ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 6,Nov. 2021,pp. 1263－1275
DOI:10. 16337/j. 1004-9037. 2021. 06. 019

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

融合U-Net改进模型与超像素优化的语义分割方法

王振奇[1]，邵　清[1]，张　生[1]，杨　振[2]，何国春[1]

(1. 上海理工大学光电信息与计算机工程学院，上海 200093；2. 上海外高桥造船有限公司工艺研究所，上海 200120)

摘　要：基于现有的语义分割方法在面对不受限制的开放词汇量和多样多变的场景时表现出的分割不够精细、语义信息提取不充分和收敛时间长的问题，提出一种融合U-Net改进模型与超像素优化的语义分割方法。U-Net改进模型中结合空间金字塔模块(Atrous spatial Pyramid pooling，ASPP)和Xception结构，在ASPP模块的分支网络中加入扩张卷积(Dilated convolutions，DC)形成模块本身的串并联结构，以增强图像特征提取能力；在Xception模块中添加注意力通道以及使用大的卷积核重构Xception模块，以减少数据的参数量并提高收敛速率，在此改进基础上再对图像进行超像素分割处理。最后使用条件随机场对分割结果施加全局约束，进一步优化像素的语义信息。本文方法在PASCAL VOC 2012测试集上进行验证并与DeepLab V3等主流网络进行对比，结果表明本文方法准确率提高了2.4%，证明了该方法在适应多变场景和应对精细语义分割上的有效性。

关键词：图像语义分割；空间金字塔池化；U-Net模型；超像素分割；条件随机场

中图分类号：TP391.9　　**文献标志码：**A

Semantic Segmentation Method Integrating U-Net Improvement Model and Superpixel Optimization

WANG Zhenqi[1]，SHAO Qing[1]，ZHANG Sheng[1]，YANG Zhen[2]，HE Guochun[1]

(1. School of Optoelectronic Information and Computer Engineering，Shanghai University of Technology，Shanghai 200093，China；2. Institute of Technology，Shanghai Waigaoqiao Shipbuilding Co.，Ltd.，Shanghai 200120，China)

Abstract：Facing unrestricted open vocabularies and diverse scenes，present semantic segmentation methods have the problems of insufficient segmentation，insufficient semantic information extraction and long convergence time. Therefore，this paper proposes a semantic segmentation method that combines U-Net improvement model and superpixel optimization. The U-Net improvement model combines the atrous spatial pyramid pooling (ASPP) and the Xception structure. Firstly，the dilated convolutions (DC) is added to the branch network of the ASPP module to form the serial-parallel structure of the module itself，thus enhancing the image feature extraction capability. And the attention channels are added to the Xception module and a large convolution kernel is used to reconstruct the Xception module，thus reducing the amount of data parameters and increasing the convergence rate. On the basis of the above improvements，the image is then subjected to the super pixel segmentation processing. Finally，conditional

基金项目：国家自然科学基金(61703278)资助项目；上海市科学技术委员会科研计划(19511105103)资助项目。
收稿日期：2020-11-06；**修订日期：**2021-01-12

random fields are used to impose global constraints on the segmentation results to further optimize the semantic information of pixels. The proposed method is verified on the PASCAL VOC 2012 test set and compared with mainstream networks such as DeepLab V3. Experimental results show that the performance accuracy of the proposed method is increased by 2.4%, which proves the effectiveness of the proposed method in adapting to diverse scenes and dealing with the fine semantic segmentation.

Key words: image semantic segmentation; spatial Pyramid pooling; U-Net model; superpixel segmentation; conditional random field

引　言

图像语义分割是计算机视觉范畴中一个重要的研究方向。研究初期,图像语义分割主要依靠人工标注特征[1-3],但是此类方法过于依赖研究人员的主观判断,难以广泛表达图像特征,在实际应用过程中具有相当大的局限性。随着深度学习技术的发展,基于深度卷积网络[4]的图像分割应运而生并得到广泛应用。文献[5]采用全卷积神经网络(Fully convolution network, FCN)进行像素级别分类和端到端的分割,此网络可以达到在接收不同尺寸图像的同时极大地提高分割效率,但仍存在分割不准确的问题。文献[6]提出新的分割网络SegNet,在全卷积网络的基础上增加了解码器,将模型明确划分成编码网络和解码网络,形成目前分割任务中普遍盛行的编解码结构。文献[7]提出反卷积网络,网络中上采样的方法是学习一个与卷积网络成镜像结构的反卷积网络,该网络较好地解决了简单上采样方法所导致的信息丢失问题。文献[8]提出U-Net分割网络,采用U形结构得到高效的全卷积神经网络,可以实现仅对少量图片训练达到较高的准确率,在医学图像分割上表现较为突出。文献[9]提出PSPNet分割网络,运用空间金字塔池化(Spatial Pyramid pooling, SPP)辅助实现背景融合的同时利用基于相异像素块的前后文聚集,充分发挥整体综合信息的能力从而产生优异的分割效果。文献[10]提出DeepLab V2分割网络,首次采用ASPP模块提高了特征信息的保有程度,进一步提高分割的准确度。文献[11]提出了图卷积网络(Graph convolutional network,GCN),设计了一种带有大型卷积核的编解码器结构来收集较小范围的特征信息,同时改进分割模型解决了分割过程中的“像素分类”和“像素本地化”问题。文献[12]提出带有可分离的编解码器网络结构,在特征提取网络中融合空间金字塔卷积模块与编解码器结构进行语义分割,可以有效恢复边界信息。

