计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(9):239−245 [doi: 10.15888/j.cnki.csa.007066]

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

基于多级特征整合的图像语义分割研究①

徐天宇, 孟朝晖

(河海大学 计算机与信息学院, 南京 211100)
通讯作者: 徐天宇, E-mail: hhuxty0720@163.com

摘　要: 传统的全卷积神经网络由于不断的池化和下采样操作使得最后的特征热点图分辨率损失严重, 导致了分割结果的细节刻画能力的缺失, 为了弥补这一缺陷, 往往通过跳跃连接融合中层的特征图以恢复空间信息. 由于无法充分利用网络的低层特征信息, 传统全卷积网络的特征融合阶段存在相当的缺陷, 本文对这一现象进行了深入的分析. 本文在上采样路径之前采用基于特征金字塔的特征信息增强方法, 克服了浅层特征图语义信息匮乏这一缺点, 使得整个网络能更充分的利用前向计算产生的特征图, 输出的分割结果也更为精确. 本文提出的算法在 Pascal VOC 数据集上取得了 75.8% 的均像素精度和 83.9% 的权频交并比, 有效的提高了分类精度.

关键词: 深度学习; 语义分割; 特征金字塔; 多级特征整合

引用格式: 徐天宇,孟朝晖.基于多级特征整合的图像语义分割研究.计算机系统应用,2019,28(9):239–245. http://www.c-s-a.org.cn/1003-3254/7066.html

Image Semantic Segmentation Based on Multi-Level Feature Integration

XU Tian-Yu, MENG Zhao-Hui

(College of Computer and Information, Hohai University, Nanjing 211100, China)

Abstract: Due to continuous pooling and down sampling, the resolution of the final feature hotspot map is seriously lost in the traditional full-convolution neural network, which leads to the loss of detail characterization ability of segmentation results. In order to make up for this defect, the feature map of middle layer is often fused by jumping connection to restore spatial information. Due to the failure to make full use of the low-level feature information of the network, the feature fusion stage of the traditional full-convolution network has some defects. This study makes an in-depth analysis of this phenomenon. The feature information enhancement method based on feature pyramid is adopted before the upper sampling path to overcome the deficiency of semantic information of shallow feature graph, so that the entire network can make full use of the feature graph generated by forward calculation and improve the segmentation result. The algorithm proposed in this study achieves 75.8% average pixel accuracy and 83.9% weight frequency crossover ratio on the Pascal VOC data set, effectively improved the classification accuracy.

Key words: deep learning; semantic segmentation; feature pyramid; multi-level feature integration

图像语义分割 (image semantic segmentation) 是图像处理和计算机视觉领域的一个关键问题, 在自动驾驶、场景理解、目标检测[1]等相关领域都有着广阔的应用前景, 是图像理解的基石性技术. 传统的图像分割旨在将分属于不同物体的像素区域分隔开, 而语义分割则更进一步, 在图像分割的基础上按照语义为每一块像素区域做分类, 在精度和细度上都是图像分割的一个重大升级, 可以理解为像素级别的分类任务.

① 收稿时间: 2019-03-01; 修改时间: 2019-03-29; 采用时间: 2019-04-04; csa 在线出版时间: 2019-09-05

在深度学习方法涌现之前, 语义分割主要依赖于传统的分割算法, 根据图像的颜色、纹理等底层信息进行区域划分, 同时需要一定的人工干预. 其中, 比较著名的有一种叫做“Normalizedcut”[2]的图划分方法, 简称 N-cut, 通过计算像素与像素之间的关系权重来综合考虑, 然后根据给出的阈值, 将图像一分为二. 但 N-cut 的缺陷很明显, 即需要对图像进行多次划分, 同时由于此类分割方法过于简单粗暴, 无法对图像中颜色纹理较为相似的部分进行分割, 导致分割结果并不准确. 在这之后提出[3]的“Grab-cut”方法增加了人工交互, 手工选定图像中待分割区域, 并提供一个大致的分割边界, 再通过算法进行目标区域的分割操作. 相比 N-cut 方法分割结果有所提升, 但由于增加了人工干预步骤, 使得其根本无法适应批量化的大规模分割要求. 同时, 仍然无法解决传统图像算法分割准确率低, 无法有效分割颜色纹理信息较为相似的区域的缺点.

