第 59 卷 第 16 期/2022 年 8 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于 Deeplab v3+的图像语义分割算法优化研究

孟俊熙，张莉*，曹洋，张乐天，宋倩

西安工程大学电子信息学院，陕西 西安 710600

摘要 针对目前 Deeplab v3+模型进行图像语义分割时部分细节损失严重，存在漏分割、误分割现象，在其算法基础上构建了新的语义分割模型 N-Deeplab v3+。新模型设计异感受野拼接的空洞空间金字塔池化结构，增强各层级信息间相关性；增设多次跨层特征融合，提升对图像细节的表征力；构建基于注意力机制的特征对齐模块，引导高低级特征对齐并有针对性地强化对重要通道特征的学习，提升模型学习能力。在 Cityscapes 数据集上的实验结果表明，所提改进方案能够有效提高小尺度目标关注度，缓解目标误分割问题，提升模型语义分割精度。在 PASCAL VOC 2012 数据集上进一步验证新模型的泛化能力。N-Deeplab v3+模型在 Cityscapes 数据集和 PASCAL VOC 2012 数据集上的平均交并比达 76.31% 和 81.97%，较原模型分别提升了 1.69 个百分点和 2.14 个百分点。

关键词 深度学习；图像语义分割；Deeplab v3+；注意力机制

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP202259.1610009

Optimization of Image Semantic Segmentation Algorithms Based on Deeplab v3+

Meng Junxi, Zhang Li*, Cao Yang, Zhang Letian, Song Qian

College of Electronics and Information, Xi'an Polytechnic University, Xi'an 710600, *Shaanxi, China*

Abstract Herein, a new semantic segmentation model N-Deeplab v3+ was proposed based on the existing Deeplab v3+ algorithm. The proposed model can be used to address some severe problems of Deeplab v3+ related to the loss of details, such as missing and incorrect segmentations, during image semantic segmentation. The new model designed an atrous spatial pyramid pooling structure with heteroreceptive field splicing to enhance the correlation between different-level data. The feature fusion of multiple crosslayers is performed to improve the characterization of image details. A feature alignment module based on the attention mechanism was developed to guide the alignment of high- and low-level features and enhance the learning process for important channel features in a targeted manner, thus improving the learning ability of the model. Experimental results based on the Cityscapes dataset show that the proposed model can effectively increase the attention for small-scale targets, alleviate the problem of target mis-segmentation, and show improved semantic segmentation accuracy. The generalization capability of the proposed model is further verified on the PASCAL VOC 2012 dataset. The mean intersection over union of N-Deeplab v3+ on the Cityscapes dataset and PASCAL VOC 2012 dataset reaches 76.31% and 81.97%, respectively, showing improvements of 1.69 percentage points and 2.14 percentage points, respectively, compared with the original model.

Key words deep learning; image semantic segmentation; Deeplab v3+; attentional mechanism

1 引 言

图像语义分割通过对像素点进行预测分类，实现图像内容准确定位并完整地呈现由同属性像素组成区域的语义特征，方便计算机视觉系统对图像内容进行准确理解。图像语义分割作为图像解析和场景理解的基础性技术[1]，在智能驾驶、智慧安防以及增强现实等领域具有较高的实用价值和发展前景[2-4]。由于图像语义分割的复杂性，现有的语义分割技术仍面临漏分割、误分割等问题，因此如何增强图像细节信息的表征能力、提升多尺度特征承载信息的利用率是提高语义分割精度的重点研究方向。全卷积神经网络(FCN)[5]使用卷积层取代全连接层，利用反卷积形成一种端到端的网络，将语义分割精度推向了新高，推动了语义分

收稿日期：2021-06-07；**修回日期**：2021-06-26；**录用日期**：2021-07-09

基金项目：陕西省教育厅研究项目(10JK510)、西安市科技局产业化项目(CXY1517(4))

通信作者：*dx_zhangli@126.com

割算法的快速发展。此后，图像语义分割算法大部分是基于FCN演变而来的，其中不乏能够有效提升语义分割性能的结构。

基于优化卷积结构的空洞卷积在不损失分辨率的前提下能扩大卷积核的感受野[6]。Deeplab v1模型[7]在骨干网络中引入空洞卷积，缓解一系列卷积操作导致有效信息丢失的问题。可变形卷积[8]通过对卷积核各个参数附加方向向量，自适应地调整尺度和感受野，增强模型对尺度变换的适应力。基于编解码结构的SegNet[9]在编码阶段保存了池化索引，准确恢复图像尺寸与空间信息，有效地保留高频细节完整性。U-Net[10]通过跳跃连接结构引入编码层内不同尺度特征来恢复丢失的信息，实现像素的精准定位。基于多尺度特征聚合结构的PSPNet[11]利用金字塔池化模块捕获不同区域的特征信息，充分利用图像全局和局部信息来缓解视觉要素尺度变化多样的问题。Deeplab系列模型[12-14]引入空洞空间金字塔池化(ASPP)，聚合不同扩张率的空洞卷积生成的多尺度特征，有助于增强对不同尺度目标的预测能力。

考虑到Deeplab v3+模型[14]同时拥有简单有效的编解码结构和聚合多尺度特征的ASPP模块，并在多个公开数据集上取得优异成绩，本文计划以Deeplab v3+模型为基础对其进行深入研究，改进该模型结构中尚存的不足之处。Deeplab v3+模型的编码阶段通过ASPP聚合上下文信息，但其内部并行结构使得各支路信息间相互独立，缺乏空间相关性；解码阶段只融合了骨干网络上多阶段浅层特征中的一个，造成部分有效信息损失，出现分割不连续和分割边界粗糙问题；在特征融合时，直接将高级特征输出与骨干网络中浅层特征拼接融合，忽略了高低级特征不对齐会向语义特征图内引入噪声的问题[15]，降低语义分割精度。

为改善Deeplab v3+模型的语义分割效果，针对模型的不足之处，分别提出对应的改进方案。在编码阶段，创新性地提出一种异感受野拼接ASPP模块，增强ASPP内多支路深层语义信息间的相关性，提升各支路上特征信息利用率；引入深度可分离空洞卷积取代异感受野拼接ASPP内的普通空洞卷积，缓解异感受野拼接后模型参数量增加、训练速度下降的问题。在解码阶段，进行多次高低级特征融合，充分利用骨干网络提取的多阶段有助于还原图像边缘和纹理信息的浅层特征，提升模型细节表征能力；在高低级特征跨层融合前，增设基于注意力机制的特征对齐模块，减少噪声干扰的同时对特征通道加权，抑制冗余通道信息，强化重要特征学习，增强网络学习能力。

