基于注意力机制和特征融合的语义分割网络

才 华[1],王玉瑶[1],付 强[2],马智勇[3],王伟刚[3],张晨洁[1]

(1.长春理工大学 电子信息工程学院,长春 130022;2.长春理工大学 空间光电技术研究所,长春 130022;3.吉林大学第一医院 泌尿外二科,长春 130061)

摘 要:针对DeepLabv3+网络中的多尺度目标分割错误、多尺度特征图及不同阶段特征图之间关联性差的问题,提出在DeepLabv3+基础上引入全局上下文注意力模块、级联自适应尺度感知模块及注意力优化融合模块。将全局上下文注意力模块嵌入骨干网络特征提取的初始阶段,获取丰富的上下文信息;级联自适应尺度感知模块可建模多尺度特征之间的依赖性,使其更加关注目标特征;注意力优化融合模块通过多条支路融合多层特征,以此提高解码时像素的连续性。改进网络在Cityscapes数据集以及PASCAL VOC2012增强数据集上进行验证测试,实验结果表明:该网络能弥补DeepLabv3+的不足,且平均交并比分别达到76.2%、78.7%。

关键词:语义分割;多尺度特征;上下文信息;注意力机制;特征融合

中图分类号:TP391.4 **文献标志码**:A **文章编号**:1671-5497(2025)04-1384-12

DOI:10.13229/j.cnki.jdxbgxb.20230740

Semantic segmentation network based on attention mechanism and feature fusion

CAI Hua[1],WANG Yu-yao[1],FU Qiang[2],MA Zhi-yong[3],WANG Wei-gang[3],ZHANG Chen-jie[1]

(1. *School of Electronic Information Engineer*, *Changchun University of Science and Technology*, *Changchun* 130022, *China*; 2. *School of Opto-Electronic Engineer*, *Changchun University of Science and Technology*, *Changchun* 130022, *China*; 3. *No.2 Department of Urology*, *The First Hospital of Jilin University*, *Changchun* 130061, *China*)

Abstract: To address the issues of multi-scale object segmentation errors, poor correlation between multi-scale feature maps and feature maps at different stages in the DeepLabv3+ network, the following modules are proposed to incorporate, including a global context attention module, a cascade adaptive Scale awareness module, and an attention optimized fusion module. The global context attention module is embedded in the initial stage of the backbone network for feature extraction, allowing it to capture rich contextual information. The cascade adaptive scale awareness module models the dependencies between multi-scale features, enabling a stronger focus on the features relevant to the target. The attention optimized fusion module merges multiple layers of features through multiple pathways to enhance pixel continuity during decoding. The improved network is validated on the CityScapes dataset and PASCAL VOC2012 augmented dataset, and the experimental results demonstrate its ability to overcome the limitations of DeepLabv3+. Furthermore, the mean intersection over union reaches 76.2% and 78.7%

收稿日期:2023-07-15.

基金项目:国家自然科学基金重大项目(61890963);吉林省科技发展计划项目(20210204099YY,20240302089GX).

作者简介:才华(1977-),男,副教授,博士.研究方向:图像处理,机器视觉与人工智能算法.E-mail: caihua@cust.edu.cn

respectively.

Key words: semantic segmentation; multi-scale features; contextual information; attention mechanism; feature fusion

0　引　言

语义分割是计算机视觉三大主流任务之一,其本质是将图像中的每个像素和类标签相关联,达到密集像素预测的目的,在医学分割[1,2]、遥感图像[3-5]、自动驾驶[6,7]等复杂场景中得到了广泛应用。然而,现实场景中分割目标可能会受遮挡、尺度、光照的影响,给语义分割任务带来了极大的挑战,为有效完成复杂场景下的语义分割任务,提升对特征像素级的识别能力是十分必要的。近年来,基于深度学习的语义分割网络在精度上取得了显著的提高。根据深度网络架构设计不同,可将基于深度学习的语义分割分为如下3类:基于全卷积的语义分割网络[8]、基于编解码的语义分割网络[6,9-12]、基于Transformer的语义分割网络[13-16]。

基于全卷积的语义分割网络(Fully convolutional networks for semantic segmentation, FCN)[8]是分割领域的基础架构,通过将基于分类的卷积神经网络(Convolution neural networks, CNN)的全连接层替换为卷积层,使FCN可接受任意尺寸的输入图像,首次实现了端到端的像素级预测,在性能上远超越了所有传统依赖人工设计的分割算法。但由于FCN固定的感受野获取的上下文信息有限和简单的上采样操作使各层级的特征未充分表达。研究者在FCN的基础上做出改进,提出了基于编解码的语义分割网络,编码器通过卷积提取图像中的特征,而解码器利用上采样操作恢复特征,并将其分类,进而得到预测结果。如SegNet[6]网络,该网络通过池化索引跳跃连接的方式将编码器的特征信息传递到解码器中,引导特征更好地解码。但该网络在进行特征提取时,卷积核的尺寸是固定的,无法捕获图像中出现的多尺度信息。针对图像中多尺度信息获取困难的问题,DeepLabv3+网络[11]提出空洞空间金字塔池化模块(Atrous spatial pyramid pooling, ASPP),即利用多个具有不同扩张率的卷积,以实现对目标多尺度信息的有效捕获。但由于卷积无法学习长距离像素依赖关系的局限性,导致其在语义分割领域仍受到挑战。受自然语言处理(Natural language processing, NLP)中Transformer良好表征的启发,研究者尝试把Transformer技术引入计算机视觉领域[17],提出基于Transformer的语义分割网络[13-16],该类架构利用Transformer代替原编码器中的卷积层,通过将图像分割成小的图像块,从而将图片变成一维图像序列,利用Transformer的自注意力机制捕获全局上下文信息,进而为语义分割提供了一个全新视角,在大型数据集上,基于Transformer的语义分割网络相较于基于CNN的语义分割网络,展现出更优越的性能[17]。但基于Transformer的语义分割架构将图像视为一维序列,忽略了图像数据的二维属性,在进行特征提取时,需要自注意力机制对全局上下文信息建模,使网络具有较高的计算复杂度,此外,自注意力机制仅实现了空间维度的建模,而忽略了通道维度的自适应性。

