Computer Science and Application 计算机科学与应用, 2019, 9(2), 250-255
Published Online February 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.92029

Hans汉斯

Image Depth Estimation Model Based on Fully Convolutional U-Net

Xiaokang Wang[1], Xiaoning Fu[1], Que Dong[2]

[1]School of Electromechanical Engineering, Xidian University, Xi'an Shaanxi
[2]Wuhan Guide Infrared Co., Ltd., Wuhan Hubei
Email: 763162755@qq.com

Received: Jan. 15th, 2018; accepted: Jan. 25th, 2019; published: Feb. 1st, 2019

Abstract

The problem of depth estimation from single image has been addressed. The mapping between a single image and the depth map is inherently ambiguous, and requires both global and local information. This paper presents a fully convolutional U-net whose encoder is pre-trained ResNet50 without fully connected layer or pooling layer, and uses residual up-sampling layers to enlarge the feature maps. Besides, skip connection is introduced, making the model U-net, to fuse global and local information. The network can be end-to-end trained.

Keywords

Depth Estimation, Fully Convolutional Network, Residual Up-Sampling Layers, Skip Connection

一种基于U型全卷积神经网络的深度估计模型

王小康[1]，付小宁[1]，董　憨[2]

[1]西安电子科技大学机电工程学院，陕西　西安
[2]武汉高德红外股份有限公司，湖北　武汉
Email: 763162755@qq.com

收稿日期：2019年1月15日；录用日期：2019年1月25日；发布日期：2019年2月1日

摘　要

本文解决了从单张图像估计深度信息的问题。单张图像与深度图之间的映射是是模棱两可的，它需要全局信息和局部信息。本文部署了一个全卷积U型神经网络，它用预训练的ResNet-50网络提取图像特征，然后用残差上采样模块将特征图恢复到深度图的尺寸大小，并且使用了跳跃链接，整个网络呈现U型，从而对全局信息和局部信息进行融合。整个网络可以进行端到端的训练。

文章引用: 王小康, 付小宁, 董憨. 一种基于 U 型全卷积神经网络的深度估计模型[J]. 计算机科学与应用, 2019, 9(2): 250-255. DOI: 10.12677/csa.2019.92029

关键词

单目深度估计，全卷积神经网络，残差上采样，跳跃链接

1. 引言

深度图估计是计算机视觉的一个基础性问题，它可以展现一个场景的一些几何关系。近几年提出了不少的从立体图像中估计深度的成功方法。如果有精确的图像匹配，深度的确可以从立体图像中估计出来。这方面已经被广泛研究过了，但是在实践中常常会遇到单目图像的深度估计的问题，图片有的是室内的，有的是室外的。因此在缺乏深度传感器的情况下估计深度信息是必不可少的。再者，深度信息对计算机视觉的其它任务有很大帮助，比如图像识别[1]与语义分割[2]。

众所周知，从单目图像中估计深度信息是一个不适定问题，因为一张 RGB 图像可能对应无限多种现实世界的场景，并且获取不到可靠的视觉线索。有很多工作都尝试着解决这个问题。近来，卷积神经网络(Convolutional Neural Networks, CNNs)常常用来学习像素与深度之间的一种内含的关系。本文实现了一个端到端可训练的 U 型 CNN 架构，结合残差网络，可以学习图像像素灰度值与对应深度图之间的映射。

2. 相关工作

2.1. 经典方法

经典方法依赖于对场景几何的强烈假设，依赖于手工制作的特征和概率图模型，而概率图模型利用图像的水平对齐几何信息或其他的几何信息。比如，Saxena *et al.* [3]利用线性回归和 MPF 从图像特征中预测出深度，之后又将该工作扩展到 Make3D 系统[4]。然而，这个系统依赖于图像的水平对齐。

