CMSC 733 Final Project

Sharmitha Ganesan (UID : 117518931), Madhu Narra Chittibabu (UID : 118206196), Pradnya Mundargi(UID : 117324570)
sganesa3@umd.edu, mnarrach@umd.edu,pradnya@umd.edu
University of Maryland
College Park

IMPLEMENTATION OF AN EFFICIENT HIGH RESOLUTION MONOCULAR DEPTH ESTIMATION APPROACH

1 Introduction

Monocular depth estimation is the task of estimating the depth of objects in an image using only a single camera as a sensor. This is a challenging problem because a single image does not contain enough information to accurately determine the depth of objects in the scene. However, it has attracted the attention of many owing to its low hardware dependency. There are several methods that can be used to perform monocular depth estimation. One approach is to use geometric methods to estimate depth from the perspective and relative sizes of objects in the scene. For example, if we know the size of an object and its distance from the camera, we can use simple geometry to estimate its depth. Another method used is deep learning algorithms to learn to predict depth from a set of training images. There are several advantages to using deep learning for monocular depth estimation. Deep learning algorithms are able to learn complex relationships between the visual features of an image and the corresponding depth values, and can handle a wide variety of different types of scenes and object appearances. They also do not require explicit hand-designed features, which can make them more robust and efficient.

Monodepth2 [1], MIDAS (Monocular Image Depth Sensing)[2] and LeRes (Least Squares Regression)[3] are few among many deep learning models that have achieved good results on a variety of benchmarks. However, the practicality of such monocular depth estimation models is hindered since they produce low resolution outputs, resulting in many missing details. The authors in [4] have attempted to fill this gap using a content-adaptive multi-resolution merging approach that combines the output of a depth estimation model at multiple scales to produce a high-resolution depth map. The method is designed to take advantage of the strengths of the depth estimation model at different scales, and can be used to improve the accuracy and efficiency of monocular depth estimation on high-resolution images.

For our project, we attempted to test this model using various cases such as images that are blurry, having large noise or varying intensity of lighting. As will be explained in detail in the methods and results section, these edge cases were improved using methods explored in [5] which explores improving the accuracy of depth maps by integrating depth estimation with panoptic segmentation.

2 Related Work

The initial works in the field of depth estimation, hand-crafted features were used to encode image depth cues such as object size and texture[7], vanishing points[8], as well as focus and defocus[9] to name a few. Monocular images essentially do not have reliable stereoscopic visual relationship, therefore, it is an ill-posed problem to regress depth in 3D space [10]. With the advent of large scale datasets, novel neural net architectures, loss functions and incorporation of geometric and semantics

constraints, many attempts were made to explore monocular depth estimation in combination with deep learning.

One of the first attempts to use a RGB image to estimate depth was done by authors in [10], wherein they attempted to use a coarse to fine framework. The coarse portion estimated global depth whereas the fine network attempted to focus on retrieving the details [11]. MiDaS [2], followed a similar approach and architecture to produce one of the best MDE models present. There has also been significant work using conditional random fields (CRFs) to enhance predictions. The authors in [12] combined features from two-stream CNN networks, extracted intermediate layers at different scales and fused them. This method exploited multi-scale estimations derived from CNN inner layers by combining them within a CRF framework. Furthermore, technologies such as lidar were used to produce pointcloud data which was combined with RGB image to help estimate monocular depth maps [13]. With the presence of such robust MDE models, many approaches were proposed as refinement methods for low-resolution depth estimates. These methods included using combination networks with residual training [14], pure upsampling[15] or in accordance with the paper we follow, generation of high-frequency details by altering the resolution of the input images of the network and merging multiple estimations [4]. A novel idea that has been explored by authors in [16, 17] has been the use of segmentation to improve the predictions of depth maps.

For our approach in this project, we have attempted to improve the depth estimation results of existing MDE models. Following the approach in [4], we vary the input resolution to obtain detailed high frequency as well as low frequency results at a high resolution. However some cases such as the presence of fog, transparent objects or nightlight limits the model's performance. In order to bridge this gap, we further follow the methods proposed in [5] to improvise depth estimation results using segmentation.

