Leveraging Adversarial training for Monocular Depth Estimation

Elnaz Mehrzadeh
Department of Computing Science
Simon Fraser University
elnaz_mehrzadeh@sfu.ca

Parham Yassini
Department of Computing Science
Simon Fraser University
parham_yassini@sfu.ca

Taher Ahmadi
Department of Computing Science
Simon Fraser University
taher_ahmadi@sfu.ca

Dorsa Dadjoo
Department of Computing Science
Simon Fraser University
ddadjoo@sfu.ca

Fatemeh Hasiri
Department of Computing Science
Simon Fraser University
Fhasir@sfu.ca

Abstract

In this project, we tackle the problem of estimating the depth of a scene given a single 2D RGB image. Approaches based on convolutional neural networks (CNNs) have been widely used for depth estimation in recent years. We propose a generative adversarial approach to extend the recent work done by Hu et al. [baseline]. A discriminator network is added to their proposed architecture which is fed with ground truth depth maps as well as the depth maps estimated by the generator network. The discriminator network learns to differentiate between the real depth maps and artificial depth maps that are generated by the generator. We also utilize the structural similarity between the generated depth maps and ground truth as a loss term for depth estimation training. We then evaluate our proposed approach on the NYU v2. dataset [23]. The experimental results show that when adversarial training is used, the performance of the existing method is improved. Also, we report the results of employing different combinations of loss functions in the literature to show how these terms would contribute to the performance.

1 Introduction

Estimating depth from 2D images is an important task in many areas of research such as scene reconstruction, 3D object recognition and semantic and instance segmentation. The monocular depth estimation problem can be defined as: given a single RGB image as input, predict a dense depth for each pixel. Recent approaches have employed Convolutional Neural Networks to address this problem [18, 17, 19]. In the mentioned studies, the problem is addressed as a supervised learning problem and the ground truth depth maps are used for training the networks.

On the other hand, different approaches have been proposed for estimating depth from video frame sequences or stereo images [16]. A downfall of such methods is that they are more costly and resource-demanding than monocular (single image) depth estimation. In [3] Godard et al. developed an unsupervised approach (as it does not require depth map as input during training) for depth estimation using left-right consistency. The authors employ binocular stereo footage with left and

right images of a view in training but only use one view for testing and inference. More recent works in this context [2,20], have employed generative adversarial depth estimation. In such studies, a discriminator network is added to differentiate between fake and real reconstructed left/right images and improve the performance of the image reconstruction phase. In this project, we integrate the same idea with the task of single image depth estimation by using adversarial training on depth map generation.

In this report, we propose a deep neural network solution for the single image depth estimation on top of the recent work [1] and we purpose the following contributions: An updated network architecture that enables adversarial training for generating depth maps. Evaluation of employing adversarial loss on the task of single image depth estimation. Note that this approach is not the same as the ones mentioned earlier as they have used discriminators to distinguish the reconstructed stereo images from the real ones but in our approach, we employed the same idea to distinguish real and artificial depth maps. Evaluation of different training losses that were presented in the literature. We also have employed the structural similarity index (SSIM) to compare similarities between depth maps and examined the potential benefits of applying this metric on depth maps instead of RGB images.

2 Proposed Approach

2.1 Network Structure

Figure 1 shows a diagram of the proposed network architecture. The depth map estimation network is the same network introduced in baseline [1] and a discriminator network was added to enable the adversarial training. Each input image is passed to the encoder module, which extracts multi-scale features. The decoder module is upsampling the features with residual up-convolution networks. The feature fusion module (MFF) fuses the different features at different scales (using the skip connections provided), and the output is concatenated with the features extracted by the decoder. The refinement module consisting of three convolutional layers generates the final depth map prediction. We have utilized the SENet [14] backbone network for implementation of the system, as experimental results of baseline, confirm that better results could be achieved compared to ResNet and DenseNet.

