基于可分离残差模块的精确实时语义分割

路文超*, 庞彦伟, 何宇清, 王建
天津大学电气自动化与信息工程学院, 天津 300072

摘要 针对当前智能驾驶领域场景理解中的语义分割算法无法同时满足高精度和高效率要求的问题,提出了精确高效的语义分割算法。基于可分离残差模块和降采样模块,设计了充分利用其学习能力和学习效率的高效精确语义分割网络结构。利用 Cityscapes 数据集,在图像处理效率 12 frame/s 的基础上达到分割精度 67.86%。研究结果表明,所提方法在精度和效率上均能达到较好的效果,实现了精度和效率的平衡。

关键词 图像处理; 语义分割; 卷积神经网络; 深度可分离卷积; 可分离残差模块
中图分类号 TP391.4 **文献标识码** A **doi**: 10.3788/LOP56.051005

Real-Time and Accurate Semantic Segmentation Based on Separable Residual Modules

Lu Wenchao*, Pang Yanwei, He Yuqing, Wang Jian
School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract Aiming at the problem that the current approaches of semantic segmentation cannot meet the simultaneous demands on accuracy and efficiency in scene parsing in the intelligent vehicles, an accurate and efficient algorithm for semantic segmentation is proposed. Based on the proposed separable residual module and the down-sampling module, a real-time and accurate semantic segmentation network is designed. With the Cityscapes dataset, the segmentation accuracy can reach 67.86% on the basis of the 12 frame/s efficiency. The research results demonstrate that the proposed method can achieve a good performance both in accuracy and efficiency, and makes a balance between accuracy and efficiency.

Key words image processing; semantic segmentation; convolutional neural network; depthwise separable convolution; separable residual module
OCIS codes 100.4996; 100.2960; 100.5010

1 引 言

自动驾驶是当前热门且富有挑战性的课题。其中,视觉感知模块是自动驾驶系统中不可或缺的一部分,涉及道路识别、交通标志识别、车辆和行人检测等视觉感知任务[1]。针对上述视觉任务当前都有较为成熟的算法,但发展自动驾驶汽车,需要实时精确地理解周围环境,同时完成对不同类别物体的分类和分割,这样,语义分割算法的研究就显得尤为重要。语义分割是图像理解中常用的技术,它可以预测图像中每个像素的类别,从而实现对图像的分割归类,同时可以对图像进行细致理解,这对研究自动驾驶非常重要[2]。

近年来,深度学习发展迅猛。自卷积神经网络(CNN)[3]提出以来,采用神经网络做物体分类的算法层出不穷,尤其在 Krizhevsky 等[4]提出的 AlexNet 取得 ImageNet 竞赛冠军后,深度学习逐渐在各类视觉任务中占据了主流地位。CNN 最初的设计用于图像分类任务,可以对图像进行端到端训练,实现对图像的分类。鉴于 CNN 在分类和检测任务上的巨大成功,将神经网络引入语义分割(像素级分类)逐渐进入人们视线。为将 CNN 应用到语

收稿日期: 2018-08-29; **修回日期**: 2018-09-17; **录用日期**: 2018-09-27
基金项目: 国家自然科学基金重点项目(61632081)
* **E-mail**: luwc@tju.edu.cn

义分割，Long 等[5]提出了全卷积神经网络(FCN)。FCN 将 CNN 的全连接层替换成了卷积层，然后用去卷积层对最后一个卷积层的特征图进行上采样，得到与输入图像尺寸相同的像素级分类结果。由于去掉了全连接层，FCN 可以接受任意尺寸的输入图像[6-7]。

当前，语义分割领域的研究主要集中在两个方向：1) 通过增加网络层级、增大网络复杂度提升网络分割精度；2) 通过降低网络复杂度、减少参数量提升网络效率[8]。在提升网络精度方面，金字塔场景解析网络(PSPNet)[9]采用空间金字塔池化模型聚合不同区域的上下文信息，提高网络获取全局信息的能力，充分利用上下文关系和标签间的关联提升分割精度；Chen 等[10-11]利用空洞空间金字塔池化在聚合不同区域信息的同时增大卷积核感受野，进一步利用多尺度信息强化分割，并在网络末端加入全连接条件随机场用于锐化分割结果。加入条件随机场可以提升分割精度，但也会增加网络计算复杂度，网络不能端到端训练。细化网络(RefineNet)[12]利用长距离残差连接对不同尺度的特征进行融合，充分利用底层特征边缘优势和高层特征语义优势，实现用低层信息优化高层语义，进而提升分割精度。上述几种分割算法都已取得了很高的分割精度，但在使用单块 TITAN X 图形处理器(GPU)时，单张图像处理时间均＞1 s，难以满足实际应用中的实时性需求。在提升网络效率方面，Paszke 等[13]提出了高效神经网络(ENet)，针对大分辨率特征图处理时间长的问题，在网络前端进行连续降采样降低特征图分辨率。采用瓶颈残差模块，结合大编码器、小解码器的结构进一步提升网络效率。加速挤压网络(SQ)[14]在编码阶段使用修正的挤压网络(SqueezeNet)[15]降低参数，压缩模型。在解码阶段用去卷积操作进行上采样，进一步节省参数。ENet 和 SQ 在降低网络运算时间和网络参数量方面取得了较大的成果，但同时其分割精度也大大降低，最终分割精度均＜60%，难以满足实际应用中的可靠性需求[16-17]。

针对上述分割精度和分割效率难以平衡的问题，提出了一种既能保证分割实时性，又能一定程度上保证分割精度的网络结构。与以往只注重分割精度或分割效率的算法不同，对这两个看似矛盾的因素进行了综合考虑，通过结合深度可分离卷积和残差连接，设计了一种更加高效的残差模块，称为可分离残差模块，解决了传统残差模块计算效率低的问题。为改善降采样造成的分割精度下降问题，设计了一种池化和卷积并行的降采样模块，并在网络前期就连续使用该降采样模块来提升网络效率。基于可分离残差模块和降采样模块设计了一种可以充分利用其学习能力和学习效率的精确高效语义分割网络结构，称为可分离残差网络(SRNet)。在 Cityscapes 路面场景数据集[18-19]上进行实验，最终网络能以 12 frame/s 的速度处理 2048 pixel×1024 pixel图像，并保持 67.82%的分割精度。

