Rethinking Monocular Depth Estimation with Adversarial Training

Richard Chen[1], Faisal Mahmood[2], Alan Yuille[1] and Nicholas J. Durr[2]
[1]Department of Computer Science [2]Department of Biomedical Engineering
Johns Hopkins University, Baltimore, MD
{rchen40, faisalm, ayuille, ndurr}@jhu.edu

Abstract

Monocular depth estimation is an extensively studied computer vision problem with a vast variety of applications. This work introduces a novel paradigm for monocular depth estimation using deep networks that incorporates adversarial loss. We describe a variety of deep learning architectures that include a structured loss term with conditional generative adversarial networks. In this framework, the generator learns a mapping between an RGB image and its corresponding depth map, while the discriminator learns to distinguish estimated depth maps from ground truth. We benchmark this approach on the NYUv2 and Make3D datasets, and observe that the addition of adversarial training reduces relative error significantly, achieving SOTA performance on Make3D. These results suggest that adversarial training is a powerful technique for improving depth estimation performance of deep networks.

在GAN中，生成器学习RGB图像与其对应的深度图之间的映射，而鉴别器学习区分估计的深度图与groundtruth

1. Introduction

[1] Depth estimation is one of the most extensively studied tasks by the computer vision community, largely due to its value in facilitating scene understanding [17, 20, 7]. An accurate depth map has been demonstrated to improve the performance of a number of computer vision tasks including semantic segmentation, topographical reconstruction, and activity recognition [33, 22, 11]. However, accurate and unambiguous depth sensing typically requires complex and large equipment (e.g. stereo camera pairs or time of flight sensors) or impractical constraints on the scene (e.g. time-invariant shape) that restrict applicability [17]. On the other hand, depth estimation from a monocular camera would allow more ubiquitous application, but is an ill-posed inverse problem as an infinite number of distinct 3D scenes can be drawn from a 2D image.

Cues such as texture, color, shading, intensity, and other handcrafted features have been widely exploited to estimate depth in 2D images. [30, 12, 19]. Most such models rely on semantic and geometric constraints using semantic and superpixel segmentation data respectively to infer contextual information in a scene [30, 20]. Recently, depth estimation via deep learning or a combination of deep learning and graphical model-based methods have become increasingly common as it has shown superior results to alternative approaches [21, 8, 23, 24]. However, these models are over-parameterized and require large amounts of data, which limit their applicability to a specific scene and do not adapt to a variety of real world scenes. Furthermore, it is difficult to create loss functions that can force these models to learn depth for a span of low and high frequency details in real world scenes. While hybrid CNN-CRF models such as Liu *et al.* [21] and Mahmood *et al.* [23, 24] maintain some spatial consistency between the input image and the ground truth depth map via the pairwise potential, over-segmenting the image into superpixels prevents the network from learning mid-high frequency details that describe object boundaries and foreground-background context.

Recently, conditional generative adversarial networks (conditional GANs) have become an emerging technique in learning mappings of multi-modal distributions of high-dimensional data [13]. Such methods have mainly been used for image-to-image translation tasks such as artistic style transfer, super-resolution [2], and synthetic data refinement [32]. However, they can also be used in inference tasks such as semantic segmentation, in which the generator is a fully convolutional network that learns a mapping from objects to their semantic labels in an image, and the discriminator provides feedback to the generator about its accuracy. Conditional GANs have been shown to improve the state-of-the-art methods in many vision tasks such as semantic segmentation and pose estimation, in which the generator learns from not only a task-specific loss, but also an adversarial loss in competing with the discriminator, which can be interpreted as a structured loss that penalizes the joint configuration of predictions made at the pixel level.

