Journal of System Simulation

Volume 35 | Issue 11 Article 4

11-30-2023

Image Semantic Segmentation Algorithm Based on Improved DeepLabv3+

Weiping Zhao
Liaoning General Aviation Academy, Shenyang Aerospace University, Shenyang 110034, China; College of Electronic Information Engineering, Shenyang Aerospace University, Shenyang 110034, China, 3370477370@qq.com

Yu Chen
College of Electronic Information Engineering, Shenyang Aerospace University, Shenyang 110034, China, 1009857106@qq.com

Song Xiang
Liaoning General Aviation Academy, Shenyang Aerospace University, Shenyang 110034, China

Yuanqiang Liu
Liaoning General Aviation Academy, Shenyang Aerospace University, Shenyang 110034, China

See next page for additional authors

Image Semantic Segmentation Algorithm Based on Improved DeepLabv3+

Abstract

Abstract: Mainstream image semantic segmentation networks currently face problems such as incorrec segmentation, discontinuous segmentation, and high model complexity, which cannot be flexibly and efficiently deployed in practical scenarios. To this end, an image semantic segmentation network that optimizes the DeepLabv3+ model is designed by comprehensively considering the network parameters, prediction time, and accuracy. The lightweight EfficientNetv2 is adopted to extract backbone network features and improve parameter utilization. In the atrous spatial pyramid pooling module, the mixed strip pooling is utilized to replace the global average pooling, and a depthwise separable dilated convolution is introduced to reduce parameters and improve the ability to learn multi-scale information. The attention mechanism is employed to enhance the model's representation power, and the multiple shallow features of the backbone network are extracted to enrich the image's geometric details. The experiment shows that the algorithm achieves 81.19% mIoU with a parameter size of 55.51×106, which optimizes the segmentation accuracy and model complexity and improves model generalization.

Keywords

DeepLabv3+, image semantic segmentation, atrous spatial pyramid pooling, attention mechanism, depthwise separable dilated convolution

Authors

Weiping Zhao, Yu Chen, Song Xiang, Yuanqiang Liu, and Chaoyue Wang

Recommended Citation

Zhao Weiping, Chen Yu, Xiang Song, et al. Image Semantic Segmentation Algorithm Based on Improved DeepLabv3+[J]. Journal of System Simulation, 2023, 35(11): 2333-2344.

第 35 卷第 11 期 系统仿真学报© Vol. 35 No. 11
2023 年 11 月 Journal of System Simulation Nov. 2023

基于改进的DeepLabv3+图像语义分割算法研究

赵为平[1,2]，陈雨[2*]，项松[1]，刘远强[1]，王超越[1]
(1. 沈阳航空航天大学 辽宁通航研究院，辽宁 沈阳 110034；2. 沈阳航空航天大学 电子信息工程学院，辽宁 沈阳 110034)

摘要：目前主流图像语义分割网络往往存在误分割、分割不连续和模型复杂度高的问题，不能灵活高效地部署于实际场景中。针对这一现象，通过综合考虑网络的参数量、预测时间和准确度，设计出一种优化DeepLabv3+模型的图像语义分割网络。*骨干网络改用轻量级EfficientNetv2网络提取特征，提高参数利用率；在空洞空间金字塔池化模块中使用混合条带池化模块代替全局平均池化，引入深度可分离膨胀卷积，减少参数量和提高学习多尺度信息的能力；使用注意力机制增强模型表征力，提取骨干网络多条浅层特征，丰富图像的几何细节信息。*实验表明，本文算法可达到mIoU为81.19%，参数量为55.51×10^6，有效优化了分割精度和模型复杂度，同时也提高了模型泛化性。

关键词：DeepLabv3+；图像语义分割；空洞空间金字塔池化；注意力机制；深度可分离膨胀卷积
中图分类号：TP391 文献标志码：A 文章编号：1004-731X(2023)11-2333-12
DOI: 10.16182/j.issn1004731x.joss.22-0690
引用格式: 赵为平, 陈雨, 项松, 等. 基于改进的DeepLabv3+图像语义分割算法研究[J]. 系统仿真学报, 2023, 35(11): 2333-2344.
Reference format: Zhao Weiping, Chen Yu, Xiang Song, et al. Image Semantic Segmentation Algorithm Based on Improved DeepLabv3+[J]. Journal of System Simulation, 2023, 35(11): 2333-2344.

Image Semantic Segmentation Algorithm Based on Improved DeepLabv3+

Zhao Weiping[1,2], *Chen Yu*[2*], *Xiang Song*[1], *Liu Yuanqiang*[1], *Wang Chaoyue*[1]
(1. Liaoning General Aviation Academy, Shenyang Aerospace University, Shenyang 110034, China;
2. College of Electronic Information Engineering, Shenyang Aerospace University, Shenyang 110034, China)

Abstract: Mainstream image semantic segmentation networks currently face problems such as incorrect segmentation, discontinuous segmentation, and high model complexity, which cannot be flexibly and efficiently deployed in practical scenarios. To this end, an image semantic segmentation network that optimizes the DeepLabv3+ model is designed by comprehensively considering the network parameters, prediction time, and accuracy. *The lightweight EfficientNetv2 is adopted to extract backbone network features and improve parameter utilization. In the atrous spatial pyramid pooling module, the mixed strip pooling is utilized to replace the global average pooling, and a depthwise separable dilated convolution is introduced to reduce parameters and improve the ability to learn multi-scale information. The attention mechanism is employed to enhance the model's representation power, and the multiple shallow features of the backbone network are extracted to enrich the image's geometric details. The* experiment shows that the algorithm achieves 81.19% mIoU with a parameter size of 55.51×10^6, which optimizes the segmentation accuracy and model complexity and improves model generalization.

