计算机科学
COMPUTER SCIENCE

基于混合注意力的实时图像语义分割算法

王燕, 夏创帅, 汪娜, 南佩奇

引用本文

王燕, 夏创帅, 汪娜, 南佩奇. 基于混合注意力的实时图像语义分割算法[J]. 计算机科学, 2023, 50(11A): 230200010-6.

WANG Yan, XIA Chuangshuai, WANG Na, NAN Peiqi. Real-time Image Semantic Segmentation Algorithm Based on Hybrid Attention [J]. Computer Science, 2023, 50(11A): 230200010-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230200010

基于混合注意力的实时图像语义分割算法

王 燕 夏创帅 汪 娜 南佩奇
兰州理工大学计算机与通信学院 兰州 730050
(wangyan@lut.edu.cn)

摘 要 针对现有语义分割算法因模型复杂、计算量庞大,导致算法较难部署在移动设备的问题,提出了一种基于混合注意力的实时图像语义分割算法。该算法是非对称的编码器解码器结构,编码器部分结合深度可分离卷积与扩张卷积设计出一个高效残差单元来提取不同网络深度的图像特征,在浅层较多关注空间位置信息,在深层增强语义信息提取。解码器部分设计了混合注意力特征融合模块,使用空间注意力强化浅层的空间位置信息,使用通道注意力增强深层特征图中关键信息的表达能力,能够有效融合不同层次特征图中空间信息与上下文信息,强化语义信息的表达,减小融合过程中图像信息的损失,最后使用分类器得到分割预测图。大量实验的结果表明,该算法在 Cityscapes 数据集上 PA 和 mIoU 分别达到了 93.2%和 73.2%,在 TeslaV100 图像计算显卡上以 1.62×10^6 的参数量达到 38FPS,在 Pascal VOC 2012 数据集上 PA 和 mIoU 达到了 92.4%和 74.8%。实验结果表明,该算法能够有效且实时地完成城市场景图片分割任务。

关键词:深度学习;语义分割;实时;特征融合;注意力机制
中图法分类号 TP391

Real-time Image Semantic Segmentation Algorithm Based on Hybrid Attention

WANG Yan,XIA Chuangshuai,WANG Na and NAN Peiqi
School of Computer and Communication,Lanzhou University of Technology,Lanzhou 730050,China

Abstract The existing semantic segmentation algorithms are difficult to deploy on mobile devices due to the complex model and a large amount of computation. A new semantic segmentation algorithm based on hybrid attention is proposed. This algorithm is an asymmetric encoder-decoder structure. The encoder part combines depth-wise separable convolution anddilated convolution to design an efficient residual module to extract image features at different levels of the network. It pays more attention to spatial position information in the shallow layer and enhances semantic information extraction in the deep layer. In the decoder part,a hybrid attention feature fusion module is designed,which uses spatial attention to strengthen the spatial location information in the shallow layer and channel attention to enhance the expression ability of key information in the deep feature map. It can effectively integrate the spatial information and context information in the feature map of different levels,strengthen the expression of semantic information,and reduce the loss of image information in the fusion process. Finally,the segmentation results are predicted by using the classifier. A large number of experiments show that the proposed algorithm achieves 93.2% PA and 73.2% mIoU in Cityscapes,respectively,and achieves 38FPS with 1.62×10^6 reference on Tesla V100 GPU. In Pascal VOC 2012 data set,PA and mIoU reaches 92.4% and 74.8% respectively. Experimental results show that this algorithm can effectively and quickly complete the task of city scene image segmentation.

Keywords Deep learning,Semantic segmentation,Real-time,Feature fusion,Attention mechanism

1 引言

语义分割是计算机视觉中的一个基本问题,作为图像理解的一种重要方法,语义分割要求将同一图像中不同类别的物体像素使用不同颜色标注出来。语义分割被广泛应用于医学图像分析[1]和遥感测绘[2]等领域。而实时语义分割要求以较高的分割速度对图像进行像素级分割,通常将图像处理速度是否超过每秒 30 帧作为达到实时的一个标准,它广泛应用于自动驾驶[3]、视频监控[4]等对图像处理速度要求较高的场景。在自动驾驶中通过语义分割使车辆获得目前场景具体包含何种物体、物体的具体位置和把物体类别分割到像素级的3种功能,从而为自动驾驶提供高标准的决策条件。

