湖南大學
HUNAN UNIVERSITY

《高等计算机体系结构》

题　　目：基于深度网络的图像语义分割

姓　　名：林福生

学　　号：S191000871

科　　系：19 级计算机科学与技术硕士

任课老师：吴强

完成日期：2019.12.30

基于深度网络的图像语义分割

摘要

图像语义分割不仅预测一幅图像中的不同类别，同时还定位不同语义类别的位置，具有重要的研究意义和应用价值。近年来,深度学习技术已经广泛应用到图像语义分割领域。基于深度学习的语义分割方法相对与传统方法在效果上有了很大的提升。本文主从语义分割的基本概念开始阐述，然后对语义分割的发展以及研究现状进行了分析。同时，文中还介绍了 2 种经典的语义分割网络——DeepLab 语义分割网络和 U-Net 语义分割网络。最后，对本文所叙述的内容做了简要总结。

关键词：语义分割；计算机视觉；DeepLab；U-Net

Abstract

Image semantic segmentation not only predicts different categories in an image, but also locates the location of different semantic categories, which has important research significance and application value. In recent years, deep learning technology has been widely used in the field of image semantic segmentation. Compared with the traditional methods, the semantic segmentation method based on deep learning has greatly improved the effect. This paper starts from the basic concept of semantic segmentation, and then analyzes the development and research status of semantic segmentation. At the same time, this paper also introduces two kinds of classical semantic segmentation networks—DeepLab and U-Net. In the end, the content of this paper is summarized briefly.

Keywords: semantic segmentation; computer vision;DeepLab;U-Net

目录

1 引言

语义分割是计算机视觉的基本任务之一，是计算机视觉中图像理解的一部分。在微生物自动检测领域，语义分割是基础。在生活中，语义分割可以应用于无人驾驶、地理信息系统和医疗影像分析等领域。例如，在智能汽车领域，通过对无人车前景物体图像进行语义分割可以有效地帮助计算机判断路况；在医疗领域，通过对医学图像进行语义分割可帮助医生迅速分析和判断患者病情。

图像语义分割是一个非常具有挑战性的问题，其难点主要体现在以下两个方面：一是类别层面上所面临的难点，即类内实例间的相异性和类间物体的相似性；二是复杂的背景，实际场景中的背景往往是错综复杂的，这种复杂性大大提升了图像语义分割的难度。

图像语义分割方法有传统方法和基于卷积神经网络的方法，其中传统的语义分割方法又可以分为基于统计的方法和基于几何的方法．大多数统计方法是基于多个简单的特征，这类似于图像分割方法，利用人工设计的特征提取方法得到图像底层特征，没有训练过程，分割效果不理想。基于卷积神经网络的语义分割方法与传统的语义分割方法最大不同是，网络可以自动学习图像的特征，进行端到端的分类学习，极大提升语义分割的精确度。

2 相关工作

1998 年，Lecun 最早提出了 LeNet 网络[1]，并设计了卷积神经网络的 3 层结构：卷积层、池化层、非线性层。该结构为深度学习技术在图像领域的成功应用奠定了坚实的理论基础。

当前基于深度学习的图像语义分割方法的主流思想是将图像分类的经典网络作为基网络，根据具体的应用场景，对基网络进行改进并提升语义分割性能，以适应场景理解的需要。

2012 年，Hinton 研究组提出了 AlexNet[2]，首创了深度卷积神经网络模型。该网络在 LeNet 的基础上调整了网络架构并加深 了网络深度。AlexNet 在当年的 ImageNet[3]竞赛中表现优异并获得了冠军。

2014 年，牛津视觉几何研究团队的 Simonyan 等提出卷积神经网络 VGG[4]，赢得了 2014 年 ImageNet 竞赛的冠军。

VGG 采用与 AlexNet 相似的 5 层结构，将网络分为 5 组，使用 3 × 3 过滤器，并将其组合作为一个卷积序列进行处理。VGG 网络与之前模型的主要不同在于：VGG 网络在第 1 层使用了一批小感受野(Receptive field)[5]的卷积层，使得模型的参数更少，非线性更强，也因此使 得决策函数更具区分度，模型更好训练。

为减少神经网络的计算开销，2014 年 Szegedy 等设计了第一个 Inception 架构的网络 GoogleNet[6]。Inception 的思路是减少每一层的特征过滤器的数目，从而减少运算量。这种新的方法证实了 CNN 层可以有更多的堆叠方 式，而不仅仅是标准的序列方式。

