计算机视觉经典综述 / 03

基于深度学习的图像分割综述

Image Segmentation Using Deep Learning: A Survey

Shervin Minaee（IEEE 会员）、Yuri Boykov（IEEE 会员）、Fatih Porikli（IEEE 会士）、Antonio Plaza（IEEE 会士）、Nasser Kehtarnavaz（IEEE 会士）和 Demetri Terzopoulos（IEEE 会士）

IEEE Transactions on Pattern Analysis and Machine Intelligence

第 44 卷第 7 期 · 2022 年 7 月 · 3523–3542 页 · DOI 10.1109/TPAMI.2021.3059968

摘要

图像分割是计算机视觉和图像处理中的一项关键任务，在场景理解、医学图像分析、机器人感知、视频监控、增强现实和图像压缩等方面有重要应用。现有文献已提出大量分割算法。在此背景下，深度学习（DL）的广泛成功推动了新的图像分割方法的发展，这些方法利用深度学习模型完成分割。本文全面回顾这一时期的文献，涵盖语义分割和实例分割中的各类开创性工作，包括基于卷积的像素标注网络、编码器-解码器架构、多尺度与金字塔方法、循环网络、视觉注意力模型，以及在对抗框架下使用生成模型的方法。我们考察这些深度学习分割模型之间的关系、各自优势及所面临的挑战，介绍广泛使用的数据集，对比模型性能，并讨论有前景的研究方向。

索引词　图像分割、深度学习、卷积神经网络、编码器-解码器模型、循环模型、生成模型、语义分割、实例分割、全景分割、医学图像分割

章节导航

摘要 1

1 引言 2

2 深度神经网络架构 3

3 基于深度学习的图像分割模型 5

4 数据集 20

5 深度学习分割模型的性能 23

6 挑战与机遇 29

7 结论 30

致谢 30

参考文献 31

作者简介 39

作者单位与论文信息 40

1 引言

从计算机视觉学科的早期发展起，图像分割就是其中的一个基本问题[1]（第8章）。它把图像（或视频帧）划分为多个区域和对象[2]（第5章），是许多视觉理解系统的重要组成部分，在多种应用中起核心作用[3]（第VI部分），例如医学图像分析（提取肿瘤边界和测量组织体积）、自动驾驶车辆（检测可通行表面和行人）、视频监控及增强现实。

图像分割可以表述为对像素赋予语义类别（语义分割）、把不同对象实例分开（实例分割），或同时完成两者（全景分割）。语义分割从一组对象类别中（如人、汽车、树、天空）为图像中的每个像素赋予标签；因此，它通常比仅为整幅图像预测一个标签的图像分类更难。实例分割进一步扩展了语义分割的范围，要求检测并勾画图像中每个感兴趣对象（例如把每个人分别找出来）。

文献中已有大量图像分割算法。从最早的阈值法[4]、基于直方图的分组、区域生长[5]、k 均值聚类[6]和分水岭法[7]，发展到主动轮廓[8]、图割[9]、条件随机场和马尔可夫随机场[10]以及基于稀疏性的方法[11]、[12]。近年来，深度学习（DL）模型催生了新一代图像分割模型，其性能显著提升，常在常用基准上取得最高准确率（例如图1）。这使该领域的研究范式发生转变。

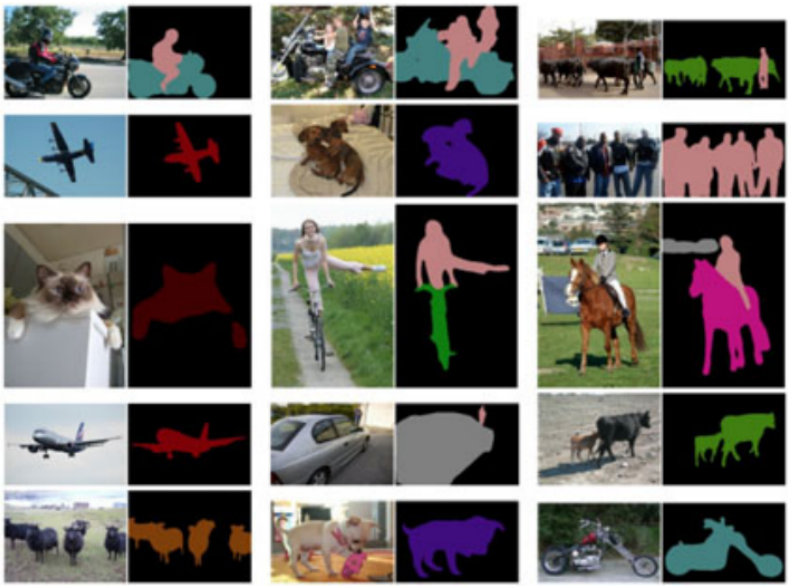

图1　DeepLabV3[13]在示例图像上的分割结果。

本文是文献[14]的修订版，回顾了基于深度学习的图像分割近期研究，涵盖截至撰稿时提出的100多种分割方法。我们从训练数据、网络架构选择、损失函数、训练策略和主要贡献等方面进行全面评述。所综述的文献分为以下几类：

1）全卷积网络
2）结合图模型的卷积模型
3）基于编码器-解码器的模型
4）基于多尺度与金字塔网络的模型
5）基于 R-CNN 的模型（用于实例分割）

6）空洞卷积模型与 DeepLab 系列
7）基于循环神经网络的模型
8）基于注意力的模型
9）生成模型与对抗训练
10）结合主动轮廓模型的卷积模型
11）其他模型

在这一分类框架下：

- 全面回顾和分析基于深度学习的图像分割算法；
- 介绍常用图像分割数据集，分为二维图像和2.5维（RGB-D）图像；
- 汇总所回顾的分割方法在常用基准上的性能；
- 讨论基于深度学习的图像分割面临的若干挑战和未来研究方向。

本文其余部分安排如下：第2节概述许多现代分割算法用作骨干网络的常用深度神经网络（DNN）架构。第3节回顾最具代表性的先进深度学习分割模型。第4节介绍若干最常用的图像分割数据集及其特点。第5节列出用于评估深度学习分割模型的常用指标，并以表格汇总模型性能。第6节讨论深度学习分割方法的主要挑战与机会。第7节给出结论。

2 深度神经网络架构

本节概述计算机视觉研究中常用的几种深度神经网络架构：卷积神经网络、循环神经网络与长短期记忆网络、编码器-解码器与自编码器模型，以及生成对抗网络。限于篇幅，本文不讨论已提出的若干其他架构，包括 Transformer、胶囊网络、门控循环单元和空间变换网络。

2.1 卷积神经网络（CNN）

CNN 是深度学习领域最成功、应用最广的架构之一，尤其适用于计算机视觉任务。Fukushima[16]最早在关于“新认知机”（Neocognitron）的奠基性论文中提出 CNN，其依据是 Hubel 和 Wiesel 关于视觉皮层分层感受野的模型。随后，Waibel 等[17]引入在时间感受野之间共享权重、用反向传播训练的 CNN，用于音素识别；LeCun 等[15]面向文档识别开发了一种实用的 CNN 架构（图2）。CNN 通常包含三类层：i）卷积层，用一组权重构成的核（或滤波器）做卷积以提取特征；ii）非线性层，对特征图施加激活函数（通常逐元素运算），使网络能够对非线性函数建模；iii）池化层，通过以统计量（均值、最大值等）代替特征图中的局部邻域来降低空间分辨率。不同层的神经元之间局部连接，即每个神经元仅接收前一层一个小邻域（即感受野）内神经元的加权输入。把多层堆叠成多分辨率金字塔后，层级越高，学习特征所用的感受野就越大。CNN 在计算上的主要优势是同一层的所有感受野都共享权重，因此参数数量显著少于全连接神经网络。最著名的 CNN 架构包括 AlexNet[18]、VGGNet[19]和 ResNet[20]。

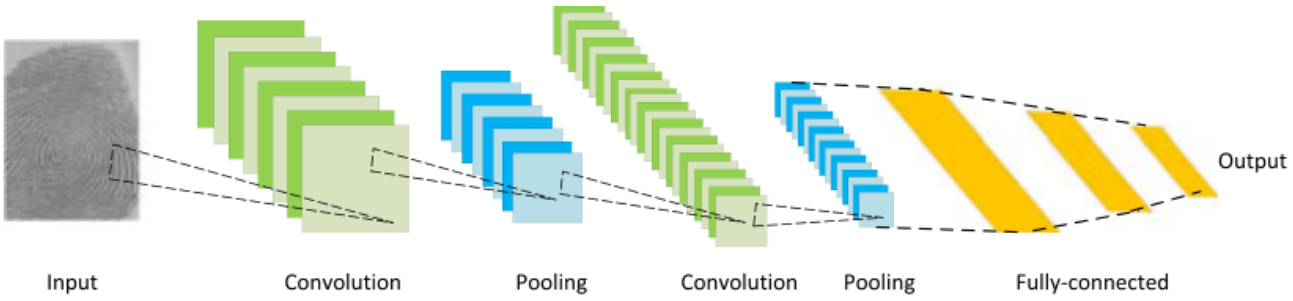

图2 卷积神经网络（CNN）的架构。引自[15]。
Input：输入；Convolution：卷积；Pooling：池化；Fully-connected：全连接；Output：输出。

2.2 循环神经网络（RNN）与长短期记忆网络（LSTM）

RNN[22]常用来处理语音、文本、视频和时间序列等序列数据。如图3所示，在每个时间步 t，模型接收当前输入 x_t 和前一步的隐藏状态 h_{t-1}，输出目标值 o_t 及下一隐藏状态 h_{t+1}。RNN 处理长序列时往往会遇到困难：在许多实际应用中难以捕捉长期依赖关系，也常出现梯度消失或梯度爆炸。长短期记忆网络（LSTM）[23]是为避免这些问题而设计的一类 RNN。LSTM 架构（图4）包含输入门、输出门和遗忘门，用于调节信息进出记忆单元；该单元可在任意长的时间间隔内保存数值。

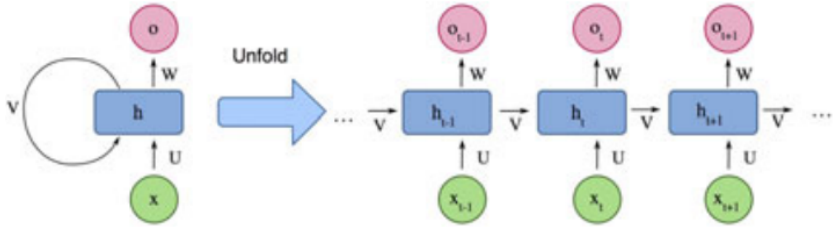

图3 简单循环神经网络（RNN）的架构。Christopher Olah 提供[21]。
Unfold：按时间展开；x、h、o 分别表示输入、隐藏状态和输出。

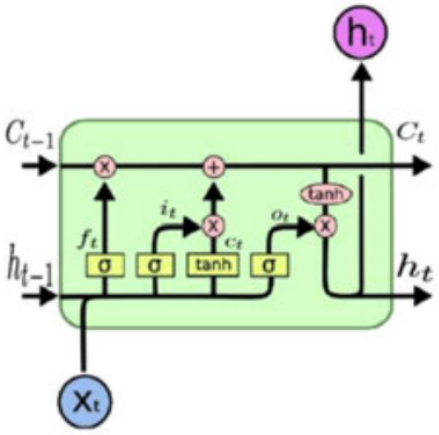

图4 标准 LSTM 模块的架构。Olah 提供[21]。

2.3 编码器-解码器与自编码器模型

编码器-解码器[24]、[25]是一类用两阶段网络学习从输入域到输出域映射的模型（图5）。编码器用编码函数 z = g(x) 把输入 x 压缩成潜在空间表示 z；解码器则用 y = f(z) 根据 z 预测输出 y。潜在表示，也称特征（向量）表示，保留了输入中有助于预测输出的语义信息。这类模型常用于自然语言处理（NLP）的序列到序列建模，也用于图像到图像的转换；在后者中，输出可以是增强后的图像（例如去模糊或超分辨率结果），也可以是分割图。自编码器是编码器-解码器模型的一种特殊情况，其输入和输出相同。

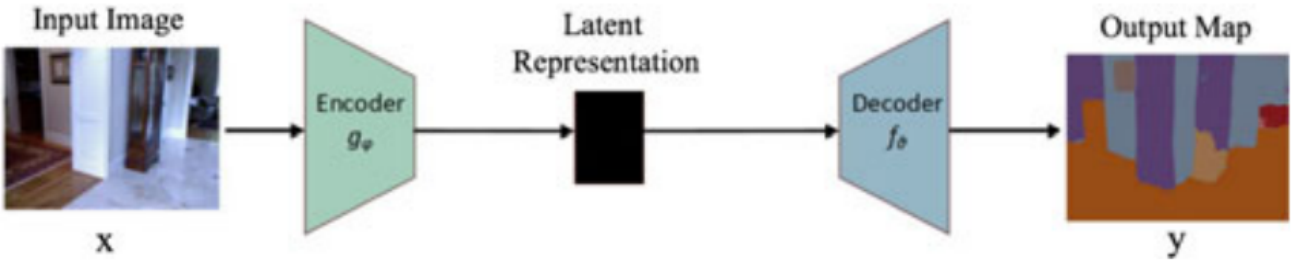

图5　简单编码器-解码器模型的架构。
从左到右：输入图像、编码器、潜在表示、解码器、输出图。

2.4 生成对抗网络（GAN）

GAN[26]是一类相对较新的深度学习模型，由生成器和判别器两个网络组成（图6）。在传统 GAN 中，生成器网络 G 学习把具有先验分布的噪声 z 映射到接近真实样本分布的目标分布 y。判别器网络 D 尝试区分生成的“假”样本与真实样本。GAN 可以看作 G 和 D 之间的一场极小极大博弈：D 试图降低辨别真假样本时的分类误差，从而最大化损失函数；G 则试图让判别器的误差最大化，从而最小化该损失函数。GAN 的变体包括卷积 GAN[27]、条件 GAN[28]和 Wasserstein GAN[29]。

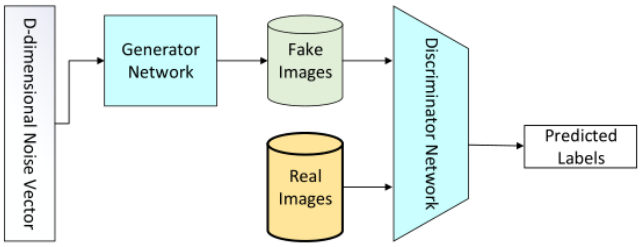

图6　生成对抗网络（GAN）的架构。Ian Goodfellow 提供。
噪声向量经生成器产生假图像，判别器接收假图像和真图像，输出预测标签。

3 基于深度学习的图像分割模型

本节回顾多种基于学习的分割方法，按模型架构归为10类。许多方法共享若干架构特征，如编码器和解码器、跳跃连接、多尺度架构，以及较晚出现的空洞卷积。根据各模型相对于既有模型的架构贡献对其归类，是一种方便的组织方式。

3.1 全卷积模型

Long 等[30]提出的全卷积网络（FCN）是深度学习语义图像分割模型的一座里程碑。FCN（图7）只包含卷积层，可以输出与输入图像尺寸相同的分割图。为了处理任意尺寸的图像，作者修改了 VGG16、GoogLeNet 等现有 CNN 架构，删去所有全连接层，使模型输出空间分割图而不是分类得分。

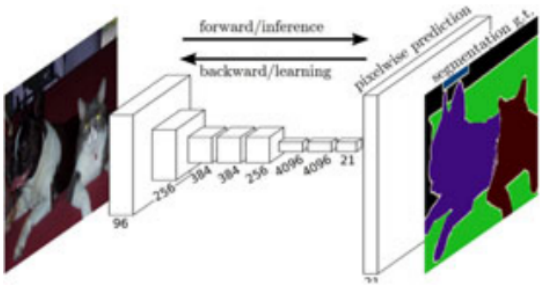

图7　全卷积网络（FCN）学习进行精确的逐像素预测。引自[30]。
Forward/inference：前向 / 推理；Backward/learning：反向 / 学习；Pixelwise prediction：逐像素预测。

