计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(1):76−86 [doi: 10.15888/j.cnki.csa.009358]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

交叉特征融合和 RASPP 驱动的场景分割方法①

朱新杰[1], 熊风光[1,2,3], 谢帅康[1], 宋宁栋[1], 李文清[1]

[1](中北大学 计算机科学与技术学院, 太原 030051)
[2](中北大学 山西省视觉信息处理及智能机器人工程研究中心, 太原 030051)
[3](中北大学 机器视觉与虚拟现实山西省重点实验室, 太原 030051)
通信作者: 熊风光, E-mail: hopenxfg@nuc.edu.cn

摘　要: 本文针对场景中目标多样性和尺度不统一等现象造成的边缘分割错误、特征不连续问题, 提出了一种交叉特征融合和 RASPP 驱动的场景分割方法. 该方法以交叉特征融合的方式合并编码器输出的多尺度特征, 在融合高层语义信息时使用复合卷积注意力模块进行处理, 避免上采样操作造成的特征信息丢失以及引入噪声的影响, 细化目标边缘分割效果. 同时提出了深度可分离残差卷积, 在此基础上设计并实现了结合残差的金字塔池化模块——RASPP, 对交叉融合后的特征进行处理, 获得不同尺度的上下文信息, 增强特征语义表达. 最后, 将 RASPP 模块处理后的特征进行合并, 提升分割效果. 在 Cityscapes 和 CamVid 数据集上的实验结果表明, 本文提出方法相比现有方法具有更好的表现, 并且对场景中的目标边缘有更好的分割效果.

关键词: 语义分割; 交叉特征融合; 金字塔池化; 注意力机制; 深度可分离卷积

引用格式: 朱新杰,熊风光,谢帅康,宋宁栋,李文清.交叉特征融合和 RASPP 驱动的场景分割方法.计算机系统应用,2024,33(1):76–86. http://www.c-s-a.org.cn/1003-3254/9358.html

Cross Feature Fusion and RASPP Driven Scene Segmentation Method

ZHU Xin-Jie[1], XIONG Feng-Guang[1,2,3], XIE Shuai-Kang[1], SONG Ning-Dong[1], LI Wen-Qing[1]

[1](School of Data Science and Technology, North University of China, Taiyuan 030051, China)
[2](Shanxi Province's Vision Information Processing and Intelligent Robot Engineering Research Center, North University of China, Taiyuan 030051, China)
[3](Shanxi Key Laboratory of Machine Vision and Virtual Reality, North University of China, Taiyuan 030051, China)

Abstract: This study proposes a cross feature fusion and RASPP-driven scene segmentation method to address the edge segmentation errors and feature discontinuity caused by target diversity and scale inconsistency in the scenes. This method combines the multi-scale features output by the encoder in the way of cross feature fusion and employs the compound convolution attention module to process high-level semantic information fusion. As a result, this avoids the feature information loss caused by the upsampling operation and the influence of noise and refines the segmentation effect of target edges. Meanwhile, this study proposes a depthwise separable convolution combining residual connections. Based on this, a pyramid pooling module RASPP combining residuals is designed and implemented to process the features after cross fusion, obtain contextual information at different scales, and enhance feature semantic expression. Finally, the features processed by the RASPP module are merged to improve the segmentation effect. The experimental results on the Cityscapes and CamVid datasets show that the proposed method outperforms existing methods and has better segmentation performance on target edges in the scenes.

Key words: semantic segmentation; cross feature fusion; pyramid pooling; attention mechanism; depthwise separable convolution

① 基金项目: 国家自然科学基金 (62272426); 山西省回国留学人员科研基金 (2020-113); 山西省科技成果转化引导专项基金 (202104021301055); 山西省科技重大专项计划"揭榜挂帅"基金 (202201150401021); 山西省自然科学基金 (202203021212138, 202303021211153, 202203021222027)
收稿时间: 2023-06-28; 修改时间: 2023-08-08; 采用时间: 2023-08-18; csa 在线出版时间: 2023-11-24
CNKI 网络首发时间: 2023-11-28

图像语义分割作为计算机视觉领域的基础性课题, 其目的是给图像中的每一个像素分配其所属的类别, 在图像理解和场景感知方面有着非常重要的意义[1]. 在医学图像处理[2]、遥感图像分割[3]以及无人驾驶[4]等方面的具体应用都展现了语义分割的强大能力.

相较于传统的基于边缘[5]、区域[6]、阈值[7]的图像分割算法, 基于深度学习的语义分割方法利用深度神经网络强大的特征表示能力, 通过对图像不同层次特征的学习, 提高了分割的精度, 同时减少了人工提取特征的工作量, 更符合当今的场景分割任务需要. 但当前主流方法往往存在着目标之间语义特征不连续, 边界分割错误等问题. 为此, 本文提出一种交叉特征融合和结合残差的金字塔池化 (atrous spatial pyramid pooling with residual, RASPP) 驱动的场景分割方法, 以优化目标边缘分割效果, 提高分割精度为导向, 从改变对编码器输出的多阶段特征图的融合方式出发, 引入复合卷积注意力模块 (composite convolutional attention module, CCAM) 自适应调整高层特征信息权重, 捕获通道和空间维度特征交互关系, 并通过提出的 RASPP 模块对融合后的特征进行处理, 获取不同大小的感受野, 改善对不同尺度目标的分割效果, 从而解决目标间边缘分割错误、分割精度较低等问题.