2.2. 基于特征的映射方法

第二种相关工作是基于特征的，给定一张 RGB 图像，在 RGB-D 的数据集中找到最近邻的图相对，检索出来的深度图会被用来产生最后的深度图。Karsch *et al.* [5]利用 SIFT 流，之后用一种全局优化的方法，而 Konrad *et al.* [6]计算检索出来的深度图的中值，之后用交叉双边滤波来平滑。Liu *et al.* [7]将优化问题建模为连续和离散的可变势能的条件随机场 Conditional Random Field (CRF)。这些方法都基于这样一个假设：RGB 图像中的区域之间的相似性也意味着相似的深度线索。

2.3. 基于卷积神经网络的方法

近来，基于 CNN 的深度估计方法开始占据主流。由于这个任务跟语义分割很相近，所以大多数工作都基于 The Image Net Large Scale Visual Recognition Challenge (ILSVRC) [8]中最成功的架构。Eigen *et al.* [9]是第一个运用 CNN 来预测单目图像深度的。他们用了两个深度网络模块，第一个做全局粗糙的速度估计，第二个在局部改善预测结果。这个想法之后被扩展[2]，三个 CNN 网络栈被用来额外预测表面法

线、类别以及深度。另外一个提高预测深度图质量的方向是将卷积神经网络与图模型结合。Liu *et al.* [10]提出用一种 CRF loss 的方式在 CNN 训练过程中学习一元的和成对势能，这种方法没有利用几何先验就达到了最好的结果.这个想法行得通是因为深度值是连续的[11]。Li *et al.* [12]和 Wang *et al.* [13]用层次 CRFs 来细化 patch-wise 的 CNN 预测结果，从超分辨到像素级别。

2.4. 基于全卷积神经网络的方法

全卷积神经网络(Fully convolutional networks, FCN)在密集预测的问题中有令人满意的表现。[14]使用 FCN，并用 CRF 来进行后处理。除了传统的卷积层，[15]使用扩张卷积来有效增加神经元的感受野，而且不用增加模型的参数以及训练所需要的数据量；[16]在语义分割任务中使用转置卷积来上采样特征图和预测图，并提出 U-net 网络结构。Laina *et al.* [17]提出一种带有残差上采样块的全卷积残差神经网络。

3. 网络结构

本文的网络结构如图 1 所示，分别为特征提取部分和上采样部分。其中第一部分用 ResNet50，删除了最后的全局池化层和全连接层，并以预训练模型的权值进行初始化；第二部分使用残差上采样[17]，使得特征图逐步放大，最终输出预测的深度图。最后，使用跳跃连接，对第一部分的特征与第二部分的特征进行融合。

1) 残差上采样层

如图 2 所示，Unpooling 层可以把特征图里的每个值都映射到一个 2*2 的矩阵的左上角处，而该矩阵的其他位置均填充数值 0，后接一个 5*5 的卷积层。这样的结构可以把输入的特征图的尺寸增大，在本例中，则可以将尺寸加倍。然而，这样的结构会在数值 0 上花费过多的计算，并且使得输出的预测图造成棋盘效应(棋盘状伪影)。为了解决这个问题，本文引入残差上采样层(Fast up-projection layer [17])，用 4 个不同的小卷积，分别是 3*3、3*2、2*3 和 2*2，代替 5*5 的卷积层。如图 3 所示，在残差上采样层的每个分支上特征图分别经过 4 个不同的卷积核，再做插入操作，使得特征图的尺寸加倍，同时通道数减半。

2) 跳跃连接

由于卷积神经网络是层级结构，也就是说，越靠近输出端的神经元拥有越大的感受野，并且可以产生越抽象的特征，而越靠近输入端的神经元其感受野越小但是包含更多的边缘信息。基于这个前提，本文将网络第一部分的特征与第二部分的特征进行级联，这种跳跃式的连接既让梯度更有效地反向传播，又保留更多的边缘信息。

