文章编号：2095-2163(2022)11-0111-07 中图分类号：TP301.6 文献标志码：A

基于注意力机制和扩张解码改进的语义分割研究

曹玉峰，高建瓴，陈 楠
（贵州大学 大数据与信息工程学院，贵阳 550025）

摘 要：随着自主系统的兴起，技术革新和生产生活中需求的不断加大，实时性计算越来越受欢迎。在本文中介绍了快速卷积神经网络（Fast-SCNN），这是一种基于高分辨率图像数据的实时语义分割模型。在该模型现有的 2 个快速分割分支基础上，将原网络解码器部分改进为扩张解码，扩大感受野，有效提升网络模型的分割精度。然后引入了 CBAM 注意力机制模块，减少对冗余信息的关注，降低了计算量，提高了分割效率。改进后的网络在 Cityscapes 数据集上获得了 73.27%的 *MIoU*，同时保证了网络的推理速度，实验结果表明改进网络较原网络性能有所提升。

关键词：语义分割；扩张解码；注意力机制

Research on semantic segmentation based on attention mechanism and extended decoding

CAO Yufeng, GAO Jianling, CHEN Nan
(College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China)

【Abstract】 With the rise of autonomous systems, technological innovation and increasing demand in production and life, real-time computation is increasingly desirable. This paper introduces fast segmentation convolutional neural network (Fast-SCNN), a real-time semantic segmentation model on high resolution images data. Building on existing two-branch methods for fast segmentation, the paper improves the original network decoder to the extended decoding, expands the receptive field, and effectively improves the segmentation accuracy of the network model. Then the attention mechanism module of CBAM is introduced to reduce the attention to redundant information, reduce the amount of calculation and improve the segmentation efficiency. The improved network obtains *MIoU* of 73.27% on Cityscapes data set, and ensures the inference speed of the network. The experimental results show that the performance of the improved network is improved compared with the original network.

【Key words】 semantic segmentation; extended decoding; attention mechanism

0 引 言

近年来，语义分割已成为计算机视觉[1]和机器学习领域的一个基本研究课题，同时也已成为备受各方关注的学术热点之一。目前的研究指出，为图像的每个像素分配单独的类别标签是构建复杂机器人系统（例如无人驾驶汽车和家用机器人）的重要步骤。语义分割的最简单定义是像素级标记。因此，认为仅查找场景中包含的类标签是远远不够的。类标签应根据上下文而变化。例如，在无人驾驶汽车中，像素标签可以是人、道路、汽车等。

目前，人们对语义分割的兴趣日益浓厚，随着当代人工智能技术的不断发展，在各种计算机视觉前沿理论研究体系中，图像语义分割已经渐渐成为最核心基础的部分，在诸多领域有着广泛的应用。例如，在医疗影像领域，图像语义分割能够分割出细节信息，使医生能够从图像中更清晰地了解到病人的身体信息，为医生诊断病情提供帮助。在拍照时，语义分割方法能够精准地识别脸部轮廓信息，辅助人脸美化算法，释放人类爱美的天性。在自动驾驶中，图像语义分割能够帮助汽车识别场景中的道路、障碍物等信息并精确定位，确保汽车行驶安全。但仍需指出的是，到目前为止，语义分割技术仍然面临重大挑战，主要表现为如下 3 点：

（1）由于各个光照、拍摄距离等各方面的差异，同样的物体在图像中表现出来的特征信息可以有极大差异，而且每个物体都存在被其他物体遮挡或割裂等现象，因此想要进行正确的语义信息标记难度较大。

（2）图像内容各不相同，且变化各异，较为复杂，相同的物体在图像中可能表现不同种情况，而不同类别的物体表现出来又可能极度相似。

（3）在现实场景中，图像中经常出现一些比较

作者简介：曹玉峰（1996-），男，硕士研究生，主要研究方向：智能算法、图像处理；高建瓴（1969-），女，副教授，硕士生导师，主要研究方向：数据库系统、数据挖掘；陈 楠（1997-），女，硕士研究生，主要研究方向：智能算法、图像处理。
通讯作者：高建瓴 Email：454965711@qq.com