因此,综合全局上下文信息和多尺度上下文信息的获取,以及计算复杂度方面的考虑,本文选用Deeplabv3+网络作为基础架构,并针对Deeplabv3+中出现的由于感受野有限及简单的通道堆叠融合方式所导致的多尺度目标分割错误、多尺度特征图及不同阶段特征图之间关联性差、目标边缘不准确的问题,提出在编码时,通过引入全局上下文注意力模块(Global context attention module, GCAM)获取更多上下文特征,并将提取的特征作为级联自适应尺度感知模块(Cascade adaptive scale awareness module, CASAM)的输入,完成多尺度特征的融合,利用注意力机制增强多尺度特征间的依赖性;在解码阶段提出注意力优化融合模块(Attention optimization fusion module, AOFM),利用通道注意力机制改善由简单通道叠加导致次优级特征被融合的问题,进一步提高分割性能,并在Cityscapes数据集和PASCAL VOC2012增强数据集上验证其分割效果,实验结果表明:本文方法进一步提升了分割性能,同时,具有很强的泛化能力。

1 相关工作

在语义分割领域,对特征提取时感受野的扩展与特征依赖性的增强已成为广大研究者研究的主流方向。目前,扩大感受野的方式主要采用空洞卷积[9-11]、池化[18,19]、大核卷积[20-22]等操作。Chen等[9-11]首次在DeepLab系列论文中提出利用空洞卷积扩大感受野,从而提高分割性能。Zhao等[18]提出的PSPNet采用金字塔池化模块获取全局上下文信息。Peng等[20]提出图卷积神经网络(GCN)方法,通过大核卷积获取较大的感受野,解决语义分割中分类和定位的问题。Guo等[22]提出视觉注意网络(Visual attention network, VAN),通过分解大核卷积策略,使分割网络具有局部性和空间、通道维度的自适应性以及较大的感受野。增强特征之间的依赖性主要是通过注意力机制的方式进行建模[23-25]。Hu等[23]提出的SENet网络是计算机视觉领域将注意力机制应用到通道维度的代表作,可通过特征重新标定的方式自适应地调整通道之间的特征响应。Wang等[24]提出ECANet采用一维卷积实现通道间的信息交互,利用自适应选择卷积核大小的方法实现局部信息交互,从而在模型复杂度和性能之间实现了较好的平衡。此外,利用自注意力机制建模远距离像素间的依赖关系。Fu等[25]提出DANet,将非局部的思想同时引入了通道域和空间域,分别将空间像素点以及通道特征作为查询语句进行上下文建模,自适应地整合局部特征和全局依赖。最近,Vision transformer(ViT)[17]在计算机视觉领域得到广泛应用,ViT是基于Transformer结构的视觉模型,主要利用自注意力机制捕获全局上下文信息。Zheng等[15]提出的SETR网络,是一种基于编解码结构的Transformer语义分割网络,但在解码过程中仍使用了卷积和上采样的操作。Xie等[16]提出的SegFormer网络是纯Transformer网络,使用轻量级的多层感知器(MLP)作为解码器,且取得了较好的分割效果,但是基于Transformer的网络参数量较大,计算复杂度较高。因此,本文通过改进上下文信息获取的方式和增强特征之间的依赖性,旨在设计一种计算复杂度低且具有良好预测效果的语义分割算法。

2 算法结构

本文算法对DeepLabv3+的编码器和解码器进行了针对性的改进。在编码器骨干网络初始特征提取阶段加入GCAM,使模型从最开始可以获取全局上下文信息。DeepLabv3+的ASPP模块并联不同扩张率的空洞卷积和池化获取局部感受野,但多尺度特征只是简单的拼接,并未考虑不同尺度特征图之间的差异。基于此,本文通过改进ASPP模块,提出CASAM,实现深层特征的多尺度信息动态融合,进一步改善由空洞卷积稀疏采样引起的信息丢失问题。DeepLabv3+网络的解码器输入由深层语义特征和浅层细节特征构成。由于语义分割是像素级别的分类任务,深层语义特征和浅层细节特征对其同样重要。深层特征分辨率较低,缺少细节特征,但含有丰富的语义信息,浅层特征分辨率较高,含有较多的细节和空间特征,但缺少语义特征。通过两者融合,使解码过程更好地恢复细节信息,实现不同层级信息的交互,得到更有判别力的特征,进而改善分割结果。但简单的通道融合导致次优特征被叠加,从而影响分割结果。本文提出AOFM,在特征融合之前加入改进的通道注意力ECA操作[24],使网络在特征融合时侧重于相对重要的特征,从而改善分割网络误割的情况。

改进分割网络总体结构如图1所示。

2.1 编码器

编码器主要是从输入图像中提取特征。本文编码器的设计如图1所示,输入图像经骨干网络ResNet101生成原图尺寸大小$\frac{1}{2}$、$\frac{1}{4}$、$\frac{1}{8}$、$\frac{1}{16}$的特征图,为获取全局感受野,在$\frac{1}{2}$、$\frac{1}{4}$、$\frac{1}{8}$、$\frac{1}{16}$特征上加入GCAM。将带有全局上下文信息的深层特征输入CASAM,得到融合后的多尺度语义特征。

2.1.1 全局上下文注意力模块

感受野尺寸决定了网络获取的上下文信息的多少,本文提出的GCAM通过大核卷积捕获长距离的像素依赖关系,从而为网络提供更全面的上下文信息,以辅助网络进行正确的判断,并改善模型的分割性能,但由于直接计算参数量较大,采取了分解策略,如图2所示。一个$K\times K$的大核卷积可分解为3部分:一个$\frac{K}{d}\times\frac{K}{d}$的深度扩张卷积(Depth-wise dilation convolution, DW-D-CONV)、一个$(2d-1)\times(2d-1)$深度卷积(Depth-wise convolution, DW-CONV)和一个1×1点卷积(Pointwise convolution),其中d为扩

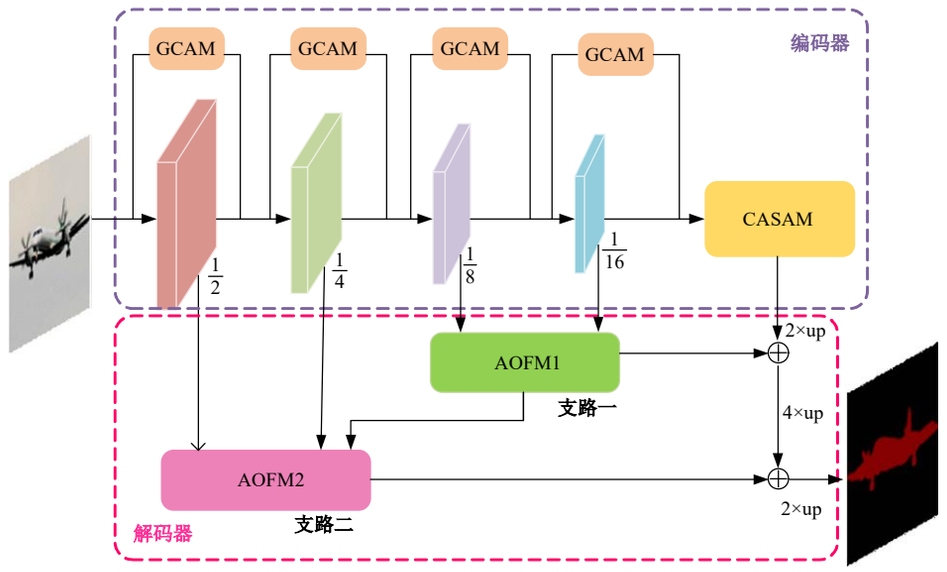


图 1　改进分割网络总体结构

Fig. 1　Improving the overall network architecture diagram

张率($d>1$)。通过上述分解,可降低计算成本,捕获远距离像素依赖关系,并通过点卷积估计每个像素点的重要性,生成相应的注意力图。本文参考VAN[22]中的参数设计,令大核卷积的尺寸为21×21,扩张率$d=3$。GCAM模块可表示为:

$$\text{Attention}=\text{Conv}_{1\times1}(\text{DW-D-Conv}(\text{DW-Conv}(F))) \tag{1}$$

$$F_{\text{A}}=\text{Attention}\otimes F \tag{2}$$

式中:$F\in\mathbb{R}^{C\times H\times W}$为输入特征,$\text{Attention}\in\mathbb{R}^{C\times H\times W}$为生成的注意力;$\otimes$表示逐元素相乘;$F_{\text{A}}\in\mathbb{R}^{C\times H\times W}$为含有注意力的特征图。

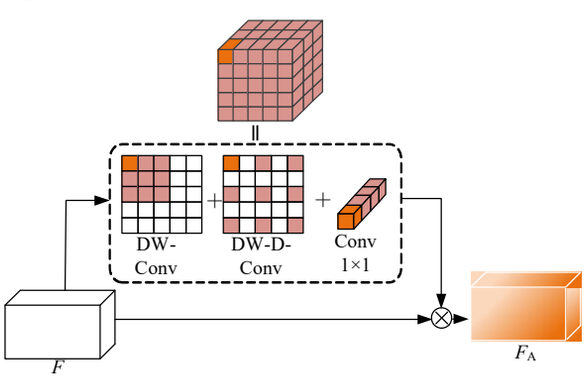


图 2　全局上下文注意力模块

Fig. 2　Global context attention module

为简化计算过程,设置有C个C层的大核卷积,未使用分解策略,直接采用大核卷积计算,参数量计算如下:

$$\text{Param}_1=K\times K\times C\times C \tag{3}$$

使用分解策略,参数量计算计算如下:

$$\text{Param}_2=C^2\times\left[\frac{K}{d}\times\frac{K}{d}+(2d-1)\times(2d-1)+1\right] \tag{4}$$

显然,$\text{Param}_1>\text{Param}_2$,因此,分解策略可减少参数量。

2.1.2　级联自适应尺度感知模块

本文提出的CASAM可自适应地融合深层特征中的多尺度上下文信息,增强多尺度特征之间的依赖性,改善由空洞卷积稀疏性引起的信息丢失问题。模块结构如图3所示。在CASAM中,使用并联的1×1卷积、空洞率分别为6、12、18的3×3空洞卷积、池化获取多尺度特征,使用尺度空间注意力模块自适应地为每个特征图生成注意力,之后选择合适的特征进行融合。具体过程如下:首先,1×1卷积和3×3($r=6$)空洞卷积的输出特征$F_{1\times1}$、$F_{r=6}$按通道叠加,再经过尺度空间注意力模块为每个特征图生成注意力$A_{1\times1}$、$A_{r=6}$,再将加有注意力的特征图$FA_{1\times1}$、$FA_{r=6}$融合,得到第一层融合后的特征图F_1。之后,F_1和3×3($r=12$)空洞卷积的输出特征$F_{r=12}$按通道叠加,重复上述过程,得到第二层融合后的特征图F_2。最后,F_2和3×3($r=18$)空洞卷积的输出特征$F_{r=18}$可得到第三层融合后的特征F_3,F_3和pooling后的输出特征F_{p}可得到第四层融合后的特征F_4,该特征也就是融合后的多尺度特征。

尺度空间注意力模块由卷积和Sigmoid操作

生成注意力,其结构图如图 3 所示。Sigmoid 在空间上为特征图 f_1 和 f_2 的每个点 f_{1i} 和 f_{2i} 生成像素级注意力图 $A_1 \in \mathbb{R}^{H\times W}$,$A_2 \in \mathbb{R}^{H\times W}$,每个像素点注意力 A_{1i},A_{2i} 计算公式如下:

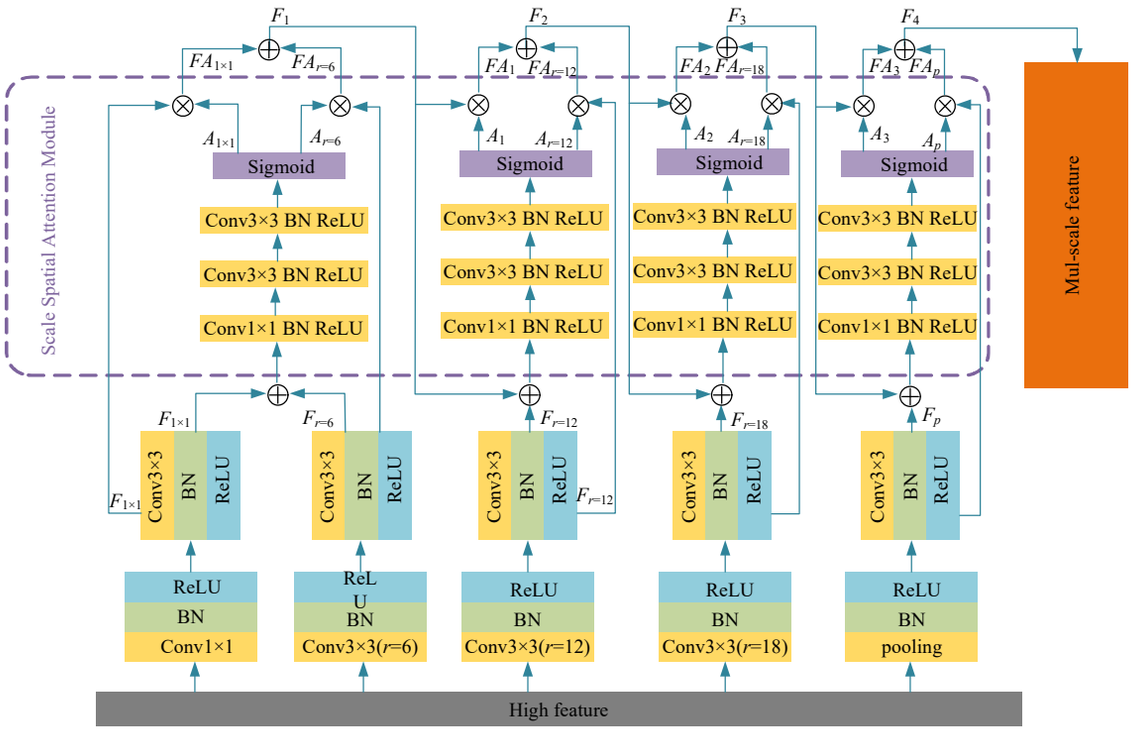


图 3 级联自适应尺度感知模块

Fig. 3 Cascade adaptive scale awareness module

$$A_{1i}=\frac{e^{f_{1i}}}{1+\sum_{i=1}^{H\times W} e^{f_{1i}}+e^{f_{2i}}},\ i=[1,2,\cdots,H\times W] \tag{5}$$