3 Method

For efficient monocular depth estimation results, Boosted Monocular Depth Estimation is done in [4] where it generates highly detailed high-resolution depth estimations from a single image by merging estimations from pre-trained network in different resolutions and different patches to generate a high-resolution estimate.MIDAS (Monocular Image Depth Sensing) has been chosen as the pre-trained network in this project. Even though the results from [4] are extremely good in most cases, a few cases like blurred images,images of reflecting and refracting surfaces,foggy images, rain effect images and night effect images did not perform well. Mostly the edges were compromised in these circumstances.

[5] had the approach of combining initial depth estimation with panoptic segmentation to preserve the edges in a depth estimation map as shown in Figure 1. Panoptic segmentation [6] is a combination of instance segmentation and object segmentation. In [5], it is assumed that any object in a given image approximately will have the same depth and hence the edges from segmented results are good to tune the depth estimation result.

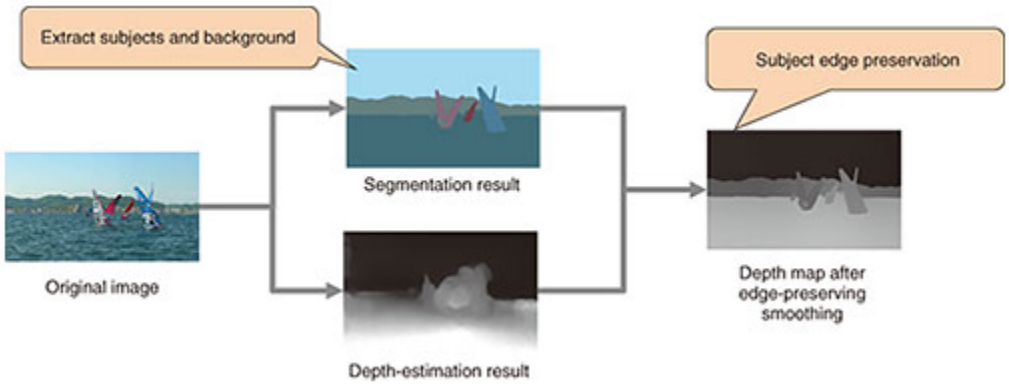

Figure 1: Optimized depth map estimation proposed in [5]

From [6], an efficient panoptic segmentation model pre-trained on coco dataset has been utilized to obtain the required segmentation results for this project. On experimenting with several images, we narrowed down to two approaches to preserve the edges of depth estimation map using panoptic segmentation.

Approach 1

As given in [5], for each segmented region, the corresponding mode of the depth map region will be the intensity of that region. This has worked considerably well for any kind of object in an image. The edges of the objects in the result are defined, however, the background depth variation has been neglected as only the mode of the entire background area is considered. The results for approach 1 in complex cases are discussed in the results section.

Approach 2

In approach 2, a simple bitwise AND operation is performed between the depth estimate and panoptic segmentation that gives out a depth estimate with significant depth variance in the background but in a layered manner. However, in most cases, each object in the foreground undergoes depth variation in layers alongside the background. This has worked considerably well for any kind of background in an image. The results for approach 2 are discussed in the results section.

4 Results

Approach 1 results

A blurred image is taken in Figure 2.

