doi:10.6041/j.issn.1000-1298.2023.09.030

基于多头自注意力机制的茶叶采摘点语义分割算法

宋　彦[1,2]　杨　帅[1]　郑子秋[1]　宁井铭[3]

(1.安徽农业大学工学院, 合肥 230036; 2.安徽省智能农机装备工程实验室, 合肥 230036;
3.茶树生物学与资源利用国家重点实验室, 合肥 230036)

摘要: 茶叶采摘点定位是茶叶选择性采摘的关键技术之一,在茶树采摘场景中,存在采摘点尺度小、背景干扰大、光照情况复杂等问题,导致准确分割茶叶采摘点成为难题。本研究针对茶园场景下采摘点精确分割问题,构建了一种基于多头自注意力机制结合多尺度特征融合的语义分割算法——RMHSA-NeXt。首先使用 ConvNeXt 卷积神经网络提取图像特征;其次构造基于残差和多头自注意力机制的注意力模块,将模型注意力集中于分割目标,增强重要特征的表达;再次通过多尺度结构(Atrous spatial pyramid pooling, ASPP)将不同尺度的特征进行融合,在其中针对采摘点特性,在融合过程中使用条状池化(Strip pooling),减少无用特征的获取;最后通过卷积以及上采样等操作完成信息的解码,得出分割结果。试验表明,茶园环境下该模型可以对采摘点进行有效分割,模型的像素准确率达到 75.20%,平均区域重合度为 70.78%,运行速度达到 8.97 f/s。基于相同测试集将本文模型与 HRNet V2、EfficientUNet++、DeeplabV3+、BiSeNet V2 模型进行对比,结果表明相比于其他模型同时具有准确性高、推理速度快、参数量小等优点,能够较好地平衡精度与速度指标。本文的研究成果可以为精准定位茶叶采摘点提供有效可靠的参考。

关键词: 茶叶采摘点; 语义分割; 注意力机制; 多尺度结构

中图分类号: TP391.9　　**文献标识码**: A　　**文章编号**: 1000-1298(2023)09-0297-09　　**OSID**:

Semantic Segmentation Algorithm Based Multi-headed Self-attention for Tea Picking Points

SONG Yan[1,2]　YANG Shuai[1]　ZHENG Ziqiu[1]　NING Jingming[3]

(1. *School of Engineering*, *Anhui Agricultural University*, *Hefei* 230036, *China*
2. *Anhui Provincial Engineering Laboratory of Intelligent Agricultural Machinery*, *Hefei* 230036, *China*
3. *State Key Laboratory of Tea Plant Biology and Utilization*, *Hefei* 230036, *China*)

Abstract: Tea picking point localization is one of the key technologies for selective tea picking. In the tea tree picking scenario, there are problems such as small scale of picking points, large background interference and complex lighting conditions, which lead to the problem of accurate segmentation of tea picking points. A semantic segmentation model based on multi-headed self-attentive mechanism combined with multi-scale feature fusion, RMHSA-NeXt, was constructed for the accurate segmentation of picking points in tea garden scenes. The attention module based on residuals and multi-headed self-attention mechanism was constructed to focus the model's attention on the segmentation target and enhance the representation of important features. The features at different scales were fused by multi-scale structure (atrous spatial pyramid pooling, ASPP), in which strip pooling was used in the fusion process for the characteristics of picking points to reduce the useless. Finally, the information was decoded by convolution and upsampling, and the segmentation results were obtained. The experiment results showed that the model can segment the picking points effectively in the tea garden environment, and the pixel accuracy of the model reached 75.20%, the average region overlap was 70.78%, and the running speed reached 8.97 f/s. The results showed that the model had the advantages of high accuracy, fast inference speed and small number of parameters, which can balance the accuracy and speed indexes well compared with other models. The research results can provide an effective and reliable reference for pinpointing tea picking points.

Key words: tea picking point; semantic segmentation; attention mechanism; multi-scale structure

收稿日期: 2023-02-24　修回日期: 2023-03-31

基金项目: 国家重点研发计划项目(2021YFD1601102)

作者简介: 宋彦(1983—),男,副教授,博士,主要从事智能农业技术与装备研究,E-mail: songyan@ahau.edu.cn

通信作者: 宁井铭(1973—),男,教授,博士,主要从事茶叶加工与品质分析研究,E-mail: ningjm@ahau.edu.cn

0 引言

目前我国名优茶采摘仍以人工为主,在春茶加工季节,需要大量的人力资源,成本高。而现有的采摘机械[1-3],嫩芽和老叶不加区分的“一刀切”,原料质量得不到保证[4]。但随着机器视觉技术的兴起,为解决茶叶选择性采摘问题提供了一种新的途径,茶芽的识别和可采摘区域定位则是实现选择性自动化采摘的前提。

早期的研究多基于图像的阈值分割方法进行茶芽的识别,如文献[5-7]中采用单一色彩阈值或联合色彩阈值的方法从图像中识别茶芽目标。然而,茶园环境具有非结构特性,包括不受控制的光照、茶叶嫩芽和老叶之间的高度相似性等,在这些条件下,阈值分割法可能面临鲁棒性不足的问题。

近年来,卷积神经网络得益于较为优秀的特征提取能力,对复杂背景的目标定位表现出较好的鲁棒性,受光照、背景影响较小。王琨等[8]、吕军等[9]选取茶叶的部分颜色和形状特征作为特征集,输入卷积神经网络模型,通过训练最终使得模型能够识别茶叶嫩芽。孙肖肖等[10]使用改进的 YOLO v3 识别茶叶,经测试平均准确率达到了 84.2%,召回率达到了 82%。吕军等[11]使用结合图像预处理的改进 YOLO v5 模型,解决由于茶园的光照变化茶芽难以准确检测的问题。由前述可知,基于卷积神经网络的目标检测技术为茶芽的识别提供了可行的技术途径。

