计算机工程
Computer Engineering
ISSN 1000-3428,CN 31-1289/TP

《计算机工程》网络首发论文

题目：基于密集卷积网络的单目图像深度估计方法
作者：王亚群，戴华林，王丽，李国燕
DOI：10.19678/j.issn.1000-3428.0059516
网络首发日期：2020-11-13
引用格式：王亚群，戴华林，王丽，李国燕．基于密集卷积网络的单目图像深度估计方法．计算机工程. https://doi.org/10.19678/j.issn.1000-3428.0059516

网络首发时间：2020-11-13 15:25:14
网络首发地址：https://kns.cnki.net/kcms/detail/31.1289.TP.20201113.1510.008.html

基于密集卷积网络的单目图像深度估计方法

王亚群，戴华林，王丽，李国燕

（天津城建大学 计算机与信息工程学院，天津 300384）

摘　要：为了解决目前单目图像深度估计方法的精度低、网络结构复杂等问题，提出一种进行单目图像深度估计的密集卷积网络，该网络采用的是端到端的编码器和解码器结构。编码器引入了密集卷积网络（Dense convolutional network, DenseNet），把前面每一层的输出作为本层的输入，加强了特征重用和前向传播的同时减少参数量和网络计算量，一定程度上避免了梯度消失问题的发生。解码器结构采用带有空洞卷积的上投影模块和双线性上采样模块，更好地表达由编码器所提取的图像特征，最终得到与输入图像相对应的估计深度图。通过在 NYU Depth V2 室内场景深度数据集训练、验证和测试，所提出的基于单目图像深度估计的密集卷积网络结构在 δ <1.25 上的精度达到了 0.851，均方根误差控制在 0.482。

关键词：密集卷积网络；单目图像；编码器；解码器；深度估计

开放科学（资源服务）标志码（OSID):

A Monocular Image Depth Estimation Method Based on Dense Convolution Network

WANG Yaqun,DAI Hualin,WANG Li,LI Guoyan

(Tianjin Chengjian University, School of Computer and Information Engineering, Tianjin 300384,China)

【Abstract】In order to solve the problems such as low accuracy and complex network structure of the current monocular image depth estimation method, a dense convolutional network for monocular image depth estimation is proposed, which adopts end-to-end encoder and decoder structure.Dense convolutional network (DenseNet) is introduced into the encoder, and the output of each previous layer is taken as the input of this layer, which enhances feature reuse and forward propagation while reducing the number of parameters and network computation, thus avoiding the occurrence of gradient disappearance to a certain extent.The decoder structure adopts the upper projection module with cavity convolution and the bilinear upper sampling module to better express the image features extracted by the encoder and finally obtain the estimated depth map corresponding to the input image.By training, verifying and testing the indoor scene depth data set of NYU DepthV2, the proposed dense convolutional network structure based on monocular image depth estimation achieves an accuracy of 0.851 and a root mean square error of 0.482 on δ <1.25.

【Key words】Dense convolutional network;Monocular image;Encoder;Decoder;The depth of the estimated

DOI:10.19678/j.issn.1000-3428.0059516

1　概述

随着科技的不断革新，人工智能方面的发展也日新月异，各种人工智能产品层出不穷，比如无人驾驶汽车[1]、智能家居、智能机器人[2]、医学成像[3]等，这些人工智能产品为了能够更好的感知周围环境，必须要对接收到的图像进行深度估计。

传统估计图像深度的方法分为两种，一种是通过硬件设备测量获得深度。现在市场上推出了各种用于感知 3D 场景的传感器，比如激光雷达或者 Kinect 硬件设备，但它们在应用过程中存在成本过高、受限较多、捕获的深度图像分辨率低等问题，所以无法广泛推广使用。另一种是基于图像处理的深度估计方法。根据视觉传感器数量的多少分为单目视觉系统和多目视觉系统，目前大多使用双目或多目方法来获取深度信息，比如利用三角测量法[2]转化两幅图像之间的视差关系为深度信息，但易

基金项目：天津市自然科学基金（17JCQNJC00500）

作者简介：王亚群(1995-)，女，硕士，研究方向为图像处理、计算机视觉；戴华林(1974-)，教授，硕士生导师；王丽（1982-），硕士，讲师；李国燕(1984-)，博士，讲师。**E-mail：**815868728@qq.com

