ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 6,Nov. 2024,pp. 1505－1516
DOI:10. 16337/j. 1004-9037. 2024. 06. 018

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

多级注意力特征优化的道路场景实时语义分割

张　鹏，彭宗举，张文瑞，罗英国，韦　玮，王培容

（重庆理工大学电气与电子工程学院，重庆 400054）

摘　要：针对复杂多变道路场景下目标重叠导致图像边缘难以分割、小目标特征提取困难等问题，提出一种多级注意力特征优化的道路场景实时语义分割方法。首先，设计深度残差注意力模块，考虑不同层级下特征权重的差异性，通过压缩注意力机制来优化图像局部特征，从而改善像素之间的边缘效应；然后，设计通道注意力和深度聚合金字塔池化模块进一步加强语义上下文信息的提取，小目标信息丢失问题得到了改善；最后，设计注意力融合模块自上而下地融合不同尺度下的特征信息，实现全局特征信息下的有效交互，增强网络对重要特征的表达。Cityscapes和CamVid道路场景数据集上进行的实验测试分别达到74.4％和67.7％的分割精度，138帧/s和148帧/s的推理速度。与近几年其他优秀方法相比，该方法改善了图像边缘信息丢失，优化了对图像中小目标的分割准确度。

关键词：道路场景；语义分割；空洞卷积；注意力机制；特征融合

中图分类号：TP391.41　　**文献标志码：**A

Real-Time Semantic Segmentation of Road Scene Based on Multi-level Attention Feature Optimization

ZHANG Peng, PENG Zongju, ZHANG Wenrui, LUO Yingguo, WEI Wei, WANG Peirong

(School of Electrical and Electronic Engineering, Chongqing University of Technology, Chongqing 400054, China)

Abstract: Aiming at the problems of overlapping targets in complex and changeable road scenes, it is difficult to segment image edges and extract small target features. A multi-level attention feature optimization method for real-time semantic segmentation of road scenes is proposed. Firstly, a lightweight residual attention module is designed, taking into account the difference in feature weights at different levels, and optimizing local features of the image through a compressed attention mechanism, thereby improving the edge effect between pixels. Then, the channel attention and depth aggregation pyramid pooling module are designed to further strengthen the extraction of semantic context information, thereby solving the problem of small target information loss. Finally, the attention fusion module is designed to fuse feature information at different scales from top to bottom. It can achieve effective interaction of global feature information and enhance the network's expression of important features. Experimental tests are carried out on the Cityscapes and CamVid road scene datasets, and the segmentation accuracy is 74.4% and 67.7%, respectively, and the inference speed are 138 frames/s and 148 frames/s. Compared with the excellent methods in recent years, this method improves the loss of image edge information and optimizes the segmentation accuracy of small objects in the image.

基金项目：国家自然科学基金(62371081)；重庆市自然科学基金(cstc2021jcyj-msxmX0411,CSTB2022NSCQ-MSX0873)。
收稿日期：2023-10-19；**修订日期：**2024-03-28

Key words: road scene; semantic segmentation; hole convolution; attention mechanism; feature fusion

引　言

语义分割是一项基本的计算机视觉任务,在许多实际应用中起着至关重要的作用,例如医学图像分割、自动驾驶和机器人等[1-2],尤其在解析复杂多变的道路场景下具有重要的研究价值[3]。随着全卷积网络[4]在图像语义分割领域的逐步推广应用,一系列新颖的网络模型相继被提出。然而,许多性能优异的语义分割模型结构极其复杂繁琐,不适合部署在计算资源有限且低延时的移动设备平台上。因此,如何在交通道路场景下进行快速准确的语义分割面临新的挑战。

随着近年无人驾驶的快速发展,交通道路场景下移动设备部署的需求不断增加,实时语义分割越来越备受关注,许多基于轻量级的卷积神经网络的实时语义分割模型[5-6]被提出,以解决精度和推理速度之间的平衡问题。首次考虑卷积神经网络效率问题的研究工作是Paszke等[7]提出的高效神经网络(Efficient neural network, ENet)模型,通过设计出较为紧凑的编解码结构,有效提升了图像语义分割效率,但该模型的感受野太小,无法捕获小目标。为了收集多尺度的上下文信息,Mehta等[8]提出高效空间金字塔网络(Efficient space pyramid network, ESPNet)模型,并采用卷积分解策略,在参数量相似的情况下,ESPNet与Enet相比取得了更高的分割精度。但此类基于传统编解码结构的实时语义分割网络模型,往往通过删减复杂的解码结构来实现网络轻量化,从而导致图像纹理细节信息丢失严重。

近年来,双分支网络结构在实时语义分割中得到了广泛应用,Li等[9]将深度可分离卷积和空洞卷积相结合,通过设置不同的空洞率以提取多尺度的图像语义信息。Yu等[10]提出一种双分支语义分割网络(Bilateral segmentation network, BiSeNet)模型,分别提取空间细节信息和语义上下文信息。文献[11]在BiSeNet的基础上优化了不同特征的融合方法,进一步提升了分割效果。Li等[12]提出一种特征重用策略,将Xception模型[13]堆叠3次,以扩大感受野和增强特征交互表达。Poudel等[14]在双分支结构的基础上引入编解码结构,在编码器阶段通过学习下采样模块来融合双分支特征。Zhao等[15]设计了一种多分辨率分支的图像级联网络,以促进局部和上下文信息的交互。Dong等[16]设计了一种高效紧凑的交互式双分支网络,采用交互模块来融合不同阶段的特征信息。上述基于双分支结构的实时语义方法忽略了图像像素间的边缘效应,导致图像小目标的边缘信息丢失。