2　相关研究

2.1　Deeplab v3+模型

Deeplab v3+模型作为经典的编解码结构，将Deeplab v3模型[13]作为编码层，在其后端级联一个简单有效的解码器，模型结构如图1所示。

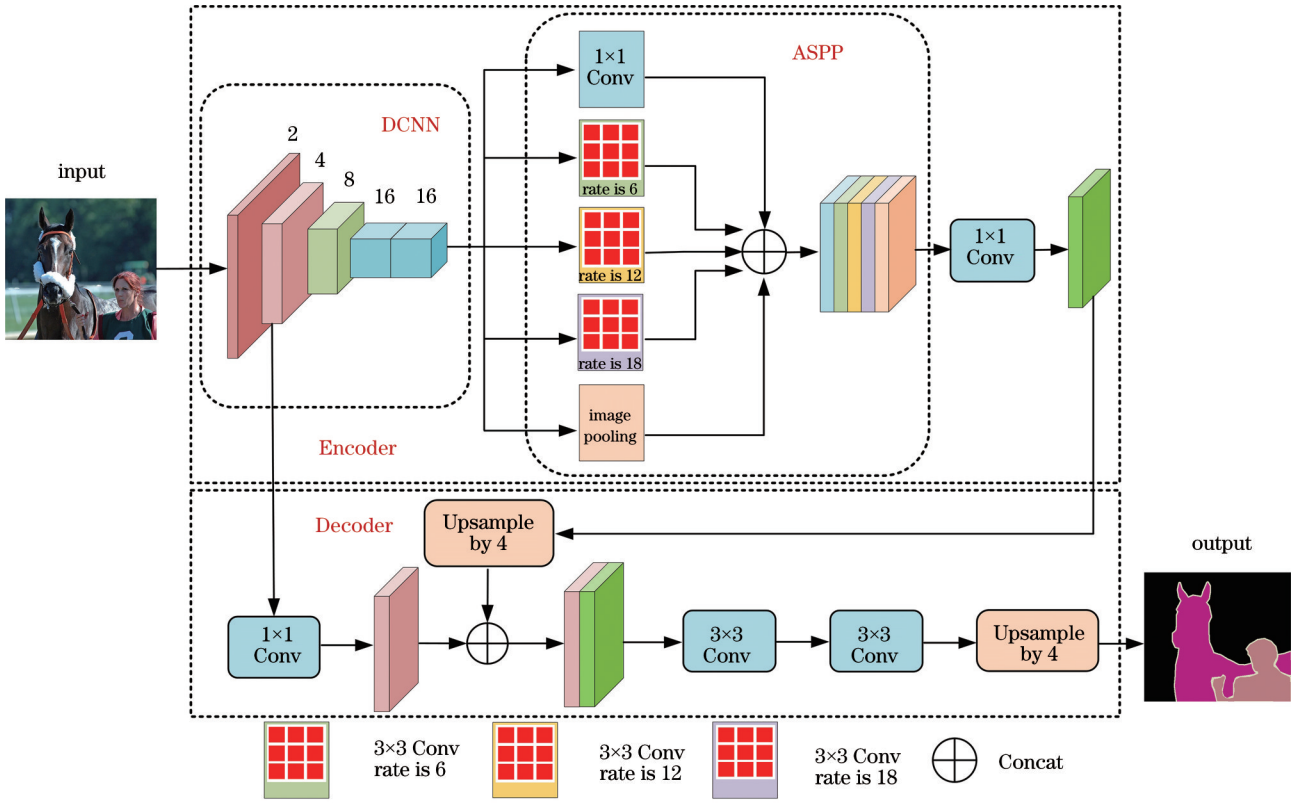

图1　Deeplab v3+模型

Fig. 1　Deeplab v3+ model

编码阶段(Encoder)内，Deeplab v3+模型开创性地将结合深度可分离卷积的轻量化网络Xception[16]用作骨干网络进行初始特征提取，并舍弃一般基础网络通过连续下采样操作扩大感受野的方法，在最后一个残差块中引入空洞卷积，在不损失图像分辨率和不增加额外计算量的同时获得更广阔的感受野；为应对分割目标尺度多样性问题，Deeplab v3+模型增设ASPP模块，该模块采用全局平均池化、1×1卷积、扩张率(用rate表示)分别为6、12、18的空洞卷积组合对图像上下文信息进行编码；接着在通道维度上将多尺度特征图拼接融合，利用1×1的卷积调整输出通道数为256，实现通道压缩，此时的特征图分辨率为原图的1/16。

解码阶段(Decoder)，对编码阶段输出的特征张量采用双线性插值4倍上采样后，与Xception上对应层级的特征图拼接，利用跨层连接捕捉浅层特征承载的细节信息，进一步丰富图像的语义信息和细节信息；经两个3×3卷积细化特征后，使用双线性插值4倍上采样将特征图尺寸逐步恢复到原始图像大小，缓解采样幅度过大导致部分特征信息丢失问题。

2.2　深度可分离卷积

深度可分离卷积[17]将标准卷积运算分解为逐通道卷积(depthwise convolution)与逐点卷积(pointwise convolution)两步操作。标准卷积通过卷积核同时对输入图像的所有通道进行加权操作[18]。两类卷积的操作流程分别如图2、3所示。深度可分离卷积先进行逐通道卷积，通过与上一层通道数相同的卷积核学习空间相关性，每个卷积核仅负责对应的一个通道；接着利用1×1卷积核对逐通道卷积的输出进行卷积操作，完成逐点卷积。深度可分离卷积分别考虑图像空间维度和通道维度，在仅损失少量精度的情况下，大幅度减少模型参数量，有效提高了模型训练效率。