受外界因素的影响，难以获取高质量的视差图。学者们开始研究如何利用深度学习的方法从单幅图像中估计场景深度信息，相比双目或者多目视觉系统，通过单目视觉来进行图像深度估计具有成本低、应用灵活方便的优点，且现实生活中提供的数据信息多为单视点数据[4-6]，但从单幅图像估计场景深度具有一定的病态性和模糊性，这就使得通过单目图像来估计深度信息变成了一项挑战。

在基于深度学习的单目图像深度估计研究过程中，Eigen[7]团队提出将深度估计、表面法线预测和语义标注三个任务统一在一个三级的神经网络中，并将结果的分辨率降为输入图像的一半。Liu[8]等将 CNN 与条件随机场（Conditional Random Filed, CRF）结合，提出一个两步框架：用深度网络提取深度特征，用 CRF 进行深度信息的优化。但这两种方法都没有实现端到端的训练，需要分布进行。Laina[9]等采用了 ResNet-50 的残差网络并设计用小卷积代替大卷积来实现上采样，并提出了新的损失函数，得到了更好的结果。

Teed L[10]针对视频中单眼深度估计的问题，提出的 DeepV2D 方法在端到端架构中结合了两种经典算法。该网络包括深度估计模块和摄像机运动模块，深度估计模块将摄像机运动作为输入并返回初始深度图，摄像机运动模块获取预测深度并输出精确的摄像机运动。此外，网络在这两个模块之间交替运行以预测最终的深度图。Godard[11]提出了一种自我监督的方法，使用完全连接的 U-Net，并利用位姿网络来估计图像对之间的位姿，使用外观匹配损失解决像素遮挡问题，使用自动遮蔽方法忽略在训练过程中没有相对移动的像素点。Wei Y[12]提出了一种利用 3D 几何约束来估计深度图的监督框架。该方法直接从深度图估计场景和表面法线来重建 3D 结构，在重建的三维空间中随机抽取三个点确定的虚拟法向，可以显著提高深度预测精度，深度图鲁棒性强且具有强大的全局约束。Lee J[13]在网络结构解码阶段插入局部平面指导层，将各层的输出串接到最后的卷积层中，得到深度图。Tosi F[14]提出了一种自我监督的框架，使用端对端的单眼残差匹配来估计深度图。使用结构相似性重建损失，具有边缘感知项的视差平滑度损失和反向 Huber 损失来训练网络。Guizilini[15]将有监督和自监督两种方法有效结合，提出一个新的有监督损失项，利用它来训练鲁棒的半监督单目深度估计模型。此外评估训练精确的尺度感知单目深度模型所需的实际监督程度。

目前在单目图像深度估计的各项研究过程中，随着网络层数的不断加深，感受野越来越小，虽然能够提高网络分类的准确性，但是网络的训练越来越困难，同时所预测出的深度图存在深度信息不准确、误差偏大等问题。为了解决以上各种问题，本文提出了一种基于密集卷积网络的单目图像深度估计方法，编码器采用密集卷积网络(DenseNet)[16]，每一层都与它之前的所有层进行连接，前面所有层的输出都作为这一层的输入，实现了特征重用，减少参数，同时防止梯度消失问题的发生。解码器采用上投影模块和双线性插值模块来放大特征图，上投影模块引入空洞卷积[17]，在不增加参数量的条件下比标准卷积拥有更大的感受野，避免信息丢失，使得高层次的特征信息能够更有效地传播，并逐步提高特征图的分辨率，适应网络的输出。

2 深度估计网络模型

本文提出了一种密集卷积的编解码网络结构来对单目图像进行深度估计。编码层使用 DenseNet 网络对输入的图像进行特征提取，解码层结构采用加入了空洞卷积的上投影模块对编码层提取的特征进行上采样。不仅解决了因网络加深而引起的梯度消失问题，而且提高精度的同时简化了网络结构，减少了网络复杂度。整体的网络结构如图 1 所示。

