计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(7):63−73 [doi: 10.15888/j.cnki.csa.009588]

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

基于边缘特征和注意力机制的图像语义分割①

王　军[1,2], 张霁云[1], 程　勇[1,2]

[1](南京信息工程大学 计算机学院, 南京 210044)
[2](南京信息工程大学 科技产业处, 南京 210044)
通信作者: 张霁云, E-mail: 202212200027@nuist.edu.cn

摘　要: 在语义分割任务中, 编码器的下采样过程会导致分辨率降低, 造成图像空间信息细节的丢失, 因此在物体边缘会出现分割不连续或者错误分割的现象, 进而对整体分割性能产生负面影响. 针对上述问题, 提出基于边缘特征和注意力机制的图像语义分割模型 EASSNet. 首先, 使用边缘检测算子计算原始图像的边缘图, 通过池化下采样和卷积运算提取边缘特征. 接着, 将边缘特征融合到经过编码器提取的深层语义特征当中, 恢复经过下采样的特征图像的空间细节信息, 并且通过注意力机制来强化有意义的信息, 从而提高物体边缘分割的准确性, 进而提升语义分割的整体性能. 最后, EASSNet 在 PASCAL VOC 2012 和 Cityscapes 数据集上的平均交并比分别达到 85.9% 和 76.7%, 与当前流行的语义分割网络相比, 整体分割性能和物体边缘的分割效果都具有明显优势.

关键词: 语义分割; 空间细节信息; 边缘特征; 特征融合; 注意力机制

引用格式: 王军,张霁云,程勇.基于边缘特征和注意力机制的图像语义分割.计算机系统应用,2024,33(7):63–73. http://www.c-s-a.org.cn/1003-3254/9588.html

Image Semantic Segmentation Based on Edge Features and Attention Mechanism

WANG Jun[1,2], ZHANG Ji-Yun[1], CHENG Yong[1,2]

[1](School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China)
[2](Science and Technology Industry Division, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: In semantic segmentation tasks, the downsampling process of the encoder can lead to a decrease in resolution, resulting in the loss of spatial information details in the image. As a result, segmentation discontinuity or incorrect segmentation may occur at object edges, which can damage overall segmentation performance. To address the above issues, an image semantic segmentation model EASSNet based on edge features and attention mechanisms is proposed. Firstly, the edge detection operator is used to calculate the edge map of the original image, and edge features are extracted through pooling downsampling and convolution operations. Next, edge features are fused into deep semantic features extracted by the encoder, restoring the spatial detail information of downsampled feature images, and strengthening meaningful information through attention mechanisms to improve the accuracy of object edge segmentation and overall semantic segmentation performance. Finally, EASSNet achieves the average intersection over the union of 85.9% and 76.7% on the PASCAL VOC 2012 and Cityscapes datasets, respectively. Compared with current popular semantic segmentation networks, EASSNet has significant advantages in overall segmentation performance and object edge segmentation.

Key words: semantic segmentation; spatial detail information; edge feature; feature fusion; attention mechanism

① 基金项目: 国家自然科学基金 (41975183)
收稿时间: 2024-02-22; 修改时间: 2024-03-19; 采用时间: 2024-03-28; csa 在线出版时间: 2024-05-31
CNKI 网络首发时间: 2024-06-04

图像语义分割[1,2]是计算机视觉领域的一项重要技术, 它为图像中的每个像素分配语义标签, 使得具有相同标签的像素具有共同的视觉属性. 语义分割通常用于定位图像中的对象和边界, 其目的是简化或改变图像的表示, 以使其更容易理解和分析. 语义分割适用于需要对图像进行精细分割和像素级分类的场景, 如自动驾驶[3]当中的道路分割、医学图像[4]中的病变分割、地理信息[5]中的遥感图像土地分割等. 以自动驾驶为例, 语义分割任务需要识别道路上车可行驶的区域和车道分隔线, 以确保汽车不会进入特定区域. 这不仅要求语义分割模型能够精确识别不同类别的物体, 还要求模型具有足够快的识别速度.

早期的语义分割技术主要分为 3 种: 第 1 种是利用对象和背景之间的灰度值差来从图像中分割前景和背景像素; 第 2 种是根据纹理、对比度、灰度、光谱波段、饱和度等特性的差异, 检测图像中的边缘像素, 从而将图像分割成不同的语义类别; 第 3 种是通过从预定义的种子像素或种子区域进行传播, 将整个图像分割成不同的语义斑块. Shelhamer 等人[6]提出的全卷积网络 (fully convolutional network, FCN) 是首个将深度学习技术应用于语义分割领域的模型, 该模型将神经网络中的全连接层全部替换成卷积层, 使之能够输入任意分辨率的图片, 并输出与输入图片大小一致的结果. 但是输入图像经过一系列卷积操作后, 分辨率不断降低, 容易造成图像空间细节信息的丢失. 计算机视觉领域的一个重要贡献来自生物医学成像领域, 特别是 U-Net[7]网络, 它通过使用编码器-解码器结构连接不同级别的特征来提取输入图像的低级和高级信息. 随后出现的 DeepLab[8–10]家族语义分割网络引入空洞卷积, 通过扩大编码器中卷积层的感受野, 以保持图像的空间分辨率, 并且通过采用全连接条件随机场来优化边界. 其中 DeepLabv3+[11]将空洞卷积与空间金字塔池化 (spatial pyramid pooling, SPP) [12]模块相结合, 形成了空洞空间金字塔池化 (atrous spatial pyramid pooling, ASPP) 模块, 通过不同膨胀率的空洞卷积来扩大感受野, 获取更丰富的上下文信息. 并且用深度可分离卷积替换标准卷积, 减少了模型的参数量, 从而提升分割性能. 但是 ASPP 模块无法获取足够精细的局部信息, 同时该网络也未能解决编码器下采样过程中空间细节信息丢失的问题, 导致在物体边缘容易出现错误分割.

