基于区域与深度残差网络的图像语义分割

罗会兰*① 卢 飞① 孔繁胜②
①(江西理工大学信息工程学院 赣州 341000)
②(浙江大学计算机科学与技术学院 杭州 310027)

摘 要：该文提出了一种结合区域和深度残差网络的语义分割模型。基于区域的语义分割方法使用多尺度提取相互重叠的区域，可识别多种尺度的目标并得到精细的物体分割边界。基于全卷积网络的方法使用卷积神经网络(CNN)自主学习特征，可以针对逐像素分类任务进行端到端训练，但是这种方法通常会产生粗糙的分割边界。该文将两种方法的优点结合起来：首先使用区域生成网络在图像中生成候选区域，然后将图像通过带扩张卷积的深度残差网络进行特征提取得到特征图，结合候选区域以及特征图得到区域的特征，并将其映射到区域中每个像素上；最后使用全局平均池化层进行逐像素分类。该文还使用了多模型融合的方法，在相同的网络模型中设置不同的输入进行训练得到多个模型，然后在分类层进行特征融合，得到最终的分割结果。在SIFT FLOW和PASCAL Context数据集上的实验结果表明该文方法具有较高的平均准确率。

关键词：语义分割；区域；深度残差网络；集成
中图分类号：TP391.41 **文献标识码**：A **文章编号**：1009-5896(2019)11-2777-10
DOI: 10.11999/JEIT190056

Image Semantic Segmentation Based on Region and Deep Residual Network

LUO Huilan① LU Fei① KONG Fansheng②
①(*School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou* 341000, *China*)
②(*School of Computer Science and Technology, Zhejiang University, Hangzhou* 310027, *China*)

Abstract: An image semantic segmentation model based on region and deep residual network is proposed. Region based methods use multi-scale to create overlapping regions, which can identify multi-scale objects and obtain fine object segmentation boundary. Fully convolutional methods learn features automatically by using Convolutional Neural Network (CNN) to perform end-to-end training for pixel classification tasks, but typically produce coarse segmentation boundaries. The advantages of these two methods are combined: firstly, candidate regions are generated by region generation network, and then the image is fed through the deep residual network with dilated convolution to obtain the feature map. Then the candidate regions and the feature maps are combined to get the features of the regions, and the features are mapped to each pixel in the regions. Finally, the global average pooling layer is used to classify pixels. Multiple different models are obtained by training with different sizes of candidate region inputs. When testing, the final segmentation are obtained by fusing the classification results of these models. The experimental results on SIFT FLOW and PASCAL Context datasets show that the proposed method has higher average accuracy than some state-of-the-art algorithms.

Key words: Semantic segmentation; Region; Deep residual network; Ensemble

收稿日期：2019-01-18；改回日期：2019-04-05；网络出版：2019-04-22
*通信作者： 罗会兰 luohuilan@sina.com
基金项目：国家自然科学基金(61862031, 61462035)，江西省自然科学基金(20171BAB202014)
Foundation Items: The National Natural Science Foundation of China (61862031, 61462035), The Natural Science Foundation of Jiangxi Province (20171BAB202014)

1 引言

图像语义分割结合了图像分割和目标识别任务，其目的是将图像分割成若干组具有特定语义含义的区域，并标记出每个区域的类别，实现从底层到高层语义的推理过程，最终获得一幅具有像素语义标注的分割图像[1]，即为图像中每个像素分配一个表示其语义目标类别的标签。图像语义分割在生活中有着很广泛的应用，如自动驾驶，地理信息系统，医疗影像分析以及虚拟或增强现实等穿戴式应用设备系统。越来越多新兴的应用领域需要精确和高效的分割机制，图像语义分割是计算机视觉任务中的研究热点之一。

目前，图像语义分割的方法主要有基于候选区域的方法[2–6]和基于卷积神经网络(Convolutional Neural Network, CNN)的全卷积方法[7–9]。

基于区域的方法首先在图像中提取颜色、纹理、形状等底层特征，然后根据提取的特征将图像分割成不同的区域，最后对区域进行语义分类，从而得到语义分割图像。文献[2]使用约束参数最小割(Constrained Parametric Min-Cut, CPMC)算法首先在图像中生成一组候选区域，并学习评分函数对这些候选区域进行评分，筛选出高分区域，然后根据其对应的类别按得分顺序进行聚合，从而得到最终的分割图像。在此基础上，文献[3]提出了2阶池化与线性分类器相结合的语义分割方法，使用CPMC算法生成候选区域，然后通过捕获图像的2阶统计量，采用2阶均值池化或2阶最大池化聚合区域的局部特征，并利用基于黎曼流形的正定对称矩阵空间来表示任意形状区域的特征，结合线性分类器得到最终的分割结果。CMPC算法是通过改变偏倚量的大小和选择种子点来产生候选区域，在此过程中由于缺乏先验知识使得种子点的选择无针对性，产生过多的不合理分割以及重复分割。因此，文献[4]提出的语义分割方法在生成候选区域阶段没有采用CPMC算法，而是使用轮廓检测和分层图像分割方法，该方法能够使得生成的区域在亮度、颜色和纹理等方面有很高的内部一致性，并且生成合理数量的区域。但文献[4]中的方法过于依赖手工标注的特征库，难以广泛表示图像特征，在实际应用中有很大的局限。为此，Girshick等人[5]提出了一种用于语义分割和目标检测的卷积神经网络方法，称为R-CNN。该方法首先在一张图像上采用选择搜索(selective search)算法[10]来生成约2000个候选区域，并将候选区域全部归一化成同一尺寸(227×227)，然后在每一个候选区域上都使用卷积神经网络提取特征，最后使用CNN得到的特征为每个类别构造支持向量机(Super Vector Machine, SVM)分类器。为避免对候选区域归一化操作，He等人[11]提出了空间金字塔池化(spatial pyramid pooling)层，该层可以接收任意大小的输入并产生固定维度的输出，在R-CNN的卷积层之后加入金字塔池化层，每幅图像的特征只需要被提取1次，无需重复性地对重叠区域进行耗时的卷积操作，提升了算法效率。为了对候选区域进行端到端的分类，Girshick[6]又提出了Fast R-CNN方法，首先将生成的候选区域输入CNN中，并且在CNN最后一层卷积层后使用感兴趣区域(Region-Of-Interest, ROI)池化层得到固定大小输出，然后使用全连接层代替SVM分类器进行分类，将特征提取、分类以及边框回归整合到一起进行端到端训练。

