基于金字塔池化网络的道路场景深度估计方法

周武杰*①② 潘 婷① 顾鹏笠① 翟治年①
①(浙江科技学院信息与电子工程学院 杭州 310023)
②(浙江大学信息与电子工程学院 杭州 310027)

摘 要：针对从单目视觉图像中估计深度信息时存在的预测精度不够准确的问题，该文提出一种基于金字塔池化网络的道路场景深度估计方法。该方法利用4个残差网络块的组合提取道路场景图像特征，然后通过上采样将特征图逐渐恢复到原始图像尺寸，多个残差网络块的加入增加网络模型的深度；考虑到上采样过程中不同尺度信息的多样性，将提取特征过程中各种尺寸的特征图与上采样过程中相同尺寸的特征图进行融合，从而提高深度估计的精确度。此外，对4个残差网络块提取的高级特征采用金字塔池化网络块进行场景解析，最后将金字塔池化网络块输出的特征图恢复到原始图像尺寸并与上采样模块的输出一同输入预测层。通过在KITTI数据集上进行实验，结果表明该文所提的基于金字塔池化网络的道路场景深度估计方法优于现有的估计方法。

关键词：单目视觉；深度估计；神经网络；金字塔池化网络
中图分类号：TP391.4 **文献标识码**：A **文章编号**：1009-5896(2019)10-2509-07
DOI: 10.11999/JEIT180957

Depth Estimation of Monocular Road Images Based on Pyramid Scene Analysis Network

ZHOU Wujie①② PAN Ting① GU Pengli① ZHAI Zhinian①
①(*School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou* 310023, *China*)
②(*College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou* 310027, *China*)

Abstract: Considering the problem that the prediction accuracy is not accurate enough when the depth information is recovered from the monocular vision image, a method of depth estimation of road scenes based on pyramid pooling network is proposed. Firstly, using a combination of four residual network blocks, the road scene image features are extracted, and then through the sampling, the features are gradually restored to the original image size, and the depth of the residual block is increased. Considering the diversity of information in different scales, the features with same sizes extracted from the sampling process and the feature extraction process are merged. In addition, pyramid pooling network blocks are added to the advanced features extracted by four residual network blocks for scene analysis, and the feature graph output of pyramid pooling network blocks is finally restored to the original image size and input prediction layer together with the output of the upper sampling module. Through experiments on KITTI data set, the results show that the proposed method is superior to the existing method.

Key words: Monocular vision; Depth estimation; Neural network; Pyramid pooling network

1 引言

深度估计是使用1个或多个视点图像来预测场景图像深度信息的过程。深度信息是理解场景中几何关系的重要线索，一个具有颜色和深度通道的图像可以应用于各种任务，如立体匹配[1]、3D模型重建[2]、场景识别[3,4]、人类姿势估计[5]等。深度信息可以从包含左右视点的立体图像[6]或运动序列[7–9]中获得，它们分别从空间和时间上为理解立体结构提供了相对丰富的信息。相比之下，从单目图像[10,11]中估计深度的难度更大，也更模糊，因为它不允许在立体图像或时间框架之间进行匹配。因此，现在

收稿日期：2018-10-12；改回日期：2019-05-21；网络出版：2019-05-28
*通信作者： 周武杰 wujiezhou@163.com
基金项目：国家自然科学基金(61502429)，浙江省自然科学基金(LY18F0002)
Foundation Items: The National Natural Science Foundation of China (61502429), The Zhejiang Provincial Natural Science foundation (LY18F020012)

单目图像中主要考虑利用各种各样的几何图形组用于深度图的估计，然而，只有当相应的假设有效时，这些技术才能重构特定情况下的深度信息，其难度较大[12–14]。随着机器学习的发展，情况有所改善，由数据驱动的特征相较于人眼主要设计的特征具有更强的泛化能力。