收稿日期：2022-08-25

繁杂且凌乱的背景,这对于进行图像语义分割也是一大难点。

针对复杂场景中目标分割精度不高的问题,本文提出一种基于注意力机制[2]的目标分割方法。结合注意力机制,自适应地学习特征图通道之间的关系。强调有用信息,抑制冗余信息,提高了特征图的判别能力。实验结果表明,与同类算法相比,该算法具有较好的分割效果,并显著提高了分割效率。同时在 Fast-SCNN 网络的解码器部分对图像信息还原时,由于过度压缩图像信息,在图像信息还原过程中许多特征细节信息将会丢失,导致最后得到的分割效率不高,因此在解码器部分进行一定的改进,使用扩张解码[3],捕获更为丰富的多尺度信息,得到更密集的特征信息和更大的感受野,有效提升网络模型的分割精度。

1 语义分割相关模型

2014 年,定义提出了一个全卷积网络(Fully Convolutional Networks, FCN)[4],该网络可以接受任意大小的图像输入,避免了采用像素块带来的重复存储和计算的问题。

2015 年,Hyeonwoo 等人[5]提出的 Deconvnet 对分割的细节处理强于 FCN,位于低层的 filter 能捕获目标的形状信息,位于高层的 filter 能够捕获特定类别的细节信息,分割效果更好,但其对细节的处理难度较大。

2016 年,剑桥大学团队基于 FCN 设计了一种卷积神经网络结构 SegNet[6],使用去池化对特征图进行上采样,在分割中保持细节的完整性;去掉全连接层,使其拥有较少的参数。但当对低分辨率的特征图进行去池化时,会忽略邻近像素的信息,导致精度不高。

2017 年,TN Kipf 利用 GCN[7]模型,提出了带有大维度卷积核的编码器-解码器结构,使其提取出来的特征较为优秀。但却具有较多的参数,从而导致计算复杂度较高。

2019 年,提出了具有自注意力机制的双注意网络 DANet[8]。在传统的 FCN 上附加 2 种注意力模块,分别模拟空间和通道维度中的语义相互依赖性,有助于提高分割结果的精度。

因此,网络模型在具有较好实时性的同时,还能使分割精度也得以提升,仍然亟待进一步的探讨。受 two-branch 结构和 encoder-decoder 网络启发的 Fast-SCNN[9]网络,可用于高分辨率(1 024×2 048)图像上的实时语义分割任务。

2 本文网络模型与结构

2019 年,提出了一种快速语义分割网络 Fast-SCNN,该模型适用于低存储的嵌入式设备上的高效计算,相较于其他语义分割模型,不仅计算速度快,计算精度也不低,该优势可以满足在交通场景上语义分割的实时性和准确性。

但 Fast-SCNN 网络在解码器部分进行图像信息还原时,由于网络结构的问题,导致图像信息压缩过度,在图像信息还原过程中许多特征细节信息丢失,导致最后得到的分割效率不高,因此本文在解码器部分进行一定的改进,使用扩张解码[3],利用 3 个扩张卷积密集连接到一起,捕获更为丰富的多尺度信息,得到更密集的特征信息和更大的感受野,有效提升网络模型的分割精度。同时,本文还使用了 CBAM 注意力模块(Convolutional Block Attention Module),用于抑制关注一些不重要的信息通道,降低了时间复杂度,从而提高了识别效率。

2.1 Fast-SCNN 网络介绍

Fast-SCNN 受双分支结构和传统的编码解码结构的启发,基于一个编码器解码器结构通过学习下采样,利用全局特征提取器进行特征提取,在特征融合阶段通过一个二次线性插值进行一次上采样,继而和学习下采样输出的结果直接相加后,再次通过特征融合,最后进行像素点的分类,用于完成高分辨率(1 024×2 048)图像上的实时语义分割任务。Fast-SCNN 的整体网络架构如图 1 所示。由图 1 可知,该网络整体上主要由 4 部分组成:学习下采样模块、全局特征提取器、特征融合模块和分类器,所有模块都采用深度可分离卷积构建。对此拟做阐释分述如下。