2 精确高效的语义分割网络结构

由残差模块构建的深度卷积神经网络已经广泛应用于语义分割任务中，并取得了较好的分割精度。然而采用传统的残差模块所构建的深度卷积神经网络在解决语义分割任务时无法满足实时性的需求。作为传统卷积的改进方法，深度可分离卷积能够有效简化计算，提升计算效率。基于深度可分离卷积，并引入残差连接，设计了一种新型残差模块——可分离残差模块，以解决现有残差模块的效率限制。为进一步提升网络效率，降低降采样在分割精度上的影响，提出了池化和卷积并行的降采样模块。基于提出的可分离残差模块和并行降采样模块，构建了一种可以充分利用其学习能力和学习效率的可分离残差网络(SRNet)。将该网络应用于语义分割任务中，在保持分割精度的同时，能够有效提升分割的效率，实现实时精确的场景理解，对将语义分割研究应用于智能驾驶领域有着极为重要的意义。

2.1 可分离残差模块

深度可分离卷积是网络模型小型化的一种方式，其本质是冗余信息更少的稀疏化表达[20]。用深度可分离卷积代替标准三维(3D)卷积可以减少卷积核的冗余表达，极大程度上降低参数量和计算量，以便将网络应用到移动端平台。图 1(a)为标准 3D 卷积核，图 1(b)为深度可分离卷积核。深度可分离卷积核将标准 3D 卷积核分解成一个逐通道处理的二维(2D)卷积核和一个跨通道的 1×1 大小的 3D 卷积核。

如图 1(a)所示，标准 3D 卷积在某一层进行卷积的计算量为

$$R_1 = D_K \times D_K \times M \times N \times D_F \times D_F, \quad (1)$$

式中：$D_K \times D_K$ 为卷积核尺寸；M 为输入通道数；N 为输出通道数；$D_F \times D_F$ 为输入分辨率。

如图 1(b)所示，深度可分离卷积在某一层进行卷积的计算量主要分为两个部分：图 1(b)上部所示

的逐通道 2D 卷积核的计算量和图 1(b)下部所示的跨通道 1×1 大小卷积核的计算量。其中，2D 卷积核每次只处理一个通道，数量和输入通道数相同。1×1大小卷积核跨通道处理特征图，将输出通道数变为指定的数量。使用深度可分离卷积的计算量为

$$R_2 = D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F, \quad (2)$$

式中：$D_K \times D_K \times M \times D_F \times D_F$ 为逐通道 2D 卷积核计算量；$M \times N \times D_F \times D_F$ 为跨通道 1×1 大小卷积核计算量。所以使用深度可分离卷积处理特征图的计算量相对于使用标准 3D 卷积处理特征图的计算量比例为

$$R = \frac{R_2}{R_1} = \frac{1}{N} + \frac{1}{D_K^2}。 \quad (3)$$

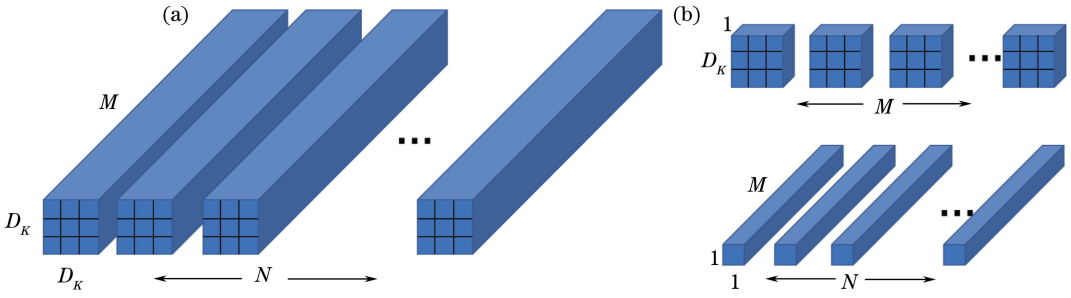

图 1 两种卷积核。(a)标准三维卷积核；(b)深度可分离卷积核

Fig. 1 Two types of convolution filters. (a) Standard 3D convolution filters; (b) depthwise separable convolution filters

假设输入 64 通道 512 pixel×256 pixel 的特征图，输出 128 通道 256 pixel×128 pixel 的特征图，卷积核尺寸为 3×3，则用深度可分离卷积的计算量仅为用标准 3D 卷积计算量的 11.9%。可见用深度可分离卷积代替标准 3D 卷积可以大大降低计算量。

理论上增加网络深度可以让网络拟合更加复杂的非线性函数，对提升网络性能有很大帮助。但在实际过程中，随着网络层数加深，会出现准确率下降的现象。经过分析，准确率下降是由于层数增加导致反向传播过程中梯度消失造成的[21]。为解决传统卷积层在网络层数加深时出现梯度消失的问题，He 等[22]提出了残差学习框架。残差学习用残差映射 $H(x)$ 表示期望得到的实际映射，用堆叠的多层非线性网络拟合另一个映射 $F(x) := H(x) - x$，实际映射关系可表示为 $F(x) + x$，可通过添加“捷径连接”(跳过一个或多个层次的连接)实现。残差学习中使用的“捷径连接”是恒等映射，没有增加新的参数，也不会增加计算复杂度。若一个映射已被优化，使其残差值趋于 0 比堆叠多个非线性网络拟合恒等映射更加容易，所以对残差映射寻优比直接对原始映射寻优更方便。

