DOI:10.19951/j.cnki.1672-9331.20221107002 文章编号:1672-9331(2024)05-0104-09
引用格式:梁小林,王欣怡,黄雅娟,等.融合注意力模块的双结构金字塔场景解析网络[J].长沙理工大学学报(自然科学版),2024,21(5):104-112.
Citation: LIANG Xiaolin, WANG Xinyi, HUANG Yajuan, et al. Two-structure pyramid scene parsing network with attention module[J]. J Changsha Univ Sci Tech (Nat Sci), 2024, 21(5): 104-112.

融合注意力模块的双结构金字塔场景解析网络

梁小林,王欣怡,黄雅娟,肖进文
(长沙理工大学 数学与统计学院,湖南 长沙 410114)

摘 要:【目的】改善原始图像信息丢失及图像分辨率下降的问题,提高图像语义分割的精度。**【方法】**提出融合注意力模块的双结构金字塔场景解析网络模型,并利用该模型对图像进行语义分割。首先,使用MobileNet V2模块提取原始图像的主干特征;其次,将特征图送入金字塔池化模块1中,获取上下文信息;然后,使用注意力模块关注重要特征,并对浅层信息进行综合,得到中间特征图;接着,将中间特征图送入金字塔池化模块2,融合局部和全局信息;最后,利用丰富的浅层和深层信息对原始图像进行分割。**【结果】**在PASCAL VOC 2007数据集上进行的试验表明,平均像素精度和平均交并比分别达到85.64%和78.12%,比金字塔场景解析网络的分别提高了4.95个百分点和12.31个百分点。**【结论】**本文模型有效解决了图像分割中信息丢失和分辨率下降问题。

关键词:图像语义分割;注意力模块;空洞卷积;深度可分离卷积;金字塔池化
中图分类号:TP18 **文献标志码:**A

Two-structure pyramid scene parsing network with attention module

LIANG Xiaolin, WANG Xinyi, HUANG Yajuan, XIAO Jinwen
(School of Mathematics and Statistics, Changsha University of Science & Technology, Changsha 410114, China)

Abstract: [**Purposes**] This study aims to improve the loss of original image information, degradation of image resolution, and accuracy of semantic segmentation of images. [**Methods**] A two-structure pyramid scene parsing network model with an attention module was proposed, and it was used for the semantic segmentation of images. Firstly, the MobileNet V2 module was used to extract the backbone feature of the original image. Secondly, the feature map was input into the pyramid pooling module 1 to obtain the context information. Then, the attention module was used to pay attention to the important features and synthesize the shallow information to obtain the intermediate feature map. The intermediate feature map was fed into the pyramid pooling module 2 to fuse local and global information. Finally, the original image was segmented by using rich shallow and deep information. [**Findings**] In PASCAL VOC 2007 dataset, the ratios of the mean pixel accuracy (MPA) and mean intersection over union (MIOU) reach 85.64% and 78.12%, respectively, which are 4.95 and 12.31 percentage points higher than the ratios of MPA and MIOU

收稿日期:2022-11-07;**修回日期:** 2022-12-10;**接受日期:** 2022-12-11
基金项目:湖南省自然科学基金资助项目(2021JJ30734);长沙理工大学研究生"实践创新与创业能力提升计划"项目(CLSJCX22124)
通信作者:梁小林(1965—)(ORCID:0000-0002-1338-2947),男,副教授,主要从事大数据分析方面的研究。
E-mail:liang@csust.edu.cn

of the pyramid scene parsing network. [**Conclusions**] The proposed model can effectively resolve the problems of information loss and resolution degradation in image segmentation.

Key words: semantic segmentation of image; attention module; atrous convolution; depthwise separable convolution; pyramid pooling

Foundation items: Project (2021JJ30734) supported by Natural Science Foundation Project of Hunan Province; Project (CLSJCX22124) supported by "the Practice Innovation and Entrepreneurial Ability Improvement Plan" for Graduate Students of Changsha University of Science and Technology

Corresponding author: LIANG Xiaolin (1965—) (ORCID: 0000-0002-1338-2947), male, associate professor, research interest: big data analysis. E-mail: liang@csust.edu.cn

0　引言

"语义分割"概念的由来可以追溯到20世纪70年代，强调分割区域在语义上有意义[1]。语义分割是利用图像的高级语义信息，预测图像中的不同类别并定位其位置的，从而将不同类别的对象分割出来，即对图像进行像素级分类[2]。作为典型的计算机视觉问题，语义分割广泛应用于虚拟现实、工业自动化、视频检测、医疗图像、农田场景、室内场景、桥梁及路面等领域中的裂纹识别、卫星图像分析和即时视觉搜索等不同方向，具有重要的研究意义和应用价值[3-6]。

早期的传统图像分割方法只能分割出图像的关键信息，只对特征明显的局部区域敏感，对图像的细节、边缘、角度等特征往往很难分割甚至不能分割，但用于图像处理的预处理是简单有效的。传统图像分割方法主要有：基于阈值的图像分割方法[7-9]、基于边缘的图像分割方法[10]、基于区域的图像分割方法[11]、基于聚类的图像分割方法[12]、基于图论的图像分割方法[13]和基于特定理论的图像分割方法[14]。这些方法对类别明显、分布整齐、类内相似度高、类间差异性较大、前景和背景差异较明显的图像，分割结果较好。但是，这些传统方法适用范围窄，对类别复杂且繁多的图像分割效果不理想，没有关注到图像的高级语义特征和空间等信息，容易受噪声干扰[15]。