本文在 DeepLabv3+的基础上, 构建了基于边缘特征和注意力机制的图像语义分割模型 EASSNet (edge attentive semantic segmentation network). 本文的主要贡献如下. 1) 设计了边缘特征提取模块, 该模块首先使用 Sobel 边缘检测算子来计算原始图像的边缘图, 随后使用最大池化下采样来降低边缘图的分辨率, 再通过卷积操作来提取边缘特征, 最后使用 Sigmoid 函数, 得到每个像素值都在 0–1 之间的单通道掩码图像. 2) 设计了边缘特征融合模块, 将边缘特征提取模块中获得的边缘特征与编码器提取的深层语义特征进行融合, 恢复经过下采样的特征图像的空间细节信息, 并且通过注意力机制来抑制无关的信息, 强化有意义的信息. 3) 对损失函数进行改进, 通过引入 Dice 损失函数的变体, 来加速语义分割任务的收敛速度, 并且提升分割性能. 4) 本文提出的模型在广泛的实验中表现出了优越的性能. 实验结果表明, EASSNet 在 PASCAL VOC 2012 和 Cityscapes 数据集上的分割精度优于当前流行的语义分割网络.

1 相关工作

1.1 语义分割模型

近年来, 随着深度学习的快速发展, 卷积神经网络技术也日益变得成熟, 并且被应用到语义分割领域. 上文提到的 FCN 是深度学习用于语义分割领域的开山鼻祖, 该模型存在经过卷积操作后特征图像分辨率降低, 导致空间细节信息丢失的问题. DenseU-Net[13]通过加深卷积层, 利用 U-Net 架构实现了小尺度特征的聚合, 从而提高了图像的分类精度. BiSeNet[14]提出了一种具有高输出分辨率的轻量级分支, 并将注意机制引入到不同分支的融合过程中, 在保持网络精度的同时, 大大提高了网络速度. DeconvNet[15]在解码器中使用堆叠的反卷积层来完成上采样操作, 逐步恢复特征图像的分辨率. APCNet[16]包含多尺度、自适应和全局指导局部亲和力这 3 种要素, 能够获取丰富的上下文信息. DFANet[17]通过对编码器下采样后获得的特征图进行上采样, 然后再次输入到编码器当中提取特征, 从而将浅层空间信息和深层语义信息进行融合. PSPNet[18]提出使用金字塔池化模块来聚合上下文信息. HRNet[19]通过多层次特征的迭代信息交换来增强特征融合, 并通过具有多尺度的卷积组合来提高空间信息的精度. 上述所有方法均基于卷积神经网络融合局部特征形成全局特征信息, 从而对图像进行像素级精度的分类.

1.2 边缘特征

图像的边缘特征指的是在图像中特性 (例如像素

灰度、纹理等) 分布出现不连续的地方, 这些地方呈现出阶跃变化或屋脊状的特征. 图像的边缘区域集中了图像大部分的信息, 因此它们通常是决定图像特性的关键部分. 图像边缘广泛存在于物体与背景之间, 以及物体与物体之间. 因此, 边缘特征在图像分割、图像理解以及图像识别中具有重要意义. 利用边缘提取算法可以有效检测出原始图像的边缘, 将边缘图像经过处理并添加到语义分割网络当中后, 可以有效恢复空间细节信息, 改善边缘分割不连续的现象.

图像分类和分割中的边缘优化和增强一直是研究的热点方向. 在一开始, 人们关注分类的后处理来解决这个问题, 比如 Zhou 等人[20]提出的 FC-RCCN 对分类结果进行边缘优化. 随后, 随着深度学习的快速发展, 人们主要关注将边缘优化与深度学习模型相结合, 以生成更准确的分类结果, 即基于边缘感知的分类和语义分割方法. GMENet[21]在语义分割的过程中, 结合了对象级上下文条件反射、部分级空间关系和形状轮廓信息. Chen 等人[22]提出了一种边缘感知卷积核, 利用深度通道中包含的几何信息, 更有效地提取 RGB-D 图像特征映射, 以提高语义分割的精度. Kuang 等人[23]提出了一种新的二维医学图像分割体和边缘感知网络 BEA-SegNet, 该网络将体分割结果与边缘特征融合, 得到最终结果.

边缘特征图像在语义分割领域已经得到了广泛应用. 本文的方法是将边缘图像经过处理后输入到编码器末端, 使经过下采样后分辨率大幅降低的特征图能够学习到有用的空间细节信息, 最终缓解边缘分割不连续的问题.

1.3 注意力机制

注意力机制可以使神经网络更有针对性地捕捉重要信息, 同时排除那些无关的信息. SENet[24]提出了压缩激励 (squeeze excitation, SE) 模块, 这是一种通道注意力机制, 可以将特征图当中更重要的通道凸显出来, 同时尽可能忽略不重要的通道. ECANet[25]为了减少模型参数放弃了 SE 模块中的全连接层, 采用一维卷积进行替代, 同时省略了通道维度减少后再恢复的步骤. Hu 等人[26]提出了 CMPE-SE 机制, 通过残差映射和身份映射之间的竞争来估计特征图的相关性. Hou 等人[27]提出坐标注意力 (coordinate attention, CA) 机制, 将像素坐标信息与通道注意力进行融合, 生成方向感知和位置敏感的注意力图, 增强移动网络在图像分类和下游任务中的表现, 计算量几乎没有增加. Woo 等人[28]提出 CBAM 模块, 通过在通道和空间两个维度上生成注意力特征权重, 然后将这两种特征权重与原始输入特征图相乘, 以实现特征图的自适应修正.

1.4 现有方法的不足

尽管国内外的研究者已经提出了许多方法来缓解在编码器的下采样过程中空间细节信息丢失的问题, 但是现有的语义分割模型仍然有需要改进的地方. 许多基于边缘特征的语义分割模型需要使用深度信息, 但是能够获取深度信息的立体相机价格较高, 图像中包含深度值的语义分割数据集也比较少见. 此外一些语义分割模型将边缘图像与原始图像进行合并以后再输入到编码器当中, 而空间细节信息仍然会在下采样的时候出现丢失.

因此, 有必要针对编码器对原始图像下采样的过程中容易丢失空间细节信息的问题进行研究, 改善在物体边缘出现分割不连续或者错误分割的现象, 提高整体分割性能.

