

基于自监督学习的玉米植株图像小样本语义分割模型

邓寒冰 刘鑫 李朝阳 苗腾
(沈阳农业大学信息与电气工程学院, 沈阳 110866)

摘要: 图像语义分割技术是获取玉米植株表型信息的重要手段之一,传统的全监督语义分割方法往往依赖大量像素级标签,但玉米在不同生长阶段形态多变,导致图像标注成本高昂,制约模型在实际生产中的应用。为了去掉模型训练中的人工标注过程,本研究提出了一种基于自监督学习的玉米植株图像小样本语义分割网络(Self-supervised few-shot semantic segmentation network for maize plant images,MSDANet),以提高不同生长时期玉米植株图像的语义分割精度和模型泛化能力。MSDANet利用基于超像素的自监督学习方法生成伪标签,无需人工标注即可为支持集图像构建初步监督信号;设计混合遮蔽机制(Mixed masking,MM),应用基于伪标签的语义遮蔽,在特征空间构建多样性遮蔽样本,促进模型学习更鲁棒性的特征表达,从而提高复杂背景下的分割精度。针对图像中玉米植株存在的弯曲、重叠、遮挡等复杂形态问题,本研究为模型设计了多尺度可变形大核卷积注意力机制(Multi-scale deformable large kernel attention,MS-DLKA),通过融合多尺度感受野和可变形卷积,能够灵活感知玉米植株在不同尺度下的重要结构信息,有效提高了语义分割精度。在小样本数据集上进行验证,在1-shot设置下,MSDANet的mIoU和FB-IoU分别达到75.63%和87.12%;在5-shot设置下,mIoU和FB-IoU分别达到76.04%和87.21%,均优于本研究给出的同类其他模型。此外,与当前主流的全监督小样本语义分割模型对比,在1-shot和5-shot设置下,mIoU分别提升2.9、2.93个百分点。结果表明,MSDANet模型能够在无人工标签和小样本的前提下,实现高精度的玉米植株图像语义分割任务,为不同生长时期的玉米图像分析与植物表型测量提供了技术支持。

关键词: 玉米图像; 植物表型; 图像处理; 深度学习; 语义分割; 自监督学习

中图分类号: S24; TP391 文献标识码: A 文章编号: 1000-1298(2026)01-0072-11 OSID: 

Self-supervised Few-shot Semantic Segmentation Model for Maize Plant Images

DENG Hanbing LIU Xin LI Chaoyang MIAO Teng
(College of Information and Electrical Engineering, Shenyang Agricultural University, Shenyang 110866, China)

Abstract: Image semantic segmentation technology is one of the key methods for obtaining phenotypic information of maize plants. Traditional fully supervised semantic segmentation methods typically rely on a large number of pixel-level labels. However, maize exhibits significant morphological variability across different growth stages, leading to high costs associated with image annotation and limiting the practical application of such models in real-world production scenarios. To eliminate the need for manual annotation during model training, a self-supervised few-shot semantic segmentation network for maize plant images (MSDANet) was proposed based on self-supervised learning, aiming to improve the semantic segmentation accuracy and model generalization capability of maize plant images across different growth stages. MSDANet utilized a superpixel-based self-supervised learning method to generate pseudo labels, enabling the construction of preliminary supervision signals for the support set images without manual annotation. It designed a mixed masking mechanism (MM) that applied pseudo label-based semantic masking to construct diverse masked samples in the feature space, promoting the model to learn more robust feature representations and thereby improving segmentation accuracy in complex backgrounds. To address the complex morphological issues of corn plants in images, such as bending, overlapping, and occlusion, a multi-scale deformable large kernel attention mechanism (MS-DLKA) for the model was designed. By integrating multi-scale receptive fields and deformable convolutions, it can

flexibly perceive important structural information of corn plants at different scales, effectively improving semantic segmentation accuracy. When validated on a small sample dataset, MSDANet achieved mIoU and FB-IoU of 75.63% and 87.12%, respectively, in the 1-shot setting; in the 5-shot setting, mIoU and FB-IoU reached 76.04% and 87.21%, respectively, both outperforming other models of the same type proposed in this study. Additionally, compared with current mainstream fully supervised few-shot semantic segmentation models, mIoU was improved by 2.9 and 2.93 percentage points under 1-shot and 5-shot settings, respectively. The results demonstrated that the MSDANet model can achieve high-precision semantic segmentation of corn plant images without human labels and with few samples, providing technical support for corn image analysis and plant phenotyping at different growth stages.

Key words: maize images; plant phenotype; image processing; deep learning; semantic segmentation; self-supervised learning

0 引言

玉米作为世界三大粮食作物之一,是全球范围内广泛种植的经济和饲料作物。为了获得高产的玉米品种,研究人员亟需快速获取不同生长时期的玉米植株表型信息的方法。随着信息技术的发展,利用视觉传感器可实现植物表型的快速、海量获取,其中图像分割在表型测量中发挥关键作用。然而,玉米在不同生长时期表现出显著的形态差异:例如苗期玉米多呈放射状分布,叶片细长、姿态多变;随着生长的推进,玉米株高增加、叶片重叠和弯曲现象加剧。传统基于人工观察和测量表型信息(株高、叶面积等)的方法效率低且精度受主观因素和环境影响,因此,基于图像的植物表型解析方法已经成为当前该领域研究热点,特别是对于玉米这种高度异质化、结构复杂的作物,其在不同时期的表型具有明显差异,更依赖于精确的图像分割技术提取关键信息。传统的计算机视觉算法(如阈值分割^[1]、基于颜色空间的图像分割^[2]、基于条件随机场的图像分割^[3]、基于统计直方图 K-means 聚类算法的图像聚类分割^[4]等)鲁棒性有限,难以应对光照变化、背景干扰及复杂结构,限制了在实际表型解析中的应用。近些年,深度学习技术迅速发展,基于卷积神经网络的图像分割方法已经广泛应用于植物表型研究中,研究人员在全监督学习、弱监督学习、半监督学习、自监督学习等方面提出了多种解决方案^[5-10]。然而,已有方法通常依赖大量像素级标注数据进行监督学习,而高质量玉米植株标签获取代价高昂,制约了其在多场景、多时期下的推广。为了缓解标注需求,越来越多的研究人员将小样本学习引入图像分割领域^[11-14]。利用先验知识,可以快速泛化到只包含少量有监督信息样本的新任务。因此,在样本数据有限或稀缺的情况下,基于小样本训练的网络模型在图像分割任务中表现突出^[11]。小样本语义分割(FSS)要求模型在极少标

注下完成新类别的像素级分割,常采用基于度量的元学习方法^[15-18],仅有少量研究工作摒弃了元学习的方法^[19]。然而此类方法仍依赖对可见类别的大量注释,降低标注依赖仍是挑战。