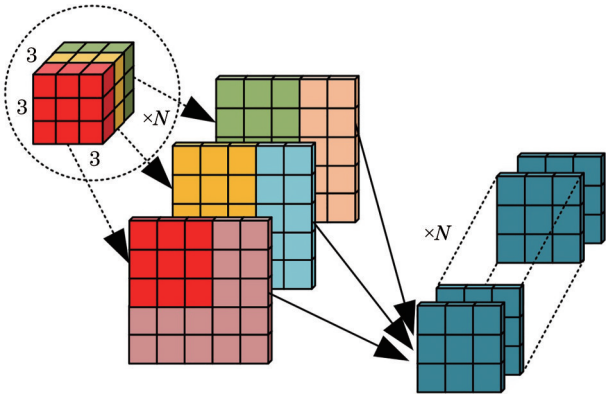

图2　标准卷积

Fig. 2　Standard convolution

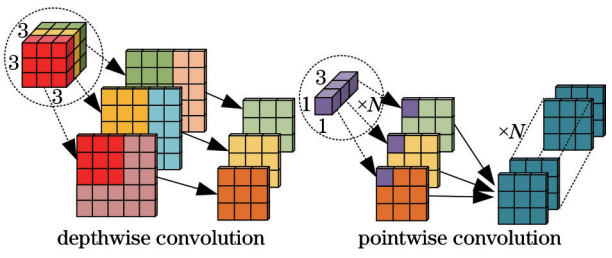

图3　深度可分离卷积

Fig. 3　Depthwise separable convolution

3　Deeplab v3+模型改进

3.1　N-Deeplab v3+模型结构

在Deeplab v3+模型基础上提出一种改进的图像语义分割模型N-Deeplab v3+，旨在解决语义分割时出现的目标漏分割、误分割和分割不连续的问题，增强语义分割精度。N-Deeplab v3+模型延续了Deeplab v3+模型的编解码结构，整体模型结构如图4所示。

编码阶段，创新性地提出一种异感受野拼接ASPP(HFS-ASPP)，以多支路卷积交互连接共享信息，增强各层支路特征间的相关性，提升信息利用率；将深度可分离卷积与空洞卷积相结合，构建深度可分离空洞卷积(DSAConv)，构成新的3×3 DSAConv代替HFS-ASPP内的普通空洞卷积，降低模型参数量，加快模型训练效率。解码阶段，在高低级特征跨层融合前引入基于注意力机制的特征对齐模块(A-FAM)，降低高低级特征直接跨层融合时噪声对特征图的影响，并通过注意力机制优化特征通道权重。为进一步提升图像细节还原度，减少特征恢复过程中的信息损失，对编码层输出先进行2倍上采样后与骨干网络对应层级的特征图融合，经1×1卷积将通道维度降为256，再进行2倍上采样后与骨干网络上尺寸为输入图像1/4大小的特征图拼接融合，最后使用两个3×3深度可分离卷积细化特征。

3.2　深度可分离空洞卷积

在深度可分离卷积的逐通道卷积环节引入空洞卷积后得到的新卷积称为DSAConv，具体操作如图5所示。深度可分离空洞卷积可以大幅降低模型参量，保证模型精度的同时提升计算速度。空洞卷积通过对卷积核进行补零操作，在未增添额外参量并保持特征图分辨率不变情况下增大卷积核的感受野，使得每个卷积输出都承载更大范围信息。深度可分离空洞卷积取代空洞卷积，以缓解异感受野拼接后模型参数量、计算量增加的问题。

3.3　异感受野拼接ASPP模块

异感受野拼接ASPP模块在保留全局平均池化和1×1卷积不变的情况下，使用3路交互连接的深度可分离空洞卷积取代3路并行空洞卷积，详细结构如图4所示。多路卷积通过交互连接的方式可以获得更大范围的感受野，感受野D的计算公式为