Contributions: In this work, we propose that deep learning-enabled monocular depth estimation can be en-

[1]Work-in-progress

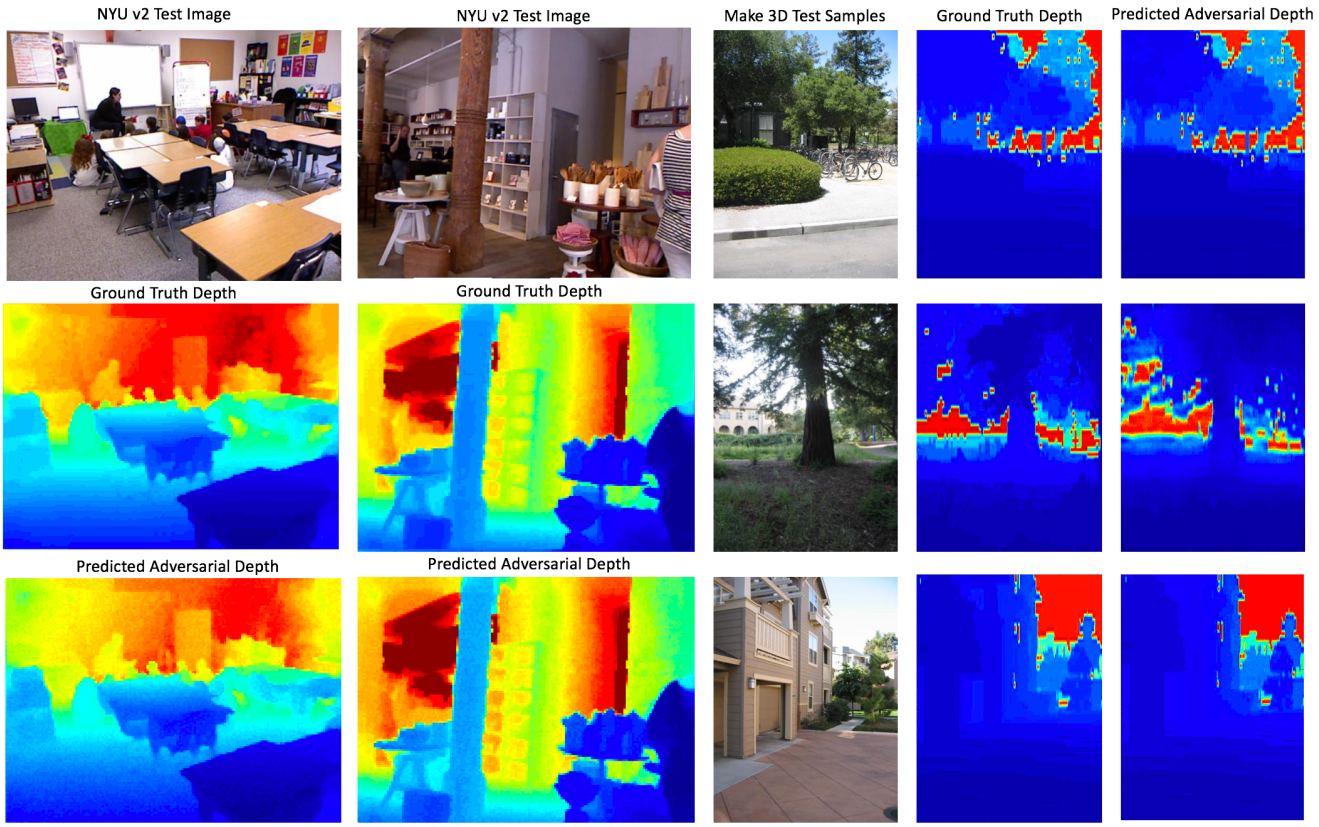

Figure 1: Representative images showing estimated depth via our proposed adversarial depth estimation paradigm. We demonstrate that using an adversarial loss can improve monocular depth estimation using a U-Net, FCRN and CNN-CRF networks. Our state-of-the-art results and comparative analysis with other methods is shown in Table 1.

hanced with adversarial training. We demonstrate an improvement over state-of-the-art depth estimation results on NYUv2 and Make3D datasets using Conditional GANs, and investigate how the addition of adversarial loss affects performance in three different settings of deep networks:

1. Encoder-decoder networks with skip connections (U-Net [26])
2. CRF-Based approaches (CNN-CRF [21])

To the best of our knowledge, this is one of the first works to benchmark and evaluate monocular depth estimation performance in deep networks with adversarial training. The specific contributions of our work are summarized below:

1. We describe a framework for training depth estimation networks that incorporates an adversarial loss term.
2. We demonstrate that, in comparison to a variety of state-of-the-art methods, adversarial training significantly improves monocular depth estimation.
3. We present new state-of-the-art benchmarks for monocular depth estimation on the NYUv2 and Make3D datasets.

2. Related Work

Monocular Depth Estimation. Previous approaches for single-image depth estimation have relied on hand-crafted features, probabilistic graphical models, and deep networks to extract multi-scale contextual information in scenes.

Prior to deep networks, depth estimation was often posed as a Markov Random Field (MRF) learning problem. Saxena *et al.* [30] was the first work to learn depth from single monocular images, using a patch-based MRF to model relations between the depth of image patches with its immediate neighbors at different scales. Liu *et al.* [20] used both semantic and superpixel segmentation information from single images to help guide depth perception, using a pixel-based MRF and superpixel-based MRF to incorporate semantic and geometric constraints respectively. Ladicky *et al.* [15] also incorporated semantic information by learning a joint classifier to predict both depth and semantic labels.

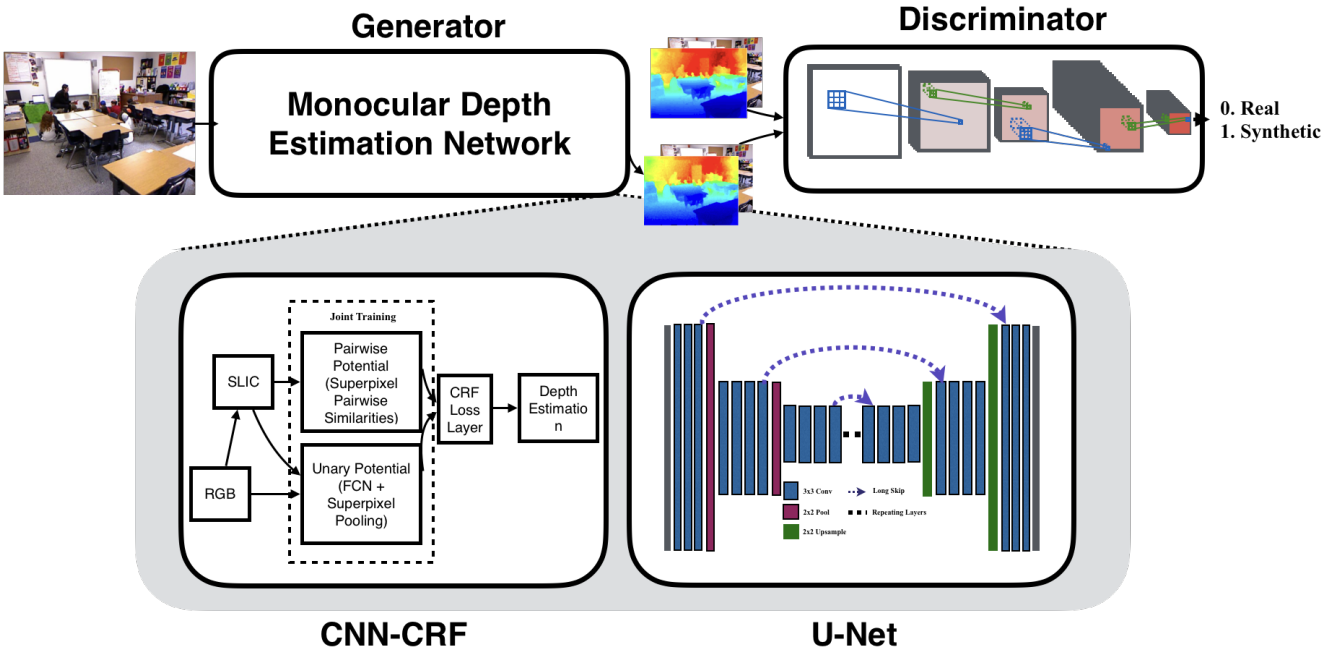

Figure 2: Conditional GAN-based depth estimation architecture with U-Net and CNN-CRF