为此,许多研究人员尝试以无监督或自监督的方式完成小样本分割^[13-14,20-22],通过显著性估计、对比学习、超像素伪标签等方式代替人工标注,取得一定成效。然而针对玉米植株图像这一高异质性场景仍缺乏有效解决方案。本研究提出一种结合自监督学习与小样本学习思想的玉米植株图像语义分割方法,构建面向不同生长时期玉米植株的小样本语义分割网络(MSDANet),以期实现不同生长时期玉米植株图像的精确分割。

1 材料与方法

1.1 试验材料和数据采集

本试验选择的玉米品种为“先玉 335”,在大田环境下采集不同背景不同光照条件的苗期阶段玉米单株俯视图像(单体株高 15~30 cm,叶片数 3~4 叶),试验数据由相机俯视平行地面拍摄获取,相机安装高度距地面 1.0 m,试验获得的原始图像分辨率为 3 024 像素×4 032 像素。在室内环境中,选取 50 株不同生长时期的玉米植株进行试验,植株高度为 20~120 cm,相机位置固定,与拍摄对象(玉米植株)保持 1.0~1.5 m,玉米植株与背景布的距离为 0.5 m,这样可以保证背景布和玉米植株都在相机的有效拍摄范围内。玉米植株被放置在可旋转底座上,通过旋转底座可以从不同角度对玉米植株进行拍摄取样。试验分别选取了 10 个拍摄角度进行拍摄,试验获取的原始图像尺寸为 1 920 像素×1 080 像素。具体的试验场景布局如图 1 所示。

1.2 构建玉米植株图像小样本数据集

本研究的数据集包含不同背景、不同株高、不同光照条件下的玉米植株,试验分别选取大田环境下的苗期玉米植株和室内环境下不同生长时期的玉米

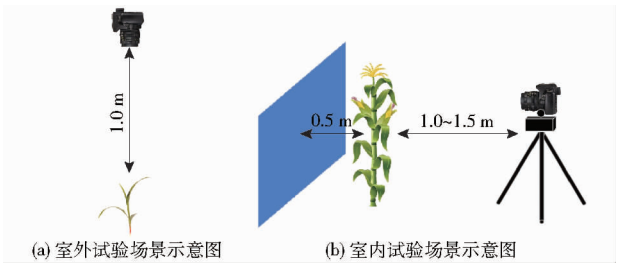


图 1 试验场景示意图

Fig. 1 Schematic of experimental scenario

植株作为研究对象,单体株高在 20 ~ 120 cm 范围内。经过初步筛选,共筛选出 500 幅大田单株玉米苗期俯视图,500 幅室内环境下不同角度、不同株高的玉米苗期图像。由于采集数据时相机的分辨率和相机与玉米植株的距离是固定的,而不同生长时期的玉米所占图像的比例不同,使得图像中包含大量非目标区域,容易干扰模型对玉米植株的准确感知。因此,采用中心裁剪的方式,从每幅图像中提取尺寸为 1 080 像素 × 1 080 像素的区域。为进一步适配主干网络 ResNet50 的输入要求,同时降低计算资源消耗,采用双线性插值法^[23] (Bilinear interpolation)将图像缩放至 224 像素 × 224 像素的标准尺寸。该方法在保留图像主要结构特征的同时,能有效降低尺寸变化引入的信息失真,提升模型对玉米植株细节结构的保留能力。为提升模型在复杂环境下的鲁棒性与泛化能力,采用多种数据增强处理,以模拟玉米植株在不同环境、角度和光照条件下的变化特征。主要采用以下数据增强方法:①随机翻转与旋转:使模型能够适应玉米植株在不同方向的形态分布。②颜色抖动:对图像亮度、对比度、饱和度和色调随机调整,增强模型对不同光照条件下图像的适应能力。③随机裁剪与缩放:增强模型对不同尺度目标及局部区域的感知能力。最后,结合本研究提出的混合遮蔽机制,对图像进行随机遮挡和基于伪标签的语义遮挡,增强模型在目标部分缺失情况下的语义推理能力,增强后的数据样本图像共 1 500 幅。

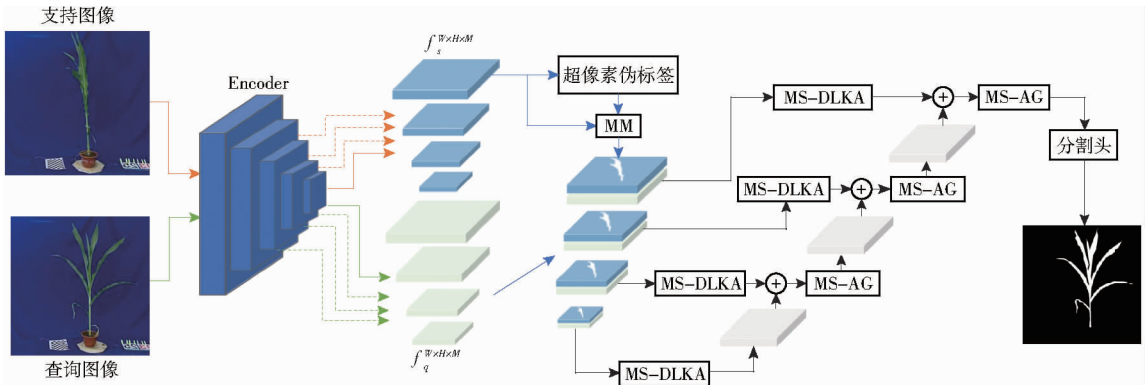


图 2 总体模型框架图

Fig. 2 Overall model framework

2 基于自监督学习的少样本语义分割模型

2.1 总体模型框架

本研究旨在构建基于自监督学习的玉米植株图像少样本语义分割模型(MSDANet),采用支持图像与查询图像作为输入,并通过深度神经网络提取多尺度特征。结合基于超像素的自监督学习,对玉米植株图像的前景部分自动生成伪标签,代替训练样本的人工标注过程。通过混合遮蔽机制,在特征空间构建多样性的遮蔽样本,使模型在特征层面增强前景区域的识别能力。同时,借助多尺度可变形大核卷积注意力机制(MS-DLKA)和多尺度注意力机制(MS-AG)模块,有效融合关键信息,提升分割精度。最终,分割头对增强后的特征进行处理,生成高质量的二值化分割结果。采用带有 ImageNet 预训练权值的 ResNet50 作为主干特征提取网络,特征提取是在骨干网络每个编码器块上的最后一个卷积层进行。这种特征提取方法产生 4 个金字塔层($p = 4$)。为确保一致性,本研究将支持图像和查询图像的空间尺寸(即图像的高与宽分别记为 H 和 W ,空间尺寸为 $H \times W$)都设置为 400 像素 × 400 像素,从而得到 $H \times W = 400$ 像素 × 400 像素,因此,得到的每个金字塔层的空间大小为: $H_1 \times W_1 = 100$ 像素 × 100 像素, $H_2 \times W_2 = 50$ 像素 × 50 像素, $H_3 \times W_3 = 25$ 像素 × 25 像素, $H_4 \times W_4 = 13$ 像素 × 13 像素。解码器组件由 4 个块组成,每个块涉及将支持原型与查询表示融合,如图 2 所示。该框架充分利用少样本学习和注意力机制,提高了玉米植株图像分割的鲁棒性和精度。

