DOI:10.16136/j.joel.2022.03.0392

一种有效融合多尺度特征的图像语义分割方法

许光宇*,汤伟建

(安徽理工大学 计算机科学与工程学院,安徽 淮南 232001)

摘要:卷积神经网络在高级计算机视觉任务中展现出强大的特征学习能力,已经在图像语义分割任务中取得了显著的效果。然而,如何有效地利用多尺度的特征信息一直是个难点。本文提出一种有效融合多尺度特征的图像语义分割方法。该方法包含4个基础模块,分别为特征融合模块(feature fusion module,FFM)、空间信息模块(spatial information module,SIM)、全局池化模块(global pooling module,GPM)和边界细化模块(boundary refinement module,BRM)。FFM采用了注意力机制和残差结构,以提高融合多尺度特征的效率,SIM由卷积和平均池化组成,为模型提供额外的空间细节信息以辅助定位对象的边缘信息,GPM提取图像的全局信息,能够显著提高模型的性能,BRM以残差结构为核心,对特征图进行边界细化。本文在全卷积神经网络中添加4个基础模块,从而有效地利用多尺度的特征信息。在PASCAL VOC 2012数据集上的实验结果表明该方法相比全卷积神经网络的平均交并比提高了8.7%,在同一框架下与其他方法的对比结果也验证了其性能的有效性。

关键词:卷积神经网络;图像语义分割;多尺度特征;特征融合;注意力机制

中图分类号:TP391 **文献标识码**:A **文章编号**:1005-0086(2022)03-0264-08

An image semantic segmentation method effectively fusing multi-scale features

XU Guangyu*, TANG Weijian

(School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, Anhui 232001, China)

Abstract: Convolutional neural networks show strong feature learning ability in advanced computer vision and have achieved remarkable effect in image semantic segmentation tasks. However, how to use the multi-scale feature information effectively is always a difficulty. This paper proposes an effective image semantic segmentation method which integrates multi-scale features. The proposed method consists of four basic modules, which are feature fusion module (FFM), spatial information module (SIM), global pooling module (GPM) and boundary refinement module (BRM). FFM adopts attention mechanism and residual structure to improve the efficiency of multi-scale feature fusion. SIM includes convolution and average pooling opearaitons, and its purpose is to assist in locating the edge information of the object by providing additional spatial details. GPM extracts the global information of the image, which can significantly improve the performance of the model. BRM takes the residual structure as the core to refine the boundary of the feature map. Four basic modules are added into the full convolutional neural network to effectively utilize the multi-scale feature information. Experimental results on PASCAL VOC 2012 dataset show that mean intersection over union of the proposed method is 8.7% higher than that of full convolutional neural network. The results of comparison with other methods in the same framework also verify the effectiveness of the proposed method.

Key words: convolutional neural network; image semantic segmentation; multi-scale feature; feature fusion; attention mechanism

* **E-mail**: xgy761220@163.com

收稿日期:2021-06-07 **修订日期**:2021-07-05

基金项目:国家自然科学基金(61471004)和安徽理工大学博士专项基金(ZX942)资助项目

1 引　言

图像语义分割是一个计算机视觉任务,目的是对图像中每个像素标注一个代表其语义类别的标签,从而将图像分割为若干个不同的类别区域。图像语义分割技术可应用于无人驾驶汽车、医疗影像、地理信息系统等众多领域,为自动驾驶过程中准确掌握道路信息、医疗辅助诊断、高精度位置定位信息等提供有力保障。随着卷积神经网络的快速发展,结合卷积神经网络的图像语义分割技术比传统算法取得了更好的分割结果。因此,更多的研究者们将研究重心放在卷积神经网络结构的设计上[1-3]。

基于卷积神经网络的图像语义分割结构一般可分为编码器网络和解码器网络,编码器网络负责对输入图像提取多尺度特征信息,而解码器网络利用这些特征信息对输入图像进行预测。很多图像语义分割方法选择 VGG[4] 和 ResNet[5] 作为编码器网络,但很难有效地利用编码器网络所产生的多尺度特征信息。VGG 和 ResNet 在浅层提取到的特征图具有更多的空间细节信息,这有助于解码器网络恢复图像中处理对象的边缘,而在深层网络提取到的特征图具有更多的语义信息,这有助于对图像中对象区域的分类识别。随着编码器网络不断深入,下采样操作会使特征图的分辨率逐层减小,这也会使同一对象在不同层次的特征图中具有不同大小的尺度,即多尺度特征信息,而多尺度特征信息对分割结果有着重要作用。因此,如何有效利用多尺度特征信息已成为图像语义分割领域的研究热点之一。