(1)学习下采样模块。在学习下采样模块中,采用了 3 层的结构。只使用了 3 层来确保低级特征共享的有效性和高效实施。第一层是标准卷积层(Conv2D),其余两层是沿深度可分离的卷积层(DSConv)。学习下采样模块中所有 3 层都使用了步长为 2、卷积核大小为 3×3 的卷积层,每一个卷积层后都会紧接着一个 *BN* 层和一个 *ReLU* 激活函数。

(2)全局特征提取器。全局特征提取器模块旨在捕获用于图像分割的全局上下文。在传统的深度分支中,其输入图像的分辨率一般很低,而全局特征提取器的输入是在经过学习下采样模块处理后所得到的特征图。当输入和输出大小相同时,通过残差连接,以及使用了高效的深度可分离卷积,从而减少了参数和浮点运算的数量。此外,在末尾添加了一个金字塔池模块(PPM),以聚合基于不同区域的上下文信息。

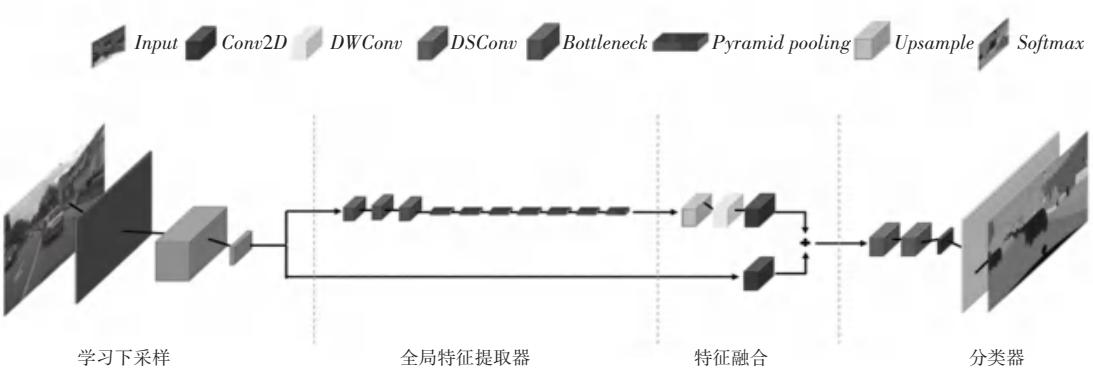


图 1　Fast-SCNN 网络结构

Fig. 1　Fast-SCNN network structure

(3)特征融合模块。通过简单地融合不同分支的特征以确保有效性。或者可以运行时以性能为代价使用更复杂的特征融合模块(例如 Bisenet),以达到更高的精度。

(4)分类器。在分类器中,采用了 2 个深度可分离卷积和逐点卷积。在特征融合模块后增加数层可以提高准确率。在训练期间使用了 *Softmax* 来进行损耗的运算。

2.2　Fast-SCNN 网络的改进

Fast-SCNN 模型中也存在着一些缺点和不足,图像最终分割的效果与感受野大小有明显的差距,需要增大感受野,原模型中还存在一些信息冗余,会导致模型计算速度较慢,因此想要获得更好的效果,必须对其进行优化。本文对 Fast-SCNN 网络从以下 2 个方面进行改进。

(1)改进解码器:在解码器部分进行一定的改进,使用扩张解码,利用 3 个扩张卷积密集连接到一起,捕获更为丰富的多尺度信息,得到更密集的特征信息和更大的感受野,有效提升网络模型的分割精度。

(2)增加注意力模块:使用了 CBAM 注意力模块(Convolutional Block Attention Module)用于抑制关注一些不重要的信息通道,降低了时间复杂度,提高了识别效率。

2.2.1　扩张解码

本文使用 3 个不同扩张率(2、3、5)的扩张卷积连接起来进行卷积操作,为了获取到更加丰富的多尺度信息,进行更为密集的采样以及获取到更大的感受野。

本文使用的扩张解码是将 3 个扩张卷积两两连接到一起,一开始输入的特征映射被传输到模块中的各个扩张卷积上作为输入,每一个扩张卷积的输出都作为此后的扩张卷积的输入,最终,将所有卷积层的输出级联到一起作为结果输出。该扩张卷积有着更大、也更为密集的空间采样。扩张解码结构如图 2 所示。