图像中的目标存在不同的尺度，基于区域的方法首先使用区域生成算法将图像分割成大小不同的候选区域，因此更容易捕获各个尺度的目标，同时使得目标更容易被识别，并且在物体边界处可以得到精细的分割结果[12]。但是这类方法的效果受到生成的候选区域质量的影响，而且生成候选区域过程比较耗时。因此，一些学者提出了基于全卷积神经网络(Fully Convolutional Network, FCN)的方法来弥补基于区域方法的不足，利用大量有像素级标注的图像作为训练样本，直接学习从像素到像素类别标签的映射。

2015年，文献[7]提出了用于语义分割的全卷积网络，该方法设计了一种可以接受任意大小的输入图像并且进行端到端训练的全卷积网络框架，实现对每一个像素分类，开创了使用全卷积网络来实现图像语义分割的先河。全卷积网络在原本用于图像分类的卷积神经网络的基础上，使用卷积层代替最后的全连接层，将分类网络转化为生成分割图像的网络。由于深度卷积网络中深层卷积特征分辨率低并且缺少空间位置信息，文献[7]通过反卷积将特征图上采样到原图大小，并使用跳层结构组合中间层输出的特征图。在文献[7]的基础上，Chen等人[8]提出了DeepLab V1模型框架，移除最后两层最大池化层，并用空洞卷积(atrous convolution)代替普通卷积，避免了局部信息的丢失，得到了更精确的预测。为了捕获不同尺度的目标和充分利用图像中的上下文信息，Chen等人在2016年进一步提出了DeepLab V2[9]方法，在文献[8]的基础上增加了多尺度输入以及多孔空间金字塔池化(atrous spatial pyramid pooling)方法，首先将图像缩放为3个不同的尺度输入到并行CNN分支中进行特征提取，然后使用不同采样率的空间金字塔池化层在多个尺度

上捕捉目标和图像内容，最后通过双线性插值将特征图还原到原图大小，得到最终的分割结果。

一些文献结合利用基于区域方法更容易捕获多尺度目标的优点，以及CNN自动特征提取的优点，将基于区域方法和基于CNN方法结合起来用于语义分割[13–15]。Hariharan等人[13]将每个候选区域从原图中裁剪出来，归一化之后输入CNN中提取区域和区域前景两种特征，串联起来训练SVM分类器，并进行区域精细化处理，但是该方法在特征提取阶段耗时太长。Dai等人[14]提出卷积特征掩膜(Convolutional Feature Masking, CFM)方法，只需要对原图提取1次特征，然后将原图特征映射到每个候选区域上，利用CFM层得到区域前景特征，但是该层参数固定，不能通过反向传播来更新参数。Caesar等人[15]在Fast-RCNN方法[6]和CFM方法[14]的基础上，提出了一种端到端的框架，通过将生成的候选区域缩放成7×7大小，结合自由形状的ROI池化层(free-form Region-Of-Interest)来获得候选区域及其前景的特征，并通过区域到像素(region-to-pixel)层得到像素级的预测，得到了更加准确的分割结果。

将候选区域方法与CNN方法结合起来可得到精细的分割边界，并可以进行端到端训练。在Caesar等人[15]方法的基础上，结合文献[16]提出的深度残差网络，本文提出一种基于区域和深度残差网络的语义分割方法。首先使用候选区域生成算法在原图像上生成一定数量的候选区域，由于候选区域大小并不固定，为了适应不同尺度的候选区域以及充分保留候选区域的细节信息，本文将每个候选区域缩放成多种不同尺寸，结合自由形状的ROI池化层，得到多种尺度的特征，输入相同的网络中训练学习得到多个模型，并在最终阶段进行融合。然后使用带扩张卷积的深度残差网络进行特征提取，并使用全局平均池化层对每一个像素分类，得到最终的分割结果。

2　本文方法

本文提出一种结合区域和深度残差网络的语义分割模型，框架如图1所示。网络分为3个部分，第1部分生成候选区域，第2部分是全卷积网络，用于特征提取。本文使用残差网络ResNet-50前5层卷积层作为基础网络，在基础网络中的部分卷积层使用不同的扩张率，在最后一层卷积层得到高分辨率的特征图。第3部分为分割网络，输入第1部分生成的候选区域和第2部分得到的特征图，输出分割图像。具体流程如下：在训练阶段，首先在输入图像上生成候选区域集，同时将输入图像归一化后输入到带扩张卷积的深度残差网络中进行特征提取，得到相应的特征图。然后将候选区域以及特征图输入到自由形状的ROI池化层中，得到候选区域特征。通过全局平均池化层对候选区域分类，并使用区域到像素层将区域类别信息映射到区域内每个像素，最终得到像素级预测结果。考虑到不同尺寸的特征图包含的细节信息不一致，将候选区域缩放成不同的尺度，并在ROI池化层中得到相应尺度的特征图。所以通过每次选择不同尺度进行缩放，可以训练学习得到不同的模型。在测试阶段，将测试图像同时输入这些模型中，将在全局平均池化分类层得到的特征进行融合，旨在得到一个更鲁棒的结果。

