ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 2,Mar. 2022,pp. 247－278
DOI:10. 16337/j. 1004-9037. 2022. 02. 001

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于深度学习的计算机视觉研究新进展

卢宏涛，罗沐昆

（上海交通大学计算机科学与工程系，上海 200240）

摘　要：近年来，深度学习在计算机视觉各个领域中的应用成效显著，新的深度学习方法和深度神经网络模型不断涌现，算法性能被不断刷新。本文着眼于2016年以来的一些典型网络和模型，对基于深度学习的计算机视觉研究新进展进行综述。首先总结了针对图像分类的主流深度神经网络模型，包括标准模型及轻量化模型等；然后总结了针对不同计算机视觉领域的主流方法和模型，包括目标检测、图像分割和图像超分辨率等；最后总结了深度神经网络搜索方法。

关键词：深度学习；目标检测；图像分割；超分辨率；计算机视觉

中图分类号：TP391　　**文献标志码：**A

Survey on New Progresses of Deep Learning Based Computer Vision

LU Hongtao, LUO Mukun

(Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: Deep learning has recently achieved great breakthroughs in some fields of computer vision. Various new deep learning methods and deep neural network models were proposed, and their performance was constantly updated. This paper makes a survey on the new progresses of applications of deep learning on computer vision since 2016 with emphases on some typical networks and models. We first investigate the mainstream deep neural network models for image classification including standard models and light-weight models. Then, we introduce some main methods and models for different computer vision fields including object detection, image segmentation and image super-resolution. Finally, we summarize deep neural network architecture searching methods.

Key words: deep learning; object detection; image segmentation; super-resolution; computer vision

引　　言

近20年来，随着深度学习技术的迅猛发展和图形处理器（Graphics processing unit, GPU）等硬件计算设备的广泛普及，深度学习技术几乎已经应用到计算机视觉的各个领域，如目标检测、图像分割、超分辨率重建及人脸识别等，并在图像搜索、自动驾驶、用户行为分析、文字识别、虚拟现实和激光雷达等产品中具有不可估量的商业价值和广阔的应用前景[1]。基于深度学习技术的计算机视觉同时可以对其他学科领域产生深远的影响，如在计算机图形学中的动画仿真和实时渲染技术、材料领域的显微图像分析技术、医学图像分析处理技术、实时评估师生课堂表现和考场行为的智慧教育、分析运动员比赛表

收稿日期：2022-02-10；**修订日期：**2022-03-01

现和技术统计的智能系统等。

深度学习早在1986年就被Dechter[2]引入机器学习领域，2000年Aizenberg等[3]又在机器学习领域引入了人工神经网络(Artificial neural network，ANN)[4]。深度学习方法由多层组成，用于学习多层次抽象的数据特征[5]。在人工神经网络领域中，深度学习又被称为分层学习[6]，是一种通过在不同计算阶段精确地分配分数来调节网络激活的技术[4]。深度学习常常用多种抽象结构来学习复杂的映射关系，如2009年蒙特利尔大学的Bengio教授提出的带隐藏层的ANN[7]等。深度学习技术可以被视作一种表征学习，是机器学习的一个分支。

2005年多伦多大学的Hinton教授团队试图用图模型模拟人类的大脑[8]，在文献[9]中提出了一种逐层贪婪算法来预训练深度信念网，克服了深度网络难以训练的弊端，并用自编码器降低数据维度[10]，开启了深度学习的热潮，使其被广泛应用在语音识别、计算机视觉和自然语言处理等领域。2011—2012年，深度学习技术在语音识别领域中最先取得重大突破，Dahl团队[11]和Hinton团队[12]先后将识别错误率降至20%～30%。在2012年的ImageNet大规模视觉识别挑战竞赛(ImageNet large scale visual recognition challenge，ILSVRC)中，Hinton的学生提出的AlexNet[13]以超过第二名准确率10%的巨大优势夺得冠军，深度学习正式进入了爆发期。近年来各大互联网科技公司，如Google、Microsoft、Facebook、百度、阿里巴巴和腾讯等也争相投入大规模深度学习系统的研发中。

笔者在2016年发表“深度卷积神经网络在计算机视觉中的应用研究综述”[1]，总结了2016年之前深度卷积神经网络在计算机视觉中的研究成果。本文在文献[1]的基础上，重点综述2016年以后基于深度学习的计算机视觉研究新进展。但为了表述的完整和逻辑的严谨，本文与文献[1]内容有少量重合。

1 通用深度神经网络模型

本文将解决图像分类任务的神经网络模型称为通用网络，这类模型通常是解决其他视觉任务的基础模型。1989年AT&T贝尔实验室的研究员LeCun通过反向传播算法成功地训练了卷积神经网络[14]，这项工作代表了20世纪80年代神经网络的研究成果。1998年LeCun等基于前人的工作提出了LeNet[15]，由2个卷积层和3个全连接层组成，因此也被称为LeNet-5，其结构如图1所示。但LeNet-5的复杂度远远无法和今天的深度网络模型相比，性能也相差悬殊，但在当时取得了和支持向量机相媲美的效果，并被广泛应用于识别手写数字，受到了广泛的关注。

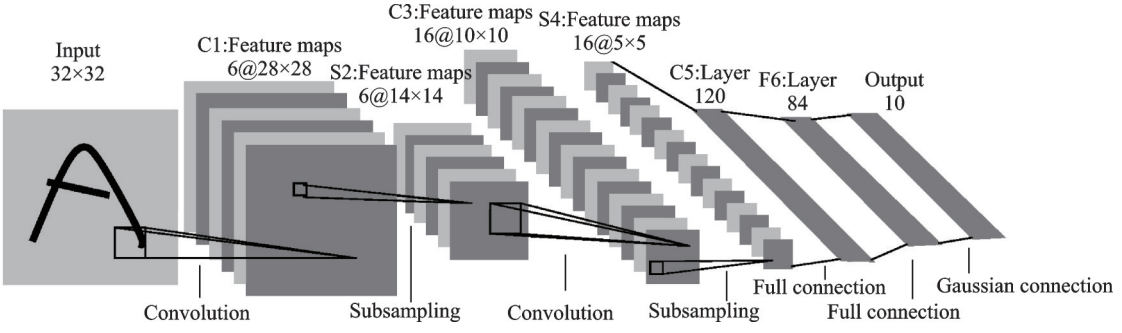

图1 LeNet-5结构示意图[15]

Fig.1 Structure of LeNet-5[15]

在LeNet提出后，很长一段时间卷积神经网络并不是计算机视觉领域的主流方法，因为LeNet只在小数据集上表现良好，在规模更大、更真实的数据集上表现一般。由于当时未普及高性能的神经网络加速硬件设备，卷积神经网络训练的时间成本和空间开销太大。因此在2012年AlexNet[13]提出之前，

大多数研究者都采用SIFT[16]、HOG[17]和SURF[18]等手工方法提取特征,并花费大量的精力进行数据整理。

2007年,普林斯顿大学李飞飞团队基于WordNet的层级结构开始搭建ImageNet数据集[19],通过网络抓取、人力标注和众包平台等各种方式,最终在2009年公开。如今ImageNet数据集包含超过14 000 000张带标签的高清图像、超过22 000个类别。从2010年开始举办的ILSVRC图像分类比赛成为计算机视觉领域的重要赛事,用于评估图像分类算法的准确率。ILSVRC比赛数据集是ImageNet的一个子集,包含1 000类、数百万张图片。来自NEC实验室的林元庆带领NEC-UIUC团队以28.2%的top-5错误率赢得了2010年ILSVRC冠军。2010和2011这两年的冠军方案主要采用HOG[17]、LBP[20-21]等算法手动提取特征再输入到特征向量机进行分类。

2012年的冠军AlexNet[13]首次将深度学习技术应用到大规模图像分类领域,证明了深度学习技术学习到的特征可以超越手工设计的特征,开启了计算机视觉领域中的深度学习热潮。AlexNet和LeNet结构理念相似,采用5层卷积层和3层全连接层,激活函数用ReLU取代了sigmoid,用dropout方法取代了权重衰减缓解过拟合,结构如图2所示。AlexNet取得了17.0%的top-5错误率。

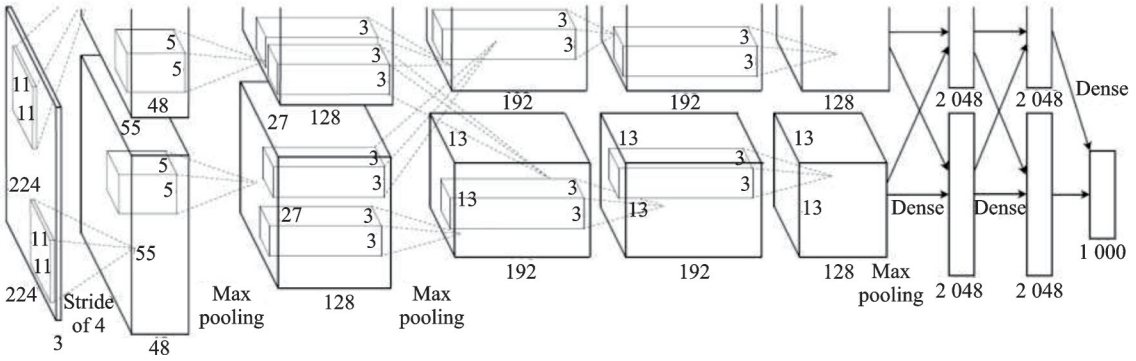

图2 AlexNet结构示意图[13]

Fig.2 Structure of AlexNet[13]

2014年的冠军团队提出的ZFNet[22]通过反卷积可视化CNN学习到的特征,取得了11.7%的错误率。2015年的冠军团队Szegedy等提出的GoogLeNet[23]将错误率降到了6.7%。GoogLeNet提出了一种Inception模块,如图3所示。这种结构基于网络中的网络(Network in network,NiN)的思想[24],有4条分支,通过不同尺寸的卷积层和最大池化层并行提取信息,1×1卷积层可以显著减少参数量,降低模型复杂度。GoogLeNet一共使用9个Inception模块,和全局平均池化层、卷积层及全连接层串联。

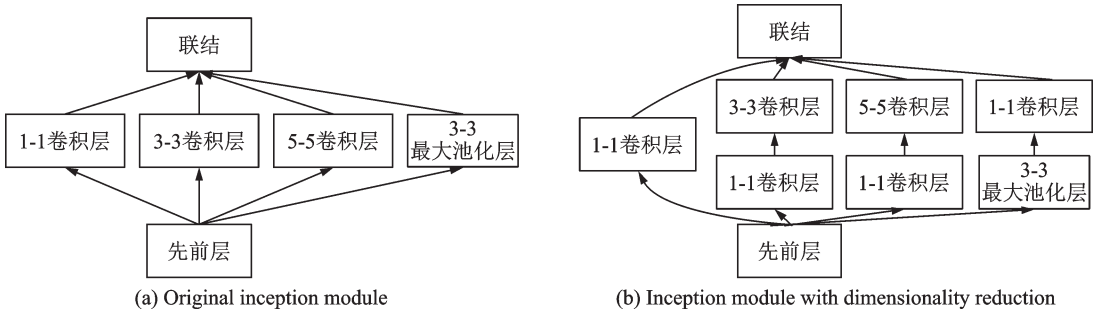

图3 Inception模块示意图[23]

Fig.3 Inception block[23]

Szegedy提出很多改进的Inception版本，陆续使用了Batch Normalization[25]、Label Smoothing[26]和残差连接[27]等方法。

2015年的ILSVRC亚军是由牛津大学视觉几何团队提出的VGGNet[28]。VGGNet重复使用了3×3的卷积核和2×2的池化层，将深度网络加深到16～19层，如图4所示。

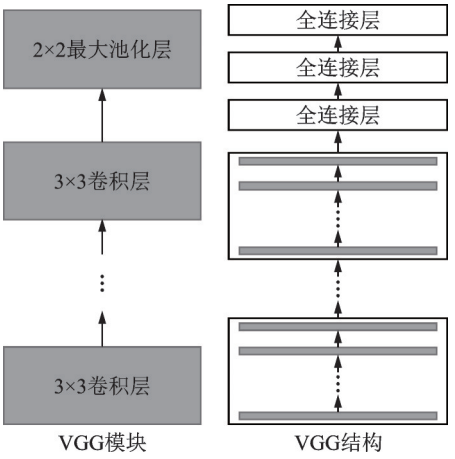

图4 VGG模块和VGG结构示意图

Fig.4 Block and structure of VGG

2016年，微软亚洲研究院He等提出的ResNet[29]夺得了ILSVRC冠军，将top-5错误率降至3.6%。ResNet最深可达152层，以绝对优势获得了目标检测、分类和定位3个赛道的冠军。该研究提出了残差模块的跳接结构，网络学习残差映射$f(x)-x$，每1个残差模块里有2个相同输出通道的3×3卷积层，每个卷积层后接1个BN(Batch normalization)层和ReLU激活函数。跳接结构可以使数据更快地向前传播，保证网络沿着正确的方向深化，准确率可以不断提高。ResNet的思想产生了深远的影响，是深度学习领域的一个重要进步，奠定了训练更深的深度网络的基础，其结构如图5所示。

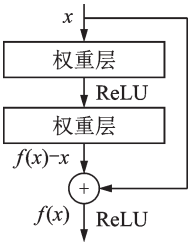

图5 残差模块

Fig.5 Residual block

2017年提出的DenseNet[30]和ResNeXt[31]都是受ResNet[29]的启发。DenseNet的目标不仅仅是学习残差映射，而且是学习类似泰勒展开的更高阶的项。因此DenseNet的跳接结构没有用加法，而是用了联结，如图6所示。

ResNeXt[31]则是结合了ResNet[29]和Inception v4[27]，采用GoogLeNet分组卷积的思想，在简化的Inception结构中加入残差连接，并通过一个超参数“基数”调整ResNeXt模块中分支的数量。这种简化的Inception结构不需要人工设计每个分支，而是全部采用相同的拓扑结构，结构如图7所示。ResNeXt在2016年ILSVRC的分类任务上获得了亚军。

和ResNeXt同年提出的Xception[32]也是一种基于Inception分组卷积思想的模型。分组卷积的核心思想是将通道拆分成不同大小感受野的子通道，不仅可以提取多尺寸的特征，还可以减少参数量，降低模型复杂度。Xception模块可以视为一种极端情况的Inception模块，它的输入先经过一个1×1的卷积层后进入多个完全相同的3×3卷积层分支，如图8所示。

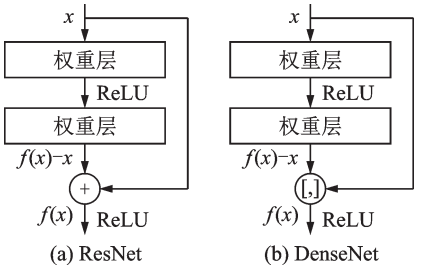

图6 ResNet和DenseNet结构比较

Fig.6 Structures of ResNet and DenseNet

ImageNet数据规模大，图像类别多，因此在ImageNet上训练的模型泛化能力较好。如今很多模型都是在ImageNet上预训练后进行微调，有些模型微调后准确率可以超过只在目标训练集上训练模型的20%。受ImageNet自由开放思想的影响，很多科技巨头也陆续开放了自己的大规模数据集：2018年谷歌发布了Open Image数据集[33]，包含了被分为6 000多类的900万张带有目标位置信息的图片；JFT-300M数据集[34]包含300万张非精确标注的图像；DeepMind也公开了Kinetics数据集[35-36]，包含650 000张人体动作的视频截图。这些大规模数据集增强了深度学习模型的泛化能力，为全世界深度学习工作者和数据科学家提供了数据支持，保障了深度学习领域的蓬勃发展。

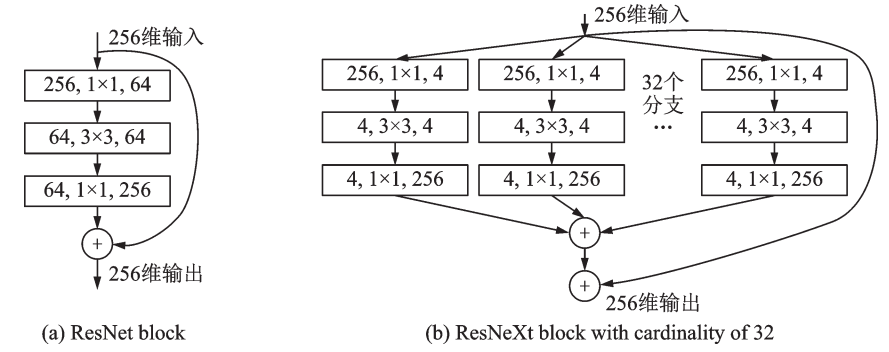

图7 ResNet残差模块和基数为32的ResNeXt模块[31]

Fig.7 ResNet block and ResNeXt block with cardinality of 32[31]

