计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(9):41−52 [doi: 10.15888/j.cnki.csa.008078]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于全卷积网络的图像语义分割方法综述①

李梦怡, 朱定局

(华南师范大学 计算机学院, 广州 510631)
通讯作者: 朱定局, E-mail: zhudingju@m.scnu.edu.cn

摘　要: 自全卷积网络 (Fully Convolutional Network, FCN) 提出以后, 应用深度学习技术在图像语义分割领域受到了许多计算机视觉和机器学习研究者的关注, 现在这一方向已经成为人工智能方向的研究热点. FCN 的核心思想是搭建一个全卷积网络, 输入任意尺寸的图像, 经过模型的有效学习和推理得到相同尺寸的输出. FCN 的提出给图像语义分割领域提供了新的思路, 但也存在很多的缺点, 比如特征分辨率低、对象存在多尺度问题等. 随着研究者不断的钻研, 卷积神经网络在图像分割领域逐渐得到了优化和拓展, 基于 FCN 的主流分割框架也层出不穷. 图像语义分割对于场景理解的重要性日渐突出, 被广泛应用到无人驾驶技术、无人机领域和医疗影像检测与分析等任务中. 因此, 对图像语义分割领域的研究将值得深入研究, 使其能够更好在实际应用中大放异彩.

关键词: 图像语义分割; 全卷积网络; 深度学习; 医疗影像

引用格式: 李梦怡,朱定局.基于全卷积网络的图像语义分割方法综述.计算机系统应用,2021,30(9):41–52. http://www.c-s-a.org.cn/1003-3254/8078.html

Review on Image Semantic Segmentation Based on Fully Convolutional Network

LI Meng-Yi, ZHU Ding-Ju

(School of Computer Science, South China Normal University, Guangzhou 510631, China)

Abstract: Since the proposal of Fully Convolutional Network (FCN), applying deep learning to image semantic segmentation has attracted extensive attention from researchers in the field of computer vision and machine learning, becoming a research hotspot of artificial intelligence. The core idea of FCN is to build a fully convolutional network that accepts the input of arbitrary sizes and produces the output of the same sizes through efficient inference and learning. FCN provides a new idea for image semantic segmentation, but it also has many shortcomings, such as low feature resolution and the objects at multiple scales. As research progresses, the convolutional neural network has been gradually optimized and expanded in the field of image segmentation. In addition, the mainstream segmentation frameworks based on FCN have emerged one after another. Image semantic segmentation plays an increasingly important role in scene understanding, which is widely applied to the self-driving technique, the UAV field, detection and analysis of medical images, and other tasks. Therefore, image semantic segmentation is worth further study to better serve practical applications.

Key words: image semantic segmentation; Fully Convolutional Network (FCN); deep learning; medical image segmentation

① 基金项目: 广东省普通高校“人工智能”重点领域专项 (2019KZDZX1027); 中国高等教育学会专项课题 (2020JXD01); 广东高校省级重点平台和重大科研项目 (2017KTSCX048); 广东省中医药局科研项目 (20191411)
Foundation item: Special Project of Artificial Intelligence for Ordinary Universities of Guangdong Province (2019KZDZX1027); Special Project of China Association of Higher Education (2020JXD01); Provincial Major Science and Technology Research Program and Key Platform of Higher Education of Guangdong Province (2017KTSCX048); Research Project of Traditional Chinese Medicine Bureau of Guangdong Province (20191411)
收稿时间: 2020-12-07; 修改时间: 2021-01-11; 采用时间: 2021-01-20; csa 在线出版时间: 2021-09-02

图像语义分割是计算机视觉领域的核心任务之一, 语义分割从微观的角度可以理解为将图像中的每一个像素点进行分类, 是一种像素级的空间密集型预测任务. 换句话说, 语义分割试图在语义上理解图像中每个像素的所代表的含义, 比如识别它是汽车、楼房还是行人等; 从宏观的角度则可以将语义分割看作是将一致的语义标签分配给一类事物[1], 而不是每个像素.

与其他计算机视觉任务一样, 卷积神经网络在图像语义分割领域发挥了重大作用, 目前许多性能优异的图像分割模型都以卷积神经网络作为基础. 语义分割对于场景理解的重要性日渐突出, 被广泛应用到无人驾驶技术、地理信息系统和医疗影像检测与分析等任务中.

1 卷积神经网络

近些年来, 随着深度学习领域研究的不断发展和进步, 卷积神经网络也在许多学者的努力下得到了不断地发展和进步. 2012 年, 卷积神经网络的发展取得了历史性的突破, Kkrizhevsky 等人提出了经典的 AlexNet 模型[2], 一种更深更宽的卷积神经网络. AlexNet 的提出, 奠定了深度卷积神经网络在计算机视觉领域的核心地位, 成为图像分类、对象检测、目标跟踪和图像分割等计算机视觉领域任务中性能优异、应用广泛的深度神经网络学习模型, 推动了深度学习在语音识别、机器翻译等其他领域的拓展.

卷积神经网络主要由输入层 (input layer)、卷积层 (convolution layer)、池化层 (pooling layer)、全连接层 (fully connected layer) 和输出层 (output layer) 组成, 如图 1 所示. 通过堆叠多层卷积层和池化层, 可以得到更深的网络结构, 如 VGGNet[3]、ResNet[4] 等, 其后的全连接层也可以是多层的.

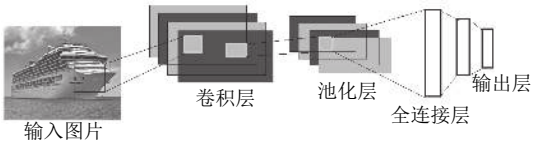


图 1 卷积神经网络结构

卷积神经网络的核心概念是卷积操作 (convolution), 卷积操作则可以看作是输入样本和卷积核的内积运算. 除了卷积层之外, 常用的卷积网络中通常还包含池化层, 作用是减小卷积层产生的特征图的尺寸. 最常用的池化操作主要包括: 最大池化 (max-pooling)、平均池化 (avg-pooling) 和全局池化 (global-pooling).

此外, 卷积神经网络在输出层之前会连接全连接层, 通过全连接层完成分类输出. 在全连接层中通常使用激活函数计算输出值, 激活函数通常包括 Sigmoid、tanh、ReLU 等函数.

卷积神经网络的出现和发展使得人工智能领域攀登上了一个新的台阶, 但是卷积神经网络存在自身的缺陷, 如何更好的完善和改进卷积神经网络仍然任重而道远.