Following the breakthrough performance of CNNs in image classification, depth estimation is often now posed as a regression problem using end-to-end trained deep networks, with some recent efforts being made to combine both deep networks and graphical models. Eigen *et al.* [7] was the first to use CNNs for monocular depth estimation, in which they proposed a multi-scale deep network that first generates a coarse depth using fully connected layer, followed by a refinement network that recovers high frequency information. Laina *et al.* [16] adopted a fully convolutional architecture that learns an upsampling convolution layer instead of a fully-connected layer to obtain finer depth estimates at higher resolutions. Liu *et al.* [21] presented a CNN-CRF network where the unary potential is a regression term that predicts depth for a given superpixel using fully convolutional layers, and the pairwise potential is a smoothness term measures intensity, color and texture differences between neighboring superpixels. Wang *et al.* [33] introduced a hierarchical CNN-CRF that jointly predicts depth and semantic segmentation from the same features, and was able to refine superpixel-wise CNN depth predictions. Most recently, Xu *et al.* [34] harnesses multi-scale representations by recovering depth maps at each side output of an encoder-decoder network using a continuous CRF framework.

Generative Adversarial Networks.

The GAN framework was first presented by Goodfellow *et. al.* in [29, 9, 10] and was based on the idea of training two networks, a generator and a discriminator simultaneously with competing losses. While the generator learns to generate realistic data from a random vector, the discriminator classifies the generated image as real or fake and gives feedback to the generator. GANs have recently been used for a variety of different applications [31, 5, 3, 1]. GANs have recently been used for image-to-image translation and style-transfer and synthetic data generation [23], conditional GANs have been used to improve state-of-the-art results in traditional vision tasks such as semantic segmentation and human pose estimation.

3. Conditional GAN Framework for Depth Estimation

In this section, we describe the Conditional GAN objective for training depth estimation networks with adversarial loss, followed by network architecture details. We denote A and B as the RGB image and depth image domains respectively, and a and b as training examples in A and B. Additionally, we denote G as a mapping function $G : A \rightarrow B$ that learns a mapping from RGB to depth, and D as the discriminator network for G.

3.1. Conditional GAN Objective

The Conditional GAN framework consists of two networks that compete against each other in a minimax game to respectively minimize and maximize the objective, $\min_G \max_D \mathcal{L}(G, D)$. The generator G would aim to learn a mapping from A to B, and a discriminator D seeks to distinguish between real and synthesized pairs. To train this framework for semantic segmentation for paired data, the

Conditional GAN objective consists of an adversarial loss term $\mathcal{L}_{\text{GAN}}$ and a per-pixel loss term $\mathcal{L}_{\text{L1}}$ to penalize both the joint configuration of pixels and accuracy of the estimated map.

The adversarial loss is used to match the distribution of generated samples to that of the target distribution. For the mapping $G : A \rightarrow B$, we can express the adversarial objective as the binary cross entropy loss of D in classifying real/synthesized pairs. We can express this loss as:

$$\min_G \max_D \mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{a,b\sim p_{\text{data}}(a,b)}[\log D(A, B)] + \mathbb{E}_{a\sim p_{\text{data}}(a)}[\log(1 - D(G(A), B)]$$

The L_1 loss term is used to score the accuracy of the depth estimation by G.

$$\mathcal{L}_{L_1}(G) = \mathbb{E}_{a,b\sim p_{\text{data}}}(a, b)[||b - S(a)||_1]$$

