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. Deep learning-based methods have demonstrated promise for both supervised and unsupervised depth estimation from monocular images. Most existing approaches treat depth estimation as a regression problem with a local pixel-wise loss function. In this work, we innovate beyond existing approaches by using adversarial training to learn a context-aware, non-local loss function. Such an approach penalizes the joint configuration of predicted depth values at the patch-level instead of the pixel-level, which allows networks to incorporate more global information. In this framework, the generator learns a mapping between RGB images and its corresponding depth map, while the discriminator learns to distinguish depth map and RGB pairs from ground truth. This conditional GAN depth estimation framework is stabilized using spectral normalization to prevent mode collapse when learning from diverse datasets. We test this approach using a diverse set of generators that include U-Net and joint CNN-CRF. We benchmark this approach on the NYUv2, Make3D and KITTI datasets, and observe that adversarial training reduces relative error by several fold, achieving state-of-the-art performance.

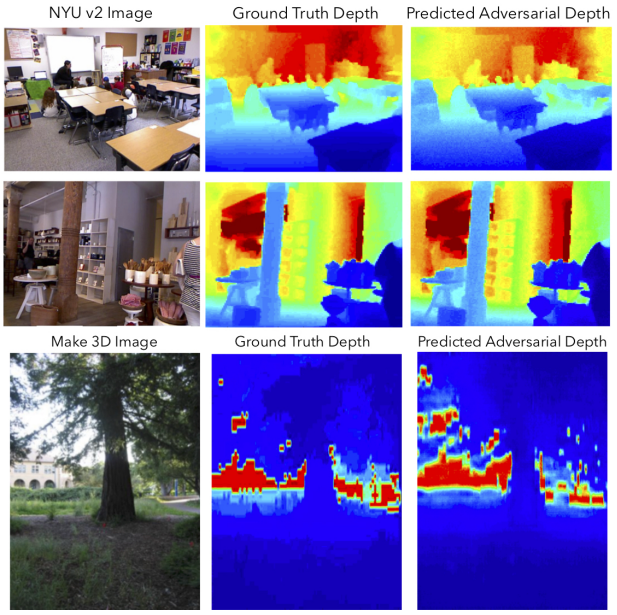

Figure 1: Representative images showing estimated depth on NYUv2 and Make 3D datasets via our proposed adversarial depth estimation paradigm.

1. Introduction

Depth estimation is one of the most extensively studied tasks by the computer vision community, largely due to its value in facilitating scene understanding and geometric relations between objects [27, 29, 10]. Fusing depth has demonstrated improved performance on a number of computer vision tasks including semantic segmentation, topographical reconstruction, and activity recognition [54, 33, 17]. Previously, the computer vision community relied on multiview methods such as stereo vision and structure-from-motion for depth estimation [40, 4, 3, 38, 55, 2]. However, situations where multiple measurements from the same scene may not be available or difficult to acquire motivate the need for developing monocular depth estimation methods.

Though deep networks have shown promise in estimating depth from monocular images, many methods rely on local pixel-wise loss functions that do not capture higher-order statistics of the training data. To make these networks more context-aware, many loss functions calculate image gradients to capture changes in depth and preserve structural details [10, 48, 19]. This problem has also been partially addressed by the many combinations of deep learning and graphical model-based methods [31, 11, 34, 35]. CNN-graphical model setups such as jointly trained CNN-CRF methods are more context-aware as compared to regular CNNs. While hybrid CNN-CRF models such as Liu *et al.* [31] and Mahmood *et al.* [34, 35] maintain some spa-

tial consistency between the prediction and the ground truth depth map via the pairwise potential, over-segmenting the image into super-pixels prevents the network from learning higher-order statistics that may describe depth cues in the image.

Recently, conditional generative adversarial networks (cGANs) have become an emerging technique in learning mapping distributions of high-dimensional data [21]. Such methods have mainly been used for image-to-image translation tasks such as artistic style transfer, super-resolution [5], and synthetic data refinement [50]. However, they can also be used in inference tasks such as semantic segmentation[32], in which the generator learns a mapping from objects to their semantic labels in an image, and the discriminator provides feedback to the generator about its accuracy. As argued by Luc *et al.* and Isola *et al.* [32, 21], this adversarial term can be interpreted as a non-local loss that penalizes the joint configuration of pixel values. We argue further that this non-local loss, when calculated at the patch-level, is beneficial for learning depth cues. From our experiments, the benefits of conditional adversarial learning are two fold: a) networks can learn a loss function for depth estimation, which promotes the recovery of features that would be generally lost due to the limitations of a local pixel-wise loss function, and (b) such a setup is more context-aware, as the discriminator forces the generator to generate realistic pixel configurations of predicted depth that would be indistinguishable from ground truth depth maps.