2.2 Adversarial Learning

The last module in our architecture is a neural discriminator network. processing both ground truth and reconstructed depth map, the discriminator tries to classify the former as a real depth map and the latter as an artificial (fake) depth map. In the training phase, the discriminator network gets improve to better distinguish between the real and fake depth maps. At the same time, the adversarial loss from the discriminator module should force the depth map generator to produce outputs that are more realistic. In general, we are facing a problem of generating a random variable with respect to a specific probability distribution. We assume that this training would force the generator to produce depth maps that follow the right distribution (according to the training depth map distributions). We have implemented the same 4-layer discriminator described in the CycleGAN paper [13], for a single channel depth map input. Also, to stabilize the discriminator training we employed the same policy for buffering generated images. The latest of 50 images produced by the generator is buffered and used to train the discriminator. The slope and hyperparameters for discriminator are exactly the same as their implementation.

2.3 Loss Functions

In the baseline method [1], they have suggested three loss functions, which we will first describe each of them briefly and later we will discuss other loss functions that we use in the literature for depth estimation. The total training loss in our work consists of the above-mentioned loss functions, added by the loss from the adversarial training (output of the discriminator module).

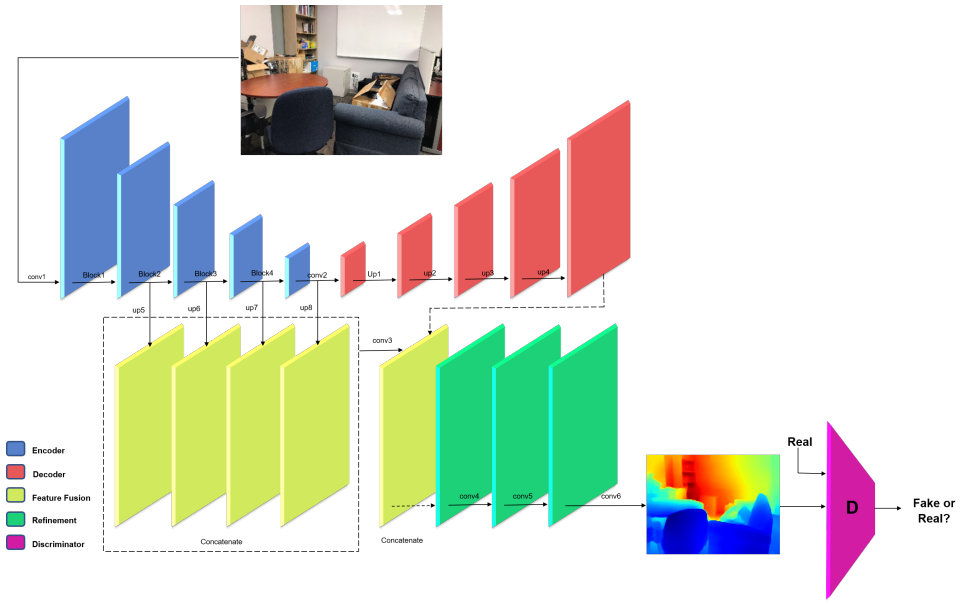

Figure 1: network architecture. by giving an image to the network, the encoder extract multi-scale feature(1/4, 1/8, 1/16, 1/32), and the decoder upscale feature to 1/2. From the decoder we up project the features to 1/2 using MFF and give the result along with the decoder output to refinement module. we give the result of this step with the ground truth to the discriminator and it produce label Fake or Real for the generated depth map.

2.3.1 Depth Loss

Many studies have used l_1 or l_2 norm loss on difference between the ground truth and estimated depth maps. A downfall of such approaches is that a unit of depth difference will have equal contribution in nearby and distant objects. The suggested loss function would penalize error on nearby objects more than error on far objects.

$$l_1 = \frac{1}{n}\sum_{i=1}^{n} F(||d_i - \hat{d}_i||^2)$$

where $F(d) = ln(d + \alpha)$ and $a > 0$ is a hyper-parameter.

The logarithm function would damp any error on nearby or far objects, thus we believe that this function is not the best option for the given intuition. We have employed another loss function introduced by [12] with the same intuition.

$$l_{DBE} = \frac{1}{2n}\sum_{i=1}^{n} g(\tilde{d}_i) - g(d_i)$$

where $g(d) = a_1 d + \frac{a_2}{2} d^2$ and a_1 and a_2 are set to 1.5 and -0.1 respectively.