2.2 基于超像素分割的伪标签生成方法

使用 LabelMe 软件标注单幅大田玉米苗期图像(俯视图),时间约为 60 s,而标注室内玉米植株图像(株高 20 ~ 120 cm,侧视图)的平均标注时间约为 650 s。为降低人工标注成本,本研究设计了基于超像素分割的伪标签生成方法,可在无需人

工标注的前提下,自动提取图像中玉米植株的前景区域,生成伪掩膜用于后续小样本语义分割任务。

在本研究中,借鉴了 OUYANG 等^[21] 基于超像素的自监督学习的思想,使用超像素消除人工标注的需求。具体而言,使用了 Felzenszwalb^[24] 超像素分割算法,该算法结合超像素方法 SLIC (Simple linear iterative cluster)^[25],在构造无向图前先对原图进行预分割,得到形状尺寸基本一致且贴合图像边界的 SLIC 超像素;然后以超像素的颜色平均值对其进行特征描述,将每个超像素视为无向图节点,构造无向权重图,进而实现基于图的无监督分割。首先,本试验将输入的玉米植株 RGB 图像转换为 Lab 颜色空间,通过获取数据集在 L 、 a 、 b 通道的分布,获

得前景阈值。图 3 为本试验中大田苗期玉米和室内玉米图像分别在 L 、 a 、 b 分量上的像素值分布,为了减少极端值对阈值选择的影响,选择 5% ~ 95% 分位数,设置大田苗期图像 L 的阈值范围为 $[10,90]$, a 的阈值范围为 $[-25,0]$, b 的阈值范围为 $[0,30]$,设置室内玉米图像 L 的阈值范围为 $[20,80]$, a 的阈值范围为 $[-25,0]$, b 的阈值范围为 $[-20,25]$ 。使用 Felzenszwalb 超像素分割,根据设置的 L 、 a 、 b 阈值,计算每个超像素区域的前景像素比例,将前景像素比例超过 50% 的超像素区域判定为前景区域。基于超像素分割生成伪标签的结果如图 4 所示,图像中存在一些离散的小面积噪声,因此在超像素分割后进行一次去噪和膨胀处理,去除冗余噪声,平滑边界,分割结果如图 4d 所示。

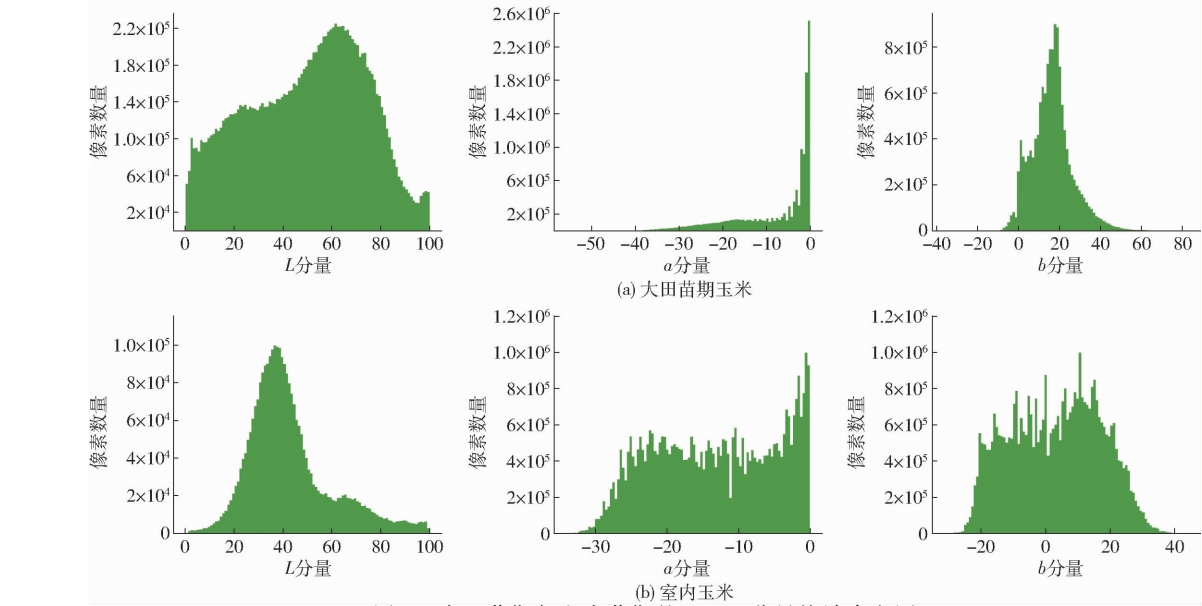


图 3 大田苗期与室内苗期的 L 、 a 、 b 分量统计直方图

Fig. 3 Statistical histograms of L , a and b components of field seedling stage and indoor seedling stage

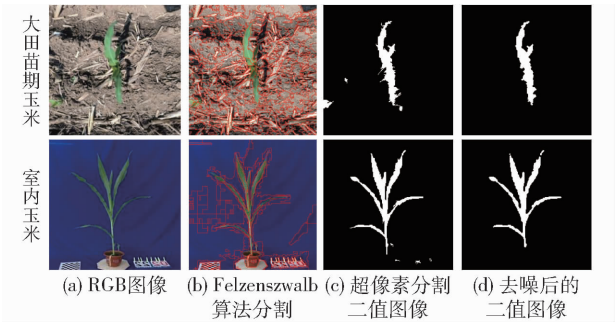


图 4 基于超像素分割的前景伪掩膜

Fig. 4 Foreground pseudo-mask based on superpixel segmentation

为了评估超像素生成的伪标签的可靠性与有效性,将基于超像素方法生成的大田环境下和室内环境下的伪标签与其对应的真实标签进行对比,采用平均交并比(mIoU)、Dice 系数、像素精度(Pixel accuracy, PA)作为评价指标进行定量分析,结果如表 1 所示。

从表 1 结果可以看出,伪标签与真实标签之间具有较强的空间结构一致性和区域覆盖准确性。尤其是 Dice 系数达到 83%,表明在区域边界保持方面表现良好;同时较高的 PA 也进一步佐证了伪标签在像素层面具有极高的准确性。本研究采用的基于超像素分割的伪标签生成方法在无需人工标注的前提下,能够获得与真实标签高度一致的语义信息,证明其在实际应用中具有较高的可行性与有效性。

表 1 伪标签与真实标签质量对比			
Tab. 1 Comparison of pseudo labels and ground truth label quality			%
环境	mIoU	Dice 系数	PA
大田	71.68	83.08	98.23
室内	72.37	83.27	98.24