传统语义分割方法主要有基于聚类[5]、阈值[6]、边缘检测[7]和图论[8]的方法,传统分割算法对计算性能考虑得较少,针对不同场景需要手动设计不同的特征提取方法,对复杂场景的适应性较低,也未充分利用现阶段计算资源丰富的优势。

随着深度学习和高性能 GPU 的蓬勃发展,以卷积神经网络 CNN(Convolutional Neural Networks)为代表的分割算法也取得了优异的成绩。Long 等提出了全卷积网络 FCN[9](Fully Convolution Network),将经典 CNN 图像分类网络最后的全连接层替换为一系列反卷积层,使得经过下采样的特征图可以通过反卷积层、上采样恢复为输入图像的分辨率

基金项目:国家自然科学基金(61863025)
This work was supported by the National Natural Science Foundation of China(61863025).
通信作者:夏创帅(xiachuangshuai@163.com)

尺寸，从而完成像素级的物体分类任务。SegNet[10]以 FCN 为基础，引入了编码器-解码器结构，成为最早的高效分割模型之一。继 SegNet 之后，ENet[11]还设计了一种层数较少的编码器-解码器，以降低计算成本。UNet[12]的编码器在提取特征的过程中不断地下采样，在解码器阶段，通过将浅层特征和深层特征相融合，逐步上采样，最终获得高分辨率的输出预测图，现在被广泛应用于医疗图像分割领域。

虽然基于深度学习的 CNN 分割算法对图像分割效果越来越好，但 CNN 分割算法的体积越来越大，结构越来越复杂，预测和训练需要的硬件资源也逐步增多，往往只能在高算力的服务器中运行。而移动设备中硬件、算力的资源限制也促使神经网络向小型化发展，即在保证算法准确率的同时降低对硬件资源的要求，同时加快图像分割的速度。谷歌提出的 MobileNets[13]算法使用深度可分离卷积来替换 CNN 中的标准卷积，有效减少了分割算法的参数与计算量。深度可分离卷积通过将空间滤波与特征生成机制分离，有效地替换了标准卷积。Zhang 等使用分组逐点卷积和通道重排(Channel Shuffle)两种方法设计出了 ShuffleNet[14]。基于以上对 CNN 的优化方法，研究人员又设计出了 ICNet[15]、BiseNet[16-17]系列算法、LEDNet[18]等实时语义分割算法，使得在移动终端、嵌入式设备上部署语义分割算法成为可能。然而，上述的实时语义分割算法依赖轻量化的主干网络提取图像特征，精简的主干网可以使分割算法有较快速的图像推理能力，但也弱化了网络的特征提取能力；神经网络浅层与深层分别提取到的特征往往包含不同类型的图像特征信息，解码器中需要融合不同深度的图像特征，现有方法在特征融合过程中容易损失语义信息。

为了能更好地平衡分割任务中的准确性和效率，本文提出了基于混合注意力的实时语义分割算法(Real-time Image Semantic Segmentation Algorithm Based on Hybrid Attention)，该算法主要有以下两个方面的贡献：

1)设计了高效残差模块(Efficient Residual Module，Efficient-res)，将其作为特征提取主干网的基本单元，该残差结构不仅保证了提取图像特征信息的有效性，而且减少了计算复杂度以及参数量，提升了模型的分割速度。

2)在解码器中设计了混合空间注意力与通道注意力的特征融合模块(Hybrid Spatial Attention and Channel Attention Modules，HSC)，促进了包含较多空间信息的浅层特征与包含较多抽象语义信息的深层特征充分融合，提升了分割精度。

2 本文算法

2.1 本文算法的结构

本文算法的详细结构如图 1 所示，该算法采用非对称的编码器-解码器的结构，其中编码器部分与 VGG16 特征提取主干结构相似，由下采样模块(Down Sampling Module)和如图 2(d)所示的高效残差模块组成。解码器部分由混合通道与空间注意力的特征融合模块和分类器(Classify)组成。