Contributions: In this work, we propose that deep learning-enabled monocular depth estimation can be enhanced with adversarial training. We demonstrate an improvement over state-of-the-art depth estimation results on NYUv2, Make3D and KITTI datasets using conditional GANs, and investigate how the addition of an adversarial term affects performance when using encoder-decoder network and CNN-graphical model setups as generators. The specific contributions of our work are summarized below:

1. We describe a framework for context-aware depth estimation with stable adversarial training. Our framework learns a non-local loss function for depth estimation by incorporating a patch-level adversarial term, in which the discriminator classifies regions in the depth map predictions as synthetic or realistic. Depth regions that are predicted as synthetic penalize the generator for producing unrealistic depth configurations that fail to mimic real depth regions.

2. We present state-of-the-art algorithms and models for monocular depth estimation on the NYUv2, Make3D and KITTI datasets, and demonstrate how adversarial training can be adapted for different model types.

2. Related Work

Monocular Depth Estimation.
Historically, depth estimation has been approached by multiview methods such as stereopsis [55]. A large body of knowledge has also focused on recovering depth from shading [43], texture [2], and focus [38]. Many approaches for monocular depth estimation rely on hand-crafted features, probabilistic graphical models, and deep networks to extract multi-scale contextual information in scenes. Prior to deep networks, depth estimation became posed as a Markov Random Field (MRF) learning problem. Saxena *et al.* [48] used a patch-based MRF to model relations between the depth of image patches with its immediate neighbors at different scales. Liu *et al.* [29] 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.* [25] also incorporated semantic information by learning a joint classifier to predict both depth and semantic labels.

Following the breakthrough performance of CNNs for regression and classification tasks, depth estimation is often posed as a regression problem using end-to-end trained deep networks, with some recent efforts being made combine deep networks with graphical models. Eigen *et al.* [10] 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 a fully connected layer, followed by a refinement network that recovers texture details. Laina *et al.* [26] 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, and exploits network depth to capture global information in an image. Liu *et al.* [31] 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.* [54] 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. Xu *et al.* [57] learned multi-scale representations by recovering depth maps at each side output of an encoder-decoder network using a continuous CRF framework, and later followed their work by incorporating attention modules at the bottleneck of their encoder-decoder network [58].

In addition to advancements made in neural network architectures to incorporate context, there is also interest in engineering novel loss functions for recovering depth beyond $\mathcal{L}_{\text{L1}}$, $\mathcal{L}_{\text{L2}}$, and Huber (Smooth $\mathcal{L}_{\text{L1}}$). Eigen *et al.* [10] was the first to use $\mathcal{L}_{\text{L1}}$ loss in *log*-space and

gradient loss terms in deep networks. The use of $log\mathcal{L}_{\text{L1}}$ downweights the contribution of depth errors at background pixel indices which tend to have less rich information, and the use of gradient terms help preserve details on local structure and surface regions. Laina *et al.* [26] introduced the BerHu loss, which penalized low and high errors by an ℓ_1-norm and ℓ_2-norm respectively. Jiao et al. [22] observed that on some depth estimation benchmarks, the distribution of depth values are skewed towards the foreground, which motivated a different loss function than Eigen *et al.* that weighted depth errors at the background indices more heavily than those at the foreground indices.

Conditional Generative Adversarial Networks.

The GAN framework was first presented by Goodfellow *et. al.* in [47, 14, 15] 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 [49, 8, 6, 1] including image-to-image translation [21] and style-transfer and synthetic data generation [34]. Although, GANs have a generative and artistic ability, in order to harness the benefits of the GAN framework for specific vision applications they must be conditioned by additional information. This auxiliary information can be class labels, images or any other kind of data. Such a setup is termed conditional GANs (cGANs) and was first introduced by Mirza *et al.* [36]. In cGANs, the noise vector typical to GAN problems is combined with this auxiliary conditioning information resulting in generative models that are capable of transferring between domains. This approach has been used for paired [21] and unpaired image-to-image translation [60]. Since their advent, cGANs have been used for a variety of computer vision tasks, most notably in semantic segmentation [32]. Recently, Krishna *et al.* [44] used cGANs for cross view image synthesis. Wang *et al.* [53] have used cGANs for jointly learning shadow detection and removal and Hong *et al.* [20] used it for structured domain adaptation. In the joint landscape of both monocular depth estimation and GANs, Pilzer *et al.* [42] describes an unsupervised approach for depth estimation using cycle-consistent adversarial training. Because this approach uses unpaired data, the adversarial training does not explicitly preserve structural information and surface regions from depth prediction and its ground truth [16].

