Depth Estimation using Single Images

(Project Report for EE610 Image Processing, 2021)

Ankit Chakraborty (204070003), Deepshikha (204070014), A.V.S. Srinidhi (213079003)

Abstract—**Depth estimation from single image is a challenging task with many uncertainties arising from different scales. It is basically figuring out the distance of different objects captured in a scene with respect to a reference (camera). Information obtained from the depth is important for various systems and tasks which help in perceiving environment so that the systems can figure out the state of their own. As the depth is to be estimated from a single image, that means, we do not have many image sequence which restricts the easy estimation. Recently, many methods are being studied and researched throughout the world for the depth estimation from single image by the usage of deep learning techniques. Not only the estimation but also, the accuracy of the estimation of depth is also important and plays a crucial role in significant usage of the methods in real time. For the improvement of accuracy, many types of architectures, training methods or even the loss functions are being rigorously studied. This work proposes a U-Net based architecture which follows a supervised learning technique. Further, the model is modified with different number of layers and also different ways of training the model are implemented to obtain a good and accurate depth estimate. The model is trained over the CityScapes dataset which has high resolution street view images. The proposed model gives depth estimate with very less error and almost accurate maps with reduced complexity in the computations as well as model.**

Index Terms—**Depth estimation, Supervised learning, U-Net, Skip connections.**

I. Introduction

Estimation of depth from a single image is an effortless task for the human eyes but same is not the case for computers. Computers find the depth estimation from a single image a herculean task to do, constrained by the high accuracy and low resource and computations complexity requirements. The task of estimating depth from single (RGB) image is referred to as monocular depth estimation. Depth estimation is one of the important tasks in digital image processing and computer vision that helps to decipher geometric relationship in the scene (image, video, etc). In the field of image processing & computer vision, the depth estimation using two or multiple images has been researched already. From the last decade itself, depth estimation from a single image has become an active area of research. Estimating relative or absolute depth from single image is pretty much ill-posed because of the absence of a second image at the input which restricts triangulation. Estimated depth information from images can be used in a number of applications like object detection, semantic segmentation, navigation, etc. With the rapid development in the field of deep learning, they have shown to be a promising option for giving good performances in image processing tasks like, classification, detection, segmentation, etc.

Above all, the recent developments have shown that the pixel-level depth map can be recovered from a single image possibly in an end-to-end manner based on deep learning techniques. Various Neural Networks like CNNs, RNNs, VAEs, GANs have proved to give promising results while estimating depth from images. Usually, CNN-based models formulate the estimation of depth as a per-pixel regression problem that is further calculated and given by using fully convolutional neural networks. For these networks to learn, the depth sensors are most oftenly used to gather the ground-truth depth values. However, the results obtained from the different models and algorithms vary largely making it arduous to compare the results obtained from them directly. For example, resolution of the images used affects the results obtained largely. Also, some attempts have been made for estimating depth with closely related data. Some attempts were made by combining CNNs with conditional random field models which resulted to more edge conforming maps of the depth. Estimating depth from stereo images is relatively less ambiguous when compared to estimation of depth from single image, which is under-constrained.

In this project work, we study various aspects which are included in the monocular depth estimation. As a result of which, here, a CNN based architecture U-Net which includes multiple encoder layers and decoder layers is implemented to estimate depth from single images. The architectural complexity is made simpler while also, obtain high quality depth maps.

II. Background Work

In the early days, prior to the adoption and extensive usage of CNNs, the problem of monocular depth estimation were mainly based on the hand-crafted features. Several geometry-based methods such as recovering 3D structures from a series of 2D image sequences by using the concept of structure from motion has been applied successfully. Structure from Motion also handled the depth of sparse features. However, it suffered from monocular scale ambiguity. Saxena et al. [3] proposed the usage of global features and also the usage of multi-scale features with Markov Random Field (MRF). In this method, it predicted depth from a set of image features which used linear regression with MRF. Hoiem et al. [5] did not predict the depth explicitly, and instead the image regions were categorized into geometric structures and categorized in different categories. For example. ground, sky, etc.

