基于双分支特征提取的轻量级图像分割算法

孙红，杨晨，莫光萍，朱江明
（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘要：目的 为了提升彩色图像的分割精度，解决彩色图像分割中存在庞大计算成本和冗余参数的问题，本文提出一种双分支特征提取网络来解决上述问题。**方法** 双分支特征提取网络主要由语义信息分支和空间细节分支组成。语义信息分支通过在非对称残差模块中设置不同的空洞卷积率来获取输入图像不同尺度的上下文信息。空间细节分支是一个浅层且简单的网络，用于建立每个像素间的局部依赖关系以保留细节。在双分支之后连接一个特征聚合模块来有效地结合这 2 个分支的输出。**结果** 在没有任何预训练和后处理的情况下，在单块 RTX2080Ti GPU 上仅用 0.91 M 参数在 Cityscapes 数据集上以 97 帧/s 的速度实现 75.1%的分割准确性，在 Camvid 数据集上以 107 帧/s 的推理速度取得了 70.5%的分割效果。**结论** 通过大量实验证明，本文模型在分割准确性和效率之间取得了较好的平衡。

关键词：语义分割；特征提取；注意力机制；空间细节
中图分类号：TP391 **文献标识码：**A **文章编号：**1001-3563(2023)11-0299-10
DOI：10.19554/j.cnki.1001-3563.2023.11.035

Lightweight Image Segmentation Algorithm Based on Two-branch Feature Extraction

SUN Hong, *YANG Chen*, *MO Guang-ping*, *ZHU Jiang-ming*

(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

ABSTRACT: The work aims to propose a two-branch feature extraction network to improve the segmentation accuracy of color images and solve the problem of huge computing costs and redundant parameters in color image segmentation. The two-branch feature extraction network was mainly composed of semantic information branch and spatial detail branch. The semantic information branch obtained the context information of different scales of the input image by setting different hole convolution rates in the asymmetric residual module. The spatial detail branch was a shallow and simple network, which was used to establish the local dependency of each pixel to retain the details. A feature aggregation module was connected behind the two branches to effectively combine the outputs of the two branches. Without any pre-training and post-processing, on a single RTX2080Ti GPU, only 0.91 M parameters were used to achieve 75.1% segmentation accuracy on Cityscapes dataset at the speed of 97 FPS, and 70.5% segmentation effect was achieved on Camvid dataset at the reasoning speed of 107 FPS. A large number of experiments have proved that the proposed model achieves a good balance between segmentation accuracy and efficiency.

KEY WORDS: semantic segmentation; feature extraction; attention mechanism; spatial detail

收稿日期：2022−08−20
基金项目：国家自然科学基金（61703277）
作者简介：孙红（1964—），女，博士，副教授，主要研究方向为模式识别与智能系统、计算机视觉与图像处理。

语义分割是计算机视觉中的一个重要研究领域，它通过对图像执行像素级标签预测实现分割目标。近年来，语义分割在彩色图像分割等领域都受到广泛关注[1-3]，这些应用领域对能够实时运行的场景理解系统要求很高，不仅需要具有低能耗和低内存的竞争性能，而且对模型的实时性有严格的要求。因此设计一个用于实时语义分割的高效神经网络成为一个具有挑战性的问题。

近年来许多实时语义分割领域的优秀研究工作试图在准确性、轻量级和高速率之间达到平衡。Paszke 等[4]提出了一种高效的实时语义分割网络 ENet，通过通道裁剪实现了一个紧凑的编码器解码器框架，但是该模型的感受野太小，无法捕捉到大物体的特征信息，导致分割精度的损失。为了提取多尺度的上下文信息，Mehta 等[5]提出了高效空间金字塔网络 ESPNet，采用高效空间金字塔模块和卷积分解策略。图像级联网络 ICNet[6]使用 3 个级联分支来高效处理图像，以推理速度的降低为代价提升分割精度。ERFNet[7]通过编码器阶段的轻量化来提取特征信息，虽然提升了分割精度，但推理速度大幅下降。此外，许多研究工作在网络结构方面作出了很多努力。Ronneberger 等[8]使用了对称的编码器–解码器结构，其策略是合并相应阶段的特征图，然而这种网络会带来巨大的额外计算成本。文献[9-11]采用双分支结构，在编码器阶段分别进行语义信息和空间信息的提取，最后在预测前使用特征融合的方法整合特征，但是这种方式仍然缺乏 2 个分支之间的交互，所以还有很大的改进空间。

针对上述出现的问题，本文提出一个基于双分支特征提取的实时语义分割网络（TBFENet）。本文主要的工作和创新点如下：

1）双分支由语义信息分支（SIB）和空间细节信息分支（SDI）组成，语义信息分支具有对称的编码器–解码器结构，可以有效地提取深层语义信息；空间细节信息分支能很好地保留没有下采样操作的浅层边界细节。

2）在语义信息分支设计一个非对称残差模块（ARM），自适应地融合注意力特征，提升模型分割的准确性；在空间细节分支提出一种空间特征提取模块（SFM），以更好地获得浅层空间特征，补偿语义信息分支中丢失的空间信息细节，同时在双分支使用深度可分离卷积实现轻量化。

3）为了提高融合特征的表示能力，使用特征融合模块（FFM）来有效地融合来自语义和空间级别上的图像特征，增强网络对全局和局部特征信息的提取能力，提高网络整体分割效果。