3. Conditional GAN Framework for Depth Estimation

In this section, we describe the conditional GAN objective for training depth estimation networks with non-local adversarial loss, followed by network architecture details.

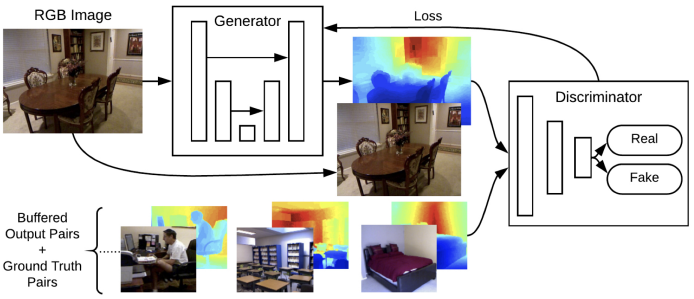

Figure 2: Conditional GAN-based depth estimation architecture with U-Net as a generator.

We denote A and A_d as the RGB and depth image domains respectively, and a and a_d as training examples in A and A_d. Additionally, we denote G as a mapping function $G : A \rightarrow A_d$ that learns a mapping from RGB to depth, and D as the discriminator network for G.

3.0.1 Conditional GAN Objective

The conditional GAN framework consists of two networks that compete against each other in a *min-max* game to respectively minimize and maximize the objective, $\min_G \max_D \mathcal{L}(G, D)$. The generator G learns a mapping from A to A_d, and the discriminator D distinguishes between real and synthesized pairs of depth and RGB. To train this framework for depth estimation 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 depth maps.

The adversarial loss is used to match the distribution of generated samples to that of the target distribution. For the mapping $G : A \rightarrow A_d$, 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:

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

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

$$\mathcal{L}_{\text{L1}}(G) = \mathbb{E}_{a,a_d \sim p_{\text{data}}}(a, a_d)[||a_d - G(a)||_1] \quad (2)$$

The motivation for using an adversarial loss is to incorporate non-local information, which has been shown to be instrumental 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 on the patch-level. Per-pixel losses are generally local, 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 $\mathcal{L}_{\text{L1}}$ and $\mathcal{L}_{\text{L2}}$ tend to produce blurry results, as the total relative error is averaged across all pixels which. An adversarial loss, on the other hand, penalizes the joint configuration of pixel predictions made in an image. The adversarial loss can be interpreted as a non-local loss that can help preserve more details. The adversarial loss comes from the discriminator, which classifies overlapping pairs of image and depth patches as being real or synthetic. By controlling the size of the patch, we can control the size of the non-locality, with bigger patches incorporating more global information in the image. Experimentally, we observed that predicting on 70×70-sized patches allowed the generator to make fine-grained depth predictions. Thus, we can write the loss function for the conditional GAN framework as:

$$\arg\min_{G} \arg\max_{D} \mathcal{L}_{\text{GAN}}(G, D) + \lambda \mathcal{L}_{\text{L1}}(G) \tag{3}$$

where, λ is a mixing parameter. As argued in Isola *et al.* [21], such an adversarial loss setup can be thought as a learned loss function where the adversarial loss is learned as the discriminator and generator are trained.

3.1. Stabilizing GAN training

As noted in previous works, the training procedure for GANs can be very unstable and lead to mode collapse and gradient artifacts in the depth predictions[14]. To address this, we apply spectral normalization in both the generator and the discriminator. Spectral normalization was first introduced in Miyamoto et al. [37], and was used to control the Lipschitz constant of the discriminator such that the spectral norm σ of the convolution weights W in the network would be bounded by the Lipschitz constraint: $\sigma(W) = 1$. As a result, the discriminator is more stable during training and can avoid exploding gradients. Following the two-timescale update rule in Heusal *et al.* [18], the learning rate for the discriminator was set to be four times the learning rate of the generator to increase speed of convergence. In subsequent experiments by [59], spectral normalization was empirically determined to be also beneficial for stabilizing the generator, allowing for fewer discriminator updates per generator update. We also further stabilize the discriminator by using a buffered data input from the generator, which consists of previously generated and classified pairs and ground truth data. This approach to stabilizing the GAN training procedure was presented in Shrivastava et al.[50], and has been used in several proceeding works [21, 34]. In our observations, we found that these techniques reduced visual artifacts made by the generator, which resulted in more smoothly-varying depth estimates.