Where a1 is set to 1.5 and a2 is set to -0.1.

We also proposed a simple function for the same goal to penalize errors in the lower layers which contain more details.

$$l_{depth} = \frac{1}{n}\sum_{i=1}^{n} ||\sqrt{\tilde{d}_i} - \sqrt{d_i}||$$

2.3.2 Gradient Loss

To penalize error around the object edges, the following loss term is considered in the baseline method and our method:

$$l_{grad} = \frac{1}{n}\sum_{i=1}^{n}(F(\nabla x(|d_i - \tilde{d}_i|)) + F(\nabla y(|d_i - \tilde{d}_i|)))$$

where $\tilde{d}_i$ is the ground truth of the pixel.

2.3.3 Surface Normal Loss

The gradient loss can not penalize small structural errors and therefore, the following loss term is introduced by [baseline]. This error term measures the accuracy of the normal to the surface of an estimated depth map with respect to the ground truth depth.

$$l_{normal} = \frac{1}{n}\sum_{i=1}^{n}(1 - \frac{<n_i^d, n_i^g>}{\sqrt{<n_i^d, n_i^d>}, \sqrt{<n_i^g, n_i^g>}})$$

where $< ., . >$ denotes the inner product of vectors, $n_i^d \equiv [-\nabla_x(d_i), -\nabla_y(d_i), 1]^\top$ denotes the surface normal of an estimated depth map and $n_i^g \equiv [-\nabla_x(g_i), -\nabla_y(g_i), 1]^\top$ denotes the surface normal of the ground truth.

2.3.4 Structural Similarity Loss

Inspired by [3], we use the structural similarity index (SSIM) to penalize dissimilarities between the reconstructed depth map and the ground truth. It should be noted that in the previous works that SSIM was employed, they use it to find structural similarity between the reconstructed images and the ground truth image. In this work, we investigate the benefits of using SSIM on finding similarities between two depth maps. We treat depth maps as a single channel image in our case.

$$l_{SSIM} = \frac{1}{n}\alpha\frac{1 - SSIM(d_{ij}, \tilde{d}_{ij})}{2} + (1 - \alpha)||d_{ij} - \tilde{d}_{ij}||$$

Where we set the hyper-parameter α to 0.6 in order to put more weight on the SSIM term rather than the l_1 loss term.

3 Experimental results

3.1 Network Training

For the experiments, we used NVIDIA GeForce GTX TITAN X, with 12 GB Memory and we trained our model on the NYU-Depth V2 dataset. The dataset consists of a variety of indoor scenes from which we used 50K samples for the training phase and 1K samples for testing. The data augmentation is identical to the baseline method: Flip, Rotation and Color Jitter were applied on each pair in the dataset. Due to the time limit, we trained our network only in 5 epochs while in the baseline paper they run their experiments with 20 epochs. In this regard and to be able to compare the results, we also trained the baseline network for 5 epochs. The initial learning rate is 0.0001, and it is reduced to 10% for every epoch after the second iteration.

3.2 Metrics

The problem of estimating depth from a single image, can be regarded as a continuous regression problem [10]. Therefore, mean squared error is widely used in the literature for performance metrics. The relative error and average log error allow a relatively large error on a larger depth values. The definition of each criterion is as follow:

- average relative error(rel): $\frac{1}{n}\sum_p^n \frac{|y_p - \hat{y}_p|}{y}$
- root mean squared error(rms): $\sqrt{\frac{1}{n}\sum_p^n (y_p - \hat{y}_p)^2}$
- average (log10) error: $\frac{1}{n}\sum_p^n |log_{10}(y_p) - log_{10}(\hat{y}_p)|$
- threshold accuracy(δ_i): % of y_p such that $max(\frac{y_p}{\hat{y}_p}, \frac{\hat{y}_p}{y_p}) = \delta < thr$ for $thr = 1.25, 1.25^2, 1.25^3$.

where y_p is a pixel in depth image y, $\hat{y}_p$ is a pixel in the predicted depth image $\hat{y}$, and n is the total number of pixels for each depth image.