随着图像分割的应用领域日益复杂化，对分割结果的要求越来越高，一系列基于卷积神经网络的图像分类[16]和分割方法不断出现。与传统的图像分割方法相比，基于卷积神经网络的语义分割方法不是采用手工设计的特征，而是从输入图像中自动学习特征，利用深度神经网络就可以实现端到端的语义分割预测[17]。深度学习的阶段性成功对语义分割算法产生了前所未有的影响，极大地提高了准确性[18]，解决了很多传统方法不能解决的问题，拓宽了语义分割的实际应用范围。

SHELHAMER 等[19]提出全卷积网络(fully convolutional networks, FCN)并将深度学习用于语义分割。FCN是首个端到端的分割模型，奠定了图像语义分割模型框架的基础。但是，因连续的下采样使图像空间位置信息丢失较多，FCN的分割过程对空间信息利用率低，导致分割结果比较粗糙，以及对不同像素之间的关系、边界和细节信息敏感性偏低。针对这两个问题，基于编码器-解码器(encoder-decoder)结构的不同对称模型应运而生，如用于生物医学图像分割的卷积网络(convolutional networks for biomedical image segmentation, U-Net)[20]和用于图像分割的深度卷积编码器-解码器架构[21]。这种编码器-解码器的架构利用下采样块构成的编码器获取像素位置信息和图像特征，再利用反卷积等操作组成的解码器对特征进行分析，确定图像的类别并精确定位像素的位置，从而对图像上下文和空间信息利用更加充分，减少了信息损失，有效缓解了物体边缘分割粗糙的问题[20-21]。但下采样过程中的信息丢失和物体边界定位问题还须进一步解决。因此，CHEN 等[22-24]和 NING 等[25]提出了 DeepLab 系列，其中 DeepLab V1 将深度卷积神经网络(deep convolutional neural networks, DCNN)和条件随机场(conditional random field, CRF)相结合，定位图像边界并细化分割边缘，并利用空洞卷积(atrous convolution)来扩大感受野，解决 DCNN 在连续池化和下采样过程中丢失大量细节信息的问题；

DeepLab V2提出了空洞空间金字塔池化(atrous spatial pyramid pooling,ASPP)模块,设置了不同的空洞率,使用了多个并行的空洞卷积来获取物体多尺度信息,对空间变换有较高的不变性[23];DeepLab V3继续改进ASPP模块,在每个卷积后加入批归一化(batch normalization,BN)处理,在ASPP模块最后加入图像级特征(image-level feature)用于捕获全局特征,同时去除之前的CRF模块[25];DeepLab V3+在DeepLab V3的基础上增加了一个简单的解码模块以提升分割精度,并将深度可分离卷积(depthwise separable convolution)应用于ASPP和解码模块,大大减少了训练参数[24]。DeepLab系列在融合多尺度上下文和空间特征方面取得了很大进步,但图像信息利用率还需要提高。为了能够充分利用图像信息,受空间金字塔池化(spatial pyramid pooling,SPP)网络的启发,ZHAO等[26]提出了金字塔场景解析网络(pyramid scene parsing network,PSPNet),如图1所示。PSPNet将输入图像的局部信息和全局信息融合,在一定程度上缓解了训练过程中信息丢失的问题,但在边缘分割上还需要优化。

针对PSPNet对小物体和边界不明显的物体分割不理想的问题,本文提出融合了卷积块注意力模块(convolutional block attention module,CBAM)(下文简称"注意力模块")的双结构金字塔场景解析网络(two-structure pyramid scene parsing network,TSPSPNet)模型,旨在改善语义分割网络中由于连续下采样导致的原始图像信息丢失和分辨率下降问题,并在PASCAL VOC(PASCAL即pattern analysis, statistical modelling and computational learning,是一种结构化编程语言;VOC即visual object classes,是一个世界级的计算机视觉挑战赛)2007数据集上进行试验。

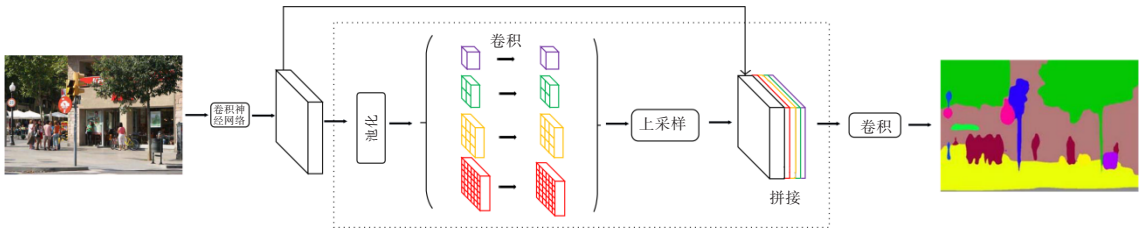

图1 PSPNet结构[26]

Fig. 1 Structure of PSPNet[26]