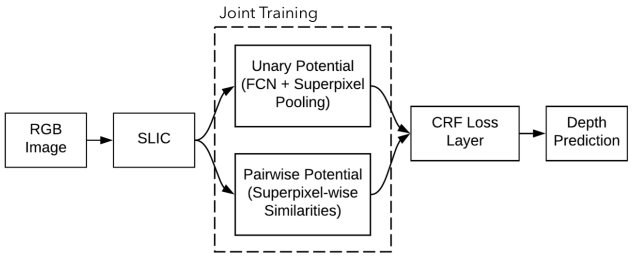

Figure 3: Joint CNN-CRF training paradigm. CNN-CRF models are more context aware as compared to regular CNN models. However, due to computational complexity we used this model as a generator on a super-pixel level.

3.2. Network Architectures

In this section, we provide details of the two different types of network architectures used as a generator for depth estimation with adversarial training.

Encoder-Decoder Networks. Encoder-decoder networks are commonly used in many deep network approaches for monocular depth estimation [26, 57, 58, 22, 39]. One formulation, the U-Net achitecture by Ronneberger *et al.* [45], draws skip connections between convolution layers on the encoder path and up-sampling layers on the decoder path that have the same spatial size. These connections are made between feature maps 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 [9]. We demonstrate that introducing an adversarial term can recover higher order information, while also preserving object boundaries and shape details (Fig. 4, Table 1). Our implementation of U-Net (Fig. 2) assumes that input images are 256×256, as the inputs are down sampled to 1×1 pixel at the bottleneck. Pooling and up-sampling operations are replaced with 4×4 convolution filters with stride 2×2 and transposed convolutions respectively. The U-Net loss can be defined as,

$$\arg\min_{G=\text{U-Net}} \arg\max_{D} \mathcal{L}_{\text{GAN}}(G_{\text{U-Net}}, D) + \lambda \mathcal{L}_{\text{L1}}(G_{\text{U-Net}}). \tag{4}$$

Joint CNN-CRF Network. In this section we explain the how an adversarial loss can be used 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,

$$Pr(\boldsymbol{y}|\boldsymbol{x}) = \frac{exp(E(\boldsymbol{y}, \boldsymbol{x}))}{\int_{-\infty}^{\infty} exp(E(\boldsymbol{y}, \boldsymbol{x}))d\boldsymbol{y}}. \tag{5}$$

and E is the energy function. In order to predict the depth

Method	rel ↓	log_{10} ↓	rms ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Make3D[48]	0.349	-	1.214	0.447	0.745	0.897
DepthTransfer[23]	0.350	0.131	1.20	-	-	-
Liu *et al.* [29]	0.335	0.127	1.06	-	-	-
Ladicky *et al.* [25]	-	-	-	0.542	0.829	0.941
Li *et al.* [27]	0.232	0.094	0.821	0.621	0.886	0.968
Wang *et al.* [54]	0.220	-	0.824	0.605	0.890	0.970
Roy *et al.* [46]	0.187	-	0.744	-	-	-
Liu *et al.* [31]	0.213	0.087	0.759	0.650	0.906	0.976
Eigen *et al.* [10]	0.158	-	0.641	0.769	0.950	0.988
Chakrabarti *et al.* [7]	0.149	-	0.620	0.806	0.958	0.987
Laina *et al.* [26]	0.194	0.083	0.790	0.629	0.889	0.971
Li *et al.* [28]	0.152	0.064	0.611	0.789	0.955	0.988
MS-CRF *et al.* [57]	0.121	0.052	0.586	0.811	0.954	0.987
DORN [11]	0.115	0.051	0.509	0.828	0.965	0.992
CNN-CRF	0.232	0.094	0.824	0.614	0.883	0.971
CNN-CRF-Adv.	0.202	0.081	0.755	0.658	0.901	0.962
U-Net [45]	0.327	0.124	0.981	0.508	0.783	0.815
U-Net-Adv.	**0.114**	**0.050**	**0.4871**	**0.852**	**0.971**	**0.997**

Table 1: Performance on NYUv2. All methods are evaluated on the test split by Eigen *et al.* [10]

of a new image we must solve the maximum aposteriori (MAP) problem, $\widehat{\boldsymbol{y}} = \text{argmax}_y \, Pr(\boldsymbol{y}|\boldsymbol{x})$.