传统残差模块常用图 2(a)所示的无瓶颈残差模块和图 2(b)所示的瓶颈残差模块两种设计方式，其中 BN 和 ReLU 分别为批归一化模块和修正线性单元模块。网络层数较少时，两种设计方式的参数

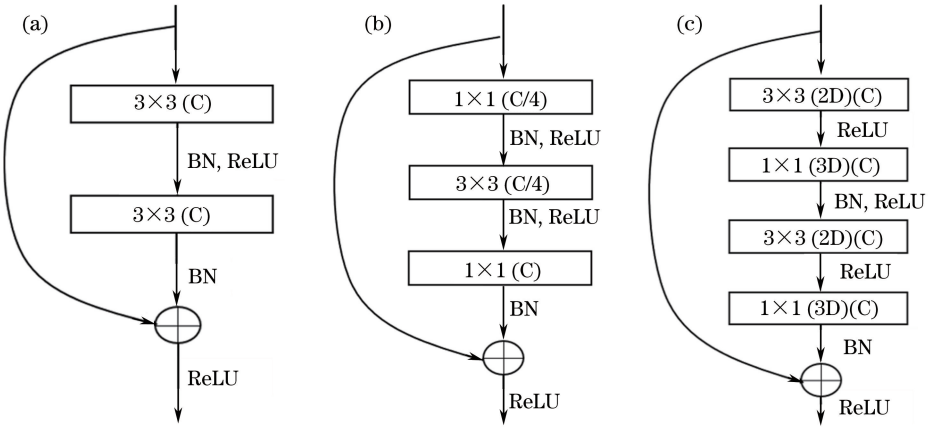

图 2 3 种残差模块。(a)无瓶颈残差模块；(b)瓶颈残差模块；(c)可分离残差模块

Fig. 2 Three types of residual modules. (a) Non-bottleneck residual module; (b) bottleneck residual module; (c) depthwise separable residual module

量和精度几乎相同。但随着网络深度加深，瓶颈残差模块需要增加的计算资源更少，无瓶颈残差模块提升的精度更多[23]。为了在利用无瓶颈残差模块精度优势的同时减少参数量和计算量，基于深度可分离卷积和残差连接，设计了一种新型残差模块，称为可分离残差模块，如图 2(c)所示。可分离残差模块采用 3×3 逐通道卷积替换 3×3 标准卷积，保证网络在深度较大的情况下依然能保持较低的参数量。用 1×1 的跨通道卷积进行通道间信息融合，可以保证提取更有效的特征。可分离残差模块在进行 3×3 卷积前未采用 1×1 卷积进行通道降维以保持网络宽度，最终实现了学习性能和学习效率间的良好平衡。

瓶颈残差模块和无瓶颈残差模块都可以引入深度可分离卷积来降低参数量和计算量。表 1 对比了输入输出均为 256 个通道和均为 64 个通道时不同类型残差模块的参数量。其中，In_Out_C 代表输入输出通道数；Bt 代表瓶颈残差模块；Non-Bt 代表无瓶颈残差模块；DS-Bt 和 DS-Non-Bt 分别代表引入深度可分离卷积后的可分离瓶颈残差模块和可分离无瓶颈残差模块。输入输出 64 通道时，DS-Bt 的参数量为 2.77 k，其是 Bt 的 63.7%；DS-Non-Bt 的参数量为 4.67 k，其是 Non-Bt 的 12.7%；输入输出 256 通道时 DS-Bt 的参数量为 35.65 k，是 Bt 的 51.2%；DS-Non-Bt 的参数量为 67.84 k，其是 Non-Bt 的 11.5%。所以，无瓶颈残差模块中引入深度可分离卷积比瓶颈残差模块中引入深度可分离卷积在降低参数量方面取得的效果更强，而且输入、输出通道数越多，引入深度可分离卷积在降低参数量方面取得的效果越强。

表 1　不同残差模块的参数量

Table 1　Weight sizes of different residual blocks

Residual block		Bt /k	Non-Bt /k	DS-Bt /k	DS-Non-Bt /k
In_Out_C	64	4.35	36.86	2.77	4.67
	256	69.63	589.82	35.65	67.84

结合表 3 的实验结果，在通道数较多的情况下，输入相同通道数的特征图，Bt、DS-Bt 的参数量和单张图片测试时间均低于 Non-Bt、DS-Non-Bt。其中，DS-Bt 的参数量和单张图片测试时间最低，但其精度也最低，难以满足算法的精确度需求。Non-Bt 的精度最高，但其参数量和单张图片测试时间也最高，难以满足算法实时性需求。Non-Bt 的精度最高说明增加网络宽度对提升分割精度有巨大作用，但增加网络宽度也会增加网络参数量，影响网络效率。DS-Non-Bt 的精度与 Non-Bt 相近，但其参数量与单张图片测试时间明显低于 Non-Bt，与 Bt 相近。所以，为平衡精度和效率，采用 DS-Non-Bt，即深度可分离卷积与无瓶颈结构结合作为可分离残差模块。虽然针对的是语义分割任务，但提出的可分离残差模块可以直接迁移到其他任务的网络架构上。

2.2　并行降采样模块

降采样是卷积神经网络中的一项重要内容，可以极大地降低计算量，还可以增大卷积核感受野，让较深的卷积层聚合更多的上下文信息，从而提升分类精度[24]。但在语义分割任务中，降采样会使特征图丢失边缘信息，造成分割像素精度下降。而且与降采样相对的上采样操作也会增大网络计算量，降低网络效率。为了进一步提高网络效率，减小降采样在分割像素精度方面的影响，设计了一种将步长为 2 的最大池化和步长为 2 的 3×3 卷积输出并行连接的新型降采样模块，如图 3 所示。