1　网络模块设计

1.1　总体设计

整个网络结构可以分为 3 个部分：初始块、双分支主干和特征融合模块。完整的网络结构如图 1 所示。

初始块包括 3 个 3×3 卷积层，将第 1 个卷积层的步幅设置为 2 来收集初始特征。为了更好地保留空间特征信息，只在初始块中执行一次下采样操作。本文通过将初始块作为 2 个分支的分界点，使语义和空间信息部分相关，便于后续的特征融合。双分支主干由语义信息分支（SIB）和空间细节信息分支（SDI）组成。为了减少模型的参数，在语义信息分支的深度可分离卷积层中采用空洞卷积来扩大感受野提取特征信息，同时在空间细节信息分支使用特征提取模块（SFM），以较小的计算成本最大程度地保留空间细节。此外，在语义信息分支的不同阶段使用通道注意力来增强通道之间的长距离依赖关系。为了弥补 SIB 中丢失的空间细节信息，使用空间注意力模块生成注意力图来关注有用的空间信息，而忽略空间细节分支中的噪声等无用信息。最后在 2 个分支的末尾使用特征融合模块（FFM）来增强语义和空间双分支的特征融合。

1.2　非对称残差模块

轻量级网络见证了许多残差模块设计，其中图 2a 为基础的残差设计。此外，如图 2b 所示，LEDNet[12]的 SS–nbt（Split-shuffle-non-bottleneck）中所展现的

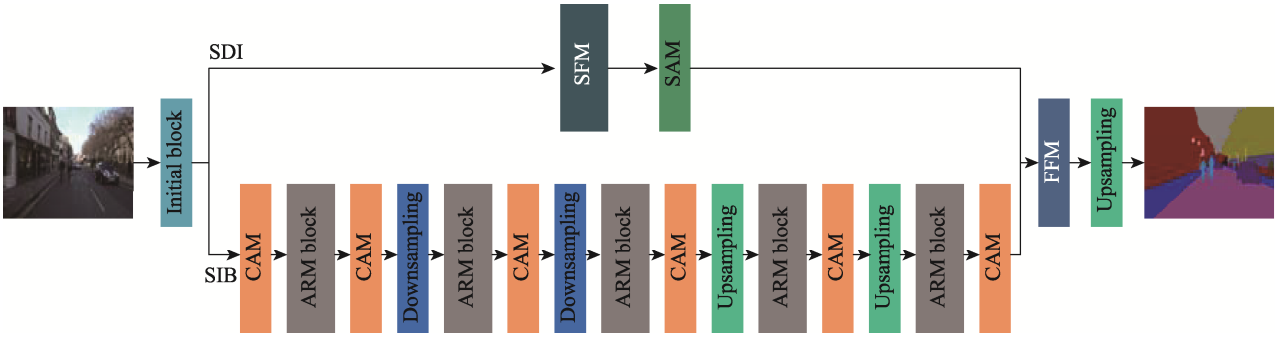


图 1　整体网络框架

Fig.1 Overall network framework

通道分割和通道混洗操作。尽管 LEDNet 在性能和速度之间取得了相对令人满意的平衡，但仍有一定的提升空间。受这些残差设计的启发，本文设计了高效的非对称残差（ARM）模块，利用非对称残差模块的共同优点，在计算能力有限的情况下获得更好的结果。非对称残差模块如图 2c 所示。首先在瓶颈处通过 1×1 卷积减少输入通道数。1×1 卷积后是双分支结构。一个分支使用分解卷积收集局部特征信息，另一个分支采用空洞卷积进一步扩大深度分离卷积的感受野，以捕获复杂和远程的特征信息。此外通过在不同的不对称残差模块中使用不同的膨胀率来降低网格化伪影的影响。

为了实现不同分支之间的信息共享，将特征交互操作放在只含分解卷积（3×1 和 1×3 卷积）分支和添加膨胀卷积分支之间，这样 2 个分支提取的上下文信息可以相互补充。然后将来自 2 个分支的特征图分别发送到通道注意力模块，以更好地提取判别特征。再将通道注意力模块的输出注入 2 个分支中进一步提升模块的特征提取能力。之后将 2 个分支提取的特征信息经过一个 1×1 的逐点卷积，恢复相关通道的特征图后融合并馈送到通道注意力模块中。最后使用通道混洗对双通道特征信息进行进一步交换和共享，减少深度卷积导致的通道间信息独立的影响。上述操作可以表示如下：

$$x_{\mathrm{o}} = C_{1\times1}(x_{\mathrm{ARM_{in}}}) \tag{1}$$

$$y_{\mathrm{o1}} = CA(C_{1\times3}(C_{3\times1}(x_{\mathrm{o}}) + C_{3\times1,R}(x_{\mathrm{o}}))) \tag{2}$$