早期的图像语义分割技术通常使用基于像素自身低阶视觉信息的无监督策略,或采用人工提取特征并与分类器相结合的传统机器学习方法。文献[6]提出一种基于卷积神经网络的 AlexNet 深度学习模型,并且在 2012 年的大规模图像识别挑战赛(ILSVRC)中以 15.3%的最低 top-5 错误率赢得第一名。AlexNet 的成功标志着计算机视觉任务进入了一个以深度学习方法为主的研究阶段,也为深度学习方法在图像语义分割任务中的应用拉开了序幕。文献[7]在 VGG16 分类网络基础上提出了一种全卷积神经网络(fully convolutional network,FCN),其以卷积层代替 VGG16 分类网络的全连接层,以接受任意大小的图像输入。FCN 是一个经典的编码器-解码器网络,其端到端的特性在一定程度上解决了传统机器学习方法中人工提取特征困难且提取的特征表达能力受限等问题。FCN 在图像语义分割任务中具有里程碑的意义,端到端的方法由此成为主流。

FCN 方法的灵活和强大在于其拥有充分的学习分层特征的能力,然而 FCN 的解码器融合多尺度的特征信息所带来的增益很小,进而导致分割结果较差。基于 FCN 的 SegNet 结构[8]是一个严格对称的编码器-解码器网络。其编码器中保留了最大池化索引,而解码器利用这些最大池化索引来进行反池化,一定程度上提高了解码器融合多尺度特征的能力。ZHANG 等[9]采用语义嵌入分支方法进行特征融合,其先将编码器的高级特征图做卷积运算,再通过双线性插值上采样,最后与编码器低级特征做乘法。文献[10]以 ResNet101 为编码器,在解码器部分设计了一个复杂的优化模块以融合编码器的多尺度特征。增强特征融合的解码器(enhanced feature fusion decoder,EFFD)[11]较好地解决了 FCN 解码器部分特征融合低效的问题。与 FCN 相比,EFFD 引入自身平方项的注意力机制后,分割质量有了很大的提升。

本文针对 FCN 无法有效地利用编码器网络的多尺度特征的问题,提出一种有效地融合多尺度特征的图像语义分割方法。该方法包含 4 个基础模块,分别为特征融合模块(feature fusion module,FFM)、空间信息模块(spatial information module,SIM)、全局池化模块(global pooling module,GPM)和边界细化模块(boundary refinement module,BRM)。本文在 FCN 网络中添加 4 个基础模块,从而有效地利用多尺度特征信息。在公开数据集上的实验结果表明该方法的有效性。

2 相关研究

2.1 残差结构

从经验上来看,网络的深度对模型的性能至关重要,当增加网络的层数后,网络可以进行更加复杂的特征模式的提取,所以在理论上当网络的深度更深时可以获得更好的结果。但在实验中发现了深层网络的退化问题,随着网络深度的增加,网络的精度出现饱和,甚至出现了精度下降的现象。为了解决网络的退化问题,文献[5]提出了残差结构。残差结构如图 1 所示,该结构与电路中“短路”相似,所以也称为短路连接。

对于一个堆叠层结构(几个非线性层堆叠而成),假设输入为 x,期望输出是 $H(x)$。如果将输入 x 传到输出 $H(x)$作为初始结果,则学习的目标为 $F(x)=H(x)-x$,这也称为残差学习。当残差为 0 时,此时堆叠层仅仅做了恒等映射(identity mapping),网络性能不会下降。但实际上残差不会为 0,这也使得堆叠层在输入特征的基础上学习到新的特

征，从而获得更好的性能。

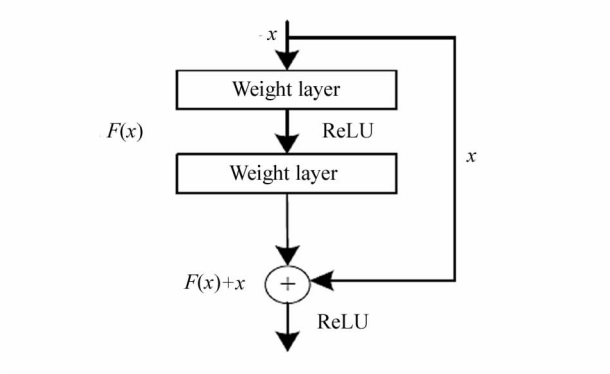

图 1　残差结构

Fig. 1　Residual structure

2.2　通道注意力机制

注意力机制最早被应用于机器翻译领域，如今已经成为计算机视觉任务的研究热点。注意力机制可以理解为通过模型学习来确定输入信息的那些部分需要更加关注，或者从输入信息的重要部分进行特征提取，以获得重要的信息。