2012 年以来, 深度学习几乎席卷了整个计算机视觉领域, 如 AlexNet[4]、GoogLeNet[5]和 ResNet[6]在图像分类和目标检测等相关任务中取得了全面优于传统方法的表现, 语义分割当然也不例外. 卷积神经网络 (Convolutional Neural Network, CNN) 能够自动从图像中提取特征的分布式表示, 避免了手工设计各类算子, 并且相较于传统方法, CNN 能够学习到更高维度的特征表示, 很多解决语义分割的网络都是以 CNN 为基础改进而来. Long 等[7]提出的全卷积网络 (Fully Convolution Network, FCN) 斩获了当年 CVPR 的最佳会议论文, 是第一个用于解决语义分割问题的、可端到端训练的深度网络. FCN 网络以 VGG-16[8]作为主干网络, 用卷积层替换原本的全连接层, 最后一层由原本的 1000 个神经元改为 21 通道的卷积层 (PASCAL VOC 数据集提供的类别数为 20 类, 加上背景类别即为 21), 最后将输出的预测结果上采样到和原图像分辨率相同大小, 在 PASCAL VOC 数据集上取得了当时最好的结果. 文献[9]提出的 SegNet 网络遵循了 FCN 的基本框架, 但在编码阶段记录每一个最大池化的位置信息, 解码阶段使用最大池化索引上采样, 将对应参数恢复至原先的位置, 以更好的恢复边缘信息, 效果相比 FCN 有所提升. Chen 等[10]提出的 DeepLab 主张特征图应保留一定的分辨率以适应语义分割此类稠密预测任务, 在解码层使用扩张卷积, 最后得到分辨率更高的特征热点图 (HeatMap), 再利用 CRF 对分割结果进行锐化, 得到了很好的分割结果.

相比于传统的分割方法, 基于卷积神经网络的语义分割方法得到的结果更为准确, 而且能直接得到图像中物体的语义信息. 目前用于解决语义分割问题的网络大致遵循全卷积网络的编码-解码结构, 编码部分利用卷积层提取图像的语义信息, 解码部分引入编码部分的信息以修复因为下采样而损失的空间信息, 或者直接进行上采样. 以现在的眼光看来, FCN 的分割结果相较而言略显粗糙, 很多视觉外观相似的物体会被误分割, 而且对小尺度物体的分割效果并不能令人满意. 一种提升网络多尺度分割能力的方法是对输入图像进行缩放, 将原始图像的不同分辨率版本输入到多通道的网络中, 再在顶部进行多尺度的特征融合, 在一定程度上能提升网络的分割表现, 但由于增加了输入通道的缘故, 增加了网络参数.

对于低分辨率的小目标的识别和检测一直是计算机视觉领域的难点问题, 传统的图像算法存在模型复杂度高、且泛化性能差等问题, 而早期的深度网络对于小目标的检测也完成的不够好. 早期的语义分割模型也存在同样的问题, 对于大尺度目标的分割较为准确, 而对于图像中的小目标分割不够精细, 甚至无法分割, 直接将其误判为背景. 为了解决这一问题, Raj 等[11]人提出了一种双路全卷积网络, 以 VGG-16 作为主干网络, 将原图像送入其中一路网络, 再将原图像上采样两倍之后送入另一路网络, 最后将两个网络的结果进行融合, 经过卷积操作消除混叠效应之后进行上采样得到最后的输出, 其结果也显示出了网络对于多尺度变化的鲁棒性. 在这之后也有很多利用图像金字塔结构来获取图像的多尺度信息, 以提高网络对多尺度物体的分割性能的工作. 文献[12]提出了一种多尺度的 FCN, 训练多个 FCN 来提取不同尺度的特征, 在网络的尾部进行特征融合, 再对融合后的特征热点图反卷积操作得到上采样之后的输出分割图, 创新性的使用了分阶段的训练方法. 首先单独训练提取不同尺度特征的 FCN, 再将其和最后的卷积层连接, 最后对最后的卷积层进行微调, 且结果本身具有良好的拓展性, 可以方便的添加任意数量的训练模型.