2.1　生成候选区域

本文使用选择搜索(selective search)算法[10]生成候选区域，生成的候选区域由3个部分描述：边界框(bounding box)、前景掩膜(mask)、前景大小(size)。其中边界框是一个4维坐标，表示候选区域

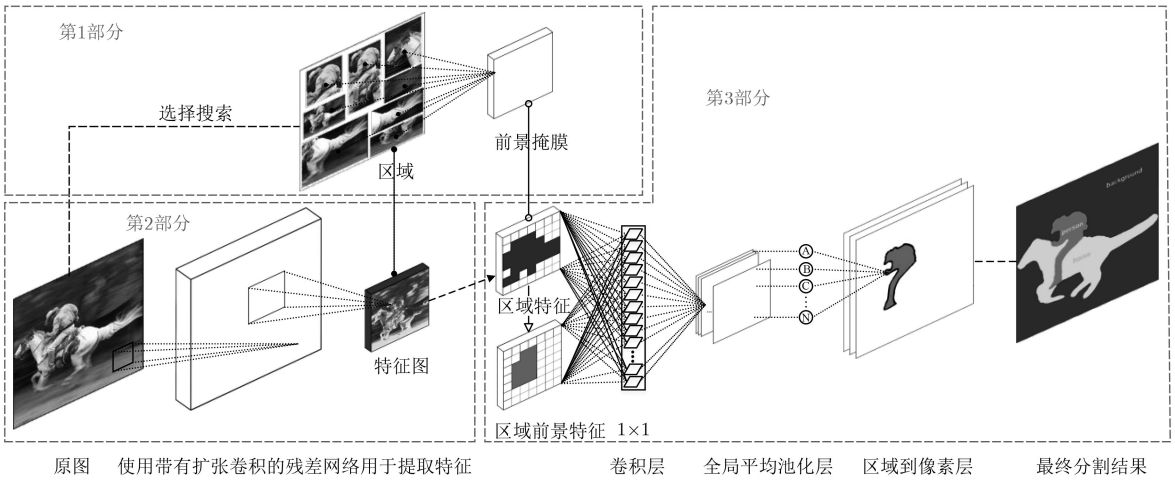


图 1 本文所提模型框架

在原图上的位置；前景掩膜是覆盖在候选区域上表示区域前景的2进制掩码。将区域特征在每个通道上与其对应的前景掩膜相乘可得到区域前景特征，如图1所示。为了充分保留图像中候选区域的空间细节信息，本文将每个候选区域缩放至4种不同尺度(7×7, 9×9, 13×13, 15×15)输入到网络中。虽然本文所提模型理论上可以接收任意尺寸的候选区域输入，但考虑到候选区域尺度太大会造成计算量剧增的问题，以及目标尺寸的分布情况，故只考虑使用上述4种合理的尺度进行实验。

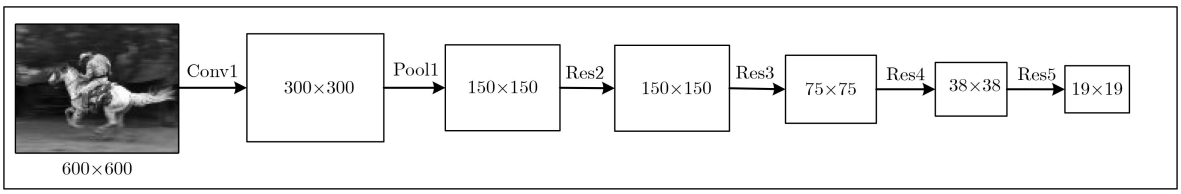
2.2 改进的带扩张卷积的残差网络

本文在He等人[16]提出的深度残差网络的基础上，结合文献[17]提出的扩张卷积结构，设计出适用于语义分割任务的带扩张卷积的深度残差网络，用于特征提取。使用50层残差网络(ResNet-50)的前5层卷积层作为本文特征提取网络的基础网络，网络的输入为经过归一化大小的图像(600×600)，输出为特征图(75×75)，如图2所示。

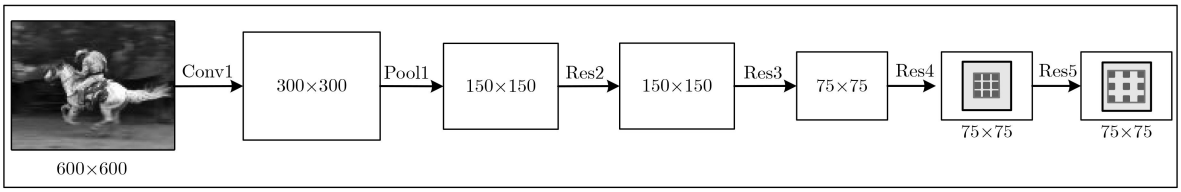
在基础网络的第4层Res4和第5层卷积层Res5中引入扩张卷积核。具体做法如下，首先将第4层和5层的卷积步长设置为1，并且设置第4层Res4的扩张率dilated=2，第5层Res5的扩张率dilated=4。原ResNet-50网络中Res5层分辨率分别相对于Res4层和Res3层输出下降了2倍和4倍，但是经过扩张卷积操作，特征图的尺度没有发生改变，最终输出的特征图大小为75×75，保留了更多的图像空间信息，结构如图2所示。

