Monocular Depth Estimation using U-Nets

Atharv Bhat ARB881@NYU.EDU
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012

Valay Shah VALAY.SHAH@NYU.EDU
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012

Mayukh Ghosh MAYUKH.GHOSH@NYU.EDU
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012

Abstract

In this paper, we explore monocular depth prediction from am RGB image using Generative Adversarial Networks (GANs). Estimating depth from a single image is a difficult task and it generally requires looking at multiple perspectives of a scene. Multiple different arrangements of different objects in 3 dimensions may have the same 2d representation which makes it this problem many to one mapping. The paper compares the predictions output by a U-NET trained with different loss functions like ℓ_2 loss, Scale invariant loss proposed by Eigen and Fergus (2015) and a combination of scale invariant loss and adversarial loss. It is seen that the models trained with ℓ_2 and adversarial loss tend to be influenced by RGB pixel values instead of taking the depth into consideration.

1. Introduction

Depth estimation is a extensively studied task in Computer Vision. In particular, the community relied on multi-view methods such as stereo vision and structure-from-motion for depth estimation. However, there may be situations where multiple measurements from the same scene may not be available. This gives us a motivation to develop methods to estimate depth through a single image.

There are several applications where estimating depth from a single RGB image is helpful. It can be used in augmented reality where it can be used to know the dimensions of the product. It can be used in determining trajectory of an object from a single video. The portrait modes in modern smartphone cameras also use application of depth estimation where the image background blurred based on the estimated depth of an object in an image.

Human visual system recognizes visual signals such as size, texture, motion and various other factors known as Depth Cues to estimate depth. Humans also posses stereo vision which allows us to estimate depth of a scene more accurately. The task of estimating depth from a single RGB image is a difficult one. Many three dimensional scenes can have the same two dimensional projection, thus making depth estimation from a single RGB image, a many to one mapping and solving such an under-determined system is a difficult task.

Our work is inspired from Eigen et al. (2014) where the authors present the NYU-Depth dataset to estimate depth from a single RGB input image. In their paper, the authors propose a multi-scale model which first outputs depth predictions on a coarser scale. A local fine-scale model then takes these coarse predictions as an input and then tires to come up with finer predictions. The authors also propose a novel loss function which operates on the output predictions in log scale to counter for the fundamental scale ambiguity that exists in this task.

In this paper, we have used a Unet model proposed by Ronneberger et al. (2015) to estimate the depth in an image in a single forward pass through a single model instead of using two different models that work on two different scales. We have also explored different loss functions such as ℓ_2 loss, scale invariant loss proposed by Eigen and Fergus (2015) as a follow up to their earlier paper and a conditional adversarial loss first proposed by Isola et al. (2018).

2. Methods

For this paper, we followed two approaches. One approach is where we used the UNet architecture with ℓ_2 loss in order to estimate the depths in the images. Then, we used the UNet architecture with the scale invariant loss proposed by Eigen and Fergus (2015). It takes an RGB image and that image is passed to UNet which will estimate the depth and penalize it if it is too different from the Ground truth images.

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

Here D represents the predicted log depths maps and D* represents the ground truth log depth maps. Here, d = D - D* and n is the number of valid pixels. Here $\nabla_x d_i$ and $\nabla_y d_i$ represents the horizontal and vertical differences in the image gradient respectively.

