文章编号:1671-4598(2019)02-0233-04 DOI:10.16526/j.cnki.11—4762/tp.2019.02.051 中图分类号:TP391.4 文献标识码:A

基于 DenseNet 的单目图像深度估计

何通能，尤加庚，陈德富
（浙江工业大学 信息工程学院，杭州 310023）

摘要：深度信息的获取是场景解析中是非常重要的环节，主要分为传感器获取与图像处理两种方法；传感器技术对环境要求很高，因此图像处理为更通用的方法；传统的方法通过双目立体标定，利用几何关系得到深度，但仍因为环境因素限制诸多；因此，作为最贴近实际情况的方法，单目图像深度估计具有极大研究价值；为此，针对单目图像深度估计，提出了一种基于 DenseNet 的单目图像深度估计方法，该方法利用多尺度卷积神经网络分别采集全局特征、局部特征；加入了 DenseNet 结构，利用 DenseNet 强特征传递、特征重用等特点，优化特征采集过程；通过 NYU Depth V2 数据集上验证了模型的有效性，实验结果表明，该方法的预测结果平均相对误差为 0.119，均方根误差为 0.547，对数空间平均误差为 0.052。

关键词：深度估计；卷积神经网络；多尺度；DenseNet

Depth Estimation from Single Monocular Images Based on DenseNet

He Tongneng, You Jiageng, Chen Defu
(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: Depth estimation is an important part in scene analysis, mainly composed of senor acquisition and image processing. Sensor technology requires demanding environment, thus image processing is a more general approach. The traditional way is using binocular stereo calibration to obtain depth information by geometric calculations, but it is still limited by environment factors. Therefore, as the closest approach to the actual situation, depth estimation from single monocular images has great research value. Consequently, a DensetNet based method is proposed, the method use multi-scale convolutional neural networks to acquire global features and local features. At the same time, it joins DenseNet structure for strong features propagation、features reuse to optimize features gathering. The experiments on NYU Depth V2 dataset demonstrate the effectiveness of this method. The average relative error of the prediction of this method is 0.119, the root mean squared error is 0.547, and the average $\log_{10}$ error is 0.052.

Keywords: depth estimation; convolutional neural network; multi-scale; DenseNet

0 引言

随着人工智能的迅猛发展，各类人工智能产品（如无人驾驶汽车、医疗机器人、巡检机器人）应运而生，在其工作过程中需要根据外界环境因素自动做出决策，通常的实现方法是利用计算机视觉技术对周围环境 3 维结构进行感知，实现 3 维重建，从而进行决策。因此，3D 场景解析是人工智能领域目前最火热、最重要的研究课题之一。3D 场景解析的重要基础为深度信息的获取，而单目图像获取深度信息是其中的重要方法。

当前获取场景深度信息主要有硬件实现与软件实现两种方法。硬件实现方法是利用传感器技术，例如微软开发的 3D 体感摄像机 Kinect，利用 ToF（time of fly）原理，通过给不可见光打码、测距光线强弱随时间变化等手段，根据光线的反射时间计算距离，特点是实时性好，算法开发工作量低。但是，尚未成熟的传感器技术导致输出图像的分辨率过低，仅仅适合室内小范围环境测量。

软件实现方法是通过图像处理获取深度信息。目前最为常见的为双目立体标定方法[1]，通过算法匹配左、右两幅图像中相同的特征点得到视差，再根据几何关系算出深度，但其匹配精度较低，受光线影响较大。在此基础上提出了结构光方法，其利用有编码的光源，主动将光源特征打在图像上进行匹配，精度有显著提升，但仍对环境光线等条件有一定要求。而单目图像对设备数量、环境条件要求较低，因此，通过单目图像获取深度信息是最贴近实际情况、应用场景最为灵活的方法。

利用单目图像获取深度信息本身是一个病态问题，一张图像理论上可以对应无限张深度图。因此，传统的单目图像获取深度信息多是基于物体的移动、相机焦距变化[2]等方法，将深度估计作为一个概率模型来处理，精度相对较低。SFM（Structure-from-Motion）是其中经典的方法之一[3]，通过相机不同时间间隔的抓拍图像，估计相机的姿态，进一步获取场景中物体深度信息。