2.3 全局平均池化分类

传统的卷积神经网络中使用全连接层进行分类，如图3(a)所示。但全连接层参数冗余，容易出现过拟合现象。受文献[18]启发，本文提出适用于语义分割任务的全局平均池化层结构，来代替全连接层进行分类，结构如图3(b)所示，输入特征向量的大小为(H, W)，通道数为D, C为类别总数。实现过程为：首先使用C个$1\times1\times D$的卷积核对输入特征向量$H\times W\times D$进行卷积，得到$H\times W\times C$特征图，

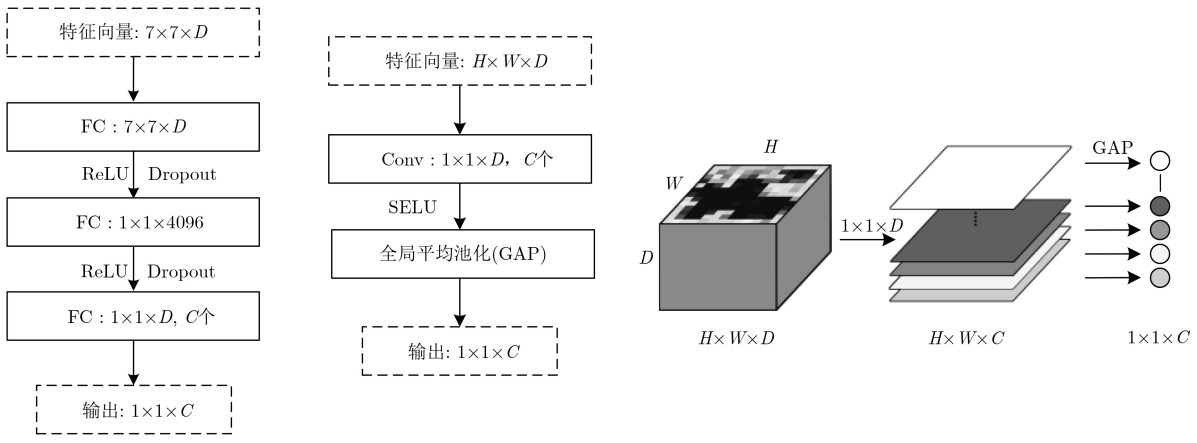


(a) ResNet-50网络中特征图大小变化情况(不含扩张卷积)



(b) 本文框架:(在ResNet-50原始网络中使用带有不同扩张率的扩张卷积核)

图 2 带扩张卷积的卷积神经网络结构



(a) 为VGG-16中全连接层结构　(b) 本文所提全局平均池化分类结构　(c) 全局平均池化分类过程示例

图 3 全局平均池化层结构示意图

这里使用$1\times1\times D$卷积核进行卷积的目的是实现跨通道的信息整合。然后使用与特征图大小相同的池化核对其进行平均池化，如图3(c)所示，得到最终的类别预测值。本文所提全局平均池化分类层可以接收任意尺度输入，对每张特征图使用全局平均池化得到一个输出，这个输出即表示类别预测值。相较于普通的全连接层，全局平均池化层更符合卷积结构，加强了特征映射与分类的对应关系，同时由于没有需要优化的参数，大大减少了参数量，从而可以加速训练过程和减轻过拟合风险。

2.4 基于区域和深度残差网络的语义分割

使用ROI池化层可以将原图经过卷积神经网络得到的特征映射到每个候选区域上，得到区域特征。为了更加突显区域前景，本文考虑第2种特征：区域前景特征，即把区域前景的特征从区域特征中提取出来，实现方法是在区域特征的每个通道上乘以其对应的前景掩膜，即可得到候选区域的前景特征。由于语义分割目标是给每一个像素分配一个语义标签，相当于多分类问题，故采用SOFTMAX回归进行分类。

对于候选区域r，经网络的全局平均池化分类层得到激活值F_r，参照文献[15]方法，使用区域到像素层从所有包含像素p的区域中选取激活值最大的作为像素p的激活值，即

$$F_p=\max_{p\in r}F_r \tag{1}$$

经过SOFTMAX层得到像素p属于第i类的概率$\theta_{p,i}$为

$$\theta_{p,i}=\frac{\exp(F_p^{(i)})}{\sum_{i=1}^{C}\exp(F_p^{(i)})} \tag{2}$$

从而p的语义类别l_p可由式(3)预测

$$l_p=\arg\max_{i\in C}\theta_{p,i} \tag{3}$$

本文模型的损失函数采用对数似然函数，如式(4)所示，其中y为像素p的真实标签，P为训练集中的像素总数，C为数据集的类别总数。

$$\text{Loss}=-\frac{1}{P}\left[\sum_{y=1}^{C}\sum_{p=1}^{P}1\{l_p=y\}\lg\left(\frac{\exp(F_p^{(l_p)})}{\sum_{i=1}^{C}\exp(F_p^{(l_p)})}\right)\right]+\frac{\lambda}{2}\|W\|^2 \tag{4}$$

其中，$1\{l_p=y\}$为示性函数，当$l_p=y$时$1\{l_p=y\}$为1，其他情况为0。$\frac{\lambda}{2}\|W\|^2$项是权重衰减项，其中λ为衰减因子，表示网络中可更新参数层的网络参数。