本文以提升网络对多尺度物体的分割能力为目的, 创新性的提出了一种利用特征金字塔网络 (feature pyramid network) 整合多级特征的语义分割网络, 以基于 VGG-16 的 FCN 网络为主干, 将顶层的具有丰富的

语义信息的特征图送入前一阶段进行融合 (将特征图尺寸不变的层称为一个阶段, 每次抽取的特征都是每一个阶段的最后一层的特征图), 自顶向下的抽取特征图送入前一阶段进行融合, 对融合后的特征图进行卷积操作以消除融合的混叠效应, 利用高层的语义信息帮助低层的具有高分辨率的特征图判别图像中物体的语义类别, 然后将最后得到的特征热点图进行上采样, 直至恢复到输入图像尺寸大小. 经过实验验证, 本文提出的方法对于分割小尺度物体具有一定的先进性.

1 全卷积网络

传统的卷积神经网络由卷积层和全连接层组成, 卷积层进行特征提取, 全连接层提供了强大的分类能力. 但全连接层的存在限制了输入图像的尺寸, 使得其必须固定. 等提出了一种基于像素块的深度学习语义分割方法, 将截取目标像素点周围邻域内一固定大小的像素块送入卷积神经网络进行训练, 从而得到像素点的分类, 将语义分割问题完全转换成了分类问题, 而忽略了语义分割本身是一个像素稠密预测问题这一特性. 显然, 基于像素块的语义分割方法极大程度上受到全连接层需要固定输入尺寸这一特性的制约. 在全卷积网络提出之前, 就有学者考虑将全连接层转换为卷积层的尝试, 因为全连接层本身可以视作以一个和特征图同样大小的卷积核对特征图进行扫描, 而且因为卷积层权重共享的缘故, 转换为卷积形式之后网络在进行前向计算时效率更高, 而且摒弃了全连接层需要固定输入大小的缺点. Long 等人首次将全卷积网络应用于语义分割任务, 将传统卷积网络尾端的全连接层替换成卷积层, 这样输出的就不再是对应的类别, 而是对应的特征热点图, 同时为了解决因为卷积和池化操作对图像分辨率造成的影响而使用上采样操作恢复图像尺寸. 在上采样的过程中, 作者发现直接将最后一层输出的特征热点图 (尺寸为原图像的 1/32) 进行上采样得到的分割图不够精确, 一些细节无法恢复, 于是作者引入了一种跳跃连接的结构, 将 heatmap 上采样两倍之后与前一阶段的缩小为 1/16 的特征图相融合, 在将其上采样为输入图像的 1/8 大小, 与前一阶段的特征图融合, 再将其上采样至原图大小得到最后的分割结果, 由于更好的兼顾了全局信息和局部信息的缘故, 使得最后的分割结果更为精细.