一旦图像中的茶芽目标被定位,准确采摘的下一步就是分割可采摘点或可采摘区域。然而,由于采摘点区域不尽相同,形态大多不规则,而茶叶选择性采摘又需要准确的像素坐标,此时再使用目标检测不再合适。语义分割是机器视觉中的典型任务,可用于像素级的对象检测,能够准确定位茶叶采摘点的区域。该方法已被应用于需精准定位目标像素坐标的相关农业领域,如作物采摘。LI 等[12]使用 DeepLabv3 + 模型[13]分割龙眼果串的主果枝,分割结果的像素准确率达到了 94.52%。ZHU 等[14]为自动准确地识别甜椒采摘区域,使用自建的全分辨率残差网络分割彩椒图像,在测试集的像素准确率达到了 97.94%。YU 等[15]使用 Mask R-CNN 模型分割草莓图像中的可采摘区域,100 幅测试图像的检测结果显示,平均检测准确率为 95.78%,召回率为 95.41%,分割的平均区域重合度为 89.85%。由于语义分割结果能精准获取采摘区域,极少包含无关信息,在采摘点定位领域,可以给予后续处理器有效的目标信息,简化下一步需要数据过滤工作。但上述研究的试验环境,部分目标与背景相似程度较小,如草莓果实与背景的颜色特征差异明显;部分研究工作在室内环境或实验室环境开展,可能难以直接应用于复杂的茶园环境。

本文针对茶芽采摘点分割时面临的目标尺度小、背景复杂等问题,提出 RMHSA-NeXt 语义分割模型,其特点在于使用残差多头自注意力模块和结合条形池化的 ASPP[13]模块,能够提高对关键特征的关注程度,并降低对背景与目标的干扰,达到准确分割采摘点的目的,以期为自然环境下茶叶机械选择性采摘提供可靠识别依据。

1 材料和方法

1.1 茶叶采摘点数据来源与数据集构建

茶叶图像数据采集于六安市金寨县青山镇抱儿村,采集时间为 2022 年 4 月上旬到 10 月下旬。图像采集设备为 STEREOLABS 公司生产的 ZED 双目相机,为了使图像数据更加贴合采茶机械的工作状况,采用视频录制的模式,视频分辨率为 1 920 像素 × 1 080 像素。所有数据采集完成后通过 STEREOLABS 公司提供的视频转换软件,将视频数据逐帧转换为图像数据,筛选后,得到原始数据集图像数量为 3 192 幅。

为了开展茶叶采摘点语义分割模型研究,首先采用目标检测模型对原始数据集做茶芽目标检测,经过对目标检测结果的筛选,最终采摘点数据集图像数量为 26 793 幅,本文使用 Labelme 软件标注出茶叶可采摘区域,按照 9∶1 的比例将数据划分为训练集和测试集。

分析采摘点数据集后,发现本文的分割目标的特点在于:茶叶采摘点目标较小(普遍尺度约 100 像素)、光照强度变化范围大、采摘点的背景多为茶叶,纹理和色彩差距不明显、采摘点形状普遍为长条形等,图 1 为茶叶目标检测与分割的场景分析图。



(a) 目标较小 (b) 采摘点长条形 (c) 光照情况复杂 (d) 背景复杂

图 1 茶叶目标检测与分割的场景分析

Fig. 1 Scenario analysis of tea object detection and segmentation

1.2　RMHSA－NeXt 语义分割模型

通用的语义分割模型由编码器和解码器构成，编码器进行特征提取，解码器利用反卷积或者上采样运算将编码器输出的低分辨率特征映射到高分辨率像素空间，得到密集的像素预测分类[16]。为解决茶叶采摘点较小、光照变化范围大、采摘点与老叶较为相似等问题，本文使用编码-解码架构设计提出了一种新的语义分割模型——RMHSA－NeXt，其网络结构如图 2 所示。为了准确分割茶芽采摘点，该模型提出了以下方法：①在模型的编码阶段，采用 ConvNeXt 作为特征提取单元，其可以有效地筛除无用特征，增强有效特征的表达，减少茶园环境中的光照变化和茶芽采摘点与背景相似度较大的影响。②提出残差多头自注意力模块（Residual multi-head self-attention，RMHSA），其根据茶芽特征相关性为各特征分配权重，将模型注意力集中于茶叶采摘点目标，减少不相关的老叶背景的干扰，进一步强化模型的特征提取能力。③设计结合条形池化的 ASPP 结构，由于茶芽采摘点多为长条形，故将常用的正方形池化改变为条形池化，便于抑制背景干扰，且由于其内部的多尺度融合机制，有望解决茶芽采摘点小、目标特征信息不足的问题。

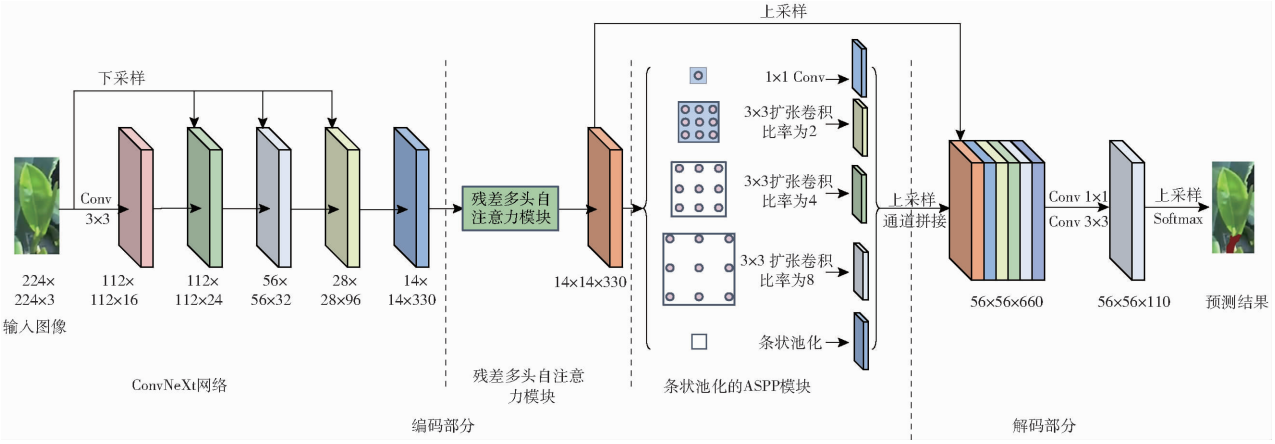

图 2　RMHSA－NeXt 语义分割模型

Fig. 2　RMHSA－NeXt semantic segmentation model

1.2.1　特征提取网络

近年来，Swin Transformer[17] 越来越多地被应用于计算机视觉领域，并且展现了优异的性能。ConvNeXt 吸收 Swin Transformer 的优点，通过结合 Swin Transformer 的层结构、倒置瓶颈和深度可分离卷积（Depthwise Convolution）等技巧，进一步提高了模型的特征提取能力，因此，本文的语义分割模型使用 ConvNeXt 作为初始特征的提取单元。