2 经典深度学习分类网络

许多性能优越的语义分割模型都是在一些经典的深层网络基础上提出和改进的, 例如 AlexNet、VGGNet、GoogLeNet 和 ResNet 等, 因此在本节主要回顾一下这些经典的分类网络.

2.1 AlexNet

卷积神经网络受到学术界和工业界的广泛关注和研究得益于 AlexNet 的提出和应用. 2012 年, Krizhevsky 等人提出了有重要影响力的卷积神经网络模型 AlexNet[2], 并用该模型参加了 ImageNet ILSVRC 比赛, 在比赛中得到了 top-1 中 37.5% 和 top-5 中 17.0% 的错误率, 远超于当时其他人所提出的模型结果, 比赛成绩斐然.

AlexNet 模型的结构图如图 2 所示, 从图中可以看出 AlexNet 模型包含输入层、5 个卷积层和 3 个全连接层, 其中有 3 个卷积层进行了最大池化, 并且首次采用修正线性单元 (Rectified Linear Unit, ReLU) 作为激活函数.

AlexNet 模型的出现, 为人们在提取特征进行图像检索、图像分类、图像识别等方面提供了一种深度卷积神经网络的思路, 也为后来的研究学者在设计卷积神经网络模型提供了启发:

(1) 模型的深度与宽度可决定网络的能力;

(2) 使用强大的 GPU 以及大规模的数据量可以进一步提高网络的训练能力;

(3) 数据增强可以有效地人工增大数据集, 减少过拟合;

(4) ReLU 作为激活函数不需要对输入进行标准化处理来防止饱和现象;

(5) 采用 LRN、Dropout 层和重叠池化等训练技巧可以减少过拟合, 提高精度和模型的泛化能力, 提升了特征的丰富性;

(6) 网络结构具体相关性, 不可轻易移除某一层.

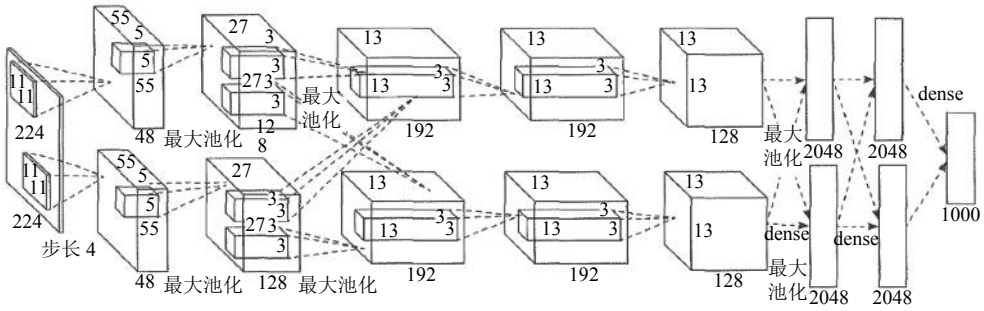


图 2 AlexNet 模型结构示意图

2.2 VGGNet

AlexNet 的成功极大地推进了卷积神经网络的发展, 越来越多的研究者开始纷纷将注意力转向卷积神经网络的研究中, 许多人尝试改进 AlexNet 模型, 以得到更好的准确率[5–7]. 2014 年 Simonyan 等人[3] 发现通过增加网络的深度可以达到提升网络的性能, 提出了新的模型——VGGNet. VGGNet 的核心思想是利用较小的卷积核, 通过反复堆叠的方式来增加网络的深度, 有两种基本类型: VGGNet-16 和 VGGNet-19. 由此卷积神经网络也开始不断向纵深化方向发展, 相继出现了 GoogLeNet[8] 和 ResNet[4] 等深层网络模型.

VGGNet 网络根据其核心思想可知在模型结构中网络全部采用 3×3 这类较小的卷积核, 以及 2×2 的池化核, 其模型结构图如图 3 所示. VGGNet 模型开启了小卷积核的时代, 3×3 的卷积核成为主流模型的关键, 同时 VGGNet 的成功也相继成为了各类图像任务的骨干网络结构, 应用在包括图像分类、目标定位、对象检测和图像分割等一系列图像大型任务中.

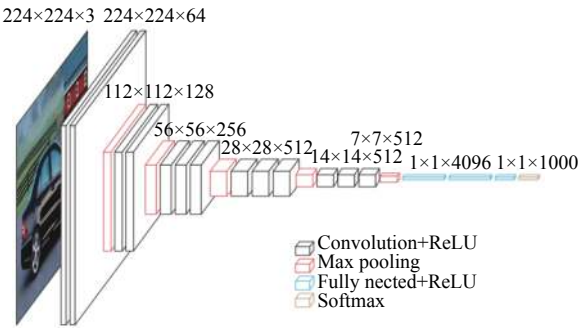


图 3 VGGNet 模型结构示意图

2.3 GoogLeNet

VGGNet 在 2014 年的 ImageNet 大规模视觉识别挑战赛 (ILSVRC14) 中, 通过采用堆叠小卷积核以扩大感受域的方式赢得了目标定位 (object localization) 比赛的冠军, 图像分类 (classification) 的亚军, 验证了加深模型结构有助于提升网络的性能. 同年, Szegedy 等人提出的 GoogLeNet[8] 则是专注于如何构建更深的网络结构, 在这场比赛中, GoogLeNet 通过引入新型的基本结构——降维 Inception 模块, 以增加网络的宽度, 该模型赢得了图像分类和目标检测 (object detection) 的冠军, 目标定位的亚军.

GoogLeNet 提出了降维 Inception 模块, 即 Inception V1 模块, 如图 4 所示. Inception V1 最大的优点就是控制了计算量和参数量的同时, 获得了非常好的分类性能. GoogLeNet V1 是一种卷积层、池化层和 Inception V1 模块的堆叠模型, GoogLeNet V1 共包含 9 个 Inception V1 模块, 其中, 所有卷积层均采用 ReLU 激活函数. 自 2014 年之后, Inception V1 模块经过不断的改进, 现已经发展到 GoogLeNet V4 版本, GoogLeNet 也成为一代经典的分类卷积神经网络.

2.4 ResNet

VGGNet 和 GoogLeNet 的提出, 无疑给许多研究者提供了一个猜想: 是不是网络越深越好, 那么是否可以无止境地加深网络层数? 然而事实却恰恰相反, 随着网络层数的增加, 网络将越来越难训练. 如果只是简单地对网络进行加深或加宽, 并不一定能带来网络性能的提升. 相反, 网络的加深或加宽, 会导致网络的参数数量变大, 计算量增大. 在不断增加网络深度时, 会出

现精度下降和梯度消失或梯度爆炸的问题, 而这些问题并不是由过拟合引起的, 仅因为网络增加了更多的层数[9].