随着近些年深度学习的发展，卷积神经网络（Convolutional Neural Networks，CNN）在图像处理、语音识别等领域发挥着越来越重要的作用。CNN 在特征提取、结果分类等方面无需人工干预，大大提高了模型的通用性。因此，学者们开始广泛使用 CNN 来研究单目图像的深度信息估计课题。鉴于 CRF（（Conditional Random Fields）在语义分割上的优异表现，Liu 等人[4]提出了将 CRF 应用到单目图

收稿日期:2018-09-03； **修回日期**:2018-09-26。

作者简介:何通能(1962-)，男，浙江省义乌人，副教授，主要从事模式识别与计算机智能控制方向的研究。

尤加庚(1994-)，男，浙江省永嘉人，硕士研究生，主要从事控制科学嵌入式方向的研究。

像深度估计的方法，与通常直接设定 CRF 的一元、二元势函数方法不同，他们将势函数本身也作为学习对象之一，通过训练系数优化势函数，从而优化训练结果，但其文中提到的超像素仍需手工实现划分。相比之下，Eigen 等人[5]提出了一种基于多尺度网络结构的深度学习方法，通过两个尺度对图像分别进行全局与局部的特征采样，获得最终输出，此方法无需提供任何人工分类特征，直接在原始图片上进行训练并获得像素级别的深度信息结果。在 Eigen 等人的基础上，Laina 等人[6]提出了一种应用 ResNet[7]的网络模型，该模型利用 ResNet 特征前向传递高效的特性，并结合更深、更复杂的网络结构，有效提高了结果精度。Godard 等人[8]提出了一种利用左右图一致性的无监督学习方法，该方法原理类似双目立体标定，利用左图样本得到深度图，再根据深度图与右图生成左图预测结果，与左图样本进行比较，实现了自编码、自解码，但上述方法在结果精度提升空间。

针对上述问题，在此提出了一种基于深度学习的单目图像深度估计方法，根据输入的 RGB 图像，直接得到图中各个像素对应的深度信息。本文的主要创新点是：首先，提出了一种新的 CNN 结构，对图像进行 3 个尺度的特征提取与融合，兼顾全局特征与局部特征，优化输出结果；其次，网络结构中融合了 DenseNet[9]，加强了特征前向传播，缓解了深层网络梯度消失问题，并且加强了特征重用，实现多级综合高效利用，同时还减少了参数数量。最后，通过实验证明，多尺度的网络结构与 DenseNet 的加入有效提升了输出深度图的精度。

1 网络模型

1.1 模型概述

为了进一步研究单目图像深度估计方法，本文提出了一种基于 DenseNet 的多尺度 CNN 网络模型。首先，将网络结构分为三个尺度，各个尺度对数据集进行不同程度的缩放，其中第一个尺度（Scale1）输入图像尺寸最大，第三个尺度（Scale3）最小。第一个尺度对图像特征进行全局粗糙采样，其输出结果与第二个尺度（Scale2）的输入图像尺寸相同。Scale2 的输入图像在原数据集基础上结合了 Scale1 的输出，Scale2 更注重图像中局部信息的采集，在上一尺度全局、粗糙的结果上进行优化，获得更具局部特征的结果。类似的，Scale3 的输入是由原数据集与 Scale2 的输出相结合，在进一步优化输出深度图的同时，提高了深度图的分辨率，获得高分辨率的输出结果。具体网络结构如图 1 所示。

1.2 多尺度网络结构

1.2.1 全局特征粗糙采样

Scale1 的主要目的是对图像进行全局的采样，在全局的层面提取图像特征，网络由 DenseNet 模块、上采样模块与卷积层组成。Scale1 的输入图像尺寸为 240×320，首先通过两个卷积层丰富信息特征采集，随后利用卷积核 3×3，步长为 2 的卷积层（简称 3×3 卷积层）代替传统的池化层对图片进行降采样，得到的结果作为 DenseNet 模块输入。