图 3 是 ConvNeXt 网络结构图，其借鉴了 ResNet 的设计思想，在 ResNet 传统结构上，增加了下采样方法，该网络主要由 ConvNeXt block 组成。图 4 是 ConvNeXt block 的结构图，首先将初始特征与提取后的特征相融合；其次在特征提取过程中使用深度可分离卷积，只在每个通道上进行空间信息的交互，减少了参数，同时激活函数使用 GeLU，相较于 ReLU 更加平滑和连续，也有更高的收敛几率；而后使用层归一化（Layer normalization），将这一中间层的神经元参数进行归一化，对神经网络中隐藏层的输入进行归一化，从而使得网络更容易训练[18]；最后使用 Drop path 随机地将深度学习中的多分支结构删除，即让某些神经元失效，添加正则化能力，防止模型过拟合。

1.2.2　残差多头自注意力模块

注意力机制主要根据上下文内容或像素间的相关性快速提取数据或者图像中的重要特征。多头自注意力模块[19]（Multi-head self-attention，MHSA）是多个自注意力模块[20]平行堆叠而来，其可以根据采摘点目标相关性为各特征动态地分配权重，将网络注意力集中于茶叶采摘点目标，减少茶园中不相关背景的干扰，提高网络特征提取性能。

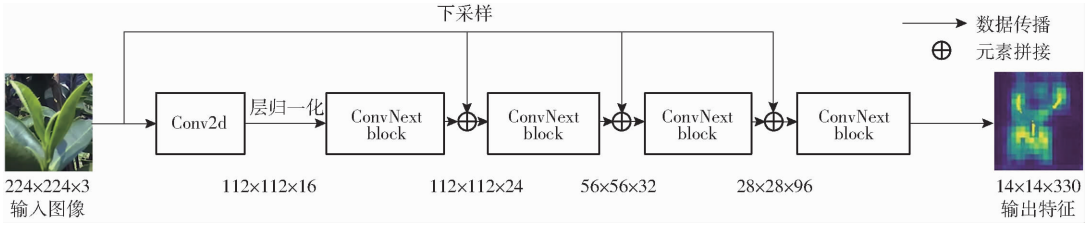

图 3　ConvNeXt 网络结构图

Fig. 3　ConvNeXt network structure diagram

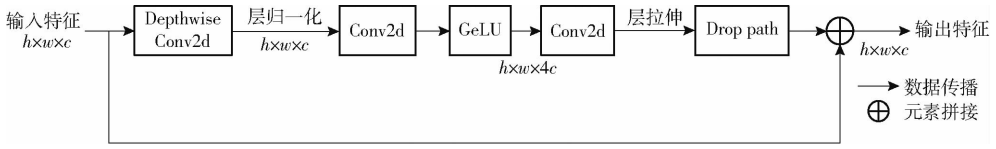

图 4 ConvNeXt block 网络结构图

Fig. 4 ConvNeXt block network structure diagram

本文基于残差结构和多头自注意力机制,构建了如图 5 所示的 RMHSA 结构,其中多头自注意力模块代替了原本残差结构中的卷积模块,同时使用两次的跳连接,让未被处理过的原始特征与被多头自注意力模块处理过的特征充分融合。在图 5 中多头自注意力机制部分是由多个自注意力机制平行计算,最后拼接而来,自注意力机制作用在于减少了对外部信息的依赖,更有利于捕捉茶芽采摘点数据或特征的内部相关性,图 6 展示了自注意力机制的结构。

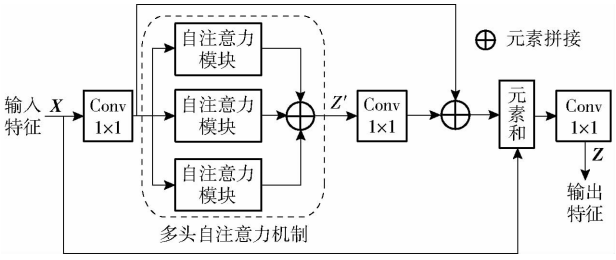

图 5 残差多头自注意力模块结构示意图

Fig. 5 Schematic of structure of residual multi-head self-attentive module

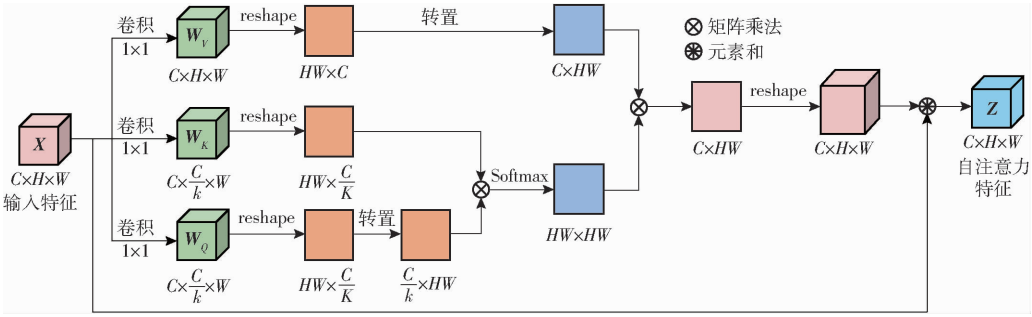

图 6 自注意力机制结构示意图

Fig. 6 Schematic of structure of self-attentive mechanism

模块的输出 $\boldsymbol{Z}$ 为基于查询向量 $\boldsymbol{W}_Q$、键向量 $\boldsymbol{W}_K$、值向量 $\boldsymbol{W}_V$ 的加权和,计算式为

$$\boldsymbol{Z}=\mathrm{Softmax}(\boldsymbol{W}_K\boldsymbol{W}_Q)\boldsymbol{W}_V+\boldsymbol{X} \tag{1}$$