此外，2016 年 He 等提出的 ResNet 网络[7]以其高达 152 层的深度以及引入的残差模块而闻名。残差模块使得网络下一层可以同时掌握前一层的

输出以及原始的输入，从而调整学习；该模块的连接方式也协助解决了梯度消失问题。

鉴于神经网络的优良表现，研究学者纷纷将深度学习应用到语义分割领域，随后又提出了FCN(Fully Convolution Network,FCN)[8]，SegNet[9],DeepLab V1[10],DeepLab V2[11]，DeepLab V3[12],DeepLab V3+[13]，RefineNet[14]，PSPNet[15]，BiSeNet[16]等模型。其中FCN不仅开启了像素级语义分割，更开拓了之后语义分割算法使用全卷积网络的新思路，使语义分割打破了传统方法的限制，从而提高了分割精度。另外，针对语义分割中孢子图像的样本不平衡及难易样本问题，Y Zhao等人提出了Constrained Focal Loss[17]用以提高孢子图像分割的性能。

3 语义分割

语义分割是计算机视觉应用的研究热点之一。语义分割中的“语义”指的图像里面的内容；“分割”

(a) (b)

图1 语义分割示例

如图1所示为一个图像语义分割的例子，在图1(a)中有3个类别：人、自行车和背景，语义分割所要做的便是推断图像中每一个像素点所属的类别并打标签。某个像素点是人这一类的，打上人的标签，属于自行车的像素点便打上自行车的标签，标签的形式可以表现为不同的颜色或是不

针对当前基于深度学习的散乱点云语义特征提取方法通用性差以及特征提取不足导致的分割精度和可靠性差的难题，彭秀平等[18]提出了一种散乱点云语义分割深度残差-特征金字塔网络框架。汪梓艺等[19]提出了一种改进的改进的Deeplab V3烟雾分割算法用于烟雾检测。宋国杰等[20]使用九点双三次卷积插值方法替换DeepLab v3模型的双线性插值方法，以获得更精确的分割图像。陈天华等[21]提出一种改进的全卷积网络，其以为网络前端，结合Inception结构，在不降低特征提取能力的前提下，通过减少网络参数数量，降低网络运算复杂度，在一定程度上提升了网络的训练速度。

即要把图像中的不同内容分割开、区分开。语义分割以类别为单位对图像进行像素级别的理解，其要准确地预测出图像中每一个像素点所属的类别并为每一个像素点打上类别标签。

同的灰度值。例如可以把属于人这一类的像素点标成粉红色，把属于自行车的像素点标成绿色，背景为黑色。

4 经典语义分割网络

4.1 deeplab v3+网络架构

Deeplab v3+是由谷歌公司所提出并开源的语义分割网络，发展至今已有deeplab v1、v2、v3和v3+四个版本。在4个版本中，分割性能逐渐提升，deeplab v3+的性能最优。其在Cityscapes数据集上的平均交并比为82.1%；在Cityscapes PASCAL VOC 2012数据集上测试平均交并比已达

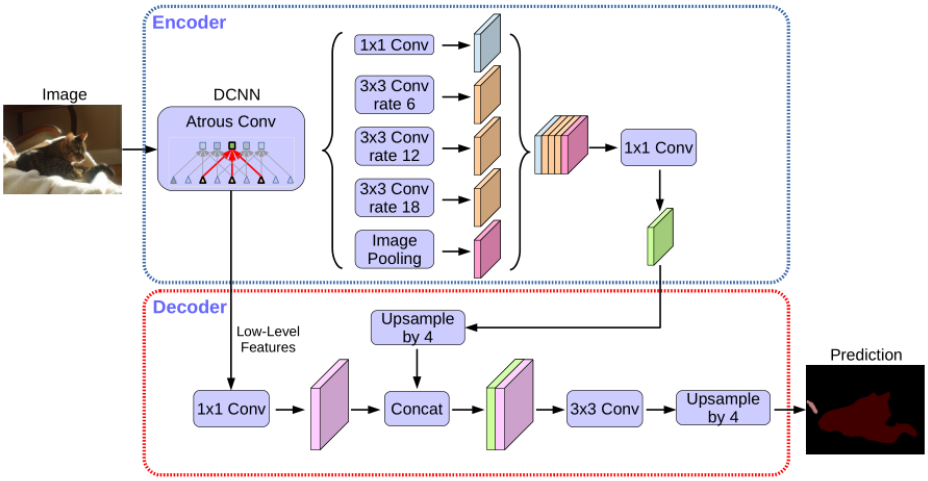

图 2 deeplab v3+整体网络架构图[13]