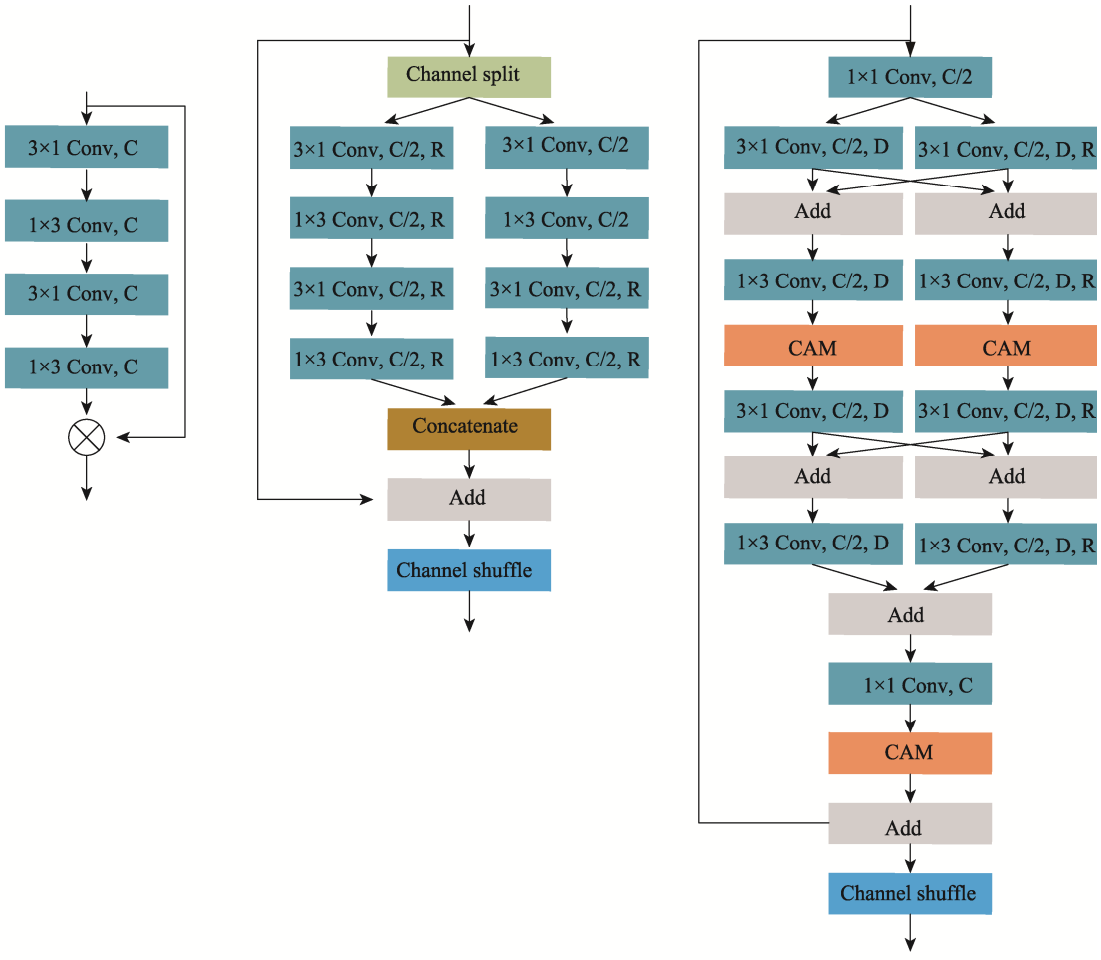
$$y_{\mathrm{o2}} = CA(C_{1\times3,R}(C_{3\times1}(x_{o}) + C_{3\times1,R}(x_{\mathrm{o}}))) \tag{3}$$

$$y'_{\mathrm{o1}} = C_{1\times3}(C_{3\times1}(y_{\mathrm{o1}}) + C_{3\times1,R}(y_{\mathrm{o2}})) \tag{4}$$

$$y'_{\mathrm{o2}} = C_{1\times3,\mathrm{R}}(C_{3\times1}(y_{\mathrm{o1}}) + C_{3\times1,R}(y_{\mathrm{o2}})) \tag{5}$$

$$y_{\mathrm{ARM_{out}}} = CS(CA(C_{1\times1}(y'_{\mathrm{o1}} + y'_{\mathrm{o2}})) + x_{\mathrm{ARM_{in}}}) \tag{6}$$

式中：$x_{\mathrm{ARM_{in}}}$ 和 $y_{\mathrm{ARM_{out}}}$ 为 ARM 模块的输入和输出；x_{o} 为 3×3 卷积的输出；y_{o1} 和 y_{o2} 为 ARM 模块中第 1 轮特征交互 2 个分支的输出；y'_{o1} 和 y'_{o2} 为 ARM 模块中第 2 轮特征交互 2 个分支的输出；$C_{m\times n}$ 为核大小为 $m\times n$ 的卷积运算；D 为可分离卷积；R 为膨胀卷积；$CS(\ )$为通道混洗操作。



a ERFNet的Non–bt–1D　　b LEDNet的SS–nbt　　c 本文的非对称残差模块ARM

图 2　残差模块对比

Fig.2 Comparison of various residual modules

1.3 语义信息分支

本文使用非对称残差模块构建了一个深度语义信息分支，这样既保证了能捕获到更多的语义信息，得到更大的感受野，同时又保证了参数的数量和计算成本非常低。非对称残差模块在不同阶段具有不同的表示能力：在网络浅层保留了丰富的空间信息，例如边缘和角落；而在网络深层阶段具有足够的语义一致性，但预测比较粗略。因此，在分支的不同阶段，在不对称残差模块中设置不同的空洞卷积率。将第 1 个到第 5 个 ARM 块中非对称残差模块数量分别设置为{1,2,3,4,5}。每个模块的扩张率分别依次设置为 r={1}、r={1,2}、r={1,2,5}、r={2,5,7,9}、r={2,5,7,9,17}。

本文在非对称残差模块和语义信息分支中都使用通道注意力模块（CAM）来强调需要突出显示的特征。同时该方法可以抑制干扰噪声，有利于特征提取。本文采用的通道注意力来源于 ECANet[13]，它只占用很少的计算资源，但相比之下明显提升了分割效果。CAM 使用全局平均池化来获取全局上下文，并生成注意力图来指导特征提取，计算成本可以忽略不计，这是提高模型性能的好方法。该过程可以表示为式（7）。

$$CA(F)=\delta(f_{K\times K}(T(AvgP(F))))\times F \tag{7}$$

式中：T 表示张量维度的压缩、转置和扩展操作；$f_{K\times K}$ 表示卷积核大小为 K 的标准卷积；$CA(F)$是通道注意力输出；F 表示输入特征；$AvgP(\ )$表示平均池化操作；δ 表示 Sigmoid 激活函数。

受 ERFNet 中下采样模块的启发，本文使用的下采样模块有 2 个替代输出，一个是步长为 2 的 3×3 卷积，另一个是步长为 2 的 2×2 最大池化。如果输入通道的数量大于或等于输出通道的数量，下采样模块使用单个 3×3 卷积。否则利用最大池化操作将这 2 个分支的连接形成最终的下采样输出。具体过程如图 3 所示。