语义分割发展到现在依旧面临难以把握图像特征和图像语义信息之间平衡的问题。一方面,图像的特征往往需要较大的卷积核进行提取,然而小的卷积核却能获得更多的语义信息[13]。另一方面大的卷积核会带来相对多的参数从而增大计算机运算负担。池化操作可以提取底层信息却会丢失特征分辨率。恢复语义分割对象的结构有两种常用的方法:(1)从不同的卷积神经网络(Convolutional neural network, CNN)层组合信息,并在图像或特征空间中构造更多的上下文关系;(2)从不同的CNN层组合信息,并跳过层架构,获取来自不同层聚合的多粒度信息。减少层间结构信息的丢失,提高插值的有效性,是恢复更多结构信息的关键。

基于以上分析,本文提出一种融合U-Net改进模型与超像素优化的语义分割方法。该方法通过设计特征提取与超像素(Simple linear iterative clustering, SLIC)[14]的双通道网络以及全连接条件随机场(Condition random field, CRF)[15]的后处理,提高分割效率,恢复分割过程中丢失的细节,同时大幅度降低图像语义分割的算法复杂度。本文方法主要有以下3点贡献:

(1)改进ASPP模块。通过在ASPP的分支网络基础上结合扩张卷积(Dilated convolution,DC)[16],形成模块本身的串并联结构,改进ASPP模块可以达到扩展感受野而减少分辨率损失的目的。

(2)改进Xception模块。在Xception结构上结合大的卷积核,融入注意力精炼模块结构,改进后的Xception模块可以保留图像更多的细节信息,减少数据的参数量并提高收敛速率。

(3)基于以上创新,提出新的语义分割架构,结合U-Net模型与SLIC算法并通过CRF后端处理进一步提高物体边界附近的定位性能,提高分割精细度。

1 相关方法

1.1 U-Net模型

U-Net模型将编解码器结构和跳跃连接相结合,只需要很少的注释图像便可以取得较好的分割效果。在U-Net模型中收缩路径的结构遵循卷积网络的典型构造,包括2个卷积的重复应用,并在卷积后连接1个激活单元(ReLU[17]激活)和1个池化操作用于下采样。在下采样流程中,特征通道数翻倍,上采样经过反卷积运算后通道数收缩。扩张路径包括以下5个阶段:特征映射上采样;同收缩路径的对应特征映射进行串联;2个卷积叠用;每个卷积由ReLU激活;扩张网络与收缩网络基本对称并形成U形结构。

1.2 ASPP模块

DeepLab V2提出的ASPP模块在指定的网络层并行应用多个差别采样率的卷积,空洞用0填充,相当于利用了多尺度视野的多个核滤波器来获取图像中的不同特征,因而提升了对多尺度物体的分割能力。如图1所示,ASPP解决了FCN语义分割中向上卷积不能完全还原池化导致的细节损失问题。ASPP模块基于扩张卷积支持指数感受野扩充,系统地聚合了多尺度上下文信息。空洞卷积感受野F的计算公式为

$$F=k+(k-1)(r-1) \tag{1}$$

式中:k为卷积核大小;r为其对应空洞卷积采样率大小。

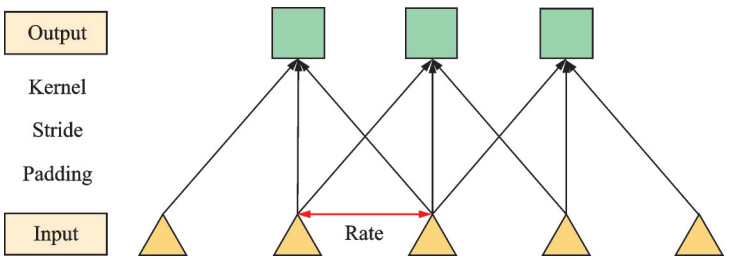


图1 ASPP稀疏特征提取结构图

Fig.1 Structure graph of ASPP sparse feature extraction

1.3 SLIC超像素分割

为了优化经过特征提取后图像的粗糙边缘,本文采用SLIC算法以提高图像边缘的分割准确率。SLIC算法是简略线性迭代聚类,它将色彩图像转为XY坐标下的特征向量,而后对特征向量构建间隔衡量标准,并对图像内像素点采用局部聚类。具体步骤如下:

(1)种子点初始设置。假定待分割的图像有N个像素,将其处理为p个同样大小的超像素种子,则每一个超像素的尺寸为N/p,而且相邻种子间间隔近似为S,计算公式为

$$S=\sqrt{N/p} \tag{2}$$

为了防止种子点位于图像的边际位置对后续的聚类进程造成干扰,种子距离在领域内选择。本文方法将种子点的领域设为3×3,同时为每个种子分配单独的标签。

(2) 类似度权衡。对每个搜索到的像素点计算相对间隔最近的种子间类似度,将与之最类似的种子标签赋予该像素。持续迭代此过程直至收敛。类似度权衡的距离指标包含像素间色彩差异 d_{lab} 和像素间的空间间隔 d_{xy},计算公式为

$$d_{\text{lab}}=\sqrt{(l_p-l_i)^2+(a_p-a_i)^2+(b_p-b_i)^2} \tag{3}$$

$$d_{xy}=\sqrt{(x_j-x_i)^2+(y_j-y_i)^2} \tag{4}$$

$$D_i=d_{\text{lab}}+\frac{m}{S}d_{xy} \tag{5}$$

式中:l_p、l_i、a_p、a_i、b_p、b_i 分别表示每个像素点聚类前后的颜色值;x_j 和 y_j 分别为像素点在 X 轴和 Y 轴两个坐标方向上的值,同样 x_i 和 y_i 分别为聚类中心在2个坐标轴上的坐标值;D_i 为种子之间的类似度距离;m 为均衡参数,用来权衡色彩值与空间信息在类似度权衡中的比重。如果 m 的取值越大则表明生成的超像素形状越规则,D_i 越大则类似度越低。