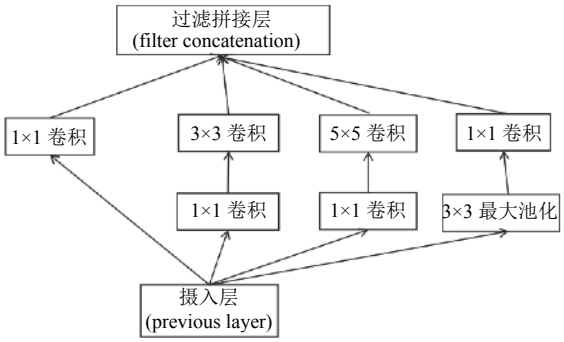


图 4 降维 Inception 模块 (Inception V1 模块)

深度残差 (deep residual learning) 模块[4] 的提出, 主要是为了解决上述提出的两个问题, 以便能够成功训练成百上千层的残差网络 (ResNet). 在 ResNet 网络中引入了跨层连接 (shortcut connection), 构造了残差模块, 如图 5 所示. 基于残差模块, 可以构建非常深的深度残差网络, 深度甚至可达 1000 层以上. 文献 [4] 中大量的实验表明 ResNet 网络可以有效地降低深层网络在训练集上误差增大的现象.

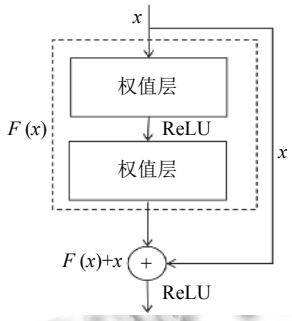


图 5 残差模块结构示意图

ResNet 为卷积神经网络往纵深化方向发展奠定了基础, 从理论上证明了加深网络比加宽网络更加有效[10], 也相继在 ResNet 网络基础上产生了一些新的变种模型, 比如 ResNeXt[11], 特征金字塔网络 (Feature Pyramid Network, FPN)[12]、宽度残差网络 (Wide Residual Network, WRN)[13].

3 全卷积网络 FCN

2015 年, 加州大学伯克利分校的 Long 等人在经典分类网络的基础上提出了全卷积神经网络 (Fully Convolution Network, FCN)[14], 该模型摒弃了全连接层, 加入了上采样层和反卷积层这类具有空间平移不变形式的层. 不同于传统的基于图像块的分割方法, FCN 证明了端到端、像素到像素训练方式下的卷积神经网络可以显著提高语义分割的计算效率和预测性能, 端到端训练为后续语义分割算法的发展铺平了道路.

FCN 的构造方法是, 把传统卷积网络的所有全连接层都改编成相应大小的密集卷积层. 例如, 在 VGGNet 基础上, FCN 把 VGGNet 网络后面 3 层全部改编为 1×1 的卷积核所对应等同向量长度的多通道卷积层, 整个网络模型全部都是由卷积层组成, 没有全连接层产生的向量, 改编过程如图 6 所示. 从图中可以看出, FCN 的输入为 224×224×3 的图片, 输出层仍为同尺寸大小的热力图.

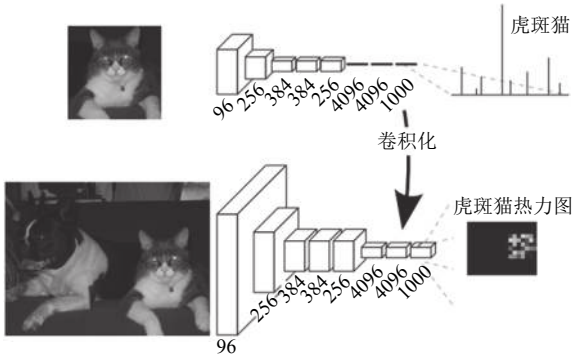


图 6 FCN 在 VGGNet 的基础上的改编过程

在改编全连接层之后, FCN 通过两种方式产生密集输出, 一种是直接放大, 通过放大变化 (上采样和反卷积), 直接把特征图放大成一个与输入大小相同的输出图像, 如图 7 所示. 在图中, FCN-32s 直接把池化层 pool5 通过上采样或反卷积方式放大 32 倍, 产生一个密集输出 (dense output), 但是直接放大 32 倍得到的结果不够精确, 一些细节无法恢复.

另一种方法是通过设计一个跳跃连接, 将全局信息和局部信息连接起来, 相互补偿来产生更加准确和精细的分割结果. 在图 7 中, FCN-16s 把池化层 pool4 和 2 倍的池化层 pool5 进行拼接, 再通过上采样或反卷积放大 16 倍, 得到另一个密集输出; 此外, FCN-8s 中先把池化层 pool4 和 2 倍的 pool5 进行拼接后的结果通过上采样或反卷积进行 2 倍放大, 然后, 与池化层 pool3 进行拼接, 最后通过上采样或反卷积放大 8 倍, 产生一个密集输出.

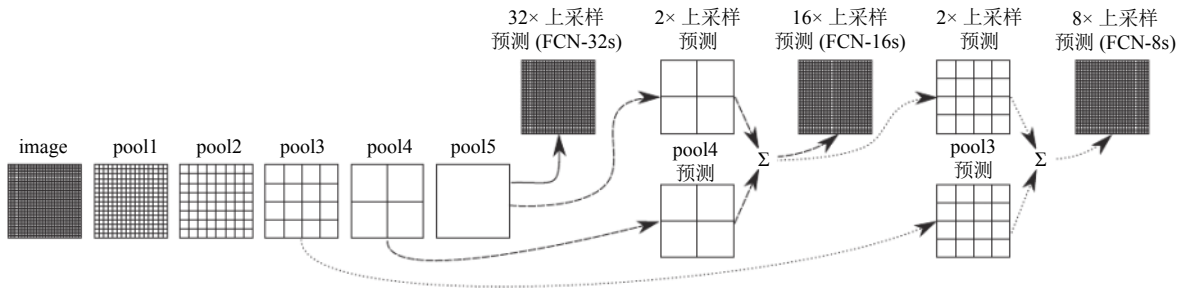


图 7 FCN 产生密集输出的方式示意图

FCN 在 PASCAL VOC 等大型数据集上都得到了 state-of-the-art 的出色结果, 可以说, FCN 是语义分割领域的开山之作, 为后续许多经典的语义分割模型的提出奠定了思想基础, 推动了图像语义分割任务的进一步发展.

