中图法分类号:TP301.6　**文献标识码**:A　**文章编号**:

论文引用格式:

单目深度估计技术进展综述

黄军,王聪,刘越*,毕天腾

2019.12.16f发布

北京理工大学光电学院,北京 100081

北京电影学院未来影像高精尖创新中心，北京 100088

中国工业互联网研究院，北京 100846

摘　要：**目的** 单幅图像深度估计是计算机视觉中的经典问题，对场景的三维重建、增强现实中的遮挡及光照处理具有重要意义。**方法** 本文回顾了近些年单幅图像深度估计技术的相关工作，介绍了单幅图像深度估计常用的数据集及模型方法。根据场景类型的不同，数据集可分为室内数据集、室外数据集与虚拟场景数据集。按照数学模型的不同，单目深度估计方法可分为基于传统机器学习的方法与基于深度学习的方法。基于传统机器学习的单目深度估计方法一般使用马尔可夫随机场(Markov random field，MRF)或条件随机场(conditional random field，CRF)对深度关系进行建模，在最大后验概率框架下，通过能量函数最小化求解深度。依据模型是否包含参数，该方法又可进一步分为参数学习方法与非参数学习方法，前者假定的模型包含未知参数，训练过程即是对未知参数进行求解；后者使用现有的数据集进行相似性检索推测深度，不需要通过学习来获得的参数。近年来，深度学习推动了计算机视觉各个领域的发展。本文阐述了基于深度学习的单目深度估计方法的国内外研究现状及优缺点，依据不同的分类标准，自底向上逐层级将其归类。第一层级为仅预测深度的单任务方法与同时预测深度及语义等信息的多任务方法。图片的深度和语义等信息关联密切，因此有部分工作研究多任务的联合预测方法。第二层级为绝对深度预测方法与相对深度关系预测方法。绝对深度是指场景中的物体到摄像机的实际距离，而相对深度关注图片中物体的相对远近关系。给定任意图片，人的视觉更擅于判断场景中物体的相对远近关系。第三层级包含有监督回归方法、有监督分类方法及无监督方法。对于单张图片深度估计任务，大部分工作都关注绝对深度的预测，而早期的大多数方法采用有监督回归模型，即模型训练数据带有标签，且对连续的深度值进行回归拟合。考虑到场景由远及近的特性，也有用分类的思想解决深度估计问题的方法。有监督学习方法要求每张 RGB 图片都有其对应的深度标签，而深度标签的采集通常需要深度相机或激光雷达，前者范围受限后者成本昂贵。而且采集的原始深度标签通常是一些稀疏的点，不能与原图很好的匹配。因此不用深度标签的无监督估计方法是近年的研究趋势，其基本思路是利用左右视图，结合对极几何与自动编码机的思想求解深度。**结果** 本文详细分析了各研究方法的研究动机、优缺点及适用范围。最后对单幅图像深度估计方法进行总结。 **结论** 展望了基于深度学习的单目深度估计方法未来的研究趋势与重点。

关键词：机器学习；深度估计；三维重建；深度学习.

Survey on the progress of monocular depth estimation technology

Huang Jun, Wang Cong, Liu Yue*, Bi Tianteng

School of Optoelectronics, Beijing Institute of Technology, Beijing 100081

Advanced Innovation Center for Future Visual Entertainment, Beijing Film Academy, Beijing 100088

收稿日期:　　；　**修回日期**:

基金项目:

*通信作者：liuyue@bit.edu.cn

Supported by the National Key Research and Development Program of China（No. 2018YFF0300802）and the National Natural Science Foundation of China (No. 61661146002) and the 111 Project (B18005).

China Academy of Industrial Internet, Beijing 100846

Abstract: Objective Depth estimation from a single image is a classical problem in computer vision. It is of great significance for scene reconstruction, occlusion and illumination processing in augmented reality. **Method** In this paper, the recent related literatures of single image depth estimation are reviewed and the commonly used data sets and methods are introduced. According to different types of scenes, the datasets can be divided into indoor datasets, outdoor datasets and virtual scene datasets. Considering the different mathematical models, monocular depth estimation methods can be divided into traditional machine learning based methods and deep learning based methods. Traditional machine learning based methods usually use Markov random field or conditional random field to model the depth relationships of pixels in an image. In the framework of maximum a posteriori probability, the depth can be obtained by minimizing the energy function. According to whether the model contains parameters or not, traditional machine learning based methods can be further divided into parameter learning methods and non-parameter learning methods. The former assumes that the model contains unknown parameters, and the training process is the way to obtain these unknown parameters; the latter uses existing data sets for similarity retrieval to infer depth, with which there is no parameters need to be solved. In recent years, deep learning has promoted the development of computer vision in many fields. The current research situations of deep learning based monocular depth estimation methods in China abroad are analyzed with their advantages and disadvantages. We classify these methods hierarchically in bottom up paradigm with reference to different classification criteria. The depth and semantics of images are closely related, some works focus on the topic of multi-task joint learning. Hence in the first level, we separate single depth estimation methods into single task methods of predicting depth only and the multi-task methods which predict depth and semantics at the same time. The second level contains absolute depth prediction methods and relative depth prediction methods. Absolute depth refers to the actual distance between the object in the scene and the camera, while relative depth focuses on the relative distance of the object in the picture. Given arbitrary images, people are often better at judging the relative distances of objects in the scene. The third level consists of supervised regression method, supervised classification method and unsupervised method. For single image depth estimation task, most of the work focuses on the prediction of absolute depth, among which most of the early methods used supervised regression model. In this way, the model regress on continuous depth values and the training data should contain depth labels. Considering the characteristics of the scene from far to near, there are also some researches to solve the problem of depth estimation with classification methods. Supervised learning methods require each RGB image to have its corresponding depth label, whose acquisition usually requires a depth camera or radar. While the depth camera is limited in scope and the radar is expensive. Furthermore, the original depth collected by depth camera are usually sparse, which cannot precisely match the original image. Therefore, the unsupervised depth estimation methods which need not depth label is the research trend in recent years. The basic idea is to combine the polar geometry based on left-right consistency with automatic coding machine to obtain depth. **Result** The motivation, advantages, disadvantages and application scope of each research method are analyzed in details in this paper. Finally, the depth estimation methods of single image are summarized. **Conclusion** The focus and trend of future research on depth estimation from a single image are prospected.

Key words: machine learning; depth estimation; 3D reconstruction; deep learning.

0 引 言

场景的深度估计是计算机视觉中的经典问题，对三维重建、遮挡处理与光照估计等问题有重要作用[30, 49]。普通摄像机在成像过程中只能记录场景的颜色信息，无法记录实际物体到摄像机之间的距离信息，即在三维空间投影到二维平面的过程中，丢失了深度信息。为了获取深度，研究者们尝试直接从RGB图像中估计深度，利用摄像机拍摄的图像估计每个像素对应的物点到相机的距离[6, 26]，在重建过程中不与重建物体接触[16]。相较于通过激光测距仪等各种硬件设备获取物体表面上一定数量点的深度，基于图像的深度估计方法由于不需要昂贵的设备仪器和专业人员，具有更广的应用范围。

基于图像的深度估计方法根据不同的输入图像数量可分为多幅图像深度估计方法与单幅图像深度估计方法。基于多幅图像的深度估计方法包括多视立体几何(Multi View Stereo，MVS)算法[2, 8]、运动中恢复结构(Structure From Motion，SFM)算法[7, 43]与从阴影中恢复形状(shape from shading，SFS)算法[50]等。MVS利用三角测量法对左右视图进行匹配计算深度，其原理类似人眼的双目立体成像过程，SFM则利用单摄像机捕获的时间序列图像获取深度，SFS一般利用灰度图像中变化的阴影恢复物体表面形状。但这些算法对输入图片均有各自的要求，普适性不足。