1 研究方法

TSPSPNet语义分割模型将原来的金字塔池化模块(pyramid pooling module,PPM)记为PPM1,并在PSPNet基础上再增加一个金字塔池化模块2(PPM2),以更好地整合图像的上下文信息和空间信息。同时,在两个PPM中间插入CBAM,在分割时突出重要特征、抑制非必要特征,由此形成两个PPM与CBAM串联的结构。

1.1 MobileNet V2与CBAM

1.1.1 深度可分离卷积

深度可分离卷积操作将普通的卷积过程分为两个步骤:逐深度卷积和逐点卷积,如图2所示。其中,逐深度卷积使用三个单通道的3×3卷积核分别作用于输入图像的三个通道,得到三个中间特征图,再使用4个三通道的1×1卷积核作用于中间特征图,得到最终的4个输出特征图。

卷积和池化等操作堆叠可得到传统卷积神经网络。该网络特征提取能力依赖于不断增加的网络深度。由于网络参数量不断增加,导致训练耗时增加和计算复杂程度提高。本文以SANDLER等[27]提出的MobileNet V2为主干特征来提取图像特征。该网络将传统卷积替换为深度可分离卷积,有效减少了训练的参数量,加速了网络的运算速度。假设输入图像大小为$C_1 \times H_1 \times W_1$,输出图像大小为$C_2 \times H_2 \times W_2$(其中的$C$、$H$、$W$分别为图像的通道数、宽和高),卷积核的大小为$k \times k$(正整数$k$=1,2,3,…,$n$),则普通卷积的参数量$P$为

$$P = k^2 C_1 C_2 \tag{1}$$

深度可分离卷积的参数量P_d为

$$P_d = k^2 C_1 + C_1 C_2 \tag{2}$$

两者之比为

$$\frac{P}{P_d}=\frac{k^2C_1C_2}{k^2C_1+C_1C_2}=\frac{k^2C_2}{k^2+C_2} \tag{3}$$

由计算可知，普通卷积的参数量是深度可分离卷积参数量的倍数。在深度学习的网络中，采用后者可以加快网络的训练速度。

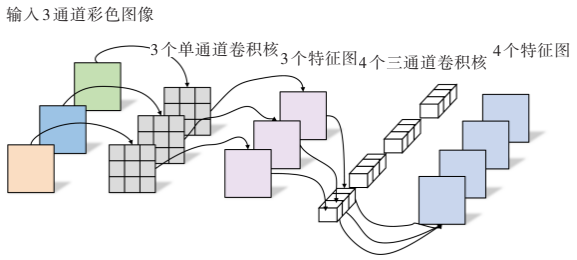

图 2　深度可分离卷积

Fig. 2　Depthwise separable convolution

1.1.2　MobileNet V2

MobileNet V2[27]以深度可分离卷积为基础，其结构如图 3 所示。本文使用 MobileNet V2 网络构建主干特征提取器来提取图像特征。

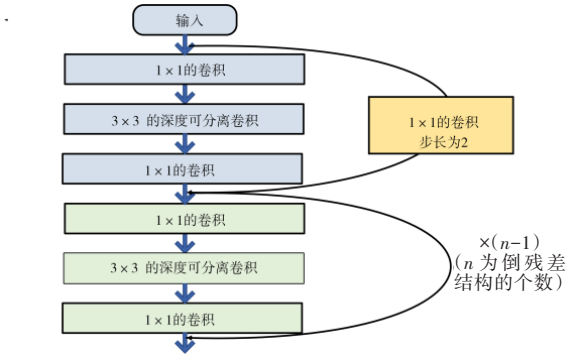

图 3　MobileNet V2 的倒残差结构

Fig. 3　Inverted residuals of MobileNet V2

1.1.3　注意力模块（CBAM）

CBAM 由通道注意力模块（channel attention module，CAM）和空间注意力模块（spatial attention module，SAM）组成，如图 4 所示。

给定一个输入特征图 $\boldsymbol{F}\in\boldsymbol{R}^{C\times H\times W}$，经 CBAM 生成一维通道注意力特征图 $\boldsymbol{M}_c\in\boldsymbol{R}^{C\times 1\times 1}$ 和二维空间注意力特征图 $\boldsymbol{M}_s\in\boldsymbol{R}^{1\times H\times W}$。其中，$\boldsymbol{R}$ 为实数集。整个过程为

1）通道注意力模块（CAM）如图 5 所示。

CAM 模块关注的是“输入图像是什么”，输入图像的每个特征通道都可以视为一个类别检测器。首先，CAM 对输入特征图 $\boldsymbol{F}$ 逐通道进行最大池化和平均池化，压缩特征图的空间信息，得到特征图 $\boldsymbol{F}_{\max,c}\in\boldsymbol{R}^{C\times 1\times 1}$ 和 $\boldsymbol{F}_{\text{avg},c}\in\boldsymbol{R}^{C\times 1\times 1}$；然后，将池化后的结果传入只有一个隐藏层的多层感知器（multilayer perceptron，MLP）中进行相关的训练，多层感知器的参数是共享的；接着，使用逐元素求和法合并输出的特征向量，再经 sigmoid 激活函数生成通道注意力特征图 $\boldsymbol{M}_c$[28]；最后，将这个输出结果乘原图，变回原图像大小，得到输出特征图 $\boldsymbol{F}_1$。