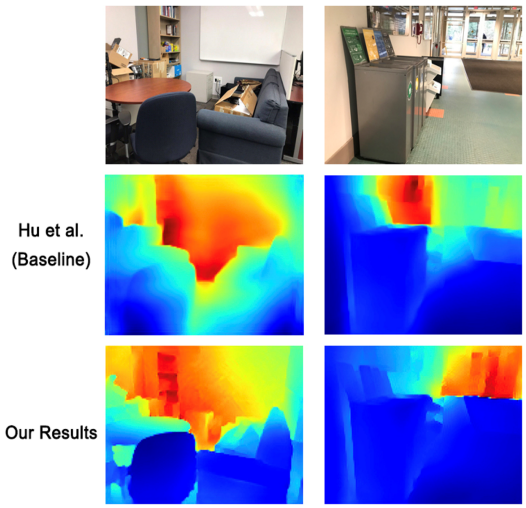

Figure 2: Example comparison of estimated depths for given input images using our method and the Hu et. al [reference] method.

3.3 Performance Compression

Figure 2 shows the result of our method as well as the baseline method on two test samples. Significant improvement in preserving object boundaries and depth details can be observed. In the first image (left), the sofa, chair, and storage cabinet boundaries are distorted in the baseline output while they are completely visible ours. In the second sample (right), the boundaries of the white trash bins are visible in our result and the baseline method has a significant error on the depth of the wall compared to the depth of the doors.

Table 1 compares the quantitative results of our method against state-of-the-art methods on the NYU v2 dataset. As shown in the table, when adversarial loss is added on top of the baseline method, we could get a slightly better result in terms of RMSE and thresholded accuracy metrics. Other rows in the table, demonstrate how different loss functions can impact the performance. When the SSIM loss term is added to the baseline method, the performance is decreased and no improvement is observed in other combinations. This confirms the findings in the literature [15] that SSIM of the depth maps does not correlate very well with the depth map quality, and thus it is not suitable to measure the distortion in depth maps.

Also, our results revealed that using the l_{depth} term proposed by [baseline] can lead to a better performance results compared to the DBE loss function and our proposed depth loss function. Figure 3 shows depth maps estimated by our method and the baseline method as well as the ground truth depth maps.

Finally, we compare the proposed approach with previous methods considering the computational cost in the test phase. It is worth mentioning that in our approach we did not add any computational cost to the test phase of the algorithm. The extended network layers for the discriminator and the loss functions only impact the training computational cost. Figure 4 shows the root mean squared error vs. running time for a single image across different approaches. As demonstrated, our running time is similar to the baseline method while the RMSE metric is improved. Our approach provides a reasonable trade-off between inference time and accuracy.

4 Conclusion

This work has sought to investigate the impact of exploiting adversarial loss and different complementary loss functions on estimating depth from a single image. Our proposed approach achieved better performance results compared to the state-of-the-art baseline method [1]. The experimental results confirm that adversarial training can be beneficial for single image depth estimation. Our

Table 1: Evaluation of depth map estimation for our method and previous methods on NYU v2 dataset. Rows shows result for using different combination of the suggested loss functions.

method	**error**			**accuracy**		
	rms	rel	log10	δ <1.25	δ <1.25^2	δ<1.25^3
Hu et al. (Baseline) [1]	0.530	0.115	0.050	0.866	0.975	0.933
(Ours) Baseline+adversarial	**0.527**	0.117	0.050	0.866	0.974	**0.944**
(Ours) Baseline+adversarial+SSIM	0.537	0.117	0.051	0.860	0.972	0.933
(Ours) Baseline (Depth Priority)+adversarial	0.538	0.121	0.052	0.854	0.971	0.990
(Ours) Baseline (DBE)+adverserial+SSIM	0.544	0.118	0.052	0.858	0.971	0.992
(Ours) Baseline (DBE)+adverserial	0.533	0.123	0.053	0.848	0.968	0.992
(Ours) Baseline+SSIM	0.535	0.117	0.051	0.862	0.974	0.933
(Ours) Baseline (DBE)+SSIM	0.540	0.122	0.053	0.853	0.972	0.993
Eigen et al. [4]	0.907	0.215	-	0.611	0.877	0.971
Xu et al. [5]	0.586	0.121	0.052	0.811	0.954	0.987
Xu et al. [6]	0.593	0.125	0.057	0.806	0.952	0.986
Fu et al. [7]	0.509	0.115	0.051	0.828	0.965	0.992
Qi et al. [8]	0.569	0.128	0.057	0.834	0.960	0.990
Lei et al. [9]	0.821	0.232	0.094	0.621	0.886	0.968

