Depth Estimation from Monocular Image and Coarse Depth Points based on Conditional GAN

Yaoxin Li[1], Keyuan Qian[2, a], Tao Huang[2] and Jingkun Zhou[1]

[1]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[2]Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China

Abstract. Depth estimation has achieved considerable success with the development of the depth sensor devices and deep learning method. However, depth estimation from monocular RGB-based image will increase ambiguity and is prone to error. In this paper, we present a novel approach to produce dense depth map from a single image coupled with coarse point-cloud samples. Our approach learns to fit the distribution of the depth map from source data using conditional adversarial networks and convert the sparse point clouds to dense maps. Our experiments show that the use of the conditional adversarial networks can add full image information to the predicted depth maps and the effectiveness of our approach to predict depth in NYU-Depth-v2 indoor dataset.

1 Introduction

Depth estimation is a central problem to many industrial applications, such as simultaneous localization and mapping (SLAM), robotic systems, autonomous driving and augmented reality (AR). Recently, many literatures have utilized single RGB image to predict the depth due to the low-cost and practical value [1, 2]. Nevertheless, the monocular image depth estimation has its own limitation and is a well-known ill-posed problem, since the colour pixel in the image has inherent ambiguity to map as a depth value. And the current methods are far away from practical usage and low accuracy and unreliability of the single image depth estimation fail to apply to robotic utilizations such as obstacle detection.

Thanks to the invention of the depth sensors including LIDAR, stereo camera and time-of-flight based depth camera, we have some devices to directly measure the depth of the environment. However, such depth sensors have their own drawbacks: the limited scope, light sensitivity, high price for high depth accuracy about time-of-flight based depth camera (e.g. Kinect v2), and high-cost, extreme low resolution (with only a few lines resolution in vertical direction. E.g. Velodyne VLS-128) about LIDAR. As for stereo cameras, careful calibration and large amount of computation are required for precise estimation, which usually fails to estimate under certain circumstances. Owing to such disadvantages, we describe an approach based on conditional Generative Adversarial Network (GAN) to reconstruct the depth map to high resolution with the limitation of the depth sensors.

In this situation, we consider whether additional depth information can increase the depth accuracy and reduce the ambiguity. So, the goal of this paper is to reconstruct high resolution depth map from coarse point cloud and single RGB image, and we can get coarse depth point from low-cost LIDARs (e.g. Velodyne VLP-16), SLAM or visual-based odometry methods. In this paper, we introduce a conditional GAN to convert the sparse depth map to dense depth image like image-to-image translation. Conditional GAN learn a conditional generative model that tries to classify whether the image input is true or fake, and the effect of the blur can be weakened because of the dissimilarity among the real depth images and generated depth images. The main contribution of this paper are as follows. We construct a conditional GAN to reconstruct the high-resolution depth image from a single image additional with sparse depth points. And we conduct our experiments on NYU-Depth-v2 indoor dataset and demonstrate the effectiveness and potential usage of our depth estimation method.

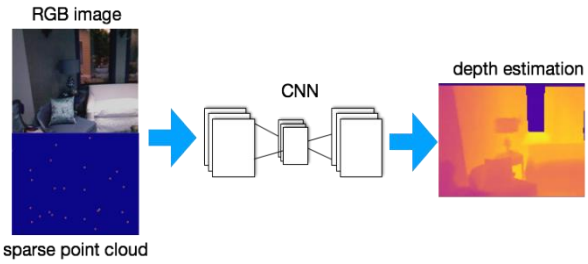


Figure 1. We propose conditional GAN for depth estimation from monocular RGB image and sampled sparse point cloud. Our method shows the effectiveness of the depth estimation.