收稿日期：2022-06-17 修回日期：2022-08-16
基金项目：辽宁省教育厅重点公关项目(JYT2020162)；电动水上飞机可靠性设计技术研究(JYT2020162)
第一作者：赵为平(1968-)，男，副教授，博士，研究方向为飞行器设计、图像处理。E-mail：3370477370@qq.com
通讯作者：陈雨(1996-)，男，硕士生，研究方向为深度学习、图像分割。E-mail：1009857106@qq.com

Keywords: DeepLabv3+; image semantic segmentation; atrous spatial pyramid pooling; attention mechanism; depthwise separable dilated convolution

0 引言

在计算机视觉领域，语义分割工作占据着举足轻重的地位[1-2]。微观来看，语义分割任务是针对各像素对应的类别进行解析，通俗的说就是将图像中某一像素识别出是汽车、建筑、树木还是地面等，并为不同标签的像素设定不同色彩。宏观解释，语义分割任务就是从底层语义向高层语义推理的过程，获取到逐像素分割的图像。目前语义分割算法在智能医学图像分析、遥感图像技术、无人驾驶等众多领域均成为了热点研究内容[3-6]。

在图像处理研究早期，传统图像分割方法有结构化随机森林、Normalized-cut 和 SVM(support vector machine)等[7-10]。单独使用这些方法，分割效果和泛化能力较差，很难应用于实际复杂场景中。近年来，随着计算机硬件的支持和深度学习的兴起[11-13]，学术界设计出大量新的高效语义分割算法，获得了不菲的效果[14-18]。其中文献[19]开创性把卷积神经网络(convolutional nerual networks, CNN)的全连接改为卷积操作，得到全卷积神经网络(full convolutional networks, FCN)。FCN 作为第一个端到端、像素到像素的分割网络，也被誉为使用深度学习进行语义分割任务的首创佳作[20]，给后续研究者提供了不容小觑的灵感启发。剑桥大学提出的 SegNet 网络[21]由编码器、解码器以及 softmax 分类层组成，在 FCN 的基础上微调 VGG-16 用于特征的提取，且利用编码器中对应的池化索引做非线性上采样，降低网络计算量和模型参数量，改善了计算效率。U-Net 模型[22]结构酷似英文字母 U，使用编码–解码结构融合低维特征，有效处理了由下采样导致的细节损失(如边界信息)，从而帮助网络完成更精确的定位，在医疗影像分析中颇受欢迎[23-24]。

对于网络模型忽略了全局信息和像素空间一致性的问题，人们把目光转向了基于空洞卷积的分割算法。文献[25]的 DeepLabv1 模型，在深层卷积神经网络 VGG[26](visual geometry group)基础上引入空洞卷积来扩大卷积感受野，感知更多的坐标信息和位置信息。同时通过全连接 CRF 概率图模型做后处理，进而得到相对精确的轮廓。DeepLabv2[27]对 DeepLabv1 进行了改进，通过空洞空间金字塔池化(atrous spatial pyramid pooling, ASPP)模块，进行多个分支异扩张率的膨胀卷积，来抽取不同大小感受野的多尺度特征。DeepLabv3[28]在 DeepLabv2 基础上，使用级联模块，去除 CRF 模块，并且在 ASPP 模块中引入批量归一化(batch normalization, BN)，利用全局平均池化缓解了远距离下重要权重损失的情况。DeepLabv3+[29]仿照编码器–解码器结构，同时充分考虑浅层和深层的语义信息，来优化物体边缘细节。

为了进一步满足模型能够应用于各种嵌入式设备，更灵活高效地完成社会生活中各方面需求。本文提出一种基于改进的 DeepLabv3+语义分割网络，将轻量级 EfficientNetv2 网络[30]作为 DeepLabv3+模型的主干网络，在 ASPP 模块中使用混合条带池化模块和深度可分离膨胀卷积，降低模型参数量、提高推理速度的同时，学习丰富的全局语义和局部纹理、边缘等细节信息。并将融合的多条浅层特征和 N-ASPP 输出的高级特征进行注意力机制操作，使融合后的特征图追踪到更丰富的特征信息，从而使模型更好地兼顾分割精度和模型复杂度。

1 优化的语义分割算法

1.1 总体框架设计

本文基于 DeepLabv3+网络在编码区和解码区

DOI: 10.16182/j.issn1004731x.joss.22-0690

均做了一些改进，总体模型结构如图1所示。

(1) 编码区。首先将原DeepLabv3+网络模型的骨干网络换为Efficientv2网络进行提取特征，然后对ASPP模块进行了改进，引入深度可分离膨胀卷积(depthwise separable dilated convolution, DSDConv)，来综合标准扩张卷积和深度可分离卷积的优势，同时使用混合条带池化模块(mixed strip pooling module, MSPM)代替全局平均池化，帮助模型进一步捕获全局和本地上下文信息，从而形成新的N-ASPP模块，经过N-ASPP模块5个分支不同程度的特征提取，使语义特征有效聚合多尺度的上下文信息。然后再利用基于归一化的注意力模块(normalization-based attention module, NAM)，通过稀疏的权重惩罚判断各通道的显著程度，并在空间注意力子模块中对像素进行归一化，最终在编码区得到一个包含更加详细的语义信息的高级特征图。

(2) 解码区。为了丰富图像局部细节信息，首先提取骨干网络EfficientNetv2中的2条浅层特征，并分别经过NAM的空间注意力子模块，然后对2个初级特征进行通道维度的拼接，完成浅层特征融合(shallow feature fusion, SFF)，获得更详细的图像几何信息，细化模型分割精度。接着将编码区得到的高级特征图做4倍双线性插值上采样，将特征尺寸大小调整为和浅层特征一样。然后将高级特征和融合后的浅层特征图进行拼接，最后再进行一次3×3的卷积和4倍上采样，将分割结果恢复到原图像尺寸大小。