2.3 混合遮蔽机制

在小样本语义分割 (Few-shot semantic

segmentation, FSS)任务中,旨在通过有限的带标签支持图像引导模型,实现对查询图像中目标区域的精准分割。支持图像中仅部分区域包含目标对象,其余区域为背景,直接使用全图特征作为支持特征,容易引入无关噪声,从而干扰查询图像的目标预测。因此,如何通过支持图像捕获判别性区域特征成为提升分割性能的关键。现有方法普遍采用特征遮蔽(Feature masking, FM)和输入遮蔽(Input masking, IM)来增强对支持图像的信息捕获。特征遮蔽(FM)的核心思想是对输入特征随机进行部分遮蔽,来去除无关的特征激活,防止模型对某些特征的过度依赖;输入遮蔽(IM)通过将支持图像与其对应支持掩膜相乘来消除背景像素对目标判别的影响。

为兼顾特征遮蔽和输入遮蔽各自优点,本研究提出了一种混合遮蔽机制(Mixed masking, MM),该方法在目标区域判别特征的同时,利用支持掩膜遮蔽生成混合图像,从而更稳健地填补非目标区域特征。该机制在支持图像的特征构建阶段引入像素级目标强化与背景抑制,有效地提升模型对目标区域的感知能力。混合遮蔽的核心思想是通过输入遮蔽

来补偿特征遮蔽中目标对象的空间细节丢失,特别是在目标边界和纹理方面,在去除背景信息的同时,保留更多目标对象的细节信息。具体而言,首先计算特征遮蔽和输入遮蔽特征,然后将混合遮蔽特征中的未激活值替换为输入遮蔽 IM 特性,其他激活值保持不变,无需替换以保持特征遮蔽的功能。同时,在计算特征遮蔽的特征时,将掩膜拟合到特征尺寸上。

此外,本研究采用超像素分割生成的支持图像掩膜在边界处往往存在缺失或不精确的情况,这种掩膜不完整会在特征引导阶段掩盖目标的真实边界信息,导致模型在处理边缘细节时出现模糊或遗漏。因此,利用背景中的相关性上下文信息来补偿有限的支持掩膜信息,计算原始支持特征 F^s 和混合遮蔽后的支持特征 F^{MM} 与查询特征 F^q 的余弦相似度来获得多层相关性,引导查询图像中目标区域的精准识别。混合遮蔽机制通过融合空间结构与语义内容,在保留目标形态的同时强化边界引导,从而在查询图像中仍能有效定位目标区域,提升整体的分割精度和边界细粒度表达能力,具体如图 5 所示。

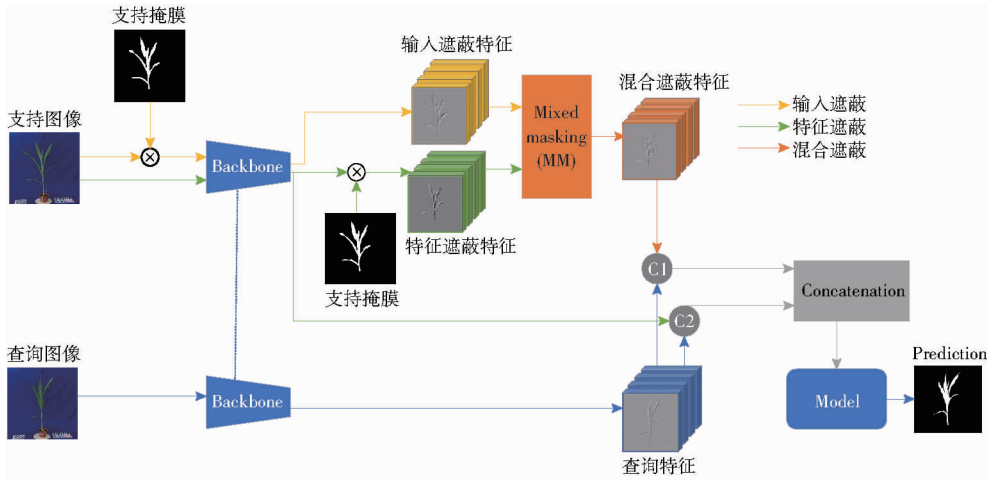


图 5 混合遮蔽 (MM) 的整体架构
Fig. 5 Mixed masking (MM) for overall architecture

混合遮蔽机制通过 2 条路径提取支持图像特征:特征遮蔽特征和输入遮蔽特征。最终在非目标区域使用输入遮蔽特征替代特征遮蔽中缺失的值,从而构建一个兼具语义精度与结构完整性的混合特征表示。此外,利用原始支持特征和混合遮蔽特征计算它们与查询特征的余弦相似度,并以此来获得它们的多层相关性表示,将相关性拼接回查询特征中。该混合遮蔽机制包括 5 个关键步骤:

(1) 查询与支持特征提取

使用主干网络对查询图像和支持图像进行特征提取,得到多尺度特征

$$F^q = \text{Backbone}(I^q) \tag{1}$$

$$F^s = \text{Backbone}(I^s) \tag{2}$$

其中 $F^q = \{F^q_l\}_{l=1}^L$ $F^s = \{F^s_l\}_{l=1}^L$
式中 F^q 、 F^s ——查询图像和支持图像从浅层到深层的多尺度特征

Backbone()——主干网络

I^q 、 I^s ——查询图像特征和支持图像特征

(2) 特征遮蔽 (FM)

利用支持掩膜 $M^s \in \{0, 1\}^{h \times w}$ 对支持特征进行遮蔽处理,仅保留前景目标区域的特征表示,剔除背景干扰信息。对每一层 l 执行逐像素乘法

$$\hat{F}^s_l = F^s_l \odot \text{Up}(M^s) \tag{3}$$

式中 $\text{Up}(\cdot)$ ——将掩膜上采样至与特征图相同尺寸

$\odot$ ——Hadamard 积

$\hat{\mathbf{F}}_l^s$ ——特征遮蔽特征

(3) 输入遮蔽 (IM)

构建遮蔽后的支持图像,仅保留 RGB 图像中目标区域,并通过相同的主干网络提取其特征。这些特征主要保留了前景区域的纹理和结构信息,能够补充原始支持图像中被遮蔽区域的上下文表示。

通过将支持图像与掩膜相乘,仅保留前景图像内容,构造遮蔽图像

$$\hat{\mathbf{I}} = \mathbf{I} \odot \mathbf{M}^s \quad (4)$$

式中 $\hat{\mathbf{I}}$ ——输入遮蔽

然后提取其结构化特征

$$\hat{\mathbf{F}}^s = \text{Backbone}(\hat{\mathbf{I}}^s) \quad (5)$$

其中 $\hat{\mathbf{F}}^s = \{\hat{\mathbf{F}}_l^s\}_{l=1}^L$

式中 $\hat{\mathbf{F}}^s$ ——输入遮蔽特征

(4) 混合遮蔽特征融合 (MM)