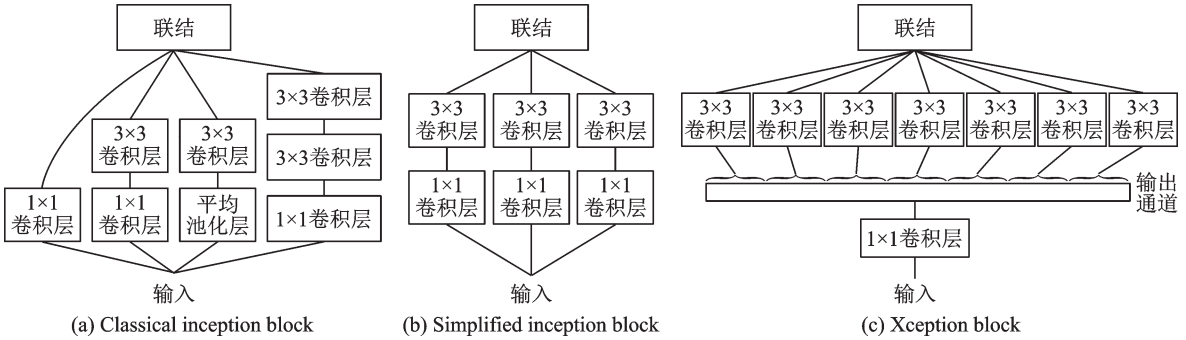

图8 经典及简化的Inception模块和Xception模块[32]

Fig.8 Classical and simplified Inception blocks and Xception block[32]

生成模型可以学习数据中隐含的特征并对数据分布进行建模,它的应用非常广泛,可以对图像、文本、语音等不同数据建模真实的分布,然后基于这一分布通过采样生成新的数据。在深度学习之前就已经有许多生成模型被提出,但由于生成模型往往难以建模,因此科研人员遇到了许多挑战。变分自编码器(Variational autoencoder, VAE)[37]是一种当前主流的基于深度学习技术的生成模型,它是对标准自编码器的一种变形。自编码器将真实样本的高级特征通过编码器映射到低级特征,被称为隐向量(或潜向量),然后又通过解码器生成相同样本的高级特征。标准自编码器和变分自编码器的区别在于对隐向量的约束不同。标准自编码器关注重构损失,即

$$\mathcal{L}(\boldsymbol{X},\boldsymbol{X}')=\left\|\boldsymbol{X}-\boldsymbol{X}'\right\|_2^2 \tag{1}$$

式中:$\boldsymbol{X}$和$\boldsymbol{X}'$分别为输入图像和重构图像。

变分自编码器则强迫隐变量服从单位高斯分布,优化如下损失函数

$$\mathcal{L}(\boldsymbol{X})=E_{z\sim q}[\lg P(\boldsymbol{X}|z)]-\mathrm{KL}(q(z|\boldsymbol{X})\|p(z)) \tag{2}$$

式中:E表示期望;z为隐变量;$q(z|\boldsymbol{X})$表示隐变量的建议分布,即编码器输出的隐变量的分布;$p(z)$表示标准高斯分布;$P(\boldsymbol{X}|z)$表示解码器分布;KL表示KL散度。式(2)等号右边第1项表示重构图片的精确度,用均方误差度量;第2项表示图片的潜变量分布和单位高斯分布之间的差异,用KL散度来度量。为了优化KL散度,变分自编码器生成1个均值向量和1个标准差向量用于参数重构。此时在隐向量分布中采样就可以生成新的图片。自编码器和变分自编码器示意图如图9、10所示。

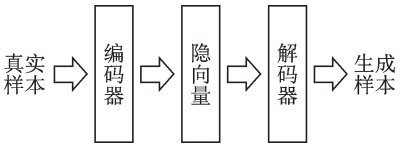

图9 自编码器示意图

Fig.9 Autoencoder

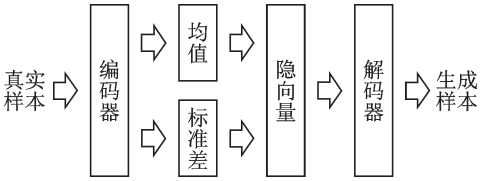

图10 变分自编码器示意图

Fig.10 Variational autoencoder

生成对抗网络(Generative adversarial net, GAN)[38]是另一种十分常见的基于深度学习技术的生成模型,它包括2个同时进行的组件:生成器和判别器,其结构如图11所示。生成器从隐向量生成图像,判别器对真伪图像进行分类,二者相互对抗,互相促进。

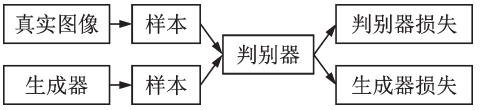

图11 生成对抗网络示意图

Fig.11 Generative adversarial net

变分自编码器和生成对抗网络近年来有了显著的发展[39]。在计算机视觉领域中,变分自编码器和生成对抗网络已经被广泛应用于图像翻译、超分辨率、目标检测、视频生成和图像分割等领域,具有广阔的研究价值和应用前景。

2 轻量化网络

随着网络层数的加深,各种深度网络模型的性能变得越来越好,随之而来的问题是模型巨大的参数量和缓慢的推理速度,因此轻量化网络的需求变得愈加强烈。轻量化网络的设计核心是在尽可能保证模型精度的前提下,降低模型的计算复杂度和空间复杂度,从而使得深度神经网络可以被部署在计算性能和存储空间有限的嵌入式边缘设备上,实现从学术界到工业界的跃迁。在分布式训练中,小模型使得服务器之间通信产生的带宽负担也相对较小。目前学术界和工业界设计轻量化的深度网络模型主要有4种方法:人工设计的轻量化神经网络、基于神经网络架构搜索(Neural architecture search,NAS)的自动设计神经网络技术、卷积神经网络压缩和基于AutoML的自动模型压缩。

2016年由伯克利和斯坦福的研究者提出的SqueezeNet[40]是最早进行深度模型轻量化的工作之一,其结构如图12所示。SqueezeNet提出了一种Fire模块用来减少参数量,其结构如图13所示。它分成Squeeze和Expand两部分:Squeeze层只由数个1×1卷积层构成;Expand层则包含数个1×1和3×3卷积层。Fire模块和Inception模块的结构很相近,二者都使用了1×1和3×3组合的拓扑结构,在使用了不同尺寸的卷积层后进行连结。在网络结构上,SqueezeNet借鉴了VGG堆叠的形式,在2层卷积层和池化层中间堆叠了8个Fire模

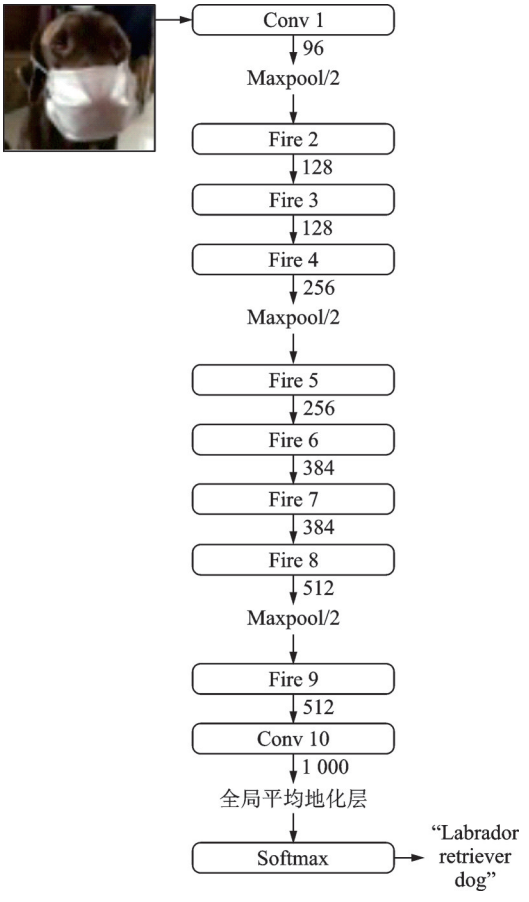

图12 SqueezeNet网络结构示意图[40]

Fig.12 Structure of SqueezeNet[40]

块。最终SqueezeNet在ImageNet上实现了AlexNet级别的精确度,参数减少到原来的1/50。通过使用Deep Compression模型压缩技术,SqueezeNet的参数量仅有50万个,约为AlexNet的1/500。

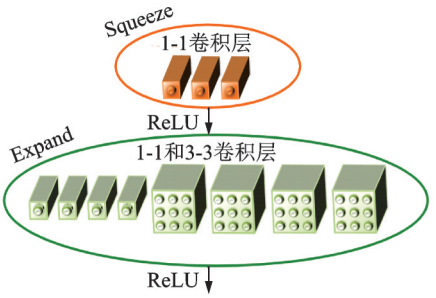

图13 SqueezeNet的Fire模块[40]

Fig.13 Fire block in SqueezeNet[40]

MobileNet[41]是谷歌于2017年提出的轻量化网络,核心是通过用深度可分离卷积代替标准的卷积。深度可分离卷积将标准卷积拆成1个深度卷积和1个逐点卷积(也就是1×1卷积),可以将计算量降低至原来的1/8~1/9。标准卷积和深度可分离卷积+BN+ReLU结构如图14所示。

深度可分离卷积的结构成为了很多轻量化网络设计的参照,这种结构的有效性自从被Xception[32]证明后成为轻量化网络设计的主流思想。比MobileNet晚2个月由Face++团队提出的ShuffleNet[42]基于这一思想,使用了Channel Shuffle和分组卷积。分组卷积的思想最早由AlexNet[13]提出,初衷是为了降低单张GPU的占用,将输入通道分成相同的几条分支然后连结,从而减少训练参数量。之后的Inception模块将这一思想发扬光大,ResNeXt[31]的成功也证明了分组卷积的有效性。由于分组卷积会让信息的流通不当,ShuffleNet设计了Channel Shuffle,将各组通道均分并进行混洗,然后依次重新构成特征图,示意图如图15所示。

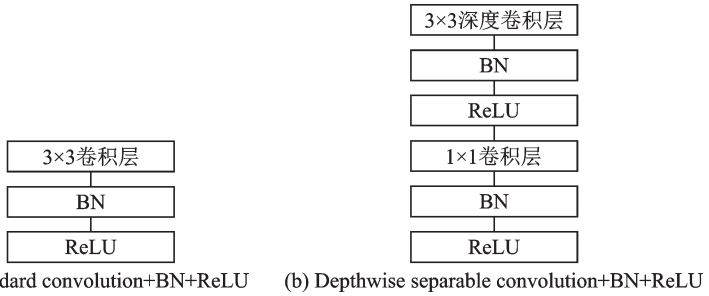

图14 标准卷积+BN+ReLU网络和深度可分离卷积+BN+ReLU网络[41]

Fig.14 Standard convolution+BN+ReLU network and depthwise separable convolution+BN+ReLU network[41]

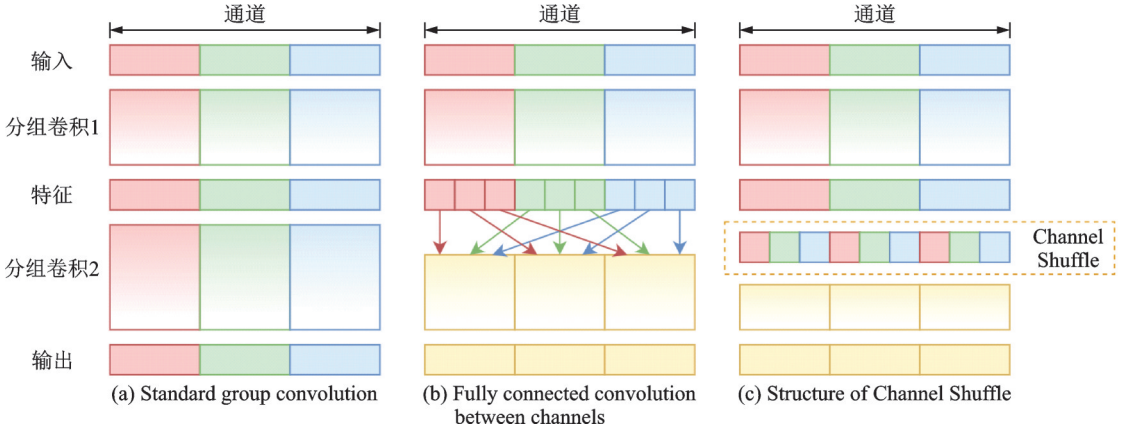

图15 Channel Shuffle示意图[42]

Fig.15 Diagrammatic sketch of Channel Shuffle[42]

图15中，Channel Shuffle后第2个组卷积GConv2的输入信息来自各个通道，图15(c,b)达到了一样的效果。ShuffleNet模块的设计借鉴了ResNet bottleneck的结构，如图16所示。

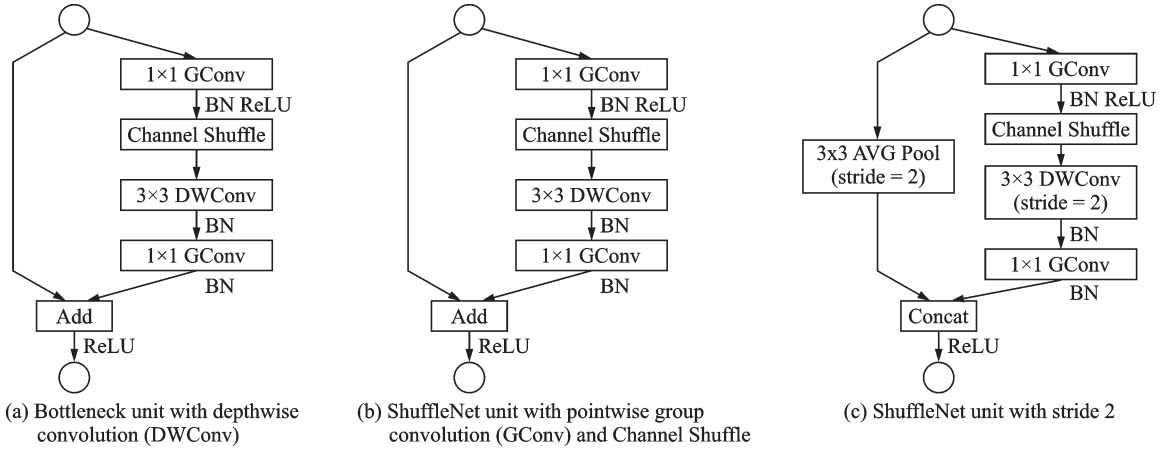

图16 ShuffleNet模块[42]

Fig.16 ShuffleNet block[42]

ShuffleNet模块摒弃了Pointwise卷积，因为对于输入维度较高的小型网络，1×1卷积的开销巨大。例如在ResNeXt模块中，1×1卷积占据了93.4%的计算量。在网络拓扑上，SqueezeNet和MobileNet都采用了VGG(Visual geometry group)的堆叠结构，而ShuffleNet采用了ResNet的跳接结构。

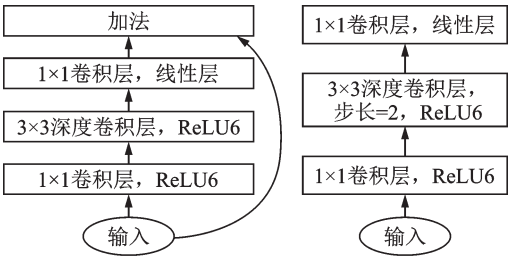

图17 MobileNet v2模块[43]

Fig.17 MobileNet v2 block[43]

2018年，MobileNet和ShuffleNet又相继提出了改进版本。MobileNet v2[43]结构如图17所示，采用了效率更高的残差结构，提出了一种逆残差模块，并将MobileNet v1模块的最后一个ReLU6层改成线性层。ShuffleNet v2[44]用更直接的运算速度评估模型，摒弃了之前如每秒浮点运算次数(FLOPS)等间接的指标。结构上ShuffleNet v2采用了一种Channel Split操作，将输入的特征图分到2个分支里，最后通过连结和Channel Shuffle合并分支并输出。ShuffleNet v1和ShuffleNet v2结构如图18所示。

2020年华为诺亚方舟实验室的团队提出了GhostNet[45]，如图19所示，可以用更少的参数量提取更多的特征图。首先对输入特征图进行卷积操作，然后进行一系列简单的线性操作生成特征图，从而在实现了传统卷积层效果的同时降低了参数量和计算量。该团队认为性能较好的主流卷积神经网络如ResNet-50通常存在大量冗余的特征图，正是这些特征图保证了网络对数据深刻的理解。Ghost模块用更小的代价模拟了传统卷积层的效果。

人工设计的轻量化网络MobileNet系列[41,43]和ShuffleNet系列[42,44]的基本思想主要是通过分离卷积操作减少运算量，再采用残差跳接结构和Channel Shuffle等混合通道的操作促进分支间的交流，提高信息利用率。随着模型规模的扩大，硬件资源变得更加稀缺，在保证精度的前提下压缩并加速模型将会是经久不衰的热门研究方向，也是信息化时代发展的必经之路。近年来大量的关于模型压缩和结构优化的工作不断涌现，如网络剪枝[46]、张量分解[47-48]和知识迁移[49]等。轻量化模型的发展有助于深度学习技术的推广和应用，推动深度学习技术的产业化发展。

