基于深度学习的单目图像深度估计的研究进展

李阳*, 陈秀万, 王媛, 刘茂林
遥感与地理信息系统研究所, 北京大学地球与空间科学学院, 北京 100871

摘要 利用二维图像来进行场景的深度估计是计算机视觉领域的经典问题之一,也是实现三维重建、场景感知的重要环节。近年来基于深度学习的单目图像深度估计发展迅速,各种新算法层出不穷。介绍了深度学习在这一领域的应用历程与研究进展,采用监督与无监督两类方式分别系统地分析了有代表性的算法与框架,综述了深度学习在单目图像深度估计领域的研究进展与变化趋势,总结了当前研究的缺陷与不足,展望了未来研究的热点。

关键词 视觉光学; 单目视觉; 场景感知; 深度学习; 深度估计; 三维重建
中图分类号 TP391 **文献标识码** A **doi**: 10.3788/LOP56.190001

Progress in Deep Learning Based Monocular Image Depth Estimation

Li Yang*, Chen Xiuwan, Wang Yuan, Liu Maolin
Institute of Remote Sensing and Geographic Information System, School of Earth and Space Sciences, Peking University, Beijing 100871, *China*

Abstract Obtaining depth estimation of a scene from a two-dimensional image is a classic computer vision problem that plays an important role in three-dimensional reconstruction and scene perception. Monocular image depth estimation based on deep learning has been developing rapidly in recent years with new methods being proposed rapidly. This study discusses the application history and research progress in deep learning-based monocular depth estimation and analyzes several representative deep learning algorithms and network architectures in detail for both supervised and unsupervised learning. Finally, the research progress and trend of the deep learning in the monocular depth estimation field are summarized. Existing problems and future research priorities are discussed as well.

Key words visual optics; monocular vision; scene perception; deep learning; depth estimation; three-dimensional reconstruction
OCIS codes 330.7310; 200.4260

1 引 言

场景深度信息在当下许多研究课题中都起着至关重要的作用,如三维(3D)立体重建[1]、障碍物检测[2-3]、视觉导航[4]等。利用激光、结构光等在物体表面的反射获取深度点云,完成表面建模与场景深度估算的方法,在一些专业场景下的应用已经相当成熟[5-6],然而要获取稠密而精确的深度信息通常需要极高的成本,甚至难以实现。相比之下,基于图像进行深度估算的方法,无需价格相对高昂的仪器设备和专业人员,可应用的范围更广[7]。

要通过普通的二维(2D)图像恢复场景深度,立体视觉方法是解决这一问题的常用手段之一。使用两个摄像头观测同一个场景获得两幅图像,利用三角测量法从两幅图像间的视差得到深度信息[8-9]。毫无疑问,立体视觉方法需要至少两个相机,且它们的相对位置关系必须保持固定,这限制了其应用范围。对于纹理较稀疏的场景,图像中难以找到足够多的特征用于匹配,这样将会导致双目图像的深度估算误差显著增大甚至失效[10]。因此,目前有许多研究人员将目光转向了单目图像的深度估算上。由于单目图像缺少诸如运动、立体视觉关系等可靠的

收稿日期: 2019-03-20; **修回日期**: 2019-04-03; **录用日期**: 2019-04-11
基金项目: 国家重点研发项目(2017YFC1500900)
* **E-mail**: yang.li2012@pku.edu.cn

深度线索，要恢复出三维空间中才有的深度在本质上是一个不适定的问题，单目图像上一点的真实深度在理论上可以有无数个解[11]。为此，研究人员提出了各种不同的方式来实现单目图像的深度估计。

早期研究人员使用诸如图像消隐点[12]、对焦与离焦[13]、阴影[14]等深度线索从单目图像中获取深度信息，然而这些方法大多对图像有特殊要求，随着计算机视觉的不断发展，许多人工设计特征被相继提出，诸如尺度不变特征变换(SIFT)、分层梯度方向直方图(PHOG)等人工设计特征也都被用于单目图像的深度估计中[15]，由于人工设计特征本身往往只能捕获到图像的局部信息，因此以 Saxena 等[16-17]为代表的研究人员利用这些特征构建条件随机场(CRF)、马尔可夫随机场(MRF)等概率图模型，并考虑全局信息和长距离信息，将问题转化为一个随机场下的学习问题，取得了许多成果[18]。

随着卷积神经网络(CNN)以及深度学习方法的兴起与发展，深度学习在语义分割[19]、目标跟踪[20]等领域的优异表现吸引了许多研究人员，深度学习方法在单目图像深度估计这一问题上的应用也已开始被探索。本文着眼于深度学习在单目图像深度估计问题上的应用，采用监督型与无监督型两类方式分别综述其研究历史与进展，介绍了其中的典型模型与算法，探讨了目前使用深度学习估计单目图像深度的难点与热点，在此基础上进行总结并对未来进行展望。

2 深度学习在单目图像深度估计中的应用概况

现在普遍认为，当前所广泛使用的深度学习概念最早由 Hinton[21]在 2006 年提出，而深度学习真正成为学术界与产业界的热点是在 2012 年 Hinton 团队[22]构建的 CNN 网络 AlexNet 以极大优势获得 ImageNet 图像识别比赛第一名后。从此，各种神经网络结构层出不穷，对于深度学习的研究也炙手可热。2014 年，Eigen 团队[23]首次利用卷积神经网络对单目图像进行深度估计，他们设计了一种包含两个尺度的卷积神经网络结构，将整个深度估计过程分为两步：1)在整幅图像上对场景全局深度的粗估计；2)通过图像局部特征优化粗估计深度图的精估计。这一方法获得了相当精确的深度估计结果。Eigen 等的工作开创了深度学习在单目图像深度估计领域的先河。

在此之后，不少研究人员在 Eigen 团队工作的基础上，通过设计不同的神经网络结构、使用新的约束条件与损失函数对单目图像的深度估计进行优化改进。Eigen 团队[24]于 2015 年将深度估计、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，并将结果的分辨率提升至输入图像分辨率的一半。Grigorev 等[25]将长短期记忆(LSTM)用于循环网络以获取图像全局信息，将其与一般的卷积神经网络混合使用实现了端到端的单目图像深度估计。Liu 等[26]把 CNN 与 CRF 统一于一个框架内，将两个卷积神经网络分别对应能量函数中包含超像素内深度信息的项和关于相邻超像素关系的项，计算其最大化后验概率。Laina 等[27]采用了更深的残差网络并使用小卷积代替大卷积来实现上采样，使得深度估计更为高效，并且提出了新颖的损失函数，从而可以得到更好的结果。Cao 等[28]将原本连续的图像深度离散化为一定深度范围的类别，并将深度估计问题转化为分类问题，使用全卷积的深度残差网络实现分类，最后使用条件随机场优化结果得到最终的深度估计值。

上述方法在网络的训练阶段都需要图像的参考深度，是监督型的学习，然而精确的深度信息并不容易获取，在监督型深度学习方法发展的同时，也有众多研究人员采用无监督的深度学习方法来解决单目图像的深度估计问题。Xie 等[29]于 2016 年提出一种通过深度卷积神经网络将单张图像生成有一定视差的新视角图的方法以实现 2D 转 3D，以此为基础，众多研究人员开始使用左右视图来训练神经网络，如 Garg 等[30]提出了无监督的框架，在编码阶段，通过全卷积神经网络生成深度图，在解码阶段，使用传统的双目摄像头测距原理重构源输入图像，对比原输入图像，构建目标函数，从而反向训练网络。Godard 等[31]使用类似的方法，利用左视图同时产生左右两张视图的视差图，并通过引入左右视图的一致性损失提高最后输出结果的质量。Zhou 等[32]则利用两个全卷积神经网络，用视频连续帧的不同视角的几何信息作为监督信息，同时完成单目图像深度估计和相机的运动估计。Casser 等[33]在此基础上，先将场景内包括相机在内的物体分解成一个个的 3D 目标，在去除静态掩模后，为每个动态物体的运动进行独立建模，然后进行深度估计，这一方法在获取诸如汽车、行人等动态物体的深度时优势明显。

3 监督型深度学习方法

监督型的方法是指基于目标的真实值来进行网

络训练的深度学习方法[34]。在单目图像的深度估计问题上，需要向神经网络输入通过激光或其他方法得到的作为真值的参考深度图来训练网络。相比无监督的深度学习方法，监督学习通常可以获得更高的精度，因此有大量研究人员从网络结构、损失函数、约束条件、问题转化等角度开展了许多研究，提出了许多不同的算法。

3.1 全卷积残差网络的应用

要想让深度学习的效果更优，一个简单的想法就是增加网络深度，然而随着网络的加深，会出现梯度消失或梯度爆炸现象[35]。且随着网络深度进一步加深，准确率会饱和并逐渐降低[36]，这一问题并不是由过拟合造成的，而是因为随机梯度下降(SGD)在模型过于复杂时的优化变得十分困难，导致训练误差增大。因此 He 等[37]在 2016 年提出了深度残差网络 ResNet，引入残差的概念，设计跳跃连接。这样在没有新增额外参数与计算复杂度下，整个网络仍然可以通过由带有反向传播的 SGD 进行端到端的训练，且可以有效缓解准确率饱和的情况。

Laina 等[27]首次将 ResNet 应用于单目图像的深度估计中，相比于之前已经广泛使用的 AlexNet 和 VGG 模型[23-24, 38]，ResNet 更深的网络深度带来的最直观的优势就是其具有更大的感受野，因此可以接收分辨率更高的输入图像，保证其全局信息可以更好地得到保持。Laina 等提出的网络结构如图 1 所示，分为两部分，第一部分是基于 ResNet-50 的卷积神经网络，第二层是一连串的上池化和反卷积层，用于上采样，从而提升最终输出结果的分辨率。

