基于上下文和浅层空间编解码网络的图像语义分割方法

罗会兰 [1] 黎 宵 [1]

摘 要 当前图像语义分割研究基本围绕如何提取有效的语义上下文信息和还原空间细节信息两个因素来设计更有效算法. 现有的语义分割模型, 有的采用全卷积网络结构以获取有效的语义上下文信息, 而忽视了网络浅层的空间细节信息; 有的采用 U 型结构, 通过复杂的网络连接利用编码端的空间细节信息, 但没有获取高质量的语义上下文特征. 针对此问题, 本文提出了一种新的基于上下文和浅层空间编解码网络的语义分割解决方案. 在编码端, 采用二分支策略, 其中上下文分支设计了一个新的语义上下文模块来获取高质量的语义上下文信息, 而空间分支设计成反 U 型结构, 并结合链式反置残差模块, 在保留空间细节信息的同时提升语义信息. 在解码端, 本文设计了优化模块对融合后的上下文信息与空间信息进一步优化. 所提出的方法在 3 个基准数据集 CamVid、SUN RGB-D 和 Cityscapes 上取得了有竞争力的结果.

关键词 语义分割, 二分支策略, 语义上下文信息, 浅层空间细节信息, 反 U 型结构

引用格式 罗会兰, 黎宵. 基于上下文和浅层空间编解码网络的图像语义分割方法. 自动化学报, 2022, **48**(7): 1834−1846

DOI 10.16383/j.aas.c190372

Image Semantic Segmentation Method Based on Context and Shallow Space Encoder-decoder Network

LUO Hui-Lan[1] LI Xiao[1]

Abstract Recently, the research on image semantic segmentation basically focuses on how to extract effective semantic context information and restore spatial information in order to get more efficient algorithms. Some models use a fully convolutional network structure to obtain effective semantic context information. This kind of framework does not use the spatial details in the shallow layers of the networks. To effectively restore the spatial details for the decoder, some researches utilize the U-shape structure of complex network connections. But they could not obtain high-quality semantic context features. To better combine context information and space information, a novel semantic segmentation framework is proposed in this paper. A two-branch strategy is adopted on the encoder. One branch is called contextual branch, which is constructed with a proposed semantic context module to obtain high-quality semantic context information. And the other branch is spatial branch, which is designed as an inverse U-shaped structure with the proposed chain-reverse residual module to enhance semantic information and preserve spatial details. Moreover, a refinement module is proposed to add to the decoder to further refine the fusion features of context information and spatial information. The proposed approach achieves competitive results on the CamVid, SUN RGB-D and Cityscapes benchmarks.

Key words Semantic segmentation, two-branch strategy, context semantic information, shallow space details, inverse U-shaped structure

Citation Luo Hui-Lan, Li Xiao. Image semantic segmentation method based on context and shallow space encoder-decoder network. *Acta Automatica Sinica*, 2022, **48**(7): 1834−1846

语义分割是计算机视觉基本任务之一, 其研究目的是为图像中的每一个像素点分配与之相对应的类别标记, 所以可以认为其属于像素级分类. 它主要用在多个具有挑战性的应用领域, 例如: 自动驾驶、医疗图像分割、图像编辑等. 因为语义分割涉及像素级分类和目标定位, 所以如何获取有效的语义上下文信息和如何利用原始图像中的空间细节信息是处理语义分割问题必须考虑的两个因素.

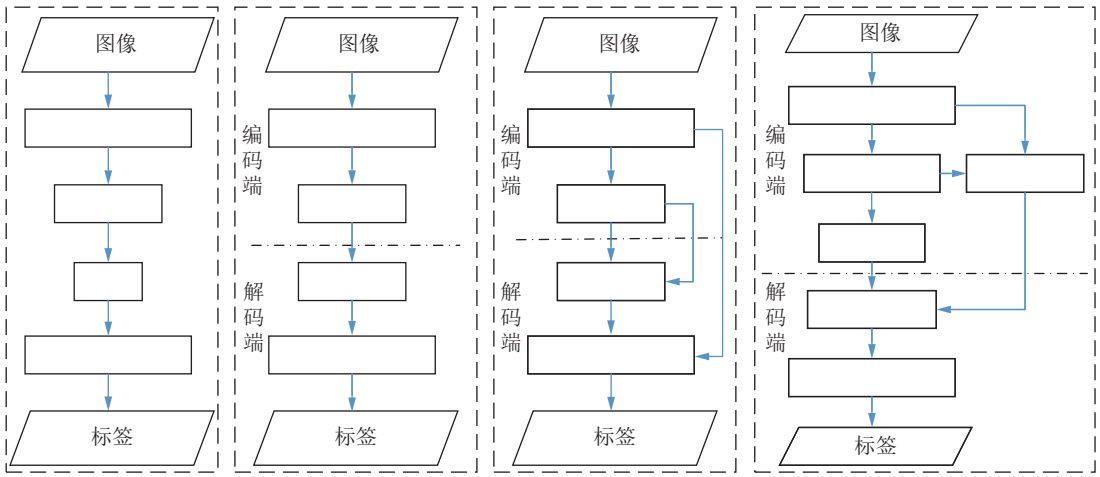
目前, 语义分割最流行的算法是采用类似全卷积网络 (Fully convolutional network, FCN)[1] 的形式, 如图 1(a) 所示, 采用这种形式的分割网络模型是将研究的重点放在提取图像的丰富语义上下文信

收稿日期 2019-05-15 录用日期 2019-10-11
Manuscript received May 15, 2019; accepted October 11, 2019
国家自然科学基金 (61862031, 61462035), 江西省自然科学基金 (20171BAB202014), 江西省主要学科学术和技术带头人培养计划 (20213BCJ22004) 资助
Supported by National Natural Science Foundation of China (61862031, 61462035), Natural Science Foundation of Jiangxi Province (20171BAB202014), and Training Plan for Academic and Technical Leaders of Major Disciplines in Jiangxi Province (20213BCJ22004)
本文责任编委 白翔
Recommended by Associate Editor BAI Xiang
1. 江西理工大学信息工程学院 赣州 341000
1. School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000

息上. 在深度卷积网络中, 感受野大小决定着网络可以获得多大范围的语义上下文信息, 扩张卷积被用来增加网络感受野从而提升分割性能[2−4]. 为了捕捉到图像中不同尺度的目标, PSPNet (Pyramid scene parsing network)[5] 通过空间金字塔方式的全局池化操作来获取多个不同大小的子区域的特征信息, DeepLabV3[6] 则采用空间金字塔方式的扩张卷积. 全卷积网络结构虽然能有效获得语义上下文信息, 但它是通过池化操作或带有步长的卷积来获得, 这会导致空间细节信息的丢失, 从而影响语义分割的精度.

为了弥补空间细节信息的丢失, 许多研究工作采用编码器–解码器结构[7−9], 如图 1(b) 所示. 编码端通常是分类网络, 它采用一系列下采样操作来编码语义上下文信息, 而解码端通过使用上采样处理来恢复空间细节信息. 为了更好地恢复编码过程丢失的空间细节信息, 一些工作[10−12] 采用了 U 型网络结构, 如图 1(c) 所示, LRN (Label refinement network)[10] 和 FC-DenseNet[11] 在解码端通过横向连接的方式, 使用各编码块的特征信息, 联合高层语义信息恢复出图像的空间细节信息, SegNet[12] 则是使用各编码块产生的最大池化索引来辅助解码端上采样特征信息. 这种结构的编码端采用传统的分类网络完成特征提取, 没有显式上下文信息提取模板, 学习到的特征可能缺少语义分割任务所须的属性. 同时, 根据可视化卷积神经网络结构[13], 网络高层特征含有极少空间细节信息, 所以在解码端过度使用编码端高层特征, 不仅不能有效地利用编码端的空间信息, 还会提升网络模型的复杂度以及计算冗余, 不利于分割算法的实时应用.