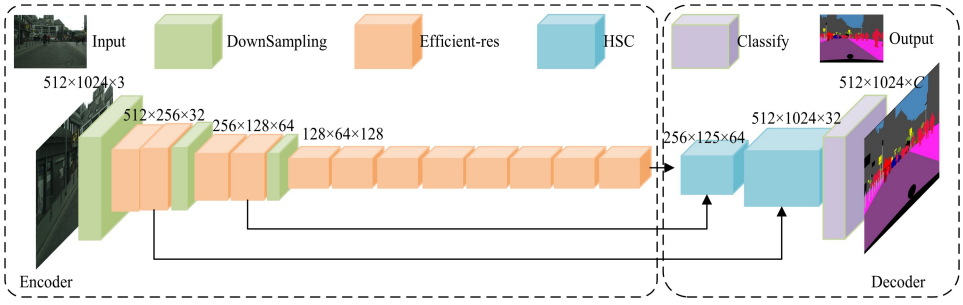

图 1 本文算法的结构

Fig. 1 Structure of the proposed algorithm

在编码器的特征提取阶段，首先对输入图像进行下采样来缩小特征图的分辨率，该下采样操作由步长为 2 的 3×3 卷积、2×2 的最大池化和 2×2 的平均池化组成。通过减小输入图像的尺寸来减少算法的计算量，由于下采样的过程中会损失语义信息，因此结合 Efficient-res 模块进一步提取特征。为保证在网络深层拥有较大的感受野(Receptive Fiele，RF)，在深度可分离卷积中加入带有空洞率的扩张卷积，在减少网络参数量的同时维持了较高的感受野，从而提取网络 3 种不同层次的图像特征。

解码器部分使用两个 HSC 模块融合编码器在 3 种不同网络深度提取到的特征图，这些特征图分别包含了不同空间信息与语义信息。将经过两次融合语义信息的特征图由分类器完成语义信息到像素类别的映射。其中在分类器中首先通过 2D 卷积将通道数调整至与对应数据集语义类别数一致，使用归一化计算得到每个像素点的所属类别概率，从而完成端对端的输出。

2.2 基于通道拆分与通道重排的高效残差结构

本节主要为了解决编码器中残差结构的效率与特征提取能力的平衡问题。由于残差结构具有防止网络退化的优点，已有多个轻量化的残差结构被成功应用在图像分类任务中，其结构如图 2 所示。其中图 2(a)来自 ResNet[19]的瓶颈残差结构，该类瓶颈残差结构通过第一个和最后一个卷积控制通道数量，使中间卷积层的通道数减少，形成类似瓶颈的结构，从而达到减少计算量的目的。但是随着网络深度的加深，它的特征提取能力也开始下降，基于该瓶颈残差结构，ShuffleNet 通过引入通道重排方法，设计了如图 2(b)所示的残差结构单元，还在该结构中使用点卷积(1×1 卷积)来减少参数量和计算量，然而点卷积占用大量的计算复杂度，这种策略对轻量级模型的设计是不利的。LedNet 构建了一个轻量级的残差结构 SS-nbt，如图 2(c)所示，它的通道使用 1D 卷积重新调整通道数量，从而减少参数量。受到 shuffleNet v2 中通道重排与通道切分的启发，本文将扩张卷积与深度可分离卷积结合，设计了带有通道切分(Channel Split)与通道重排的高效残差结构 Efficient-res，结构如图 2(d)所示。