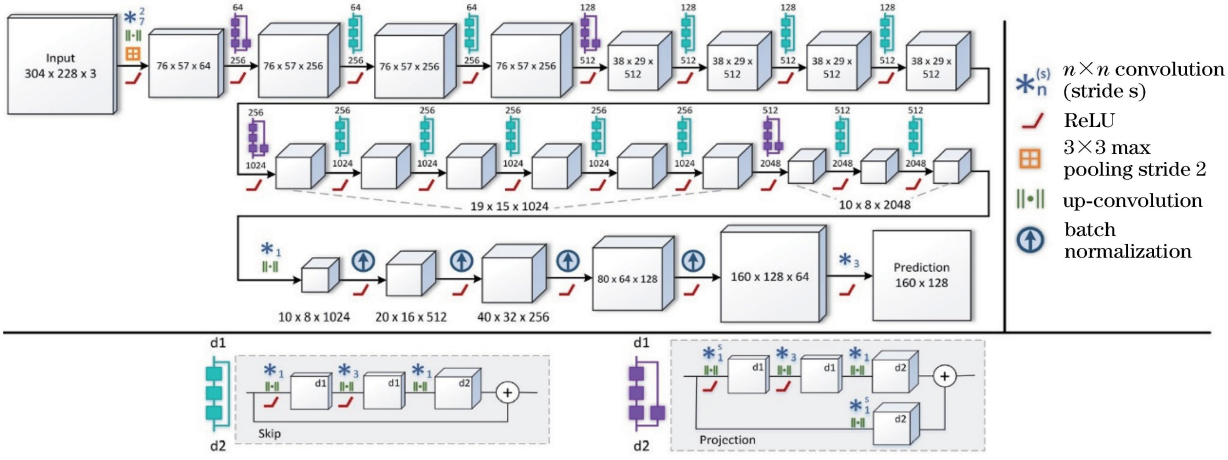

图 1 Laina 等[27]提出的网络结构图

Fig. 1 Network architecture proposed by Laina *et al*. [27]

在上采样的过程中，作者提出了“上投射(up-projection)”的概念，即在上卷积之后添加一个 3×3 的卷积，随后再加上一个从低质量分辨率特征映射至最终结果的投射连接，如图 2 (a)所示，这样可以使高层的信息在网络中的传递更为高效。为了提升这一过程的效率，作者将 5×5 的卷积核替换成了 4 个更小的卷积核，如图 3 所示，这样不仅可以避免原来上池化结果的特征图中包含大量零值的情况，同时使用小卷积核也节省了运算时间。

此外，为了进一步改善训练的结果，使用逆

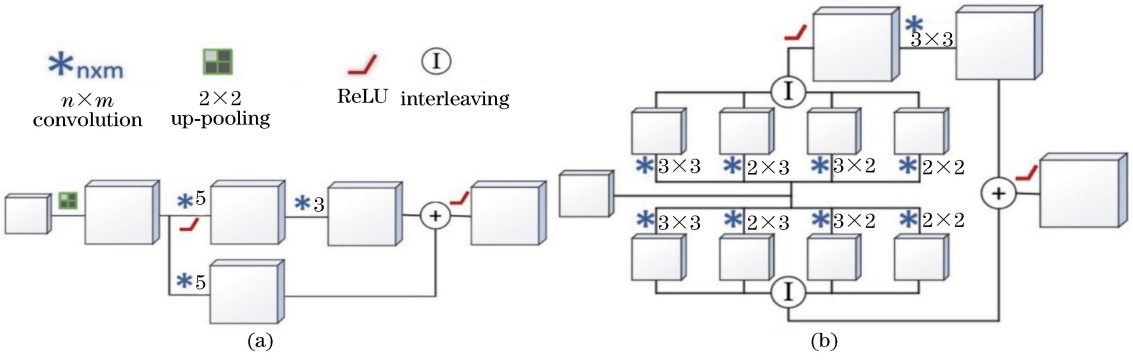

图 2 上投射结构示意图。(a)上投射；(b)快速上投射

Fig. 2 Schematics of up-projections. (a) Up-projection; (b) fast up-projection

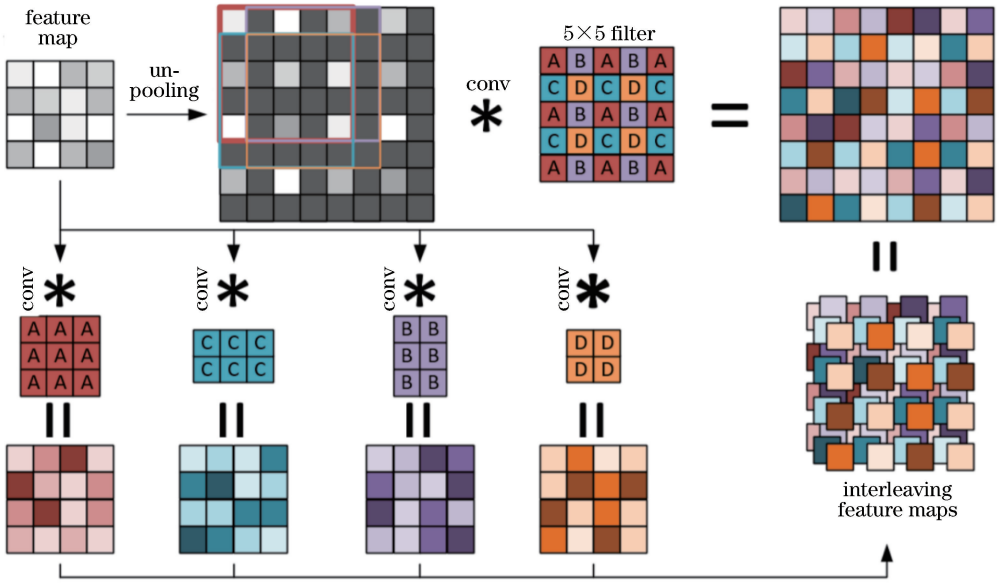

图 3 使用 4 个小卷积核替换传统 5×5 的卷积核

Fig. 3 Replacing tranditional 5×5 convolution kernels with four small convolution kernels

Huber(berHu)代替常用的 L_2 范数作为损失函数,即

$$B(x)=\begin{cases}|x|, & |x|\leqslant c\\ \dfrac{x^2+c^2}{2c}, & |x|>c\end{cases},\tag{1}$$

式中:$c=\dfrac{1}{5}\max_i(|\tilde{y}_i-y_i|)$,即每批数据深度估计误差最大值的 20%,$\tilde{y}_i$ 为估计的深度值,y_i 为真实深度值,i 为像素位置;B 为 berHu 损失;x 为自变量。berHu 损失函数综合了 $\mathscr{L}_1$ 和 $\mathscr{L}_2$ 范数的优点,保留了高残差项的高权重,对于低残差项采用效果更好的 $\mathscr{L}_1$ 范数。ResNet 以及以此为基础发展出来的 DenseNet 可以有效地解决神经网络效果随层数增加的退化问题,获得更平滑的结果,相比于 Eigen 等[23]于 2014 年的实验结果,其全局误差明显降低。在这之后,大量研究人员均采用这类神经网络结构取得了不错的效果[10,39]。

3.2 结合语义信息的深度估计

场景感知包含了许多方面的内容,其中深度估计描述了空间上的几何关系,而语义信息表征了场景内不同部分的实体含义,这些任务共享相似的上下文信息[40]。于是许多研究人员希望将语义信息与深度信息结合起来,在同一框架下使用神经网络来处理。如:Eigen 团队[24]在 2015 年就与 Fergus 团队合作将深度估计、表面法线预测和语义标注三个任务统一在一起;Liu 等[18]则将两项任务拆解,先完成语义分割,然后用语义分割的结果指导图像的深度估计;Chen 等[41]在无人驾驶场景下使用 ResNet-50 神经网络对图像进行目标识别,并同时获得目标的深度,然后对图像进行语义分割。

Jiao 等[42]通过统计分析,指出图像的像素个数在像素深度和语义标签上都呈现长尾分布。如图 4 所示,图片中大部分像素集中于一个较小的深度区间内,而深度值较大的像素点则以长尾形式零散地分布。相似的情况也体现在语义的分布上。这一问题是由透视投影的性质所决定的,因此单纯地通过增加训练样本的方式无法解决这一问题。在之前的大部分研究中,损失函数对所有像素一视同仁,没有考虑深度值的分布。由于低深度值的像素点更多,损失函数会被这些低深度值的像素点主导,故模型更容易预测出偏小的深度值,尤其是在高深度值的区域这一表现更加明显。

针对这个问题,作者提出一个协同作用的网络,将深度估计与语义分割两个任务结合起来,使用一个注意力驱动的损失来指导训练,解决上述的数据长尾分布的偏差。网络结构如图 5 所示。输入的 RGB 图像经过 VGG/ResNet 的主干编码器被转变为高维的特征图,然后输入至深度估计和语义分割两个子网络,信息在两个子网络间通过横向共享单元(LSU)传递,在子网络内则通过半稠密上跳跃连接(SUC)传递,整个训练过程由一个包含深度感知与注意力焦点的损失来监督。

在协同网络中,LSU 以一个动态路由的方式学习共享策略。每隔两个反卷积层,添加一个LSU模

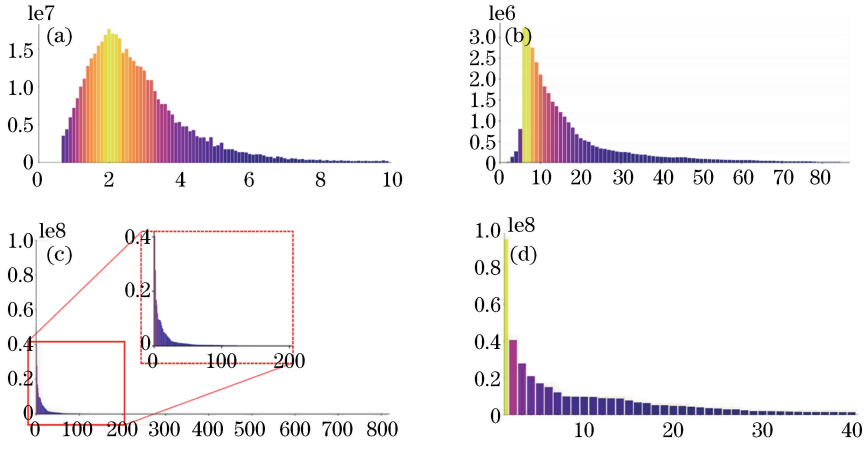

图 4 深度与语义标签的长尾分布。(a) NYU Depth V2 数据集的像素深度分布;(b) KITTI 数据集的像素深度分布;(c)(d) NYU Depth V2 数据集的像素-语义标签分布(全类别/40 类)