基于以上分析, 本文提出了一种基于上下文和浅层空间编解码网络的图像语义分割方法, 如图 1(d) 所示. 整个模型采用编解码框架, 本文的动机是在编码端即能获取高质量的语义上下文信息, 同时又能充分保留原始图像中的空间细节信息. 受 BiSeNet[14] 启发, 本文在编码端使用二分支策略, 上下文路径用于获取有效的上下文语义信息, 而空间路径则充分保留图像的空间细节信息, 将语义上下文信息的提取与空间细节信息的保留进行分离. 根据可视化深度卷积神经网络[13], 深度卷积网络的浅层携带大量的空间细节信息, 而高层特征基本不包含空间细节信息. 本文将空间路径设计为反 U 型结构, 这样能将编码网络的浅层和中层特征进行从上到下的融合, 以充分利用编码网络浅中层特征所携带的空间细节信息. 受 MobileNetV2[15] 启发, 本文设计了链式反置残差模块, 对编码网络浅中层特征所携带的空间细节信息进行处理, 达到保留空间信息的同时提升特征的语义表达能力. 在编码网络的上下文路径, 本文设计了语义上下文模块, 它由混合扩张卷积模块和残差金字塔特征提取模块组成. 使用混合扩张卷积模块是为了进一步提升网络感受野, 而残差金字塔特征提取模块可以获取多尺度特征信息. 在解码端, 首先对编码端的空间信息和语义上下文信息进行融合, 受 R2U-Net[16] 启发, 本文设计了带有残差的循环卷积网络优化模块对融合的特征进一步优化, 最后采用可学习的反卷积将优化的分割图还原到原始图像大小.



(a) FCN 结构 (a) Structure of FCN　(b) 编解码结构 (b) Structure of encoder-decoder　(c) U 型结构 (c) Structure of U-shaped network　(d) 本文结构 (d) Structure of ours

图 1　本文提出的网络结构与其他网络结构

Fig. 1　The network structures of our method and other methods

本文主要贡献为:

1) 提出了基于上下文和浅层空间信息结合的编解码网络用于图像语义分割, 即能获取高质量的上下文语义信息又能保留有效的空间细节信息.

2) 为了获得高效的语义上下文信息, 本文组合混合扩张卷积模块和残差金字塔特征提取模块, 以提升网络感受野以及获取周围特征信息和多尺度特征信息; 对于浅层空间信息的使用, 本文设计了反 U 型结构的空间路径以利用编码端浅中层特征所携带的大量空间信息. 针对编码端不同层的特征差异, 在空间路径中设计了链式反置残差模块以保留空间细节信息并提升特征的语义表达能力, 这样不仅可以弥补高层语义信息中丢失的位置信息还使得模型轻量化.

3) 本文设计了残差循环卷积模块, 对语义特征和空间信息融合后的分割特征进一步优化, 提升分割性能. 本文方法在 3 个基准数据集 CamVid、SUN RGB-D 和 Cityscapes 上取得了有竞争力的结果.

1 相关工作

1.1 编码器–解码器

编码器–解码器网络已经在语义分割任务中得到广泛应用, 它由编码器模块和解码器模块组成, 编码器模块逐渐减小特征图来编码高层语义信息, 而解码器模块是逐渐恢复空间细节信息. ENet[7] 没有使用任何来自于编码端的信息, 直接在解码端恢复空间信息. 文献 [11−12, 17] 则是通过跳层连接在解码端使用编码端特征恢复空间信息. G-FRNet[18] 通过门控机制对编码端各相邻模块特征进行门控选择, 再用于解码端恢复出空间信息. 为了更有效地利用编码端空间信息, 本文设计了一种反 U 型结构的空间路径, 以充分利用编码端浅中层特征信息中所携带的空间细节信息, 有效地提升分割性能.

1.2 二分支结构

最近的一些研究工作 [14, 19−20] 采用二分支网络结构, 即将空间信息保留和上下文信息提取放在网络中的不同分支. 网络的深层分支采用可分离卷积等轻量化操作来获取语义上下文信息, 浅层分支采用简单的卷积操作以保留有效的空间细节信息. 这种结构的网络模型更轻量化, 推动了语义分割的实时应用, 但很难提取到有效的语义上下文信息, 再加上两种特征信息差距较大, 进行融合并不能产生很好的效果. 为此, 本文使用参数适中、分类表达能力较好的 ResNet-34[21] 作为骨干网络, 结合本文的语义上下文模块以获取更强的语义表达能力. 对于空间细节信息, 本文充分利用编码端浅层和中层特征中所携带的空间细节信息, 而不使用高层来恢复空间细节信息, 这样能高效地使用编码端有用的空间细节信息, 并且有利于与语义上下文信息融合, 同时减小了模型复杂度.

1.3 空间金字塔

空间金字塔模块是一种学习上下文语义信息的有效模块, 利用平行的空间金字塔池化以捕获多尺度的语义信息, 成功用于不同计算机视觉任务中, 如目标检测和语义分割等. PSPNet[5] 以空间金字塔方式的池化操作, 来获取特征图中不同子区域的全局信息. 由于池化是一种下采样操作, 会严重丢失特征信息, DeepLabV3[6] 以并联的方式利用不同扩张率的扩张卷积来获取多尺度上下文信息, 有效提升了分割性能, 但由于使用的扩张卷积的扩张率非常大, 最大为 18, 这使得扩张卷积稀疏化严重, 提取到的特征缺少细节信息. 与他们利用空间金字塔结构的方式不同, 本文设计了混合扩张卷积模块和残差金字塔特征提取模块, 先使用混合扩张卷积模块提升网络感受野, 在获取上下文信息的同时减少扩张卷积稀疏化, 再使用残差金字塔特征提取模块, 以并联较小扩张率的扩张卷积来获取多尺度信息, 避免特征信息的丢失, 即可以增加网络感受野又可以获取高质量的多尺度特征信息.

2 本文提出的方法

本文提出了一种新的语义分割方法, 称为基于上下文和浅层空间编解码网络的图像语义分割方法, 能够学习丰富的上下文语义信息以及获取更加有效的空间信息. 在解码端采用了优化模块进一步优化语义上下文信息与空间信息的融合特征, 帮助解码端恢复更加精准的像素级预测分割图. 整个网络结构如图 2 所示. 第 2.1 节对提出的网络结构进行整体阐述, 第 2.2 节论述其中的混合扩张卷积模块 (Hybrid atrous convolution block, HAB), 第 2.3 节对残差金字塔特征提取模块 (Residual pyramid feature block, RPB) 进行论述, 第 2.4 节阐述了链式反置残差模块 (Chain inverted residual block, CRB), 最后, 第 2.5 节论述了残差循环卷积模块 (Residual recurrent convolution block, RRB).

2.1 网络结构

本文网络模型采用编解码网络框架, 在编码端采用两分支方式分别获取有效的高层上下文信息和

低层的空间信息. 由于深度卷积网络随着网络层数的不断加深会产生梯度消失或爆炸的现象, 这不利于深度卷积网络的学习和训练, 而 ResNet[21] 网络通过在每个模块之间添加跳层连接方式, 避免了梯度消失问题同时加速了网络的收敛. 所以在编码端, 本文的骨干网络使用了在 ImageNet[22] 数据集上预训练的 ResNet-34, 去除了最大池化层和全连接层以适应语义分割任务. 为了区分 ResNet-34 的层级特征, 本文将 ResNet-34 分为 5 个模块, 其结构如图 3 所示, 用 Conv、Block 1 表示浅层, Block 2 表示中层, Block 3 和 Block 4 表示高层和特高层特征提取模块, 浅层和中层特征用于空间信息提取路径, 而高层特征作为语义上下文信息提取模块的输入特征. 为了提升网络的感受野, 本文将 ResNet-34 网络的后两个模块 Block 3 和 Block 4 中的普通卷积替换为扩张卷积, 这里扩张卷积与普通卷积具有相同的参数, 扩张率分别为 2 和 4. 在骨干网络 ResNet-34 中, 除 Block 1 外, 其他各模块存在一个步