2 模型改进

2.1 模型总体结构

本文采用 DeepLabv3+作为基线语义分割模型, 其结构如图 1 所示. 在编码器部分, 输入图像首先被输入到深度卷积神经网络 (deep convolutional neural network, DCNN) 当中, 分别提取出经过 4 倍下采样的浅层特征图和经过 16 倍下采样的深层特征图. 接下来, 将深层特征图送入 ASPP 模块, 使用具有不同膨胀率的空洞卷积以获取不同感受野的特征图. 通过 concat 操作将这些特征图沿通道维度连接在一起, 并通过 1×1 卷积进行通道数的调整. 在解码器部分, 首先对浅层特征图进行 1×1 卷积, 进行通道数的调整, 然后对经过 ASPP 模块处理的深层次特征图进行 4 倍上采样操作. 接着, 将上述两个特征图在通道维度上连接, 以融合浅层空间信息和深层语义信息. 最后, 通过 3×3 卷积和 4 倍上采样将特征图的分辨率还原至原始图像的大小, 从而得到语义分割的预测结果.

DeepLabv3+虽然通过使用 ASPP 模块来扩大感受野, 提高了网络对全局特征的感知能力, 但是无法获取足够精细的局部特征, 也无法恢复编码器下采样过程中丢失的空间细节信息, 导致物体的边缘部分容易出现错误分割的问题. 本文针对上述问题, 对 DeepLabv3+网络进行改进, 提出了语义分割模型 EASSNet, 其总体结构如图 2 所示. 首先使用 Sobel 边缘检测算子来计算原始图像的边缘图, 再将边缘图输入到边缘特征提取

(edge feature extraction, EFE) 模块中, 提取出边缘特征, 最后使用边缘特征融合 (edge feature fusion, EFF) 模块, 结合注意力机制将 ASPP 输出的主干特征与边缘特征进行融合. 融合后的特征不仅保留了编码器提取的深层语义信息, 还包含了边缘特征所贡献的空间细节内容. 通过注意力机制, 有意义的信息得到强化突显, 无意义的冗余信息则被不被网络关注, 模型整体的学习能力和泛化能力得到提高.

EASSNet 语义分割模型通过将边缘特征融合到主干特征当中, 缓解了物体边缘错误分割的问题, 并且通过改进损失函数, 使得模型在训练时更容易得到接近真实分布的参数, 最终提升了语义分割任务的整体性能.

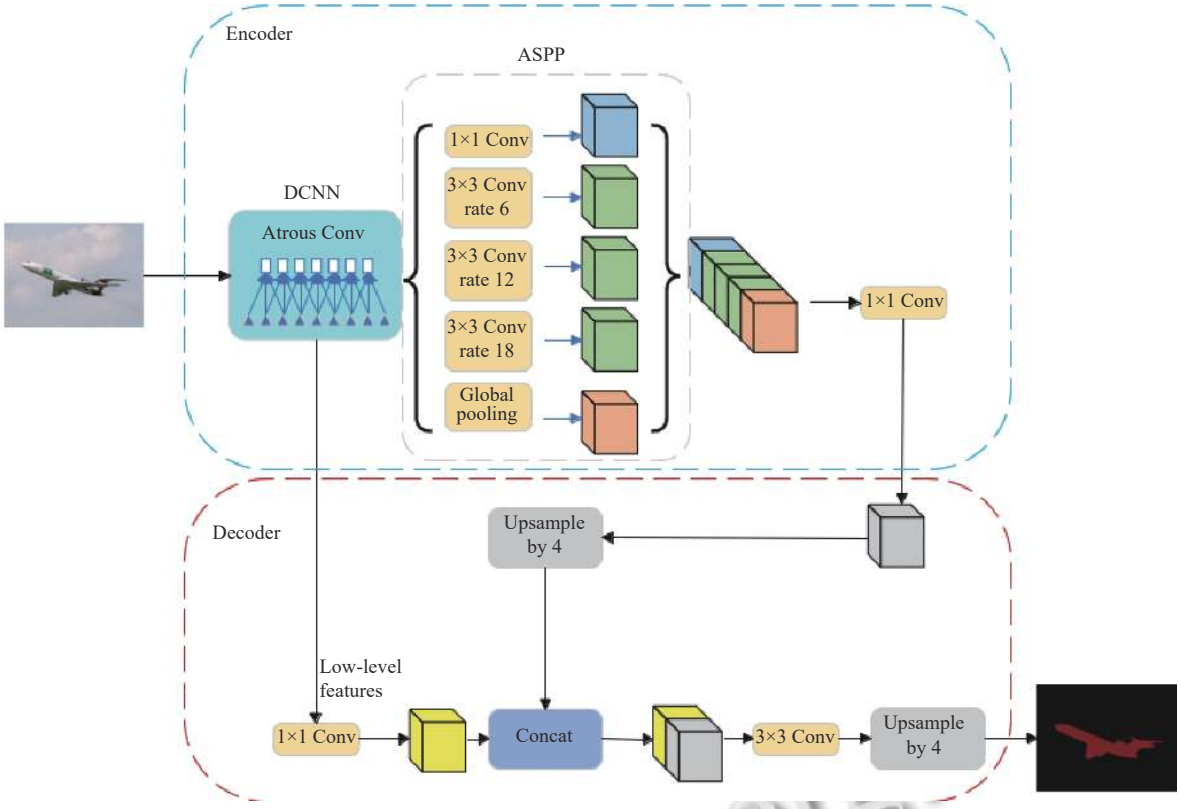

图 1 DeepLabv3+模型的总体结构

2.2 边缘特征提取模块

经过编码器下采样后的特征图像空间细节信息损失严重, 需要从原始输入图像中提取信息并且补充到特征图像中. 针对该问题, 本文设计了 EFE 模块, 用以提取输入图像中的边缘特征, 其结构如图 3 所示. 首先使用传统的边缘检测方法 Sobel 算子, 计算出原始图像的边缘图像, 然后对该图像进行 16 倍最大池化下采样. 下采样后的边缘图像虽然分辨率降低, 但是基本保存了原始图像的空间细节信息. 随后使用连续 4 个卷积块来提取边缘图像中的深层语义信息, 每个卷积块包含一个 7×7 卷积、BatchNorm 和 GELU 激活函数. 接着使用 1×1 卷积恢复通道数, 最后使用 Sigmoid 激活函数计算出最终的边缘特征.

EFE 模块提取的边缘特征是一个单通道掩码图像, 每个像素值都在 0–1 之间. 该特征图包含了边缘图像中的空间细节信息和深层语义信息, 可以为后续的语义分割任务提供帮助.