$$A_{2i}=\frac{e^{f_{2i}}}{1+\sum_{i=1}^{H\times W} e^{f_{1i}}+e^{f_{2i}}},\ i=[1,2,\cdots,H\times W] \tag{6}$$

注意力加权融合过程可表示为:

$$F=A_1\otimes f_1+A_2\otimes f_2 \tag{7}$$

式中:$\otimes$ 表示逐元素相乘。

2.2 解码器

解码器将经过编码器处理的输入所得到的特征做进一步特征优化和处理,得到分割图。本文的解码器设计如图 1 所示。将从编码器得到的各层特征作为融合模块的输入,解码过程有两条支路。支路一,第三层和第四层特征经 AOFM1 模块融合,CASAM 模块的输出进行 2 倍上采样,两者输出做跳跃连接得到支路一的输出。支路二,第一层特征、第二层特征和 AOFM1 输出特征经 AOFM2 模块融合,支路一的输出做 4 倍上采样操作,同样,两者输出通过跳跃连接融合,最后将支路二的输出做 2 倍上采样操作,得到分割效果图。

浅层信息和深层信息融合,可以提高网络特征表达能力,有效改善分割效果。简单的通道叠加导致次优特征被融合,且浅层信息中含有无用信息会造成冗余,进而影响分割结果。本文提出的 AOFM1、AOFM2 通过计算输入特征的通道注意力改善上述问题。首先,浅层特征经过 Conv1×1 操作升维,深层特征通过上采样操作得到与浅层特征相同尺寸大小的特征;然后,浅层特征和深层特征分别经过高效通道注意力(Efficient channel attention,ECA)模块计算通道注意力再加权;最后,将添加通道注意力的浅层特征和深层特征按通道拼接的方式融合,获取含有丰富语义信息和空间信息的特征。其结构如图 4 所示,模块中 ECA 是在传统通道注意力机制上的改进,通过不降维的局部跨信道交互策略,可有效提高网络的预测精度。

ECA 操作计算过程可表示为:

$$\mathrm{Atte}_{F_l}=\sigma(\mathrm{GAP}(\mathrm{Conv}_{1\times1}(F_l))) \tag{8}$$

$$\mathrm{Atte}_{F_h}=\sigma(\mathrm{GAP}(\mathrm{up}(F_h))) \tag{9}$$

式中:F_h、F_l 为深层特征和浅层特征;up 为上采样

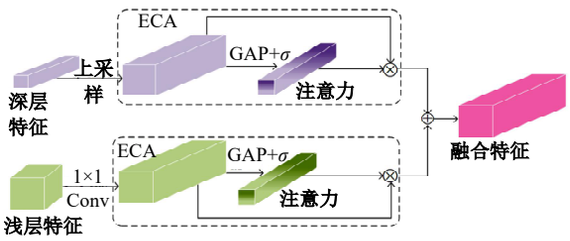


图 4　注意力优化融合模块

Fig. 4　Attention Optimization Fusion Module

操作;$\mathrm{Conv}_{1\times1}$为 1×1 卷积;GAP 为全局平均池化;σ为 Sigmoid 激活函数;Atte_{F_h}、Atte_{F_l}为深层特征和浅层特征的通道注意力。

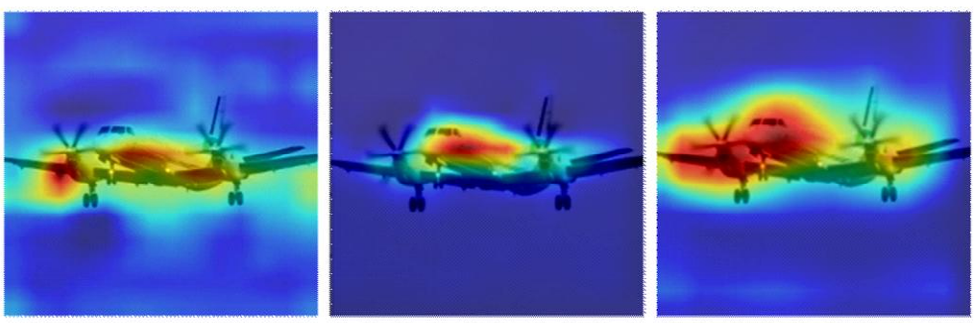
融合过程可表示为:

$$F_{\mathrm{fusion}}=\mathrm{Concat}(F_l\otimes\mathrm{Atte}_{F_l},F_h\otimes\mathrm{Atte}_{F_h}) \tag{10}$$

式中:F_{fusion}为深层特征和浅层特征融合后的特征;Concat 表示按通道叠加方式融合;$\otimes$表示逐通道相乘。

2.3　改进网络注意力及各模块输出可视化

为了对改进算法做进一步解释说明,本小节对算法中的注意力及各模块输出进行了可视化分析,可视化结果如图 5 所示,图中颜色的深浅表示目标区域在网络中获得的注意力高低。改进网络在特征提取的初始阶段加入 GCAM 模块,利用全局注意力对上下文信息加权,加权后的结果图如 5(a)所示,可以看出,GCAM 能够较好地获取以目标"飞机"为中心的全局上下文信息,其中,全局上下文信息中不仅含有目标特征,还包括背景等无用特征。CASAM 通过多尺度注意力机制动态从全局上下文信息中选择目标特征进行融合,可从图 5(b)可以看到,网络更加关注目标特征。如图 5(c)所示,AOFM 利用 ECA 注意力将浅层特征和深层特征加权融合,得到解码后的注意力图。