文献[12]提出一种特征通道注意力机制，也称为“压缩和激活”(squeeze and excitation，SE)。该方法通过损失函数值的变化来学习特征图通道的权重参数，使有效特征图的权重变大，而无效或作用不大的权重变小。SE 的原理图如图 2 所示。

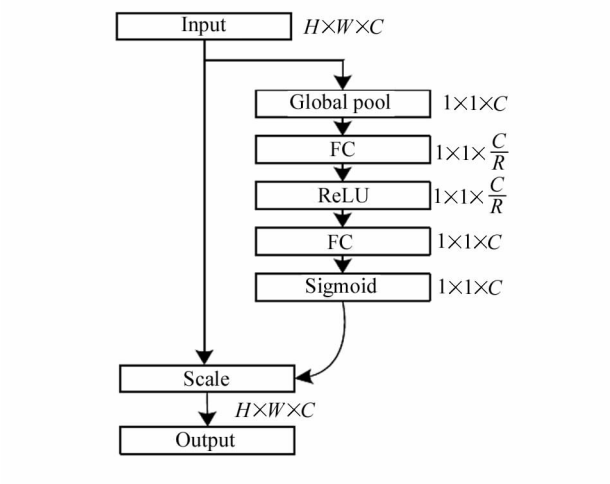

图 2　SE 的原理图

Fig. 2　Schematic diagram of SE

在编码器网络提取特征信息时，每个学习到的卷积核都有一个局部感受野，如 3×3 卷积核的感受野为 3×3 大小的矩形，这使得编码器输出的每个特征图都不能有效利用这个区域之外的上下文信息。为了解决这个问题，SE 通过全局平均池化将每个 2 维的特征图压缩成具有全局感受野的 1 维向量，这被称为压缩操作。

为了有效地利用压缩操作中聚合的信息，SE 采用了激活操作以捕获特征通道之间的相关性。输入特征经过全局平均池化得到具有特征图全局信息的 1 维向量，1 维向量依次经过 2 个全连接层后得到特征通道的权重参数，这 2 个全连接层的作用是融合各个通道的特征图信息。由于特征通道的权重参数需要通过全连接层和非线性层不断学习得到，因此可以端到端训练。其表达式为：

$$s=\sigma(W_2\delta(W_1\boldsymbol{y})), \tag{1}$$

式中，$\boldsymbol{s}$ 表示输入特征图通道的权重信息，$\boldsymbol{y}$ 表示由全局平局池化压缩的一维向量，维度为 $1\times1\times C$(C 为特征图的通道数)，W_1 和 W_2 表示全连接层，二者的维度分别为 $\frac{C}{R}\times C$、$C\times\frac{C}{R}$，其中参数的作用是降低全连接层的计算量($R=16$)，δ 和 σ 分别表示 Sigmoid 激活函数和 ReLU 激活函数。

SE 在获得输入特征图通道的权重信息之后，再利用缩放操作(Scale)将权重信息与输入信息按照对应的通道相乘。其计算式为：

$$\boldsymbol{X}=F_{\text{scale}}(\boldsymbol{s},\boldsymbol{u}), \tag{2}$$

式中，$\boldsymbol{X}$ 表示带有权重信息的原始输入特征，F_{scale} 表示逐通道的乘法操作，$\boldsymbol{s}$ 表示输入特征图通道的权重信息，$\boldsymbol{u}$ 表示原始输入信息。

2.3　全局池化

在深层卷积神经网络中，感受野的大小可以大致表示模型包含多少上下文信息。虽然从理论上讲，ResNet50 的感受野已经大于输入图像，但卷积神经网络在高层的感受野比理论上要小。这使得许多模型没有充分整合的图像中对象的全局上下文信息，会造成预测对象不全面，进而导致预测结果粗糙。全局池化能有效地增强模型的表现性能和泛化能力，并且很多方法已经证明了全局池化的有效性。文献[13]提出金字塔池化模块，能够聚合不同区域的上下文信息，从而提高模型获取全局信息的能力。文献[14]在空间金字塔模块中应用全局图像级池化，并在多个数据集中获得出色的结果。

3　本文方法

本文提出的方法的网络结构由多条路径和多个基础模块组成，可分为编码器网络和解码器网络 2 个部分。网络结构如图 3 所示(图 3 中左边的虚线框为编码器网络，右边的虚线框为解码器网络)。

