DOI：10.13245/j.hust.200502

基于多尺度特征融合的单目图像深度估计

王泉德 张松涛

(武汉大学电子信息学院，湖北 武汉 430072)

摘要 为解决从单目图像中很难恢复出准确、有效深度信息的问题，提出一种多尺度特征融合的单目图像深度估计算法．算法采用端对端训练的卷积神经网络(CNN)结构，引入从图像编码器到解码器的跳层连接来实现在不同尺度上特征的提取和表达，设计了一种多尺度的损失函数来提升卷积神经网络的训练效果．通过在NYU Depth V2室内场景深度数据集和KITTI室外场景深度数据集上的训练、验证和测试，实验结果表明：提出的多尺度特征融合方法得到的深度图边缘清晰、层次分明，且在室内场景和室外场景中均能适用，具有较强的泛化性，可以适应多种实际场景的需求．

关键词 计算机视觉；深度学习；卷积神经网络；单目图像深度估计；多尺度特征融合

中图分类号 TP391.41 **文献标志码** A **文章编号** 1671-4512(2020)05-0007-06

Monocular depth estimation with multi-scale feature fusion

WANG Quande ZHANG Songtao

(School of Electrical Information，Wuhan University，Wuhan 430072，China)

Abstract To solve the problem that it is difficult to recover accurate and effective depth information from monocular images，a monocular image depth estimation algorithm based on multi-scale feature fusion was proposed．End-to-end trained convolutional neural network (CNN) structure was applied to the algorithm，and the skip layer connection from image encoder to decoder was introduced to realize feature extraction and expression on different scales．A multi-scale loss function was designed to improve the training effect of the convolutional neural network．Through training，verification and testing on the NYU Depth V2 indoor scene depth dataset and KITTI outdoor scene depth dataset，experimental results show that the proposed multi-scale feature fusion method can obtain clear，sharp-edged edges in the depth map，and is applicable to both indoor and outdoor scenes with strong generalization，which can adapt to the demands of a variety of actual scenes．

Key words computer vision；deep learning；convolutional neural network；monocular depth estimation；multi-scale feature fusion

深度信息对理解场景中的三维几何关系具有十分重要的作用，准确有效的深度信息能提升图像分类、目标识别、语义分割等诸多计算机视觉任务的结果．目前基于图像的深度信息提取方法大都采用多视角立体匹配的方式来获得图像的深度信息，例如运动恢复结构(SFM)[1]和双目立体匹配[2]等．虽然这些方法都取得了不错的结果，但是依然存在匹配困难，难以适应深度变化剧烈的场景等问题．为了改善这些问题，根据单幅图像估计场景深度信息的研究受到了越来越多研究者的重视．但是相机成像是将三维空间投影到像平面上，这必然会造成深度信息的丢失，因此对于计算机视觉系统来说，仅通过单张图像来恢复深度信息长期以来一直被视为病态问题，难以实现．

近些年来深度学习发展迅猛，卷积神经网络(CNN)凭借其高效的图像特征提取和表达能力，不

收稿日期 2019-06-22．

作者简介 王泉德(1973-)，男，副教授，E-mail：wqd@whu.edu.cn．

基金项目 国家自然科学基金青年基金资助项目(61701351)．

断刷新计算机视觉各领域的记录，这为图像深度信息的估计提供了新的思路．目前许多基于 CNN 来估计图像深度信息的方法都取得了不错的结果．文献[3]最早将 CNN 应用于单目图像深度估计，所提出的网络结构分两步在不同的尺度进行训练，得到了效果较好的深度图．文献[4]通过将条件随机场(CRF)和 CNN 网络模型相结合的方法来提升深度估计效果．但是这两种方法没有实现端对端训练，须要分步进行．文献[5]通过一系列上采样和卷积组成的单元来得到分辨率较高的深度图，但是在上采样过程中造成了图像细节的丢失．考虑到图像对应场景的深度信息采集较为困难，文献[6]提出了一种基于无监督学习的方法，利用双目相机采集并标定好的图像对，学习左(右)图像到右(左)图像的映射关系，从而估计到场景深度信息，将单目图像深度估计问题转换为图像重建问题．文献[7]在文献[6]的基础上加入了左右视图一致性损失和增强视差平滑性损失，进一步提升了神经网络训练效果和深度信息估计精度．但是这些无监督的方法训练较困难，仍旧没有解决所获得深度图中物体轮廓不清晰、深度变化不平滑等问题．