1 相关工作

1.1 基于深度学习的场景分割

基于深度学习的场景分割方法往往采用编码器-解码器结构, 首先将图像输入到特征提取网络, 然后将提取到的语义特征送入解码器, 解析语义特征并得到分割图. 文献[8]提出了一种全卷积网络 (fully convolutional networks, FCN), 实现了端到端的图像分割网络, 奠定了主流的卷积神经网络分割架构. 文献[9]使用带有池化索引的上采样操作来降低池化层造成的信息丢失, 提升了分割效果. 文献[10]提出一种使用跳跃连接 (skip-connection) 结构融合不同层次特征的方法, 用于医学图像分割并提高了分割精度. 文献[11]使用金字塔池化模块 (pyramid pooling module, PPM) 提取不同尺度的上下文信息, 缓解了由于场景中存在不同大小的目标导致的分割精度降低问题. 文献[12]使用空洞卷积 (atrous convolution) 解决编码过程中由于多次下采样导致细节特征丢失的问题; 并使用全连接条件随机场 (fully-connected conditional random field) 提高模型捕获结构信息的能力. 文献[13]提出了空洞空间金字塔池化模块 (atrous spatial pyramid pooling, ASPP), 在金字塔池化模块的基础上结合空洞卷积, 在不降低特征图分辨率的同时增大网络的感受野, 有效缓解了分割不连续和边界粗糙等问题. DeepLabV3[14]将 ASPP 中的大小为 3×3, dilation=24 的空洞卷积替换为 1×1 的普通卷积, 保留了卷积块的有效权重. 文献[15]在 DeepLabV3 的基础上, 引入解码结构, 融合低层特征以恢复丢失的细节信息. 文献[16]提出一种新型上采样方法, 针对物体边缘分割进行优化, 在物体密集场景中有更好的分割效果. 文献[17]提出的 ViT 模型在分类任务中的优异表现证明了 Transformer[18]结构在 CV 领域的巨大潜力. 文献[19]在 ViT 的基础上提出了 SETR, 将 Transformer 结构应用于图像分割任务. Xie 等人针对 ViT 计算过程中特征图尺度单一问题, 提出了 SegFormer 模型[20], 舍弃了 ViT 中的位置编码, 避免了因为测试图像分辨率与训练图像不一致, 对位置编码进行上采样而导致的模型精度下降问题.

虽然, 主流的基于深度学习的场景分割方法在场景语义分割与解析方面取得了较好结果, 并有效提升了分割精度. 但是针对复杂场景中目标尺度差异较大、环境复杂多变等现象, 上述方法仍存在着目标边界分割效果较差, 分割精度不高等问题.

1.2 注意力机制

受人类视觉系统注意力机制启发, 注意力机制的基本思想就是让网络能够聚焦于重要信息, 抑制或忽略冗余信息. 注意力机制通常分为空间注意力机制、通道注意力机制以及空间-通道混合注意力机制, 能够有效地捕获不同位置或不同通道间的相关性. SENet[21]关注不同通道之间的重要程度, 将输入特征图通过全局平均池化压缩空间信息, 然后使用全连接层和激活函数, 获取每个通道的权重信息. CBAM[22]包含两个子注意力机制, 分别在通道维度和空间维度上计算权重, 更好地突出部分通道和空间的影响. DANet[23]提出一种双重注意力网络, 自适应地集成局部特征和全局依赖, 显著提升了分割效果. Non-Local[24]使用非局部信息注意力机制, 计算特征自相关性, 有效地融合了全局信息, 但是计算量较大. TAM[25]使用三重注意力机制来捕获通道维度和空间维度之间的相互作用, 获取全局感受野. DNLNet[26]在 Non-Local 的基础上, 将基于点乘的注意力分离为两个模块, 用于解决在 Non-Local

中的性能问题.

2 交叉特征融合和 RASPP 驱动的场景分割方法

本文方法采用 SegFormer 网络编码器结构, 主要针对多阶段特征融合方式进行优化, 结合复合卷积注意力机制突出重要特征信息, 设计并实现了 RASPP 模块用于改善特征不连续问题, 提升网络的分割精度.

SegFormer 网络采用编码器-解码器结构, 编码器结构由 4 个阶段组成, 每个阶段由多个 Transformer block 堆叠而成. 在每个 Transformer block 中采用重叠切片嵌入 (overlap patch embeddings) 模块对输入图片进行特征提取, 然后将提取到的特征通过高效多头自注意力 (efficient multi-head self-attention) 模块计算特征相关性, 最后将计算后的特征通过混合前馈网络 (mix feed forward network, MixFFN) 模块. SegFormer 在 MixFFN 中取代了位置编码, 而改用卷积核大小为 3×3 的深度可分离卷积来提供位置信息. 在网络解码器部分, 将编码器的 4 个阶段输出的分辨率分别为原图像 1/4、1/8、1/16、1/32 大小的特征图, 通过卷积核大小为 1×1 的卷积层将特征图的通道数统一调整为 256, 然后通过双线性插值将特征图上采样到原图像的 1/4 大小, 将多个同样大小的特征图进行拼接, 通过两个卷积核大小为 1×1 的卷积层完成像素级预测, 最后通过双线性插值将图片还原为原图像大小, 输出最终的分割结果, 分割流程如图 1 所示.

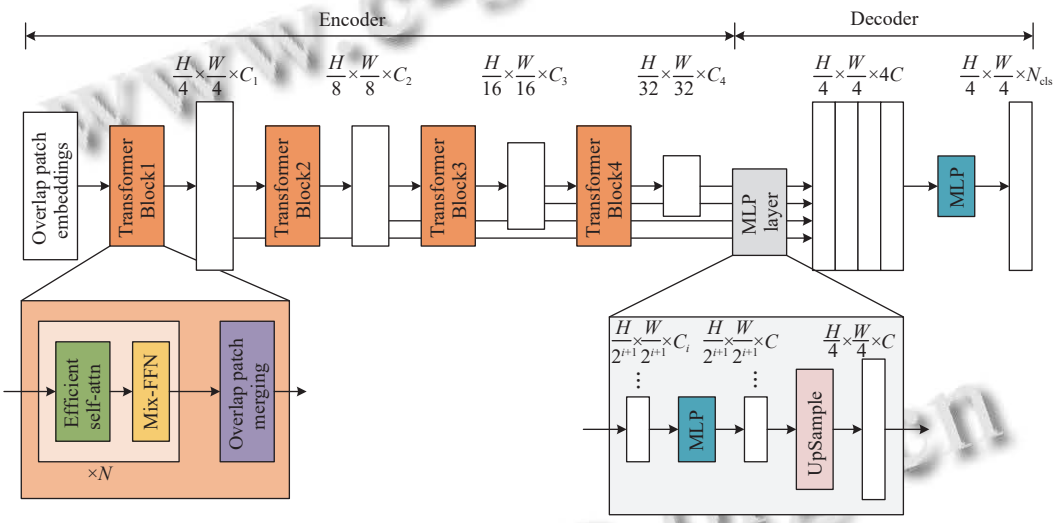


图 1 SegFormer 网络结构图

SegFormer 网络虽然在场景分割方面表现良好, 但在其解码器部分将多个阶段特征图一次性上采样到同样大小, 容易造成低层细节信息与高层语义信息融合不足, 同时直接将高层特征与低层特征进行拼接, 难免引入噪声, 导致分割精度下降. 除此之外, SegFormer 网络未引入多尺度上下文处理结构, 易导致不同尺度目标之间特征不连续、边缘分割错误等情况, 造成分割精度问题.