$$D=\sum_{i=1}^{n}D_i-(n-1),\tag{1}$$

式中：n表示级联卷积个数；D_i表示第i个卷积的感受野范围。

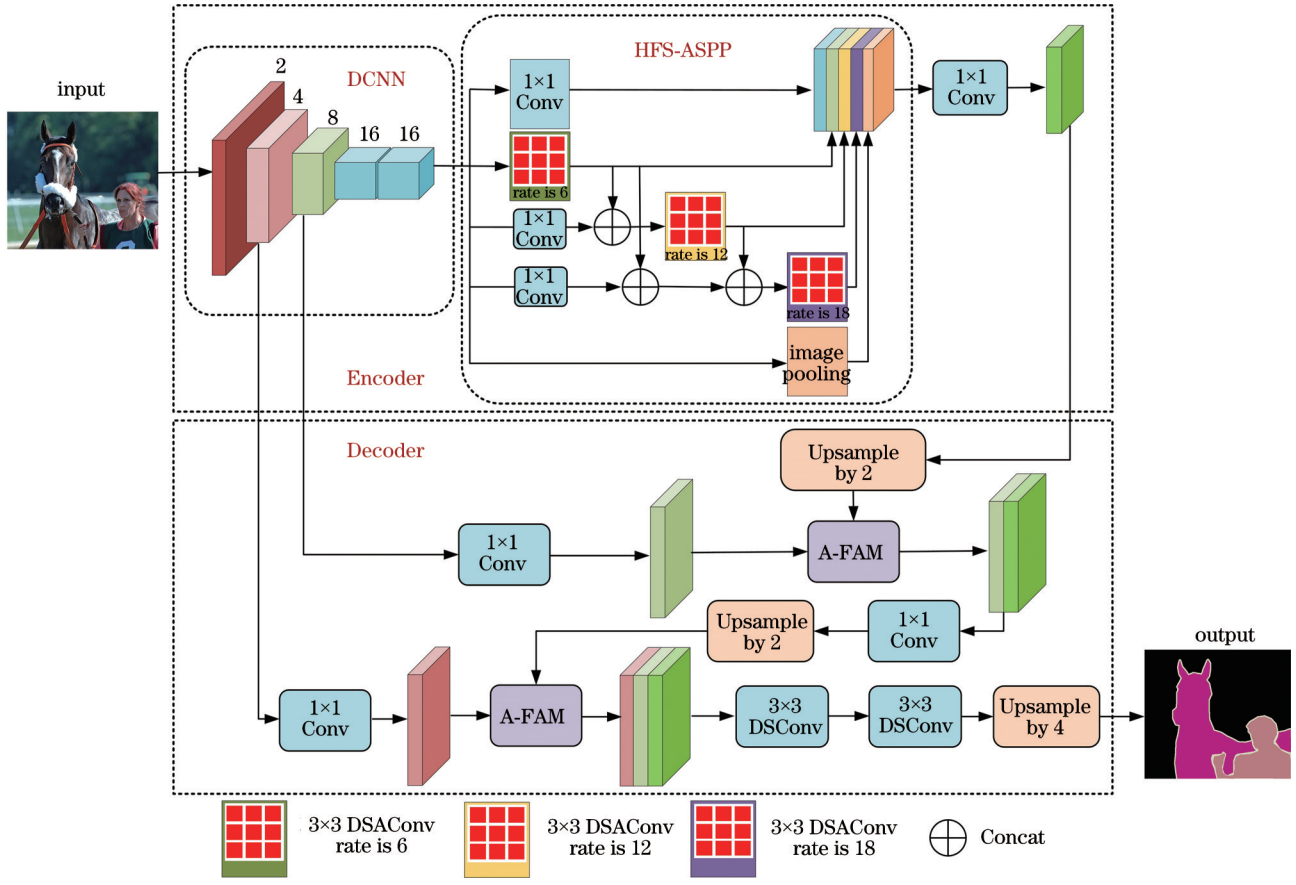

图 4　N-Deeplab v3+模型总体结构

Fig. 4　Overview of N-Deeplab v3+ model architecture

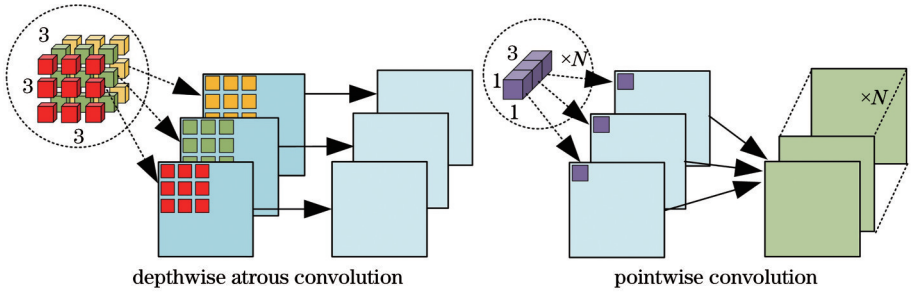

图 5　深度可分离空洞卷积

Fig. 5　Depthwise separable atrous convolution

图 6 更为直观地展示了异感受野拼接的优势，其中，图 6(a)表示扩张率为 12 的单层卷积对原始特征图的采样点一维分布，图 6(b)表示扩张率为 12 的卷积级联扩张率为 6 的卷积后在原始特征图上的采样点一维

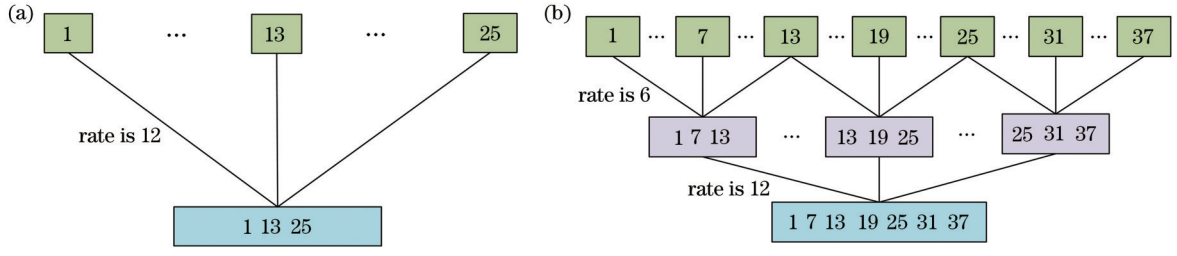

图 6　异感受野拼接对采样点的影响。(a)扩张率为 12 的卷积采样点分布；(b)扩张率为(6 12)的级联卷积采样点分布

Fig. 6　Influence of hetero-receptive field splicing on sampling points. (a) Sampling point distribution of convolution with rate of 12; (b) sampling point distribution of cascaded convolution with rates of (6 12)

分布。级联拼接后的卷积组合可以在原始特征图上获取更多采样点信息，增强信息交互关系。

将特征图上有效运算的像素点个数与感受野范围内全部像素点的比值定义为信息利用率，不同扩张率的卷积组合经异感受野拼接前后在特征图上的表现如表1所示。分析表1可知，异感受野拼接ASPP模块能够提供7种范围的感受野，扩张率为12和18的卷积经异感受野拼接后有效提升像素信息利用率。结果表明，异感受野拼接可以提供更为丰富的尺度多样性，进一步增强像素间的交互关系，获得更稠密的上下文信息，有效提升信息利用率。

表1　异感受野拼接的影响

Table 1　Influence of heterogeneous field splicing

Dilation rate	Receptive field	Effective operation pixel	Information utilization /%
6	13×13	3×3	5.326
12	25×25	3×3	1.440
18	37×37	3×3	0.657
6+12	37×37	7×7	3.579
6+18	49×49	9×9	3.374
12+18	61×61	9×9	2.177
6+12+18	73×73	13×13	3.171

异感受野拼接方式增加了模型复杂度，为防止网络过宽、参数量过大影响模型效率，对此通过调整通道数降低异感受野拼接ASPP模块的参数量。ASPP模块输入特征图通道数2048，输出通道数256，输入与输出尺寸一致，此时的参数量为$n_1=3^2\times2048\times256\times3=14155776$。对于改进ASPP模块，在扩张率为12和18的深度可分离空洞卷积前引入1×1卷积减小网络宽度，通道数分别调整为1792和1536，同时1×1卷积能够增加更多的非线性因素，此时参数量为$n_2=(3^2\times2048+2048\times256)+\{2048\times1792+[3^2\times(1792+256)+(1792+256)\times256]\}+\{2048\times1536+[3^2\times(1536+256\times2)+(1536+256\times2)\times256]\}=8443904$，引入深度可分离空洞卷积并调节网络宽度后，异感受野拼接ASPP模块的模型参数量与ASPP相比降低40.35%，证明改进后的ASPP拥有更小的网络复杂度，有利于提升模型训练效率。