Ladicky et al. gives the method for the integration of semantic object labels alongwith the monocular depth features so as to improve the performance but it also depends on the handcrafted features. A kNN transfer mechanism is used by

Karsh et al. for the estimation of depth of the backgrounds which are static from the given single images. They further augmented this information with the information of motion that helped in better estimation of the foreground subjects that were in motion and not static in the videos. Driven by the efficient results obtained by the usage of deep learning, the performance of estimation of depth has been hugely improved with embedding of CNNs based methods. These models are generally trained with supervised inputs which are obtained from the depth sensors. Eigen et al. [1] introduced the usage of Convolutional Neural Networks (CNNs) for the task of depth estimation. In this method, two deep network stacks were employed: a coarse global prediction network based on the entire mage, and another one that refines the prediction (made by the global network) locally. This work was further extended and expanded to a three stack architecture which not only performed the depth estimation task efficiently but also performed semantic segmentation and normal estimation in conjunction with the depth estimation.

Roy et al.[6] proposed a model which incorporated shallow convolution neural networks into a regression forest. Further, ResNet architecture was implemented by Laina et al. which was used to perform depth estimation and also alongwith the estimation of depth, it increased the resolution of the depth maps obtained from the model by the introduction of an up-projection module.

Later, the state-of-the-art was given by Fu et al.[9] that proposed a deep ordinal regression network which converts the regression task of depth estimation into a classification task. Further, the depth estimation information was extended and implemented at the super pixel level which was also composed of CNNs for the estimation part and further was refined to enhance the resolution at the pixel level based on the model of convolutional random field.

III. Data and Methodology

Datasets: CityScapes, consists images of street view which are of high resolution. This dataset is used here for the depth estimation task. In order to make the training faster, we used the pre-processed images of size 128 × 256 for all the images used in training and validation. For the pre-processing of the images in the dataset, the images were first cropped on different scales like 1.0, 1.2 and 1.5 and then followed by interpolation by considerng both height and width of the images. The images are RGB therefore results to 3 channels for each image for R, G and B individually.

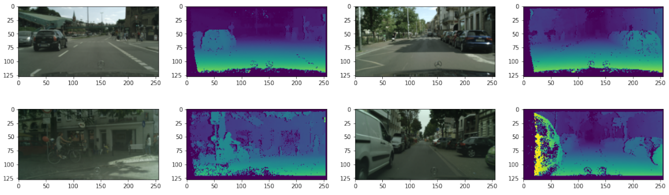

Fig. 1. Images from the CityScapes dataset with their ground truth depth map.

Methodology: The work is done by implementing the model using TensorFlow. The model architecture is shown in the Fig. 2 below.

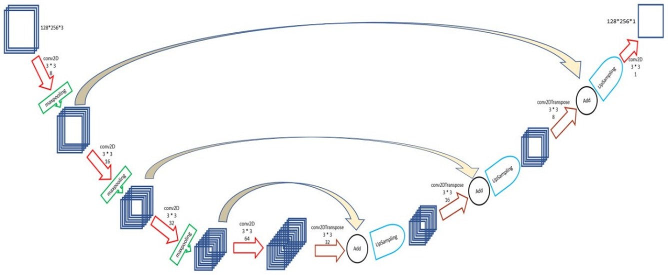

Fig. 2. Proposed U-Net model architecture.

The encoder has 3 layers, the decoder has 3 layers alongwith a bottleneck. Here, adam optimizer is used. The batch size is set to 32. The total number of trainable parameters for the whole network is nearly 50k parameters. Training is performed for 10 iterations for the given CityScapes dataset which approximately took more than half an hour to train fully with 2975 train data.

Several metrics are used to evaluate the performance of our proposed model. They are: mse, logcosh, accuracy.
The metric mse used to evaluate the model performance is shown by equation (1) below.

$$\text{rmse} = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2}\text{.....................(1)}$$

Here, n is the total number of pixels available for which there exists a valid ground-truth depth and also, a valid predicted truth with xi as the predicted depth and yi as the ground-truth depth.