基于此, 本文在两方面提出创新性改进: 一是使用交叉特征融合方式替换原结构中直接上采样然后合并的方式, 将低层特征分阶段与高层特征进行融合, 并且在高低层特征融合之前使用复合卷积注意力模块对高层语义信息进行特征校准, 抑制低层冗余信息以及上采样过程引入的噪声的干扰; 二是提出深度可分离残差卷积模块 (depthwise separable convolution with residual, RDSConv), 在此基础上设计并实现了结合残差的金字塔池化模块 RASPP, 将交叉特征融合后的语义信息通过不同尺度的 RASPP 模块获取不同阶段的上下文信息, 加强特征之间的语义关联, 改善特征不连续问题, 提升图像分割精度.

2.1 交叉特征融合

本文遵循编码器-解码器结构搭建网络模型, 编码器部分采用 SegFormer 模型的 MiT-B2[20]作为主干网络. 模型整体架构如图 2 所示. 首先将各个阶段特征图进行交叉特征融合, 在保留低层特征丰富细节信息的同时减少噪声对高层语义的干扰, 同时使用复合卷积注意力模块校准高层特征权重, 然后将融合后的两部

分结果送入不同大小的 RASPP 多尺度特征提取模块, 提取不同粒度的多尺度信息, 然后将多尺度信息进行合并, 提升对不同尺度的目标分割效果, 将合并后的多尺度信息进行双线性插值上采样, 还原为输入图像大小, 最后通过 1×1 的卷积层调整通道数为类别数, 得到网络的输出结果.

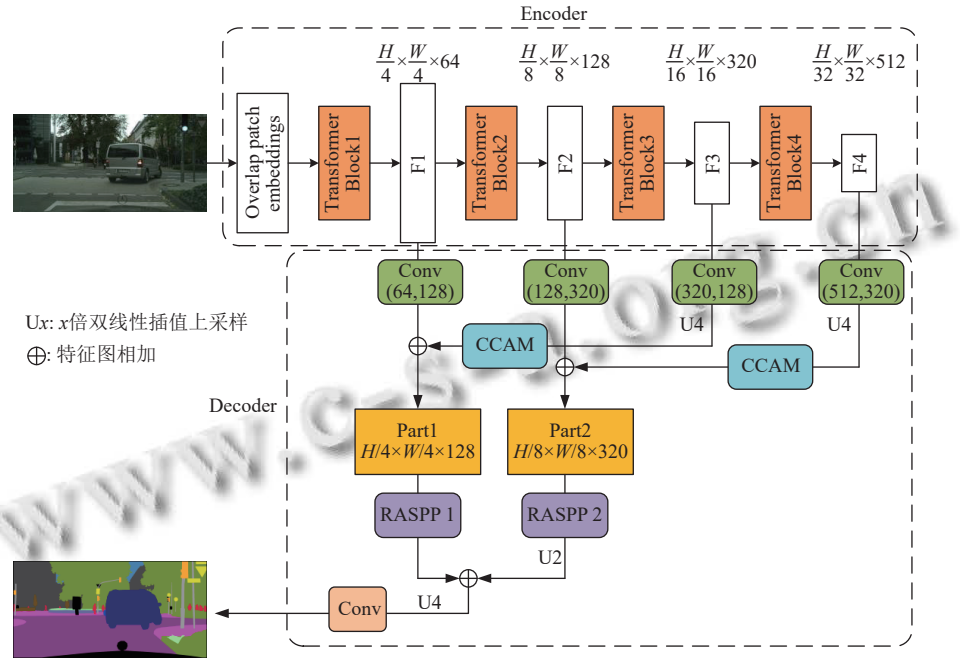


图 2 本文模型整体结构图

如图 1 所示, 在 SegFormer 原始网络结构的解码器部分, 将编码器各个阶段输出的不同大小的特征图上采样到统一尺寸, 然后直接拼接起来, 并通过 1×1 卷积调整通道数得到最后的分割图. 这种做法虽然具有较少的参数量和较低的运算成本, 但是将低层细节特征和高层语义特征直接融合, 在丰富了细节信息的同时不可避免地引入了大量噪声, 影响分割精度. 为此, 本文在 MiT-B2 主干网络的基础上, 对提取到的各个阶段特征图进行交叉融合. 如图 2 所示, 将主干网络得到的不同阶段的特征图命名为 F1、F2、F3 和 F4. 交叉特征融合分为两部分, 第 1 部分将特征图 F3 先通过 1×1 卷积调整通道数, 得到特征图大小为[H/16, W/16, 128], 然后进行 4 倍双线性插值上采样, 特征图分辨率由原图像的 1/16 上采样到原图像的 1/4, 再通过注意力模块 CCAM 调整特征间关系, 增强高层特征通道和空间信息; 同时将特征图 F1 通过 1×1 卷积调整通道数, 得到特征图大小为[H/4, W/4, 128], 将这两部分特征图相加, 得到第 1 部分的输出. 同理, 交叉融合的第 2 部分将特征图 F4 先使用 1×1 卷积调整通道数为 320, 然后进行 4 倍双线性插值上采样, 通过注意力模块 CCAM 后与调整通道数为 320 之后的特征图 F2 相加, 得到第 2 部分的输出. 由于编码器不同阶段输出的特征图包含的特征信息不同, 相比较而言, 特征图 F1 和 F2 的空间位置细节信息丰富, 有助于提升目标边缘等细节的分割效果, 但是语义信息相对不足, 而特征图 F3 和 F4 所含的抽象语义信息丰富, 但缺少空间细节. 如果将多个特征图直接进行融合, 虽然丰富了高层特征的细节信息, 但是低层特征图除细节信息外同样存在大量噪声, 直接进行融合反而不利于提高精度. 本文提出的交叉特征融合方式, 将特征图 F1 和 F4 与 F3 和 F2 这两个过渡阶段的特征图进行融合, 并使用注意力模块 CCAM 自适应调整高层特征权重, 既实现了高低层特征的融合, 又避免了低层噪声对分割结果的干扰.

2.2 复合卷积注意力模块

在将编码器输出的多阶段特征进行融合时, 往往会出现高低层特征图大小不匹配的问题, 因此引入特征图上采样操作, 将高层特征图上采样到与低层特征图同一尺寸, 方便特征间的融合. 但是上采样操作不可

避免会引入噪声信息, 对高层特征图富含的语义信息造成干扰. 因此, 本文提出在对高层特征图上采样操作之后, 通过复合卷积注意力模块 CCAM 调整特征空间关系, 增强上下文依赖, 减少噪声信息的干扰.