对每层支持特征执行逐像素替换操作:若特征遮蔽特征中该位置为 0 (即背景),则使用输入遮蔽特征进行补全,从而生成混合特征。混合策略为

$$\mathbf{F}_l^{MM} = \begin{cases} \hat{\mathbf{F}}_l^s(x, y) & (\hat{\mathbf{F}}_l^s(x, y) > 0) \\ \tilde{\mathbf{F}}_l^s(x, y) & (\text{其他}) \end{cases} \quad (6)$$

式中 $\mathbf{F}_l^{MM}$ ——混合遮蔽特征

最终得到融合后的支持特征

$$\mathbf{F}^{MM} = \{\mathbf{F}_l^{MM}\}_{l=1}^L$$

(5) 多层相关性计算

为了进一步挖掘查询特征 $\mathbf{F}^q$ 与混合遮蔽特征融合后的支持特征 $\mathbf{F}^{MM}$ 在语义空间的对应关系,在混合遮蔽特征基础上设计了多层相关性计算模块,并用余弦相似度作为相似度度量方式,增强对前景目标区域的特征补偿。计算方法为:

① 特征归一化

对于每一层 l , 令 $\mathbf{F}_l^q \in \mathbf{R}^{C_l \times H_l \times W_l}$, $\mathbf{F}_l^s \in \mathbf{R}^{C_l \times H_l \times W_l}$, $\mathbf{F}_l^{MM} \in \mathbf{R}^{C_l \times H_l \times W_l}$, 将空间位置 (x, y) 处的通道向量分别记为 $\mathbf{Q}_l(x, y) \in \mathbf{R}^{C_l}$, $\mathbf{S}_l(x, y) \in \mathbf{R}^{C_l}$, $\mathbf{M}_l(x, y) \in \mathbf{R}^{C_l}$, 对它们做逐向量归一化

$$\hat{\mathbf{Q}}_l(x, y) = \frac{\mathbf{Q}_l(x, y)}{\|\mathbf{Q}_l(x, y)\|_2} \quad (7)$$

$$\hat{\mathbf{S}}_l(x, y) = \frac{\mathbf{S}_l(x, y)}{\|\mathbf{S}_l(x, y)\|_2} \quad (8)$$

$$\hat{\mathbf{M}}_l(x, y) = \frac{\mathbf{M}_l(x, y)}{\|\mathbf{M}_l(x, y)\|_2} \quad (9)$$

② 双路径余弦相似度计算

在层 l 的每个空间位置 (x, y) , 分别计算原始支持特征和混合遮蔽特征与查询图像特征的余弦相似

度,得到 2 个相似度图 C_s, C_m

$$C_s(x, y) = \langle \hat{\mathbf{Q}}_l(x, y), \hat{\mathbf{S}}_l(x, y) \rangle = \frac{\hat{\mathbf{Q}}_l(x, y) \hat{\mathbf{S}}_l(x, y)}{\|\hat{\mathbf{Q}}_l(x, y)\|_2 \|\hat{\mathbf{S}}_l(x, y)\|_2} \quad (10)$$

$$C_m(x, y) = \langle \hat{\mathbf{Q}}_l(x, y), \hat{\mathbf{M}}_l(x, y) \rangle = \frac{\hat{\mathbf{Q}}_l(x, y) \hat{\mathbf{M}}_l(x, y)}{\|\hat{\mathbf{Q}}_l(x, y)\|_2 \|\hat{\mathbf{M}}_l(x, y)\|_2} \quad (11)$$

③ 相似度图拼接

将同层的两幅相似度图在通道维度拼接,得到融合后的相似度特征图

$$C_l(x, y) = \text{Concat}[C_s(x, y), C_m(x, y)] \quad (12)$$

其中 $C_l(x, y) \in \mathbf{R}^{2 \times H_l \times W_l}$

式中 Concat——特征图拼接

$C_l(x, y)$ ——拼接后的特征图

④ 多层融合

对所有选定层 $l \in \mathcal{L}$ 重复上述计算,得到一组 $\{C_l\}_{l \in \mathcal{L}}$, 在通道维度继续拼接,形成多层相关性特征

$$C_{\text{multi}} = \text{Concat}[C_{l_1}, C_{l_2}, \dots] \quad (13)$$

$(C_{\text{multi}} \in \mathbf{R}^{2^{|\mathcal{L}|} \times H' \times W'})$

式中 C_{multi} ——多层相关性特征

⑤ 查询特征增强

最终将多层相关性特征与查询特征拼接

$$\hat{\mathbf{F}}_l^q = \text{Concat}(\mathbf{F}_l^q, C_{\text{multi}}) \quad (14)$$

得到的融合相似图不仅考虑了不同分辨率下的匹配信息,还引入了从纹理细节到空间信息的多层相关性,抑制背景区域与前景区域之间的特征偏差,有效增强了模型对小样本场景下前景位置的判别能力。

在针对玉米图像的小样本语义分割任务中,玉米植株往往具有形态多变、背景复杂、遮挡严重等问题。MM 机制通过保留植株的关键结构信息(如叶片轮廓、纹理分布)并抑制背景冗余(如田间杂草、阴影等),在特征提取阶段更准确地聚焦于支持图像的判别区域,从而有效提升模型在复杂环境中的分割性能与泛化能力。

2.4 多尺度可变形大核卷积注意力机制

由于玉米植株在不同生长阶段的形态差异较大,其叶片存在不同程度的弯曲、重叠、遮挡等情况,传统的固定形状卷积核难以精准捕捉这些变形结构。此外,在大田环境背景复杂、植株与杂草等存在相似的颜色和纹理,传统的 CNN 结构难以有效区分目标与背景。而在少样本语义分割任务中,由于训练样本有限,模型需要具备更强的特征提取能力,以有效学习玉米植株的形态特征和背景区域的区分

性。因此,在这种背景下,本研究设计了一种多尺度可变形大核卷积注意力机制(Multi-scale deformable large kernel attention, MS-DLKA),模型不仅提高了对局部特征和多尺度特征的感知能力,还能够自适应地处理不同光照和环境变化下的玉米图像。特别是对不同生长阶段的玉米植株,使模型可以有效应对玉米形态和结构的变化,从而提高了分割精度和鲁棒性。

本研究提出的多尺度可变形大核卷积注意力机制融合多尺度感受野、可变形卷积和大核注意力机制,旨在增强模型对不同尺度特征的感知能力,同时利用可变形卷积动态调整感受野,以适应玉米不同生长时期的形态变化。多尺度可变形大核卷积注意力机制的核心思想是:使用可变形卷积来增强局部特征的感知能力,采用多尺度卷积提取不同感受野的信息,使用 1×1 卷积进行通道融合,提高特征表达能力,加入残差连接,增强梯度流动,提高训练稳定性。模块的体系结构如图6所示。