整个过程可概括为

$$\boldsymbol{F}_1=\boldsymbol{M}_c(\boldsymbol{F})\otimes\boldsymbol{F} \tag{4}$$

$$\boldsymbol{M}_c(\boldsymbol{F})=\sigma\{\mathrm{MLP}[\mathrm{AvgP}(\boldsymbol{F})]+\mathrm{MLP}[\mathrm{MaxP}(\boldsymbol{F})]\}=\sigma\{W_1[W_0(\boldsymbol{F}_{\text{avg},c})]+W_1[W_0(\boldsymbol{F}_{\max,c})]\} \tag{5}$$

式中：σ 为 sigmoid 激活函数；$\boldsymbol{W}_0$ 和 $\boldsymbol{W}_1$ 为 MLP 的共享权重，$\boldsymbol{W}_0\in\boldsymbol{R}^{C/r\times C}$，$\boldsymbol{W}_1\in\boldsymbol{R}^{C\times C/r}$，$r$ 为通道缩减率；AvgP、MaxP 分别为平均池化、最大池化操作。

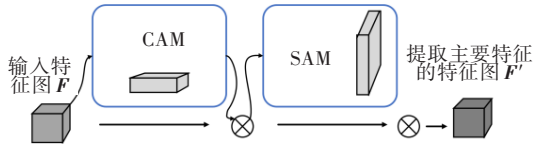

图 4　注意力模块（CBAM）结构[28]

Fig. 4　Structure of convolutional block attention module (CBAM)[28]

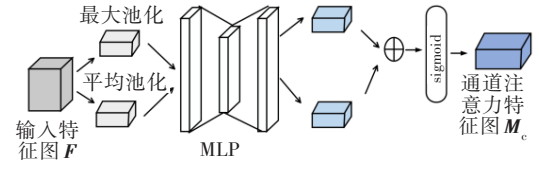

图 5　通道注意力模块结构[28]

Fig. 5　Structure of channel attention module (CAM)[28]

2）空间注意力模块（SAM）如图 6 所示。

SAM 模块关注的是“输入图像的有用信息在哪里”，其中 $\boldsymbol{F}_1$ 是 SAM 模块的输入特征图。首先，对 $\boldsymbol{F}_1$ 沿通道进行逐位置最大池化和平均池化，压缩特征图的通道信息，沿着特征图的通道维度进行池化可以有效地突出目标信息区域，得到池化后的特征图 $\boldsymbol{F}_{\max,c}\in\boldsymbol{R}^{1\times H\times W}$ 和 $\boldsymbol{F}_{\text{avg},c}\in\boldsymbol{R}^{1\times H\times W}$；然后，将池化后的结果连接起来；接着，经 sigmoid 激活函数得到空间注意力特征图 $\boldsymbol{M}_s$[28]；最后，再将这个输出结果乘原图，变回原图像大小，得到输出特征图 $\boldsymbol{F}_2$。

整个过程可以概括为

$$\boldsymbol{F}_2=\boldsymbol{M}_s(\boldsymbol{F}_1)\otimes\boldsymbol{F}_1 \tag{6}$$

$$\boldsymbol{M}_s(\boldsymbol{F}_1)=\sigma\{\boldsymbol{f}^{7\times7}[\mathrm{AvgP}(\boldsymbol{F}_1);\ \mathrm{MaxP}(\boldsymbol{F}_1)]\}$$
$$=\sigma\{\boldsymbol{f}^{7\times7}[\boldsymbol{F}_{\mathrm{avg},s};\ \boldsymbol{F}_{\mathrm{max},s}]\}\quad(7)$$

式中:σ为sigmoid激活函数;$\boldsymbol{f}^{7\times7}$为7×7的卷积核。

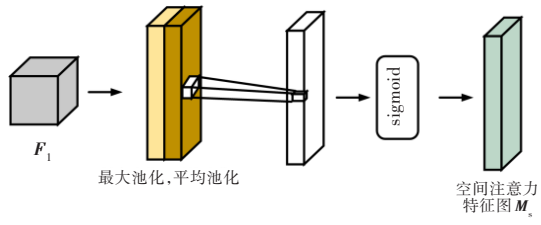

图6 空间注意力模块(SAM)结构[28]

Fig. 6 Structure of spatial attention module (SAM)[28]

CBAM结构结合了图像的特征通道和空间位置,通过注意力机制提高整个网络的泛化能力,突出重要特征,抑制非必要特征。本文利用CAM和SAM模块可以灵活地嵌入两个特征图之间的优点,在PPM1和PPM2之间嵌入CBAM结构,既能够充分利用原始图像的全局上下文信息,也能把提取到的特征同原始图像结合,减少信息损失。对于CAM在前、SAM在后的串行结构,WOO等[28]已经证明其效果比并行结构的或者单独使用的更好,并且最大池化和平均池化可以在全局范围内聚合图像特征,有利于图像目标特征的分割。