针对上述问题，本研究提出了一种端对端训练多尺度特征融合的单目图像深度估计算法．通过引入从图像编码器到解码器的跳层结构，使训练时能够充分学习到了图像中的多尺度的特征，并设计了一种多尺度的损失函数来提升卷积神经网络的训练效果．

1　卷积神经网络结构

1.1　方法和思路

在单目图像深度估计问题中，一幅图像对应场景的深度信息一般由相同大小的灰度图来描述，灰度图中每个像素的灰度值描述该点对应场景的深度值，该灰度图又称为深度图．本研究借助 CNN 高效的图像特征提取和表达能力，通过大量彩色图和深度图图像对的训练，得到彩色图像到对应深度图之间的映射关系，从而估计到图像对应场景的深度信息，提出一种端对端训练的单目图像深度估计网络模型．该模型加入了由图像编码器各层到对应图像解码器各层的跳层结构，实现多尺度的特征融合，如图 1 所示为总体网络结构图，图中的立方体表示网络不同层的特征图，其旁边的数字表示特征图的特征维度数．此外，通过构建图像金字塔在多个尺度计算网络训练损失，使最终得到的深度图不断被优化．通过在常用室内数据集 NYU Depth V2 和室外数据集 KITTI 上的测试表明：与目前已有的单目深度估计方法相比，由于引入多尺度策略，因此本研究获取的深度图中物体轮廓清晰、深度变化平滑，主客观评价指标都更优，且具有较强的泛化性．

图像领域的编码解码网络结构[8]最早应用在图像语义分割方面，但是由于其只是编码器的最后一层与解码器的第一层进行连接，因此若网络层数较

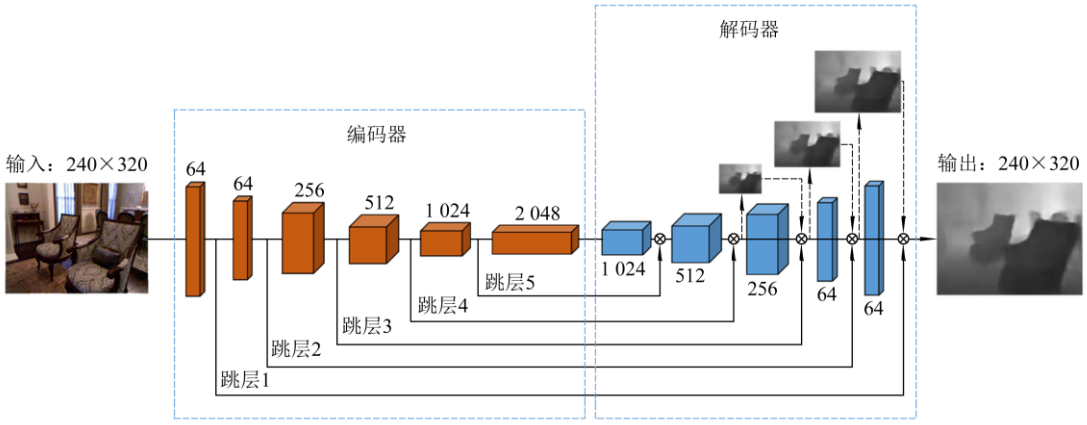

图 1　总体网络模型结构图

深，则会使编码器中浅层提取到的特征不能较好地表达．受 U-net[9]的启发，本研究的网络结构引入了从编码器到对应解码器的跳层结构，如图 1 中跳层 1～跳层 5 所示，这样不仅可以解决浅层特征不能表达的问题，而且可以将从编码器不同层中提取到的特征在解码器中进行多尺度特征融合，从而提高深度图的精度．

1.2　编码器

编码器的作用主要是对输入的图像进行特征提取．近些年来，随着计算性能的显著提高和深度学习的不断发展，为了解决更加复杂的实际问题，CNN 的层数也不断加深．但是越来越深的网络会导致梯度消失和梯度爆炸，使训练困难，并且会影响最终的实验结果．而残差网络(Resnet)[10]通过在残差