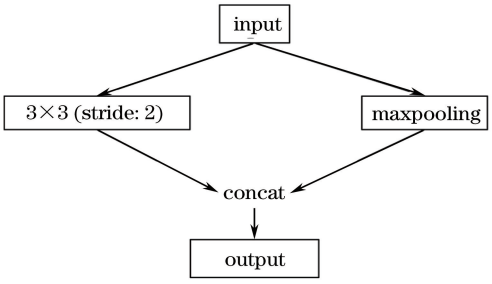

图 3　并行降采样模块

Fig. 3　Parallel down-sampling block

该降采样模块利用 3×3 卷积核的感受野提升了精度，同时利用最大池化的高效性提升了速度。为了减小多次降采样造成的分割精度下降以及降低多次上采样操作需要的计算时间，本文采用三次降采样操作。为了弥补因降采样次数减少造成的感受野下降问题，引入了膨胀卷积[25-26]。膨胀卷积又称空洞卷积，通过在卷积核中插“0”的方式，让卷积核在不增加额外参数量的情况下增大感受野，从而聚合更多的上下文信息。图 4(a)为标准 3×3 卷积核，感受野为 3×3；图 4(b)为膨胀率为 2 的 3×3 卷积核，感受野为 5×5。加入膨胀卷积，可以在不增加额外参数的情况下扩大卷积核感受野，聚合更多上下文信息，从而提升网络精度；图 4(c)为可分离

残差模块与膨胀卷积结合的方式，r 代表膨胀率，本文采用了 4 种膨胀卷积，膨胀率分别为 2、4、8、16。

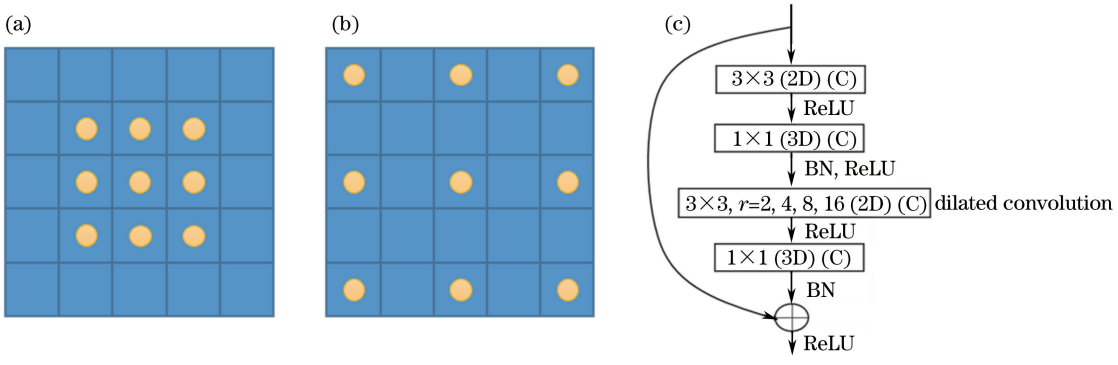

图 4 膨胀卷积。(a)标准卷积核；(b)膨胀率为 2 的卷积核；(c)与膨胀卷积结合的可分离残差模块

Fig. 4 Dilated convolution. (a) Standard convolution filters; (b) 2-dilated convolution filters; (c) separable residual module combined with dilated convolution

2.3 可分离残差语义分割网络

为解决当前语义分割算法无法同时满足高精度和高效率要求的问题，结合深度可分离卷积和残差连接设计了一种残差模块——可分离残差模块，解决了传统残差模块学习能力和运算效率低的问题。为进一步提升网络效率，降低降采样对分割像素精度的影响，设计了一种卷积和池化并行的降采样模块。以可分离残差模块和并行降采样模块为基础，设计了一种精确高效的语义分割网络，称为可分离残差网络。

如图 5 所示，采用编码-解码架构进行端到端网络训练。网络由连续的编码器和连续的解码器构成。编码器负责提取特征，并产生降采样的特征图。解码器负责进一步对特征进行提取，并对编码器的输出进行上采样，产生分割结果。以输入 1024 pixel×512 pixel 的三通道彩色图像为例，输入图像在编码阶段经过多次降采样模块和可分离残差模块提取特征后，进入解码阶段经过上采样模块产生与输入图像大小相同的分割结果。

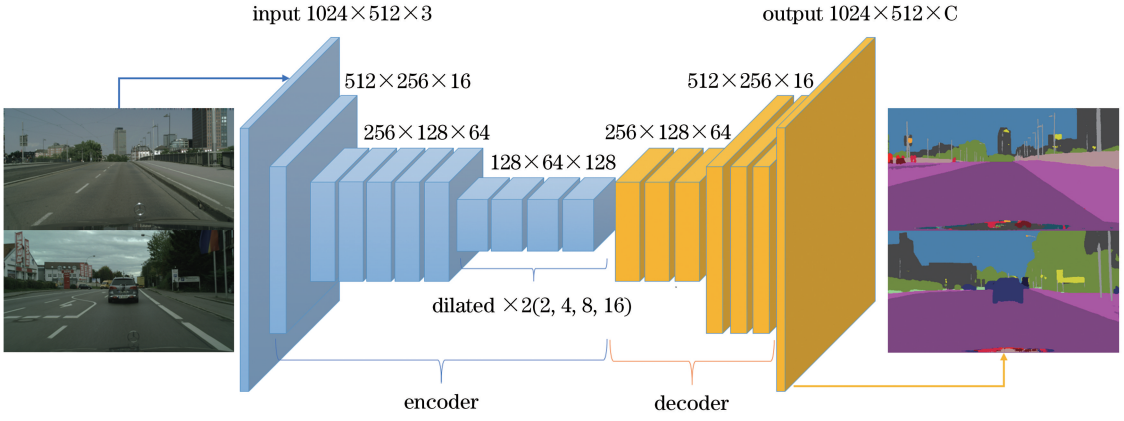

图 5 网络结构

Fig. 5 Network architecture

表 2 展示了可分离残差网络的具体细节，其中 Block 表示模块序列，Type 表示模块类别，In-Res 和 In-C 分别表示输入分辨率和输入通道数，Out-Res 和 Out-C 分别表示输出分辨率和输出通道数。Down-sampling 为降采样模块，DS-Non-Bt 为可分离残差模块，Deconvolution 为去卷积操作，用于上采样。C 为最终的输出通道数，也是数据集的类别数。网络共包括 23 个模块，1～16 个模块属于编码器，17～23 个模块属于解码器。编码器主要由残差模块和降采样模块组成。降采样模块会降低分割像素精度，同时也能降低计算消耗。为平衡分割效率和分割精度，仅在第 1、2、8 个模块使用降采样操作。由于网络只采用 3 次降采样操作，较深的卷积层感受野较小，不能聚合足够的上下文信息用于分类。为了让较深的卷积层能聚合更多的上下文信息，在 9～16 个可分离残差模块的 2D 卷积部分分别加入了膨胀率为 2、4、8、16、2、4、8、16 的膨胀卷积，