For monocular depth estimation, we train a generator to learn a mapping between a RGB image and its corresponding depth map, and a discriminator to distinguish between ground truth and predicted depth conditioned on the RGB image. In training the generator, we mix a per-pixel loss with an adversarial loss to model both smooth depth transitions for pixels that are similar in intensity and color, and sharp depth transitions at object boundaries that give foreground and background context. Per-pixel losses by themselves are "unstructured", in that each output pixel is considered conditionally independent from all other pixels given the image. When used in depth estimation, per-pixel losses such as ℓ_1 and ℓ_2 tend to produce blurry results, as the total relative error is averaged across all pixels which can prevent the network from learning from inaccurate depth measurements. An adversarial loss, on the other hand, penalizes the joint configuration of pixel predictions made in an image or image patch. The adversarial loss can additionally be interpretted as a structural loss that can help preserve realism. Pixel configurations such as image blur at the corner of a table would look unrealistic to the discriminator when sharp object boundaries are present in the scene. The generator's objective can then be rewritten as:

$$\arg\min_G \arg\max_D \mathcal{L}_{\text{GAN}}(G, D) + \lambda \mathcal{L}_{L_1}(G)$$

where λ is the mixing parameter. To validate this concept we introduce an adversarial loss for depth estimation in three different networks and demonstrate that state-of-the-art results can be achieved with NYUv2 and Make3D datasets.

3.2. Network Architectures

In this section, we provide details on the three network architectures used for depth estimation with adversarial training.

Encoder-Decoder Networks with Skip Connections. Encoder-decoder architectures are commonly used in image-to-image translation tasks where the input and output image share roughly the same structure [25]. One particular formulation of the encoder-decoder architecture is U-Net [27], which draws skip connections between convolution layers on the encoder path and upsampling layers on the decoder path that have the same spatial size. Skip connections are used to recover and enforce spatial information across multiple resolutions and enforce spatial consistency on the output image, where the input and outputs are expected to align channel-wise [6]. In practice, however, U-Nets do not produce realistic depth estimate they introduce too much high frequency and color information at the end of the network. We demonstrate that introducing an adversarial loss can remove high frequency information introduced by the skip connection, while also preserving object boundaries and other mid frequency information (Fig. 1, Table 1). Our particular formulation of U-Net (Fig. 2) assumes that input images are 256×256, as the inputs are eventually downsampled to 1×1 pixel at the bottleneck. Pooling and upsampling operations are replaced with 4×4 convolution filters with stride 2×2 and transposed convolutions respectively.

$$\arg\min_{G=\text{U-Net}} \arg\max_D \mathcal{L}_{\text{GAN}}(G_{\text{U-Net}}, D) + \lambda \mathcal{L}_1(G_{\text{U-Net}})$$

Joint CNN-CRF Network. In this section we explain the concept of using an adversarial loss in a joint CNN-CRF network. Assuming $\boldsymbol{x} \in \mathbb{R}^{n\times m}$ be an image which has been divided into g superpixels and $\boldsymbol{y} = [y_1, y_2, ..., y_g] \in \mathbb{R}$ be the depth vector corresponding each superpixel. In this case the conditional probability distribution of the raw data can be defined as,

$$Pr(\boldsymbol{y}|\boldsymbol{x}) = \frac{exp(E(\boldsymbol{y}, \boldsymbol{x}))}{\int_{-\infty}^{\infty} exp(E(\boldsymbol{y}, \boldsymbol{x}))d\boldsymbol{y}}. \quad (1)$$

and E is the energy function. In order to predict the depth of a new image the following maximum aposteriori (MAP) problem has to be solved, $\widehat{\boldsymbol{y}} = \text{argmax}_y \, Pr(\boldsymbol{y}|\boldsymbol{x})$.