块中引入跳层连接，很好地解决了梯度消失和梯度爆炸的问题．后来密集连接网络结构(Densenet)[11]又对 Resnet 网络的改进．该网络结构的密集连接块中的每一层不仅与相邻层相连，而且与所有后续层相连．因此在密集连接块中每一层卷积都可以从先前所有卷积层中获取特征．所以相比于 Resnet，Densenet 的鲁棒性和抗过拟合性更好．编码器尝试使用了基于 Densenet-121[11]的网络结构，而不同的是该网络结构移除了最后的全连接层，直接将最后一个密集连接块得到的特征图输入到解码器中．并且网络结构中采用的激活函数是 ELU，相比于常用的 RELU 激活函数，ELU 的负半轴不再是零，而是以指数函数的形式存在，这使得它对输入变化或噪声更为鲁棒，并且 ELU 的输出平均值接近于零，收敛速度较快．

1.3　解码器

解码器主要的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．现在主要有上采样(upsampling)、上池化(unpooling)和反卷积(deconvolution)三种方法来实现这一过程．上采样一般采用最近邻插值、双线性插值和三次卷积插值等方法来实现．上池化一般视为最大池化的逆过程，通过进行 0 填充来扩充特征图．但是由于这两种方法都要进行插值，会导致最终得到的深度图不够精细，而基于反卷积的方法，其参数是通过训练学习得到的，在训练过程中深度图会不断被优化，因此解码器尝试使用反卷积的方式，由 6 个连续的步长为 2 的 3×3 的反卷积单元构成．通过解码器最终得到与输入图像尺寸相同的深度图．

2　多尺度损失函数

为了提升网络的训练效果，本研究采用多尺度的损失函数，分别在最后 4 个尺度计算损失并估计深度，并且将前一个尺度得到深度图上采样后，与下一尺度的输入相级联，从而实现对下一尺度深度估计结果的优化．在这个过程中，最终得到的深度图不断地被优化．

选择合适的损失函数对优化网络模型十分重要.回归问题中常用的损失函数为最小平方误差(L_2)损失，但文献[5]的研究表明在单目图像深度估计问题中，Berhu 损失得到的结果比 L_2 损失更好．Berhu 损失函数 $L(\mu)$ 的表达式为

$$L(\mu)=\begin{cases}|\mu| & (|\mu|\leqslant\sigma);\\ (\mu^2-\sigma^2)/2\sigma & (|\mu|>\sigma),\end{cases} \tag{1}$$

式中：μ 为预测值与真实值之间的差；σ 为阈值．由式(1)可以看出：当预测值与真实值之间的差小于阈值时，Berhu 损失表现为最小绝对值(L_1)损失；当预测值与真实值之间的差大于阈值时，Berhu 损失表现为 L_2 损失．所以当误差小于阈值时，Berhu 损失反向传播的梯度较小且为常数 1；而大于阈值时，反向传播的梯度与误差成正比.因此当误差较小时，权值变化幅度较小；当误差较大时，权值变化幅度随误差增大而增大．这就是 Berhu 损失能取得较好结果的主要原因，所以选用 Berhu 损失作为多尺度损失函数的基本单元．本研究的多尺度损失函数 $L_j(d,\ d^*)$的表达式为

$$L_j(d,\ d^*)=\begin{cases}\dfrac{1}{n_j}\displaystyle\sum_i^{n_j}|d_i-d_i^*| & (|d_i-d_i^*|\leqslant\sigma);\\ \dfrac{1}{2\sigma n_j}\displaystyle\sum_i^{n_j}(d_i-d_i^*)^2-\dfrac{\sigma}{2} & (|d_i-d_i^*|>\sigma),\end{cases}$$

式中：d_i 和 d_i^* 分别为每个像素点的预测深度值和真实深度值；n_j 分别为最后四个尺度特征图的像素个数．为了在最后 4 个尺度上计算损失，本研究将数据集中的真实深度图下采样构成 4 个不同尺度的图像金字塔，训练中总的损失 $L(d,\ d^*)$的表达式为

$$L(d,\ d^*)=\sum_{j=1}^{4}L_j(d,\ d^*).$$

但是每个尺度损失的反向传播是单独进行的，前一个尺度的损失函数不影响接下来网络的权重更新．此外，阈值 σ 对最终的实验结果也有较大的影响，一些研究[12]表明：当 σ=1.35 时，可以得到较好的实验结果，因此最终将阈值 σ 设置为 1.35．