FCN 利用跳跃连接（图8），将模型末端层的特征图上采样，并与前面层的特征图融合。这样就能结合来自深层、低分辨率层的语义信息与来自浅层、高分辨率层的外观信息，从而获得准确且细致的分割结果。在 PASCAL VOC、NYUDv2 和 SIFT Flow 上测试时，该模型达到了当时最先进的分割性能。

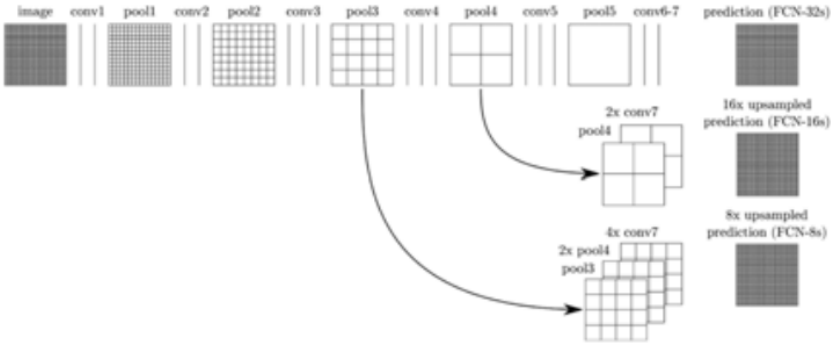

图8　跳跃连接融合粗粒度和细粒度信息。引自[30]。

FCN 已用于多种分割问题，例如脑肿瘤分割[31]、实例感知语义分割[32]、皮肤病灶分割[33]和虹膜分割[34]。它表明，深度神经网络可以通过端到端训练对不同尺寸的图像进行语义分割。不过，传统 FCN 有一些局限：计算开销太大，难以用于实时推理；无法有效纳入全局上下文信息；也不易推广到三维图像。一些研究尝试克服这些局限。例如，Liu 等[35]提出 ParseNet（图9），利用某层的平均特征增强各位置处的特征，把全局上下文引入 FCN。做法是对该层的特征图在整幅图像上池化，得到一个上下文向量，再把该向量归一化并反池化，生成与原特征图同样大小的新特征图。随后将新、旧特征图拼接；这相当于用所述模块替代 FCN 的卷积层（图9e）。

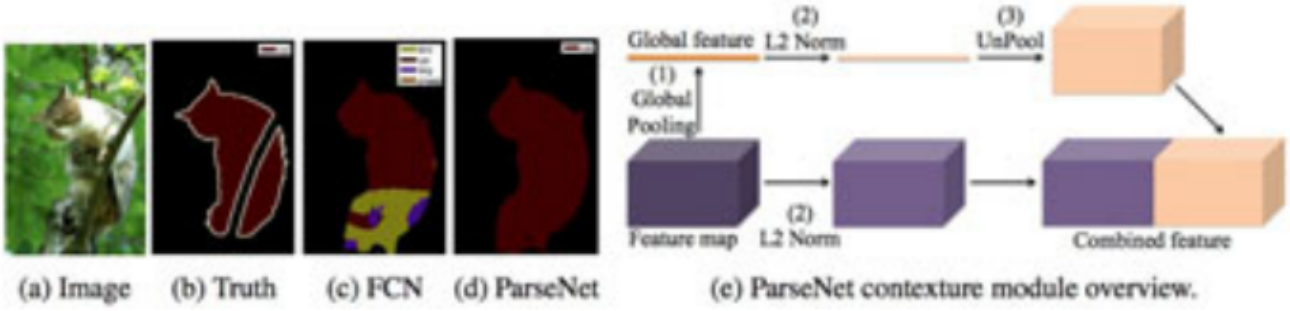

图9 ParseNet 的结构见 (e)。它引入额外的全局上下文，得到的分割结果 (d) 比 FCN 的结果 (c) 更平滑。引自[35]。
图中 Image：原图；Truth：真值；Global Average Pooling：全局平均池化；L2 Norm：L2 归一化；UnPool：反池化。

3.2 结合图模型的 CNN

如前所述，FCN 忽略了可能有用的场景级语义上下文。为了利用更多上下文，一些方法把条件随机场（CRF）和马尔可夫随机场（MRF）等概率图模型纳入深度学习架构。

不变性使 CNN 擅长分类等高层任务，却也意味着深层 CNN 后部各层的响应定位不够精确，因而不利于准确分割对象。为弥补这一不足，Chen 等[36]提出将 CNN 与全连接 CRF 相结合的语义分割算法（图10）。结果表明，该模型定位分割边界的准确率高于此前的方法。

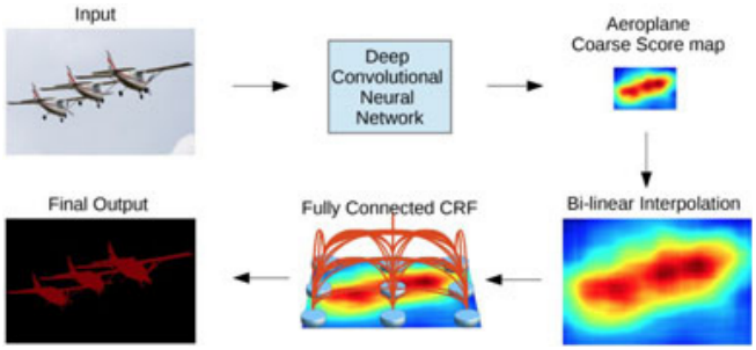

图10 结合 CNN 与 CRF 的模型。引自[36]。
Coarse Score Map：粗分辨率得分图；Bi-linear Interpolation：双线性插值；Fully Connected CRF：全连接 CRF。

Schwing 和 Urtasun[37]提出一种用于图像分割的全连接深度结构化网络。他们联合训练 CNN 和全连接 CRF 来做图像语义分割，并在颇具挑战性的 PASCAL VOC 2012 数据集上取得了令人鼓舞的结果。Zheng 等[38]提出了类似的语义分割方法。在相关工作中，Lin 等[39]提出了一种基于上下文深度 CRF 的高效语义分割模型。他们研究了图像区域之间的“区域-区域”上下文和“区域-背景”上下文，用这些上下文信息改进语义分割。

Liu 等[40]提出一种语义分割算法，把高阶关系和多种标签上下文的混合等丰富信息纳入 MRF。以往方法通常用迭代算法优化 MRF；他们则提出一种名为解析网络（Parsing Network）的 CNN 模型，可在一次前向过程中完成确定性的端到端计算。

3.3 基于编码器-解码器的模型

大多数常用的深度学习分割模型都采用某种形式的编码器-解码器架构。我们将这类模型按用途分为通用图像分割和医学图像分割两类。

3.3.1 通用图像分割

Noh 等[41]引入了基于反卷积（也称转置卷积）的语义分割。其模型 DeConvNet（图11）由两部分组成：采用16层 VGG 网络卷积层的编码器，以及接收特征向量并生成逐像素类别概率图的多层反卷积网络。后者包含反卷积层和反池化层，用来识别逐像素类别标签并预测分割掩膜。

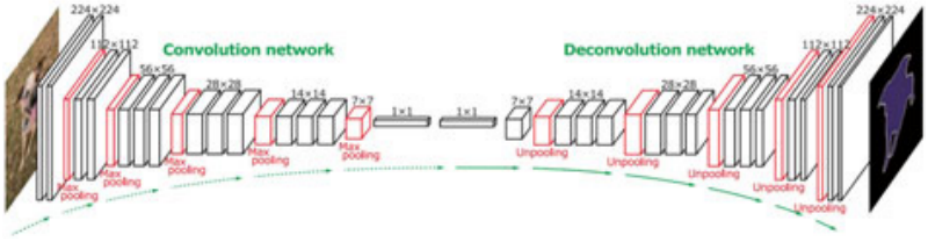

图11　基于反卷积的语义分割。引自[41]。

Badrinarayanan 等[25]提出用于图像分割的全卷积编码器-解码器架构 SegNet（图12）。与上述反卷积网络类似，SegNet 可训练的核心分割模块由编码器、对应的解码器以及其后的逐像素分类层组成；编码器的拓扑结构和 VGG16 的13个卷积层相同。SegNet 的主要创新在于解码器如何对低分辨率的输入特征图执行上采样：它利用对应编码器在最大池化步骤中得到的池化索引进行非线性上采样。

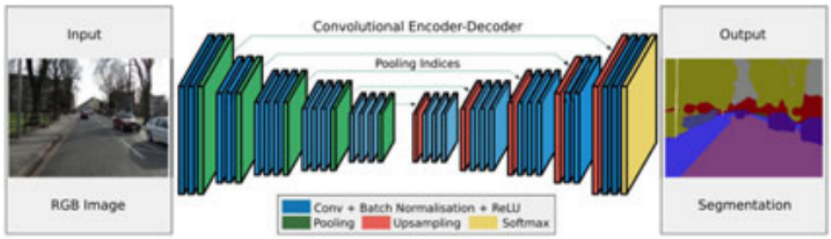

图12　SegNet 模型。引自[25]。
Pooling Indices：池化索引；绿色为池化，红色为上采样，黄色为 Softmax。

基于编码器-解码器的模型有一个局限：编码会降低分辨率，致使图像的细粒度信息丢失。HRNet [42]（图13）解决了这一问题。不同于 DeConvNet、SegNet 等模型靠恢复来获得高分辨率表示，HRNet 把从高分辨率到低分辨率的不同卷积分支并联，并让各分辨率分支反复交换信息，从而在整个编码过程中维持高分辨率表示。HRNet 有四个阶段：第1阶段采用高分辨率卷积，第2/3/4阶段分别重复使用含2/3/4种分辨率的模块。近年有多种语义分割模型以 HRNet 为骨干网络。

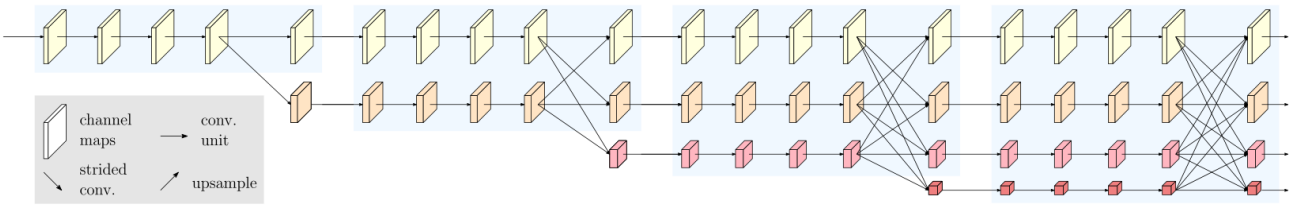

图13 HRNet 架构。引自[42]。
channel maps：通道图；conv. unit：卷积单元；strided conv.：带步幅卷积；upsample：上采样。

还有一些工作采用转置卷积或编码器-解码器进行图像分割，如堆叠式反卷积网络（SDN）[43]、LinkNet[44]、W-Net[45]，以及用于 RGB-D 分割的局部敏感反卷积网络[46]。

3.3.2 医学与生物医学图像分割

受 FCN 与编码器-解码器网络启发，一些模型起初是为医学或生物医学图像分割开发的，目前也用于医学之外的领域。

Ronneberger 等[47]提出 U-Net（图14），用于高效分割生物显微图像。U-Net 由两部分组成：用于捕捉上下文的收缩路径，以及结构对称、用于精确定位的扩张路径。U-Net 的训练策略通过数据增强使模型能够从极少量带标注的图像中有效学习。该模型用30幅透射光显微图像进行训练，并以较大优势赢得了2015年 ISBI 细胞跟踪挑战赛。

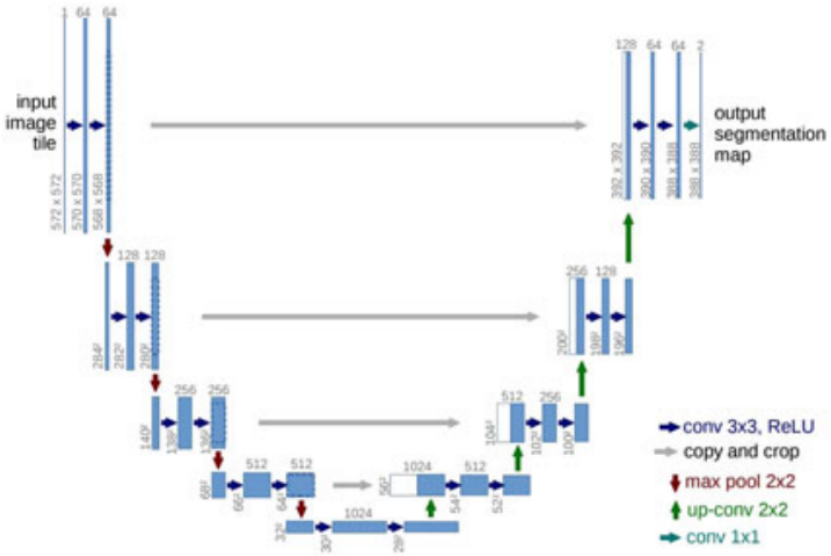

图14 U-Net 模型。引自[47]。
copy and crop：复制并裁剪；max pool：最大池化；up-conv：上卷积。

针对不同类型的图像和问题领域，研究人员提出了多种 U-Net 扩展。例如，Zhou 等[48]开发了嵌套式 U-Net 架构，Zhang 等[49]开发了基于 U-Net 的道路分割算法，Çiçek 等[50]则提出了适用于三维图像的 U-Net 架构。

Milletari 等[51]提出的 V-Net（图15）是另一种著名的 FCN 模型，用于三维医学图像分割。作者引入基于 Dice 系数的新损失函数，使模型能处理前景与背景体素数量严重不均衡的情况。该网络在前列腺磁共振（MRI）图像上以端到端的方式训练，可一次预测整个体数据的分割结果。相关的医学图

像分割工作还包括使用 Progressive Dense V-Net 等方法从胸部 CT 图像自动分割肺叶，以及用于病灶分割的 3D-CNN 编码器[52]。

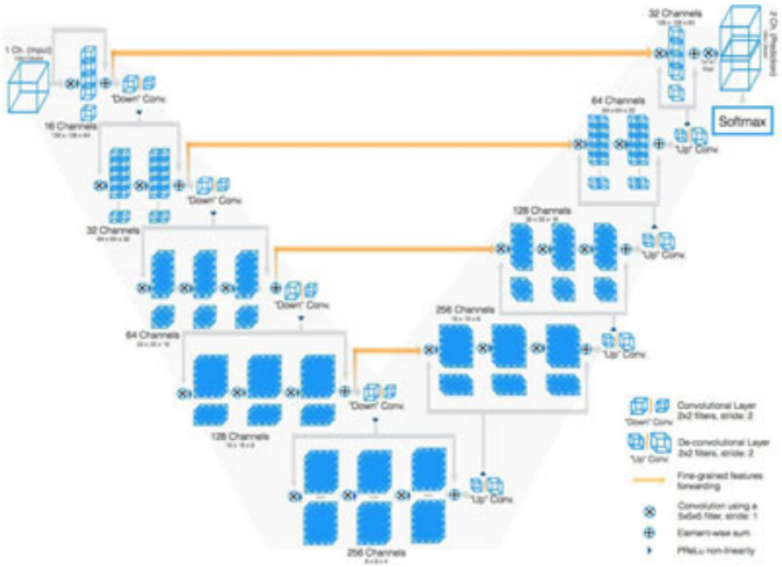

图15　用于三维图像分割的 V-Net 模型。引自[51]。

3.4 基于多尺度与金字塔网络的模型

多尺度分析是图像处理中的成熟方法，已被用于多种神经网络架构。其中最具代表性的模型之一是 Lin 等[53]提出的特征金字塔网络（FPN）。它最初为目标检测开发，也应用于分割。FPN 利用深层 CNN 固有的多尺度金字塔层级结构构建特征金字塔，只增加很少的开销。为融合低分辨率和高分辨率特征，FPN 由自底向上路径、自顶向下路径和横向连接组成。拼接后的特征图经过 3 × 3 卷积，得到每个阶段的输出。最后，自顶向下路径的每一阶段都会生成用于目标检测的预测。在图像分割中，作者使用两个多层感知机（MLP）生成掩膜。

