Monocular Depth Estimation

—

Solving single image depth estimation using Image-to-Image translation

Ishan Prakash

Problem Statement

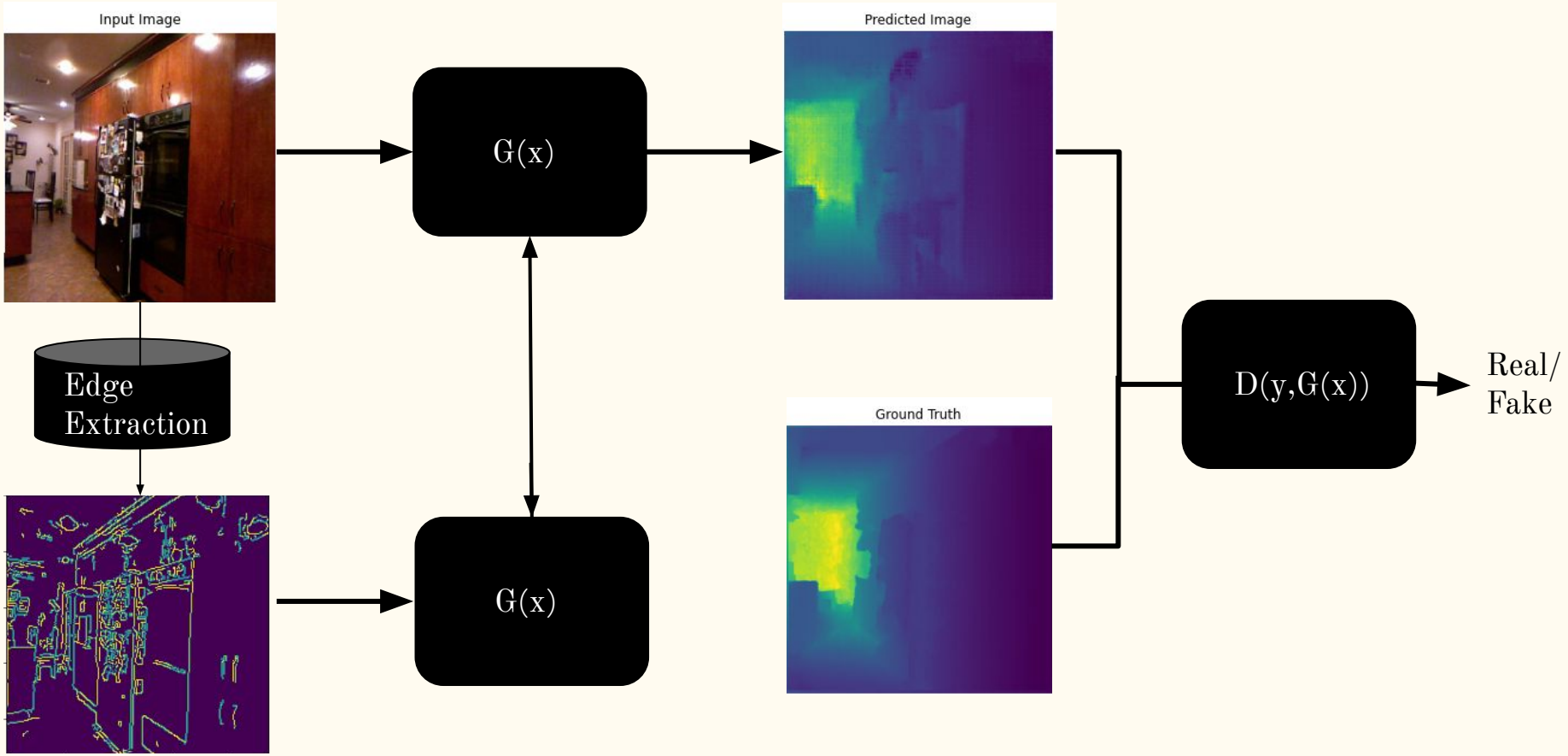
A Deep Learning solution that aims at reducing the cost of Depth estimation in Autonomous systems. LiDAR technology is expensive (₹50,000-₹1,00,000), whereas a single camera (GoPro - ₹10,000) is way cheaper.