迄今为止，研究人员对于从单目图像中深度估计进行了各种尝试。早期主要利用几何图形的内在特征，例如，Saxena等人[15]采用了多尺度的马尔可夫随机场 (Markov Random Field, MRF)，其利用图像的全局上下文和局部特征进行深度评估；之后，Saxena等人[15]将图像分割成均匀的块，并利用马尔科夫随机场MRF获得每个块的立体参数加以重建深度信息。Karsch等人[16]提出将参考彩色加深度图像的深度特征转换成一个输入的彩色图像。最近，基于深度学习的技术已经越来越成熟。Eigen等人[17]提出将两种深度网络结合的方式，将图像进行粗糙-精细两个尺度的特征提取，先在粗糙尺度获得低分辨率特征以预测全局分布，之后输入到精细尺度网络框架，从而完善局部的深度特征对于整体特征的优化；Laina等人[18]设计了一个基于ResNet架构的深度评估网络，通过加入自定义的上采样的结构来改善深度图的分辨率，网络训练阶段加入自定义的损失函数，通过阈值实现自适应的L1和L2两种函数的切换；Fu等人[19]引入了一个空间增加的离散化策略来离散深度，并将深度网络学习重新定义为一个有序回归问题且获得了较好的深度估计结果。之后，在此基础上，深度估计领域也提出较多的网络框架[20,21]。此外，深度估计领域引入了无监督或半监督学习等方式以学习深度估计网络[22–24]。这些方法从左视图中恢复一个正确的视图来设计重建损失来估计差异图。同时，近年有人提出了一些弱监督的方法，考虑了两排序的信息，并对深度进行了比较[25,26]。

针对当前深度估计精度不够准确的问题，本文提出了一种基于金字塔池化网络的道路场景深度估计方法。在网络模块构建中，加入了残差网络块的上采样模块从而使特征恢复到原始尺寸；利用金字塔池化操作变换到不同尺寸中以获取更精细的特征信息。本网络具体特点如下：(1)在特征提取阶段利用ResNet网络框架[27]提取特征，通过较深的网络框架提升了估计精度；(2)对于高级特征加入了金字塔池化网络块，通过利用全局上下文信息以及子区域的上下文信息对于每一个像素的深度信息都能起到较好的估计效果，并且在金字塔池化网络块中，通过加入插孔卷积扩大了卷积层的感受野，有效减少了计算的参数量；(3)上采样阶段采用上采样块的形式，通过扩大上采样神经网络的深度有效提高了深度估计的精度。在KITTI数据集上进行实验，结果表明所提基于金字塔池化网络的道路图像深度估计方法优于现有的估计方法。

2 神经网络框架模型

本文网络框架模型如图1所描述，本模型主要包含3部分：特征提取部分，上采样尺度恢复部分和金字塔池化网络部分。特征提取部分利用了ResNet[27]网络中的2种残差网络块，将原始图像不断变小直到原始尺寸的1/16。上采样恢复尺度部分通过上采样操作恢复尺寸，过程中融入特征提取部分中对应尺寸的特征图，从而充分利用特征信息。为了提高估计精度，对于残差网络块提取的高级特征图加入金字塔池化网在不同感受野下获取深度信息，最后一起输入到预测网络层进行深度图的估计。

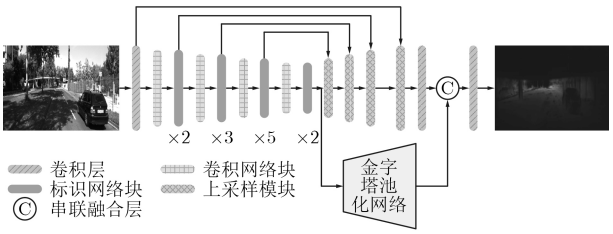

图 1 本文提出的神经网络框架

在本文卷积神经网络中使用了大量的跳跃连接(skip connection)，跳跃连接是融合特征提取模块的输出中低级特征和尺寸恢复模块的输出中高级特征，对于经过多层特征提取后高级特征中已经丢失了很多原始图像中物体的边缘轮廓信息，这对最终的深度预测结果是很不利的，而使用跳跃连接的方法就是为了将特征提取模块的输出中具有丰富轮廓信息的低级特征与尺寸恢复模块的输出中高级特征串联融合在一起以提高整个网络的预测结果精度。

