Leveraging Adversarial training for Monocular Depth Estimation

Parham Yasini, Taher Ahmadi, Elnaz Mehrzadeh, Fatemeh Hasiri, Dorsa Dadjoo

I. Introduction

Estimating depth from 2D images is a crucial step in many applications such as scene reconstruction, 3D object recognition, segmentation, and detection. The problem can be framed as: given a single RGB image as input, predict a depth map for each pixel.
Most Existing methods:
-suffer from loss of spatial resolution in the estimated depth maps
-have distorted and blurry reconstruction of object boundaries

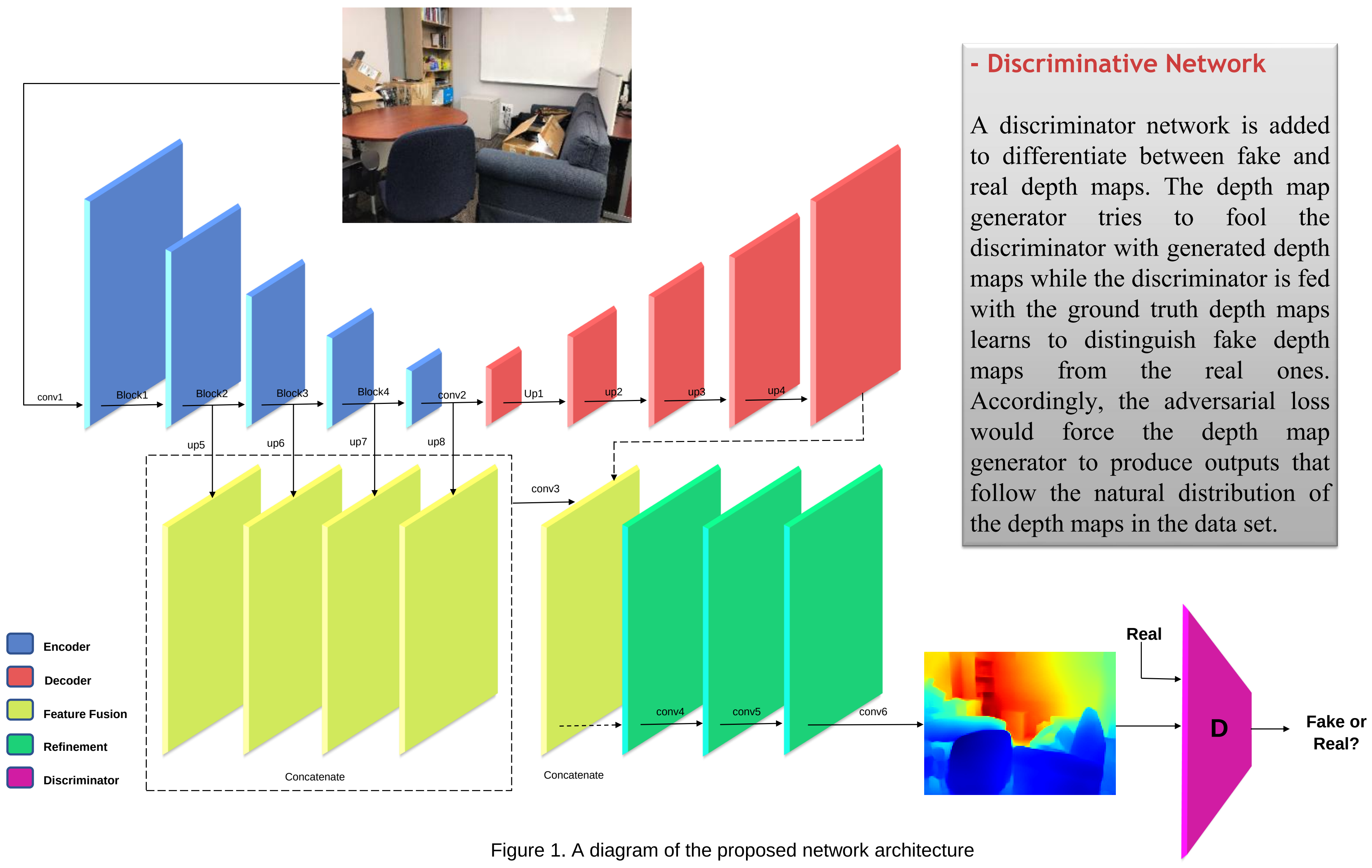
II. Contribution

We extend the baseline method of Hu et al. (2018)[1] by adding a discriminator network and introducing two new loss functions. The discriminator network (D) is trained using ground truth depth maps from the dataset and reconstructed depth maps from the depth estimation module (generator). The discriminator forces the depth estimator to generate depth maps that are more similar to the real depth maps.
Also, we added an error term to compare the structural similarity of the reconstructed depth map and ground truth depth. Another error term is added to penalize the depth error on nearby objects depth more than distant objects.
Our experimental results show the effect of using different combinations of these ideas on the accuracy of the network.

III. Method

- Architecture