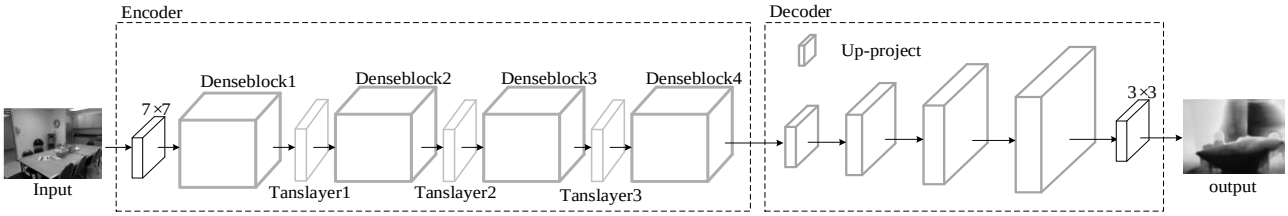


图 1 整体网络结构图

Fig.1 Overall network structure

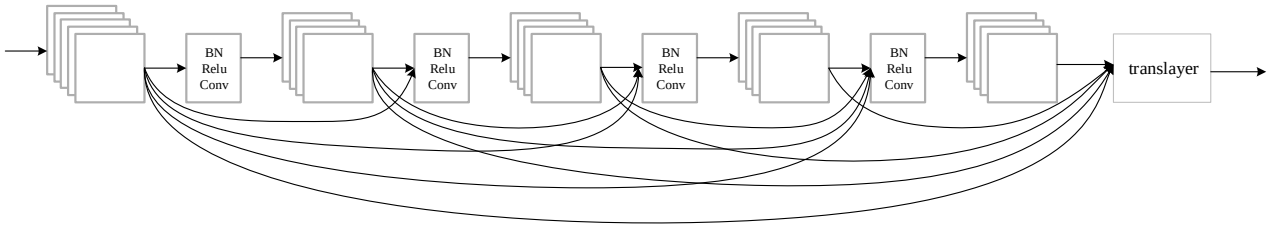


图 2 密集连接块（Dense block）

Fig.2 Dense connection block

2.1 编码器结构

目前，针对图像预测深度的研究中，编码器结构大多采用 ResNet 网络[18]，因为其跳跃连接的结构特点，在一定程度上解决了梯度消失和网络退化问题，但在运行过程中易产生大量冗余且每一层学习的特征信息较少，在实际应用场景中成本较高。而 DenseNet 网络使用的密集连接，能够互相连接每一层。传统的 N 层卷积网络在每层及其后续层之间有 N 个连接，DenseNet 网络有 N×(N-1)/2 个连接，这种连接方式称为密集连接。其非线性变换方程如公式（1）所示：

$$x_l = \boldsymbol{H}_l\left(\left[x_0, x_1, ..., x_{l\text{-}1}\right]\right) \quad (1)$$

其中，$\boldsymbol{H}_l$ 代表非线性转化函数，是一个组合操作，其可能包括一系列的批标准化(Batch Normalization，BN)[19]，线性整流函数（Rectified Linear Unit,Relu），池化(Pooling)及卷积(Conv)操作。其中 l 层与 l-1层之间可能包含多个卷积层。

DenseNet 网络包括卷积层、密集连接块（Dense block）、过渡层（Transition layer）、全连接层。

Dense block 是一个紧密连接的模块，如图 2 所示。在这个模块内，每一层的输入来自于这个模块内这一层之前的所有层的输出，一个密集连接块包含 BN、Relu、Conv 和 Dropout[20]操作，是 DenseNet 的重要结构。本文在每个密集连接块中 3×3 卷积的前面都加入一个 1×1 的卷积操作，称之为瓶颈层（Bottleneck layer）。原始的密集连接模块:BN+Relu+Conv(3×3)+Dropout，如图 3 的（a）所示；加入 1×1 卷积之后的密集连接模块表示为：BN+Relu+Conv(1×1)+Dropout+BN+Relu+Conv(3×3)+Dropout，如图 3 的（b）所示。

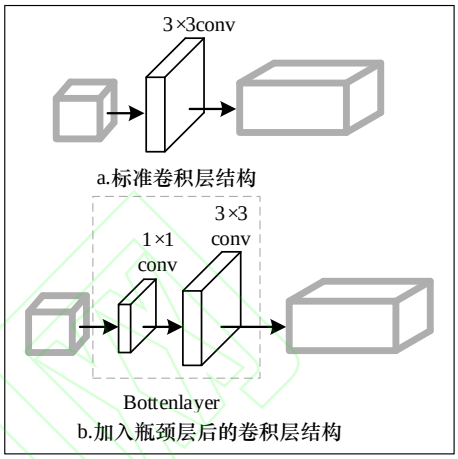


图 3 不同的卷积层结构

Fig.3 Different convolutional layer structures

过渡层由 1×1 卷积层和 2×2 平均池化层构成，如图 4 所示。它负责连接相邻的密集连接块，对其输出执行 BN、Relu、Dropout、Pooling 等处理。过渡层中压缩系数的参数 θ∈（0，1），当 θ=1 时，特征图的数量保持不变。本文实验将参数设置为 0.5，使传递到下一个密集块的通道数减半，这样在不影响准确率的情况下降低网络复杂度，提高计算效率，加快网络模型的收敛。