如图 3 所示, CCAM 模块首先将输入特征图 F_{in} 通过 1×1 的卷积模块以及 GELU 激活函数, 然后通过核心的 CCA Block, 再通过一个 1×1 的卷积块, 然后将得到的结果与输入特征图相加, 得到输出结果 F_{out}, 计算过程如式 (1) 所示:

$$F_{out} = F_{in} + Conv(CCA(GELU(Conv(F_{in})))) \quad (1)$$

其中, CCA block 的结构如图 4 所示. 其主要由 3 部分组成, 深度可分离卷积模块 *DSConv*、深度可分离膨胀卷积 *DSDConv* 以及点卷积 *PWConv*. 其中, *DSConv* 卷积核大小设置为 5×5, padding 设置为 2; *DSDConv* 卷积核大小设置为 7×7, padding 设置为 9, 膨胀率设置为 3; 点卷积则使用卷积核大小为 1×1 的普通卷积实现. 在该模块中, 输入特征图 X_{in} 依次通过 *DSConv*、*DSDConv*、*PWConv* 计算得到注意力权重, 然后将输入特征图与注意力权重相乘, 对输入特征图 X_{in} 重新进行加权, 得到输出特征 X_{out}, 计算过程如式 (2) 所示, 其中, *操作表示特征图间逐元素相乘.

$$X_{out} = X_{in} * PWConv(DSDConv(DSConv(X_{in}))) \quad (2)$$

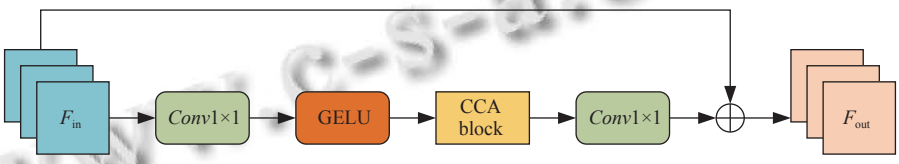


图 3 CCAM 模块结构图

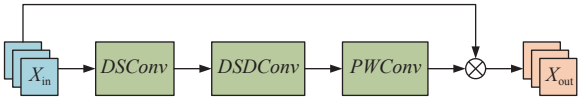


图 4 CCA block 结构图

本文提出的 CCA Block 并没有使用普通通道注意力机制使用的自适应平均池化或自适应最大池化来获取通道间特征关系, 而是通过使用 3 种卷积模块复合使用的方法增强特征信息. 首先使用深度可分离卷积将空间和通道两个维度的操作分离开, 然后使用深度可分离膨胀卷积获取输入特征图较大的感受野内容, 突出特征空间位置关系, 最后使用点卷积校准不同通道间的权重关系. 整体来说, CCA block 兼顾了通道维度和空间维度, 并且不需要使用激活函数, 简化了注意力图的计算, 能够做到即插即用.

2.3 RASPP 模块

由于场景中存在目标尺度不统一现象, 容易造成分割精度不高、边缘分割错误等问题. 因此, 获取特征图多尺度上下文信息, 从而提取不同尺度特征, 对于提高分割精度, 缓解不同尺度目标之间的差异造成的特征不连续问题非常重要. 为此, 本文设计并实现了 RASPP 模块, 如图 5 所示. RASPP 模块主要使用本文提出的深度可分离残差卷积模块进行搭建. 首先, 使用深度可分离卷积可以减少模型的参数量, 有利于模型的轻量化, 但同时深度可分离卷积将空间维度操作和通道维度操作分离开来, 易导致特征之间不连续问题. 因此本文提出深度可分离残差卷积, 将空间与通道维度特征统一起来, 在不引入额外参数的前提下增强特征间的语义表达, 同时避免出现由网络层次加深导致的梯度消失或梯度爆炸现象.

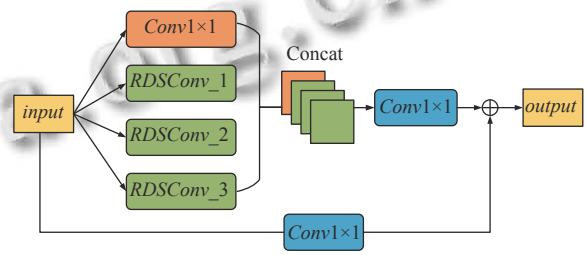


图 5 RASPP 模块示意图

RASPP 模块主要由 1×1 卷积和 3 个不同膨胀率的 RDSConv 模块组成. 普通深度可分离卷积 (depth-wise separable convolution, *DSConv*)[27]操作由两部分组成, 首先对输入特征图逐通道进行卷积核大小为 3×3 的膨胀卷积, 使其通道分离, 然后再通过 1×1 卷积进行通道维度的合并. *DSConv* 具体计算公式如式 (3) 所示:

$$output = Conv_{1\times1}(DConv_{3\times3}(input)) \quad (3)$$

其中, *input* 表示模块的输入特征图, *output* 表示其输出

特征图, $Conv_{1\times1}$ 表示卷积核大小为 1×1 的普通卷积, $DConv_{3\times3}$ 表示卷积核大小为 3×3 的膨胀卷积.

本文提出在普通深度可分离卷积中引入残差操作, 即对特征图进行 3×3 的膨胀卷积操作之后, 将卷积后的结果与输入相加, 然后再通过 1×1 的卷积得到输出. 不同于 *DSConv* 中将空间和通道两个维度分开来进行操作, *RDSConv* 将空间维度操作与通道维度操作结合起来, 增强语义特征关联, 缓解目标分割不连续、边缘分割不清晰问题. *RDSConv* 具体计算过程如式 (4) 所示:

$$output = Conv_{1\times1}(input + DConv_{3\times3}(input)) \quad (4)$$

在 RASPP 模块中, 将输入特征图通过 1×1 卷积和 3 个不同膨胀率的 *RDSConv*, 将得到的结果进行通道维度的拼接, 然后使用 1×1 卷积调整至指定通道数. 同时使用 1×1 卷积调整输入特征图至指定通道数, 将两者相加得到 RASPP 模块的输出, 具体操作过程如图 5 所示.