2.3 边缘特征融合模块

通过 EFE 模块提取的边缘特征保留了很多浅层空间细节信息, 而编码器下采样后的主干特征图像具有丰富的高层语义信息, 需要将这两个特征进行融合, 使边缘特征成为主干特征的有益补充, 并且对其中有意义的特征信息进行强化. 针对该问题, 本文设计了 EFF 模块, 用以将边缘特征融合进编码器下采样后的特征图像当中, 其结构如图 4 所示. 首先将编码器输出的主干特征与边缘特征进行逐像素相乘, 然后与原始的主干特征进行残差连接, 得到新的主干特征. 该步骤可以用式 (1) 来表示:

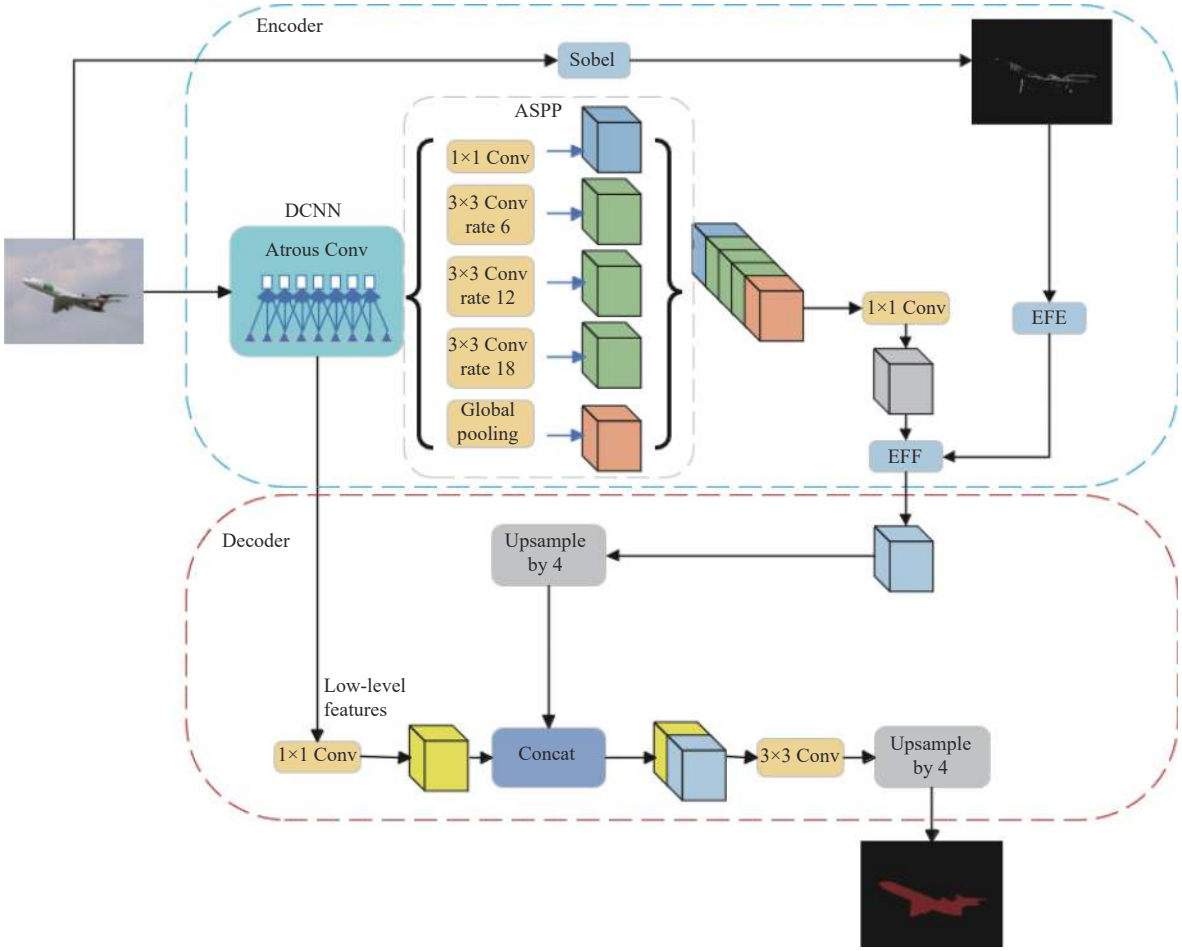

图 2 EASSNet 模型结构

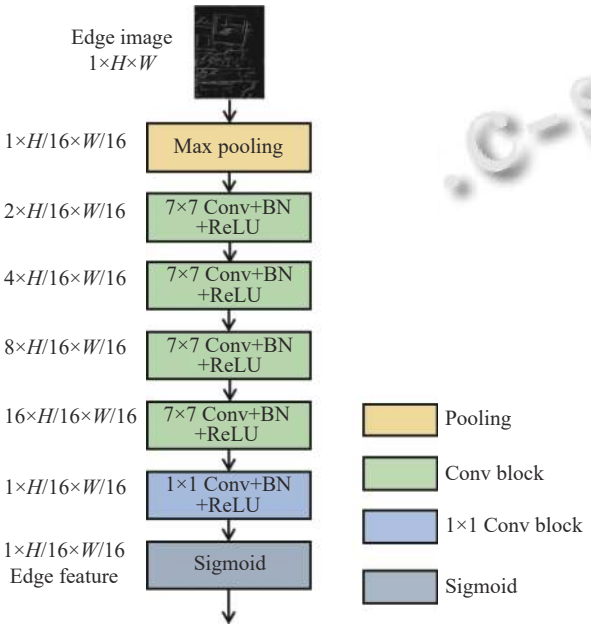

图 3 边缘特征提取模块

$$F = F_E \times F_B + F_B \tag{1}$$

其中, F_B为主干特征, F_E为边缘特征. 进行残差连接是为了防止边缘特征喧宾夺主, 对编码器从原始图像中提取的主干特征造成太大的影响, 导致语义分割的性能降低. 式 (1) 得到的新特征图当中不仅包含了原始图像的深层语义信息, 而且包含了边缘特征提供的空间细节信息.