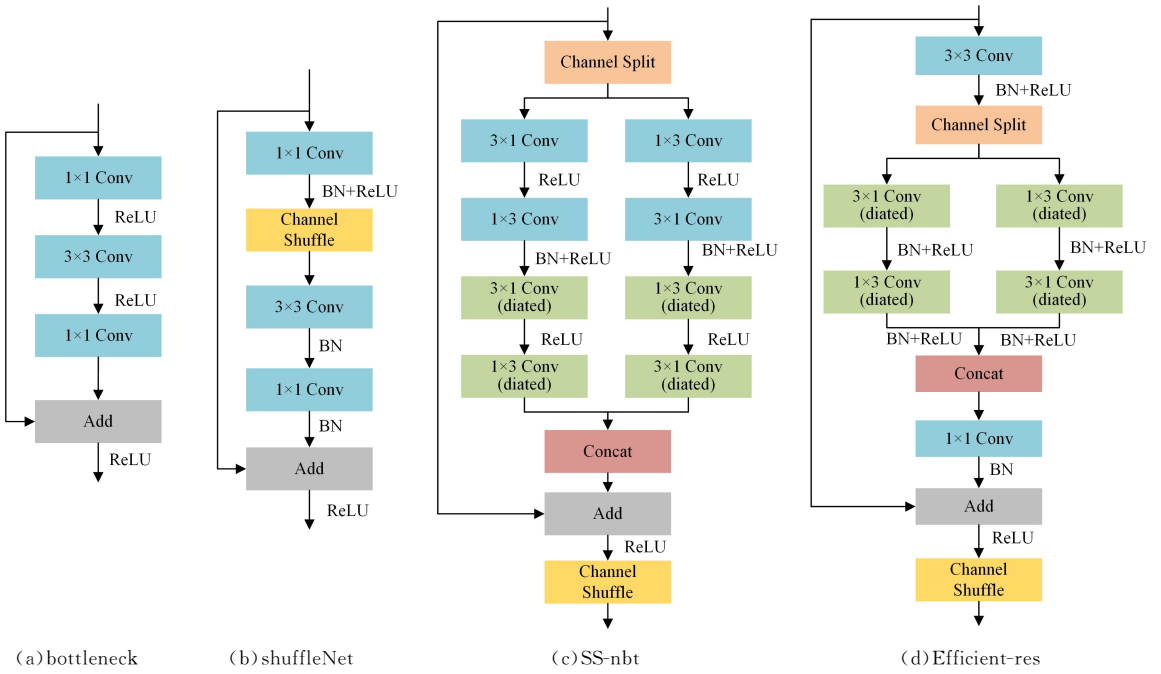

图 2　残差结构图

Fig. 2　Residual structure diagram

Efficient-res 结构采用“切分—转换—合并—重排”的设计策略。其中对通道“重排”和“切分”的操作过程如图 3(a)和图 3(b)所示。Efficient-res 首先经过 2D 卷积层聚合特征并调整通道数量，达到减少参数的目的。在通道切分的过程中，过多的分支会增加对内存的重复访问，也会在一定程度上降低计算效率，因此本文中切分为两个子分支。为了能更高效地提取图像特征，简单使用深度可分离卷积替换标准卷积会导致性能下降，本文在分解后的卷积中融合了带有空洞率的卷积运算，扩张卷积是提升感受野的有效工具，它在不增加计算资源和参数量的情况下拥有更大的感受野，在神经网络中第$(l+1)$卷积层的感受野的计算式如式(1)－式(3)所示：

$$RF_{l+1}=RF_l+(k'-1)\cdot S_l \tag{1}$$

$$k'=k+(k-1)(d-1) \tag{2}$$

$$S_l=\prod_{i=1}^{l} Stride_i \tag{3}$$

其中，k'，k 和 d 分别表示扩张卷积的大小、普通卷积的大小和空洞率，$Stride_i$ 表示之前所有卷积核的移动步长。

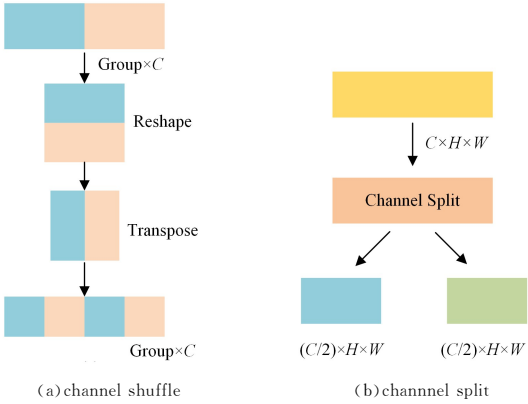

图 3　通道重排与通道拆分执行流程

Fig. 3　Channel shuffle and split working process

经过结合了扩张卷积的卷积运算后，使用 concat 操作在通道维度上拼接两个子分支得到的特征图，再使用一个 2D 卷积将输出通道数恢复至与输入通道数调整一致。由于同一组的各个通道特征图可能包含相同的语义信息，使用通道重排操作来交换不同通道的语义信息，使得每个组的语义信息更加丰富。

本文方法在保证残差结构特征提取能力的基础上能提升残差结构的运行效率，使用通道切分与通道重排的操作交换不同通道的语义信息，使得不同组的语义信息更加丰富，可以提取到更多的图像特征，在计算资源有限的情况下提升了算法的拟合能力，对最终的分割结果也有着正向的作用。