1.4 基于CRF的分割后处理

本文将CRF引入语义分割的后处理。在CRF模型中,以图片像素点为单位的标注作为随机变量,将像素与像素之间的对映关系作为边,如此便可构成一个CRF。具体来说,CRF中拥有输入图像的 N 个像素点,体现了整体观测 I。随后给定图 $G(V,E)$,其中 V 和 E 分别为给定图的对应顶点和边。假设 $\boldsymbol{X}$ 是由随机变量 $\{X_1,X_2,\cdots,X_n\}$ 组成的向量,其中 X_i 为随机变量,体现为给像素 i 分配的标注。CRF符合Gibbs分布,且元组 $(I,\boldsymbol{X})$ 可以被建模为

$$P(\boldsymbol{X}|I)=\frac{1}{Z(I)}\exp\left(-\sum_{c\in C_G}\phi_c(X_c|I)\right) \tag{6}$$

式中:$G=(\nu,\varepsilon)$ 为 $\boldsymbol{X}$ 向量上的图;$E(\boldsymbol{X}|I)=\sum_{c\in C_G}\phi_c(X_c|I)$ 为标记 $x\in\boldsymbol{\Gamma}^N$ 的Gibbs能量;$Z(I)$ 为分割函数。而CRF应用的能量函数为

$$E(x)=\sum_i\varphi_u(x_i)+\sum_{i<j}\varphi_p(x_i,x_j) \tag{7}$$

式中:$\varphi_u(x_i)$ 为一元势能;$\varphi_p(x_i,x_j)$ 为二元势能。二元势能进而对相近像素点之间的联系进行建模描述,并由色彩之间的类似性进行加权,其表达式为

$$\varphi_p(x_i,x_j)=\mu(x_i,x_j)\sum_{m=1}^{K}\omega^{(m)}k^{(m)}(\boldsymbol{f}_i,\boldsymbol{f}_j) \tag{8}$$

$$K^{(m)}(\boldsymbol{f}_i,\boldsymbol{f}_j)=\exp\left(-\frac{1}{2}(\boldsymbol{f}_i-\boldsymbol{f}_j)^{\mathrm{T}}\boldsymbol{\Lambda}^{(m)}(\boldsymbol{f}_i-\boldsymbol{f}_j)\right) \tag{9}$$

式中 $\boldsymbol{f}_i$ 与 $\boldsymbol{f}_j$ 为像素 i 和 j 在相应位置自由维度的特征向量,且核函数 $k(m)$ 的规模由对称矩阵 $\boldsymbol{\Lambda}^{(m)}$ 判断。

在使用CRF进行精确边缘恢复过程中,一元势能是由边缘优化后的语义标签,相较于本文基于U-Net改进模型提取的粗糙特征,经过边缘优化后的像素级语义标签更加利于CRF模型的性能发挥。

2 本文模型

2.1 模型框架

本文方法的基本框架如图2所示,其中XceptionP表示改进的Xception网络。图像首先通过基于U-Net模型改进的场景分析特征提取网络来提取特征信息获得语义标签,之后运用SLIC提取边缘信

息,通过结合特征提取网络获得的粗糙特征优化粗糙分割结果。最后使用条件随机场对结果施加整体约束,进一步优化每个像素的语义信息,从而得到兼具像素高级语义信息和较好边缘贴合度的图像语义分割结果。

