DEPTH PREDICTION FROM A SINGLE IMAGE WITH CONDITIONAL ADVERSARIAL NETWORKS

Hyungjoo Jung[1], *Youngjung Kim*[1], *Dongbo Min*[2], *Changjae Oh*[1], *and Kwanghoon Sohn*[1]

[1]School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
[2]Department of Computer Science and Engineering, Chungnam National University, Daejeon, Korea

ABSTRACT

Recent works on machine learning have greatly advanced the accuracy of depth estimation from a single image. However, resulting depth images are still visually unsatisfactory, often producing poor boundary localization and spurious regions. In this paper, we formulate this problem from single images as a deep adversarial learning framework. A two-stage convolutional network is designed as a generator to sequentially predict global and local structures of the depth image. At the heart of our approach is a training criterion based on adversarial discriminator which attempts to distinguish between real and generated depth images as accurately as possible. Our model enables more realistic and structure-preserving depth prediction from a single image, compared to state-of-the-arts approaches. An experimental comparison demonstrates the effectiveness of our approach on large RGB-D dataset.

Index Terms— Depth from a single image, deep neural network, generative adversarial learning, RGB-D database.

1. INTRODUCTION

Predicting 3D structure from a single monocular image has remained an active research topic in computer vision. This can be attributed to the fact that depth information often leads to significant improvements on a number of challenging vision applications, including robotics [1], intrinsic image decomposition [2], pose recognition [3], and scene understanding [4]. Traditional approaches to estimating a depth map from a single image exploited various monocular cues like parallax, motion [13], or shading [5]. However, strict assumptions imposed on the depth prediction model limit their application to some restricted environments, e.g., translational camera motion and static scenes. Meanwhile, the human visual system (HVS) has no difficulty in perceiving depth from a monocular input, thanks to long accumulated experience and data [6]. Machine learning approaches to replicate this capability would open new eras, motivating various state-of-the-art methods [11]– [16].

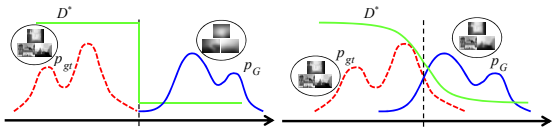

Fig. 1. When G is very poor, D can reject samples with high confidence. In this case, the adversarial model of (1) may not provide sufficient information for G to learn well (left). We propose two-stage learning process, so that p_G and p_{gt} are initially to be concentrated on similar manifolds (right). D^* denotes the optimal Bayes discriminator between p_G and p_{gt}.

We roughly classify existing methods into two categories: non-parametric sampling and parametric learning methods. The first group addresses the question of whether it would be possible to correctly transfer depth from a large RGB-D database to a single query image. Karsch *et al.* [7] devised the depth transfer algorithm using candidates method from a training dataset. The retrieved candidates are densely warped to the input query image via SIFT Flow [9], and they are then fused for final depth estimation. Contrary to [7], Konrad *et al.* directly fused matching candidates by computing a median depth value for each pixel [8]. A joint bilateral filtering [10] is then applied to align depth discontinuities at the input query image. Instead of selecting depth values from training data, the depth analogy [11] transfers depth gradients as reconstruction cues, followed by the Poisson reconstruction. However, these methods usually assume that pixels classified similarly using hand-crafted descriptors are likely to have similar values. This assumption is often violated in real-world data, limiting their applications.

The second category casts the monocular depth estimation as a parametric learning process. Saxena *et al.* learned a Markov Random Field (MRF) for mapping between RGB color and depth spaces [12]. Recently, convolutional neural networks (CNNs) have been applied with great success in a single image depth estimation. Liu *et al.* [14] learned unary and pairwise potentials with the CNNs, and explicitly modeled the relation among neighboring superpixels in a continuous conditional random fields (CRF). Eigen *et al.* applied the

This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. R0115-16-1007, High quality 2d-to-multiview contents generation from large-scale RGB+D database).