如图 4(c)所示。为了防止过拟合,对所有的残差模块采用 Dropout 正则化,并将 Dropout 概率设为 0.3。

表 2 网络具体细节

Table 2 Detailed descriptions of our network

Network	Block	Type	In-Res /(pixel×pixel)	In-C	Out-Res /(pixel×pixel)	Out-C
Encoder	1	Down-sampling	1024×512	3	512×256	16
	2	Down-sampling	512×256	16	256×128	64
	3-7	5×DS-Non-Bt	256×128	64	256×128	64
	8	Down-sampling	256×128	64	128×64	128
	9-16	2×DS-Non-Bt (rate=2,4,8,16)	128×64	128	128×64	128
Decoder	17	Deconvolution	128×64	128	128×64	64
	18-19	2×DS-Non-Bt	256×128	64	256×128	64
	20	Deconvolution	256×128	64	512×256	16
	21-22	2×DS-Non-Bt	512×256	16	512×256	16
	23	Deconvolution	512×256	16	1024×512	C

解码器主要用于进一步提取特征,并对编码器的输出进行上采样,产生分割结果。为保证解码阶段的像素精度,SegNet[27-28] 和 ENet[13] 在降采样模块中进行池化操作时,除了储存结果元素外,还可额外储存结果元素的元素索引。在上采样模块中调用储存的结果元素和元素索引进行去池化操作。由于要额外存储结果元素的元素索引,这种编码-解码方式需要占用较多的内存。为了降低内存需要,本文网络在进行上采样时用去卷积操作代替了上述的去池化操作。

该网络可以充分利用可分离残差模块的学习能力和学习效率。通过控制降采样模块的数量并结合膨胀卷积,网络可以在分割效率和分割精度之间实现良好的平衡。采用去卷积操作代替常用的基于元素索引的去池化操作,可以大大降低网络的内存需求。最终网络在基本满足实时的情况下,可以实现较高的分割精度。

3 实验结果与分析

3.1 实验平台与数据集

基于 PyTorch 深度学习框架构建可分离残差网络,硬件环境为 GPU GeForce GTX TITAN X。为验证所提方法的有效性,在 Cityscapes 数据集上进行语义分割实验。Cityscapes 数据集是图像语义分割领域中广泛采用的城市路面场景数据集,包含 5000 张精标注的图像数据和 19998 张粗标注的图像数据。实验是基于精标注数据进行的,未使用粗标注数据进行预训练。5000 张精标注数据分为 2975 张训练数据、500 张验证数据和 1525 张测试数据。其中,测试数据的真实标签未提供,但可以在其官方测试服务器上进行测试。Cityscapes 路面场景数据集的图像分辨率为 2048 pixel×1024 pixel,所有像素可标注为 30 类,选择其中常用的 19 类(路面、汽车、行人等)进行训练和测试。最终使用平均交并比(M_{iou})作为分割精度的评价标准,使用每秒处理图像张数,即帧率(f_{ps})作为分割速度的评价标准。M_{iou} 反映预测值和真实值之间的相关度,相关度越高,M_{iou} 越大。

$$M_{\mathrm{iou}}=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}+F_{\mathrm{N}}},\tag{4}$$

式中 T_{P}、F_{P}、F_{N} 分别为真正率(标签为正,预测结果为正)、假正率(标签为负,预测结果为正)和假负率(标签为正,预测结果为负)。

3.2 不同残差模块对比

对编码器和解码器分开进行训练,所有实验均未使用预训练模型。训练时先独立训练编码器,编码器训练完成后,再结合解码器继续训练整个网络。在编码器末端附加一个额外卷积层来训练编码器。编码器训练完成后,移除额外添加的卷积层,并加上解码器训练整个网络。训练过程中通过随机水平翻转和 0~2 pixel 的平移变换进行数据增强。

为验证可分离残差模块的有效性,对瓶颈残差模块和无瓶颈残差模块均引入深度可分离卷积,并对比 4 种残差模块的精度(M_{iou})、参数量和单张图片测试时间,结果如表 3 所示。Bt 和 Non-Bt 分别代表瓶颈残差模块和无瓶颈残差模块。DS-Bt 和 DS-Non-Bt 分别代表引入可分离卷积后的瓶颈残差模块和无瓶颈残差模块。M_{iou} 用来分析网络模型的学习能力。参数量用来记录模型大小,是网络权重和偏置的总和。单张图像测试时间用来评估网络效

率，在该组实验中表示 1024 pixel×512 pixel 分辨率图像的测试时间。为保证实验效率，这部分实验用 512 pixel×256 pixel 的图像进行训练。

在输入通道数相同的情况下 Bt、DS-Bt 的参数量和单张图片测试时间均低于 Non-Bt、DS-Non-Bt。其中 DS-Bt 的参数量仅为 0.22 ×10^6，单张图片测试时间仅为 15 ms，在所有残差模块中效率最高。但 DS-Bt 精度仅为 54.36%，难以满足算法的精确度需求。Non-Bt 的精度能达到 62.19%，但其参数量相当于 Bt 的 10 倍，单张图片测试时间也相当于 Bt 的 2 倍，难以满足算法实时性需求。DS-Non-Bt 的参数量为 0.49×10^6，与 Bt 相近，仅为 Non-Bt 的 16.2%，但其精度能达到 61.37%，与 Non-Bt 相近，明显高于 Bt。为平衡精度和效率，采用 DS-Non-Bt，即深度可分离卷积与无瓶颈结构结合作为可分离残差模块。