Zhao 等[54]提出金字塔场景解析网络（PSPN），用多尺度结构改善对场景全局上下文表示的学习（图16）。它使用残差网络（ResNet）作为特征提取器，结合空洞卷积网络，从输入图像中提取多种模式。随后将特征图送入金字塔池化模块，以区分不同尺度的模式。特征图在4种尺度上池化，每种尺度对应金字塔的一级；再经 1 × 1 卷积层降维。各金字塔层级的输出上采样后与初始特征图拼接，同时捕捉局部与全局上下文信息。最后由卷积层输出逐像素预测。

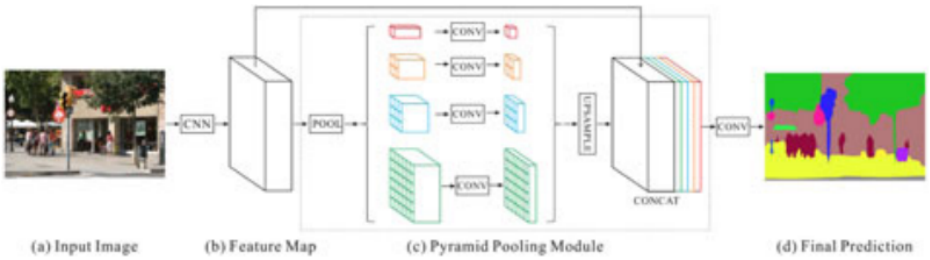

图16　金字塔场景解析网络（PSPN）的架构。引自[54]。
从左到右：输入图像、特征图、金字塔池化模块、最终预测。

Ghiasi 和 Fowlkes[55]基于拉普拉斯金字塔开发了多分辨率重建架构。它利用来自较高分辨率特征图的跳跃连接和乘法门控，逐步细化从较低分辨率特征图重建的分割边界。作者表明，尽管卷积特征图的表观空间分辨率较低，高维特征表示仍包含大量亚像素定位信息。

其他使用多尺度分析进行分割的模型和方法包括动态多尺度滤波器网络（DM-Net）[56]、上下文对比网络与门控多尺度聚合（CCN）[57]、自适应金字塔上下文网络（APC-Net）[58]、多尺度上下文交织（MSCI）[59]，以及显著对象分割[60]。

3.5 基于 R-CNN 的模型

基于区域的 CNN（R-CNN）及其扩展在目标检测中取得了成功。其中，Faster R-CNN[61]架构（图17）引入区域提议网络（RPN），用来给出候选边界框。RPN 提取感兴趣区域（RoI），RoIPool 层再从这些候选区域计算特征，用于推断对象的边界框坐标和类别。R-CNN 的一些扩展被用来处理实例分割，即同时执行目标检测和语义分割的任务。

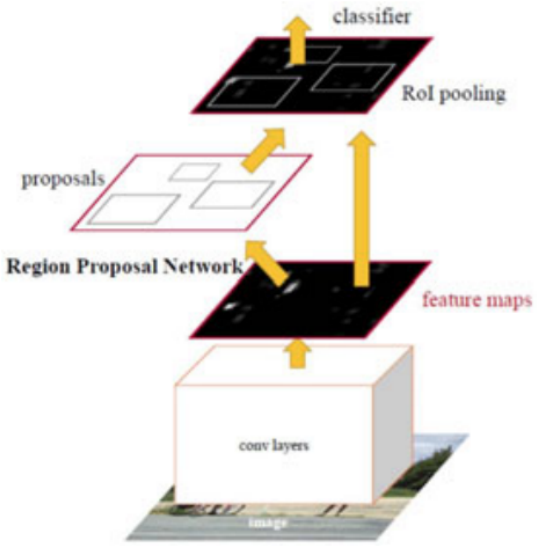

图17　Faster R-CNN 架构。卷积层处理每幅图像并提取其特征。区域提议网络（RPN）在特征图的每个位置使用滑窗；每个位置采用 k 个锚框（k = 9），即 128、256、512三种尺度和 1:1、1:2、2:1 三种宽高比的组合，以生成区域提议。分类层（cls）针对 k 个框是否含对象输出 2k 个得分；回归层（reg）为 k 个框输出 4k 个坐标参数，即框的中心坐标、宽和高。引自[61]。
Region Proposal Network：区域提议网络；RoI pooling：感兴趣区域池化；classifier：分类器。

He 等[62]提出 Mask R-CNN（图18）。它在多项 COCO 对象实例分割挑战中超过此前的基准表现（图19），能高效检测图像中的对象，同时为每个实例生成高质量分割掩膜。其本质是一个有3个输出分支的 Faster R-CNN：第一个分支计算边界框坐标，第二个分支计算对应类别，第三个分支计算用于分割对象的二值掩膜。Mask R-CNN 的损失函数结合边界框坐标、预测类别和分割掩膜的损失，并对三个分支进行联合训练。

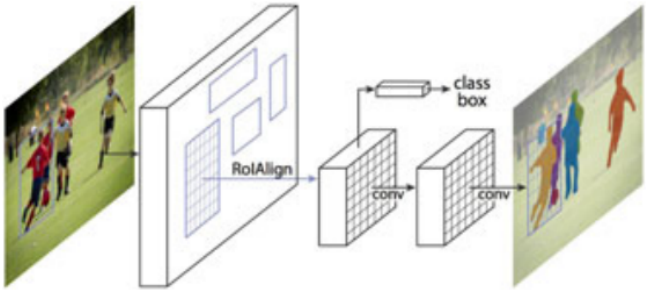

图18　Mask R-CNN 架构。引自[62]。
RoIAlign：感兴趣区域对齐；class、box、mask 分别对应类别、边界框及掩膜输出。

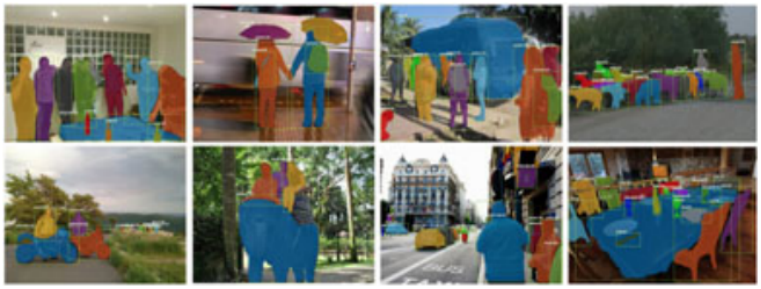

图19　Mask R-CNN 的实例分割结果。引自[62]。

Liu 等[63]提出路径聚合网络（PANet），以 Mask R-CNN 和 FPN 为基础（图20）。该网络的特征提取器以 FPN 为骨干网络，增加一条自底向上的增强路径，以改善低层特征的传播。这条第三路径的每个阶段都以前一阶段的特征图为输入，用 3 × 3 卷积层处理。横向连接把该输出加到自顶向下路径同一阶段的特征图上，叠加后的特征图再输入下一阶段。

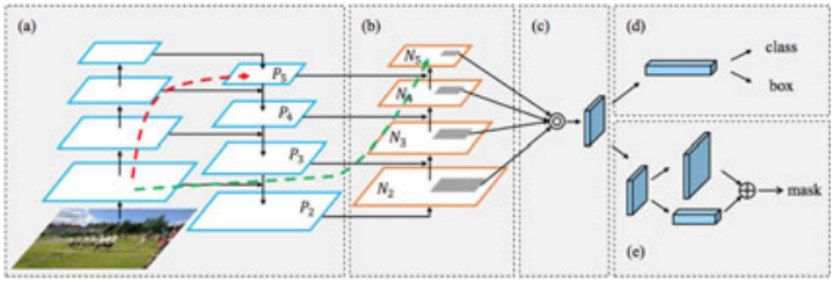

图20　路径聚合网络。（a）FPN 骨干网络；（b）自底向上的路径增强；（c）自适应特征池化；（d）边界框分支；（e）全连接融合。引自[63]。

Dai 等[64]开发了面向实例感知语义分割的多任务网络，由三个网络分别区分实例、估计掩膜和对对象分类。这三个网络构成级联结构，并共享卷积特征。Hu 等[65]提出一种新的部分监督训练范式及权重迁移函数。这使实例分割模型能够在大量类别上训练：这些类别均有边界框标注，但只有很小一部分有掩膜标注。

Chen 等[66]在 Faster R-CNN 的基础上，用语义特征和方向特征改善目标检测，由此开发了实例分割模型 MaskLab。模型有三种输出（图21）：边界框检测结果、用于逐像素分类的语义分割 logits，以及用于预测每个像素朝向其所属实例中心的方向 logits。借助 Faster R-CNN 检测器，预测的边界

框可以准确定位对象实例。在每个感兴趣区域内，MaskLab 结合语义预测和方向预测进行前景与背景分割。

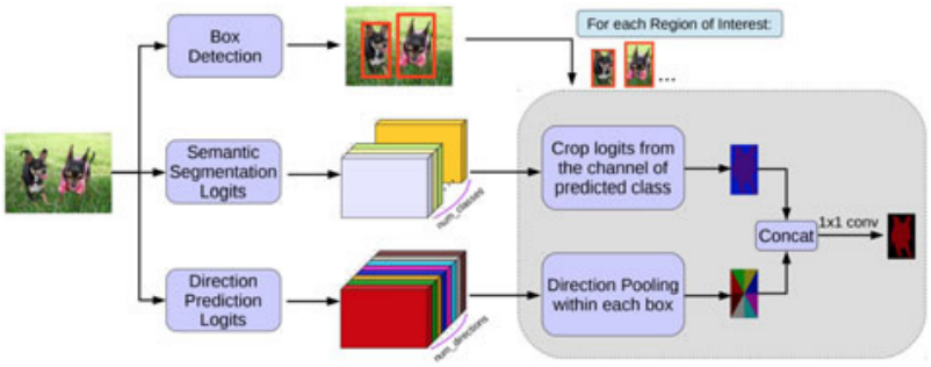

图21　MaskLab 模型。引自[66]。
三个分支依次是边界框检测、语义分割 logits、方向预测 logits；模型在感兴趣区域内融合分割信号。

Chen 等[67]提出的 TensorMask 基于密集滑窗做实例分割。作者把密集实例分割视作四维张量上的预测任务，给出一个通用框架，可以在四维张量上实现新型算子。他们证明，张量方法相较基线取得了较大提升，结果与 Mask R-CNN 相当。

其他基于 R-CNN 的实例分割模型包括面向掩膜提议开发的 R-FCN[68]、DeepMask[69]、PolarMask[70]、边界感知实例分割[71]和 CenterMask[72]。另一条有前景的路线是学习自底向上分割所需的分组线索来处理实例分割，例如深度分水岭变换[73]、实时实例分割[74]，以及基于深度度量学习的语义实例分割[75]。

3.6 空洞卷积模型

空洞卷积（dilated convolution，也称 atrous convolution）为卷积层增加了一个参数：膨胀率。例如，膨胀率为2的 3 × 3 卷积核（图22）与 5 × 5 卷积核有同样大的感受野，却只使用9个参数，因此可以在不增加计算成本的情况下扩大感受野。

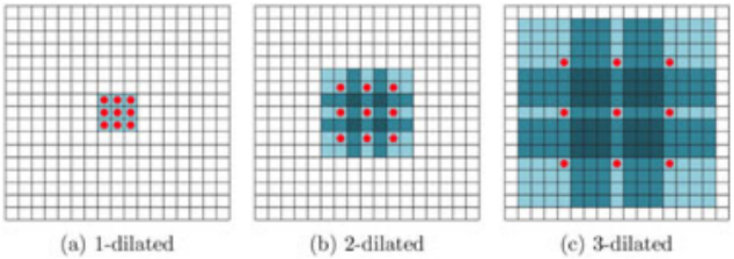

图22　空洞卷积：3 × 3 卷积核在不同膨胀率下的采样位置。

空洞卷积广泛用于实时分割，许多近期论文都报告使用了这一技术。具有代表性的有 DeepLab 系列[76]、多尺度上下文聚合[77]、密集上采样卷积与混合空洞卷积（DUC-HDC）[78]、密集连接的空洞空间金字塔池化（DenseASPP）[79]和高效网络（ENet）[80]。

Chen 等提出的 DeepLabv1[36]与 DeepLabv2[76]属于最常用的图像分割模型。后者有三个关键特征（图23）。第一，采用空洞卷积来解决最大池化和带步幅运算导致的网络分辨率下降。

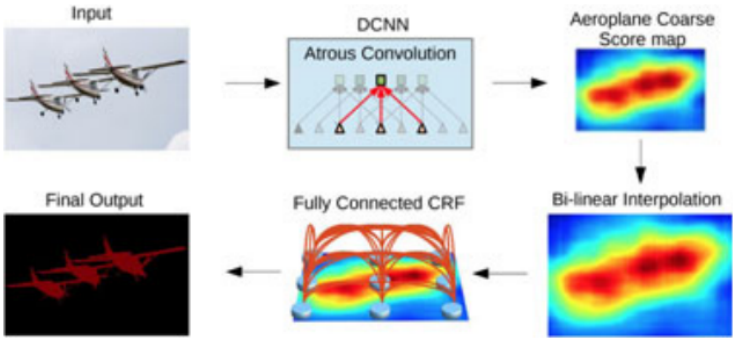

图23 DeepLab 模型。引自[76]。

第二，采用空洞空间金字塔池化（ASPP），使用不同采样率的滤波器探测输入的卷积特征层，同时捕捉对象和多尺度图像上下文，从而稳健地分割不同尺度的对象。第三，把全卷积 VGG-16 或 ResNet-101 等深层 CNN 方法和全连接 CRF 等概率图模型结合起来，提高对象边界的定位精度。

随后，Chen 等[13]提出 DeepLabv3，把级联和并行的空洞卷积模块结合起来。并行卷积模块编入 ASPP，其中还加入了 1 × 1 卷积和批归一化。把各路输出拼接后，再经另一个 1 × 1 卷积，得到每个像素的 logits。Chen 等[81]接着发布了 DeepLabv3+（图24），采用编码器-解码器架构，其中的空洞可分离卷积由逐通道卷积（对输入的每个通道分别进行空间卷积）和逐点卷积（以逐通道卷积的输出为输入的 1 × 1 卷积）组成。作者用 DeepLabv3 框架作为编码器。文中着重介绍的版本采用经改造的 Xception 骨干网络，层数更多，并采用空洞深度可分离卷积来替代最大池化，同时使用批归一化。