2.1 特征提取部分

本文在特征提取部分主要采用ResNet模型引入了包含跳跃层的网络块，这种网络块主要由两个并行的网络层构成：(1)包含两个或更多的卷积；(2)第2种的网络层又分为两种：无操作和通过一个卷积层，最后将：(1)和(2)两个并行的网络层的输出求和。其中，根据是否对于输入另行卷积或者无操作将网络块分为两种，分别叫做：标识网络块(Identity_block)和卷积网络块(Conv_block)，其中，由于卷积网络块的第一个卷积层的步长都设为2，输出为输入图像尺寸的1/2。如图2所示为特征提取阶段用到的两种残差网络块。因此，本文的特

征提取实际上就是由多个标识网络块和卷积网络块相互连接组成，具体个数依次为：1个卷积网络块和2个标识网络块，1个卷积网络块和3个标识网络块，1个卷积网络块和4个标识网络块，以及1个卷积网络块和2个标识网络块。通过这个设计可以构建更深入的网络，不用担心出现网络的退化或由于数据量造成的欠拟合现象。这种极深的网络框架可以接受较大尺寸图片的输入，本文原始图像的输入尺寸为320×512，因此可以直接将原始图像输入。当原始图像通过4个神经网络块的卷积操作进行特征提取时，特征图的数量会在通过每个卷积网络块时增加1倍，最终产生的特征图的输出尺寸为20×32。在训练过程中，本文利用ResNet50在ImageNet数据集上的预训练权重作为初始权重，有效加快了整个网络的训练速度。

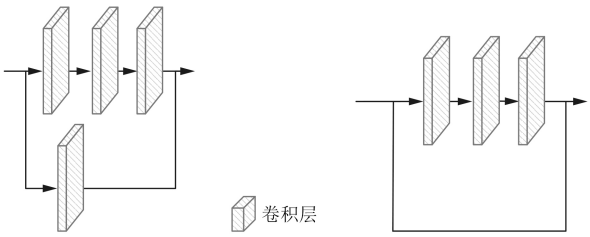

图 2 两种残差网络块块的结构图

2.2 恢复尺度部分

在恢复尺度部分，特征图的张量每经过一个上采样模块都会减半，从而保证了与特征提取部分网络的对称性。上采样模块主要由上采样层和标识网络块组成，在每个残差网络块的前后都连接了两个卷积层，考虑到来自上一层的特征图的数量可能与来自特征提取部分同尺度的特征图的数量不同，此时这些卷积层充当输入特征图和输出到上采样层的特征图的连接器，也保证了整个网络变得完全对称。最后，通过4个上采样模块将特征图恢复到原始图像的尺寸，为后续输入到预测层做准备，每个上采样模块具体如图3所示。实验中，考虑到当前主流的恢复尺度方法有3种：上采样层，反卷积层和利用卷积块的方式替代反卷积的操作，分别做实验最终发现上采样层的结果最好。