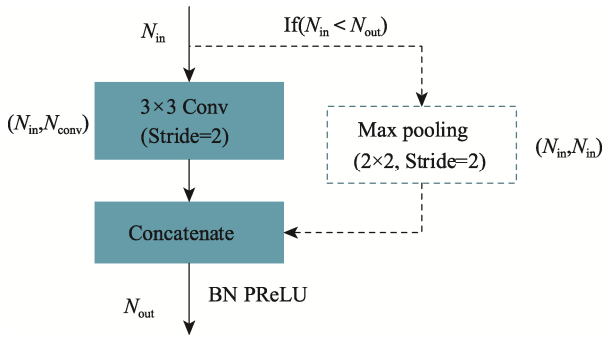


图 3 下采样模块

Fig.3 Downsampling module

1.4 空间细节分支

在语义信息分支的处理过程中，空间信息不可避免地会丢失。原因是深层语义信息的提取与浅层边界信息的保留是一对矛盾的关系。为了解决这个问题，本文设计了空间细节分支，它实际上是对语义丢失的细节信息的补充信息分支，以帮助模型在预测过程中实现更好的准确性。与深度语义信息分支不同，在这个分支中只使用了一个简单有效的空间特征提取模块（SFM）和一个空间注意力模块。SFM 是专门为补充语义分支中丢失的细节而设计的，如图 4 所示。它由 3 个 3×3 的卷积层和一个 1×1 的逐点卷积层组成。为了获取更多的特征信息，将第 2 和第 3 卷积层的通道数增加到原始输入的 4 倍（4C）。最后使用一个 1×1 的卷积层再将通道数减少到 C，该操作可以去除冗余特征并提取有效特征。为了减少参数数量和计算成本，将后面的 2 个 3×3 卷积层替换为深度可分离卷积，因此空间特征提取模块可以以较少的参数和计算成本提取丰富的浅层空间特征。

空间注意力模块用于提取和保存整个模型的浅层空间特征信息。空间特征提取模块输出的特征图作为输入，通过最大池化和平均池化进行池化处理，然后将池化后的结果进行融合后经过一个卷积层将双通道的特征信息降维为一维特征信息，经过激活函数生成空间注意力特征图。空间注意力的过程如式（8）所示。

$$SA(F)=\delta(f_{7\times 7}(Concat[AvgP(F),MaxP(F)]))\times F \tag{8}$$

式中：$f_{7\times 7}$ 为卷积核大小为 7 的标准卷积；$SA(F)$为空间注意力特征图；F 为输入特征；$Concat[\ ]$为连接操作；$AvgP(\ )$为平均池化操作；$MaxP(\ )$为平均池化操作；δ 为 Sigmoid 激活函数。

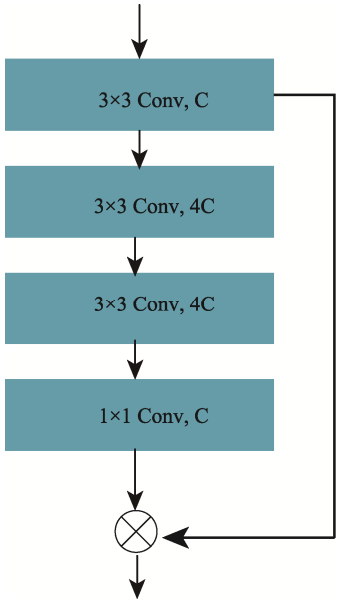


图 4 空间特征提取模块

Fig.4 Spatial feature extraction module

1.5　特征融合模块

如何有效地整合语义分支和空间分支的信息是双分支结构的关键问题。最广泛使用的方法是直接按元素添加或者直接连接它们。但是这些方法忽略了 2 个分支提供的功能之间的差异。为了解决这个问题，本文使用了由注意力机制驱动的方法[14]构建特征融合模块。该方法不仅可以捕获跨通道信息，还可以获取方向和位置感知信息，最重要的是它的计算成本较小，这意味着更少的参数可以换取更多的收益。

特征融合模块通过 2 个过程实现对通道关系和远程依赖进行编码：坐标信息嵌入和坐标注意生成。特征融合模块（FFM）的结构如图 5 所示。给定一个输入 $X\in R^{C\times H\times W}$，使用池化内核的 2 个空间维度$(1,W)$和$(H,1)$分别沿水平坐标和垂直坐标对每个通道进行编码。高度 h 处的第 c 个通道的输出可以表示为式（9）；长度为 w 的第 c 个通道的输出见式（10）。

$$A_c^h(h)=\frac{1}{W}\sum_{0\leqslant i<W}x_c(h,i) \tag{9}$$

$$A_c^w(w)=\frac{1}{H}\sum_{0\leqslant j<H}x_c(j,w) \tag{10}$$

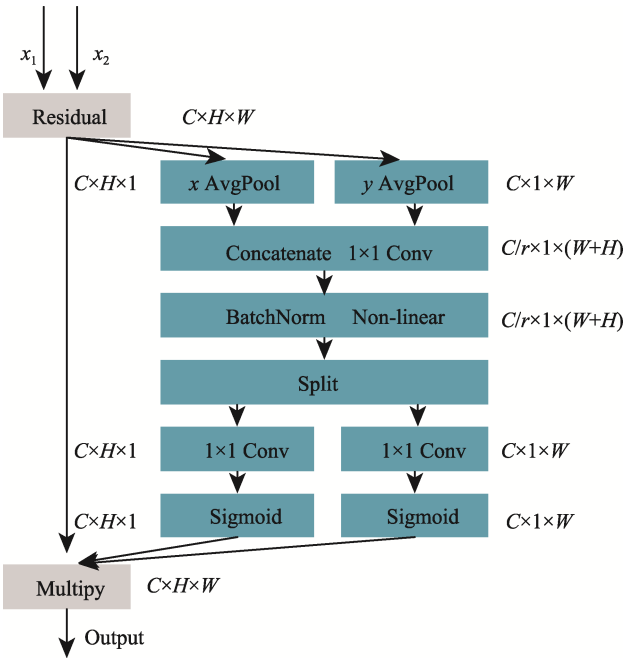


图 5　特征融合模块

Fig.5 Feature fusion module

上述 2 个变换分别沿 2 个空间方向聚合特征，产生一对方向感知特征图。这 2 个转换使用注意力模块一个沿空间方向捕获远程依赖关系，另一个沿空间方向保留精确的位置信息。由给定式（9）和式（10）的步骤生成聚合特征图，首先将它们连接起来，然后将它们发送到一个共享的 1×1 卷积变换函数 $f_{1\times1}$，具体过程如式（11）所示。