式中 $\boldsymbol{X}$——输入特征

自注意力模块的输入通过线性变换得到 $\boldsymbol{W}_Q$、$\boldsymbol{W}_K$、$\boldsymbol{W}_V$。其中 $\boldsymbol{W}_Q$ 的目的是计算当前位置与其他位置之间的相似度;$\boldsymbol{W}_K$ 用于计算当前位置和与其他位置之间的相似度,在运算过程中,$\boldsymbol{W}_K$ 通常会与 $\boldsymbol{W}_Q$ 进行矩阵乘法,得到每个位置向量的权值。而 $\boldsymbol{W}_V$ 与 $\boldsymbol{W}_Q$、$\boldsymbol{W}_K$ 计算出的权值矩阵进行矩阵乘法,得到特征的加权平均结果。

1.2.3 结合条形池化的 ASPP

多尺度结构可以通过不同尺度的特征提取,帮助网络获得不同感受野下的图像信息,以增强模型对小目标的感知能力。因此,多尺度信息捕获能力对于解决茶叶采摘点图像有效信息过少具有重要意义,本文选择了 ASPP 结构作为模型中的多尺度结构。ASPP 通常由 1 个 1×1 卷积、3 个不同采样率的 3×3 扩张卷积和 1 个空间池化组成。在这种结构中,通过设置不同的采样率可以得到不同比例的特征图。

由于初始 ASPP 中的空间池化的采样窗口为正方形,而本文中的茶叶采摘区域大部分为长条形,此时正方形窗口不可避免地会包含其他不相关区域的干扰信息。因此,本文采用了条形池化结构,由于长条形采样窗口采样时会减少不相关信息的获取,从而降低无关信息的干扰[21]。改进后的 ASPP 结构如图 7 所示,而条形池化计算过程如图 8 所示。

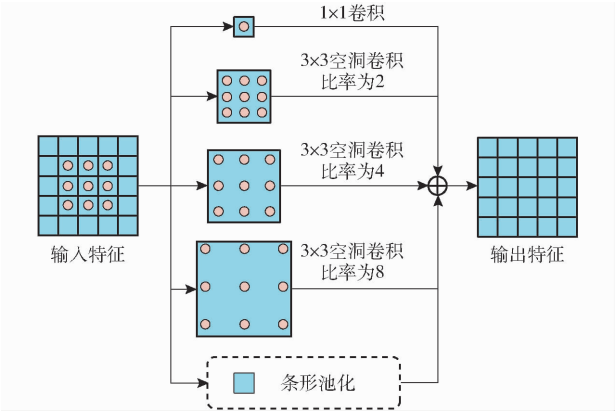

图 7 改进后的 ASPP 结构示意图

Fig. 7 Schematic of improved ASPP structure

在条形池化时,首先将输入特征图进行水平和竖直条形池化后变为 $H\times1$ 和 $1\times W$ 两个特征图,随后经过卷积核为 3 的一维卷积对 2 个特征图分别沿

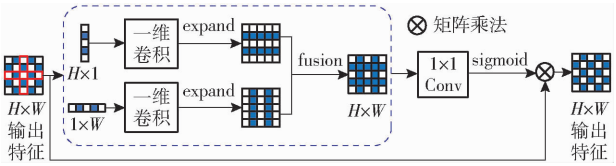

图 8　条形池化示意图

Fig. 8　Schematic of stripe pooling

着左右和上下进行扩容，扩容后两个特征图尺寸相同，扩容后的两特征图对应相同位置求和得到$H \times W$的特征图，之后通过 1 × 1 的卷积与 sigmoid 处理后与原输入图对应像素相乘得到了条状池化输出结果，在本文中 H 和 W 均为 5。

1.2.4　解码部分

在编解码结构中，编码器的作用是提取特征，而解码器的作用是将经过编码器处理后的特征进行采样和映射，最终实现逐像素分类。在本文模型中，首先会将从改进 ASPP 中获得的多尺度特征和从残差多头自注意力机制中获得的特征进行通道拼接，在此过程中，部分多尺度结构会被上采样以保证特征图大小一致；而后会经过一次 1 × 1 卷积和 3 × 3 处理，降低特征图的维度，筛选特征；最后使用上采样将图像恢复到与输入图像一样的大小，再使用 Softmax 函数进行结果预测，确定每一个特征图的类别。

2　实验结果与分析

2.1　实验平台

计算机配置 CPU 型号为 Intel Core i7 - 9700 CPU，GPU 型号为 NVIDIA GTX2080Ti，内存为 32 GB，1 TB 固态硬盘。计算机操作系统为 Ubuntu 20.04，配置 PyTorch 深度学习框架用于所有模型的训练和测试，模型评价等程序均在 Python 语言环境下编写。

2.2　模型评价指标

本文语义分割模型采用准确率、检测速度、参数量等指标进行性能评价。准确率是根据网络模型预测图像与人工标注图像之间的像素误差计算得到，设语义类别总数为 $k+1$（k 个目标类与 1 个背景类），P_{ii}表示属于第 i 类且被预测为第 i 类的像素数，P_{ij}表示属于第 i 类却被预测为第 j 类的像素数，在本文中 $k+1$ 为 2，即采摘区域和非采摘区域。

像素准确率（Pixel accuracy，PA）为正确预测像素数量与图像像素总量的比值，计算公式为

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \times 100\% \tag{2}$$

平均区域重合度（Mean intersection over union，MIoU）为每类预测像素数量与真实像素数量交集与并集比值，然后取所有类别的平均值。平均区域重合度反映了预测结果与图像真实结果的重合程度，是语义分割模型常采用的准确率度量标准，计算公式为

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \times 100\% \tag{3}$$

检测速度（每秒检测帧数，FPS）用来表明每秒能检测的数据帧数，其数值越高表明训练好的模型检测速度越快。

2.3　结果分析

2.3.1　消融实验

为测试本文所提出的语义分割算法的有效性，设计消融实验分析各功能模块对模型性能的影响。构建基础模型，由 ConvNeXt 网络和解码部分组成。在基础模型上逐步加入残差多头自注意力机制、改进的 ASPP 等结构构成基础模型 + RMHSA、基础模型 + ASPP、基础模型 + 改进 ASPP 以及基础模型 + RMHSA + 改进 ASPP 模型等。通过像素准确率、平均区域重合度、参数量和检测速度对模型性能进行分析，表 1 为测试集在上述 5 种模型的运行结果。