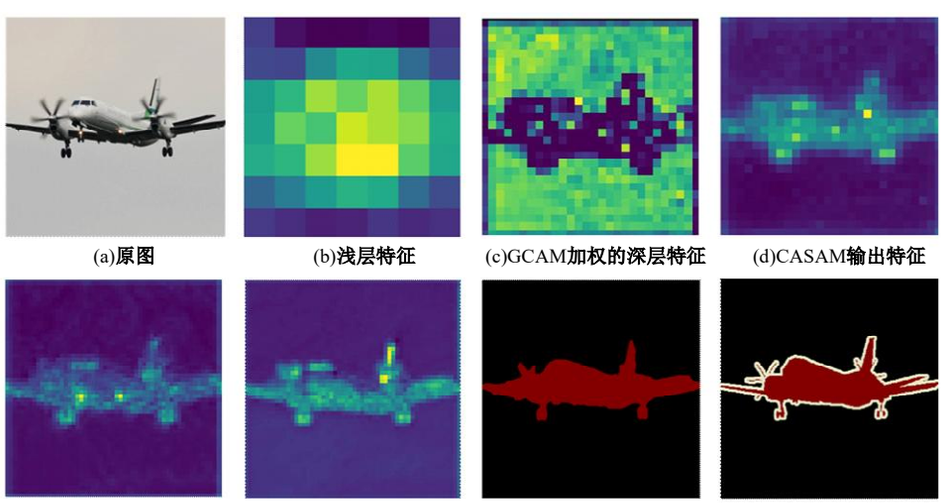


(a)GCAM全局注意力　(b)CASAM优化后的注意力　(c)AOFM解码后的注意力

图 5　改进网络中注意力可视化结果图

Fig. 5　Visualization results of attention in the improved network

各模块输出可视化结果如图 6(b)所示,浅层特征包含颜色、纹理等特征,但其语义性低、含噪声较多。骨干网络 ResNet101 的输出也就是经 GCAM 加权后的深层特征,如图 6(c)所示,能够



(a)原图　(b)浅层特征　(c)GCAM加权的深层特征　(d)CASAM输出特征

(e)未加AOFM输出特征　(f)加有AOFM输出特征　(g)预测结果　(h)标签

图 6　改进网络的特征可视化结果图

Fig. 6　Visualization results of feature maps in the improved network

简单看出图中目标为“飞机”。CASAM的输出特征是对加权GCAM的深层特征进一步的提取、选择得到的,可从图6(d)看到,去除了背景等无用特征,更加关注目标区域,且目标区域连续无孔洞。图6(e)是未加AOFM的解码器输出,是直接融合浅层特征和深层特征经上采样操作得到的,可以看出,解码后的目标边缘不准确,这是由于浅层特征和深层特征含有噪声、背景等次优特征。图6(f)是加有AOFM的解码器输出,与6(e)相比,图6(f)特征表示更加清晰且目标边缘更加准确。

3 实验结果和分析

3.1 实验设计

3.1.1 数据集与参数设置

实验采用语义分割任务中广泛使用的权威数据集Cityscapes[26]和PASCAL VOC2012增强数据集[27]训练网络。

Cityscapes数据集是一个基于19个类别像素级标注的大规模城市景观数据集。该数据集包含5 000张像素级标注的图片和20 000张粗糙标注的图像,每张图片分辨率为1 024×2 048。其中,训练集有精细标注图2 975张,验证集有精细标注图500张,测试集有精细标注图1 525张。

PASCAL VOC2012增强数据集中包含20个物体语义类别和一个背景类。该数据集有10 582张图片做训练集,1 449张图片做验证集,1 456张图片做测试集。数据集中每张图片的分辨率是不同的。

为了提高网络的收敛速度以及训练稳定性,首先将特征提取网络在Imagenet分类数据集上进行预训练,然后使用预训练模型对改进的DeepLabv3+网络进行训练。实验中超参数的设置借鉴了DeepLabv3+网络,batch size大小设置为4,权重衰减weight decay为1×10^{-4},梯度优化采用随机梯度下降策略(Stochastic gradient descent,SGD),初始学习率lr为1×10^{-2},动量momentum为0.9,学习率衰减采用Poly策略。

3.1.2 实验环境

实验操作平台为Ubuntu 18.04的Linux系统、CPU为Intel Xeon(R) CPU E5—2660 V2 @ 2.20 GHz×40,GPU为GeForce GTX 2080Ti,深度学习框架为Pytorch1.8.0, Cuda10.1, Python3.8.5。

3.1.3 评价指标

本文采用语义分割最常用的评价指标平均交并比(Mean intersection over union,MIoU)评估所提方法的性能,值越高,则网络性能越好。MIoU的计算分为两个步骤:首先,计算每个类别真实目标掩膜和预测掩膜的交集、并集的比例,之后再对所有类别的计算结果求平均。计算公式如下:

$$\mathrm{MIoU}=\frac{1}{K+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}\sum_{j=0}^{k}(\mathrm{p}_{ji}-\mathrm{p}_{ii})} \tag{11}$$

式中:K为像素类别;p_{ii}为正确分类的像素;p_{ij}表示真实类别为i、预测类别为j的像素。

3.2 实验结果与性能分析

为了对所提方法进行验证与评估,本文在Cityscapes数据集和PASCAL VOC2012增强数据集上进行了系统实验,并对实验结果进行了分析,以此验证改进网络的有效性和广泛适用性。分割性能评价指标均采用MIoU。

3.2.1 不同骨干网络的比较

骨干网络的性能直接影响到特征的提取,选用合适的骨干网络在很大程度上可提高最终分割预测效果。表1列出了基于不同骨干网络的DeepLabv3+网络及改进后的DeepLabv3+语义分割网络的分割评价结果。

表1实验数据表明:在Cityscapes验证集上,以ResNet101为骨干网络的改进网络分割效果最佳,MIoU为76.2%,与Baseline相比,MIoU提高了2.3%,相比于以MobileNetV2、ResNet18、ResNet50为骨干网络的改进网络,MIoU分别提高了3.7%、1.9%、0.6%,表明ResNet101相较于其他骨干网络具有更强的特征提取能力,原因在于ResNet101是一个很深的网络,可以很好地根据复杂性捕捉特征细节,并使用残差连接解决深层网络梯度爆炸和梯度消失的问题。因此,选用ResNet101作为改进网络的骨干网络提取特征。

3.2.2 消融实验

本文训练的网络模型是基于DeepLabv3+网络改进的,主要在DeepLabv3+网络上加入了GCAM、CASAM、AOFM,用以捕获全局上下文信息,并自适应选择目标特征,去除冗余,增强特征之间的依赖性。为验证各组成模块的有效性,本文在Cityscapes验证集上进行了如下消融实

表 1　不同骨干网络在 Cityscapes 验证集上的比较

Table 1　Comparison of different backbone networks on the Cityscapes validation set