Given an input image, the encoder extracts multi-scale features. The decoder converts the last 1/32 scale feature to get a 1/2 scale feature. Each of the multi-scale features is up-scaled to 1/2 scale and fused by the multi-scale feature fusion module (MFF). The outputs of D and MFF and are refined by the refinement module (R) to obtain the final depth map. Each box named "blockn" denotes a block of multiple convolutional layers, such as residual block of ResNet; each box named "upn" denotes a up-projection layer.



- Discriminative Network

A discriminator network is added to differentiate between fake and real depth maps. The depth map generator tries to fool the discriminator with generated depth maps while the discriminator is fed with the ground truth depth maps learns to distinguish fake depth maps from the real ones. Accordingly, the adversarial loss would force the depth map generator to produce outputs that follow the natural distribution of the depth maps in the data set.

Figure 1. A diagram of the proposed network architecture

- Loss Functions

$$L = l_{normal} + l_{grad} + l_{dept\ h} + l_{SSIM}$$

Penalize small structural errors such as those of high frequency undulation of a surface:

$$l_{norma\ l} = \frac{1}{n}\sum_{n=1}^{n}\left(1 - \frac{< n_i^d, n_i^g >}{\sqrt{< n_i^d, n_i^d >}, \sqrt{< n_i^g, n_i^g >}}\right)$$

Penalize error around distorted or blurry edges :

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

It stands to reason that the error function should be perceptually motivated :

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

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

Sensitive to changes in the depth direction but insensitive to changes in x and y direction:

$$l_{DBE} = \frac{1}{2n}\sum_{i=1}^{n}\left(g(\tilde{d}) - g(d)\right)^2 \qquad g(d) = a_1 d + \frac{a_2}{2}d^2$$