从单幅RGB图像中估计深度的方法也称单目深度估计方法，是计算机视觉领域近年来热门的研究课题，但该问题是一个病态问题[10]，其原因在于单张RGB图片对应的真实场景可能有无数个，而图像中没有稳定的线索来约束这些可能性。受人类能够轻易地利用经验和图像中的线索推断出单张图像对应的深度信息的启发，早期的研究根据光学原理，利用图像中的离焦信息恢复深度(Depth From Defocus，DFD)[1, 11, 36, 37]，其基本假设是图像中焦点所在位置景物最为清晰，离焦点越远模糊程度越深，但是这种算法仅适用于小景深图像。

近些年机器学习方法在物体识别[38]、场景理解[45]及手势识别[44]等领域得到了广泛的应用。基于RGB图像与深度图之间存在着某种映射关系这一基本假设，基于数据驱动的机器学习方法逐渐应用于单目深度估计问题中。本文主要对国内外最近十年的基于机器学习方法的单幅图像深度估计进行综述性论述，阐述了基于传统机器学习的单目深度估计的研究进展，分析总结了最新的基于深度学习的单目深度估计方法，最后对未来的研究趋势进行了展望。

1 研究基础与国内外研究现状

1.1 问题建模

机器学习旨在赋予计算机学习的能力，通过研究并构建算法使得计算机能够从特定的数据集中学习规律并做出预测和判断[3]。根据训练数据是否带标签，机器学习方法可分为监督学习方法与无监督学习方法。监督学习是将样本数据和其对应的输出数据作为输入，从中学习出一个将输入映射到输出的规则并对新的样本做出预测和判断。无监督学习方法则不需要训练标签。

在深度估计的研究中，由于室内外场景类型与深度范围具有较大的差异，对应不同的场景分别会构造不同的数据集。纽约大学 Silberman 等[42]提供的NUY depthv2是常用的室内数据集之一。作者选取了464个不同的场景，利用RGB相机和微软的Kinect深度相机同时采集室内场景的RGB信息和深度信息，收集了407024帧RGBD图像对构建数据集。由于红外相机和摄像机之间的位置偏差，深度相机采集的原始深度图存在缺失部分或是噪点，作者从中选取了1449张图片，利用着色算法[24]对深度图进行填充得到稠密深度图，同时人工标注语义信息。图1所示是NYU depth v2数据集中的图像示例，数据集中的深度范围从0.5m到10m。作者将1449个样本划分为795个训练样本与654个测试样本。

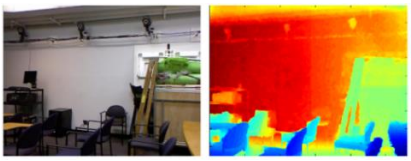

图1 NYU depth V2 数据集示例

Fig.1 Example of NYU depth V2 dataset

对于室外场景，常用的数据集是斯坦福大学Saxena等构建的Make3D数据集[40, 41]。作者使用激光扫描仪采集室外场景的深度信息，选取的场景类型为白天的城市和自然风光，深度范围是 5-81m，大于该范围统一映射为81m。数据集共包含534张RGBD图像对，其中400张用于训练，134张用于测试。RGB图像的原始分辨率为2272*1704，深度图的分辨率为55*305，图2示出了数据集的样本示

例。

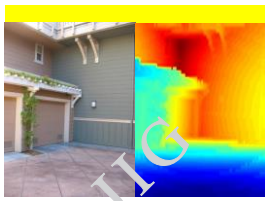

图 2 Make3D 数据集的样本示例

Fig.2 Example of Make3D dataset

自动驾驶领域常用的数据集 KITTI[13]也包含深度数据标签。KITTI 数据集由德国卡尔斯鲁厄理工学院和美国丰田技术研究院基于自动驾驶平台 Annieway 联合开发，通过一辆装配有 2 台高分辨率彩色摄像机、2 台灰度摄像机、激光扫描仪和 GPS 定位系统的汽车采集数据，其中激光扫描仪的最大测量距离为 120 米。图像场景包括卡尔斯鲁厄市、野外地区以及高速公路。数据集共包含 93000 个 RGBD 训练样本。

密歇根大学的 Chen 等[5]以相对深度作为标签构建了 Depth in the Wild 数据集。作者从词典中随机选取单词作为搜索关键字，从互联网中收集得到原始的 RGB 图片。标注工作外包给专门的工作人员，为了更加高效，每一张图片作者选取了两个高亮的点，工作人员只需判定两个点的远近关系即可。对于采样点对之间的位置关系，作者采用 50%随机采样，另外 50%对称采样的方法以使构建的数据集尽可能平衡。最终获得的有效标注图片约 500000M 张。

以上数据集都采集自真实场景，还有一些利用计算机生成的虚拟场景采集数据如 SceneNet RGB-D 数据集、SYNTHIA 数据集等 [32, 39]。由于是通过虚拟场景生成，数据集中包括更多天气、环境及光照，场景类型多样。各数据集有各自的优缺点，在实际研究中，应根据具体研究问题来选择合适的数据集。

除了训练数据以外，模型也是决定算法性能优劣的关键因素。传统的机器学习方法一般使用马尔可夫随机场(Markov random field，MRF)或条件随机场(conditional random field，CRF)对深度关系进行建模，在最大后验概率框架下，通过能量函数最小化求解深度。深度学习方法大多使用卷积神经网络(Convolutional Neural Networks，CNN)，通过逐层特征抽象拟合输入数据与输出深度之间的关系。

1.2 基于传统机器学习的单目深度估计

基于传统机器学习的单目深度估计方法可分为参数学习方法与非参数学习方法。参数学习方法是指能量函数中含有未知参数的方法，训练的过程是对这些参数的求解。2005 年斯坦福大学的 Saxena 等[22]利用 MRF 学习输入图像特征与输出深度之间的映射关系。作者利用图像中多尺度的纹理、模糊等深度线索，分别构建了高斯 MRF 和拉普拉斯 MRF 模型对单幅图像的深度进行估计。在特征提取阶段，作者在三个不同的尺度下进行了特征提取操作，分别将不同大小的图像块作为深度估计的基本单元，使用一组卷积滤波器与图像块进行卷积操作，将卷积输出及平方后的卷积输出作为绝对特征，用于估计每个图像块的深度。此外，作者使用相邻区域的滤波器组的输出产生的直方图的差作为相对特征，估计相邻区域间深度的差别。基于提取的特征向量，作者分别构建了高斯和拉普拉斯两种 MRF 模型。利用训练数据对模型参数进行估计求解后，即可通过最大后验概率来对测试图像进行深度估计。如图 3 所示，Saxena 等改进后的方法[41]在最大化后验概率框架下以超像素为单元，利用 MRF 拟合特征与深度、不同尺度的深度之间的关系，进而实现对深度的估计。Liu 等[3]以语义标签为辅助，使用不同的语义标签，分别以像素和超像素为节点构造双层 MRF 模型优化深度图。Wang 等[48]使用非线性空间中的核函数描述 RGB 图像和其深度图之间的关系，利用图像块学习参数进行深度估计。上述方法均需人为假设 RGB 图像与深度之间的关系满足某种参数模型，而假设模型难以模拟真实世界的映射关系，因此预测精度有限。