如图 2 所示, 对编码器多个阶段特征图进行交叉特征融合后得到两部分输出, 第 1 部分由特征图 F1 和特征图 F3 进行融合得到, 第 2 部分由特征图 F2 和特征图 F4 进行融合得到. 针对这两部分特征所含信息的不同, 使用不同膨胀率的 RASPP 模块进行多尺度特征的提取. 对于第 1 部分, 采用的 RASPP 模块中膨胀率为 (1, 3, 6, 9); 对于第 2 部分, 采用的膨胀率为 (1, 6, 12, 24), 膨胀率为 1 代表使用的是 1×1 的普通卷积. 其中每个卷积模块的输出通道均设置为 256. 如表 1 所示.

表 1 RASPP 模块参数设置

不同阶段	输入特征图大小	膨胀率设置	输入通道数	输出通道数
Part1	(*H*/4, *W*/4)	(1, 3, 6, 9)	128	256
Part2	(*H*/8, *W*/8)	(1, 6, 12, 24)	320	256

对于第 1 部分, 使用较小的膨胀率来构建 RASPP 模块, 主要是针对其特征中细节信息较为丰富的特点, 尽量避免冗余信息的干扰; 而对于语义信息丰富的第 2 部分, 则使用较大的膨胀率, 增大网络的感受野, 获取不同尺度下的语义信息, 提高网络的分割精度.

对于交叉特征融合得到的两部分特征图, 经过各自的 RASPP 模块之后, 得到大小为[*H*/4, *W*/4, 256]和[*H*/8, *W*/8, 256]的两部分特征图. 然后将第 2 部分特征图进行 2 倍双线性插值上采样, 得到的特征图大小为[*H*/4, *W*/4, 256], 与第 1 部分特征图维度相同, 最后将两部分融合之后的特征图进行相加, 通过双线性插值上采样, 再通过 1×1 的卷积调整特征图通道数, 得到网络的输出结果.

3 实验结果及分析

3.1 实验数据集

为验证本文方法的有效性, 本文使用 Cityscapes 和 CamVid 公开数据集进行实验. 数据集具体信息如下.

(1) Cityscapes 数据集提供了包括 50 多个城市在内的不同季节、不同天气情况下的街道场景共 5 000 张精细标注的图片, 数据集分为训练集、验证集和测试集, 包含图片数量分别为 2 975、1 525 和 500 张, 每张图片及其标注图像分辨率均为 1024×2048. 该数据集共分为天空、行人、汽车、道路和建筑物等 19 个语义类别.

(2) CamVid 数据集是经典城市场景分割数据集之一, 该数据集由车载摄像头拍摄获得, 提供了不同时段的街景图像共 701 张, 分为训练集 367 张, 验证集 101 张, 测试集 233 张, 每张图片及其标注图像分辨率均为 720×960. 该数据集共分为 32 个语义类别, 但由于该数据集类别较多且部分类别与其他类别的界限相对模糊, 因此本文从中选择包含背景在内的共 12 个类别进行训练与验证.

3.2 实验环境及训练配置

本文实验环境配置如下: GPU 显卡型号为 NVIDIA GeForce RTX 3090, 显存大小为 24 GB, 网络模型使用 PyTorch 1.10 版本开源框架进行搭建, Python 环境版本为 3.8.15. 各数据集训练配置如表 2 所示.

表 2 数据集训练配置

配置项	Cityscapes	CamVid
input size (原图像大小)	1024×2048	720×960
train size (训练图像大小)	512×1024	768×768
test size (测试图像大小)	1024×2048	720×960
iterations (训练迭代次数)	80 000	48 000
batch size (批次大小)	8	8
optimizer (优化器)	AdamW	AdamW
learning policy (学习策略)	poly	poly

为了保证模型在训练时的稳定性, 实验采用 AdamW 优化器[28], 使用 poly 学习策略, 初始学习率 lr_{base} 为 0.000 06, 动量 *power* 为 1.0, 权重衰减 *weight_decay* 为 0.01. poly 学习策略计算公式如式 (5) 所示. 其中, *iter* 为当前迭代次数, *totaliter* 为最大迭代次数.

$$lr = lr_{\text{base}} \times \left(1 - \frac{iter}{totaliter}\right)^{power} \tag{5}$$

为了便于查看模型训练时的状态, 以及统计模型训练信息, 采用 iteration 而不是 epoch 作为参数训练迭代单位.

3.3 评价指标

为了更好地评估本文提出方法的有效性, 采用平均交并比 (mean intersection over union, *mIoU*)、平均像素准确率 (mean pixel accuracy, *mPA*) 以及平均 *F*1-score (mean *F*1-score, *mFScore*) 作为模型的评价指标. *mIoU* 通过计算图像标签集合与网络预测结果集合的交集与并集之比, 并计算不同类别的平均值, 计算方式如式 (6) 和式 (7) 所示. *mPA* 通过计算网络预测正确像素点数与总预测像素个数之比, 并计算不同类别的平均值, 计算方式如式 (8) 所示.

$$IoU_i = \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{6}$$

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} IoU_i \tag{7}$$

$$mPA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{8}$$

mFScore 则是通过计算每个类别的 *F*1-score, 并计算不同类别的平均值. *F*1-score 的计算方式如式 (9) 所示:

$$F1_i = \frac{2 \times Recall_i \times Precision_i}{Recall_i + Precision_i} \tag{9}$$

其中, *Recall* 为召回率, *Precision* 为精确率, 计算方式分别为式 (10) 和式 (11) 所示:

$$Recall_i = \frac{p_{ii}}{\sum_{j=0}^{k} p_{ji} + p_{ii}} \tag{10}$$

$$Precision_i = \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + p_{ii}} \tag{11}$$

mFScore 的计算方式如式 (12) 所示:

$$mFScore = \frac{1}{k+1} \sum_{i=0}^{k} F1_i \tag{12}$$

式 (6)–式 (12) 中, k 为图像中有效标签类别数 (不包含背景类别), k+1 表示图像中的类别总数. p_{ii} 表示真实像素类别为 i, 并且被预测为类别 i 的像素数量, 即预测正确的像素个数; p_{ij} 和 p_{ji} 分别表示真实像素类别为 i, 被预测为类别 j 的像素数量和真实像素类别为 j, 被预测为类别 i 的像素数量.

3.4 实验结果分析

为测试本文提出的方法的有效性, 分别在 CamVid 和 Cityscapes 数据集上进行相关模块的消融实验, 并结合表 2 中的训练配置信息, 将本文方法与场景分割领域常用基准方法进行对比实验.