$$F=\delta(f_{1\times1}(Concat[z^h,z^w])) \tag{11}$$

式中：$Concat[\ ,\ ]$为沿空间维度的连接操作；δ 为非线性激活函数；$F\in R^{C/r\times(H+W)}$为在水平方向和垂直方向 2 个方向上编码空间信息获得的中间特征图；r 为用于控制块大小的缩小率。将 F 沿空间维度拆分为 2 个单独的张量 $f_h\in R^{C/r\times H}$ 和 $f_w\in R^{C/r\times W}$。使用 2 个 1×1 卷积变换分别将 f_h 和 f_w 分别变换为与输入 X 具有相同通道数的张量，具体过程见式（12）—（13）。

$$g_h=\delta(f_{1\times1}(f_h)) \tag{12}$$

$$g_w=\delta(f_{1\times1}(f_w)) \tag{13}$$

为了降低模型的复杂性，将缩小率 r 设置为 32 来减少 F 的通道数。然后将 g_h 和 g_w 分别用作注意力权重，通过该方法可以将 2 个分支的特征充分融合，同时在通道和空间方向下自适应突出特征信息。最终得到坐标注意块的输出，见式（14）。

$$y(i,j)=x(i,j)\times g_w(i)\times g_h(j) \tag{14}$$

1.6　网络详细结构参数

整个网络结构主要可分为 3 个部分：初始块、双分支主干和特征融合模块。完整的网络结构见图 1，详细的网络结构组成见图 6。语义信息分支 SIB 为编码器–解码器结构，而在空间细节分支 SDI 中空间特征提取模块 SFM 完成了“编码–解码”的过程。在图 6 中体现为空间细节分支的整个过程对应语义信息分支的编码器过程，最终 2 个分支输出相同尺寸的特征图。

2　实验

本文提出的模型将会在公开数据集 Camvid 和 Cityscapes 上进行分割效果和推理速度的实验，采用的评价指标分别为类交并比（class IoU）、均交互比（mIoU）、帧率（FPS）、参数量（parameters）。mIoU 的计算公式如式（15）所示。

$$P=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{i=0}^{k}p_{ij}+\sum_{i=0}^{k}p_{ji}-p_{ii}} \tag{15}$$

式中：p_{ij} 表示将 i 预测为 j，为假负（FN）；p_{ji} 表示将 j 预测为 i，为假正（FP）；p_{ii} 表示将 i 预测为 i，为真正（TP）。

2.1　实验环境

本文使用 PyTorch 深度学习框架实现训练，所有的实验都是在单块 RTX2080Ti GPU 上执行的。对 CamVid 数据集进行训练时，由于输入分辨率不同，采用 Adam 优化器训练神经网络，batch_size 设置为 8，权重衰减设置为 2×10^{-4}，此外将初始学习率设置为 1×10^{-3}。对于 Cityscapes 数据集，通过随机梯度下降的方法来训练本文提出的算法。batch_size 设置为 4，权重衰减设置为 1×10^{-4}，初始学习率配置为 4.5×10^{-2}，超参数 momentum 设置为 0.9。为了保证实验结果具有可对比性，本文所有实验均使用 CrossEntropy 损失函数，采用 poly 学习策略来动态调整学习率。

Type (size: C×H×W)	
Initial Block (16×256×512)	
SIB	SDI
CAM (16×256×512)	SFM
ARM(r={1}) (16×256×512)	(16×256×512)
CAM (16×256×512)	(64×256×512)
Downsample Block (64×128×256)	(64×256×512)
ARM(r={1,2}) (64×128×256)	(16×256×512)
CAM (64×128×256)	
Downsample Block (128×64×128)	
ARM(r={1,2,5}) (128×64×128)	
CAM (128×64×128)	SAM (16×256×512)
Upsample Block (64×128×256)	
ARM(r={2,5,7,9}) (64×128×256)	
CAM (64×128×256)	
Upsample Block (16×256×512)	
ARM(r={2,5,7,9,17}) (16×256×512)	
CAM (16×256×512)	
FFM (16×256×512)	
Projection Layer (19×512×1024)	

图 6 双分支特征提取网络细节

Fig.6 Details of two-branch feature extraction network

2.2 数据集

Camvid 是一个从驾驶汽车角度拍摄的街景数据集，它总共包括 701 张图片，其中 367 张图片用于训练，101 张用于验证，233 张用于测试。这些图像的分辨率为 960×720，共有 11 个语义类别，在训练前将这些图片尺寸调整为 360×480 的大小。

Cityscapes 是一个城市景观数据集。它包含 5 000 张精细标注和 20 000 张粗标注图像。该数据集是从 50 个不同城市在不同季节和天气中捕获的。对于精细标注集，它包含 2 975 张训练图像、500 张验证图像和 1 525 张测试图像。原始图像的分辨率为 1 024×2 048。整个数据集包含 19 个类别，在训练前将这些图片尺寸调整为 512×1 024 的大小。

2.3 消融实验

为了验证本文提出的网络的可行性和有效性，对各个模块的结构细节和分割效果在 Camvid 数据集上进行对比实验。在未加入其他模块的情况下，保证网络其余结构参数不变进行消融实验，最终结果如表 1 所示。