Fig. 4 Long-tail distributions on depth and semantic labels. (a) Pixel-depth distribution of NYU Depth V2 dataset; (b) pixel-depth distribution of KITTI dataset; (c)(d) pixel-semantic label distributions of NYU Depth V2 dataset (all categories/40 categories)

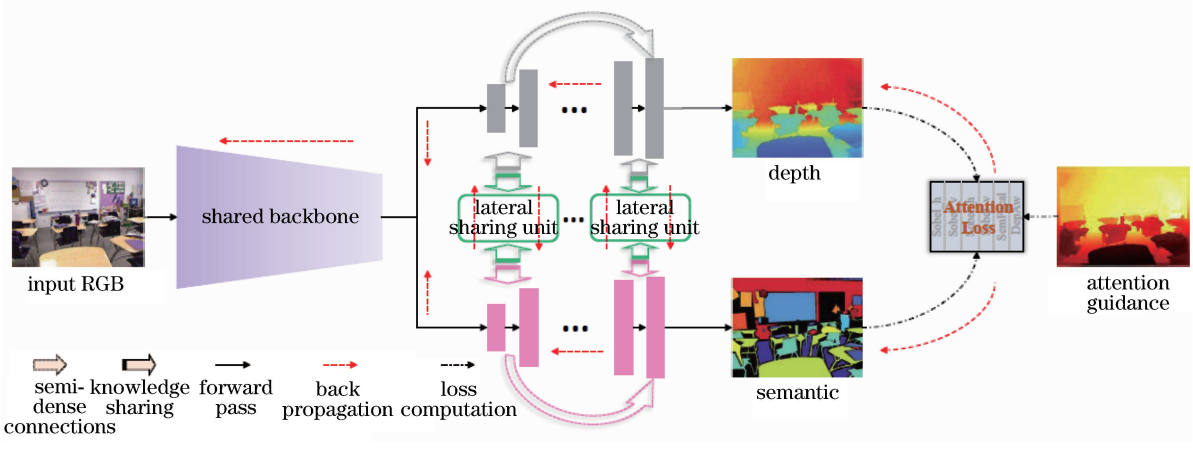

图 5 Jiao 等[42]提出的网络结构示意图

Fig. 5 Schematic of network architecture proposed by Jiao *et al.* [42]

块,以获得另一个任务共享的残差知识,这一过程在前向和后向传播中均会生效。LSU 的结构如图 6(a)所示,在输入到下一层进行上卷积之前,对该层结果加上一项来自于代表协同深度估计和语义分割这两个任务的融合信息。融合信息中两部分的比例通过两个参数 φ 和 γ 来控制。这一结构其实与残差模型的结构比较相似,如果控制信息共享项的权重都为 0,那么输入到下一层的就是单纯的原输出了。虽然所有 LSU 的结构都是一致的,但是每个 LSU 的权重都是在训练过程中自动学习而来,这种动态的方式使得信息能够更灵活地在两个子网络间传递。

子网络内部的 SUC 的目的是为了保持长时记忆的传递,最终的输出 f_{out} 不仅取决于上一层结果,还包含了之前每层输出的上卷积,如图 6(b)所示。作者通过插值保证最终的维度都相同,然后把插值后各个结果直接求和,即

$$f_{\text{out}}=\hbar(f_{\text{in}})+\sum\hbar(f_k), \tag{2}$$

式中:$\hbar(\cdot)$ 是一个由上卷积和卷积核为 3 的卷积层组成的结构;f_{out} 为 SUC 的输出;f_{in} 为 SUC 的输入;f_k 为网络中第 k 层的中间结果。

这一长-短连接过程与之前许多单目深度估计所用的多尺度网络策略有着异曲同工之处,即将低维特征关注的全局信息与高维特征关注的局部信息相融合。

最后,作者提出的注意力驱动的损失,包含三个部分,分别是考虑到深度分布差异的深度感知损失 L_{DA}、考察训练值与真实值差异的联合梯度损失 L_{JG} 和借鉴焦点损失[43]的语义焦点损失 L_{semF}。这三者

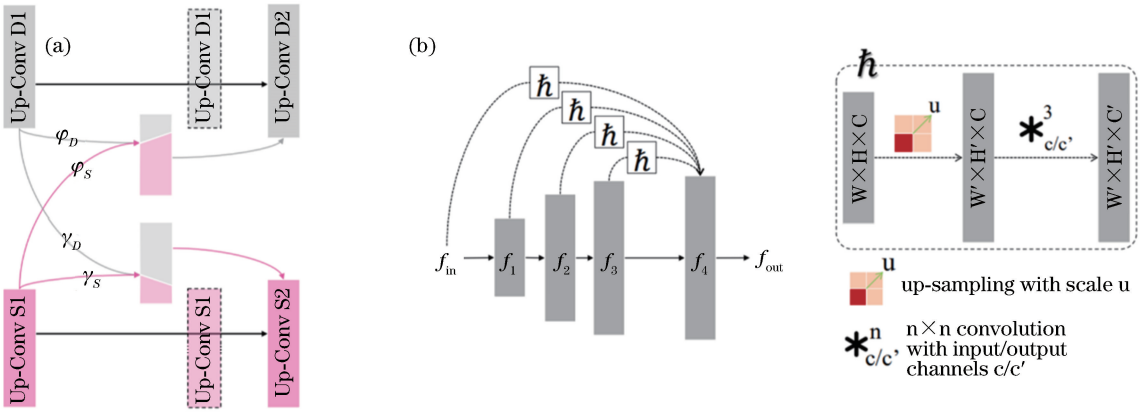

图 6 横向共享单元(LSU)和半稠密上跳跃连接(SUC)结构示意图。(a) LSU;(b) SUC

Fig. 6 Schematics of proposed LSU and SUC connections. (a) LSU; (b) SUC

之和即为总损失。其中,深度感知损失 L_{DA} 应重点关注,其表达式为

$$L_{\mathrm{DA}}=\frac{1}{N}\sum_{i=1}^{N}(\alpha_{\mathrm{D}}+\lambda_{\mathrm{D}})\cdot D(d_i,d_i^{\mathrm{GT}}),\tag{3}$$

式中:N 为图像像素数;α_{D} 是深度感知关注项,用于消除深度长尾分布的偏差,它与深度值正相关,使网络可以关注到深度值相对更大的区域,作者在实验中使 $\alpha_{\mathrm{D}}=d^{\mathrm{GTn}}$,$d^{\mathrm{GTn}}$ 即归一化后的参考深度值;λ_{D} 是为了避免在训练开始时出现梯度消失,以及在邻近区域学习时产生截断而添加的正则化项,类似于学习率,若训练中预测深度值越接近参考值,λ_{D} 越接近 0,反之越接近 1;D_i 和 d_i^{GT} 分别是该像素预测的深度值与参考真值,在实验中,使用逆平滑的 L_1 范式[27]作为距离度量 D。

作者将该方法与其他单目图像的深度估计方法在 NYU Depth V2 数据集上进行了对比,其总体精度有一定提升,相对误差可控制在 0.1 以内,但是这一方法对于距离较远的像素结果要明显优于距离相对较近的像素。此外,由于引入的 SUC 和使用的损失都强调了全局信息的长程传递,因此整体过于平滑,在细节的表现上有所缺憾。

3.3 从分类角度出发进行深度估计

单目图像深度估计问题的核心是构建一个关联图像信息和深度信息的模型[44],其本质上是一个回归问题。理想的预测结果是连续的深度值,然而以往的从回归角度出发的深度估计方法,都面临着分辨率相对较低、网络结构日益复杂、计算成本高昂等问题。因此有部分研究人员尝试通过把深度离散化,将深度估计问题转变为分类问题来实现对深度的估计。Cao 等[28]使用全卷积的深度残差网络来解决深度“分类”问题,并利用 CRF 对输出进行后处理优化,取得了不错的效果。Li 等[45]则率先使用空洞卷积[46],并在网络结构中通过分层融合策略融合不同尺度的信息,使用多分类的逻辑回归损失,更快地获得了相当精度的深度估计结果。

更进一步地,Fu 等[47]提出使用深度有序回归网络来处理这一问题,因为对深度进行分类不同于一般的图像分类问题,深度值的分类是有次序的。根据远近关系,所有的类别可以按照由远至近或由近至远的次序排列,即所谓的有序回归(ordinal regression)[48]。将回归问题转化为分类问题的首要工作就是要将连续的深度值离散化。考虑到预测的不确定度将随着深度值的增大而升高,所以预测较大深度值时可以允许相对大一些的误差。据此,作者提出了一种名为空间递增的离散化方法(SID),SID 是在对数空间上的线性采样,以避免使用平均采样时,大深度值的部分在训练过程中对整体结果产生过大的影响。当深度值分布范围为[α,β],且需要分割成 K 个区间时,SID 采样可表述为

$$t_l=\exp\left[\ln\alpha+\frac{\ln(\beta/\alpha)\cdot l}{K}\right],\tag{4}$$

式中:l 为第 l 段区间;t_l 为该段的下限深度值。

如图 7 所示,作者所提出的网络结构大致可以分为两部分:一个是用于获得特征图的深度卷积神经网络,另一个是由全图编码器、空间金字塔的空洞池化(ASPP)[49]和跨通道信息学习机三个平行通道组成的场景理解模块,其结果经过 softmax 转化为概率形式,而整个网络通过有序回归损失进行优化。

在第一部分卷积神经网络的最后几层,作者采用空洞卷积来代替之前被广泛使用的普通卷积,这样做可以在不增加参数或降低图像分辨率的情况下,增大感受野的面积,避免多次降采样导致图像分辨

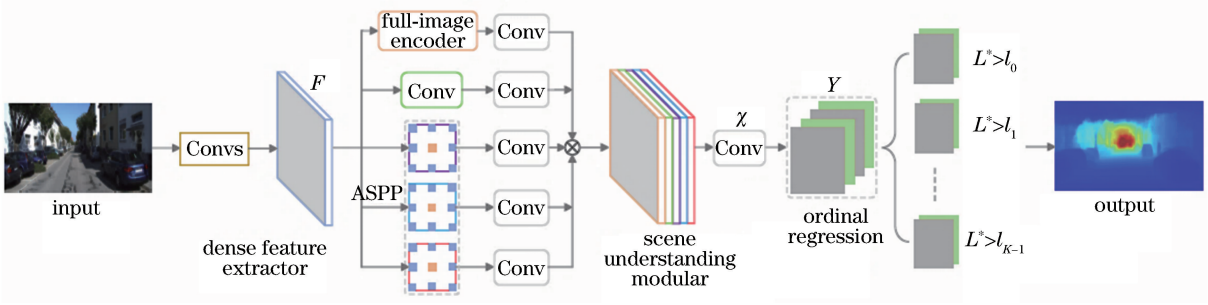

图 7 Fu 等[47]提出的网络结构示意图

Fig. 7 Schematic of network architecture proposed by Fu *et al*. [47]