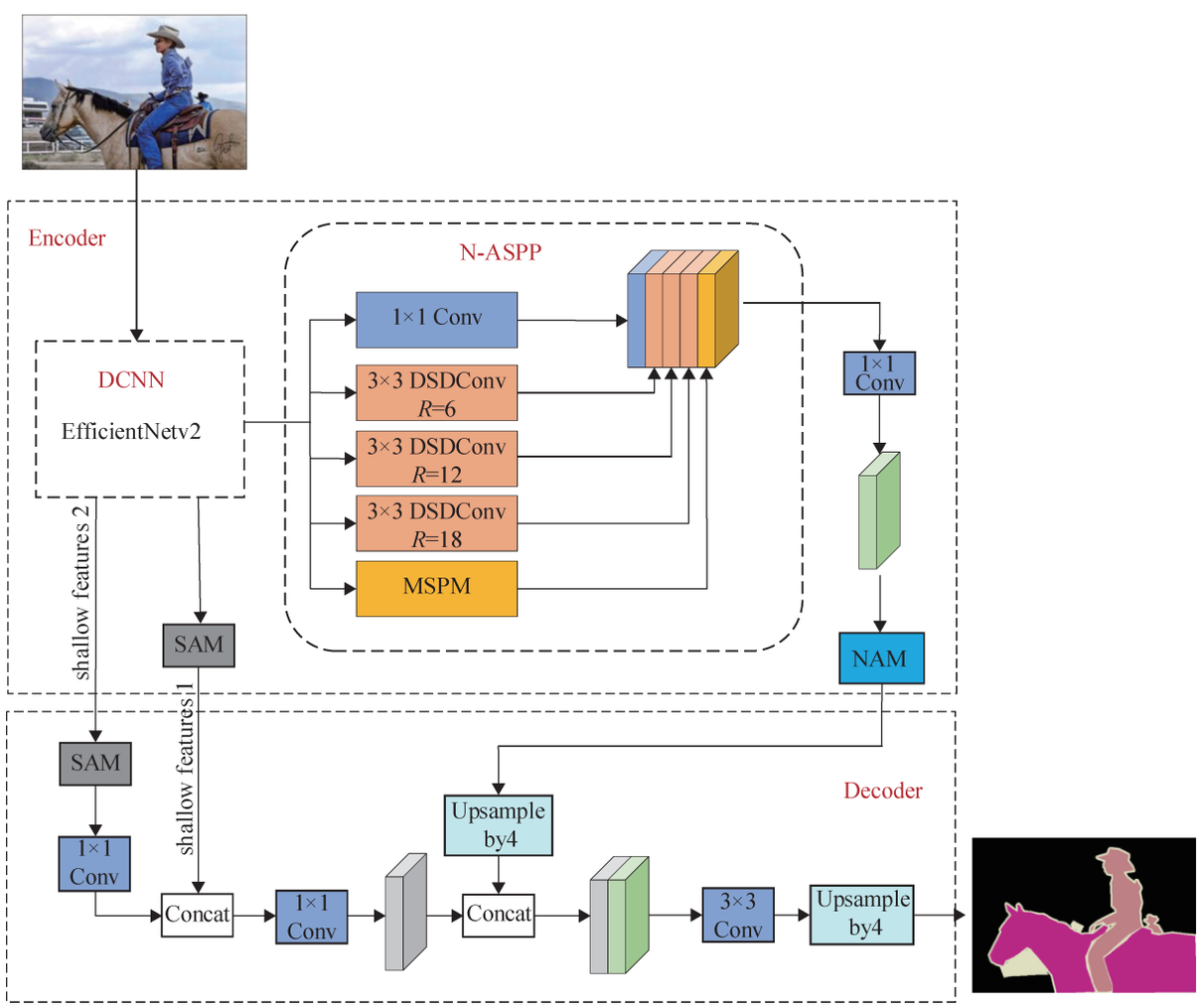


图1　优化的DeepLabv3+网络结构

Fig. 1　Optimized DeepLabv3+ network structure

1.2 骨干网络

EfficientNet[31]是谷歌提出的一种新的轻量级卷积网络，本文采用 EfficientNetv2[30] 作为 DeepLabv3+的特征提取骨干网络，其网络结构参数如表 1 所示。

表 1 EfficientNetv2 网络结构参数
Table 1 Network structure parameters of EfficientNetv2

网络结构形式	图像尺寸	通道数	层数
3×3Conv	224×224	24	1
3×3Fused-MBConv1	112×112	24	2
3×3Fused-MBConv4	112×112	48	4
3×3Fused-MBConv4	56×56	64	4
3×3MBConv4	28×28	128	6
3×3MBConv6	14×14	160	9
3×3MBConv6	14×14	272	15
Conv2D&Pooling&FC	7×7	1 792	1

EfficientNetv2 在 MBConv 的基础上引入了 Fused-MBConv，如图 2 展示了 Fused-MBConv 和常规 MBConv 的具体结构。通过训练感知神经架构搜索(nerual architecture search, NAS)和缩放技术，大幅度改善模型参数的利用率。NAS 是一种搜索最优网络结构的算法，可动态设计 Fused-MBConv 和普通 MBConv 的最优策略，可改善模型精度、参数利用率和硬件 GPU/CPU 效率。并且通过去除非必要的搜索选项，来减小模型的搜索空间，提高训练效率。EfficientNetv2 搜索奖励函数为

$$r = A \cdot S^{\omega} \cdot P^{\upsilon} \tag{1}$$

式中：A 为模型准确率；S 为每个训练 step 的时长；P 为参数量；ω 和 υ 为控制奖励比例的两个超参，ω=0.07，υ=−0.05。

EfficientNetv2 采用新的渐进式学习方法对正则化因子进行自适应调节，有效缓解了极度正则化造成的模型欠拟合和过拟合的情况，主要有两步：①训练处于前期时，选择分辨率较小的输入和较弱的正则化；②逐步扩大输入的尺寸大小和更强的正则化尺度。这一方法可以很好地提高训练速度，同时优化了模型精度和泛化性能。

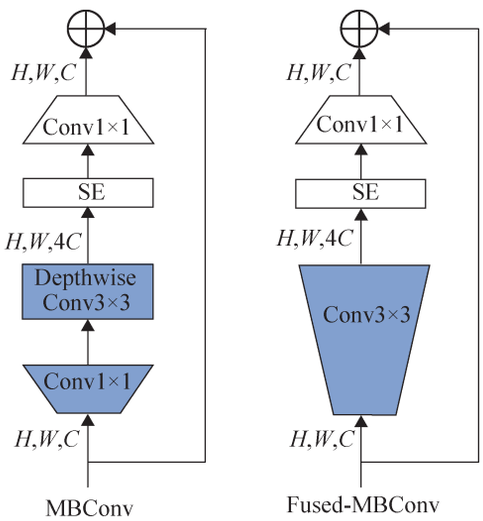


图 2 MBConv 和 Fused-MBConv 结构对比图
Fig. 2 Structure comparison diagram of MBConv and Fused-MBConv

1.3 N-ASPP 模块

通过在 ASPP 模块的基础上引入混合条带池化和深度可分离膨胀卷积，构建出 N-ASPP 模块。将骨干网络提取出来的特征输入到 N-ASPP 模块中，分别经过 1 个 1×1 卷积、膨胀率为 6，12，18 的 3 个深度可分离膨胀卷积、混合条带池化模块等 5 条支路，能更加丰富高效地提取深层语义特征。下面将对混合条带池化模块和深度可分离膨胀卷积分别介绍。

1.3.1 混合条带池化模块

传统空间平均池化是正方形池化窗口，在提取空间位置较复杂的特征时，往往不能收集到各向空间尺度的相关性信息，从而包含许多不相关的像素区域。为有效捕获空间长程依赖关系的同时，学习到丰富的物体几何细节，本模型在 N-ASPP 中将全局平均池化换为 MSPM，如图 3 所示。