Another metric used to evaluate the model performance is shown below in equation (2)

$$\text{logcosh = log((exp(x) + exp(-x))/2) ....(2)}$$

Here, x is the error evaluated by using (y$_{pred} - y_{true})$

IV. Experiments and Results

Firstly, we made the UNet model architecture with 3 layers in the encoder and 3 layers in the decoder with a layer in the bottleneck. We trained this model while keeping the kernel size same for all the layers. For this architecture, we have a total no. of 49073 trainable parameters. It is optimized by using the adam optimizer with mean square error loss and the metrics as accuracy. Training is performed for 10 iterations. The depth map obtained from model 1 UNet with 3 encoder layers and 3 decoder layers is shown below in Fig. 2.

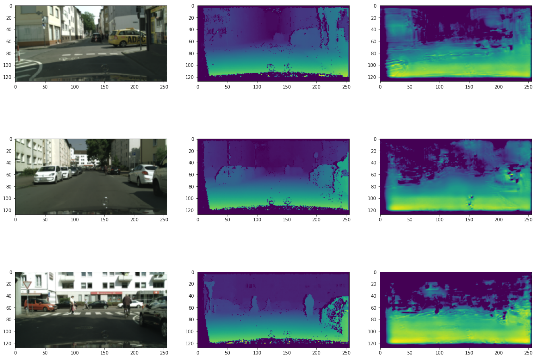

Fig. 3. Depth map obtained from model 1.

Further, we modified model 1 by adding a few layers to it which results to model 2. Model 2 consists of 4 layers in encoder and 4 layers in decoder with same kernel size in all the layers. Total number of trainable parameters here are 196,977. The output obtained from model 2 can be seen in Fig. 3 below.

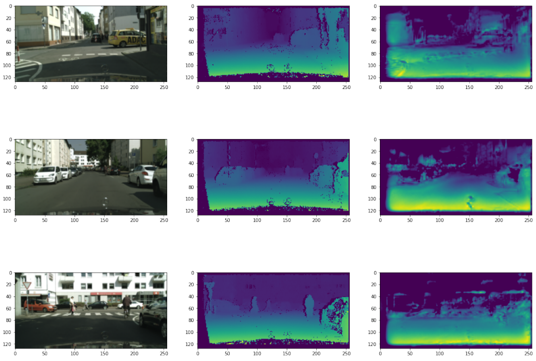

Fig. 4. Depth map obtained from model 2.

In model 3, the architecture is same as that of model 1 and it is trained with adam optimizer with mse loss but the metrics used here is logCosh which resulted in a different depth map which can be seen in Fig. 4 below. Here, the total number of trainable parameters is 49,073. The training is done for 5 iterations.

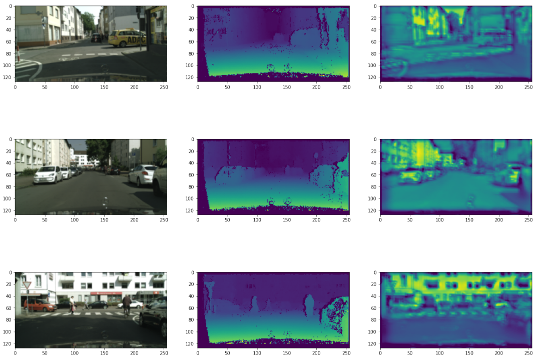

Fig. 5. Depth map obtained from model 3.

In model 4, the architecture is same as that of model 1 and it is trained with adam optimizer with mse loss but the metrics used here is logCosh which resulted in a different depth map which can be seen in Fig. 5 below. Here, the total number of trainable parameters is 49,073. The training is done for 20 iterations.