率过低的问题，同时无需为融合多尺度信息及上采样设计复杂的网络结构。与之类似地，作者在场景理解模块中采用 ASPP 来在更大的感知野上提取特征。

在全图编码的场景理解上，为了避免完全使用全连接层而造成的训练难度增大以及计算效率较低的问题，设计了新的编码器结构，如图 8 所示，输入图片 F 首先通过一个核为 k 的池化层，得到一个池化之后的特征图 F'，对这个特征图通过全连接层得到一个 Q 维向量的元素，这 Q 个元素可以看作是 $1\times1\times Q$ 的特征图，通过 1×1 的卷积就可得到混合 Q 个通道的特征，然后将这个特征复制得到新的 $w\times h\times Q$ 的特征。此外，作者额外添加了两个 1×1 的卷积层，分别用于降低特征维度，将特征转变为多通道的稠密有序标签。

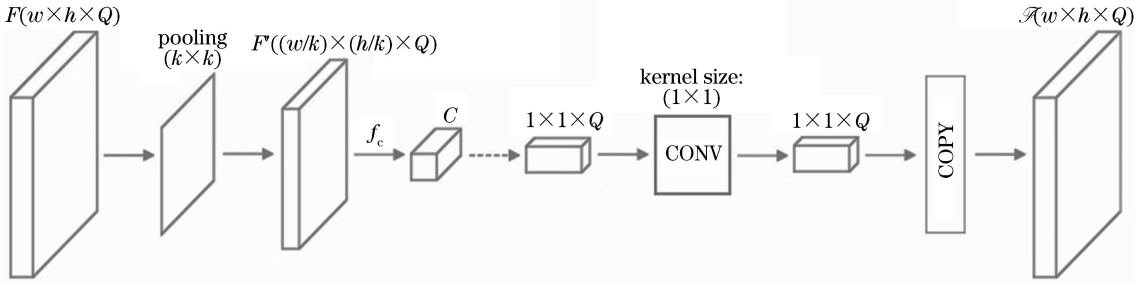

图 8 全图编码器结构图

Fig. 8 Structural schematic of full-image encoders

经典的多分类问题通常将网络输出用一个 softmax 层归一化为概率，使用逻辑回归损失对其进行训练。但是对于深度值这一有序类别，损失函数也需要做出相应调整。作者将整个特征图 χ 的损失 L 定义为其每个像素点的有序损失 Ψ 的平均值，即

$$L(\chi,\Theta)=-\frac{1}{N}\sum_{w=0}^{W-1}\sum_{h=0}^{H-1}\Psi(w,h,\chi,\Theta),\quad(5)$$

式中：Θ 为权重参数向量；W 为图像宽度；H 为图像高度。每个像素点 (w,h) 的有序损失可表示为

$$\Psi(w,h,\chi,\Theta)=\sum_{k=0}^{l_{(w,h)}-1}\ln(P_{(w,h)}^{(k)})+\sum_{k=l_{(w,h)}}^{K-1}\ln(1-P_{(w,h)}^{(k)}),\quad(6)$$

式中：$l_{(w,h)}$ 为像素点 (w,h) 在离散化深度序列中所属的位置；$P_{(w,h)}^{(k)}$ 代表了预测的该像素点的深度值大于离散化深度序列中第 k 个阈值的概率。为了使得到的概率值归一化，通过 softmax 函数来计算这一概率，即

$$P_{(w,h)}^{(k)}=\frac{\exp[y_{(w,h,2k+1)}]}{\exp[y_{(w,h,2k)}]+\exp[y_{(w,h,2k+1)}]},\quad(7)$$

式中：$y_{(w,h,2k+1)}$ 为像素点 (w,h) 在有序回归中的预测值。因为深度的有序性，某点的深度值要么大于等于第 k 个阈值，要么小于第 k 个阈值。其实，这一损失函数类似于二分类问题的交叉熵损失函数。

值得注意的是，在实验过程中，作者将每幅图像按照训练阶段的分割方式分割成了一些互有重叠的子窗口，针对重叠区域的最终结果取不同取值的平均值，这一策略丰富了离散化深度的取值，使得结果更为平滑。截至本文完成之时，该方法在 KITTI 数据集下公开的深度估计在线评分排行中排名第一。

4 无监督型深度学习方法

监督学习的方法在训练神经网络的过程中，需要向神经网络输入大量有真实深度值标注的图片，以真实值作为训练的约束对神经网络进行反向传播，优化参数，从而训练出可在相似场景下完成单目图像深度估计任务的神经网络[34]。然而目前高质量的有深度标注的公开数据还比较有限，在现实情况下，要获取场景所对应的深度值相比获取图片更难。为此研究人员在近些年开始探索可在无参考深度值时获得深度估计的神经网络。目前研究人员对于这一问题主要有两种解决方案，一种是借助立体视觉对神经网络进行约束，另一种则是使用视频，通过考察相机运动来计算深度。

4.1 立体视觉型无监督学习方法

在计算机视觉还未发展之时，人们就已经用摄影测量的原理解释如何通过双目视觉来恢复场景深度。对双目图像深度估计的研究也一直未曾中断，因此研究人员首先想到将双目立体像对应用于单目图像深度估计的无监督学习中，Garg 等[30]在 2016 年的工作开创了这方面的先河，其总体流程如图 9 所示，分为三个部分：第一部分是基于传统的卷积神经网络（作者使用的是有跳跃连接结构的全卷积网络）构成编码器，全卷积网络的输入为左视图 $\mathscr{I}_1$，通过网络后得到其对应的预测深度图；第二部分将编码输出的深度图与右视图 $\mathscr{I}_2$ 结合，基于右视图沿扫描线移动的相应距离来获得反向变形的图像 $\mathscr{I}_w$；第三部分将原左视图 $\mathscr{I}_1$ 与变形图 $\mathscr{I}_w$ 的光流误差作为目标函数，匹配重建的图像。

为了计算后向传播的梯度，Garg 等在变形重建左图时使用 Taylor 展开并进行线性化处理，这使得该方法的结果并不完全可微，导致训练的结果有可能陷入次优解。为此 Godard 等[31]使用可微的重建损失函数，提出了一种引入左右深度一致性的方法，如图 10 所示。其网络结构与上述方法的最大不同在于：向神经网络输入左图 $\mathscr{I}^l$，同时为左右两个视图输出视差 d^l 和 d^r，然后分别将右左视差应用于左右视图，得到重建的左右视图。这里作者在编码时使用的是改进的全卷积网络 DispNet[50]，在解码器的最外层，作者估计了当前特征维度下对应的视差值，并且将其经过上采样后传递给了下层，相当于做了一个多尺度的长程预测。最后的输出是 4 个尺度上的视差预测，每一个尺度的空间分辨率是前一个尺度的 2 倍，而在每个尺度下，会有左至右和右至左两个输出结果，充分体现了左右深度一致原则。

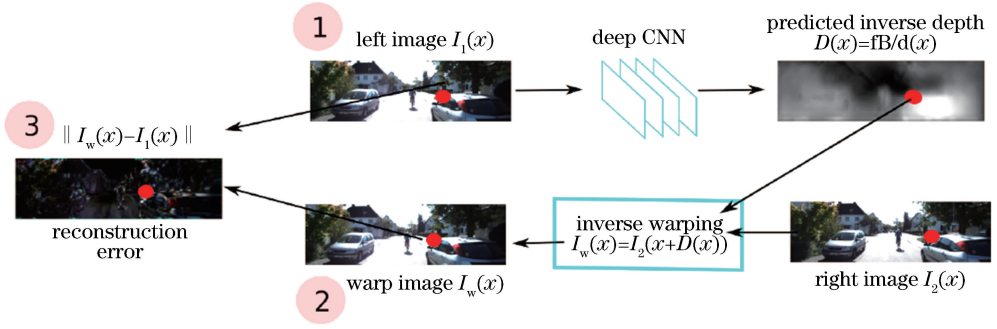

图 9 Garg 等[30]所提算法的总体流程图

Fig. 9 Overall flow chart of algorithm proposed by Garg *et al*.[30]

为此，该方法的损失函数相应也包含了左右一致性的部分，总的损失 C 是 4 个尺度下的各损失 C_S（$S\in\{1,2,3,4\}$）之和，具体的损失由三部分组成，分别是重建外观匹配损失 C_{ap}、视差平滑损失 C_{ds} 和左右一致性损失 C_{lr}，每一项损失既包含左视图部分，也包含交换左右视图在另一方向采样的右视图部分，即

$$C_S=\alpha_{ap}(C_{ap}^l+C_{ap}^r)+\alpha_{ds}(C_{ds}^l+C_{ds}^r)+\alpha_{lr}(C_{lr}^l+C_{lr}^r),\tag{8}$$

式中：α_{ap}，α_{ds}，α_{lr} 分别为调配上述三项损失的参数；上标 l 和 r 分别代表左视图和右视图。

由于作者使用的是双线性采样，梯度的范围始终来自于周围的 4 个坐标点，局部完全可微，这意味着不需要对损失函数进行简化或近似化处理，因此作者使用了 $\boldsymbol{L}_1$ 范式结合单尺度的结构相似性（SSIM）[51]来研究重建外观匹配损失 C_{ap}；对于视差平滑损失 C_{ds}，使用了 $\boldsymbol{L}_1$ 范式来进行视差的局部平滑，而针对出现在边缘区域的深度不连续情况，采用文献[52]的方式利用图像梯度 $\partial\mathscr{I}$ 来感知边缘，保证得到深度图的光滑性与图像梯度一致；由于在应用