2 特征金字塔网络

特征金字塔网络 (Feature Pyramid Networks, FPN) 是由何凯明等提出的, 具体结构如图 1 所示. 原先是为了应对多尺度目标检测这一具有挑战性的问题, 对小目标的检测取得了很好的效果. 物体的尺度变化带来的挑战几乎是所有的计算机视觉任务都要面对的难题, 语义分割也不例外[13]. 现有的基于 CNN 的语义分割网络当中都会包含池化层, 池化层有降低特征图尺寸同时聚合感受野的作用, 但这也带来了分辨率的损失, 并不利于语义分割此类像素稠密型的预测任务, 逐级的下采样会使得物体尺寸不断缩小, 原图中的小尺度物体在传递到深层网络时很有可能已经完全消失. 文献[14]提出了一种基于图像金字塔的方法, 将不同分辨率版本的输入图片送入网络, 再将各自得到的最后的特征热点图进行融合, 如此得到最终分割结果, 在一定程度上解决了多尺度的物体分割. 但这种基于图像金字塔的方法的缺点在于会引入额外的计算量, 降低了模型的效率. FPN 提出了一种基于特征金字塔的特征聚合方法, 除了自底向上的特征提取路径, 还增加了一路自上向下的语义信息传递路径, 实现了深层语义信息的传递, 使得浅层的特征图也具有了一定的语义信息, 每一层都可以输出对应的检测结果, 实现了端到端的多尺度检测任务. 同样的结果也可以用于语义分割任务中.

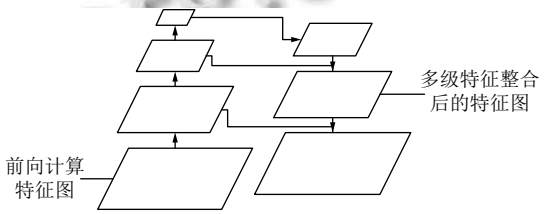

图 1 特征金字塔结构

3 多级特征融合的语义分割

3.1 多级特征融合

传统的卷积神经网络包含了卷积层和全连接层, 卷积层进行特征提取, 全连接层作为分类器. 和手工设计的传统算子不同的是, 作为特征提取器的卷积核并不需要很强的先验知识, 而是自动的从数据中学习规律, 提取出分类效果最好的特征, 这样的好处在于卷积核可以提取出人类无法理解的高维度特征, 且实验证明高维度的特征确实更有利于分类任务, 从而打破了

人类先验知识的一个束缚.

一般意义上, 研究人员认为卷积操作的特征提取是由浅到深、由低维到高维的一个过程. 浅层的卷积层提取到的一般是低层的特征信息, 比如颜色、线条、焦点等; 再往后就可以学习到边缘、纹理等具有一定区分度的信息; 深层的卷积层学习到的特征就更加完整, 具有明显的语义信息, 比如物体的具体轮廓、目标的位置信息等. 毫无疑问, 深层的、语义信息明确的特征更有利于我们的分类任务. 语义分割作为一种空间稠密型的预测任务, 不仅需要对图像中的像素进行正确的分类, 图像中的空间信息同样重要, 而在特征提取阶段往往需要进行多阶段的下采样, 导致了空间信息的损失. FCN 选择在上采样阶段逐级的融合上一阶段的特征信息, 以恢复损失的空间信息, 但这种融合方式存在一定的限制[15]. 文献[16]发现, 如果直接将浅层的特征图与深层的特征图进行融合, 不仅没有起到恢复空间信息的作用, 反而使得原有的分割结果更为糟糕. 这是因为浅层的特征图虽然很好的保留了空间信息, 但由于缺少明确的语义信息, 反而会对像素的分类带来干扰. 所以 FCN 的特征融合只进行到第三阶段的卷积层, 即只到原图像八分之一大小的特征图为止, 这也从侧面印证了我们只要保证进行融合的特征图具有相对程度的语义信息[17], 即能实现融合的有效性.