表 1　不同模型的分割结果

Tab. 1　Segmentation results of different functional units

模型	像素准确率/%	平均区域重合度/%	检测速度/（f·s^{-1}）	参数量
基础模型	39.46	32.88	17.19	1.06×10^{6}
基础模型 + ASPP	51.35	46.22	13.89	1.42×10^{6}
基础模型 + 改进 ASPP	54.92	49.42	13.35	1.43×10^{6}
基础模型 + RMHSA	57.14	51.42	13.27	1.35×10^{6}
基础模型 + RMHSA + 改进 ASPP	75.20	70.78	8.97	2.06×10^{6}

由表 1 可知，改进 ASPP 通过聚合不同尺度的信息增强了模型的判别能力，模型的 PA 与 MIoU 相较于基础模型也增加 15.46% 和 16.54%，且由于采摘点目标形状均为长条形，使用条形池化减少了不相关信息的获取，提高了分割指标；对比基础模型和增加 RMHSA 模型的结果发现，增加 RMHSA 模型的 PA 与 MIoU 提升到 57.14%、51.42%，表明增加 RMHSA 模型后，由于高效注意力机制的引入，模型更加倾向于寻找更有显著性的特征，实现强化目标、弱化背景的目的；对比基础模型 + RMHSA + 改进 ASPP 与基础模型的结果发现，相较于基础模型，本文模型的 PA 和 MIoU 增加 35.74% 和 37.90%，相

较于基础模型 + 改进 ASPP 与基础模型 + RMHSA，本文模型在准确率上有了较大的提升，且在高效注意力机制以及针对性多尺度结构的同时作用下，模型最大程度地保留了有效特征，保证了其向后传播，减少无用信息的干扰，促使网络更加关注目标物体不同部位的细节特征，也提升了目标区域的定位精度，有效组合不同尺度下的特征信息，提高分割精度。

在实时性方面，随着功能模块的加入，分割模型包含的参数不断增加，检测速度逐渐降低。其中，基础模型的参数数量最小，检测速度最快，RMHSA 模块的加入使基础模型参数量增加 25.35%，检测速度降低 12.99%；改进 ASPP 模块的增加使参数量增加 34.9%，检测速度降低 22.80%。基于上述数据定性分析，可以得知改进 ASPP 模块对模型的运行效率和计算开销影响最大，RHMSA 模块次之。

图 9 为表 1 中 5 种类别模型识别结果图，针对于分割目标的 3 个难点：目标较小、光照强度变化范围大、背景复杂，本文选取 7 幅图像展示。由图 9 可知，基础模型由于缺少多尺度结构以及有效的注意力机制，7 幅图像均出现了较大面积的错检和漏检。基础模型 + RMHSA 虽然增加了高效的注意力机制，相较于基础模型有较大的进步，如图 9 中的第 1 行和第 5 行，但由于网络结构中的多次下采样，模型损失较多的有效特征，所以造成结果仍有较多的漏检和错检。而基础模型 + ASPP 和基础模型 + 改进 ASPP 结果中，由于增加了多尺度结构，减少了下采样的特征丢失，但是由于缺少高效的注意力机制的引导，仍然效果不佳，如图 9 中的第 2、3、6、7 行。而本文所提出的模型结合残差多头自注意力机制以及改进的 ASPP 的优点，其识别结果相较于标注图像没有明显差距，没有大面积的漏检和错检，证明使用