3.4　基于注意力机制的特征对齐模块

Deeplab v3+模型在预测分割结果过程中，选择直接将高级特征与4倍下采样的低级特征拼接，添加额外空间信息，未考虑高低级特征是否对齐情况。实际上不同深度网络特征图各通道承载的特征信息各不相同，其与目标关联程度不尽相同。网络浅层特征图包含大量有助于为高分辨率预测生成清晰边界的细节信息，随着网络加深，特征图涵盖更多有助于图像区域分类识别的抽象语义信息，二者直接拼接会引入噪声，影响后续特征学习。因此，引入基于注意力机制的特征对齐模块(A-FAM)来引导高级特征与低级特征对齐，并依据各个特征通道的承载信息对目标预测贡献的大小，对各特征通道附加权重系数，突出对目标预测有重要作用的特征，抑制冗余通道信息，有针对性地强化特征学习，进一步提升模型的学习能力和泛化能力。

A-FAM参考了压缩激励网络(SENet)[19]中的通道注意力模块(CAM)，但不仅仅使用全局平均池化获得通道更新权重向量，额外引入一路使用全局最大值池化的通道注意力，全局最大值池化能够有效降低特征中噪声的影响[20]。A-FAM结构如图7所示。

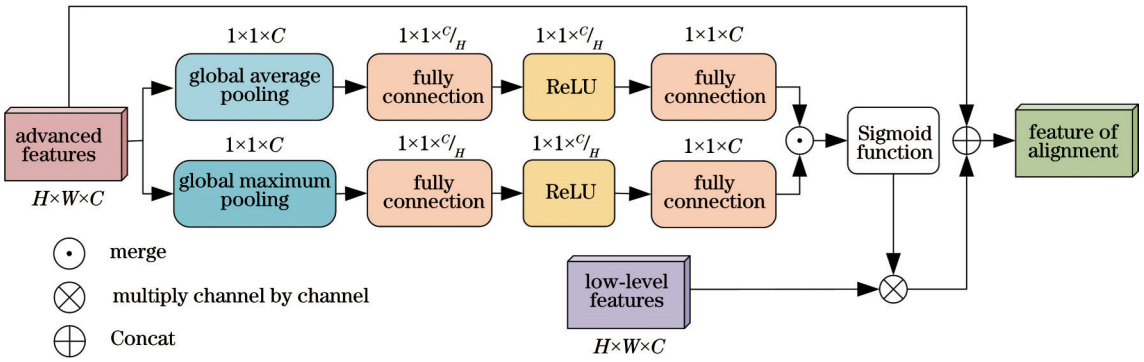

图7　基于注意力机制的特征对齐模块

Fig. 7　Feature alignment module based on attention mechanism

两路并行通道注意力都包含压缩和激励部分。压缩部分，分别通过全局平均池化和全局最大值池化统计特征图通道信息，获得1×1×C的一维向量$\boldsymbol{z}$。激励部分，通过全连接得到通道间的关系表达，公式为

$$\boldsymbol{s}=\sigma\left[g(\boldsymbol{z},\boldsymbol{\omega})\right]=\sigma\left[\boldsymbol{\omega}_2\delta(\boldsymbol{\omega}_1\boldsymbol{z})\right]。\tag{2}$$

首先，通过权重为$\boldsymbol{\omega}_1$的全连接层降低通道数，为原来的1/H，经中间层ReLU函数δ激活后传递给第二个权重为$\boldsymbol{\omega}_2$的全连接层，恢复通道数；其次，两路升维后的特征向量相加并经Sigmoid函数σ生成归一化通道权重向量$\boldsymbol{s}\in\mathbf{R}^{1\times1\times C}$，其与低级特征$\boldsymbol{F}_1$逐通道相乘，得到通道权重修正后的特征$\boldsymbol{F}_m$，表达式为

$$\boldsymbol{F}_{\mathrm{m}}=\boldsymbol{F}_{\mathrm{l}}\cdot\sigma\left\{\boldsymbol{\omega}_2\delta\left\{\boldsymbol{\omega}_1\left[A_{\mathrm{vgpool}}(\boldsymbol{F}_{\mathrm{h}})\right]\right\}+\boldsymbol{\omega}_2\delta\left\{\boldsymbol{\omega}_1\left[M_{\mathrm{axpool}}(\boldsymbol{F}_{\mathrm{h}})\right]\right\}\right\},\tag{3}$$

最后与高级特征 $\boldsymbol{F}_{\mathrm{h}}$ 合并后输出带有“通道注意力”的高低级对齐特征。

结合 Deeplab v3+模型改进实际情况，解码层输出特征与骨干网络上低级特征的两次融合前采用通道衰减率 H 为 16 的 A-FAM。为降低模型参数量，A-FAM 前端两路并行通道内的全连接层实行参数共享。

4　实　　验

4.1　实验设计

对 N-Deeplab v3+算法进行实验验证，使用城市街景数据集 Cityscapes 和标准公共数据集 PASCAL VOC 2012 验证所提改进模型的有效性和泛化性。Cityscapes 是基于驾驶场景的大规模城市街景数据集，拥有 50 个城市内不同街道上 5000 张经高质量标注的街景图，用于训练、测试和验证的图片数量分别为 2975、1525 和 500。图片分辨率统一为 1024 pixel×2048 pixel，共计 19 个语义类别用于模型训练与评估。PASCAL VOC 2012 是常用于计算机视觉领域的公共标准数据集，用于训练、测试和验证的图片数量分别为 1464、1456 和 1449，共计 21 个语义类别。

采用平均交并比(mIoU)作为语义分割质量的评价指标。mIoU 可直观理解为预测值与真实值间交集与并集的比值，表示预测值与真实值的重合度，表达式为

$$P_{\mathrm{mIoU}}=\frac{1}{K}\left(\sum_{x=1}^{K}\frac{T_{xx}}{\sum_{y=1}^{K}T_{xy}+\sum_{y=1}^{K}T_{yx}-T_{xx}}\right),\tag{4}$$

式中：K 代表图像内像素的总类别数；T_{xx} 表示像素点实际类别是 x 类、预测类别也是 x 类的像素总数；T_{xy} 表示像素实际类别是 x 类、预测类别是 y 类的像素总数；T_{yx} 表示像素实际类别是 y 类、预测类别是 x 类的像素总数。

实验基于 TensorFlow 网络框架，使用 python3.6 编写实现，在 Ubuntu18.04 系统下使用 4 块 NVIDIA GeForce GTX 1080Ti 图形处理器计算，Cuda10.0 库加速。采用迁移学习的方式初始化权重，通过 Cityscapes 和 PASCAL VOC 2012 数据集分别对模型进行微调，增加模型收敛速度。模型使用带动量 momentum 的随机梯度下降法进行训练；采用随迭代次数增加学习率逐渐衰减的“Poly”学习策略，设置基础学习率为 1×10^{-4}，动量为 0.9；将输入图片尺寸均裁剪为 513×513。针对上述两个数据集，迭代步数分别选择 130×10^3 和 50×10^3。