本文算法提出, 将特征金字塔结构嵌入初始的 FCN 中, 利用特征金字塔模块实现深层语义信息由深至浅的传递, 使得浅层的特征图在很大程度的保留空间信息的同时也具有了一定的语义信息, 能够参与到下阶段的上采样之中. 具体做法是, 由后往前的将深层的特征图进行传递, 通过上采样和 1×1 卷积核调整特征图的尺寸和通道数, 最后使得浅层特征图也具有一定程度的语义信息. 需要注意的一点是, 特征图的上采样在一定程度上会带来图像的混叠效应, 本文在每一个融合阶段之后采用了一个 3×3 的卷积核来消除混叠效应, 以得到语义信息更为明确的特征图. 最后将经过修正的特征图加入跳层连接的上采样路径, 逐层的恢复下采样过程中损失的空间信息, 得到最后的分割图. 有别于传统的图像金字塔方法, 基于特征金字塔的方法在聚合多级特征的同时有效的减少了模型本身的计算量.

3.2 语义分割网络结构

在上一小节中, 我们详细介绍了特征金字塔结构实现深层信息传递的具体方式, 本节中我们会详细介绍本文网络的具体结构.

FCN 以 VGG 为主干网络, 用三层卷积层代替了原本的全连接层, 使得网络具有了全卷积的结构, 紧接着添加一层卷积核尺寸为1×1的卷积层[18]调整输出通道数, 再利用 softmax 分类器预测每个像素位置上的得分, 然后对生成的特征热点图进行上采样恢复至输入图像大小. 本文的上采样方法选用的是双线性插值法, 公式如下所示

$$E(p)=\sum_{q}G(q,p)E(q) \tag{1}$$

其中, E 代表像素值, $G(,)$ 为双线性插值系数, p 代表上采样之后的像素位置, q 代表与其相邻的四个像素点位置. 双线性插值的好处在于可微, 便于网络进行端到端训练.

但对特征热点图进行一次直接上采样得到的结果比较粗糙[19], 原因在于池化层在获取更大感受野的同时丢失了一定的空间信息[20], 研究人员考虑到这一问题, 选择利用跳跃连接引入前一阶段的特征图修复空间信息, 这个过程只进行到中间层的特征图为止, 因为浅层的特征图语义信息不够明确, 会对像素预测造成干扰. 本文提出利用特征金字塔模型, 将深层的语义信息传递至浅层, 生成语义信息更为明确的浅层特征图, 再通过逐阶段的特征图融合和上采样生成最后的分割图. 网络具体结构如图 2 所示.

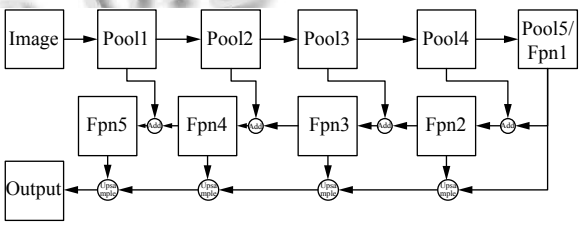

图 2 网络结构图

图 2 中, 由左至右的箭头代表下采样, 由右至左的箭头代表上采样, 由上至下的箭头代表对应阶段特征图的融合操作, 本文采用像素值直接相加的融合方式.

当训练样本进入网络之后, 经过五层池化层的下采样, 每次池化的步长为 2, 图像细粒度会变为原来的 1/32, 图 2 中的 pool1 阶段包含了两层卷积层, 经过两层的特征提取之后进行池化下采样, 聚合特征的同时扩大感受野, 相似的, 后面紧跟的四个池化阶段进行同