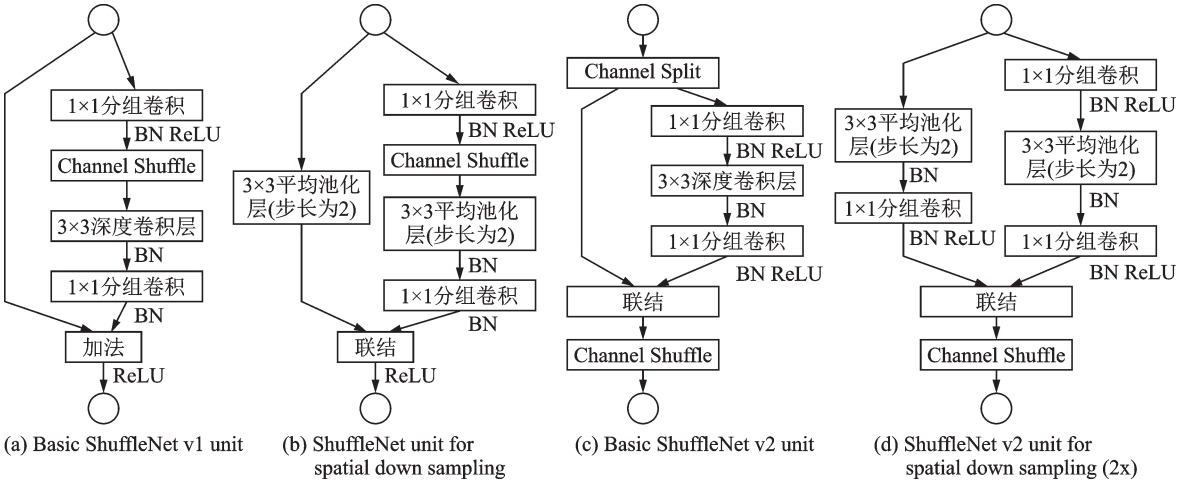

图 18 ShuffleNet v1 和 ShuffleNet v2 结构[44]

Fig.18 Structures of ShuffleNet v1 and ShuffleNet v2[44]

3 面向特定任务的深度网络模型

计算机视觉任务众多,深度学习最开始在图像分类实现突破,当前深度学习几乎深入到了计算机视觉的各个领域。本节将针对目标检测、图像分割、图像超分辨率和神经架构搜索等其他计算机视觉任务简要总结深度学习方法。

3.1 目标检测

目标检测任务作为计算机视觉的基本任务之一,包含物体的分类、定位和检测。近年来随着深度学习技术的发展,目标检测算法已经从基于手工特征的HOG[17]、SIFT[16]及LBP[20-21]等传统算法转向了基于深度神经网络的机器学习技术。自2014年Girshick等提出了R-CNN[50]模型以来,目标检测就成为了计算机视觉最受人关注的领域之一。在R-CNN之后,Girshick团队相继提出了Fast R-CNN[51]、Faster R-CNN[52]等一系列模型,这些模型均将目标检测问题归结为如何提出可能包含目标的候选区域和如何对这些区域分类两个阶段,因此这类模型也被称作两阶段模型。

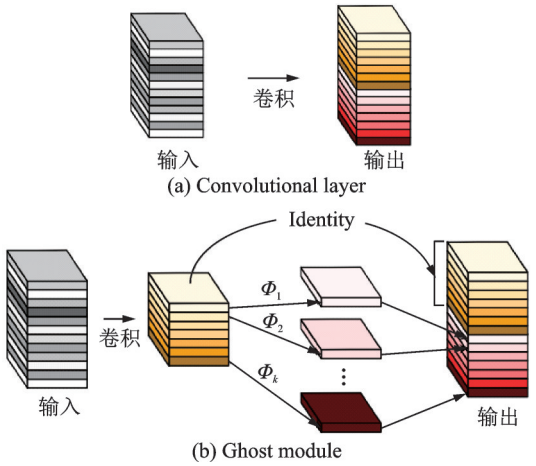

图 19 卷积层和 Ghost 模块[45]

Fig.19 Convolutional layer and Ghost module[45]

受当时性能最好的图像分类网络,如AlexNet[13]和VGG[28]等的影响,R-CNN系列模型的网络结构由2个子网组成:第1个子网用普通分类网络的卷积层提取共享特征;第2个子网的全连接层进行感兴趣区域(Region of interest,RoI)的预测和回归,中间用一个RoI池化层连接。这些网络的结构在文献[1]中已做介绍,这里不再赘述。在ResNet[29]、GoogLeNet[23]等性能更强的分类网络出现后,这种全卷积网络结构也被应用到了目标检测任务上。然而,由于卷积层并不能有针对性地保留位置信息,这种全卷积结构的检测精度远低于它的分类精度。R-FCN[53]提出了一种位置敏感分数图来增强网络对于位置信息的表达能力,提高网络的检测精度,其结构如图20所示。R-FCN[53]在PASCAL VOC 2007数据集上平均精度均值(mean Average precision, mAP)达到了83.6%,单张图片的推理速度达到170 ms。

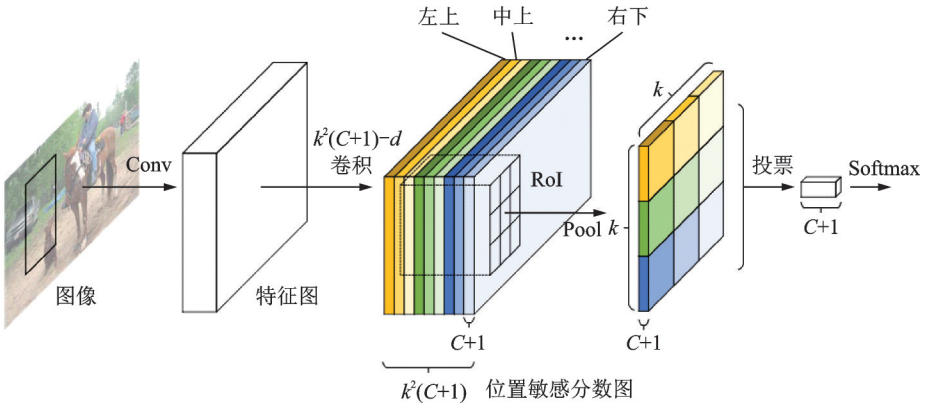

图 20 R-FCN结构示意图[53]

Fig.20 Structure of R-FCN[53]

如何准确识别不同尺寸的物体是目标检测任务的难点之一。图 21(a)中的方法通过对不同尺寸的图片提取不同尺度特征来增强不同尺度特征的语义信息,但时间和计算成本太高。图 21(b)中的单一特征图方法即为SPPnet[54]、Fast R-CNN[51]和Faster R-CNN[52]使用的方法,即在最后一层的特征图上进行预测。尽管速度较快,但包含的语义信息很少,不能准确地预测目标的位置。图 21(c)是SSD[55]采用的多尺度融合方法,从网络的不同层抽取不同尺度的特征分别进行预测,这种方法不需要额外的计算,但不能很好地提取小目标敏感的浅层高分辨率特征。

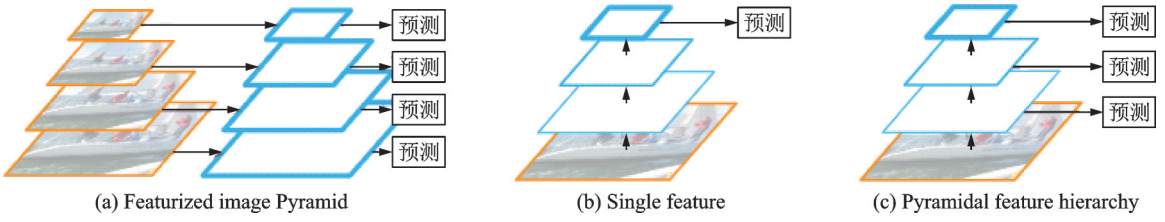

图 21 多尺度检测的常见结构[56]

Fig.21 Common structures of multiscale detection[56]

特征金字塔网络(Feature Pyramid network, FPN)[56]借鉴了ResNet跳接的思想,结合了层间特征融合与多分辨率预测,其结构如图 22 所示。文献[56]将FPN用于Faster R-CNN的区域候选网络(Region proposal network, RPN),在每层金字塔后面接一个RPN头。由于输入了多尺度的特征,因此不需要生成多尺度的锚框,只需要在每个尺度上设置不同的宽高比,并共享参数。以ResNet-101为骨干网络的Faster R-CNN+FPN在COCO test-dev上AP@0.5达到了59.1%,超过不用FPN的Faster R-CNN 3.4%。实验证明对于基于区域的目标检测器,该特征金字塔结构的特征提取效果优于单尺度的特征提取效果。

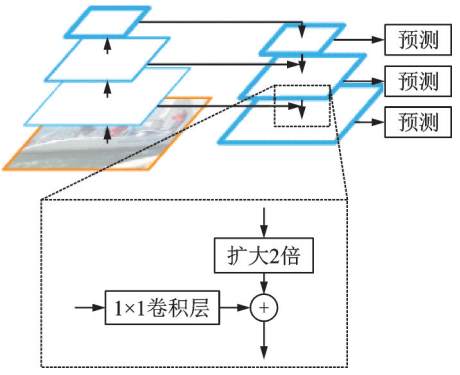

图 22 FPN结构示意图[56]

Fig.22 Structure of FPN[56]

YOLO[57]是单阶段模型的代表,它没有提出候选区域的过程,而是直接将提出候选区域和分类统一为一个边界框回归的问题,将整张图片作为网络的输入,在输出层对边界框

位置信息和类别进行回归,实现了端到端的学习过程,其示意图如图23所示。它首先将图片缩放并划分为等分的网格,然后在每张图片上运行单独的卷积网络,最后用非极大值抑制得到最后的预测框。损失函数被分为3部分:坐标误差、物体误差和类别误差。为了平衡类别不均衡和大小物体等带来的影响,损失函数中添加了权重并将长宽取根号。

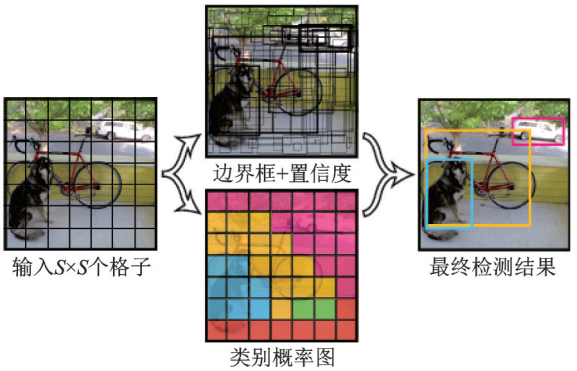

图23 YOLO示意图[57]

Fig.23 Pipeline of YOLO[57]

YOLO的网络结构借鉴了GoogLeNet的结构,用24层卷积层后接2层全连接层,将Inception模块替换为类似网中网[24]中的1×1卷积层后接3×3卷积层,并在ImageNet上预训练,其结构如图24所示。在PASCAL VOC 07+12数据集上,YOLO在达到最高帧率155帧/s时mAP可以达到52.7%,在mAP最高达到63.4%时帧率可达45帧/s。YOLO在保证了准确率的同时拥有极高的推理速度,远超当时的两阶段模型。

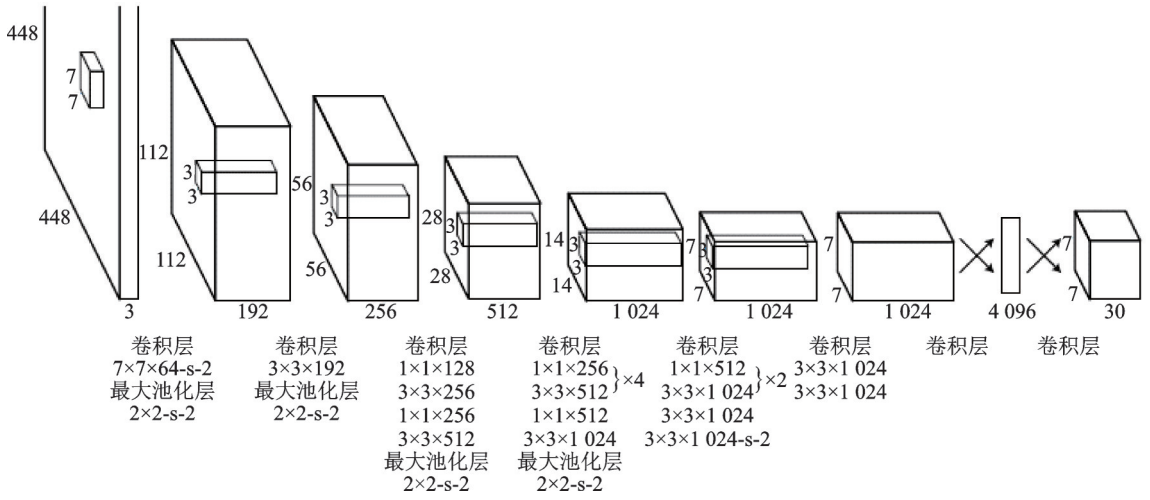

图24 YOLO网络结构图[57]

Fig.24 Structure of YOLO[57]

YOLOv1的训练流程简单,背景误检率低,但由于只选择交并比最高的边界框作为输出,每个格子最多只能预测出一个物体。当每个格子包含多个物体时,YOLOv1只能检测出1个目标。YOLOv2[58]在YOLOv1的基础上,骨干网络采用了以VGG16为基础的Darknet19,使用了批量归一化缓解了梯度爆炸和消失的问题。YOLOv2借鉴了Faster R-CNN锚框的设计,将YOLOv1的全连接层替换为锚框预测边界框的位置,解耦了位置和类别的回归计算。YOLOv2[58]同时采用了多尺度训练,提升了模型的健壮性。后续的YOLOv3[59]骨干网络采用了Darknet53,使用了ResNet的跳接结构,并引入了FPN,一定程度上解决了YOLOv2小目标检测精度较差的问题。YOLOv3在分辨率320像素×320像素的输入上以22 ms的推理时间使得mAP达到28.2%,和当时最好的单阶段检测器SSD达到相同精度,但拥有3倍的推理速度。YOLOv3以51 ms的推理时间使得AP@0.5达到57.9%,相较于以198 ms的推理时

间AP@0.5达到57.5%的RetinaNet[60]，精度相近但YOLOv3的速度是RetinaNet[60]的近4倍。

SSD[55]是最早达到两阶段模型精度的单阶段模型之一，对后期的单阶段工作影响很深，其结构如图25所示。为解决YOLOv1小目标检测精度低的问题，基于VGG不同的卷积段采用了多尺度的特征图，并在每个网格点生成更多的不同大小和长宽比的预测框。SSD在PASCAL VOC 2007数据集上，对于300像素×300像素的输入mAP达到了74.3%，512像素×512像素的输入mAP达到了76.9%。在COCO trainval35k数据集上预训练再在PASCAL VOC 07+12上微调后，SSD最终mAP达到了81.6%。

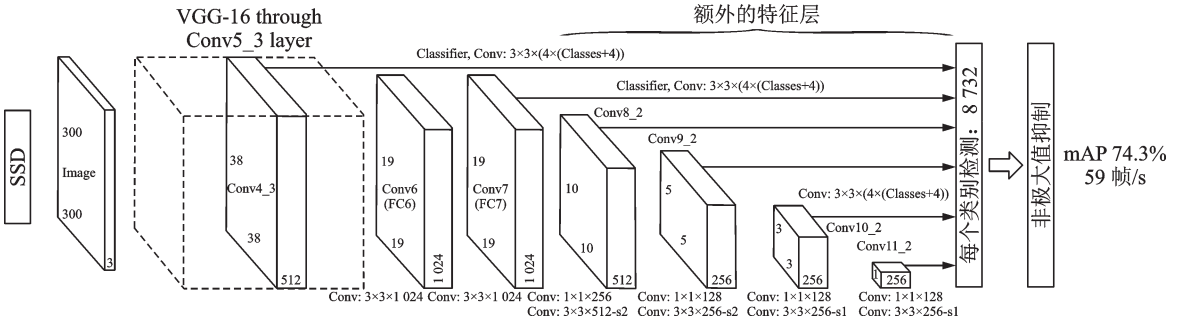

图25 SSD网络结构图[55]

Fig.25 Structure of SSD[55]

和两阶段模型相比，单阶段模型只需要进行一次类别预测和位置回归，因此卷积运算的共享程度更高，拥有更快的速度和更小的内存占用。最新的单阶段模型如FCOS[61]、VFNet[62]等工作已经可以达到接近两阶段模型精度，同时拥有更好的实时性，更适合在移动端部署。