Solution:

Monocular Depth Estimation using adversarial machine learning, with emphasis on edge/discontinuity depth estimation.

https://github.com/ishan14/Depth-Estimation

Model Framework



Model Architecture

For our experiment we take:

- **Modified Autoencoder** as Generator. Multiple skip connections are added so that the bottleneck is evaded smoothly, and we are able to describe features at multiple levels.
- The generator is modelled like a **U-Net**, with skip connections from layers(l) to L-l (L-total layers)
- For the discriminator we've used a **PatchGAN**, this penalizes patches that are smaller than the image size and and has fewer parameters, so it can be used in real time.
- This discriminator also models a Markov random field, where pixel outside a certain patch are independent of the patch. This penalty focusses on **texture** and **style**.
- The edge map is derived using a Canny edge detector, it is fed into a similar model as the Input Image. All upsampled layers are **concatenated to the generator** to **preserve edge features**. Several convolution operations are added in the end for the seamless integration of both images.

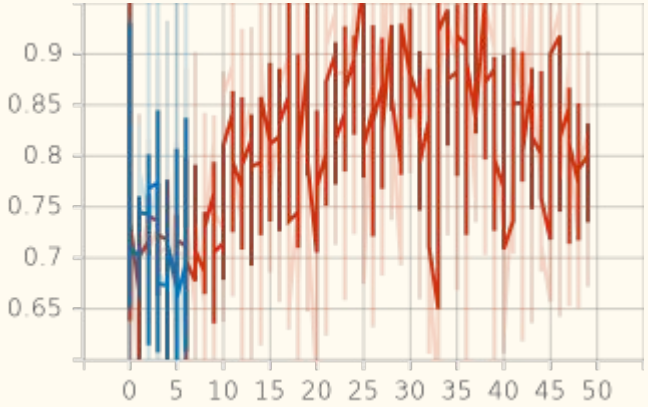
Dataset

We use the NYUv2 data where each sample comprises of a real image and the ground truth map, which is obtained using a LiDAR.

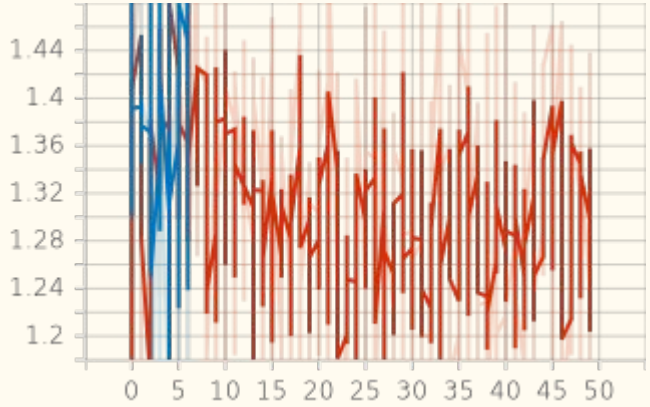
Loss Function

- We are using the conditional GAN loss in which the generator aim to minimize and the discriminator maximizes.
- For the generated output to have structural similarity with the real image we use L1/L2 losses in addition to the Conditional GAN loss.

