End-to-End Depth Prediction of RGB-D Images

Matsuura Ryuki
5122F083
Waseda University

Smita Priyadarshani
5122FG16
Waseda University

Wang Tianchen
5121FG60
Waseda University

ABSTRACT

In our project we have proposed a deep learning based model to generate dense depth maps for the outdoor scenes using surface normal and occlusion boundary.

The model uses an end-to-end encoder-decoder architecture which is similar to the one used in [4], but with more simplified approach, because of the time constraint that we have. Moreover the authors in [4] have worked of depth completion for outdoor scenes. However, we have worked on depth estimation of RGB images for the outdoor scene. The code is available here: https://github.com/RusCucumber/cgo_groupwork.

ACM Reference Format:
Matsuura Ryuki, Smita Priyadarshani, and Wang Tianchen. 2022. End-to-End Depth Prediction of RGB-D Images. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In depth estimation and prediction distance of each pixel is measured relative to the camera. It has several applications in indoor as well as outdoor environment.Some of the examples are autonomous driving, drones, augmented reality, scene reconstruction, etc.

Traditional depth estimation methods tend to fail sometimes because of the strong lightning interference or lower resolution of distant areas. LiDAR is a reliable solution for this purpose, however they are expensive. Therefore deep learning based solutions are being developed these days which minimizes the loss function or learns to generate a novel view from the sequence to directly estimate the depth.

2 RELATED WORK

Yinda Zhang et al. [4] proposed a model which uses color images to predict surface normal and occlusion boundary for indoor scene. They further used there predictions to solve completed depths.

Jiaxiong Qiu et al. [3] proposed a deep learning based model called DeepLiDAR which uses single color image and sparse depth to produce dense depth for the outdoor scene. It consists of two pathways: color pathway and normal pathway to produce the final output.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 NETWORK

Our model accepts RGB and sparse depth image as input and output dense depth map. The model utilizes an encoder-decoder structure, where the encoder consists of two streams(figure1): (1)A surface normal stream that consumes RGB, sparse depth image, and binary mask to extract the surface normal features. (2)An occlusion boundary stream that consumes RGB image to extract the occlusion boundary features. Upon the output of two feature tensors from the encoder, we simply sum the tensors and then use them as input into the decoder to generate dense depth map.

3.1 Occlusion Boundary Encoder

To obtain encoder which extracts latent features including occlusion boundary information, we first trained SegNet[1]. We used the KITTI segmentation dataset that separates the outdoor images with the same meaningful objects because occlusion boundary dataset of outdoor images were not provided. Then, we only use the encoder part of the model as the occlusion boundary encoder.

3.2 Surface Normal Encoder

Since we do not obtain a dataset with labeled surface normal, we use a pre-trained model from DeepLiDAR[3]. Because DeepLiDAR's surface normal module is trained on a synthetic dataset based on CARLA autonomous driving simulator, we believe that the same pre-training module can be used on KITTI, the open dataset for autonomous driving scenarios. Although this pre-trained model has three encoder-decoder modules, we only use the surface normal module's encoder part because only its latent features are necessary.

4 IMPLEMENTATION DETAILS

4.1 Dataset

As mentioned previously, we train our occlusion boundary encoder on the KITTI segmentation dataset. As for the surface normal encoder, same as DeepLiDAR, we generated binary masks using sparse depth maps to indicate the pixel-wise availability of the LiDAR depth. Next We train the entire model on the KITTI depth completion dataset.

4.2 Experiment Settings

For the suface normal encoder, we use series of ResNet blocks followed by convolution with stride to downsize the feature resolution to 1/16 of the input. For the occlusion boundary encoder, we use 13 convolution layer corresponding to first 13 layers of VGG16.

For the decoder, we use 13 convolution layer corresponding to reversed layers of VGG16.

The optimizer we use for the occlusion boundary encoder and decoder is Adam [2], the loss function is MSE(Mean squared error), and the evaluation metrics is RMSE(Root mean square error). In