表 3 各残差模块的精度和效率

Table 3 Accuracy and efficiency of each residual module

Module	M_{iou} /%	Parameter /10^6	Time /ms
Bt	57.12	0.31	18
Non-Bt	62.19	3.03	35
DS-Bt	54.36	0.22	15
DS-Non-Bt	61.37	0.49	24

为进一步说明可分离残差模块有效性，在参数量相近的情况下对几种残差模块进行比较。将无瓶颈残差模块的输入通道数设为瓶颈残差模块的1/4，进行了两组对比实验，如表 4 所示，其中 n 代表按照表 2 结构设计的输入通道数。第一组实验对比了瓶颈残差模块输入通道数不变，无瓶颈残差模块输入通道数变为 1/4 时 3 种残差模块的结果。第二组实验对比了无瓶颈残差模块输入通道数不变，瓶颈残差模块输入通道数变为 4 倍时 3 种残差模块的结果。这部分实验同样用 512 pixel×256 pixel 的图像进行训练。

如表 4 所示，第一组实验中 Bt 的精度高于 Non-Bt 和 DS-Non-Bt，同时参数量也高于 Non-Bt 和 DS-Non-Bt，单张图像测试时间也为三者最长。第二组实验中 Non-Bt 和 DS-Non-Bt 的精度高于 Bt，而且参数量均比 Bt 少，单张图像测试时间也低于 Bt。分析原因为第一组实验中 Non-Bt 和 DW-Non-Bt 的输入通道数过少，难以拟合复杂的非线性函数，通道数增加后，Non-Bt 和 DW-Non-Bt 的精度就会大幅上升，并超过 Bt。第二组实验精度均明显高于第一组，说明特征图通道数越多，网络越能近似需要学习的损失函数，因此让网络更宽是增强网络学习能力的有效方法。但网络加宽明显会使网络参数量增加，运算时间增加。第二组实验中 DS-Non-Bt 的参数量为 0.49×10^6，仅为 Non-Bt 的 16.2%，且单张图像测试时间为 24 ms，低于 Non-Bt。这说明所提的 DS-Non-Bt 可以让网络宽度增加的同时保持较低的参数量和较快的运算速度。

表 4 通道数不同时各残差模块的精度和效率

Table 4 Accuracy and efficiency of each residual module with different channels

Module	Channel	M_{iou} /%	Parameter /10^6	Time /ms
Bt	n	57.12	0.31	18
Non-Bt	n/4	52.38	0.20	13
DS-Non-Bt	n/4	53.23	0.05	11
Bt	4n	60.81	4.71	45
Non-Bt	n	62.19	3.03	35
DS-Non-Bt	n	61.37	0.49	24

3.3 探究通道降维的影响

为进一步提升可分离残差模块效率，对可分离残差模块内部通道进行降维。在图 2(c)中可以将 1×1卷积用于降维和升维，构建更高效的可分离残差模块。实验分别测试了图 2(c)中不用 1×1 卷积降通道，图 6(a)中用 1×1 卷积将通道数降为 1/2 和图 6(b)中用 1×1 卷积将通道数降为 1/4 时的精度(M_{iou})、参数量和单张图像测试时间，结果如表 5 所示。其中，DS-Non-Bt 代表所用的可分离残差模块，DS-Bt-1/2 和 DS-Bt-1/4 分别代表将通道数降为 1/2 和 1/4 时的可分离残差模块。

如表 5 所示，不降通道的可分离残差模块精度为 67.82%，参数量为 491 k，单张图片测试时间为 88 ms。通道数降为 1/2 的可分离残差模块相对于不降通道的可分离残差模块，精度下降 3.35%，参数量降低 34.6%，单张图像测试时间减少 18 ms。通道数降为 1/4 的可分离残差模块相对于不降通道的可分离残差模块，精度下降了 6.93%，参数量降低了 51.9%，单张图像测试时间减少了 26 ms。对可分离残差模块内部通道进行降维，网络运算时间和参

数量下降，网络效率提升，但效率提升效果不显著。加入通道降维会使网络分割精度显著下降，这说明在实际应用中应减少使用可分离残差模块中的通道降维，所以本文模块采用不降维的设计方式。

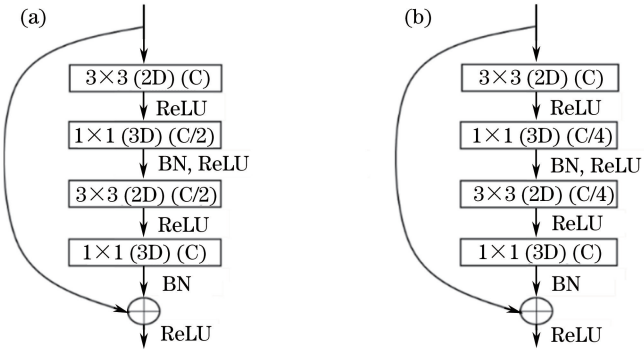

图 6 与通道降维结合的可分离残差模块。(a)通道数降为 1/2；(b)通道数降为 1/4

Fig. 6 Separable residual module combined with channel reduction. (a) 1/2 channels; (b) 1/4 channels

表 5 与通道降维结合后的残差模块精度和效率

Table 5 Separable residual module combined with channel reduction

Module	M_{iou} /%	Parameter /k	Time /ms
DW-Non-Bt	67.82	491	88
DW-Bt-1/2	64.47	321	70
DW-Bt-1/4	60.89	236	62