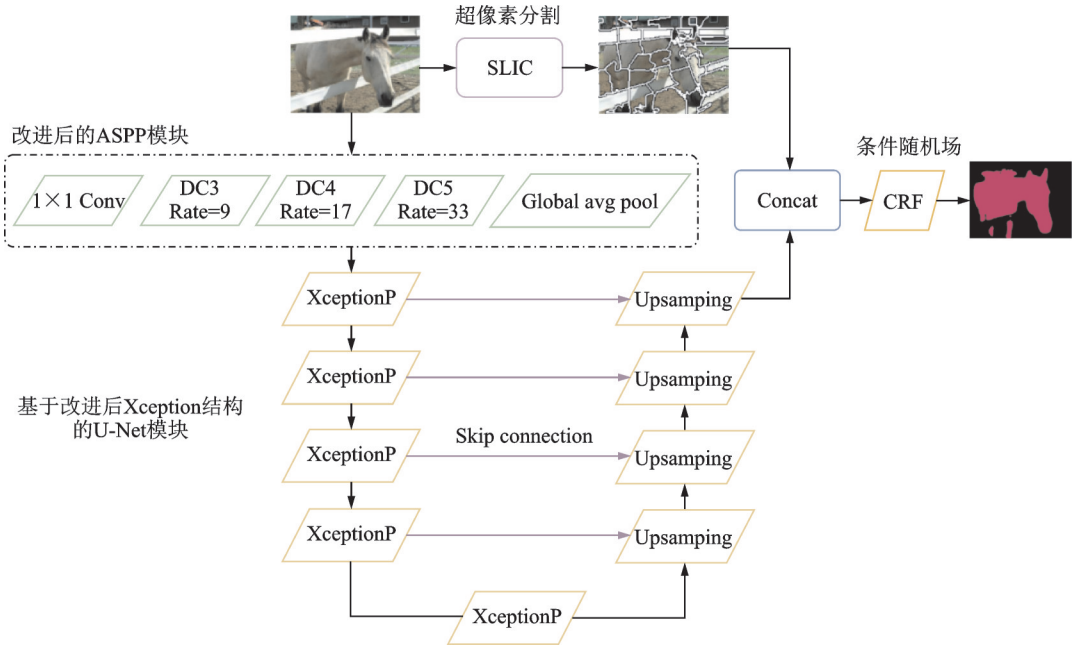


图2 融合U-Net模型和SLIC的语义分割架构

Fig.2 Semantic segmentation architecture combining U-Net model and SLIC

2.2 相关模块的改进

2.2.1 融合U-Net模型的特征提取网络

场景分析特征提取网络的核心是U-Net模型,基于改进后Xception结构的U-Net模块如图3所示,分为上采样和下采样。

(1)网络模型的下采样。在特征提取网络中首先通过改进后的ASPP模块对图像以多尺度、多层次的方式提取特征。然后对处理后的特征数据进行连续下采样。下采样中的卷积使用改进Xception模块提取图像中像素流信息,实现多个通道相互关联以及图像特征空间充分解耦。

(2)网络模型的上采样。因为考虑到反卷积可能产生的网格效应,本文恢复图像运用上采样中的双线性插值法,期间使用1×1卷积恢复通道数目。在上采样过程中取下采样对应相同分辨率的特征图进行合并,每次合并后再经过2个3×3卷积继续细化特征图,依次上采样直至将编码器中所提取的特征还原到输入图片尺寸时的大小。

2.2.2 ASPP模块的改进

本文对ASPP模块的改进借鉴了DC扩张卷积网络结构。DC扩张卷积网络基础的上下文模块由多层3×3的不同膨胀系数的空洞卷积组成,膨胀系数分别为{1,1,2,4,8},不同膨胀系数的卷积感受野也不同,通过融合不同感受野的卷积组成串联结构,实现不降低感受野的同时尽量多地保留特征信息。

为了防止向上下文网络提供输入的特征模块生成过低分辨率的特征图像,结构中选择停止了第6

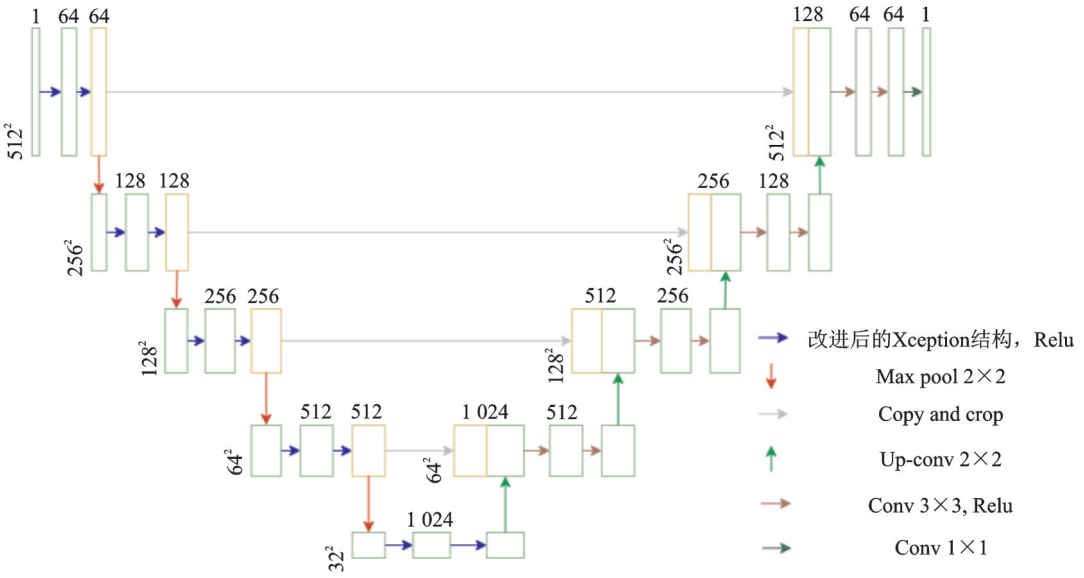


图3 基于改进后Xception结构的U-Net结构图

Fig.3 Structure diagram of U-Net based on improved Xception structure

层之后感受野的指数扩展。考虑到参数量问题本文截取前5层的上下文网络结构，详细参数如表1所示，表中C为基础通道数。

表1 上下文网络结构

Table 1 Context network structure

卷积层	1	2	3	4	5
卷积核	3×3	3×3	3×3	3×3	3×3
扩张倍率	1	1	2	4	8
感受野	3×3	5×5	9×9	17×17	33×33
输出通道(基础通道)	C	C	C	C	C
输出通道(扩张通道)	$2C$	$2C$	$4C$	$8C$	$16C$

改进的ASPP模块通过将DC结构中的{3,4,5}层与ASPP模块相结合组成各分支网络串联、各分支间并联的结构，达到扩大感受野而不损失分辨率的目的。改进后的模块结构如图4所示。在该模块中将带有全局内容的图像信息先通过Relu函数激活，之后经过5个卷积通道，再次由Relu函数激活并通过1×1卷积修改通道数。