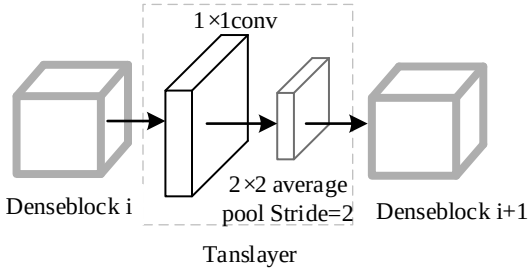


图 4 过渡层

Fig 4 Transition layer

本文采用的 DenseNet 包含 4 个密集连接块和 3 个过渡层，同时在 DenseNet 原有网络的基础上去除了末尾的全局平均池化层和用于分类的全连接层，直接输出图像作为解码器模块的输入，如图 1 中 Encoder 模块所示。网络结构参数如表 1 所示。

表 1　Encoder 网络结构参数（其中 Conv 代表 BN-Relu-Conv）

Table 1 Encoder Network structure parameters (where Conv stands for BN-Relu-Conv)

Layers	DenseNet
Convolution	7×7conv，stride2
Pooling	3×3max pool，stride2
Dense block（1）	$\begin{matrix}1\times1\text{conv}\\3\times3\text{conv}\end{matrix}\Big\}\times6$
Transition layer（1）	1×1conv 2×2average pool，stride2
Dense block（2）	$\begin{matrix}1\times1\text{conv}\\3\times3\text{conv}\end{matrix}\Big\}\times12$
Transition layer（2）	1×1conv 2×2average pool，stride2
Dense block（3）	$\begin{matrix}1\times1\text{conv}\\3\times3\text{conv}\end{matrix}\Big\}\times24$
Transition layer（3）	1×1conv 2×2average pool，stride2
Dense block（4）	$\begin{matrix}1\times1\text{conv}\\3\times3\text{conv}\end{matrix}\Big\}\times16$

2.2　解码器结构

解码器的主要作用是表达编码器所提取到的图像特征，得到与输入图像相对应的深度图，从而实现端对端训练。目前主要有上采样(Up-sampling)、上池化(Un-pooling)和反卷积(Deconvolution)三种方法[21]来实现这一过程。

本文采用的解码器结构属于上采样的一种，它包括四个带有空洞卷积的上投影模块（Up-projection），上投影模块是在标准的上卷积模块的基础上进行了改进。标准的上卷积模块如图 5 所示：首先经过一个 2×2 的上池化层，进行最大池化的反向操作，除最大值位置外其他位置补零，使得特征图扩大一倍，紧接着经过卷积核为 5×5 的卷积操作，之后进行 Relu 函数激活。本文采用的上投影模块结构如图 6 所示：包括两个扩张率（dilatation）为 1 的 5×5 的空洞卷积[22]和 3×3 的卷积操作。在相同的计算条件下，使用空洞卷积能够提供更大的感受野，使得每次输出的内容保留更多的特征，减少信息丢失。上投影模块的连接使高级信息在网络中更有效地向前传递，同时逐步增加特征图的大小。

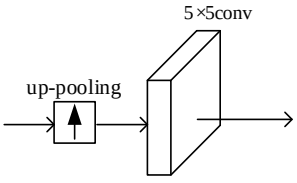


图 5　上卷积模块

Fig. 5 Up-convolution module

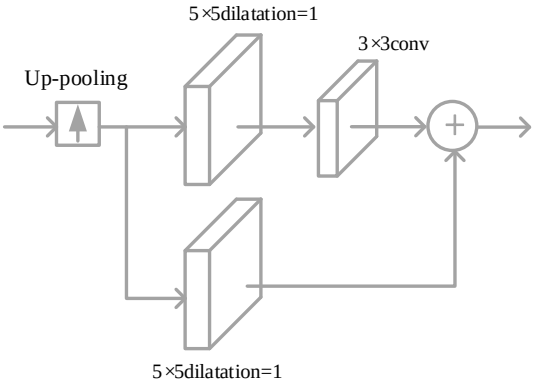


图 6　上投影模块

Fig. 6 Up-projection module

本文在 4 个带有空洞卷积的上投影模块后采用 3×3 的卷积操作，实现降维，之后采用双线性插值操作，输出预测深度图。

双线性插值计算方法：假设我们已知函数 f 在 $\boldsymbol{Q}_{11}=(x_1,y_1)$，$\boldsymbol{Q}_{12}=(x_1,y_2)$，$\boldsymbol{Q}_{21}=(x_2,y_1)$ 以及 $\boldsymbol{Q}_{22}=(x_2,y_2)$ 四个点的值。