3　实验结果

本研究分别在 NYU Depth V2 和 KITTI 这两个典型的深度数据集上进行实验，计算相应的评价指标，与现有的一些效果较好的方法进行对比．

3.1　实验设置

本研究的网络结构是通过主流的深度学习框架 Tensorflow 实现的，训练是在内存为 11 GiB 的 NVIDIA GeForce GTX 1080Ti 显卡上进行．根据网络模型结构以及显卡的性能，将批尺寸(batch size)设置为 8，初始学习率 l 设置为 0.001，学习率衰减

因子 α 设置为 0.9，并且每进行 1 000 次迭代，学习率乘以一次衰减因子. 网络的优化器采用动量为 0.9 的随机梯度下降法(SGD). 最终在 NYU Depth V2 数据集上训练网络花费了约 28 h，在 KITTI 数据上花费了约 33 h.

3.2　评价指标

由于仅通过人眼的观察来评判深度图的优劣太过于主观，因此利用 R，S 和 M 这三个常用的误差计算公式来评判预测深度与真实深度的差别，即：

$$R = \frac{1}{N}\sum_{i}^{N}\frac{|d_i - d_i^*|}{d_i^*};$$

$$S = \frac{1}{N}\sum_{i}^{N}\frac{|d_i - d_i^*|^2}{d_i^*};$$

$$M = \sqrt{\frac{1}{N}\sum_{i}^{N}|d_i - d_i^*|^2},$$

式中 N 为总的像素个数. 这三个值越小得到的深度图质量越好.

此外，借助相似度 F 来衡量预测深度和真实深度的相似度，即

$$F = \sum_{i}^{N}\left[\max\left(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}\right) < t\right] \Big/ N,$$

式中 t 为相似度判定阈值，本研究中 t 取 1.25，1.25^2 和 1.25^3. 分别在三个不同阈值来计算相似度 F，F 值越高，深度图质量越好.

3.3　NYU Depth V2 数据集结果

NYU Depth V2 数据集是最常用的室内深度数据集，其原始数据集图像尺寸均为 640×480 像素，涵盖了 464 个场景，其中训练场景 249 个，测试场景 215 个.

数据集中的深度图是通过微软的 Kinect 深度相机采集到的，此外数据集官网还提供了标记好的数据集，它由 1 449 对彩色图深度图图像对构成. 根据先前的划分方法[1]，其中的 795 对作为训练集，654 对作为测试集. 因为训练集的数据量太少，所以须要进行数据扩充，本研究的数据扩充方法总结如下. **a.** 左右翻转：将输入的图像水平地左右翻转. **b.** 颜色变换：将输入图像的颜色通道随机乘以因子 $\beta \in [0.8, 1.2]$. **c.** 尺度变换：将输入图像随机按因子 $\gamma \in [1.2, 1.6]$进行尺度变换，裁剪图像的中心面片，使之与原始图像匹配.

通过上述数据扩充的方法，最终生成 $1.192\,5\times10^4$ 对训练图像对，然后将其下采样到 320×240 像素，作为网络结构的输入.

如图 2 所示，将本算法得到的深度图与文献[5]得到的深度图进行比较，图中的深度图进行了伪彩色处理，越接近红色表示深度值越大，越接近蓝色表示深度值越小. 可以明显看出：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，层次更为分明.

将本算法测得的评价指标与已有的单目深度估计的主要方法进行对比，如表 1 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)，可以发现本算法的客观评价指标也更优.