到了 89.0%的水平。

Deeplab v3+整体的网络架构如图 2 所示。deeplab v3 +大体上可以划分为编码模块和解码模块。在编码模块中，deeplab v3+利用深层残差网络[7] (Residual network,Resnet) Resnet-101 对图像进行有效的特征提取，并且 deeplab v3+在 Resnet-101 上融合了具有不同空洞率的孔洞卷积[22] (Dilated c onvolution) 捕获图像的多尺度特征；受空间金字塔池化 (Spatial Pyramid Pooling,SPP) 和孔洞卷积的启发，deeplab v3+引入了孔洞空间金字塔池化 (Atrous Spatial Pyramid Pooling,ASPP)，同时又融合了深度可分离卷积[17] (Depthwise Separable Convolutions) 技术，减少网络计算量的同时提高了模型的性能，最后通过一个简单而有效的解码模块恢复图像的尺寸，能够在一定程度上提升物体边缘的预测性能。

4.1.1 残差网络

在一定程度上，神经网络的深度与网络性能存在着一定的正比关系，即越深的网络代表着越强的拟合能力及更好的性能。然而并不是简单地把网络堆叠得更深就能得到性能更加优异的网络模型，实验研究表明，当网络的深度增加到一定的阈值后，网络性能将会发生退化：伴随着网络的加深，精准率将会达到饱和状态，随后精准率会快速地下降并且随着网络的进一步加深，精准率会进一步地降低。可以知道的是，网络退化问题并不是因为模型的过拟合。因为过拟合表现为在训练集上效果好而在测试集上的效果差，但过深的网络在测试集和训练集上的表现都很差，因此可以判断这并不是过拟合。

神经网络采用误差反向传播算法进行梯度更新，神经网络从输出层开始，不断地向后传播梯度。梯度在向后传导的过程中会渐渐降低，导致后面网络层的权重无法得到有效的更新，这就是所谓的梯度消失问题。网络越深梯度消失问题就越严重。

之前说到，随着网络深度的增加，模型的性能会达到一个饱和值，再继续加深网络的深度模型的性能将会明显的下降，该过程可用图 3 进行

描述：

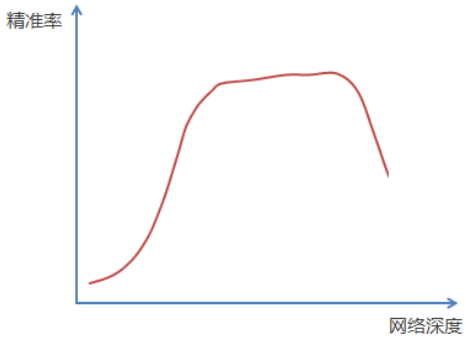

图 3 网络深度与准确率的关系

残差网络通过引入恒等变换解决了网络退化的问题。残差网络的基本元素为残差块，残差网络主要由多个残差块串联构成。残差块的结构如图 4 所示，与普通网络结构相比，残差块多了一个跨层的快捷连接，因为这个快捷连接的存在，使得网络可以设计得更深、拟合能力可以更强。

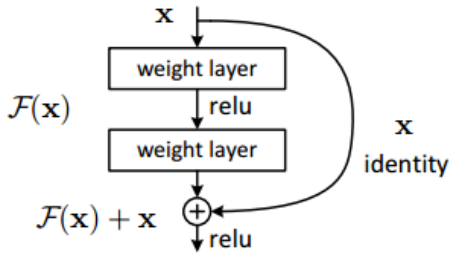

图 4 残差块结构[7]

4.1.2 空洞卷积

语义分割是一项极具挑战性的任务因为它需要结合上下文信息进行像素级别的推理预测。传统的卷积网络通常会通过池化操作或者下采样操作整合多尺度上下文信息，而池化操作或者下采样会降低特征图的尺寸。语义分割是像素级别的预测，输出预测图的尺寸必须与原图尺寸相同，通过池化操作或下采样提取特征再通过上采样恢复为原图大小必然会丢失掉一些信息，这将会损失一部分语义分割的正确率。