表 1 消融对比实验

Tab.1 Ablation contrast experiment

Models	SIB	SDI	CA	SA	Add	Concat	FFM	Para 值/M	mIoU 值/%
TBFENet	✓							0.872 9	68.27
TBFENet(CA)	✓		✓					0.873 1	68.95
TBFENet	✓	✓	✓		✓			0.903 2	69.12
TBFENet	✓	✓	✓			✓		0.905 6	69.56
TBFENet(FFM)	✓	✓	✓				✓	0.905 0	70.13
TBFENet	✓	✓	✓				✓	0.905 0	70.13
TBFENet(SA)	✓	✓	✓	✓			✓	0.905 1	**70.58**
TBFENet（r={1,1,2,1,2,2,5,7,9,17}）	✓	✓	✓	✓			✓	0.528 6	65.75
TBFENet(r={1,1,2,2,5,1,1,2,2,4,4,8,8,16,16})	✓	✓	✓	✓			✓	0.904 7	70.16
TBFENet(r={1,1,2,1,2,5,2,5,7,9,2,5,7,9,17})	✓	✓	✓	✓			✓	0.905 1	**70.58**

注：✓表示添加此模块。

2.3.1　通道注意力

从表 1 实验的前 2 行可以看出，如果不使用通道注意力模块，网络的预测结果会更差。CAM 可以提升网络 0.68%的分割精度，而计算成本几乎没有增加。实验证明了通道注意力模块的添加增强了网络的特征提取能力。

2.3.2　特征融合

特征融合方法一直是多语义特征聚合的重点研究课题，其中“添加”和“连接”操作是使用最广泛的方法。在表 1 中提供“Add”“Concat”和 FFM 的比较。根据表格第 6 行可知，FFM 达到了 70.13%的局部最佳性能，分别比“添加”和“连接”操作高出 1.01%和 0.57%。与“Add”操作相比，特征整合模块只增加了极少的参数（0.001 8 M）。此外，与“Concat”直接连接操作相比，FFM 以更少的参数实现了更好的分割结果，在不增加模型复杂度的情况下有效提升模型的性能。

2.3.3　空间注意力

空间注意力机制（SA）的添加使得网络的分割准确率提升了 0.45%，达到了 70.58%的最佳性能，而增加的参数量几乎可以忽略不计。说明浅层空间的特征信息提取对网络性能的提升有很大的作用。

2.3.4　扩张率

如表 1 实验第 4 部分所示，本文设计了 3 个实验来验证非对称残差模块中空洞卷积率的设置对模型分割精度的影响。首先将第 1 个到第 5 个 ARM 块中将非对称残差模块数量分别设置为{1,1,2,2,4}，每个模块的扩张率分别依次设置为 r={1,1,2,1,2,2,5,7,9,17}；第 2 和第 3 个实验将非对称残差模块的重复次数都分别设置为{1,2,3,4,5}，其中将第 2 个实验的扩张率依次设置为 r={1,1,2,2,5,1,1,2,2,4,4,8,8,16,16},第 3 个实验的扩张率设置为 r={1,1,2,2,5,1,2,5,7,9,2,5,7,9,17}。得益于模型框架的优异性，TBFENet 在第 1 个实验中仅用 0.52 M 参数就取得了 65.75%的分割准确性。增加非对称残差模块后实验结果显著提升，证明更多的 ARM 模块可以提升性能，而空洞卷积的使用进一步增强了网络的特征提取能力。在第 3 个实验中实现了 70.58%的最优分割结果。

2.4　Camvid 数据集测试实验

为了进一步验证本文网络的分割性能，在 CamVid 测试数据集上提供了与其他优秀分割方法的定量比较，实验结果如表 2 所示。根据表 2 可以明显看出，与类似模型大小的方法相比，本文分割网络达到了最佳的分割效果，均交互比达到了 70.5%。虽然在参数量表现上不如 DABNet[15]，但在分割精度上高出 DABNet 4.0%。相比于 LEDNet，本文模型得益于空间细节信息的保留分割更加精确。与其他大型模型相比，本文网络以更少的参数取得了最优的分割结果。在推理速度方面，本文模型推理速度达到了 107 帧/s。本文模型在实现轻量化的同时分割准确性表现依旧出色。充分证明本文网络在准确性和效率之间取得了很好的平衡。为了更清晰地体现本文模型在 Camvid 数据集上分割的效果，将本文模型得到的语义分割掩码，并与其他优秀网络模型进行对比，对比效果如图 7 所示。通过图 7 中本文网络分割图圈出的部分可以明显看出，本文模型在边界细节特征信息的提取明显优于 DABNet 和 BiseNet v2 模型在边界细节特征信息的提取，充分证明空间特征提取模块的有效性。

表 2　Camvid 数据集测试结果对比
Tab.2 Comparison of Camvid dataset test results

Model	Inputsize	Para/M	FLOPs/G	Speed/（帧·s^{-1}）	mIoU 值/%
ENet	360×480	**0.36**	3.8	103	51.3
SegNet[16]	360×480	29.50	286	37	55.6
SwiftNet[17]	720×960	11.80	104	31	62.2
DFANet[18]	720×960	7.80	**3.4**	**121**	63.9
ICNet	720×960	26.50	28.3	29	64.2
EDANet[19]	360×480	0.68	—	81	65.7
BiseNet	720×960	5.80	14.8	89	65.9
DABNet[20]	360×480	0.76	10.5	27	66.5
LEDNet	360×480	0.95	—	75	67.1
BiseNet v2	720×960	49.00	55.3	103	69.3
ours	360×480	0.91	9.3	107	**70.5**