2.3　混合注意力特征融合模块

为解决不同层次的特征在融合过程中容易损失上下文信息的问题，本文在解码器的特征融合阶段设计了混合通道与空间注意力的特征融合模块 HSC，如图 4 所示。

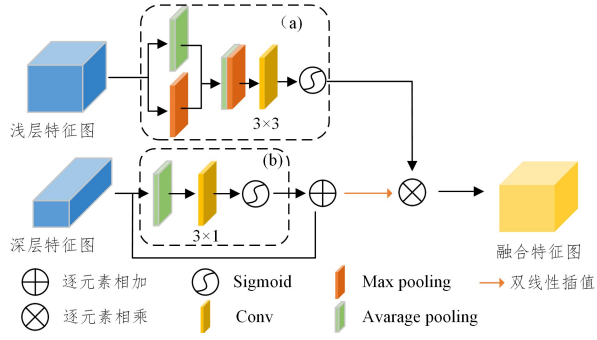

图 4　HSC 模块结构图

Fig. 4　HSC module structure diagram

空间注意力可以对不同像素点的特征图赋予不同的注意力权重，强化包含更多语义信息的通道表示能力。空间注意力的详细结构如图 4(a)所示，空间注意力结构中首先使用步长为 2 的自适应平均池化和自适应最大池化，在通道的维度上拼接两部分池化结果，然后经过一个 2D 卷积结合信息，

使用 Sigmoid 激活神经节点，最终得到包含空间位置信息的特征图。空间注意力的计算式如式(4)所示：

$$x_s = conv\left(\frac{1}{c}\sum_{k=1}^{c} u_{ij}(k) + \max_{k\in[1,c]}(u_{ij}(k))\right) \tag{4}$$

其中，x_s 表示空间注意力机制计算的特征图，$conv$ 为卷积归一化激活函数操作，u_{ij} 表示像素权重，h 和 w 分别为特征图的高度和宽度，c 表示通道数。

通道注意力如图 4(b)所示，通道注意力中使用全局平均池化来提取通道权重，使用 1D 卷积结构聚合通道权重信息，使用 Sigmoid 激活函数激活神经结点，从而得到通道注意力特征图，计算式如式(5)所示：

$$x_c = \frac{1}{h\times w}\sum_{i=1}^{h}\sum_{i=1}^{w} u_c(i,j) \tag{5}$$

其中，x_c 表示通道注意力得到的通道权重分布，h，w 和 c 分别为特征图的高度、宽度和通道数。

图 4 给出了 HSC 融合模块的详细结构。因为浅层的特征图包含更多的空间位置信息，首先对浅层特征使用空间注意力强化重点关注的空间位置信息，同时对深层次特征使用通道注意力，对包含更多语义信息的部分通道进行强化。由于深层次图像特征的分辨率较小，使用双线性插值的方法将特征图分辨率上采样至与浅层特征图一致。将经过注意力强化信息表达的浅层特征图与深层特征图逐点相乘，得到包含更加丰富语义信息的特征图。

3 实验及结果分析

3.1 数据集及评价指标

Cityscapes[20]城市景观数据集是一个用于语义场景解析的数据集，它包含了 5000 张精确标注的城市街景图，其中训练集、验证集和测试集的图片分别为 2975，500 和 1525 张，由于测试集不公开，我们使用验证集进行模型评估。因为该数据集的图片分辨率较大，在该数据集上进行 FPS 的分析。并且在实验过程中对图片进行翻转、随机剪裁和添加噪声等，以进行数据扩张。

Pascal VOC 2012[21](下文简称 VOC)是计算机视觉领域非常流行的数据集，有注释的图像可用于 5 种任务：分类、分割、检测、动作识别和人物布局。对于分割任务，有 21 个标记的对象类，如果像素不属于这些类中的任何一个，则标记为背景。数据集被分为两个集，即训练和验证，分别有 1 464 和 1 449 张图像，以及一个用于实际挑战的私有测试集。

本文采用平均交并比(mean Intersection over Union，mIoU)作为衡量模型分割效果的评价指标，计算式如式(6)所示：

$$mIou = \frac{1}{N}\sum_{i=0}^{N}\frac{p_{ij}}{\sum_{j=0}^{N}p_{ij} + \sum_{j=0}^{N}p_{ji} - p_{ii}} \tag{6}$$

使用平均像素精度(Pixel Accuracy，PA)来衡量模型的分割准确度，其代表预测分类正确的像素数量占标注空间所有像素的百分比，其计算式如式(7)所示：

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \tag{7}$$

其中，N 为所有像素种类的个数，p_{ij} 表示实际类别为 i 类但预测为 j 类的像素总数，p_{ii} 表示实际类别为 i 类预测也为第 i 类的像素总数。

使用帧每秒(Frames Per Second，FPS)来表示模型处理图片的速度，常见测量方法为网络一次性推理 100 张图像之后，求得其平均推理每张图像的耗时，从而推算网络每秒处理帧数 FPS。还通过模型需要的浮点运算量(FLOPs)和参数量(Params)来评估语义分割网络模型的复杂度。