4.2　实验结果与性能分析

4.2.1　模型性能对比

为准确衡量新模型性能，验证模型改进的有效性。对所提 N-Deeplab v3+与 FCN-8S、SegNet、Deeplab v2、PSPNet 和 Deeplab v3+等模型在 Cityscapes 验证集上进行实验验证，预测结果如表 2 所示。表 3 展示了 Deeplab v3+模型改进前后的详细量化信息对比，其中 T_0 表示预测单张图片所需时间。

表 2　不同模型在 Cityscapes 数据集上的性能对比

Table 2　Performance comparison of different models on the Cityscapes dataset

Model	Backbone network	mIoU /%
FCN-8S	VGG-16	62.21
SegNet	VGG-16	62.64
Deeplab v2	ResNet101	68.52
PSPNet	ResNet101	73.98
Deeplab v3+	Xception	74.62
N-Deeplab v3+	Xception	76.31

结合表 2、3 可知：Deeplab v3+相比其他语义分割模型取得了优异的预测结果；构建的新模型获得了更有竞争性的成果，mIoU 为 76.31%，比 Deeplab v3+的预测结果高出 1.69 个百分点，以较小的预测速度为代价，换得分割精度的显著提升，较好地权衡了二者之间的关系，并在提升模型分割精度的同时，降低计算机内存占用，一定程度上提升了工程实用性。

表 3　Deeplab v3+改进前后量化信息对比

Table 3　Quantifying information comparison of Deeplab v3+ before and after improvement

Model	Number of parameters /10^6	Model size /Mbit	T_0 /ms	Speed /(frame·s^{-1})
Deeplab v3+	43.51	165.62	275.3	3.632
N-Deeplab v3+	37.38	142.28	302.7	3.466

为了更加直观地展现新模型的优越性，可视化模型改进前后在 Cityscapes 验证集上的预测结果，可视化结果对比如图 8 所示。观察图 8 第一列图片可知：Deeplab v3+模型对图片内左侧路灯和公交车后视镜分割不连续，将左侧道路区域错误地理解为人行道，且未能预测出右侧路标杆，细节丢失严重；N-Deeplab v3+模型妥善处理了上述不足之处，准确地表征出图像细节信息，边缘预测更为准确清晰，解决了漏分割和分割不连续问题。对比图 8 第二列图片，Deeplab v3+模型将左侧公交车后视镜误分割为交通标志类，错误地将右侧交通标志预测为公交车的一部分；N-Deeplab v3+模型可以正确地预测出物体相对应类

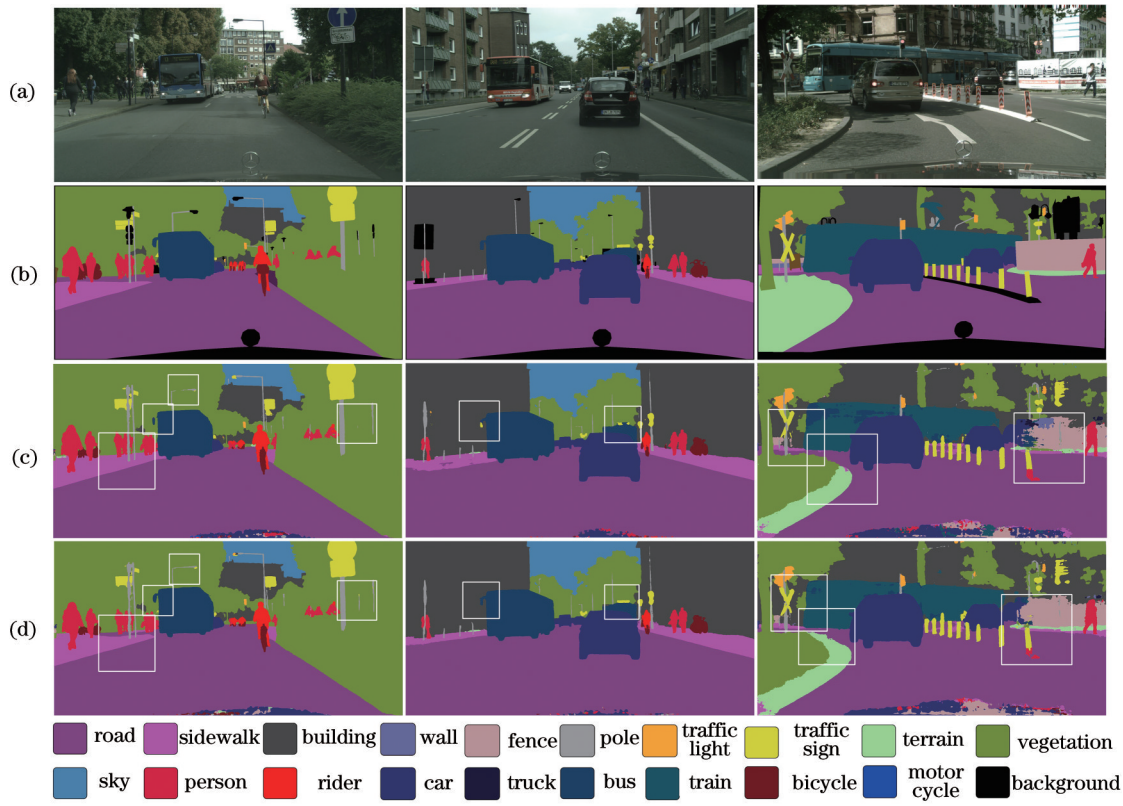

图 8　Cityscapes 验证集上分割结果对比。(a)输入图片;(b)标签图片;(c) Deeplab v3+分割结果;(d) N-Deeplab v3+分割结果

Fig. 8　Comparison of segmentation results on Cityscapes validation set. (a) Input images; (b) label images; (c) segmentation results of Deeplab v3+; (d) segmentation results of N-Deeplab v3+

别,成功避免了误分割问题。对比图 8 第三列图片,Deeplab v3+模型对交通标志和墙面等细节部分分割较为粗糙,相比较而言,N-Deeplab v3+模型更好地保留图像细节信息,预测结果更加准确和全面。