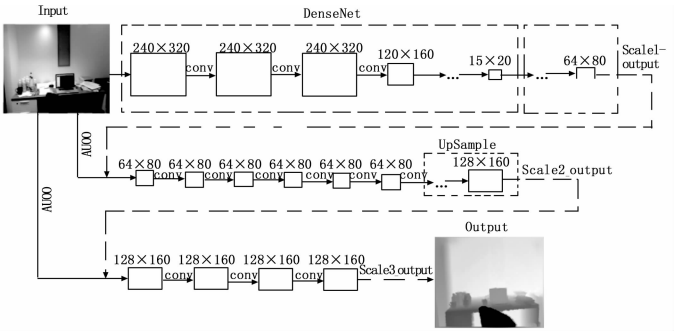


图 1 网络模型结构

图像通过 DenseNet 模块后，由上采样模块对输出结果进行上采样，最终图像输出尺寸为 64×80，与 Scale2 输入尺寸相同。由于 Scale1 的输入图像尺寸最大，因此 Scale1 的视野更广，采集到的特征最为丰富、最为原始，在全局层面对图像完成了一个粗糙的采样。

1.2.2 局部特征细致采样

Scale2 通过连接层（concatenate layer）将缩放的数据集与 Scale1 的结果输出组合，扩展了输入样本数量，输入图像尺寸为 64×80。与 Scale1 相比，Scale2 考虑到图像中像素与周边像素深度信息的相关性，更注重局部信息，利用局部特征优化输出深度图。Scale2 网络结构由卷积层、上采样模块组成，卷积层负责丰富特征信息采样，上采样模块负责获得合适分辨率的输出。对于卷积层部分，采用了多个 3×3 卷积核替换传统 5×5、7×7 卷积核的做法，在获得相同感受野的情况下，多个 3×3 卷积核相比大尺寸卷积核存在更多的非线性因子，使得判决函数更有判决性。同时，3×3 卷积核的实现方式可以显著减少参数量，提升训练速度。最终，Scale2 输出尺寸为 128×160 的深度图像，与 Scale3 输入尺寸相同。

1.2.3 提高分辨率，获得预测结果

Scale3 通过 concatenate layer 将缩放的数据集与 Scale2 的结果输出组合，网络结构与 Scale2 类似。Scale3 负责进一步优化图像输出并提高分辨率，得到空间上连贯、详细的输出结果。

1.3 DenseNet 模块

随着深度学习的发展，各类 CNN 网络模型被提出，从开始的 AlexNet、VGGNet，到后来的 ResNet 等等。DenseNet 于 2017 年由刘壮等人提出，网络结构具有以下特点：第一，DenseNet 有效地缓解了网络过深带来的梯度弥散问题，DenseNet 每层都能获取前面各层的损失函数，有效加强了特征前向传播，因此可以训练更深的网络。第二，相比 ResNet 采用求和的方式传播特征，DenseNet 采用拼接的方式，将前面所有层输出拼接在一起作为当前输入，显著提高特征传播效率，其非线性变换方程如式 1 所示。第三，DenseNet 有效减少了网络参数，从特征重用的角度提升网络性能。例如，基于 ResNet 的网络模型通过随机丢弃层防止过拟合，表明并非所有层都是必须的，网络中存在大量冗余，造成运算量的浪费。通常，对于同样的预测

精度，DenseNet 只需要 ResNet 一半的参数量。

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]) \tag{1}$$

DenseNet 由 Dense Block、Transition Layer、池化层、卷积层以及全连接层组合，Dense Block 为稠密连接的 highway 模块，是各层 Bottleneck Layer 输出的集合。Bottleneck Layer 包括卷积层、激活函数、归一化函数以及 Dropout 组成。Transition Layer 负责连接相邻 Dense Block，对 Dense Block 的输出进一步执行归一化、Dropout 等处理。

