# ECE 740 Project Midway Report

Mahdi Abdollah Chalaki
University of Alberta
Edmonton, AB, CA
Abdolla2@ualberta.ca

Masoud Jafaripour
University of Alberta
Edmonton, AB, CA
jafaripo@ualberta.ca

## Abstract

*In this report, we will discuss what we have achieved so far in our project, including a discussion about various datasets, a baseline model implementation and our obstacles. We will also mention some solutions that might help us to improve our model performance, and a timeline that we think we can schedule our efforts towards reaching our goal.*

## 1. Introduction

In this report, we will begin by proposing a summary of the work we have done so far and contribution of each team member, and the we will discuss our efforts, including introducing a number of datasets within the area of monocular depth estimation field and also, we will show the results of our implementation of a baseline model.

### 1.1. Our efforts

In the past month, we started learning more about Convolutional neural networks. Although we were both familiar with simple neural networks, but through the literature review process, we found out that it is essential to use convolutional networks to reach a superior performance. Therefore, we audited the first three courses of the Deep learning specialization (available on Coursera) and tried to get hands on experience through some small projects.

In addition, we tried to split the work and assign each member a specific task. Masoud went through a literature review on the available and most popular datasets used for depth estimation. He also investigated possible future works which could be done on Monocular Depth Estimation (MDE), in particular those associated with Unsupervised Learning (UL) approaches. Meanwhile, Mahdi worked on the implementation of a rather simple convolutional network, to make a baseline model. The results are discussed in this report.

At last, it is noteworthy to say that in spite of the fact that we were working on different parallel tasks, we shared our information and helped each other to reach our shared goal.

## 2. Datasets

Datasets play a vital role in training and assessing monocular depth estimation. Quite a few datasets can be classified based on the scenario, sensors, annotation, and type of data stored in them. Here, some of the most frequent and popular ones are highlighted. In Table 1, a summary of these datasets is provided.[1][2]

### 2.1. KITTI

The KITTI [3] is not only the most commonly implemented dataset in computer vision but is the most used dataset for unsupervised and semi-supervised MDE as well. The KITTI includes 44K images and their associated depth maps collected in the driving scenario from outdoor scenes in the city, rural areas, and roads. It is common to use Eigen split method for dividing this dataset into two sections for DL use: 32 scenes for training and 29 scenes for testing [4][1][5]

### 2.2. NYU Depth

The NYU Depth [6] is one the most popular dataset for depth estimation and the primary source in the training stage of supervised MDE. This dataset includes 1449 densely annotated images divided into two sections for testing and training. The NYU Depth dataset focuses on indoor environments and scenes, including indoor and depth images corresponding to them. Some studies reduced its image resolution to speed up training. [4][1]

### 2.3. DIODE

The DIODE (the Dense Indoor/Outdoor Depth) dataset [7] is a monocular depth dataset that includes diverse indoor and outdoor scenes collected with the same hardware setup. In this dataset, 8574 indoor and 16,884 outdoor samples were collected from 20 scans for training, and 325 indoor and 446 outdoor samples were collected from 10 scans for validation. In the Middlebury 2014 Middlebury dataset, 33 images have a resolution of 6 megapixels, with indoor and outdoor ranges of 50 m and 300 m, respectively. A stereo DSLR camera and two point-and-shoot cameras were used to capture the images.

Table 1 A summary of most popular datasets

| Dataset | Sensors | Annotation | Type | Scenario | Images | Resolution | Year |
|---|---|---|---|---|---|---|---|
| KITTI | LIDAR | Sparse | Real | Driving | 44K | 1024×320 | 2013 |
| NYU-Depth | Kinect V1 | Dense | Real | Indoor | 1449 | 640×480 | 2012 |
| DIODE | Laser Scanner | Dense | Real | In/Outdoor | 25.5K | 768×1024 | 2014 |
| Make3D | Laser Scanner | Dense | Real | Outdoor | 534 | 1024×2048 | 2016 |
| Cityscapes | Stereo Camera | Disparity | Real | Driving | 5K | 2272×1704 | 2008 |
| Driving Stereo | LIDAR | Sparse | Real | Driving | 182K | 1762×800 | 2019 |

With a 6 megapixels resolution, the disparity ranges from 200 to 800 pixels.

### 2.4. Make3D

The Make3D dataset [8] includes 534 images taken by laser scanners during an outdoor scenario. The Make3D contains monocular RGB and depth images but does not include stereo images. Therefore, it is suitable for training steps in the supervised method. It usually is not applied to semi-supervised and unsupervised learning training and is mainly used for evaluating unsupervised learning networks over disparate datasets as a testing set. Like some of the other datasets with high-resolution images, down-sampling its images can speed up the training step.

### 2.5. Cityscapes

The Cityscapes datasets [9] possess 25K of fine- and course-annotation images mainly focused on semantic segmentation.