2 Related work

2.1 Depth estimation from monocular images

* Corresponding author: qianky@sz.tsinghua.edu.cn

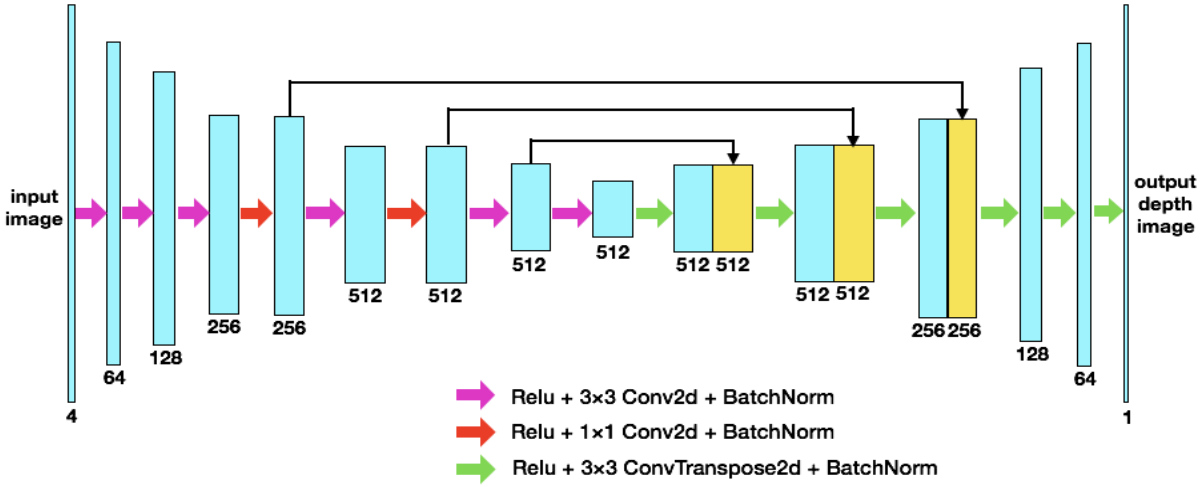


Figure 2. Generator network architecture

These days, with the development of convolutional neural network (CNN), depth estimation using single image has experienced a strong interest in both computer vision and robotic field. Deep learning-based method are proved to be more sufficient comparing to the traditional approaches, such as graphical models and hand-crafted features [3, 4]. Eigen et al.[5] first developed a multiscale convolutional neural network to learn three different tasks including depth prediction from a single image. The network regressed depth values from the input RGB-based image and caught many image details to investigate the validity of the depth prediction. Liu et al. [6] presented a deep convolutional neural field model combining the capacity of deep CNN with continuous conditional random field (CRF) to estimate depths from single monocular images.

More recently, Cao et al. [8] and Laina et al. [7] dealt with the problems by using Residual Neural Network (ResNet) [14]. Cao et al. converted depth estimation task to a pixel-wise classification task, while Laina et al. trained an end-to-end network to learn up-sampling feature maps without the help of refinement and post-processing approaches like CRF respectively.

2.2 Depth reconstruction with additional sparse samples

The advanced depth sensors invented in recently time inspire many researchers to predict dense depth maps from coarse point clouds. For example, Liao et al. [9] introduced 2D laser range data to construct a dense reference map by ResNet and combined the regression loss and classification loss to improve the depth accuracy. Ma. and Karaman. [10] learned directly from monocular image and additional sparse depth samples by a single regression network for depth prediction in scenes.

2.3 Conditional GAN

Another field of related work is Conditional GANs [11]. Our model utilizes conditional GAN to reconstruct dense depth map from RGB images and coarse depth points, and there are some other works use conditional GANs on label [11], text [13] and images. Ledig et al. [12] used conditional GANs for super-resolution on photo-realistic natural images. And Isola et al. [14] presented conditional adversarial networks to solve image-to-image translation problems such as reconstructing objects by edge maps and synthesizing images. Unlike previous work, we learn the mapping between the single image additional with coarse depth sample and dense depth image by conditional GAN.

3 Methodology

Our conditional GAN models learn a mapping from image x, sampled depth points d, and random noise vector z to dense depth image y: G: {x, d, z} →y. The generator is to produce images as similar as possible to the real images and cannot be distinguished by discriminator, while the discriminator is to detect whether the image is real or fake. In this section, we describe our depth point sample method, the architecture of our conditional GAN and the training stage.

3.1. Depth Point Sampling

In this part, we introduce the how we create the sparse depth point from the ground truth. In order to sample randomly in the depth ground truth, we design a probability model to get the sample points. For any target pixel, we calculate the Bernoulli probability $\mathrm{p} = \frac{n}{m}$, where m is the total number of the depth points, and n represents the number of the pixel we consider as the training data. And the formulation of any pixel (i, j) in image is as follows.

$$D(i,j) = \begin{cases} D^*(i,j), & with\ the\ probility\ p \\ 0, & with\ the\ probility\ (1-p) \end{cases} \quad (1)$$

Where $D^*(i,j)$ is the real depth in ground truth on pixel (i,j). By such sampling approach, we can get the sparse depth point cloud differently in each training step, and the points are uniformly distributed in the image. So, we can increase the robustness of our model using such sampled points as our training data.