3.4.1 RASPP 模块消融实验

为验证本文提出的 RASPP 模块的有效性, 针对 RASPP 模块进行了消融实验. 首先定义 Baseline, 其计算流程为: 将交叉特征融合第 2 阶段得到的特征图进行上采样后, 将其与第 1 阶段得到的特征图通过 1×1 卷积调整通道数为 256, 然后直接进行相加, 并通过双线性插值上采样还原为输入图像大小, 得到分割结果. 在其基础上, 结合使用深度可分离残差卷积搭建的 RASPP 模块, 模型记为 Baseline+RASPP. 结合使用深度可分离卷积搭建的 ASPP 模块, 模型记为 Baseline+ASPP. 在 Cityscapes 和 CamVid 数据集上进行的消融实验结果如表 3 所示.

表 3 RASPP 模块消融实验结果 (%)

模型	Cityscapes		CamVid	
	mIoU	*mPA*	*mIoU*	*mPA*
Baseline	79.02	86.44	66.34	74.21
Baseline+ASPP	80.42	87.41	67.26	75.32
Baseline+RASPP	**81.16**	**87.89**	**68.01**	**75.88**

由表 3 可以看出, Baseline 将交叉特征融合后的两部分特征直接进行相加, 并进行上采样得到分割图, 没有考虑到不同尺度的上下文信息对分割精度的影响, *mIoU* 较低; 而使用 Baseline+RASPP, 针对不同阶段的特征图所含语义信息的不同, 采用不同膨胀率的 *RDSConv* 模块获取不同尺度信息, 有效提高了分割精度, 在 Cityscapes 和 CamVid 数据集上 *mIoU* 较 Baseline 分别提高了 2.14% 和 1.67%, 并且 *mPA* 较 Baseline 分别提高了 1.45% 和 1.67%. 此外, 还将 RASPP 与 ASPP 模块进行了性能比较. 可以看到, Baseline+RASPP 的 *mIoU* 较 Baseline+ASPP 在 Cityscapes 和 CamVid 数据集上分别提高了 0.74% 和 0.75%, *mPA* 则分别提高了 0.48% 和 0.56%. 这是因为 RASPP 模块中采用的 *RDSConv*

模块将空间和通道两个维度上的特征结合起来, 增强了空间和通道特征间的语义表达, 缓解了特征不连续问题, 从而提高了分割精度.

3.4.2 CCAM 模块消融实验

为验证本文提出的复合卷积注意力模块 (CCAM) 的有效性, 针对该模块进行了消融实验. 由于 CCAM 是即插即用的, 直接在网络模型中去掉该模块不会影响计算过程. 因此在本文网络模型中去掉该模块, 其余部分保持不变记为 Baseline, 在其基础上添加该模块记为 Baseline+CCAM. 在 Cityscapes 和 CamVid 数据集上进行的消融实验结果如表 4 所示.

表 4 CCAM 模块消融实验 (%)

模型	Cityscapes		CamVid	
	mIoU	*mPA*	*mIoU*	*mPA*
Baseline	80.78	87.55	67.71	75.43
Baseline+CCAM	**81.16**	**87.89**	**68.01**	**75.88**

由表 4 可以看出, Baseline 没有使用 CCAM 模块, 即将高层语义特征上采样后直接与低层细节特征进行融合, 没有考虑到由于两方面特征图所含信息的不同导致的信息干扰, 同时由于上采样操作, 引入了部分噪声, 容易影响特征图内部维度相关性, *mIoU* 较低; 而通过使用 CCAM, 重新校准高层特征的通道和空间维度权重, 减少对低层细节信息的干扰, 同时减少了噪声的影响, 在 Cityscapes 和 CamVid 数据集上 *mIoU* 较 Baseline 分别提高了 0.38% 和 0.30%, *mPA* 较 Baseline 分别提高了 0.34% 和 0.45%, 证明了该模块的有效性.

3.4.3 不同特征融合方式对比实验

为验证本文使用的交叉特征融合方式的有效性, 针对不同的特征融合方式设计了对比实验. 以本文使用的交叉特征融合方式为基准, 如图 6(a) 所示, 即将特征图 F1 和特征图 F3、特征图 F2 和特征图 F4 分阶段进行融合. 此外还设计了两组不同的特征融合方式, 如图 6(b) 和图 6(c) 所示, 将特征图 F1 和特征图 F2、特征图 F3 和特征图 F4 分阶段进行融合, 以及将特征图 F1 和特征图 F4、特征图 F2 和特征图 F3 分阶段进行融合. 针对语义分割中常见的逐级特征融合方式, 本文也进行了对比实验, 如图 6(d) 所示. 图 6 中 U 代表上采样操作, 后面的数字代表上采样的倍数. 在 Cityscapes 和 CamVid 数据集上的实验结果如表 5 所示.

由表 5 的实验结果可以看出, 采用将特征图 F1 和特征图 F4、特征图 F2 和特征图 F3 分阶段进行融合的融合方式②, 相比其他融合方式的效果最差, 这是因为将低层细节信息与高层语义信息直接融合, 引入了大量噪声, 对语义特征产生了干扰; 而将特征图 F1 和特征图 F2、特征图 F3 和特征图 F4 分阶段进行融合的融合方式①, 虽然效果相比融合方式②有略微提升, 但由于没有与低层细节信息充分融合, 分割效果仍然不尽人意; 在语义分割中常用的逐级特征融合方式较二者效果均有所提高, 但不及本文使用的交叉特征融合方式. 相比以上 3 种融合方式来说, 本文使用的交叉特征融合对高低层特征信息进行权重校准, 调整不同层次特征关联, 取得了最好的效果.

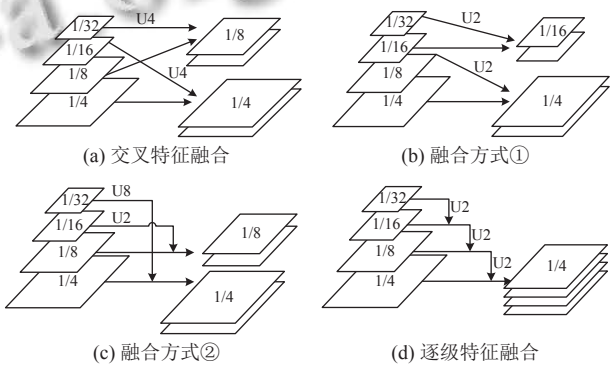


图 6 不同特征融合方式

表 5 特征融合方式对比实验 (%)

模型	Cityscapes		CamVid	
	mIoU	*mPA*	*mIoU*	*mPA*
交叉特征融合	**81.16**	**87.89**	**68.01**	**75.88**
融合方式①	80.63	87.64	66.72	74.09
融合方式②	80.42	87.41	66.14	74.06
逐级特征融合	80.98	87.78	67.23	75.32

3.4.4 与当前流行方法的对比实验

为验证本文方法的有效性, 将本文方法与场景分割领域常用基准方法进行对比实验, 实验结果如表 6 所示.