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

$$E(\boldsymbol{y}, \boldsymbol{x}) = \sum_{i \in \mathcal{N}} \psi(y_i, \boldsymbol{x}; \boldsymbol{\gamma}) + \sum_{(i,j) \in \mathcal{S}} \phi(y_i, y_j, \boldsymbol{x}; \boldsymbol{\beta}), \quad (6)$$

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 is shown in Fig. 3. The unary part takes a single image superpixel patch as an input and feeds it to a CNN which outputs the regressed depth of that superpixel. Based on [31, 35] the unary potential can be defined as,

$$\psi(y_i, \boldsymbol{x}; \boldsymbol{\gamma}) = -(y_i - h_i(\boldsymbol{\gamma}))^2 \quad (7)$$

where h_i is the regressed depth of superpixel $i \in N$ and γ represents CNN parameters. The pairwise potential function is based on the standard CRF vertex and edge feature function studied in [34]. 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 the intensity difference and grayscale histogram as pairwise similarity metrics using ℓ_2-norm. The pairwise potential can be defined as,

$$\phi(y_i, y_j; \boldsymbol{\beta}) = -\frac{1}{2} \sum_{k=1}^{K} \beta_k S^k_{i,j} (y_i - y_j)^2. \quad (8)$$

The overall energy function is,

$$E = -\sum_{i \in \mathcal{N}} (y_i - h_i(\boldsymbol{\gamma}))^2 - \frac{1}{2} \sum_{(i,j) \in \mathcal{S}} \sum_{k=1}^{K} \beta_k S^k_{i,j} (y_i - y_j)^2. \quad (9)$$

During training the negative log likelihood of the probability density function is calculated from Eq. 5 and minimized with respect to the two learning parameters, γ and β. 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,

$$\min_{\gamma, \beta \geq 0} - \sum_{1}^{N} \log Pr(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\gamma}, \boldsymbol{\beta}) + \frac{\lambda_1}{2} \|\gamma\|_2^2 + \frac{\lambda_2}{2} \|\beta\|_2^2 . \quad (10)$$

As in section 3.1, 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,

$$\arg\min_{G=\text{CNN-CRF}} \arg\max_{D} \mathcal{L}_{\text{GAN}}(G_{\text{CNN-CRF}}, D) + \lambda \mathcal{L}_{\text{L1}}(G_{\text{CNN-CRF}}).$$

Implementation Details We implemented our encoder-decoder network and CNN-CRF model in PyTorch and