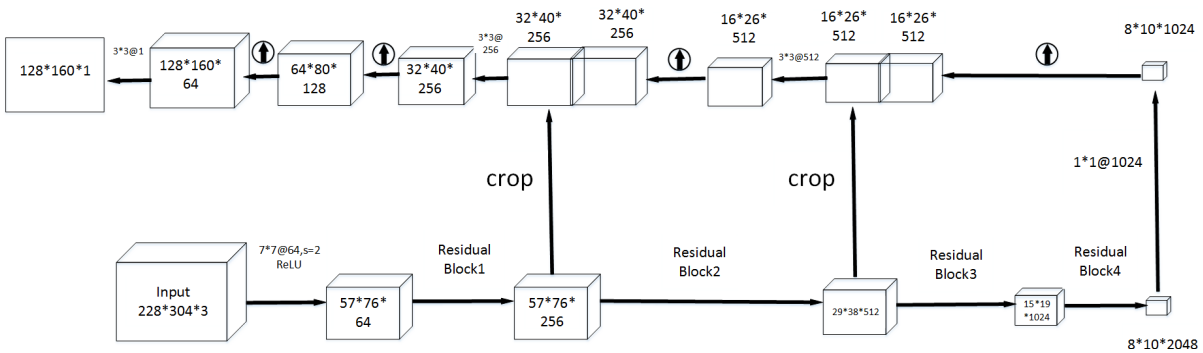


Figure 1. The architecture of the proposed network
图 1. 本文提出的网络结构

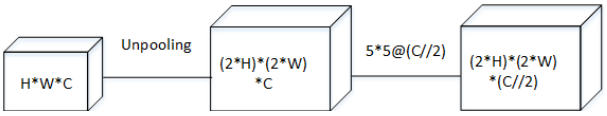


Figure 2. Up-sampling layer
图 2. 上采样块

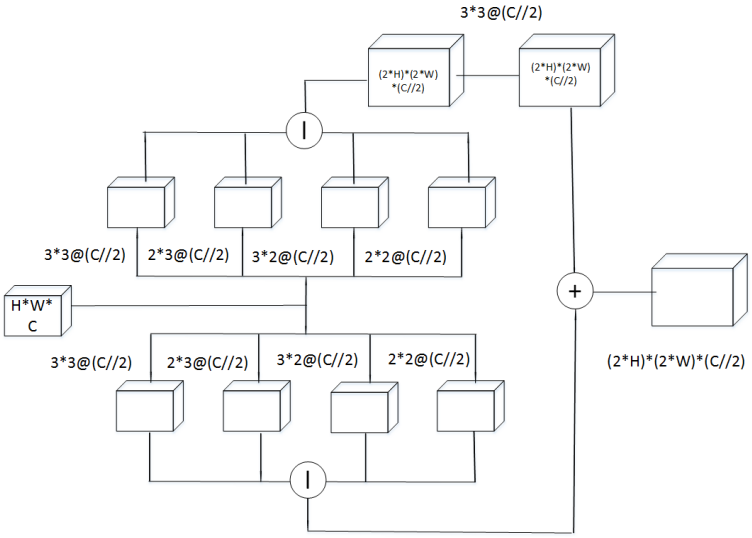


Figure 3. Fast up-projection layer
图 3. 残差上采样块

4. 实验

本文将提出的模型在 NYU V2 数据集[18]上训练和测试。包含 1449 个室内场景 RGB 图像及其对应的深度图，其中 795 个训练样本对，654 个测试样本对。在对训练集中拿出 10%作为验证集，剩余的训练样本当做开发集，并对开发集离线做了随机旋转、改变尺寸以及剪切，最后数据扩增到了大约 5000 个样本对，并且保证 RGB 图与深度图是同步变换的。在训练的过程中，又对训练集中每一对样本进行同步的随机左右翻转和随机改变 RGB 图的颜色，这样的在线随机变换相当于进一步离线扩增数据集而且不用占用额外的磁盘容量，最后可以推算出数据扩增到了 10,000 个样本对。

本文模型实现基于 PyTorch，电脑配置为 Intel core i5 处理器、NVIDIA GTX 1060 显卡和 8 GB 内存。使用 ADAM 优化器，学习率初始化为 0.003，并且每 5 个 epoch 将学习率乘于 0.1，权重衰减系数为 0.0001。