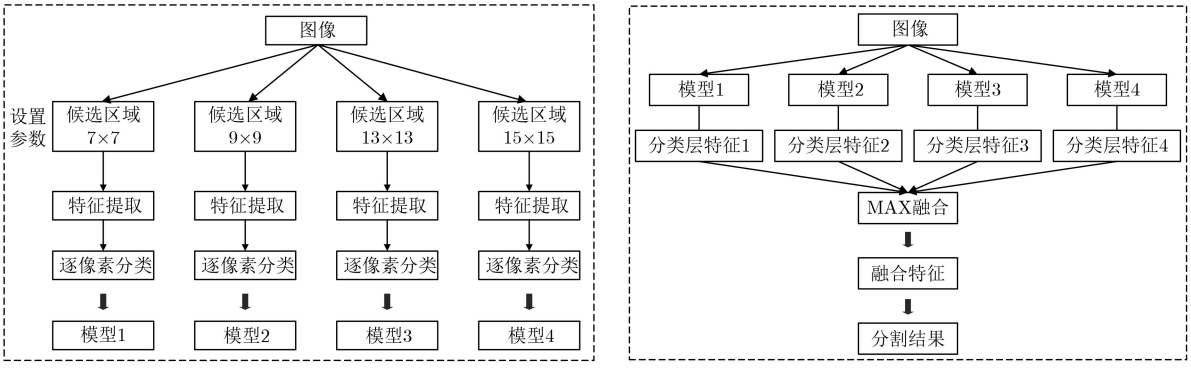
2.5 模型融合

为了得到平均性能更好的语义分割结果，将候选区域缩放成4种尺度：7×7，9×9，13×13，15×15，分别训练学习得到4个不同的模型。在测试阶段，按照每个模型不同的候选区域尺寸参数设置，将测试图像分别输入到这4个模型中，然后将图像在全局平均池化分类层得到的激活值进行融合，融合方式为取对应激活值的最大值，框架图如图4所示。

3 实验结果与分析

3.1 实验数据集及评价指标

为了验证本文所提方法的有效性，在语义分割研究领域常用的2个数据集SIFT FLOW[19]和PASCAL Context[20]上进行实验分析。SIFT FLOW包含2688张图像，每张图像带有33种语义类别(如“桥”“山”“太阳”等)的像素标签和3种几何位置分类(“水平”“垂直”“天空”)，本文使用该数据集提供的划分标准文件将其分为训练集2488张图像，测试集200张图像，该数据集类别极



(a) 训练阶段　　(b) 测试阶段

图 4 模型融合框架示意图

度不平衡，给分割算法带来了极大的挑战。PASCAL Context为当前流行的PASCAL VOC 2010数据集中的物体和物体类提供完整的像素级注释，共10103张图像，包括4998张训练图像和5105张验证图像，加上背景共有60个类别。由于该数据集没有专用的测试图像，所以本文参照文献[15]方法使用验证图像进行测试。

图像语义分割方法中基于像素的评价指标有像素准确率(Pixel Accuracy, PA)、平均准确率(Mean Accuracy, MA)和平均交并比(Mean Intersection-over-Union, MIoU)。PA是指图像中分类正确的像素数和图像中总像素数的比例；MA是指所有类别的像素准确率的平均值；而MIoU是对于每个类，计算预测目标类和真实标注类的交并比，然后取类平均。

3.2 实验设置

实验的硬件环境是：CPU为Intel(R) Xeon(R) CPU E5-1630 v4 @ 3.70 GHz，内存32 GB，显卡是4 GB的NVIDIA Quadro M2000，系统类型为64位WINDOWS 7 操作系统，仿真软件为MATLAB R2016a。整个深度模型基于MatConvNet框架及其工具包实现，版本为1.0-beta24。

网络采用带动量的随机梯度下降法(Stochastic Gradient Descent, SGD)进行训练。总共迭代次数为30代，在前20代，使用了0.001的学习速率，后10代学习率降为0.0001，每一次学习的样本数(batch size)设置为32，动量设置为0.9，权重衰减系数λ设置为0.0005。特征提取网络的权重参数来源于ResNet-50[16]分类网络在ImageNet数据集上预训练得到的参数。

3.3 实验对比

3.3.1 在SIFT FLOW数据集上的对比

在SIFT FLOW数据集上，比较了本文算法与Yang[21], Long[7], Eigen[22], Caesar[15]4种近年来语义分割的主流方法，定量实验结果如表1所示。

参与比较的方法中Yang等人[21]使用了基于图像检索和超像素匹配的场景解析算法，该方法更注重数据集中的稀有类别，并且通过反馈机制结合图像中的上下文信息，该方法得到的PA和MA为48.7%和79.8%，分别比本文低了17.5%和5.9%。Long等人[7]提出的全卷积方法，通过将分类网络中的全连接层转化为卷积层，得到语义分割网络。虽然该方法使用了跳层连接组合中间层的特征向量，但最终输出层的分辨率相较于原图还是下降了8倍，使用上采样还原到原图分辨率时得到的结果比较粗糙。该方法在SIFT FLOW数据集中得到的MA和PA分别为51.7%和85.2%，分别比本文低了14.5%和0.5%。文献[22]在Long等人的基础上，使用3种不同的尺度训练出3个CNN模型，并对输出结果进行组合，该方法在SIFT Flow数据集中的MA和PA分别为55.7%和86.8%，在MA上比本文方法低10.5%，PA略高于本文方法。文献[15]结合了区域和卷积神经网络的方法，但该方法局限于全连接层的固定输入，使用7×7大小的候选区域丢失了大量的目标细节信息。该方法取得的MA和PA准确度分别比本文低了2.2%和1.4%。

表 1 本文算法与其他先进方法在SIFT FLOW数据集上的实验对比(%)

方法	平均准确率(MA)	像素准确率(PA)
Yang[21]	48.70	79.80
Long[7]	51.70	85.20
Eigen[22]	55.70	**86.80**
Caesar[15]	64.00	84.30
本文算法	**66.20**	**85.70**