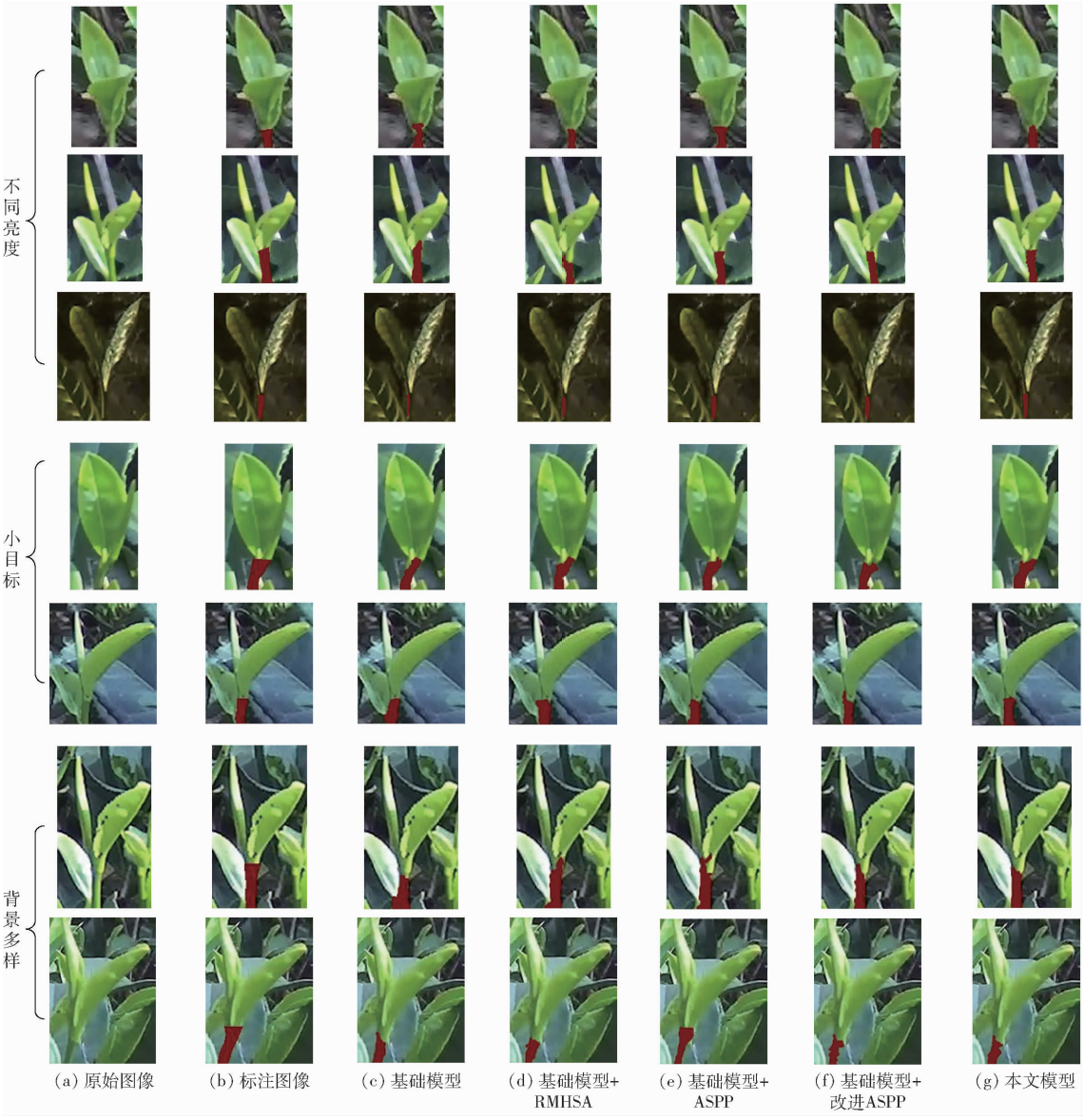

图 9 不同模型的语义分割结果

Fig. 9 Semantic segmentation results for different functional units

这两种结构组合，可有效加强模型的茶芽采摘点的分割能力。

2.3.2　不同语义分割模型性能对比

为验证本文所提出的模型具体性能情况，选择 HRNet V2[22]、EfficientUNet + +[23]、DeeplabV3 +[24]、BiSeNet V2[25]等模型与本文模型进行对比测试，通过像素准确率、区域重合度、检测速度和参数数量等指标对模型性能做出评价。表 2 为不同语义分割模型性能参数对比。

表 2　不同网络模型性能对比

Tab. 2　Performance comparison of different network models

模型	像素准确率/%	平均区域重合度/%	检测速度/(f·s^{-1})	参数量
DeeplabV3 +	67.92	65.42	3.19	4.12×10^7
EfficientUNet + +	74.64	69.86	4.14	1.28×10^7
BiSeNet V2	65.35	63.91	6.54	6.89×10^6
HRNet V2	77.43	72.62	4.27	1.43×10^7
本文模型	75.20	70.78	8.97	2.06×10^6

由表 2 可以看出，在像素准确率方面，本文模型的平均像素准确率与平均区域重合度分别为 75.20%、70.78%，比 DeeplabV3 + 分别高 7.28、5.36 个百分点；比 EfficientUNet + + 分别高 0.54、0.92 个百分点；比 BiSeNet V2 分别高 9.85、6.87 个百分点，主要原因是本文模型引入的残差多头自注意力机制和条形池化的 ASPP 模块能够强化模型各阶段判别特征的能力，聚合不同尺度池化区域获取有效的全局上下文信息，增加模型对茶芽采摘点对象的分割精度；尽管 DeeplabV3 + 也采用空洞卷积金字塔模块聚合多尺度特征以提高分割准确性，但由于数据集的特殊性，目标形态均为长条形，造成 DeeplabV3 + 存在像素预测不一致性，因此在分割精度方面略低于本文模型。在实时性方面，本文采用 ConvNeXt 作为前置基础网络降低模型计算量，检测速度达到 8.97 f/s，为 DeeplabV3 +、EfficientUNet + +、BiSeNet V2、HRNet V2 模型的 2.81、2.17、1.37、2.10 倍。在参数量方面，本文模型为 2.06×10^6，低于其他网络模型。

在所有对比模型中，BiSeNet V2 的特点是双路同时计算，模型的设计过程中缺乏多个尺度的图像特征交换，所以 BiSeNet V2 并不能很好地解决本文问题；HRNet V2 以及通过多次高分辨率特征浅层网络与快速下采样深度网络融合，对于小目标的识别有很好的效果，但是由于自身设计的原因，参数量约为本文模型的 7 倍，虽在识别精度上超过了本文模型，但是在识别速度上确实有所差距；而 EfficientUNet + + 使用 EfficientNet[26] 作为每层的特征提取网络，同时使用 U - Net[27] 的结构来构建网络，U - Net 结构的特点是同时存在多层特征提取网络，各层之间通过下采样和上采样相互联通，从各项指标来看，虽然 EfficientUNet + + 的精度指标与 RMHSA - NeXt 相近，但是由于 EfficientNet 的参数量较高，造成整体的检测速度不佳。综合评价上述精度和速度指标后，可得 MHSA - NeXt 模型可以实现精度与速度的均衡，具有良好的分割性能。

针对茶叶采摘点存在的目标较小、光照强度变化范围大和背景复杂问题，选出了 7 幅图像，图 10 为 5 种模型对测试集的分割结果。

从图 10 中可知，DeeplabV3 + 虽然在背景较为简单的场景如第 2、7 行分割效果与标注结果覆盖区域相似，但是在较为复杂的场景如第 1、3、4、5 行中，均有较大的误差，尤其在第 3 行出现了较大面积的错检，基于结果，本文认为 DeeplabV3 + 分割效果不佳的原因在于缺少高效的注意力机制和针对于茶叶采摘点特性的针对性设计。BiSeNet V2 识别效果不佳，虽然在第 6 行没有出现大面积漏检，但是在第 1、3 行等都有较为明显的误差，说明 BiSeNet V2 的双路并行计算的方式并不能很好地解决目标小、背景复杂、光照强度变化范围大的问题。HRNet V2 的识别效果与本文模型识别效果也如表 2 所示，检测结果与标注差别较小，但其在 PA 和 MIoU 指标上略高于 RMHSA - NeXt，表明多次的高低层级特征融合能有效利用小目标的图像信息，但是由于层次堆叠过多模型的整体数据量较高，检测速度较慢。从本文模型的识别结果来看，图 10 中除了在第 3 行中出现部分漏检，在其他目标中识别结果与标注图像基本相差较小，其根据目标的实际特点，设计了残差多头自注意力模块和结合条形池化的 ASPP 模块，以较小的参数量，获得了较好分割性能。