3.2 Network Architecture

Our network includes a generator and a discriminator, and both of them follows the modules that contains Convolution-BatchNorm-Relu layers. Figure 2 shows the overall architecture of our model and the details of the architecture are discussed below.

3.2.1 Generator with skips

The goal of our generator is to map the low-resolution depth point clouds with a supplementary information about single RGB image to high resolution depth images. In this area, we design an encoder-decoder network to reach our aim. Our encoder part is developed based on CNN architecture with a kernel size of 3-by 3. Many previous methods have exploited ResNet Block [14] to down sample the input image and produced feature maps in different scales. Since there are some low-level features shared between the input and output images, we design to introduce such information directly to the up-sample layers. Thus, we add a skip connection between the down-sample layers and up-sample layers to avoid the bottleneck layer information, following the architecture 'U-net' [15]. As for decoder part, our model contains 5 up-sample layers including deconvolution, BatchNorm, Relu and dropout layers.

3.2.2 Discriminator

Since the L1 and L2 loss produces burry results on image, we design a discriminator architecture to reduce such effect. In spite of the failure in strengthening the high frequency features in images, L1 loss can encourage the low frequency features in images. Therefore, we create a discriminator network to punish the high frequency dissimilarity and try to classify whether the image is real or fake. Our discriminator architecture is much smaller than the generator architecture, only contains a few convolutional layers.

3.3 Training stage

The full objective of our model can be described as follows.

$$L_{cGAN}(G,D) = E_{x,y\sim P_{data}(x,y)}[logD(x,y)] + E_{x\sim P_{data}(x),z\sim P_z(z)}[log(1-D(x,G(x,y))] \quad (2)$$

And we add a more traditional pixel-level content loss, for example L1 loss, to this model and increase the performance of our model by making our generated image similar to the target images. Thus, our generator creates the images not only to fool the discriminator, but also to be as near as possible to ground truth in an L1 loss distance.

$$L_{L1}(G) = E_{x,y\sim P_{data}(x,y),z\sim P_z(z)}[\|y - G(x,z)\|_1] \quad (3)$$

Our final model is expressed as follows.

$$G^* = arg\ \min_G \max_D \lambda_A L_{cGAN}(G,D) + \lambda_B L_{L1}(G) \quad (4)$$

Where discriminator D tries to maximize the GAN loss, while the generator G tries to minimize the objective against D. λ_A and λ_B controls the relative weights of the two parts.

4 Experiments

We implement our model in python 3.4 and cuDNN 6.0 using pyTorch [16] and use NYU-Depth-v2 dataset to train our model. As for training, we use an NVIDIA TITAN xp with 12GB memory. The batch size of our model is 2. Different to the image-to-image translation problem, the model is not expected to produce uncertain depth value in the image. So, we increase the weight of L1 loss leading to the low relative weight on GAN loss. After numerous trials our model preforms well when λ_A=1, λ_B=1000. The initial learning rate of our model is 0.002. We use this learning rate in the first 150 epochs, and slightly decay to zero in the next 250 epochs.

4.1 Evaluation metrics

In this section, we introduce the evaluation metrics for our approach. The standard metrics for depth estimation evaluation are Root Mean Squared Error (RMSE), Mean Absolute Relative Error (MARE) and the percentage δ_i of the estimated pixels whose relative error are in a threshold δ_k.

$$\text{RMSE:} \quad \sqrt{\frac{1}{N}\sum_i^N(\overline{y_i} - y_i)^2} \quad (5)$$