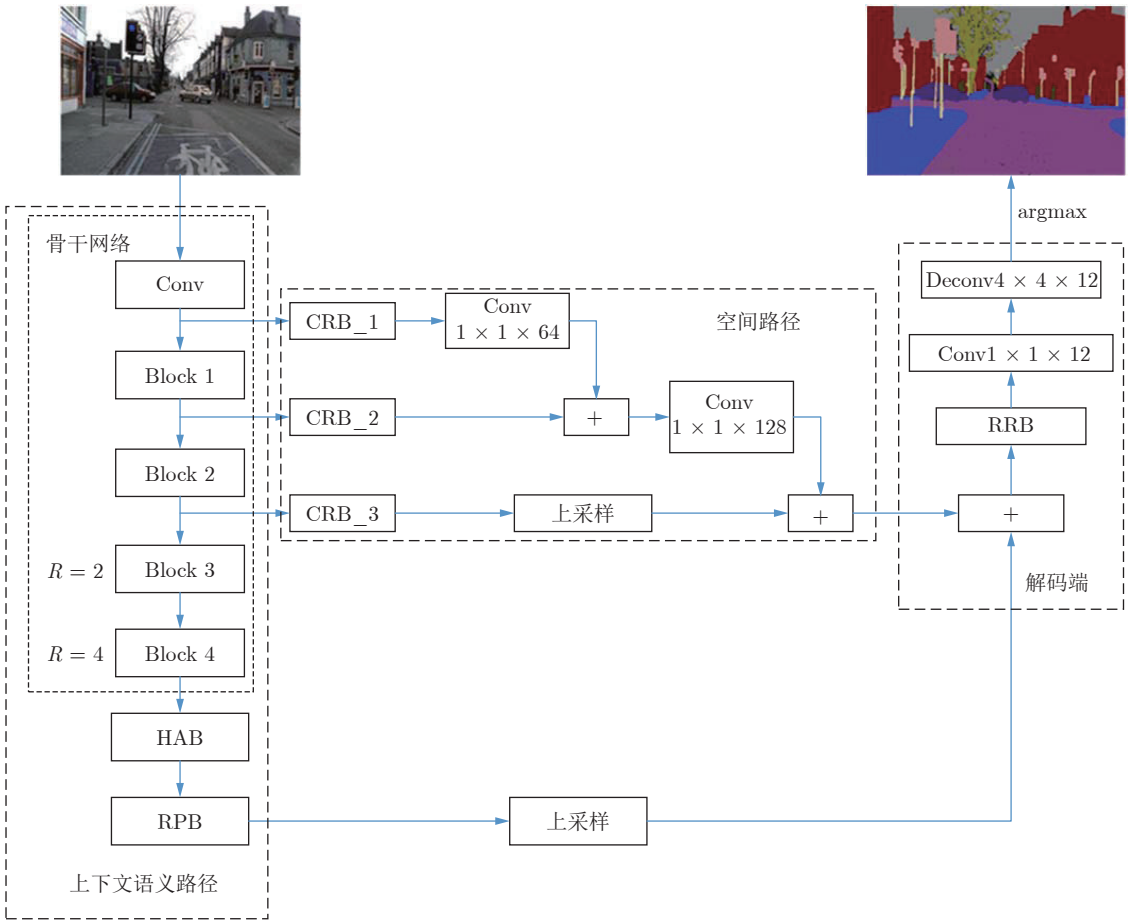


图 2　本文网络框架 (HAB: 混合扩张卷积模块; RPB: 残差金字塔特征提取模块; CRB: 链式残差模块; RRB: 残差循环卷积模块; Deconv: 转置卷积; R: 扩张率)

Fig. 2　The network framework of our method (HAB: hybrid atrous convolution block; RPB: residual pyramid feature block; CRB: chain inverted residual block; RRB: residual recurrent convolution block; Deconv: transposed convolution; R: atrous rate)

Conv	Block 1	Block 2	Block 3	Block 4
$7 \times 7 \times 64$	$\begin{pmatrix} 3 \times 3 \times 64 \\ 3 \times 3 \times 64 \end{pmatrix} \times 3$	$\begin{pmatrix} 3 \times 3 \times 128 \\ 3 \times 3 \times 128 \end{pmatrix} \times 4$	$\begin{pmatrix} 3 \times 3 \times 256 \\ 3 \times 3 \times 256 \end{pmatrix} \times 6$	$\begin{pmatrix} 3 \times 3 \times 512 \\ 3 \times 3 \times 512 \end{pmatrix} \times 3$

图 3　ResNet-34 骨干网络结构

Fig. 3　The backbone structure of ResNet-34

长为 2 的卷积, 使得骨干网络最终输出的特征图大小为输入图像的 1/16.

为了获取高质量的语义上下文信息, 本文设计了语义上下文信息模块, 它由混合扩张卷积模块和残差金字塔特征提取模块组成. 在利用空间信息方面, 与 U 型结构不同的是, 本文没有利用编码端不含有细粒度空间信息的高层特征来恢复空间信息, 这样不仅高效利用了编码端空间细节信息还节省了模型内存开销, 也不需要像 ContextNet[19] 一样去另外设计空间信息获取路径, 而是共享编码端浅、中层特征, 使得获取的空间信息最有效. 本文将空间信息路径设计为反 U 型结构, 结合本文设计的链式反置残差模块, 在保留浅层空间信息的同时提升特征的语义信息. 在解码端, 将编码端的高级语义上下文信息进行双线性上采样与空间细节信息以逐像素点求和方式进行融合, 再对融合的特征进一步优化, 设计了残差循环卷积网络优化模块, 最后对得到的分割图使用转置卷积恢复出原始图像大小.

为了使网络有效地收敛, 与 PSPNet[5] 和 BiSeNet[14] 类似, 本文在上下文语义路径的末端加入监督信息, 即引入额外的辅助损失函数对上下文语义路径产生的初始分割结果进行监督学习. 辅助损失函数和最终分割结果的主损失函数均是使用多元交叉熵损失函数, 如式 (1), 其中 $softmax$ 为 $softmax\left(z_i\right)=\mathrm{e}^{z_i}/\sum_j \mathrm{e}^{z_j}$ 函数, $pred$ 是预测分割图, Y 是真值分割图, $Cost$ 表示多元交叉熵损失函数, 其定义如式 (2) 所示, 其中 N 是样本数.

$$Loss\left(pred, Y\right)=Cost\left(softmax\left(pred\right), Y\right) \tag{1}$$

$$Cost=-\frac{1}{N}\sum_i((1-Y)\times\log(1-softmax(pred))+ Y\times\log\left(softmax\left(pred\right)\right)) \tag{2}$$

网络训练时, 总的损失函数如式 (3) 所示, $loss1$ 是主损失函数, $loss2$ 是辅助损失函数, 引入辅助损失函数有助于优化学习过程, 并且为辅助损失函数添加权重因子 α 来平衡辅助损失与主损失函数对网络的表达能力. 本文实验中将权重因子设为 0.05.

$$Loss=loss1+\alpha\times loss2 \tag{3}$$

2.2 混合扩张卷积模块

扩张卷积根据扩张率在卷积核中相邻两个权值之间插入相应的零, 因此通过增加扩张率可以增大卷积核对特征图的局部计算区域, 从而可以识别更大范围的图像特征信息. 扩张卷积在二维信号中的定义如式 (4), 其中输入特征图 $x(m, n)$ 与卷积核 $w(i, j)$ 进行卷积操作产生输出 $y(m, n)$, 卷积核的长度和宽度为 M、N, r 是扩张率, 后面部分出现的 r 都表示扩张率, 它控制卷积核对输入 x 的采样大小, 这相当于在卷积核中相邻两个权值之间插入 $r-1$ 个零. 当 $r=1$ 时就是普通卷积, 扩张卷积通过修改扩张率可以自适应地改变感受野大小. 不同扩张率的扩张卷积如图 4 所示. 相比于传统卷积, 扩张卷积在没有增加网络参数的情况下就可以获得更大的感受野. 扩张卷积是一种稀疏计算, 即当扩张率很大时, 卷积核的参数数量没有变化, 但对特征图的作用区域却很大, 这就导致扩张卷积从特征图中提取到的有用信息量很少, 从而使扩张卷积失去了建模能力.