假设输入特征为 $\boldsymbol{x} \in \mathbb{R}^{C\times H\times W}$，首先对其进行池化核为 H×1 的垂直池化(vertical pooling，V_pooling)，即对特征图 $\boldsymbol{x}$ 中每一列像素值进行相加再求均值，输出 $\boldsymbol{y}^{v}$ 为 C×1×W 的行向量，其元素

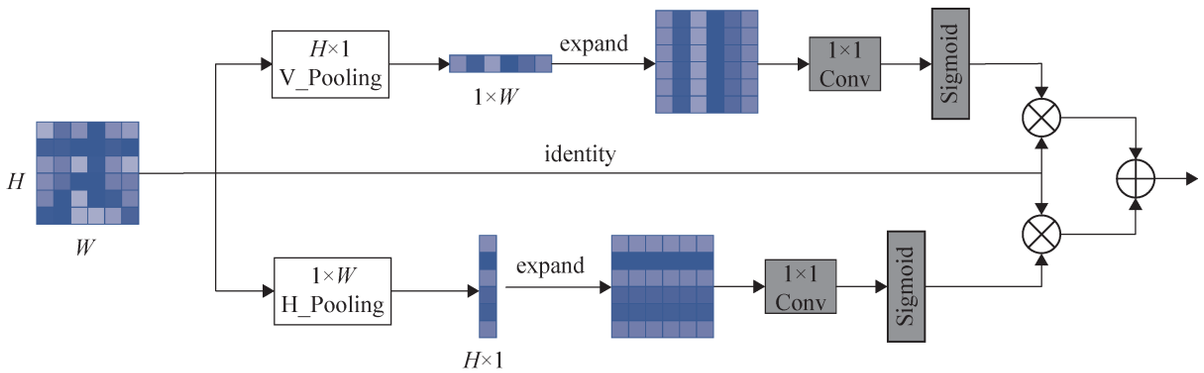


图3　混合条带池化模块
Fig. 3　Mixed strip pooling module

表示为

$$\boldsymbol{y}_{c,j}^{v}=\frac{1}{H}\sum_{0\leqslant i<H}\boldsymbol{x}_{c,i,j} \tag{2}$$

同样地，在水平池化(horizontal pooling，H_pooling)过程中，即对特征图$\boldsymbol{x}$中每一行像素值进行相加再求均值。进行池化核为$1\times W$的水平池化后，输出$\boldsymbol{y}^{h}$为$C\times H\times 1$的列向量，其元素表示为

$$\boldsymbol{y}_{c,i}^{h}=\frac{1}{W}\sum_{0\leqslant j<W}\boldsymbol{x}_{c,i,j} \tag{3}$$

式中：c为通道数；H，W分别为特征图的高和宽；i，j分别为特征图的第i行和第j列。

为了获得包含更有用的全局先验的输出$\boldsymbol{z}$，分别对垂直池化和水平池化的结果进行expand操作得到$\boldsymbol{y}^{h}$和$\boldsymbol{y}^{v}$，并分别与输入特征图结合，最后再相加。输出z为

$$\boldsymbol{y}_1=Scale(\boldsymbol{x},\sigma(f(\boldsymbol{y}^{h}))) \tag{4}$$

$$\boldsymbol{y}_2=Scale(\boldsymbol{x},\sigma(f(\boldsymbol{y}^{v}))) \tag{5}$$

$$\boldsymbol{z}=\boldsymbol{y}_1+\boldsymbol{y}_2 \tag{6}$$

式中：$Scale()$为元素之间相乘；σ为Sigmoid激活函数；f为1×1卷积。

条带池化[32]核呈长条姿态，能有效创建水平或垂直远程关系，进一步帮助搜索全局信息。而且由于条带池化另一个维度较窄，还有助物体细节的捕获。因此MSPM可以收集图像中不同维度的远程上下文，同时兼顾全局和局部信息，使特征更具代表性，更有利后续图像的分割。

1.3.2 深度可分离膨胀卷积

标准空洞卷积能够在保持特征分辨率和像素相对空间不变的前提下，增大卷积感受野。ASPP模块通过多个并行异膨胀系数的扩张卷积，获得不同大小的卷积视野，来追踪多尺度的上下文特征。从图4可见，深度可分离膨胀卷积(DSDConv)首先进行逐通道膨胀卷积(depthwise dDilated convolution，DWDConv)，特征的每个通道只被对应的膨胀卷积核卷积，获取空间维度相关性和局部信息一致性。然后再进行逐点卷积(pointwise convolution，PWConv)，使用n个大小为$1\times 1\times C$的卷积核，对特征图中不同通道的相同空间位置的像素进行卷积，实现不同通道之间的信息交流，最终输出通道数就是逐点卷积的卷积核个数n。深度可分离膨胀卷积融合了膨胀卷积及深度可分离卷积的优点，获取多个较大感受野的同时，还降低了参数量和计算复杂度，提高了模型的高效性。

1.4 注意力机制

在图像语义分割中，特征图的各通道表达不同的特征信息，为充分考虑每个通道之间的重要性关系，本模型在编码区采用了轻量且高效的NAM[33]。NAM是在CBAM的基础上进行改进的

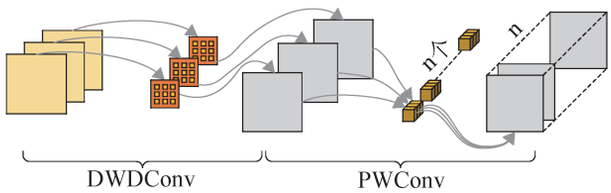


图4 深度可分离膨胀卷积
Fig. 4 Depthwise separable dilated convolution

一种注意力机制，图5为CBAM的结构图，图6和图7为改进后的NAM的两个子模块，分别为通道注意力子模块(channel attention module，CAM)和空间注意力子模块(spatial attention module，SAM)。NAM应用稀疏的权重惩罚并通过训练模型权值参数的方差来评价通道或空间特征的显著程度，规避了SE[34] (squeeze and excitation)和CBAM[35]等一些注意力机制引入全连接的操作，保证性能的同时提高了计算效率。通过不同注意力机制的对比试验，得出NAM的性能最好。