通道设置过程中，前4个通道分别是1×1卷积、9×9感受野的DC3卷积(DC3指3个3×3卷积的串联结构，共同作用达到9×9感受野的效果)、17×17感受野的DC4卷积(DC4指4个3×3卷积的串联结构，共同作用达到17×17感受野的效果)、33×33感受野的DC5卷积(DC5指5个3×3卷积的串联结构，共同作用达到33×33感受野的效果)，每个卷积的扩张倍率如表1所示；第5个通道是全局池化层用以提高模型性能、减少过拟合。最后将分支处理好的图像特征细节进行拼接再共同通过1×1卷积聚合，并且在每一个卷积后接正则化层操作(Batch normalization，BN)以加快网络训练和收敛的速度。改

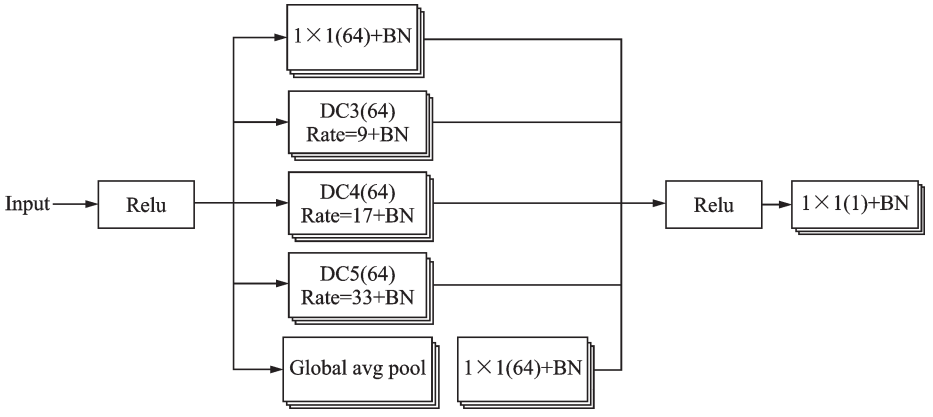


图4 改进后的ASPP模块结构图

Fig.4 Improved ASPP module structure diagram

进后的ASPP结构中滤波器个数为64。

2.2.3 Xception模块的改进

为了进一步提取图像中像素流信息,本文对Xception模块[18]加以改进。Xception模块是分类网络中常用的一种并行网络结构,其结合ResNet结构[19]并对大的卷积核充分解耦[20-21],能够兼顾准确精度和计算效率。本文将原有的Xception结构进行扩展,提高了特征细节的提取能力。

改进后结构的右侧由5条并行的路径组成,其中(2,3,4,5)条路径使用卷积核大小分别是(1×1,3×3,5×5,7×7)的卷积以差别提取图像特征尺寸下的信息(5×5和7×7的卷积分别使用2个3×3以及3个3×3卷积进行替代,可以明显减少网络参数量且保证感受野不会变小),在每个路径的卷积前会对输入的图像特征信息先通过1×1卷积来降低输入的通道数目。第1条路径是与ResNet相结合的ShortCut结构,解决了深度卷积过程中的梯度消失问题。所有路径的卷积中均使用Same填充来确保输入与输出的统一,而后将每条路径的结果在末端通道维上进行聚合,并输出到相连的下一层中继续提取特征。

改进后结构的左侧先将上层输入的图像特征信息连续进行4倍、8倍、16倍和32倍的下采样,选取其中16倍和32倍下采样结果,分别通过注意力精炼模块(Attention refinement module,ARM)[22]优化输出特征,整合整体语境信息后将全局平均池化的输出与保留更多细节的9×9大核卷积输出相结合,进而优化语义路径的输出结果。

改进后的Xception模块结构如图5所示。模块拥有更多的池化通道,加入注意力精炼模块并采用较大的9×9卷积核,可以解决全连接以及全局池化等相关操作失去位置信息的问题,模块中卷积使用$1\times k+k\times 1$和$k\times 1+1\times k$代替$k\times k$,可以显著减小参数量以及网络计算成本,同时保有更多的特征细节,获得更好的分割效果。

2.2.4 SLIC边缘优化算法

在融合通过特征提取网络获得的特征基础上,通过SLIC算法细腻贴合的边缘信息来重新标注每个超像素内的语义标签,对图像边缘进行优化,从而更好地还原图像的边缘信息。具体优化算法流程如下。