4.2.2　ASPP 模块改进实验

为检验异感受野拼接 ASPP 对增强算法性能的有效性,采用控制变量法验证各改进方案的效果,对比 ASPP 分别选择不同扩张率组合、连接方式以及引入深度可分离空洞卷积的情况,各改进方案在 Cityscapes 验证集上的测试结果如表 4 所示,其中 Train time 表示模型训练时长。分析表 4 信息可知,异感受野拼接在显著提升预测准确率的同时,也伴随着模型复杂度提升,影响了图片预测速度。对比实验 3、5 和实验 4、6 可知,引入深度可分离空洞卷积后,mIoU 保持基本不变的情况下模型训练时长和图片预测时间都有明显缩短。综合分析,异感受野拼接 ASPP 内扩张率组合为(6 12 18 24)时分割效果最佳,但预测单张图片消耗的时间比原模型多出 13.5%;而扩张率为(6 12 18)的卷积组能够做到 mIoU 提升 0.74 个百分点的同时,预测速度提升 8.03%。因此,异感受野拼接 ASPP 模块内 3 路深度可分离空洞卷积组合的扩张率选择(6 12 18)。

表 4　ASPP 模块改进方案测试结果对比

Table 4　Comparison of test results of improved scheme in ASPP module

Group	Dilation rate	HFS	DSAConv	mIoU /%	Train time /h	T_0 /ms
1	(6 12 18)			74.62	23.82	275.3
2	(6 12 18 24)			74.91	25.64	310.8
3	(6 12 18)	√		75.39	27.27	322.4
4	(6 12 18 24)	√		75.71	30.44	372.0
5	(6 12 18)	√	√	75.36	21.59	253.2
6	(6 12 18 24)	√	√	75.62	25.60	312.5

4.2.3　消融实验

为验证异感受野拼接 ASPP、多级特征融合(MFF)模块和基于注意力机制的特征对齐模块等方案的有效性,在 Cityscapes 数据集上进行逐层的消融实验,以 mIoU 和速度为对比指标,实验结果如表 5 所示。

表 5 不同改进方案在 Cityscapes 数据集上性能分析

Table 5 Performance comparison of different improvement schemes on the Cityscapes dataset

Group	HFS-ASPP	MFF	A-FAM	mIoU /%	Speed /(frame·s^{-1})
7				74.62	3.632
8	√			75.36	3.949
9	√	√		75.69	3.832
10	√	√	√	76.31	3.466

对比实验 7、8 可知，使用异感受野拼接 ASPP 模块代替原模型中 ASPP 模块后，mIoU 提升了 0.74 个百分点，速度提升 8.73%，该方法增强了各支路间信息的相关性，并有效扩大卷积组的感受野，提高信息利用率，并降低了模型参数量和计算量，提升模型预测效率。对比实验 8、9 可知，多级特征融合操作后的 mIoU 升高 0.33 个百分点，速度下降 0.117 frame/s，表明多尺度跨层融合操作在丰富细节信息和语义信息的同时，增加了部分计算量。对比实验 9、10 可知，引入基于注意力机制的特征对齐模块后模型的 mIoU 升高了 0.62 个百分点，证明该模块对进一步提高分割精度有着积极作用；速度下降 0.366 frame/s，反映出高低级特征对齐和通道权重优化等操作增加了模型参数量和计算量，分割效率受到影响。综合分析表 5 信息，若要兼顾预测精度与检测速率，对模型进行改进的难度较大，N-Deeplab v3+模型整体以较小的检测速度为代价，换得模型分割精度显著提升，较好地平衡了分割精度与效率，体现了所提方法的优越性。

4.2.4 泛化实验

此外，为进一步检验 N-Deeplab v3+模型的泛化能力，另外在 PASCAL VOC 2012 数据集上进行实验验证，以 mIoU 和速度为评价指标，模型改进前后的实验详细量化信息如表 6 所示。数据显示，改进后的模型与原模型相比，速度仅受到微弱影响，分割精度有效提升了 2.14 个百分点。

表 6 Deeplab v3+改进前后在 PASCAL VOC 2012 数据集上的性能对比

Table 6 Performance comparison of Deeplab v3+ before and after improvement on the PASCAL VOC 2012 dataset

Method	mIoU /%	T_0 /ms	Speed /(frame·s^{-1})
Deeplab v3+	79.83	46.86	21.340
N-Deeplab v3+	81.97	48.35	20.612

所提 N-Deeplab v3+模型与 Deeplab v3+模型在 PASCAL VOC 2012 验证集上的预测信息可视化图片如图 9 所示。通过观察第 1、2 列图片可知，N-Deeplab v3+