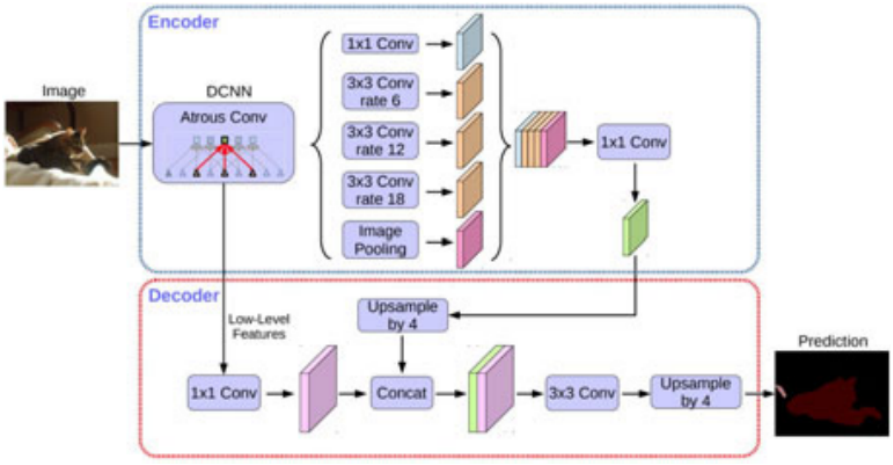

图24 DeepLabv3+ 模型。引自[81]。
Encoder：编码器；Decoder：解码器；Concat：拼接；Upsample by 4：上采样至 4 倍尺寸。

3.7 基于 RNN 的模型

CNN 很适合计算机视觉问题，但不是唯一选择。RNN 可以对像素间的短期、长期依赖关系建模，有望改善分割图的估计。采用 RNN 时，可把像素相连后按序处理，以对全局上下文建模，改进语义分割。然而，图像天然具有二维结构，这给 RNN 带来了挑战。

Visin 等[82]提出名为 ReSeg 的 RNN 语义分割模型（图25），主要以为图像分类开发的 ReNet [83]为基础。每个 ReNet 层由四个 RNN 组成，在水平和垂直方向上分别对图像做双向扫描，对图像块或激活值编码，提供相应的全局信息。将 ReNet 层堆叠在预训练的 VGG-16 卷积层上，由这些卷积层提取通用的局部特征；其后接上采样层，使最终分割预测恢复到原始图像分辨率。

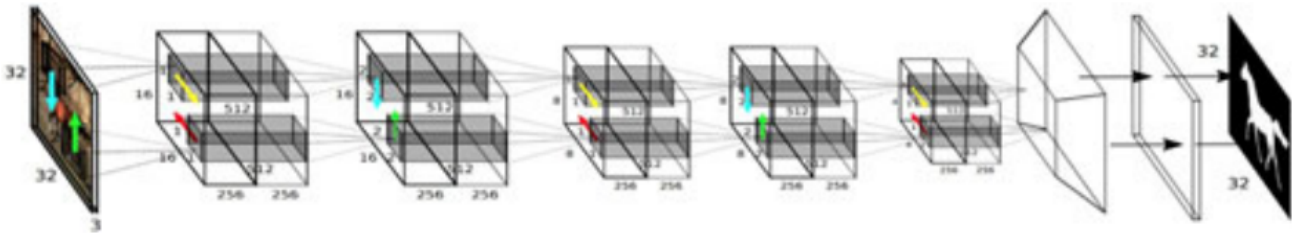

图25　ReSeg 模型（未画出预训练的 VGG-16 特征提取器）。引自[82]。

Byeon 等[84]使用二维 LSTM 网络对自然场景图像逐像素分割和分类。该网络在同一个模型中学习纹理及标签间复杂的空间依赖关系，同时完成分类、分割与上下文整合。

Liang 等[85]提出基于图 LSTM 网络的语义分割模型（图26），在卷积层之外加入以超像素图为基础的图 LSTM 层，引入更全局的结构上下文。这样就把用于规则阵列数据的 LSTM（按行、按网格或沿对角线处理）推广到了不规则的图结构数据：形状不规则但语义一致的超像素作为节点，超像素之间的相邻关系作为边，共同构成无向图（图27）。

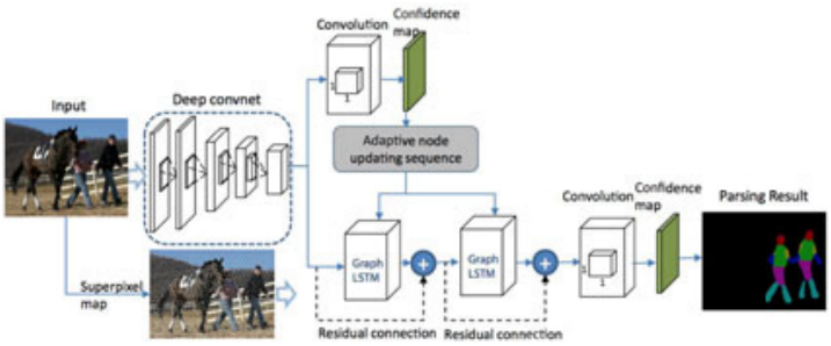

图26　用于语义分割的图 LSTM 模型。引自[85]。

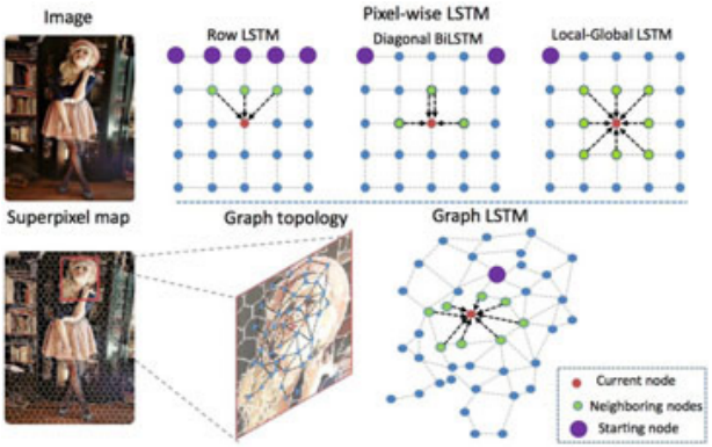

图27　常规 RNN 模型与图 LSTM 的比较。引自[85]。
图中分别比较按行的 LSTM、沿对角线的双向 LSTM、局部-全局 LSTM，以及按超像素邻接关系运行的图 LSTM。

Xiang 和 Fox[86]提出数据关联循环神经网络（DA-RNN），联合进行三维场景建图和语义标注。DA-RNN 采用新的循环神经网络架构对 RGB-D 视频做语义标注，并将网络输出与 Kinect-Fusion 等建图技术结合，把语义信息注入重建的三维场景。

Hu 等[87]提出一种语义分割算法，结合 CNN 对图像编码、LSTM 对该图像的语言描述编码。为根据语言输入得到逐像素的图像分割，他们提出一个可端到端训练的循环-卷积模型，联合学习对视觉和语言信息的处理（图28）。这不同于在一组预先定义的语义类别上进行的传统语义分割；例如，给出“坐在右边长椅上的两名男子”，只应分割出坐在右边长椅上的这两个人，不分割坐在另一张长椅上或站立的其他人。图29给出该模型的一个分割示例。

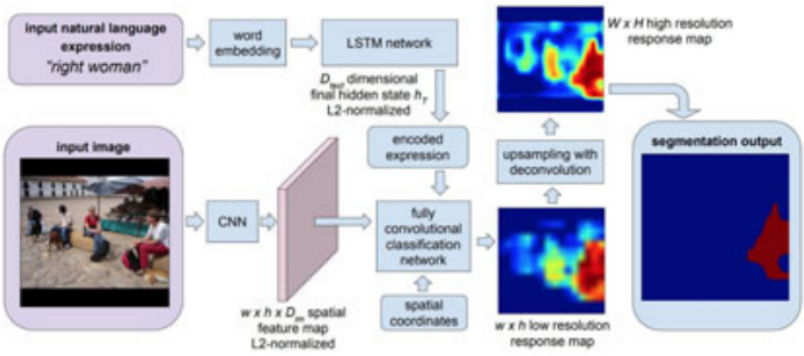

图28　根据自然语言表达式进行语义分割的 CNN+LSTM 架构。引自[87]。
上方为自然语言表达式输入及其词嵌入、LSTM 编码，下方为图像和 CNN 特征；联合得到分割结果。

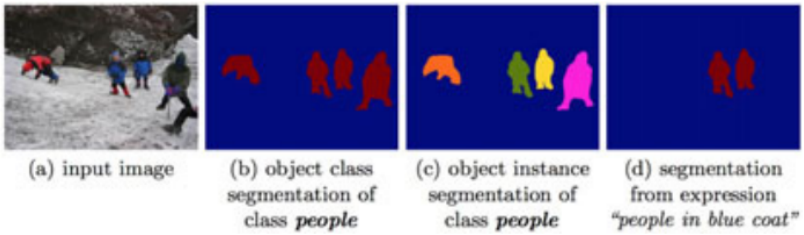

图29　CNN+LSTM 根据“穿蓝色外套的人”这一查询生成的分割掩膜。引自[87]。
(a) 输入；(b) 人物类别分割；(c) 人物实例分割；(d) 按语言表达式分割。

基于 RNN 的模型有一个缺点：它们的序列式处理不易并行化，因此通常比对应的 CNN 模型慢。

3.8 基于注意力的模型

多年来，计算机视觉领域一直在研究注意力机制，也已有论文将其用于语义分割。

Chen 等[88]提出一种注意力机制，学习对每个像素位置的多尺度特征进行软加权。他们改造了一个性能较强的语义分割模型，使用多尺度图像对它和注意力模型进行联合训练。如图30所示，对1.0尺度的特征，模型给背景中的人物（绿色虚线圈）分配较高权重；对0.5尺度的特征，则给画面上较大的儿童（品红色虚线圈）分配较高权重。注意力机制让模型能判断不同位置、不同尺度的特征有多重要，效果优于平均池化和最大池化。

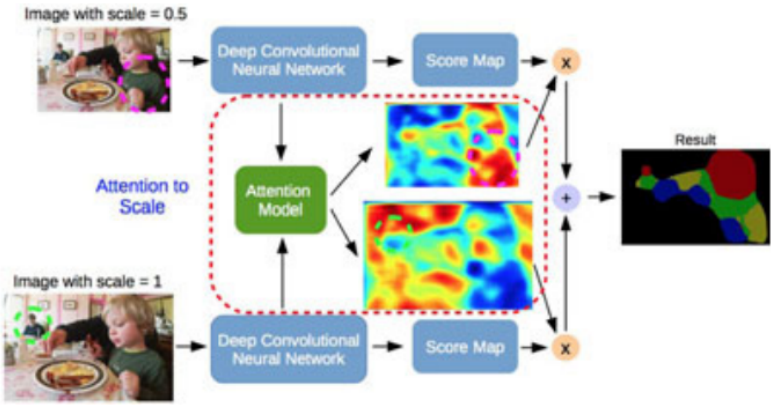

图30　基于注意力的语义分割模型。引自[88]。
Score Map：得分图；Attention Model：注意力模型；Result：融合结果。

以往方法训练卷积分类器来学习带标注对象的代表性语义特征。与之不同，Huang 等[89]为语义分割提出反向注意力网络（RAN）架构（图31），同时运用反向注意力机制，使模型也能捕捉相反的概念，即哪些特征与目标类别无关。RAN 同时进行正向和反向注意力学习。

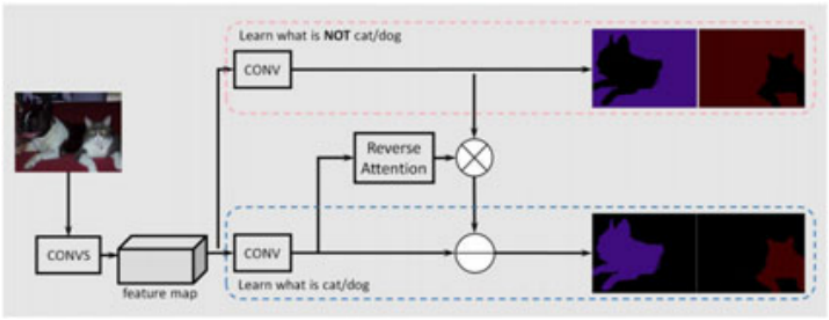

图31　反向注意力网络（RAN）的架构。引自[89]。
上下两支分别学习“不是猫 / 狗”和“是猫 / 狗”的特征。

Li 等[90]为语义分割开发了金字塔注意力网络，利用全局上下文信息进行分割。他们不采用复杂的空洞卷积和解码器网络，而是结合注意力机制与空间金字塔来提取可用于像素标注的精确稠密特征。Fu 等[91]提出用于场景分割的双重注意力网络，可借助自注意力机制捕捉丰富的上下文依赖关系。具体而言，他们在空洞 FCN 之上接入两种注意力模块，分别对空间维度和通道维度上的语义依赖关系建模。其中，位置注意力模块通过加权求和有选择地汇聚各位置的特征。

注意力机制在语义分割中的其他应用包括：受自注意力启发、使用对象上下文池化的 OCNet[92]，ResNeSt 分组注意力网络[93]，高度驱动注意力网络[94]，期望最大化注意力网络（EMANet）[95]，十字交叉注意力网络（CCNet）[96]，基于循环注意力的端到端实例分割[97]，用于场景解析的逐点空间注意力网络[98]，以及判别特征网络（DFN）[99]。

3.9 生成模型与对抗训练

GAN 已应用于计算机视觉的多种任务，图像分割也在其中。

Luc 等[100]提出语义分割的对抗训练方法：在训练卷积语义分割网络的同时（图32），训练一个对抗网络来区分真值分割图与分割网络生成的图。他们的结果表明，对抗训练提高了模型在 Stanford Background 和 PASCAL VOC 2012 数据集上的准确率。

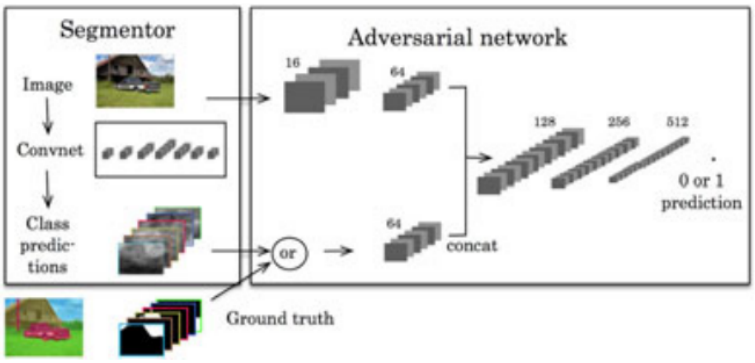

图32　用于语义分割的生成对抗网络。引自[100]。
Segmentor：分割网络；Adversarial network：对抗网络；Ground truth：真值。

Souly 等[101]提出使用 GAN 的半弱监督语义分割方法。该模型中，生成器为一个多类分类器提供额外的训练样本，该分类器在 GAN 中担任判别器：要么为样本分配已有类别中的标签，要么把它标记为假样本（额外类别）。

Hung 等[102]用对抗网络构建了半监督语义分割框架。他们设计了一个 FCN 判别器，在兼顾空间分辨率的同时，区分预测的概率图和真值分割分布。模型的损失函数由三项组成：以分割真值计算的交叉熵损失、判别网络的对抗损失，以及根据判别器输出的置信图计算的半监督损失。

Xue 等[103]提出用于医学图像分割的多尺度 L1 损失对抗网络。他们用 FCN 作为分割网络，生成分割标签图；并提出一个使用多尺度 L1 损失函数的新对抗评判网络，以促使评判网络与分割网络都学习全局和局部特征，捕捉像素间长距离和短距离的空间关系。

基于对抗训练的其他方法包括使用 GAN 分割细胞图像[104]，以及分割并生成对象中不可见的部分[105]。

3.10 结合主动轮廓模型的 CNN

近来，研究人员开始关注全卷积网络（FCN）和主动轮廓模型（ACM）[8]如何相互补充。

一种做法是根据 ACM 的原理构造新的损失函数。例如，受文献[106]全局能量表达式启发，Chen 等[107]提出一个有监督的损失层，在训练 FCN 时纳入预测掩膜的面积和大小信息，并用它处理心脏 MRI 中的心室分割。类似地，Gur 等[108]利用无边缘的形态学主动轮廓[109]构造无监督损失函数，用于微血管图像分割。