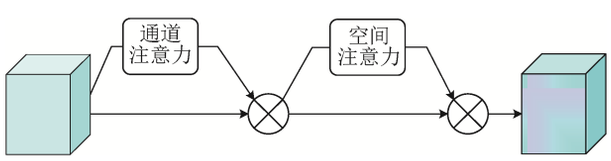


图5 CBAM网络结构
Fig. 5 CBAM network structure

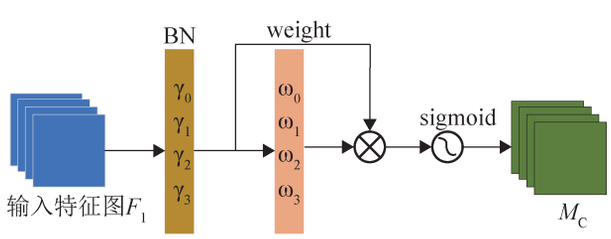


图6 通道注意力模块
Fig. 6 Channel attention module

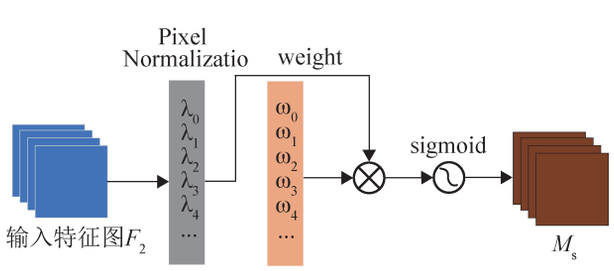


图7 空间注意力模块
Fig. 7 Spatial attention module

NAM引入BN中的比例因子，如式(7)所示，比例因子映射出通道的方差，表示各通道的变化程度，方差越大，代表该通道变化越强烈，即特征信息越显著。反之，该通道的特征较单一。

$$B_{\text{out}}=\text{BN}(B_{\text{in}})=\gamma\frac{B_{\text{in}}-\mu_B}{\sqrt{\sigma_B{}^2+\varepsilon}}+\beta \tag{7}$$

式中：μ_B为小批量样本B的均值；σ_B为B的标准差；γ和β分别为尺度因子和位移；ε为一个超参，一般为一个非常小的数，用来防止分母为0。

CAM的输出特征为

$$\boldsymbol{M}_C=\text{sigmoid}(\boldsymbol{W}_\gamma(\text{BN}_c(\boldsymbol{F}_1))) \tag{8}$$

$$\boldsymbol{W}_\gamma=\frac{\gamma_i}{\sum_{j=0}\gamma_j} \tag{9}$$

式中：γ为每个通道的比例因子；$\boldsymbol{F}_1$为通道注意力模块的输入特征矩阵；BN()为批标准化函数；$\boldsymbol{W}_\gamma$为该通道的权重。

相似地，在空间维度也引入比例因子来计算像素的显著程度，称为像素归一化。SAM的输出特征为

$$\boldsymbol{M}_S=\text{sigmoid}(\boldsymbol{W}_\lambda(\text{BN}_s(\boldsymbol{F}_2))) \tag{10}$$

$$\boldsymbol{W}_\lambda=\frac{\lambda_i}{\sum_{j=0}\lambda_j} \tag{11}$$

式中：λ为比例因子；$\boldsymbol{F}_2$为空间注意力模块等输入牲矩阵；$\boldsymbol{W}_\lambda$为权重。同时将NAM中的SAM应用到SFF模块中，对骨干网络提取出来的两条浅层特征分别使用SAM，可以充分学习两条浅层特征中的空间相关性，从而帮助模型提高图像分割精度。

2 实验

2.1 实验介绍

本论文所完成的实验均是在Linux系统上实现的，系统硬件设备CPU为Intel(R) Xeon(R) Gold 5218 CPU @ 2.30 GHz、GPU为Tesla V100 16 GB。模型基于PyTorch1.10深度学习框架进行搭建,所用

编程语言为python3.8。

本文使用PASCAL VOC 2012公开数据集和WHU遥感建筑物数据集分别进行一系列对比实验和泛化实验，更全面地验证模型的高效性能和泛化性能。首先通过在PASCAL VOC 2012数据集上验证模型的高效性能，该数据集拥有21种类别，其中1 464张图片作为训练数据集，1 449张为验证数据集，还有1 456张图像用来测试。然后通过训练遥感建筑物数据集验证所提网络的泛化性能，该数据集包含3 788张训练图片，948张验证图片。

2.2 评价指标

本文将平均交并比(mean Intersection over Union，mIoU)、参数量作为实验的评价指标。mIoU表示分割网络的整体准确性；参数量反映网络模型的空间复杂度。

mIoU指的是对所有类别像素点的IoU求平均，如图8所示。IoU是通过混淆矩阵取图像中各像素点预测值和真实标签值的交集和并集的比值：

$$mIoU=\frac{1}{n+1}\sum_{i=0}^{n}IoU \tag{12}$$

$$IoU=\frac{TP}{TP+FN+FP}=\frac{p_{ii}}{\sum_{j=0}^{n}p_{ij}+\sum_{j=0}^{n}p_{ji}-p_{ii}} \tag{13}$$

式中：TP为标签为正例，预测亦为正例；FN为标签为反例，而预测为正例；FP为标签为正例，而预测为反例；n为所有像素的类别数目；p_{ii}为将第i类别的像素预测为第i类别的像素总数目；p_{ij}为将第i类别的像素预测为第j类别的像素总数目；p_{ji}为将第j类别的像素预测为第i类别的像素总数目。

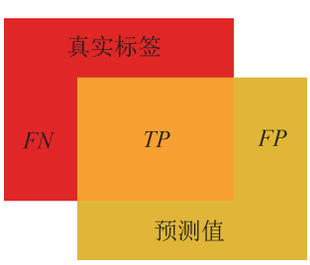


图8　像素预测值和真实标签
Fig. 8　Pixel predictions and ground truth labels

2.3 网络模型改进实验

所提算法主要是在原模型DeepLabv3+的基础上，对主干网络和ASPP模块进行优化，并利用融合多级浅层特征和引入注意力机制，使模型追踪浅层特征和深层语义特征中更加重要的语义信息。通过在PASCAL VOC 2012数据集上进行不同骨干网络、ASPP模块改进策略和不同注意力机制等对比实验，来验证本模型的高效性。

2.3.1 不同骨干网络的对比试验

在语义分割任务中，特征的提取对于最终分割效果起着关键性作用，因此考虑对DeepLabv3+模型的特征提取骨干网络进行更换。在PASCAL VOC 2012数据集上一共进行了5组对比实验，分别测试了不同骨干网络基于改进后的DeepLabv3+算法的性能，用mIOU和参数量评价不同骨干网络的适配性，实验数据见表2所示。

表2　基于不同骨干网络的性能对比
Table 2　Performance comparison based on different backbone networks

实验	骨干网络	*mIoU*/%	参数量/M
1	MobileNetv2	76.90	5.13
2	ResNet101	80.23	56.85
3	Xception	79.71	71.30
4	SwinTransformer	83.68	92.93
5	EfficientNetv2	81.19	55.51