本文中的 DenseNet 模块由 4 个 Dense Block 和 3 个 Transition Layer 组成，并在 DenseNet 原有结构上对其进行了一定改善，是 Scale1 网络的主要结构。首先，对于图像初始化部分，用多个 3×3 卷积层替换 7×7 卷积层、池化层，用小卷积核替换大卷积核，减少参数与计算量；其次，针对各个 Dense Block 的输出通道数，通过实验对比，最终将各层输出通道数从上至下确定为 6、12、48、32；然后，对于 Transition Layer，采用了 3×3 卷积层替换池化层；最后，删除了 DenseNet 网络结构末尾的全局池化层和用于分类的全连接层，替换为性能更好的 1×1 的卷积层。

1.4　上采样模块

上采样的主要目的是放大原图像，获得分辨率更高的输出结果。放大图像基本都是采用内插值的方法，在原有图像像素的基础上在像素点之间采用插值算法插入新的像素。

在深度学习模型中，传统的上采样实现方式大多是先将图像分辨率扩展为 2 倍，用 0 填充没有数据的像素，然后用 5×5 卷积层处理扩展后的图像实现插值，由于有很多 0 值，因此存在很多无用计算。本文中的上采样模块将 5×5 卷积核拆分为 4 个大小为 3×3、2×3、3×2、2×2 的卷积核[6]，直接在原图上进行操作，跳过 0 值的处理，然后将 4 个输出结果进行交错连接，获得输出结果。同时，在此基础上，添加了激活函数（Relu）、归一化函数以及卷积层，组合成上采样模块，有效提高了上采样的效率。

1.5　损失函数

损失函数的选择是深度学习网络模型训练过程重要环节之一，选择正确的损失函数能提供更好的收敛方向，获得更好的训练结果。目前通用的损失函数为 L_2 损失函数，以最小化预测结果与真实值差的欧式范数平方作为收敛方向，函数定义如式（2）所示：

$$l_2(\overline{y} - y) = \|\overline{y} - y\|_2^2 \tag{2}$$

容易看出，当预测值与真实值差异较大时，L_2 损失函数能够迅速下降梯度，收敛速度很快，但是当预测值接近真实值时，L_2 损失函数收敛速度大大减小，梯度下降缓慢。因此，本文采用了 *BerHu* 损失函数[6]，巧妙地将 L_1 损失函数与 L_2 损失函数结合，获得了更好的收敛性能。其定义如式（3）所示：

$$B(x) = \begin{cases} |x| & |x| \leqslant c, \\ \dfrac{x^2 + c^2}{2c} & |x| > c \end{cases} \tag{3}$$

Berhu 函数以 c 作为界限，在大于 c 时以 L_2 损失函数工作，保证梯度迅速下降，在小于 c 时以 L_1 损失函数工作，保证预测值与真实值相近时也能保持一定的梯度下降速度。本文将 c 设定为 $c = k\max_i(|\overline{y} - y|)$，并对 k 分别取 $\frac{1}{10}$、$\frac{1}{2}$、$\frac{1}{5}$ 进行尝试，在 k 取 $\frac{1}{5}$ 时获得了最好输出结果[6]。同时，本文在模型中分别尝试了 Berhu、L_2 以及 Eigen 等人定义的损失函数，对比结果如表 1 所示。其中，Eigen 等人定义的损失函数如式 4 所示。从对比结果可以看出，Berhu 损失函数有效地提高了预测精度。

$$L_{depth}(D, D^*) = \frac{1}{n}\sum_i d_i^2 - \frac{1}{2n^2}\left(\sum_i d_i\right)^2 + \frac{1}{n}\sum_i\left[(\nabla_x d_i)^2 + (\nabla_y d_i)^2\right] \tag{4}$$

表 1　损失函数比较

损失函数	rel	rms	$\log_{10}$
L_2	0.209	0.845	0.090
L_{depth}	0.215	0.907	—
Berhu	0.119	0.547	0.052