3.2 实验参数

本文算法使用 python3.8、pytorch 1.13.0 编程实现。Ubuntu 18.04 操作系统，CPU 为 Intel(R) Xeon(R) Gold，GPU 为 TeslaV100，CUDA11.0 加速运算。

Cityscapes 数据集中 crop-size 为 1 024，batch-size 设置为 6，优化器为随机梯度下降(SDG)，初始学习率 0.001，学习率衰减为 0.000 5，epoch 为 200。VOC 数据集中 crop-size 为 512，batch-size 设置为 8，优化器与上述一致，学习初始学习率为 0.01，学习率衰减为 0.001，epoch 为 80。

3.3 Cityscapes 数据集实验结果评估

图 5 给出了本文算法与其他实时和非实时算法的分割结果对比，每一列分别为地面真实场景、分割标签图，以及各种算法的分割效果图。从分割结果中可以看出，由于本文算法使用了混合注意力的特征融合结构，有效融合了图像的空间位置信息与上下文信息，在分割目标的连续性上表现较好，例如对于道路、建筑、车辆等较大体积目标的边界曲线较为光滑和连续，分割的整体性也较高。对路灯、广告牌、交通标志等小物体分割效果存在一些小的识别错误问题，原因是小物体的空间位置包含在较大目标中，边界的分割不是非常精准，但整体效果表现良好。

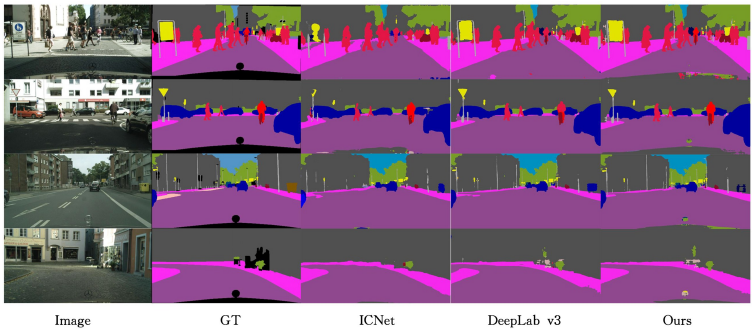

图 5 不同算法在 Cityscapes 数据集上的分割结果图

Fig. 5 Different algorithms segmentation results on Cityscapes datasets

表 1 列出了各种算法在 Cityscapes 数据集上的实验对比结果。相对于 PSPNet[22]，DeepLab v3[23]非实时的分割算法，

本文算法的参数量仅有 DeepLab v3 参数量的 1/20，但取得了与 DeepLab v3 算法较为接近的分割结果，并且在实时性上提升较大。相较于实时算法 LEDNet 等，本文算法虽然参数量增加了 0.68×10^6，但是在 mIoU 上提升了 3.2%的，在 PA 上提升了 1.1%，且在实时性方面提升了 6FPS。与其他实时分割算法ICNet，Fast-SCNN[24] 和 FPANet[25] 等相比，虽然参数量略有增加，但是在精度和图片推理速度上获得了一定的提升，虽然 ENet 的推理速度表现较好，但作为早期经典算法在分割精度方面表现有所欠缺。因此本文算法在参数量、计算复杂度和分割精度中取得了平衡。

表 1　不同算法在 Cityscapes 数据集上的实验结果

Table 1　Experimental results of different algorithms on Cityscapes dataset

Type	Method	Pre-trained	Backbone	Params	Speed/FPS	GFLOPs	mIoU/%	PA/%
non-real time	PSPNet	Y	ResNet-101	250.8×10^6	0.78	412.20	74.00	94.20
	DeepLab v3	Y	ResNet-50	38×10^6	0.56	84.30	73.40	95.50
real time	SegNet	Y	VGG-16	29.5×10^6	13.00	286.00	56.20	89.32
	ENet	N	—	0.36×10^6	53.60	3.80	61.10	90.73
	BiSeNet v1	Y	Xception-39	5.8×10^6	34.40	14.80	62.30	91.50
	ICNet	N	PSPNet50	7.8×10^6	30.30	28.30	69.50	92.50
	LEDNet	N	—	0.94×10^6	32.00	11.50	70.00	92.10
	Fast-SCNN	N	—	1.11×10^6	43.00	—	72.14	93.00
	FPANet-A	N	ResNet-18	14.11×10^6	69.21	30.70	72.00	92.80
	Ours	N	—	1.62×10^6	38.00	11.34	73.20	93.20