接着, 使用注意力机制对特征进行优化, 使网络更多地关注有意义的特征, 抑制无意义的特征. 注意力机制由串行连接的一个通道注意力模块和一个空间注意力模块组成. 在通道注意力模块中, 首先对特征图分别进行全局平均池化 (global average pooling, GAP) 和全局最大池化 (global max pooling, GMP) 操作, 得到两个 C×1×1 的张量, 然后将它们同时输入由两个 1×1 卷积块组成的人工神经网络当中. 将人工神经网络输出的两个张量相加后, 使用 Sigmoid 函数计算出通道注意力权重 M_C. 最后通过矩阵点乘运算, 将通道注意权

重映射到输入的特征图当中, 得到新的特征图, 实现通道维度的自适应修正. 该步骤可以用式 (2) 和式 (3) 来表示:

$$M_C(F)=\sigma(W_1(W_0(F_{c,\text{avg}}))+W_1(W_0(F_{c,\max}))) \quad (2)$$

$$F=F\times M_C(F) \quad (3)$$

其中, $F_{c,\text{avg}}$ 表示全局平均池化后的张量, $F_{c,\max}$ 表示全局最大池化后的张量, W_0、W_1 分别表示两个 1×1 卷积块, σ 为 Sigmoid 激活函数.

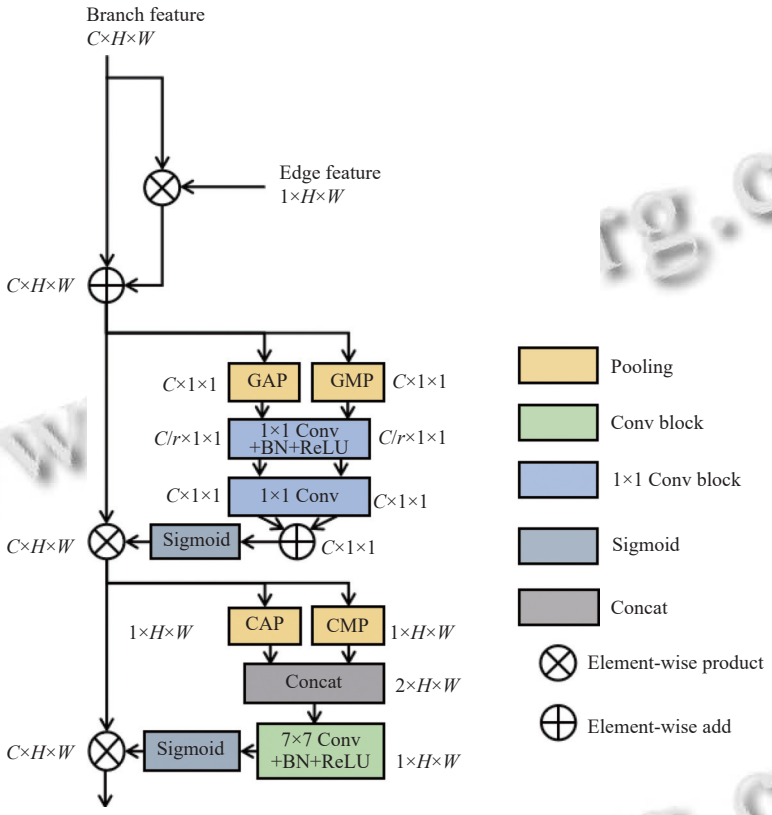

图 4 边缘特征融合模块

随后, 在空间注意力模块当中, 首先对通道注意力模块输出的特征图在通道维度进行最大池化和平均池化操作, 得到两个 1×H×W 的张量, 将这两个张量沿着通道维度连接在一起, 再通过 7×7 卷积块提取特征, 并且将通道数恢复为 1, 随后使用 Sigmoid 函数计算出空间注意力权重 M_S. 最后通过矩阵点乘运算将空间注意力权重映射到输入的特征图当中, 得到新的特征图, 实现空间维度的自适应修正. 该步骤可以用式 (4) 和式 (5) 来表示:

$$M_S(F)=\sigma(f_{7\times7}(F_{s,\text{avg}};F_{s,\max})) \quad (4)$$

$$F=F\times M_S(F) \quad (5)$$

其中, $F_{s,\text{avg}}$ 表示沿着通道维度进行全局平均池化后的张量, $F_{s,\max}$ 表示沿着通道维度进行全局最大池化后的张量, “;”符号表示沿着通道维度的连接操作, σ 为 Sigmoid 激活函数.

经过 EFF 模块融合后的特征图能够有效恢复在编码器的下采样过程中丢失的空间细节信息, 并且通过注意力机制重点突出了有意义的特征信息. 将特征图输入到解码器进行上采样, 并进行像素级精度的分类后, 得到的语义分割预测图可以有效改善在物体的边缘分割不连续的现象.

2.4 改进损失函数

在语义分割领域中, 损失函数有多种形式, 其中最常用的是交叉熵损失函数. 该函数通过衡量预测分布与真实分布的相似性来计算损失, 预测分布越接近真实分布, 损失函数的值越小, 反之越大, 其表达式如式 (6) 所示:

$$Loss_{\text{CE}}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}y_{ij}\log(p_{ij}) \quad (6)$$

其中, N 是样本数量, C 是类别数量, y_{ij} 是样本 i 属于类别 j 的标签, 其值为 0 或 1, p_{ij} 是模型对样本 i 预测为类别 j 的概率, 其值在 0–1 之间. 交叉熵损失函数的局限性在于没有考虑到标签分布不平衡的情况, 当不同类别的像素数量差异很大时, 损失函数的训练会变得较为困难. 此外, 交叉熵损失函数只是离散地计算每个像素的损失值然后取平均, 而不是全局考虑整幅图像的预测结果. 为了弥补交叉熵损失函数的不足, 引入 Dice 损失函数及其变体 Tanimoto 损失函数, 其表达式分别如式 (7) 和式 (8) 所示:

$$Loss_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N}\sum_{j=1}^{C}(y_{ij}p_{ij})}{\sum_{i=1}^{N}\sum_{j=1}^{C}(y_{ij}+p_{ij})} \tag{7}$$

$$Loss_{\text{Tanimoto}} = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1}^{C}(y_{ij}p_{ij})}{\sum_{i=1}^{N}\sum_{j=1}^{C}(y_{ij}+p_{ij}-y_{ij}p_{ij})} \tag{8}$$

其中, N 是样本数量, C 是类别数量, y_{ij} 是样本 i 属于类别 j 的标签, 其值为 0 或 1, p_{ij} 是模型对样本 i 预测为类别 j 的概率, 其值在 0–1 之间. Dice 损失函数和 Tanimoto 损失函数在数值上是等价的, 并且都有助于解决标签分布不平衡时训练困难的问题, 但是当小目标较多时, 该损失函数容易出现震荡, 极端情况下甚至会出现梯度饱和的情况. 此外, 根据经验, 无论权重的随即初始值如何, 分母上有二次项的损失函数都更容易让预测结果接近于真实值. 因此, 选择交叉熵损失函数和 Tanimoto 损失函数的加权和为总体损失函数, 其表达式如式 (9) 所示:

$$Loss_{\text{total}} = Loss_{\text{CE}} + \partial Loss_{\text{Tanimoto}} \tag{9}$$

其中, α 为对交叉熵损失函数和 Tanimoto 损失函数的影响进行平衡的参数, 其取值范围为 $(0, +\infty)$.