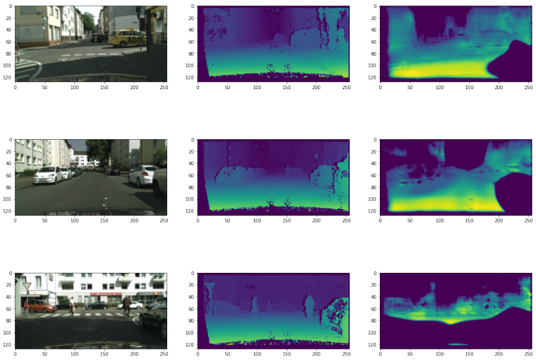

Fig. 6. Depth map obtained from model 4.

In model 5, the architecture has 3 encoder and 3 decoder layers with no batcj normalization and no dropout layer and it is trained with adam optimizer with mse loss but the metrics used here is logCosh. Here, the total number of trainable parameters is 48,849. The training is done for 10 iterations. The depth map can be seen in Fig. 6 below.

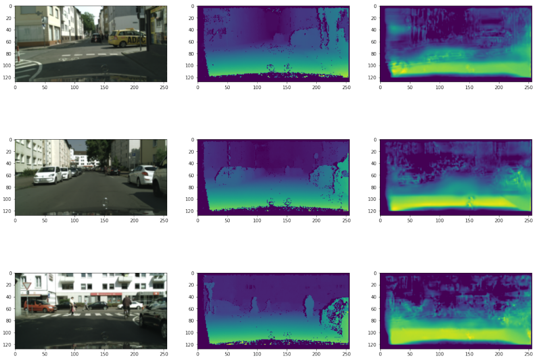

Fig. 7. Depth map obtained from model 5.

Model number	Depth error value	Metric value
Model 1	0.0074	0.1803
Model 2	0.0079	0.1803
Model 3	0.0571	0.0278
Model 4	0.0092	0.0046
Model 5	0.0034	0.0017

TABLE I
PERFORMANCE OF DIFFERENT MODELS ON THE CITYSCAPES DATASET

From the above Table I, we see that the lowest value of mse obtained is 0.0034 and the lowest value of logcosh is 0.0017. Here, the depth error value is given by mean square error and metric value is given by logcosh and accuracy. Accuracy here, is found out by taking the max of the two ratios found out with the predicted depth and true depth values. One ratio is taken as predicted depth by true depth and the other ratio is taken as true depth by predicted depth.

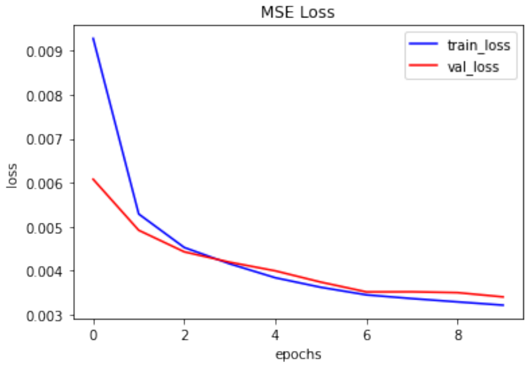

Fig. 8. Performance based on MSE loss.

The above Fig. 8 shows the variation of MSE with the increasing number of epochs. We can see that with the increasing number of epochs, the value of mse is decreasing and also, the model is a good fit. The mse obtained is as low as 0.0034. It should be as low as possible.

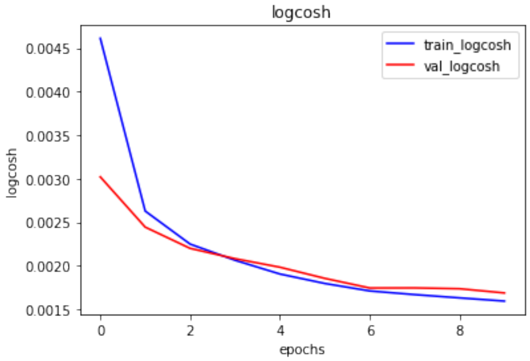

Fig. 9. Performance based on logCosh.