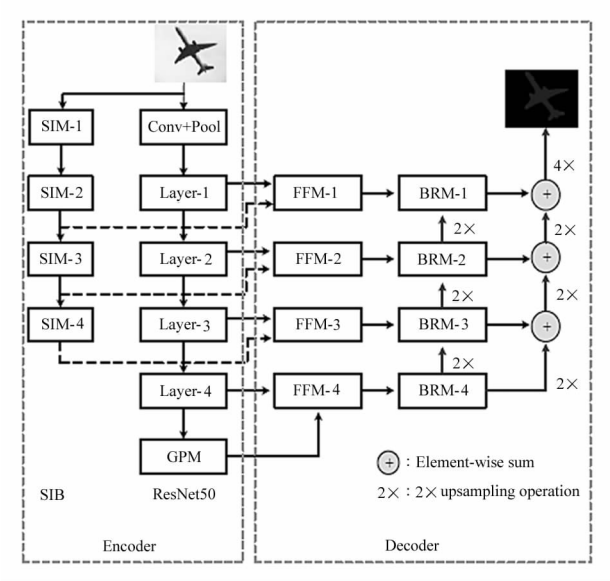

图 3　网络结构

Fig. 3　Network structure

3.1　编码器网络

本文的编码器网络分为 ResNet50、空间信息分支(spatial information branch,SIB)和 GPM 等 3 个部分。输入图像分为 2 个路径,一个路径经过空间信息分支获取输入图像的空间细节信息,为解码器网络补充图像中对象的边缘信息。另一个路径经过 ResNet50 提取输入图像的多尺度特征信息,该路径最后通过 GPM 获取图像的全局信息。

3.1.1　ResNet50 基础网络

本文将 ResNet50 网络尾部的池化层、全连接层和 softmax 分类层转化为卷积层,以接受任意尺度的图像输入。一般根据 ResNet50 所产生特征图的大小和特征图的数量不同,可将该网络分为五个阶段,分别为 Conv + Pool、Layer-1、Layer-2、Layer-3 和 Layer-4(每个阶段对应的特征图分辨率分别是输入图像分辨率的 1/4、1/4、1/8、1/16 和 1/32)。

表 1 为 ResNet50 的配置情况,其中 i 表示输入特征图通道的数量,o 表示输出特征图通道的数量,b 表示该层中残差单元的数量。

表 1　ResNet50 配置

Tab. 1　ResNet50 configuration

Block	i	o	b
Conv+Pool	3	64	—
Layer-1	64	256	3
Layer-2	256	512	4
Layer-3	512	1 024	6
Layer-4	1 024	2 048	3

3.1.2　空间信息分支

本文在编码器网络中添加了一条 SIB,如图 3 所示。空间信息分支由 4 个 SIM 组成,为解码器网络提供额外的空间细节信息以辅助定位对象的边缘信息。

SIM 由步长为 2 的 3×3 卷积层和步长为 2 的 3×3 平均池化层组成,由于卷积和平均池化的特性不同,因此二者在一定程度上能够提取到不同的空间细节信息。随着输入图像下采样次数的增加,特征图中包含空间信息将不断减少,而抽象的语义信息会不断增加。因此,为了在图像中提取到更多的空间细节信息,本文仅采用 4 个 SIM。SIM 的原理图,如图 4 所示。SIM 的输入特征分别经过平均池化层和卷积层下采样,将二者提取到的特征信息通过级联操作(Concat)合并在一起,最后输出特征。

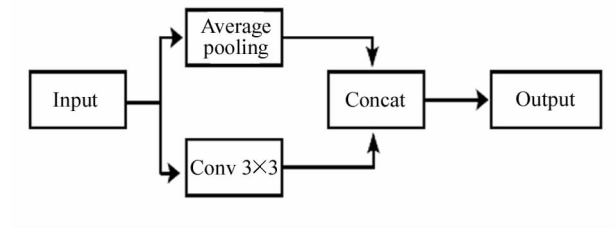

图 4　SIM 的原理图

Fig. 4　Schematic diagram of SIM

3.1.3　全局池化模块

本文在 ResNet50 网络底部添加一个 GPM,以获取图像的全局信息。GPM 的原理图如图 5 所示,GPM 对输入特征图进行全局平均池化,以得到图像的全局信息,再通过 1×1 卷积层降低特征图通道的数量,最后通过双线性插值将特征图上采样至原始输入图像分辨率的 1/32。