图5示例了本文方法在SIFT FLOW测试集的一些测试图片上得到的分割效果图。从图5可以看出，本文方法在物体边界处分割精确，分割边缘几乎与真实标注结果一致，如图5(a)中的建筑物，图5(d)，图5(f)和图5(h)中的树木。由于生成候选区域算法使用了多尺度提取相互重叠的区域，并且后续使用了多尺度融合，使得各个尺度的目标都有概率被识别出来，故本文方法在一些小物体上也有不错的识别效果，如图5(g)中的路灯与标志物，图5(i)与图5(j)中的电线杆。甚至在真实标注中未精准标注的类别都被识别出来，如图5(b)，图5(c)和图5(e)中的草坪，真实标注为田野，而本文算法识别为草地，但这反而会影响本文算法在某些类别上的准确度。

3.3.2 在PASCAL Context数据集上的对比

将本文方法和其他先进方法在PASCAL Context数据集上进行定量实验比较，结果如表2所示。

文献[3]使用了2阶均值池化的方法，MIoU为18.1%。Dai等人[23]提出的Boxsup的方法，使用边界框代替分割掩模(segment mask)进行像素级的分割，MIoU为34.4%。文献[15]在该数据集中得到的MA, PA和MIoU分别比本文低2.3%, 3.9%和2.2%。而文献[7]在MIoU上略高于本文0.4%，但在MA和PA上分别比本文低了5.7%和0.4%。

图6示例了一些本文方法在PASCAL Context验证集上得到的分割结果。本文算法得到的分割结果接近于真实标注，如图6(e)中的狗，有些结果甚

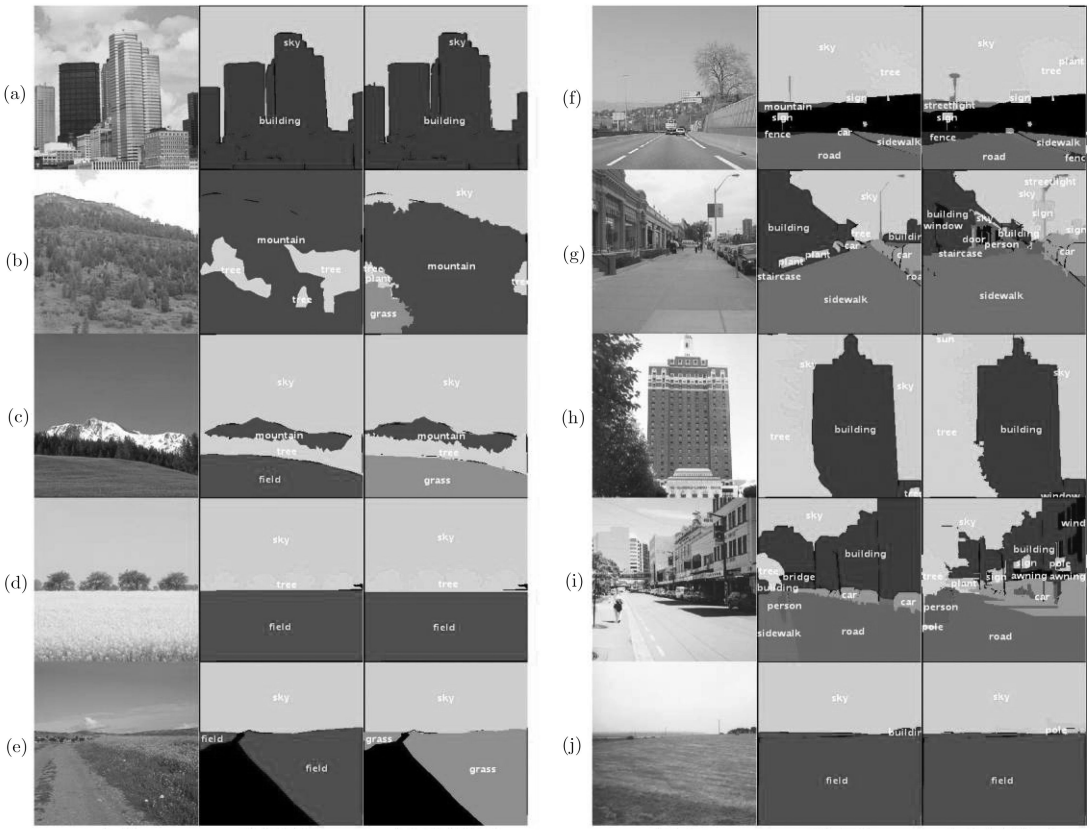


图 5 SIFT FLOW数据集上图像分割效果

表 2　本文算法与其他先进方法在PASCAL Context数据集上的实验对比(%)

方法	平均准确率(MA)	像素准确率(PA)	平均交并比(MIoU)
O2P[3]	–	–	18.10
Dai[23]	–	–	34.40
Long[7]	46.50	65.90	**35.10**
Caesar[15]	49.90	62.40	32.50
本文	**52.20**	**66.30**	**34.70**

至要优于真实标注，如图6(f)中的猫胡须。在一些小物体识别上也取得了很好的识别效果，如图6(d)中的食物以及杯子，图6(c)中的雪，图6(b)中的标志物。在图6(a)中，本文提出的方法把被围栏分割的天空完整识别出来了，而真实标注忽略了这个目标，并且在该图中本文方法也将椅子的轮廓较完整地分割了出来。

3.3.3　不同层上使用扩张卷积核的对比

为了分析在不同层中使用扩张卷积对实验结果的影响，本实验比较了单独在第4层卷积层Res4、单独在第5层卷积层Res5，以及同时在第4层和第5层使用扩张卷积对实验结果的影响。表3所示是这3种情况下本文方法在SIFT FLOW数据集中取得的平均准确度MA。表3中的“无操作”是指按照原残差网络ResNet-50[16]设置Res4和Res5层中stride=2, dilated=1；“仅移除stride操作”是指设置stride=1, dilated=1。