4 FCN 的衍生—主流分割框架

目前, 许多最成功的, 最先进的深度学习图像语义分割模型思想都来源于 2015 年 Long 等人提出的全卷积神经网络 (FCN) 模型[14]. 全卷积神经网络可以说是深度学习在图像语义分割领域的基石, 也是深度学习技术应用于图像语义分割的一个里程碑, 它展示了如何端到端训练卷积神经网络来进行图像语义分割, 分割精度比传统方法有了显著的提高.

尽管全卷积神经网络模型具有强大功能和足够的灵活性, 它仍然存在某些不足, 阻碍了它在某些问题和情况下的应用: 卷积网络所具有的平移不变性使其没有考虑有用的全局上下文信息. 研究表明, 全局特征信息或上下文信息相互作用有助于正确地分类像素进行语义分割[15,16]. 在本节中将介绍 4 种基于 FCN 的卷积神经网络分割框架, 它们利用上下文信息进行语义分割.

4.1 编码器-解码器

编码器-解码器: 该模型由两部分组成: 编码器和解码器. 其中编码器主要有卷积层和下采样层组成, 通过卷积操作逐渐减小特征图大小并捕获更高层次的语义信息; 而解码器主要由上采样层或反卷积、卷积层和融合层组成, 通过上采样或反卷积的方式逐渐恢复对象细节信息和空间维度来进行分割. 整个结构利用来自编码器模块的多尺度特征, 并从解码器模块恢复空间分辨率[17]. U-Net 模型[18] 和 SegNet 模型[19] 就是编码器-解码器结构的典型代表之一.

U-Net 模型结构如图 8 所示, 该结构主要包括一个捕获上下文信息的收缩路径和一个用以精确定位的对称拓展路径. 因此, 该模型左侧可视为编码器部分, 编码器中主要由 4 个子模块组成, 每个子模块包含两个卷积层, 每个子模块之后通过最大池化层 (max-pool) 实现的下采样; 模型的右侧可视为解码器部分, 解码器同样由 4 个子模块组成, 通过上采样操作恢复对象细节和空间维度, 直到与输入图像的分辨率一致. 此外, U-Net 网络还设计了一个跳跃连接, 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. U-Net 模型提出, 成功实现了使用非常少的数据完成端到端的训练, 并获得了出色的医学图像分割效果, 成为大多数医疗影像语义分割任务的基线.

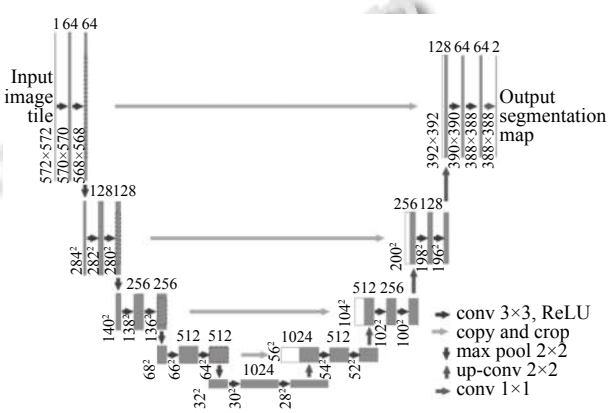


图 8 U-Net 模型结构示意图

SegNet 模型如图 9 所示, 该模型主要由编码网络 (encoder network)、解码网络 (decoder network) 和逐像素分类层 (pixel-wise classification layer) 组成. 其中编码网络是将高维向量转换成低维向量, 同时在池化过程中记录最大池化索引信息, 保存了最大特征值所在的位置, 以保存边界信息. 解码网络与编码网络是相对称的, 通过解码网络可以将低分辨率的特征图映射到

高空间分辨率的特征图, 实现了低维向量到高维向量的重构, 最后通过 Softmax 激活函数用于输出与输入图像具有相同分辨率的像素级标签. SegNet 模型的提出, 使得编码器-解码器网络结构普适化.

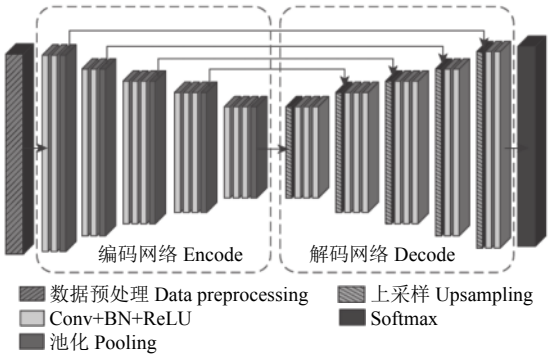


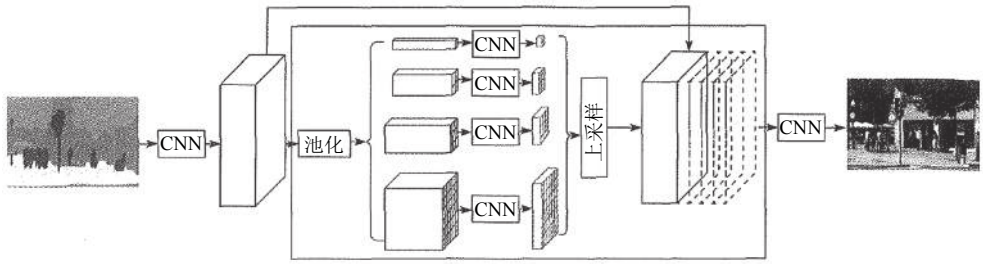
图 9 SegNet 模型结构示意图

基于编码器-解码器结构的网络模型还包括 LRR[20]、Refine Net[21] 和 DeepLab v3_plus[22] 等. 这些模型的提出证明了该结构在图像语义分割领域的高效性, 此外, 该结构也广泛应用于对象检测方面并取得了一定的成果[23,24].

4.2 特征融合

特征融合模型通常采用空间金字塔池化在多个范围内捕获上下文信息, 然后进行多尺度特征融合. 该模型结构在一定程度上克服了用全卷积网络 (FCN) 进行图像语义分割时没有考虑全局上下文信息的主要缺点. 其中比较具有代表性的模型是金字塔场景分析网络 (Pyramid Scene Parsing Network, PSPNet)[25] 和 DeepLab v2 网络[26].

PSPNet 网络在 FCN 的基础上通过对不同区域的上下文信息进行聚合, 充分利用了全局上下文信息来提高最终预测的可靠性, 其总体结构如图 10 所示, 它的特点在于使用了金字塔池化模块 (pyramid pooling module), 将 4 个不同的粗细尺度进行特征融合. 在金字塔模块中, 不同尺度级别的输出包含不同大小的特征图, 通过采用 1×1 卷积层把上下文表示的维数降低为原来的 1/N, 以保证全局权重. 然后, 通过双线性插值对低维特征图进行上采样以获得相同大小的特征, 不同级别的特征被拼接为最终的金字塔池化全局特征. 最后再经过一个卷积层输出最终的逐像素预测结果.