为了定量评估模型，本文使用如下指标。d 为深度标签值，$\hat{d}$ 为深度预测值，T 为图像中所有的像素的集合，δ 为准确率，threshold 为准确率阈值(分别取 1.25、1.25^2 和 1.25^3)。

$$\delta = \max\left(\frac{\hat{d}_i}{d_i}, \frac{d_i}{\hat{d}_i}\right) < threshold$$

平均绝对误差(rel)：$\frac{1}{|T|}\sum_{d\in T}\left|\hat{d}-d\right|/d$

平均 $\log_{10}$ 误差：$\frac{1}{|T|}\sum_{d\in T}\left|\log_{10}\hat{d}-\log_{10}d\right|$

均方根误差：$\sqrt{\frac{1}{|T|}\sum_{d\in T}\left\|\hat{d}-d\right\|^2}$

实验结果：

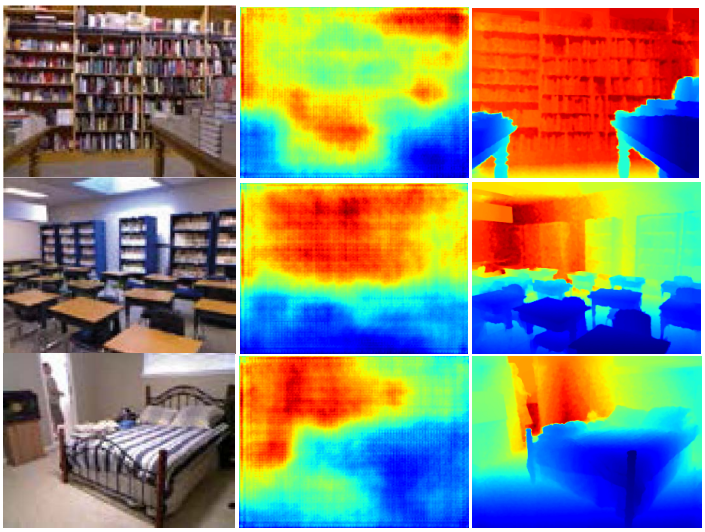
Table 1. Comparison of evaluation index on NYU v2 dataset
表 1. NYU v2 数据集上实验评价指标对比

方法	准确率(越高越好)			误差(越低越好)		
	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$	Rel	RMS	log10
Karsch [5]	-	-	-	0.374	1.12	0.134
Liu [7]	-	-	-	0.335	1.06	0.127
Eigen [9]	0.611	0.887	0.971	0.215	0.907	-
Liu [12]	0.614	0.883	0.971	0.230	0.824	0.095
本文方法	0.625	0.898	0.980	0.225	0.720	0.100

NYU V2 数据集上的实验结果如表 1 所示，通过结果对比可以得知本文算法结果更优。与 Liu [12] 的结果相比，本文模型的准确率分别提高 1.79%、1.7%和 0.9%，同时平均相对误差和均方根误差分别降低了 2.17%和 12.6%。

分析：

在测试集中随机选取 3 张图片用于训练好的模型进行深度估计，本文方法的结果如图 4 所示。与 Eigen 等[9]的方法比较，本文没有先粗略预测再细致调优的多尺度 CNN 结构，而是单尺度的网络；与 Liu 等[12]提出的方法相比，本文在像素级别进行了密集预测，并且不需要用 CRF 作为后处理方式对预测图进行细节优化。



(a) 原图 (b) 本文方法 (c) 标签图

Figure 4. The depth estimation of NYU v2 dataset
图 4. NYU v2 图像深度估计结果

5. 结论

本文提出了一种新的深度估计模型，以解决单目图像的深度估计问题。该模型利用预训练的 ResNet50 作为编码器，利用残差上采样块构造出解码器，并且利用跳跃连接将底层特征与高层特征融合起来，从而加速网络收敛和保留更多的边缘信息，是一个端到端的全卷积网络，大大减少了参数的个数与训练所

需的样本数量。与文献[5]、[7]、[9]和[12]相比，本文提出的模型误差更小、准确率更高。在未来的研究中，将会引入扩张卷积，利用这种卷积增加神经元的感受野，从而得到更优的局部特征。

参考文献

[1] Ren, X., Bo, L. and Fox, D. (2012) Rgb-(d) Scene Labeling: Features and Algorithms. 2012 *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Providence, 16-21 June 2012, 2759-2766.