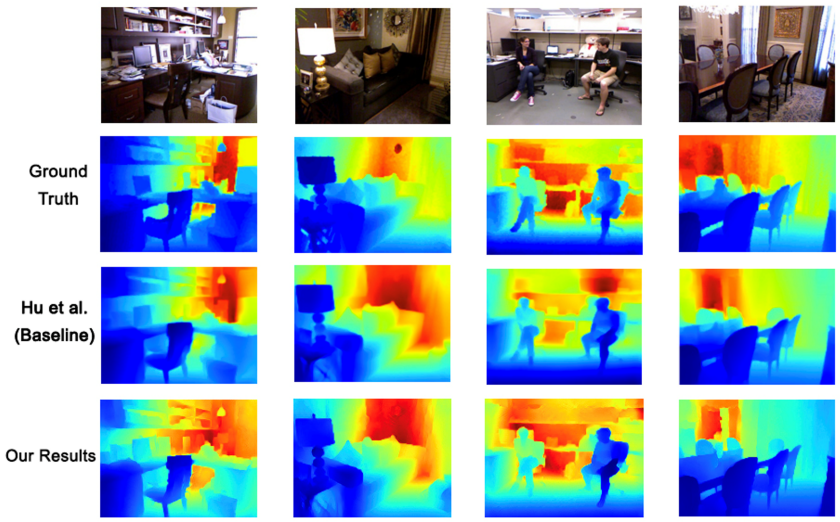

Figure 3: Visual comparison of the result for our purposed method (trained with adversarial loss) and the Hu et. al [reference] method on four test images from the dataset.

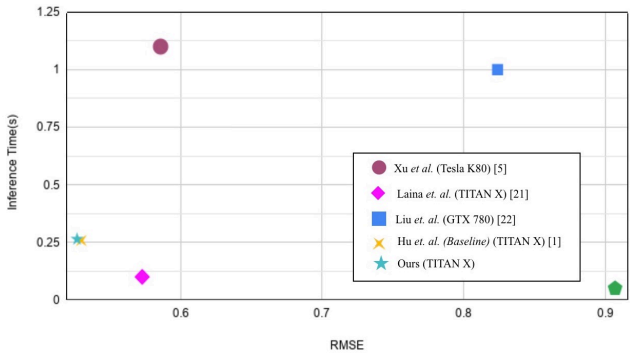

Figure 4: Scatter plot for comparison with previous methods: Inference time for a single frame vs. RMSE

results also reveal that no significant improvement in the depth map accuracy is observed when the SSIM loss is used. We can conclude that when other restricted constraints such as gradient loss on edges and surface normal loss are involved, the SSIM loss term hardly contributes to the quality of the predicted depth map.

5 Member Contributions

Elnaz Mehrzadeh: Reviewed the literature of depth estimation applications. Selected the baseline method and proposed ideas about extending it.developed depth loss. Prepared the report and poster manuscripts. Also, contributed to designing experiments for evaluation of the proposed methods.

Parham Yasini: Reviewed the literature of monocular depth estimation methods and generative adversarial models. Suggested the idea of adding the discriminator network based on the literature review and implemented the proposed approach. Also, participated in preparing the report manuscript and analyzing the experimental results.

Dorsa Dadjoo: Reviewed the literature of monocular depth estimation using video sequences. Proposed the idea of using SSIM loss function. Contributed to the preparation of the poster content and report manuscript.