(a) 输入图像 (b) 特征图 (c) 金字塔池化模型 (d) 最终预测

图 10 PSPNet 的总体结构

DeepLab v2 网络中提出了基于空洞卷积的空间金字塔池化 (Atrous Spatial Pyramid Pooling, ASPP) 模块, 其思想来源于 SPPNet[27], 主要是为了解决目标的多尺度问题. 通过设计不同的空洞率的多个并行卷积核, 组成类似金字塔方式在给定的特征层上进行有效的重采样, 空洞卷积可以保证在参数量的不变的情况下, 通过有效地增大卷积核尺寸来扩大感受域的大小, 如图 11 所示. 在这个基础上, DeepLab v3_plus 提出了包含 ASPP 模块的编码器-解码器模型, 并在几个语义分割基准数据集上表现出优异的性能. 空间金字塔池化不仅仅应用于图像分割领域, 也应用于目标检测领域[28].

4.3 上下文模块

上下文模块包含级联的其他模块, 以对长距离上下文进行编码. 语义分割需要对多种空间尺度的信息进行整合, 同时也需要对局部信息与全局上下文信息进行平衡. 其原因在于, 一方面, 局部信息对于提高像素级别的标注的正确率是非常重要的; 另一方面, 整合图像全局的上下文信息对于解决局部模糊性问题来说也同样重要的. 然而, 普通的卷积神经网络具有的平移不变性容易造成其忽略了高分辨率的特征图, 会导致边缘信息的丢失. 一种常见的改进是使用条件随机场 (Conditional Random Field, CRF) 作为后处理过程来调

优结果.

其中一种有效的方法是将全连接 CRF 与 DCNN[29] 结合, 通过计算任意两个像素之间的概率值来判断它们之间的相似性从而达到判断它们是否属于同一个类. 通过这种方式来实现利用全局上下文信息, 而不是局部信息来进行语义分割. 全连接 CRF 可以捕获物体的边缘信息, 弥补了 DCNN 带来的边界平滑问题.

4.4 图像金字塔

图像金字塔是图像中多尺度表达的一种, 最主要用于图像分割, 是一种通过提取多分辨率图像以对图像进行解释的有效但概念简单的结构. 图像金字塔最初用于机器视觉和图像压缩, 将图像以不同的分辨率以金字塔形状进行排列从而形成图像的金字塔, 其通过梯次向下采样获得, 直到达到某个终止条件才停止采样. 在图像金字塔中, 金字塔的顶层图像分辨率最低, 底层图像分辨率最高. 常见的图像金字塔有高斯金字塔和拉普拉斯金字塔两种[30].

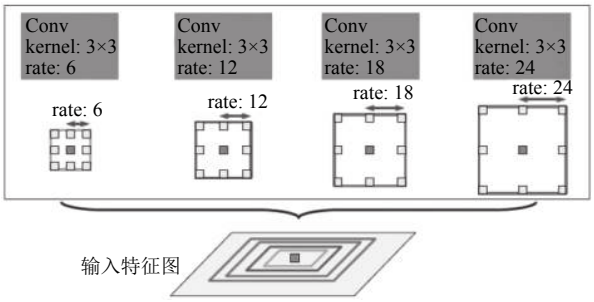


图 11 基于空洞卷积的空间金字塔池化

典型的例子是 Farabet 等人[31] 通过 Laplacian 金字塔转换输入图像, 以多尺度的方式输入到 DCNN 中, 并从所有尺度中合并特征图. 但是这类模型的主要缺点是计算机量大, 对于较深的 DCNNs 不能很好地进行缩放, 因此通常应用在理论分析.

5 图像分割的评价指标

目前已经有许多专注于图像语义分割的模型与基准数据集, 这些基准数据集为评价模型的性能提供了一套统一的标准. 通常对分割模型进行评价会从执行时间、内存使用率和算法精度等方面进行考虑. 在这节中, 我们主要介绍语义分割模型的算法精度评价指标.

类比二分类问题[32], 在图像分割中, 我们也引入“混淆矩阵”, 用 *PA* 和 *IoU*[33] 的值来评估语义分割技术的准确性. 假设: 共有 k+1 个类, P_{ij} 表示本属于类 i 但被预测为类 j 的像素数量. 即, P_{ii} 表示真正的数量 (*TP*+*TN*, *TP*: 真正例, *TN*: 真反例), 而 P_{ij} 和 P_{ji} 则分别被解释为假正例 (*FP*) 和假负例 (*FN*), 当 $i\neq j$ 时, P_{ii} 表示 *TP*, P_{jj} 表示 *TN*, P_{ij} 表示 *FP*, P_{ji} 表示 *FN*.

PA 像素精度 (Pixel Accuracy): 标记正确的像素占总像素的比例, 等价于准确率, 公式如下:

$$PA=\frac{\sum_{i=0}^{k}P_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}P_{ij}}=\frac{TP+TN}{TP+TN+FN+FP} \tag{1}$$

Intersection over Union (*IoU*): 模型对某一类别预测结果和真实值的交集与并集的比值, 一种测量在特定数据集中检测相应物体准确度的一个标准, 如式 (2):

$$IoU=\frac{A_{\text{pred}}\cap A_{\text{true}}}{A_{\text{pred}}\cup A_{\text{true}}}=\frac{TP}{(TP+FP+FN)} \tag{2}$$

MPA 均像素精度 (Mean Pixel Accuracy): 计算每个类内被正确分类像素数的比例, 再求所有类的平均, 公式如式 (3):

$$MPA=\frac{1}{K+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{j=0}^{k}P_{ij}} \tag{3}$$

均交并比 (Mean Intersection over Union, *MIoU*): 计算真实值和预测值的交集和并集, 公式如式 (4):

$$MIoU=\frac{1}{K+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{j=0}^{k}P_{ij}+\sum_{j=0}^{k}P_{ji}-P_{ii}} \tag{4}$$

频权交并比 (Frequency Weighted Intersection over Union, *FWIoU*): 是根据每一类出现的频率设置权重, 权重乘以每一类的 *IoU* 并进行求和, 公式如式 (5):

$$FWIoU=\frac{1}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ij}}\sum_{i=0}^{k}\frac{\sum_{j=0}^{k}p_{ij}p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{5}$$

利用混淆矩阵计算: 每个类别的真实数目为 *TP*+*FN*, 总数为 *TP*+*FP*+*TN*+*FN*, 其中每一类的权重和其 *IoU* 的乘积计算公式如下, 在将所有类别的求和即可, 公式如式 (6):

$$FWIoU=\left[\frac{(TP+FN)}{(TP+FP+TN+FN)}\right]\times\left[\frac{TP}{(TP+FP+FN)}\right] \tag{6}$$

在上面描述的所有度量中, *MIoU* 由于其代表性和

简单性而脱颖而出, 成为最常用的度量. 大多数挑战和研究人员利用这一指标来展示他们的结果.