道路场景下目标多样繁杂,为了提升对图像中小目标的分割效果,目前学者们提出了基于注意力机制的双分支卷积神经网络语义分割模型,注意力机制可进一步对提取到的特征进行校正,以更好地保留有价值的特征信息。Hu等[17]提出一种学习通道注意力的有效机制,在不增加模型复杂度的同时取得了良好的语义分割性能。Gao等[18]通过在网络的不同阶段中引入通道注意力和空间注意力模块,有效地提高了语义特征的表征能力。此外,Gao等[19]通过引入特征聚合模块有效地实现了不同分支间的特征信息交互。Lu等[20]设计了多特征融合网络,采用注意力引导融合不同分支间的特征信息。然而,上述方法未能充分考虑不同层级下特征权重的差异性,对小目标的特征信息提取仍不理想。

针对复杂多变道路场景下目标重叠导致图像边缘难以分割、小目标特征提取困难等问题,本文在双分支网络结构基础上,提出一种多级注意力特征优化的道路场景实时语义分割方法(Real-time semantic segmentation network of road scene based on multi-level attention feature optimization, MANet),通过设计多级注意力特征优化模块,并考虑不同层级下特征权重的差异性,可进一步提升复杂道路场景小目标的分割精度。本文方法的主要贡献如下:(1)构建MANet模型,引入多级注意力机制来优化道路场景图像的局部特征和全局特征,在特征提取阶段设计了压缩注意力,在特征加强阶段设计了通

道注意力和深度聚合金字塔池化,在特征融合阶段设计了注意力融合,并充分利用双分支结构的特征表达能力,提高了复杂道路场景下目标重叠时对小目标的分割鲁棒性;(2)设计了一种深度残差注意力模块,将深度可分离卷积和空洞卷积相结合,设置不同的空洞率以逐渐增大感受野,并引入压缩注意力子模块,采用局部、全局加权机制来提取更多的显著性特征,改善图像像素间的边缘效应,捕获图像中多层次的局部特征信息,增强网络对重要特征的表达。

1　本文方法

1.1　网络结构设计

本文提出的基于多级注意力特征优化的道路场景实时语义分割网络采用双分支结构,总体架构如图1所示,主要包括浅层特征提取阶段、空间细节分支(Spatial detail branch, SDB)、语义信息分支(Semantic information branch, SIB)和注意力融合模块(Attention fusion module, AFM)。

在浅层特征提取阶段,首先通过共享的3个3×3卷积提取输入图像特征,并将第一个卷积步长S设置为2以获取输入图像1/2分辨率特征图,由于浅层特征提取阶段只进行一次下采样操作,道路场景原始图像的空间信息得到了有效保留。

在双分支结构中,通过SDB和SIB分别提取空间信息和语义信息,可实现不同分支间的信息交互,便于后续的特征融合。SDB是一个特征图分辨率高、网络层数较浅且简单的分支,由3个3×3卷积层组成,为了获取更多的特征,将中间卷积层通道数量进行扩充,该操作可以去除冗余并增强空间信息特征的提取。SIB是一个特征图分辨率低、网络层数较深且复杂的分支,主要由通道注意力模块(Channel attention module, CAM)、深度残差注意力模块(Depth residual attention module, DRA)和深度聚合金字塔池化模块(Deep aggregation pyramid pool module, DAPPM)组成。

在SIB分支中,首先引入CAM来强调需要突出显示的特征,同时抑制干扰噪声;然后采用本文设计的DRA模块提取更精细的语义上下文信息;为了聚合不同区域的上下文信息,在特征提取阶段后引入DAPPM,从而提高模型对全局信息的捕获能力并保存图像边缘特征。另外,该分支分别引入两次下采样和上采样操作对特征图进行处理,以改变不同阶段特征图的分辨率大小,保证在提取语义特征过程中去除图像冗余信息。

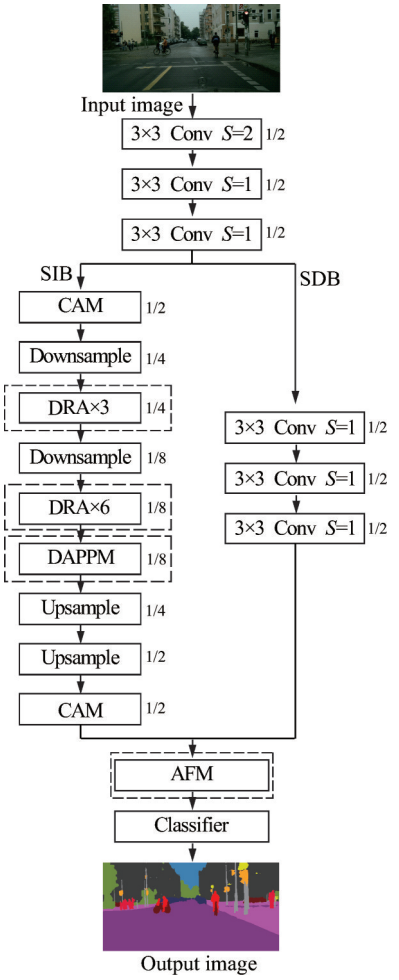

图1　网络总体架构

Fig.1　Network architecture

在注意力融合模块,传统级联方式不可避免地忽略了特征间的相关性,为了更好地融合不同分支间的特征,本文采用AFM模块自上而下地引导空间信息和上下文信息的有效融合。该模块可充分融合两个分支的特征,并在通道和空间上自适应地突出全局特征信息,最后在分类器上预测输出精细的道路场景分割图像。