非参数学习方法使用现有的数据集进行相似性检索推测深度，不需要通过学习来获得的参数。2012 年，Karsch 等提出了 Depth Transfer 方法[19]，该方法利用 GIST 特征检索与输入图像最相似的图像集，然后基于 SIFT 流得到变形后的深度结果，最后对深度图进行优化。Mo 等[33]在上述方法的基础上提出基于前景背景融合的单目图像深度估计方法，其中前景深度主要反映场景显著性区域内的深度，背景深度反映场景整体的深度分布趋势，综合两种估计结果获得最终的深度图。澳大利亚国立大学的 Liu 等[29]于 2014 年提出了以超像素为节点的离散-连续 CRF 模型，用离散变量表达相邻超像素间的关系，用连续变量表达深度的连续变化。作者以超像素中心的深度及三维平面法线共四个参数表征超像素，训练高斯过程回归器对每个超像素的参数进行预测，进而计算得到每个超像素初始的深度值。然后

使用粒子信仰传播方法对离散-连续 CRF 模型的优化，相比于 Depth Transfer，该方法用高斯过程回归器替换了基于 SIFT 流的变形过程，提高了算法的效率并且通过结合离散变量和连续变量提高了深度估计的精度。Konrad 等[20]改进了上述方法，作者先将检索到的相似图像进行中值滤波产生初始深度图，然后用双边交叉滤波对初始深度图进行平滑。最后使用获得的深度图得到立体图像对中的右眼图像，完成 2D 到 3D 的转换，以此避免了复杂的 SIFT 流计算。非参数化方法不需要对模型进行人为假设，但由于依赖于图像检索，计算量大、耗时高，难以实际应用。

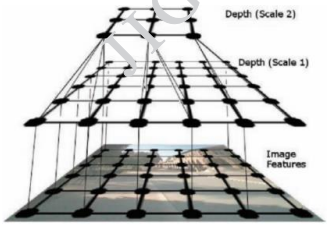

图 3 多尺度 MRF 的单目深度估计方法[41]

Fig.3 Multi-scale MRF for monocular depth estimation [41]

1.3 基于深度学习的单目深度估计

参数学习方法和非参数学习方法都实现了单幅图像的深度估计，但总体看来，两种方法的缺点主要体现在人为假设较多、处理过程烦琐等方面，因而更自然、更统一的模型框架是提高算法的发展方向。

近年来，深度学习发展迅速，深度神经网络以其强大的特征拟合能力和优异的性能，在计算机视觉、自然语言处理和语音识别等各个领域发挥了重要作用。神经网络由诸多神经元按照一定的拓扑结构连接而成。对于高维的图像数据，实验中常利用 Lecun 在 Lenet[23]架构中提出的 CNN 进行处理。CNN 利用权值共享的策略让一组神经元共享参数来节省计算开销。类似于动物视觉系统的多层抽象机制[18]，CNN 利用卷积核提取图像特征，通过深度神经网络对特征逐层抽象来完成高级的视觉任务。基于深度学习的单张图片估计方法，按照不同的分类标准，可以分为以下几类：

1) 有监督\半监督\无监督
2) 绝对深度\相对深度
3) 有后处理\无后处理
4) 单任务\多任务
5) 回归\分类

1.3.1 单任务

1) 绝对深度

(1) 有监督回归模型

图片的深度和语义等信息关联密切，因此有部分工作研究多任务的联合预测方法，在此先对单任务的深度估计方法进行讨论。绝对深度是指场景中的物体到摄像机的实际距离，而相对深度关注图片中物体的相对远近关系。对于单张图片深度估计任务，大部分工作都关注绝对深度的预测，而早期的大多数方法采用有监督回归模型，即模型训练数据带有标签，且对连续的深度值进行回归拟合。

2014 年 Eigen et al[10]首次将深度神经网络用于单目深度估计任务。作者提出使用两个尺度的神经网络对单张图片的深度进行估计：粗尺度网络预测图片的全局深度，细尺度网络优化局部细节。网络由两个堆栈组成，如图 4 所示，两个网络均以 RGB 图片作为输入，原始图片输入粗尺度网络后，得到全局尺度下场景深度的粗略估计结果。然后将粗尺度网络的输出作为附加的第一层图像特征传递给细尺度网络，对全局预测进行局部优化以添加更多的细节信息。粗尺度网络的任务是预测场景的全局深度，有效地利捕获诸如消失点、目标位置和空间对齐等深度线索。为了获得足够的感受野，网络上层使用的是全连接层。底层和中间层使用 max-pooling 将来自图像不同区域的信息组合成一个维度更小的特征图，以此整合对整个场景的理解。粗尺度网络的结构包含五个特征提取层以及两个全连接层，输出图片的分辨率是输入的 1/4。细尺度网络的任务是对全局预测的结果进行局部优化，网络结构仅包含卷积层和池化层，输出的神经元的感受野为 45*45 像素。通过设计，全局输出与细尺度网络第一层输出的尺寸大小相同，两者连接在一起，后续层使用零填充卷积以保持输出尺寸大小。除了最后一个卷积层是线性的，其它层均使用 ReLU 作为激活函数。训练时先对粗尺度网络进行训练，然后固定粗尺度网络的参数训练细尺度网络。作者还提出了一个尺度不变的误差评价函数，如下式。不同于考虑单个点的预测值与真值的误差，尺度不变误差考虑的是点对之间的误差，以此消除全局尺度的影响。

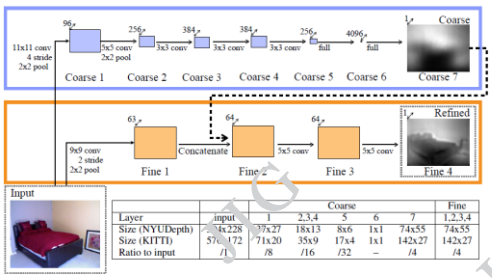

图 4 多尺度网络结构[10]

Fig.4 The structure of multi-scale network [10]

Eigen 等[9]基于上述工作改进后提出了一个统一的多尺度网络框架，分别将其用于深度预测，表面法向量估计和语义分割三个任务。值得一提的是，这里是将同一框架独立应用于不同任务，并不是多任务统一学习，因此将此归为单任务方法。不同的任务设定不同的损失函数，使用不同的数据集训练。网络模型是端到端的，不需要后处理。网络结构如图 5 所示，共包含三个尺度的网络，scale1 网络对整张图片做粗略估计，然后用 scale2 和 scale3 网络对全局预测进行细节优化。相较于方法[10]，网络模型作了如下改进：使用了更深的基础网络 VGG；利用第三个细尺度的网络进一步增添细节信息，提高分辨率；将 scale1 网络的多通道特征图输入 scale2 网络，联合训练前面两个尺度的网络，简化训练过程，提高网络性能。scale1 网络为了保证全局感受野使用了两层全连接层，输出特征图尺寸是输入图片的 1/16，通道数为 64。Scale1 输出经 4 倍上采样后输入到 scale2 网络中，最终 scale3 网络的输出尺寸是原图的一半。训练方法上，作者联合深度预测与法向量估计，共享 scale1 网络的参数。对深度估计任务的损失函数，作者在尺度不变损失的基础上添加了梯度正则项，计算预测梯度和真实梯度值的差异，对输出结果进行平滑优化。

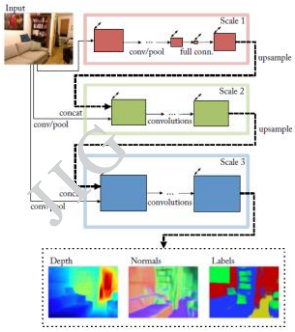

图 5 三尺度网络结构[9]