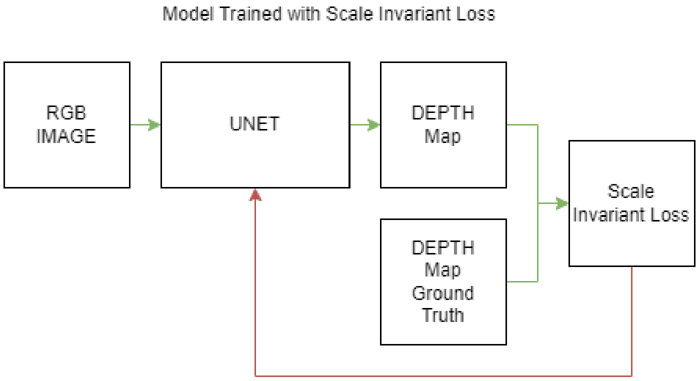

Figure 1: Model architecture for model trained with ℓ_2 or scale invariant loss

The second approach that we used was depth estimation using conditional GAN. Like a traditional GAN, our model has two components; the first one being generator model which would generate the depth map of an image. The second component is a discriminator, a binary classifier, which will classify the depth maps as real or fake. The real depth maps are the ground truth depth maps from the dataset while the ones generated by the generator should be classified as fake. Based on the loss values, it will penalize the generator and the discriminator will keep training till the discriminator cannot accurately classify the depth maps. The generators task is to fool the discriminator into falsely classifying the generated image as real. The discriminator is a fully Convolutional 9 layer neural network. This allows us to use the discriminator for images of any input dimensions. The flowchart shown in Figure 2 describes the training process.

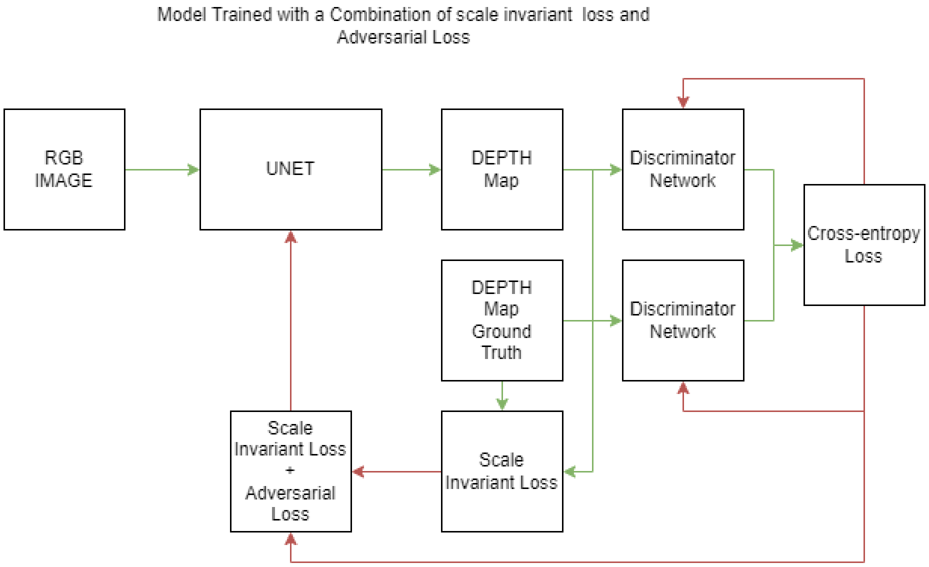

Figure 2: Model architecture for model trained with a combination of scale invariant loss and adversarial loss

In the case of training a model where no adversarial component is involved, we use the ADAM optimizer with a learning rate scheduler with starting learning rate set to $1e^{-3}$ and weight decay of $1e^{-5}$. For training the conditional GAN based method, we use the ADAM optimizer with a fixed learning rate of $1e^{5}$. Data augmentations such as Random Cropping, Horizontal flipping and colour, hue and saturation jitters were used in the training both the methods. When training the model with scale invariant loss, we used weights of the model trained using ℓ_2 loss to initialize the weights of the model. Using the model weights of the model trained on ℓ_2 loss avoids outputting NaN's as the outputs are always positive non zero real numbers as required by the log operation in the scale invariant loss. [1]

[1] The code for this project can be found at : https://github.com/AtharvBhat/EstimateDepth

3. Results

We observed that the models trained with ℓ_2 loss wasn't able to accurately predict the depth of the input image. The output depth values are highly correlated to the pixel intensity values. Brighter objects in the scene always end up being predicted as being far and the model shows complete disregard for the actual scene of the image and focuses too much on the RGB pixel intensities.