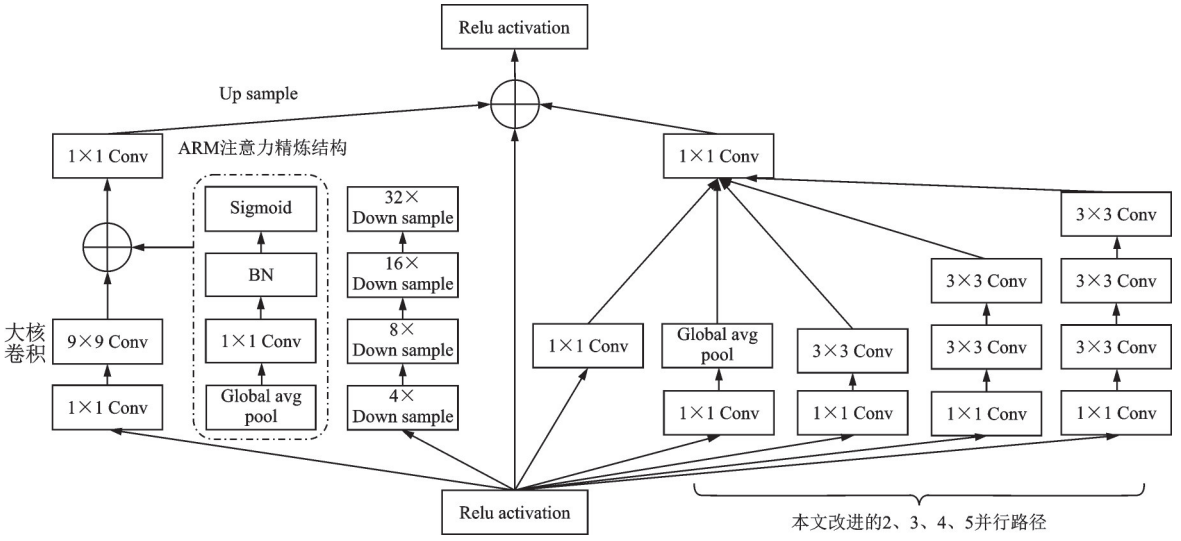


图5 改进后的Xception模块结构图

Fig.5 Structure diagram of improved Xception module

算法 SLIC边缘优化

设定输入图像为I,经过特征提取网络提取出的粗糙特征为L。

利用SLIC算法获得超像素p个,$O=\{O_1,O_2,\cdots,O_p\}$,其中O_i表示第i个超像素。

外部:while $i<p$ do :

(1)用$C=\{C_1,C_2,\cdots,C_p\}$体现O_i中的全部像素,其中C_j为被类别j所标记的像素点。

(2)从特征提取网络获得C中每个像素的特征进而初始化权重W_c。

(3)内部:while $j<Q_i$:

在while循环中判断C_j在C中属于哪一种语义分类,如果C_j表示C_k,令Q_k增加1,其中Q_k表示指向C_k的总像素数。

将C_j的特征标签保存为L_{C_j},而后更换超像素内一切标签的权重$W_{C_j}=W'_{C_j}+\frac{1}{N}$,其中$W'_{C_j}$表示$W_{C_j}$的上一个值。若$W_{C_j}>0.8$,则退出内部循环,否则持续循环。

end while

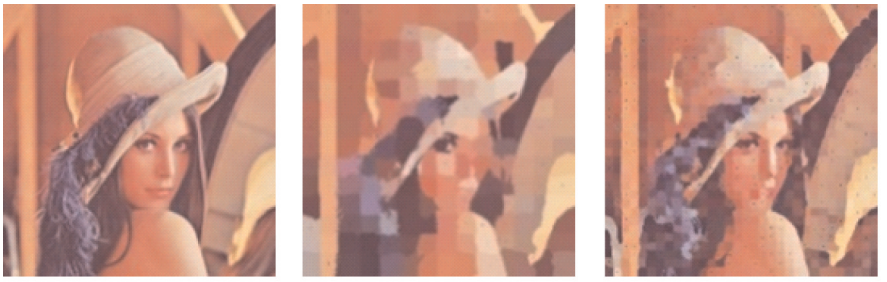
(4)搜查所有的W_C并判断是否具有某一W_{C_j}值大于0.8。若有,执行下一步;否则,持续寻觅最大的$W_{\max}$和第二大的W_{sub},而后判断它们之间的插值是否大于0.2。如果是,则进行下一步;否则持续外部循环。

(5)应用目前超像素内几率最大的分类$L_{C_{\max}}$,而后再次标注目前超像素的语义标签。

end while

得到最后经过超像素优化后的分割结果并且输出$\tilde{I}$。

图6显示了超像素迭代分割的结果。可以看出,经过超像素处理的图像内容边缘更加清晰。为提高边缘分割的准确度,本文设置基础p值为1 000。



(a) Input image (b) Segmentation result (p=500) (c) Segmentation result (p=1 000)

图6 超像素迭代分割结果

Fig.6 Segmentation results of superpixel iteration

3 实验设计与分析

3.1 实验数据集与评估指标

本文采用PASCAL VOC2012数据集[23]作为基准数据进行实验,可以达到较好的训练分割网络并准确从场景中分割对象的目的。该数据集提供了1组带有标签的相关图像的训练集,是基准测试最广泛使用的语义分段数据集。它由20个前景对象类和1个背景类组成。实验中使用其中的5 000张注释图像,并将其分为2 975/500/1 525张图像分别作为本文模型的训练、验证和测试使用。

为了更好地评估模型方法的分割精度,本文采用均交并比(Mean intersection over union,mIoU)和像素精度(Pixel accuracy,PA)作为评估规范。在图像语义分割领域mIoU值是权衡图像分割精度的重要指标,它体现了计算真实数值和预测数值2个集合的交并集之比,也即在每个类别上计算交并比(Intersection over union,IoU)值。mIoU和PA计算公式为

$$\mathrm{mIoU}=\frac{1}{K+1}\sum_{i=0}^{K}\frac{p_{ii}}{\sum_{j=0}^{K}p_{ij}+\sum_{j=0}^{K}p_{ji}-p_{ii}} \tag{10}$$

$$\mathrm{PA}=\frac{\sum_{i=0}^{K}p_{ii}}{\sum_{i=0}^{K}\sum_{j=0}^{K}p_{ij}} \tag{11}$$