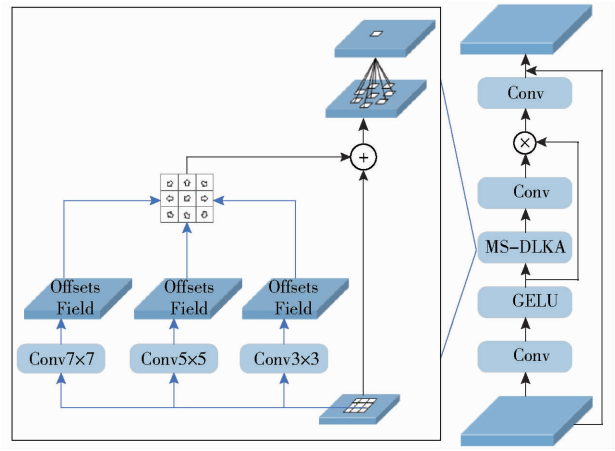


图6 多尺度可变形大核卷积注意力机制
Fig.6 Multi-scale deformable large kernel attention

多尺度可变形大核卷积注意力机制包含3个模块:可变形卷积(Deformable convolution, DCN)^[26]、多尺度大核卷积(Multi-scale large kernel)和多尺度可变形大核卷积注意力机制。

可变形卷积通过引入学习得到的偏移量(offsets)使卷积核能够对输入特征图的不同位置进行自适应采样,从而提升对目标形状的适应性。传统卷积的计算方式为

$$y(p) = \sum_k w_k x(p + p_k) \tag{15}$$

式中 $y(p)$ ——输出特征图在位置 p 的值
 x ——输入特征图
 w_k ——卷积核权重
 p_k ——固定的采样点
在可变形卷积中,加入了一个学习得到的偏移量

Δp_k ,即

$$y(p) = \sum_k w_k x(p + p_k + \Delta p_k) \tag{16}$$

Δp_k 使得卷积核可以根据输入特征动态调整感受野。训练过程中,偏移量 Δp_k 通过额外的偏移预测网络(offset_net)计算到

$$\Delta p_k = f_{\text{offset}}(X) \tag{17}$$

式中 f_{offset} ——卷积层
这样,每个位置的卷积核采样点都可以自适应调整,使得采集位置能适应玉米植株的不同形态,如叶片弯曲、重叠等情况。

为了提升模型对不同尺度的适应能力,采用多尺度大核卷积(Multi-scale large kernel)结合了可变形卷积模块,采用了3种不同感受野的可变形卷积,分别采用 3×3 、 5×5 、 7×7 3种卷积核尺寸,并采用不同的扩张率进行感受野控制

$$y_3 = \sum_k w_k^{(3)} x(p + p_k + \Delta p_k^{(3)}) \tag{18}$$

$$y_5 = \sum_k w_k^{(5)} x(p + p_k + \Delta p_k^{(5)}) \tag{19}$$

$$y_7 = \sum_k w_k^{(7)} x(p + p_k + \Delta p_k^{(7)}) \tag{20}$$

式中 y_3 、 y_5 、 y_7 ——卷积核为3、5、7的可变形卷积
为了增强非线性表示能力,采用不同的膨胀率(dilation)进行多尺度建模; 3×3 采用1; 5×5 采用2; 7×7 采用3。最终,将3种尺度的特征进行拼接后通过 1×1 卷积进行融合

$$y_{\text{final}} = \text{Conv}_{1\times 1}(\text{Concat}(y_3, y_5, y_7)) \tag{21}$$

MS-DLKA模块作为本模型的核心结构之一,引入可变形卷积,提高对玉米边缘和复杂结构的建模能力,通过多尺度的可变形卷积核,提高对不同尺度玉米植株特征的感知能力,有效提升了分割网络在玉米小样本场景下的表达能力与泛化能力。

3 结果与分析

3.1 试验设置

考虑到本研究数据集样本种类单一,为减少网络过深造成的过拟合风险,采用带有ImageNet预训练权值的ResNet50作为主干特征提取网络。为防止预训练的骨干网络从训练数据中学习特定类的标志,冻结了编码器的权值。整个模型基于PyTorch框架实现,使用Adam优化器进行优化,选择mIoU(平均交并比)和FB-IoU(前景-背景交并比)作为模型评价指标,采用交叉熵损失函数,图像的输入尺寸统一为400像素 $\times$ 400像素,学习率为 10^{-4} ,batch size为16,epoch为500。为了保证模型训练、测试的公平性,将模型对应的训练任务放到同一硬件平台上完成。试验平台内存为128 GB;中央处理器为

i9-12900k,主频为 5.1 GHz,核心数为 16;图形处理器为 NVIDIA RTX A6000,显存为 48 GB,核心数为 10 752。

3.2 不同主干网络对比试验分析

为评估本研究提出的 MSDANet 模型在不同主干网络结构下的语义分割性能,选取了 ResNet50、ResNet101、ResNet152 共 3 种经典的残差网络作为主干网络进行对比试验。在相同的试验配置和数据集下,分别记录模型在 1-shot 和 5-shot 设置下的 mIoU 和 FB-IoU 指标,试验结果如表 2 所示。

表 2 MSDANet 模型在不同主干网络下的平均精度

Tab.2 Average accuracy of MSDANet model under different backbone networks

主干网络	1-shot		5-shot	
	mIoU	FB-IoU	mIoU	FB-IoU
ResNet50	75.63	87.12	76.04	87.21
ResNet101	74.04	86.42	74.12	86.47
ResNet152	66.95	82.44	67.24	82.62

从表 2 可以看出,在 1-shot 和 5-shot 2 种小样本学习设定中,ResNet50 主干网络均表现出最优的分割精度。具体而言,ResNet50 在 1-shot 设置下 mIoU 达到 75.63%,FB-IoU 达到 87.12%;在 5-shot 设置下,mIoU 和 FB-IoU 分别达到 76.04%和 87.21%。模型在 3 种主干网络下的 1-shot 可视化结果如图 7 所示。

图 7 展示了不同主干网络 (ResNet50、ResNet101、ResNet152) 下的不同生长时期玉米植株图像的分割结果,每一行代表一种玉米株高的支持-查询图像对,红色区域表示预测掩膜,蓝色区域为伪标签。从图 7 可以看出,ResNet50 能够较好地分割出不同生长时期的玉米植株,边界清晰、结构完整,能准确捕捉叶片轮廓、主干和分支结构,整体预测与真实形态高度一致,仅在细长或弯曲叶片处出现轻微模糊或漏分,表现出良好的鲁棒性与泛化能力。相比之下,ResNet101 和 ResNet152 在处理复杂形态和尺度变化时,常出现边缘模糊、误分背景、叶片断裂或粘连等问题,整体分割效果不及 ResNet50。