空洞卷积的最大贡献在于无需降低图像的分辨率或者对图像进行缩放就能很好地聚合多尺度上下文信息。并且空洞卷积能够应用到现有的任一网络中而无需考虑图片分辨率的问题。空洞卷积抛弃传统的池化和下采样操作，能做到在不损失图像分辨率的状态下以指数增长方式增加感受野。

相比传统的卷积方式，空洞卷积增加了空洞率这一参数。在实际应用中，可以根据需要设置不同的空洞率，空洞率越大，感受野越大。空洞卷积可以使感受野呈指数方式增加，如图 5 所示，(a) 为传统的卷积，其感受野大小为 3×3；(b) 为空洞率为 2 的空洞卷积，其感受野为 7×7；(c) 为空洞率为 4 的卷积，其感受野为 15×15。

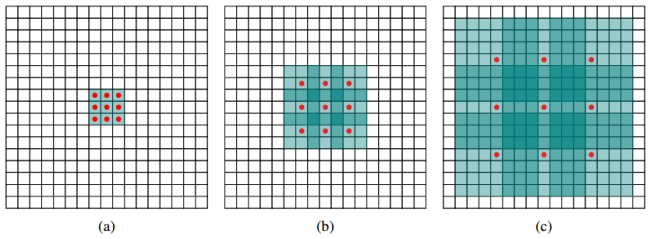

图 5 不同空洞率的孔洞卷积示意图[22]

4.1.3 深度可分离卷积

一个有效的、性能较好的网络模型往往拥有大量的权重参数，这意味着需要消耗大量时间和计算资源。深度可分离卷积的应用能够在一定程度上减少网络模型的参数，提高模型速度，并且能略微提升模型的性能。传统的卷积方式往往同时考虑区域和通道，与传统卷积方式不同，深度可分离卷积采用先通道，再区域的方式，这种方

式能够大大地减少参数量，使得模型拥有更快的速度。

4.1.4 孔洞空间金字塔池化

通过在同一个输出上应用不同的孔洞率而捕获图像的多尺度信息。空间金字塔池化展示了不同尺度的重采样能够进一步改善网络的性能，受空间金字塔池化启发，deeplab v3+采用了 SPP 的变体，孔洞空间金字塔池化 (ASPP)。在 deeplab v3+中，ASPP 有着孔洞率为 (6,8,12) 的 3 个并行空洞卷积。实验表明，拥有不同空洞率的 ASPP 可以很好地获取图像中的多尺度信息。除了三个并行的空洞卷积之外，ASPP 还加入了一个 1×1 的卷积和一个图像级别的池化操作。

4.1.5 编解码结构

编解码结构已经被成功地应用到许多计算机视觉任务中，包括目标检测、人体关键点检测和语义分割等。编解码结构主要由编码模块和解码模块两部分组成，直观上讲，编码模块主要用来提取图像的高级特征，显然经过编码模块图像的尺寸将大大减小；解码模块则以编码模块提取到的小尺寸特征图为输入，逐渐地恢复成原图大小。

Deeplab v3 +采用了一个简单而有效的解码结构，如图 1 下半部分所示，该解码结构先经过 4 倍的双线性上采样，得到的特征图与经过 Resnet-101 提取的低级特征相融合，在此之前，经 Resnet-101 提取的特征经过一个 1×1 的卷积以减少特征图通道数，此后再经过一个 3×3 的卷积对特征图作进一步的校准，最后经过一个 4 倍的双线性上采样恢复成原图大小并输出预测图。

4.2 Unet 语义分割网络

Unet 是一个简单而有效的语义分割网络，因其网络结构形状类似于英文字母“U”，因此称为 U-net。训练一个有效的网络模型通常需要成千上万的训练样本，由于种种因素的限制，我们往往无法获取足够多的训练样本集来训练网络模型，尤其在医学影像处理方面数据集尤为匮乏，无法得到足够多的样本也是深度学习面临的难题之一。

Unet 在某些领域例如医疗影像分割方面的表现优异，其在充分做好数据扩充的情况下，只需少量的训练样本便能达到很好的分割效果，并且在速度上也不落后于其他网络。

如图 6 所示，Unet 网络结构较为简单。Unet 可“掰”为左右两部分，左半部分用来不断地压缩特征图尺寸，提取更高级的特征，可称为压缩路径，主要用来捕获图像的上下文信息。由于语义分割是在像素级别上的预测，因此必须把图片尺寸恢复成原图大小，这一部分工作主要由 Unet 的右半部分完成。右半部分通过不断地上采样逐渐恢复图片尺寸为原输入大小和还原图像信息。左右两半部分合起来类似一个“U”型形状。