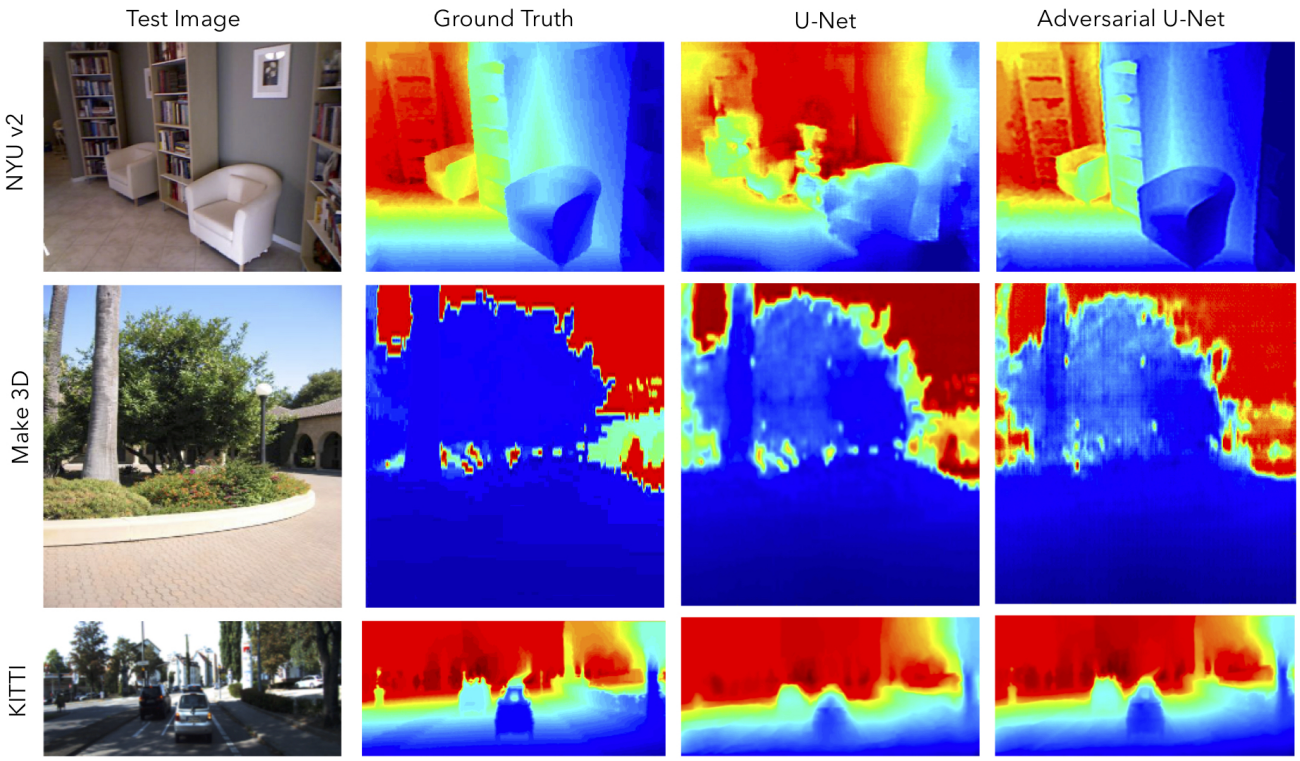

Figure 4: Representative images showing estimated depth on NYUv2, Make 3D and KITTI datasets via our proposed adversarial depth estimation paradigm. We demonstrate that model trained via adversarial training can predict relievedly more accurate depths..

Method	C1			C2		
	rel ↓	log_{10} ↓	rms ↓	rel ↓	$\log_{10}$ ↓	rms ↓
Make3D [48]	-	-	-	0.370	0.187	-
Liu *et al.* [29]	-	-	-	0.379	0.148	-
DepthTransfer [23]	0.355	0.127	9.20	0.361	0.148	15.10
Liu *et al.* [31]	0.355	0.137	9.49	0.338	0.134	12.60
Li *et al.* [27]	0.278	0.092	7.12	0.279	0.102	10.27
Liu *et al.* [31]	0.287	0.109	7.36	0.287	0.122	14.09
Roy *et al.* [46]	-	-	-	0.260	0.119	12.40
Laina *et al.* [26]	0.176	0.072	4.46	-	-	-
LRC-Deep3D [56]	1.000	2.527	0.981	-	-	-
LRC *et al.* [13]	0.443	0.156	11.513	-	-	-
U-Net [45]	0.428	0.142	5.127	0.446	0.164	6.38
Kuzietsov *et al.* [54]	0.421	0.190	8.24	-	-	-
MS-CRF *et al.* [57]	0.184	0.065	4.38	0.198	-	8,56
DORN (ResNet) [11]	0.157	0.062	3.97	0.162	0.067	7.32
CNN-CRF [30]	0.314	0.119	8.603	0.307	0.125	12.89
CNN-CRF Adv.	0.287	0.103	6.188	0.266	0.098	9.148
U-Net [45]	0.428	0.142	5.127	0.446	0.164	6.38
U-Net-Adv.	**0.0646**	**0.0277**	**1.812**	**0.0817**	**0.0493**	**4.163**

Table 2: Performance on Make3D. All methods are evaluated on the test split by Make3D [48]

Method	cap	abs rel ↓	squared rel ↓	rms ↓	$\text{rms}_{\log}$ ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Make3D [48]	0-80m	0.280	3.012	8.734	0.361	0.601	0.820	0.926
Eigen *et al.* [10]	0-80m	0.190	1.515	7.156	0.270	0.692	0.899	0.967
Liu *et al.* [31]	0-80m	0.217	1.841	6.986	0.289	0.647	0.882	0.961
LRC (CS+K) [13]	0-50m	0.114	0.898	4.935	0.206	0.861	0.949	0.976
Kuzietsov *et al.* [24]	0-50m	0.113	0.741	4.621	0.189	0.862	0.960	0.986
DORN (VGG) [11]	0-80m	0.081	0.376	3.056	0.132	0.915	0.980	0.993
DORN (ResNet) [11]	0-80m	0.072	0.307	2.727	0.120	0.932	0.984	0.994
CNN-CRF [30]	0-80m	0.217	1.793	7.421	0.261	0.656	0.881	0.958
CNN-CRF Adv.	0-80m	0.164	1.279	6.138	0.246	0.692	0.921	0.967
U-Net [45]	0-80m	0.097	0.411	3.349	0.158	0.626	0.832	0.902
U-Net Adv.	0-80m	**0.061**	**0.282**	**2.349**	**0.106**	**0.943**	**0.988**	**0.996**