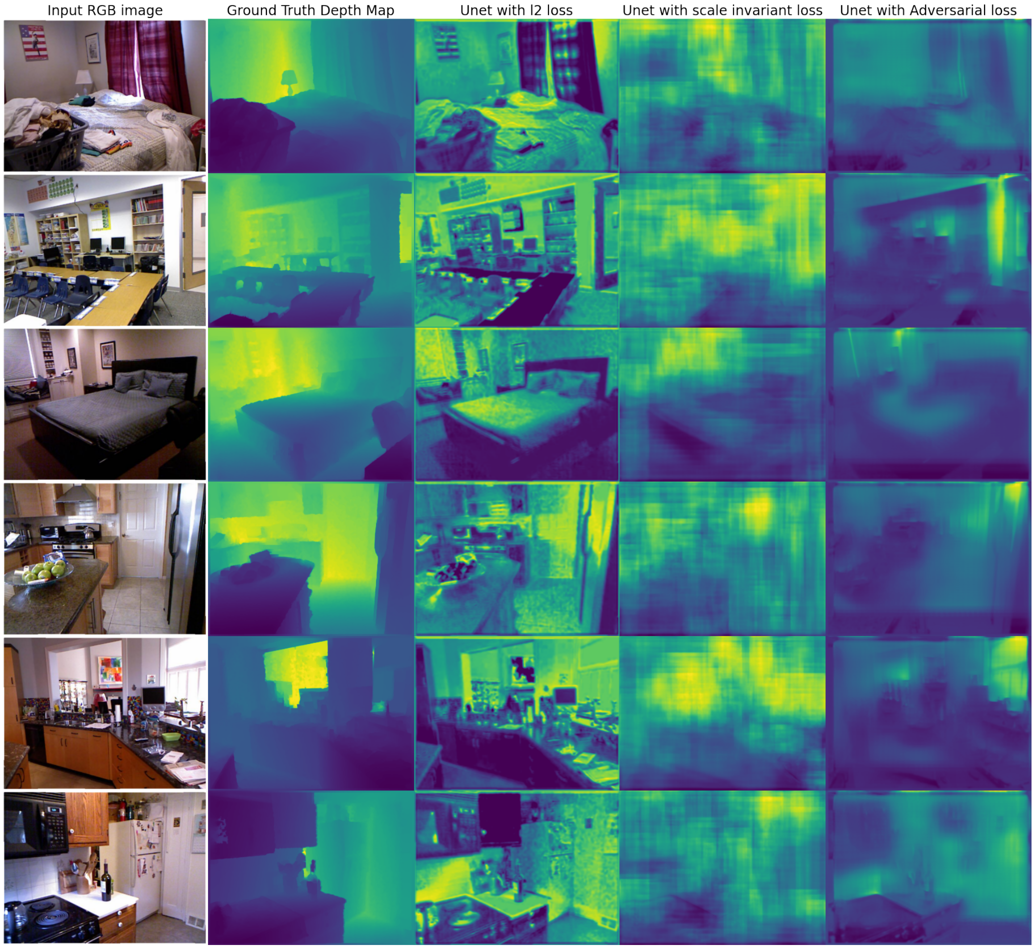

Figure 3: Qualitative Comparison between models trained with ℓ_2 loss, Scale Invariant loss and Adversarial loss. The different columns, from the left, show the RGB image, ground truth, outputs of model trained with ℓ_2 loss, Scale invariant loss and adversarial loss respectively.

As shown in figure 3, we can see that the UNet model trained with scale invariant loss is better than the UNet model trained with ℓ_2 loss at predicting the depth of the scene but the outputs lack detail and are blurry. The depths predicted by the UNet Model trained with a combination of scale invariant and Adversarial loss have a much clearer and less blurry depth map prediction. Unfortunately, when comparing the metrics in Table 1, while the outputs look more plausible and realistic, they are inaccurate.

Metric	Unet(ℓ_2)	Unet(scale inv.)	Unet(Adversarial)
RMSE (linear)	2.0761	1.6827	4.2640
RMSE (log)	0.2648	0.2308	0.5531
RMSE (log, scale inv.)	0.1436	0.1116	0.1354

Table 1: Comparison between Models trained with different loss functions

4. Conclusions

In summary, the Unet model trained with the scale invariant loss performed the best in all the metrics. The model trained with adversarial loss and scale invariant loss was good at producing highly detailed depth maps which seem plausible but are inaccurate.

Given enough time and tweaking of hyperparemeters, model architecture and loss functions, we believe the GAN has the potential to perform better. GAN training is difficult and unstable and we did not have enough time to implement the more state of the art GAN techniques such as Wasserstein GAN or use spectral norm to stabilize the training.

Acknowledgments

We would like to acknowledge support for this project from Professor Rob Fergus for his valuable guidance. We would also like to thank Nan Wu, Nikhil Verma and Yi Max who constantly helped us in this project and showed us the right directions.

References

David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, 2015.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network, 2014.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.