式中:K表示除背景类别之外的总类别个数;$K+1$表示包括背景类别在内的语义类别总数;i为真实值;j为预测值;p_{ji}表示将i预测为j。

3.2 实验环境与参数设置

3.2.1 实验环境

本文实验中CPU为Intel Core i7;内存为24 GB;GPU为NVIDIA GTX 1050Ti 4 GB;语言环境为Python 3.7;机器学习环境为PyTorch 1.3;Cuda版本为Cuda 10.0 with cudnn。

3.2.2 参数配置

模型训练采用“poly”学习率策略,其中当前学习率等于基数乘以$\left(1-\frac{\mathrm{iter}}{\max_i \mathrm{ter}}\right)^{\mathrm{power}}$。为了防止过拟合,本文将基础学习率设置为0.007,将power设置为0.9。通过适当提高迭代次数可以增强性能,PAS-

CAL VOC设置为30 000次。本文对PASCAL VOC数据集采取随机镜像并让图像的随机大小在0.5～2之间进行调整;然后在－15°～15°之间加入随机旋转,对数据集使用随机高斯模糊,这种综合的数据加强方案使网络具备抗过拟合能力。

3.3 实验结果分析

3.3.1 实验结果

实验中用到2个重要参数:(1) 裁剪尺寸;(2) 批次标准化处理层的批量尺寸。这2个参数的大小设置会影响到分割网络的性能。本文在训练过程中将批量大小设置为8,能够取得较好的效果。为了找到最好的结合效果,进行了以下对比实验。

实验1为不同设置的分割模型对比。针对多变多样的场景,本文将分割模型的不同设置进行对比,对比结果如表2所示。表2中:Xception-Baseline为基于ResNet-101的网络结构;XceptionP-Baseline为基于改进后Xception模块的U-Net扩张网络结构;Xception＋A0＋MAX为DeepLab V3前端网络的基本配置;A0表示未改进的ASPP空洞卷积结构;A1表示改进后的ASPP空洞卷积结构;MAX和AVE分别为最大池化操作与平均池化操作;BR表示在池化之后进行维度适当缩减。实验1结果在PASCAL VOC数据集上用单标度输入进行测试。由表2中mIoU和PA值可知,本文网络在特征提取准确率方面具有优势,而平均池化相较最大池化对本文网络的结合工作更好且使用改进后的金字塔空洞卷积,随着特征图像的缩小以及网络深度的增加,网络性能可以得到进一步提升。

表2 不同设置的分割模型对比

Table 2 Comparison of segmentation models with different settings %

方法	mIoU	PA
Xception-Baseline	38.23	72.30
XceptionP-Baseline	40.34	78.02
XceptionP＋A1＋MAX	42.75	79.36
XceptionP＋A1＋AVE	43.05	79.85
XceptionP＋A1＋MAX＋BR	42.12	79.68
XceptionP＋A1＋AVE＋BR(本文)	43.67	81.05
Xception＋A0＋MAX(DeepLab V3)	41.25	78.73

实验2为采用不同的图像数据流提取模块时,数据参数量以及收敛速度的对比,对比结果如表3所示。表3中:XceptionP-Non-Residual为本文改进后且去掉残差结构的Xception模块;Resnet-152表示Resnet网络的152层实现,增加了Resnet网络深度且不会过拟合。由表3可知"XceptionP-Non-Residual"的数据参数量最低,但同时它收敛速率最慢,本文采用的"XceptionP"模块在收敛速率上最快,同时数据参数量也做到了比"Resnet-152"要低。这一方面说明了残差网络有助于提升网络学习的效率,另一方面也说明了本文"XceptionP"模块可以一定程度降低数据的参数量。

表3 不同模块的参数量与收敛速度对比

Table 3 Comparison of parameters and convergence speed of different modules

结构	参数量/10^7	测试时间/s
XceptionP	1.27	0.065
XceptionP-Non-Residual	1.19	0.073
Resnet-152	1.37	0.069

3.3.2 主观评估

图7给出了本文方法与DeepLab V3的对比结果,可以看出本文方法分割的精细程度较高,在第1行图片中的人腿部分可以完整凸显,第2行的飞机下部侧翼得以保留,第3行的鸟类则减少了错误识别的情况,第4行中可以看到在复杂细节场景中本文方法更具优势。

(a) Input image (b) Label image (c) DeepLab V3 (d) The proposed model

图7 不同分割模型产生的语义分段效果

Fig.7 Semantic segmentation effects produced by different segmentation models

3.3.3 客观评估

表4对当前主流模型(FCN, PSPNet, DeepLab V3)在PASCAL VOC 2012测试集上的图像精度以及每类结果的准确率进行详细对比。从结果可以看到,本文方法在对比实验中具有一定优势:mIoU值达到84.2%,较DeepLab V3提高了2.4%;而PA值为95.42%,较DeepLab V3提高了1.57%,这主要得益于前端网络的高级特征提取与超像素边缘优化相结合取得的效果。

表4 不同分割模型评估指标对比

Table 4 Performances of different segmentation models %

方法	mIoU	PA
FCN	62.20	71.43
PSPNet	79.24	81.69
DeepLab V3	81.80	93.95
本文方法	84.20	95.42

表5给出了30 000次迭代次数下不同分割模型在测试集中每类结果的对比,可以看到本文方法在准确率上整体优于其他方法,在测试集所有20个分类中多数类别的准确率较高。