目标检测技术从传统的手工特征算法到如今的深度学习算法，精度越来越高的同时速度也越来越快。在过去几年中，工业界已经出现了成熟的基于目标检测技术的应用，如人脸检测识别、行人检测、交通信号检测、文本检测和遥感目标检测等。这些应用不仅便利了人们的生活，也为学术界提供了启发和指导。

在未来的研究工作中，小目标检测和视频目标检测依旧是研究的热点问题。同时，为了加快推理速度并在移动端嵌入式设备部署模型，目标检测的轻量化一直备受工业界的关注。在采集到多模态的信息（如文字、图像、点云等）后，如何通过更好的信息融合来提高检测性能也是未来的一个重点研究方向。

3.2 图像分割

本文的图像分割指图像语义分割任务，其要求将整张图片的所有像素分类为预先定义的多个类别之一。由于是像素级的稠密分类任务，相比图像分类和目标检测更加困难，是图像处理和计算机视觉中的一个重要课题，在场景理解、医学图像分析、机器人感知及视频监控等领域有着广泛的应用。近年来，由于深度学习技术在计算机视觉领域应用中取得的成功，人们也进行了大量的工作研究基于深度学习模型的图像分割方法。

U-Net[63]和全卷积网络（Fully convolutional network, FCN）[64]都是在2015年提出的网络，启发了后来的很多图像分割和目标检测的工作。FCN已在文献[1]中进行介绍，此处不再赘述。U-Net最初是一个用于医学图像分割的卷积神经网络，分别赢得了ISBI 2015细胞追踪挑战赛和龋齿检测挑战赛的冠军。U-Net可视为一个编码器-解码器结构，编码器有4个子模块，每个子模块通过一个最大池化层下采样，解码器再通过上采样的4个子模块增大分辨率直到与输入图像的分辨率保持一致，其结构如

图26所示。由于卷积采用的是Valid模式,实际输出图像的分辨率低于输入图像的分辨率。U-Net网络同时还采取了跳接结构(即图26中的灰色箭头),将上采样结果与编码器中具有相同分辨率的子模块的输出进行连接,作为解码器中下一个子模块的输入。

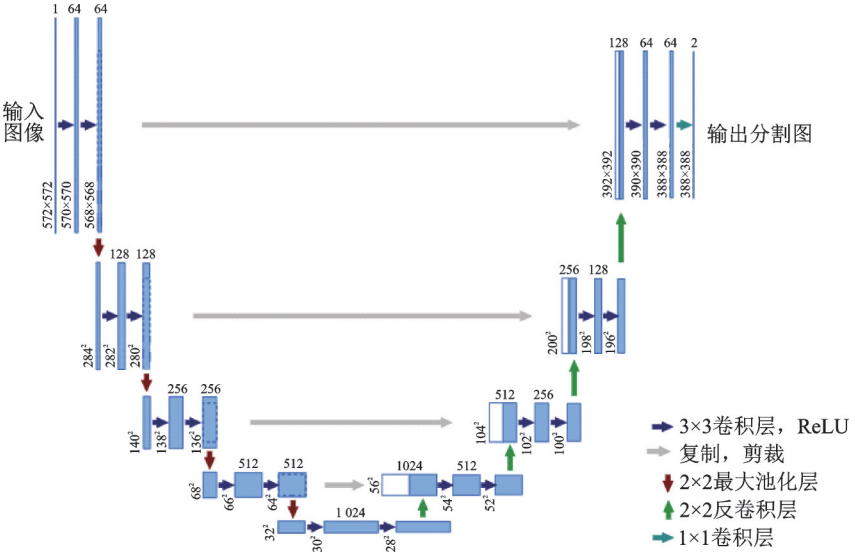

图26 U-Net结构示意图[63]

Fig.26 Structure of U-Net[63]

由于人体结构相对固定,分割目标在图像内的分布很有规律,医学图像大多语义明确,需要低分辨率的信息用于目标物体的识别。同时医学图像形态复杂,往往要求高精度的分割,需要高分辨率的信息用于精准分割。U-Net融合了高低分辨率的信息,因此对医学图像分割的效果很好。

Mask R-CNN[65]是R-CNN团队的又一次探索,他们在之前Faster R-CNN[52]的基础上,将其扩展到更精细的像素级别的分类,从而从目标检测领域拓展到图像分割领域。通过使用RoIAlign代替RoIPooling,得到更好的定位效果,并在Faster R-CNN上添加了二进制的Mask,表征像素是否在目标范围内完成图像分割的任务。Mask R-CNN网络结构图和分支结构图如图27、28所示。

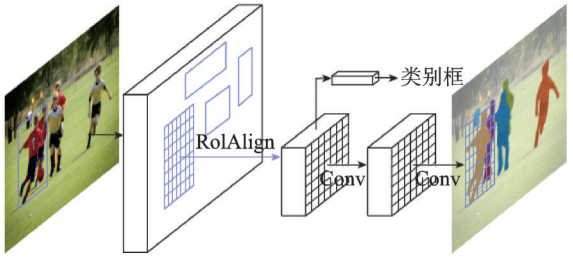

图27 Mask R-CNN网络示意图[65]

Fig.27 Structure of Mask R-CNN[65]

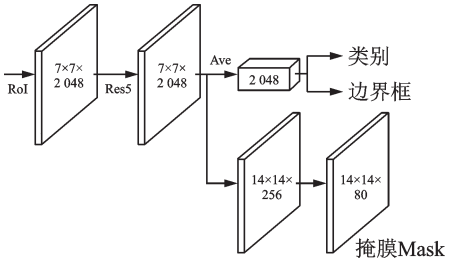

图28 Mask R-CNN分支示意图[65]

Fig.28 Structure of Mask R-CNN's branches[65]

深度卷积神经网络中池化层和上采样层的设计对于图像分割的设计有致命缺陷。因为参数不可学习,而且池化会导致像素的空间信息和内部的数据结构丢失,上采样也无法重建小物体信息,因此图像分割的精度一直处于瓶颈。针对这一问题,2016年的DeepLab[66]又提出了一种空洞卷积,避免了池

化层带来的信息损失，并使用全连接的条件随机场（Conditional random field，CRF）优化分割精度，其结构如图29所示。

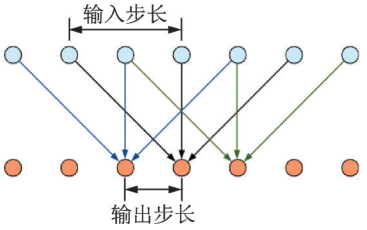

图29 空洞卷积示意图（卷积核尺寸为3，输入步长为2，输出步长为1）[66]

Fig.29 Dilated convolution (kernel size=3, input stride=2, output stride=1)[66]

空洞卷积可以在避免使用池化层损失信息的情况下增大感受野，同时不增加参数数量。作为后处理，DeepLabv1将每个像素点作为节点，像素之间的关系作为节点间的连线，构成一个条件随机场，再用一个二元势函数描述像素点之间的关系，将相似像素分配相同的标签，从而在分割边界取得良好的效果。DeepLabv1速度很快，帧率达到8帧/s，在PASCAL VOC 2012数据集上平均交并比（Mean intersection over union，mIoU）达到了71.6%，它的“深度卷积神经网络＋条件随机场”结构对之后很多工作产生了深远的影响。

2017年剑桥大学提出的SegNet[67]的主要动机是针对道路和室内场景理解，设计一个像素级别的图像分割网络，同时保证内存和计算时间方面上的高效。SegNet采用“编码器-解码器”的全卷积结构，编码网络采用VGG16[28]的卷积层，解码器从相应的编码器获取最大池化索引后上采样，产生稀疏特征映射。复用池化索引减少了端到端训练的参数量，并改善了边界的划分。SegNet在道路场景分割数据集CamVid 11 Road Class Segmentation[68]上mIoU达到60.1%，边界F_1得分（Boundary F_1 score，BF）达到46.84%；在室内场景分割数据集SUN RGB-D Indoor Scenes[69]上几乎所有当时的深层网络结构都表现不佳，但SegNet依然在绝大多数的指标上超过了其他网络。SegNet结构如图30所示。

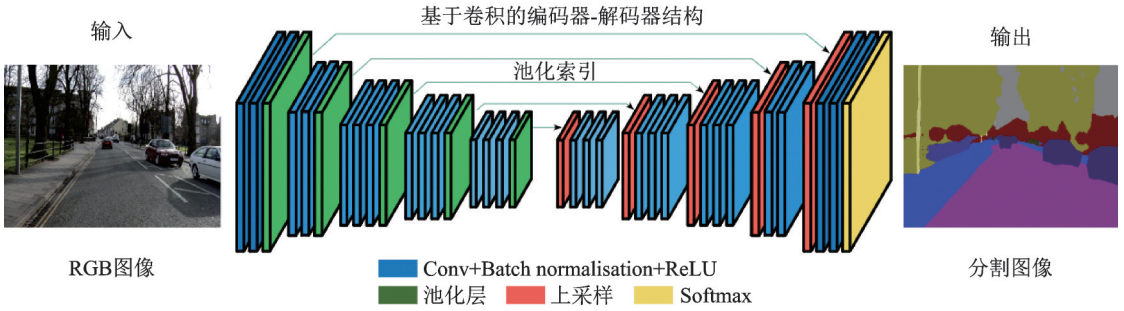

图30 SegNet结构示意图[67]

Fig.30 Structure of SegNet[67]

2017年香港中文大学提出了PSPNet[70]，该网络采用金字塔池化模块，用大小为1×1、2×2、3×3和6×6的4层金字塔分别提取不同尺度的信息，然后通过双线性插值恢复长宽，把不同层的特征连结起来得到全局信息，这种结构比全局池化更具有代表性，融合了多尺度的信息。PSPNet在PASCAL VOC 2012数据集上mIoU达到了82.6%，在MS COCO数据集上预训练后达到85.4%。PSPNet结构如图31所示。

DeepLabv2[71]在DeepLabv1[66]和PSPNet[70]的基础上用ResNet101代替VGG16，并提出了一种带有空洞卷积的空间金字塔池化模块（Atrous spatial Pyramid pooling，ASPP），用多尺度的方法以不同的速率并行地提取特征图信息，极大地增加了感受野，其结构如图32所示。DeepLabv2使用不同的学习率，相比DeepLabv1，mIoU达到了79.7%，提升了8.1%，但二者都使用了全连接条件随机场模块。

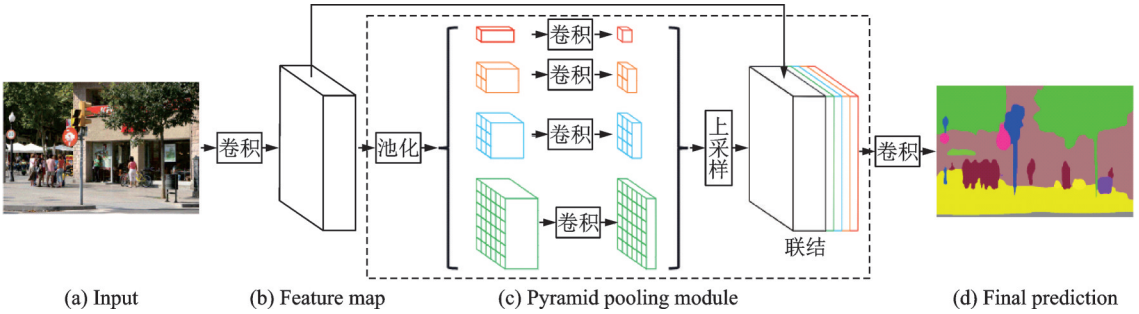

图31 PSPNet结构示意图[70]

Fig.31 Structure of PSPNet[70]

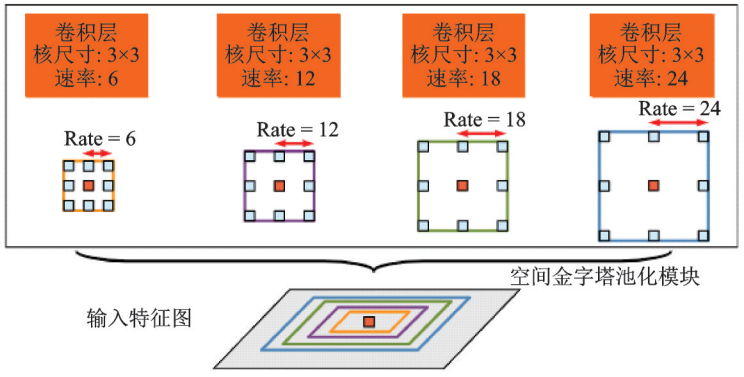

图32 空洞空间金字塔池化示意图[71]

Fig.32 Structure of ASPP[71]

DeepLabv3[72]重新审视了空洞卷积的作用,将其级联模块应用在ResNet最后一个模块之后。不使用空洞卷积和使用空洞卷积的级联模块示意图如图33所示。

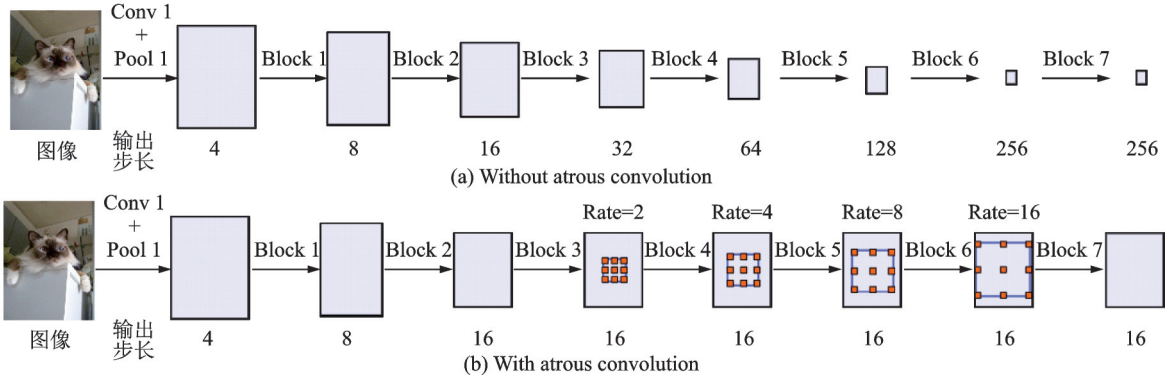

图33 不使用和使用空洞卷积的级联模块示意图[72]

Fig.33 Structures of cascade modules without and with atrous convolution[72]

DeepLabv3改进了ASPP模块,应用BN层,并将DeepLabv2中Rate=24的3×3卷积模块替换为1×1卷积模块和全局池化模块,克服了像素点相对距离增大时有效权重减少的问题。DeepLabv3去掉了后处理的DenseCRF模块,并最终在PASCAL VOC 2012数据集上mIoU达到了86.9%,相较DeepLabv2进一步提升了7.2%。改进的ASPP模块示意图如图34所示。

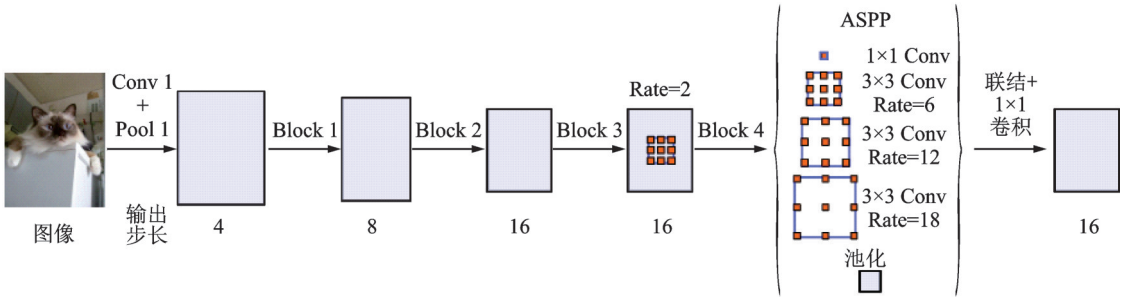

图 34 改进的 ASPP 模块示意图[72]

Fig.34 Improved ASPP module[72]

DeepLabv3+[73]相对于 DeepLabv3，采用了“编码器-解码器”的结构，编码器中包含丰富的语义信息，解码器则输出图像的边缘细节信息。空间金字塔池化模块，“编码器-解码器”结构和带有空洞卷积的“编码器-解码器”结构如图 35 所示，DeepLabv3+结构如图 36 所示。