Unet 是基于全卷积网络[8] (Fully Convolutional Networks, FCN) 而改造的语义分割网络，Unet 在全卷积网络的基础上进行了修改和扩展使其运用少量的训练样本便能对图片进行精准地分割。在 Unet 网络架构中，Unet 在上采样部分也设置大数量的特征通道，有利于上下文信息更好地传播到更深的网络层，这也是 Unet 的一个重要改进。Unet 中不含任何的全连接层，只由

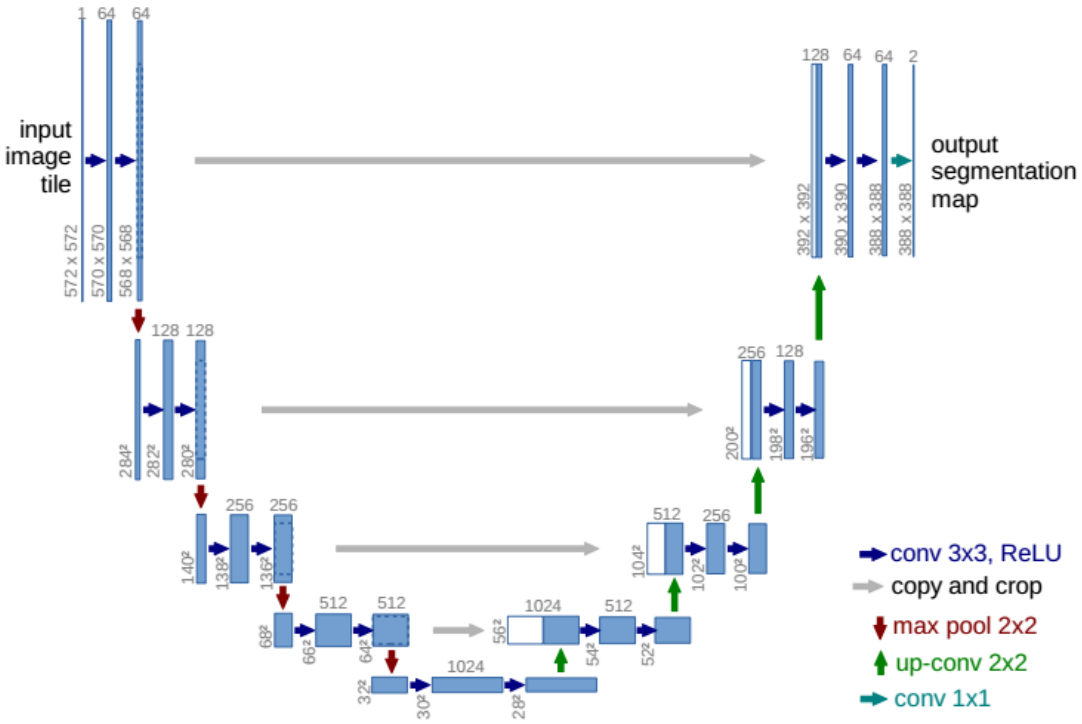

图 6 Unet 网络结构

单一的卷积层组成，是“全卷积”网络架构，因此能够接受任意大小的图像输入，不受图片大小限制，灵活性较好。

在 Unet 中，不管在压缩或是扩展路径中，输出的特征图都比输入图片的尺寸小。这种策略的优点在于处理大尺寸图像的时候有优势，即通过这种方式能够减少计算量和 GPU 显存占用，特别是在 GPU 条件受限的情况下。但这种方式也存在着边缘信息的丢失问题，Unet 通过在压缩路径中复制一份输入镜像到扩展路径中以恢复边缘信息，如图 6 中间长箭头所示。

如图 6 所示为 Unet 的整体网络架构，左半部分为压缩路径，右半部分为扩展路径，压缩路径的设计遵循传统的经典的卷积神经网络。压缩路径主要由几个重复的模块：2 个串联 3×3 卷积+线性整流单元（Rectified Linear Unit,ReLU）+1 个最大池化操作组成，每一次经最大池化下采样都使特征通道数翻倍。与左半部分压缩路径类似，右半部分的模块头部仍由 2 个串联的卷积层组成，但第 2 个卷积层的通道数是第 1 个的一半，且在进行卷积之前压缩路径的特征图会与从压缩路径传过来的特征图“镜像”相融合，这一步骤不能或缺因其有助于恢复丢失的边界信息；卷积后接一个线性整流单元，最后进行上采样操作逐渐恢复图像大小，每次上采样操作特征图通道数也会减半。最后有一个 1×1 的卷积层进行预测输出。