另一种做法起初只把 ACM 用于 FCN 输出的后处理，也有一些工作尝试通过预训练 FCN 实现有限程度的联合学习。在自然图像语义分割中，一个采用 ACM 后处理的例子是 Le 等[110]的工作，他们以 RNN 实现水平集 ACM。Rupprecht 等的深度主动轮廓[111]是另一个例子。对于医学图像分割，Hatamizadeh 等[112]提出集成式深度主动病灶分割（DALS）模型，训练 FCN 骨干网络来预测一种新的局部参数化水平集能量泛函的参数函数。在另一项相关工作中，Marcos 等[113]提出深度结构化主动轮廓（DSAC），用结构化预测框架结合 ACM 与预训练的 FCN，对航拍图像中的建筑物做实例分割，不过需要人工初始化。针对同一应用，Cheng 等[114]提出深度主动射线网络（DarNet），其思路与 DSAC 类似，但采用了基于极坐标的另一种显式 ACM 形式，以避免轮廓自相交。

Hatamizadeh 等[115]近期提出可训练的深度主动轮廓（TDAC），真正做到了把 FCN 和 ACM 完全集成，并用反向传播端到端训练。在文献[112]基础上，他们把局部参数化的水平集 ACM 实现为 FCN 骨干网络之后的若干附加卷积层，并借助 TensorFlow 的自动微分机制，让训练误差的梯度在整个 DCAC 框架中反向传播。这个全自动模型在训练或分割时均不需要人工干预，可以直接分割多个感兴趣对象实例，并能处理任意形状的对象，包括具有尖角的对象。

3.11 其他模型

用于图像分割的其他常用深度学习架构包括下列模型。

上下文编码网络（EncNet）[116]通过基础特征提取器得到特征图，再把它们输入上下文编码模块。RefineNet[117]是多路径细化网络，它显式利用下采样过程中可用的所有信息，并通过长距离残差连接实现高分辨率预测。SeedNet[118]提出基于深度强化学习的自动种子生成技术，学习解决交互式分割问题。对象上下文表示（OCR）[42]学习对象区域以及每个像素与每个对象区域之间的关系，用对象上下文表示增强像素的表示。其他模型和方法包括 BoxSup[119]、图卷积网络（GCN）[120]、Wide ResNet[121]、Exfuse[122]（增强低层与高层特征的融合）、Feedforward-Net[123]、用于测地视频分割的显著性感知模型[124]、双重图像分割（DIS）[125]、FoveaNet[126]（透视感知场景解析）、Ladder DenseNet[127]、双边分割网络（BiSeNet）[128]、用于场景解析的语义预测引导网络（SPGNet）[129]、门控形状 CNN[130]、自适应上下文网络（AC-Net）[131]、动态结构语义传播网络（DSSPN）[132]、符号图推理（SGR）[133]、CascadeNet[134]、尺度自适应卷积（SAC）[135]、统一感知解析网络（UPerNet）[136]、基于重新训练和自训练的分割[137]、密集连接神经架构搜索[138]、分层多尺度注意力[139]、高效 RGB-D 语义分割（ESA-Net）[140]、迭代金字塔上下文[141]，以及为语义分割学习动态路由[142]。

全景分割[143]正受到越来越多关注。相关工作包括全景特征金字塔网络（PFPN）[144]、注意力引导的全景分割网络[145]、无缝场景分割[146]、Panoptic DeepLab[147]、统一全景分割网络[148]和高效全景分割[149]。

图33按时间列出了自2014年以来部分最具代表性的深度学习图像分割模型。

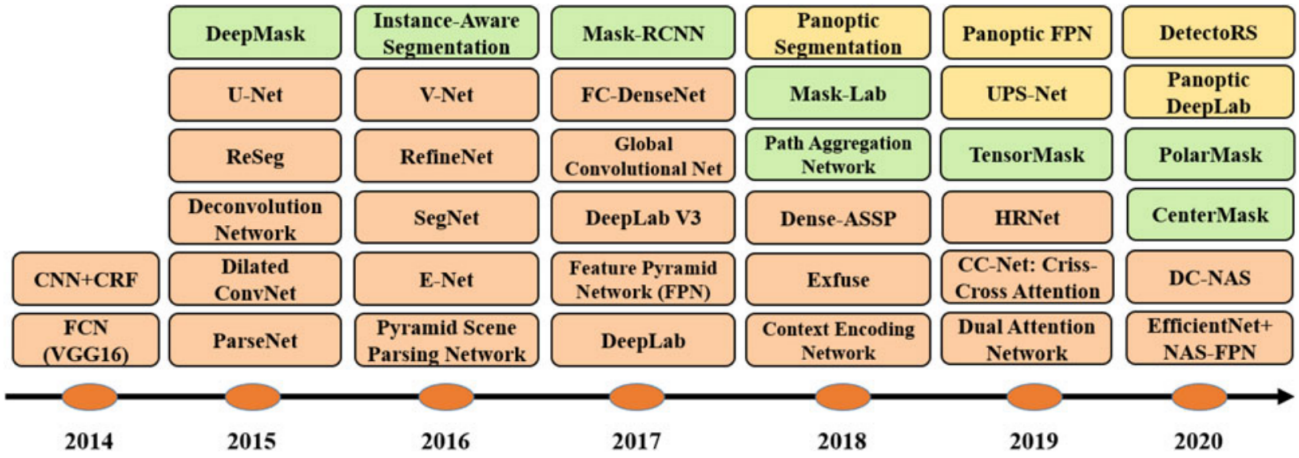

图33 代表性深度学习图像分割算法的时间线。橙色、绿色、黄色分别表示语义分割、实例分割和全景分割。

4 数据集

本节回顾训练和测试深度学习图像分割模型时最常用的图像数据集，按二维图像（像素）、2.5维RGB-D 图像（彩色+深度）和三维图像（体素）分为3类，并具体介绍各数据集的特点。

数据增强常用来增加带标注的样本数量，尤其适用于医学影像等领域的小型数据集，从而提高深度学习分割模型的性能。可在数据空间或特征空间实施一系列变换，也可以同时进行（即同时变换图像和分割图）。常见变换包括平移、镜像、旋转、几何形变、缩放、颜色空间偏移、裁剪和向主成分投影。数据增强还有助于加快收敛、降低过拟合风险并提高泛化能力。在某些小型数据集上，数据增强已被证明可以让模型性能提升超过20%。

4.1 二维图像数据集

图像分割研究大多集中在二维图像上，因此已有许多二维图像分割数据集。下面介绍其中最常用的一些。

PASCAL Visual Object Classes（VOC）[150]是计算机视觉中十分常用的数据集，为5种任务提供带标注的图像：分类、分割、检测、动作识别和人体结构布局。对于分割任务，有21个带标注的对象类别；不属于这些类别的像素被标为背景。数据集分为训练集和验证集，分别包含1,464幅和1,449幅图像，另有不公开的测试集用于正式挑战赛。图34展示一幅示例图像与它的逐像素标签。

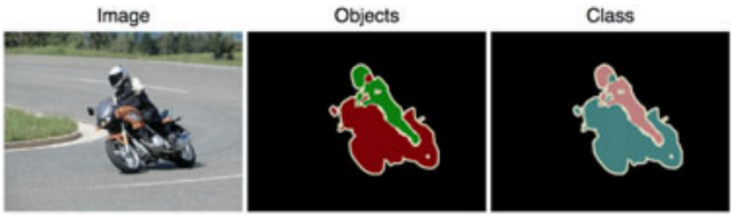

图34　PASCAL VOC 数据集中的图像示例。引自[151]。
左至右：原图、实例标注、语义类别标注。

PASCAL Context[152]扩展了 PASCAL VOC 2010 检测挑战赛的数据，为所有训练图像提供逐像素标签。它包含400多个类别（包括 PASCAL VOC 分割的原有20个类别和背景），归为三组：可数对象（things）、不可数区域（stuff）和混合类别。许多类别的样本过于稀少，因此一般选用包含59个类别的子集。

微软 Common Objects in Context（MS COCO）[153]是一个面向目标检测、分割和图像描述的大规模数据集。COCO 的图像来自复杂的日常场景，呈现常见对象在自然环境中的样子。数据集含91类对象的照片，328K 幅图像中共标注了250万个实例。图35以一幅图像为例，将 MS COCO 的标签与此前数据集的标签进行对比。

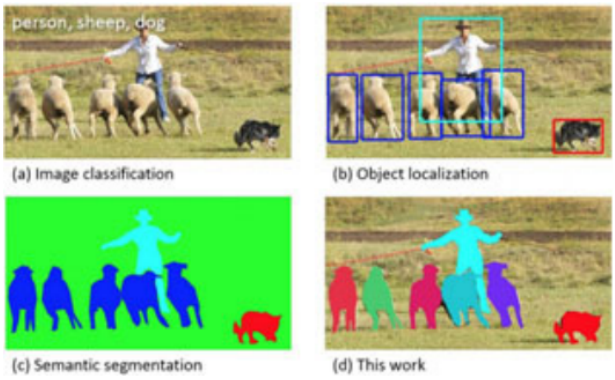

图35　COCO 中的样本图像和分割图。引自[153]。
(a) 图像分类；(b) 对象定位；(c) 语义分割；(d) MS COCO 的实例级分割标注。

Cityscapes[154]是以城市街景语义理解为重点的大型数据库。它包含在50座城市的街景中采集的多样化双目视频序列，其中5K帧具有高质量的像素级标注，另有20K帧带弱标注。数据集对30个类别提供稠密的像素级语义标注，分为8组：平坦表面、人、车辆、建筑构筑物、物品、自然、天空和空类（void）。图36展示了该数据集中的分割图示例。



图36　Cityscapes 数据集中的分割图。引自[154]。

ADE20K / MIT Scene Parsing（SceneParse150）为场景解析算法提供训练和评估平台。该基准基于 ADE20K 数据集[134]，后者包含20K多幅以场景为中心的图像，对其中的对象及其组成部分都有

详尽标注。该基准用20K幅图像训练，以2K幅做验证，并另外留出一批测试图像。数据集共有150个语义类别。

SiftFlow[155]来自 LabelMe 数据库的一个子集，包含2,688幅带标注图像，涵盖街道、山地、田野、海滩和建筑等8种室外场景，使用33个语义类别之一进行标注。

Stanford Background[156]汇集 LabelMe、MSRC 和 PASCAL VOC 等现有数据集的室外场景图像，包含715幅至少有一个前景对象的图像。这些图像有逐像素标注，可用于语义场景理解。

伯克利分割数据集（BSD）[157]包含30名受试者为 Corel 数据集的1,000幅图像给出的12,000份人工分割标注，旨在为图像分割和边界检测研究提供经验依据。半数标注是在给受试者呈现彩色图像时获得的，另一半对应灰度图像。基于这些数据的公开基准收录了300幅图像的全部灰度和彩色分割标注，其中200幅用于训练、100幅用于测试。

YouTube-Objects[158]包含从 YouTube 收集的视频，涉及 PASCAL VOC 的10个类别：飞机、鸟、船、汽车、猫、牛、狗、马、摩托车和火车。该数据集最初为目标检测开发，只带弱标注，没有逐像素标注。Jain 等[159]随后人工标注了其中126个序列，再抽取部分帧生成语义标签。该数据集共提供约10,167帧带标注图像，分辨率为 480 × 360 像素。

CamVid 是另一个场景理解数据库，侧重道路和驾驶场景。最初使用安装在汽车仪表台上的相机录制了5段视频，从中采样出701帧，并以32个类别对这些视频帧做了人工标注。

KITTI[160]是最常用的自动驾驶数据集之一，包含以多种传感器模态录制的交通场景视频，例如高分辨率 RGB 相机、灰度双目相机和三维激光扫描仪。原始数据集没有语义分割真值，研究人员对其中一部分做了人工标注。例如，Alvarez 等[161]为道路检测挑战赛的323幅图像制作了3类真值：道路、竖直区域和天空。

其他用于图像分割的数据集包括 Semantic Boundaries Dataset（SBD）[162]、PASCAL Part[163]、SYNTHIA[164]和 Adobe 的人像分割数据集[165]。

4.2 2.5维数据集

随着价格可负担的测距扫描仪日益普及，RGB-D 图像的使用越来越广泛。以下是部分最常用的 RGB-D 数据集。

NYU-Depth V2[166]由微软 Kinect 的 RGB 相机和深度相机在多种室内场景中录制的视频序列组成。它收录来自3座城市、450多个场景的1,449对带稠密标注的 RGB 图像和深度图像。每个对象都有类别和实例编号，例如 cup1、cup2、cup3 等；数据集另外包含407,024帧无标注图像。图37展示一个 RGB-D 样本及其标签图。

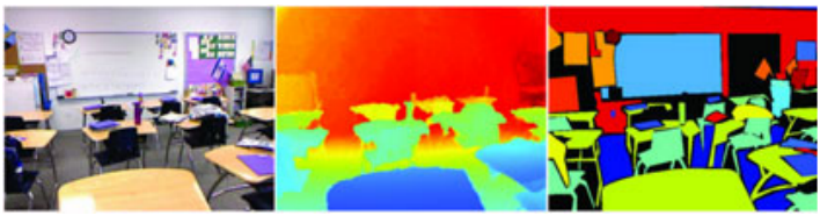

图37　NYU V2 数据集的一个样本。从左到右依次为 RGB 图像、预处理后的深度图像、类别标签图。引自[166]。

SUN-3D[167]是大型 RGB-D 视频数据集，包含从41栋不同建筑内的254处空间采集的415个序列。其中8个序列已经标注，之后还将标注更多序列。每个带标注的帧提供场景中对象的语义分割结果和相机位姿信息。

SUN RGB-D[168]提供 RGB-D 基准，以推动各项主要场景理解任务的性能上限。数据由四种不同的传感器采集，包含10,000幅 RGB-D 图像，规模与 PASCAL VOC 相当。

ScanNet[169]是 RGB-D 视频数据集，包含1,500多次扫描中的250万个视图，标注包括三维相机位姿、表面重建和实例级语义分割结果。研究者为此设计了易用、可扩展的 RGB-D 采集系统，能自动重建表面，并通过众包获得语义标注。使用这些数据有助于在若干三维场景理解任务上达到当时的先进水平，包括三维对象分类、语义体素标注和 CAD 模型检索。

Stanford 2D-3D[170]提供采自6处室内区域、彼此配准的多种二维、2.5维及三维模态，并带有实例级语义和几何标注。数据集包含70,000多幅 RGB 图像，以及相应的深度、表面法向量和语义标注，还包括全局 XYZ 图像、相机信息，以及经配准的原始三维网格和点云、带语义标注的三维网格和点云。

另一个常用的2.5维数据集是 UW RGB-D Object Dataset[171]，包含用 Kinect 式传感器采集的300种常见家居物品。

5 深度学习分割模型的性能

本节汇总评估分割模型性能时的常用指标，并报告深度学习分割模型在基准数据集上的表现。

5.1 图像分割模型的评价指标

理想情况下，评估图像分割模型应考虑多个方面，包括量化的准确率、视觉质量、速度（推理时间）和存储需求（内存占用）。不过，截至撰稿时，大多数研究主要关注量化模型准确率的指标。最常用的指标如下。

像素准确率是正确分类的像素数占像素总数的比例。对 K + 1 个类别（K 个前景类别和一个背景类别），像素准确率定义如下：

$$\mathrm{PA} = \frac{\sum_{i=0}^{K} p_{ii}}{\sum_{i=0}^{K} \sum_{j=0}^{K} p_{ij}}, \qquad (1)$$