### 2.6. Driving Stereo

Among the new large-scale stereo driving datasets, driving stereo contains 182k images. In addition to disparity images, LIDAR can also capture them. MDE stereo matching is evaluated using two new metrics, a distance-aware metric, and a semantic-aware metric. 1762*800 pixels are the dimensions of this dataset.[10]

## 3. Model Implementation

As we have recently started learning about Neural Networks and the convolutional networks, we tried to implement a very simple model so we can have a baseline model. We have borrowed our model from [11], which uses a simple U-net [12] as the backbone. In this model, we used the indoor image section of the DIODE dataset which we discussed before. There is roughly 81GB of data in the training set, compared to 2.6GB in the validation set. Due to the fact that this is not the final model, and training on the training data requires a lot of computational resources, we just used the validation set as our training set. Other datasets also could be used for training. In figure 1, a few samples of this dataset are plotted.

### 3.1. U-Net

U-net architecture is symmetric and consists of two major parts. A general convolutional process constitutes the first part, which is called the contracting path. The second part layers (expansive path) are formed by transposed 2d convolutional. In Figure 2, the original architecture of a U-Net is displayed.

### 3.2. Our Model

Our model is a bit different from this one and it is a smaller version. Input files include RGB images, depth images, and depth mask images. The images are first resized to 256*256, and then fed into the network.

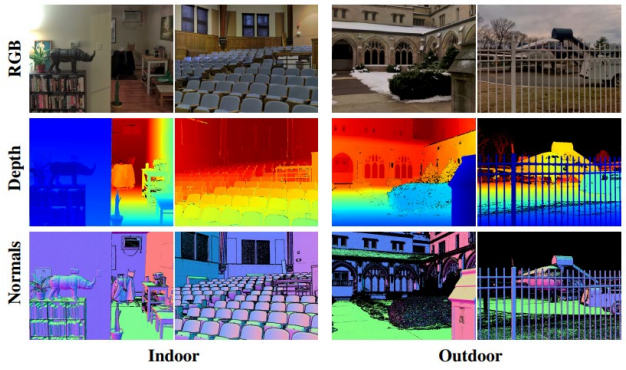


Figure 1: Random examples of indoor and outdoor pictures sampled from the DIODE dataset

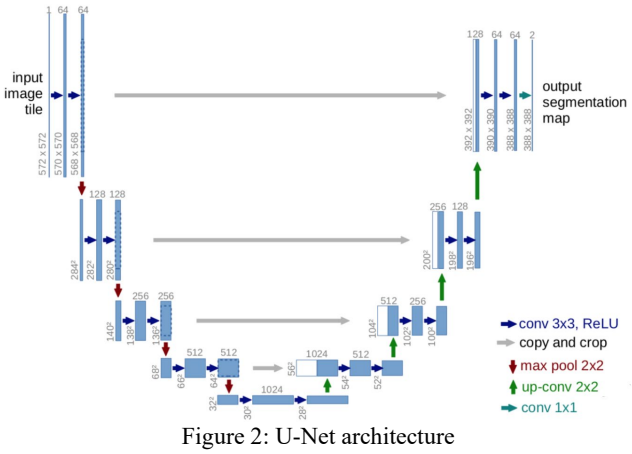


Figure 2: U-Net architecture

There are 8 convolutional layers in the downscaling part, with 16,32,64, and 128 filters per pair. In the upscaling part, it is just a reversed version. After each conv-layer, we implemented a batch-normalization layer, and after each pair, there exists a dropout as large as 0.25. The activation functions used in these blocks are Leaky ReLU, and the last layer has a Tanh activation function. We also have to mention that the kernel sizes are 3×3, padding is set to be the same, and stride is 1.

### 3.3. Loss Function

The loss function also consists of three parts. The first one in our case, is a point-wise depth loss, or L1-loss. The second part is a depth smoothness loss and the third part (which is more important) is the Structural similarity index (SSIM).

### 3.4. Training and Results

After defining the model structure and loss function, we trained our model in 50 epochs with a learning rate of 0.0001 and a batch size of 32. We could reach a validation loss of 0.1949 and test loss of 0.034 during this training process. In figure 3, a picture from the dataset is shown along the actual depth map and the corresponding predicted map.

### 3.5. Discussion

As it could be seen from the results, we still have not reached a good prediction. Our model completely relies on the amount of light and shadows. Also, the edges of each part are not clear enough. Actually, we somehow expected this model not to perform very well, because we did not use complicated structures (such as residual blocks), or transfer-learning (using pre-trained networks and fine-tuning them).

### 3.6. Future Works

The majority of our efforts have so far been devoted to learning more about deep learning and CNNs. We could implement a baseline model and train it to predict depth maps of monocular images. In the future, we will work on improving our model's accuracy and performance. The literature review process is underway, and we are learning how to reconstruct the results of newer papers. We have faced many challenges, including our limited skills in implementing complicated neural networks and the amount of computing resources needed to train a network with millions of parameters. However, we are planning to find a better model with a superior performance that could be implemented with a reasonable resource (by Nov 10th), and then reconstruct and test it on a variety of images and compare it with other models (by Nov 30th). We will also prepare a presentation for the due dates.