3.4　VOC 数据集结果评估

图 6 给出了本文算法与其他语义分割算法在 VOC 数据集上分割结果的对比图，每一列分别为地面真实场景、分割标签图、各种算法的分割效果图。从图 6 中可以发现，本文算法对该数据集的分割特点与 Cityscapes 较为相似，在分割结果中物体边界的分割较为连续和光滑。VOC 数据集中输入图片包含的类别数量少，单张图片中除背景之外的目标类只有一种或者两种，包含 3 类以上的复杂图片非常少。本文算法在只包含一类的图片中分割效果较好，边界连续，分割完整，在包含两类及以上的图片分割中会存在边界不清晰的情况，但整体的分割效果与非实时分割算法较为相似，在部分类别的分割完整性上略有提升。详细实验结果如表 2 所列。

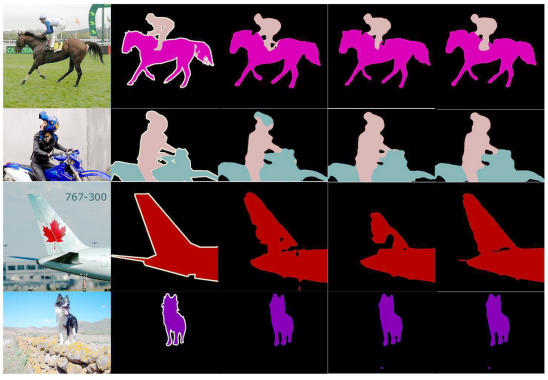

图 6　不同算法在 VOC 数据集上的分割结果图

Fig. 6　Segmentation results of different algorithms on VOC dataset

表 2　不同算法在 VOC 数据集上的实验结果

Table 2　Experimental results of different algorithms on VOC dataset

Method	Backbone	Pre-trained	*mIoU*/%	PA/%
DeepLabv3	ResNet-50	Y	73.7	92.1
PSPNet	ResNet-101	Y	74.5	91.3
ICNet	PSPNet50	N	69.5	90.7
BiSeNetv1	Xception39	Y	67.3	90.0
Ours	—	N	74.8	92.4

本文算法在 VOC 数据集上取得了 74.8%的 mIoU，在 PA 上相比以 DeepLab v3 为代表的非实时分割算法提升了 1%，与经典实时分割算法 ICNet 和 BiSeNet v1 相比，在两种精度评价指标上均有提升。

3.5　消融实验

本文算法用 3 个下采样模块将编码器分为 3 个特征提取块，在第 3 个特征提取块使用带有空洞率的扩张卷积来进一步提取特征。由于相同卷积下不同空洞率会影响感受野，空洞率较小时更关注微小物体的分割，较大的空洞率对空间位置信息提取能力更强，选择合适的空洞率组合对编码器相当重要。为验证使用不同的空洞率组合对编码器特征提取能力的影响，与其他采用相似编码器的算法中使用的两组空洞率进行对比验证，详细结果如表 3 所列。其中分组 1 是 DSA-Net[26] 算法采用的空洞率组合，分组 2 是 LRDNet[27] 采用的组合，分组 3 是本文算法选取的空洞率组合，将 3 组空洞率组合在本文算法中进行实验对比，可以得出本文中使用的空洞率组合取得了较优结果，其 mIoU 达到了 73.2%，相比组合 1、组合 2 平均提升了 0.8%，由此可以证明不空的洞率组合会通过改变深层网络感受野的大小来影响算法的特征提取能力。

表 3　不同空洞率组合实验对比结果

Table 3　Comparative experiment results of different cavity ratio combination

分组	空洞率组合	*mIoU*/%
1	1,3,6,12,3,6,12,34	72.1
2	1,2,5,9,2,5,9,17	72.7
3	1,3,7,11,2,5,13,17	73.2

本文还验证了在特征融合模块中单独使用某一种注意力机制和使用混合注意力机制对本文算法性能的影响，详细实验结果如表 4 所列。

表 4　不同注意力机制组合实验结果

Table 4　Experimental results of combination of different attention mechanisms

Method	Ca	Sa	*mIoU*/%	FPS
Ours	√	—	70.4	41
	—	√	71.5	42
	√	√	73.2	38

从实验结果可以发现，当仅仅使用通道注意力 Ca 或者使用空间注意力 Sa 时，本文算法在 Cityscapes 数据集上在相似的 FPS 下获得了平均 71.1%的 mIoU，而当使用了本文提出的混合注意力时，在 FPS 性能降低 2～3 帧的情况下将 mIoU 提升了 2%，因此同时使用通道注意力 Ca 和空间注意力 Sa 的混合注意力能有效提升算法分割精度，单独使用空间或者