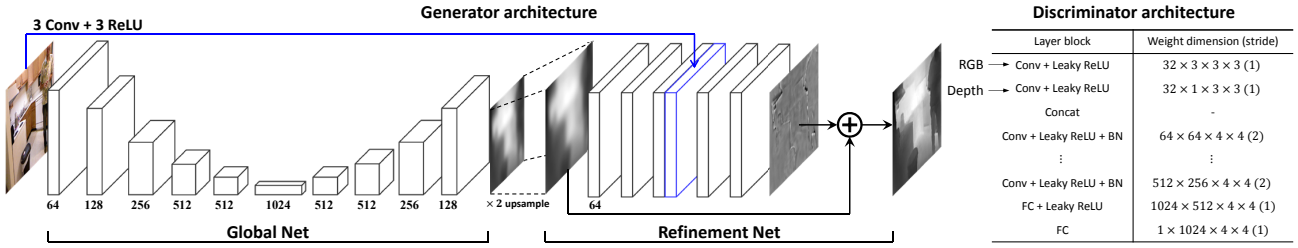

Fig. 2. The generator G consists of two major components: the global net and the refinement net (left). The global net is first pre-trained with the single L_1 loss function. In the refinement net, we model high-frequency structure of the depth image using the L_1 and adversarial loss functions. We follow the architectural guidelines introduced in [20] for the discriminator D (right).

CNNs in multiple stages to sequentially generate features and refine the depth prediction to a higher spatial resolution [15]. In [16], a set of relative depth annotations rather than metric depth are used to improve depth perception in an unconstrained setting. The CNNs-based approaches achieved state-of-the-art performance on benchmarks, but the results often lack fine details and are still perceptually unsatisfying.

It would be desirable to make the depth prediction results indistinguishable from natural depth images. To this end, we explore a generative adversarial network (GAN) [17]. We learn a discriminative model to determine whether the prediction is *natural* or not, while training a generative model to generate a faithful depth image. We demonstrate that our model is able to generate realistic and structure-preserving depth image from a single image, without using any low-level segmentation or superpixels.

The primary contributions of our work are the following:

- We present a generative adversarial network for single image depth estimation. Experimental results demonstrate that the proposed method outperforms the-state-of-the-arts significantly.
- We design a fully convolutional multi-scale network with a symmetric architecture, which sequentially estimates global and local structures of depth image.

2. PROPOSED METHOD

Our goal is to infer a perceptually plausible depth image d from a single RGB image I. We specifically design a deep adversarial network, so that the estimates are forced to have statistical characteristics of the natural depth image. The detailed network architectures will be described in Section 2.2.

2.1. Preliminaries: Adversarial Network

In the conditional setting, we briefly introduce a deep adversarial model from the perspective of single image depth estimation. Let us consider a discriminator network $D(\cdot)$. The adversarial model alternatively optimizes D[1] along with a generator network $G(\cdot)$ to solve the following min-max problem [17, 23]:

$$\min_G \max_D \mathbb{E}_{I\sim p_G}[\log(1 - D((I, G(I))))] + \mathbb{E}_{I',d'\sim p_{\text{gt}}}[\log D((I', d'))] + \lambda \mathbb{E}_{I\sim p_G, d\sim p_{\text{gt}}}[\|G(I) - d\|_1], \quad (1)$$

where (I, d) and (I', d') pairs are sampled from RGB-D dataset, and λ is a relative weighting factor. p is a conditional probability distribution that has class 1 when a depth map is from the ground-truth RGB-D dataset, or class 0 from G. (1) consists of an adversarial loss (the first two terms) and a standard pixel-wise reconstruction loss (the last term). The discriminator D takes as inputs the depth image from the generator G or the ground-truth depth image with I, and determines whether these are network output or not. The formulation of (1) enables one to train the generator network G, deceiving the discriminator network D. In this manner, we can produce results that are highly similar to the natural depth images or are indistinguishable by D. Technical details for solving the min-max problem (1) will be described in Section 2.3.

2.2. Network Architecture

Here, we describe the proposed network in detail. The overall network architecture is illustrated in Fig. 2. Note that the discriminator is used during training only.