6 语义分割的常用实验数据集

6.1 常用实验数据集

PASCAL VOC 挑战赛主要是为图像分类、目标检测和图像分割 3 类任务的基准测试比赛, 主要有 4 个大类别, 分别是人、常见动物、交通车辆、室内家具用品, 数据集标注质量高、场景复杂、目标多样、检测难度大、数据量小但是场景丰富, 相比 ImageNet 等更加考验人工智能算法的设计和创新能力[34]. 其官方网址: http://host.robots.ox.ac.uk/pascal/VOC/.

Cityscapes 数据集专注于对城市街道场景的语义理解, 共包含来自 50 个城市的不同场景、不同背景、不同街道的 24 998 张图片, 以及包含 30 种类别涵盖地面、建筑、交通标志、自然、天空、人和车辆等物体标注, 以关注真实场景下的环境理解著称, 任务难度大. 此外, 图像根据标记质量分为两组, 其中 5000 个是精细注释, 其余 19 998 个是粗注释. 将 5000 张精细标注的图像进一步分组为 2975 张训练图像, 500 张验证图像和 1525 张测试图像, 其官方网站: https://www.cityscapes-dataset.com/.

CamVid 是第一个具有目标类别语义标签的视频集合. 该数据库提供 32 个 ground truth 语义标签, 将每个像素与 32 个语义类之一相关联. 数据主要通过在固定位置架设 CCTV 式摄像机进行拍摄和从驾驶汽车的角度拍摄的两种方式进行获取, 驾驶场景增加了观察目标的数量和异质性. 其官方网站: http://mi.eng.cam.ac.uk/research/projects/VideoRec/Cam-Vid/.

SUN RGB-D 数据集是普灵斯顿大学的 Vision & Robotics Group 公开的一个有关场景理解的数据集[35], 拥有 5285 张训练图像和 5050 张测试图像. 这些图像是由不同的传感器捕获的, 因此具有不同的分辨率. 在图像类别中涵盖 37 个室内场景类, 包括墙, 地板, 天花板, 桌子, 椅子, 沙发等. 由于对象类具有不同的形状、大小和不同的姿态, 因此对这些图像进行语义分割的任务是十分困难和复杂的, 极具有挑战性, 官方网站: http://rgbd.cs.princeton.edu.

NYUD 同样是关于室内场景的数据集, 该数据集分为两大类型: NYU Depth V1 和 NYU Depth V2. 其中 NYU-Depth V2 数据集通过摄像机记录的各种室内场景的视频序列组成, 它包含来自 3 个城市的 464 种不同的室内场景和 26 种场景类型, 407 024 个未标记的帧以及 1449 个密集标记的帧. 而 NYU-Depth V1 包含 64 种不同的室内场景和 7 种场景类型, 108 617 张未标记的帧和 2347 个密集标记的帧. 其官方网站: https://cs.nyu.edu/~silberman/datasets/.

最后, 通过表 1 来对 PASCAL VOC、CityScapes、Camvid、SUN RGB-D 和 NYUDV2 这些常用语义分割数据集做简单的归纳总结. 除此之外, 大规模图像分割数据集还包括 Semantic Boundaries Dataset(SBD)、Microsoft Common Objects in COntext (COCO)、KITTI、Adobe's Portrait Segmentation、Youtube-Objects、Materials IN Context (MINC)、Densely-Annotated VIdeo Segmentation (DAVIS)、Stanford background、SiftFlow 以及 3D 数据集, 包括 ShapeNet Part、Stanford 2D-3D-S、A Benchmark for 3D Mesh Segmentation、Sydney Urban Objects Dataset、Large-Scale Point Cloud Classification Benchmark 等.

表 1 常见语义分割数据集归纳总结

数据集	场景	类别	训练	验证	测试	总数
SUNRGB-D	室内	37	2666	2619	5050	10335
NYUD-V2	室内	40	795	654	—	1449
PASCALVOC	综合	21	1464	1449	—	2913
Citysacpes	道路	30	2975	500	1525	5000
CamVid	道路	32	367	100	233	700

6.2 实验结果分析与对比

为了更好地说明网络的性能, 本节选取其中一个具有挑战性的数据集: CamVid, 主要针对 FCN[14]、U-Net[18]、PSPNet[26] 和 DeepLab v3_Plus[22] 模型进行实验, 并对实验结果进行分析对比. CamVid 数据集主要是道路场景, 物体涉及行人、路牌、路灯、车辆、建筑物、树木等, 物体种类丰富, 背景复杂多变, 共包含 701 张图像, 其中 367 张图像用于训练,101 张用于验证, 233 张用于测试. 在训练 CamVid 数据集时, 没有考虑类别之间的平衡问题, 所以没有做类别不平衡修正.

我们的实验系统包括预训练网络都是基于 PyTorch 框架实现, 所有实验都是基于 NVIDIA-RTX 2070S (8 G) 显卡和 Python3.7 版本上完成. 为了加快网络收敛速度和根据计算机配置, 将循环次数设置为 50, 初始学习率设置为 1×10^{-4}, 学习率随训练次数的增加而线性递减, mini-batch size 设置为 4, 优化器采用 SGD, 且动量设

置为 0.9, 输入图像大小为 512×512. 我们在类别像素上采用 Negative Log Liklihood Loss (NLLLoss) 损失函数, 该函数适用于二分类和多分类任务, 也适用于训练数据集类别不平衡任务. 将通过 Mean-IoU 记录模型的性能表现. CamVid 数据集上的实验对比如表 2 所示, 主要比较因素包括基础网络、核心技术和评价指标 *MIoU*.