样的操作, 只是内部卷积层的层数有所不同, pool2 阶段同样包含了两层卷积层, pool3、pool4、pool5 阶段都包含了四层卷积层, 整个网络的卷积核尺寸统一为 3×3, 激活函数采用 Relu 函数. 我们将 pool5 层输出的特征图称为特征热点图, 直接对热点图进行上采样得到的结果较为粗糙. 为了得到更为精细的分割结果, 逐级的融合前一阶段的特征图是有必要的, 本文利用深层特征图的反向融合来来增强浅层特征图的语义信息. 如图 2 所示, 将 pool5 即特征提取网络提取到的最深层特征和前一阶段的特征图, 即 pool4 融合, 利用双线性插值法把 pool5 的特征图上采样到 pool4阶段相同大小, 利用 1×1 卷积核调整 pool5 的通道数, 融合方式本文采用对应通道特征图的像素加, 我们将融合之后的特征图称为 fpn2, pool5 即为 fpn1. 同样的, 再将融合之后的 fpn2 往浅层进行逐级的传递和融合操作, 便得到了相应的 fpn3、fpn4 和 fpn5, 这便是进行信息增强之后的特征图. 进行完上述操作之后, 即便是浅层的特征图也具有了一定的语义信息, 可以参与到之后的上采样过程之中. 上采样阶段, 还是从 pool5 即 fpn1 开始, 逐级的使用双线性插值法[21]扩大热点图尺寸, 使用 1×1卷积核调整特征通道数, 需要注意的是, 这时候我们将每一级的通道数都调整为$n+1$, 最后得到了一组通道数和标签类别数相匹配的特征图, 最后在每个像素点上利用 softmax 分类器进行分类, 即能得到最终的分割图.

4 实验分析

本文选用的 TensorFlow-1.4.0 版本作为实验框架, 运算平台为 6 核 i7 处理器, 利用 GTX 1060 显卡进行 GPU 加速, 选择 Pascal VOC 作为实验数据集. Pascal VOC 是在检测和分割任务中常被用到的一个基准数据集, 包含了 20 个语义类别和 1 个背景类别[22]. 本文选取了其中了 1747 张样本作为训练集、874 张样本作为验证集合、1165 张样本作为测试集并可视化实验结果以对本文提出的算法进行评估.

本文采用两个指标来评价分割结果的好坏, 分别是均像素精度 (Mean Pixel Accuracy, MPA)

$$MPA=\frac{1}{k}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}} \tag{2}$$

和权频交并比 (Frequency Weight Intersectionover Union, FWIU)

$$FWIU=\frac{1}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ij}}\sum_{i=0}^{k}\frac{\sum_{j=0}^{k}p_{ij}p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{3}$$

前者指的是每一类像素的精度的平均值[23], 后者指的是在每一类出现的频率作为权重的条件下真实值和预测值的交集比上两者并集的平均值.

在进行实验时, 本文将所有图片统一尺寸为 1024×1024 再送入网络进行训练, 需要注意的一点是, 由于分割任务的标注方式是逐像素的, 所以本文采用裁剪而不是缩放的方式来改变输入图像的尺寸[24], 后者会导致标注信息的失效. 本文与未经过多级特征整合的 3 个版本的全卷积网络架构分割准确率进行比较, 结果如表 1 所示.

表 1 本文提出算法与各版本 FCN 准确度比较 (%)

方法	FCN-32s	FCN-16s	FCN-8s	本文
MPA	73.3	75.3	75.3	75.8
FWIU	81.5	83.1	83.6	83.9

传统的全卷积网络在经过第三层池化层之后特征图尺寸缩小为原图的 1/8, 经过第四、第五层池化层之后尺寸相应的缩小为 1/16 和 1/32, 将 1/32 的特征热点图经过双线性插值直接上采样变得到了 FCN-32s. 但研究者发现直接上采样的方法得到的结果不够精细, 于是在上采样路径中引入了“跳层连接”, 融合了第四层池化层和第三层池化层的特征图以补全空间信息, 得到了 FCN-16s 和 FCN-8s 版本的分割结果. 从结果上看, 融合的特征图越多, 分割出的结果更加精细, 对边缘等细节信息的刻画更为准确. 本文的算法在上采样路径之前进行多级特征的信息整合, 利用特征金字塔结构增强浅层特征图的语义信息, 在融合了在 FCN-8s 的基础上了, 又和第一层和第二层池化层的特征图进行融合, 利用浅层的特征图对分割结果进行空间信息的补全, 克服了传统全卷积网络无法充分利用特征图空间信息的缺陷. 从表 1 可以看出, 本文提出的算法在均像素精度 (MPA) 和权频交并比 (FWIU) 这两个标准上都要优于传统全卷积网络. 几种算法的分割结果如图 3 所示, 从图中我们可以看出, 经过特征整合之后,

分割结果的边缘更为平滑, 对细节的勾勒更为清晰, 可见本文提出的算法具有一定程度的先进性.