首先在 x 方向进行线性插值，当 $\boldsymbol{R}_1=(x,y_1)$，

$$f(\boldsymbol{R}_1)\approx\frac{x_2-x}{x_2-x_1}f(\boldsymbol{Q}_{11})+\frac{x-x_1}{x_2-x_1}f(\boldsymbol{Q}_{21}) \tag{2}$$

当 $\boldsymbol{R}_2=(x,y_2)$，

$$f(\boldsymbol{R}_2)\approx\frac{x_2-x}{x_2-x_1}f(\boldsymbol{Q}_{12})+\frac{x-x_1}{x_2-x_1}f(\boldsymbol{Q}_{22}) \tag{3}$$

然后在 y 方向进行线性插值，得到

$$f(\boldsymbol{P})\approx\frac{y_2-x}{y_2-y_1}f(\boldsymbol{R}_1)+\frac{y-x_1}{y_2-y_1}f(\boldsymbol{R}_2) \tag{4}$$

这样就得到所要的结果 $f(x,y)$

$$\begin{aligned}f(x,y)\approx&\frac{f(\boldsymbol{Q}_{11})}{(x_2-x_1)(y_2-y_1)}(x_2-x)(y_2-y)\\&+\frac{f(\boldsymbol{Q}_{21})}{(x_2-x_1)(y_2-y_1)}(x-x_1)(y_2-y)\\&+\frac{f(\boldsymbol{Q}_{12})}{(x_2-x_1)(y_2-y_1)}(x_2-x)(y-y_1)\\&+\frac{f(\boldsymbol{Q}_{22})}{(x_2-x_1)(y_2-y_1)}(x-x_1)(y-y_1)\end{aligned} \tag{5}$$

2.3　损失函数

损失函数的选择对优化网络模型具有重要意义。本文采用 Berhu 损失函数对网络进行训练，公式如式（6）所示。以 c 值作为界限，在 c 点的值是连续且一阶可微的，当 $x\in[-c,c]$ 时，Berhu 损失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，获得更好的收敛性

能，在大于 c 时保证梯度迅速下降。设置 $c=\frac{1}{5}\max_i\left(\left|y_i-y_i\right|\right)$ 能够取得最好的训练。

$$\boldsymbol{B}(x)=\begin{cases}|x|, & |x|\le c\\ \dfrac{x^2+c^2}{2c}, & |x|>c\end{cases} \tag{6}$$

对于损失函数的选择，本文通过实验对比，使用 Berhu 损失函数均优于 $\boldsymbol{L}_1$、$\boldsymbol{L}_2$ 等其他损失函数。

3 仿真分析

3.1 数据集

NYU Depth V2 是最常用的室内深度预测图数据集，其原始数据集图像尺寸均为 640×480 像素，涵盖了 464 个场景，这些场景均为微软公司采用 Kinect 相机采集的视频帧序列，包含 RGB 图像及深度图。其中训练场景 249 个，测试场景 215 个。因为训练集的数据量太少，所以需要对采样所得的数据通过采用左右翻转、颜色变换、尺度变换的方法进行扩充。

我们使用双线性插值将图像从原始大小 640×480 降采样为 320×240 像素，然后裁剪其中心部分以获得 304×228 像素的图像。为了进行训练，深度图会下采样为 152×114 以适合输出的大小。

3.2 实验设置

实验硬件配置为 Intel(R) Core(TM) i7-7700CPU@3.60GHz 处理器，模型在具 16GB 内存的 NVIDIA Tesla K20M GPU 上进行训练。网络的训练是基于 RGB 图像的输入，用对应的真实深度进行训练。实验将批处理大小(batch size)设置为 8，最大迭代次数(max epoch)设置为 1000，初始学习率（learing rate）设置为 0.001，随着迭代的增加，学习速率会逐渐减小，直到网络收敛。