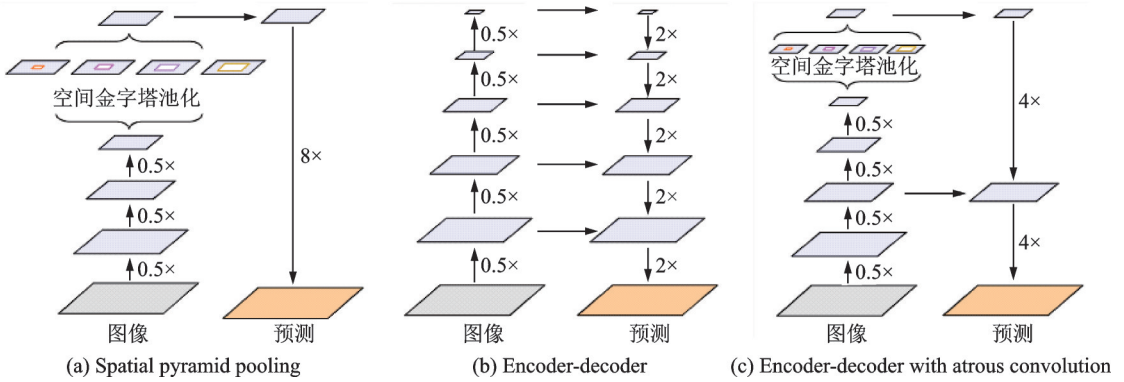

图 35 DeepLabv3+使用了空间金字塔池化模块，“编码器-解码器”结构和空洞卷积[73]

Fig.35 DeepLabv3+ employing spatial Pyramid pooling, encoder-decoder and atrous convolution[73]

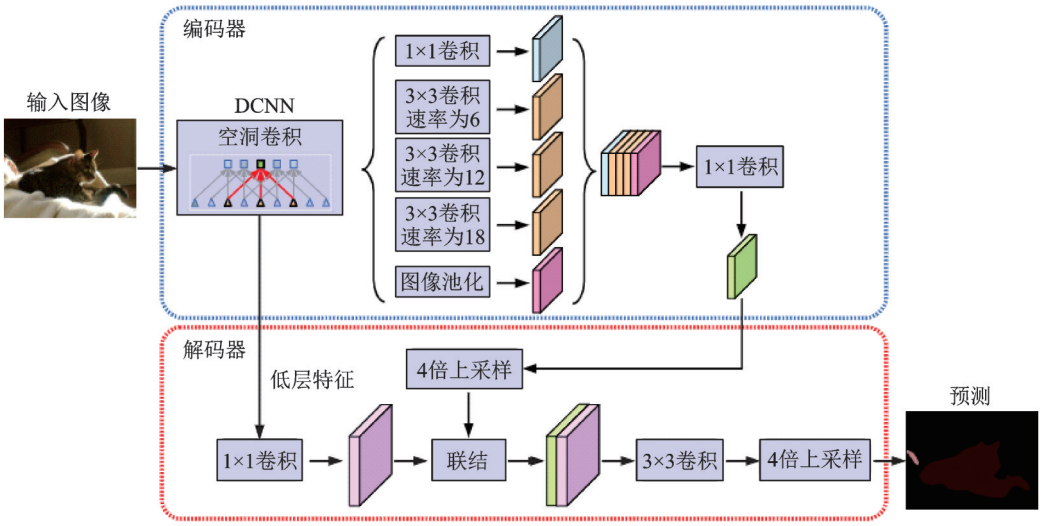

图 36 DeepLabv3+示意图[73]

Fig.36 Structure of DeepLabv3+[73]

DeepLabv3+将之前的骨干网络ResNet101替换为Xception,并结合深度可分离卷积的思想提出了空洞深度可分离卷积,在减少参数量的同时进一步增大感受野。和DeepLabv3一样,DeepLabv3+也没有使用DenseCRF后处理模块。最终DeepLabv3+在PASCAL VOC 2012数据集上mIoU达到了89.0%,相较DeepLabv3提升了2.1%。深度卷积、逐点卷积和空洞深度可分离卷积示意图如图37所示。

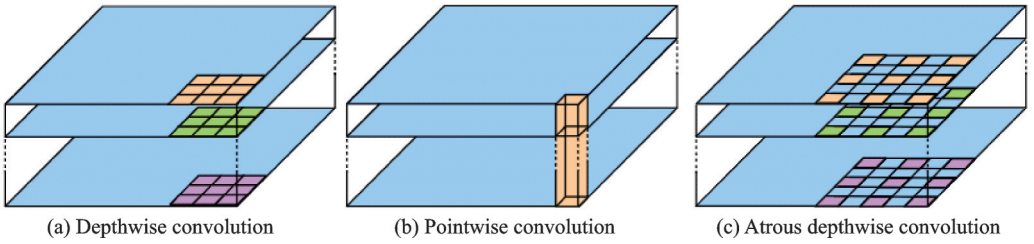

图37 空洞深度可分离卷积示意图[73]

Fig.37 Structure of atrous depthwise separable convolution[73]

2019年旷视科技提出了一种名为DFANet[74]的高效CNN架构,通过子网和子级联的方式聚合多尺度特征,极大地减少了参数量,其结构如图38所示。DFANet采用"编码器-解码器"结构,解码器的骨干网络采用3个改良的轻量级Xception融合结构,编码器则是一个高效的上采样模块,用于融合高层和底层的语义信息。在CityScapes[75]测试数据集上,对于1 024像素×1 024像素的输入图片,DFANet在一块NVIDIA Titan X上mIoU达到71.3%,FLOPS仅为3.4×10^9,帧率达到100帧/s;在CamVid[68]测试数据集上,对于960像素×720像素的输入图片,DFANet在8 ms的计算时间内mIoU达到64.7%,帧率达到120帧/s。

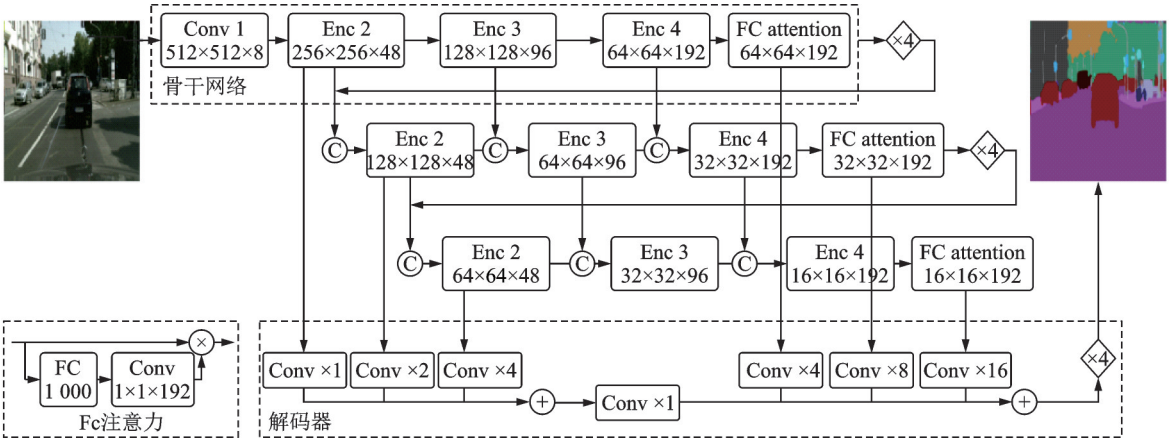

图38 DFANet结构示意图[74]

Fig.38 Structure of DFANet[74]

2020年笔者提出一种轻量级网络LRNNet[76]。其中分解卷积块FCB(图39(a))利用1×3和3×1的空间分解卷积处理短距离特征,并利用空洞深度分离卷积处理远距离特征,实现了参数量和计算量更少、深度更快、准确率更高的特征提取;高效的简化Non-Local模块LRN(图39(b))利用区域主奇异向量作为Non-Local模块的Key和Value,在降低Non-Local模块的计算量和内存占用的同时,保持其处理远距离关联的效果。在Cityscapes[75]测试集上,LRNNet的mIoU达到了72.2%,而网络仅有68万个参数,并在1张GTX 1080Ti卡上达到71帧/s的推理速度;在CamVid[68]测试集上,对于360像素×480像

素的输入，LRNNet的mIoU达到了69.2%，参数量也为68万个，在1张GTX 1080Ti卡上帧率达到76.5帧/s。

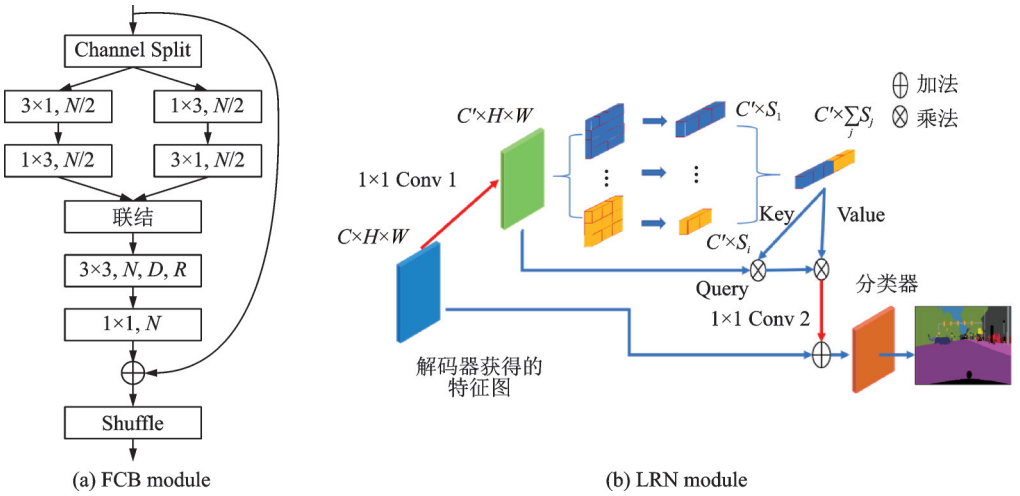

图39 LRNNet中的FCB和LRN模块[76]

Fig.39 FCB and LRN modules in LRNNet[76]

图像分割是像素级的稠密分类任务，在搜集数据集时需要真值标注每个像素，但由于这个要求极其耗时且非常昂贵，许多研究人员开始用弱监督学习和半监督学习的方法训练网络。常见的弱标注有图像类别标签、边界框、显著图和类激活图(Class activation map，CAM)等。

2015年谷歌和UCLA团队的工作[77]是最早开始研究基于弱监督学习技术的图像分割算法之一。该工作基于DeepLab模型[66]，研究了弱标注(类别标签、边界框等)与少量强标注(像素级标签)和大量弱标注混合对DCNN图像分割模型的影响，并在半监督和弱监督的设定下提出了一种期望最大化方法(Expectation-maximization，EM)。这项工作证实了仅使用图像级标签的弱标注存在性能差距，而在半监督设定下使用少量强标注和大量弱标注混合可以获得优越的性能，在MS COCO数据集上使用5 000张强标注图片和118 287张弱标注图片mIoU超过70%。

尽管类别标签的获取成本很低，但这类标注信息仅仅标明某类目标存在，不能表示出目标的位置和形状，这往往会导致分割效果不够理想，存在边界模糊等问题。当出现目标遮挡的情况时，仅使用图像级标签获取完整的目标边界会更加困难。为了补充监督信息中缺少的位置和形状信息，使用图像的显著性信息是一种常见的手段。文献[78]提出了一个仅使用类别标签和显著图信息的图像分割模型，其结构如图40所示。该模型将图像的显著图定义为一个人最有可能先看到的目标的二进制掩膜，用预训练的目标检测网络提取出显著性区域，通过种子信息确定目标的类别和位置。该工作同样基于DeepLab[66]的网络结构，提出的模型测试精度mIoU达到56.7%，实现了全监督模型80%的性能。

定位线索的另一个流行的选择是使用CAM。主流的弱监督方法通过将CAM作为分割种子，突出局部的显著部分，然后逐渐生长直到覆盖整个目标区域，从而补充了缺失的目标形状信息。2018年提出的AffinityNet[79]结合了类别标签和CAM信息，首先计算图像的CAM作为监督源训练AffinityNet，通过构建图像的语义相似度矩阵，结合随机游走进行扩散，不断奖励或惩罚从而修改CAM，最终恢复出目标的形状。AffinityNet流程如图41所示。

深度学习技术在图像分割领域取得了显著成就，但仍然面临不小的挑战。当前的大规模数据集如

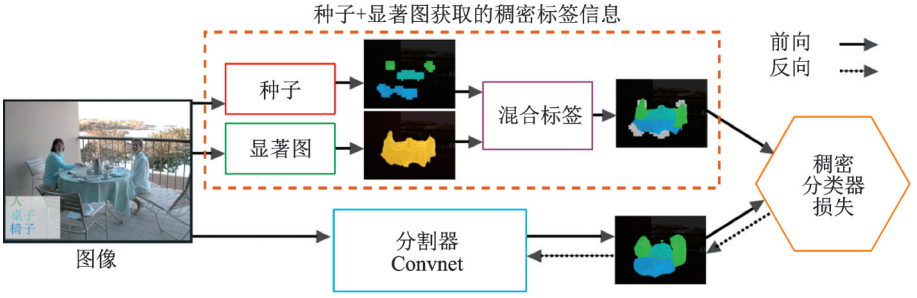

图40 高层信息指导的图像分割网络结构图[78]

Fig.40 High-level guided segmentation architecture[78]

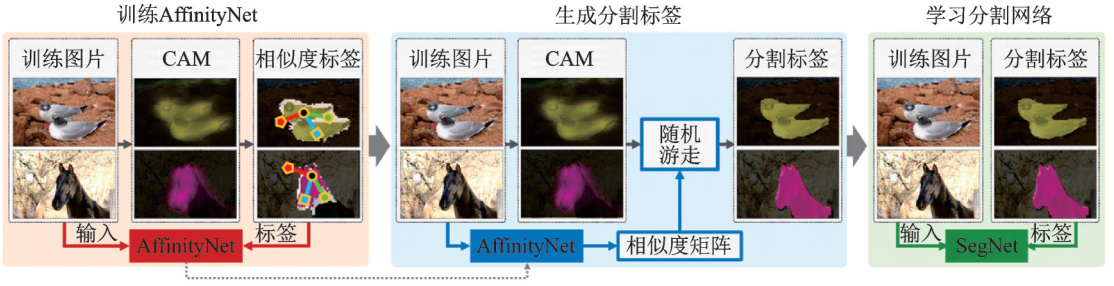

图41 AffinityNet流程示意图[79]

Fig.41 Pipeline of AffinityNet[79]

MS COCO[80]和PASCAL VOC[81]并不能满足工业界的需求,而具有多目标和重叠目标的数据集对于图像分割而言更具有应用价值,这可以使得图像分割技术更好地处理密集目标场景和现实生活中常见的重叠目标场景。基于小样本学习技术的图像分割算法同样具有广阔的前景,因为在许多应用领域,例如医学图像分析领域,获取学习样本的成本较高,难度也较大。图像分割技术的实时性也是一个难题,目前大多数模型并不能达到实时性的要求,但在很多应用场景下,速度的重要性远高于精度。

3.3 超分辨率

超分辨率技术是计算机视觉领域提高图像和视频分辨率的重要处理技术之一,研究如何将低分辨率的图像或图像序列恢复出具有更多细节信息的高分辨率图像或图像序列,在高清电视、监控视频、医学成像、遥感卫星成像、显微成像及老旧图像视频修复等领域有着重要的应用价值。传统上超分辨率属于底层视觉领域,但本文叙述顺序从图像分类、目标检测、图像分割到超分辨率,输出逐级复杂,依次为图像标签、目标位置和类别标签、与输入同大小的分割图、比输入图像大的高分辨率图像等。与前几个任务不同,超分辨率需要生成和恢复输入中不存在的信息。

超分辨率的概念最早出现在光学领域,1952年Francia第一次提出了用于提高光学分辨率的超分辨率的概念[82]。1964年前后,Harris[83]和Goodman[84]分别提出了后来称为Harris-Goodman频谱外推的方法,这被认为是最早的图像复原方法,但这种技术只能在一些理想情况下进行仿真,实际效果不太理想,因此并未得到推广。1984年Tsai等[85]首次利用单幅低分辨率图像的频域信息重建出高分辨率图像后,超分辨率重建技术才得到广泛的认可和应用,如今它已经成为图像增强和计算机视觉领域中最重要的研究方向之一。

传统的超分辨率方法包括基于预测、基于边缘、基于统计、基于块和基于稀疏表示等方法。根据输

入输出的不同，超分辨率问题可以分为基于重建的超分辨率问题、视频超分辨率问题和单幅图像超分辨率问题。根据是否依赖训练样本，超分辨率问题则又可以分为增强边缘的超分辨率问题（无训练样本）和基于学习的超分辨率问题（有训练样本）。

最简单、应用最广泛的经典单幅图像超分辨率方法是插值法，包括Lanczos、Bicubic、Bilinear和Nearest等，这种方法操作简单、实施性好，但并不能恢复出清晰的边缘和细节信息，因此很多其他用于增强细节的传统算法相继被提出。文献[86]提出了基于块的方法，也被称为基于邻域嵌入的方法。这种方法使用流形学习中的局部线性嵌入，假设高、低维度中图像块的线性关系可以保持，用低分辨率图像的特征（梯度等）重构高分辨率图像。文献[87-88]提出了基于稀疏表示的方法，也被成为字典学习。这种方法将低分辨率图像和高分辨率图像表示为字典D与原子α，高分辨率图像可表示为$x=D_{\text{high}}$，低分辨率图像为$y=D_{\text{low}}$，假设不同分辨率的同一幅图像的原子α，在训练完字典D_{high}和D_{low}后，用低分辨率的图像得到α，随后得到重构的高清图像。基于学习的超分辨率技术[89]如图42所示，上、下采样方法示意图[90]如图43所示。