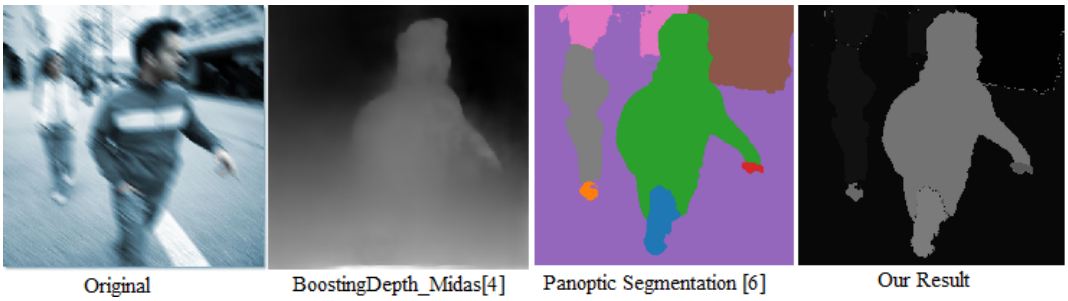

Figure 2: Blurred image

We can observe that the original image in Figure 2 is a person in motion and hence the edges of the person are not very clear hence the approach mentioned in [4] was unable to generate a high-resolution depth map. So we considered this case to boost the depth map using the approach mentioned in [5], in which the segmented image (third image in Figure 2) with clear edge definition has been used to boost the depth map. We can notice in the image (right-most in Figure 2) that the depth map has clear edge definitions for the human in front, also defines the estimate for human in the back who is not even visible with [4] result and the depth of the surrounding area has been assigned to the entire area.

A transparent glass image is taken in Figure 3.

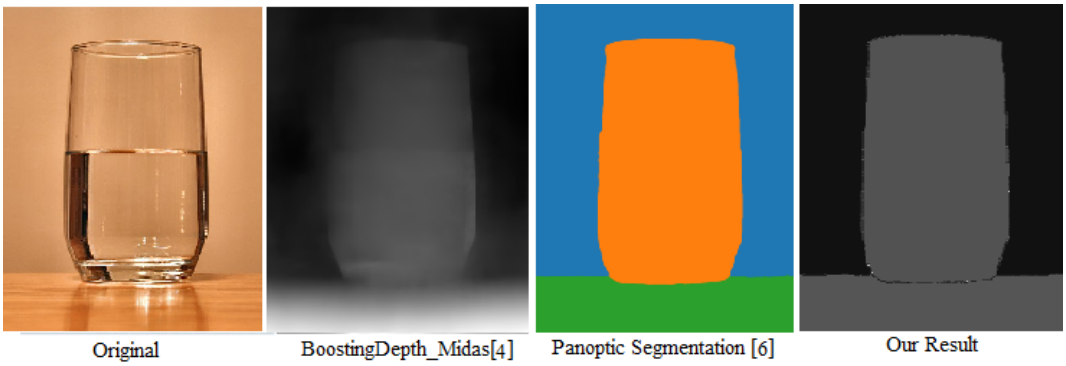

Figure 3: Transparent image

In the case of the image in Figure 3, we tried to generate the boosted MIDAS depth using [4] and we can notice in the second image of Figure 3 which is the result of [4] does not have clear edges for the top part of the glass as the glass and water both are transparent. We leveraged the approach mentioned in [5] to improve the depth map of this image and the generated depth map is the fourth image in Figure 3 which has a clear edge definition and also the depth value.

A glared image is taken in Figure 4.

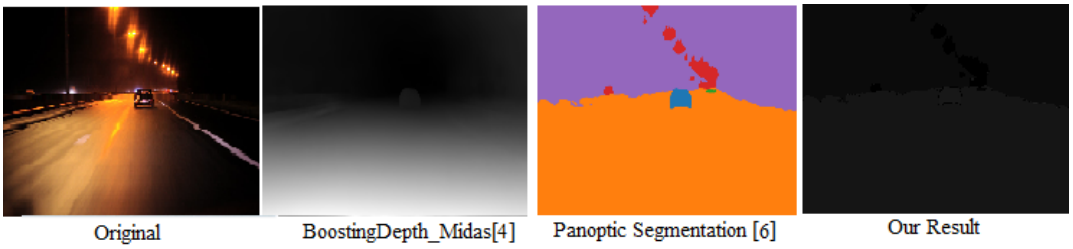

Figure 4: Glare image

One of the drawbacks of [4] is that, in a night light setting, if there are sources of light such as street lamps, headlights of cars, etc. present, then the network performs poorly. We see the output of [4] for a night light setting in the second image of Figure 4. The ground plane definition is lost completely and hence the structural definition is lost. This is because of the receptive field of the network in [4]. Therefore we used the segmented output from [6] to generate a better depth map which is the right-most image in Figure 4.