4 结束语

本文结合U-Net模型、ASPP模块、Xception模块和SLIC超像素设计了一个融合U-Net改进模型与超像素优化的语义分割新架构,用于解决面对多变场景时分割不精细、语义信息提取不足的问题。在PASCAL VOC2012数据集上的实验验证了本文方法的有效性,提出的模型在与多种主流模型对比中均表现出极佳的性能,获得更加优异的分割结果,包括更低的参数量、更快的收敛速度和更精细准确的边界。如何在分割中更好地结合超像素以及如何更好地设计语义信息提取网络是下一步工作需要继续研究的重点。

表5 30 000迭代次数下不同分割模型对比

Table 5 Comparison of different segmentation models under 30 000 iterations %

分割模型	FCN	PSPNet	DeepLab V3	本文模型
aero	76.8	92.6	91.8	96.2
bike	34.2	60.4	71.9	73.9
bird	38.9	91.6	94.7	96.0
boat	49.4	67.4	71.2	74.1
bottle	60.3	76.3	75.8	76.1
bus	75.3	95	95.2	96.7
car	76.7	88.4	89.9	87.9
cat	77.6	92.6	95.9	96.8
chair	21.8	32.7	39.3	44.1
cow	62.5	88.5	90.7	92.6
table	46.8	67.6	71.7	82.3
dog	71.4	89.6	90.5	91.2
horse	63.9	92.1	94.5	94.2
mbike	76.5	87.5	88.8	94.1
person	73.9	87.4	89.6	89.7
plant	45.2	63.3	72.8	71.2
sheep	72.4	88.3	89.6	93.0
sofa	37.4	60.0	60.4	68.2
train	70.9	86.8	85.1	88.4
tv	55.1	74.5	76.3	76.5
mIoU	62.2	79.2	81.8	84.2

参考文献：

[1] KADOTA R, SUGANO H, HIROMOTO M, et al. Hardware architecture for HOG feature extraction[C]//Proceedings of IEEE Intelligent Information Hiding and Multimedia Signal Processing. Piscataway: IEEE, 2009: 1330-1333.

[2] ZHOU H, YUAN Y, SHI C. Object tracking using SIFT features and mean shift[J]. Computer Vision and Image Understanding, 2009, 113(3): 345-352.

[3] 田萱, 王亮, 丁琪. 基于深度学习的图像语义分割方法综述[J]. 软件学报, 2019, 30(2): 440-468.
TIAN Xuan, WANG Liang, DING Qi. An overview of image semantic segmentation based on deep learning[J]. Journal of Software, 2019, 30(2): 440-468.

[4] LIU Y, CHENG M M, HU X, et al. Richer convolutional features for edge detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 3000-3009.

[5] JONATHAN L, EVAN S, TREVOR D. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39(4): 640-651.

[6] 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.

[7] NOH H, HONG S, HAN B. Learning deconvolution network for semantic segmentation[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2015: 1520-1528.

[8] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]//Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer,

2015: 234-241.

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

[10] 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.

[11] PENG C, ZHANG X, YU G, et al. Large kernel matters-improve semantic segmentation by global convolutional network [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 4353-4361.

[12] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: IEEE, 2018: 801-818.

[13] HAN H, FAN L. Excavating effective information in different stage of backbone to improve semantic segmentation results[J]. Journal of Physics Conference Series, 2019 (1325): 012081.

[14] ACHANTA R, SHAJI A, SMITH K, et al. SLIC Superpixels compared to state-of-the-art superpixel methods[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2012, 34(11): 2274-2282.

[15] KRAHENBÜHL P, KOLTUN V. Efficient inference in fully connected CRFS with Gaussian edge potentials[C]//Proceedings of Advances in Neural Information Processing Systems. Massachusetts: MIT Press, 2011: 109-117.

[16] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[EB/OL]. (2016-04-30)[2020-10-30]. https://arxiv.org/abs/1511.07122.

[17] FARABET C, COUPRIE C, NAJMAN L, et al. Learning hierarchical features for scene labeling[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(8): 1915-1929.

[18] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//Proceedings of Thirty-First AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2017.

[19] YANG Qiangpeng, CHENG Mengli, ZHOU Wenmeng, et al. IncepText: A new inception-text module with deformable PSROI pooling for multi-oriented scene text detection[EB/OL]. (2018-05-08)[2020-10-30]. https://arxiv.org/abs/1805.01167.

[20] HOU F, LIU B, ZHUO L, et al. Remote sensing image retrieval with deep features encoding of inception V4 and Largevis dimensionality reduction[J]. Sensing and Imaging, 2021. DOI: https://doi.org/10.1007/s11220-021-00341-7.

[21] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-12-05)[2020-10-30]. https://arxiv.org/abs/1706.05587v1.

[22] 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). [S.l.]: Springer, 2018: 325-341.

[23] EVERINGHAM M, ESLAMI S M A, GOOL L V, et al. The pascal visual object classes challenge: Aretrospective[J]. Internatiaonal Journal of Computer Vision, 2015, 111(1): 98-136.

作者简介:

王振奇(1994-),男,硕士研究生,研究方向:图像处理,E-mail:1141121548@qq.com。

邵清(1970-),**通信作者**,女,博士,副教授,研究方向:网络智能与自然语言处理,E-mail:sq_usst@126.com。

张生(1968-),男,高级工程师,研究方向:智能制造、计算机应用,E-mail:zhangsheng@usst.edu.cn。

杨振(1987-),男,高级工程师,硕士,研究方向:船舶智能制造,E-mail:350442501@163.com。

何国春(1999-),男,本科,研究方向:自然语言处理,E-mail:528881398@qq.com。

(编辑:张黄群)