3 实验结果与分析

3.1 数据集

在 PASCAL VOC 2012 和 Cityscapes 这两个数据集上评估了所提出的语义分割模型的性能. 其中 PASCAL VOC 2012 数据集用于模型的训练和性能评估, Cityscapes 数据集用于模型的泛化性能测试.

PASCAL VOC 2012 是计算机视觉领域中被广泛使用的公共图像数据集. 该数据集有 21 个语义类别, 包括 20 个物体种类和 1 个背景类. 总共有 2 913 张有标签的图像, 其中随机挑选的 2 622 张照片作为训练集, 291 张照片作为验证集. 输入语义分割网络的图像大小设定为 512×512.

Cityscapes 是城市环境中自动驾驶场景的著名数据集之一. 该数据集有 34 个语义类别, 根据前人的工作, 只使用其中 19 个类别. 总共有 5 000 张精细标注的图像, 每张照片的分辨率都是 1024×2048, 为了与 VOC 2012 数据集保持一致, 将数据集中随机挑选的 4 500 张照片作为训练集, 500 张照片作为验证集. 输入语义分割网络的图像大小设定为 512×1024.

3.2 评价指标

采用平均交并比 (mean intersection over union, *MIoU*) 和平均精度 (mean accuracy, *MAcc*) 来评价实验结果, 其表达式如式 (10) 和式 (11) 所示:

$$MIoU = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i + FN_i} \tag{10}$$

$$MAcc = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i} \tag{11}$$

其中, N 表示类别总数, TP_i 类别 i 正确预测的像素数量, FP_i 是将其他类别预测为类别 i 的像素数量, FN_i 是将类别 i 预测为其他类别的像素数量. *MIoU* 是对每个语义类别分别计算出交并比, 求和后计算平均值, *MAcc* 是对每个语义类别分别计算出精度, 求和后计算平均值.

3.3 实验设置

本实验在 PyTorch 框架上实现, 所用操作系统为 Windows 11 64 位操作系统, 处理器为 Intel(R) Xeon(R) Gold 5218R, 显卡为 NVIDIA A10, 内存为 128 GB, 硬盘为 2 TB.

采用适应性矩估计 (adaptive moment estimation, Adam) 算法来作为优化器, 动量设置为 0.9. 应用“poly”学习速率策略, 学习率随着迭代次数的增加而逐渐减少, 其表达式如式 (12) 所示:

$$r = r_{\text{base}} \times \left(1 - \frac{T}{T_{\max}}\right)^{power} \tag{12}$$

其中, r 表示当前学习率, r_{base} 表示初始学习率, 设置为

5×10^{-4}, T 表示当前迭代次数, T_{max} 表示最大迭代次数, *power* 表示动量, 设置为 0.9. 此外, 将每轮训练的批量大小设置为 16, 训练轮数设置为 200, 其中在前 100 轮训练当中冻结主干网络的参数, 使其不参与训练, 在第 101–200 轮训练中对主干网络进行解冻.

在数据增强方面, 采用[0.5, 2]的随机缩放、[−10°, 11°]的随机旋转、随机反转和随机高斯模糊的措施, 来增强网络的鲁棒性.

3.4 消融实验

首先在 DeepLabv3+基线模型的基础上, 分别选用 MobileNetv2、ResNet101 和 Xception 这 3 种深度卷积神经网络作为主干网络进行实验, 结果如表 1 所示. 可以看出, Xception 的 *MIoU* 和 *MAcc* 最高, 分割效果最好, 因此选择 Xception 作为模型的主干网络.

表 1 不同主干网络实验结果 (%)

Backbone	*MIoU*	*MAcc*
MobileNetv2	77.23	86.12
ResNet101	80.26	89.03
Xception	82.63	90.39

接下来, 对模型采用的不同模块进行消融实验, 结果如表 2 所示. 将没有注意力机制, 只有将主干特征和边缘特征进行逐像素相乘, 再与原始主干特征进行残差连接的边缘特征融合模块称为 EFF_1, 将采用压缩激励模块 (squeeze and excitation, SE) 注意力机制的边缘特征融合模块称为 EFF_2, 将采用卷积块注意力模块 (convolutional block attention module, CBAM) 注意力机制的边缘特征融合模块称为 EFF_3. 由于 EFE 模块提取边缘特征后必须要有特征融合的步骤, 故不对该模块进行单独的消融实验. 可以看出在基线模型的基础上添加 EFE 模块和 EFF_3 模块的情况下, 模型的 *MIoU* 和 *MAcc* 均为最高, 故采用该方法作为最终的网络结构.

表 2 消融实验结果 (%)

方法	*MIoU*	*MAcc*
Baseline	82.63	90.39
Baseline+SE	84.38	90.75
Baseline+CBAM	84.62	90.83
Baseline+EFE+EFF_1	84.47	90.62
Baseline+EFE+EFF_2	85.67	91.34
Baseline+EFE+EFF_3 (EASSNet)	85.44	91.31

在上述消融实验中, 模型使用的损失函数均为交叉熵损失函数. 最后, 在 EASSNet 模型的基础上, 采用不同的损失函数进行训练, 结果如表 3 所示. 可以看出, 当总体损失函数 $Loss_{total}=Loss_{CE}+\frac{1}{3}Loss_{Tanimoto}$ 时, 训练效果最好, 说明采用该损失函数更容易使模型参数收敛至最优值.