对于特征提取对应尺寸的特征图与上采样恢复尺度模块的融合，本文是使用像素相加融合层来完成的。

2.3 金字塔池化网络部分

通过观察发现在深度估计中，许多错误可能与上下文的联系和不同接受域的全局信息有部分关系。因此，本文加入了金字塔池化网络[28]，可以提

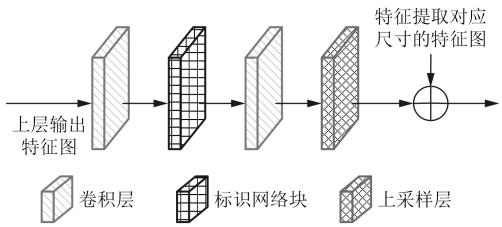

图 3 上采样恢复尺度模块

高场景解析的性能，它已经被证明是一种有效的获得全局特征的方式。在一个深度神经网络中，特征图的接受域的大小可以粗略地表示本文使用上下文信息的程度。尽管从理论上讲，ResNet的接受域已经大于输入图像，但Zhou等人[29]表示，卷积神经网络中的接受域实际比理论层面小得多，尤其是在高层特征上，这使得许多网络没有充分考虑到全局特征。本文利用金字塔池化网络模块来解决这个问题。全局平均池化是一个很好的用于全局联系上下文的基础，之前通常用于分类[30]。但是，在复杂场景图像中如果只用全局平均池化会漏掉部分必要的信息。因此，为了进一步减少不同子区域之间的上下文信息丢失，本文采用了一个具有层次化的全局神经网络模块，本模块包含了不同级别的信息，在不同的子区域之间通过池化操作进行改变，本文称它为金字塔池化网络模块，在深度神经网络的最终层之前的输出中，金字塔池化网络模块融合了4种不同的金字塔型。第1种是全局池化操作，以产生单个的块输出。之后的其他3种金字塔层将特征图划分为不同的子区域，并将不同的位置通过不同的集合加以表示。金字塔池化网络模块中不同级别的输出包含不同大小的特性图。例如为了能够输出全局特征，本文用卷积尺寸为1×1的卷积层将每个经过金字塔池化网络模块的特征图的维度降为原来的的$1/N$级。此外，金字塔的数量和每个层中卷积核的尺寸大小主要由输入的特征图的尺寸决定，考虑到此时输入的特征图的尺寸大小为20×32，本文将金字塔的数量设为4，每一层中的池化的尺寸大小分别为：1×1, 2×2, 3×3和6×6。通过池化操作之后，需要通过一个卷积层对于特征图提取全局特征，本文采用了带孔卷积的方式以扩大感受野。引入带孔卷积的原因有两个，一是在不损失图像分辨率大小的前提下扩大感知域范围。二是带孔卷积并不会增加参数量，不降低训练的速度。之后，本文将此网络模块输出的特征图通过双线性插值的方式上采样到原始图像尺寸，具体如图4所示，其中第1列代表不同尺寸的平均池化，第2列代表不同大小的插孔卷积层。使用融合层(concatenate)将4个经过上采样恢复到原始图像尺寸的金字塔特征图与上

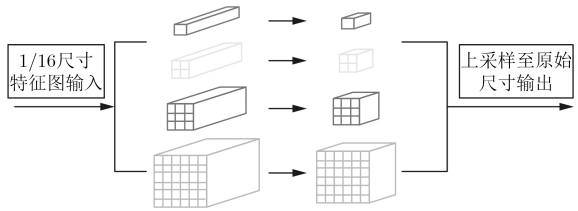

图 4 金字塔池化模块

采样恢复尺度部分输出的特征图串联融合后一起输入预测层，通过预测层的输出即为预测获得的道路图像的深度图。

本文将特征提取模块的输出经过一个金字塔变化网络再与尺寸恢复模块连接是为了借助于变化网络的金字塔池化进一步强化高级特征，以用于最终融合中加强具有准确度高的深度特征信息从而改善其输出结果。金字塔池化由于把一个特征图从不同的角度进行特征提取，再聚合的特点，本文使用它来对最终的高级特征从不同角度以不同尺寸再一次进行特征提取并直接分别上采样到原始图像的尺寸大小后与尺寸恢复模块连接，这样增强了准确度高的深度特征信息的作用，显著地提高了本文所提神经网络的性能。

2.4 损失函数

对于深度预测，考虑到欧几里得距离具有较强的实际意义，本文采用了一个包含不同阶数的损失函数作为边界条件来比较预测和真值的深度图的值，损失函数设置如式(1)

$$L\left(y_i, y_i^*\right)=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-y_i^*\right)^2-\frac{1}{2n^2}\sum_{i=1}^{n}\left(y_i-y_i^*\right)^2 \quad (1)$$

其中，n代表每张图片中的像素点的个数，y_i和y_i^*分别代表对应的原始图像以及预测图像中的像素点的深度值。

3 实验与分析

3.1 实验数据与实验环境

本文在实验中使用了KITTI官网[31]提供的训练集用于训练数据。由于数据集为视频剪辑的序列数据集，分别在原始的训练集序列和验证集序列中的每个序列中每隔10张选出1张，构成了本文当前使用的训练集和测试集；在训练集和测试集构造过程中，已经考虑了在不同的场景中获取，它们之间的相似性很小。其中，训练集包含4286张图片及其对应的深度图，测试集包含434张图片及其对应的深度图，考虑到模型复杂度和提升计算性能，将网络框架的输入尺寸定为320×512，同时由于对比的5个框架的输入尺寸都不一致，为了控制变量，其它方法的图像输入尺寸也为320×512。在Linux环境下完成了本次实验。实验在Ubuntu 16.04系统下进行，硬件设施如下：16 G内存，GPU为NVIDIA 1080Ti的显卡，包含11 G显存；通过Python语言完成实验。