其中 p_{ij} 表示真实类别为 i、被预测为类别 j 的像素数。

平均像素准确率（MPA）是像素准确率（PA）的扩展：先按类别计算像素分类正确的比例，再对所有类别取平均：

$$\mathrm{MPA} = \frac{1}{K+1}\sum_{i=0}^{K}\frac{p_{ii}}{\sum_{j=0}^{K} p_{ij}}. \tag{2}$$

交并比（IoU），也称 Jaccard 指数，定义为预测分割图 A 与真值图 B 交集的面积除以并集的面积，取值在0和1之间：

$$\mathrm{IoU} = J(A,B) = \frac{|A \cap B|}{|A \cup B|}. \tag{3}$$

平均交并比（mIoU）是所有类别的 IoU 的平均值。

精确率、召回率和 F1 分数既可按类别定义，也可在总体层面定义，如下所示：

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}; \qquad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{4}$$

其中 TP 表示真阳性比例，FP 表示假阳性比例，FN 表示假阴性比例。通常需要结合考察精确率与召回率，因此将 F1 分数定义为二者的调和平均数：

$$\mathrm{F1} = \frac{2\ \mathrm{Precision}\ \mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}. \tag{5}$$

Dice 系数常用于医学图像分析，可以定义为预测分割图与真值图重叠面积的两倍除以两图像素数之和。

$$\mathrm{Dice} = \frac{2|A \cap B|}{|A|+|B|}. \tag{6}$$

它与交并比（式3）很相似。对于把前景视为正类的二值图，Dice 系数与 F1 分数相同，见式（7）。

$$\mathrm{Dice} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} = \mathrm{F1}. \tag{7}$$

5.2 基于深度学习的模型：量化性能

本节以表格列出此前讨论的若干算法在常用分割基准上的表现。多数论文在标准数据集上使用标准指标报告模型性能，但也有一些并未这样做，因此难以全面比较。此外，只有少数论文以可复现的方式提供了运行时间、内存占用等额外信息。这些信息对工业应用（如无人机、自动驾驶汽车、机器人等）很重要：这类应用可能在算力和存储有限的嵌入式系统上运行，因而需要轻量化模型。

下列表格汇总了若干具有代表性的深度学习分割模型在不同数据集上的性能。

表1关注 PASCAL VOC 测试集。自引入第一个基于深度学习的图像分割模型 FCN 以来，模型的准确率明显提高。

表1　分割模型在 PASCAL VOC 测试集上的准确率

Method	Backbone	mIoU
FCN [30]	VGG-16	62.2
CRF-RNN [38]	-	72.0
CRF-RNN* [38]	-	74.7
BoxSup* [119]	-	75.1
Piecewise* [39]	-	78.0
DPN* [40]	-	77.5
DeepLab-CRF [76]	ResNet-101	79.7
GCN* [120]	ResNet-152	82.2
Dynamic Routing [142]	-	84.0
RefineNet [117]	ResNet-152	84.2
Wide ResNet [121]	WideResNet-38	84.9
PSPNet [54]	ResNet-101	85.4
DeeplabV3 [13]	ResNet-101	85.7
PSANet [98]	ResNet-101	85.7
EncNet [116]	ResNet-101	85.9
DFN* [99]	ResNet-101	86.2
Exfuse [122]	ResNet-101	86.2
SDN* [43]	DenseNet-161	86.6
DIS [125]	ResNet-101	86.8
APC-Net* [58]	ResNet-101	87.1
EMANet [95]	ResNet-101	87.7
DeeplabV3+ [81]	Xception-71	87.8
Exfuse [122]	ResNeXt-131	87.9
MSCI [59]	ResNet-152	88.0
EMANet [95]	ResNet-152	88.2
DeeplabV3+* [81]	Xception-71	89.0
EfficientNet+NAS-FPN [137]	-	90.5

表头：方法、骨干网络、平均交并比（mIoU）。

* 表示在其他数据集（MS COCO、ImageNet 等）上预训练的模型。

表2关注 Cityscapes 测试集。在这一数据集上，最新模型比开创性的 FCN 模型相对提升约23%。

表2 分割模型在 Cityscapes 数据集上的准确率

Method	Backbone	mIoU
SegNet [25]	-	57.0
FCN-8s [30]	-	65.3
DPN [40]	-	66.8
Dilation10 [77]	-	67.1
DeeplabV2 [76]	ResNet-101	70.4
RefineNet [117]	ResNet-101	73.6
FoveaNet [126]	ResNet-101	74.1
Ladder DenseNet [127]	Ladder DenseNet-169	73.7
GCN [120]	ResNet-101	76.9
DUC-HDC [78]	ResNet-101	77.6
Wide ResNet [121]	WideResNet-38	78.4
PSPNet [54]	ResNet-101	85.4
BiSeNet [128]	ResNet-101	78.9
DFN [99]	ResNet-101	79.3
PSANet [98]	ResNet-101	80.1
DenseASPP [79]	DenseNet-161	80.6
Dynamic Routing [142]	-	80.7
SPGNet [129]	2xResNet-50	81.1
DANet [91]	ResNet-101	81.5
CCNet [96]	ResNet-101	81.4
DeeplabV3 [13]	ResNet-101	81.3
IPC [141]	ResNet-101	81.8
AC-Net [131]	ResNet-101	82.3
OCR [42]	ResNet-101	82.4
ResNeSt200 [93]	ResNeSt-200	82.7
GS-CNN [130]	WideResNet	82.8
HA-Net [94]	ResNext-101	83.2
HRNetV2+OCR [42]	HRNetV2-W48	83.7
Hierarchical MSA [139]	HRNet-OCR	85.1

表头：方法、骨干网络、平均交并比（mIoU）。

表3关注 MS COCO Stuff 测试集。该数据集比 PASCAL VOC 和 Cityscapes 更具挑战性，因为最高 mIoU 约为40%。

表3 分割模型在 MS COCO Stuff 数据集上的准确率

Method	Backbone	mIoU
RefineNet [117]	ResNet-101	33.6
CCN [57]	Ladder DenseNet-101	35.7
DANet [91]	ResNet-50	37.9
DSSPN [132]	ResNet-101	37.3
EMA-Net [95]	ResNet-50	37.5
SGR [133]	ResNet-101	39.1
OCR [42]	ResNet-101	39.5
DANet [91]	ResNet-101	39.7
EMA-Net [95]	ResNet-50	39.9
AC-Net [131]	ResNet-101	40.1
OCR [42]	HRNetV2-W48	40.5

表头：方法、骨干网络、平均交并比（mIoU）。

表4关注 ADE20K 验证集。该数据集也比 PASCAL VOC 和 Cityscapes 更具挑战性。

表4　分割模型在 ADE20K 验证集上的准确率

Method	Backbone	mIoU
FCN [30]	-	29.39
DilatedNet [77]	-	32.31
CascadeNet [134]	-	34.90
RefineNet [117]	ResNet-152	40.7
PSPNet [54]	ResNet-101	43.29
PSPNet [54]	ResNet-269	44.94
EncNet [116]	ResNet-101	44.64
SAC [135]	ResNet-101	44.3
PSANet [98]	ResNet-101	43.70
UperNet [136]	ResNet-101	42.66
DSSPN [132]	ResNet-101	43.68
DM-Net [56]	ResNet-101	45.50
AC-Net [131]	ResNet-101	45.90
ResNeSt-101 [93]	ResNeSt-101	46.91
ResNeSt-200 [93]	ResNeSt-200	48.36

表头：方法、骨干网络、平均交并比（mIoU）。

表5给出主要实例分割算法在 COCO test-dev 2017 数据集上的平均精度和运行速度。

表5　实例分割模型在 COCO Test-Dev 2017 上的表现

Method	Backbone	FPS	AP
YOLACT-550 [74]	R-101-FPN	33.5	29.8
YOLACT-700 [74]	R-101-FPN	23.8	31.2
RetinaMask [172]	R-101-FPN	10.2	34.7
TensorMask [67]	R-101-FPN	2.6	37.1
SharpMask [173]	R-101-FPN	8.0	37.4
Mask-RCNN [62]	R-101-FPN	10.6	37.9
CenterMask [72]	R-101-FPN	13.2	38.3

表头：方法、骨干网络、每秒帧数（FPS）、平均精度（AP）。

表6以全景质量[143]为指标，列出主要全景分割算法在 MS COCO 验证集上的表现。

表6　全景分割模型在 MS COCO 验证集上的表现

Method	Backbone	PQ
Panoptic FPN [144]	ResNet-50	39.0
Panoptic FPN [144]	ResNet-101	40.3
AU-Net [145]	ResNet-50	39.6
Panoptic-DeepLab [147]	Xception-71	39.7
OANet [174]	ResNet-50	39.0
OANet [174]	ResNet-101	40.7
AdaptIS [175]	ResNet-50	35.9
AdaptIS [175]	ResNet-101	37.0
UPSNet* [148]	ResNet-50	42.5
OCFusion* [176]	ResNet-50	41.3
OCFusion* [176]	ResNet-101	43.0
OCFusion* [176]	ResNeXt-101	45.7

表头：方法、骨干网络、全景质量（PQ）。

* 表示使用可变形卷积。

最后，表7汇总了几种主要 RGB-D 分割模型在 NYUDv2 和 SUN-RGBD 数据集上的表现。

表7　分割模型在 NYUD-v2 和 SUN-RGBD 数据集上的表现

	NYUD-v2		SUN-RGBD	
Method	m-Acc	m-IoU	m-Acc	m-IoU
Mutex [177]	-	31.5	-	-
MS-CNN [178]	45.1	34.1	-	-
FCN [30]	46.1	34.0	-	-
Joint-Seg [179]	52.3	39.2	-	-
SegNet [25]	-	-	44.76	31.84
Structured Net [39]	53.6	40.6	53.4	42.3
B-SegNet [180]	-	-	45.9	30.7
3D-GNN [181]	55.7	43.1	57.0	45.9
LSD-Net [46]	60.7	45.9	58.0	-
RefineNet [117]	58.9	46.5	58.5	45.9
D-aware CNN [182]	61.1	48.4	53.5	42.0
MTI-Net [183]	62.9	49	-	-
RDFNet [184]	62.8	50.1	60.1	47.7
ESANet-R34-NBt1D [140]	-	50.3	-	48.17
G-Aware Net [185]	68.7	59.6	74.9	54.5

两组指标分别对应 NYUD-v2、SUN-RGBD；m-Acc 为平均准确率，m-IoU 为平均交并比。

总体来看，在过去5–6年里，深度分割模型的性能显著进步，在不同数据集上的 mIoU 相对提升了25%–42%。不过，一些论文由于种种原因难以复现：有的在非标准的基准或数据库上报告性能，有的仅在常用基准测试集上任意选取子集，有的没有充分描述实验设置，有时还只在一部分对象类别上评估模型。最重要的是，许多论文没有提供模型实现的源代码。随着深度学习模型日益普及，情况有所好转：许多研究团队开始采用可复现的框架，并公开其实现的源代码。

6 挑战与机遇

图像分割无疑从深度学习中获益良多，但还面临若干挑战。接下来，我们讨论一些有前景的研究方向。我们认为这些方向有助于进一步推进图像分割算法。

6.1 更具挑战性的数据集

已有多个为语义分割和实例分割构建的大规模图像数据集，不过仍需要更具挑战性的数据集，以及覆盖不同图像类型的数据集。对于静态图像，包含大量对象且对象彼此重叠的数据集将很有价值。真实场景中的对象常有大面积重叠，这类数据有助于训练出更善于处理密集对象和大量重叠的模型。随着三维图像分割越来越普及，特别是在医学图像分析中，对大规模带标注三维图像数据集的需求也很强；这类数据集比低维数据集更难构建。

6.2 结合深度学习与早期分割模型

目前已有广泛共识：深度学习分割算法的性能提升正在趋于停滞，医学图像分析等特定应用领域尤其如此。若要进一步提高性能，必须更深入地探索怎样将基于 CNN 的图像分割模型与有代表性的基于模型的“经典”图像分割方法结合起来。已有研究探讨 CNN 与图模型的结合，但把 CNN 与主动轮廓、图割及其他分割模型相结合的工作出现较晚，值得继续研究。

6.3 可解释的深度模型

虽然深度学习模型已在有挑战性的基准上表现出良好的潜力，但仍有一些问题没有解答。例如，深度模型究竟学到了什么？该如何解释模型学到的特征？对于给定数据集，要达到某一分割准确率，最小的神经网络架构是什么？尽管已有一些技术能把这些模型学到的卷积核可视化，人们还缺乏对模型内部行为与动态过程的全面研究。更深入理解其理论基础，有助于为不同的分割场景开发更合适的模型。

6.4 弱监督学习与无监督学习

弱监督学习（原文此处亦称 few-shot learning，即少样本学习）[186]与无监督学习[187]正成为非常活跃的研究方向。它们有望对图像分割特别有价值，因为许多应用领域很难收集逐像素精确标注的训练图像，医学图像分析尤其如此。迁移学习的做法是先用大量有标注的样本（可能来自公开基准）训练一个通用分割模型，再用目标应用中的少量样本微调。自监督学习是另一个有前景的方向，在不同领域受到广泛关注。借助自监督学习，可以提取图像中的许多细节，从而用少得多的训练样本来训练分割模型。基于强化学习的模型也可能是一个值得探索的方向，因为它们在图像分割中迄今很少受到关注。例如，MOREL[188]提出使用深度强化学习来分割视频中的运动对象。

6.5 面向各类应用的实时模型

许多应用首要关注准确率，但有些应用同样要求分割模型近实时运行，或达到常用相机的帧率（至少25帧/秒）。例如，这对部署在自动驾驶车辆上的计算机视觉系统很有用。当时大多数模型远达不到这一帧率：例如 FCN-8 处理一幅低分辨率图像约需100 ms。基于空洞卷积的模型在一定程度上提高了分割速度，但仍有很大改进空间。

6.6 节省内存的模型

许多现代分割模型即使在推理阶段也需要大量内存。以往的大量工作集中在提高这类模型的准确率；但要让模型在手机等特定设备上运行，必须简化网络。可以使用较简单的模型，或采用模型压缩技术；也可以先训练复杂模型，再用知识蒸馏将其压缩为一个能够模仿复杂模型、体积更小且更节省内存的网络。

6.7 应用

基于深度学习的分割已成功应用于遥感卫星图像[189]，如支持城市规划[190]和精准农业[191]。研究人员也用这类方法分割机载平台[192]和无人机[193]采集的影像，以解决重要环境问题，包括与气候变化有关的问题。遥感领域的主要挑战在于影像体量往往很大（通常由具有数百乃至数千个光谱波段的成像光谱仪采集），而评估分割算法准确率所需的真值信息又很有限。类似地，在建筑材料评估[194]中应用深度学习分割时，也面临相关图像数据量巨大、可用于验证的参考信息有限的问题。生物医学影像[195]也是深度学习分割的一个重要应用领域。在该领域，一个机会是设计标准化图像数据库，用于评估新发传染病和追踪大流行[196]。

7 结论

本文综述了基于深度学习模型的图像分割算法。这些算法在多种分割任务和基准上取得了显著成果。我们按照架构对算法分类，包括 CNN 与 FCN、RNN、R-CNN、空洞 CNN、基于注意力的模型、生成与对抗模型等。我们汇总了这些模型在 PASCAL VOC、MS COCO、Cityscapes 和 ADE20K 等常用基准上的量化性能，并讨论了基于深度学习的图像分割仍待解决的若干挑战及未来几年的有前景的研究方向。

致谢

作者感谢 Google Brain 的 Tsung-Yi Lin，以及微软亚洲研究院的 Jingdong Wang 和 Yuhui Yuan。他们提出的有益意见帮助改进了本文。

参考文献

[1] A. Rosenfeld and A. C. Kak, Digital Picture Processing. Cambridge, MA, USA: Academic Press, 1976.

[2] R. Szeliski, Computer Vision: Algorithms and Applications. Berlin, Germany: Springer, 2010.