1.2 双结构PSPNet语义分割网络模型

受YANG等[29]的双边架构网络的启发,本文提出双结构PSPNet网络,其网络结构分为PPM1、PPM2和CBAM三部分,如图7所示。

1)经过MobileNet V2之后,输出不同维度的特征图$\boldsymbol{M}_i(i=1,2,\cdots,n)$。

2)将特征图划分为1×1、2×2、3×3和6×6的不同区域,分别进行平均池化;然后做卷积核大小为1×1的卷积运算,此时特征图的通道数变为1/4$\boldsymbol{M}_i$;接着将得到的4个特征图进行双线性插值上采样至尺寸与$\boldsymbol{M}_i$相同;最后与$\boldsymbol{M}_i$进行拼接。

3)对拼接后的特征图作1×1的卷积运算,将通道数降到1/2$\boldsymbol{M}_i$(即与$\boldsymbol{M}_{i-1}$通道数保持一致);然后进行上采样至尺寸与$\boldsymbol{M}_{i-1}$相同,得到$\boldsymbol{M}_{i+1}$;接着与$\boldsymbol{M}_{i-1}$拼接;最后经过CBR和CBAM得到$\boldsymbol{M}_{i+2}$。

4)将$\boldsymbol{M}_{i+2}$作为PPM2的输入,PPM2的操作与PPM1相同。

5)经softmax生成最终的分割图像,各特征图的通道数见表1。

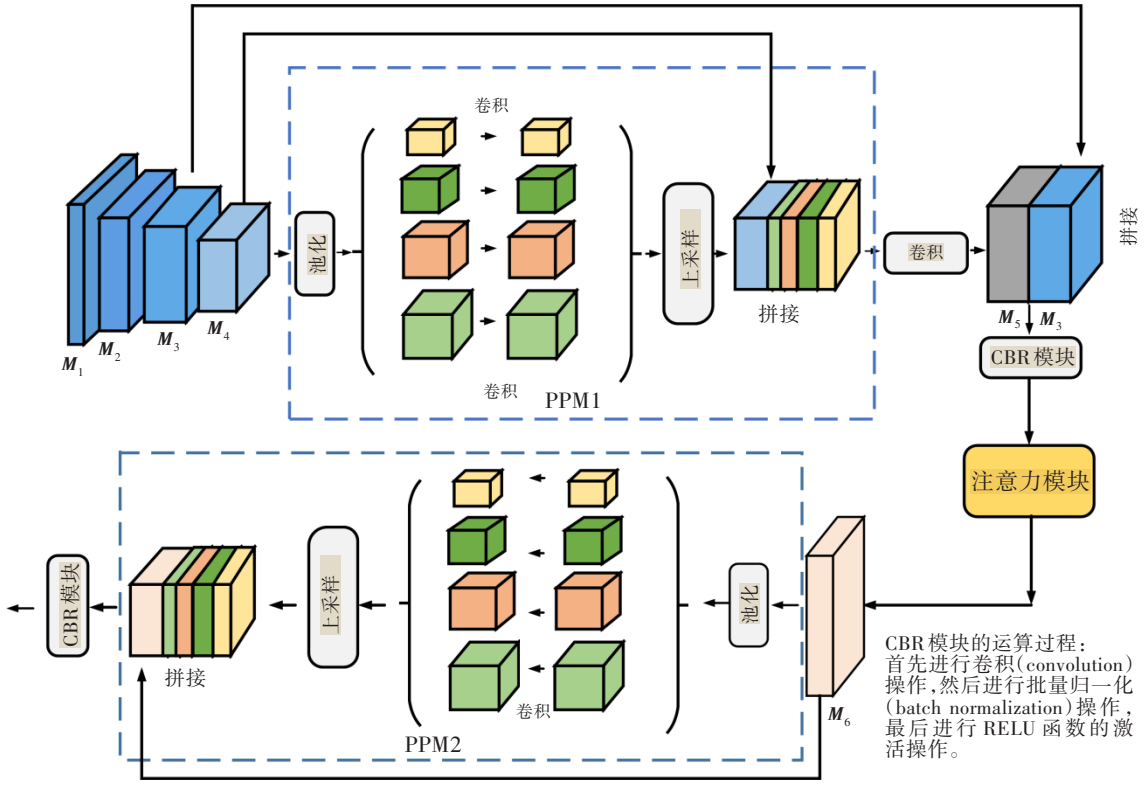

图7 融合注意力模块的双结构金字塔场景解析网络结构

Fig. 7 Two-structure pyramid scene parsing network with attention module

表1　各特征图的通道数

Table 1　Number of channels of each feature map

特征图	通道数
$\boldsymbol{M}_{i-1}$	$C/2$
$\boldsymbol{M}_i$	C
$\boldsymbol{M}_{i+1}$	$C/2$
$\boldsymbol{M}_{i+2}$	$C/2$

2　试验结果与性能评价

2.1　试验细节

PASCAL VOC 2007包含汽车、飞机、人等20个类别10 582张图像。本文以PASCAL VOC 2007中的1 449张图像为研究对象。

本研究在训练前，将输入图像的尺寸调整为473×473像素，选择自适应动量（adaptive momentum，ADAM）随机优化算法进行优化。该优化算法的训练分为冻结主干网络阶段和解冻阶段。本研究借鉴迁移学习的思想，考虑到主干网络提取到的特征是通用的，因此采用冻结训练来加快训练效率，这在一定程度上也可有效防止权值被破坏。冻结阶段的周期（epoch）指整个训练数据集被神经网络完整地遍历一遍的次数，为50；批大小（batch-size）指每次模型权重更新时的使用样本数，为8。解冻阶段的epoch也为50（总的epoch为100），batch-size为4，学习率为5×10^{-4}。本研究利用验证集进行验证，每个epoch保存一次训练结果。