Results

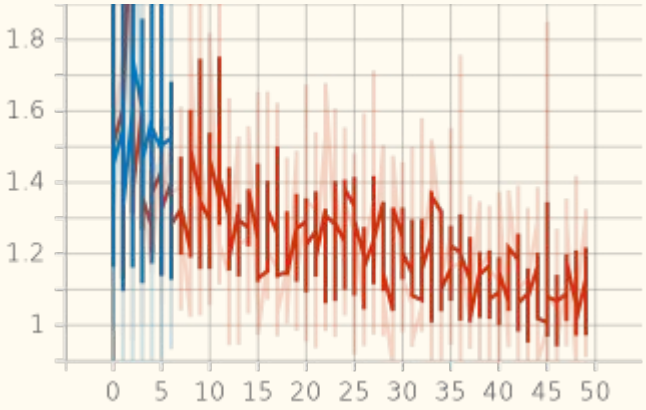


Generator Gan Loss

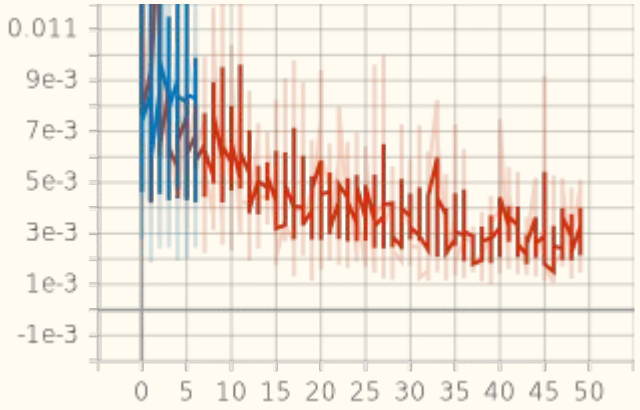


Discriminator Loss

- Here both losses are not settling therefore we can conclude that the model has been trained successfully.
- The discriminator is clearly not able to distinguish between real and fake



Generator total loss

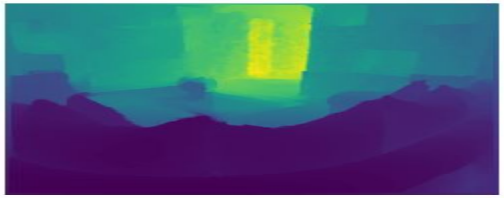


Generator L1 loss

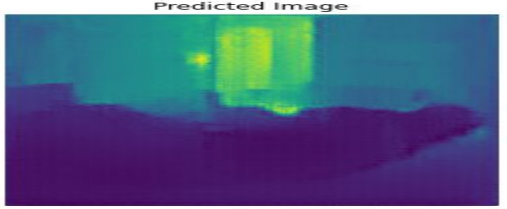
- Both the losses are steadily decreasing therefore model converges effectively.
- Generator L1 loss signifies that the generated output is similar to the the target structurally.

Input Image

Ground Truth

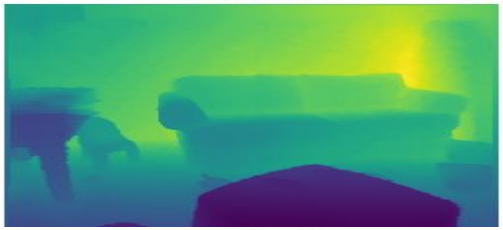


Predicted Image



Input Image

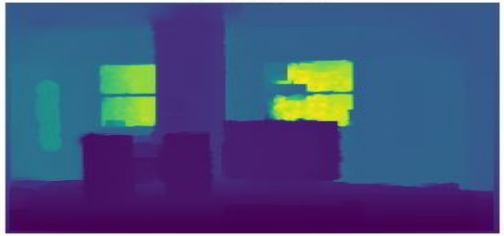
Ground Truth



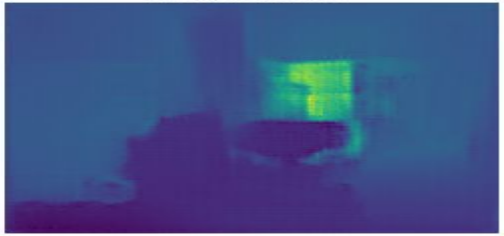
Predicted Image

Input Image

Ground Truth

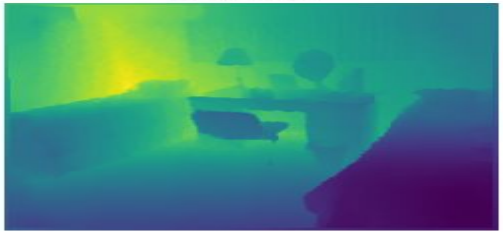


Predicted Image



Input Image

Ground Truth



Predicted Image

Future Work

- Training on the full NYUv2 dataset for better accuracy.
- Implementing **pose estimation** network along with the depth network.
- Supervised Learning relies **extensively** on **labelled data**, but a UGV may experience some **unknown truth**, which the model won't be able to understand. Therefore, relying on methods like **self-supervised learning**, which only need a stream of image data is better.
- Exploring and devising better custom cost functions suited for the given problem statement.

References

- CS Kumar, Arun, Suchendra M. Bhandarkar, and Mukta Prasad. "Monocular depth prediction using generative adversarial networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018.
- Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
- Zheng, Chuanxia, Tat-Jen Cham, and Jianfei Cai. "T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.