Taher Ahmadi: Reviewed the literature of unsupervised and semi-supervised depth estimation methods. Proposed a difference of square root of depth-based loss term. Performing training due to different loss terms. Implemented the live demo for the poster session.

Fatemeh Hasiri: Reviewed the literature of single image depth estimation. Proposed the idea of using the DBE loss function and contributed to the implementation of the error terms. Prepared the scatter chart for running time and accuracy comparison.

References

[1] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In IEEE Winter Conf. on Applications of Comp. Vis., 2019.

[2] Groenendijk, Rick & Karaoglu, Sezer & Gevers, T. & Mensink, Thomas. (2019). On the benefit of adversarial training for monocular depth estimation. Computer Vision and Image Understanding. 102848. 10.1016/j.cviu.2019.102848.

[3] Godard, Clement & Aodha, Oisin & Brostow, Gabriel. (2017). Unsupervised Monocular Depth Estimation with Left-Right Consistency. 10.1109/CVPR.2017.699.

[4] D.Eigen,C.Puhrsch,andR.Fergus,DepthMapPrediction from a Single Image using a Multi-Scale Deep Network, NIPS, 2014

[5] D.Xu,E.Ricci,W.Ouyang,X.Wang,andN.Sebe,Monocular depth estimation using multi-scale continuous crfs as sequential deep networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.

[6] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, Structured attention guided convolutional neural elds for monocular depth estimation, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 39173925.

[7] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, Deep ordinal regression network for monocular depth estimation, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 20022011.

[8] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In CVPR, pages 283291, 2018. [9] 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. CVPR, pages 11191127, 2015.

[10] Bhoi, Amlaan. (2019). Monocular Depth Estimation: A Survey.

[11] Zhou, Lingtao, Jiaojiao Fang, and Guizhong Liu. "Unsupervised Video Depth Estimation Based on Ego-motion and Disparity Consensus." arXiv preprint arXiv:1909.01028 (2019).

[12] Lee, Jaehan & Heo, Minhyeok & Kim, Kyung-Rae & Kim, Chang-Su. (2018). Single-Image Depth Estimation Based on Fourier Domain Analysis. 330-339. 10.1109/CVPR.2018.00042.

[13]Zhu, Jun-Yan et al. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. 2017 IEEE International Conference on Computer Vision (ICCV)(2017): 2242-2251.

[14] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 7132-7141.

[15] D. V. S. X. De Silva, W. A. C. Fernando, S. T. Worrall and A. M. Kondoz, "A novel depth map quality metric and its usage in depth map coding," 2011 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), Antalya, 2011, pp. 1-4.

[16] Casser, Vincent & Pirk, Sren & Mahjourian, Reza & Angelova, Anelia. (2019). Depth Prediction without the Sensors: Leveraging Structure for Unsupervised Learning from Monocular Videos. Proceedings of the AAAI Conference on Artificial Intelligence. 33. 8001-8008. 10.1609/aaai.v33i01.33018001.

[17] Xu, Dan et al. Structured Attention Guided Convolutional Neural Fields for Monocular Depth Estimation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018): 3917-3925. https://arxiv.org/pdf/1803.11029.pdf

[18] Liu, Fayao et al. Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 38.10 (2016): 20242039. Crossref. Web. https://arxiv.org/pdf/1502.07411.pdf

[19] D. Eigen and R. Fergus, "Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 2650-2658.

[20] Kumar, Aran & Bhandarkar, Suchendra & Prasad, Mukta. (2018). Monocular Depth Prediction Using Generative Adversarial Networks. 413-4138. 10.1109/CVPRW.2018.00068.

[21] L. Iro, R. Christian, B. Vasileios, T. Federico, and N. Nas- sir. Deeper depth prediction with fully convolutional residual networks. In 3DV, pages 239248, 2016.

[22] F.Liu,C.Shen,andG.Lin.Deepconvolutionalneuralfields for depth estimation from a single image. In CVPR, pages 51625170, 2015.

[23] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In ECCV, 2012.