3.3 评价指标

目前，对图像深度进行预测时，为了评估预测深度与真实深度之间的差异，常采用以下的评价指标进行判断：

均方根误差（Root mean squarederror,RMSE）：

$$\boldsymbol{RMSE}=\sqrt{\frac{1}{|\boldsymbol{N}|}\sum_{i=1}^{N}\left|y_i-y_i^*\right|^2} \tag{7}$$

平均相对误差（Averange relative error,REL）：

$$\boldsymbol{REL}=\frac{\frac{1}{|\boldsymbol{N}|}\sum_{i=1}^{N}\left|y_i-y_i^*\right|}{y^*} \tag{8}$$

对数平均误差（Averange relative error,Log10）：

$$\mathrm{Log10}=\frac{1}{|\boldsymbol{N}|}\sum_{i=1}^{N}\left|\lg y_i-\lg y_i^*\right| \tag{9}$$

不同阈值下的准确率

$$\max\left(\frac{y_i}{y_i^*},\frac{y_i^*}{y_i}\right)=\delta<threshold \tag{10}$$

其中 δ =1.25，1.25^2，1.25^3

其中，y_i 为像素 i 的真实深度值，y_i^* 为预测深度值，N 为所有测试图像的像素数目总和。

3.4 实验结果与分析

本文通过三个对比实验，分别验证了选择 Berhu 作为损失函数的有效性、编码器结构采用 DenseNet 网络的有效性和解码器结构采用带有空洞卷积的上投影模块的有效性。

3.4.1 损失函数的有效性

本文在网络结构相同的条件下，分别使用 $\boldsymbol{L}_1$、$\boldsymbol{L}_2$ 和 Berhu 三种不同的损失函数进行了对比实验，实验结果表示：使用 Berhu 作为损失函数预测出的图像深度误差更小。实验对比如表 2 所示。这是因为 Berhu 损失函数的权值能够根据误差的大小发生较明显的幅度变化。

表 2 损失函数的对比

Table 2 Comparison of loss functions

损失函数	RMSE	REL	Log10
$\boldsymbol{L}_1$	0.844	0.208	0.090
$\boldsymbol{L}_2$	0.903	0.214	—
Berhu	**0.546**	**0.201**	**0.075**

3.4.2 DenseNet 模块的有效性

为了验证 DenseNet 模块的有效性，本文将编码器结构分别替换为 VGG 网络、Resnet 网络与本文提出的 DenseNet 网络进行对比，实验结果如表 3 所示。由于 DenseNet 网络自身密集连接的结构特点，保留了更多的特征信息，加强了特征的前向传播。使得实验结果误差更小，准确率更高。

表 3 DenseNet 模块的有效性

Table 3 Effectiveness of DenseNet module

编码器	误差		准确率		
	RMSE	REL	δ <1.25	δ <1.25^2	δ <1.25^3
VGG	0.641	0.350	0.811	0.953	0.968
Resnet	0.534	0.152	0.803	0.956	0.979
DenseNet	**0.551**	**0.144**	**0.832**	**0.958**	**0.985**

3.4.3 上投影模块的有效性

本文在其他条件相同的情况下对不同的上采样模块

进行了对比，包括采用：（1）3×3 的反卷积；（2）上卷积；（3）没有添加空洞卷积的标准上投影模块；（4）本文提出的带有空洞卷积的上投影模块，结果如表 4 所示。实验结果表示：使用本文提出的带有空洞卷积的上投影模块作为解码层结构时均方根误差 RMSE 为 0.553，δ<1.25 时的准确率为 0.840。这是因为空洞卷积扩大感受野，提取特征的范围更广；分辨率提高的同时精确定位目标。

表 4 上投影模块的有效性

Table 4 Effectiveness of the up-projection module

模型	误差		准确率		
	RMSE	REL	δ <1.25	δ <1.25^2	δ <1.25^3
（1）	0.556	0.168	0.828	0.954	0.988
（2）	0.565	0.215	0.821	0.940	0.987
（3）	0.572	**0.125**	0.831	0.958	0.989
（4）	**0.553**	0.136	**0.840**	**0.958**	**0.989**