IV. Results

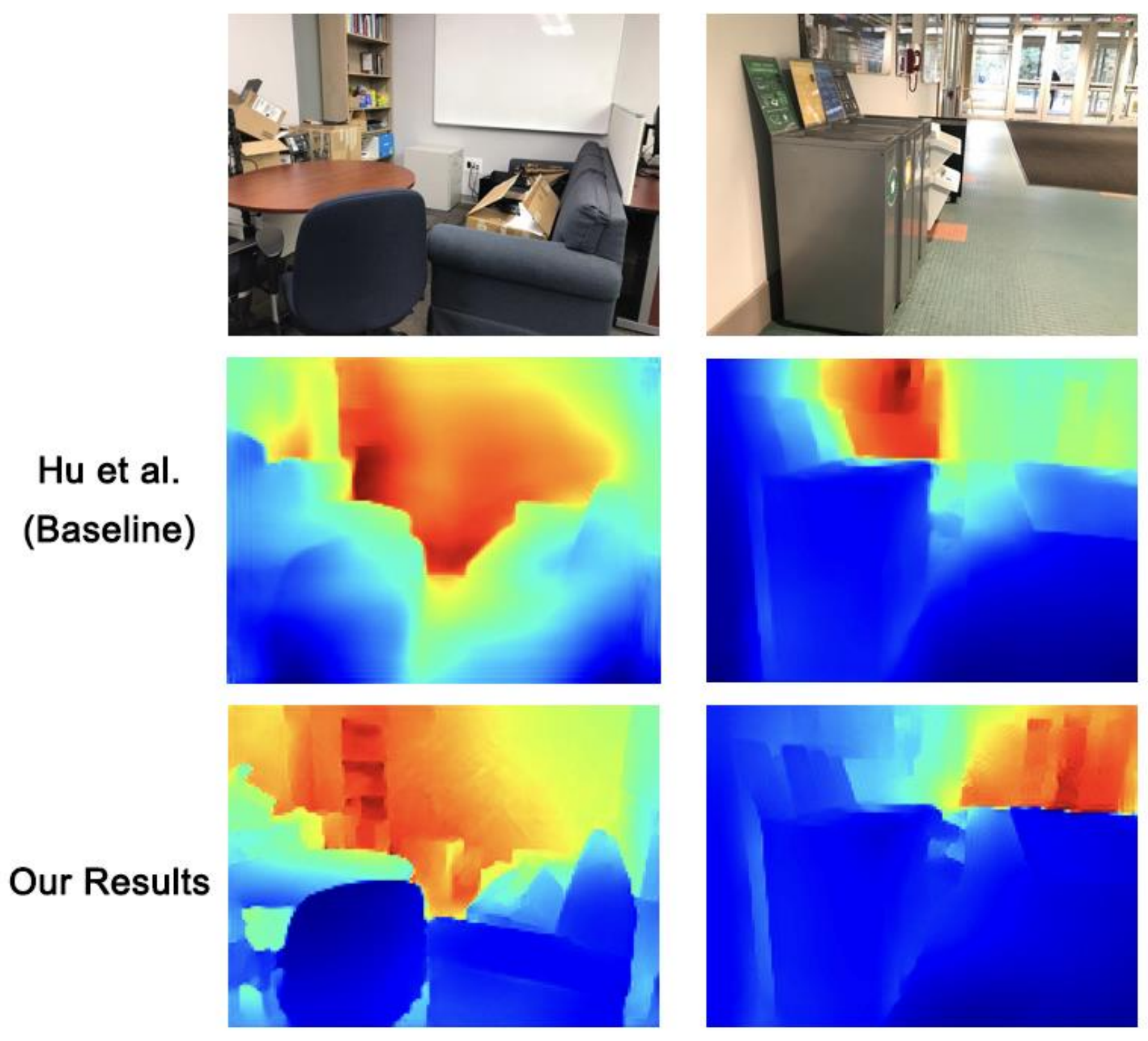


Figure 3. Visual comparison of estimated depth maps

We use 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 5 epochs instead of 20 epochs that the baseline method used. We also, trained the baseline network for 5 epochs to perform a fair comparison between results. The initial learning rate is 0.0001, and reduce it to 10% for every epoch after the second iteration.