表 2 基于 CamVid 数据集实验对比

模型	基础网络	核心技术	*MIoU* (%)
FCN-8s	Vgg-16	上采样跳跃连接	45.3
U-Net	—	编码器-解码器 跳跃连接	47.8
PSPNet	ResNet	多尺度特征融合 空间金字塔池化	51.6
DeepLab v3_plus	Xception-Net	带孔卷积 ASPP模块	52.4

从表 2 中可以看到 U-Net、PSPNet、DeepLab v3_plus 三种模型在 CamVid 数据集上 *MIoU* 与 FCN 相比均有所提升. 其中, PSPNet、DeepLab v3_plus 两种模型的 *MIoU* 均达到了 50% 左右, 能够识别图像中不同尺度的物体, 分割结果对比其他模型, 更加接近真实分割, 边界信息更为准确, 是性能出色, 具有代表性的图像语义分割模型. DeepLab v3_plus 集成了 DeepLab v2、DeepLab v3 等众多网络的优点, 通过逐渐恢复空间信息来捕获更清晰的边界信息; 而 PSPNet 通过空间金字塔池化模块, 利用多尺度的方式对图像特征进行融合, 有效地捕获了图像丰富的上下文信息, 二者均提高了图像的识别效果. 此外, U-Net 通过跳跃连接将高级和低级图像特征进行融合, 最大化提取了细节信息, 更常用在数据量少、边界模糊, 需要高分辨率信息用于精准分割的医学图像. 实验结果进一步阐释了语义分割需要丰富的上下文信息才能产生高质量的分割结果.

7 图像分割的应用

图像语义分割技术越来越受到计算机视觉和机器学习研究者的关注. 许多正在兴起的产业应用需要精确有效的分割技术, 包括自动驾驶、室内导航, 医疗影像分析等. 这些需求与深度学习技术在计算机视觉领域的基本任务中研究不谋而合, 包括语义分割或场景理解[33], 下面介绍主要的语义分割的 3 个方面的应用.

7.1 医疗影像分割

医学图像是临床诊断与医学研究中不可缺少的工具, 在医学图像处理与分析领域, 医学图像分割是对医学图像进一步分析和处理的基础[36]. 医学图像分割的目的就是通过提取描述对象的特征, 把感兴趣对象从周围环境中分离出来, 分析和计算分割对象的解裁、病理、生理和物理等方面的信息[37]. 医学图像分割对医学研究、临床诊断、病例分析、手术计划、影像信息处理和计算机辅助手术等医学研究与实践领域有着广泛的应用和研究价值.

然而, 构建自动分割框架一直是医学界极具挑战性的课题. 随着深度学习的不断发展, 神经网络应用于医学图像分割成为主流趋势, 极大地提高了医疗诊断的准确性和可靠性. 在基于深度学习的医学图像分割方法中, 最具代表性的模型是 U-Net[17], 其通过含有跳跃连接的编码—解码结构的网络结构, 有效地适用于医学图像分割, 在此基础上也涌现了许多适合医学图像分割的模型并且得到了广泛的应用.

7.2 地理信息系统

遥感图像是人们获取地球表层各类地物信息的重要数据来源. 遥感图像语义分割, 是指对遥感图像进行处理、分析, 目的是像素级地识别图像, 即标注遥感图像中每个像素所属的对象类别, 是图像分析的第一步. 通过图像语义分割技术, 可以同时提取遥感图像中的多种地物, 可以在土地利用调查、自然灾害探测、环境监测、精确植被、城市规划等一系列潜在的实际应用中发挥重要作用.

然而, 由于遥感影像的成像特征, 地物之间可能会因“同物异谱, 同谱异物”造成类内光谱差异大, 类间光谱特征相互重叠, 使得物体之间难以区分. 因此, 很难提出一种从遥感图像中以合理的精度和可靠性定位多个目标的方法. 随着深度学习的发展, 目前已经有大量文献证明了神经网络在遥感图像语义分割中取得了显著进展[38,39].

7.3 无人车驾驶

自动驾驶汽车作为人工智能和汽车工业结合的产物, 毫无疑问是当今的研究热点. 对于无人驾驶汽车而言, 最重要的是自动驾驶的汽车对行驶道路上环境的感知与物体的识别, 车辆行驶在过程中, 通过车载摄像头采集车辆所处的位置以及车辆周围的物体. 在目前的自动驾驶技术中, 处理好图像感知到的信息, 将有助于提高车辆行驶决策的准确度. 图像语义分割是将图像中每一个像素按照标签进行分类, 从而从图像中提

取出丰富的驾驶环境信息, 辅助决策[40]. 目前已经有许多对城市景观的语义分割开源数据集, 这些数据集可以帮助语义分割模型的训练与测试.

与图像分类或目标检测相比, 语义分割使我们对图像有更加细致的了解. 这种了解在诸如自动驾驶、医疗影像分析、机器人以及图像搜索引擎等许多领域都是非常重要的.

8 总结与展望

2015 年, FCN 的提出极大地推动了图像分割的发展, 所取得的成就已经使图像分割任务迈入盛况空前的新阶段.

尽管如此, 图像语义分割方面仍然存在挑战, 可以从两个角度出发看待这些挑战:

(1) 从深度卷积神经网络的角度看待语义分割面临的挑战[26]:

① 特征分辨率降低: 主要是由深度卷积神经网络结构中设计重复采用最大池化操作和下采样操作, 重复池化和下采样降低了空间分辨率. 其中一种解决方法是采用转置卷积 (deconvolutional layer), 但是该方法需要额外的空间的同时也增大了计算量. 另一种解决方法是采用将最大池化层用空洞卷积来替代下采样, 保证了不增加参数量的同时有效地扩大了感受域的大小, 以更高的采样密度计算特征图;

② 对象在多尺度的存在: 主要是目标在多尺度图像中的状态造成的, 因为在同一种尺度下, 不同目标的特征往往响应并不相同. 已有的解决该问题的一个方法是将图片缩放成不同尺寸, 汇总特征得到结果. 这种方法可以提高性能, 但是增加了计算成本. 除此之外, 受 SPPNet 启发, DeepLab v2 提出一个 ASPP 模块, 对给定输入使用不同采样率的空洞卷积并行采样, 以多比例捕捉图像上下文;

③ 深度卷积神经网络的空间不变性造成定位精度的下降: 现存的解决该问题的一个方法是使用跳跃连接结构, 通过融合不同层的特征图从而计算最终的分割结果. 此外, 另一种高效的方法是采用条件随机场增强模型捕捉细节的能力.

(2) 从分割结果的角度看待语义分割面临的挑战:

语义分割可以看成是一个逐像素分类的任务, 它包含分类和定位两个挑战[41].