From the above figure 9, we see that the metric logcosh's value is continuously decreasing. It is ideally 0 for a good performance and here we see that it is tending to 0.

Finally, we find that the model with 4 encoder layers and 4 decoder layers i.e., 1 layer added to the model architecture shown in Fig. 2 gives the best depth map which is shown in Fig. 4.

V. Conclusions

In this work, a U-Net based architecture for depth estimation from monocular images is designed. Also, the initial model is further modified by changing the number of layers, metrics, removing batch normalization or even dropout layers. Models are also trained for different number of iterations (or epochs) and the observations are recorded. The experimental results evaluate the performance of the models based on depth error value (MSE, MAE) and metric value (accuracy, logCosh). The base model designed shows a sharp enhancement for the depth estimation task.

Future Work: This work can further be taken to several directions for improvement of resolution of the depth map obtained by incorporating different modules to the presented model architecture. Also, the task complexity can further be decreased by making architectures resulting to very less number of parameters with really good results.

Contribution of team members

Project idea was proposed by Deepshikha. Final problem statement was defined and paper reading for the background of work was done by Deepshikha and Ankit. Data Collection by Deepshikha and Srinidhi. Coding by Ankit and Srinidhi. Experimentation,report writing and video making by Deepshikha and Ankit.

Acknowledgment

We would like to thank Prof. Amit Sethi, Department of EE (IIT Bombay) for giving us the opportunity to work on this interesting project topic.

References

[1] Eigen, David, Christian Puhrsch, and Rob Fergus. ”Depth map prediction from a single image using a multi-scale deep network.” arXiv preprint https://arxiv.org/abs/1406.2283v1.

[2] Zhao, C., Sun, Q., Zhang, C. et al. Monocular depth estimation based on deep learning: An overview. Sci. China Technol. Sci. 63, 1612–1627 (2020). https://doi.org/10.1007/s11431-020-1582-8.

[3] Saxena, Ashutosh, Sung H. Chung, and Andrew Y. Ng. ”3-d depth reconstruction from a single still image.” International journal of computer vision 76.1 (2008): 53-69. https://doi.org/10.1007/s11263-007-0071-y

[4] C. Godard, O. Mac Aodha, and G. J. Brostow, “Unsupervised monocu-lar depth estimation with left-right consistency,” inProceedings of theIEEE Conference on Computer Vision and Pattern Recognition, 2017,pp. 270–279. https://arxiv.org/abs/1609.03677v3

[5] S. Ghosh, N. Das, I. Das, and U. Maulik, “Understanding deep learningtechniques for image segmentation,”ACM Computing Surveys (CSUR),vol. 52, no. 4, pp. 1–35, 2019. https://doi.org/10.1145/3329784

[6] Roy, Anirban, and Sinisa Todorovic. ”Monocular depth estimation using neural regression forest.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[7] Dataset: Pre-processed dataset of CityScapes used by Shikun Liu for the work of MUlti Task Attention Network paper.

[8] Y. Cao, Z. Wu and C. Shen, ”Estimating Depth From Monocular Images as Classification Using Deep Fully Convolutional Residual Networks,” in IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 11, pp. 3174-3182, Nov. 2018,

https://doi.org/10.1109/TCSVT.2017.2740321.

[9] Fu H, Gong M, Wang C, Batmanghelich K, Tao D. Deep Ordinal Regression Network for Monocular Depth Estimation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2018;2018:2002-2011. doi:10.1109/CVPR.2018.00214.

[10] J. Lee, M. Heo, K. Kim and C. Kim, ”Single-Image Depth Estimation Based on Fourier Domain Analysis,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 330-339, doi: https://doi.org/10.1109/CVPR.2018.00042

[11] Shi, J., Sun, Y., Bai, S. et al. A self-supervised method of single-image depth estimation by feeding forward information using max-pooling layers. Vis Comput 37, 815–829 (2021). https://doi.org/10.1007/s00371-020-01832-6