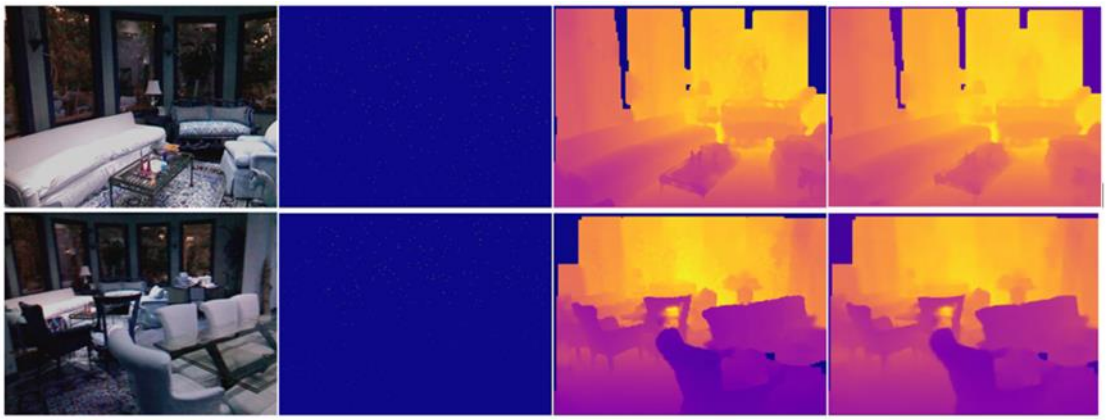
$$\text{MARE:} \quad \frac{1}{N}\sum_i^N \frac{|\overline{y_i} - y_i|}{y_i} \quad (6)$$

$$\delta_i: \quad \frac{\{max\left(\frac{\overline{y_i}}{y_i}, \frac{y_i}{\overline{y_i}}\right) < \delta^k\}}{\{y_i\}} \quad (7)$$

where $\overline{y_i}$ and y_i represent the predicted depth value and ground truth depth value respectively. And in threshold δ_i, $\delta = 1.25$ and $k = 1, 2, 3$.

4.2 Results

NYU-Depth-v2 dataset consists of RGB image and depth images captured by Microsoft Kinect on 464 different indoor scenes. We use 300 of the scenes for training and 164 of the scenes for testing. And Figure 3 shows some examples of our results. (a), (b), (c) are RGB images, coarse depth point cloud images and ground truth depth maps separately. And Figure 3 (d) are our depth prediction results.



(a) RGB image (b) coarse depth points (c) ground truth (d) prediction

Figure 3. Depth prediction outputs of our method on NYU-Depth-v2 indoor dataset.

In the evaluation part, we first present the performance of our method with different pixel-level content loss, such as L1 loss, L2 loss and smooth L1 loss, we compare the result on different content loss with the fixed number of sampled points (200 points). In order to compare the loss function, we use the same generator network architecture and the same weights of the ratio between the adversary loss and the content loss. Then we evaluate the relations between the number of the depth samples and the depth map estimation accuracy. We train a network for each different input number of depth samples and optimize each sampled number separately.

The conditional GANs are trained with an RGB image an average of 200 depth sampled points. The results are listed on Table 1 and Figure 4 and we compare our method with the existing approaches. From Table 1 we know that our method is better than the prior approaches using RGB images only and outperforms the previous method using 225 depth points on RMSE, MARE and the percentage δ_i. And we can also learn that the L2 loss and Smooth L1 loss will add the blurry to the depth images and cannot perform as well as using L1 loss.

Table 1. The comparison of different methods and different loss functions

Samples	Method	RMSE	MARE	δ_1	δ_2
0	Roy et al. [17]	0.744	0.187	-	-
0	Eigen et al. [5]	0.641	0.158	76.9	95.0
0	Laina et al [7]	0.573	0.127	81.0	95.3
225	Liao et al. [9]	0.442	0.104	87.8	96.4
200	Ours-L1 loss	**0.256**	**0.046**	**98.3**	**99.7**
200	Ours-L2 loss	0.943	0.572	99.5	99.6
200	Ours-smoothL1 loss	1.278	0.726	95.7	95.8

In the next step, we compare the impact on the number of the depth sample points. We express the relationship between the number of the sampled depth points and the depth images estimation accuracy. From Figure 4 we can conclude that with the increase of the number of depth points input to the network, the depth prediction accuracy of the depth maps has raised, and the error has rapidly diminished. And we can see from the figure that our RMSE when a set of 1000 the depth sampled points are inputted to the network, the RMSE can decrease to 0.2m. Thus, our method can be applied on LIDAR super-resolution tasks as well as SLAM odometry algorithms.