[3] D. Forsyth and J. Ponce, Computer Vision: A Modern Approach. Upper Saddle River, NJ, USA: Prentice Hall, 2002.

[4] N. Otsu, “A threshold selection method from gray-level histograms,” IEEE Trans. Syst. Man Cybern., vol. SMC-9, no. 1, pp. 62– 66, Jan. 1979.

[5] R. Nock and F. Nielsen, “Statistical region merging,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 11, pp. 1452–1458, Nov. 2004.

[6] N. Dhanachandra, K. Manglem, and Y. J. Chanu, “Image segmentation using K-means clustering algorithm and subtractive clustering algorithm,” Procedia Comput. Sci., vol. 54, pp. 764–771, 2015.

[7] L. Najman and M. Schmitt, “Watershed of a continuous function,” Signal Process., vol. 38, no. 1, pp. 99–112, 1994.

[8] M. Kass, A. Witkin, and D. Terzopoulos, “Snakes: Active contour models,” Int. J. Comput. Vis., vol. 1, no. 4, pp. 321–331, 1988.

[9] Y. Boykov, O. Veksler, and R. Zabih, “Fast approximate energy minimization via graph cuts,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[10] N. Plath, M. Toussaint, and S. Nakajima, “Multi-class image segmentation using conditional random fields and global classification,” in Proc. 26th Int. Conf. Mach. Learn., 2009, pp. 817–824.

[11] J.-L. Starck, M. Elad, and D. L. Donoho, “Image decomposition via the combination of sparse representations and a variational approach,” IEEE Trans. Image Process., vol. 14, no. 10, pp. 1570– 1582, Oct. 2005.

[12] S. Minaee and Y. Wang, “An ADMM approach to masked signal decomposition using subspace representation,” IEEE Trans. Image Process., vol. 28, no. 7, pp. 3192–3204, Jul. 2019.

[13] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017, arXiv: 1706.05587.

[14] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, “Image segmentation using deep learning: A survey,” 2020, arXiv: 2001.05566.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[16] K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” Biol. Cybern., vol. 36, no. 4, pp. 193–202, 1980.

[17] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, “Phoneme recognition using time-delay neural networks,” IEEE Trans. Acoust. Speech Signal Process., vol. 37, no. 3, pp. 328–339, Mar. 1989.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. 25th Int. Conf. Neural Inf. Process. Syst., 2012, pp. 1097–1105.

[19] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556.

[20] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[21] 2015. [Online]. Available: https://colah.github.io/posts/2015– 08-Understanding-LSTMs/

[22] D. E. Rumelhart et al., “Learning representations by back-propagating errors,” Cogn. Model., vol. 5, no. 3, 1988, Art. no. 1.

[23] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[24] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[25] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[26] I. Goodfellow et al., “Generative adversarial nets,” in Proc. 27th Int. Conf. Neural Inf. Process. Syst., 2014, pp. 2672–2680.

[27] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” 2015, arXiv:1511.06434.

[28] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” 2014, arXiv:1411.1784.

[29] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN,” 2017, arXiv: 1701.07875.

[30] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 3431–3440.

[31] G. Wang, W. Li, S. Ourselin, and T. Vercauteren, “Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks,” in Proc. Int. MICCAI Brainlesion Workshop, 2017, pp. 178–190.

[32] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, “Fully convolutional instance-aware semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2359–2367.

[33] Y. Yuan, M. Chao, and Y.-C. Lo, “Automatic skin lesion segmentation using deep fully convolutional networks with Jaccard distance,” IEEE Trans. Med. Imag., vol. 36, no. 9, pp. 1876–1886, Sep. 2017.

[34] N. Liu, H. Li, M. Zhang, J. Liu, Z. Sun, and T. Tan, “Accurate iris segmentation in non-cooperative environments using fully convolutional networks,” in Proc. Int. Conf. Biometrics, 2016, pp. 1–8.

[35] W. Liu, A. Rabinovich, and A. C. Berg, “ParseNet: Looking wider to see better,” 2015, arXiv:1506.04579.

[36] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected CRFs,” 2014, arXiv:1412.7062.

[37] A. G. Schwing and R. Urtasun, “Fully connected deep structured networks,” 2015, arXiv:1503.02351.

[38] S. Zheng et al., “Conditional random fields as recurrent neural networks,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1529– 1537.

[39] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid, “Efficient piece-wise training of deep structured models for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 3194–3203.

[40] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, “Semantic image segmentation via deep parsing network,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1377–1385.

[41] H. Noh, S. Hong, and B. Han, “Learning deconvolution network for semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1520–1528.

[42] Y. Yuan, X. Chen, and J. Wang, “Object-contextual representations for semantic segmentation,” 2019, arXiv: 1909.11065.

[43] J. Fu, J. Liu, Y. Wang, J. Zhou, C. Wang, and H. Lu, “Stacked deconvolutional network for semantic segmentation,” IEEE Trans. Image Process., early access, Jan. 25, 2019, doi: 10.1109/ TIP.2019.2895460.

[44] A. Chaurasia and E. Culurciello, “LinkNet: Exploiting encoder representations for efficient semantic segmentation,” in Proc. IEEE Int. Conf. Visual Commun. Image Process., 2017, pp. 1–4.

[45] X. Xia and B. Kulis, “W-Net: A deep model for fully unsupervised image segmentation,” 2017, arXiv: 1711.08506.

[46] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, “Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 3029–3037.

[47] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention, 2015, pp. 234– 241.

[48] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A nested U-Net architecture for medical image segmentation,” in Proc. Int. Workshop Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support, 2018, pp. 3–11.

[49] Z. Zhang, Q. Liu, and Y. Wang, “Road extraction by deep residual U-Net,” IEEE Geosci. Remote Sens. Lett., vol. 15, no. 5, pp. 749– 753, May 2018. Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronne-

[50] berger, “3D U-Net: Learning dense volumetric segmentation from sparse annotation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention, 2016, pp. 424–432.

[51] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” in Proc. 4th Int. Conf. 3D Vis., 2016, pp. 565–571.

[52] T. Brosch, L. Y. Tang, Y. Yoo, D. K. Li, A. Traboulsee, and R. Tam, “Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation,” IEEE Trans. Med. Imag., vol. 35, no. 5, pp. 1229– 1239, May 2016.

[53] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection ,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2117– 2125.

[54] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2881–2890.

[55] G. Ghiasi and C. C. Fowlkes, “Laplacian pyramid reconstruction and refinement for semantic segmentation,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 519–534.

[56] J. He, Z. Deng, and Y. Qiao, “Dynamic multi-scale filters for semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 3562–3572.

[57] H. Ding, X. Jiang, B. Shuai, A. Qun Liu, and G. Wang, “Context contrasted feature and gated multi-scale aggregation for scene segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 2393–2402.

[58] J. He, Z. Deng, L. Zhou, Y. Wang, and Y. Qiao, “Adaptive pyramid context network for semantic segmentation,” in Proc. Conf. Comput. Vis. Pattern Recognit., 2019, pp. 7519–7528.

[59] D. Lin, Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang, “Multi-scale context intertwining for semantic segmentation,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 603–619.

[60] G. Li, Y. Xie, L. Lin, and Y. Yu, “Instance-level salient object segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2386–2395.

[61] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. 28th Int. Conf. Neural Inf. Process. Syst., 2015, pp. 91–99.

[62] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2961–2969.

[63] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path aggregation network for instance segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8759–8768.

[64] J. Dai, K. He, and J. Sun, “Instance-aware semantic segmentation via multi-task network cascades,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 3150–3158.

[65] R. Hu, P. Dollár, K. He, T. Darrell, and R. Girshick, “Learning to segment every thing,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4233–4241.

[66] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam, “MaskLab: Instance segmentation by refining object detection with semantic and direction features,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018 , pp. 4013–4022.

[67] X. Chen, R. Girshick, K. He, and P. Dollár, “TensorMask: A foundation for dense object segmentation,” 2019, arXiv: 1903.12174.

[68] J. Dai, Y. Li, K. He, and J. Sun, “R-FCN: Object detection via region-based fully convolutional networks,” in Proc. 30th Int. Conf. Neural Inf. Process. Syst., 2016, pp. 379–387.

[69] P. O. Pinheiro, R. Collobert, and P. Dollár, “Learning to segment object candidates,” in Proc. 28th Int. Conf. Neural Inf. Process. Syst., 2015, pp. 1990–1998.

[70] E. Xie et al., “PolarMask: Single shot instance segmentation with polar representation,” 2019, arXiv: 1909.13226.

[71] Z. Hayder, X. He, and M. Salzmann, “Boundary-aware instance segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5696–5704.

[72] Y. Lee and J. Park, “CenterMask: Real-time anchor-free instance segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 13 906–13 915.

[73] M. Bai and R. Urtasun, “Deep watershed transform for instance segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5221–5229.

[74] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, “YOLACT: Real-time instance segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 9157–9166.

[75] A. Fathi et al., “Semantic instance segmentation via deep metric learning,” 2017, arXiv: 1703.10277.

[76] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, Apr. 2018.

[77] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” 2015, arXiv:1511.07122.

[78] P. Wang et al., “Understanding convolution for semantic segmentation,” in Proc. IEEE Winter Conf. Appl. Comput. Vis., 2018, pp. 1451–1460.

[79] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, “DenseASPP for semantic segmentation in street scenes,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 3684–3692.

[80] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, “ENet: A deep neural network architecture for real-time semantic segmentation,” 2016, arXiv:1606.02147.

[81] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 801–818.

[82] F. Visin et al., “ReSeg: A recurrent neural network-based model for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2016, pp. 41–48.

[83] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. Courville, and Y. Bengio, “ReNet: A recurrent neural network based alternative to convolutional networks,” 2015, arXiv:1505.00393.

[84] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki, “Scene labeling with LSTM recurrent neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 3547–3555.

[85] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, “Semantic object parsing with graph LSTM,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 125–143.

[86] Y. Xiang and D. Fox, “DA-RNN: Semantic mapping with data associated recurrent neural networks,” 2017, arXiv: 1703.03098.

[87] R. Hu, M. Rohrbach, and T. Darrell, “Segmentation from natural language expressions,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 108–124.

[88] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, “Attention to scale: Scale-aware semantic image segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 3640–3649.

[89] Q. Huang et al., “Semantic segmentation with reverse attention,” 2017, arXiv: 1707.06426.

[90] H. Li, P. Xiong, J. An, and L. Wang, “Pyramid attention network for semantic segmentation,” 2018, arXiv: 1805.10180.

[91] J. Fu et al., “Dual attention network for scene segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 3146– 3154.

[92] Y. Yuan and J. Wang, “OCNet: Object context network for scene parsing,” 2018, arXiv: 1809.00916.

[93] H. Zhang et al., “ResNeSt: Split-attention networks,” 2020, arXiv: 2004.08955.

[94] S. Choi, J. T. Kim, and J. Choo, “Cars can’t fly up in the sky: Improving urban-scene segmentation via height-driven attention networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 9373–9383.

[95] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, “Expectation-maximization attention networks for semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 9167–9176.

[96] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “CCNet: Criss-cross attention for semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 603–612.

[97] M. Ren and R. S. Zemel, “End-to-end instance segmentation with recurrent attention,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6656–6664.

[98] H. Zhao et al., “PSANet: Point-wise spatial attention network for scene parsing,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 267–283.

[99] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, “Learning a discriminative feature network for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 1857– 1866.

[100] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, “Semantic segmentation using adversarial networks,” 2016, arXiv:1611.08408.

[101] N. Souly, C. Spampinato, and M. Shah, “Semi supervised semantic segmentation using generative adversarial network,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 5688–5696.

[102] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, “Adversarial learning for semi-supervised semantic segmentation,” 2018, arXiv: 1802.07934.

[103] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang, “SegAN: Adversarial network with multi-scale L1 loss for medical image segmentation,” Neuroinformatics, vol. 16, no. 3/4, pp. 383–392, 2018.

[104] M. Majurski et al., “Cell image segmentation using generative adversarial networks, transfer learning, and augmentations,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2019, pp. 1114–1122.

[105] K. Ehsani, R. Mottaghi, and A. Farhadi, “SegAN: Segmenting and generating the invisible,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6144–6153.

[106] T. F. Chan and L. A. Vese, “Active contours without edges,” IEEE Trans. Image Process., vol. 10, no. 2, pp. 266–277, Feb. 2001.

[107] X. Chen, B. M. Williams, S. R. Vallabhaneni, G. Czanner, R. Williams, and Y. Zheng, “Learning active contour models for medical image segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 11 632–11 640.

[108] S. Gur, L. Wolf, L. Golgher, and P. Blinder, “Unsupervised microvascular image segmentation using an active contours mimicking neural network,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 10 722–10 731.

[109] P. Marquez-Neila, L. Baumela, and L. Alvarez, “A morphological approach to curvature-based evolution of curves and surfaces,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 1, pp. 2–17, Jan. 2014.

[110] T. H. N. Le, K. G. Quach, K. Luu, C. N. Duong, and M. Savvides, “Reformulating level sets as deep recurrent neural network approach to semantic segmentation,” IEEE Trans. Image Process., vol. 27, no. 5, pp. 2393–2407, May 2018.

[111] C. Rupprecht, E. Huaroc, M. Baust, and N. Navab, “Deep active contours,” 2016, arXiv:1607.05074.

[112] A. Hatamizadeh et al., “Deep active lesion segmentation,” in Proc. Int. Workshop Mach. Learn. Med. Imag., 2019, vol. 11861, pp. 98–105.

[113] D. Marcos et al., “Learning deep structured active contours end-to-end,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8877–8885.

[114] D. Cheng, R. Liao, S. Fidler, and R. Urtasun, “DARNet: Deep active ray network for building segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 7431–7439.

[115] A. Hatamizadeh, D. Sengupta, and D. Terzopoulos, “End-to-end trainable deep active contour models for automated image segmentation: Delineating buildings in aerial imagery,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 730–746.

[116] H. Zhang et al., “Context encoding for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7151– 7160.

[117] G. Lin, A. Milan, C. Shen, and I. Reid, “RefineNet: Multi-path refinement networks for high-resolution semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1925– 1934.

[118] G. Song, H. Myeong, and K. M. Lee, “SeedNet: Automatic seed generation with deep reinforcement learning for robust interactive segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 1760–1768.

[119] J. Dai, K. He, and J. Sun, “BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1635–1643.

[120] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, “Large kernel matters — Improve semantic segmentation by global convolutional network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4353–4361.

[121] Z. Wu, C. Shen, and A. Van Den Hengel, “Wider or deeper: Revisiting the ResNet model for visual recognition,” Pattern Recognit., vol. 90, pp. 119–133, 2019.

[122] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, “ExFuse: Enhancing feature fusion for semantic segmentation,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 269–284.

[123] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, “Feedforward semantic segmentation with zoom-out features,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 3376– 3385.

[124] W. Wang, J. Shen, and F. Porikli, “Saliency-aware geodesic video object segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 3395–3402.

[125] P. Luo, G. Wang, L. Lin, and X. Wang, “Deep dual learning for semantic image segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2718–2726.

[126] X. Li et al., “FoveaNet: Perspective-aware urban scene parsing,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 784–792.

[127] I. Kreso, S. Segvic, and J. Krapac, “Ladder-style densenets for semantic segmentation of large natural images,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 238–245.

[128] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, “BiSeNet: Bilateral segmentation network for real-time semantic segmentation,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 325–341.

[129] B. Cheng et al., “SPGNet: Semantic prediction guidance for scene parsing,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 5218–5228.

[130] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, “Gated-SCNN: Gated shape CNNs for semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 5229–5238.

[131] J. Fu et al., “Adaptive context network for scene parsing,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 6748–6757.

[132] X. Liang, H. Zhou, and E. Xing, “Dynamic-structured semantic propagation network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 752–761.