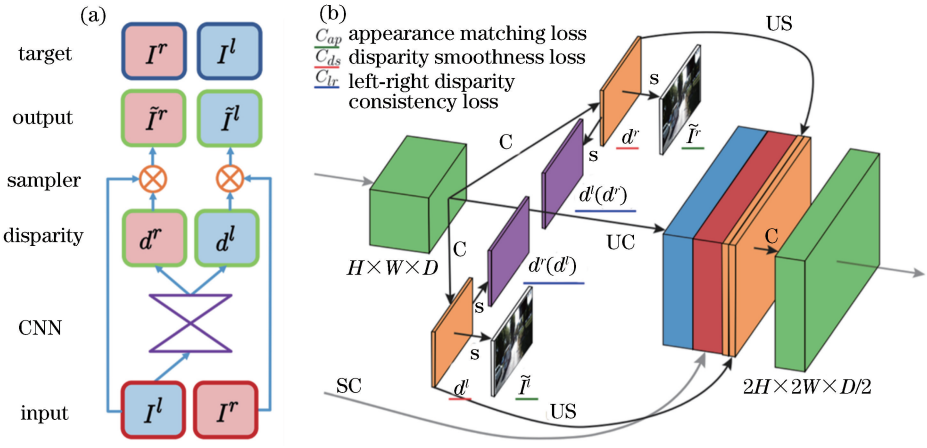

图 10 Godard 等[31]提出的网络结构示意图。(a)左右一致性采样策略;(b)损失函数

Fig. 10 Schematics of network architecture proposed by Godard *et al*[31]. (a) Sampling strategy with left-right consistency; (b) loss function

时只有左视图输入,为了确保左右一致性,需引入左右一致性损失 C_{lr}。与前两项类似,同样使用 $\boldsymbol{L}_1$ 范式,这一损失用于监督左视图的视差图尽可能与对应的投射的右视图视差相等。

Godard 等[31]在对比实验中甚至取得了优于之前使用真实深度值进行监督学习的结果,同时模型在数据方面还展示出较强的泛化能力。Garg 等[30]的工作大大激励了之后研究人员在无监督学习方面的研究。Kuznietsov 等[53]尝试将传感器得到的稀疏的深度作为参考标准,与上述的完全无监督的方法联合起来,以半监督的方式共同实现单目图像的深度估计。作者采取的方式是对整个图像都进行无监督学习,在部分有参考深度值的地方使用监督学习,其基本框架与损失如图 11 所示。在训练过程中,不仅需要输入图像的立体像对,还需要将激光获得的稀疏深度图严格按照几何关系投射至左右视图中,得到两幅参考的深度图。

该方法的重点在于其损失函数的构成,由于半监督学习包括有真实深度的监督学习与无监督学习两部分,故总的损失函数 L_θ 中也包括监督损失 L_θ^{S}、无监督损失 L_θ^{U} 两部分,此外还包含深度域上预测深度值梯度的正则项 L_θ^{R},即

$$L_\theta(\boldsymbol{I}_{\mathrm{l}},\boldsymbol{I}_{\mathrm{r}},Z_{\mathrm{l}},Z_{\mathrm{r}})=\lambda L_\theta^{\mathrm{S}}(\boldsymbol{I}_{\mathrm{l}},\boldsymbol{I}_{\mathrm{r}},Z_{\mathrm{l}},Z_{\mathrm{r}})+\gamma L_\theta^{\mathrm{U}}(\boldsymbol{I}_{\mathrm{l}},\boldsymbol{I}_{\mathrm{r}})+L_\theta^{\mathrm{R}}(\boldsymbol{I}_{\mathrm{l}},\boldsymbol{I}_{\mathrm{r}}),\tag{9}$$

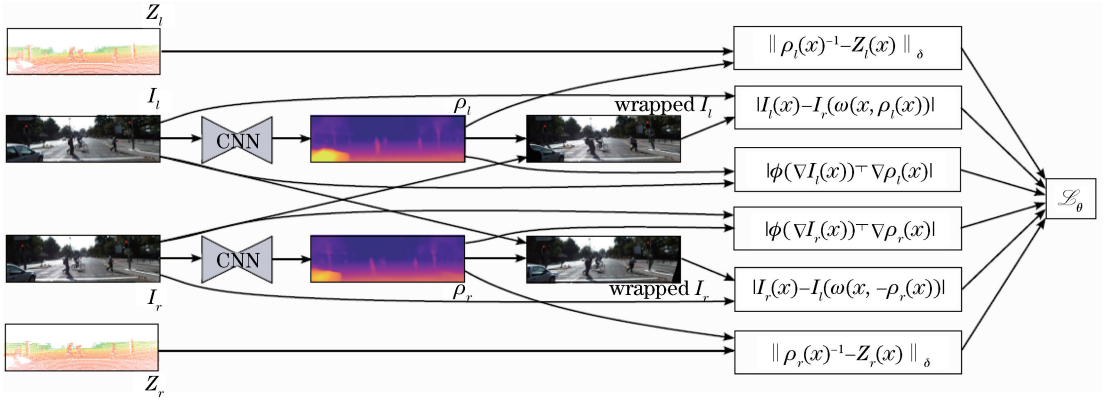

图 11 Kuznietsov 等[53]提出的半监督网络框架与损失函数

Fig. 11 Schematic of semi-supervised network architecture and loss function proposed by Kuznietsov *et al*. [53]

式中:$\mathscr{I}_{\mathrm{l}}$ 为左视图;$\mathscr{I}_{\mathrm{r}}$ 为右视图;Z_{l} 为逆合成恢复的左视图;Z_{r} 为逆合成恢复的右视图;λ 与 γ 为控制监督与无监督损失比例的参数;θ 为参数集合。

有监督的损失部分,作者参考 Laina 等[27]在其监督学习的网络中使用了 berHu 函数 $|d|_\delta$,其中 d 是深度值,下标 δ 是确定损失的阈值以更加有效地压制高深度值的残差项;无监督的损失部分也包括了左至右与右至左两部分。在计算投影点误差时,作者先对左右视图和生成的左右视图进行了一次高斯平滑以消除噪声;梯度正则项上,作者采用了类似

Godard 等[31]视差平滑损失中的方式，使其既能够平滑深度变化，同时保持在边缘处深度不连续的特性。作者对图像中的每个像素从水平与竖直方向取梯度，考虑到边缘梯度的不连续，取其自然指数倒数的函数作为权重，当梯度太大时权重就小。具体的损失函数计算表述在此就不再展开。

上述三项损失中的每一项都对称地包括了左图与右图的加和，因此，作者认为没有必要再像文献[31]那样单独显式地使用一项损失来表征左右一致性的损失。此外，值得注意的是，作者以预测逆深度为目标。这样做比起直接求解深度，能够减轻深度值由近及远的长尾分布效应的影响，且不额外增添其他的参数或调整网络结构。

4.2 视频型无监督学习方法

上述的通过立体像对进行无监督学习的方法，在训练过程中主要依赖立体像对间的关系进行约束。可否仅依靠单个相机就能实现约束引起了学者们的关注。早在 21 世纪初，就已经有研究人员研究利用连续的图像序列来还原相机的运动状态与位姿变化，即提出视觉里程计的概念[54]。视觉里程计并不关心相机到图像上每一点的绝对距离，只考察位置变化带来的相对位移。在移动过程中，考虑到通过相机位姿的变化来感知场景的绝对深度，展开了从单目视频中恢复场景深度的工作。

利用单目视频来进行深度估计的一大难点在于每帧图像成像时相机的位姿是不确定的。Zhou 等[32]用两个网络分别独自无监督地估计单帧的深度和视频序列中相机位姿的变化，其大致框架如图 12 所示。深度估计网络的输入只有 t 时刻的目标帧图像 $\mathcal{I}_t$，输出为其深度图 $\hat{D}_t(p)$，p 是图像序列中的同名点；位姿估计网络不仅需要目标帧图像，还要输入其前序图像 $\mathcal{I}_{t-1}$ 和后序图像 $\mathcal{I}_{t+1}$，得到从 t 时刻到 $t-1$ 时刻以及 t 时刻到 $t+1$ 时刻的相机相对位姿 $\hat{t}_{t\to t-1}$ 和 $\hat{t}_{t\to t+1}$。类似用立体像对的无监督学习方法，利用两个网络的输出合成出 $t-1$ 时刻与 $t+1$ 时刻的重建图，以其光度差作为损失监督网络的训练。其核心思路与上述使用立体像对进行无监督学习的思路有着相似之处。

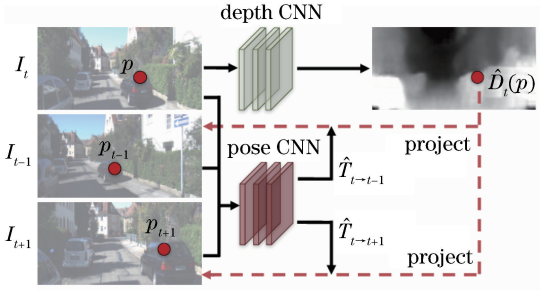

图 12 基于视角合成的视频深度估计网络框架示意图[32]

Fig. 12 Schematic of network frame of video depth estimation based on view synthesis[32]

由于每幅图像的拍摄出现了时间差，因此除了要求表面是朗伯的，以保证光度误差有效外，还需要场景是静态的，场景中没有移动物体，在多个图像间不存在遮蔽与解遮蔽现象。然而现实场景是复杂的，有许多动态物体。为此，作者在上述两个网络结构之外，加入了一个解释性预测网络，其为每个源图片与目标图片对生成一个掩模，用这个掩模来过滤无用信息。于是在表征光度差的视图合成损失函数中增添了一个权重项，且为了避免掩模 E_s 最终被优化为 0，在最终的损失函数中添加了形如交叉熵的正则项 L_{reg}，以及用类似文献[31]中的方法来保持梯度平滑的损失项 L_{smooth}，作者采用二阶梯度的 $\boldsymbol{L}_1$ 范式，故最终的损失函数为

$$L_{\text{final}}=\sum_m L_{\text{vs}}^m+\lambda_s L_{\text{vs}}^m+\lambda_e\sum_s L_{\text{reg}}(\hat{E}_s^m)\,, \tag{10}$$

式中：L_{final} 为最终总的损失；L_{vs}^m 为视图合成损失；λ_s 与 λ_e 为调节参数；$\hat{E}_s^m$ 为掩模；m 为不同的图像尺度。

最终在实验过程中，其相机位姿估计精度与 ORB-SLAM2[55]的效果相当，但其深度估计的精度不如文献[31]中方法以及一些监督学习方法。但其工作为后续的研究人员提供了许多有益的参考，Yang 等[56]在利用单目视频估计深度时加入了深度-法向量一致性的约束，Zhou 等[57]吸取了视觉 SLAM 中图优化的思路，将其应用于相机位姿网络的训练中，取得了较好的结果。