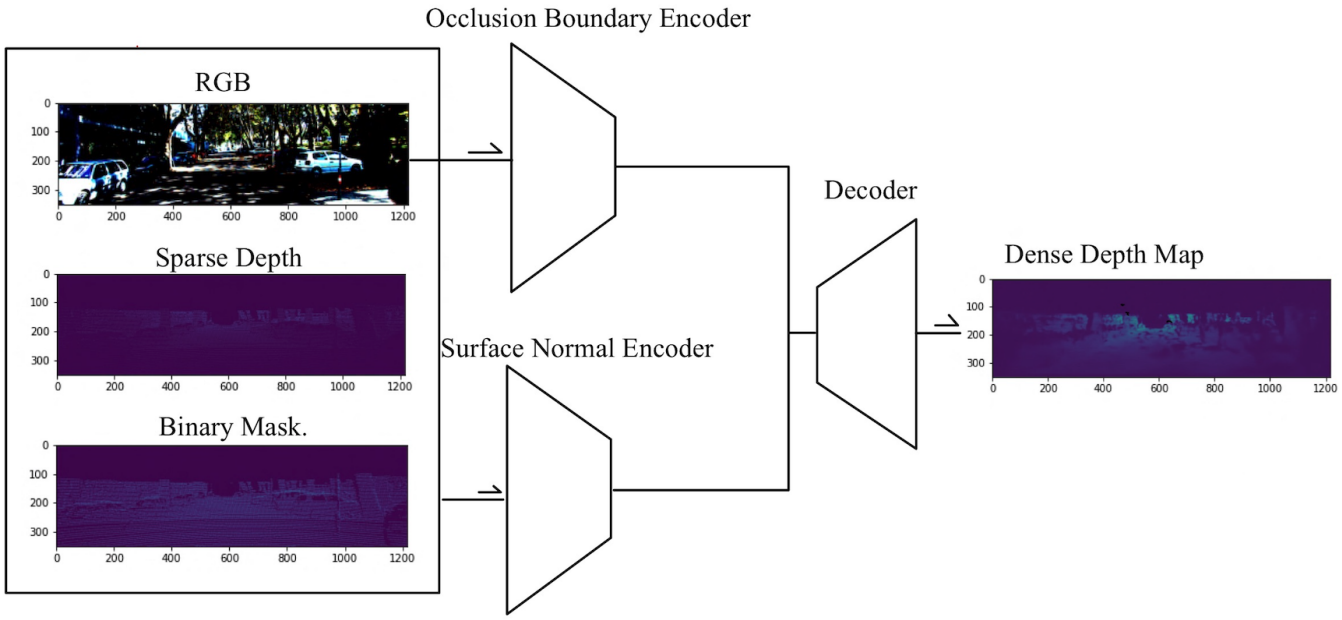

Figure 1: Network overview.

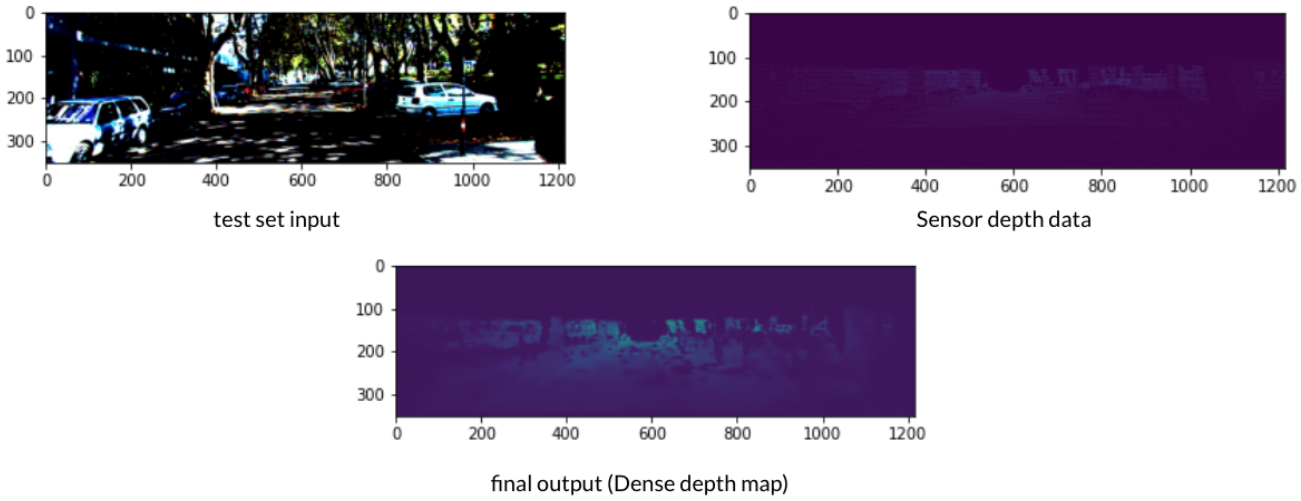

Figure 2: Result.

addition, to avoid overfitting, we also utilize the early stopping training method.

5 RESULT

A sample result on test set is shown in Figure2. After evaluation, the RMSE of our model is 1444.

6 CONCLUSION

There is still a big gap between the performance of our model and that of the state-of-the-art model. We think one of the possible reason is because the fusion method used in our model is relatively coarse, simply summing the tensor of the two encoder outputs. We believe better results will be achieved if we use the encoder-decoder structure on both streams to generate dense depth maps respectively, and then use the attention-based strategy to fuse the depths generated by the two streams.

REFERENCES

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2015. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. https://doi.org/10.48550/ARXIV.1511.00561

[2] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. https://doi.org/10.48550/ARXIV.1412.6980

[3] Jiaxiong Qiu, Zhaopeng Cui, Yinda Zhang, Xingdi Zhang, Shuaicheng Liu, Bing Zeng, and Marc Pollefeys. 2019. DeepLiDAR: Deep Surface Normal Guided Depth Prediction for Outdoor Scene From Sparse LiDAR Data and Single Color Image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3308–3317. https://doi.org/10.1109/CVPR.2019.00343

[4] Yinda Zhang and Thomas Funkhouser. 2018. Deep Depth Completion of a Single RGB-D Image. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 175–185. https://doi.org/10.1109/CVPR.2018.00026