由表2可见，将骨干网络换为EfficientNetv2后，mIoU达到了81.19%，参数量为55.51M。虽然实验1中MobileNetv2拥有更少的参数量，但mIoU比EfficientNetv2少了4.29%。实验2中使用ResNet101的参数量和EfficientNetv2相差不大，但EfficientNetv2的提取精度更高。实验4的mIoU虽然达到了83.68%，但其参数量却远远大于EfficientNetv2。经过综合分析，EfficientNetv2有效优化了模型复杂度和特征提取精度，对于部署到实际场景更具有优势。

2.3.2 ASPP改进实验

本文在ASPP模块中引入了深度可分离膨胀卷积和MSPM，为选出最佳组合的ASPP模块，基于

EfficientNetv2 骨干网络进行 4 组对比试验，其中 SPTT 为预测单张图片的时间。实验对比结果如表 3 所示。

表 3　ASPP 改进实验对比结果
Table 3　Comparison results of ASPP improvement experiments

组别	GAP	MSPM	DSDConv	*mIoU*/%	*SPPT*/ms
1	√			78.64	59.84
2	√	√		79.85	65.10
3		√		79.99	51.63
4		√	√	79.74	43.46

通过表 3 结果可知，联合使用 MSPM 和原 ASPP 模块的 GAP，虽然 mIoU 有所提升，但预测时间也有所下降。从组别 3 发现单独使用 MSPM，不仅 mIoU 进一步得到了提升，而且预测时间也得到了降低。最后通过引入深度可分离膨胀卷积，以牺牲 0.25% 的精度，使预测时间得到显著提升。故本文选择组别 4 的组合方式作为 N-ASPP 模块。

2.3.3 不同注意力机制的对比实验

注意力机制本质是通过计算相应的权重值，让卷积神经网络识别出需要重点关注的有用特征向量，忽略不重要的特征信息。从而在避免无用特征干扰拟合结果的同时，还对运算速度有一定的改善。本文基于 EfficientNetv2 骨干网络和原版 ASPP 模块，引入 4 个不同的注意力机制进行对比。对比结果如表 4 所示。

表 4　不同注意力机制性能对比
Table 4　Performance comparison of different attention mechanisms

实验	Backbone	Attention	*mIoU*/%	*FPS*/(frame/s)
1	EfficientNetv2	ECA	79.58	16.71
2	EfficientNetv2	SE	79.43	15.79
3	EfficientNetv2	CBAM	80.20	14.32
4	EfficientNetv2	NAM	80.25	16.63

由表 4 可知，NAM 相比 SE 和 CBAM，性能均有明显改善。相比 ECA,虽然速度慢了一点，但 mIoU 提升了 0.63%。综合分析，本文选择了性能更强大的 NAM，帮助模型更好地完成分割任务。

2.3.4 不同模块的消融实验

为证明 N-ASPP 模块、NAM、浅层特征融合 (shallow feature fusion，SFF)等各模块的有效性，利用控制变量法设计了 5 组消融实验，以 mIoU 和 SPTT 作为实验评价指标，实验数据如表 5 所示。

表 5　不同模块的消融实验结果
Table 5　Ablation experiment results of different modules

组别	Efficient-Netv2	N-ASPP	NAM	SFF	*mIoU*/%	*SPPT*/ms
1	√				78.94	59.84
2	√	√			79.74	43.46
3	√		√		80.25	60.13
4	√	√	√		80.80	44.40
5	√	√	√	√	81.19	44.92

通过比较表 5 的组别 1，2 可知，将 ASPP 的全局平均池化换为混合条带池化模块并引入深度可分离膨胀卷积，不仅 mIoU 提升了 0.8%，预测时间也大幅度减少了。组别 1，3 进行对比可看出，对编码区的高级语义特征图使用 NAM，虽然预测时间会有所增加，但 mIoU 有明显提升。通过组别 3，4 的比较可得出，同时使用改进后的 N-ASPP 模块和 NAM，mIoU 和预测时间均继续有所改善。通过组别 5 可看出，对骨干网络进行 SFF 作为解码区的输入，仅仅以损失较少预测时间为代价，mIoU 提高到了 81.19%。综合分析表 5，验证了所提模块均起到了一定作用。

2.3.5 不同模型之间的对比实验

将本文提出的模型与 FCN、SegNet、PSPNet、DeepLabv3+等经典语义分割模型在 PASCAL VOC 2012 数据集上进行对比实验，对比结果如表 6 所示，又通过表 7 和图 9 展现了 DeepLabv3+改进前后的详细对比信息。

由表 6，7 可看出，相较于其他的经典语义分割模型，本文所提的网络模型取得了更优异的分割效果，不仅在精度方面相比原 DeepLabv3+提升了 2.88%，参数量也减少了 21%，从而能更好地满足实际场景。

DOI: 10.16182/j.issn1004731x.joss.22-0690

表6 不同模型在PASCAL VOC 2012上的性能对比结果
Table 6 Performance comparison results of different models on PASCAL VOC 2012

算法	骨干网络	*mIoU*/%
FCN-8s	VGG-16	68.17
SegNet	VGG-16	69.31
PSPNet	ResNet101	73.66
DeepLabv3+	MobileNetv2	74.89
DeepLabv3+	Xception	78.31
本文算法	EfficientNetv2	81.19

表7 DeepLabv3+改进前后在PASCAL VOC 2012数据集上对比结果
Table 7 Comparison of results on the PASCAL VOC 2012 dataset before and after DeepLabv3+ improvement

算法	骨干网络	*mIoU*/%	参数量/M
原DeepLabv3+	Xception	78.31	70.25
本文算法	EfficientNetv2	81.19	55.51

为了更直观地对比DeepLabv3+改进前后的分割性能，对5组分割图进行可视化对比分析，如图10所示。从第一行中发现，原网络没有将猫和人正确分割出来，而本文模型在猫腿部位得到了有效的改进；第二行原网络在处理牛后蹄时存在漏分割现象，而且人头部、人腿等细节部位处理的也比较粗糙，本文模型在以上问题中都有一定的完善；通过比较第三行分割马头和第四行分割人的跳跃动作，明显发现改进后的网络在处理马耳和人腿、胳膊时，更加细腻完整；通过第五行的对比，可看出原模型存在漏分割、分割不连续、细节模糊等问题，本文模型在分割大狗尾巴时，明显更加连续，而且大狗后肢和小狗四肢都得了更细化的分割效果。