如上文中所提到的，利用单目视频来进行深度估计的另一大难点是场景中可能存在物体运动，文献[32]使用的解释性预测掩模并不能很好地解决这一问题。Vijayanarasimhan 等[58]尝试通过一个物体掩模将运动物体区分开，并在位姿估计网络中，同时处理相机的运动与物体的运动。Yin 等[59]设计了一种级联的网络结构，对道路、房屋等刚性静态表面和行人、汽车等动态对象，采取分治的策略，分别自适应地学习刚体流和目标运动，这种策略将整个运动场逐步细化，使得学习更加容易。对于静态场景而言，从目标帧 $\boldsymbol{I}_t$ 到源帧 $\boldsymbol{I}_s$ 的相对二维刚性流可通过相机的运动 $t_{t\to s}$ 和深度图 D_t 来表述，即

$$f_{t\to s}^{\text{rig}}(p_t)=KT_{t\to s}D_t(p_t)K^{-1}p_t-p_t\,, \tag{11}$$

式中：K 是相机内参数；p_t 是图像 $\boldsymbol{I}_t$ 中像素的齐次坐标。而对于无约束的自由物体移动，作者使用二维位移向量的方式和经典的光流概念进行建模，其网络框架如图 13 所示，第一部分负责刚性结构重建，第二部分则是动态目标定位部分。

在第一部分中，与之前的方法类似，由深度网络与位姿网络两个平行部分组成。其中深度估计网络以图像序列作为独立图像的批输入，而位姿估计网络为了能更好地利用不同时序下的特征关联，将整个图像序列在不同维度的通道上串联输入，最后输出的是从源图像得到的重建目标图像$\tilde{\boldsymbol{I}}\mathrm{rig}_s$，这一部分网络则由目标图像与重建图像的光度差来监督，其损失函数采用了对于异常值较稳定的可微的结构相似性（SSIM）度量，即

$$L_{\mathrm{rw}}=\alpha\frac{1-S_{\mathrm{SSIM}}(\boldsymbol{I}_t,\tilde{\boldsymbol{I}}_s^{\mathrm{rig}})}{2}+(1-\alpha)\parallel\boldsymbol{I}_t-\tilde{\boldsymbol{I}}_s^{\mathrm{rig}}\parallel_1,\tag{12}$$

式中：L_{rw} 为刚体流重建损失；α 为调节参数，文中取 0.85；S_{SSIM} 为结构相似性度量指数。

在此之外，考虑与前述方法类似的边缘感知的梯度平滑损失 L_{ds}。在动态目标定位部分，与刚性流不同，仅使用残差 $f_{t\to s}$ 的概念来代表物体与静态平面的相对运动，使用 ResFlowNet 学习，作者参照[60]的方式将 ResFlowNet 网络级联在第一部分之后。最终整个预测流的结果为两段的加和，而总的损失也与上一部分类似，只是用 $f_{t\to s}$ 来替换 $f_{t\to s}$。

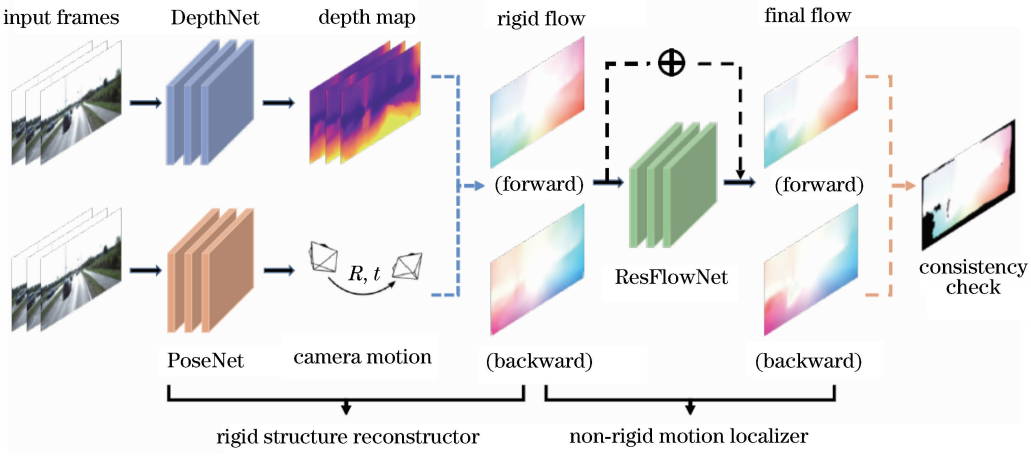

图 13 Yin 等[59]提出的 GeoNet 网络结构示意图

Fig. 13 Schematic of GeoNet network architecture proposed by Yin *et al.* [59]

作者用以上的架构解决了动态物体的问题，然而非朗伯面与视角遮蔽的问题仍然在实际中制约着方法的使用，为此作者在不改变网络结构的前提下在网络框架中采用了几何前向-后向一致性检验，这与 Godard 等[31]所提出的左右一致性检验有着相似之处，一致性检验通过几何一致性损失来执行，表达式为

$$L_{\mathrm{gc}}=\sum_{p_t}[\delta(p_t)]\cdot\parallel\Delta f_{t\to s}^{\mathrm{full}}(p_t)\parallel_1,\tag{13}$$

式中：δ 为判断这一检验是否有必要的函数；$\Delta f_{t\to s}^{\mathrm{full}}(p_t)$ 为在像素点 p_t 处进行前向-后向一致性检验得到的整体光流变化的微分，但这一检验只应对未遮蔽区域进行，因此当 $\parallel\Delta f_{t\to s}^{\mathrm{full}}(p_t)\parallel_2<\max\{\alpha,\beta\parallel\Delta f_{t\to s}^{\mathrm{full}}(p_t)\parallel_2\}$ 时，(13)式才有意义，否则为 0，其中的两个参数 α 和 β，作者给出了他们实验中所用的值以供参考。对于那些在前向-后向过程中即不满足光度相似，又未能通过几何前后一致性检验的点，只能作为异常点处理，改进后的总损失可表示为

$$L=\sum_m\sum_{\langle t,s\rangle}(L_{\mathrm{rw}}+\lambda_{\mathrm{ds}}L_{\mathrm{ds}}+L_{\mathrm{fw}}+\lambda_{\mathrm{fs}}L_{\mathrm{fs}}+\lambda_{\mathrm{gc}}L_{\mathrm{gc}}),\tag{14}$$

式中：L 为总的损失；m 为图像尺度；λ_{ds} 与 λ_{fs} 为调节参数；L_{fw} 为光流变换损失；L_{fs} 为平滑损失。

最终，改进方法在实验中取得了较高的精度，但在深度估计上，虽然要略逊于上述 Godard 的方法[31]，但作者认为这是由于精准校对过的左右像对和单目视频序列这两个训练数据上的巨大差异所导致。

Casser 等[33]于美国人工智能协会 2018 年年会（AAAI 2018）提出了一种新的方法，其网络结构以每一个目标在三维空间中的运动单独建模。这样一来，重建的目标图像不再是一个单一的仅考虑相机位姿变化的图像，而是相机自运动和所有运动物体投射的集合。这一方法在实验阶段展现出较高的精度，然而值得注意的是，在训练阶段，作者使用了预先计算好的实体分割掩模与静态背景掩模，而其对

每一个物体都使用一个卷积网络来估算其运动，使其网络复杂程度远超其他方法。

5 单目深度估计方法评价与对比

为了直观、定量地评价近年来提出的基于深度学习解决单目图像深度估计问题方法的精度，表 1 与表 2 分别展示了部分算法在目前最常用的 NYU Depth v2 和 KITTI 这两个代表室内与室外场景的公开数据集上的表现。列举了目前主流的 5 种定量评价指标，即相对误差（Abs rel，e_{Abs}）、均方根误差（RMSE，e_{RMS}）、常用对数误差（lg，e_{lg}）、均方根对数误差（$\mathrm{RMSE}_{\mathrm{ln}}$，$e_{\mathrm{RMS\_ln}}$）、阈值误差 δ，表达式分别为

$$e_{\mathrm{Abs}}=\frac{1}{N}\sum_{i=1}^{N}\frac{|D_i-D_i^*|}{D_i^*},\tag{15}$$

$$e_{\mathrm{RMS}}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}|D_i-D_i^*|^2},\tag{16}$$

$$e_{\mathrm{lg}}=\frac{1}{N}\sum_{i=1}^{N}|\lg D_i-\lg D_i^*|,\tag{17}$$

$$e_{\mathrm{RMS\_ln}}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}|\ln D_i-\ln D_i^*|^2},\tag{18}$$

$$\delta=\max\left\{\frac{D_i^*}{D_i},\frac{D_i}{D_i^*}\right\}<t_{\mathrm{threshold}},\tag{19}$$

式中：D_i 为 i 位置的预测深度值；D_i^* 为 i 位置的真实参考深度值；$t_{\mathrm{threshold}}$ 为准确度的阈值。

表 1 和表 2 中，Sup. 与 Unsup. 分别表示监督学习与无监督学习。每项指标中表现最好的用粗体标识，次好的用斜体标识。从表 1 和表 2 可以看到，从 2014 年 Eigen 团队首次使用深度学习对单目图像恢复场景深度开始，在短短不到 5 年的时间内，获得的深度精度相比最初已提高 1 倍左右。借助于深度学习的方法，在原本病态的单目图像的深度估计问题上，可以得到精度相当高的深度估计结果，由此也证明深度学习方法在这一问题上的适用性。

表 3 列举了近年来单目图像深度估计中具有代表性的若干深度学习方法，展示了其使用的数据类型（包括普通单目 RGB 图像 RGB、双目 RGB 图像 binocular RGB、深度数据 depth、稀疏深度数据 sparse depth 以及单目视频 video）、包含的损失类型及亮点。

表 1 部分算法在 NYU Depth V2 数据集上的定量评价

Table 1 Quantitative evaluation of selected algorithms on NYU Depth V2 dataset

Method	Year	Type	e_{Abs}	e_{RMS}	e_{lg}	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Method in Ref. [23]	2014	Sup.	0.215	0.907	—	0.611	0.887	0.971
Method in Ref. [26]	2016	Sup.	0.213	0.759	0.087	0.650	0.906	0.976
Method in Ref. [27]	2016	Sup.	0.127	0.573	0.055	0.811	0.953	0.988
Method in Ref. [61]	2017	Sup.	0.121	0.586	0.052	0.811	0.954	0.987
Method in Ref. [47]	2018	Sup.	*0.115*	0.509	*0.051*	*0.828*	*0.965*	*0.992*
Method in Ref. [42]	2018	Sup.	**0.098**	**0.329**	**0.040**	**0.917**	**0.983**	**0.996**
Method in Ref. [45]	2018	Sup.	0.139	*0.505*	0.058	0.820	0.960	0.989
Method in Ref. [10]	2019	Sup.	0.128	0.523	0.059	0.813	0.964	*0.992*

Note: bold means best; italic means second best.