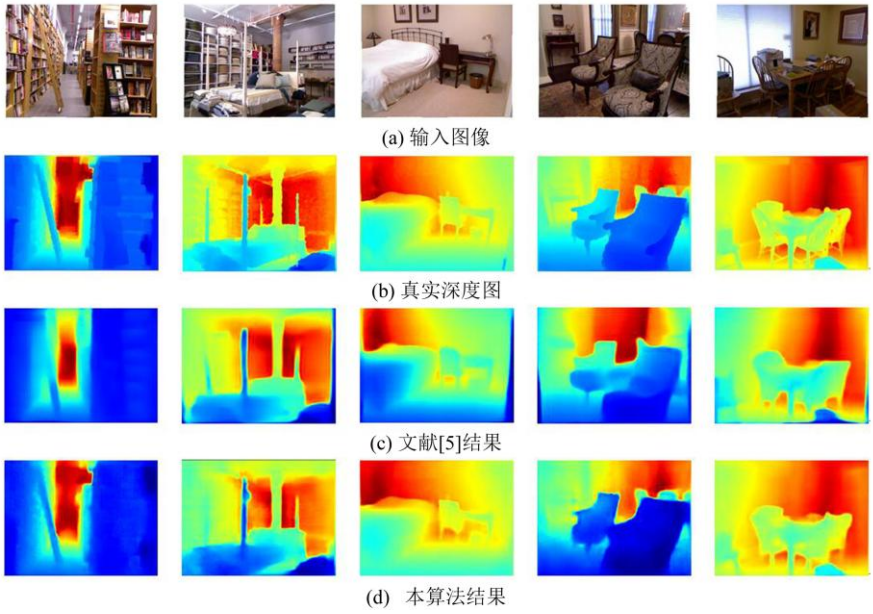

图 2　NYU Depth V2 数据集上得到的深度图

表 1　在 NYU Depth V2 数据集上测得评价指标

方法	是否数据增强	误差			相似度		
		R	*S*	*M*	t=1.25	t=1.25^2	t=1.25^3
文献[3]	是	0.215	0.212	0.907	0.611	0.887	0.971
文献[13]	是	0.134	0.095	0.540	0.832	0.965	0.989
文献[5]	是	0.127		0.573	0.801	0.953	0.988
文献[14]	是	0.121		0.586	0.811	0.954	0.987
本文	是	0.118	0.093	0.486	0.828	0.954	0.989

3.4　KITTI 数据集结果

KITTI 数据集是最常用的自动驾驶领域的视觉算法评估数据集，其原始数据集包含大量从城市地区、村庄和公路等场景收集的室外街道场景图像. 数据集中还包含了用 Velodyne 激光传感器捕获的稀疏三维激光测量数据，但是原始数据集中没有相应的深度图，所以首先须要将激光测量数据投影到图像空间，以生成相应的深度图. 图像的原始尺寸为 376×1 242 像素，将其下采样到 188×621 像素，作为网络的输入. 为了能与先前的方法进行比较，本研究也使用了文献[1]提出的数据集分割方法，将 2.260 0×10^4 对图像对用来训练，将 697 对图像对用来测试. 如图 3 所示，将得的深度图与文献[7]得到的深度图进行比较，图中的深度图也进行了伪彩色处理，从中也可以看出：本文算法得到的深度图细节表现更好，且与真实深度图更接近. 在表 2 中，为了与不同的方法进行比较，本研究计算了两个不同深度范围(50 m 和 80 m)内的评价指标，可以发现本文算法在这两个不同的深度范围内测得的评价指标要优于目前已有的单目深度估计的主要方法.

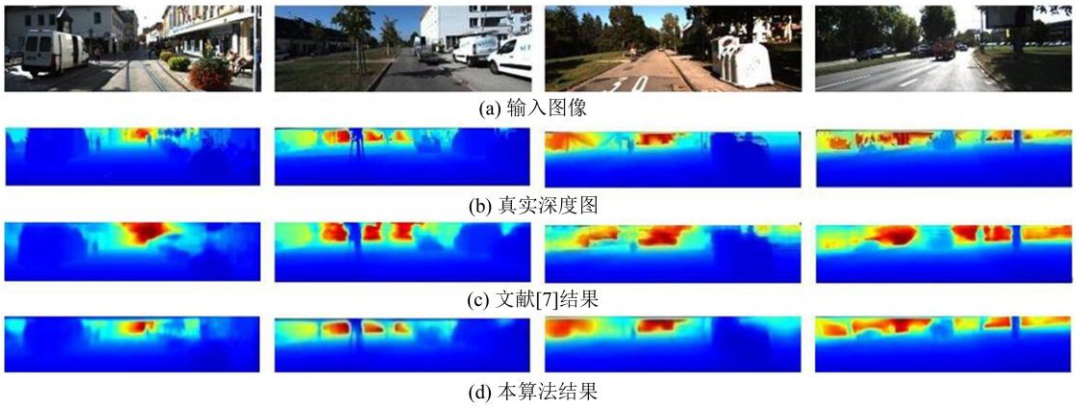

图 3　KITTI 数据集上得到的深度图

表 2　在 KITTI 数据集上测得的评价指标

方法	深度范围/m	误差			相似度		
		R	*S*	*M*	t=1.25	t=1.25^2	t=1.25^3
文献[4]	80	0.217		7.046	0.656	0.881	0.958
文献[3]	80	0.190	1.515	7.156	0.692	0.899	0.967
本文	80	0.123	0.828	4.567	0.880	0.959	0.983
文献[7]	50	0.140	0.976	4.471	0.818	0.931	0.969
文献[15]	50	0.115		3.605	0.887	0.963	0.982
本文	50	0.113	0.778	4.055	0.899	0.962	0.987