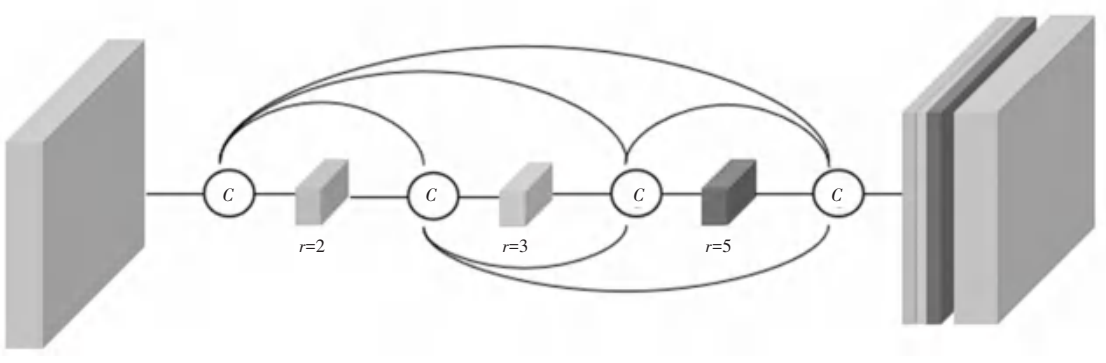


图 2　扩张解码模块结构图

Fig. 2　Structure diagram of extended decoding module

传统特征提取的步骤可概述为:首先,对特征图进行下采样操作来获取到分辨率较低的特征图;然后,进行普通卷积操作来提取图像特征;最后,再进行一次上采样操作,该操作主要作用是恢复特征图像的分辨率,并将其输出。而本文使用的扩张解码可以直接应用于高分辨率的输入图片,获取更大的感受野,使用更为密集的采样方式来获取到更多的空间细节信息。

2.2.2 注意力机制

CBAM 注意力机制具有轻量性和通用性的优点,在前馈卷积式神经网络中,该模块的应用更高效简洁。CBAM 包含 2 个独立的子模块:通道注意力模块(Channel Attention Module,CAM)和空间注意力模块(Spartial Attention Module,SAM),分别进行通道与空间上的注意力操作。CBAM 注意力机制如图 3 所示。

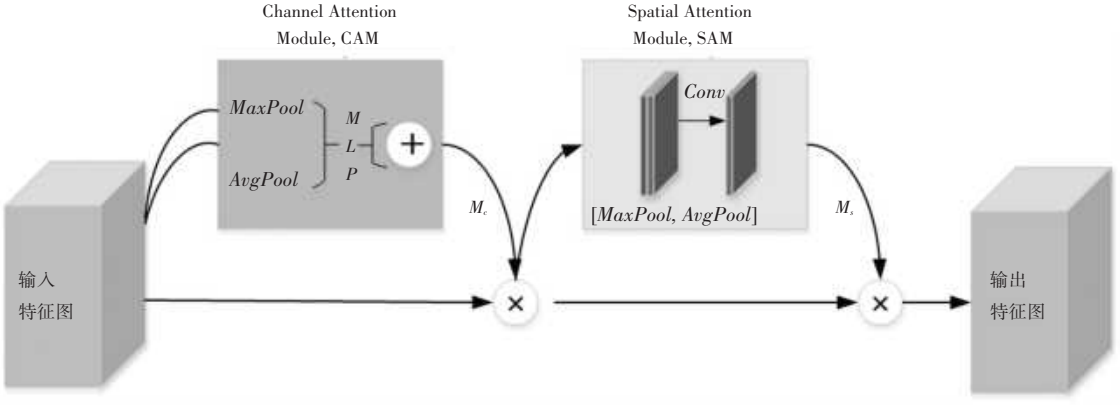


图 3 CBAM 注意力机制模型结构

Fig. 3 CBAM model structure of attention mechanism

CBAM 整体结构为:输入特征图后,卷积层输出的结构首先要通过一个通道注意力模块,将输出进行加权后,再通过一个空间注意力模块,最后则进行一次加权得到完整的输出特征图。