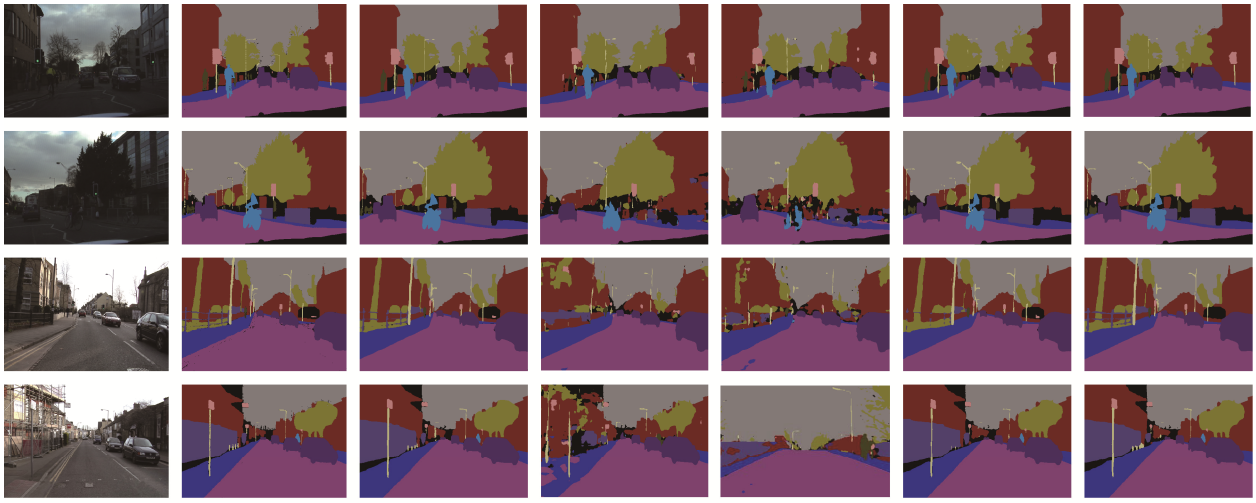


图 7　Camvid 数据集分割效果对比

Fig.7 Comparison of Camvid data set segmentation effect

2.5　Cityscapes 数据集测试实验

表 3 中提供了在 Cityscapes 测试数据集上与其他最先进的图像语义分割方法的定量比较。根据这些实验结果可以发现，当使用更少的参数时，本文网络可以实现更好的准确性和更快的运行速度。与本文方法具有相似数量参数的模型达不到相同的实时效果，即使实时效果更优，在分割精度上也大幅落后于本文算法。具有相同分割和实时效果的模型往往需要更多的参数运算。从参数量的角度看，ENet、ESPNet、CGNet[21]、NDNet[22]的参数量较少，但它们的分割精度分别比本文网络的低 16.8%、14.9%、10.8%和 10%，这在分割领域是一个很大的差距。本文算法的参数数量最多只比上述网络的多 0.55 M，相对于分割精度的提升，参数量的增加是在可接受范围之内的。从实时性的角度来看，本文算法推理速度达到了 97 帧/s，满足实时处理街景画面的条件。就均交互比来说，本文模型取得了 75.1%的最好分割效果，本文模型不仅在分割准确性上大幅领先其他优秀网络，在网络轻量化层面，参数数量也仅有 0.91 M，与分割效果较好的 BiseNet v2 相比，参数量仅约为 BiseNet v2 的 1/50。本文模型参数较少但推理速度较慢的原因是在网络中使用了注意力机制，而这些注意力机制会带来一些计算开销，导致推理速度变慢，但这些性能损失是在可以接受的范围之内的。

此外在表 4 中提供了城市景观的所有类 IoU(%)的结果。本文算法在 13 个类别中的分割精度领先于其他优秀网络在 13 个类别中的分割精度，而在交通标志类（Tsi）和自行车（Bic）类分割准确性只比最佳模型略低了 0.2%。通过实验证明了本文算法在模型的准确性、模型大小和推理速度之间实现了最佳平衡。

表 3　Cityscapess 数据集测试结果对比

Tab.3 Comparison of Cityscapess dataset test results

Models	Inputsize	Backbone	Pretrained	Para/M	Speed/（帧·s^{-1}）	mIoU 值/%
SegNet	640×360	VGG16	ImageNet	29.50	17	56.5
ENet	512×1 024	No	No	**0.36**	98	58.3
ESPNet	512×1 024	ESPNet	No	**0.36**	113	60.2
CGNet	360×640	No	No	0.50	50	64.3
NDNet	1 024×2 048	No	No	0.50	40	65.1
EDANet	512×1 024	No	No	0.68	81	66.7
ERFNet	512×1 024	No	No	2.10	42	67.5
Fast-SCNN	1 024×2 048	No	—	1.11	120	68.0
BiseNet	768×1 536	Xception	ImageNet	5.80	106	68.4
ICNet	1 024×2 048	PSPNet50	ImageNet	26.50	30	69.5
DABNet	1 024×2 048	No	No	0.80	28	70.1
LEDNet	512×1 024	No	No	0.90	52	70.3
MSCFNet	512×1 024	No	No	1.15	50	70.9
DFANet	512×1 024	Xception	No	7.80	108	71.5
BiseNet v2	512×1 024	No	ImageNet	3.40	**126**	73.3
ours	512×1 024	No	No	0.91	97	**75.1**