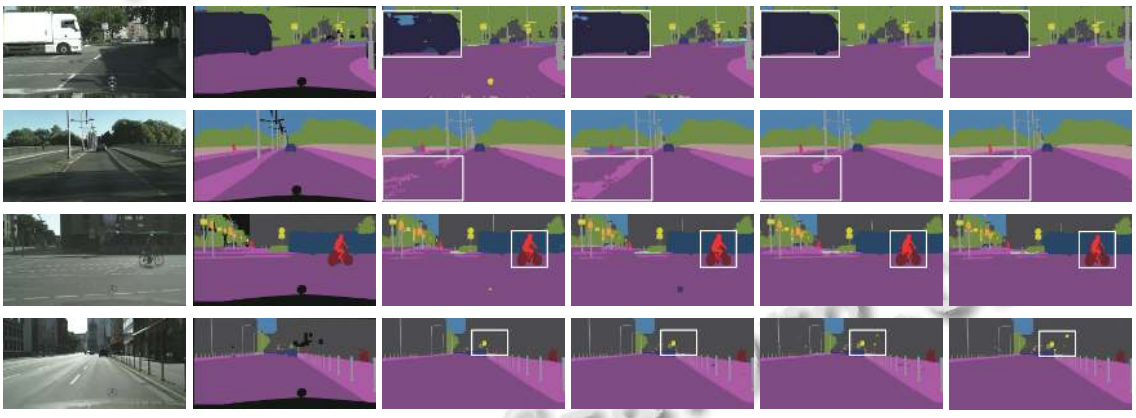
由表 6 可以看出, 本文提出的方法在 CamVid 数据集上优于主流场景分割方法, 并且较 FCN、PSANet、DeepLabV3+、DANet、DNLNet、PointRend、STDC 以及 SegFormer 方法在 *mIoU* 上分别提高了 1.96%, 0.96%, 0.56%, 0.88%, 1.70%, 1.64%, 1.43%, 0.48%, 在 *mPA* 上分别提高了 1.06%, 0.50%, 0.43%, 0.93%, 1.91%, 1.58%, 1.75%, 0.22%, 在 *mFScore* 上分别提高了 1.83%, 1.44%, 0.76%, 0.74%, 1.72%, 1.84%, 1.57%, 0.48%, 表明了本文方法的有效性. 同时还可以

看出, 本文提出方法不仅在 CamVid 数据集上优于主流分割方法, 而且在 Cityscapes 数据集上 *mIoU* 分别超过了 FCN、PSANet、DeepLabV3+、DANet、DNLNet、PointRend、STDC 等方法 7.62%, 3.92%, 1.15%, 1.89%, 1.83%, 4.69%, 4.49%, 与原 SegFormer 网络相比, *mIoU* 提高了 1.20%, 并且在 *mPA* 和 *mFScore* 上均有所提高, 表明本文方法具有较好的泛化性.

为更好地展示网络模型分割效果, 本文选择了在实验中效果比较好的 DeepLabV3+、DANet 和 SegFormer 模型以及本文提出的方法在 Cityscapes 数据集和 CamVid 数据集上的分割结果进行可视化展示. 分割结果对比如图 7 和图 8 所示.

表 6　不同方法性能对比实验

网络模型	主干网络	参数量 (M)	*mIoU* (%)		*mPA* (%)		*mFScore* (%)	
			Cityscapes	CamVid	Cityscapes	CamVid	Cityscapes	CamVid
FCN[8]	ResNet-50	49.50	73.54	66.05	80.65	74.82	83.60	76.95
PSANet[29]	ResNet-50	59.14	77.24	67.05	83.95	75.38	86.33	77.34
DeepLabV3+[15]	ResNet-50	43.59	80.01	67.45	86.94	75.45	88.29	78.02
DANet[23]	ResNet-50	49.85	79.27	67.13	86.98	74.95	87.83	78.04
DNLNet[26]	ResNet-50	50.02	79.33	66.31	86.32	73.97	87.86	77.06
PointRend[16]	ResNet-50	28.72	76.47	66.37	84.05	74.30	85.92	76.94
STDC[30]	STDC2	12.47	76.67	66.58	84.02	74.13	86.09	77.21
SegFormer[20]	MiT-B2	24.99	79.96	67.53	86.86	75.66	88.26	78.30
Ours	MiT-B2	25.97	**81.16**	**68.01**	**87.89**	**75.88**	**89.11**	**78.78**