2.2　评价指标

本研究采用平均像素精度（mean pixel accuracy，MPA）的指标E_{MPA}和平均交并比（mean intersection over union，MIOU）的指标E_{MIOU}来评价模型的分割效果。

$$E_{\mathrm{MPA}}=\frac{1}{m}\sum_{i=1}^{m}\frac{p_{ii}}{\sum_{j=1}^{m}\sum_{j=1}^{m}p_{ij}} \tag{8}$$

$$E_{\mathrm{MIOU}}=\frac{1}{m}\sum_{i=1}^{m}\frac{p_{ii}}{\sum_{j=1}^{m}p_{ij}+\sum_{j=1}^{m}(p_{ji}-p_{ii})} \tag{9}$$

式中：正整数m为标签类别数；i=1，2，3，…，m；j=1，2，3，…，m；p_{ii}、p_{ij}和p_{ji}分别为真正例、假正例和假负例。

2.3　结果分析

为了验证本文所提模型的有效性，在其他参数和试验设置条件均相同时进行对比试验。第一组验证加入CBAM的有效性；第二组验证加入PPM2的有效性。

2.3.1　第一组验证

将加入PPM2的网络结构和本文提出的模型分别记为模型Ⅰ和模型ours。本文将两者进行对比，结果如表2所示。从表2可以看出，CBAM提升了总体平均像素精度和平均交并比。模型ours分割结果中的平均像素精度和平均交并比分别达到85.64%和78.12%，比模型Ⅰ的分别提高了3.93、8.30个百分点。

在加入CBAM进行分割时，本文在不同阶段对不同的特征赋予不同的权重。首先，计算输入图像中每个通道的权重；接着，对包含重要信息多的通道给予更多的关注，对无重要信息的通道给予更少的关注；然后，找到关键通道，进一步确定重要信息在特征图中的最密集位置；最后，突出重要特征，弱化不重要特征，从而提高模型的特征表达能力。

2.3.2　第二组验证

1）将加入CBAM的PSPNet记为模型Ⅱ，并与模型ours进行对比，如表2所示。从表2可以看出，加入PPM2提升了总体平均像素精度和平均交并比；模型ours的总体平均像素精度、平均交并比比模型Ⅱ的分别提高了2.60、8.16个百分点。

2）将原始PSPNet记为模型Ⅲ，并与模型Ⅰ对比，如表2所示。显然，加入PPM2的模型的总体平均像素精度和平均交并比均略有提升。

表2　有无CBAM和PPM2分割结果对比

Table 2　Comparison of segmented results with/without CBAM and with/without PPM2　　%

网络模型	PPM2	CBAM	E_{MPA}	E_{MIOU}
Ⅱ	无	有	83.04	69.96
ours	有	有	85.64	78.12
Ⅰ	有	无	81.71	69.82
Ⅲ	无	无	80.69	65.81

由以上分析可知，加入PPM2的分割效果有所提升。由于深度神经网络的下采样使图像细节信息和空间信息出现损失，所以本文通过加入PPM2来融合上下文信息，可以更好地结合全局信息和局部信息，保留更多的图像细节特征，从而使模型学到的特征更丰富，以及对图像分割的结果更好。

本文模型注重简单对象分割和小且复杂对象分割之间的平衡。不同模型在部分对象上的分割结果如表3所示。由表3可知,本文模型对研究对象的分割结果,即对汽车、飞机和牛的分割平均交并比分别达到81.83%、81.49%和85.13%,以及对餐桌、椅子、自行车和盆栽的分割平均交并比分别达到45.27%、32.14%、38.43%和54.33%,均优于不加CBAM的模型I的结果。表3中前三列的值都高于后4列的,这说明,对于轮廓清晰、特征简单、类间差异较小的图像,两种模型的分割效果均较好;对于外观相似、边界不明显、较小的不同类别图像,两种模型的分割均比较困难。但与模型I相比,模型ours的分割效果比较好,尤其对盆栽的分割结果,其平均交并比提高了5.45个百分点。

由表3可知,在不同对象的分割中,模型ours与模型Ⅱ相比,其平均交并比有提升也有下降,但总体提升的类别数比较多,且总体提升的幅度大于下降的幅度。模型ours与模型I、Ⅲ相比,其总体的平均交并比更高一些,说明在同时加入CBAM和PPM2后的图像分割性能较好;网络能够在分割中充分利用图像信息,能结合图像深层、浅层信息更好地区分不同类别和细化同类别边界。

表3 有无CBAM和PPM2不同类别对象的E_{MIOU}

Table 3 E_{MIOU} of different categories with/without CBAM and with/without PPM2 %

网络模型	PPM2	CBAM	E_{MIOU}						
			汽车	飞机	牛	餐桌	椅子	自行车	盆栽
Ⅱ	无	有	82.58	80.43	78.70	48.23	32.74	36.78	51.24
ours	有	有	81.83	81.49	85.13	45.27	32.14	38.43	54.33
Ⅲ	无	无	81.79	82.87	80.45	36.83	33.21	37.83	52.72
I	有	无	79.89	80.52	82.46	38.12	31.65	36.88	48.88