3.2 评测指标与实验分析

通过在KITTI数据集上实验，采用6个指标比较预测的深度图与真实深度图之间的关联性，这6个指标分别为：均方根误差(Root Mean Square Error, RMSE)，对数平均误差(average Lg10 error, Lg)，对数均方根误差(Log root mean square error, Lg_rms)以及阈值下的精确度：$a1$, $a2$, $a3$。其中$a1$, $a2$, $a3$分别代表式(5)中a为1, 2, 3时的取值。

6个评测指标的计算公式为

$$\mathrm{RMSE}=\sqrt{\frac{1}{|T|}\sum_{i=1}^{|T|}\left(y_i-y_i^*\right)^2} \quad (2)$$

$$\mathrm{Lg}=\frac{1}{|T|}\sum_{i=1}^{|T|}\left|\lg y_i-\lg y_i^*\right| \quad (3)$$

$$\mathrm{Lg\_rms}=\sqrt{\frac{1}{|T|}\sum_{i=1}^{|T|}\left\|\lg y_i-\lg y_i^*\right\|^2} \quad (4)$$

y^*的有阈值的精度

$$\max\left(\frac{y}{y^*},\frac{y^*}{y}\right)<\delta_a, \delta_a=1.25^a, a=1,2,3 \quad (5)$$

其中，y是测试图像的真实深度值，y^*是测试图像的预测深度值，T是测试数据集中图像的数量。

为了优化模型，使用了Adam[32]作为优化器，并且用accurancy和loss函数，相较于随机梯度下降法，Adam作为优化器能够更快地找到下降的方向并且能够在尽可能少的迭代次数下下降到局部最小值点，充分提高了学习和训练的效率，能有效提高了深度估计的结果。

表1所示为本文方法与其它5个经典的深度学习方法(Fine_coarse[17], ResNet50[18], ResNet_fcn50[19], D_U[20]和UVD_fcn[21])在深度估计上的误差对比。对数平均误差，均方根误差和对数均方根误差这3个指标反应的是预测的深度图与实际深度图之间的误差，应该越小越好。相较于其他5种方法，本文所提的网络框架的误差都是最小的，实现了预测性能较大提升，尤其在均方根误差和对数平均误差这两个误差上都取得了比较好的改进。通过计算预测的深度图像与真实的深度图像的比值结果来比较预测的深度图像相对于真实数据的预测精度，结果表明本文方法在指标值为a1时，在精度预

表 1　深度图像的预测值与真实值之间的误差和相关性

	RMSE	Lg	Lg_rms	a1	a2	a3
Fine_coarse[17]	2.6440	0.272	0.167	0.488	0.948	0.972
ResNet50[18]	2.4618	0.243	0.126	0.674	0.943	0.972
ResNet_fcn50[19]	2.5284	0.247	0.134	0.636	0.950	0.979
D_U[20]	2.8246	0.305	0.127	0.634	0.916	0.945
UVD_fcn[21]	2.6507	0.264	0.145	0.566	0.945	0.970
本文方法	2.3504	0.230	0.120	0.684	0.949	0.975

测上获得了较好的结果。在指标值为a2和a3情况下接近加入了CRF层的文献中提出的方法，但是综合比较多个指标可以发现本文方法仍然取得了较好的结果。

在此基础之上，考虑到当前主流的恢复尺度方法有3种：上采样层，反卷积层和利用卷积块的方式替代反卷积的操作，分别进行实验之后，深度估计的对比结果如表2所示。从上述实验结果中，可以发现在考虑了需要计算的参数量和估计精度等两个因素后，本文采用了上采样层恢复尺度的方法性能最好，同时上采样对图直接插值操作，无需训练，因此相较于其他两种方式大大减少了参数量。