1.2　深度残差注意力模块

目前,常用的残差特征提取模块包括一维非瓶颈模块(Non-bottleneck-1D)和深度非对称瓶颈模块

(Depth-wise asymmetric bottleneck, DAB),分别如图2(a,b)所示,图中C为通道数,r为空洞卷积中的空洞率。Non-bottleneck-1D采用一维分解卷积来加速和降低参数量,但没有考虑到多尺度信息的获取,对小目标可能会出现分类错误。DAB采用深度可分离卷积和空洞卷积设计成双分支结构,以残差累计相加(Adding, Add)的方式连接,但同时也未充分考虑上下文信息之间的关联性,导致图像细节边缘信息易于丢失。

为了克服Non-bottleneck-1D和DAB的不足,本文设计了深度残差注意力模块DRA,如图2(c)所示,将深度可分离卷积和空洞卷积间的连接方式设计成串联形式,可有效提升特征信息的表达。本文方法充分考虑了不同层级下特征权重的差异性,在SIB分支的1/4和1/8分辨率特征图中,分别采用3个空洞率为2和6个空洞率序列为{4,4,8,8,16,16}的DRA模块,逐渐增大感受野,增强对图像细节特征的表示,这种结构设计不仅有利于提取局部特征、增强像素间的上下文联系,还能以较少的参数量在浅层获得与深层网络相似的上下文信息。同时,引入压缩注意力(Squeeze-and-attention, SA)子模块[21],通过采用局部相乘(Multiply, Mul)、全局加权机制来提取更多的显著性特征,并建立通道上下文依赖关系,以改善图像像素间的边缘效应,捕获复杂多变道路场景下目标重叠时多层次的图像局部特征信息。

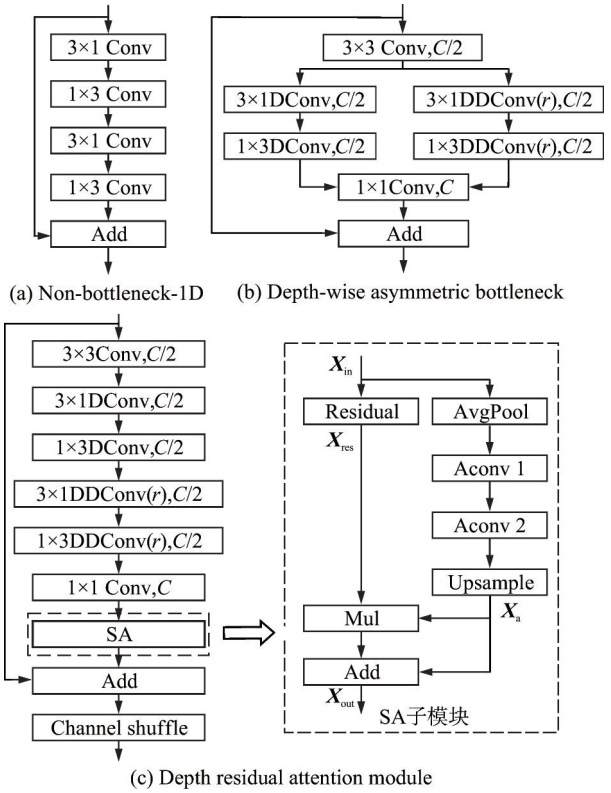

图2 不同类型模块的比较

Fig.2 Comparison of different types of modules

在DRA模块中,首先,对于输入特征图,使用3×3卷积将通道数降为原来的一半。然后,对降维后的特征图依次执行3×1和1×3的深度可分离卷积和深度空洞卷积,将两者相结合以更大程度减少参数量,并在不同网络层采用不同大小空洞率的空洞卷积,可扩大感受野,以获取多尺度的上下文信息。最后,采用1×1卷积恢复通道数,并引入SA子模块来加强特征表示,同时采用通道混洗来增强信息交互,图像卷积后得到的特征图中的各个通道通常表达着不同的特征,这些特征对语义分割的影响不同,若各通道都保持相同的权重,不能体现不同通道之间的依赖关系,不利于小目标特征信息的提取。

SA子模块如图2(c)所示。首先,对于输入特征图$\boldsymbol{X}_{\text{in}}$,使用全局平均池化压缩每个通道的二维特征;然后,设计一种注意力卷积层(Aconv)来对通道特征分配权重,并设计了额外的残差路径学习权重,以重新校准输出特征$\boldsymbol{X}_{\text{out}}$的通道,具体为

$$\boldsymbol{X}_{\text{out}}=\boldsymbol{X}_{\text{a}}\times\boldsymbol{X}_{\text{res}}+\boldsymbol{X}_{\text{a}} \tag{1}$$

式中:$\boldsymbol{X}_{\text{a}}$和$\boldsymbol{X}_{\text{res}}$计算方法分别为

$$\boldsymbol{X}_{\text{a}}=\text{Up}(P_{\text{Aconv}}(P_{\text{Aconv}}(P_{\text{APool}}(\boldsymbol{X}_{\text{in}})))) \tag{2}$$

$$\boldsymbol{X}_{\text{res}}=P_{\text{Conv}}(\boldsymbol{X}_{\text{in}}) \tag{3}$$

式中:P_{APool}表示全局平均池化,P_{Aconv}表示通道注意卷积,Up(•)表示上采样,P_{Conv}表示残差卷积,$\boldsymbol{X}_{\text{a}}$表

示注意卷积通道的输出,$\boldsymbol{X}_{\text{res}}$表示残差卷积的输出。

本文设计的DRA模块,通过引入SA子模块,对特征映射通道进行选择性加权,可有效地提高残差模块的表征能力。同时,由于SA对空间信息未完全压缩,保留了图像细节特征,更适用于层数较少、分割精度较高的轻量级网络。

1.3　深度聚合金字塔池化模块

为更好地捕捉图像中不同尺度的语义信息,受DDRNet[22]启发,本文采用DAPPM模块进一步从低分辨率特征图中提取上下文信息,特征图的高度为H,宽度为W,结构如图3所示。