3.4.4 NYU Depth v2 数据集结果

将本算法测得的评价指标与近年其他的单目深度估计的方法进行对比，实验结果如表 5 所示(表中各文献方法的评价指标直接引用文献中的值，空白处表示该文献未计算该评价指标)，可以发现本算法的均方根误差 RMSE 为 0.482，与文献[8]相比，降低了 34.2%；预测深度图的精确度 δ<1.25 时，指标为 0.851，与文献[7]相比，提高了 8.2%，与文献[8]相比提高了 23.7%。由此可见，本文的算法预测的图像深度误差更小，准确度更高。

在 NYU Depth v2 数据集上的测试结果如图 7 所示，将本算法得到的深度图与文献[7]和文献[9]进行比较，可以看出，本算法得到的深度图与真实深度图更为接近，且图中物体轮廓更为清晰，层次更加分明。

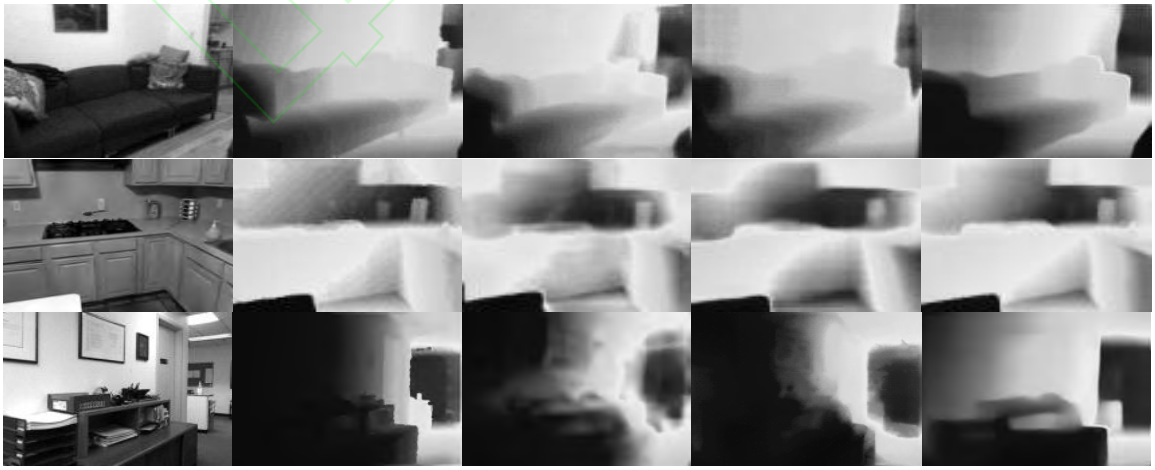
表 5 NYU Depth v2 数据集上的评价指标

Table 5 Evaluation indicators on NYU Depth v2 dataset

方法	误差		准确率		
	RMSE	REL	δ <1.25	δ <1.25^2	δ <1.25^3
文献[8]	0.824	0.231	0.614	0.883	0.971
文献[7]	0.641	0.158	0.769	0.950	0.988
文献[23]	0.579	**0.108**	0.823	0.957	0.987
文献[9]	0.573	0.127	0.811	0.953	0.988
文献[12]	0.485	0.134	0.829	0.956	0.980
本文结果	**0.482**	0.112	**0.851**	**0.958**	**0.988**

4 结束语

本文提出了一种对单目图像进行深度估计的密集卷积网络结构，编码层由 DenseNet 组成，采用密集连接的方式加强了特征重用和特征的前向传播；解码层采用上投影模块和双线性插值操作，逐步提高分辨率，降低通道数，上投影模块负责恢复深度信息，其中引入的空洞卷积扩大了感受野，双线性插值模块映射至目标输出深度图，在不需要引入其他参数的情况下放大特征图。通过在 NYU Depth V2 室内数据集上的广泛评估，结果表明：本方法与一些经典的方法相比获得了较优的实验结果，在 δ <1.25 上的精度达到了 0.851，均方根误差 RMSE 控制在 0.482，且减少了网络参数。但目前本文的模型对数据样本有较高的要求，需要在有监督的情况下进行，下一步将尝试构建适用于无监督方式的网络模型，并改善损失函数来加强边缘信息，改善边缘失真情况。



输入图像　　真实值　　文献[7]　　文献[9]　　本文方法

图 7 NYU Depth v2 数据集上得到的深度图

Fig 7. The depth map obtained on the dataset of NYU Depth v2

参考文献

[1] Shifeng ZENG, Jinjun WU, Zhiwen YE, et al.ROS based Driverless intelligent car [J]. Internet of Things Technology, 2020,10(06):62-63+66.
曾仕峰,吴锦均,叶智文,等.基于 ROS 的无人驾驶智能车[J].物联网技术,2020,10(06):62-63+66.