通道注意力模块(Channel Attention Module,CAM)[10] 如图 4 所示,将输入特征图分别经过全局最大池化和全局平均池化,再将输出结果送入一个 2 层神经网络(MLP)。经过 MLP 处理后,获得的结果将由逐个元素依次进行加和操作,并经 *Sigmoid* 激活操作,生成最终输出的通道注意力特征图。将所获得的特征图与初始输入的特征图的所有元素逐个进行乘法运算作为 Spatial attention 的输入特征。

空间注意力(Spartial Attention Module,SAM)模块[11] 如图 5 所示,将通道注意力模块的输出作为空间注意力的输入,输入特征经过一次最大池化后,再经一次平均池化,得到 2 个特征图,然后将这 2 个特征图通道拼接起来,随后经过一个 7×7 的卷积核对通道进行降维处理,使通道数量变成一个,接着又通过 *Sigmoid* 激活函数将其激活,最后直接输出空间注意力特征。

综上所述,本文结合扩张解码和注意力机制对 Fast-SCNN 模型进行改进,获得更好的语义分割效果,通过改进提升了分割精度和效率。改进后的整体网络模型结构如图 6 所示。本文使用的扩张解码以 DDModule 命名。

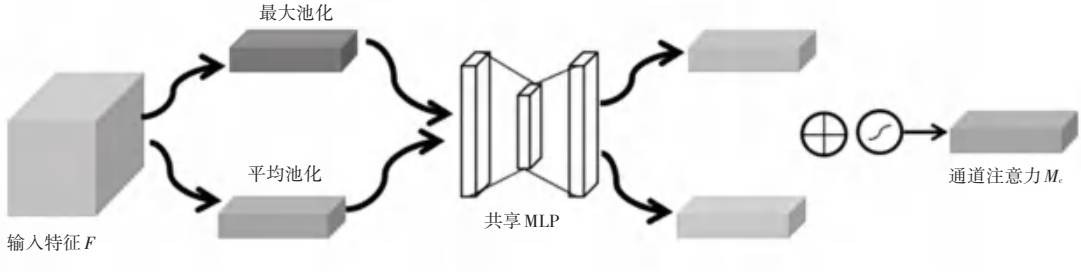


图 4 通道注意力模块

Fig. 4 Channel Attention Module(CAM)

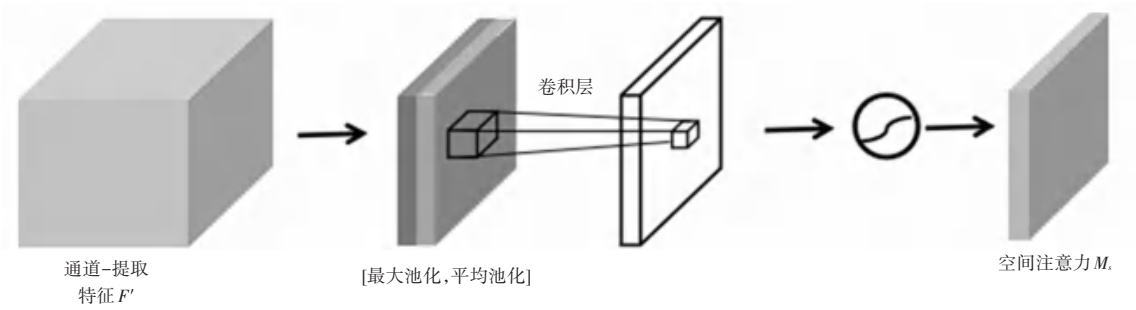


图 5　空间注意力模块

Fig. 5　Spatial Attention Module(SAM)