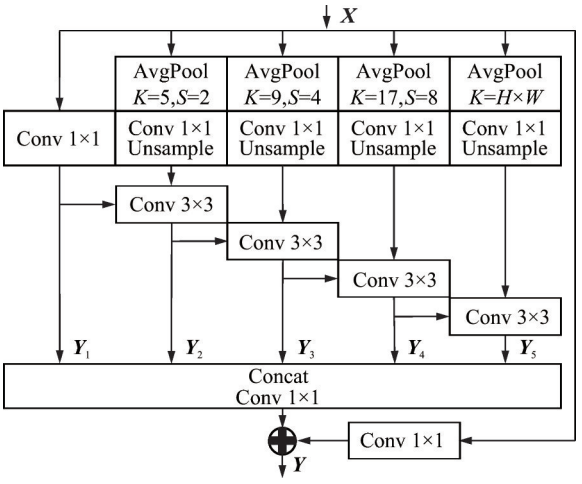

图3　深度聚合金字塔池化模块

Fig.3　Deep aggregation pyramid pool module

通过不同尺度大小的池化操作对特征图进行处理,得到不同尺度下的特征信息。考虑到单个3×3或1×1卷积混合所有多尺度上下文信息不够充分,本文首先对特征图进行上采样,然后设计更多的3×3卷积,以分层残差的方式融合不同尺度的上下文信息。若输入特征图为$\boldsymbol{X}$,不同尺度下的输出$\boldsymbol{Y}_i$可写成

$$\boldsymbol{Y}_i=\begin{cases}C_{1\times1}(\boldsymbol{X}) & i=1\\ C_{3\times3}(\text{Up}(C_{1\times1}(P_{2^i+1,2^{i-1}}(\boldsymbol{X})))+\boldsymbol{Y}_{i-1} & 1<i<n\\ C_{3\times3}(\text{Up}(C_{1\times1}(P_{\text{APool}}(\boldsymbol{X})))+\boldsymbol{Y}_{i-1} & i=n\end{cases}\tag{4}$$

式中:n表示尺度大小,$C_{1\times1}$表示1×1卷积,$C_{3\times3}$表示3×3卷积,Up表示上采样,$P_{j,k}$表示核大小为j和步长为k的池化层,P_{APool}为全局平均池化。最后,通过1×1卷积将所有特征映射连接并压缩,同时在输入端添加一个1×1卷积,同时在输入端添加一个1×1卷积进行通道合并(Concatenate, Concat),以进一步优化输出特征。

1.4　注意力融合模块

如何有效地融合不同分支间特征信息是双分支网络结构中的关键问题,常用的特征融合方式主要有Concat或Add,但这些方式忽略了各层特征之间的位置差异性和信息多样性,同时也未考虑像素间的关联,易导致分类错误,从而降低算法的分割性能。本文设计了注意力融合模块来高效融合双分支的局部特征和全局特征,结构如图4所示。

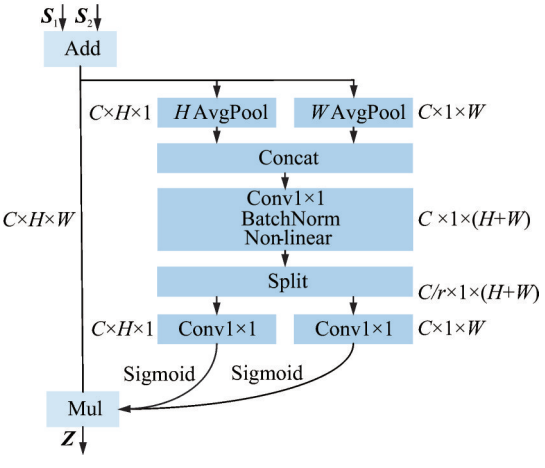

图4　注意力融合模块

Fig.4　Attention fusion module

AFM将不同尺度的特征信息聚合,同时保留全局和局部特征信息,在不引入过多计算量的同时可有效提升分割性能。首先通过全局平均池化在两个维度上进行编码[23],一个维度获得丰富的上下文信息,另一个维度保留空间位置信息,具体为

$$\boldsymbol{S}^{H,W}=C_{1\times1}([f^{H}_{\text{Avg}}(\boldsymbol{S}_1+\boldsymbol{S}_2),f^{W}_{\text{Avg}}(\boldsymbol{S}_1+\boldsymbol{S}_2)])\tag{5}$$

然后,对生成的特征图进行通道缩减,利用缩减因子r来降低通道维度,生成注意力权重,具体为

$$T^{H,W}=\delta(C_{1\times1}(S^{H,W})) \tag{6}$$

式中:$C_{1\times1}$表示1×1卷积,[·]表示级联操作,$f_{\mathrm{Avg}}^{H}(\cdot)$和$f_{\mathrm{Avg}}^{W}(\cdot)$分别表示不同方向的池化,$\delta$为Sigmoid处理,$T$为生成的注意力权重。

最后,AFM的输出Z计算方式如下

$$Z=(S_1+S_2)\times T^{H}\times T^{W} \tag{7}$$

式中S_1和S_2为不同尺度的输入。通过这种方式,AFM可以充分融合两个分支的特征,并同时在通道和空间维度上自适应地突出特征信息。

2 实验结果与分析

2.1 数据集及评价指标

本文采用Cityscapes和CamVid两个常用道路场景数据集进行实验并分析结果,验证算法的有效性。Cityscapes是一个大型城市街道场景数据集,被广泛应用于语义分割领域。本方法采用5 000张精细注释图像进行训练、验证和测试,数量分别为2 975、500和1 525,包含19个类别。CamVid是一个驾驶汽车的街景数据集,包含11个类别和701个精细注释图像,这些图像被分为367个训练样本,101个验证样本和233个测试样本。