Table 3: Performance on KITTI. All the methods are evaluated on the test split by Geiger et al. [12]

MatConvNet respectively [41, 52], and trained on Nvidia P100 GPUs using Google Cloud. Both networks trained for 150 epochs with a base learning rate of 0.0002 with ADAM optimization in both the generator and the discriminator, followed by a linear step decay for 150 epochs. Both networks were trained from scratch, and used Xavier weight initialization and Spectral normalization in the generator and the discriminator, with the input also normalized to be between [-1, 1]. To prevent mode collapse, patches were pooled and fed to the discriminator in batches rather than individual images in each iteration. A pooling history of randomly selected 50 patch pairs was used in the discriminator. Dropout was used the bottleneck layer of our U-Net architecture, but was not added in our CNN-CRF setup.

4. Experiments

To demonstrate the improvements made by adversarial training relative to state-of-the-art methods, which do not use adversaries, we evaluate our methods on three standard datasets for for depth estimation: NYU Depth v2 [51], Make3D [48], and KITTI [12]. We also perform an ablation study for comparative analysis with and without adversarial loss.

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 [51]. We worked with a 1449 aligned subset of images from NYUv2, with 795 pairs for training and 654 pairs for testing of resolution 640×320. During training, we downsampled the images to 386×288, and performed random horizontal flips with random crops of size 256×256. We report our scores on a pre-definend test and train split created by Eigen. [10].

Make3D. The Make3D Range Image Dataset contains image pairs of outdoor scenes (1704×2272) and ground truth laser depths (55×305), with 400 pairs for training and 134 images for testing. During training, we resized the images to 400×300, and performed similar random horizontal flips with random crops of 256×256 crop. On this dataset, we report $C1$ and $C2$ errors (depth ranges for $0 - 80$m and $0 - 70$m respectively) using the same model.

KITTI. The KITTI Vision Dataset contains image pairs of outdoor scenes (375×1241) and raw LiDaR scans for 61 scenes, ranging from "residential" to "city" scenes. We trained and tested with the 0-80 depth range for 32 scenes and 29 scenes respectively. In both splits, we preprocesed the images by resizing them to be 256×256.

Evaluation Metrics. Following previous works [26, 57, 29, 31], 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. Accuracy with threshold t: Percentage of y_i s.t. $\max(\frac{y_i^*}{y_i}\frac{y_i}{y_i^*} = \delta < t(t \in [1.25, 1.25^2, 1.25^3])$, where y_i is the estimated depth, y_i^* is the corresponding ground truth

In these series of comparisons, we evaluate the proposed adversarial depth estimation networks with their non-adversarial counterparts. We observed a decrease in relative error and increase in accuracy when an adversarial loss was added. In the U-Net vs. adversarial U-Net comparison, despite enforcing strong spatial consistency between the convolution and upsampling layers with long skip connections, we achieved relative errors of 0.114, 0.0646, and 0.061 on NYUv2, Make3D and KITTI respectively, which improves over the current state-of-the-art relative error by Xu *et al.*

[57]. In addition, Adversarial U-Net also improves over accuracy on these three benchmarks, which suggests that the addition of adversarial training helped U-Net learn depth cues and more context-aware features that preserved structural details of the original image. Qualitatively, we can see how regular U-Net produced blurry results on the NYUv2 and KITTI datasets, however, after adding adversarial training, the network produced sharper edge details for objects in both the foreground and background. In contrast, 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 the CNN-CRF can be attributed to how using a $\mathcal{L}_{\text{L1}}$ loss directly optimizes for the relative error rather than accuracy, which penalizes greater deviations in per-pixel predictions from the ground truth. In addition, the relatively low accuracy of adversarial CNN-CRF may be due to the small training set used to train CNN-CRF and the fact that the loss was computed on the super-pixel level rather than the entire image.

5. Conclusions

In this paper, we demonstrate the effectiveness of adversarial training for monocular depth estimation from a single image across two kinds of neural network architectures: encoder-decoder style U-Net and joint CNN-CRFs. Our method approaches the depth estimation problem by incorporating an adversarial loss which captures non-local information as compared to local per-pixel losses. Unlike more complex multi-scale, deep architectures used for capturing global interactions for understanding local and non-local context, our approach is relatively more simple and robust. The global information is incorporated by a discriminator which aims to discriminate patches of an estimated depth prediction as real or fake. In our findings, we show how adversarial training can improve depth predictions as compared to state-of-the-art methods. This improvement is particularly pronounced for U-Net which performs weakly by itself, but outperforms state-of-the-art when adversarial loss is used.