5 语义分割评价指标

语义分割的性能指标一般用平均交并比（Mean intersection over union,mIoU）来衡量。mIoU 用来计算真实值和预测值的重叠程度，即准确率，其值域在 0 和 1 之间，mIoU 为 1 则表示预测完全正确，反之 mIoU 则为 0。

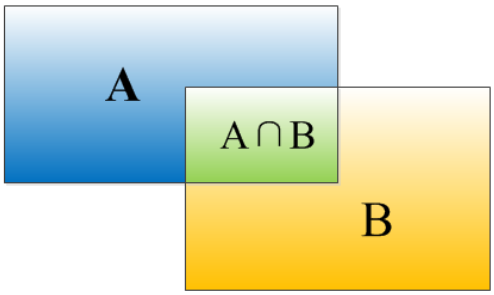

图 7 交并比示意图

如图 7 所示，$A \cap B$ 表示预测正确的部分，除 $A \cap B$ 的其它部分表示预测错误的部分，则交并比为 $A \cap B$（预测正确的部分）与 $A \cup B$（预测正确+预测错误）的比值，其计算方式如式（1）所示：

$$IoU = \frac{A \cap B}{A \cup B} \qquad (1)$$

6 总结

如今，深度学习技术已经广泛应用到图像语义分割领域.本文主要对语义分割及其国内外研究现状进行了较为详细的阐述。同时也对基于深度学习的图像语义分割的经典网络进行了较为细致的分析。其中，Deeplab v3+利用深度卷积网络对图像特征进行有效提取，同时又融合了空洞卷积而且引入多孔空间金字塔池化提取多尺度特征；利用深度可分离卷积的优点减少模型的参数量和计算量，在一定程度上提升了模型的速度；deeplab v3+又加入了编解码技术，能够更好地分割物体边缘。相比 deeplab 网络，Unet 的设计显得格外简单。Unet 主要是由一条压缩路径和一条扩展路径组成，类似一个“U”型，压缩路径对图片进行编码提取特征，扩展路径对编码后的特征进行解码恢复图像尺寸。

语义分割在深度学习时代下取得了飞速的进步，但是语义分割仍然有很多问题需要克服，目前还远称不上已经解决，更准确的分割边界，小物体的分割，实时性语义分割等问题仍然是一个挑战。

参考文献

[1] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[2] Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural[C]//Neural Information Processing Systems. 2014: 1-9.

[3] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International journal of computer vision, 2015, 115(3): 211-252.

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

[5] Liu Y, Yu J, Han Y. Understanding the effective receptive field in semantic image segmentation[J]. Multimedia Tools and Applications, 2018, 77(17): 22159-22171.

[6] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[7] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[8] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[9] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.

[10] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv preprint arXiv:1412.7062, 2014.

[11] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.

[12] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv:1706.05587, 2017.

[13] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

[14] Lin G, Milan A, Shen C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1925-1934.

[15] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.

[16] Yu C, Wang J, Peng C, et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 325-341.

[17] Zhao Y, Lin F, Liu S, et al. Constrained-Focal-Loss Based Deep Learning for Segmentation of Spores[J]. IEEE Access, 2019, 7: 165029-165038.

[18] 彭秀平,仝其胜,林洪彬,冯超,郑武.一种面向散乱点云语义分割的深度残差-特征金字塔网络框架[J/OL].自动化学报:1-10[2019-12-28].

[19] 汪梓艺,苏育挺,刘艳艳,张为.一种改进DeeplabV3 网络的烟雾分割算法[J/OL].西安电子科技大学学报:1-8[2019-12-28].

[20] 宋国杰,黄佳芳,陈普春,陈亚丽.使用九点双三次卷积插值方法改进的 Deep Lab-v3 模型[J/OL].计算机应用研究:1-6[2019-12-28].

[21] 陈天华, 郑司群, 于峻川. 采用改进 DeepLab 网络的遥感图像分割[J]. 测控技术, 2018, 37(11): 34-39.

[22] Yu F , Koltun V . Multi-Scale Context Aggregation by Dilated Convolutions[C]// ICLR. 2016.

[23] Chollet, François. Xception: Deep Learning with Depthwise Separable Convolutions[J]. 2016.