3.4 与经典网络对比

与 SegNet[27]、ENet[13]、SQ[14]几种常用的实时语义分割网络在分割精度和运算时间上进行了比较。表 6 为不同网络在 Cityscapes 数据集 2048 pixel×1024 pixel 分辨率图像上每一类的分割精度。表 7 为不同网络针对不同分辨率输入的运算时间和帧率。从表 7 可看出，SRNet 在输入为 2048 pixel×1024 pixel 超大分辨率图像时能保持 12 frame/s 的处理速度。SRNet 处理一张 2048 pixel×1024 pixel 分辨率图像的时间为 88 ms，慢于 ENet 和 SQ，但仍可基本满足算法实时性。在分割精度方面，SRNet 能达到 67.86%的分割精度，比 ENet 高 9.58%，比 SQ 高 8.02%。综合分析，所提基于可分离残差模块的 SRNet 可以在分割速度和分割精度上都达到较好的结果，可以实现精确高效的语义分割。

表 6 各网络的分割精度

Table 6 Separation accuracy of each network %

Model	Class	Roa	Sid	Bui	Wal	Fen	Pol	TLi	TSi	Veg	Ter	Sky	Per	Rid	Car	Tru	Bus	Tra	Mot	Bic
SegNet	56.95	96.4	73.2	84.0	28.4	29.0	35.7	39.8	45.1	87.0	63.8	91.8	62.8	42.8	89.3	38.1	43.1	44.1	35.8	51.9
SQ	59.84	96.9	75.4	87.8	31.6	35.7	50.9	52.0	61.7	**90.9**	**65.8**	**93.0**	**73.8**	42.6	91.5	18.8	41.2	33.3	34.0	59.9
ENet	58.28	96.3	74.2	75.0	32.2	33.2	43.4	34.1	44.0	88.6	61.4	90.6	65.5	38.4	90.6	36.9	50.5	48.1	38.8	55.4
SRNet	**67.86**	**97.1**	**78.6**	**89.6**	**49.3**	**51.2**	**56.9**	**57.5**	**66.3**	90.4	57.0	92.2	71.8	**48.6**	**91.7**	**55.7**	**70.2**	**58.3**	**40.3**	**66.0**

表 7 各网络的分割效率

Table 7 Separation efficiency of each network

Model	2048×1024		1024×512		512×256		1920×1080		1280×720		640×360	
	Time /ms	Frame rate /(frame·s^{-1})	Time /ms	Frame rate /(frame·s^{-1})	Time /ms	Frame rate /(frame·s^{-1})	Time /ms	Frame rate /(frame·s^{-1})	Time /ms	Frame rate /(frame·s^{-1})	Time /ms	Frame rate /(frame·s^{-1})
SegNet	641	2	169	6	41	24	637	1	289	3	69	14
SQ	59	17	19	53	6	167	58	17	33	30	9	111
ENet	49	20	13	77	7	143	46	21	21	46	7	135
SRNet	88	12	24	42	6	167	88	12	37	27	9	111

3.5 与 ENet 网络定性对比

为直观展示 SRNet 的优越性，与经典网络 ENet[13]进行了定性对比。图 7 为 SRNet 和 ENet 的定性对比结果。图 7(a)为输入图像，图 7(b)为真实标签，图 7(c)为 ENet 输出结果，图 7(d)为 SRNet 输出结果，图中圈出部分为着重对比部分。

在第一列实例中，左侧多辆汽车中间有几辆自行车，ENet 将其全部分割成汽车，SRNet 可分割出自行车部分；在第二列实例中，对于一个骑自行车的人，ENet 仅将其分割成了行人，SRNet 既分割出了行人，也分割出了自行车；在第三列实例中，SRNet 对汽车尾部的分割效果明显优于 ENet；在第四列实例中，SRNet 对电线杆等边缘细节要求较高的类别分割效果明显优于 ENet 的分割效果。由此可知，在处理边缘细节和较小目标时，SRNet 展现了比 ENet 更精确的分割结果。因此所提 SRNet 在分割效果上相较于 ENet 有了较大提升。

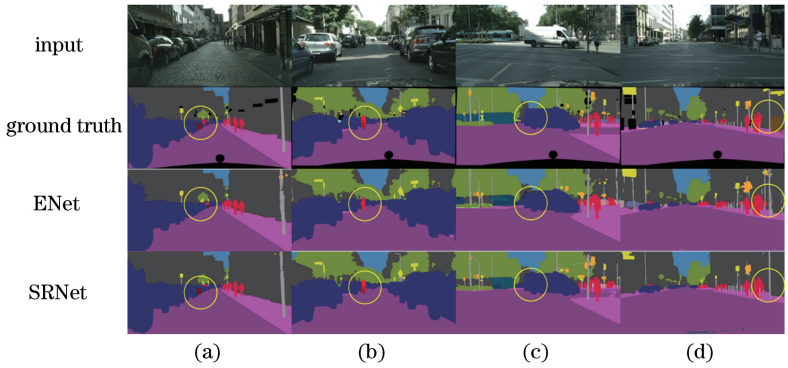

图 7 SRNet 和 ENet 的定性对比。(a)输入图像；(b)真实标签；(c) ENet 输出结果；(d) SRNet 输出结果

Fig. 7 Qualitative comparison between SRNet and ENet. (a) Input image; (b) ground truth; (c) ENet result; (d) SRNet result

4 结 论

基于深度可分离卷积和残差连接，设计了一种更高效的残差模块，称为可分离残差模块，解决了传统残差模块的效率限制问题。提出了一种池化和卷积并行的降采样模块，并引入膨胀卷积，解决降采样造成的分割像素精度下降的问题。基于所提可分离残差模块和降采样模块，设计了一种可以充分利用其学习效率和学习能力的精确高效语义分割网络。与以往通过增大网络复杂度、计算消耗量提升精度的方法或通过牺牲网络精度获取网络效率的方法不同，SRNet 综合考虑分割精度和分割效率，实现精确高效的语义分割。在 Cityscapes 数据集上进行实验，SRNet 可在 12 frame/s 的处理效率基础上达到 67.86%的分割精度。实验结果表明，本文方法能够实现分割精度和分割效率的平衡，可以进一步应用到智能驾驶场景理解等需要保持算法稳健性和实时性的智能视觉应用中。

未来将在降低模型功耗方面进行更加深入的实验，还会通过权重二值化等模型压缩技术进一步降低模型的计算资源需求，以促进语义分割在智能视觉中的应用。

参 考 文 献

[1] Romera E, Bergasa L M, Arroyo R. Can we unify monocular detectors for autonomous driving by using the pixel-wise semantic segmentation of CNNs?[EB/OL]. (2016-07-04)[2018-08-10]. https://arxiv.org/abs/1607.00971.