本文采用平均交并比(Mean intersection over union, MIoU)和每秒处理帧数(Frames per second, FPS)作为精度和推理速度的度量指标,同时采用参数量(Params)来评估模型大小。此外,受LEANet[24]启发,本文设计了一个平衡权重I_i来衡量网络的性能,其表示如下

$$I_i=\frac{a\times M_i^{*}+b\times F_i^{*}}{P_{\mathrm{avg}}^{*}} \tag{8}$$

$$M_i^{*}=\frac{M_i}{M_{\max}},\quad F_i^{*}=\frac{F_i}{F_{\max}},\quad P_i^{*}=\frac{P_i}{P_{\min}} \tag{9}$$

$$a+b=1,\quad a=\frac{F_{\max}^{*}-F_{\min}^{*}}{M_{\max}^{*}-M_{\min}^{*}}\times b \tag{10}$$

式中:M_i、F_i和P_i分别为模型的MIoU、FPS和Params的值,i为不同模型的索引值,$M_{\max}$、$F_{\max}$和$P_{\min}$分别为M_i、F_i和P_i的最优值,M_i^{*}、F_i^{*}和P_i^{*}分别为M_i、F_i和P_i与$M_{\max}$、$F_{\max}$和$P_{\min}$的比值大小,$M_{\max}^{*}$、$M_{\min}^{*}$和$F_{\max}^{*}$、$F_{\min}^{*}$为对应的最大最小值,P_{avg}^{*}为P_i^{*}的平均值,a和b为计算得到的权重。

2.2 实验设置

本文实验环境基于Pytorch1.11.0,Python3.7.9,在单个GTX 4090 GPU上进行实验。实验设置如下:采用小批次随机梯度下降优化器SGD优化网络,批大小为8,最大训练Epoch为1 000,动量和权重衰减分别设置为0.9和1e−4,学习率采用“poly”策略,每次迭代后自适应学习率调整如下

$$\mathrm{lr}=\mathrm{lr}_{\mathrm{init}}\times\left(1-\frac{\mathrm{iter}}{\mathrm{iter}_{\max}}\right)^{\mathrm{power}} \tag{11}$$

式中:lr为每次迭代后的学习率,$\mathrm{lr}_{\mathrm{init}}$为初始学习率,iter为当前迭代索引,$\mathrm{iter}_{\max}$为每个epoch中的最大迭代次数,power为动量,初始学习率设置为4.5e−2。

关于数据增强,在训练时对输入图像使用随机水平翻转、均值衰减和随机裁剪,随机值为{0.75, 1.0, 1.25, 1.5, 1.75, 2.0}中的数。将Cityscapes数据集随机裁剪为512×1 024分辨率进行训练。

2.3　实验分析

2.3.1　算法消融实验

为了验证各个模块的性能,采用MIoU和FPS为评价标准对各模块分别进行实验对比分析,实验结果如表1所示。包括SIB、SDB和特征融合部分,其中SIB主要有CAM和DAPPM,SDB主要是考虑额外添加的3个3×3卷积,特征融合部分主要有Cat、Add和AFM。此外,由于本文的特征提取模块DRA采用了空洞卷积,为进一步验证不同空洞率设置下的影响,另添两组对比实验。

表1　在Cityscapes数据集上消融实验结果

Table 1　Results of ablation experiment on Cityscapes dataset

模型	SIB		SDB	特征融合			MIoU/%	FPS/(帧·s^{-1})
	CAM	DAPPM		Cat	Add	AFM		
MANet						√	72.9	163
MANet			√			√	73.1	155
MANet	√		√			√	73.4	147
MANet		√	√			√	73.6	144
MANet	√	√				√	74.1	149
MANet	√	√	√	√			73.6	146
MANet	√	√	√		√		72.4	152
MANet	√	√	√			√	74.4	138
MANet {r=4,4,4,4,4,4}	√	√	√			√	70.9	132
MANet {r=3,3,7,7,13,13}	√	√	√			√	73.0	141
MANet {r=4,4,8,8,16,16}	√	√	√			√	74.4	138

首先,仅采用本文设计的DRA模块提取图像局部特征并进行特征融合,得到72.9%的精度和163帧/s的速度。其次,在此基础上引入SDB中的额外卷积层提取更加丰富的全局信息,并融合双分支的特征信息,最终得到73.1%的精度和155帧/s的速度。然后,分别考虑引入CAM和DAPPM,优化对局部特征通道和全局特征信息的表达,仅引入CAM时,得到73.4%的精度和147帧/s的速度,仅引入DAPPM时,得到73.6%的精度和144帧/s的速度。最后,在引入CAM和DAPPM的基础上,去除SDB中卷积层的影响,得到74.1%的精度和149帧/s的速度。

为更直观地显示各模块的性能,图5展示了加入不同模块的道路场景可视化分割图。可见,当在SDB中加入额外卷积后,图像的全局信息表达更加准确;加入CAM后,当图像中出现相似目标时,可减

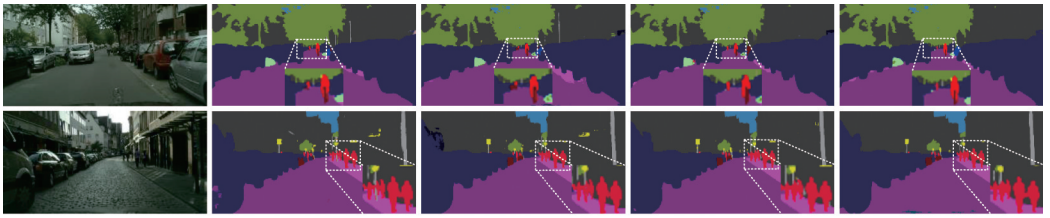

图5　引入各模块的结果对比

Fig.5　Comparison of results of introducing various modules

少目标之间的互相干扰,避免出现分类错误;加入DAPPM后,当图像中出现目标重叠时,通过获取多尺度的特征信息,一些遮挡的小目标可以在一定程度上正确分割出来。

在特征融合阶段,为探索AFM的影响,设计两组对比实验,采用常用的特征融合方式Cat和Add。采用Cat的融合方式,得到73.6% 的精度和146帧/s的速度,采用Add的融合方式,得到72.4% 的精度和152帧/s的速度。本文采用的AFM融合方式可达到74.4% 的精度和138帧/s的速度,由此可见,AFM考虑了不同分支特征之间的差异性和多样性,分割精度更高,且无需牺牲过多推理速度。