$$y\left(m, n\right)=\sum_{i=1}^{M}\sum_{j=1}^{N}x\left(m+r\times i, n+r\times j\right)w\left(i, j\right) \tag{4}$$

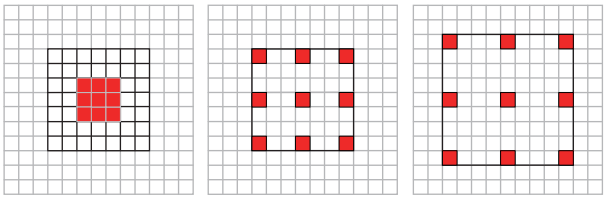


图 4 3 种不同扩张率的扩张卷积, 从左到右分别为 $r=1, 3, 4$

Fig. 4 Illustrations of the atrous convolution with three different atrous rates, $r=1, 3, 4$

本文提出的混合扩张卷积模块的设计动机是获取像素点周围特征信息的同时可以提升网络感受野, 并且减少特征信息的丢失. 根据 CGNet[23], 通过融合小感受野和大感受野特征能够获取同一像素点的周围特征信息. 受 Inception-v4[24] 启发, 本文提出了混合扩张卷积模块, 通过混合叠加的方式来获取周围特征信息以及增加网络感受野. 如图 5 所示, 整个模块分为两个分支, 首先特征图通过一个 1×1 的卷积, 目的是减少特征通道数, 从而减少网络参数. 然后, 一个分支通过 3×3 的卷积, 另一分支进入 5 种不同的扩张卷积: 扩张率为 2 的 3×3 卷积层、扩张率为 4 的 3×3 卷积层以及扩张率为 3 的 5×5 卷积层进行融合, 再融合扩张率为 2 的 5×5 卷积层与扩张率为 2 的 7×7 卷积层, 目的是获取像素点的周围特征信息. 最后将两个分支进行融合, 最终可以获得周围特征信息和大感受野的同时, 信息丢失较少. 这里每一个卷积层后面跟着批归一化处理 (Batch normalization)[25] 和 $Relu(x)=\max(0,\ x)$ 激活函数.

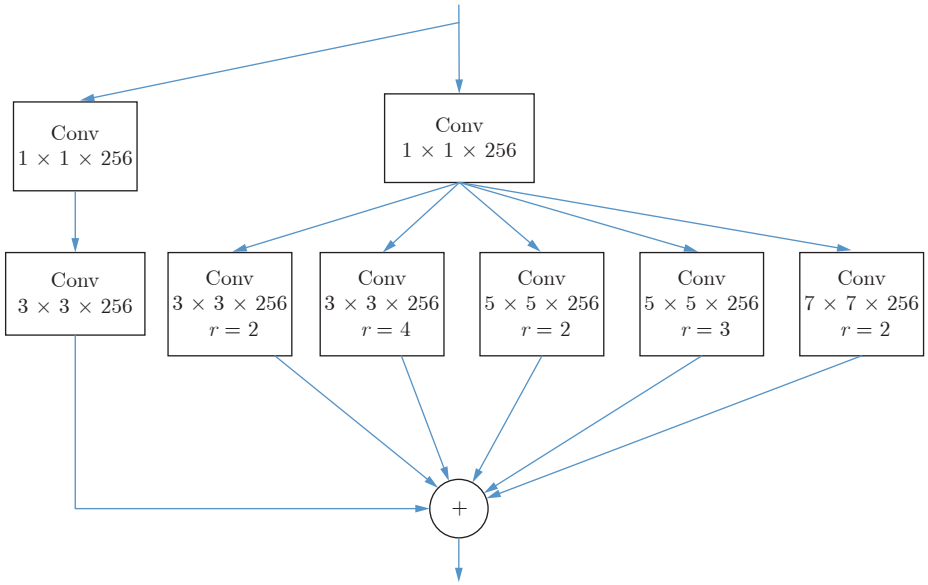


图 5　混合扩张卷积模块

Fig. 5　Hybrid atrous convolution block

2.3　残差金字塔特征提取模块

在语义分割场景中, 物体大小存在多样性, 如果使用单一尺寸的图像特征, 可能丢失图像中小物体或不显著物体的特征信息. 为了分割不同尺度的物体, 本文提出残差金字塔特征提取模块来获得具有判别力的多尺度特征, 通过使用多个不同感受野大小的扩张卷积来提取不同尺度的图像特征信息, 从而识别不同大小的物体. 本文采用了 4 个不同扩张率的扩张卷积, 它们的扩张率分别为: 2, 3, 5, 7. 同时, 为了利用全局场景上下文信息, 本文将全局池化操作扩展到扩张卷积空间金字塔池化中.

本文提出的残差金字塔特征提取模块结构如图 6 所示, 输入特征首先进入扩张卷积金字塔模块, 它由 4 个不同扩张率的扩张卷积和全局平局池化以并联的方式组成, 其中 4 个扩张卷积输出特征通道数都相同. 对全局平均池化的结果进行 1×1 卷积操作和双线性上采样操作, 使其与扩张卷积的输出大小相同. 然后对它们进行拼接操作以获取多尺度特征信息. 最后与残差进行融合, 提升语义表达能力的同时加速梯度反向传播.

2.4　链式反置残差模块

本文提出链式反置残差模块构造从编码端到解码端的空间路径, 实现原始图像的空间信息与高层语义上下文信息融合. 考虑到高层特征已经不包含细粒度的空间信息, 本文将空间路径设计为反 U 型结构, 只将含有丰富空间信息的低中层特征与语义信息进行融合. 受 MobileNetv2[15] 启发, 本文提出的链式反置残差模块的结构如图 7 所示. 每个链式反置残差模块将多个反置残差结构以链式结构相结合, 目的是保留空间信息的同时提升特征图的语义表达能力. 反置残差结构由两个 1×1 点级卷积层和一个 3×3 分组卷积层组成. 输入特征首先进入 1×1 点级卷积层来增加特征通道数, 再进入 3×3 的分组卷积层, 其中分组数等于输入通道数, 最后经过一个 1×1 点级卷积层来降低特征通道数. 需要注意的是本文所用的 3 个链式残差模块的链长不一样, 如图 1 所示, 连接低层特征的 CRB_1 的链长为 3, 即由 3 个反转残差结构链接而成, CRB_2 的链长为 2, 而连接中层特征的 CRB_3 链长为 1. 通过链长的不同设置, 可以有针对性地提升浅层特征的语义表达能力. 反置残差结构使用点级卷积和分组卷积, 将通道操作和空间操作进行分离, 避免通道操作对空间信息的影响. 分组卷积与普通卷积相比, 参数量也更少. 同时, 设计了残差学习以避免梯度消失和爆炸. 整个链式残差优化模块可以抽象为

$$L_{l+1} = f(L_l) + L_l \tag{5}$$

其中, $f(\cdot)$ 表示反置残差块的函数形式, 从函数形式中可以发现, 下一层特征信息除了与反转残差模块有关, 还与上一层特征信息相关, 这样既可以保留空间信息又可以提升特征的语义信息.