3　结论

（1）本文模型在实际茶园场景下，检测结果的像素准确率达到 75.20%，MIoU 为 70.78%，运行速度达到 8.97 f/s，解决了该场景下茶叶采摘点目标较小、背景复杂、光照强度变化范围大等困难，较好地完成了茶叶采摘点语义分割的任务。

（2）选择 HRNet V2、EfficientUNet + +、DeeplabV3 +、BiSeNet V2 模型与本文模型进行对比测试，通过比较像素准确率、区域重合度、检测速度、参数数量等指标，可以发现多次不同阶段的高低层

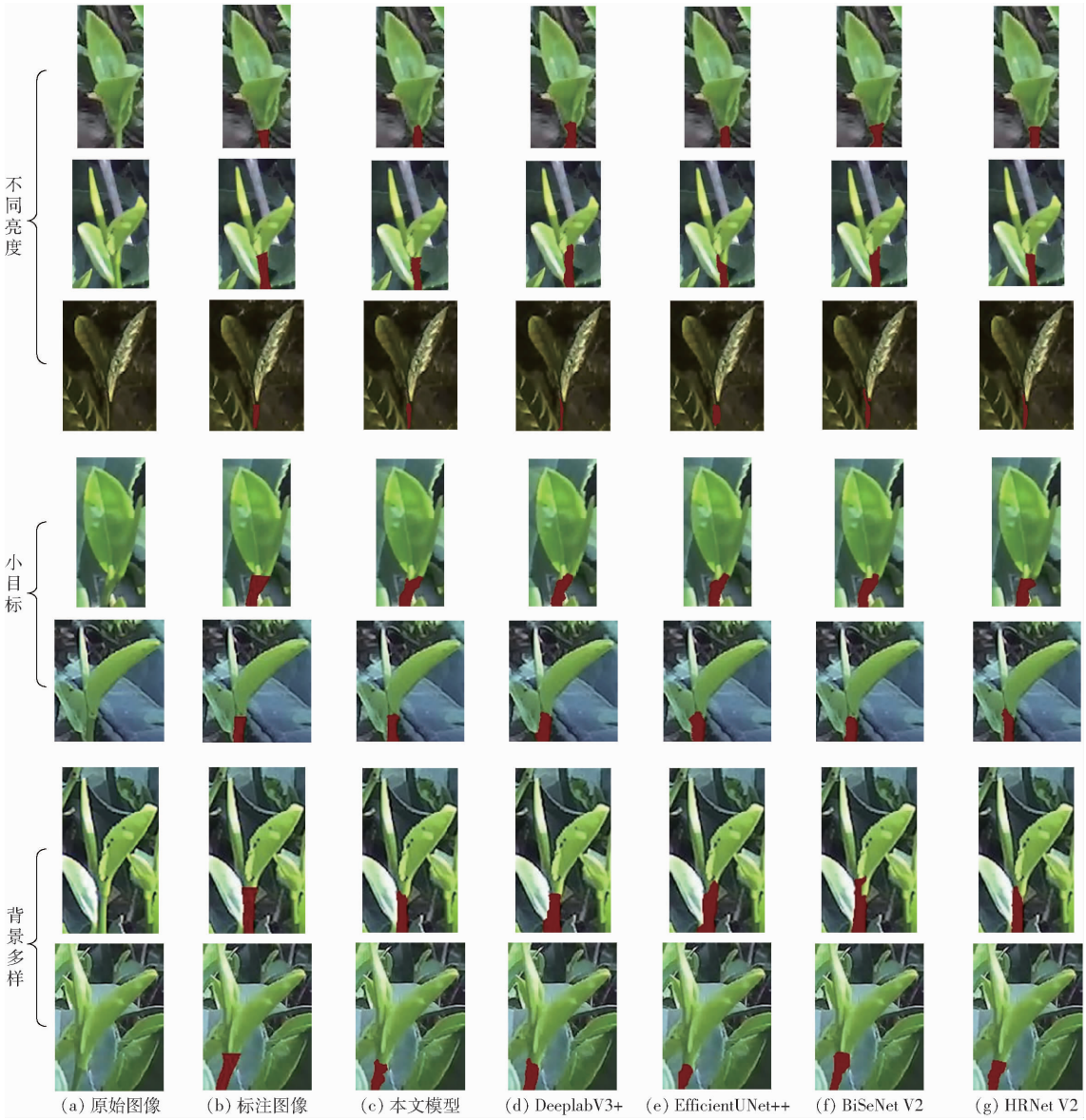

图 10 不同网络模型语义分割结果对比

Fig. 10 Comparison of semantic segmentation results of different network models

特征融合可以有效提取小目标的显著特征,但是检测速度就会下降,而本文中提出的两种特殊结构,以相对较少的参数量达到了较高的准确率,平衡了模型的检测速度与准确率。

参 考 文 献

[1] 王攀, 易文裕, 熊昌国, 等. 4CJZ - 1000 型自走式采茶机的设计与试验[J]. 西南大学学报(自然科学版), 2022, 44(11): 228 - 233.
WANG Pan, YI Wenyu, XIONG Changguo, et al. Design and test of 4CJZ - 1000 self-propelled tea picker[J]. Journal of Southwest University(Natural Science Edition), 2022, 44(11): 228 - 233. (in Chinese)

[2] 刘浩, 贺福强, 李荣隆, 等. 便携式真空吸附采茶机设计与试验[J]. 农机化研究, 2021, 43(7): 110 - 114.
LIU Hao, HE Fuqiang, LI Ronglong, et al. Design and test of portable vacuum absorption tea collector[J]. Journal of Agricultural Mechanization Research, 2021, 43(7): 110 - 114. (in Chinese)

[3] 赵润茂, 卞贤炳, 陈建能, 等. 分布控制的乘坐式仿形采茶原型机研制与试验[J]. 茶叶科学, 2022, 42(2): 263 - 276.
ZHAO Runmao, BIAN Xianbing, CHEN Jianneng, et al. Development and test for distributed control prototype of the riding profiling tea harvester[J]. Journal of Tea Science, 2022, 42(2): 263 - 276. (in Chinese)

[4] TIAN Jun, ZHU Honglin, LIANG Wenjiang, et al. Research on the application of machine vision in tea autonomous picking [J]. Journal of Physics: Conference Series, 2021, 1952(2): 22 - 31.