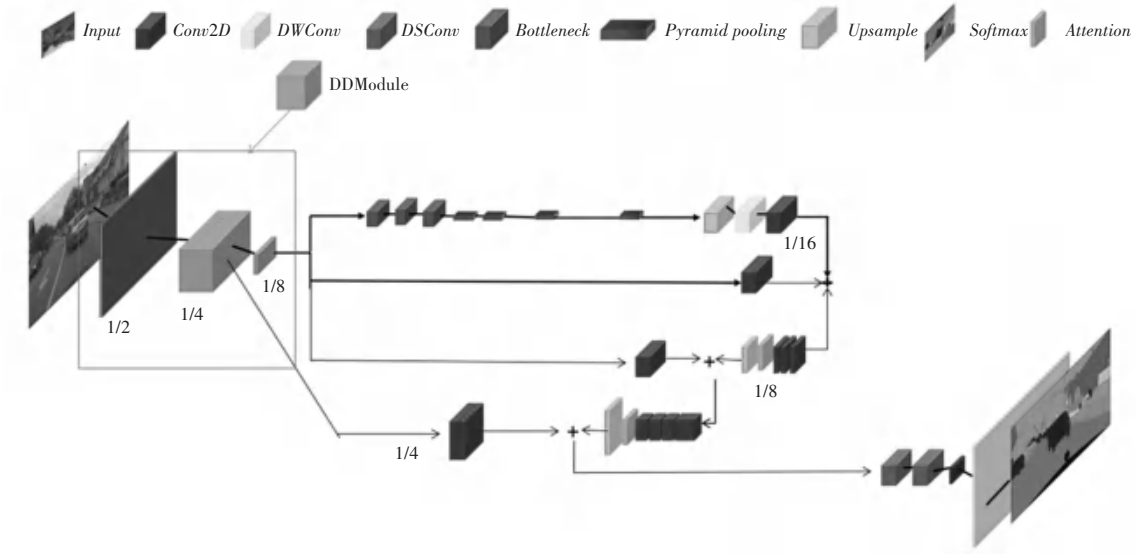


图 6　改进后的整体网络结构图

Fig. 6　Overall network structure diagram after improvement

改进后的特点及优势可阐释如下：

(1)使用扩张解码，增大感受野，获得更丰富的的特征信息，提高了分割精度。

(2)引入 CBAM 注意力机制，减少对冗余信息的关注，减少计算量。提高了分割效率。

3　实验结果与分析

3.1　数据集

Cityscapes 数据集、即城市景观数据集，是一个道路场景分割中常用的大型公开数据集，其中包含从 50 个不同城市的街景中记录的各种立体视频序列，除了更大的 20 000 个弱注释帧之外，还有高质量的 5 000 帧像素级注释。数据集中包括 8 个类别。这八个大类涵盖了 19 个子类别。常用的 19 个语义类别是：道路(Road)、人行道(Sidewalk)、建筑物(Building)、墙壁(Wall)、栅栏(Fence)、栏杆(Pole)、交通灯(Traffic light)、交通标志(Traffic sign)、草丛(Vegetation)、地面(Terrain)、天空(Sky)、行人(Person)、骑行者(Rider)、汽车(Car)、卡车(Truck)、公交车(Bus)、火车(Train)、摩托车(Motorcycle)、自行车(Bicycle)。

3.2　评价标准

本文选用 *MIoU*(Mean Intersection over Union)语义分割评价指标，表示平均交并比，即数据集上每一个类别的 *IoU* 值的平均。

以 Cityscapes 为例，共包含 19 个类别，分别对每个类别求 *IoU*。研究推得的 *MIoU* 的计算公式为：

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{j=0}^{k}P_{ij}+\sum_{j=0}^{k}P_{ji}-P_{ii}} \tag{1}$$

其中，k 表示类别；$(k+1)$ 表示加上背景类；i 表示真实值；j 表示预测值。

此外，选用像素精度(*PA*)描述图像像素的准确度，选用平均像素精度(*MPA*)对改进实验加以评定，此处需用到的公式可写为：

$$PA = \frac{\sum_{i=0}^{k}P_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}P_{jj}} \tag{2}$$

$$MPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{j=0}^{k}P_{ij}} \quad (3)$$

3.3 实验环境

参数设置:初始学习率为 0.001,动量为 0.9,权重衰减 0.000 5,批尺寸 4,选择扩张率为(2,3,5)的扩张解码。具体的环境硬件以及软件配置见表 1。

表 1 实验环境配置

Tab. 1 Experimental environment configuration

实验环境	配置参数
CPU	Intel I5-7500
内存	64 G
显卡	GeForce RTX2080TI
CUDA	CUDA11.0
Pytorch	1.0.0