表 3　3种不同扩张卷积核使用方案的性能比较

实验	操作		最后卷积层输出大小	SIFT FLOW MA(%)
1	无操作		19×19	64.50
2	仅移除stride操作	仅Res4 (stride=1)	38×38	26.61
3		仅Res5 (stride=1)	38×38	37.47
4		Res4 (stride=1)+Res5 (stride=1)	75×75	39.76
5	+设置dilated	仅Res4(dilated=2)	38×38	64.20
6		仅Res5(dilated=4)	38×38	63.60
7		Res4(dilated=2)+Res5(dilated=4)	75×75	**65.50**

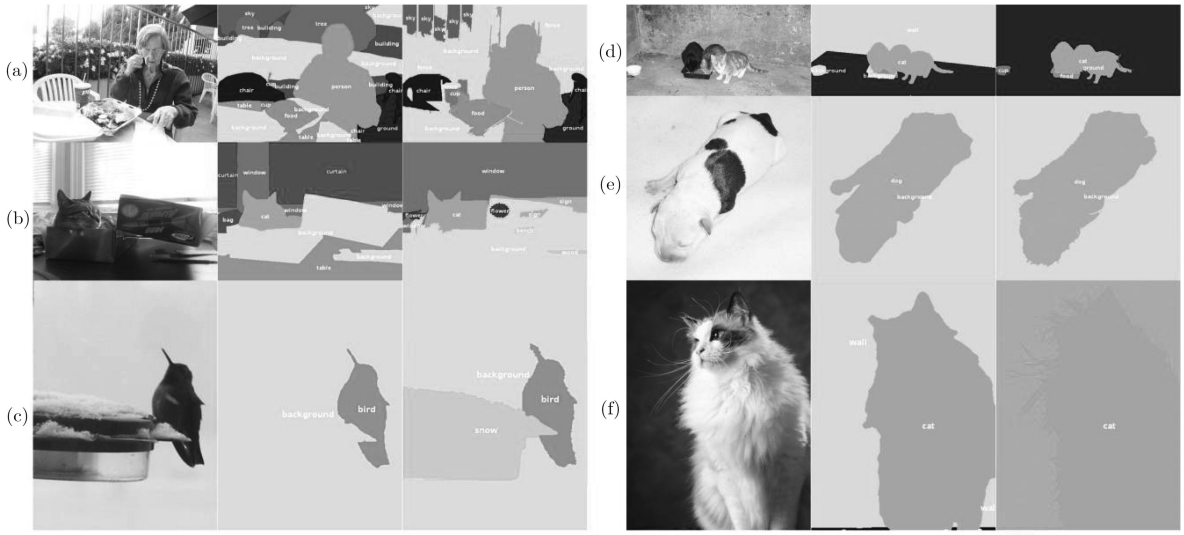


图 6 PASCAL Context数据集上图像分割效果

从表3的实验结果可以看出，当分别在Res4层和Res5层仅移除stride操作时，实验结果相对于无操作得到的精度分别下降了37.89%和27.03%，这表明通过设置卷积层的stride为1，虽然使得特征提取网络最后一层卷积层输出的特征图分辨率变大了1倍，但是容易导致对图像中部分区域的特征重复提取，影响了最后的分割效果。在Res4和Res5卷积层同时使用扩张卷积的效果要优于不使用或者单独在Res4或Res5层使用扩张卷积，原因可能在于虽然使用扩张卷积可以保持分辨率不变，但由于扩张卷积是在特征图上跳跃连接，单独在某层(如Res4层或Res5层)使用扩张卷积时并没有完全提取上层的所有特征信息，导致传入下层的特征图缺失空间信息，故效果略差于同时在Res4和Res5上使用扩张卷积。同时在Res4和Res5层中使用扩张卷积，能够让这两层卷积层中提取到的特征相互补充，得到更好的效果。对比实验4和实验7，虽然通过同时移除Res4层和Res5层的stride操作也可以使得最终卷积层输出大小为75×75，但实验4在SIFT FLOW数据集中得到的精度却远小于实验7，说明扩张卷积不仅保持了特征映射的大小不变，还保留了更多的图像空间信息，从而获得了更好的分割精度。

3.3.4 多模型融合

在生成候选区域时，保持其他参数不变，使用4种不同尺寸(7×7, 9×9, 13×13, 15×15)的候选区域输入到网络中，分别训练模型，并将得到的4个模型及本文提出的融合方法在SIFT FLOW数据集中进行测试，表4中分析比较了单独训练1个模型和融合模型的测试结果。

由表4可知，当候选区域尺寸在7×7～13×13之间时，模型的平均准确率随着区域尺寸的增大而增加。但当候选区域尺寸为15×15时，相较于13×13准确率有了略微的下降。模型的融合效果优于单个模型的效果，并且融合4个模型的效果最好，得到了更稳定的平均性能。

表 4 4个单模型以及融合模型在SIFT FLOW上的效果比较

模型序号	候选区域尺寸	SIFT FLOW MA(%)
1	7×7	64.20
2	9×9	64.80
3	13×13	65.30
4	15×15	65.20
融合模型3 4	–	65.70
融合模型3 4	–	66.00
融合模型1 2 3 4	–	**66.20**