图 8 展示了不同主干网络下的损失曲线图。尽

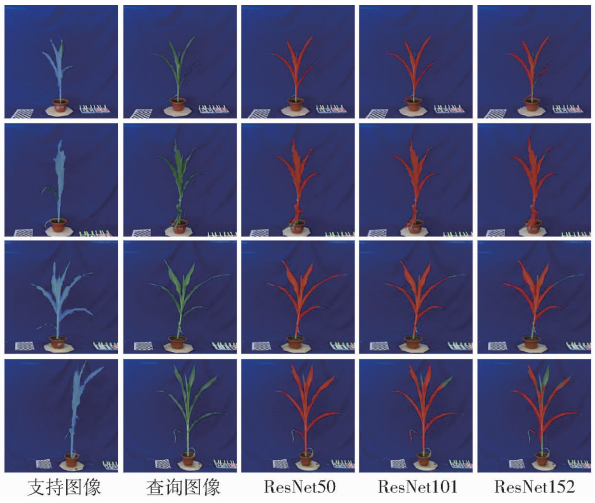


图 7 3 种不同主干网络下的模型预测结果

Fig.7 Prediction results of model using three different backbone networks

管 ResNet152 网络更深,理论上具有更强的表达能力和更大的感受野,但玉米植株图像小样本数据集样本单一,训练数据不足会导致过拟合,同时过深结构忽略了边缘与纹理等关键信息,在本研究中表现最差。ResNet101 虽优于 ResNet152,但仍不及 ResNet50。相比之下,ResNet50 在网络深度与特征提取能力之间实现了良好平衡,训练稳定、收敛快速,不易过拟合,并能有效捕捉叶片边缘和茎秆等关键特征。同时结合混合遮蔽机制和多尺度可变形大核卷积注意力机制,ResNet50 在保证精度的同时显著降低计算开销。综合精度、效率与泛化能力,本研究最终选用 ResNet50 作为 MSDANet 主干网络。

3.3 不同模型对比试验分析

为了进一步验证本研究提出的 MSDANet 模型在玉米小样本语义分割任务中的有效性,与当下主流的全监督和自监督小样本语义分割模型 (FewshotNet^[14]、DCAMA^[27]、HSNet^[28]、VAT^[29]) 进行对比试验。试验在 1-shot 与 5-shot 2 种训练方式下展开,试验结果如表 3 所示。

从表 3 可以看出,本研究提出的方法在 1-shot 和 5-shot 下均取得最高性能,优于绝大多数对比模型。在自监督任务中,1-shot 设置下的 mIoU 和 FB-IoU 分别达到 75.63% 和 87.12%,较 HSNet 和

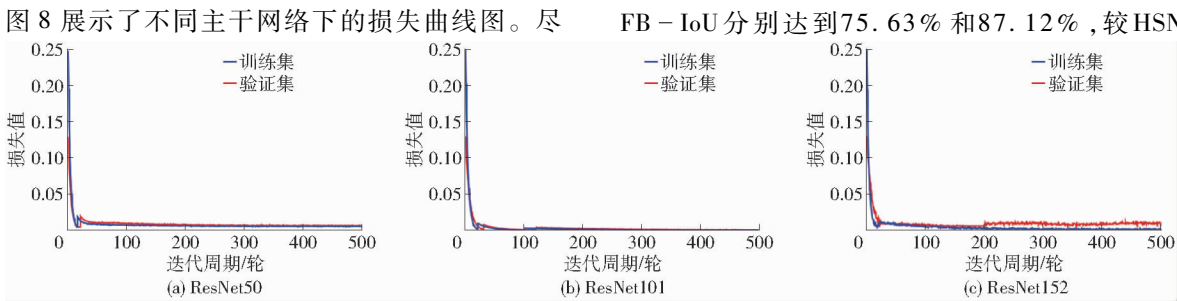


图 8 3 种不同主干网络下的损失曲线

Fig.8 Loss curves under three different backbone networks

表 3 MSDANet 与有监督和自监督模型的平均精度对比
Tab.3 Comparison of average accuracy between MSDANet and supervised and self-supervised models %

方法	模型	1-shot		5-shot	
		mIoU	FB-IoU	mIoU	FB-IoU
有监督	VAT	66.86	83.09	67.06	83.20
	HSNet	71.00	84.97	71.48	85.22
	DCAMA	72.73	87.18	73.11	87.42
自监督	HSNet	62.85	80.52	64.04	81.09
	FewshotNet	65.43	80.96	64.54	81.01
	MSDANet(本文方法)	75.63	87.12	76.04	87.21

FewshotNet 的 mIoU 提升约 12 个百分点,FB-IoU 提升 6 个百分点以上;5-shot 设置下同样保持领先,mIoU 提升超 11 个百分点。在全监督对比中,尽管 MSDANet 未使用任何人工标签,其在 1-shot 和 5-shot 下的 mIoU 分别比 DCAMA 高出 2.90、2.93 个百分点。这充分说明 MSDANet 在无需标签的前提下,依然能高效地捕获图像的结构信息与语义特征,表现出更强的特征表达能力与泛化能力,达到或超过全监督方法的效果。不同模型在 1-shot 设置下的结果如图 9 所示。