图 3 分割结果图

5 总结与展望

本文在传统全卷积网络的基础上, 在上采样路径之前, 利用特征金字塔网络进行多级特征信息的整合, 再利用特征融合之后的特征图补全初始分割结果的空间信息, 克服了传统全卷积网络无法充分利用浅层特征信息的缺点[25], 实现了对图像空间信息更好的恢复. 在 Pascal VOC 数据集上取得了 75.8% 的均像素精度和 83.9% 的权频交并比. 在保证更高的精度基础之上, 如何兼顾网络整体的运行速度, 以及在实际应用过程中的鲁棒性, 也是亟待解决的问题, 需要更多的努力.

参考文献

1 Garcia-Garcia A, Orts-Escolano S, Oprea S, *et al*. A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv: 1704.06857, 2017.

2 Shi JB, Malik J. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888. [doi: 10.1109/34.868688]

3 Rother C, Kolmogorov V, Blake A. "GrabCut": Interactive foreground extraction using iterated graph cuts. ACM SIGGRAPH 2004 Papers. Los Angeles, CA, USA. 2004. 309–314.

4 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2012. 1097–1105.

5 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 1–9.

6 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2015. 770–778.

7 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 3431–3440.

8 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014.

9 Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv: 1511.00561, 2015.

10 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. [doi: 10.1109/TPAMI.2017.2699184]

11 Raj A, Maturana D, Scherer S. Multi-scale convolutional architecture for semantic segmentation. Robotics Institute, Carnegie Mellon University, Technical Report. CMU-RITR-15-21. Robotics Institute, Carnegie Mellon University, 2015.

12 Chen LC, Papandreou G, Schroff F, *et al*. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv: 1706.05587, 2017.

13 Lin TY, Dollár P, Girshick RB, *et al*. Feature pyramid networks for object detection. arXiv: 1612.03144, 2017.

14 Chen LC, Zhu Y K, Papandreou G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European Conference on Computer Vision (ECCV). Cham, Switzerland. 2018. 801–818.

15 Zhao HS, Shi JP, Qi XJ, *et al*. Pyramid scene parsing network. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 2881–2890.

16 Zhang ZL, Zhang XY, Peng C, *et al*. ExFuse: Enhancing feature fusion for semantic segmentation. arXiv preprint arXiv: 1804.03821, 2018.

17 Yu CQ, Wang JB, Peng C, *et al*. Learning a discriminative feature network for semantic segmentation. arXiv preprint arXiv: 1804.09337, 2018.

18 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 7132–7141.

19 Peng C, Zhang XY, Yu G, *et al*. Large kernel matters — Improve semantic segmentation by global convolutional network. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 1743–1751.

20 Chen LC, Yang Y, Wang J, *et al*. Attention to scale: Scale-aware semantic image segmentation. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2015.

21 Zhou BL, Zhao H, Puig X, *et al*. Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision, 2019, 127(3): 302–321. [doi: 10.1007/s11263-018-1140-0]

22 Rakelly K, Shelhamer E, Darrell T, *et al*. Conditional networks for few-shot semantic segmentation. 6th International Conference on Learning Representations, ICLR 2018 Workshop. 2018.

23 Rakelly K, Shelhamer E, Darrell T, *et al*. Few-shot segmentation propagation with guided networks. arXiv preprint arXiv: 1806.07373, 2018.

24 Cipolla R, Gal Y, Kendall A. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 7482–7491.

25 Kirillov A, Girshick R, He KM, *et al*. Panoptic feature pyramid networks. arXiv preprint arXiv: 1901.02446, 2019.