Fig.5 Three-scale network [9]

Liu 等[27]将深度卷积神经网络与连续条件随机场结合，提出深度卷积神经场，用以从单张图片中估计深度。对于深度卷积神经场，使用深度结构化的学习策略，在统一的神经网络框架中学习连续 CRF 的一元势能项和成对势能项。通过解析地求解配分函数的积分，可以精确的求解似然概率优化问题。对输入图像使用超像素分割，得到的超像素作为图模型中的节点，超像素利用其质心的深度描述。设 X 是输入图像，超像素个数为 n，Y 是包含 n 个连续深度值的矢量。在 CRF 模型，中，能量函数设定为在结点 N 的一元势能 U 和在连线处的成对势能 V 的组合，一元势能 U 的目的在于回归单个超像素的深度值，成对势能 V 鼓励相似外观的相邻超像素预测相近的深度。U 和 V 即是深度卷积神经场需要优化的目标函数。图 6 是深度卷积神经场的示意图。整个网络由一元势能部分，成对势能部分和 CRF 损失层组成。首先将输入图像分割成 N 个超像素，对每个超像素围绕其质心的截取一个图像块。一元部分将所有图像块作为输入，经过 CNN 后输出超像素中心对应的回归深度值，其网络结构由 5 个卷积和 4 全连接层组成。成对部分把所有邻近超像素对的相似向量作为输入，经过全连接层后输出包含一维相似性的向量。一元部分和成对部分的输出作为 CRF 损失层的输入，通过最小化负对数似然函数进行训练。一元势能为 CNN 输出的最小二乘损失，其网络基础架构为 AlexNet，由 5 个卷积层和 4 个全连接层组成。以超像素质心为中心的输入图像块大小为 224*224 像素。五个卷积层和前两个全连接层以 ReLU 作为激活函数，对于第三个全连接层以 Sigmoid 作为激活函数，最后一个全连接层没有激活函数。网络最终输出超像素质心的预测深度值。成对势能利用邻近超像素的一致性信息对结果施加平滑约束。为了使深度卷积随机场用反向传播方法进行优化，文中还对似然函数的导数进行了推导。

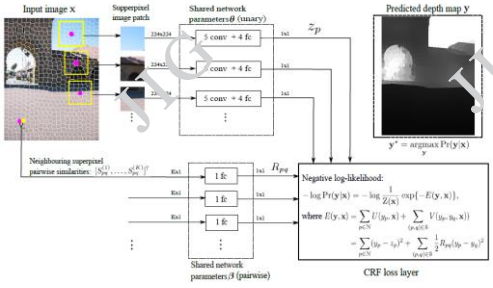

图 6 深度卷积场模型[27]

Fig.6 Deep convolutional field [27]

Liu 等[28]基于方法[44]改进后的工作整体框架类似，但由于原方法对原图需要生成众多超像素，对每一个超像素需要截取一个图片块输入网络进行训练，整体计算耗时。为此作者将超像素信息编码进神经网络中以提高计算效率。对于原 CRF 中的一元部分，作者改进后以原图作为输入，输出包含对应超像素个数的 n 维特征向量。其中，全卷积网络由卷积层，池化层和 ReLU 激活函数层组成，输出 d 维的卷积特征图。超像素池化的具体操作如图 7，首先将原图分割成 n 个超像素，得到超像素对应的 mask。同时将原图输入全卷积网络并利用反卷积网络上采样得到与原图同等尺寸大小的特征图。然后将超像素的 mask 作用在特征图上进行池化。超像素池化层的前传过程通过计算对应某类超像素出现的频率，作为特征图某个像素的权重，相乘累加后得到输出的特征向量。

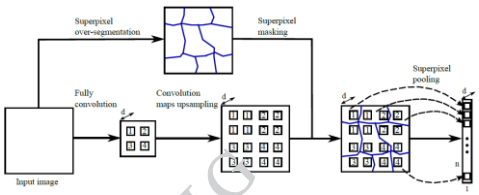

图 7 超像素池化层[28]

Fig.7 Super-pixel pooling layer [28]

Li 等[25]提出多尺度深度估计方法，首先用深度神经网络对超像素尺度的深度进行回归，然后用多层条件随机场后处理，结合超像素尺度与像素尺度的深度进行优化。具体过程如下：对于给定输入图片，用 SLIC 算法生成超像素；对每一个超像素，以其质心为中心截取多个尺度的图片块；用神经网络拟合输入图像块和对应深度之间的关系；最后用多层 CRF 对深度结果进行优化。网络结构如图 8 所示，以不同尺度的图片作为输入，前端以训练好的网络进行参数初始化，最后通过两个全连接层输出结果。实验表明多尺度图片作为输入有利于网络学习全局的深度信息。CRF 的能量函数由超像素尺度的一元势能，成对势能，以及像素尺度的自回归势能组成。第一项计算回归深度与网络预测深度的欧式距离，第二项用于约束相邻超像素之间的一致性，第三项基于某像素的深度可以用其局部领域的深度表示这一假设，得到相应的平滑约束项。值得一提的是，这也是将同一框架用于不同的任务的文献，作者也将此框架用于表面法向量预测。

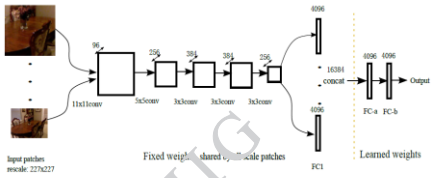

图 8 多层条件随机场处理框架[25]

Fig.8 The framework of multi-layer CRF [25]

Laina 等[22]提出了一种基于残差学习的全卷积网络架构用于单目深度估计，网络结构更深，并且不需要后处理。为了提高输出分辨率同时优化效率，文中提出了一种新的上采样方法；考虑到深度的数值分布特性，作者还引入了逆 Huber Loss 作为优化函数。网络整体架构如图 9，前端基于预训练好的残差网络 ResNet，后端使用新的上采样结构。在损失函数上，不同于传统的 L2 损失，作者提出用 reverse Huber 作为损失函数可以得到更好的结果。试验结果表明 BerHu 损失是两个范数之间的很好的平衡：因为有 L2 项，所以提高同一个像素中有高的残差项的权重；对于小的残差项 L1 项比 L2 项有更大影响。另外考虑到数据集中深度值的重尾分布，BerHu 损失也优于 L2 损失。

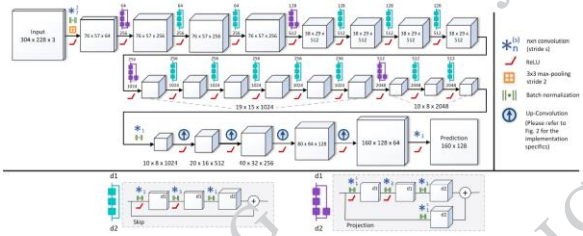

图 9 深度残差网络结构[22]

Fig.9 Deep residual network for depth estimation [22]

(2) 有监督分类模型

深度估计与语义分割任务具有一定的相似性，两者都是像素级的预测。不同之处在于深度值是连续的，常被当作回归处理，而语义标签是离散的，常被看作分类问题处理。考虑到场景由远及近的特性，也有用分类的思想解决深度估计问题的方法。

Cao 等[4]将深度估计问题当作像素级的分类问题处理。首先将深度值离散化，然后训练一个深度残差网络预测每个像素对应的类别。分类之于离散不同之处在于能够得出一个概率分布，便于后面使用条件随机场作为后处理进行细节优化。整体框架如图 10 所示，输入图片先经过一个全连接的残差网络，输出概率得分图，然后通过一个全连接的 CRF