表 3 不同损失函数实验结果 (%)

$Loss_{total}$	*MIoU*	*MAcc*
$Loss_{CE}$	85.44	91.31
$Loss_{CE}+Loss_{Dice}$	85.62	91.38
$Loss_{CE}+Loss_{Tanimoto}$	85.65	91.43
$Loss_{CE}+\frac{1}{2}Loss_{Tanimoto}$	85.79	91.53
$Loss_{CE}+\frac{1}{3}Loss_{Tanimoto}$	85.91	91.62
$Loss_{CE}+\frac{1}{4}Loss_{Tanimoto}$	85.83	91.46

3.5 在 PASCAL VOC 2012 数据集上的对比实验结果

在 PASCAL VOC 2012 数据集上将 EASSNet 与当前流行的语义分割模型进行对比, 结果如表 4 所示. 可以看出, 本文提出的 EASSNet 在 *MIoU* 和 *MAcc* 这两个指标上均取得了最好的结果. 与性能次优的模型 DMNet 相比, EASSNet 的 *MIoU* 提升了 1.32 个百分点, *MAcc* 提升了 0.87 个百分点; 相比于 HRNet 和 PSPNet, *MIoU* 分别提升了 2.11 和 3.72 个百分点, *MAcc* 分别提升了 2.50 和 3.05 个百分点. 可见, EASSNet 在语义分割的性能上普遍优于当前流行的语义分割模型.

表 4 在 PASCAL VOC 2012 数据集上与其他方法的对比结果 (%)

方法	*MIoU*	*MAcc*
SegNet[29]	60.82	73.42
FCN	62.39	75.20
DeepLab	67.03	76.61
U-Net	72.94	80.58
DeconvNet	74.35	84.66
BiSeNet	79.86	87.54
APCNet	80.56	87.11
PSPNet	82.19	88.57
HRNet	83.80	89.12
DMNet[30]	84.55	90.75
EASSNet	85.91	91.62

EASSNet 与 U-Net、PSPNet 和 HRNet 的可视化结果如图 5 所示, 其中第 1 列为输入图像, 第 2 列为标签图像, 第 3–6 列分别为 U-Net、PSPNet、HRNet 和 EASSNet 的分割结果, 在图 5 中用红色方框标出了其他模型的分割错误之处, 以及本文提出的模型的改进之处. 从图 5 第 1 行中可以看出, U-Net、PSPNet 在对

绵羊的腿部进行分割时均出现了错误, HRNet 在绵羊的头部出现了典型的分割不连续现象, 而 EASSSNet 均无这些错误. 在第 2 行中, U-Net 将图片左侧的背景部分误分割为沙发类, PSPNet、HRNet 在对猫的尾部进行分割时均出现了错误, 而 EASSNet 较为准确地完成了猫尾的分割. 在第 3 行中, U-Net 对狗的身体进行分割时出现了空洞, PSPNet、HRNet 在分割狗的嘴部以及腿和身体连接的部位时出现错误, 而 EASSNet 在这些部位均取得了良好的分割效果. 在第 4 行中, U-Net、PSPNet、HRNet 在对飞机的尾翼进行分割时都有明显错误, 其中 HRNet 再次出现了分割不连续的现象, 只有 EASSNet 在飞机尾翼部分未出现分割错误的现象. 这是因为本文提出的 EASSNet 直接从原始输入图像中提取边缘特征, 与编码器下采样之后输出的特征图像进行融合, 使特征图像同时具有丰富的浅层空间信息和深层语义信息, 并且通过注意力机制强化有意义的信息, 而其他 3 种语义分割模型都没有将空间细节信息补充到经过下采样后的特征图像当中. 因此, EASSNet 模型在物体边缘的分割当中具有显著优势, 整体分割性能也更加出色.

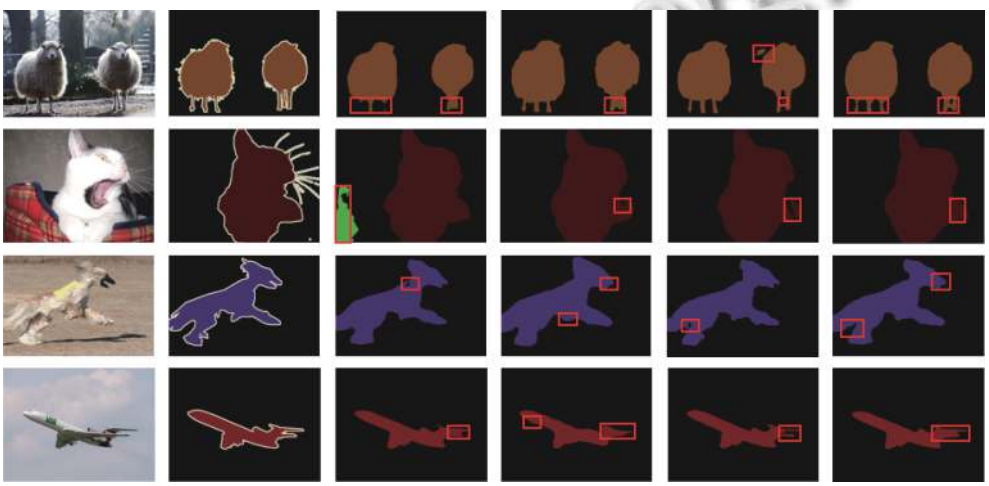

图 5 在 PASCAL VOC 2012 数据集上的可视化结果对比

总而言之, 通过构建边缘特征提取模块和边缘特征融合模块, 本文提出的 EASSNet 语义分割模型可以在一定程度上恢复经过编码器下采样之后的特征图像的空间细节信息, 增强物体边缘分割的准确性, 并且更加关注有意义的信息. 此外, 通过改进损失函数, 本文提出的模型参数更容易收敛至最优值, 最终使得语义分割的整体效果取得一定程度的提高. 实验结果表明, 基于边缘特征和注意力机制的 EASSNet 模型在语义分割性能上取得了显著的进展, 尤其在物体边缘部分的分割方面表现出明显的优势.

3.6 在 Cityscapes 数据集上的对比实验结果

为了验证 EASSNet 模型的泛化能力, 采用 Cityscapes 数据集进行泛化实验, 结果如表 5 所示. EASSNet 在 *MIoU* 和 *MAcc* 这两个指标上均取得了最好的结果, 比基线模型 DeepLabv3+分别高了 2.16 和 2.31 个百分点.