模型	骨干网络	MIoU/%
Baseline	MobileNetV2	72.1
+GCAM+CASAM+AOFM1,2	MobileNetV2	72.5
Baseline	ResNet18	73.2
+GCAM+CASAM+AOFM1,2	ResNet18	74.3
Baseline	ResNet50	73.8
+GCAM+CASAM+AOFM1,2	ResNet50	75.6
Baseline	ResNet101	73.9
+GCAM+CASAM+AOFM1,2	ResNet101	76.2

验,并用 ResNet101 作为消融实验的骨干网络。由表 2 可看出,CASAM 对特征的提取影响最大,与 DeepLabv3+网络相比,MIoU 提高了 1.9%,同样,加入 GCAM、AOFM1、AOFM2 使网络的分割性能在不同程度上得到了提升,尽管 AOFM2 模块对精度的改善仅为 0.1%,但在保持模型参数量基本不变的情况下,仍有助于优化网络的性能。

表 2　改进网络在 Cityscapes 验证集上的消融实验

Table 2　Ablation experiments of the improved network on the Cityscapes validation set

Deep-Labv3+	GCAM	CASAM	AOFM1	AOFM2	参数量/M	MIoU/%
√					52.7	73.9
√	√				58.2	74.2
√	√	√			64.3	75.8
√	√	√	√		66.2	76.1
√	√	√	√	√	67.5	76.2

为进一步直观描述 CASAM 的有效性,在 Cityscapes 测试集上对 CASAM 和 ASPP 的预测结果进行可视化对比,对比结果如图 7 所示。由图 7 可见,CASAM 可以有效融合不同尺度的特征,通过增强多尺度特征之间的关联性,从而改善近处的大目标如路面、行人等分割不连续的问题,并且进一步优化了目标的边缘。

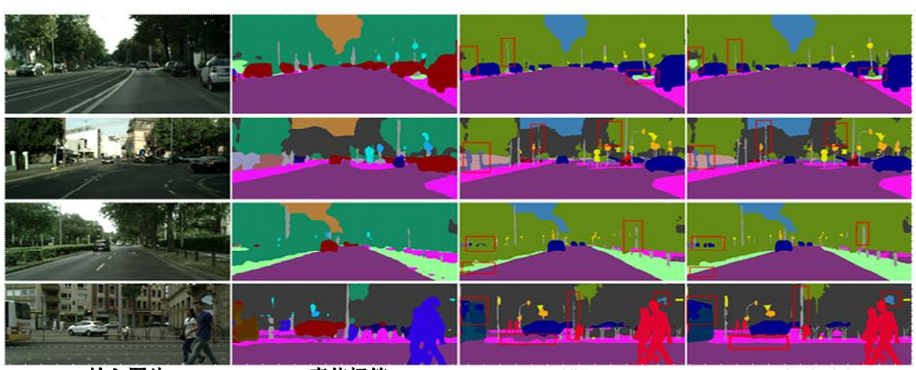


图 7　CASAM 与 ASPP 在 Cityscapes 测试集上的对比

Fig. 7　Comparison of CAMSM and ASPP on the Cityscapes test set

3.2.3　不同算法在 Cityscapes 验证集上的比较

为评价本文网络模型的性能,在 Cityscapes 验证集上与其他先进网络进行了性能比较,如表 3 所示。结果表明:本文所提模型的分割效果优于其他算法,比 SegNet、DANet、SETR-MLA、DeepLabv3+、CCNet、UperNe、OCRNet、Segformer 分别提高了 9.5%、2.3%、0.1%、2%、4.8%、3.7%、2.8%、0.3%,本文算法在精度上虽与 SETR-MLA 基本持平,但参数量却显著减少,综合来看,本文所改进的算法在性能上表现更加优异。此外,表 4 验证了不同算法在 Cityscapes 验证集上的各类别预测结果,所提方法在多个类别上处于优势。与 DeepLabv3+相比,本文方法在识别图像中出现的多尺度目标如“汽车”“火车”时,精度提高了 2.9%、3.1%,这是因为 CASAM 对多尺度特征进行了增强,使网络能够更好地捕捉到目

表 3　不同算法在 Cityscapes 验证集上的对比

Table 3　Comparison of different algorithms on the Cityscapes validation set

算法	骨干网络	参数量/M	MIoU/%
SegNet	VGG16	50.85	66.7
CCNet	ResNet50	60.8	71.4
UperNet	ResNet101	86.4	72.5
OCRNet	HRNet	72.8	73.4
DANet	ResNet50	58.9	73.9
Segformer	MiT	101.2	75.9
SETR-MLA	T-Small	311.8	76.1
DeepLabv3+	ResNet101	52.7	74.2
本文	ResNet101	67.5	**76.2**

标的细节,通过引入AOFM,利用ECA对特征加权,使网络具有更强的解码能力。对语义相似的目标如“行人”和“骑行者”,GCAM的引入使网络能够更好地获取全局上下文信息,从而提高了目标的区分度,精度分别提高了2.4%和2.2%。

为更加直观地展现分割网络的性能,将其他算法和本文所提方法的预测结果进行可视化,可视化结果如图8所示,从上往下依次是输入图片、

表4 不同算法在Cityscapes验证集上的各类别预测结果

Table 4 Prediction results for each category of different algorithms on the Cityscapes validation set

算法	道路	人行道	建筑物	墙体	围栏	交通灯	交通标识	电线杆	植被	地形	天空	行人	骑行者	汽车	卡车	公交车	火车	摩托车	自行车	MIoU/%
SegNet	89.8	75.4	81.7	38.7	58.7	60.3	57.9	59.8	70.1	56.7	80.6	72.4	58.3	89.6	62.3	68.4	67.6	51.3	67.9	66.7
CCNet	91.1	78.8	86.3	50.9	60.2	61.5	66.8	61.4	73.2	58.0	84.4	75.2	59.1	79.2	73.7	83.1	76.3	66.2	71.2	71.4
UperNet	92.4	79.6	87.4	51.2	58.1	62.9	67.4	63.5	73.4	59.1	86.6	76.3	60.3	85.6	77.1	83.1	77.5	64.6	71.4	72.5
OCRNet	93.3	79.4	88.1	51.8	59.6	63.7	68.3	63.7	74.2	58.9	87.3	77.8	61.4	86.9	78.4	84.2	78.6	66.7	72.3	73.4
DANet	93.6	80.1	88.3	47.4	58.9	64.5	69.1	64.5	74.1	60.7	87.6	78.3	60.6	87.9	79.9	86.7	79.2	68.6	73.3	73.9
Segformer	93.8	80.5	88.9	55.4	61.9	66.1	72.6	66.0	78.4	61.3	87.9	79.4	62.7	90.6	82.2	86.9	81.6	69.5	76.4	75.9
SETR-MLA	94.2	80.3	89.6	55.9	62.3	66.4	71.8	64.3	77.5	62.8	88.7	80.2	62.9	90.4	82.6	86.7	81.9	69.4	77.1	76.1
DeepLabv3+	93.8	80.2	88.6	53.2	59.6	64.5	69.5	63.4	76.1	60.4	87.3	78.5	60.3	89.1	79.3	85.6	79.9	67.6	73.7	74.2
本文	94.3	81.1	89.4	56.6	62.6	66.7	71.2	66.7	77.2	62.7	88.2	80.7	62.8	90.8	82.7	87.2	82.3	69.5	75.2	76.2