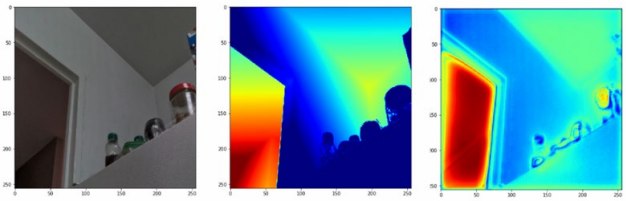

Figure 3: (left) sample image (middle) ground truth depth map (right) predicted depth map

At the end, we would like to mention a brief literature review we have carried out on unsupervised learning methods, that we hope we will have enough time to implement some of them in our project:

The issue of collecting ground truth labels prevents supervised learning methods from being used in the real world. To address this issue, research has focused on unsupervised monocular depth estimation. There are lots of works done on implementing unsupervised learning method for MDE and here we just refer to some of the most recent one that can be extended for our project purpose.

One of the studies done by [13] proposes a convolutional neural network for depth map estimation that uses high performance pre-trained networks for feature extraction. This algorithm with standard encoder-decoder models and transfer learning can achieve high quality depth estimation with fewer parameters and training time. In a work done by Cesser et. al [14] the authors propose a novel architecture in which object motion is used to model dynamic scenes, and refinement can be applied online to refine the learning strategy. Video itself is the only source of supervision in this algorithm. A novel framework, DynaDepth, is developed by Zhang [15] to combine IMU dynamical with visual information for depth and ego-motion estimation. The authors propose an IMU photometric loss and a cross-sensor photometric consistency loss. In addition, under unsupervised learning, ego-motion uncertainty is measured using a camera-centric EKF framework. By integrating IMU dynamics, DynaDepth overcomes scale ambiguity, a common challenge for unsupervised learning methods, while attaining state-of-the-art accuracy even when only lightweight networks are used.

## References


[1] A. Masoumian, H. A. Rashwan, J. Cristiano, M. S. Asif, and D. Puig, “Monocular Depth Estimation Using Deep Learning: A Review,” *Sensors*, vol. 22, no. 14, pp. 1–24, 2022, doi: 10.3390/s22145353.

[2] X. Dong, M. A. Garratt, S. G. Anavatti, and H. A. Abbass, “Towards Real-Time Monocular Depth Estimation for Robotics: A Survey[-5pt],” *IEEE Trans. Intell. Transp. Syst.*, pp. 1–22, 2022, doi: 10.1109/TITS.2022.3160741.

[3] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research,” *Int. J. Rob. Res.*, no. October, pp. 1–6, 2013.

[4] A. Bhoi, “Monocular Depth Estimation: A Survey,” 2019, [Online]. Available: http://arxiv.org/abs/1901.09402.

[5] A. Mertan, D. J. Duff, and G. Unal, “Single image depth estimation: An overview,” *Digit. Signal Process. A Rev. J.*, vol. 123, pp. 1–18, 2022, doi: 10.1016/j.dsp.2022.103441.

[6] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from RGBD images,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7576 LNCS, no. PART 5, pp. 746–760, 2012, doi: 10.1007/978-3-642-33715-4_54.

[7] I. Vasiljevic *et al.*, “DIODE: A Dense Indoor and Outdoor DEpth Dataset,” Aug. 2019, doi: 10.48550/arxiv.1908.00463.

[8] A. Saxena, M. Sun, and A. Ng, “Make3D: Learning 3D Scene Structure from a Single Still Image,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, pp. 824–840, 2009, doi: 10.1109/TPAMI.2008.132.

[9] M. Cordts *et al.*, “The Cityscapes Dataset for Semantic Urban Scene Understanding,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 3213–3223, 2016, doi: 10.1109/CVPR.2016.350.

[10] G. Yang, X. Song, C. Huang, Z. Deng, J. Shi, and B. Zhou, “Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 899–908, 2019, doi: 10.1109/CVPR.2019.00099.

[11] V. Basu, “Monocular depth estimation,” 2021. https://keras.io/examples/vision/depth_estimation/.

[12] W. Weng and X. Zhu, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” *IEEE Access*, vol. 9, pp. 16591–16603, May 2015, doi: 10.48550/arxiv.1505.04597.

[13] I. Alhashim and P. Wonka, “High Quality Monocular Depth Estimation via Transfer Learning,” 2018, [Online]. Available: http://arxiv.org/abs/1812.11941.

[14] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, “Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos,” *33rd AAAI Conf. Artif. Intell. AAAI 2019, 31st Innov. Appl. Artif. Intell. Conf. IAAI 2019 9th AAAI Symp. Educ. Adv. Artif. Intell. EAAI 2019*, pp. 8001–8008, 2019, doi: 10.1609/aaai.v33i01.33018001.

[15] S. Zhang, J. Zhang, and D. Tao, “Towards Scale-Aware, Robust, and Generalizable Unsupervised Monocular Depth Estimation by Integrating IMU Motion Dynamics,” 2022, [Online]. Available: http://arxiv.org/abs/2207.04680.