在MANet特征提取阶段,1/8特征图分辨率时采用6个空洞率序列为{r=4,4,8,8,16,16}的DRA模块,为验证该设计的有效性,另添加两组对比实验,空洞率分别设置为固定空洞率{r=4,4,4,4,4,4}和互质空洞率{r=3,3,7,7,13,13}。采用固定空洞率{r=4,4,4,4,4,4}的DRA模块,得到70.9% 的精度和132帧/s的速度。采用互质空洞率{r=3,3,7,7,13,13}的DRA模块,得到73.0% 的精度和141帧/s的速度。由此可见,固定空洞率受其感受野大小的限制,不利于捕获多尺度的特征信息,且采用多个相同空洞率的卷积会产生网格效应,降低分割准确性。而互质空洞率相比本方案感受野依然较小,不利于提取更大范围的上下文信息。

经上述实验研究证明,MANet中各模块的设计对模型整体的分割精度都有所提升,且提升了模型的鲁棒性,充分验证了MANet网络结构设计的有效性和合理性。

2.3.2 在Cityscapes数据集的性能分析

为验证本文方法的有效性,在Cityscapes道路场景数据集上将MANet与其他先进的实时语义分割算法进行对比分析,结果如表2所示。实验结果表明,由于MANet引入了多级注意力机制特征优化,并在不同尺度下实现特征交互,MANet在分割精度方面达到最优,且推理速度也具有较为优异的性能,有效地实现了精度和速度之间的平衡,尽管CIDNet速度最快,但其在精度方面比MANet低0.9%,特征表示能力不如MANet。Enet和ESPNet在模型参数上表现最优,但在解码阶段丢失大量图像边缘细节信息,分割精度和速度都明显低于MANet。此外,另外几种在参数量上低于MANet的对比方法,它们在分割精度和速度上的表现都不如MANet,整体性能和MANet存在差异。综合3个性能指标和计算得到的平衡权重进行对比分析,本文方法MANet在保持模型轻量的同时兼顾了分割精度和推理速度,在准确度和实时性之间取得了有效平衡。

图6展示了DABNet、FBSNet和MANet在Cityscapes道路场景数据集上的主观对比效果。从第1行的分割结果可以看出,MANet引入压缩注意力优化图像边缘特征表达后,对道路两边的电线杆能

表2 不同算法在Cityscapes数据集上的对比结果

Table 2 Comparison results of different algorithms on Cityscapes dataset

模型	输入尺寸	预训练	MIoU/%	FPS/(帧·s^{-1})	参数量/MB	权重I_i
ENet[7]	512×1 024	无	58.3	135	0.36	0.14
ESPNet[8]	512×1 024	无	60.3	155	0.36	0.15
DABNet[9]	512×1 024	无	70.1	104	0.76	0.15
FBSNet[19]	512×1 024	无	70.9	90	0.62	0.15
LEANet[24]	512×1 024	无	71.9	109	0.74	0.16
MSCFNet[18]	512×1 024	无	71.9	50	1.15	0.14
MFNet[20]	512×1 024	无	72.1	116	1.34	0.16
BiSeNet-V2[11]	512×1 024	无	72.6	156	3.40	0.16
CIDNet[16]	512×1 024	无	73.5	164	6.50	0.17
MANet	512×1 024	无	74.4	138	5.58	0.17

够清晰准确地分割出来;而DABNet和FBSNet由于缺少对图像中边缘特征的表达,分割结果中出现了分割错误或无法分割,表明本文方法对近距离小目标的分割效果更加准确。从第2行和第3行的分割结果可以看出,MANet对远处的红绿灯能够准确地分割出来,且对图像中的一群人也能很好地分割出来,相互之间不会造成干扰;DABNet和FBSNet对远处的红绿灯无法分割,且DABNet在对人群分割时抗干扰能力弱,分割结果会出现伪影等现象,MANet采用的通道注意力加强了对小目标的突出表达,在对多目标重叠进行分割时抗干扰能力强。从第4行的分割结果可以看出,MANet对图像中出现的遮挡和重叠部分能够准确地分割出来,避免了不同类别之间的相互干扰;DABNet和FBSNet在对重叠部分分割时出现了分类错误,且对电线杆等细小类别的分割效果很差。实验结果表明,本文方法在道路实际场景展现出更加优异的语义分割能力和分类识别能力,且满足道路场景下实时性的要求。