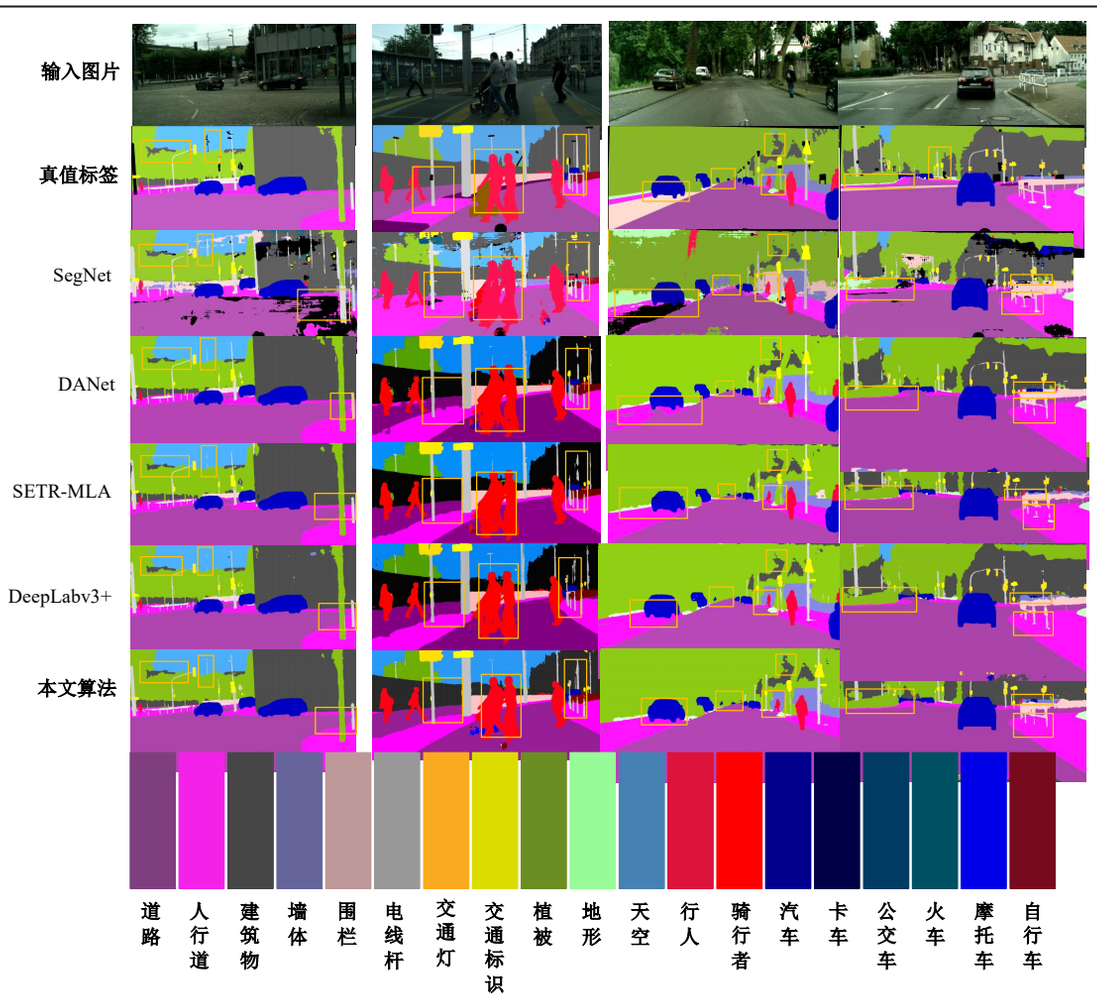


图8 不同算法在Cityscapes验证集上的预测结果可视化

Fig. 8 Visualization of prediction results of different networks on the Cityscapes validation set

真值标签、SegNet、DANet、SETR_MLA、DeepLabv3+以及本文改进网络,关注图中黄色方框圈出的部分,可以看出,相较于DeepLabv3+,本文所提方法可以有效识别图中的小尺度目标如“远处的车”等,对覆盖面积较大的目标的边缘如“树叶”“车”“路面”等可以进行很好的优化,使其孔洞减少,更具完整性,且包含更少的错误,进一步验证了各模块的有效性。

3.2.4　不同算法在PASCAL VOC2012验证集上的比较

为验证本文网络的泛化性,本文在PASCAL VOC2012增强数据集上进行了实验,并比较了不同算法的性能,如表5所示,本文算法精度相对较高,比SegNet、DANet、SETR-MLA、DeepLabv3+、CCNet、UperNe、OCRNet、Segformer分别提升了8.6%、2.4%、0.4%、5.8%、7.2%、6.3%、5%、0.1%,并对比了精度、计算复杂度和模型参数量,本文的改进网络实现了三者之间很好的平衡,验证了本文方法的泛化性能。

为更加直观地对比不同算法在PASCAL VOC2012验证集上的预测效果,图9对预测结果进行可视化。结果表明:在室内复杂场景中,本文方法可有效捕获多类别信息,使大尺度目标分割更加完整,减少错误分割,在一定程度上优化目标边缘。

表5　不同算法在PASCAL VOC2012验证集的对比

Table 5　Comparison of different algorithms on the PASCAL VOC2012 validation set

算法	骨干网络	参数量/M	Flops/G	MIoU/%
SegNet	VGG16	20.32	22.3	70.1
CCNet	ResNet50	47.2	60.1	71.5
UperNet	ResNet101	53.4	64.5	72.4
OCRNet	ResNet101	69.4	68.4	73.7
DANet	ResNet50	10.34	51.4	76.3
SETR-MLA	T-Small	180.62	65.2	78.3
Segformer	MiT	100.4	63.9	78.6
DeepLabv3+	ResNet101	32.7	56.6	72.9
本文	ResNet101	40.2	62.3	78.7