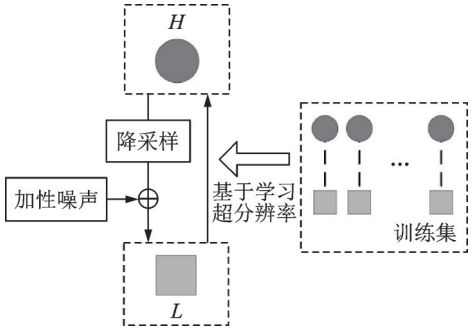

图42 基于学习的超分辨率技术[89]

Fig.42 Learning-based super-resolution[89]

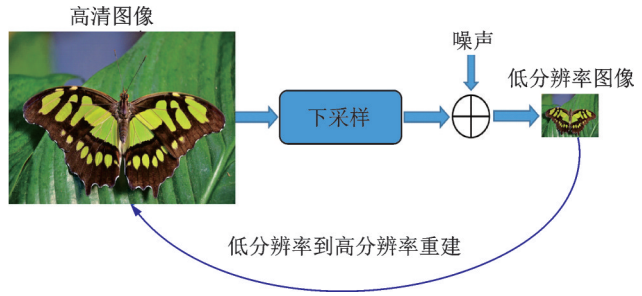

图43 超分辨率问题中的上采样和下采样方法[90]

Fig.43 Upsampling and downsampling in super-resolution[90]

经典的超分辨率方法要求研究者具备深厚的相关领域先验知识。随着深度学习技术的兴起，用神经网络方法重建的图像质量超过了传统方法，速度也更快，这使得大批学者转向对深度学习技术在超分辨率领域的应用研究。香港中文大学Dong等于2015年首次将卷积神经网络用于单幅图像超分辨率重建，提出了SRCNN[91]，该网络仅仅用了3个卷积层，利用传统稀疏编码，依次进行图像块提取、非线性映射和图像重建，实现了从低分辨率图像到高分辨率图像的端到端映射，流程图如图44所示。SRCNN激活函数采用ReLU，损失函数采用均方误差。

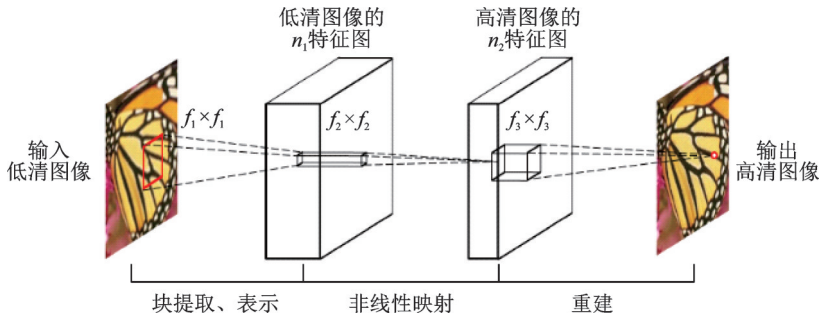

图44 SRCNN流程图[91]

Fig.44 Pipeline of SRCNN[91]

2016年Dong团队在之前SRCNN的基础上提出了更快、实时性更好的FSRCNN[92],在原始网络的最后加入反卷积层放大尺寸,摒弃了Bicubic插值方法,使用了更多的映射层和更小的卷积核,改变了特征维度,并共享其中的映射层,FSRCNN改进示意图如图45所示。训练时FSRCNN只需要微调最后的反卷积层,因此训练速度很快。FSRCNN激活函数采用PReLU,损失函数仍采用均方误差。

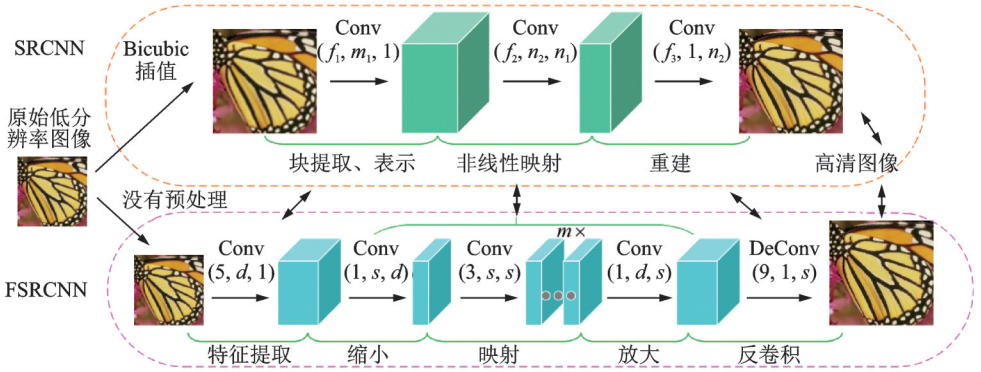

图45 FSRCNN对SRCNN的改进[92]

Fig.45 FSRCNN's improvement on SRCNN[92]

2016年提出的ESPCN[93]在SRCNN基础上进一步提高了速度,其结构如图46所示。该工作提出了一种亚像素卷积层,可以直接在低分辨率图像上提取特征,从而避免在高分辨率图像上进行卷积,降低了计算复杂度。ESPCN激活函数采用tanh,损失函数仍然采用均方误差。

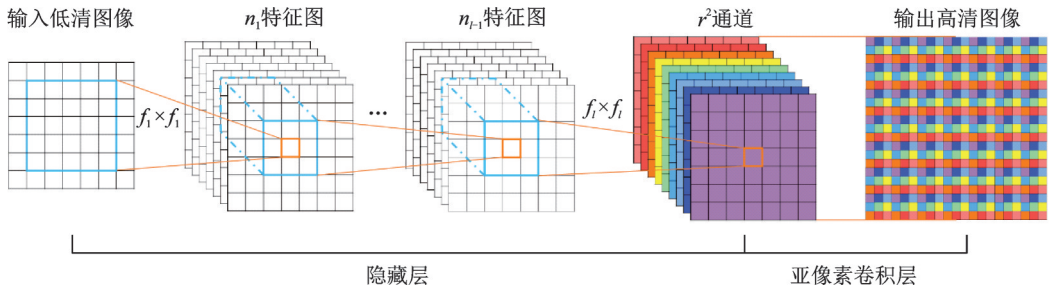

图46 ESPCN示意图[93]

Fig.46 Structure of ESPCN[93]

SRCNN的网络输入是经过上采样的低分辨率图像,计算复杂度很高,因此FSRCNN和ESPCN都选择在网络末端上采样以降低计算复杂度。但如果在上采样后没有足够深的网络提取特征,图像信息就会损失。为了更好地使用更深的网络,很多工作引入了残差网络。2016年首尔国立大学Kim等提出的VDSR[94]是第一个引入全局残差的模型,其结构如图47所示。Kim等指出,高低分辨率图像携带的低频信息很相近,因此事实上网络只需要学习高频信息之间的残差即可。VSDR思想启发了很多之后利用残差结构的工作。

CARN[95]是NTIRE2018超分辨率挑战赛的冠军方案,该方案使用全局和局部级联,将ResNet的残差块替换成级联模块和1×1卷积模块组合,并提出了一种残差-E模块,可以提升CARN的效率。CARN的改进如图48所示,其局部级联模块如图49所示。

EDVR[96]是商汤科技2019年提出的一种用于视频修复的通用框架,在NITRE 2019的4个赛道中均以较大的优势获得了冠军。视频修复任务包括超分辨率、去噪声等任务,早期的研究者们简单地将

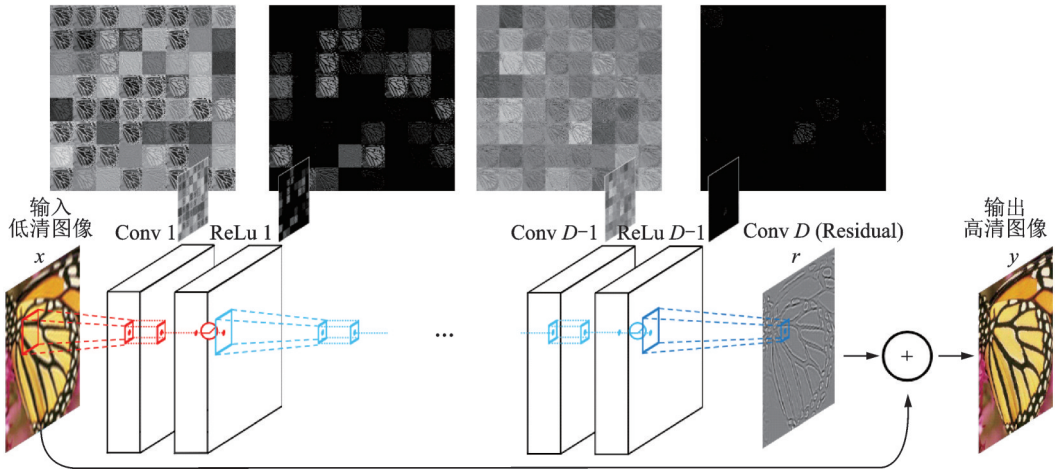

图 47 VSDR 网络结构图[94]

Fig.47 Structure of VSDR[94]

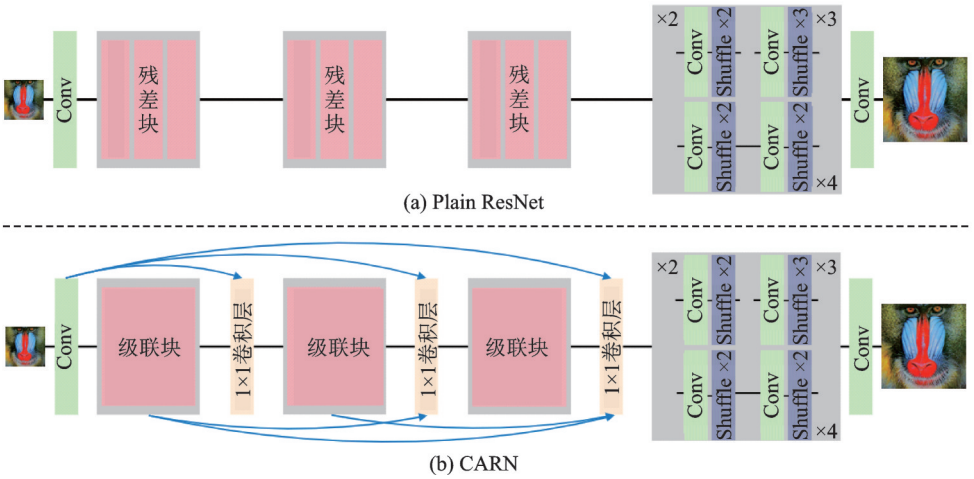

图 48 CARN 对于 ResNet 的改进[95]

Fig.48 Improvement of CARN based on ResNet[95]

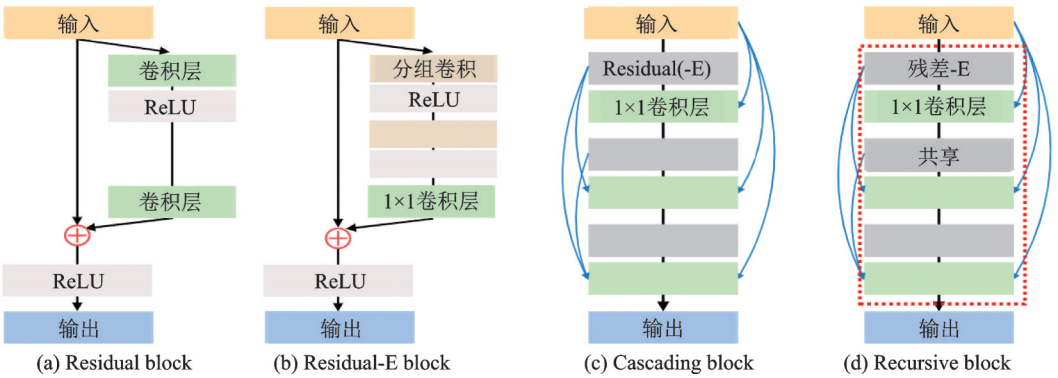

图 49 残差-E 模块与其他常见模块的对比[95]

Fig.49 Comparison between residual-E block and other common blocks[95]

视频修复视作图像修复的延伸,帧间冗余的时间信息并没能被充分利用。EDVR通过增强的可变形卷积网络实现视频的修复和增强,适用于各种视频修复任务,如超分辨率、去模糊等任务。EDVR框架示意图如图50所示。

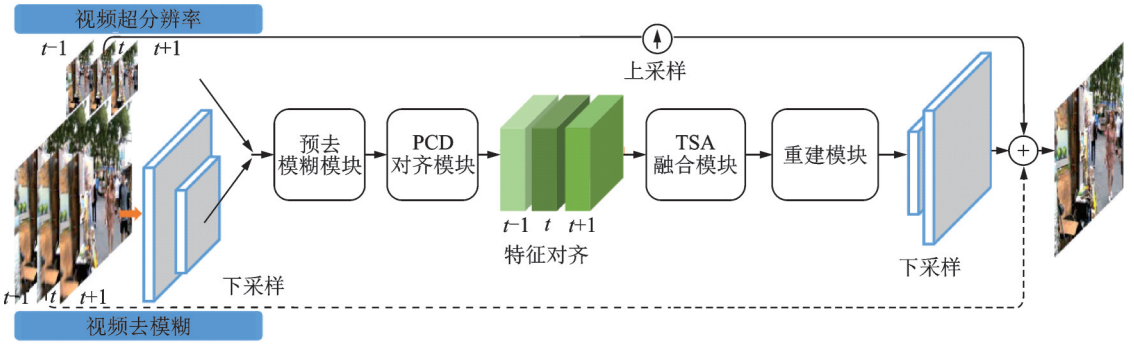

图50 EVDR框架示意图[96]

Fig.50 Pipeline of EDVR[96]

EDVR提出了PCD(Pyramid, cascading and deformable)对齐模块和TSA(Temporal and spatial attention)融合模块,其结构如图51所示。PCD模块受TDAN[97]的启发,用一种金字塔结构从低尺度到高尺度使用可变形卷积将每个相邻帧与参考帧对齐。TSA模块则用于在多个对齐的特征层之间融合信息,通过计算每个相邻帧与参考帧特征之间的元素相关性引入时间注意力机制,相关系数代表每个位置上相邻帧特征信息量的大小。在融合时间特征后进一步应用空间注意力机制,从而更有效地利用跨通道空间信息。

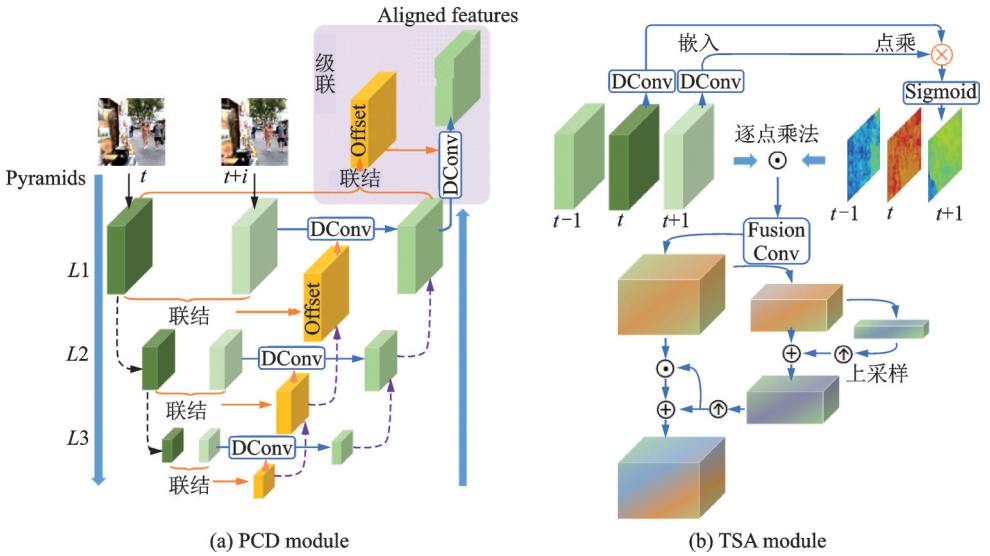

图51 EVDR中的PCD模块和TSA模块[96]

Fig.51 PCD and TSA modules in EVDR[96]