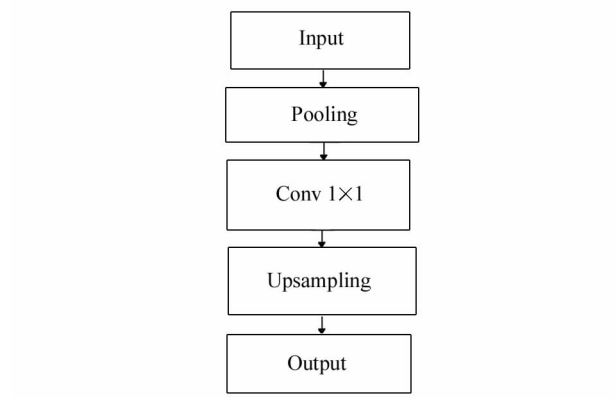

图 5　GPM 的原理图

Fig. 5　Schematic diagram of GPM

GPM 的核心是全局平均池化(global average pooling),其先将每一个通道的特征图按照每个位置产生的响应求和,最后取平均值,表达式为:

$$y_i = \frac{1}{H \times W} \sum_{j}^{H} \sum_{k}^{W} x_{i,j,k}, \tag{3}$$

式中,y_i 表示第 i 个输入特征图的全局平均池化后的结果,H 代表输入特征图的高,W 代表输入特征图的宽,$x_{i,j,k}$代表输入特征图第 i 个通道第 j 行第 k 列的响应值。

3.2　解码器网络

解码器网络负责将编码器网络在图像中提取的空间细节信息、多尺度特征信息及全局信息进行融合和预测,通过双线性插值逐步恢复特征图的分辨率,其主要由 FFM 和 BRM 组成。

3.2.1　特征融合模块

ResNet50 在不同阶段(Layer-1 至 Layer-4)的特征具有不同的特点,浅层特征包含更多的空间信息,而深层特征包含更多抽象的语义信息。此外,空间信息分支产生的特征信息也包含较多的空间信息。因此,本文面临着一个问题,即如何将 ResNet50 产生的多尺度特征信息与空间信息分支产生的空间细节信息充分地整合在一起。为了解决这个问题,本文在注意力机制和残差结构的基础上设计了一种 FFM。

FFM 的原理图如图 6 所示,FFM 接收两个同尺度大小的输入特征。X 为 ResNet50 基础网络所产生的多尺度特征信息(Layer-1至Layer-4),而Y为

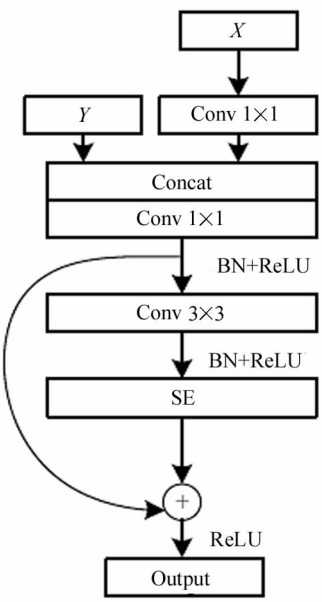

图 6　FFM 的原理图

Fig. 6　Schematic diagram of FFM

SIM 或 GPM 所产生的特征信息。X 先经过 1×1 卷积层,以压缩特征图的数量。压缩后的特征图与 Y 进行级联,再通过 1×1 卷积层产生子阶段结果。子阶段结果进入到结合通道注意力机制的残差结构,在学习输入特征通道之间相关性时,以权重的形式重点强调重要特征通道。

3.2.2　边界细化模块

BRM 的原理图如图 7 所示,本文设计了一种整合语义分类信息的 BRM。BRM 一般有两个输入,分别为 FFM 输出结果和 BRM 自身高层次的输出结果(BRM-4 仅有 FFM 的输入)。2 个输入通过级联操作整合在一起,之后 1×1 卷积层实现对特征图的预测,最后利用残差结构完成对象边界的细化。