表 2 部分算法在 KITTI 数据集上的定量评价

Table 2 Quantitative evaluation of selected algorithms on KITTI dataset

Method	Year	Type	e_{Abs}	e_{RMS}	e_{lg}	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Method in Ref. [23]	2014	Sup.	0.190	7.156	0.270	0.692	0.899	0.967
Method in Ref. [26]	2016	Sup.	0.217	7.046	—	0.656	0.881	0.958
Method in Ref. [53]	2017	Semi. (stereo)	0.113	4.621	0.189	0.862	0.960	0.986
Method in Ref. [31]	2017	Unsup. (stereo)	0.114	4.935	0.206	0.861	0.949	0.976
Method in Ref. [47]	2018	Sup.	**0.072**	**2.727**	**0.120**	**0.932**	**0.984**	**0.994**
Method in Ref. [42]	2018	Sup.	—	5.110	0.215	0.843	0.950	0.981
Method in Ref. [45]	2018	Sup.	0.113	4.687	—	0.856	0.962	0.988
Method in Ref. [33]	2018	Unsup. (video)	0.109	4.750	0.187	0.874	0.958	0.982
Method in Ref. [62]	2018	Unsup. (stereo)	*0.095*	*4.316*	*0.177*	*0.892*	*0.966*	*0.984*
Method in Ref. [59]	2018	Unsup. (video)	0.153	5.737	0.232	0.802	0.934	0.972
Method in Ref. [57]	2019	Unsup. (video)	0.139	5.160	0.215	0.833	0.939	0.975

Note: bold means best; italic means second best.

表 3 部分典型算法的概要总结

Table 3 Summary of selected representative algorithms

Method	Year	Type	Data type	Loss	Main contributions
Method in Ref. [23]	2014	Sup.	RGB+depth	Inference error (original)	First to use deep learning on monocular depth estimation (MDE)
Method in Ref. [27]	2016	Sup.	RGB+depth	Inference error (berHu loss)	Introduction of residual learning to MDE with optimized up-convolutions
Method in Ref. [61]	2017	Sup.	RGB+depth	Inference error (square loss)	Achievement of end-to-end MDE with CNN layers fused within CRF
Method in Ref. [53]	2017	Semi.	Binocular RGB + sparse depth	Berhu loss (supervised loss), image alignment error (unsupervised loss), regularization loss	Introduction of a semi-supervised deep learning approach of MDE
Method in Ref. [47]	2018	Sup.	RGB+depth	Ordinal regression loss	Ordinal regression method for MDE with dilated convolutions
Method in Ref. [59]	2018	Unsup.	Video	Wrapping loss, depth smoothness loss, geometric consistency loss	Cascaded architecture to resolve rigid flow and object motion separately in depth estimation from monocular video
Method in Ref. [33]	2018	Unsup	Video	Reconstruction loss, wrapping loss, depth smoothness loss object size constraints	MDE in highly dynamic scenes with explicit modeling of 3D motions of moving objects and camera itself

6 结束语

通过总结近年来的研究历程，可以发现以下研究趋势：

1）监督学习到无监督学习的转变：使用深度学习进行单目图像深度估计的早期工作都需要有精确的参考深度来进行监督训练神经网络。常用的包含深度标注的公开数据集有 NYU Depth V2、KITTI、Make3D 等，然而由于深度获取的难度大，这三个数据集的数据也相对有限，虽然训练样本并非越多越好，但随着网络结构日益复杂和网络深度日益加深，对数据的需求也更为迫切。于是无需参考深度的无监督学习获得了青睐，其中相比立体像对，更容易获取的视频颇受欢迎，因此在 2018 年，使用单目视频的无监督深度估计的有关研究明显增多，成为无监督学习的主流。而得益于生成对抗网络的提出与发展，最近有研究人员开始将生成对抗网络（GAN）应用于这一问题，有望进一步缓解训练样本缺乏的问题[63-64]。

2）多任务的协同处理：场景的深度估计可以认为是场景感知中的重要内容之一，除此之外，还有语义分割、目标检测等工作也是这一领域的研究热点。对于视频而言，还有运动跟踪等。这些任务之间有着密切的关联，而这些联系不能简单地用几组公式来表述。因此，研究人员希望通过设计更为复杂的网络框架，同时处理多个有关联的任务，以协同优化每个任务的结果。其中最具代表性的是在监督学习中同时实现深度估计与语义分割，以及在使用视频的无监督学习中同时实现深度估计与视觉里程计。

3）更复杂的约束与损失：由于单目图像深度估计问题本身具有病态性，为了获得更好的结果，更多的约束条件与更复杂的损失函数是必然的。在这些约束中有一个趋势：越来越多的约束已经不单单是几何上的约束。结合上述多任务协同处理的发展趋势，可以预见在未来的一段时间内，将会看到更多不一样的约束条件与更复杂的损失函数。

本文主要阐述了深度学习在单目图像恢复场景深度方面的应用，介绍其发展历程，总结最新研究进展，根据是否有真实的深度值参与训练进行分类，分别简要介绍监督与无监督两类方法中的一些代表性算法，总结归纳了近年来这一领域研究的发展趋势。总的来说，单目图像的深度估计问题历史悠久，而深度学习方法的应用才刚刚 5 年就已大放异彩。虽然获得了不少进展，但目前依然还有以下几个问题有待解决，这也将会是下一阶段研究人员们要突破的重点：

1）动态物体及遮挡问题的解决。动态物体及遮挡的问题一直是影响通过单目视频来进行深度估计的一大重要问题，之前的许多算法中直接通过掩模将这一问题忽略，文献[33,65]中的算法开始正视

这一问题，取得了一定进展，然而如何进一步优化动态物体表面深度和处理遮蔽问题依然有许多工作要做。

2）算法泛化能力的提升。限制使用双目图像来估计深度的一大原因在于目前尚没有算法能很好地处理弱纹理场景下的匹配问题，而受限于目前公开数据集的数据，尚未有工作对其算法在复杂光照、弱纹理等环境条件下的模型泛化能力进行考察。

3）网络框架的整合与优化。目前的无监督学习框架中普遍有多个神经网络，以获得深度估计和位姿估计结果。而许多监督学习框架中，会增加完成语义分割、法向量检测等任务的网络，目前这些网络并没有很好地关联起来，能否对网络框架进行整合优化以实现更好的结果将成为新的探索方向。

4）高分辨率的深度图输出。为了提升计算效率与结果精度，当前各方法所得到的输出结果的分辨率普遍是输入的一半。而受限于公开数据集的数据质量，输入图像的分辨率普遍在百万像素以下，这样的结果显然难以满足增强现实（AR）等任务的需要，已经有研究人员开始研究通过神经网络提升深度图分辨率[66]。如何获得高分辨率输出将是研究人员面临的一大问题。

参 考 文 献

[1] Zeng A, Song S R, NieBner M, *et al*. 3DMatch: learning local geometric descriptors from RGB-D reconstructions [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 199-208.

[2] Wang Z, Liu H, Wang X D, *et al*. Segment and label indoor scene based on RGB-D for the visually impaired[M] // Gurrin C, Hopfgartner F, Hurst W, *et al*. MultiMedia modeling. Lecture notes in computer science. Cham: Springer, 2014, 8325: 449-460.

[3] Mancini M, Costante G, Valigi P, *et al*. Fast robust monocular depth estimation for Obstacle Detection with fully convolutional networks[C] // 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 9-14, 2016, Daejeon, Korea. New York: IEEE, 2016: 4296-4303.

[4] Chen Z H, Hong Y, Wang J K, *et al*. Monocular visual odometry based on recurrent convolutional neural networks[J]. Robot, 2019, 41(2): 147-155.
陈宗海, 洪洋, 王纪凯, 等. 基于循环卷积神经网络的单目视觉里程计[J]. 机器人, 2019, 41(2): 147-155.

[5] Li X Z, Yang A L, Qin B L, *et al*. Monocular camera three dimensional reconstruction based on optical flow feedback[J]. Acta Optica Sinica, 2015, 35(5): 0515001.
李秀智, 杨爱林, 秦宝岭, 等. 基于光流反馈的单目视觉三维重建[J]. 光学学报, 2015, 35(5): 0515001.

[6] Zhan K F, Chen W J, Li W S, *et al*. Line laser 3D scene reconstruction system and error analysis[J]. Chinese Journal of Lasers, 2018, 45(12): 1204004.
詹坤烽, 陈文建, 李武森, 等. 线激光三维场景重建系统及误差分析[J]. 中国激光, 2018, 45(12): 1204004.

[7] Bi T T, Liu Y, Weng D D, *et al*. Survey on supervised learning based depth estimation from a single image[J]. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(8): 3-13.
毕天腾, 刘越, 翁冬冬, 等. 基于监督学习的单幅图像深度估计综述[J]. 计算机辅助设计与图形学学报, 2018, 30(8): 3-13.

[8] Žbontar J, LeCun Y. Stereo matching by training a convolutional neural network to compare image patches [J]. The Journal of Machine Learning Research, 2016, 17: 2287-2318.

[9] Hirschmuller H. Accurate and efficient stereo processing by semi-global matching and mutual information [C] // 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR′05), June 20-25, 2005, San Diego, CA, USA. New York: IEEE, 2005, 2: 807-814.

[10] Zhao S Y, Zhang L, Shen Y, *et al*. Super-resolution for monocular depth estimation with multi-scale sub-pixel convolutions and a smoothness constraint[J]. IEEE Access, 2019, 7: 16323-16335.