2.5　残差循环卷积模块

在解码端, 需要将编码端产生的高级语义信息

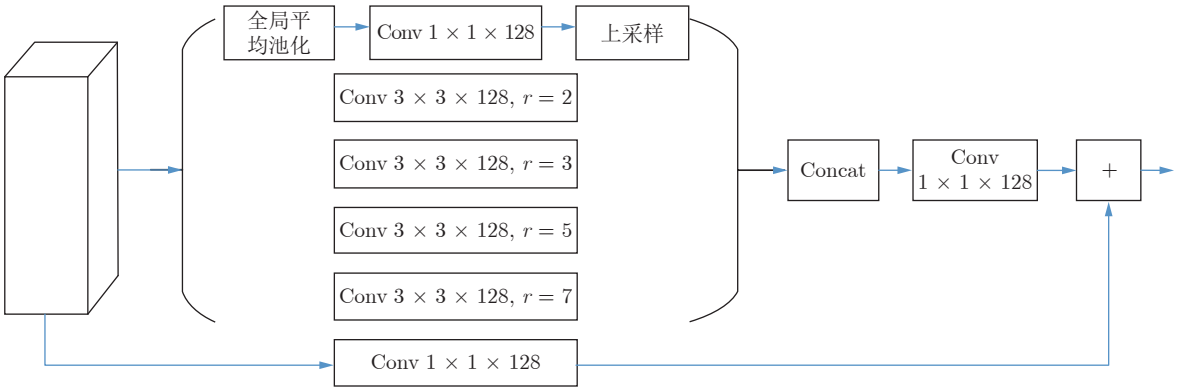


图 6 残差金字塔特征提取模块

Fig. 6 Residual pyramid feature block

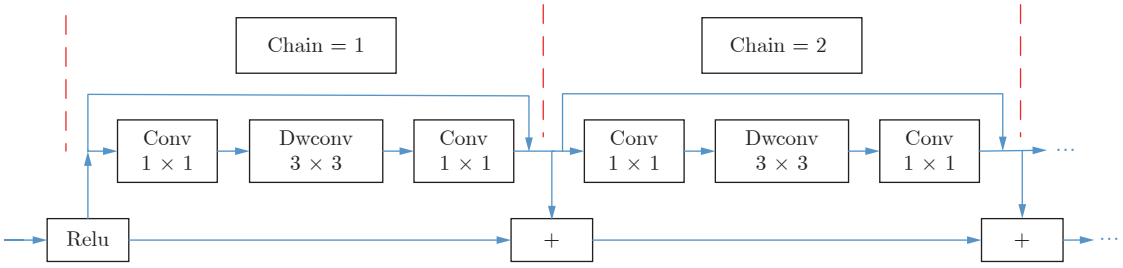


图 7 链式反置残差模块

Fig. 7 Chain inverted residual block

与空间细节信息进行融合, 本文以简单的求和方式进行融合. 为了对融合后的特征进一步优化, 本文设计了优化模块. 如图 8 所示, 优化模块由两个 3×3 的循环卷积网络以及残差组成, 其中每个 3×3 循环卷积都含有批归一化处理和 Relu 激活函数. 循环卷积网络有助于特征积累, 相当于一个自学习的过程, 用于提升网络的表达能力. 除此之外, 循环卷积相当于对卷积层的重复利用, 减少了参数量. 整个模块在提升语义识别能力的同时保留了空间信息. 再是使用残差结构加速网络的信息流动, 同时有助于梯度的反向传播. 优化模块可以抽象为式 (6), 其中, $f(\cdot)$ 为循环卷积的函数表示.

$$x_{l+1} = f\left(f\left(x_l\right) + x_l\right) \tag{6}$$

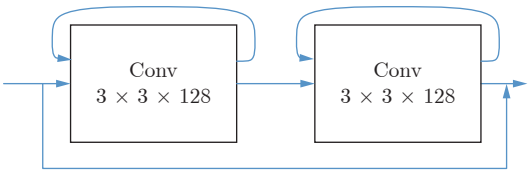


图 8 残差循环卷积模块

Fig. 8 Residual recurrent convolution block

3 实验

本文提出的方法在 3 个基准数据集上进行评估: CamVid[26]、SUN RGB-D[27] 和 Cityscapes[28]. 实验环境配置: 操作系统为 64 位的 Windows10、CPU 为 Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60 GHz, 内存为 512 GB, 显卡是 16 GB 的 NVIDIA Tesla P100-PCIE. 本文实验系统实现是基于深度学习开源框架 Pytorch.

3.1 实验设置

除预训练的 ResNet-34 外, 网络的语义上下文模块参数初始化是基于 Kaiming[29] 初始化方法. 对于训练过程, 本文使用 Adam[30] 优化器来优化网络模型参数, 按照之前的工作 PSPNet[5] 和 DeepLabV3[6], 采用 Poly 学习率策略动态改变网络学习率大小. Poly 学习策略的定义是 $lr = initial\_lr \times \left(\dfrac{iter}{max\_iter}\right)^{\text{power}}$, 其中, $initial\_lr$, max_iter, $iter$ 分别表示初始学习率、最大迭代次数、当前迭代次数, 初始化学习率设为 0.0001, power 设为 0.9, 在 CamVid 数据集上训练迭代 150 次, 在 SUN RGB-D 数据集上训练迭代 80 次, 在 Cityscapes 数据集上训练迭代 250 次.

本文使用像素级交叉熵损失函数作为目标函数来优化网络参数, 同时忽略未标记的像素点.

本文使用当前最有效且普遍使用的平均交并比 (Mean of class-wise intersection over union, MIoU) 作为语义分割评估指标[1], 它用于计算预测分割图与真实标记图之间的相似度, 也就是通过计算预测分割图与真实标记图之间的交集除于它们的并集, 计算式为

$$MIoU = \frac{1}{n_c}\sum_i\left(\frac{n_{ii}}{t_i + \sum\limits_{j=1} n_{ji} - n_{ii}}\right),\ t_i = \sum_j n_{ij} \tag{7}$$

其中, n_c 表示图像中包含的类别总数, n_{ij} 表示实际类别为 i 而被预测为类别 j 的像素点数目, t_i 表示实际类别为 i 的像素点数目. $MIoU$ 的取值范围为 [0, 1], 当 $MIoU$ 的值越大, 说明预测分割图与真值标记图的重叠部分越大, 也即预测的分割图越准确.

3.2　CamVid 数据集上的结果

CamVid 数据集是自动驾驶领域中的道路场景数据集, 这个数据集包含 376 幅训练图片、101 幅验证图片、233 幅测试图片, 图片分辨率为 360×480 像素, 共有 11 个语义类别. 按照 ENet[7], 本文使用带权类别交叉熵损失以弥补数据集中小数目的类别, 即为每个语义类别分配不同的权重, 用于解决 CamVid数据集中类别不平衡问题. 类别权重计算如式 (8), 其中, $class$ 表示类别, p_{class} 表示类别 $class$ 在图像像素中出现的概率, t 是超参数, 这里设置为 1.02.

$$W_{\text{class}} = \frac{1}{\ln\left(t + p_{\text{class}}\right)} \tag{8}$$

本文方法在 CamVid 测试集上的结果与当前分割方法的比较如表 1 所示, 本文方法在测试时没有采用后置处理以及一些测试技巧, 像多尺度. 从实验结果可以看出, 本文方法比使用 U 型结构和二分支结构的语义分割方法的性能要好, 说明本文方法能够获取高质量的上下文语义特征和有效使用浅层的空间细节信息. 从图 9 中的实验效果图可以看出, 本文方法基本能够准确识别图像中物体位置并且分割出物体, 而 SegNet 在第 1 行第 3 列的分割图中未能识别出路灯; CGNet 在第 2 行第 4 列的分割图中将建筑物错误识别为树木以及 BiSeNet (xception) 在第 2 行第 5 列的分割图中将道路错误识别为汽车类型以及对远距离小物体路灯未能识别出来.

为了比较本文方法的效率, 在表 2 中比较了本文模型与其他方法的参数量, 以及在 NVIDIA Tesla P100 显卡上测试的处理速度. 如表 2 中结果所示, 相比 SegNet, 本文方法在分割精度上有很大的提升. 对比 BiSeNet (xception), 本文方法在分割精度上取得了明显的提升. 相比 BiSeNet (ResNet18), 本文方法参数量更少.

3.3　SUN RGB-D 数据集上的结果

SUN RGB-D 是一个非常大的室内场景数据集, 它包含 5 285 幅训练图片和 5 050 幅测试图片, 并且有 37 个室内物体类别, 如墙壁、地板、桌子、椅子、沙发、床等. 由于图片中的物体有多个不同的形状、大小以及摆放的位置, 这些因素对语义分割来说是一个很大的挑战. 本文只使用了 RGB 数据而没有用深度信息. 在实验过程中, 本文从训练集中抽出 1 000 幅图片作为验证数据集, 用于检测模型性能并选择泛化能力最好的训练模型.