EASSNet 与 U-Net、DeepLabv3+的可视化结果如图 6 所示, 其中第 1 列为输入图像, 第 2 列为标签图像, 第 3–5 列分别是 U-Net、DeepLabv3+和 EASSNet 的分割结果. 可以看出, EASSNet 可以较为准确地分割出物体的边缘部分, 分割结果较为完整清晰, 总体性能更优.

表 5 在 Cityscapes 数据集上与其他方法的对比结果 (%)

方法	*MIoU*	*MAcc*
FCN	62.50	70.35
DeepLab	63.07	72.17
U-Net	63.58	71.30
BiSeNet	69.27	76.81
DeepLabv3+	73.76	79.48
PSPNet	74.61	80.66
EASSNet	75.92	81.79

4 结论

本文在 DeepLabv3+的基础上进行改进, 提出了基于边缘特征和注意力机制的 EASSNet 模型. 首先, 设

计了 EFE 模块, 对原始图像的边缘图进行下采样和卷积操作, 以捕获关键的边缘特征. 接下来, 设计了 EFF 模块, 将 EFE 模块获取的边缘特征融合到编码器提取的主干特征中, 并通过注意力机制对融合后的特征进行优化, 使网络更加聚焦于有意义的特征. 最后, 对损失函数进行改进, 使得模型参数更容易收敛至最优值. 通过这些改进步骤, EASSNet 能够有效地恢复下采样图像的空间信息细节, 从而增强分割图像的边缘连续性, 改善物体边缘错误分割的问题, 最终提升整体的语义分割性能. 大量的实验结果表明所提出的方法在两个普遍使用的语义分割数据集上具有更优的性能, 证明了改进的有效性. 在未来的工作中, 将聚焦于对模型进行轻量化改进, 减少模型的参数, 提高语义分割的速度, 使其更适用于自动驾驶等需要实时性的领域.

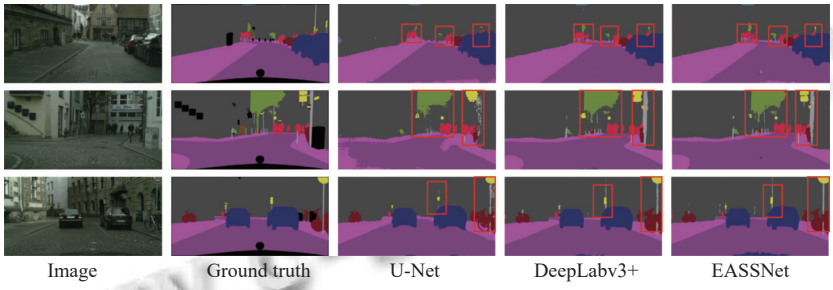

图 6 在 Cityscapes 数据集上的可视化结果对比

参考文献

1 王龙飞, 严春满. 道路场景语义分割综述. 激光与光电子学进展, 2021, 58(12): 1200002.

2 Jiang HJ, Wang RP, Shan SG, *et al*. Adaptive metric learning for zero-shot recognition. IEEE Signal Processing Letters, 2019, 26(9): 1270–1274. [doi: 10.1109/LSP.2019.2917148]

3 Kong YY, Zhang BW, Yan BY, *et al*. Affiliated fusion conditional random field for urban UAV image semantic segmentation. Sensors, 2020, 20(4): 993. [doi: 10.3390/s20040993]

4 Jiang F, Grigorev A, Rho S, *et al*. Medical image semantic segmentation based on deep learning. Neural Computing and Applications, 2018, 29(5): 1257–1265. [doi: 10.1007/s00521-017-3158-6]

5 Xiao AR, Yang XF, Lu SJ, *et al*. FPS-Net: A convolutional fusion network for large-scale LiDAR point cloud segmentation. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 176: 237–249. [doi: 10.1016/j.isprsjprs.2021.04.011]

6 Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640–651. [doi: 10.1109/TPAMI.2016.2572683]

7 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.

8 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, 2015.

9 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]

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

11 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. Munich: Springer, 2018. 833–851.

12 He KM, Zhang XY, Ren SQ, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904–1916. [doi: 10.1109/TPAMI.2015.2389824]

13 Dong RS, Pan XQ, Li FY. DenseU-Net-based semantic segmentation of small objects in urban remote sensing images. IEEE Access, 2019, 7: 65347–65356. [doi: 10.1109/ACCESS.2019.2917952]

14 Yu CQ, Wang JB, Peng C, *et al*. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 334–349.

15 Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015. 1520–1528.
16 He JJ, Deng ZY, Zhou L, *et al*. Adaptive pyramid context network for semantic segmentation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 7511–7520.
17 Li HC, Xiong PF, Fan HQ, *et al*. DFANet: Deep feature aggregation for real-time semantic segmentation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 9514–9523.
18 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.
19 Wang JD, Sun K, Cheng TH, *et al*. Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3349–3364. [doi: 10.1109/TPAMI.2020.2983686]
20 Zhou L, Kong XY, Gong C, *et al*. FC-RCCN: Fully convolutional residual continuous CRF network for semantic segmentation. Pattern Recognition Letters, 2020, 130: 54–63. [doi: 10.1016/j.patrec.2018.08.030]
21 Michieli U, Zanuttigh P. Edge-aware graph matching network for part-based semantic segmentation. International Journal of Computer Vision, 2022, 130(11): 2797–2821. [doi: 10.1007/s11263-022-01671-z]
22 Chen RS, Zhang FL, Rhee T. Edge-aware convolution for RGB-D image segmentation. Proceedings of the 35th International Conference on Image and Vision Computing New Zealand. Wellington: IEEE, 2020. 1–6.
23 Kuang HL, Liang YX, Liu N, *et al*. BEA-SegNet: Body and edge aware network for medical image segmentation. Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine. Houston: IEEE, 2021. 939–944.
24 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.
25 Wang QL, Wu BG, Zhu PF, *et al*. ECA-Net: Efficient channel attention for deep convolutional neural networks. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 11531–11539.
26 Hu Y, Wen GH, Luo MN, *et al*. Competitive inner-imaging squeeze and excitation for residual network. arXiv:1807.08920, 2018.
27 Hou QB, Zhou DQ, Feng JS. Coordinate attention for efficient mobile network design. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 13708–13717.
28 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.
29 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]
30 He JJ, Deng ZY, Qiao Y. Dynamic multi-scale filters for semantic segmentation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 3561–3571.

(校对责编: 孙君艳)