表 4　Cityscapes 类别分割精度对比实验

Tab.4 Comparative experiment on segmentation accuracy of Cityscapes

测试集	SegNet	ENet	ESPNet	ERFNet	ICNet	DABNet	LEDNet	DFANet	ours
Roa	96.4	96.3	97	97.7	97.1	97.9	98.1	98.1	**98.9**
Sky	91.8	90.6	92.6	94.2	93.5	92.8	94.9	94.9	**95.5**
Car	89.3	90.6	92.3	92.8	92.6	93.7	90.9	94.3	**95.3**
Veg	87	88.6	90.8	91.4	91.5	91.8	92.6	92.4	**93.5**
Bui	84	75	76.2	89.8	89.7	90.6	91.6	91.6	**92.4**
Sid	73.2	74.2	77.5	81	79.2	82	79.5	83.1	**84**
Ped	62.8	65.5	67	76.8	74.6	78.1	76.2	80.4	**83.9**
Bus	43.1	50.5	52.5	60.1	**72.7**	63.7	64	60.9	67.3
Tsi	45.1	44	46.3	65.3	63.4	67.7	**72.8**	71.4	72.6
Bic	51.9	55.4	57.2	61.7	70.5	66.8	**71.6**	67.7	71.4
Ter	63.8	61.4	63.2	68.2	68.3	70.1	61.2	69.7	**71.4**
Tli	39.8	34.1	35.6	59.8	60.4	63.5	61.3	67.2	**68.3**
Rid	42.8	38.4	40.9	57.1	56.1	57.8	53.7	61.1	**63.9**
Pol	35.7	43.4	45	56.3	61.5	59.3	**62.8**	62.6	62.4
Tra	44.1	48.1	50.1	51.8	51.3	**56**	52.7	52.5	53.1
Mot	35.8	38.8	41.8	47.3	53.6	51.3	44.4	55.3	**58.8**
Wal	28.4	32.2	35	42.5	43.2	45.5	47.7	45.4	**50.2**
Fen	29	33.2	36.1	48	48.9	50.1	49.9	50.6	**53.7**
Tru	38.1	36.9	38.1	50.8	51.3	52.8	**64.4**	50	52.1
Avg	57	58.3	60.3	68	69.5	70.1	70.3	71.5	**75.1**

3　结语

本文提出了一个用于彩色图像分割的双分支特征提取网络。本文算法主要侧重于在分割精度、模型参数和推理速度之间取得较好的平衡。实验证明，本文提出的非对称残差模块通过深度可分离卷积和空洞卷积在减少参数计算的情况下扩大感受野，全面地提取语义信息。语义信息分支和空间细节分支可以分别提取深层语义信息并保留各边界细节。本文模型在只有 0.91 M 参数的情况下，在 Cityscapes 数据集上以 97 帧/s 速度实现 75.1%的最佳分割准确性，在 Camvid 数据集上以 107 帧/s 的速度取得了 70.5%的最优分割效果。通过大量实验证明本文模型在准确性和效率之间取得了较好的平衡。

参考文献：

[1] 卢印举，郝志萍，戴曙光. 融合双特征的玻璃缺陷图像分割算法[J]. 包装工程, 2021, 42(23): 162-169.
LU Yin-ju, HAO Zhi-ping, DAI Shu-guang. Glass Defect Image Segmentation Algorithm Fused with Dual Features[J]. Packaging Engineering, 2021, 42(23): 162-169.

[2] 孙刘杰，张煜森，王文举，等. 基于注意力机制的轻量级 RGB-D 图像语义分割网络[J]. 包装工程, 2022, 43(3): 264-273.
SUN Liu-jie, ZHANG Yu-sen, WANG Wen-ju, et al. Lightweight RGB-D Image Semantic Segmentation Network Based on Attention Mechanism[J]. Packaging Engineering, 2022, 43(3): 264-273.

[3] 吴世燃，严国平，杨小俊. 纸塑复合袋表面缺陷图像分割算法的设计与实现[J]. 包装工程, 2021, 42(1): 244-249.
WU Shi-ran, YAN Guo-ping, YANG Xiao-jun. Design and Implementation of Image Segmentation Algorithm for Surface Defects of Paper Plastic Composite Bag[J]. Packaging Engineering, 2021, 42(1): 244-249.

[4] PASZKE A, CHAURASIA A, KIM S, et al. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation[EB/OL]. 2016: arXiv: 1606.02147. https://arxiv.org/abs/1606.02147.

[5] 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), Munich, 2018: 552-568.

[6] 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), Munich, 2018: 405-420.

[7] ROMERA E, ALVARE J M, BERGASA L M, et al. Erfnet: Efficient Residual Factorized Convnet for Real-Time Semantic Segmentation[J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 19(1): 263-272.

[8] 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, Springer, Cham, 2015: 234-241.

[9] YU Chang-qian, WANG Jing-bo, PENG Chao, et al. BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation[C]// Proceedings of the European conference on computer vision (ECCV), Munich, 2018: 325-341.

[10] POUDEL R P K, LIWICKI S, CIPOLLA R. Fast-SCNN: Fast Semantic Segmentation Network[EB/OL]. 2019: arXiv: 1902.04502. https://arxiv.org/abs/1902.04502.

[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] WANG Yu, ZHOU Quan, LIU Jia, et al. LEDNet: A Lightweight Encoder-Decoder Network for Real-Time Semantic Segmentation[C]// IEEE International Conference on Image Processing (ICIP), Taipei, 2019: 1860-1864.

[13] WANG Q, WU B, ZHU P, et al. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, IEEE, 2020.

[14] 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, 2021: 13713-13722.

[15] GAO G, YU Y, XIE J, et al. Constructing Multilayer Locality-Constrained Matrix Regression Framework for Noise Robust Face Super-Resolution[J]. Pattern Recognition, 2021, 110: 107539.

[16] 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.

[17] ORSIC M, KRESO I, BEVANDIC P, et al. In Defense of pre-Trained Imagenet Architectures for Real-Time Semantic Segmentation of Road-Driving Images[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, 2019: 12607-12616.

[18] 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, Long Beach, 2019: 9522-9531.

[19] LO S Y, HANG H M, CHAN S W, et al. Efficient Dense Modules of Asymmetric Convolution for Real-Time Semantic Segmentation[M]. Proceedings of the ACM Multimedia Asia, 2019: 1-6.

[20] LI Gen, YUN I, KIM J, et al. DABNet: Depth-Wise Asymmetric Bottleneck for Real-Time Semantic Segmentation[EB/OL]. 2019: arXiv: 1907.11357. https://arxiv.org/ abs/1907.11357.

[21] WU Tian-yi, TANG Sheng, ZHANG Rui, et al. CGNet: A Light-Weight Context Guided Network for Semantic Segmentation[J]. IEEE Transactions on Image Processing, 2020, 30: 1169-1179.

[22] YANG Z, YU H, FU Q, et al. Ndnet: Narrow While Deep Network for Real-Time Semantic Segmentation[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(9): 5508-5519.

责任编辑：曾钰婵