3D-Imaging using Monocular Depth-Prediction Networks

Abstract

Correct depth estimation plays a vital role in video understanding, scene recognition, simultaneous localization and mapping (SLAM), and 3D reconstruction. Applications based on depth estimation have excellent development potential in many fields, such as archaeology, agriculture, autonomous vehicles. Common methods to measure the depth involve epipolar geometry algorithms based on stereovision, deep learning-based methods, and imaging radar or Time of Flight (ToF) sensors. In this paper, a convolutional neural network (CNN) with encoder-decoder as the architecture is implemented for obtaining the depth information from a monocular RGB image. The framework uses densenet as the pre-trained base network. With the help of Transfer Learning and training strategies, accurate depth prediction of high-resolution input images is made. The network has the advantages of narrower layers, low parameters requirement, and low computational cost while using the model. The model has remarkable performance and generates quantitatively good results on NYU open datasets.

Keywords—Computer vision, monocular depth estimation, deep learning network, densenet, depthwise separable convolution

1. Introduction

Depth estimation plays a fundamental role in computer vision. The concept of depth prediction refers to the processes of keeping 3D information of the scene with the input of 2D information captured by cameras.

The classical depth prediction approaches rely heavily on epipolar geometry, which is also called the multi-view method. Multiple photos from different views are captured simultaneously by placing cameras parallelly. Depth information is estimated after the feature comparison of the object in different views. The results generated by this method usually are very accurate. However, it is limited by the size of the equipment, the high price, the relatively large computational time, and storage requirements [1]. Also, for some objects with smooth surfaces or simple textures or in some scenes with special light conditions, this method often requires algorithms to compensate for errors [2], [3].

The basic idea of radar and ToF sensors is to determine an object's 3D shape by sending out a radio wave signal and detecting the direction and delay of the reflected signal [4]. However, such sensors' limitations are also obvious: apart from the expensive commercial sensor and the operational skill requirement, this method also has relatively low scan resolution and short perception depth.

Since a large amount of data is generated on the consumer side, methods like ToF sensors or multi-view-based technologies to achieve depth prediction are not suitable for consumer products. It has been shown recently that monocular depth prediction methods could be a potential solution to meet the needs [5]. These solutions perform well in efficiency and do not require specific equipment, complex setting up, and operational skills.

Recent developments on depth prediction focus on using deep learning techniques, such as convolutional neural networks (CNNs), to achieve 2D to 3D data mapping, more specifically, to explore the relationship between depth and pixel relevance via deep learning. Although these methods still have room for improvement: it always takes a long time to train models, it is undeniable that deep learning-based monocular depth estimation has excellent advantages in ease of use, efficiency, and lower cost [6]

2. Proposed Method & Implementation

Architecture

Figure 1 shows an overall architecture of our encoder-decoder framework for depth estimation from a single image. On the encoder side, a monocular RGB high-resolution image is encoded by pretrained feed-forward network - DenseNet-169 into a vector of features [7]. The result of the depth map is generated at half of the original resolution after outputting from series of up-sampling layers [8]. The decoder part consists of these layers and their skip connections. Depth-wise separable convolutions, as a state-of-the-art method, also be used in the decoder part in order to decrease the requirement of parameters. Detailed loss functions and data augmentation strategies can be found in the paper.

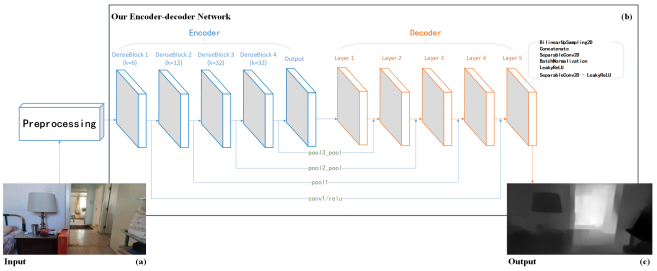

Figure 1 **Proposed Network Architecture.** From (a) to (c), they are respectively the input image, our proposed architecture and the output depth map. The encoder-decoder architecture is employed with skip connections. A pretrained state-of-the-art sub network, DenseNet-169, is adopted in the encoder part without modification.

Datasets

NYU Depth V2 is an open dataset published by Silberman *et al.* [9] that consists of video sequences (RGB Images and corresponding depth maps) captured from 464 different indoor scenes at the resolution of 640 × 480, see Figure 2. We feed our model with training samples of ground truth depth at half of the initial resolution (down-sampling to 320×240 without any trimming)

Figure 2 **NYU Dataset.** Images in the first row are examples and their depth maps from the training set. Images in the second row are from testing set for validation use.