Let ψ and ϕ be unary and pairwise potentials over nodes $\mathcal{N}$ and edges $\mathcal{S}$ of $\boldsymbol{x}$, then the energy function can be formulated as,

$$E(\boldsymbol{y}, \boldsymbol{x}) = \sum_{i\in\mathcal{N}} \psi(y_i, \boldsymbol{x}; \boldsymbol{\gamma}) + \sum_{(i,j)\in\mathcal{S}} \phi(y_i, y_j, \boldsymbol{x}; \boldsymbol{\beta}), \quad (2)$$

where, ψ regresses the depth from a single superpixel and ϕ encourages smoothness between neighboring superpixels. The objective is to learn the two potentials in a unified convolutional neural network (CNN) framework. This setup has been described in Fig. 2. The unary part takes a single image superpixel patch as an input and feeds it to

a CNN which outputs a regressed depth of that superpixel. Based on [21, 24] the unary potential can be defined as,

$$\psi(y_i, \boldsymbol{x}; \boldsymbol{\gamma}) = -(y_i - h_i(\boldsymbol{\gamma}))^2 \tag{3}$$

where h_i is the regressed depth of superpixel and γ represents CNN parameters. The pairwise potential function is based on standard CRF vertex and edge feature function studied in [**?**]. Let $\boldsymbol{\beta}$ be the network parameters and $\boldsymbol{S}$ be the similarity matrix where $S^k_{i,j}$ represents k similarity metrics between the i^{th} and j^{th} superpixel. We use intensity difference and grayscale histogram as pairwise similarity metrics expressed in the general ℓ_2. The pairwise potential can be defined as,

$$\phi(y_i, y_j; \boldsymbol{\beta}) = -\frac{1}{2}\sum_{k=1}^{K} \beta_k S^k_{i,j}(y_i - y_j)^2. \tag{4}$$

The overall energy function can now be written as,

$$E = -\sum_{i \in \mathcal{N}} (y_i - h_i(\boldsymbol{\gamma}))^2 - \frac{1}{2}\sum_{(i,j)\in\mathcal{S}} \sum_{k=1}^{K} \beta_k S^k_{i,j}(y_i - y_j)^2. \tag{5}$$

For training the negative log likelihood of the probability density function which can be calculated from Eq. 6 is minimized with respect to the two learning parameters. Two regularization terms are added to the objective function to penalize heavily weighted vectors. Assuming N is the number of images in the training data,

$$\min_{\gamma,\beta\geq 0} -\sum_{1}^{N} \log Pr(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\gamma}, \boldsymbol{\beta}) + \frac{\lambda_1}{2}\left\|\gamma\right\|_2^2 + \frac{\lambda_2}{2}\left\|\beta\right\|_2^2. \tag{6}$$

Incorporating adversarial loss in this setup means treating the objective function above as the generator and jointly training the discriminator and generator using the following loss,

$$\arg\min_{G=\text{CNN-CRF}} \arg\max_{D} \mathcal{L}_{\text{GAN}}(G_{\text{CNN-CRF}}, D) + \lambda \mathcal{L}_1(G_{\text{CNN-CRF}})$$

3.3. Experimental Setup

In this section, we evaluate our method against current state-of-the-art monocular depth estimation methods on two standard benchmark datasets for depth prediction, i.e., NYU Depth v2 [7] and Make3D [30].

NYUv2. NYUv2 is one of the largest RGB-D datasets for indoor scene reconstruction, with over 120K unique pairs of RGB and depth images acquired from 464 scenes with a Microsoft Kinect. We worked with a 1449 aligned subset of images from NYUv2, with 795 pairs for training and 654 pairs for testing, and downsamples the images from 640×480 to 320×240. To train our adversarial U-Net, we further downsampled the training set to 386×288, and performed random horizontal flips with random 256×256 crop, To train our adversarial FRCN, we performed random horizontal flips with 304×228 crops.