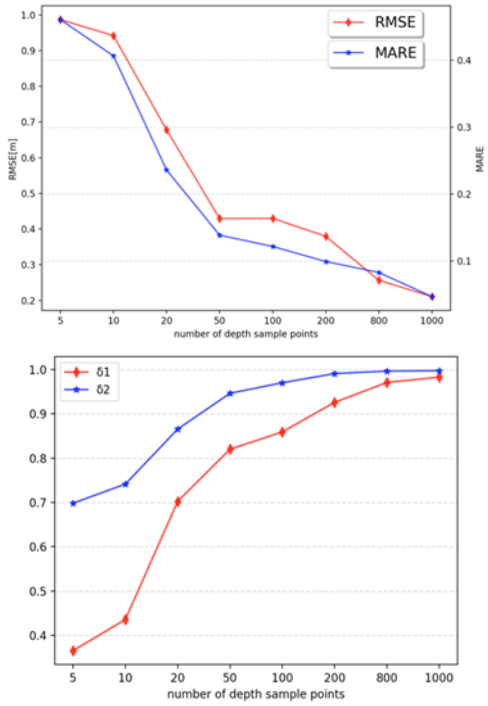


Figure 4. The influence of the number of the sampled depth points on the evaluation metrics

5 Conclusions

In this paper, we introduce a novel depth estimation method for dense depth maps from monocular RGB images and coarse depth point clouds. By offering supplementary information of sparse depth point value, our approach can relieve the ambiguity and unreliability of the methods predicting depth from single RGB images. And by using the conditional GAN, our approach can reduce the obscure depth values produced by regression model. We conduct the experiments on NYU-Depth-v2

indoor dataset, and our experiments shows the effectiveness of our method. We demonstrate our method can outperform many other depth estimation method and can be used to many engineering applications such as SLAM and robotic systems.

Acknowledgments

This study was financially supported by Basic Research Priorities Program of Shenzhen (Grants JCYJ20160608170030295).

References

1. Mancini, M., Costante, G., Valigi, P., & Ciarfuglia, T. A. (2016, October). Fast robust monocular depth estimation for obstacle detection with fully convolutional networks. In Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on (pp. 4296-4303). IEEE.
2. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., & Navab, N. (2016, October). Deeper depth prediction with fully convolutional residual networks. In 3D Vision (3DV), 2016 Fourth International Conference on (pp. 239-248). IEEE.
3. Saxena, A., Chung, S. H., & Ng, A. Y. (2006). Learning depth from single monocular images. In Advances in neural information processing systems (pp. 1161-1168).
4. Saxena, A., Sun, M., & Ng, A. Y. (2009). Make3d: Learning 3d scene structure from a single still image. IEEE transactions on pattern analysis and machine intelligence, 31(5), 824-840.
5. Eigen, D., & Fergus, R. (2015). Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2650-2658).
6. Liu, F., Shen, C., Lin, G., & Reid, I. (2016). Learning depth from single monocular images using deep convolutional neural fields. IEEE transactions on pattern analysis and machine intelligence, 38(10), 2024-2039.
7. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., & Navab, N. (2016, October). Deeper depth prediction with fully convolutional residual networks. In 3D Vision (3DV), 2016 Fourth International Conference on (pp. 239-248). IEEE.
8. Cao, Y., Wu, Z., & Shen, C. (2017). Estimating depth from monocular images as classification using deep fully convolutional residual networks. IEEE Transactions on Circuits and Systems for Video Technology.
9. Liao, Y., Huang, L., Wang, Y., Kodagoda, S., Yu, Y., & Liu, Y. (2017, May). Parse geometry from a line: Monocular depth estimation with partial laser observation. In Robotics and Automation (ICRA), 2017 IEEE International Conference on (pp. 5059-5066). IEEE.
10. Ma, F., & Karaman, S. (2017). Sparse-to-dense: Depth prediction from sparse depth samples and a single image. arXiv preprint arXiv:1709.07492.
11. Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.
12. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., ... & Shi, W. (2016). Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint.
13. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396.
14. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
15. Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.
16. Ketkar, N. (2017). Introduction to pytorch. In Deep Learning with Python (pp. 195-208). Apress, Berkeley, CA.
17. Roy, A., & Todorovic, S. (2016). Monocular Depth Estimation Using Neural Regression Forest. Computer Vision and Pattern Recognition (pp.5506-5514). IEEE.