本文方法在 SUN RGB-D 测试集上的实验结果与当前分割方法的比较如表 3 所示. 可以看出,

表 1　本文方法与其他方法在 CamVid 测试集上的 MIoU 比较 (%)

Table 1　Comparison of MIoU between our method and the state-of-the-art methods on the CamVid test set (%)

方法	Tree	Sky	Building	Car	Sign	Road	Pedestrian	Fence	Pole	Sidewalk	Bicyclist	MIoU
FCN-8[1]	–	–	–	–	–	–	–	–	–	–	–	52.0
DeconvNet[8]	–	–	–	–	–	–	–	–	–	–	–	48.9
SegNet[12]	52.0	87.0	68.7	58.5	13.4	86.2	25.3	17.9	16.0	60.5	24.8	50.2
ENet[7]	77.8	95.1	74.7	82.4	51.0	95.1	67.2	51.7	35.4	86.7	34.1	51.3
Dilation[2]	76.2	89.9	82.6	84.0	46.9	92.2	56.3	35.8	23.4	75.3	55.5	65.29
LRN[10]	73.6	76.4	78.6	75.2	40.1	91.7	43.5	41.0	30.4	80.1	46.5	61.7
FC-DenseNet103[11]	77.3	93.0	83.0	77.3	43.9	94.5	59.6	37.1	37.8	82.2	50.5	66.9
G-FRNet[18]	76.8	92.1	82.5	81.8	43.0	94.5	54.6	47.1	33.4	82.3	59.4	68.0
BiSeNet (xception)[14]	74.4	91.9	82.2	80.8	42.8	93.3	53.8	49.7	31.9	81.4	54.0	65.6
CGNet[23]	–	–	–	–	–	–	–	–	–	–	–	65.6
本文方法	75.8	92.4	81.9	82.2	43.3	94.3	59.0	42.3	37.3	80.2	**61.3**	**68.26**

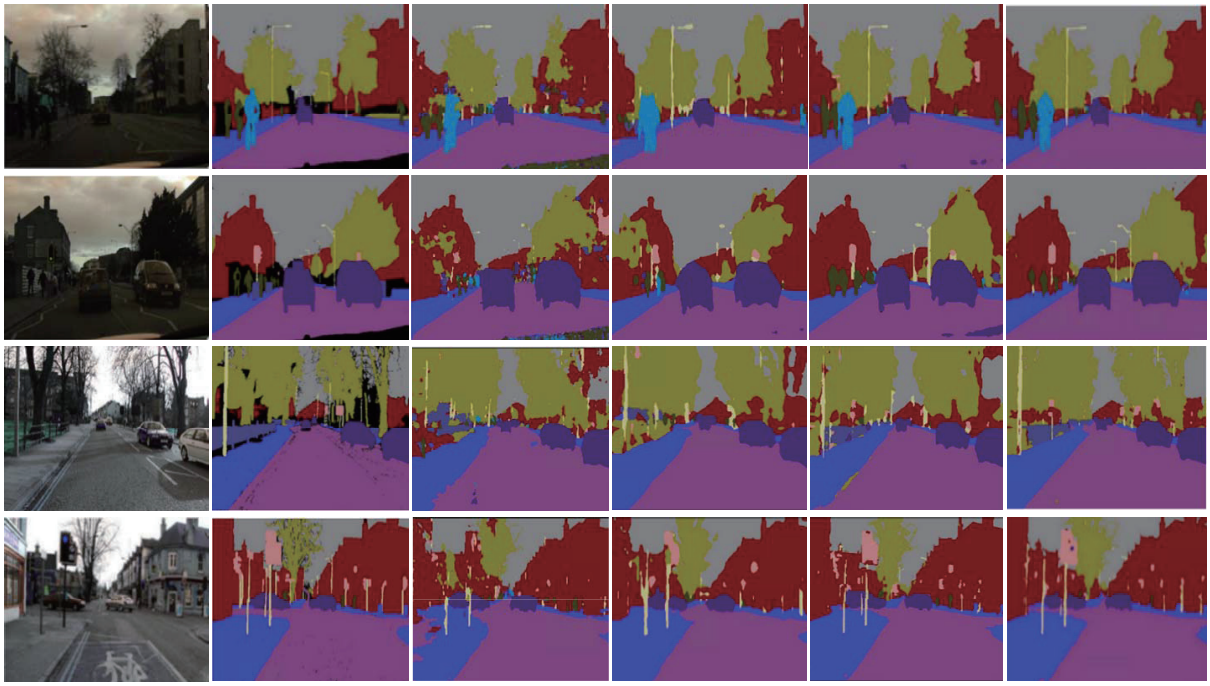


图 9 在 Camvid 测试集上本文方法与 SegNet[12]、CGNet[23] 和 BiSeNet (xception)[14] 方法的定性比较

Fig. 9 Qualitative comparisons with SegNet[12], CGNet[23] and BiSeNet (xception)[14] on the CamVid test set

表 2 在 CamVid 数据集上的性能比较

Table 2 Performance comparisons of our method and the state-of-the-art methods on the CamVid dataset

方法	参数量 (MB)	运行时间 (ms)	帧速率 (帧/s)	MIoU (%)
SegNet[12]	29	23.2	42	50.2
BiSeNet (xception)[14]	5.8	12.1	82	65.6
BiSeNet (ResNet18)[14]	49	64.8	15	68.7
本文方法	31	39.5	25	68.26

表 3 本文方法与其他方法在 SUN RGB-D 测试集上的 MIoU 比较 (%)

Table 3 Comparison of MIoU between our method and the state-of-the-art methods on the SUN RGB-D test set (%)

方法	MIoU
FCN-8[1]	27.4
DeconvNet[8]	22.6
ENet[7]	19.7
SegNet[12]	31.8
DeepLab[4]	32.1
本文方法	**40.79**

本文方法在 SUN RGB-D 数据集上有比较大的性能提升, 从而验证了本文方法的有效性. 可视化分割结果如图 10 所示, 可以发现本文方法可以有效地分割图像中的物体.

3.4 Cityscapes 数据集上的结果

Cityscapes 数据集是一个高分辨率城市道路场景分析数据集, 每幅图像的分辨率为 2 048×1 024 像素, 具有 19 个语义类别, 包含 5 000 幅高质量标记图和 2 万幅粗略标记图, 本文只使用具有 5 000 幅精准标记图的图像用于实验. 按照官方标准, 这个数据集分成 3 个子集, 2 975 幅训练集、500 幅验证集、1 525 幅测试集. 在训练时, 本文将图像裁剪为 512×768 像素大小, 然后在验证集上评估性能, 并将测试集上得到的结果提交到官方评估系统中.

本文方法在 Cityscapes 测试集上的实验结果与当前分割方法的比较如表 4 所示, 可以看出本文取得了有竞争的分割结果, 本文没有采用任何测试技巧, 像 PSPNet 中多尺度. 虽然 PSPNet 取得了最好的分割效果, 但其使用 ResNet101 作为骨干网络, 这导致它的网络复杂度最高, 达到了 65 MB 参数量, 在一般设备中基本无法运行. 本文采用参数量适中的 ResNet34 作为骨干网络并取得较好的性能, 虽然 BiSeNet (ResNet18) 的骨干网络只采用了 ResNet18, 但它使用了多尺度训练的前置处理, 还使用了通道注意力机制模块来优化语义特征, 所

图 10　本文方法在 SUN RGB-D 测试集上的定性结果

Fig. 10　Qualitative results of our method on the SUN RGB-D test set

表 4　本文方法与其他方法在 Cityscapes 测试集上的比较

Table 4　Comparisons of our method with the state-of-the-art methods on the Cityscapes test set

方法	参数量 (MB)	MIoU (%)
FCN-8[1]	134.5	65.3
ENet[7]	0.4	58.3
SegNet[12]	29.5	56.1
DeepLab[4]	44.04	70.4
Dilation[2]	–	67.1
PSPNet[5]	65.7	78.4
CGNet[23]	0.5	64.8
BiSeNet (xception)[14]	5.8	68.4
BiSeNet (ResNet18)[14]	49	74.7
本文方法	31	**73.1**

以也取得了比本文更好一些的性能. 而且由于本文提出的模型使用了分组卷积和点级卷积等轻量型模块, 参数量比 BiSeNet (ResNet18) 更少. 图 11 是在验证集上的可视化分割图效果, 可以看出本文方法基本可以准确地分割图像中的物体.