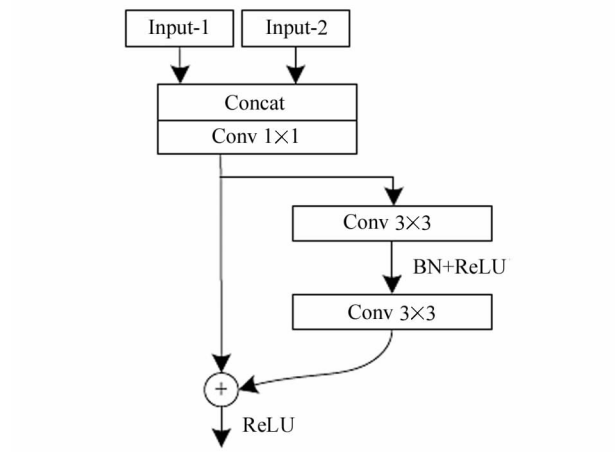

图 7　BRM 的原理图

Fig. 7　Schematic diagram of BRM

表 2 为 BRM 与其他边界细化方法的对比实验结果,其中 BR 为文献[15]的边界细化方法(boundary refinement),RRB 为文献[16]的细化残差块(refinement residual block),3 种方法的实验环境和模型均一致。平均交并比(mean intersection over union,$mIoU$)为实验的评估指标,$mIoU$ 值越大,则说明网络的分割质量越好。表中的实验数据说明 BRM 比 BR 和 RRB 更加适合本文的网络结构。

表 2　不同方法的对比

Tab. 2　Comparison of methods

BR	RRB	BRM	$mIoU$
			0.762 0
√			0.764 3
	√		0.768 0
		√	0.772 0

4　实验结果与分析

4.1　实验设置

1）实验环境：操作系统 64 位 Windows 10，显卡为 NVIDIA GeForce RTX 3080(10 GB)，处理器为 Intel.CoreTM i9-9900K CPU，深度学习框架为 Pytorch。

2）数据集：实验选用的数据集是 PASCAL VOC 2012[17]的语义分割部分，该数据集是一个公开的语义分割基准数据集，包含 20 个对象类别和 1 个背景类别，涉及 1,464 张训练图像、1,449 张验证图像和 1,456 张测试图像。为了获得更多的训练数据，本文对原始数据集进行了扩充，产生了10,582张训练图像。

3）评估指标：实验的评估指标为 $mIoU$，即先计算每一类真实标签图和预测图的两个像素点集合的交集和并集之比，然后在所有类上取一个平均值。$mIoU$ 的取值范围为[0,1]，平均交并比的值越大，说明网络的分割质量越好。计算式如下：

$$mIoU=\frac{1}{k}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}},\quad(4)$$

式中，k 是数据集中的目标类别数，p_{ii} 表示第 i 类物体的像素被分到第 i 类的像素的数量，p_{ij} 表示第 i 类的物体的像素被分到第 j 类的像素的数量。

4）数据增强：对输入图像做随机水平翻转、随机缩放和随机旋转的数据增强方法。随机缩放比例在[0.5,2]之间，随机水平翻转的概率为 0.5，随机旋转范围在[−10°,10°]之间。

5）训练策略：采用 ResNet50 作为编码器的基础网络，加载在 ImageNet 上训练参数。ResNet50 基础网络的学习率为 0.001，SIM、GPM 和解码器网络的学习率均为 0.01。采用随机梯度下降法对网络进行优化，batch size 取 8，权重衰减取 0.000 1，训练 60 个，近 80,000 次迭代达到收敛。目标损失函数为交叉熵，学习率调整策略设置为“poly”。poly 的计算式如下：

$$LR=lr\times\left(1-\frac{n}{N}\right)^{p},\quad(5)$$

式中，lr 为初始学习率，LR 为模型学习率，n 为当前迭代次数，N 为最大迭代次数，p 为学习率参数(设置为 0.9)。在训练阶段，输入图像在随机缩放后再随机裁剪为 380×380 像素大小。

4.2　结果对比与分析

本文在实验部分选择经典的 FCN 作为对照组，并将 FCN 的编码器网络调整为 ResNet50、上采样方法调整为双线性插值、学习率设置为 0.01。

表 3 为添加不同模块对分割结果的影响。当加入 SIM、GPM、FFM 和 BRM 后，模型都能得到一定的性能提升，说明了 4 种基础模块的有效性。尤其当添加 GPM 后，提高了 5.32%，达到 0.746 2，这表明全局信息对模型性能的提升有很大帮助。

表 3　添加不同模块对分割结果的影响

Tab.3　Influence of different modules on segmentation results

SIM	GPM	FFM	BRM	$mIoU$
				0.685 0
√				0.693 0
√	√			0.746 2
√	√	√		0.762 0
√	√	√	√	0.772 0

本文方法与 FCN 的对比如表 4 所示，通过计算 $mIoU$ 的方式，对二者进行比较。通过对比实验，可以明显看出本文方法所取得的 $mIoU$ 远远高于 FCN，相比 FCN，本文方法通过使用 SIM、FFM、GPM 和 BRM 能够更加有效地利用编码器网络所产生的多尺度特征信息，从而获得比其他方法更优的结果。

