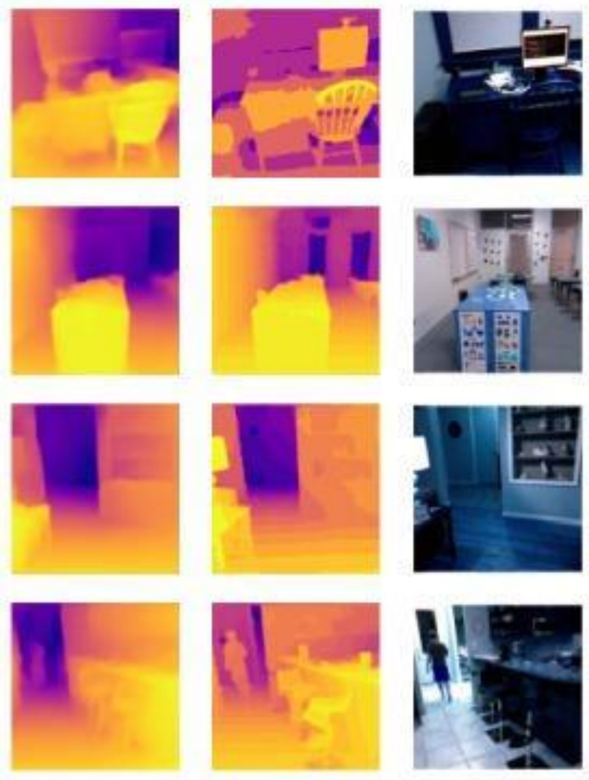
- 4 Upsampling blocks each consisting two 3X3 conv (same padding), Leaky Relu, Batch Norm each
- A ‘bottleneck’ of 3X3 conv and Leaky Relu
- 4 Upsampling blocks each consisting 3X3 conv (same padding), Leaky Relu, Batch Norm

Resnet 34 Architecture

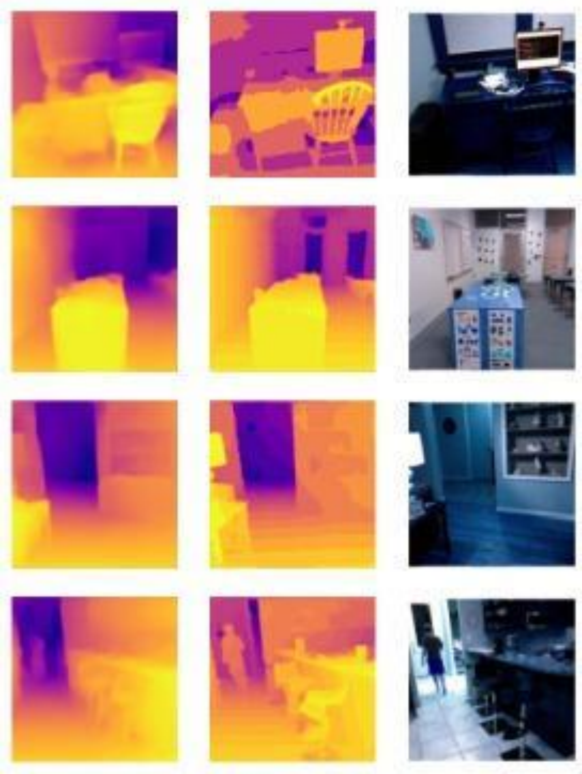


Experimental Evaluations - Qualitative

Depth maps results of of UNet based CNN model



Depth maps results of transfer learning based model.



Depth Accuracy of 0.76 (UNet) vs 0.82 (Transfer Learning)

Experimental Evaluations - Quantitative

Loss Function

$$L(y,\hat{y}) = \lambda L_{depth}(y,\hat{y}) + L_{grad}(y,\hat{y}) + L_{SSIM}(y,\hat{y}). \tag{1}$$

$$L_{depth}(y,\hat{y}) = \frac{1}{n}\sum_{p}^{n}|y_p - \hat{y}_p|. \tag{2}$$

$$L_{grad}(y,\hat{y}) = \frac{1}{n}\sum_{p}^{n}|\boldsymbol{g}_{\mathbf{x}}(y_p,\hat{y}_p)| + |\boldsymbol{g}_{\mathbf{y}}(y_p,\hat{y}_p)| \tag{3}$$

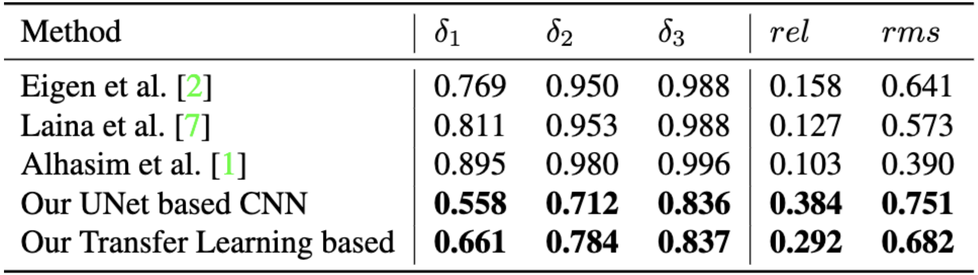
$$L_{SSIM}(y,\hat{y}) = \frac{1 - SSIM(y,\hat{y})}{2}. \tag{4}$$

Results

average relative error (rel): $\frac{1}{n}\sum_{p}^{n}\frac{|y_p - \hat{y}_p|}{y}$

root mean squared error (rms): $\sqrt{\frac{1}{n}\sum_{p}^{n}(y_p - \hat{y}_p)^2)}$

threshold accuracy (δ_i): % of y_p s.t. $\max(\frac{y_p}{\hat{y}_p}, \frac{\hat{y}_p}{y_p}) = \delta < thr$ for $thr = 1.25, 1.25^2, 1.25^3$;



Method	δ_1	δ_2	δ_3	rel	rms
Eigen et al. [2]	0.769	0.950	0.988	0.158	0.641
Laina et al. [7]	0.811	0.953	0.988	0.127	0.573
Alhasim et al. [1]	0.895	0.980	0.996	0.103	0.390
Our UNet based CNN	**0.558**	**0.712**	**0.836**	**0.384**	**0.751**
Our Transfer Learning based	**0.661**	**0.784**	**0.837**	**0.292**	**0.682**

References

[1] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning.
arXiv arXiv preprint 1812.11941
[2] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems
[3] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction
with fully convolutional residual networks. In 2016 Fourth international conference on 3D vision (3DV), pages 239–248. IEEE, 2016.
[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox.
U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241, Springer, 2015. 1, 2