3.5　消融实验

为了验证所提出模块的有效性, 本小节对所提出的方法在 CamVid 数据集上进行了消融实验.

3.5.1　验证混合扩张卷积模块和残差金字塔特征提取模块的有效性

本文采用 4 种方案来评估混合扩张模块和残差金字塔特征提取模块性能: 1) 在编码端的上下文路径只使用混合扩张卷积模块; 2) 在编码端的上下文路径只使用残差金字塔特征提取模块; 3) 在编码端上下文路径没有混合扩张卷积和残差金字塔特征提取模块; 4) 在编码端上下文路径使用混合扩张卷积模块和残差金字塔特征提取模块. 实验结果如表 5 所示, 从表中可以看出, 同时使用混合扩张卷积模块和残差金字塔特征提取模块时获得的分割性能最好, 说明这两个模块能获取有效的周围特征信息以及多尺度特征, 从而提升网络分割性能.

3.5.2　验证混合扩张卷积的有效性

本文采用 4 种方案来评估混合扩张卷积的有效性: 1) 只使用所有分支扩张率为 1 的混合扩张卷积模块; 2) 只使用分支扩张率分别为 2, 3, 4 的混合扩张卷积模块; 3) 所有分支扩张率为 1 的混合扩张卷积模块加残差金字塔特征提取模块; 4) 本文方法, 即分支扩张率分别为 2, 3, 4 的混合扩张卷积模块加残差金字塔特征提取模块. 实验结果如表 6 所示, 只使用所有分支扩张率为 1 的混合扩张卷积模块虽然比只使用分支扩张率分别为 2, 3, 4 的混合扩张卷积模块分割性能好, 但加入残差金字塔特征提取模块后, 本文设计的分支扩张率分别为 2, 3, 4 的混合扩张卷积模块加残差金字塔特征提取模块的模型获得了最好的效果.

3.5.3　验证混合扩张卷积模块与残差金字塔特征提取模块的结构顺序

本文采用 4 种方案来验证上下文模块顺序: 1)

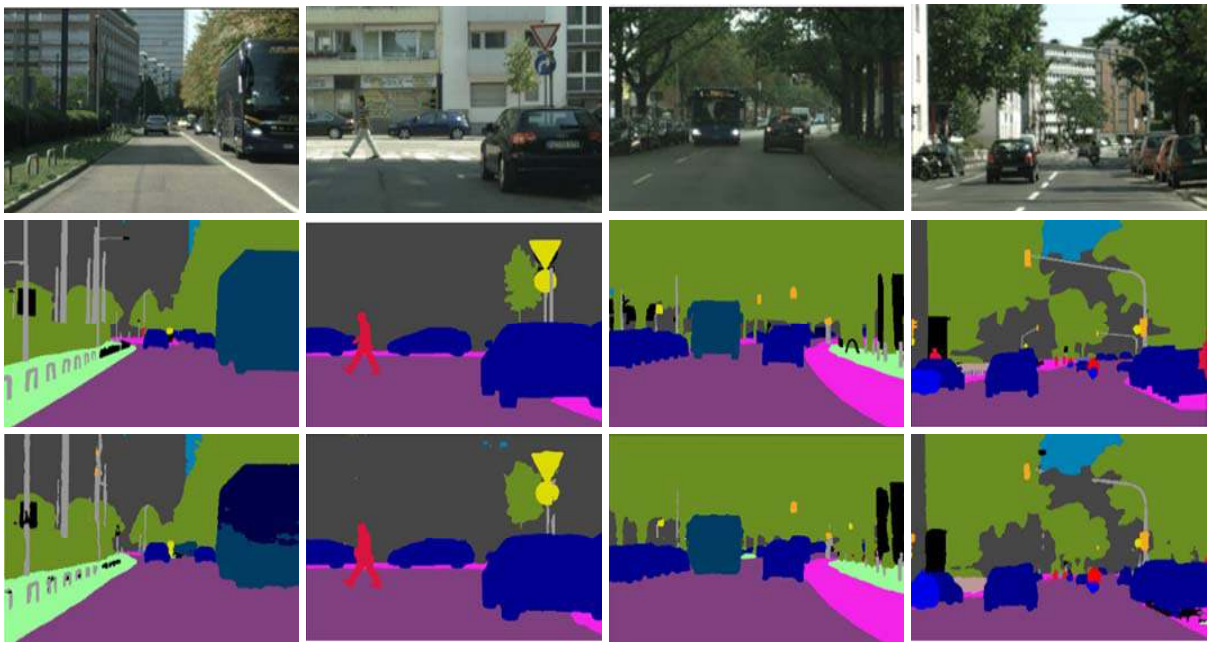


图 11 本文方法在 Cityscapes 验证集上的定性结果

Fig. 11 Qualitative results of our method on the Cityscapes val set

表 5 混合扩张卷积和残差金字塔特征提取模块对性能的影响 (HAB: 混合扩张卷积模块; RPB: 残差金字塔特征提取模块)

Table 5 The influence of HAB and RPB on performance (HAB: hybrid atrous convolution block; RPB: residual pyramid feature block)

HAB	RPB	MIoU (%)
×	×	66.57
√	×	66.22
×	√	67.51
√	√	**68.26**

表 6 混合扩张卷积模块对性能的影响 (HAB: 混合扩张卷积模块; RPB: 残差金字塔特征提取模块)

Table 6 The influence of HAB on performance (HAB: hybrid atrous convolution block; RPB: residual pyramid feature block)

HAB	RPB	HAB各分支扩张率	MIoU (%)
√	×	2, 3, 4	66.22
√	×	1	67.84
√	√	1	68.16
√	√	2, 3, 4	**68.26**

混合扩张卷积模块加混合扩张卷积模块; 2) 残差金字塔特征提取模块加残差金字塔特征提取模块; 3) 残差金字塔特征提取模块加混合扩张卷积模块; 4) 本文方法, 即混合扩张卷积模块加残差金字塔特征提取模块. 实验结果如表 7 所示, 在语义上下文特征提取模块中同时使用混合扩张卷积模块和残差金字塔特征提取模块能够获得最佳的性能, 由于残差金字塔特征提取模块的拼接操作会影响处理速度, 从处理速度和性能上, 混合扩张卷积模块加残差金字塔特征提取模块的组合更加有效.

表 7 混合扩张卷积模块与残差金字塔特征提取模块的结构顺序对性能的影响 (HAB: 混合扩张卷积模块; RPB: 残差金字塔特征提取模块)

Table 7 The influence of the structural order of HAB and RPB on performance (HAB: hybrid atrous convolution block; RPB: residual pyramid feature block)

方法	MIoU (%)
HAB+HAB	67.29
RPB+RPB	66.95
RPB+HAB	68.29
本文 (HAB+RPB)	68.26

3.5.4 验证空间路径的有效性

本文采用了 5 种方案来评估编码端空间路径的有效性: 1) 没有空间路径; 2) 使用编码端骨干网络浅层特征作为空间路径; 3) 使用编码端骨干网络的高层特征作为空间路径; 4) 反 U 型结构的空间路径但其中每个路径没有链式反置残差模块; 5) 本文方法, 即反 U 型结构的空间路径, 其中每个路径使用链式反置残差模块. 实验结果如表 8 所示, 使用空间路径可以将性能从 63.55% 提升到 66.79%, 说明

表 8　不同空间路径对性能的影响 (CRB: 链式反置残差模块; SP: 空间路径; LFP: 浅层特征作为空间路径; HFP: 高层特征作为空间路径; RUP: 反 U 型空间路径)
Table 8　The influence of different spatial paths on performance (CRB: chain inverted residual block; SP: spatial path; LFP: low-level feature as spatial path; HFP: high-level feature as spatial path; RUP: reverse u-shaped spatial path)

方法	MIoU (%)
No SP	63.51
LFP	66.06
HFP	67.49
RUP	66.79
RUP+CRB	**68.26**

使用反 U 型结构能够非常有效地利用编码端浅、中层特征, 这也说明编码端的浅、中层特征包含了解码时所需的空间细节信息. 使用链式反置残差块能使性能从 66.79% 提升到 68.26%, 说明链式反置残差模块可以保留空间细节信息的同时提升其语义表达能力, 从而提升语义分割性能.