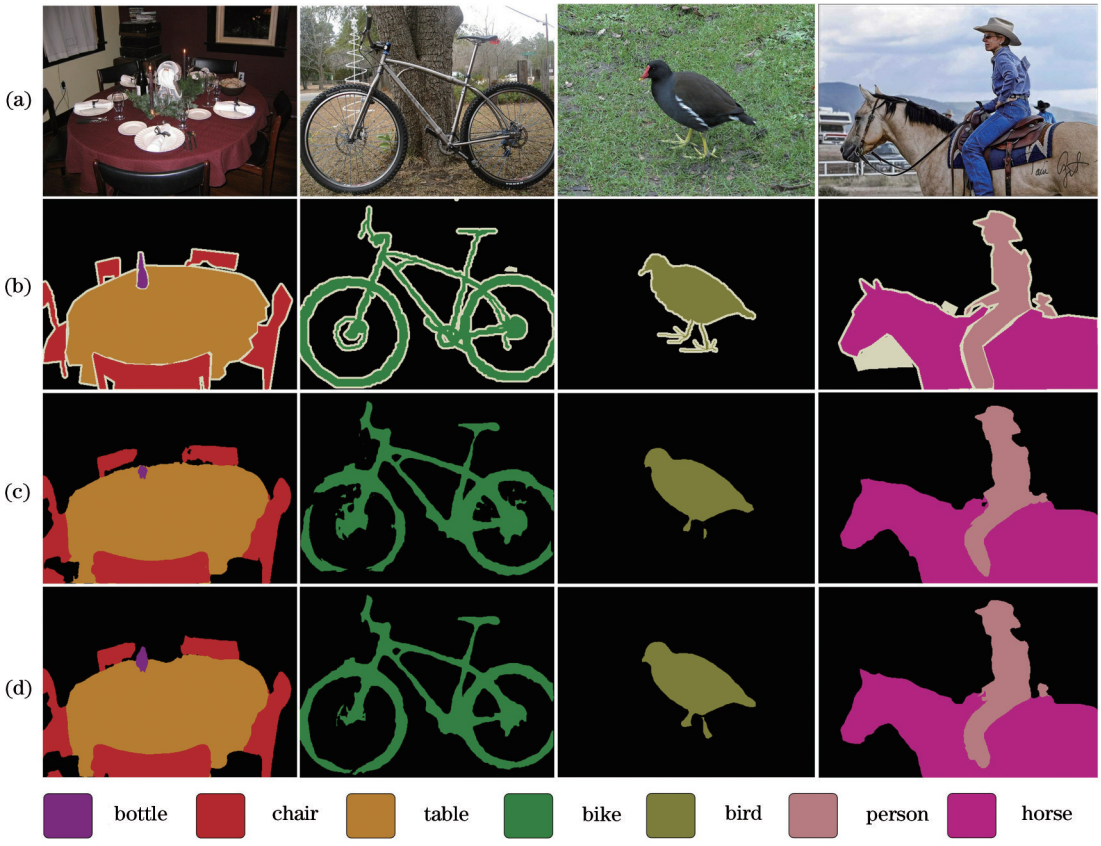

图 9 PASCAL VOC 2012 验证集上分割结果对比。(a)输入图片；(b)标签图片；(c)Deeplab v3+分割结果；(d)N-Deeplab v3+分割结果

Fig. 9 Comparison of segmentation results on PASCAL VOC 2012 verification set. (a) Input images; (b) label images; (c) segmentation results of Deeplab v3+; (d) segmentation results of N-Deeplab v3+

对图像区域预测更为连续和准确；通过第3、4列图片可知，N-Deeplab v3+对鸟喙、鸟腿和马耳等细节部分的预测更加细腻，预测的整体轮廓更为平滑。综上，所提N-Deeplab v3+模型在PASCAL VOC 2012数据集上依旧取得了比原模型更优异的分割性能，进一步验证所提改进模型具有一定的泛化性。

5 结 论

针对Deeplab v3+模型细节表征能力较弱，存在漏分割、误分割问题，提出相应的模型改进方案。通过设计异感受野拼接ASPP模块，提供更为丰富的尺度多样性，进一步增强了各层级信息间的交互关系，提升信息利用率；在异感受野拼接ASPP内引入深度可分离空洞卷积取代普通空洞卷积，加快模型训练速度；增设多级高低特征融合操作，尽可能多地恢复在降采样过程中损失的空间维度信息和像素位置信息；最后，创新性地构建基于注意力机制的特征对齐模块，引导高低级特征对齐并强化特征学习，增强模型学习能力。在两个公开数据集上的实验数据证明，改进后的模型结构不仅对边缘的语义类别刻画更为准确，并且更加关注图像位置与纹理信息的提取，提升了模型结构的表征力，成功改善了模型分割效果。在后续工作中，将深入研究兼顾预测精度与实时性的高性能网络，进一步增强语义分割算法在工程应用中的实用性。

参 考 文 献

[1] 陈浩, 杨恺伦, 胡伟健, 等. 基于全景环带成像的语义视觉里程计[J]. 光学学报, 2021, 41(22): 2215002
Chen H, Yang K L, Hu W J, et al. Semantic visual odometry based on panoramic annular imaging[J]. Acta Optica Sinica, 2021, 41(22): 2215002.

[2] Chen B K, Gong C, Yang J. Importance-aware semantic segmentation for autonomous vehicles[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(1): 137-148.

[3] Sáez Á, Bergasa L M, López-Guillén E, et al. Real-time semantic segmentation for fisheye urban driving images based on ERFNet[J]. Sensors, 2019, 19(3): 503.

[4] 张哲晗, 方薇, 杜丽丽, 等. 基于编码-解码卷积神经网络的遥感图像语义分割[J]. 光学学报, 2020, 40(3): 0310001.
Zhang Z H, Fang W, Du L L, et al. Semantic segmentation of remote sensing image based on encoder-decoder convolutional neural network[J]. Acta Optica Sinica, 2020, 40(3): 0310001.

[5] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 3431-3440.

[6] Yu F, Koltun V, Funkhouser T. Dilated residual networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 636-644.

[7] 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) [2021-03-02]. https://arxiv.org/abs/1412.7062.

[8] Dai J F, Qi H Z, Xiong Y W, et al. Deformable convolutional networks[C]//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 764-773.

[9] 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.

[10] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation[M]//Navab N, Hornegger J, Wells W M, et al. Medical image computing and computer-assisted intervention-MICCAI 2015. Lecture notes in computer science. Cham: Springer, 2015, 9351: 234-241.

[11] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6230-6239.

[12] 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.

[13] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-12-05) [2020-10-15]. https://arxiv.org/abs/1706.05587.

[14] Chen L C, Zhu Y K, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 833-851.

[15] 张蕊, 李锦涛. 基于深度学习的场景分割算法研究综述[J]. 计算机研究与发展, 2020, 57(4): 859-875.
Zhang R, Li J T. A survey on algorithm research of scene parsing based on deep learning[J]. Journal of Computer Research and Development, 2020, 57(4): 859-875.

[16] Chollet F. Xception: deep learning with depthwise separable convolutions[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1800-1807.

[17] Chollet F. Xception: deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1251-1258.

[18] 刘文, 王海荣, 周北京. DeepLabv3plus-IRCNet: 小目标

特征提取的图像语义分割[J]. 中国图象图形学报, 2021, 26(2): 391-401.
Liu W, Wang H R, Zhou B J. DeepLabv3plus-IRCNet: an image semantic segmentation method for small target feature extraction[J]. Journal of Image and Graphics, 2021, 26(2): 391-401.

[19] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7132-7141.

[20] 郭列, 张团善, 孙威振, 等. 融合空间注意力机制的图像语义描述算法[J]. 激光与光电子学进展, 2021, 58(12): 1210030.
Guo L, Zhang T S, Sun W Z, et al. Image semantic description algorithm with integrated spatial attention mechanism[J]. Laser & Optoelectronics Progress, 2021, 58(12): 1210030.