[2] Eigen, D. and Fergus, R. (2015) Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture. *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, 7-13 December 2015, 2650-2658. https://doi.org/10.1109/ICCV.2015.304

[3] Saxena, A., Chung, S.H. and Ng, A.Y. (2006) Learning Depth from Single Monocular Images. *Advances in Neural Information Processing Systems*, **18**, 1161-1168.

[4] Saxena, A., Sun, M. and Ng, A.Y. (2009) Make3d: Learning 3d Scene Structure from a Single Still Image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**, 824-840. https://doi.org/10.1109/TPAMI.2008.132

[5] Karsch, K., Liu, C. and Kang, S. (2012) Depth Extraction from Video Using Non-Parametric Sampling. *Proceedings of the* 12*th European Conference on Computer Vision—Volume Part V*, Florence, 7-13 October 2012, 775-788. https://doi.org/10.1007/978-3-642-33715-4_56

[6] Konrad, J., Wang, M. and Ishwar, P. (2012) 2D-to-3D Image Conversion by Learning Depth from Examples. 2012 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Providence, 16-21 June 2012, 16-22. https://doi.org/10.1109/CVPRW.2012.6238903

[7] Liu, M., Salzmann, M. and He, X. (2014) Discrete-Continuous Depth Estimation from a Single Image. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 716-723. https://doi.org/10.1109/CVPR.2014.97

[8] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., *et al.* (2015) Imagenet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, **115**, 211-252. https://doi.org/10.1007/s11263-015-0816-y

[9] Eigen, D., Puhrsch, C. and Fergus, R. (2014) Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network. *Advances in Neural Information Processing Systems*, 2366-2374.

[10] Liu, F., Shen, C. and Lin, G. (2015) Deep Convolutional Neural Fields for Depth Estimation from a Single Image. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 5162-5170. https://doi.org/10.1109/CVPR.2015.7299152

[11] Liu, F., Shen, C., Lin, G. and Reid, I. (2016) Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**, 2024-2039. https://doi.org/10.1109/TPAMI.2015.2505283

[12] Li, B., Shen, C., Dai, Y., van den Hengel, A. and He, M. (2015) Depth and Surface Normal Estimation from Monocular Images Using Regression on Deep Features and Hierarchical CRFS. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 1119-1127.

[13] Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B. and Yuille, A.L. (2015) Towards Unified Depth and Semantic Prediction from a Single Image. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 2800-2809.

[14] Cao, Y., Wu, Z. and Shen, C. (2016) Estimating Depth from Monocular Images as Classification Using Deep Fully Convolutional Residual Networks. arXiv:1605.02305 [cs.CV]

[15] Li, B., Dai, Y., Chen, H. and He, M. (2017) Single Image Depth Estimation by Dilated Deep Residual Convolutional Neural Network and Soft-Weight-Sum Inference. arXiv:1705.00534 [cs.CV]

[16] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, 5-9 October 2015, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[17] Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F. and Navab, N. (2016) Deeper Depth Prediction with Fully Convolutional Residual Networks. 2016 *Fourth International Conference on* 3*D Vision* (3*DV*), Stanford, 25-28 October 2016, 239-248.

[18] Silberman, N., Hoiem, D., Kohli, P. and Fergus, R. (2012) Indoor Segmentation and Support Inference from RGBD Images. *Computer Vision— ECCV* 2012, Florence, 7-13 October 2012, 746-760. https://doi.org/10.1007/978-3-642-33715-4_54