[133] X. Liang, Z. Hu, H. Zhang, L. Lin, and E. P. Xing, “Symbolic graph reasoning meets convolutions,” in Proc. 32nd Int. Conf. Neural Inf. Process. Syst., 2018, pp. 1853–1863.

[134] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, “Scene parsing through ADE20K dataset,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1858–1868.

[135] R. Zhang, S. Tang, Y. Zhang, J. Li, and S. Yan, “Scale-adaptive convolutions for scene parsing,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2031–2039.

[136] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, “Unified perceptual parsing for scene understanding,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 418–434.

[137] B. Zoph et al., “Rethinking pre-training and self-training,” 2020, arXiv: 2006.06882.

[138] X. Zhang, H. Xu, H. Mo, J. Tan, C. Yang, and W. Ren, “DCNAS: Densely connected neural architecture search for semantic image segmentation,” 2020, arXiv: 2003.11883.

[139] A. Tao, K. Sapra, and B. Catanzaro, “Hierarchical multi-scale attention for semantic segmentation,” 2020, arXiv: 2005.10821.

[140] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H.-M. Gross, “Efficient RGB-D semantic segmentation for indoor scene analysis,” 2020, arXiv: 2011.06961.

[141] M. Zhen et al., “Joint semantic segmentation and boundary detection using iterative pyramid contexts,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 13 666–13 675.

[142] Y. Li et al., “Learning dynamic routing for semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 8553–8562.

[143] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, “Panoptic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 9404–9413.

[144] A. Kirillov, R. Girshick, K. He, and P. Dollar, “Panoptic feature pyramid networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 6399–6408.

[145] Y. Li et al., “Attention-guided unified network for panoptic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 7019–7028.

[146] L. Porzi, S. R. Bulo, A. Colovic, and P. Kontschieder, “Seamless scene segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 8277–8286.

[147] B. Cheng et al., “Panoptic-DeepLab,” 2019, arXiv: 1910.04751.

[148] Y. Xiong et al., “UPSNet: A unified panoptic segmentation network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 8818–8826.

[149] R. Mohan and A. Valada, “EfficientPS: Efficient panoptic segmentation,” 2020, arXiv: 2004.02307.

[150] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The PASCAL visual object classes (VOC) challenge,” Int. J. Comput. Vis., vol. 88, pp. 303–338, 2010.

[151] 2012. [Online]. Available: http://host.robots.ox.ac.uk/pascal/ VOC/voc2012/

[152] R. Mottaghi et al., “The role of context for object detection and semantic segmentation in the wild,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 891–898.

[153] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 740–755.

[154] M. Cordts et al., “The cityscapes dataset for semantic urban scene understanding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 3213–3223.

[155] C. Liu, J. Yuen, and A. Torralba, “Nonparametric scene parsing: Label transfer via dense scene alignment,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 1972–1979.

[156] S. Gould, R. Fulton, and D. Koller, “Decomposing a scene into geometric and semantically consistent regions,” in Proc. Int. Conf. Comput. Vis., 2009, pp. 1–8.

[157] D. Martin, C. Fowlkes, D. Tal, and J. Malik, “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics,” in Proc. Int. Conf. Comput. Vis., 2001, vol. 2, pp. 416– 423.

[158] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, “Learning object class detectors from weakly annotated video,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2012, pp. 3282– 3289.

[159] S. D. Jain and K. Grauman, “Supervoxel-consistent foreground propagation in video,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 656–671.

[160] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The KITTI dataset,” The Int. J. Robot. Res., vol. 32, no. 11, pp. 1231–1237, 2013.

[161] J. M. Alvarez, T. Gevers, Y. LeCun, and A. M. Lopez, “Road scene segmentation from a single image,” in Proc. Eur. Conf. Comput. Vis., 2012, pp. 376–389.

[162] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, “Semantic contours from inverse detectors,” in Proc. Int. Conf. Comput. Vis., 2011, pp. 991–998.

[163] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, “Detect what you can: Detecting and representing objects using holistic models and body parts,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 1971–1978.

[164] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, “The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 3234–3243.

[165] X. Shen et al., “Automatic portrait segmentation for image stylization,” Comput. Graph. Forum, vol. 35, no. 2, pp. 93–102, 2016.

[166] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from RGBD images,” in Proc. Eur. Conf. Comput. Vis., 2012, pp. 746–760.

[167] J. Xiao, A. Owens, and A. Torralba, “Sun3D: A database of big spaces reconstructed using SFM and object labels,” in Proc. IEEE Int. Conf. Comput. Vis., 2013, pp. 1625–1632.

[168] S. Song, S. P. Lichtenberg, and J. Xiao, “Sun RGB-D: A RGB-D scene understanding benchmark suite,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 567–576.

[169] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, “ScanNet: Richly-annotated 3D reconstructions of indoor scenes,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5828–5839.

[170] I. Armeni, A. Sax, A. Zamir, and S. Savarese, “Joint 2D-3D-semantic data for indoor scene understanding,” Feb. 2017, ArXiv e-prints.

[171] K. Lai, L. Bo, X. Ren, and D. Fox, “A large-scale hierarchical multi-view RGB-D object dataset,” in Proc. IEEE Int. Conf. Robot. Autom., 2011, pp. 1817–1824.

[172] C.-Y. Fu, M. Shvets, and A. C. Berg, “RetinaMask: Learning to predict masks improves state-of-the-art single-shot detection for free,” 2019, arXiv: 1901.03353.

[173] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár, “Learning to refine object segments,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 75–91.

[174] H. Liu et al., “An end-to-end network for panoptic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 6172–6181.

[175] K. Sofiiuk, O. Barinova, and A. Konushin, “AdaptIS: Adaptive instance selection network,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 7355–7363.

[176] J. Lazarow, K. Lee, K. Shi, and Z. Tu, “Learning instance occlusion for panoptic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 10 720–10 729.

[177] Z. Deng, S. Todorovic, and L. J. Latecki, “Semantic segmentation of RGBD images with mutex constraints,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1733–1741.

[178] D. Eigen and R. Fergus, “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 2650–2658.

[179] A. Mousavian, H. Pirsiavash, and J. Kosecka, “Joint semantic segmentation and depth estimation with deep convolutional networks,” in Proc. 4th Int. Conf. 3D Vis., 2016, pp. 611–619.

[180] A. Kendall, V. Badrinarayanan, and R. Cipolla, “Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding,” 2015, arXiv:1511.02680.

[181] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, “3D graph neural networks for RGBD semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 5199–5208.

[182] W. Wang and U. Neumann, “Depth-aware CNN for RGB-D segmentation,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 135–150.

[183] S. Vandenhende, S. Georgoulis, and L. Van Gool, “MTI-net: Multi-scale task interaction networks for multi-task learning,” 2020, arXiv: 2001.06902.

[184] S.-J. Park, K.-S. Hong, and S. Lee, “RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 4980–4989.

[185] J. Jiao, Y. Wei, Z. Jie, H. Shi, R. W. Lau, and T. S. Huang, “Geometry-aware distillation for indoor semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 2869–2878.

[186] Z.-H. Zhou, “A brief introduction to weakly supervised learning,” Nat. Sci. Rev., vol. 5, no. 1, pp. 44–53, 2018.

[187] L. Jing and Y. Tian, “Self-supervised visual feature learning with deep neural networks: A survey,” IEEE Trans. Pattern Anal. Mach. Intell., early access, May 4, 2020, doi: 10.1109/TPAMI.2020.2992393.

[188] V. Goel, J. Weng, and P. Poupart, “Unsupervised video object segmentation for deep reinforcement learning,” in Proc. 32nd Int. Conf. Neural Inf. Process. Syst., 2018, pp. 5683–5694.

[189] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, “Deep learning in remote sensing applications: A meta-analysis and review,” ISPRS J. Photogrammetry Remote Sens., vol. 152, pp. 166– 177, 2019.

[190] L. Gao, Y. Zhang, F. Zou, J. Shao, and J. Lai, “Unsupervised urban scene segmentation via domain adaptation,” Neurocomputing, vol. 406, pp. 295–301, 2020.

[191] M. Paoletti, J. Haut, J. Plaza, and A. Plaza, “Deep learning classifiers for hyperspectral imaging: A review,” ISPRS J. Photogrammetry Remote Sens., vol. 158, pp. 279–317, 2019.

[192] J. F. Abrams et al., “Habitat-Net: Segmentation of habitat images using deep learning,” Ecological Informat., vol. 51 , pp. 121–128, 2019.

[193] M. Kerkech, A. Hafiane, and R. Canals, “Vine disease detection in UAV multispectral images using optimized image registration and deep learning segmentation approach,” Comput. Electron. Agriculture, vol. 174, 2020, Art. no. 105446.

[194] Y. Song, Z. Huang, C. Shen, H. Shi, and D. A. Lange, “Deep learning-based automated image segmentation for concrete petrographic analysis,” Cement Concrete Res., vol. 135, 2020, Art. no. 106118.

[195] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, “Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation,” Med. Image Anal., vol. 63, 2020, Art. no. 101693.

[196] A. Amyar, R. Modzelewski, H. Li, and S. Ruan, “Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation,” Comput. Biol. Med., vol. 126, 2020, Art. no. 104037.

作者简介



Shervin Minaee

Shervin Minaee（IEEE 会员）于2018年在纽约大学获得电气工程与计算机科学博士学位。论文发表时，他是 Snapchat 公司计算机视觉团队的机器学习负责人。他的研究兴趣包括计算机视觉、图像分割、生物特征识别和深度学习应用。读博期间，他发表论文和专利共40余项。他曾任三星研究院、AT&T 实验室和华为实验室的研究科学家，也曾任 Expedia 集团数据科学家。他为 IEEE、ACM、Elsevier 和 Springer 的20多种计算机视觉相关期刊担任审稿人。他获得多项荣誉，包括2017年三星美国研究院最佳研究报告奖和2016年 Verizon 开放创新挑战奖。



Yuri Boykov

Yuri Boykov（IEEE 会员）在论文发表时任加拿大滑铁卢大学 Cheriton 计算机科学学院教授。他的研究领域包括计算机视觉和生物医学图像分析，重点研究结构化分割、复原、配准、立体视觉、运动、模型拟合、识别、照片与视频编辑以及其他数据分析问题中的建模与优化。他是《International Journal of Computer Vision》（IJCV）的编辑。他的工作被《IEEE Transactions on Pattern Analysis and Machine Intelligence》选为30年十大最具影响力的论文之一。2017年，Google Scholar 将他2006年发表的分割论文列为“计算机视觉与模式识别领域的经典论文”。2011年，他获 IEEE 的 Helmholtz 奖和国际计算机视觉会议的时间检验奖（Test of Time Award）。



Fatih Porikli

Fatih Porikli（IEEE 会士）于2002年在纽约大学获得博士学位。论文发表时，他担任美国圣迭戈高通公司的高级总监。他曾任澳大利亚国立大学工程研究学院正教授，且在此之前不久还担任圣迭戈华为消费者业务集团（CBG）终端硬件部门副总裁。他曾领导澳大利亚 NICTA 的计算机视觉研究组，并任马萨诸塞州剑桥三菱电机研究实验室的杰出研究科学家。他获2006年 R&D 100 年度科学家奖及6项最佳论文奖，发表250多篇论文，合编2部书，并有100多项发明专利。截至论文发表时，他在过去15年里担任过多场 IEEE 会议的大会主席及技术程序主席，并担任过 IEEE 和 Springer 领先期刊的副编辑。



Antonio Plaza

Antonio Plaza（IEEE 会士）分别于1999年和2002年在西班牙埃斯特雷马杜拉大学计算机与通信技术系获得计算机工程硕士和博士学位，论文发表时任该系教授。他发表了600余项学术成果，其中包括300篇 JCR 期刊论文（超过170篇发表于 IEEE 期刊）、24个书籍章节和300多篇经同行评审的会议论文。他曾获《IEEE Signal Processing Magazine》2015年最佳专栏奖、《IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing》2013年最佳论文奖，并在《Journal of Parallel and Distributed Computing》上发表了该刊2005–2010年间被引次数最多的论文。他入选2018、2019及2020年的全球高被引研究人员名单。



Nasser Kehtarnavaz

Nasser Kehtarnavaz（IEEE 会士）在论文发表时担任美国得克萨斯州理查森市得克萨斯大学达拉斯分校电气与计算机工程系 Erik Jonsson 杰出教授，同时担任嵌入式机器学习实验室主任。他的研究兴趣包括信号与图像处理、机器学习、深度学习及在嵌入式处理器上的实时实现。他在这些领域独著或合著10部书，以及400余项期刊论文、会议论文、专利、手册及社论。他是 SPIE 会士和持照专业工程师，并任《Journal of Real-Time Image Processing》主编。



Demetri Terzopoulos

Demetri Terzopoulos（IEEE 会士）于1984年在美国马萨诸塞州剑桥的麻省理工学院（MIT）获得人工智能博士学位。论文发表时，他在加利福尼亚大学洛杉矶分校任计算机科学杰出教授及校长讲席教授，领导该校的计算机图形与视觉实验室，同时是 VoxelCloud 公司联合创始人兼首席科学家。他曾获古根海姆研究基金，是或曾是 ACM、IETI、加拿大皇家学会和伦敦皇家学会的会士，以及欧洲科学院、纽约科学院和 Sigma Xi 的成员。他获得多项荣誉，其中包括美国电影艺术与科学学院颁发的奥斯卡奖，表彰他在基于物理的计算机动画方面的开创性工作；IEEE 也颁给他计算机先驱奖、Helmholtz 奖与首届计算机视觉杰出研究者奖，表彰他对可变形模型及其应用所作的开创性、持续性研究。“可变形模型”一词由他提出，已收入 IEEE 术语分类体系。

如需了解本主题或其他计算机主题，请访问 IEEE 计算机学会数字图书馆：www.computer.org/csdl。

作者单位与论文信息

Shervin Minaee 任职于 Snapchat 机器学习研究部门，美国加利福尼亚州威尼斯，邮编 90405。电子邮箱：shervin.minaee@nyu.edu。

Yuri Boykov 任职于滑铁卢大学，加拿大安大略省滑铁卢市，邮编 N2L 3G1。电子邮箱：yboykov@uwaterloo.ca。

Fatih Porikli 任职于澳大利亚国立大学（澳大利亚首都领地堪培拉，邮编 0200），并任职于华为（美国加利福尼亚州圣迭戈，邮编 92121）。电子邮箱：fatih.porikli@anu.edu.au。

Antonio Plaza 任职于埃斯特雷马杜拉大学，西班牙巴达霍斯，邮编 06006。电子邮箱：aplaza@unex.es。

Nasser Kehtarnavaz 任职于得克萨斯大学达拉斯分校，美国得克萨斯州理查森市，邮编 75080。电子邮箱：kehtar@utdallas.edu。

Demetri Terzopoulos 任职于加利福尼亚大学洛杉矶分校，美国加利福尼亚州洛杉矶，邮编 90095。电子邮箱：dt@cs.ucla.edu。

ORCID（原文作者顺序，前五位）：0000-0001-6689-9221 · 0000-0001-6374-1736 · 0000-0002-1520-4466 · 0000-0002-9613-1659 · 0000-0001-5183-6359

稿件于2020年1月18日收到，2021年1月28日修订，2021年2月7日录用。在线发布日期：2021年2月17日；当前版本日期：2022年6月3日。通信作者：Shervin Minaee。L. Wang 推荐录用。数字对象标识符（DOI）：10.1109/TPAMI.2021.3059968。

0162-8828 © 2021 IEEE。允许个人使用；重新出版或再分发须获 IEEE 许可。详见 https://www.ieee.org/publications/rights/index.html。

授权使用仅限加利福尼亚大学洛杉矶分校图书馆。于2022年6月5日 04:20:43 UTC 从 IEEE Xplore 下载，适用相关限制。