模型进行优化。网络前端基于深度残差网络 ResNet，输入图片先经过 7*7 的卷积，3*3 的最大池化，然后通过 4 个卷积块，每个卷积块包含若干个残差块，基础架构不同残差块的个数不同，残差块包含两种连接，恒等映射与线性投影，后者只在匹配维度时使用。接着在通过 7*7 的平均池化，3 个全卷积层与 softmax 层得到输出的概率分布，得到的特征图大小为原来的 1/8，后端用诸如双线性插值的方法上采样，使得输出图片尺寸大小与输入相同。作者将深度值投影到对数空间，然后按照深度范围将连续的深度值离散化为一个个类别标签。损失函数使用的是包含信息增益的多项逻辑函数，不同于一般的多项式逻辑损失函数，改进后的损失函数对离真值越远的点会施加越大的惩罚，另外网络更加关注预测概率小于一定值，即难以训练的点。其后作者使用 CRF 进行后处理，其能量函数包含一元势能和成对势能。一元势能即为神经网络输出的类别概率，成对势能考虑各个像素对之间的关系。

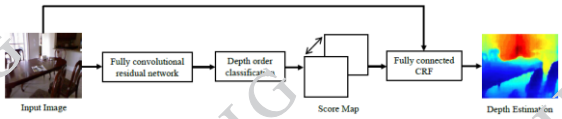

图 10 分类网络架构[4]

Fig.10 The framework of classification method [4]

(3) 无监督模型

有监督学习方法要求每张 RGB 图片都有其对应的深度标签，而深度标签采集通常需要深度相机或激光雷达，前者范围受限后者成本昂贵。再者，采集的原始深度标签通常是一些稀疏的点，不能与原图很好的匹配。因此不用深度标签的无监督估计方法是近年的研究趋势，其基本思路是利用左右视图，结合对极几何与自动编码机的思想求解深度。

Garg 等[12]提出利用立体图像对实现无监督单目深度估计，不需要深度标签，其工作原理类似于自动编码机。训练时利用原图和目标图片构成的立体图像对，首先利用编码器预测原图的深度图，然后用解码器结合目标图片和预测的深度图重构原图，将重构的图片与原图对比计算损失。整体框架如图 11 所示，利用两个已知焦距和偏移的相机获取立体图像对，编码过程为左图通过 CNN 得到深度图，利用深度视差公式计算深度图对应的视差图，解码过程结合视差图与右图利用几何知识解算得到重建的左图，两者计算重建误差，另考虑平滑项，使用简单的梯度 L2 正则项。为了使误差能够反传，作者利用泰勒展开计算数值导数。为了获取更好的细节信息，网络结构使用了 skip 结构与上采样层。网络训练时需要左右图像对，预测时只需要一张图片即可。

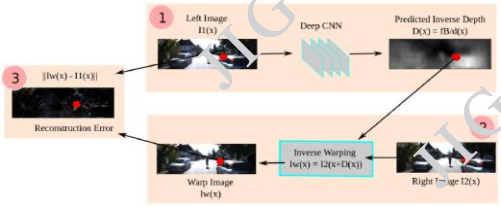

图 11 无监督深度估计基本框架[12]

Fig.11 The framework of unsupervised depth estimation [12]

Godard 等[14]对上述方法进行了进一步改进，作者利用左右视图的一致性实现无监督的深度预测。利用对极几何约束生成视差图，再利用左右视差一致性优化性能，提升鲁棒性。图 12 所示为三种不同的框架，首先为 naïve 版本，即将左图输入神经网络，输出基于右图的视差图，再由视差图及左图逆推右图，最后用重建的右图与源右图比较计算误差。但是这样得出的是基于右图的视差图，于是有了方案二，即文献[12]的工作。为了对结果进一步优化，作者提出利用左右视差一致性，即将左图输入网络同时输出左视差图与右视差图。然后利用左图与右视差图重构右图，右图与左视差图重构左图，最后联合统计误差。损失函数包含三项，第一项比较重建图像与原图的误差，第二项利用梯度信息平滑边缘细节，第三项用于保证左右视差图的一致性，即左视差图投影后应与右视差图一致。

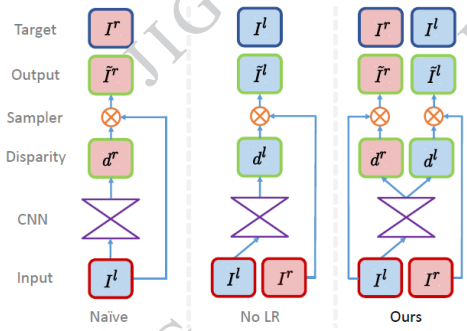

图 12 三种无监督学习框架[14]

Fig.12 Three types of unsupervised learning methods [14]

RGBD 数据集采集时通常需要同步使用 RGB 相机和传感器同时获取 RGB 图像和深度图像，但传感器中心和相机中心通常无法对齐。而且传感器采集的深度是稀疏的，存在噪点。Kuznietsov 等[21]提出将以稀疏深度图为标签的监督学习方法和无监督的学习方法相结合，即半监督学习，来进一步提高性

能。具体体现在损失函数上，如图 13 所示，监督学习的损失函数鼓励网络预测的深度图与传感器获得的深度图一致；无监督学习的损失函数鼓励重建图片与原图的一致性；梯度正则项用以平滑优化。

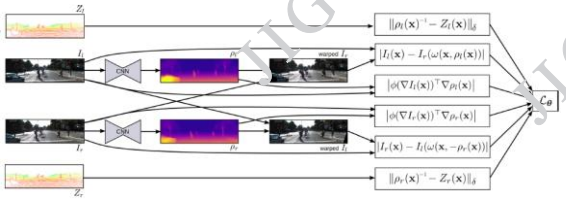

图 13 半监督学习深度估计框架[21]

Fig.13 Semi-supervised framework for depth estimation [21]

2）相对深度

给定任意图片，人们往往更善于判断场景中物体的相对远近关系。无监督学习方法不需要深度图作为标签，同样地，利用相对远近关系，不用深度值标签也可对深度进行估计。

Zoran 等[52]关注相对深度关系，不同于使用深度值作为标签训练网络的方法，作者提出利用图像中点对之间的相对关系推断深度信息。网络输出点对之间的相对关系，然后利用数值优化方法将稀疏的输出稠密化为最终结果。利用相对关系有如下几个优点：比数值回归更加简单；人们能够很容易判断相对关系，训练数据集获取成本低；相对关系不受数据的单应变换影响，因此系统更加鲁棒。作者也把该框架应用于反射率预测等其它任务上。整体框架由三部分组成，第一部分从图像中选择点对，第二部分估计每一个点对的相对关系，提取相关信息并做三分类，第三部分将点对之间的相对关系扩展至全局，得到稠密输出。第一部分，选取的点对最好是远离边缘，在某一块同质区域的中心，而两点之间的连线最好能够跨过边缘区域，这样才能尽可能多的考虑相对关系，一条边上的两个点即是需要比较的点对。作者利用超像素分割，选取超像素的质点作为基点。同时为了考虑更长距离的点对，作者结合了不同尺度的超像素信息。第二部分从点对到相对关系估计，相对关系仅有三种，更近更远或相等，即可以转化为三分类问题。训练神经网络做三分类，网络的输入由两部分组成，一是两个点对的局部特征与局部关系，二是全局信息与 ROI（感兴趣区域）。对点对中的每一个点，截取其周边的一小块作为局部特征，每个点相对于 ROI 的位置作为附加标记；将原图降采样后作为全局输入，ROI 相对于全图的位置作为附加标记。第三部分将点对之间的关系扩展至全局，一是需要保证全局关系的一致性，二是要填充稀疏的点。前者可作为一个带约束的优化问题加以解决，后者在超像素尺寸足够小的条件下可假设同一超像素中的结果均为同一常数。网络结构如图 14 所示，输入图片下采样至 64*64 像素，ROI 区域为一个矩形框，图像块的大小为 16*16，ROI 缩放到 32*32 的大小，附加的位置标记是高斯形状的斑点。网络使用 ReLU 作为激活函数，最后一层用 softmax 完成分类。损失函数设定为根据两个点之间的深度值关系赋予不同的类别标签。