Make3D. The Make3D dataset contains 534 RGB-D image pairs of outdoor scenes, with 400 pairs for training and 134 images for testing. To train our adversarial U-net, we performed a similar data augmentation to NYUv2 by downsampling the images to 400×300, and performed random horizontal flips with random 256×256 crop.

Evaluation Metrics. Following previous works [16, 34, 20, 21], we considered the following performance metrics for accurate depth estimation:

1. Relative Error (rel): $\frac{1}{N}\sum_y \frac{|y_{gt} - y_{est}|}{y_{gt}}$
2. Average log_{10} Error (log_{10}): $\frac{1}{N}\sum_y |\log_{10} y_{gt} - \log_{10} y_{est}|$
3. Root Mean Square Error (rms): $\sqrt{\frac{1}{N}\sum_y (y_{gt} - y_{est})^2}$

4. Discussion

In these series of comparisons, we evaluate the proposed adversarial depth estimation networks with their non-adversarial counterparts, in which we overall observed a decrease in relative error and increase in accuracy when an adversarial loss was added. In the traditional U-Net vs. adversarial U-Net comparison, despite enforcing strong spatial consistency between the convolution and upsampling layers in the long skip connections, we achieved a relative error of 0.1136 and 0.0646 on NYUv2 and Make3D respectively, which improves over the current state of the art relative error Xu *et al.*. While the relative error decreased in adding an adversarial loss for U-Net, the accuracy was not very high, though still making an improvement over traditional U-Net. The Adversarial CNN-CRF only marginally improved in relative error and accuracy with a threshold of 0.125 on NYUv2, with accuracy decreasing with a threshold of 0.125^3. The low accuracy in both networks can be attributed to how using a Ł$_1$ loss directly optimizes for the relative error than accuracy, which penalizes greater deviations in per-pixel predictions from the ground truth. In addition, the relatively low accuracy can also be attributed to the small training set used to train U-Net and CNN-CRF. In observing both relative error and accuracy improving, we hypothesize that this framework in adapting depth estimation networks to be adversarial is still useful. As more depth estimation network implementations become open source, we plan to add these networks to our benchmark.

Method	rel	log_{10}	rms	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
CNN-CRF [21]	0.213	0.087	0.759	0.650	**0.906**	**0.976**
CNN-CRF-Adv.	**0.202**	**0.081**	**0.755**	**0.658**	0.901	0.962
U-Net [27]	0.327	0.124	0.981	0.508	0.783	0.815
U-Net-Adv.	**0.1136**	**0.050**	**0.4871**	**0.634**	**0.733**	**0.852**

Table 1: A comparative analysis of monocular depth estimation methods with NYUv2 dataset.

	C1			C2		
Method	rel	log_{10}	rms	rel	log_{10}	rms
Make3D [30]	-	-	-	0.370	0.187	-
Liu *et al.* [20]	-	-	-	0.379	0.148	-
DepthTransfer [14]	0.355	0.127	9.20	0.361	0.148	15.10
Liu *et al.* † [21]	0.355	0.137	9.49	0.338	0.134	12.60
Li *et al.* [17]	0.278	0.092	7.12	0.279	0.102	10.27
Liu *et al.* [21]	0.287	0.109	7.36	0.287	0.122	14.09
Roy *et al.* [28]	-	-	-	0.260	0.119	12.40
Laina *et al.* [16]	0.176	0.072	**0.790**	-	-	-
LRC-Deep3D § [27]	1.000	2.527	0.981	-	-	-
LRC *et al.* [21]	0.443	0.156	11.513	-	-	-
U-Net § [27]	0.428	0.142	5.127	0.446	0.164	6.38
Kuzietsov *et al.* [33]	0.421	0.190	8.24	-	-	-
MS-CRF *et al.* [34]	0.184	0.065	4.38	0.198	-	8,56
DORN (ResNet) [8]	0.157	0.062	3.97	0.162	0.067	7.32
U-Net-Adv.	**0.0646**	**0.0277**	1.812	**0.0817**	**0.0493**	**4.163**

Table 2: A comparative analysis of monocular depth estimation methods with Make3D Dataset.§ U-Net implemented for depth estimation.