[11] He L, Dong Q L, Hu Z Y. The inherent ambiguity in scene depth learning from single images [J]. Scientia Sinica (Informationis), 2016, 46(7): 811-818.
何雷, 董秋雷, 胡占义. 从单幅图像学习场景深度信息固有的歧义性[J]. 中国科学:信息科学, 2016, 46(7): 811-818.

[12] Tsai Y M, Chang Y L, Chen L G. Block-based vanishing line and vanishing point detection for 3D scene reconstruction [C] // 2006 International Symposium on Intelligent Signal Processing and

Communications, December 12-15, 2006, Tottori, Japan. New York: IEEE, 2006: 586-589.

[13] Tang C, Hou C P, Song Z J. Depth recovery and refinement from a single image using defocus cues[J]. Journal of Modern Optics, 2015, 62(6): 441-448.

[14] Prados E, Faugeras O. Shape from shading[M]// Paragios N, Chen Y, Faugeras O. Handbook of mathematical models in computer ision. Boston, MA: Springer, 2009: 375-388.

[15] Karsch K, Liu C, Kang S B. Depth extraction from video using non-parametric sampling[M]// Fitzgibbon A, Lazebnik S, Perona P, *et al*. Computer vision-ECCV 2012. Lecture notes in computer science. Berlin, Heidelberg: Springer, 2012, 7576: 775-788.

[16] Saxena A, Sun M, Ng A Y. Make3D: learning 3D scene structure from a single still image[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(5): 824-840.

[17] Saxena A, Sun M, Ng A Y. Learning 3-D scene structure from a single still image[C]// 2007 IEEE 11th International Conference on Computer Vision, October 14-21, 2007, Rio de Janeiro, Brazil. New York: IEEE, 2007: 9848899.

[18] Liu B, Gould S, Koller D. Single image depth estimation from predicted semantic labels[C]// 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 13-18, 2010, San Francisco, CA, USA. New York: IEEE, 2010: 1253-1260.

[19] Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 580-587.

[20] Liu F, Liu P Y, Li B, *et al*. Deep learning model design of video target tracking based on TensorFlow platform[J]. Laser & Optoelectronics Progress, 2017, 54(9): 091501.
刘帆, 刘鹏远, 李兵, 等. TensorFlow 平台下的视频目标跟踪深度学习模型设计[J]. 激光与光电子学进展, 2017, 54(9): 091501.

[21] Hinton G E. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[22] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// Proceedings of the 25th International Conference on Neural Information Processing Systems, December 3-6, 2012, Lake Tahoe, Nevada, USA. Canada: NIPS, 2012.

[23] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C]// 27th International Conference on Neural Information Processing Systems, December 8-13, 2014, Montreal, Canada. Canada: NIPS, 2014.

[24] Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture[C]// 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 2650-2658.

[25] Grigorev A, Jiang F, Rho S, *et al*. Depth estimation from single monocular images using deep hybrid network[J]. Multimedia Tools and Applications, 2017, 76(18): 18585-18604.

[26] Liu F Y, Shen C H, Lin G S, *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): 2024-2039.

[27] 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), October 25-28, 2016, Stanford, CA, USA. New York: IEEE, 2016: 239-248.

[28] Cao Y, Wu Z F, Shen C H. Estimating depth from monocular images as classification using deep fully convolutional residual networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(11): 3174-3182.

[29] Xie J Y, Girshick R, Farhadi A. Deep3D: fully automatic 2D-to-3D video conversion with deep convolutional neural networks[M]// Leibe B, Matas J, Sebe N, *et al*. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9908: 842-857.

[30] Garg R, Kumar B G V, Carneiro G, *et al*. Unsupervised CNN for single view depth estimation: geometry to the rescue[M]// Leibe B, Matas J, Sebe N, *et al*. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9912: 740-756.

[31] Godard C, Aodha O M, Brostow G J. Unsupervised

monocular depth estimation with left-right consistency[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 6602-6611.

[32] Zhou T H, Brown M, Snavely N, *et al*. Unsupervised learning of depth and ego-motion from video[C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 6612-6619.

[33] Casser V, Pirk S, Mahjourian R, *et al*. Depth prediction without the sensors: leveraging structure for unsupervised learning from monocular videos[J/OL]. (2018-11-15)[2019-03-15]. https://arxiv.org/abs/1811.06152.

[34] Bao Z Q, Li A H, Cui Z G, *et al*. Research progress of deep learning in visual localization and three-dimensional structure recovery [J]. Laser & Optoelectronics Progress, 2018, 55(5): 050007.
鲍振强, 李艾华, 崔智高, 等. 深度学习在视觉定位与三维结构恢复中的研究进展[J]. 激光与光电子学进展, 2018, 55(5): 050007.

[35] Saxe A M, McClelland J L, Ganguli S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks[J/OL]. (2014-02-19)[2019-03-15]. https://arxiv.org/abs/1312.6120v1.

[36] Srivastava R K, Greff K, Schmidhuber J. Training very deep networks[J/OL]. (2015-11-23)[2019-03-15]. https://arxiv.org/abs/1507.06228.

[37] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[38] Roy A, Todorovic S. Monocular depth estimation using neural regression forest [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 5506-5514.

[39] He L, Wang G H, Hu Z Y. Learning depth from single images with deep neural network embedding focal length [J]. IEEE Transactions on Image Processing, 2018, 27(9): 4676-4689.

[40] Couprie C, Farabet C, Najman L, *et al*. Indoor semantic segmentation using depth information[J/OL]. (2013-03-14) [2019-03-15]. https://arxiv.org/abs/1301.3572.

[41] Chen L F, Yang Z, Ma J J, *et al*. Driving scene perception network: real-time joint detection, depth estimation and semantic segmentation [C] // 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), March 12-15, 2018, Lake Tahoe, NV, USA. New York: IEEE, 2018: 1283-1291.

[42] Jiao J B, Cao Y, Song Y B, *et al*. Look deeper into depth: monocular depth estimation with semantic booster and attention-driven loss [M] // Ferrari V, Hebert M, Sminchisescu C, *et al*. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11219: 55-71.

[43] Lin T Y, Goyal P, Girshick R, *et al*. Focal loss for dense object detection[C] // The IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 2980-2988.

[44] Saxena A, Chung S H, Ng A Y. Learning depth from single monocular images[C] // Proceedings of the 18th International Conference on Neural Information Processing Systems, December 5-8, 2005, Vancouver, British Columbia, Canada. Canada: NIPS, 2005.

[45] 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.

[46] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[J/OL]. (2016-04-30)[2019-03-15]. https://arxiv.org/abs/1511.07122.

[47] Fu H, Gong M M, Wang C H, *et al*. Deep ordinal regression network for monocular depth estimation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT. New York: IEEE, 2018: 2002-2011.

[48] Herbrich R, Graepel T, Obermayer K. Support vector learning for ordinal regression [C] // 9th International Conference on Artificial Neural Networks: ICANN '99, September 7-10, 1999, Edinburgh, UK. New York: IEEE, 1999: 97-102.

[49] Chen L C, Papandreou G, Kokkinos I, *et al*. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[50] Mayer N, Ilg E, Hausser P, *et al*. A large dataset to

train convolutional networks for disparity, optical flow, and scene flow estimation[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 4040-4048.

[51] Wang Z, Bovik A C, Sheikh H R, *et al*. Image quality assessment: from error visibility to structural similarity [J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[52] Heise P, Klose S, Jensen B, *et al*. PM-huber: PatchMatch with Huber regularization for stereo matching[C]// 2013 IEEE International Conference on Computer Vision, December 1-8, 2013, Sydney, Australia. New York: IEEE, 2013: 2360-2367.

[53] Kuznietsov Y, Stuckler J, Leibe B. Semi-supervised deep learning for monocular depth map prediction[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI. New York: IEEE, 2017: 2215-2223.

[54] Nister D, Naroditsky O, Bergen J. Visual odometry[C]// Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA. New York: IEEE, 2004: 1315094.

[55] Mur-Artal R, Tardós J D. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras[J]. IEEE Transactions on Robotics, 2017, 33(5): 1255-1262.

[56] Yang Z H, Wang P, Xu W, *et al*. Unsupervised learning of geometry with edge-aware depth-normal consistency [J/OL]. (2017-11-10) [2019-03-15]. https://arxiv.org/abs/1711.03665.

[57] Zhou L P, Ye J M, Abello M, *et al*. Unsupervised learning of monocular depth estimation with bundle adjustment, super-resolution and clip loss [J/OL]. (2018-12-08) [2019-03-15]. https://arxiv.org/abs/1812.03368.

[58] Vijayanarasimhan S, Ricco S, Schmid C, *et al*. SfM-Net: learning of structure and motion from video[J/OL]. (2017-04-25) [2019-03-15]. https://arxiv.org/abs/1704.07804.

[59] Yin Z C, Shi J P. GeoNet: unsupervised learning of dense depth, optical flow and camera pose[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT. New York: IEEE, 2018: 1983-1992.

[60] Ilg E, Mayer N, Saikia T, *et al*. FlowNet 2.0: evolution of optical flow estimation with deep networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI. New York: IEEE, 2017: 1647-1655.

[61] Xu D, Ricci E, Ouyang W L, *et al*. Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI. New York: IEEE, 2017: 161-169.

[62] Guo X Y, Li H S, Yi S, *et al*. Learning monocular depth by distilling cross-domain stereo networks[M]// Ferrari V, Hebert M, Sminchisescu C, *et al*. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11215: 506-523.

[63] Kumar A R S, Bhandarkar S M, Prasad M. Monocular depth prediction using generative adversarial networks [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 18-22, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 413-418.

[64] Almalioglu Y, Saputra M R U, de Gusmao P P B, *et al*. GANVO: unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks[J/OL]. (2019-03-05)[2019-03-15]. https://arxiv.org/abs/1809.05786v2.

[65] Teng Q R, Chen Y M, Huang C. Occlusion-aware unsupervised learning of monocular depth, optical flow and camera pose with geometric constraints[J]. Future Internet, 2018, 10(10): 92.

[66] Li S M, Lei G Q, Fan R. Depthmap super-resolution based on two-channel convolutional neural network[J]. Acta Optica Sinica, 2018, 38(10): 1010002.
李素梅, 雷国庆, 范如. 基于双通道卷积神经网络的深度图超分辨研究[J]. 光学学报, 2018, 38(10): 1010002.