A rain effect image is taken in Figure 5.

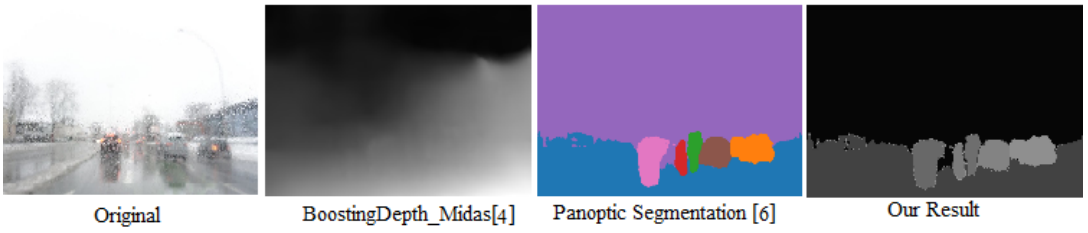

Figure 5: Rain effect image

One more complex case of image is taken where the image is taken from indoor through a glass while raining. It is evident how poorly [4] performs here. With the panoptic segmentation, the depths of cars on the road are estimated.However, the building are not segmented and hence are not estimated by

this approach. This proves that this approach is heavily dependent upon panoptic segmentation results.

A crowd - blurred image is taken in Figure 6.

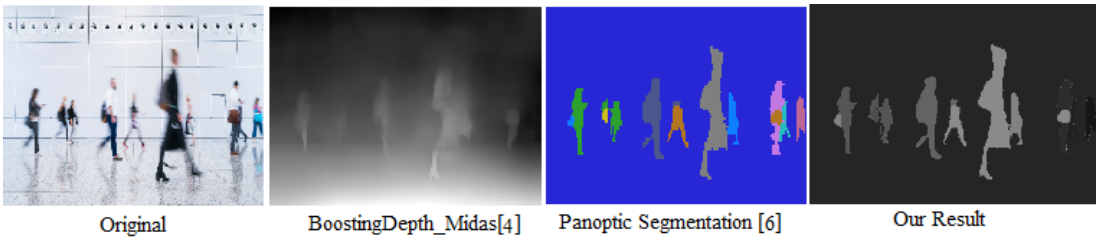

Figure 6: Crowd blurred image

Same as Figure 2, this image also includes people in motion. Approach 1 successfully produced the depth estimate for all the people in the image. However, the depth variance in the floor area has been omitted here.

In conclusion of approach 1 results, it is apparent that foreground objects are estimated accurately whereas background disparity is not considered by the algorithm.

Approach 2 results

A day light image is taken in Figure 7.

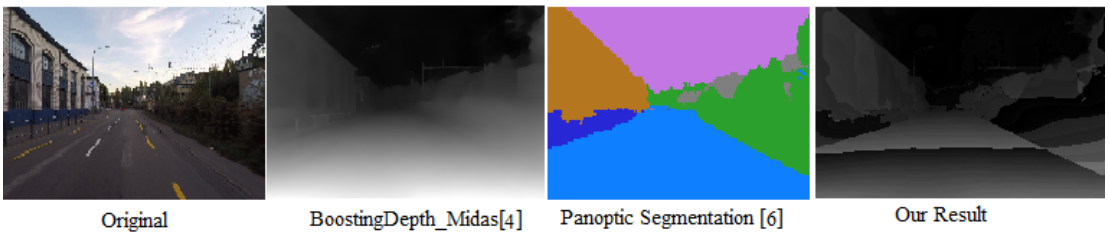

Figure 7: Day Light - Background image

It is discernible that the background i.e road and building in Figure 7 has better depth estimation than the previous approach.The intensity variation along the depth axis are structured in a layered manner which is a good approximation for real-time applications.

A night light image is taken in Figure 8.