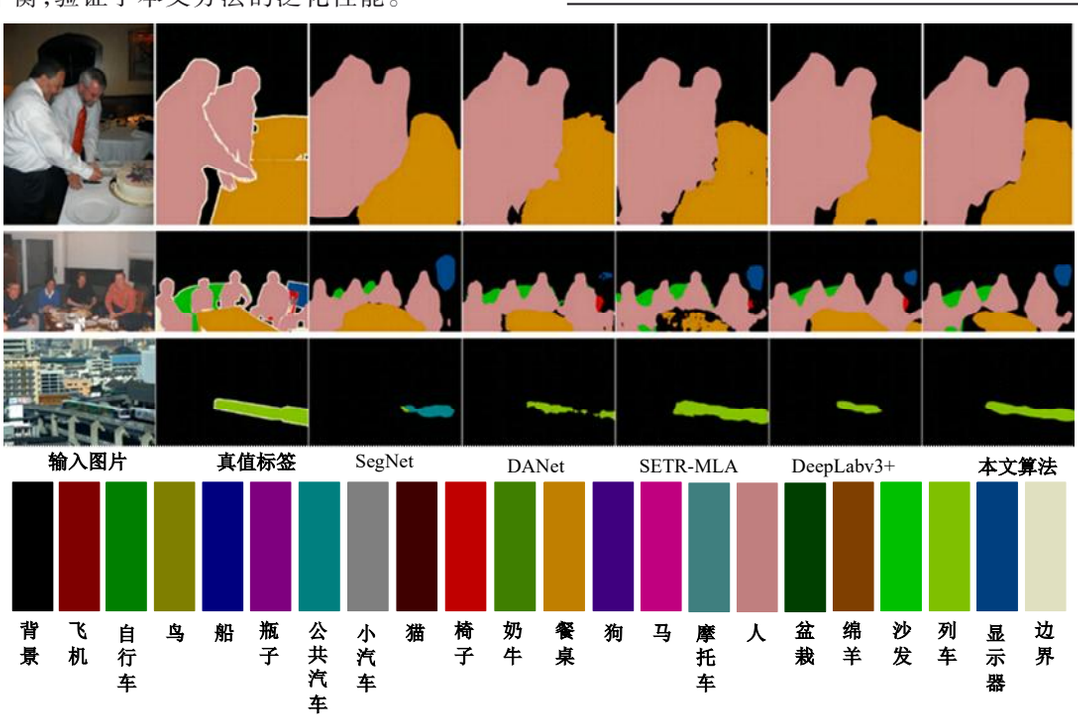


图9　不同网络在PASCAL VOC2012验证集的预测可视化

Fig. 9　Visualization of prediction results of different networks on the PASCAL VOC2012 validation set

4　结束语

本文针对DeepLabv3+中出现的多尺度目标分割错误、多尺度特征图及不同阶段特征图关联性差的问题,通过改进DeepLabv3+的编码器和解码器,提出了一种基于注意力机制和特征融合的语义分割网络。在编码器中,引入全局上下文注意力模块以获取丰富的上下文信息,并将其作为级联自适应尺度感知模块的输入,通过级联连接和空间维度上注意力的计算对多尺度特征进行

建模,增强多尺度特征间的关联性,减少特征提取时造成的冗余。在解码器中,将浅层细节特征、深层语义特征和多尺度融合特征作为解码器的输入,通过注意力优化融合模块自动选择最优特征融合,从而增强像素的语义辨析能力。实验结果表明:本文方法在分割精度上有所提升,并且在计算复杂度、精度和模型参数方面实现了良好的平衡。

参考文献:

[1] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation [C]//Medical Image Computing and Computer-Assisted Intervention-MICCAI: The 18th International Conference, Munich, Germany, 2015: 234-241.

[2] Chen J, Lu Y, Yu Q, et al. Transunet: transformers make strong encoders for medical image segmentation [J/OL]. [2023-07-02]. arXiv preprint arXiv: 2102.04306v1.

[3] Zhao T Y, Xu J D, Chen R, et al. Remote sensing image segmentation based on the fuzzy deep convolutional neural network[J]. International Journal of Remote Sensing, 2021, 42(16): 6264-6283.

[4] Yuan X H, Shi J F, Gu L C. A review of deep learning methods for semantic segmentation of remote sensing imagery[J]. Expert Systems with Applications, 2021, 169: No. 114417.

[5] Xu Z Y, Zhang W, Zhang T X, et al. Efficient transformer for remote sensing image segmentation[J]. Remote Sensing, 2021, 13(18): No. 3585.

[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] Yu C, Gao C, Wang J, et al. Bisenet v2: bilateral network with guided aggregation for real-time semantic segmentation[J]. International Journal of Computer Vision, 2021, 129: 3051-3068.

[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, Boston, USA, 2015: 3431-3440.

[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, 2017, 40(4): 834-848.

[10] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation [J/OL]. [2023-07-03]. arXiv preprint arXiv: 1706.05587v3.

[11] 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), Munich, Germany, 2018: 833-851.

[12] Wang J, Sun K, Cheng T, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(10): 3349-3364.

[13] Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 10012-10022.

[14] Wang W, Xie E, Li X, et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 568-578.

[15] Zheng S, Lu J, Zhao H, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 6881-6890.

[16] Xie E, Wang W, Yu Z, et al. SegFormer: simple and efficient design for semantic segmentation with transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 12077-12090.

[17] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J/OL]. [2023-07-04]. arXiv preprint arXiv: 2010. 11929v2.

[18] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 2881-2890.

[19] Hou Q, Zhang L, Cheng M M, et al. Strip pooling: rethinking spatial pooling for scene parsing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 4003-4012.

[20] 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, Honolulu, USA, 2017: 4353-4361.

[21] Ding X, Zhang X, Han J, et al. Scaling up your kernels to 31×31: revisiting large kernel design in CNNs [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 11963-11975.

[22] Guo M H, Lu C Z, Liu Z N, et al. Visual attention network[J/OL]. [2023-07-04]. arXiv preprint arXiv: 2202.09741.

[23] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 7132-7141.

[24] Wang Q, Wu B, Zhu P, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 11534-11542.

[25] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 3146-3154.

[26] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 3213-3223.

[27] Everingham M, Eslami S M A, Van Gool L, et al. The pascal visual object classes challenge: a retrospective[J]. International Journal of Computer Vision, 2015, 111: 98-136.

[28] 王雪, 李占山, 吕颖达. 基于多尺度感知和语义适配的医学图像分割算法[J]. 吉林大学学报: 工学版, 2022, 52(3): 640-647.
Wang Xue, Li Zhan-shan, Lyu Ying-da. Medical image segmentation algorithm based on multi-scale perception and semantic adaptation [J]. Journal of Jilin University(Engineering and Technology Edition), 2022, 52(3): 640-647.