表 4　本文方法与 FCN 的 *mIoU* 对比

Tab.4　*mIoU* comparison of our method and FCN

Model	$mIoU$
FCN	0.685 0
Ours	0.772 0

在引言中，本文认为 FCN 无法有效利用编码器网络所产生的多尺度特征信息，因此在表 5 和表 6 中对 FCN 和本文方法融合不同阶段特征的分割结果的影响进行了比较。表中的{1,2,3,4}分别代表 ResNet50 中 Layer-1、Layer-2、Layer-3 和 Layer-4 的输出特征图。根据表 5 和表 6 的实验结果，FCN 融合{3,4}，{2,3,4}和{1,2,3,4}后，分别提高了 1.71%、0.09%、0.1%，而本文方法的分别提高 2.4%、0.58%、0.52%。

FCN 的解码器部分和特征融合方式较为简单，因此其融合 ResNet50 的多尺度特征的结果较差。相比 FCN，本文方法融合 ResNet50 的多尺度特征所带来的性能提升都高于 FCN，这也证明本文方法能够有效地利用编码器网络所产生的多尺度特征信息。

表 5 FCN 融合不同阶段特征的分割结果

Tab. 5 FCN segmentation results using given feature levels

Model	Stages	*mIoU*
FCN	{4}	0.666 0
	{3,4}	0.683 1
	{2,3,4}	0.684 0
	{1,2,3,4}	0.685 0

表 6 本文方法融合不同阶段特征的分割结果

Tab. 6 Our segmentation results using given feature levels

Model	Stages	*mIoU*
Ours	{4}	0.737 0
	{3,4}	0.761 0
	{2,3,4}	0.766 8
	{1,2,3,4}	0.772 0

为了进一步证明本文方法的有效性，在保证参数量相近的基础上，在同一框架下比较不同特征融合方法对模型性能的提升。基于 ResNet50 的对比，如表 7 所示。在表 7 中，SEB 为文献[9]的语义嵌入分支，RCU 为文献[10]的残差卷积单元，EFFD 为文献[11]增强特征融合的解码器部分。实验表明本文方法比基于 SEB 的方法、基于 RCU 的方法和基于EFFD等注意力机制的方法对模型性能的提升最大。

本文从模型的参数量和 *mIoU* 角度与语义分割领域中的经典分割网络进行了对比，经过消融实验，得到了与其他方法对比的数据如表 8 所示，其中 *m* 代表不同模型的可训练参数量。

表 7 基于 ResNet50 的 *mIoU* 对比

Tab. 7 Comparison of *mIoU* based on ResNet50

Model	*mIoU*
ResNet50-SEB	0.703 0
ResNet50-RCU	0.721 0
ResNet50-EFFD	0.764 0
Ours	0.772 0

表 8 与其他方法对比

Tab. 8 Comparison of segmentation methods

Model	$m(\times 10^6)$	*mIoU*
GCN[15]	25.40	0.661 0
FCN[7]	24.23	0.685 0
SegNet[8]	53.56	0.694 0
DeepLabV3+[14]	59.34	0.773 0
PSPNet[13]	51.44	0.781 0
Ours	30.11	0.772 0

表 8 中的其他方法均是在相同的实验环境中运行，微调了学习率、batch size、编码器网络等，以获得最优结果。本文方法、FCN、GCN 和 SegNet 均采用 ResNet50 作为编码器网络，而 DeepLabV3 + 和 PSPNet 以 ResNet101 作为编码器网络。

本文方法虽然比 GCN 和 FCN 多出近 500 万参数，但其性能远超这 2 个方法。根据表 8 的实验结果可知，DeepLabV3+和 PSPNet 的指标略高于本文方法，但二者的参数量远远超过本文方法。尤其是 DeepLabV3+方法的参数量比本文方法高出近 3 000 万参数量，但只提高了 0.1%，这说明了本文方法的优越性。与其他方法相比，本文方法在参数量和精度之间找到一个平衡点，以获取更优的分割结果。

分割结果对比如图 8 所示，本文方法与其他几种语义分割领域的方法分别使用相同的样本进行测试，从而得到语义分割效果图。从图 8 中第 3 行的效果图可以看出，FCN 方法和 PSPNet 方法对乘客旁边的公交车的处理有明显的分割错误，公交车没有完全被识别，而公交车的部分区域被误识别为火车。相对而言，本文方法和 DeepLabV3+方法的预测效果更好，体现出了强大的图像语义识别能力，而本文方法在细节处理方面比 DeepLabV3+更优秀。根据对实验效果图的对比，不难发现本文方法对图像的分割效果总体上要优于其他方法。