3.4 实验结果

使用扩张率分别为(1,1,1)、(2,3,5)、(3,4,5)的扩张解码在 Cityscapes 数据集上进行实验,验证不同扩张率对扩张解码性能的影响,实验结果见表 2。

表 2 扩张解码对比实验

Tab. 2 Contrast experiment of extended decoding

Classes	(1,1,1)	(2,3,5)	(3,4,5)
Road	97.54	97.31	98.41
Sidewalk	81.52	83.39	83.38
Building	90.38	90.89	90.77
Wall	52.70	53.53	52.53
Fence	52.14	53.91	51.91
Pole	45.78	52.77	44.45
Light	58.53	59.78	59.43
Sign	57.97	58.81	58.61
Vegetation	69.49	70.53	70.55
Terrain	91.30	91.19	91.20
Sky	93.61	94.33	94.59
Person	73.35	74.41	73.88
Rider	48.41	51.72	50.29
Car	92.72	93.31	93.26
Truck	64.29	65.65	64.67
Bus	74.81	76.34	72.49
Train	51.23	63.39	47.98
Motocycle	52.89	53.27	54.62
Bicycle	69.49	70.81	70.63
MIoU	**66.41**	**68.23**	**68.97**

在扩张率为(2,3,5)下具有比扩张率(1,1,1)更高的 *MIoU*, 且比扩张率为(3,2,5)的计算速度快,综上所述研究选择扩张率为(2,3,5)的扩张解码。

将 Fast-SCNN 网络模型的浅层特征提取解码结构改为扩张解码的网络命名为 EFast-SCNN。并将加入 CBAM 注意力机制的网络模型与原 Fast-SCNN 和 EFast-SCNN 进行对比实验,实验结果见表 3。

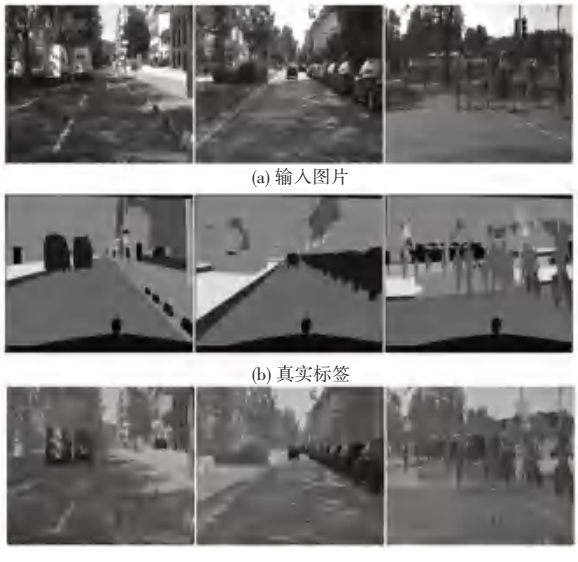
表 3 引入注意力机制对比实验

Tab.3 A comparative experiment introducing attention mechanism

模型	*MIoU* /%	单张图像处理时间/ms	*PA* /%	*MPA* /%
Fast-SCNN	60.78	19.91	94.03	69.13
EFast-SCNN	68.23	25.13	94.67	77.85
EFast-SCNN+CBAM	73.27	23.78	95.21	79.31

从表 3 的实验结果可看出将解码器结构改为扩张解码后的 EFast-SCNN 网络在 Cityscapes 上较原 Fast-SCNN 网络的 *MIoU* 值提升了 7.45%,再加入 CBAM 注意力模块后又提升了 5.04%,像素精度(*PA*)和平均像素精度(*MPA*)都有显著增加,证明了改进后的分割效果得到了提升,由于模型复杂度的提高,使得改进网络在单张图像上的处理时间较原网络有所增加,但加入轻量化注意力模块后有效缓解了该问题。综合前述分析可知,在大幅提升分割精度的同时,也不影响模型分割的实时性,证明了改进是有效的。

将改进后的 DFast-SCNN+CBAM 模型在 Cityscapes 数据集上得到的结果进行可视化分析,如图 7 所示。



(a) 输入图片

(b) 真实标签

(c) 模型输出

图 7 改进模型在 Cityscapes 数据集上结果可视化

Fig. 7 Visualization of results of the improved model on Cityscapes datasets

(下转第 121 页)