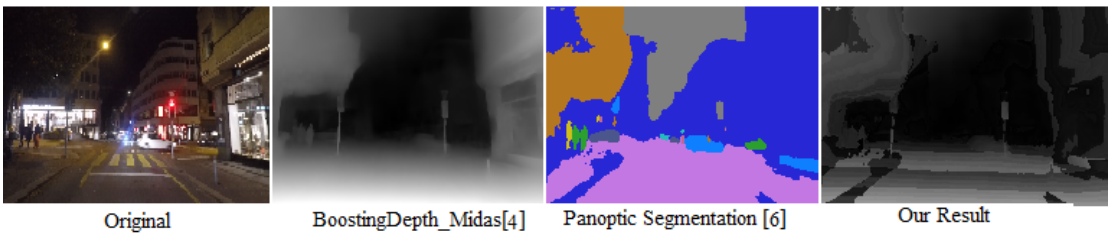

Figure 8: Night Light - Background image

Approach 2 also works well for night light circumstances even with various background entities like road,buildings and shops.

In conclusion of approach 2 results, it is apparent that background objects are estimated better.Even so, the bitwise AND approach messes up the object depth estimate in few images.

5 Conclusion

- This project has achieved considerable results in improving the efficiency of depth estimates for complex image cases with two approaches.
- These approaches can be used in depth estimation applications for autonomous navigation under complex environments.
- A way to preserve edges without ruining the depth estimation of background of the image can be found if the above solutions be combined.
- Adaptive selection of aforementioned approaches for foreground and background can be done as an extension of this project.

References

[1] Ramamonjisoa, M., Firman, M., Watson, J., Lepetit, V. and Turmukhambetov, D., 2021. Single image depth prediction with wavelet decomposition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11089-11098).

[2] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K. and Koltun, V., 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence.

[3] Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S. and Shen, C., 2021. Learning to recover 3d scene shape from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 204-213).

[4] Miangoleh, S.M.H., Dille, S., Mai, L., Paris, S. and Aksoy, Y., 2021. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9685-9694).

[5] Masato Ono et al. ,Improving Depth-map Accuracy by Integrating Depth Estimation with Image Segmentation, NTT Technical Review, 2021, 19-3.

[6] Alexander Kirillov, Qihang Yu, Yuxin Wu COCO 2018 Panoptic Segmentation Task API. https://github.com/cocodataset/panopticapi

[7] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3d: Depth perception from a single still image. In Proc. AAAI, 2008

[8] Y.M. Tsai, Y.L. Chang, L.G. Chen, Block-based vanishing line and vanishing point detection for 3d scene reconstruction, in: 2006 International Symposium on Intelligent Signal Processing and Communications, IEEE, 2006, pp. 586–589.

[9] C. Tang, C. Hou, Z. Song Depth recovery and refinement from a single image using defocus cues J. Mod. Opt., 62 (2015), pp. 441-448

[10] R. Mur-Artal, J.M.M. Montiel, J.D. Tardos Orb-slam: a versatile and accurate monocular slam system IEEE Trans. Robot., 31 (2015), pp. 1147-1163

[11] D. Eigen, C. Puhrsch, R. Fergus Depth map prediction from a single image using a multi-scale deep network Adv. Neural Inf. Process. Syst. (2014), pp. 2366-2374

[12] B. Li, C. Shen, Y. Dai, A. Van Den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015.

[13] Huynh, L., Nguyen, P., Matas, J., Rahtu, E. and Heikkilä, J., 2021. Boosting monocular depth estimation with lightweight 3d point fusion. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 12767-12776).

[14] Yifan Zuo, Yuming Fang, Ping An, Xiwu Shang, and Junnan Yang. Frequency-dependent depth map enhancement via iterative depth-guided affine transformation and intensityguided refinement. IEEE Trans. Multimed., 2020.

[15]Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3D Ken Burns effect from a single image. ACM Trans. Graph., 38(6):1–15, 2019.

[16] Saeedan, F. and Roth, S., 2021. Boosting monocular depth with panoptic segmentation maps. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 3853-3862).

[17] Kwak, D.H. and Lee, S.H., 2020. A novel method for estimating monocular depth using cycle gan and segmentation. Sensors, 20(9), p.2567.