3. Results

Qualitative results comparing to other deep learning-based depth estimation methods are summarized in Table 1 and demonstrated in Figure 3. Comparing to other deep learning models, it has been proven that our proposed network achieves a state-of-the-art performance in depth estimation.

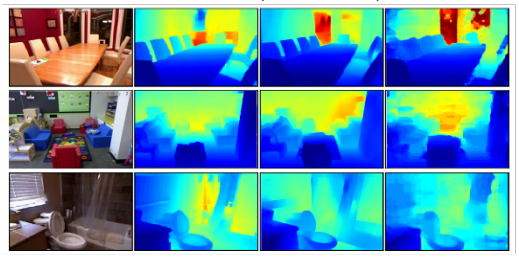

Figure 3 **Quantitative result comparison based on monocular data from NYU dataset.** From the left to the right: Original Image, Ground Truth, Result of our Encoder-Decoder Network, Result of Fu *et al.*'s Network [10]

Method	Evaluation Criteria					
	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	$Rel \downarrow$	$Rms \downarrow$	$log_{10} \downarrow$
Eigen et al. (2014)	0.769	0.950	0.988	0.158	0.641	-
Laina et al. (2016)	0.811	0.954	0.987	0.121	0.586	0.052
Xu et al. (2017)	0.811	0.953	0.988	0.127	0.573	0.055
Hao et al. (2018)	0.841	0.966	0.991	0.127	0.555	0.053
Fu et al. (2018)	0.828	0.965	0.992	**0.115**	**0.509**	**0.051**
Ours	**0.848**	**0.973**	**0.993**	0.129	0.541	0.054

Table 1 **Performance Evaluation and Comparison[10][11][12][13][14]**

4. Conclusion

We implemented a convolutional neural network for monocular depth estimation. We proved the pre-trained model's high availability and showed the great potential of a basic encoder-decoder architecture with the latest technologies integrated. We applied multiple data cleaning and augmentation strategies, trained our model on NYU Depth V2 open dataset, and achieve a comparable performance of indoor depth prediction with other state-of-the-art approaches.

In future work, we will try other practical data augmentation strategies, develop the potential of encoder-decoder on its architecture to extend our framework, and finally deploy the model in the fields.

5. References

[1] H. Javidnia and P. Corcoran, "A Depth Map Post-Processing Approach Based on Adaptive Random Walk With Restart," IEEE Access, vol. 4, pp. 5509-5519, 2016.

[2] Z. Zhang, "A flexible new technique for camera calibration," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, (11), pp. 1330-1334, 2000.

[3] J. Heikkila and O. Silven, "A four-step camera calibration procedure with implicit image correction," in 1997.

[4] M. McMahon, "What is an Imaging Radar?" wiseGEEK, Sep-2020. [Online]. Available: https://www.wisegeek.com/what-is-an-imaging-radar.htm. [Accessed: 17-Jan-2021]

[5] S. Elkerdawy, H. Zhang and N. Ray, "Lightweight monocular depth estimation model by joint end-to-end filter pruning," 2019.

[6] F. Khan, S. Salahuddin and H. Javidnia, "Deep Learning-Based Monocular Depth Estimation Methods-A State-of-the-Art Review," *Sensors* (Basel, Switzerland), vol. 20, (8), pp. 2272, 2020.

[7] G. Huang et al, "Densely Connected Convolutional Networks," Proceedings of the *IEEE conference on computer vision and pattern recognition*. 2017.

[8] J. Lehtinen et al, "Noise2Noise: Learning Image Restoration without Clean Data," *arXiv preprint* arXiv:1803.04189. 2018 Mar 12

[9] N. Silberman et al, "Indoor segmentation and support inference from RGBD images," In *European conference on computer vision*, Springer, Berlin, Heidelberg,2012,pp.746-760.

[10] H. Fu, M. Gong, C. Wang, K. Batmanghelich and D. Tao, "Deep Ordinal Regression Network for Monocular Depth Estimation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018,pp. 2002-2011.

[11] D. Eigen, C. Puhrsch and R. Fergus, "Depth Map Prediction from a Single Image using a Multi-Scale Deep Network," 2014.

[12] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari and N. Navab, "Deeper Depth Prediction with Fully Convolutional Residual Networks," *2016 Fourth International Conference on 3D Vision (3DV),* Stanford, CA, USA, 2016, pp. 239-248.

[13] D. Xu, E. Ricci, W. Ouyang, X. Wang and N. Sebe, "Multi-scale Continuous CRFs as Sequential Deep Networks for Monocular Depth Estimation," *2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp.161-169

[14] Z. Hao, Y. Li, S. You and F. Lu, "Detail Preserving Depth Estimation from a Single Image Using Attention Guided Networks," *2018 International Conference on 3D Vision (3DV),* Verona, Italy, 2018, pp. 304-313

Contact

Leave me comment