[2] Yan ZHAO, Yinggang XIE, Lili CHEN. Research on robot Intelligent Laser localization Based on SLAM algorithm [J]. Laser Journal,2019,40(07):169-173.
赵妍,解迎刚,陈莉莉.基于 SLAM 算法的机器人智能激光定位技术的研究[J].激光杂志,2019,40(07):169-173.

[3] Xingtong LIU, Sinha A, Ishii M, et al. Dense Depth Estimation in Monocular Endoscopy With Self-Supervised Learning Methods[J]. IEEE Transactions on Medical Imaging, 2019,39(5):1438-1447.

[4] Elkerdawy S, Hong ZHANG, Ray N. Lightweight Monocular Depth Estimation Model by Joint End-to-End Filter Pruning[C]//International Conference on Image Processing (ICIP). Taiwan:IEEE, 2019:4290-4294

[5] Wofk D, Fangchang MA, Yang T J, et al. Fastdepth: Fast monocular depth estimation on embedded systems[C]// International Conference on Robotics and Automation (ICRA). Montreal, QC, Canada: IEEE, 2019: 6101-6108.

[6] Poggi M, Aleotti F, Tosi F, et al. Towards real-time unsupervised monocular depth estimation on cpu[C]//RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid:IEEE, 2018: 5848-5854.

[7] Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture[C]//International conference on computer vision(ICCV). Santiago: IEEE, 2015: 2650-2658.

[8] Fayao LIU, Chunhua SHEN, Guosheng LIN, et al. Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016,38(10):2024-39.

[9] Laina I, Rupprecht C, Belagiannis V, et al. Deeper depth prediction with fully convolutional residual networks[C]//2016 Fourth international conference on 3D vision (3DV). Stanford, CA: IEEE, 2016: 239-248.

[10] Teed Z, Jia DENG. Deepv2d: Video to depth with differentiable structure from motion[J]. arXiv:1812.04605, 2018.

[11] Godard C, Mac Aodha O, Firman M, et al. Digging into self-supervised monocular depth estimation[C]// International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019: 3828-3838.

[12] Wei YIN ; Yifan LIU ; Chunhua SHEN, et al. Enforcing geometric constraints of virtual normal for depth prediction[C]// International Conference on Computer Vision.(ICCV). Seoul, Korea (South): IEEE, 2019: 5684-5693.

[13] Lee J H , Han M K , Ko D W , et al. From Big to Small: Multi-Scale Local Planar Guidance for Monocular DepthEstimation[J]. arXiv:1907.10326, 2019.

[14] Tosi F, Aleotti F, Poggi M, et al. Learning monocular depth estimation infusing traditional stereo knowledge[C]// Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 9799-9809.

[15] Guizilini V, Li J, Ambrus R, et al. Robust Semi-Supervised Monocular Depth Estimation With Reprojected Distances[C]//Conference on Robot Learning. PMLR, 2020: 503-512.

[16] Gao HUANG, Zhuang LIU, Van Der Maaten L, et al. Densely connected convolutional networks[C]// conference on computer vision and pattern recognition (CVPR), Honolulu, HI: IEEE, 2017: 4700-4708.

[17] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[J]. arXiv:1511.07122, 2015.

[18] Junjie HU, Ozay M, Yan ZHANG, et al. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries[C]//Winter Conference on Applications of Computer Vision (WACV). Waikoloa Village, HI, USA: IEEE, 2019: 1043-1051.

[19] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv:1502.03167, 2015.

[20] Hinton G E , Srivastava N , Krizhevsky A , et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer ence, 2012, 3(4): 212-223.

[21] Quande WANG, Songtao ZHANG. Monocular image depth Estimation based on Multi-scale Feature Fusion [J]. Journal of Huazhong University of Science and Technology (Natural Science edition) ,2020,48(05):7-12.
王泉德,张松涛.基于多尺度特征融合的单目图像深度估计[J].华中科技大学学报(自然科学版),2020,48(05):7-12.

[22] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous

convolution for semantic image segmentation[J]. arXiv:1706.05587, 2017.

[23] Ricci E, Wanli OUYANG, Xiaogang WANG, et al. Monocular depth estimation using multi-scale continuous CRFs as sequential deep networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(6): 1426-1440.