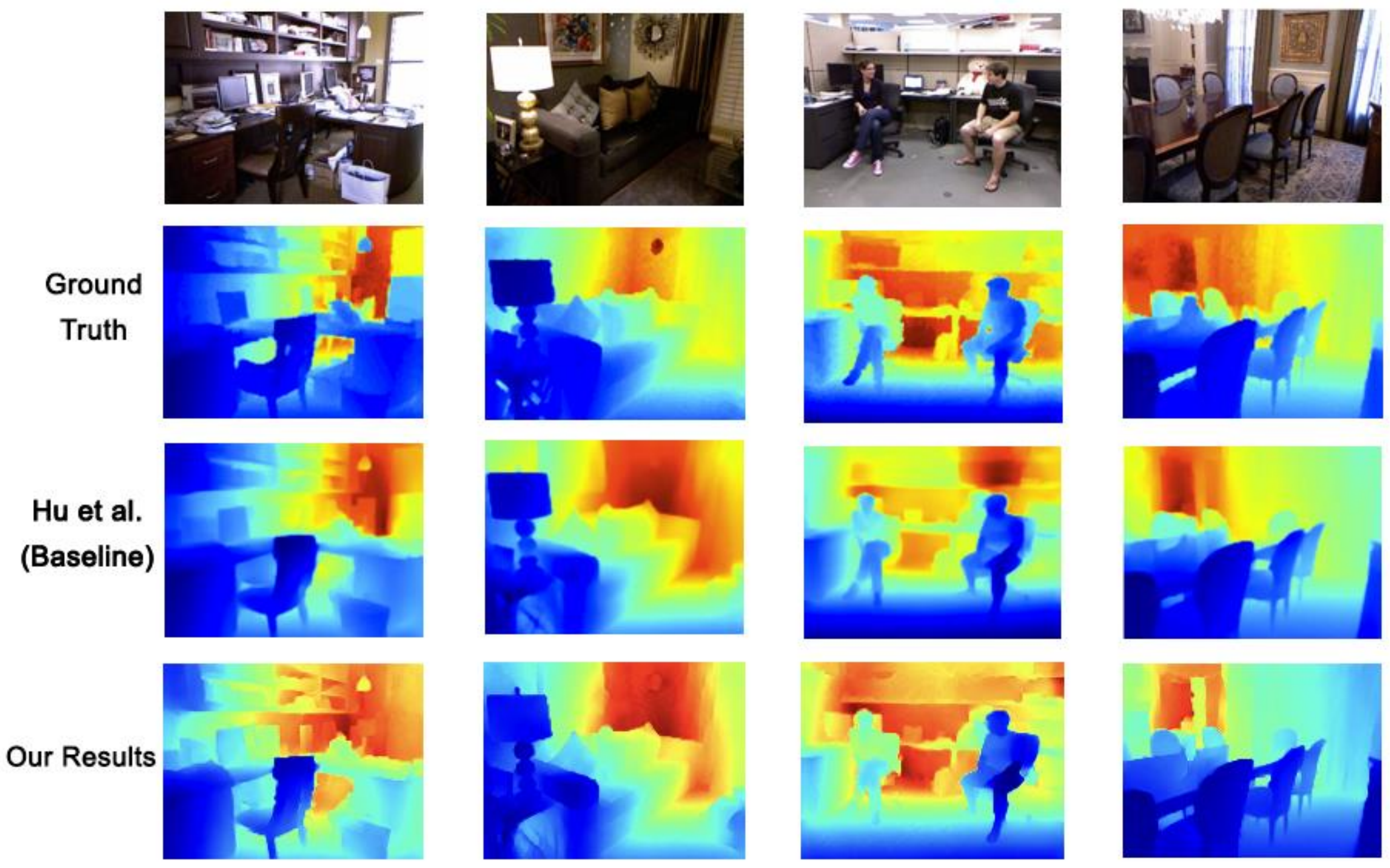


Figure 2. Example comparison of estimated depth maps

Method	Error (lower is better)			Accuracy (higher is better)		
	rms	rel	log10	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Hu et al. (Baseline) [1]	0.530	0.115	0.050	0.866	0.975	0.993
(Ours) Baseline+adversarial	**0.527**	0.117	0.050	0.866	0.974	**0.994**
(Ours)Baseline+adversarial+SSIM	0.537	0.117	0.051	0.860	0.972	0.993
(Ours) Baseline(Depth Priority)+adversarial	0.538	0.121	0.052	0.854	0.971	0.990
(Ours) Baseline(DBE)+adverserial+SSIM	0.537	0.117	0.051	0.860	0.972	0.993
(Ours) Baseline(DBE)+adverserial	0.553	0.123	0.053	0.848	0.968	0.992
(Ours) Baseline+SSIM	0.535	0.117	0.051	0.862	0.974	0.993
Eigen et al. [4]	0.907	0.215	-	0.611	0.887	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

Table 1. Comparison of different methods on the NYU-Depth V2.

V. 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 experiments confirm that adding a discriminator network can be beneficial for depth estimation as shown by our quantitative and qualitative results. Our results also reveal that no significant improvement in the depth map accuracy is observed when using the structural similarity loss. We can conclude that when other restricted constraints such as gradient loss on edges and surface normal loss are involved, the SSIM loss contributes hardly to the quality of the predicted depth map.

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, Clément, Oisin Mac Aodha, and Gabriel J. Brostow. "Unsupervised monocular depth estimation with left-right consistency." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
[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 fields for monocular depth estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3917–3925.
[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. 2002–2011.
[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 283–291, 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 1119–1127, 2015.