4　结语

本研究提出了一种单目深度估计算法，通过引入从图像编码器到解码器的跳层结构，实现了多尺度的特征融合. 此外，还引入了一个多尺度损失函数来优化最终结果. 该算法仅通过一张 RGB 图像便可以估计出图像中各个像素点的深度值. 通过在 NYU Depth V2 室内数据集和 KITTI 室外数据集上的广泛评估，结果表明：本方法几乎获得了最优的实验结果，且泛化性较好. 虽然最终测得深度的精度与传统多视角匹配的方法得到的结果仍有差距，但是由于其算法简单、使用场景不受限制，因此具有一定的研究意义和较为广阔的应用前景.

参 考 文 献

[1] 解则晓，周作琪．基于运动恢复结构的空间点定位方法[J]．激光与光电子学进展，2018，55(8)：370-377．

[2] 朱清波，王宏远．使用图像分割的遮挡恢复立体匹配算法[J]. 华中科技大学学报(自然科学版)，2010，38(1)：81-84．

[3] DAVID E，CHRISTIAN P，ROB F. Depth map prediction from a single image using a multi-scale deep network[C]// Proc of Advances in Neural Information Processing Systems．Cambridge：MIT Press，2014：2366-2374．

[4] LIU F，SHEN C，LIN G，et al. Learning depth from single monocular images using deep convolutional neural fields [J]．IEEE Transactions on Pattern Analysis & Machine Intelligence，2015，38(10)：2024-2039．

[5] LAINA I，RUPPRECHT C，BELAGIANNIS V，et al．Deeper depth prediction with fully convolutional residual networks[C]// Proc of International Conference on 3D Vision．Piscataway：IEEE，2016：239-248．

[6] GARG R，CARNEIRO G，Reid I，et al．Unsupervised CNN for single view depth estimation：geometry to the rescue[C]// Proc of European Conference on Computer Vision．Berlin：Springer，2016：740-756．

[7] GODARD C，AODHA O M，BROSTOW G J．Unsuper-vised monocular depth estimation with left-right consistency[C]// Proc of Conference on Computer Vision and Pattern Recognition．Piscataway：IEEE，2017：6602-6611.

[8] MAYER N，IIG E，FISCHER P，et al．A large dataset to train convolutional networks for disparity，optical flow，and scene flow estimation[C]// Proc of Conference on Computer Vision and Pattern Recognition．Piscataway：IEEE，2016：4040-4048．

[9] OLAF R，PHILIPP F，THOMAS B．U-Net：convolutional networks for biomedical image segmentation[C]// Proc of International Conference on Medical Image Computing and Computer-Assisted Intervention．Berlin：Springer，2015：234-241．

[10] HE K，ZHANG X，REN S，et al．Deep residual learning for image recognition[C]// Proc of Conference on Computer Vision and Pattern Recognition．Piscataway：IEEE，2016：770-778．

[11] HUANG G，LIU Z，LAURENS V D M，et al．Densely connected convolutional networks[C]// Proc of Conference on Computer Vision and Pattern Recognition．Piscat-away：IEEE，2017：2261-2269．

[12] 秦磊，谢邦昌．L_1和L_2分位数趋势滤波及其集成方法[J]．数理统计与管理，2015，34(3)：442-451．

[13] LI B，DAI Y C，HE M Y．Monocular depth estimation with hierarchical fusion of dilated CNNs and soft-weighted-sum inference[J]．Pattern Recognition，2018，83：328-339．

[14] XU D，RICCI E，OUYANG W，et al．Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation[C]// Proc of Conference on Computer Vision and Pattern Recognition．Piscataway：IEEE，2017：161-169．

[15] CAO Y，WU Z，SHEN C．Estimating depth from monocular images as classification using deep fully convolutional residual networks[J]．IEEE Transactions on Circuits & Systems for Video Technology，2018，28(11)：3174-3182．