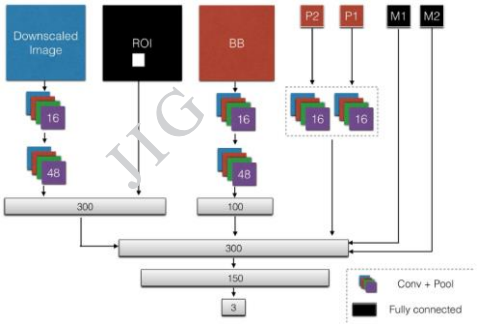

图 14 序列关系估计网络结构[52]

Fig.14 The framework of ordinal estimation network [52]

Chen 等[5]创建了一个新的数据集“Depth in the Wild”，包含任意图片以及图片中随机点对之间的相对深度关系，同时也提出了一个利用相对深度关系估计数值深度的算法。单张图片估计常用数据集诸如 NYU depth 或 Make 3D 都是在固定的摄像机参数下采集的，由此训练出的网络模型难以适用于任意测试图片。而实际场景多种多样，数据集中的图片往往无法囊括所有场景。为此作者构建了一个新的数据集 DIW，它包含 495k 图片，每一张图片都有随机采样的点对和它们的相对深度。对于数据集的构建，传统方法采集时都需要使用深度相机，场景类型受限。作者从网上收集图片，每一张图片中选取两个高亮点，通过众包方式利用人工标注它们的远近关系。考虑到图片信息与人力利用率，每张图片仅选取一对点进行标注。点位置的选取原则是一半随机选取，另一半选取对称点。方法[5]在预测时，对一张图片需要重复的计算各个点对之间的关系，计算耗时；输出图片以超像素为单位，精度有所欠缺；需要通过数值优化得到稠密输出，计算复杂。Chen 等利用相对深度关系构造损失函数，通过多尺度的神经网络直接预测像素级的深度。网络结构如下图 15 所示，采用沙漏型架构，保证输出图片的尺度大小与输出一致。

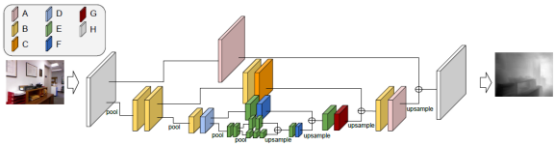

图 15 相对深度估计网络结构[5]

Fig.15 The framework of relative depth estimation net [5]

损失函数形式如下：

$$L(I,R,z)=\sum_{k=1}^{K}\psi_k(I,i_k,j_k,r,z) \qquad (1)$$

对于点对中的两个点 i 和 j，如果 i 点比 j 点近，则 r 为 1，否则为-1，近似相等则取 0。z 为网络预测深度。此损失函数的设计，让网络能够利用相对深度关系作为标签，深度值作为网络的输出结果，即将相对深度关系与连续深度值联系了起来。

基于 3D 传感器的数据集有一些限制，包括场景类型有限，如 NYU 仅含室内图像；数据集样本数量有限，如 Make3D 仅包含少量训练样本；深度标签稀疏，如 KITTI 数据集。Li 等[31]提出使用互联网上的图片，通过从运动中恢复结构（SFM）和多视点立体(MVS)方法生成训练数据，构造了一个深度数据集 MegaDepth。MVS 方法会产生噪点，图片中有些对象也无法重建，因此作者提出了新的数据清理方法，并使用语义分割生成的相对深度关系来自动增强数据。数据集的创建，作者从 Flickr 上下载图片，然后使用 SFM 或 MVS 方法重建三维模型，由此获得每张图片对应的深度图。由于瞬态物体，前背景遮挡以及深度的不连续性等原因，MVS 得到的深度图存在噪点，像素点稀疏，为此作者采用了如下方法：首先改进了 MVS 方法，迭代计算深度图，采用宁缺毋滥的原则，尽可能保证其与邻近深度图几何上一致。其次，结合语义分割，利用语义信息优化深度。使用 PSPNet 进行语义分类，将像素分为前景背景和天空三类。对前景中的某一块连通域，如果其中大于一半的像素都没有重建出深度，则将该连通域舍弃。对于一张图片，如果其中大于 30%的像素能够得到深度，则该图片将被标注深度值，否则标注相对深度关系。为了标注相对深度，在语义分割得到的结果后，前景区域选一块作为前景，背景区域选一块作为背景。网络结构上，作者比较了 VGG，ResNet 和沙漏网络，实验表明沙漏网络效果最好。损失函数由尺度不变误差项，梯度平滑项与相对深度关系三项构成。

1.3.2 多任务

由深度信息和其它信息之间的互补性，部分工作提出利用统一的框架联合多任务进行训练，不同任务提取的特征互相通信，提升最终效果。

由深度信息和语义信息之间的互补性，Wang 等[46]提出了一个统一的框架联合深度估计和语义分割任务，如图 16 所示。给定一张输入图片，用一个训练好的网络做全局预测，得到像素级的深度值和语义标签，两者联合训练比单一训练可获得更高的精度。为了进一步提高精度，作者将原图分割成各个区域，然后在全局预测的指导下得到区域级的深度和语义标签。最后利用两层的条件随机场，由像素级的全局预测和区域级的局部预测得到最后的优化结果。条件随机场的下层是由全局网络得到的像素级的深度和语义标签，下层是由局部网络得到的区域级的深度和语义标签。全局网络预测全局尺度和趋势，局部网络优化边缘细节。通过 CNN 的联合训练和 HCRF 的强化，深度估计和语义分割的结合在了一个统一的框架中。HCRF 中包含三种边，相邻像素之间，相邻区域之间，上层区域与下层对应区域的像素之间。能量函数包括像素级的一元势能，成对势能，区域级的一元势能，成对势能，区域与相应像素之间的势能。为了联合预测深度和语义标签，全局网络的损失函数包含两部分：前者计算预测深度与实际深度的损失，后者计算分类损失。训练方法上，由于数据中的语义标签较少，作者先训练深度网络，然后加上语义标签微调。最终的网络同时预测深度和语义标签。对于局部网络，其输入为图像区域，以该区域像素数占比最多的语义类别作为该区域的语义标签，以归一化后的相对深度作为该区域的深度标签。但是由于图像的全局信息缺失，局部区域的深度难以学习，但局部区域的深度类型有限，因此作者用深度模板加以代替。

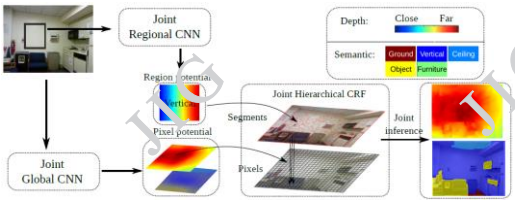

图 16 HCRF 多尺度联合深度语义网络架构[46]