本文使用MobileNet V2作为主干特征提取网络,减少了模型参数,加快了模型整体运算速度,提升了模型的运算效率。模型ours和其他模型的CEL值变化趋势如图8所示。在所有模型对比试验中,数值为0~1的交叉熵损失(cross entropy loss,CEL)函数被用于衡量模型输出与真实标签之间的差异。具体如下:当模型处理完全准确时,CEL值接近0;随着预测误差增加,CEL值也会增加。从图8可知,在epoch为60个左右时,模型ours收敛并趋于稳定;与其他模型相比,模型ours的CEL值下降最快、最平稳,这说明了CEL具有一定程度的有效性。

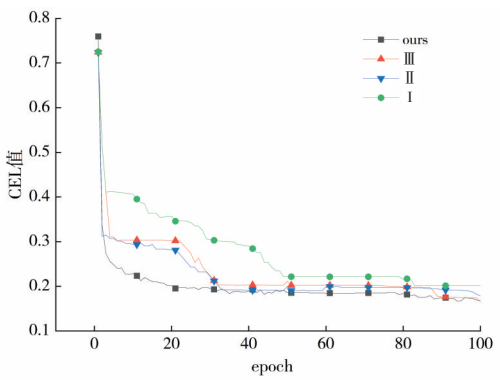

图8 不同模型的CEL值对比

Fig. 8 Comparison of CEL values of different models

3 讨论

为改善语义分割中原始图像信息丢失和图像分辨率下降造成的分割结果不准确问题,本文提出了融合CBAM及PPM2的双结构PSPNet分割模型,并通过试验验证了加入部分的有效性,证实本文提出的网络模型提升了图像分割精度。

1)使用深度可分离卷积和空洞卷积降低模型参数,加快训练速度,提高图像分辨率。本研究在使用的主干网络MobileNet V2中以深度可分离卷积代替普通卷积,并使用空洞卷积构造了轻便、快速的主干特征提取结构,为整个网络训练过程节省了时间和空间,并且为后续的图像特征融合和分割奠定了网络基础。

2)原始图像中的信息是多样化的,且任何网络都不可能在整个训练过程中同时对所有信息进行分析和综合,故分割试验在不同分割阶段必须有所取舍。本文通过在模型中加入CBAM,并在分割过程中的通道维度和空间维度方面对图像信息进行加权,达到了突出重要特征、抑制非重要特征的目的。

3)网络通过加入PPM2来融合多维度图像信

息，减少图像信息损失，从而保留更多的细节特征，使模型尽可能多地学习到更丰富的特征，进而提高原始图像信息的利用率，最终提升网络的分割精度。

本文以PSPNet为基础，使用MobileNet V2作为主干特征提取器，加入CBAM和PPM2，形成“PPM1-CBAM-PPM2”的串联结构，充分利用了图像上下文信息和空间位置信息，并在PASCAL VOC 2007数据集上进行训练。本研究结果显示，平均像素精度和平均交并比分别达到85.64%和78.12%，模型整体泛化性能较好，且在常见类别和难以分割的类别中分割性能都有所提升，同时模型收敛更快。

4　结论

语义分割是计算机视觉领域最具挑战性的问题之一。充分利用原始图像信息并增大图像中相同类别的一致性和不同类别的差异性，是提高分割精度的关键，也是一个不断优化的过程。

本文模型虽然取得了一些进步，但由于试验次数较少、试验数据单一、对比试验不多，故在网络结构、分割细节等方面还有一定改进空间。

[参考文献]

[1] YU H S, YANG Z G, TAN L, et al. Methods and datasets on semantic segmentation: a review [J]. Neurocomputing, 2018, 304: 82-103. DOI: 10.1016/j.neucom.2018.03.037.

[2] GUO Y M, LIU Y, GEORGIOU T, et al. A review of semantic segmentation using deep neural networks [J]. International Journal of Multimedia Information Retrieval, 2018, 7(2): 87-93. DOI: 10.1007/s13735-017-0141-z.

[3] 景庄伟，管海燕，彭代峰，等. 基于深度神经网络的图像语义分割研究综述[J]. 计算机工程，2020，46(10)：1-17. DOI：10.19678/j.issn.1000-3428.0058018.
JING Zhuangwei, GUAN Haiyan, PENG Daifeng, et al. Survey of research in image semantic segmentation based on deep neural network [J]. Computer Engineering, 2020, 46(10): 1-17. DOI: 10.19678/j.issn.1000-3428.0058018.

[4] 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. DOI：10.13328/j.cnki.jos.005659.
TIAN Xuan, WANG Liang, DING Qi. Review of image semantic segmentation based on deep learning [J]. Journal of Software, 2019, 30(2): 440-468. DOI: 10.13328/j.cnki.jos.005659.

[5] ZHU L Z, KANG Z Z, ZHOU M, et al. CMANet: cross-modality attention network for indoor-scene semantic segmentation [J]. Sensors, 2022, 22(21): 8520-8541. DOI: 10.3390/s22218520.

[6] HUANG S Y, HSU W L, HSU R J, et al. Fully convolutional network for the semantic segmentation of medical images: a survey [J]. Diagnostics, 2022, 12(11): 2765-2785. DOI: 10.3390/diagnostics12112765.