三维卷积是一种常见的利用视频时空间信息的方法,但这种方法往往复杂度较高,限制了模型的深度。2019年提出的FSTRN[98]通过使用一种快速时空间残差模块将三维卷积用于视频超分辨率任务,将每个三维滤波器分解为2个维数更低的3位滤波器乘积,从而降低复杂度,实现更深的网络和更好的性能。此外,FSTRN还提出了一种跨空间残差学习方法,直接连接低分辨率空间和高分辨率空

间,减轻了特征融合和上采样部分的计算负担。FSTRN结构如图52所示。

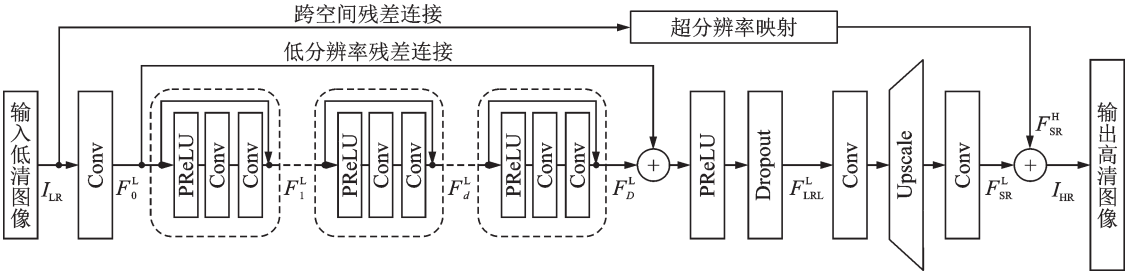

图52 FSTRN结构示意图[98]

Fig.52 Pipeline of FSTRN[98]

随着深度学习技术的兴起,近20年来超分辨率领域发展迅速,出现了很多具有优异性能的模型,但距离实际应用还有一定的距离。图像配准技术对于多帧图像超分辨率的重建效果至关重要,目前还没有成熟的解决方案。另一个难点则是大量未知的密集计算限制了视频超分辨率重建的计算效率,难以达到实时性的要求。超分辨率算法的鲁棒性和可迁移性仍然是下阶段的研究热点,现有的评价标准,如均方误差、峰值噪声比、结构相似性等还不能客观地衡量重建效果,有时甚至会出现和人眼视觉相违背的情况。

4 神经架构搜索

深度学习技术在图像分类、语音识别及机器翻译等诸多领域上取得了举世瞩目的成功,可以自动地学习数据信息,让研究人员摆脱特征工程,这离不开GoogLeNet、ResNet等经典的深度神经网络模型。然而一个具有优异性能的网络结构往往需要花费研究人员大量的时间资金投入,同时需要具备扎实的专业知识和丰富的经验。因此人们开始研究让机器代替人类,根据数据集和算法自动设计网络结构。神经架构搜索技术(Neural architecture search,NAS)设计的模型如今已经在很多任务上取得了超过人工设计深度模型的性能,如图像分割领域的Auto-DeepLab[99],目标检测领域的NAS-FPN[100]。神经架构搜索技术是机器学习自动化(Automated machine learning,AutoML)的子领域,代表了机器学习未来发展的方向。神经架构搜索技术的流程如图53所示,首先从一个搜索空间中通过某种策略搜索候选网络架构,然后对其精度、速度等指标进行评估,通过迭代不断优化直到找到最优的网络架构。

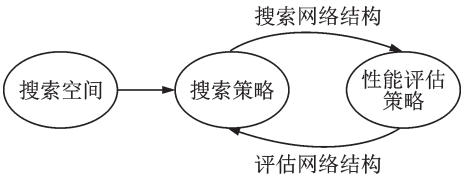

图53 神经架构搜索流程图

Fig.53 Pipeline of NAS

搜索空间内定义了优化问题的变量,如网络架构参数和超参数,这些变量决定了模型的性能。常见的网络架构有链式结构和分支结构等,每一个节点的网络架构参数包括卷积层、池化层和激活函数等,超参数包括卷积的尺寸、步长、加法或连结等。典型的网络架构[101]如图54所示。

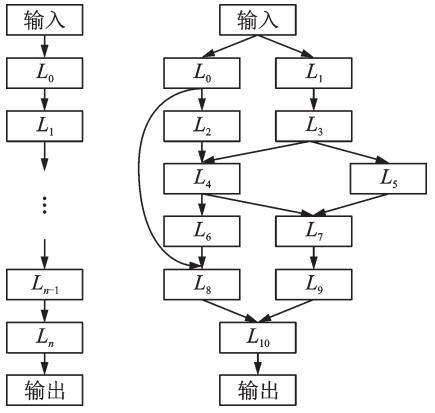

图54 网络架构[101]

Fig.54 Network architecture[101]

搜索策略被用于探索神经架构空间,常见的策略包括随机搜索、贝叶斯优化、遗传算法、强化学习[102-103]和梯度算法等,其中强化学习、遗传算法及梯度算法是目前主流

的搜索策略。在性能评估时,由于训练和验证的时间成本较高,因此常常需要采用评估策略降低评估成本,如减少迭代次数、在训练集的子集上训练、减少卷积核数量等,但这些策略往往会导致一些偏差,可能会对最终的优劣次序产生影响。更高级的策略包括权重共享、通过迭代时的表现推断最终性能以及通过模块预测网络性能等方法。

DARTS[104]是第一个基于连续松弛的搜索空间的神经网络架构技术。早期传统的NAS方法如NasNet[105]、PNAS[106]和ENAS[107]等大多在离散不可微的搜索空间上应用强化学习、进化算法等搜索策略,由于搜索空间内待搜索的参数不可导,因此一个性能优异的模型往往需要耗费大量的计算资源和时间成本。事实上,当时的研究者们将神经架构搜索技术视为一个在离散空间上的黑箱优化问题,每次架构的迭代优化都需要性能评估,效率十分低下。而DARTS使用了松弛连续的结构表示,使用梯度下降优化网络在验证集上的性能,实现了端到端的网络搜索,大大减少了迭代次数,把搜索时间从数千个GPU日降低到数个GPU日。

DARTS流程如图55所示。其中:图(a)表示边上的初始未知操作;图(b)在每条边上放置候选操作的组合,连续松弛搜索空间,不断放宽搜索条件;图(c)通过解决一个双层规划问题联合优化混合概率与网络权重;图(d)用学到的混合概率求得最终的网络架构。DARTS是一种简单的NAS方法,适用于CNN和RNN,在CIFAR-10数据集[108]上用4个GPU日达到了2.76%的测试误差,参数量仅有330万个;在PTB数据集[109]上用1个GPU日以2 300万个的参数量达到了55.7%的测试困惑度,达到了当时的最好性能。在CIFAR-10数据集上搜索出来的模型架构在ImageNet[19]数据集上以470万个的参数量达到8.7%的top-5错误率,在PTB数据集上搜索出来的模型架构在WikiText-2数据集[110]上以3 300万个的参数量达到69.6%的困惑度,优于很多手工设计的轻量化模型。

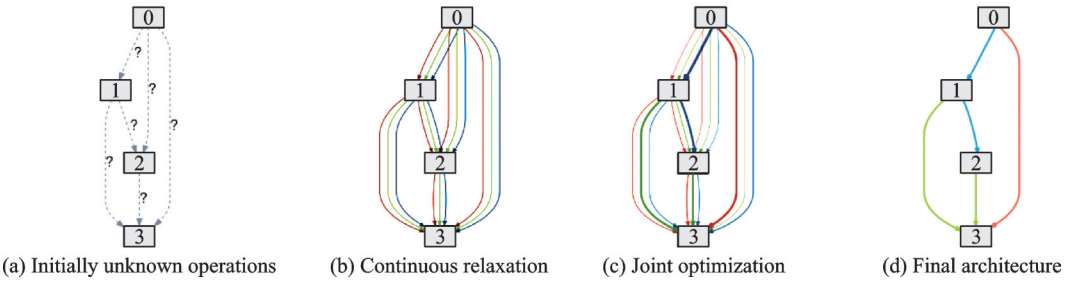

图55 DARTS流程示意图[104]

Fig.55 Pipeline of DARTS[104]

基于DARTS,一系列改进算法被相继提出。在DARTS中,搜索在一个有8个单元的网络上进行,搜索出来的架构通过堆叠在一个具有20个单元的网络上被评估,但深度网络和浅层网络的结构往往不同。例如,在代理数据集(如CIFAR-10数据集)上搜索出来的网络架构可能在目标数据集(如ImageNet数据集)上表现不理想。2019年华为诺亚方舟实验室提出P-DARTS[111],针对这一问题(被称为Depth Gap)提出了一种渐进式搜索的方法,如图56所示。搜索网络的深度从最初的5个单元增加到中期的11个和后期的17个,而候选操作的数量(用不同的颜色表示)相应地从5个减少到4个和2个。在上一阶段得分最低的操作将被丢弃,最后结合分数和可能的附加规则确定最终架构[111]。

2019年MIT提出ProxylessNAS[112],针对DARTS只能在小型代理数据集上搜索而在大型数据集上则会出现显存爆炸的问题提出了无代理神经架构搜索技术,在训练时二值化路径,用和DARTS双层规划类似的思想联合训练权重参数和架构参数,从而达到降低显存的目的,并首次提出针对不同的硬件平台搜索满足特定时延的神经网络架构方法。ProxylessNAS不再采用搜索单元然后堆叠达到更深网络的方法,而是选择主干网络,如MobileNet[41]、ShuffleNet[42]等。ProxylessNAS在CIFAR-10数据集

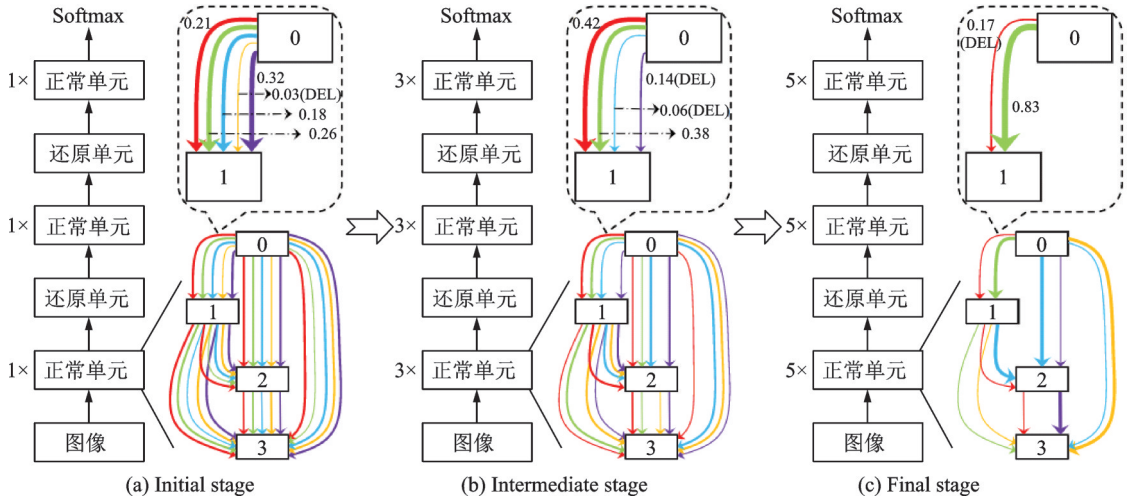

图 56 P-DARTS 流程示意图[111]

Fig.56 Pipeline of P-DARTS[111]

上以仅 570 万个的参数量达到 2.08% 的测试误差。ProxylessNAS 示意图如图 57 所示。

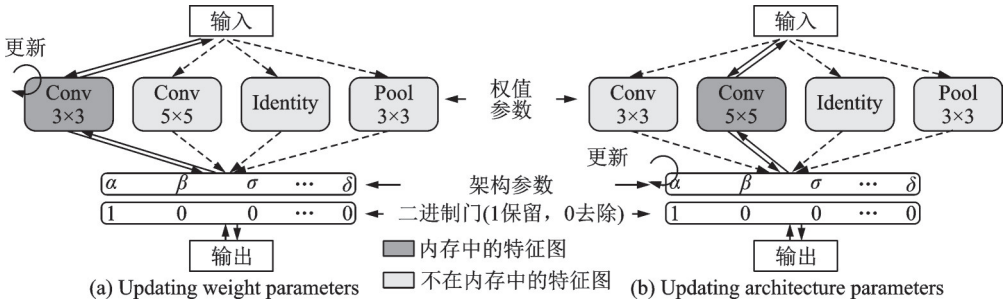

图 57 ProxylessNAS 示意图[112]

Fig.57 Pipeline of ProxylessNAS[112]

当迭代次数过大后，DARTS 设计出的网络架构会包含很多跳接结构，使得性能变得很差，称为 DARTS 的坍塌。2020 年诺亚方舟实验室提出的 DARTS+[113]通过引入早停机制，即当一个正常单元出现 2 个或 2 个以上的跳接结构时就停止搜索，缩短了 DARTS 搜索的时间，极大地提高了 DARTS 的性能，其示意图如图 58 所示。

2020 年商汤研究院提出的随机神经架构搜索 SNAS[114]也是一种可微的端到端方法，但与 DARTS 相比，SNAS 将 NAS 重新表述为在一个单元中搜索空间的联合分布参数优化问题，直接优化损失函数，偏差更小。在同一轮反向传播中 SNAS 同时训练操作参数和架构参数，并提出了一种新的搜索梯度。相比基于强化学习的神经架构搜索技术，SNAS 优化相同的目标函数，但更高效地只使用训练损失作为奖励。

PC-DARTS[115]是华为诺亚方舟实验室 2020 年提出的 NAS 技术，在 P-DARTS[111]的基础上设计了部分通道连接机制，每次只有一部分通道进行操作搜索，这节省了训练需要的显存，减少了计算量，并采用边正则化降低由于操作搜索不全造成的不确定性。PC-DARTS 在 CIFAR-10 数据集[108]上用 0.1 个 GPU 日达到了 2.57% 的测试误差，参数量仅有 360 万个；在 ImageNet 数据集[19]上用 3.8 个 GPU 日以 530 万个的参数量达到了 7.3% 的 top-5 错误率，取得了更快更好的搜索效果。PC-DARTS 结构如图 59

所示。

当前的神经架构搜索技术大多被用于图像分类任务,这促使许多研究人员试图设计出更好的人工网络。但一方面由于搜索空间的定义被局限在现有的网络结构设计经验中,使得NAS设计出的网络很难与人工网络有本质上的区别。另一方面,NAS技术设计的网络可解释性很差,由于研究人员采用的数据增强、搜索空间、训练方法及正则化策略等方法常常不同,这使得NAS设计出的架构很难被复现,不同网络架构的性能也难以比较。由此可见,神经架构搜索领域仍然存在很多挑战,如何解决这些问题将会是下一阶段的热门研究方向之一。

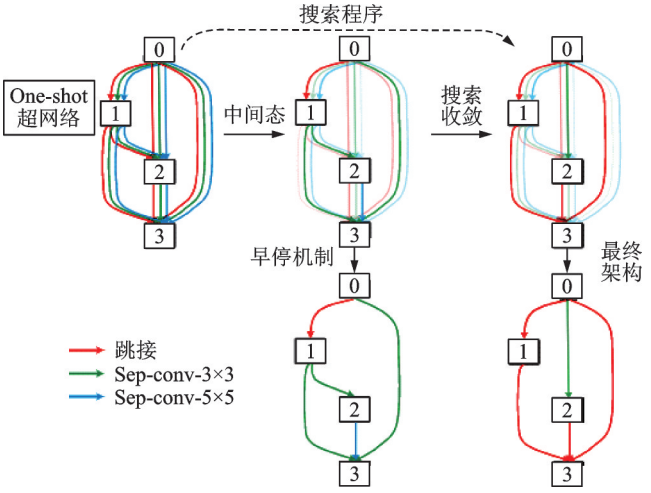

图58 DARTS+中的早停机制示意图[113]

Fig.58 Early Stopping in DARTS+[113]

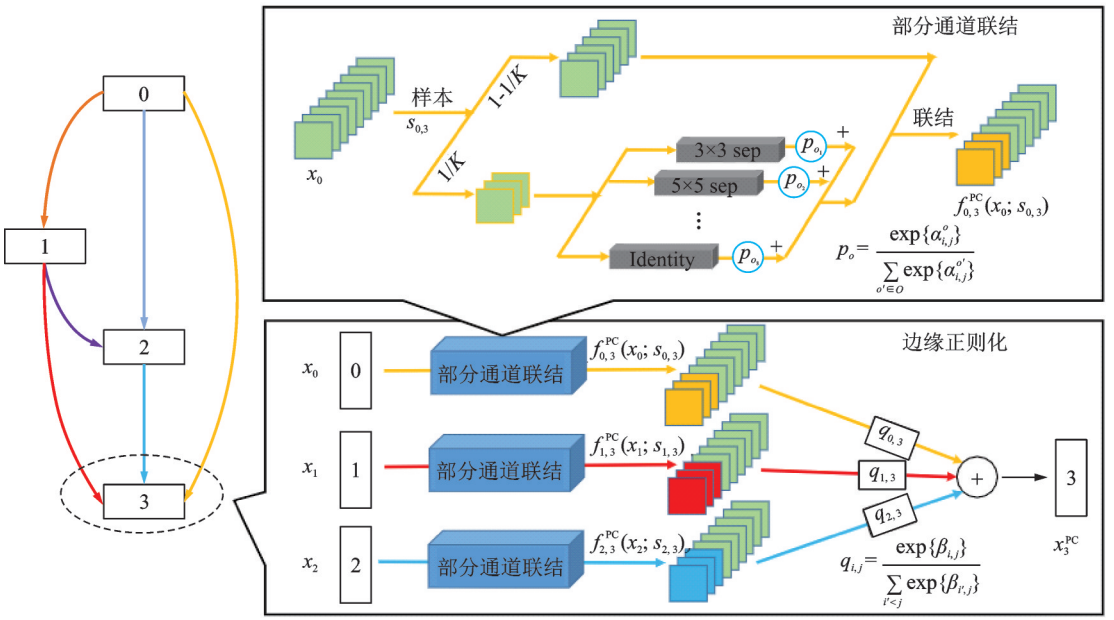

图59 PC-DARTS结构示意图[115]