[2] Li L H, Qian B, Lian J, *et al*. Study on traffic scene semantic segmentation method based on convolutional neural network[J]. Journal on Communications, 2018, 39(4): 123-130.
李琳辉, 钱波, 连静, 等. 基于卷积神经网络的交通场景语义分割方法研究[J]. 通信学报, 2018, 39(4): 123-130.

[3] Lecun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[4] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[5] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3431-3440.

[6] Fang X, Wang G H, Yang H C, *et al*. High resolution remote sensing image classification combining with mean-shift segmentation and fully convolution neural network[J]. Laser & Optoelectronics Progress, 2018, 55(2): 022802.
方旭, 王光辉, 杨化超, 等. 结合均值漂移分割与全卷积神经网络的高分辨遥感影像分类[J]. 激光与光

电子学进展, 2018, 55(2): 022802.

[7] Wang L, Liu Q. A multi-object image segmentation algorithm based on local features [J]. Laser & Optoelectronics Progress, 2018, 55(6): 061002.
王琳, 刘强. 基于局部特征的多目标图像分割算法[J]. 激光与光电子学进展, 2018, 55(6): 061002.

[8] Ji M Y, Xi X M, Yu Z L. A review of semantic segmentation based on deep learning[J]. Information Technology and Informatization, 2017(10): 137-140.
计梦予, 袭肖明, 于治楼. 基于深度学习的语义分割方法综述[J]. 信息技术与信息化, 2017(10): 137-140.

[9] Zhao H S, Shi J P, Qi X J, *et al*. Pyramid scene parsing network [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6230-6239.

[10] Chen L C, Papandreou G, Kokkinos I, *et al*. Semantic image segmentation with deep convolutional nets and fully connected CRFs[EB/OL]. (2014-12-22)[2018-08-10]. https://arxiv.org/abs/1412.7062

[11] 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, 2018, 40(4): 834-848.

[12] Lin G S, Milan A, Shen C H, *et al*. RefineNet: multi-path refinement networks for high-resolution semantic segmentation [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017: 5168-5177.

[13] Paszke A, Chaurasia A, Kim S, *et al*. Enet: a deep neural network architecture for real-time semantic segmentation[EB/OL]. (2016-06-07)[2018-08-10]. https://arxiv.org/abs/1606.02147.

[14] Yu C, Wang J, Peng C, *et al*. Bisenet: bilateral segmentation network for real-time semantic segmentation[J]. [EB/OL]. (2018-08-02)[2018-08-13]. https://arxiv.org/abs/1808.00897.

[15] Iandola F N, Han S, Moskewicz M W, *et al*. Squeezenet: alexnet-level accuracy with 50 × fewer parameters and <0.5 MB model size [EB/OL]. (2016-11-04)[2018-08-10]. https://arxiv.org/pdf/1602.07360v4.pdf.

[16] Guo C C, Yu F Q, Chen Y. Image semantic segmentation based on convolutional neural network feature and improved superpixel matching[J]. Laser & Optoelectronics Progress, 2018, 55(8): 081005.
郭呈呈, 于凤芹, 陈莹. 基于卷积神经网络特征和改进超像素匹配的图像语义分割[J]. 激光与光电子学进展, 2018, 55(8): 081005.

[17] Yao H B, Bian J W, Cong J W, *et al*. Medical image segmentation model based on local sparse shape representation [J]. Laser & Optoelectronics Progress, 2018, 55(5): 051011.
姚红兵, 卞锦文, 丛嘉伟, 等. 基于局部稀疏形状表示的医学图像分割模型[J]. 激光与光电子学进展, 2018, 55(5): 051011.

[18] Cordts M, Omran M, Ramos S, *et al*. The cityscapes dataset [C] // CVPR Workshop on the Future of Datasets in Vision, 2015, 1(2): 3.

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

[20] Howard A G, Zhu M, Chen B, *et al*. Mobilenets: efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17)[2018-08-11]. https://arxiv.org/abs/1704.04861.

[21] Szegedy C, Liu W, Jia Y Q, *et al*. Going deeper with convolutions [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1-9.

[22] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[23] Wu Z, Shen C, Hengel A. High-performance semantic segmentation using very deep fully convolutional networks [EB/OL]. (2016-04-15)[2018-08-11]. https://arxiv.org/abs/1604.04339.

[24] Giusti A, Cireşan D C, Masci J, *et al*. Fast image scanning with deep max-pooling convolutional neural networks [C] // Proceedings of IEEE International Conference on Image Processing, 2013: 4034-4038.

[25] Holschneider M, Kronland-Martinet R, Morlet J, *et al*. A real-time algorithm for signal analysis with the help of the wavelet transform [M]. Berlin, Heidelberg: Springer, 1990: 286-297.

[26] Sermanet P, Eigen D, Zhang X, *et al*. Overfeat: integrated recognition, localization and detection using convolutional networks [EB/OL]. (2014-02-24) [2018-08-12]. https: // arxiv. org/abs/1312.6229.

[27] 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.

[28] Kendall A, Badrinarayanan V, Cipolla R. Bayesian SegNet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding [EB/OL]. (2016-10-10) [2018-08-12]. https://arxiv.org/abs/1511.02680.