[7] OTSU N. A threshold selection method from gray-level histograms [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1979, 9(1): 62-66. DOI: 10.1109/TSMC.1979.4310076.

[8] PUN T. A new method for grey-level picture thresholding using the entropy of the histogram [J]. Signal Processing, 1980, 2(3): 223-237. DOI: 10.1016/0165-1684(80)90020-1.

[9] YEN J C, CHANG F J, CHANG S. A new criterion for automatic multilevel thresholding [J]. IEEE Transactions on Image Processing, 1995, 4(3): 370-378. DOI: 10.1109/83.366472.

[10] KHAN J F, BHUIYAN S M A, ADHAMI R R. Image segmentation and shape analysis for road-sign detection [J]. IEEE Transactions on Intelligent Transportation Systems, 2011, 12(1): 83-96. DOI: 10.1109/TITS.2010.2073466.

[11] TREMEAU A, BOREL N. A region growing and merging algorithm to color segmentation [J]. Pattern Recognition, 1997, 30(7): 1191-1203. DOI: 10.1016/S0031-3203(96)00147-1.

[12] CHENG Y Z. Mean shift, mode seeking, and clustering [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995, 17(8): 790-799. DOI: 10.1109/34.400568.

[13] BOYKOV Y Y, JOLLY M P. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images [C]//Proceedings of the Eighth IEEE International Conference on Computer Vision. Vancouver, BC, Canada: IEEE, 2001: 105-112. DOI: 10.1109/ICCV.2001.937505.

[14] CRESPO J, SCHAFER R W, SERRA J, et al. The flat zone approach: a general low-level region merging segmentation method [J]. Signal Processing, 1997, 62(1): 37-60. DOI: 10.1016/S0165-1684(97)00114-X.

[15] 黄鹏，郑淇，梁超. 图像分割方法综述[J]. 武汉大学学报(理学版)，2020，66(6)：519-531. DOI：10.14188/j.1671-8836.2019.0002.
HUANG Peng, ZHENG Qi, LIANG Chao. Overview of image segmentation methods [J]. Journal of Wuhan

University (Natural Science Edition), 2020, 66(6): 519-531. DOI: 10.14188/j.1671-8836.2019.0002.

[16] 龚红仿, 孙中宇, 狄俊珂. 基于滑动平均与规则决策的卷积神经网络图像分类[J]. 长沙理工大学学报(自然科学版), 2020, 17(3): 102-110. DOI: 10.3969/j.issn.1672-9331.2020.03.014.
GONG Hongfang, SUN Zhongyu, DI Junke. Convolutional neural network for image classification based on moving average and rule decision[J]. Journal of Changsha University of Science & Technology (Natural Science), 2020, 17(3): 102-110. DOI: 10.3969/j.issn.1672-9331.2020.03.014.

[17] WANG C L, MAUCH L, SAXENA M M, et al. On the contextual aspects of using deep convolutional neural network for semantic image segmentation[J]. Journal of Electronic Imaging, 2018, 27(5): 051223-051235. DOI: 10.1117/1.jei.27.5.051223.

[18] MANUGUNTA R K, MASKELIŪNAS R, DAMAŠEVIČIUS R. Deep learning based semantic image segmentation methods for classification of web page imagery[J]. Future Internet, 2022, 14(10): 277-290. DOI: 10.3390/fi14100277.

[19] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651. DOI: 10.1109/TPAMI.2016.2572683.

[20] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation [M]. Cham, Germany: Springer International Publishing, 2015: 234-241. DOI: 10.1007/978-3-319-24574-4_28.

[21] 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. DOI: 10.1109/TPAMI.2016.2644615.

[22] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs[C]//Proceedings of the 3rd International Conference on Learning Representations. San Diego, CA, USA: ICLR 2015, 2015. DOI:10.48550/arXiv.1412.7062.

[23] 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, 2018, 40(4): 834-848. DOI: 10.1109/TPAMI.2017.2699184.

[24] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation [M]. Cham, Germany: Springer International Publishing, 2018: 833-851. DOI: 10.1007/978-3-030-01234-2_49.

[25] NING Q Q, ZHU J K, CHEN C. Very fast semantic image segmentation using hierarchical dilation and feature refining[J]. Cognitive Computation, 2018, 10(1): 62-72. DOI: 10.1007/s12559-017-9530-0.

[26] ZHAO H S, SHI J P, QI X J, et al. Pyramid scene parsing network [C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 6230-6239. DOI: 10.1109/CVPR.2017.660.

[27] SANDLER M, HOWARD A, ZHU M L, et al. MobileNet V2: inverted residuals and linear bottlenecks [C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 4510-4520. DOI: 10.1109/CVPR. 2018. 00474.

[28] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module [M]. Cham, Germany: Springer International Publishing, 2018: 3-19. DOI: 10.1007/978-3-030-01234-2_1.

[29] YANG K Y, WANG Z H, YANG Z, et al. RecepNet: network with large receptive field for real-time semantic segmentation and application for blue-green algae[J]. Remote Sensing, 2022, 14(21): 5315-5339. DOI: 10.3390/rs14215315.

(责任编辑:彭三军;校对:刘平)