Fig.59 Structure of PC-DARTS[115]

5 结束语

深度学习技术近年来在计算机视觉中的目标检测、图像分割、超分辨率和模型压缩等任务上都取得了卓越的成绩,充分证明了它的价值和潜力。然而深度学习领域仍然有不少难题无法解决,如对数据的依赖性强、模型难以在不同领域之间直接迁移、深度学习模型的可解释性不强等,如何攻克这些难题将是下一阶段的发展方向。为了追求极致的性能,很多科技巨头投入了巨大的人力财力搭建巨型模型,如OpenAI发布的拥有1 750亿个参数的GPT-3,谷歌发布的拥有1.6万亿个参数的Switch Transformer,快手发布的拥有1.9万亿个参数的推荐精排模型,这些模型需要大量的训练时间和计算资源,如

何设计计算硬件、系统和算法来加速训练是一项新的挑战。深度学习技术严重依赖大规模带标签的数据集,因此无监督学习技术、自监督技术,例如表示学习、预训练模型等,仍然是重要的研究方向。同时深度学习技术带来的安全隐患也引起了重视,如何在保护用户隐私的前提下优化分布式训练是另一个具有潜力的研究方向。

参考文献:

[1] 卢宏涛, 张秦川. 深度卷积神经网络在计算机视觉中的应用研究综述[J]. 数据采集与处理, 2016, 31(1): 1-17.
LU Hongtao, ZHANG Qinchuan. Applications of deep convolutional neural network in computer vision[J]. Journal of Data Acquisition and Processing, 2016, 31(1): 1-17.

[2] DECHTER R. Learning while searching in constraint-satisfaction problems[C]//AIAA-86 Proceedings. [S.l.]: AIAA, 1986: 179-183.

[3] AIZENBERG I, AIZENBERG N, BUTAKOV C, et al. Image recognition on the neural network based on multi-valued neurons[C]//Proceedings of the 15th International Conference on Pattern Recognition. [S.l.]: IEEE, 2000: 989-992.

[4] SCHMIDHUBER J. Deep learning in neural networks: An overview[J]. Neural Networks, 2015, 61: 85-117.

[5] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[6] DENG L, YU D. Deep learning: Methods and applications[J]. Foundations and Trends in Signal Processing, 2014, 7(3/4): 197-387.

[7] BENGIO Y. Learning deep architectures for AI[M]. Boston: Now Publishers Inc. 2009.

[8] HINTON G E. What kind of graphical model is the brain?[C]//Proceedings of the 19th International Joint Conference on Artificial Intelligence. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005: 1765-1775.

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

[10] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18 (7): 1527-1554.

[11] DAHL G E, YU D, DENG L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. IEEE Transactions on Audio, Speech, and language Processing, 2011, 20(1): 30-42.

[12] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[13] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25: 1097-1105.

[14] LECUN Y, BOSER B, DENKER J, et al. Handwritten digit recognition with a back-propagation network[J]. Advances in Neural Information Processing Systems, 1989, 2: 396-404.

[15] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86: 2278-2324.

[16] LOWE D G.Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60 (2): 91-110.

[17] DALAI N, TRIGGS B.Histograms of oriented gradients for human detection[C]//Proceedings of Computer Vision and Pattern Recognition(CVPR). San Diego, USA: IEEE, 2005, 1: 886-893.

[18] BAY H, TUYTELAARS T, VAN GOOL L . SURF: Speeded up robust features[C]//Proceedings of European Conference on Computer Vision. Berlin, Heidelberg: Springer, 2006: 404-417.

[19] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2009: 248-255.

[20] OJALA T, PIETIKAINEN M, HARWOOD D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions[C]//Proceedings of the 12th International Conference on Pattern Recognition. [S.l.]: IEEE, 1994, 1: 582-585.

[21] OJALA T, PIETIKÄINEN M, HARWOOD D. A comparative study of texture measures with classification based on featured distributions[J]. Pattern Recognition, 1996, 29(1): 51-59.

[22] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]//Proceedings of European Conference on Computer Vision. Berlin, Heidelberg: Springer, 2014: 818-833.

[23] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2015: 1-9.
[24] LIN M, CHEN Q, YAN S. Network in network[EB/OL].(2013-03-08)[2022-01-20]. https://arxiv.org/abs/1312.4400.
[25] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[EB/OL]. (2015-03-13)[2022-01-20]. http://arxiv.org/asb/1502.03167.
[26] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 2818-2826.
[27] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2017.
[28] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2022-01-20]. https://arxiv.org/abs/1409.1556.
[29] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 770-778.
[30] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 4700-4708.
[31] XIE S, GIRSHICK R, DOLLÁR P, et al. Aggregated residual transformations for deep neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 1492-1500.
[32] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 1251-1258.
[33] KUZNETSOVA A, ROM H, ALLDRIN N, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale[J]. International Journal of Computer Vision, 2018. DOI: 10.1007/s11263-020-01316-z.
[34] SUN C, SHRIVASTAVA A, SINGH S, et al. Revisiting unreasonable effectiveness of data in deep learning era[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2017: 843-852.
[35] CARREIRA J, NOLAND E, BANKI-HORVATH A, et al. A short note about kinetics-600[EB/OL]. (2018-08-03)[2022-01-20]. https://arxiv.org/abs/1808.01340v1.
[36] SMAIRA L, CARREIRA J, NOLAND E, et al. A short note on the kinetics-700-2020 human action dataset[EB/OL]. (2020-10-21)[2022-01-20]. https://arxiv.org/abs/2010.10864.
[37] KINGMA D P, WELLING M. Auto-encoding variational bayes[EB/OL]. (2014-05-01)[2022-01-20]. https://arxiv.org/abs/1312.6114.
[38] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014, 27: 2672-2680.
[39] YU X, ZHANG X, CAO Y, et al. VAEGAN: A collaborative filtering framework based on adversarial variational autoencoders[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. [S.l.]: ACM, 2019: 4206-4212.
[40] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size[EB/OL]. (2016-11-04)[2022-01-20]. https://arxiv.org/abs/1602.07360.
[41] HOWARD A G, ZHU M, CHEN B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications [EB/OL]. (2017-04-17)[2022-01-20]. https://arxiv.org/abs/1704.04861.
[42] ZHANG X, ZHOU X, LIN M, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 6848-6856.
[43] SANDLER M, HOWARD A, ZHU M, et al. MobileNet v2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE Conference on computer vision and Pattern Recognition. [S.l.]: IEEE, 2018: 4510-4520.
[44] MA N, ZHANG X, ZHENG H T, et al. ShuffleNet v2: Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: Springer, 2018: 116-131.
[45] HAN K, WANG Y, TIAN Q, et al. GhostNet: More features from cheap operations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 1580-1589.
[46] LI H, KADAV A, DURDANOVIC I, et al. Pruning filters for efficient convnets[EB/OL]. (2017-03-10)[2022-01-20]. https://arxiv.org/abs/1608.08710.
[47] HITCHCOCK F L. The expression of a tensor or a polyadic as a sum of products[J]. Journal of Mathematics and Physics,

1927, 6(1/2/3/4): 164-189.

[48] TUCKER L R. Some mathematical notes on three-mode factor analysis[J]. Psychometrika, 1966, 31(3): 279-311.

[49] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. (2015-03-09)[2022-01-20]. https://arxiv.org/abs/1503.02531.

[50] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2014: 580-587.

[51] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2015: 1440-1448.

[52] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[53] DAI J, LI Y, HE K, et al. R-FCN: Object detection via region-based fully convolutional networks[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. [S.l.]: ACM, 2016: 379-387.

[54] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37: 1904-1916.

[55] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2016: 21-37.

[56] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 2117-2125.

[57] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 779-788.

[58] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 7263-7271.

[59] REDMON J, FARHADI A. YOLOV3: An incremental improvement[EB/OL]. (2018-04-08)[2022-01-20]. https://arxiv.org/abs/1804.02767.

[60] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2017: 2980-2988.

[61] TIAN Z, SHEN C, CHEN H, et al. FCOS: Fully convolutional one-stage object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2019: 9627-9636.

[62] ZHANG H, WANG Y, DAYOUB F, et al. VarifocalNet: An iou-aware dense object detector[EB/OL]. (2021-05-04)[2022-01-20]. https://arxiv.org/abs/2008.13367.

[63] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2015: 234-241.

[64] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2015: 3431-3440.

[65] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2017: 2961-2969.

[66] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs[EB/OL]. (2016-06-07)[2022-01-20]. https://arxiv.org/abs/1412.7062.

[67] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[68] BROSTOW G J, FAUQUEUR J, CIPOLLA R. Semantic object classes in video: A high-definition ground truth database[J]. Pattern Recognition Letters, 2009, 30(2): 88-97.

[69] SONG S, LICHTENBERG S P, XIAO J. Sun RGB-D: A RGB-D scene understanding benchmark suite[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2015: 567-576.

[70] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 2881-2890.

[71] 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, 2017, 40(4): 834-848.

[72] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-12-05)[2022-01-20]. https://arxiv.org/abs/1706.05587.
[73] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European Conference on Computer vision (ECCV). [S.l.]: Springer, 2018: 801-818.
[74] LI H, XIONG P, FAN H, et al. DFANet: Deep feature aggregation for real-time semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 9522-9531.
[75] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 3213-3223.
[76] JIANG W, XIE Z, LI Y, et al. LRNNet: A light-weighted network with efficient reduced non-local operation for real-time semantic segmentation[C]//Proceedings of 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). [S.l.]: IEEE, 2020: 1-6.
[77] PAPANDREOU G, CHEN L C, MURPHY K P, et al. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2015: 1742-1750.
[78] OH S J, BENENSON R, KHOREVA A, et al. Exploiting saliency for object segmentation from image level labels[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2017: 5038-5047.
[79] AHN J, KWAK S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 4981-4990.
[80] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2014: 740-755.
[81] EVERINGHAM M, ESLAMI S M A, VAN GOOL L, et al. The pascal visual object classes challenge: A retrospective[J]. International Journal of Computer Vision, 2015, 111(1): 98-136.
[82] DI FRANCIA G T. Super-gain antennas and optical resolving power[J]. IL Nuovo Cimento (1943—1954), 1952, 9(3): 426-438.
[83] HARRIS J L. Diffraction and resolving power[J]. JOSA, 1964, 54(7): 931-936.
[84] GOODMAN J W. On the origin of peritoneal fluid cells[J]. Blood, 1964, 23(1): 18-26.
[85] TSAI R Y, HUANG T S. Multiframe image restoration and registration[J]. Advance Computer Visual and Image Processing, 1984, 1: 317-339.
[86] GAO X, ZHANG K, TAO D, et al. Image super-resolution with sparse neighbor embedding[J]. IEEE Transactions on Image Processing, 2012, 21(7): 3194-3205.
[87] YANG J, WRIGHT J, HUANG T, et al. Image super-resolution as sparse representation of raw image patches[C]//Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2008: 1-8.
[88] YANG J, WRIGHT J, HUANG T S, et al. Image super-resolution via sparse representation[J]. IEEE Transactions on Image Processing, 2010, 19(11): 2861-2873.
[89] 苏衡, 周杰, 张志浩 . 超分辨率图像重建方法综述[J]. 自动化学报, 2013, 39(8): 1202-1213.
SU Heng, ZHOU Jie, ZHANG Zhihao. Survey of super-resolution image reconstruction methods[J]. Acta Automatica Sinica, 2013, 39(8): 1202-1213.
[90] BASHIR S M A, WANG Y, KHAN M, et al. A comprehensive review of deep learning-based single image super-resolution [EB/OL]. (2021-07-13)[2022-01-20]. https://arxiv.org/abs/2102.09351.
[91] DONG C, LOY C C, HE K, et al. Image super-resolution using deep convolutional networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38: 295-307.
[92] DONG C, LOY C C, TANG X. Accelerating the super-resolution convolutional neural network[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2016: 391-407.
[93] SHI W, CABALLERO J, HUSZÁR F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 1874-1883.
[94] KIM J, LEE J K, LEE K M. Accurate image super-resolution using very deep convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 1646-1654.
[95] AHN N, KANG B, SOHN K A. Fast, accurate, and lightweight super-resolution with cascading residual network[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: Springer, 2018: 252-268.

[96] WANG X, CHAN K C K, YU K, et al. EDVR: Video restoration with enhanced deformable convolutional networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. [S.l.]: IEEE, 2019.

[97] TIAN Y, ZHANG Y, FU Y, et al. TDAN: Temporally-deformable alignment network for video super-resolution[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 3360-3369.

[98] LI S, HE F, DU B, et al. Fast spatio-temporal residual network for video super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 10522-10531.

[99] LIU C, CHEN L C, SCHROFF F, et al. Auto-DeepLab: Hierarchical neural architecture search for semantic image segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 82-92.

[100] GHIASI G, LIN T Y, LE Q V. NAS-FPN: Learning scalable feature Pyramid architecture for object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 7036-7045.

[101] ELSKEN T, METZEN J H, HUTTER F. Neural architecture search: A survey[J]. The Journal of Machine Learning Research, 2019, 20(1): 1997-2017.

[102] BAKER B, GUPTA O, NAIK N, et al. Designing neural network architectures using reinforcement learning[EB/OL]. (2017-05-22)[2022-01-20]. https://arxiv.org/abs/1611.02167.

[103] ZOPH B, LE Q V. Neural architecture search with reinforcement learning[EB/OL]. (2017-11-13)[2022-01-20]. https://arxiv.org/abs/1611.01578.

[104] LIU H, SIMONYAN K, YANG Y. DARTS: Differentiable architecture search[EB/OL]. (2019-04-23)[2022-01-20]. https://arxiv.org/abs/1806.09055v1.

[105] ZOPH B, VASUDEVAN V, SHLENS J, et al. Learning transferable architectures for scalable image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 8697-8710.

[106] LIU C, ZOPH B, NEUMANN M, et al. Progressive neural architecture search[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: Springer, 2018: 19-34.

[107] PHAM H, GUAN M, ZOPH B, et al. Efficient neural architecture search via parameters sharing[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2018: 4095-4104.

[108] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[EB/OL]. [2022-01-20]. https://doi.org/10.1.1.222.9220.

[109] GOLDBERGER A L, AMARAL L A N, GLASS L, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals[J]. Circulation, 2000, 101(23): e215-e220.

[110] MERITY S, XIONG C, BRADBURY J, et al. Pointer sentinel mixture models[EB/OL]. (2016-09-26)[2022-01-20]. https://arxiv.org/abs/1609.07843.

[111] CHEN X, XIE L, WU J, et al. Progressive differentiable architecture search: Bridging the depth gap between search and evaluation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2019: 1294-1303.

[112] CAI H, ZHU L, HAN S. ProxylessNAS: Direct neural architecture search on target task and hardware[EB/OL]. (2019-02-23)[2022-01-20]. https://arxiv.org/abs/1812.00332.

[113] LIANG H, ZHANG S, SUN J, et al. DARTS+: Improved differentiable architecture search with early stopping[EB/OL]. (2020-10-20)[2022-01-20]. https://arxiv.org/abs/1909.06035v1.

[114] XIE S, ZHENG H, LIU C, et al. SNAS: Stochastic neural architecture search[EB/OL]. (2020-04-01)[2022-01-20]. https://arxiv.org/abs/1812.09926.

[115] XU Y, XIE L, ZHANG X, et al. PC-DARTS: Partial channel connections for memory-efficient architecture search[EB/OL]. (2020-04-07)[2022-01-20]. https://arxiv.org/abs/1907.05737v1.

作者简介:

卢宏涛(1967-),**通信作者**,男,长聘教授,博士生导师,研究方向:机器学习、深度学习和计算机视觉,E-mail:htlu@sjtu.edu.cn。

罗沐昆(1998-),男,硕士研究生,研究方向:深度学习、目标检测和弱监督。

(编辑:张黄群)