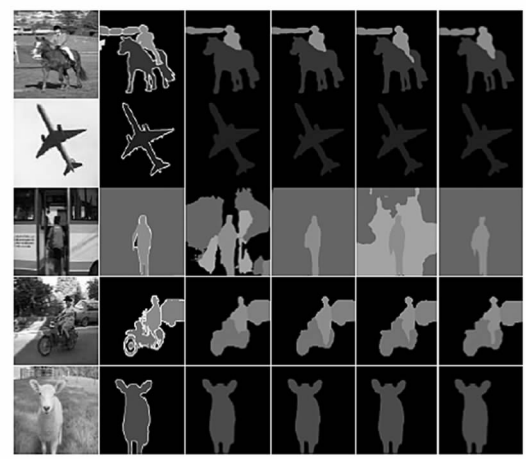

图 8 分割结果对比：(a) 原始输入图像；(b) 图像标签；(c) FCN 方法的分割结果；(d) DeepLabV3+方法的分割结果；(e) PSPNet 方法的分割结果；(f) 本文方法的分割结果

Fig. 8 Comparison of segmentation results: (a) Inputs; (b) Labels; (c) FCN; (d) DeepLabV3+; (e) PSPNet; (f) Ours

5 结　论

图像语义分割任务是计算机视觉和机器学习领域的研究热点之一,正为越来越多的视觉应用提供精确且高效的分割结果。本文针对 FCN 无法有效地利用编码器网络的多尺度特征信息的问题,提出一种有效融合多尺度特征的图像语义分割方法。本文方法在 PASCAL VOC 2012 公开数据集上进行了大量的对比实验,实验结果表明本文方法不仅能够有效地利用多尺度特征信息,而且还能在模型的参数量与精度之间找到一个平衡点。下一步将结合文献[14]的网络结构,研究性能更优的特征融合方式。

参考文献:

[1] FENG X J, SUN S J. Semantic segmentation method integrating multilevel features[J]. Application Research of Computers, 2020, 37(11): 3512-3515.
冯兴杰,孙少杰.一种融合多级特征信息的图像语义分割方法[J].计算机应用研究,2020,37(11):3512-3515.

[2] XU C H, SHI C, CHEN Y N. End-to-end dilated convolution network for document image semantic segmentation[J]. Journal of Central South University, 2021, 28(6): 1765-1774.

[3] TIAN Q C, MENG Y. Image semantic segmentation algorithm with multi-scale feature fusion and enhancement [J]. Computer Engineering and Applications, 2021, 57(2): 177-185.
田启川,孟颖.多尺度融合增强的图像语义分割算法[J].计算机工程与应用,2021,57(2):177-185.]

[4] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10) [2021-06-07]. https://arxiv.org/abs/1409.1556v6.

[5] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. Washington DC: IEEE Computer Society, 2016, 770-778.

[6] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

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

[8] 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.

[9] ZHANG Z L, ZHANG X Y, PENG C, et al. Exfuse: enhancing feature fusion for semantic segmentation[C]//The 15th European Conference on Computer Vision, September 8-14, 2018, Munich, Germany. Berlin: Springer-Verlag, 2018: 273-288.

[10] LIN G S, MILAN A, SHEN C H, et al. RefineNet: multi-path refinement networks for high-resolution semantic segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. Washington DC: IEEE, 2017: 1063-1069.

[11] MA Z H, GAO H J, LEI T. Semantic segmentation algorithm based on enhanceed feature fusion decoder[J]. Computer Engineering, 2020, 46(5): 254-258.
马震环,高洪举,雷涛.基于增强特征融合解码器的语义分割算法[J].计算机工程,2020,46(5):254-258.

[12] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[13] ZHAO H S, SHI J P, QI X J, et al. Pyramid scene parsing network[C]//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. Washington DC: IEEE, 2017: 2881-2890.

[14] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//European Conference on Computer Vision, September 8-14, 2018, Munich, Germany. Berlin: Springer, 2018: 833-851.

[15] PENG C, ZHANG X Y, YU G, et al. Large kernel matters: improve semantic segmentation by global convolutional network[C]//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. Washington DC: IEEE, 2017: 1743-1751.

[16] YU C Q, WANG J B, PENG C, et al. Learning a discriminative feature network for semantic segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. Washington DC: IEEE, 2018: 1857-1866.

[17] EVERINGHAM M, VAN GOOL L, WILLIAMS C K, et al. The pascal visual object classes challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

作者简介:

许光宇　(1976—),男,博士,硕士生导师,主要从事数字图像处理方面的研究.