表 2　不同恢复尺度方法的结果

	RMSE	Lg	Lg_rms	a1	a2	a3
使用反卷积层恢复尺度的方法	2.3716	0.237	0.125	0.673	0.946	0.973
使用卷积块恢复尺度的方法	2.4724	0.240	0.129	0.646	0.948	0.974
使用上采样层恢复尺度的方法	2.3504	0.230	0.120	0.684	0.949	0.975

总之，通过在KITTI数据集上实验结果表明，在保证训练次数，训练批次大小等其他因素一致的情况下，进行对比实验发现本文模型获得了较好的结果。相较于其它方法，通过计算6个指标结果，本文的方法都取得了较好的结果，能够保证误差尽可能的小以及精确度的值尽可能高。

4　结束语

针对从单目视觉图像中恢复深度信息时存在的预测精度不够准确的问题，本文提出了一种基于金字塔池化网络的道路场景深度估计方法。首先，利用4个残差网络块的组合对图像提取特征，之后通过上采样将图片恢复到原始尺寸，考虑到上采样过程中不同尺度信息的多样性，将提取特征过程中的特征图与上采样过程中相同尺度的特征图融合并利用残差网络块增加网络的深度，提高深度信息估计的精确度。此外，为了充分利用残差网络块提取的高级特征，利用金字塔池化操作变换到不同尺寸中以获取更精细的特征信息，最后恢复到原始尺寸与上采样模块的输出一同输入预测层。通过在KITTI数据集上进行实验，结果表明本文提出的基于金字塔池化网络的道路图像深度估计方法优于现有的估计方法。本文的创新点主要在于充分考虑到了不同尺寸的特征图中包含的信息不同，通过两次上采样和连接的操作将不同尺寸的图像包含的特征信息融合到一起，对于局部场景的深度估计起到了增强作用。在以后的工作中，将继续完善该框架以期获得更好的预测精度，并且希望能够将此输出的深度图应用到其他领域中。

参 考 文 献

[1] LUO Yue, REN J, LIN Mude, *et al*. Single view stereo matching[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 155–163.

[2] SILBERMAN N, HOIEM D, KOHLI P, *et al*. Indoor segmentation and support inference from RGBD images[C]. The 12th European Conference on Computer Vision, Florence, Italy, 2012: 746–760.

[3] REN Xiaofeng, BO Liefeng, and FOX D. RGB-(D) scene labeling: Features and algorithms[C]. 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012: 2759–2766.

[4] SHOTTON J, SHARP T, KIPMAN A, *et al*. Real-time human pose recognition in parts from single depth images[J]. *Communications of the ACM*, 2013, 56(1): 116–124. doi: 10.1145/2398356.

[5] ALP GÜLER R, NEVEROVA N, and KOKKINOS I. Densepose: Dense human pose estimation in the wild[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 7297–7306.

[6] LUO Wenjie, SCHWING A G, and URTASUN R. Efficient deep learning for stereo matching[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 5695–5703.

[7] FLINT A, MURRAY D, and REID I. Manhattan scene understanding using monocular, stereo, and 3D features[C]. 2011 International Conference on Computer Vision, Barcelona, Spain, 2011: 2228–2235.

[8] KUNDU A, LI Yin, DELLAERT F, *et al*. Joint semantic segmentation and 3D reconstruction from monocular video[C]. The 13th European Conference on Computer Vision, Zurich, Switzerland, 2014: 703–718.

[9] YAMAGUCHI K, MCALLESTER D, and URTASUN R. Efficient joint segmentation, occlusion labeling, stereo and flow estimation[C]. The 13th European Conference on Computer Vision, Zurich, Switzerland, 2014: 756–771.

[10] BAIG M H and TORRESANI L. Coupled depth learning[C]. 2016 IEEE Winter Conference on Applications of Computer Vision, Lake Placid, USA, 2016: 1–10.

[11] EIGEN D and FERGUS R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture[C]. 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 2650–2658.