3.5.5　验证链式反置残差模块链长设置的合理性

在空间路径中, 本文采用长度递减的链长分别处理深度模型的浅、中及高层特征信息, 目的是减少融合时各层的语义差异性. 表 9 中展现了在 CamVid 数据集上不同链长设置对本文提出框架分割性能的影响. 从表 9 中的结果可以看出, 本文针对浅、中及高层特征信息分别使用递减链长的反置残差模块更加有效.

表 9　链式反置残差模块不同链长对性能的影响
Table 9　The influence of CRB chain length on performance

方法	MIoU (%)
各路径链长均为 1	67.20
各路径链长均为 3	67.25
本文 (分别为 3, 2, 1)	**68.26**

3.5.6　验证残差循环卷积模块的有效性

为了表明所提出优化模块的有效性, 本文将使用优化模块和不使用优化模块进行对比. 如表 10 所示, 使用优化模块能使分割性能提升 0.76 %, 说明优化模块可以改善融合后的语义特征, 从而增强了分割性能.

表 10　残差循环卷积模块对性能的影响
Table 10　The influence of RRB on performance

方法	MIoU (%)
使用优化模块	**68.26**
不使用优化模块	67.50

4　结束语

本文深入研究了采用编解码结构和二分支结构的语义分割方法, 提出了一种新的端到端的深度学习框架用于语义分割. 在编码端采用二分支结构以获取高质量的上下文语义特征, 同时有效利用编码端浅中层的空间细节信息. 本文方法在 3 个语义分割基准数据集上取得了有竞争力的结果, 一系列消融实验也验证了本文提出的各功能模块的有效性. 通过可视化预测结果, 发现本文方法在小物体上的分割还不够精准, 进一步的工作拟研究产生这种现象的原因, 并进一步改进分割模型.

References

1　Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(4): 640−651

2　Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. In: Proceedings of the 4th International Conference on Learning Representations (ICLR). San Juan, Puerto Rico, USA: Conference Track Proceedings, 2016.

3　Yu F, Koltun V, Funkhouser T. Dilated residual networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, Hawaii, USA: IEEE, 2017. 472−480

4　Chen L C, Papandreou G, Kokkinos I, Murphy K, Yuille A L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40**(4): 834−848

5　Zhao H S, Shi J P, Qi X J, Wang X G, Jia J Y. Pyramid scene parsing network. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, Hawaii, USA: IEEE, 2017. 2881−2890

6　Chen L C, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. [Online], available: https://arxiv.org/abs/1706.05587v1, Jun 17, 2017

7　Paszke A, Chaurasia A, Kim S, Culurciello E. Enet: A deep neural network architecture for real-time semantic segmentation. [Online], available: https://arxiv.org/abs/1606.02147, Jun 7, 2016

8　Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 1520−1528

9　Huang Ting-Hong, Nie Zhuo-Yun, Wang Qing-Guo, Li Shuai, Yan Lai-Cheng, Guo Dong-Sheng. Real-time image semantic segmentation based on block adaptive feature fusion. *Acta Automatica Sinica*, 2021, **47**(5): 1137−1148
(黄庭鸿, 聂卓赟, 王庆国, 李帅, 晏来成, 郭东生. 基于区块自适应特征融合的图像实时语义分割. 自动化学报, 2021, **47**(5): 1137−1148)

10　Islam M A, Naha S, Rochan M, Bruce N, Wang Y. Label refinement network for coarse-to-fine semantic segmentation. [Online], available: https://arxiv.org/abs/1703.00551v1, Mar 1, 2017

11　Jégou S, Drozdzal M, Vazquez D, Romero A, Bengio Y. The one

hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, Hawaii, USA: IEEE, 2017. 11−19

12 Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for scene segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(12): 2481−2495

13 Zeiler M D, Fergus R. Visualizing and understanding convolutional networks. In: Proceedings of the 2014 European Conference on Computer Vision (ECCV). Zürich, Switzerland: Springer, 2014. 818−833

14 Yu C Q, Wang J B, Peng C, Gao C X, Yu G, Sang N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the 2018 European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 325−341

15 Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L C. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt lake, Utah, USA: IEEE, 2018. 4510−4520

16 Alom M Z, Hasan M, Yakopcic C, Taha T M, Asari V K. Recurrent residual convolutional neural network based on U-net (R2U-Net) for medical image segmentation. [Online], available: https://arxiv.org/abs/1802.06955v1, Feb 20, 2018

17 Chaurasia A, Culurciello E. Linknet: Exploiting encoder representations for efficient semantic segmentation. In: Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP). Saint Petersburg, Florida, USA: IEEE, 2017. 1−4

18 Amirul Islam M, Rochan M, Bruce N D, Wang Y. Gated feedback refinement network for dense image labeling. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, Hawaii, USA: IEEE, 2017. 3751−3759

19 Poudel R P, Bonde U, Liwicki S, Zach C. Contextnet: Exploring context and detail for semantic segmentation in real-time. In: Proceedings of the 2018 British Machine Vision Conference (BMVC). Northumbria University, Newcastle, UK: BMVA, 2018. 146

20 Poudel R P, Liwicki S, Cipolla R. Fast-SCNN: Fast semantic segmentation network. [Online], available: https://arxiv.org/abs/1902.04502, Feb 12, 2019

21 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 770−778

22 Deng J, Dong W, Socher R, Li J A, Li K, Li F F. Imagenet: A large-scale hierarchical image database. In: Proceedings of the 2009 IEEE conference on Computer Vision and Pattern Recognition (CVPR), Florida, USA: IEEE, 2009. 248−255

23 Wu T Y, Tang S, Zhang R, Zhang Y D. CGNet: A light-weight context guided network for semantic segmentation. [Online], available: https://arxiv.org/abs/1811.08201v1, Nov 20, 2018

24 Szegedy C, Ioffe S, Vanhoucke V, Alemi A A. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI). San Francisco, California, USA: AAAI, 2017. 4278−4284

25 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 2015 International Conference on Machine Learning (ICML). Lille, France: PMLR, 2015. 448−456

26 Brostow G J, Fauqueur J, Cipolla R. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 2009, **30**(2): 88−97

27 Song S R, Lichtenberg S P, Xiao J X. Sun RGB-D: A RGB-D scene understanding benchmark suite. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, Massachusetts, USA: IEEE, 2015. 567−576

28 Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, et al. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 3213−3223

29 He K M, Zhang X Y, Ren S Q, Sun J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, USA: IEEE, 2015. 1026−1034

30 Kingma D P, Ba J. Adam: A method for stochastic optimization. In: Proceedings of the 4th International Conference on Learning Representations ICLR). San Diego, CA, USA: Conference Track Proceedings, 2015.

罗会兰 江西理工大学信息工程学院教授. 2008 年获浙江大学计算机科学与技术博士学位. 主要研究方向为计算机视觉与机器学习. 本文通信作者.
E-mail: luohuilan@sina.com
(**LUO Hui-Lan** Professor at the School of Information Engineering, Jiangxi University of Science and Technology. She received her Ph.D. degree in computer science and technology from Zhejiang University in 2008. Her research interest covers computer vision and machine learning. Corresponding author of this paper.)

黎 宵 江西理工大学信息工程学院硕士研究生. 主要研究方向为计算机视觉与语义分割.
E-mail: williamlixiao@sina.com
(**LI Xiao** Master student at the School of Information Engineering, Jiangxi University of Science and Technology. His research interest covers computer vision and semantic segmentation.)