(a) 原图像　(b) Ground truth　(c) DANet　(d) DeepLabV3+　(e) SegFormer　(f) Ours

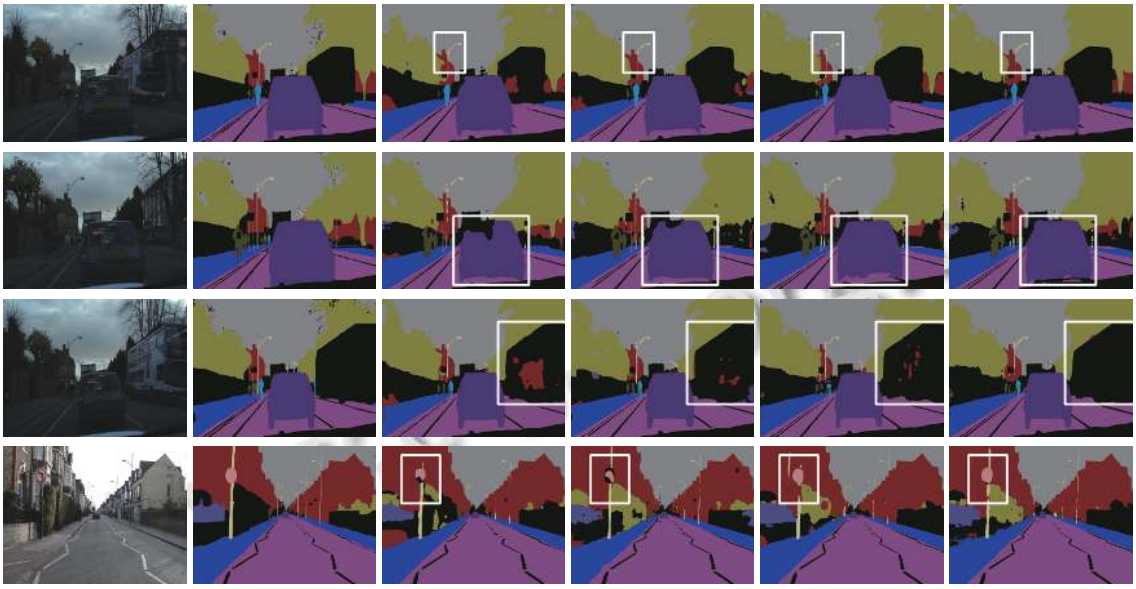
图 7　Cityscapes 上分割结果对比图

如图 7 所示, 在场景 1 (第 1 行) 中的左侧卡车部分, 原始图像中卡车属于尺度较大的目标, 特征之间连续性强, 容易导致误分割现象. DANet 和 DeepLabV3+在卡车分割时均出现了特征不连续导致的空洞问题, 而本文方法分割效果较好. 在场景 2 (第 2 行) 中左侧道路部分, 原始图像道路较长、宽且一直连续, 过长的目标对网络模型是较大的挑战. SegFormer 和 DANet 在左侧道路部分出现大面积空缺, DeepLabV3+分割效果虽然比二者较好, 但仍出现分割零散、空洞问题. 在图 8 的场景 2 (第 2 行) 和场景 3 (第 3 行) 中, 其他方法均出现了较为严重的空洞以及目标边缘参差不齐现象, 而本文方法由于使用交叉特征融合, 并结合复合卷积注意力机制抑制了低层冗余信息的干扰, 缓解了分割零散的问题, 并且使用 RASPP 模块进行不同尺度下的信息提取, 增强了特征间的语义关联, 取得了较好的分割效果.

4　结论与展望

本文针对场景分割中的目标边缘误分割、特征不连续问题, 提出了一种交叉特征融合和 RASPP 驱动的场景分割方法, 改变不同阶段的特征融合方式, 结合复合卷积注意力机制, 将高低层特征进行融合的同时抑

制冗余信息的干扰; 设计并实现了结合残差的多尺度金字塔池化模块 RASPP, 增强特征间的语义表达, 缓解特征不连续问题. 实验结果表明: 本文所提出的方法有效解决了分割空洞、目标边缘不清晰等问题, 提升了分割精度, 分割效果较好. 下一步将向着轻量化结构进行优化, 并提高模型在不同场景下的泛化能力.



(a) 原图像 (b) Ground truth (c) DANet (d) DeepLabV3+ (e) SegFormer (f) Ours

图 8 CamVid 上分割结果对比图

参考文献

1 范润泽, 刘宇红, 张荣芬, 等. 基于多尺度注意力机制的道路场景语义分割模型. 计算机工程, 2023, 49(2): 288–295. [doi: 10.19678/j.issn.1000-3428.0063257]

2 王金祥, 付立军, 尹鹏滨, 等. 基于 CNN 与 Transformer 的医学图像分割. 计算机系统应用, 2023, 32(4): 141–148. [doi: 10.15888/j.cnki.csa.009010]

3 王关茗, 胡乃平. 基于深度学习的自然灾害遥感影像语义分割. 计算机系统应用, 2023, 32(2): 322–328. [doi: 10.15888/j.cnki.csa.008994]

4 王大方, 刘磊, 曹江, 等. 基于空洞空间池化金字塔的自动驾驶图像语义分割方法. 汽车工程, 2022, 44(12): 1818–1824. [doi: 10.19562/j.chinasae.qcgc.2022.12.003]

5 Lei ZP, Zheng W, Miao YX, *et al*. Level set for semantic segmentation with edge compensation. Journal of Physics: Conference Series, 2020, 1449: 012041. [doi: 10.1088/1742-6596/1449/1/012041]

6 桑高丽, 郑增国, 闫超. 基于区域分割的表情鲁棒三维人脸识别方法. 计算机应用研究, 2020, 37(3): 914–918. [doi: 10.19734/j.issn.1001-3695.2018.07.0665]

7 张日升, 原明亭, 丁军航, 等. 基于图像阈值分割的浒苔图像提取. 自动化技术与应用, 2020, 39(2): 83–86.

8 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3431–3440.

9 Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481–2495. [doi: 10.1109/TPAMI.2016.2644615]

10 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241.

11 Zhao HS, Shi JP, Qi XJ, *et al*. Pyramid scene parsing network. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6230–6239.

12 Chen LC, Papandreou G, Kokkinos I, *et al*. Semantic image segmentation with deep convolutional nets and fully connected CRFs. Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2015.

13 Chen LC, Papandreou G, Kokkinos I, *et al*. 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. [doi: 10.1109/TPAMI.2017.2699184]

14 Chen LC, Papandreou G, Schroff F, *et al*. Rethinking atrous convolution for semantic image segmentation. arXiv: 1706.05587, 2017.

15 Chen LC, Zhu YK, Papandreou G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 833–851.

16 Kirillov A, Wu YX, He KM, *et al*. PointRend: Image segmentation as rendering. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9796–9805.

17 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. ICLR, 2021.

18 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 6000–6010.

19 Zheng SX, Lu JC, Zhao HS, *et al*. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 6877–6886.

20 Xie EZ, Wang WH, Yu ZD, *et al*. SegFormer: Simple and efficient design for semantic segmentation with transformers. Proceedings of the 35th International Conference on Neural Information Processing Systems. NeurIPS, 2021. 12077–12090.

21 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.

22 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 3–19.

23 Fu J, Liu J, Tian HJ, *et al*. Dual attention network for scene segmentation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3141–3149.

24 Wang XL, Girshick R, Gupta A, *et al*. Non-local neural networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7794–7803.

25 Misra D, Nalamada T, Arasanipalai AU, *et al*. Rotate to attend: Convolutional triplet attention module. Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2021. 3138–3147.

26 Yin MH, Yao ZL, Cao Y, *et al*. Disentangled non-local neural networks. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 191–207.

27 Chollet F. Xception: Deep learning with depthwise separable convolutions. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1800–1807.

28 Loshchilov I, Hutter F. Decoupled weight decay regularization. Proceedings of the 7th International Conference on Learning Representations. New Orleans: ICLR, 2017.

29 Zhao HS, Zhang Y, Liu S, *et al*. PSANet: Point-wise spatial attention network for scene parsing. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 270–286.

30 Fan MY, Lai SQ, Huang JS, *et al*. Rethinking BiSeNet for real-time semantic segmentation. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 9711–9720.

(校对责编: 牛欣悦)