[12] SCHARSTEIN D and SZELISKI R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms[J]. *International Journal of Computer Vision*, 2002, 47(1/3): 7–42. doi: 10.1023/A:1014573219977.

[13] UPTON K. A modern approach[J]. *Manufacturing Engineer*, 1995, 74(3): 111–113. doi: 10.1049/me:19950308.

[14] FLYNN J, NEULANDER I, PHILBIN J, *et al*. Deep stereo: Learning to predict new views from the world's imagery[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 5515–5524.

[15] SAXENA A, CHUNG S H, and NG A Y. 3-D depth reconstruction from a single still image[J]. *International Journal of Computer Vision*, 2008, 76(1): 53–69.

[16] KARSCH K, LIU Ce, and KANG S B. Depth transfer: Depth extraction from video using non-parametric sampling[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 36(11): 2144–2158. doi: 10.1109/TPAMI.2014.2316835.

[17] EIGEN D, PUHRSCH C, and FERGUS R. Depth map prediction from a single image using a multi-scale deep network[C]. The 27th International Conference on Neural Information Processing Systems, Montréal, Canada, 2014: 2366–2374.

[18] LAINA I, RUPPRECHT C, BELAGIANNIS V, *et al*. Deeper depth prediction with fully convolutional residual networks[C]. The 4th International Conference on 3D Vision, Stanford, USA, 2016: 239–248.

[19] FU Huan, GONG Mingming, WANG Chaohui, *et al*. Deep ordinal regression network for monocular depth estimation[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 2002–2011.

[20] DIMITRIEVSKI M, GOOSSENS B, VEELAERT P, *et al*. High resolution depth reconstruction from monocular images and sparse point clouds using deep convolutional neural network[J]. SPIE, 2017, 10410: 104100H.

[21] MANCINI M, COSTANTE G, VALIGI P, *et al*. Toward domain independence for learning-based monocular depth estimation[J]. *IEEE Robotics and Automation Letters*, 2017, 2(3): 1778–1785. doi: 10.1109/LRA.2017.2657002.

[22] GARG R, VIJAY KUMAR B G, CARNEIRO G, *et al*. Unsupervised CNN for single view depth estimation: Geometry to the rescue[C]. The 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 2016: 740–756.

[23] KUZNIETSOV Y, STUCKLER J, and LEIBE B. Semi-supervised deep learning for monocular depth map prediction[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 6647–6655.

[24] GODARD C, MAC AODHA O, and BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 6602–6611.

[25] ZORAN D, ISOLA P, KRISHNAN D, *et al*. Learning ordinal relationships for mid-level vision[C]. 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 388–396.

[26] CHEN Weifeng, FU Zhao, YANG Dawei, *et al*. Single-image depth perception in the wild supplementary Materia[C]. The 30th Conference on Neural Information Processing Systems, Barcelona, Spain, 2016: 730–738.

[27] HE Kaiming, ZHANG Xiangyu, Ren Shaoqing, *et al*. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778.

[28] ZHAO Hengshuang, SHI Jianping, QI Xiaojuan, *et al*. Pyramid scene parsing network[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 6230–6239.

[29] ZHOU Bolei, KHOSLA A, LAPEDRIZA A, *et al*. Object

detectors emerge in deep scene CNNs[J]. arXiv preprint arXiv: 1412.6856, 2014.

[30] SZEGEDY C, LIU Wei, JIA Yangqing, *et al*. Going deeper with convolutions[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 1–9.

[31] UHRIG J, SCHNEIDER N, SCHNEIDER L, *et al*. Sparsity invariant CNNs[C]. 2017 International Conference on 3D Vision, Qingdao, China, 2017: 11–20.

[32] KINGMA D P and BA J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv: 1412.6980, 2014.

周武杰：男，1983年生，副教授，博士，研究方向为计算机视觉与模式识别，深度学习.

潘　婷：女，1994年生，硕士，研究方向为计算机视觉与模式识别.

顾鹏笠：男，1989年生，硕士，研究方向为计算机视觉与模式识别.

翟治年：男，1977年生，讲师，博士，研究方向为深度学习.