6. Acknowledgements

This work was supported in part with funding from the NIH NIBIB Trailblazer Award (R21 EB024700).

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] J. Aloimonos. Shape from texture. *Biological Cybernetics*, 58(5):345–360, 1988.

[3] H. H. Baker and T. O. Binford. Depth from edge and intensity based stereo. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI'81)*, IJCAI'81, pages 631–636, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.

[4] S. T. Barnard and M. A. Fischler. Computational stereo. *ACM Computing Surveys*, 14(4):553–572, Dec. 1982.

[5] 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. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] 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.

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

[8] 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.

[9] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. The Importance of Skip Connections in Biomedical Image Segmentation. pages 179–187, 2016.

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

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

[12] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[13] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[14] I. Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016.

[15] 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 27 (NIPS 2014)*, pages 2672–2680, 2014.

[16] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000.

[17] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Proceedings of the 13th Asian Conference on Computer Vision (ACCV 2016)*, pages 213–228. Springer, 2016.

[18] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 6626–6637, 2017.

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

[20] W. Hong, Z. Wang, M. Yang, and J. Yuan. Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1335–1344, 2018.

[21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[22] J. Jiao, Y. Cao, Y. Song, and R. Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *Proceedings of the 2018 IEEE European Conference on Computer Vision (ECCV)*, September 2018.

[23] 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, 2014.

[24] Y. Kuznietsov, J. Stückler, and B. Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

[26] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *Proceedings of the 4th International Conference on 3D Vision (3DV 2016)*, pages 239–248. IEEE, 2016.

[27] 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 to the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1119–1127, 2015.

[28] 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 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3372–3380, 2017.

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

[30] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5162–5170, 2015.

[31] 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.

[32] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016.

[33] D. C. Luvizon, D. Picard, and H. Tabia. 2d/3d pose estimation and action recognition using multitask deep learning, 2018.

[34] F. Mahmood, R. Chen, and N. J. Durr. Unsupervised reverse domain adaption for synthetic medical images via adversarial training. *IEEE Transactions on Medical Imaging*, 2018.

[35] F. Mahmood and N. J. Durr. Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. *Medical Image Analysis*, 2018.

[36] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[37] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, 2018.

[38] S. K. Nayar and Y. Nakagawa. Shape from focus: An effective approach for rough surfaces. In *Proceedings of the 1990 IEEE International Conference on Robotics and Automation (ICRA)*, pages 218–225. IEEE, 1990.

[39] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning (PLMR 2017)*, Proceedings of Machine Learning Research, pages 2642–2651, 2017.

[40] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 15(4):353–363, Apr. 1993.

[41] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[42] A. Pilzer, D. Xu, M. P. Marian, E. Ricci, and N. Sebe. Unsupervised adversarial depth estimation using cycled generative networks. In *Proceedings of the 6th International Conference on 3D Vision (3DV 2018)*. IEEE, 2018.

[43] E. Prados and O. Faugeras. Shape from shading. In *Handbook of mathematical models in computer vision*, pages 375–388. Springer, 2006.

[44] K. Regmi and A. Borji. Cross-view image synthesis using conditional gans. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3510, 2018.

[45] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the 18th International Conference on Medical*

Image Computing and ComputerA sisted Intervention (MICCAI 2015), pages 234–241. Springer, 2015.

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

[47] 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 29 (NIPS 2016)*, pages 2234–2242, 2016.

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

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

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

[51] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the 2012 IEEE European Conference on Computer Vision (ECCV)*, pages 746–760. Springer, 2012.

[52] A. Vedaldi and K. Lenc. Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 689–692. ACM, 2015.

[53] J. Wang, X. Li, and J. Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1788–1797, 2018.

[54] 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 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2800–2809, 2015.

[55] D. Weinshall. Application of qualitative depth and shape from stereo. In *Proceedings of the 1998 IEEE International Conference on Computer Vision (ICCV)*, pages 144–148. IEEE, 1988.

[56] J. Xie, R. Girshick, and A. Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *Proceedings in the 2016 IEEE European Conference on Computer Vision (ECCV)*, pages 842–857. Springer, 2016.

[57] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe. Monocular depth estimation using multi-scale continuous crfs as sequential deep networks. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[58] 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 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[59] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[60] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. 2017.