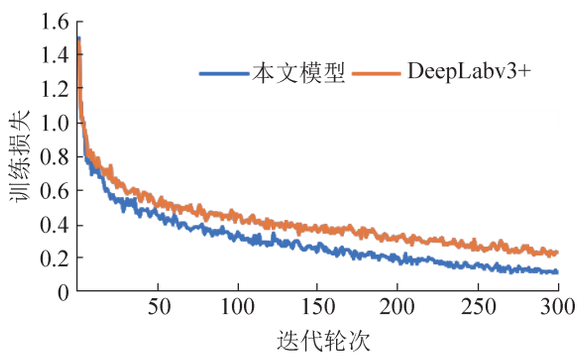


图9 不同模型的训练损失曲线
Fig. 9 Training loss curves for different models

(a) 原图像 (b) 真实标签 (c) DeepLabv3+分割图 (d) 本文模型分割图

图10 PASCAL VOC 2012数据集分割结果对比图
Fig. 10 Comparison of segmentation results on PASCAL VOC 2012 dataset

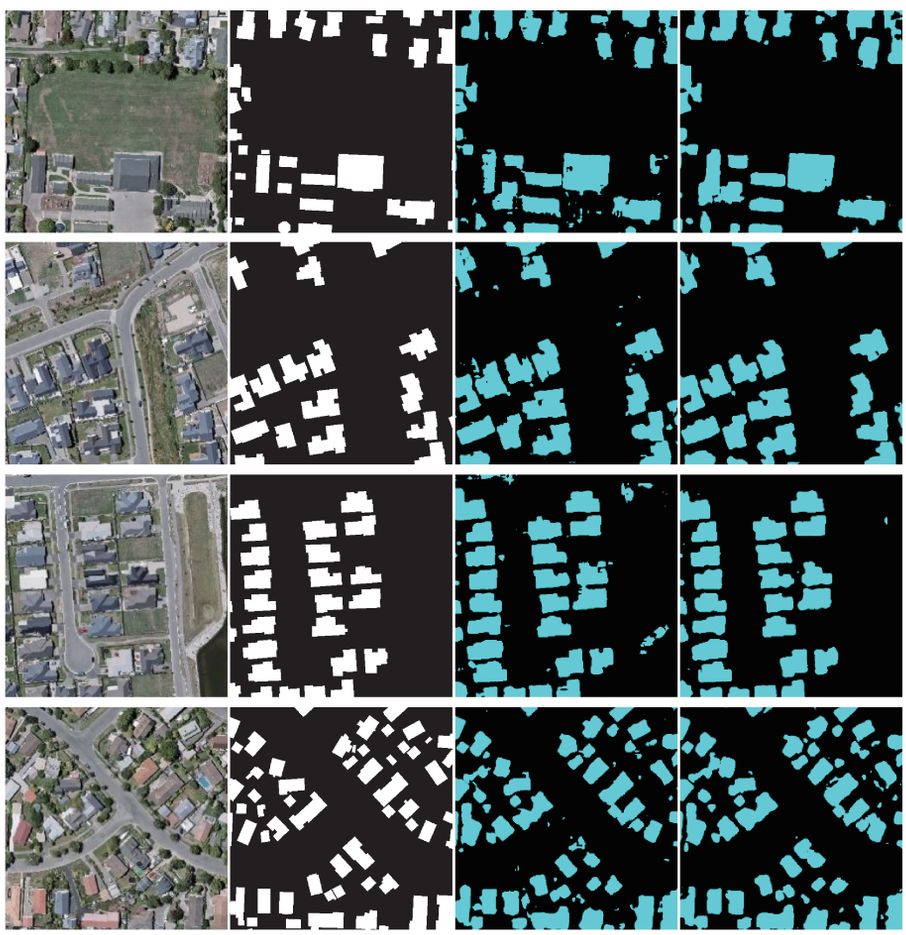
2.4 泛化实验

为了验证本文的改进算法具有很好的适用性，现将 DeepLabv3+原网络和本文网络在 WHU 遥感数据集上做泛化对比实验，对比结果见表 8。

表 8 DeepLabv3+改进前后在 WHU 数据集上对比结果
Table 8 Comparison of results on the WHU dataset before and after DeepLabv3+ improvement

算法	骨干网络	*mIoU*/%	参数量/M
原 DeepLabv3+	Xception	85.41	70.25
本文算法	EfficientNetv2	87.92	55.51

从表 8 看出，本文模型相较于原 DeepLabv3+模型在 WHU 遥感数据集上，同样在均交并比 mIoU 和参数量方面都有一定提升。通过比较 DeepLabv3+网络改进前后的模型的可视化分割结果图，如图 11 所示。可明显看出，对于遥感图像中屋顶边缘的分割，本文模型相对更加的平滑和完整。而且原 DeepLabv3+网络存在大量误分割、分割不连续的地方，优化后的网络均有了大幅度的改善。因此本文所提模型在 WHU 遥感数据集上依旧拥有更好的分割效果，在减少参数量，降低模型复杂度的同时，细化了目标的分割准确度，同时也说明了本文模型具有一定的鲁棒性和泛化能力。



(a) 原图像 (b) 真实标签 (c) DeepLabv3+分割图 (d) 本文模型分割图

图 11 WHU 数据集分割结果对比图
Fig. 11 Segmentation result comparison on WHU dataset

DOI: 10.16182/j.issn1004731x.joss.22-0690

3 结论

本文提出了一种基于改进的DeepLabv3+图像语义分割模型。将骨干特征提取网络改为EfficientNetv2网络，并提取主干网络中的多条浅层特征作为解码器的一条输入支路。在提出的N-ASPP模块中将全局平均池化换为条带池化模块，并通过深度可分离膨胀卷积融合了扩张卷积和深度可分离卷积两者的优点。最后分别对浅层细节特征和高级语义特征使用不同的注意力机制，有效提高了通道和空间维度获取信息的能力。通过对比和泛化实验发现，优化后的模型在降低模型复杂度的同时，明显改善了漏分割、误分割、分割不连续、边缘细节模糊等问题，有效提高了分割性能。但是本模型在实时性方面还有待更进一步的提高，接下来工作的重点会对精度、复杂度、推理速度综合考虑，使模型同时兼顾高精度、轻量化和强实效性。

参考文献：

[1] Wang Lei, Wu Jiaji, Liu Xunyu, et al. Semantic Segmentation of Large-scale Point Clouds Based on Dilated Nearest Neighbors Graph[J]. Complex & Intelligent Systems, 2022, 8(5): 3833-3845.