通道注意力时虽然算法的图片推理速度较快，但以个位数帧率的代价换取 2%的 mIoU 提升是值得的。

结束语 本文基于现有的实时图像语义分割算法，提出了一种新的算法，该算法采用编码器解码器结构，引入了混合注意力的特征融合方式融合图像特征信息。该算法在 Cityscapes 和 Pascal VOC 2012 数据集上取得了不错的分割性能。在后续的工作中将使用剪枝和量化的方法对本文算法的参数量进行进一步压缩，在减少算法内存消耗的同时提升了算法的分割速度。也对尝试对 Vision transformer 这类需要高算力的方法进行优化，降低其对算力的需求，构建体积更小、特征提取能力更强的神经网络结构，使算法在实时性和准确性方面均取得一定的进步，满足嵌入式设备的实际需求。

参 考 文 献

[1] ASGARI TAGHANAKI S, ABHISHEK K, COHEN J P, et al. Deep semantic segmentation of natural and medical images: a review[J]. Artificial Intelligence Review, 2021, 54: 137-178.

[2] HE X, ZHOU Y, ZHAO J, et al. Swin transformer embedding UNet for remote sensing image semantic segmentation[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-15.

[3] RIZZOLI G, BARBATO F, ZANUTTIGH P. Multimodal Semantic Segmentation in Autonomous Driving: A Review of Current Approaches and Future Perspectives [J]. Technologies, 2022, 10(4): 90.

[4] CAO X, GAO S, CHEN L, et al. Ship recognition method combined with image segmentation and deep learning feature extraction in video surveillance[J]. Multimedia Tools and Applications, 2020, 79(13): 9177-9192.

[5] MA J W, LEITE F. Performance boosting of conventional deep learning-based semantic segmentation leveraging unsupervised clustering[J]. Automation in Construction, 2022, 136: 104167.

[6] LEE M, KIM D, SHIM H. Threshold matters in WSSS: manipulating the activation for the robust and accurate segmentation model against thresholds [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4330-4339.

[7] LIU Y, CHENG M M, FAN D P, et al. Semantic edge detection with diverse deep supervision[J]. International Journal of Computer Vision, 2022, 130(1): 179-198.

[8] YU H, YANG Z, TAN L, et al. Methods and datasets on semantic segmentation: A review[J]. Neurocomputing, 2018, 304: 82-103.

[9] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3431-3440.

[10] 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.

[11] PASZKE A, CHAURASIA A, KIM S, et al. Enet: A deep neural network architecture for real-time semantic segmentation[J]. arXiv: 1606.02147, 2016.

[12] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]// International Conference on Medical Image Computing and Computer-assisted Intervention. 2015: 234-241.

[13] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv: 1704.04861, 2017.

[14] ZHANG X, ZHOU X, LIN M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6848-6856.

[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. 2018: 405-420.

[16] 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. 2018: 325-341.

[17] 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.

[18] WANG Y, ZHOU Q, LIU J, et al. Lednet: A lightweight encoder-decoder network for real-time semantic segmentation [C]// 2019 IEEE International Conference on Image Processing. 2019: 1860-1864.

[19] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[20] 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. 2016: 3213-3223.

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

[22] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2881-2890.

[23] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv: 1706.05587, 2017.

[24] POUDEL R P K, LIWICKI S, CIPOLLA R. Fast-scnn: Fast semantic segmentation network[J]. arXiv: 1902.04502, 2019.

[25] WU Y, JIANG J, HUANG Z, et al. FPANet: Feature pyramid aggregation network for real-time semantic segmentation[J]. Applied Intelligence, 2022, 52(3): 3319-3336.

[26] ELHASSAN M A M, HUANG C, YANG C, et al. DSANet: Dilated spatial attention for real-time semantic segmentation in urban street scenes[J]. Expert Systems with Applications, 2021, 183: 115090.

[27] ZHUANG M, ZHONG X, GU D, et al. LRDNet: A lightweight and efficient network with refined dual attention decorder for real-time semantic segmentation [J]. Neurocomputing, 2021, 459: 349-360.

WANG Yan, born in 1971, master, professor, is a member of China Computer Federation. Her main research interests include pattern recognition and artificial intelligence.

XIA Chuangshuai, born in 1998, master. His main research interests include pattern recognition and artificial intelligence.