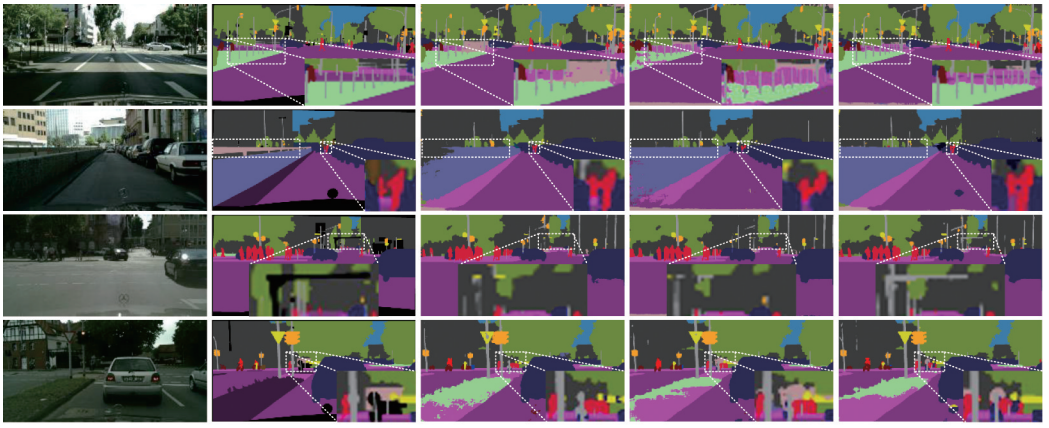

图6　不同算法在Cityscapes数据集上的实验结果对比

Fig.6　Comparison of experimental results of different algorithms on Cityscapes dataset

2.3.3　在CamVid数据集的性能分析

表3展示了MANet与其他先进的实时语义分割算法在CamVid街景数据集上的对比情况。

表3　不同算法在CamVid数据集上的对比结果

Table 3　Comparison results of different algorithms on CamVid dataset

模型	输入尺寸	预训练	MIoU/%	FPS/(帧·s^{-1})	参数量/MB	权重I_i
ENet[7]	360×480	无	51.3	136	0.36	0.12
ESPNet[8]	360×480	无	55.6	219	0.36	0.15
DABNet[9]	360×480	无	66.4	117	0.76	0.13
LEANet[24]	360×480	无	67.5	98	0.74	0.13
FBSNet[19]	360×480	无	68.9	120	0.62	0.14
MSCFNet[18]	360×480	无	69.3	—	1.15	—
MFNet[20]	512×512	无	71.5	145	1.34	0.15
BiSeNet-V2[11]	720×960	无	72.4	124	3.40	0.15
CIDNet[16]	720×960	无	73.5	130	6.50	0.15
MANet	360×480	无	67.7	148	5.58	0.14

实验结果表明,在低分辨率的CamVid数据集中,MANet结合双分支网络结构的设计优化了对图像中小目标的分割效果,引入的多级注意力机制有效地突出了对重要特征的表达,MANet在推理速度上达到次优,分割精度和参数量也有一定优势,同时在整体的平衡权重上达到次优。在几个分割精度优于MANet的网络中,MFNet、BiseNet-v2和CIDNet对输入图像的分辨率要求都比较高,需要消耗更大的计算资源,FBSNet虽然精度略高于MANet,但速度略慢。虽然ENet参数量最低,但MANet速度优于ENet,且精度比ENet高16.4%。LEANet在精度上与MANet相近,但在速度上远比MANet低,DABNet尽管参数量低于MANet,但它在分割精度和速度两方面都明显不如MANet,ESPNet尽管速度最快,但其精度过低。

图7展示了DABNet、FBSNet和MANet在CamVid道路场景数据集上的主观对比效果,可以看出,MANet在网络的不同阶段引入了注意力机制进行特征优化,抑制了图像中一些无用的特征信息,强化了对重要特征的表达。针对小目标进行分割时,例如第1行中的道路旁边的指示牌,具有更精准的分割效果。同时从第2行和第3行中可以看出,在对近距离和远距离的一些干扰目标,MANet具有更强的抗干扰能力。

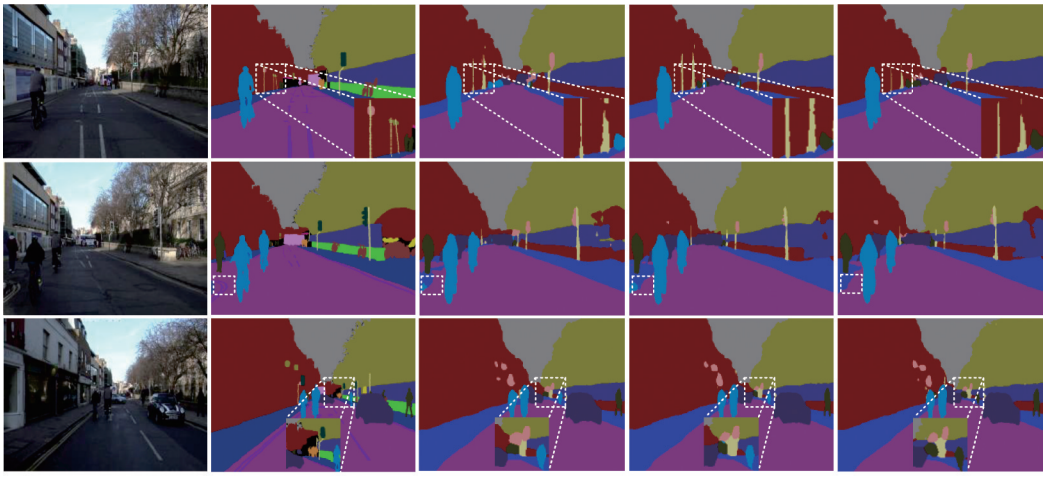

图7 不同算法在CamVid数据集上的实验结果对比

Fig.7 Comparison of experimental results of different algorithms on CamVid dataset

由此可见,MANet在CamVid数据集上同样能取得较为优异的分割性能,综合两个数据集的实验结果,分析可知,MANet虽在CamVid数据集上分割精度上有所不足,但分割速度存在优势。同时MANet在主流的Cityscapes数据集上性能优异,分割精度达到最优,有效地实现了分割精度和推理速度之间的平衡。这是由于CamVid数据集道路场景单一,场景中的类别较少,且数据量少,而Cityscapes数据集道路场景复杂,目标多样且数据量充足。这些充分地说明了MANet在不同道路场景下具有鲁棒性强的特点,且更适合在场景复杂、目标多样的道路场景中使用。