[5] 唐仙, 吴雪梅, 张富贵, 等. 基于阈值分割法的茶叶嫩芽识别研究[J]. 农业装备技术, 2013,39(6):10 - 14.
TANG Xian, WU Xuemei, ZHANG Fugui, et al. Contrastive research on tender tea recognition based on multiple threshold segmentation methods[J]. Agricultural Equipment & Technology. 2013,39(6):10 - 14. (in Chinese)

[6] 韦佳佳. 名优茶机械化采摘中嫩芽识别方法的研究[D]. 南京:南京林业大学, 2012.
WEI Jiajia. Researches on high-quality tea flushes identification for mechanical-plucking[D]. Nanjing: Nanjing Forestry University,2012. (in Chinese)

[7] 吴雪梅, 张富贵, 吕敬堂. 基于图像颜色信息的茶叶嫩叶识别方法研究[J]. 茶叶科学, 2013,33(6):584 - 589
WU Xuemei, ZHANG Fugui, LÜ Jingtang, et al. Research on recognition of tea tender leaf based on image color information [J]. Journal of Tea Science, 2013,33(6):584 - 589 (in Chinese)

[8] 王琨, 刘大茂. 基于深度学习的茶叶状态智能识别方法[J]. 重庆理工大学学报(自然科学), 2015, 29(12): 120 - 126.
WANG Kun, LIU Damao. Intelligent identification for tea state based on deep learning[J]. Journal of Chongqing University of Technology(Natural Science), 2015, 29(12): 120 - 126. (in Chinese)

[9] 吕军, 夏华鹃, 方梦瑞, 等. 基于 AlexNet 的茶叶嫩芽状态智能识别研究[J]. 黑龙江八一农垦大学学报, 2019, 31(2): 72 - 78.
LÜ Jun, XIA Huakun, FANG Mengrui, et al. Research on intelligent identification of tea sprouts state based on AlexNet[J]. Journal of Heilongjiang Bayi Agricultural University, 2019, 31(2): 72 - 78. (in Chinese)

[10] 孙肖肖, 牟少敏, 许永玉, 等. 基于深度学习的复杂背景下茶叶嫩芽检测算法[J]. 河北大学(自然科学版), 2019, 39(2): 211 - 216.
SUN Xiaoxiao, MU Shaomin, XU Yongyu, et al. Detection algorithm of tea tender buds under complex background based on deep learning[J]. Journal of Hebei University(Natural Science Edition), 2019, 39(2): 211 - 216. (in Chinese)

[11] 吕军, 方梦瑞, 姚青, 等. 基于区域亮度自适应校正的茶叶嫩芽检测模型[J]. 农业工程学报, 2021, 37(22): 278 - 285.
LÜ Jun, FANG Mengrui, YAO Qing, et al. Detection model for tea buds based on region brightness adaptive correction[J]. Transactions of the CSAE, 2021, 37(22): 278 - 285. (in Chinese)

[12] LI Denghui, SUN Xiaoxuan, LÜ Shengping, et al. A novel approach for the 3D localization of branch picking points based on deep learning applied to longan harvesting UAVs[J]. Computers and Electronics in Agriculture, 2022, 199: 107 - 191.

[13] XIE Zhongzhao, LI Yongbo, HU Zijia, et al. Deep convolutional nets for semantic image segmentation with deep Gaussian CRFs[C]// The 2017 International Workshop on Advanced Computational Intelligence and Intelligent Informatics, 2017.

[14] ZHU Weidong, SUN Jun, WANG Simin, et al. Segmentation and recognition of filed sweet pepper based on improved self-attention convolutional neural networks[J]. Multimedia Systems, 2023, 29(1): 223 - 234.

[15] YU Yang, ZHANG Kailiang, YANG Li, et al. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask R - CNN[J]. Computers and Electronics in Agriculture, 2019, 163: 104846.

[16] 孟庆宽, 杨晓霞, 张漫, 等. 基于语义分割的非结构化田间道路场景识别[J]. 农业工程学报, 2021, 37(22): 152 - 160.
MENG Qingkuan, YANG Xiaoxia, ZHANG Man, et al. Recognition of unstructured field road scene based on semantic segmentation model[J]. Transactions of the CSAE, 2021, 37(22): 152 - 160. (in Chinese)

[17] LIU Ze, LIN Yutong, CAO Yue, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021.

[18] XU Jingjing, SUN Xu, ZHANG Zhiyuan, et al. Understanding and improving layer normalization[C]// Advances in Neural Information Processing Systems,2019.

[19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems, 2017.

[20] LIN Zhouhan, FENG Minwei, SANTOS C, et al. A structured self-attentive sentence embedding[J]. arXiv: 1703.03130, 2017.

[21] HOU Qibin, ZHANG Li, CHENG Mingming, et al. Strip pooling: rethinking spatial pooling for scene parsing[J]. arXiv: 2003.13328, 2020.

[22] YUAN Yuhui, CHEN Xilin, WANG Jingdong. Object-contextual representations for semantic segmentation[C]// Computer Vision-ECCV 2020: 16th European Conference, 2020.

[23] LOURENÇO-SILVA J L, MENEZES M N, RODRIGUES T, et al. Encoder-decoder architectures for clinically relevant coronary artery segmentation[C]// Computational Advances in Bio and Medical Sciences: 11th International Conference, ICCABS 2021, 2021.

[24] CHEN Liang-Chieh, ZHU Yukun, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]// Computer Vision-ECCV 2018, 2018.

[25] YU Changqian, GAO Changxin, WANG Jinbo, et al. Bisenet v2: bilateral network with guided aggregation for real-time semantic segmentation[J]. International Journal of Computer Vision, 2021, 129: 3051 - 3068.

[26] TAN Mingxing, LE Q V. Efficientnet: rethinking model scaling for convolutional neural networks[J]. arXiv:1905.11946, 2020.

[27] 马钦, 万传峰, 卫建, 等. 基于 U - Net 和特征金字塔网络的秸秆覆盖率计算方法[J]. 农业机械学报, 2023,54(1): 224 - 234.
MA Qin, WAN Chuanfeng, WEI Jian, et al. Calculation method of straw coverage based on U - Net network and feature pyramid network[J]. Transactions of the Chinese Society for Agricultural Machinery,2023,54(1): 224 - 234. (in Chinese)