2　结果与分析

2.1　实验设置

为了验证此模型的优化效果，本文选择在 NYU Depth V2 官方数据集上对模型进行训练、测试。NYU Depth V2 是微软 Kinect 相机采集的室内场景的视频帧序列，由 1449 对深度信息与 RGB 像素对应的图片组成。该数据集含有 3 个城市的 464 个场景，共分为 26 种场景类型，1000 多种对象。本文训练集与测试集的划分比例为官方的 249：215，将原始的 480×640RGB 图与深度图降采样至 240×320 作为模型输入，并通过预处理忽略图片中深度信息缺省的像素。根据官方划分，经 49 个场景作为验证集，200 个场景作为训练集，训练完成后在官方 654 张标准验证图像上对模型进行测试。本文对训练图像进行随机缩放、目标平面内旋转、水平翻转以及改变颜色和对比度等处理来扩充数据集，避免模型过拟合，提升泛化能力。

实验设备是一台显卡为 Tesla K40C 的服务器，使用 TensorFlow 框架。网络训练使用随机训练梯度下降优化模型参数，具体的超参数如下：批处理大小（batch size）为 8，最大迭代轮数（max epoch）为 1000，学习率（learning rate）为 0.001，学习率每经过 10 轮迭代衰 90%，直至网络收敛。整个模型训练耗时约为 72 个小时，CNN 前向过程占时约为每张图 0.06 s，整个模型每张图预测时间约为 0.23 s，训练时的 loss 曲线如图 2 所示。

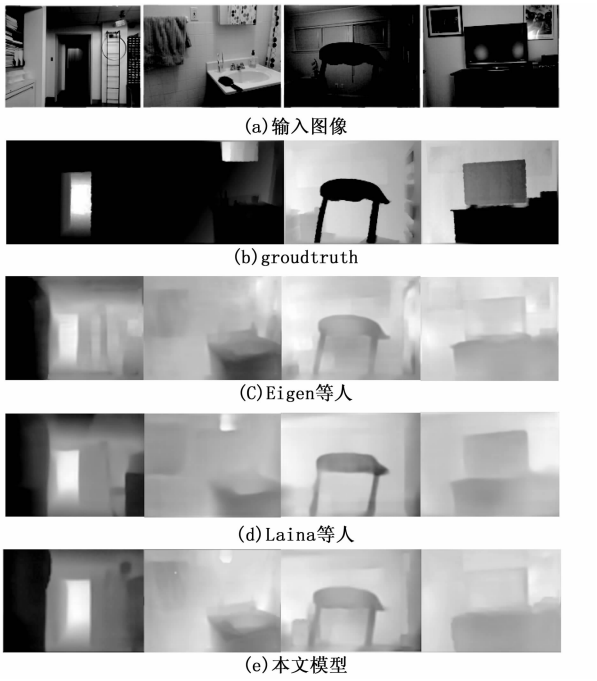
在此将模型的实验结果与同样在 NYU Depth V2 数据集上进行训练的相关工作进行对比，采用了常用的衡量指标评估结果[10]：

1）平均相对误差（average relative error，Rel）：

$$Rel = \frac{1}{T}\sum_i \frac{|y_i^* - y_i|}{y_i^*} \tag{5}$$

2）均方根误差（root mean squared error，RMS）：

$$RMS = \sqrt{\frac{1}{T}\sum_i (y_i^* - y)^2} \tag{6}$$



(a)输入图像

(b)groudtruth

(C)Eigen等人

(d)Laina等人

(e)本文模型

图 2 各模型深度预测结果图

3）对数空间平均误差（average $\log_{10}$ error，log10）：

$$\log10 = \frac{1}{T}\sum_i |\log y_i^* - \log y_i| \tag{7}$$

4）准确率：

满足 $\max(\frac{y_i^*}{y_i}, \frac{y_i}{y_i^*}) = \delta < threadhold$ 的像素占总像素的百分比。