3 结束语

为满足语义分割在道路场景下的应用,本文提出一种多级注意力特征优化的道路场景实时语义分割网络。为了改善道路场景中多目标重叠造成的相互干扰以及图像中小目标信息丢失等问题,首先在特征提取阶段设计深度残差注意力模块提取丰富的语义上下文信息,优化图像的局部特征;在特征加

强阶段设计通道注意力和深度聚合金字塔池化模块,进一步加强语义上下文信息的表示;在特征融合阶段设计注意力融合模块自上而下地融合不同层级下的特征信息,实现图像全局特征信息的有效交互。实验结果表明,MANet兼顾了道路场景下的分割精度和推理速度,整体表现优于其他对比算法。在后续工作中,针对小目标背景重叠等问题,将考虑设计轻量级的解码器替换上采样操作,减少特征图目标细节信息的丢失,进一步提高整个算法的分割性能。

参考文献:

[1] DHAMIJA T, GUPTA A, GUPTA S, et al. Semantic segmentation in medical images through transfused convolution and transformer networks[J]. Applied Intelligence, 2023, 53(1): 1132-1148.

[2] 李嘉祥, 宣士斌, 刘丽霞, 等. 学习几何结构特征的真实点云场景语义分割[J]. 数据采集与处理, 2023, 38(2): 336-349.
LI Jiaxiang, XUAN Shibin, LIU Lixia, et al. Learning semantic segmentation of real point cloud scene with geometric structure features[J]. Journal of Data Acquisition and Processing, 2023, 38(2): 336-349.

[3] XU J, XIONG Z, BHATTACHARYYA S P. PIDNet: A real-time semantic segmentation network inspired by PID controllers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 19529-19539.

[4] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39(4): 640-651.

[5] 项建弘, 刘茁, 王霖郁, 等. 基于强化语义流场和多级特征融合的道路场景分割方法[J]. 数据采集与处理, 2022, 37(2): 426-436.
XIANG Jianhong, LIU Zhuo, WANG Linyu, et al. Road scene segmentation method based on enhanced semantic flow field and multilevel feature fusion[J]. Journal of Data Acquisition and Processing, 2022, 37(2): 426-436.

[6] YANG Z, YU H, FENG M, et al. Small object augmentation of urban scenes for real-time semantic segmentation[J]. IEEE Transactions on Image Processing, 2020, 29: 5175-5190.

[7] PASZKE A, CHAURASIA A, KIM S, et al. ENet: A deep neural network architecture for real-time semantic segmentation [EB/OL]. (2016-06-07). https://arxiv. org/pdf/1606. 02147. pdf.

[8] MEHTA S, RASTEGARI M, CASPI A, et al. ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: [s.n.], 2018: 552-568.

[9] LI G, YUN I, KIM J, et al. DABNet: Depth-wise asymmetric bottleneck for real-time semantic segmentation[C]//Proceedings of the 2019 British Machine Vision Conference. Durham: BMVA Press, 2019: 186.

[10] 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.]: [s.n.], 2018: 325-341.

[11] 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(11): 3051-3068.

[12] LI H, XIONG P, FAN H, et al. DFANet: Deep feature aggregation for real-time semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 9522-9531.

[13] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 1251-1258.

[14] POUDEL R P K, LIWICKI S, CIPOLLA R. Fast-SCNN: Fast semantic segmentation network[EB/OL]. (2019-02-12). https://doi.org/10.48550/arXiv.1902.04502.

[15] ZHAO H, QI X, SHEN X, et al. ICNet for real-time semantic segmentation on high-resolution images[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: [s.n.], 2018: 405-420.

[16] DONG Y, YANG H, PEI Y, et al. Compact interactive dual-branch network for real-time semantic segmentation[J]. Complex & Intelligent Systems, 2023,9: 6177-6190.

[17] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 7132-7141.

[18] GAO G, XU G, YU Y, et al. MSCFNet: A lightweight network with multi-scale context fusion for real-time semantic segmentation[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(12): 25489-25499.

[19] GAO G, XU G, LI J, et al. FBSNet: A fast bilateral symmetrical network for real-time semantic segmentation[J]. IEEE Transactions on Multimedia, 2022, 25: 3273-3283.

[20] LU M, CHEN Z, LIU C, et al. MFNet: Multi-feature fusion network for real-time semantic segmentation in road scenes[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(11): 20991-21003.

[21] ZHONG Z, LIN Z Q, BIDART R, et al. Squeeze-and-attention networks for semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 13065-13074.

[22] HONG Y, PAN H, SUN W, et al. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes[EB/OL]. (2021-01-15). https://arxiv.org/abs/2101.06085.

[23] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2021: 13713-13722.

[24] ZHANG X L, DU B C, LUO Z C, et al. Lightweight and efficient asymmetric network design for real-time semantic segmentation[J]. Applied Intelligence, 2022, 52(1): 564-579.

作者简介:

张鹏(1999-),男,硕士研究生,研究方向:深度学习、图像语义分割,E-mail:919101490@qq.com。

彭宗举(1973-),**通信作者**,男,教授,博士生导师,研究方向:视频信号处理与编码,E-mail:pengzongju@126.com。

张文瑞(2000-),男,硕士研究生,研究方向:图像语义分割,E-mail:2218756929@qq.com.

罗英国(1999-),男,硕士研究生,研究方向:图像压缩,E-mail:2427673474@qq.com。

韦玮(1999-),男,硕士研究生,研究方向:光场图像超分辨率重建,E-mail:545645775@qq.com。

王培容(1973-),女,副教授,研究方向:图像处理,E-mail:wangpeirong@cqut.edu.cn。

(编辑:陈珺,王婕)