2.2.1. *Generator network*

The generator G implicitly defines a probability distribution p_G as the distribution of the samples $G(I)$. If p_G and p_{gt} are concentrated on very different manifolds, the discrimination problem between them becomes trivial. In this case, $\log(1 - D((I, G(I))))$ is saturated regardless of the result of G (the left part of Fig. 1). Keeping this in mind, we propose a two-stage learning process for G. The generator consists of two main components: a global net and a refinement net. The first component is a pair of encoder-decoder networks [24], as shown in Fig. 2 (left). It takes the single image I as an input and generates the initial depth image. For the encoder,

[1]The discriminator D is trained to distinguish samples from the ground-truth distribution p_{gt} and the generative distribution p_G [17].

we gradually reduce the spatial resolution with 2×2 max-pooling (stride 2) while doubling the number of channels. The decoder part predicts the initial depth estimate through a sequence of deconvolutional (a factor of 2) and convolutional layers. The output resolution of the global net is half the input image.

The refinement net maps the predictions of global net to the full resolution. It takes the input as the bilinearly-upsampled ($\times 2$) output of the global net. To provide structural guidance into the refinement net, a feature map[2] from RGB input is concatenated with the third convolutional layer of refinement net (the blue box in Fig. 2). A key aspect of this network is that it does not directly estimate the high-resolution depth images, but the residual to the input (see Fig. 2). This approach is especially beneficial to convolutional networks, as it is not needed to carry the input information through the whole network [19]. Note that our generator is fully convolutional, and thus more flexible than the previous multi-scale network [15]. We can apply the generator G to an input image of arbitrary size.

2.2.2. Discriminator network

Following the architectural guidelines introduced in [20], we build discriminator as in Fig. 2 (the right table). It uses LeakyReLU activation (0.2 slope) and four convolutional layers. The strided convolution is adopted to reduce the spatial resolution instead of max-pooling. We also add a batch normalization layer to the output of every convolutional layer. The discriminator D outputs a single scalar, representing the probability that the input comes from the ground-truth rather than p_G.

2.3. Learning Adversarial Network

We train our model in two phases using stochastic gradient method (SGM): The global net is first pre-trained, and these parameters are fixed during learning the refinement net and discriminator. The convolutional layers in the global net are initialized using the Oxford VGG-net [22].

Given M training RGB-D pairs $\{I^{(p)}, d^{(p)}\}_{p=1}^{M}$, we apply the L_1 loss directly for the global net:

$$\mathcal{L}_g = \frac{1}{M} \sum_p \|u_g^{(p)} - d^{(p)}\|_1, \tag{2}$$

where u_g denotes the output of global net. Though L_1 (or L_2) loss tends to produce blurry estimates on generation problems [25], this content loss captures the low frequency information reliably in many tasks.

After pre-training the global net, we jointly train the refinement net and the discriminator to capture high-frequency details by solving (1). The min-max problem is solved by alternatively applying gradient descent and gradient accent once. The discriminator D is trained by maximizing (1) with fixed G. The minimization problem in (1) only updates the parameters of refinement net by minimizing the following loss:

$$\mathcal{L}_r = \frac{1}{M} \sum_p \Big(\|u_r^{(p)} - d^{(p)}\|_1 + \mu \log(1 - D((I^{(p)}, u_r^{(p)}))) \Big), \tag{3}$$

where u_r is the output of refinement net ($\mu = 1/\lambda$). The adversarial loss, $\log(1 - D((I, u_r)))$, encourages our model to generate more natural and structure preserving depth predictions. The procedure for learning the refinement net and the discriminator is summarized in Algorithm 1. In practice, training the refinement net with the full resolution image is very expensive. We instead extract 64×64 local image patches from training set for learning adversarial network. As the global structure is already recovered in the global net, it is sufficient to focus the attention on reconstructing fine details of local patches [23].