上述公式中，y_i^* 与 y_i 分别为像素 i 的模型预测值与真实深度值，T 为 测试图像的像素数量总和。

整个训练模型的学习过程为：

步骤 1：输入 RGB 图像 $L = (x_1, x_j, \ldots x_n)$ 与真实深度图，并对数据集进行数据扩充。

步骤 2：RGB 图像经过网络模型后得到输出 y，与真实深度值 y^* 比较，计算 $d_i = |y - y^*|$。

步骤 3：根据 d_i 大小，计算损失函数，并更新参数。

步骤 4：达到收敛条件或迭代次数上限时终止模型运行，否则重复步骤 2)、3)。

2.2 实验结果

通过实验将得到的输出结果与 make3D、Eigen 以及 Laina 的训练方法输出结果进行对比，其结果如表 2 所示。首先，可以看出利用 CNN 卷积网络模型的输出结果明显优于传统利用先验的几何假设等方法，这主要得益于 CNN 结构在图像处理方面的强大能力，CNN 结构能够从图像中提取出足够多的特征信息，且特征提取和统计分类都不需要人工干预。在利用 CNN 结构的网络模型中，输出结果略优于 Eigen 与 Laina 的模型输出结果，这主要得益于 DenseNet 的网络特性与多尺度的网络结构，深度预测图结果对比如图 2 所示。

表 2 NYU Depth V2 数据集上实验结果对比

方法	误差(越小越好)			准确率(越大越好)		
	Rel	RMS	Log10	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
make3D	0.349	1.214	—	0.449	0.745	0.897
Eigen 等	0.158	0.641	—	0.769	0.950	0.988
Laina 等	0.127	0.573	0.055	0.811	0.953	0.988
本文	0.119	0.547	0.052	0.821	0.958	0.988

3 结论

针对单目图像深度预测问题，提出了基于深度学习的多尺度网络结构模型，利用 DenseNet 的强化特征前向传播、特征重用以及减少参数等特性，以及多尺度网络结构提供的全局预测、局部预测结果结合特性，有效提高了单目图像深度预测精度，其结果显著优于传统的处理方法，且在与其他 CNN 网络模型对比中也取得更好的输出结果。

但目前此模型仍是在有监督的条件下训练，需要提供真实深度信息，因此对数据集提出了很高的要求，笔者正在尝试构建无监督的训练模型，加大模型适用范围。同时，此模型是直接由 RGB 图训练得到深度结果，下一步考虑在模型中融入传统的深度获取方法，例如利用双目标定原理等，进一步提升预测精度。

参考文献：

[1] 张正友，马颂德. 计算机视觉 [M]. 北京：科学出版社，1998.

[2] Saxena A，Chung S H，Ng A Y. Learning depth from single monocular images [A]. Advances in Neural Information Processing System [C]. 2005：1161－1168.

[3] Szelisk R. Structure from motion [J]. Computer Science，2011：303－314

[4] Liu F，Shen C，Lin G. Deep convolutional nerural fields for depth estimation from a single image [A]. International Conference on Computer Vision and Pattern Recognition（CVPR）[C]. 2015：5162－5170.

[5] Eigen D，Puhrsch C，Fergus R. Depth map predicition from a single image using a multi－scale deep network [A]. Advances in Neural Information Processing System [C]. 2014.

[6] Laina I，Rupprecht C，Belagiannis V，et al. Deeper depth prediction with fully convolutional residual networks [A]. 3DV [C]. 2016.

[7] He K，Zhang X，Ren S，et al. Deep Residual Learning for Image Recognition [A]. Computer Vision and Pattern Recognition [C]. IEEE，2016：770－778.

[8] Godard C，Mac Aodha O，Brostow G J. Unsupervised monocular depth estimation with left－right consistency [Z]. 2016：6602－6611.

[9] Huang G，Liu Z，Maaten L V D，et al. Densely connected convolutional networks [A]. IEEE Conference on Computer Vision and Pattern Recognition [C]. IEEE Computer Society，2017：2216－2269.

[10] 李耀宇，王宏民，张一帆，等. 基于结构化深度学习的单目图像深度估计 [J]. 机器人，2017，39（6）：812－819.