Fig.16 HCRF the framework of multi-scale net for depth and semantic estimation [46]

Mousavian 等[34]提出一个同时预测深度和语义标签的模型，对每个任务先训练网络的一部分，再使

用单个损失函数对两个任务同时优化微调整个网络，最后结合 CNN 与 CRF，利用语义和深度线索对细节进行优化。这篇文章把深度信息作为辅助，主要关注语义分割性能的提升。输入一张 RGB 图片，首先对深度和语义标签做初始预测，然后将两个预测结果利用 CRF 结合得到最终的语义标签。使用一个语义和深度统一的损失函数对网络参数训练优化，最终训练好的网络模型既可以用于单任务预测，也可用于多任务预测。网络结构如图 17 所示，蓝色部分是共享参数，红绿色部分是独立参数，损失函数包含深度损失和语义损失两项，语义使用 Softmax 损失函数，深度使用尺度不变误差。初始语义标签，深度标签和输入图片最终通过一个全连接 CRF 进行优化。CRF 的能量函数包含一元势能和成对势能，一元势能即网络输出的概率，成对势能关注像素之间关系对语义标签的影响，其权值由深度值，RGB 值和空间位置三者决定。网络训练分三个阶段，第一阶段只需要语义分割网络，第二阶段在添加深度网络进一步优化，第三阶段联合 CNN 和 CRF 用反向传播进行优化。

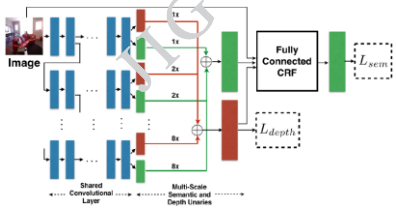

图 17 CRF 联合深度语义网络架构[34]

Fig.17 The framework of joint CRF [34]

Wang 等[47]提出将深度与多种信息结合的框架，如图 18 所示，首先用四个支路的 CNN 同时预测输入图片的表面法向量，深度，平面，边缘；然后用稠密条件随机场，以平面信息和边缘信息作为辅助，使得表面法向量和深度的预测结果兼容。该方法将 CRF 与 CNN 结合，推导了能量函数的梯度，使得整体框架能够通过反向传播算法加以优化。文章重点在于 DCRF 能量函数的设计，包括综合考虑深度与法向量的一元势能，将四种信息结合的成对势能。

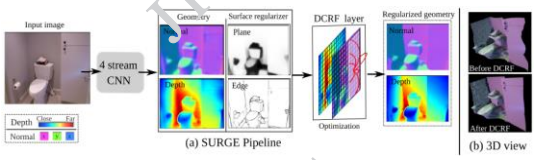

图 18 联合训练架构[47]

Fig.18 The framework of joint training method [47]

2 结 论

在单目深度估计问题中，常用的精度评估指标有相对误差(relative error，REL)、均方根误差(roor mean squared error，RMS)、对数误差(lg error，LG)及阈值误差(% correct)[51]，它们的具体形式如下：

$$REL=\frac{1}{N}\sum_{i=1}^{N}\frac{\left|D_i-D_i^*\right|}{D_i^*}, \tag{2}$$

$$RMS=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(D_i-D_i^*\right)^2}, \tag{3}$$

$$LG=\frac{1}{N}\sum_{i=1}^{N}\left|\lg D_i-\lg D_i^*\right|, \tag{4}$$

$$\%correct:\max\left(\frac{D_i}{D_i^*},\frac{D_i^*}{D_i}\right)=\delta<thr. \tag{5}$$

其中 D_i 表示图像中第 i 个像素的估计深度值，D^*表示对应的真值，N 表示像素总数。阈值精度表示估计深度的最大相对误差在指定阈值内的像素数量的多少，thr 是阈值，被设为 1.25、1.25^2、1.25^3。在定量的性能评估上，深度学习算法普遍优于传统的机器学习方法。由于神经网络强大的特征提取能力，深度学习模型也往往具有更好的鲁棒性与泛化能力。

总结全文，可以看到基于深度学习的单目深度估计是本领域的发展方向。目前，该领域的发展主要集中在数据集和深度学习模型两方面。首先，数据集的质量在很大程度上决定了模型的鲁棒性与泛化能力，深度学习要求训练数据必须有更多的数量、更多的场景类型，如何构建满足深度学习的数据集成为一个重要的研究方向。目前，基于虚拟场景生成深度数据具有不需要昂贵的深度采集设备、场景类型多样、节省人力成本等优势，结合真实场景和虚拟场景的数据共同训练也是未来深度学习方法的趋势[27]。其次，为了提高深度学习估计单幅图像深度的精度，要求更新的更复杂的深度框架。除了神经网络模型本身结构的优化，更新颖的算法设计也能有效地提升预测精度[15, 17, 35]。研究工作大多采用有监督回归模型对连续的绝对深度值进行回归拟

合。考虑到场景由远及近的特性，也有用分类模型进行绝对深度估计的方法。由深度信息和其它信息之间的互补性，部分工作结合表面法线等信息提升深度预测的精度。深度学习发展迅速，新的模型层出不穷，如何将这些模型应用于单幅图像深度估计问题中需要更加深入的研究。另外，探索神经网络在单目深度估计问题中学到的是何种特征也是一个重要的研究方向。

参考文献(References)

[1] Asada N, Fujiwara H, Matsuyama T. Edge and depth from focus[J]. International Journal of Computer Vision, 1998, 26(2):153-163.

[2] Barnard S T, Fischler M A. Computational stereo[J]. ACM Computing Surveys, 1982, 14(4):553-572.

[3] Bishop C. Pattern recognition and machine learning[M]. New York: Springer-Verlag, 2006.

[4] Cao Y, Wu Z, Shen C. Estimating depth from monocular images as classification using deep fully convolutional residual networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 28(11): 3174-3182.

[5] Chen, Weifeng, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild[C]//Advances in Neural Information Processing Systems. 2016: 730-738.

[6] Criminisi A, Reid I, Zisserman A. Single view metrology[J]. International Journal of Computer Vision, 2000, 40(2):123-148.

[7] Dellaert, Frank, Steven M. Seitz, Charles E. Thorpe, and Sebastian Thrun. Structure from motion without correspondence[C] //Proceedings of the Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2000:557-564.

[8] Dhond U R, Aggarwal J K. Structure from stereo-a review[J]. IEEE Transactions on Systems Man & Cybernetics, 1989, 19(6):1489-1510.

[9] Eigen, David, and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture[C]//Proceedings of the IEEE international conference on computer vision. 2015: 2650-2658.

[10] Eigen, David, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network[C]//Advances in neural information processing systems. 2014: 2366-2374.

[11] Favaro, Paolo, and Stefano Soatto. A geometric approach to shape from defocus[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2005, 27(3):406-417.

[12] Garg, Ravi, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue[C]//European Conference on Computer Vision. Springer, Cham, 2016: 740-756.

[13] Geiger, Andreas, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite[C] //Proceedings of the Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2012: 3354–3361.

[14] Godard, Clément, Oisin Mac Aodha, and Gabriel J. Unsupervised monocular depth estimation with left-right consistency[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 270-279.

[15] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets[C]//Advances in neural information processing systems. 2014: 2672-2680.

[16] Herbort, Steffen, and Christian Wöhler. An introduction to image-based 3D surface reconstruction and a survey of photometric stereo methods [J]. 3d Research, 2011, 2(3):1-17.