Algorithm 1 Learning procedure for the refinement net and discriminator

for number of iterations **do**
  Sample minibatch of $\{u_r^{(1)}, \ldots, u_r^{(m)}\}$
  Sample minibatch of $\{d^{(1)}, \ldots, d^{(m)}\}$
  Update the discriminator D by maximizing:
  $\frac{1}{m} \sum_p \Big(\log(D(d^{(p)}) + \log(1 - D((I^{(p)}, u_r^{(p)}))) \Big)$.
  Sample minibatch of $\{(I^{(1)}, u_g^{(1)}), \ldots, (I^{(m)}, u_g^{(m)})\}$
  Update the refinement net by minimizing:
  $\frac{1}{m} \sum_p \Big(\|u_r^{(p)} - d^{(p)}\|_1 + \mu \log(1 - D((I^{(p)}, u_r^{(p)}))) \Big)$.
end for

3. EXPERIMENTS

3.1. Implementation details and parameters

The proposed network was implemented and trained using the MatConvNet library [26] with 12GB NVIDIA Titan. We train the global net with the 0.2 million NYU v2 [21] raw distribution using the same test split as in [15]. The input to the network is resized to 320×256[3]. We perform 65k SGM iterations in total (each with a batch size of 16). Then, we train the second stage using 0.8 million RGB-D patches (size of 64×64) extracted from the training set. In this phase, the outputs from the global net are appended to the minibatch. Before training the actual discriminator, the refinement net is first trained with $\mu = 0$ (for 10k iterations) to avoid undesired local optima. We then alternately update to the refinement net and discriminator based on Algorithm 1 with $\mu = 0.01$ (for 150k iterations). Note that while training the second stage we

[2] It is generated using three convolutional and ReLU layers given RGB input.

[3] The corresponding ground-truth label is resized to 160×128

Fig. 3. Qualitative results for depth prediction: (From left to right) RGB input, ground truth, DCNF [14], MSC [15], and ours.

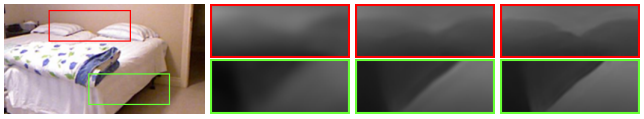

Fig. 4. The outputs of each network: (From left to right) RGB input, global net, refinement net without adversarial learning, and refinement net with adversarial learning($\mu = 0.01$).

keep all the weights of the global net fixed. In all cases, the momentum and weight decay parameters are set to 0.9 and 0.0005, respectively.

3.2. Results on NYU v2

We compare the proposed method against various recent state-of-the-art techniques, including depth transfer (DT) [7], DCNF [14], and MSC [15]. The first method is based on the non-parametric sampling and the others are CNNs-based approaches. We use several metrics from prior works for a quantitative evaluation:

- Threshold: $\max\left(\frac{d_i}{u_i}, \frac{u_i}{d_i}\right) = \delta < thr$
- abs rel: $\%$ s.t. $\frac{1}{N}\sum_i |d_i - u_i| / d_i$
- sqr rel: $\frac{1}{N}\sum_i \|d_i - u_i\|^2 / d_i$
- RMS(lin): $\sqrt{\frac{1}{N}\sum_i \|d_i - u_i\|^2}$

where u_i denotes the predicted depth at pixel indexed by i, and N is the total number of pixels. For 654 test images, the results are summarized in Table 1. The best results for each metric are highlighted in bold. It can be observed that the CNNs-based methods tend to have better performance than the non-parametric sampling [7]. We obtain the best performance in all metrics. The proposed method significantly outperforms the MSC [15] which was trained using the entire NYU v2 dataset [21] (approximately 0.5 million). This indicates the effectiveness of adversarial learning for monocular depth estimation. Note that the DCNF [14] and DT [7] were trained using 795 RGB-D pairs.

Table 1. Quantitative comparisons on the NYU v2 dataset [21]. The proposed method performs the best in all cases.

	DT [7]	DCNF [14]	MSC [15]	Ours
$\delta < 1.25$	-	0.614	0.769	**0.822**
$\delta < 1.25^2$	-	0.883	0.950	**0.971**
$\delta < 1.25^3$	-	0.971	0.988	**0.993**
abs rel	0.350	0.214	0.158	**0.134**
sqr rel	-	-	0.121	**0.103**
RMS(lin)	1.200	0.824	0.641	**0.527**

Qualitative results are shown in Fig. 3. Our results show substantial improvement especially on depth boundaries, and look more realistic. The DCNF [14] has sharper transitions thanks to superpixel segmentation, but includes false texture edges. The MSC produces blurry artifacts on depth boundaries and fails to capture main structures faithfully. We also report the effect of the adversarial learning in Fig. 4. The depth image estimated using the discriminator is superior in depth discontinuities to the result without adversarial learning.