[2] 田萱, 王亮, 丁琪. 基于深度学习的图像语义分割方法综述[J]. 软件学报, 2019, 30(2): 440-468.
Tian Xuan, Wang Liang, Ding Qi. Review of Image Semantic Segmentation Based on Deep Learning[J]. Journal of Software, 2019, 30(2): 440-468.

[3] Asgari Taghanaki S, Abhishek K, Cohen J P, et al. Deep Semantic Segmentation of Natural and Medical Images: A Review[J]. Artificial Intelligence Review, 2021, 54(1): 137-178.

[4] Yuan Xiaohui, Shi Jianfang, Gu Lichuan. A Review of Deep Learning Methods for Semantic Segmentation of Remote Sensing Imagery[J]. Expert Systems with Applications, 2021, 169: 114417.

[5] 王奕清. 基于计算机视觉的卫星云图反演降水量方法研究[D]. 成都: 电子科技大学, 2021.
Wang Yiqing. A Computer Vision Method for Precipitation Inversion With Satellite Cloud Images[D]. Chengdu: University of Electronic Science and Technology of China, 2021.

[6] Ivanovs M, Ozols K, Dobrajs A, et al. Improving Semantic Segmentation of Urban Scenes for Self-driving Cars with Synthetic Images[J]. Sensors, 2022, 22(6): 2252.

[7] Kontschieder P, Samuel Rota Bulò, Bischof H, et al. Structured Class-labels in Random Forests for Semantic Image Labelling[C]//2011 International Conference on Computer Vision. Piscataway, NJ, USA: IEEE, 2011: 2190-2197.

[8] Martijn van den Heuvel, Mandl R, Hulshoff Pol H. Normalized Cut Group Clustering of Resting-state FMRI Data[J]. PLoS One, 2008, 3(4): e2001.

[9] Cherkassky V, Ma Yunqian. Practical Selection of SVM Parameters and Noise Estimation for SVM Regression [J]. Neural Networks, 2004, 17(1): 113-126.

[10] Hu Yaosi, Chen Zhenzhong, Lin Weiyao. RGB-D Semantic Segmentation: A Review[C]//2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Piscataway, NJ, USA: IEEE, 2018: 1-6.

[11] Kamilaris A, Francesc X Prenafeta-Boldú. Deep Learning in Agriculture: A Survey[J]. Computers and Electronics in Agriculture, 2018, 147: 70-90.

[12] 刘瑞军, 王向上, 张晨, 等. 基于深度学习的视觉SLAM综述[J]. 系统仿真学报, 2020, 32(7): 1244-1256.
Liu Ruijun, Wang Xiangshang, Zhang Chen, et al. A Survey on Visual SLAM Based on Deep Learning[J]. Journal of System Simulation, 2020, 32(7): 1244-1256.

[13] 罗荣, 王亮, 肖玉杰. 深度学习技术应用现状分析与发展趋势研究[J]. 计算机教育, 2019(10): 19-22.

[14] Yu Changqian, Wang Jingbo, Peng Chao, et al. BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation[C]//Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 334-349.

[15] Zhang Fan, Chen Yanqin, Li Zhihang, et al. ACFNet: Attentional Class Feature Network for Semantic Segmentation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2019: 6797-6806.

[16] Wu Tianyi, Tang Sheng, Zhang Rui, et al. CGNet: A Light-weight Context Guided Network for Semantic Segmentation[J]. IEEE Transactions on Image Processing, 2021, 30: 1169-1179.

[17] Zhao Yaochi, Liu Shiguang, Hu Zhuhua. Focal Learning on Stranger for Imbalanced Image Segmentation[J]. IET Image Processing, 2022, 16(5): 1305-1323.

[18] Zhao Yaochi, Liu Shiguang, Hu Zhuhua. Dynamically Balancing Class Losses in Imbalanced Deep Learning[J]. Electronics Letters, 2022, 58(5): 203-206.

[19] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2015: 3431-3440.

[20] Guo Yanming, Liu Yu, Georgiou T, et al. A Review of Semantic Segmentation Using Deep Neural Networks[J]. International Journal of Multimedia Information Retrieval, 2018, 7(2): 87-93.

[21] 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.

[22] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]// Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. Cham: Springer International Publishing, 2015: 234-241.

[23] Edgar Schönfeld, Schiele B, Khoreva A. A U-net Based Discriminator for Generative Adversarial Networks[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 8204-8213.

[24] Jaeger P F, Kohl S A A, Bickelhaupt S, et al. Retina U-net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection [C]//Proceedings of the Machine Learning for Health NeurIPS Workshop. Chia Laguna Resort, Sardinia, Italy: PMLR, 2020: 171-183.

[25] Chen L C, Papandreou G, Kokkinos I, et al. Semantic Image Segmentation With Deep Convolutional Nets and Fully Connected CRFs[EB/OL]. (2016-06-07) [2022-05-30]. https://arxiv.org/abs/1412.7062.

[26] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-scale Image Recognition[EB/OL]. (2015-04-10) [2022-05-30]. https://arxiv.org/abs/1409.1556.

[27] 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.

[28] Chen L C, Papandreou G, Schroff F, et al. Rethinking Atrous Convolution for Semantic Image Segmentation [EB/OL]. (2017-12-05) [2022-05-30]. https://arxiv. org/abs/1706.05587.

[29] Chen L C, Zhu Yukun, Papandreou G, et al. Encoder-decoder With Atrous Separable Convolution for Semantic Image Segmentation[C]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 833-851.

[30] Tan Mingxing, Le Q. EfficientNetV2: Smaller Models and Faster Training[C]//Proceedings of the 38th International Conference on Machine Learning. Chia Laguna Resort, Sardinia, Italy: PMLR, 2021: 10096-10106.

[31] Tan Mingxing, Le Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks[C]// Proceedings of the 36th International Conference on Machine Learning. Chia Laguna Resort, Sardinia, Italy: PMLR, 2019: 6105-6114.

[32] Hou Qibin, Zhang Li, Cheng Mingming, et al. Strip Pooling: Rethinking Spatial Pooling for Scene Parsing [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 4002-4011.

[33] Liu Yichao, Shao Zongru, Teng Yueyang, et al. NAM: Normalization-based Attention Module[EB/OL]. (2021-11-24) [2022-05-30]. http://arxiv.org/abs/2111.12419.

[34] Hu Jie, Shen Li, Sun Gang. Squeeze-and-excitation Networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 7132-7141.

[35] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional Block Attention Module[C]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.

DOI: 10.16182/j.issn1004731x.joss.22-0690