[17] Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[18] Hubel, David H., and Torsten N. The period of susceptibility to the physiological effects of unilateral eye closure in kittens[J]. The Journal of physiology, 1970, 206(2): 419-436.

[19] Karsch, Kevin, Ce Liu, and Sing Bing Kang. Depth transfer: depth extraction from video using non-parametric sampling[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 36(11): 2144-2158

[20] Konrad, Janusz, Meng Wang, and Prakash Ishwar. 2d-to-3d image conversion by learning depth from examples[C] //Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2012: 16-22.

[21] Kuznietsov, Yevhen, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6647-6655.

[22] Laina, Iro, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks[C]//2016 Fourth international

conference on 3D vision (3DV). IEEE, 2016: 239-248.

[23] LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[24] Levin, Anat, Dani Lischinski, and Yair Weiss. Colorization using optimization[C]//ACM transactions on graphics (tog). ACM, 2004, 23(3): 689-694.

[25] Li, Bo, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1119-1127.

[26] Liu, Beyang, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels[C] //Proceedings of the Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2010:1253-1260.

[27] Liu, Fayao, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 5162-5170.

[28] Liu, Fayao, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(10): 2024-2039.

[29] Liu, Miaomiao, Mathieu Salzmann, and Xuming He. Discrete-continuous depth estimation from a single image[C] //Proceedings of the Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014: 716-723.

[30] Liu Wankui and Liu Yue. Review on illumination estimation inaugmented reality[J]. Journal of Computer-Aided Design & Computer Graphics, 2016, 28(2): 197-207.

[31] Li, Zhengqi, and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2041-2050.

[32] McCormac, John, Ankur Handa, Stefan Leutenegger, and Andrew J. Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation[C]// Proceedings of International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2697 – 2706.

[33] Mo Yiming. Depth Estimation of Monocular Video using Non-parametric Fusion of Multiple Cues[D]. Nanjing: Nanjing University of Posts and Telecommunications. Nanjing University of Posts and Telecommunications, 2014.

[34] Mousavian, Arsalan, Hamed Pirsiavash, and Jana Košecká. Joint semantic segmentation and depth estimation with deep convolutional networks[C]//2016 Fourth International Conference on 3D Vision (3DV). IEEE, 2016: 611-619.

[35] Nath Kundu, Jogendra, Phani Krishna Uppala, Anuj Pahuja, and R. Venkatesh Babu. Adadepth: Unsupervised content congruent adaptation for depth estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2656-2665.

[36] Nayar S K and Nakagawa Y. Shape from focus[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1994, 16(8):824-831.

[37] Pentland A P. A new sense for depth of field[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1987, 9(4):523-531.

[38] Pope, Arthur R., and David G. Lowe. Learning object recognition models from images[C] //Proceedings of International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 1993:296-301.

[39] Ros, German, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes[C] //Proceedings of the Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016:3234-3243.

[40] Saxena, Ashutosh, Sung H. Chung, and Andrew Y. Ng. Learning depth from single monocular images[C] //Proceedings of the 18th International Conference on Neural Information Processing Systems Conference. Cambridge: MIT press, 2005: 1161-1168.

[41] Saxena A, Schulte J, Ng A Y. Depth estimation using monocular and stereo cues[C] //Proceedings of the 20th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc., 2007: 2197-2203.

[42] Silberman, Nathan, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images[C]//European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012: 746-760.

[43] Tomasi C, Kanade T. Shape and motion from image streams under orthography: a factorization method[J]. International Journal of Computer Vision, 1992, 9(2):137-154.

[44] Trigueiros P, Ribeiro F and Reis L P. A comparison of machine learning algorithms applied to hand gesture recognition[C] //Proceedings of Information Systems and Technologies. Los

Alamitos: IEEE Computer Society Press, 2012:41-46.

[45] Wang H Y, Gould S and Koller D. Discriminative learning with latent variables for cluttered indoor scene understanding[C] //Proceedings of European Conference on Computer Vision. Berlin Heidelberg: Springer-Verlag, 2010:497-510.

[46] Wang, Peng, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price and Alan L. Yuille. Towards unified depth and semantic prediction from a single image[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2800-2809.

[47] Wang, Peng, Xiaohui Shen, Bryan Russell, Scott Cohen, Brian Price, and Alan L. Yuille. Surge: Surface regularized geometry estimation from a single image[C]//Advances in Neural Information Processing Systems. 2016: 172-180.

[48] Wang Y G, Wang R P and Dai Q H. A parametric model for describing the correlation between single color images and depth maps[J]. IEEE Signal Processing Letters, 2014, 21(7):800-803.

[49] Xu, W., Y. Wang, Y. Liu, and D. Weng. Survey on occlusion handling in augmented reality[J]. Journal of ComputerAided Design & Computer Graphics, 2013, 25(11): 1635-1642.

[50] Zhang, Ruo, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape from shading: a Survey[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1999, 21(8):690-706.

[51] Zhuo, Wei, Mathieu Salzmann, Xuming He, and Miaomiao Liu. Indoor scene structure analysis for single image depth estimation[C] //Proceedings of the Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 614-622.

[52] Zoran, Daniel, Phillip Isola, Dilip Krishnan, and William T. Freeman. Learning ordinal relationships for mid-level vision[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 388-396.

作者简介

黄 军，1995 年生，男，北京理工大学光电学院硕士研究生，研究方向为深度学习与计算机视觉，具体研究课题为基于单张图像的深度估计方法。
E-mail：huagnjun@bit.edu.cn.

刘 越，男，北京理工大学光电学院教授、博士生导师，光电信息技术与颜色工程研究所所长，北京市混合现实与新型显示工程技术研究中心副主任。2000 年获吉林大学通信与信息系统博士学位，先后在美国加州大学伯克利分校、佐治亚理工学院、天普大学以及澳大利亚国立大学等有关实验室访问研究，主要研究领域包括虚拟现实与增强现实、自然人机交互、以及计算机视觉等，兼任中国计算机学会虚拟现实专业委员会副主任；中国系统仿真学会 3D 教育与装备专业委员会副主任；中国人工智能学会智能交互专业委员会副主任；中国图形图像学会理事、副秘书长、青年委员会执行委员；北京图形图像学会常务理事、秘书长等，目前主持国家科技支撑计划、国家高技术发展计划(863 计划)和国家自然科学基金等多项课题的研究工作，已发表论文 100 余篇，申请专利 30 余项，研究成果“交互式显示关键技术及应用”曾荣获 2017 年度教育部发明奖一等奖。
E-mail：liuyue@bit.edu.cn.

王 聪，高级工程师，担任全国信息技术标准化委员会图形图像分委会（SAC/TC28/SC24）秘书长，虚拟现实产业联盟（IVRA）标准委员会主任委员，国际标准化组织 ISO/IEC JTC1 SC24 WG7（图像处理与交换）召集人、ISO/IEC JTC1 SC41（物联网及相关技术）、IEC TC124（可穿戴电子设备与技术）专家，《虚拟现实与智能硬件（中英文）》编委，国家标准化管理委员会“国家标准技术评估专家”。长期从事虚拟现实与增强现实领域的国际、国内标准化工作，制定国家“十三五”虚拟现实/增强现实领域标准体系，作为标准主要完成人（前三）制定虚拟现实/增强现实领域国家标准 10 余项，填补了我国在虚拟现实/增强现实领域标准化建设的空白，为加快增强现实、虚拟现实等技术推广应用，引导行业良性发展起到重要支撑作用。

毕天腾，目前于北京理工大学光电学院攻读博士学位，主要研究方向是单幅图像逆渲染、深度学习、虚拟现实与增强现实。