① 分类: 对于分类任务, 模型必须具有空间不变性的, 以适应目标的各种形式, 如平移和旋转, 因此在设计网络时可以采用全卷积的结构, 去掉全连接层或全局池化层;

② 定位: 对于定位任务, 模型必须是对图像变换足够的敏感的, 即能够精确定位语义类别的每个像素, 因此在涉及网络时可以采用较大的卷积核, 使得像素与特征图的结合更加紧密, 增强处理不同变换的能力, 采用较大卷积核可以保证感受野足够大, 可以覆盖了较大的目标, 有利于分类.

在现有研究成果的基础上, 展望未来, 图像语义分割技术的研究可能会从以下几个方面展开[33,42]:

(1) 应用于 3D 数据的语义分割.

(2) 应用于场景解析任务的图像语义分割.

(3) 实例级图像语义分割.

(4) 实时图像语义分割.

(5) 应用于序列数据集的语义分割.

参考文献

1 Zhang ZL, Zhang XY, Peng C, *et al*. ExFuse: Enhancing feature fusion for semantic segmentation. Proceedings of the 15th European Conference on Computer Vision. Munich, Germany. 2018. 1–2.

2 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84–90. [doi: 10.1145/3065386]

3 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proceedings of the 3rd International Conference on Learning Representations. San Diego, CA, USA. 2014. 1–5.

4 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 770–777.

5 Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland. 2014. 818–833.

6 Sermanet P, Eigen D, Zhang X, *et al*. OverFeat: Integrated recognition, localization and detection using convolutional networks. Proceedings of the 2nd International Conference on Learning Representations. Banff, AB, Canada. 2014.

7 Lin M, Chen Q, Yan SC. Network in network. arXiv: 1312.4400, 2013.

8 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2014. 1–9.
9 He KM, Sun J. Convolutional neural networks at constrained time cost. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2014. 5353–5360.
10 Eldan R, Shamir O. The power of depth for feedforward neural networks. Proceedings of the 29th Annual Conference on Learning Theory. New York, NY, USA. 2016. 907–940.
11 Xie SN, Girshick R, Dollár P, *et al*. Aggregated residual transformations for deep neural networks. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 1–2.
12 Lin TY, P Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 2117–2125. [doi: 10.1109/CVPR.2017.106]
13 Zagoruyko S, Komodakis N. Wide residual networks. Proceedings of the British Machine Vision Conference. York, UK. 2017. 87. [doi: 10.5244/C.30.87]
14 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640–651. [doi: 10.1109/TPAMI.2016.2572683]
15 Shotton J, Winn J, Rother C, *et al*. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. International Journal of Computer Vision, 2009, 81(1): 2–23. [doi: 10.1007/s11263-007-0109-1]
16 Yao J, Fidler S, Urtasun R. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA. 2012. 7.
17 Chen LC, Papandreou G, Schroff F, *et al*. Rethinking atrous convolution for semantic image segmentation. arXiv: 1706.05587, 2017.
18 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany. 2015. 234–241.
19 Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481–2495. [doi: 10.1109/TPAMI.2016.2644615]
20 Ghiasi G, Fowlkes CC. Laplacian pyramid reconstruction and refinement for semantic segmentation. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands. 2016. 519–534.
21 Lin GS, Milan A, Shen CH, *et al*. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 5168–5177. [doi: 10.1109/cvpr.2017.549]
22 Chen LC, Zhu YK, Papandreou G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the 15th European Conference on Computer Vision. Munich, Germany. 2018. 833–851.
23 Kechaou A, Martinez M, Haurilet M, *et al*. Detective: An attentive recurrent model for sparse object detection. Proceedings of 2020 25th International Conference on Pattern Recognition. Milan, Italy. 2020. 5340–5347.
24 Shrivastava A, Sukthankar R, Malik J, *et al*. Beyond skip connections: Top-down modulation for object detection. arXiv: 1612.06851, 2017.
25 Zhao HS, Shi JP, Qi XJ, *et al*. Pyramid scene parsing network. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 6230–6239.
26 Chen LC, Papandreou G, Kokkinos I, *et al*. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834–848. [doi: 10.1109/TPAMI.2017.2699184]
27 He KM, Zhang XY, Ren SQ, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904–1916. [doi: 10.1109/TPAMI.2015.2389824]
28 李成跃, 姚剑敏, 林志贤, 等. 基于改进 YOLO 轻量化网络的目标检测方法. 激光与光电子学进展, 2020, 57(14): 37–45.
29 Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with Gaussian edge potentials. Advances in Neural Information Processing Systems. Granada, Spain. 2012. 109–117.
30 毛星云, 冷雪飞, 王碧辉, 等. OpenCV3 编程入门. 北京: 电子工业出版社, 2015. 223–226.

31 Farabet C, Couprie C, Najman L, *et al*. Learning hierarchical features for scene labeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1915–1929. [doi: 10.1109/TPAMI.2012.231]
32 周志华. 机器学习. 北京: 清华大学出版社, 2016. 30–31.
33 Garcia-Garcia A, Orts-Escolano S, Oprea S, *et al*. A review on deep learning techniques applied to semantic segmentation. arXiv: 1704.06857, 2017.
34 Everingham M, Van Gool L, Williams CKI, *et al*. The PASCAL Visual Object Classes (VOC) challenge. International Journal of Computer Vision, 2010, 88(2): 303–338. [doi: 10.1007/s11263-009-0275-4]
35 Song SR, Lichtenberg SP, Xiao JX. SUN RGB-D: A RGB-D scene understanding benchmark suite. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 567–576.
36 王阳萍, 杜晓刚, 赵庶旭, 等. 医学影像图像处理. 北京: 清华大学出版社, 2012. 69–72.
37 赵于前, 柳建新, 刘剑. 基于形态学重构运算的医学图像分割. 计算机工程与应用, 2007, 43(10): 238–240. [doi: 10.3321/j.issn:1002-8331.2007.10.072]
38 Marmanis D, Wegner JD, Galliani S, *et al*. Semantic segmentation of aerial images with an ensemble of CNNS. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016, III-3: 473–480. [doi: 10.5194/isprs-annals-III-3-473-2016]
39 Dai JF, He KM, Li Y, *et al*. Instance-sensitive fully convolutional networks. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands. 2016. 534–549.
40 俞涛. 基于深度学习的图像语义分割在自动驾驶中的应用 [硕士学位论文]. 武汉: 华中科技大学, 2018.
41 Peng C, Zhang XY, Yu G, *et al*. Large kernel matters-improve semantic segmentation by global convolutional network. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 1743–1751.
42 田萱, 王亮, 丁琪. 基于深度学习的图像语义分割方法综述. 软件学报, 2019, 30(2): 440–468. [doi: 10.13328/j.cnki.jos.005659]