5. Conclusion

In this paper, we present an adversarial framework for depth estimation with benchmarks. Future work includes adding more adversarial depth networks to our benchmark.

References

[1] U. Ahsan, C. Sun, and I. Essa. Discrimnet: Semi-supervised action recognition from videos using generative adversarial networks. *arXiv preprint arXiv:1801.07230*, 2018.

[2] A. Bulat and G. Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. *arXiv preprint arXiv:1712.02765*, 2017.

[3] E. A. Burlingame, A. Margolin, J. W. Gray, and Y. H. Chang. Shift: speedy histopathological-to-immunofluorescent translation of whole slide images using conditional generative adversarial networks. In *Medical Imaging 2018: Digital Pathology*, volume 10581, page 1058105. International Society for Optics and Photonics, 2018.

[4] A. Chakrabarti and T. Zickler. Depth and deblurring from a spectrally-varying depth-of-field. In *European Conference on Computer Vision*, pages 648–661. Springer, 2012.

[5] K. Choi, S. W. Kim, and J. S. Lim. Real-time image reconstruction for low-dose ct using deep convolutional generative adversarial networks (gans). In *Medical Imaging 2018: Physics of Medical Imaging*, volume 10573, page 1057332. International Society for Optics and Photonics, 2018.

[6] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. The Importance of Skip Connections in Biomedical Image Segmentation. *ArXiv e-prints*, Aug. 2016.

[7] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

[8] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[9] I. Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[11] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian Conference on Computer Vision*, pages 213–228. Springer, 2016.

[12] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172, 2007.

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[14] K. Karsch, C. Liu, and S. B. Kang. Depth transfer: Depth extraction from videos using nonparametric sampling. In *Dense Image Correspondences for Computer Vision*, pages 173–205. Springer, 2016.

[15] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–96, 2014.

[16] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016.

[17] 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 Conference on Computer Vision and Pattern Recognition*, pages 1119–1127, 2015.

[18] J. Li, R. Klein, and A. Yao. A two-streamed network for estimating fine-scaled depth maps from single rgb images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3372–3380, 2017.

[19] X. Li, H. Qin, Y. Wang, Y. Zhang, and Q. Dai. Dept: depth estimation by parameter transfer for single still images. In *Asian Conference on Computer Vision*, pages 45–58. Springer, 2014.

[20] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1253–1260. IEEE, 2010.

[21] F. Liu, C. Shen, G. Lin, and I. Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2024–2039, 2016.

[22] D. C. Luvizon, D. Picard, and H. Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. *arXiv preprint arXiv:1802.09232*, 2018.

[23] F. Mahmood, R. Chen, and N. J. Durr. Unsupervised reverse domain adaption for synthetic medical images via adversarial training. *arXiv preprint arXiv:1711.06606*, 2017.

[24] F. Mahmood and N. J. Durr. Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. *arXiv preprint arXiv:1710.11216*, 2017.

[25] A. Odena, C. Olah, and J. Shlens. Conditional Image Synthesis With Auxiliary Classifier GANs. *ArXiv e-prints*, Oct. 2016.

[26] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI2015*, May 2015.

[27] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[28] A. Roy and S. Todorovic. Monocular depth estimation using neural regression forest. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5506–5514, 2016.

[29] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[30] A. Saxena, M. Sun, and A. Y. Ng. Learning 3-d scene structure from a single still image. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[31] O. Sbai, M. Elhoseiny, A. Bordes, Y. LeCun, and C. Couprie. Design: Design inspiration from generative networks. *arXiv preprint arXiv:1804.00921*, 2018.

[32] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, page 6, 2017.

[33] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille. Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2800–2809, 2015.

[34] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe. Monocular depth estimation using multi-scale continuous crfs as sequential deep networks. *arXiv preprint arXiv:1803.00891*, 2018.