4 结束语

本文提出一种基于区域和深度残差网络的语义分割方法，结合了基于区域方法中可以得到清晰物体边界的优点和基于全卷积网络的可进行端到端训练的优点。通过使用带扩张卷积的深度残差网络来提取特征，得到了包含更多信息的高分辨率的特征图。使用全局平均池化分类方法，从而输入到该层的特征图可以是任意尺度。在SIFT FLOW和PASCAL Context数据集上的测试结果表明本文提出的算法在语义分割任务中具有很好的性能。实验部分验证了本文方法的有效性，并且分析了候选区域大小对算法性能的影响，发现区域不宜过大，在SIFT FLOW数据集中较合适的尺度为13×13。在同一个网络结构上通过不同尺寸的输入，训练得到

多个模型并进行融合，发现融合模型的结果要优于任意单一模型，并且融合的模型越多，得到的准确率也越高。下一步将探索结合其他区域生成网络(如RPN网络)的效果。

参 考 文 献

[1] 魏云超, 赵耀. 基于DCNN的图像语义分割综述[J]. 北京交通大学学报, 2016, 40(4): 82–91. doi: 10.11860/j.issn.1673-0291.
WEI Yunchao and ZHAO Yao. A review on image semantic segmentation based on DCNN[J]. *Journal of Beijing Jiaotong University*, 2016, 40(4): 82–91. doi: 10.11860/j.issn.1673-0291.

[2] CARREIRA J, LI Fuxin, and SMINCHISESCU C. Object recognition by sequential figure-ground ranking[J]. *International Journal of Computer Vision*, 2012, 98(3): 243–262. doi: 10.1007/s11263-011-0507-2.

[3] CARREIRA J, CASEIRO R, BATISTA J, *et al*. Semantic segmentation with second-order pooling[C]. Proceedings of the 12th European Conference on Computer Vision 2012, Berlin, Germany, 2012: 430–443. doi: 10.1007/978-3-642-33786-4_32.

[4] ARBELÁEZ P, HARIHARAN B, GU Chunhui, *et al*. Semantic segmentation using regions and parts[C]. Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012: 3378–3385. doi: 10.1109/CVPR.2012.6248077.

[5] GIRSHICK R, DONAHUE J, DARRELL T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, 2014: 580–587. doi: 10.1109/CVPR.2014.81.

[6] GIRSHICK R. Fast R-CNN[C]. Proceedings of 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 1440–1448. doi: 10.1109/iccv.2015.169.

[7] SHELHAMER E, LONG J, and DARRELL T. Fully convolutional networks for semantic segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 39(4): 640–651. doi: 10.1109/TPAMI.2016.2572683.

[8] CHEN L C, PAPANDREOU G, KOKKINOS I, *et al*. Semantic image segmentation with deep convolutional nets and fully connected CRFs[J]. *Computer Science*, 2015(4): 357–361. doi: 10.1080/17476938708814211.

[9] 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*, 2016, 40(4): 834–848. doi: 10.1109/TPAMI.2017.2699184.

[10] UIJLINGS J R R, VAN DE SANDE K E A, GEVERS T, *et al*. Selective search for object recognition[J]. *International Journal of Computer Vision*, 2013, 104(2): 154–171. doi: 10.1007/s11263-013-0620-5.

[11] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(9): 1904–1916. doi: 10.1109/TPAMI.2015.2389824.

[12] YU Tianshu, YAN Junchi, ZHAO Jieyi, *et al*. Joint cuts and matching of partitions in one graph[C]. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 705–713. doi: 10.1109/cvpr.2018.00080.

[13] HARIHARAN B, ARBELÁEZ P, GIRSHICK R, *et al*. Simultaneous detection and segmentation[C]. Proceedings of the 13th Conference on Computer Vision, Zurich, Switzerland, 2014: 297–312. doi: 10.1007/978-3-319-10584-0_20.

[14] DAI Jifeng, HE Kaiming, and SUN Jian. Convolutional feature masking for joint object and stuff segmentation[C]. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 3992–4000. doi: 10.1109/CVPR.2015.7299025.

[15] CAESAR H, UIJLINGS J, and FERRARI V. Region-based semantic segmentation with end-to-end training[C]. Proceedings of the 14th European Conference on Computer Vision 2016, Amsterdam, The Netherlands, 2016: 381–397. doi: 10.1007/978-3-319-46448-0_23.

[16] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778. doi: 10.1109/CVPR.2016.90.

[17] YU F and KOLTUN V. Multi-scale context aggregation by dilated convolutions[EB/OL]. https://arxiv.org/abs/1511.07122, 2015.

[18] LIN Min, CHEN Qiang, and YAN Shuicheng. Network in network[EB/OL]. https://arxiv.org/abs/1312.4400, 2014.

[19] LIU Ce, YUEN J, and TORRALBA A. Nonparametric scene parsing via label transfer[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, 33(12): 2368–2382. doi: 10.1109/TPAMI.2011.131.

[20] MOTTAGHI R, CHEN Xianjie, LIU Xiaobai, *et al*. The

role of context for object detection and semantic segmentation in the wild[C]. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 891–898. doi: 10.1109/CVPR.2014.119.

[21] YANG Jimei, PRICE B, COHEN S, *et al*. Context driven scene parsing with attention to rare classes[C]. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 3294–3301. doi: 10.1109/CVPR.2014.415.

[22] EIGEN D and FERGUS R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture[C]. Proceedings of 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 2650–2658. doi: 10.1109/iccv.2015.304.

[23] DAI Jifeng, HE Kaiming, and SUN Jian. Boxsup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation[C]. Proceedings of 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 1635–1643. doi: 10.1109/iccv.2015.191.

罗会兰：女，1974年生，博士，教授，研究方向为机器学习和模式识别等.

卢 飞：男，1994年生，硕士，研究方向为图像语义分割.

孔繁胜：男，1946年生，博士生导师，教授，研究方向人工智能与知识发现等.