4. CONCLUSION

In this work, we have presented a generative adversarial model for estimating depth from a single monocular image. We designed a fully convolutional generator under the assumption that the RGB and depth images share underlying structures. Simultaneously, we also learned the adversarial discriminator that distinguishes whether the prediction is natural or not. Experimental results demonstrate that the proposed method outperforms state-of-the-art methods on large RGB-D dataset. In particular, our model is able to generate realistic and discontinuity-preserving depth prediction without any low-level segmentation or superpixels.

5. REFERENCES

[1] C. Kerl, J. Sturm, and D. Cremers, "Robust odometry estimation for RGB-D cameras," in *IEEE Int. Conf. on Robotics and Automation*, pp. 3748-3754, 2013.

[2] S. Kim, K. Park, K. Sohn, and S. Lin, "Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields," in *Proc. Eur. Conf. Comput. Vis.*, pp. 143-159, 2016.

[3] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, and A. Kipman, "Efficient human pose estimation from single depth images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, 2013.

[4] S. Song, S.P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 567-576, 2015.

[5] R. Zhang, P.S. Tsai, J.E. Cryer, and M. Shah, "Shape-from-shading: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 57, no. 12, 2011.

[6] W. Zhou, M. Salzmann, X. He, and M. Liu, "Indoor scene structure analysis for single image depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 614-622, 2015.

[7] K. Karsch, C. Liu, and S.B. Kang, "Depth extraction from video using non-parametric sampling," in *Proc. Eur. Conf. Comput. Vis.*, pp. 775-788, 2012.

[8] J. Konrad, M. Wang, and P. Ishwar, "Learning-based, auto-matic 2d-to-3d image and video conversion," *IEEE Trans. Image Process.*, vol. 22, no. 9, 2013.

[9] C. Liu, J. Yuen, and A. Torralba, "SIFTFLOW: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, 2011.

[10] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint Bilateral Upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, pp. 96-100, 2007.

[11] S. Choi, D. Min, B. Ham, Y. Kim, C. Oh, and K. Sohn, "Depth Analogy: Data-driven approach for single image depth estimation using gradient samples," *IEEE Trans. Image Process.*, vol. 24, no. 12, 2015.

[12] A. Saxena, M. Sun, and A.Y. Ng, "Make3D: Learning 3D-scene structure from a single still image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, 2009.

[13] D. Kim, D. Min, and K. Sohn, "A stereoscopic video generation method using stereoscopic display characterization and motion analysis," in *IEEE Trans. on Broadcast.*, vol. 54, no. 2, pp. 188-197, 2008.

[14] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 5162-5170, 2015.

[15] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *IEEE Int. Conf. on Computer Vision*, pp. 2650-2658, 2015.

[16] W. Chen, Z. Fu, D. Yang, and J. Deng, "Single image depth perception in the wild," in *Advances in Neural Information Processing Systems*, 2016.

[17] I. Goodfellow, J.P. Abadie, and M. Mirza, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, pp. 2672-2680, 2016.

[18] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrel, and A.A. Efros, "Context Encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[19] J. Kim, J. Lee, and K. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[20] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *IEEE Int. Conf. on Learning Representations*, 2016.

[21] N. Silberman, P. Kohli, D. Hoiem, and R. Fergus, "Indoor segmentation and support inference from RGB-D images," in *Proc. Eur. Conf. Comput. Vis.*, pp. 746-760, 2012.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *IEEE Int. Conf. on Learning Representations*, 2015.

[23] P. Isola, J. Zhu, T. Zhou, and A.A. Efros, "Image-to-image translation with conditional adversarial networks," in *CoRR*, abs/1611.07004, 2016.

[24] H. Noh, S. Hong, and B. Han, "Learning deconvolutional network for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015.

[25] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for neural networks for image processing," in *arXiv*, 2015.

[26] http://www.vlfeat.org/matconvnet.