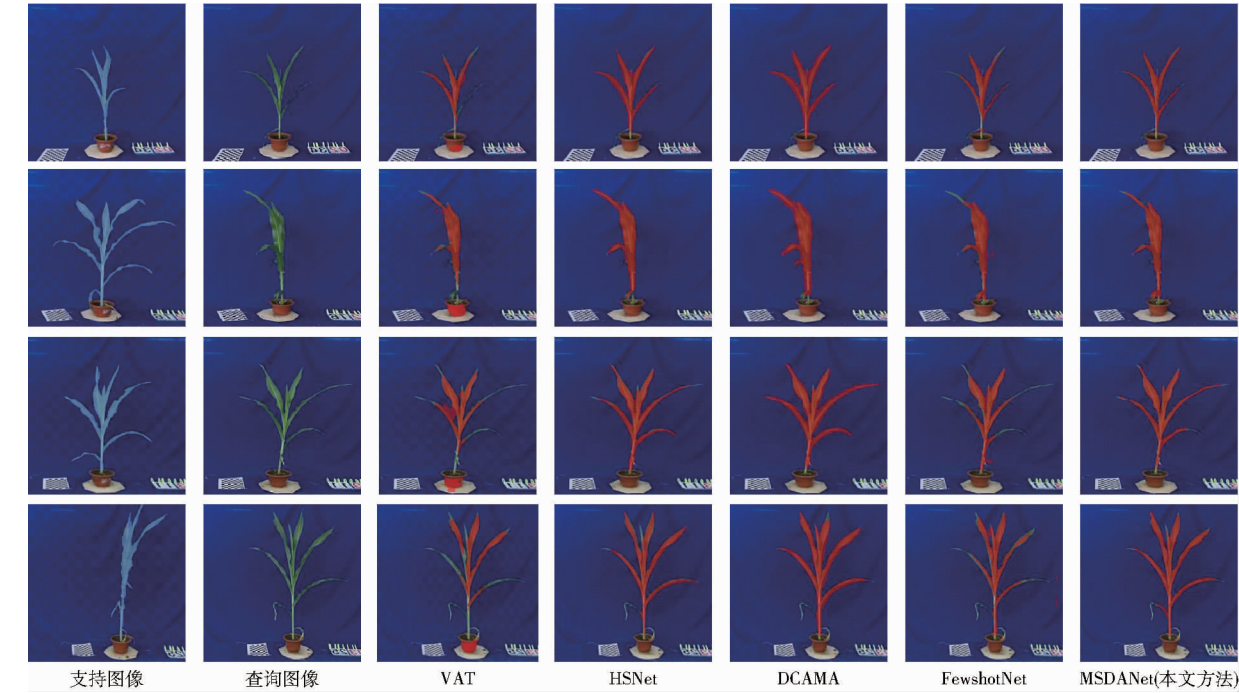


图 9 不同模型的分割结果对比

Fig.9 Comparative analysis of segmentation results obtained by different models

图 9 展示了本研究与其他主流模型在多个样本上的可视化分割结果对比。其中 VAT、HSNet 和 DCAMA 是全监督模型,FewshotNet 是自监督模型。从图 9 可以明显观察到,MSDANet 能更全面地捕捉植株的主干、分枝、叶片等结构信息,预测结果边缘清晰,形态完整。在目标轮廓提取、分割准确性、边界细节保持以及结构完整性等方面明显优于对比模型。相比之下,VAT 易出现断裂、粘连并误分花盆区域;HSNet 在细长叶片和分叉处存在断裂与漏分;DCAMA 虽具全局感知能力,但常导致粘连与过分割;FewshotNet 整体建模尚可,但细节捕捉不足,在叶片交叉或遮挡时准确性下降。而 MSDANet 凭借混合遮蔽机制和多尺度可变形大核卷积注意力机制,不仅准确捕捉了复杂结构,还有效抑制了背景干扰,确保了高质量的语义分割输出,表现出良好的鲁棒性。

3.4 消融试验

为了验证本研究所提出的混合遮蔽机制和多

尺度可变形大核卷积注意力机制的有效性,在相同试验设置下进行分析,试验结果如表 4 所示。从表 4 可以看出混合遮蔽机制和多尺度可变形大核卷积注意力机制的引入显著提升了模型分割性能。在 1-shot 设置下,MM 使 mIoU 和 FB-IoU 分别提升 4.28 个百分点和 2.73 个百分点,这表明混合遮蔽机制能够有效提升模型表征学习能力,从而有效提升了模型在复杂环境中的分割性能与泛化能力。其次,在 1-shot 设置下,MS-DLKA 使 mIoU 和 FB-IoU 分别提升 5.72、3.99 个百分点,验证了 MS-DLKA 在捕获多尺度特征和局部特征的有效性。当同时采用 MM 和 MS-DLKA 时,模型取得最佳性能,在 1-shot 设置下,mIoU 和 FB-IoU 分别提升 10.20、6.16 个百分点,优于单独引入任一模块的情况,进一步证明 MM 和 MS-DLKA 之间具有互补性,从而显著增强模型整体分割性能。

表 4 MM 与 MS-DLKA 消融试验

Tab.4 Ablation experiments of MM and MS-DLKA

		%			
模型		1-shot		5-shot	
MM	MS-DLKA	mIoU	FB-IoU	mIoU	FB-IoU
×	×	65.43	80.96	64.54	81.01
✓	×	69.71	83.69	70.08	83.95
×	✓	71.15	84.95	71.53	85.58
✓	✓	75.63	87.12	76.04	87.21

注：“✓”表示使用该模块，“×”表示未使用该模块。

4 结论

(1)本研究以不同生长时期的玉米为研究对象,设计了一种基于自监督学习的玉米植株图像小样本语义分割模型 MSDANet。

(2)本研究利用基于超像素的自监督学习,通过颜色空间转换、Felzenswalb 算法分割、去噪、膨胀等方法,自动生成目标前景的掩膜,可为图像构建初步监督信号,进而实现无人工标签的训练过程。

(3)提出基于自监督学习的小样本语义分割模型,设计混合遮蔽机制和多尺度可变形大核卷积注意力机制。通过支持掩膜遮蔽提取混合特征,实现

目标强化与背景抑制,使模型在特征提取阶段更准确地聚焦于支持图像的前景区域。同时利用多尺度特征和可变形卷积动态调整感受野,灵活感受玉米在不同尺度下的重要结构特征,增强模型的分割精度和泛化能力。模型在 1-shot 设置下 mIoU 和 FB-IoU 分别达到 75.63%、87.12%;在 5-shot 设置下,mIoU 和 FB-IoU 分别达到 76.04%、87.21%。

(4)本研究提出的小样本语义分割模型在定量评估中取得了优异成绩,在定性可视化中也展现了出色的结构完整性和边界准确性。其在 1-shot 和 5-shot 下均优于对比的主流全监督与自监督方法,相较当前最优的自监督小样本语义分割模型,在 1-shot 下的 mIoU 提升 10.20 个百分点,相较于当前主流全监督小样本语义模型,在 1-shot 下的 mIoU 提升 2.90 个百分点,验证了本方法在面对玉米植株小样本语义分割的有效性和鲁棒性。

(5)针对不同生长时期的玉米植株图形的小样本语义分割任务,MSDANet 模型可以在无人工标签的条件下实现高精度语义分割,节省了大量的人工标注成本,为不同生长时期的玉米图像分析与植物表型测量提供了解决方案和技术支持。

参 考 文 献

[1] 吴一全, 孟天亮, 吴诗姘. 图像阈值分割方法研究进展 20 年(1994—2014)[J]. 数据采集与处理, 2015, 30(1): 1-23. WU Yiquan, MENG Tianliang, WU Shihua. Research progress of image threshold segmentation methods in the past 20 years (1994—2014)[J]. Journal of Data Acquisition and Processing, 2015, 30(1): 1-23. (in Chinese)

[2] 王礼, 方陆明, 陈珣, 等. 基于 Lab 颜色空间的花朵图像分割算法[J]. 浙江万里学院学报, 2018, 31(3): 67-73. WANG Li, FANG Luming, CHEN Xun, et al. Flower image segmentation algorithm based on Lab color space[J]. Journal of Zhejiang Wanli University, 2018, 31(3): 67-73. (in Chinese)

[3] 钱生, 陈宗海, 林名强, 等. 基于条件随机场和图像分割的显著性检测[J]. 自动化学报, 2015, 41(4): 711-724. QIAN Sheng, CHEN Zonghai, LIN Mingqiang, et al. Saliency detection based on conditional random fields and image segmentation[J]. Acta Automatica Sinica, 2015, 41(4): 711-724. (in Chinese)

[4] 陈科尹, 吴崇友, 关卓怀, 等. 基于统计直方图 k-means 聚类的水稻冠层图像分割[J]. 江苏农业学报, 2021, 37(6): 1425-1435. CHEN Keyin, WU Chongyou, GUAN Zhuohuai, et al. Rice canopy image segmentation based on statistical histogram and k-means clustering[J]. Jiangsu Journal of Agricultural Sciences, 2021, 37(6): 1425-1435. (in Chinese)

[5] 丁永军, 杨文涛, 赵一龙. 基于一致性半监督学习的苹果叶片病斑分割模型研究[J]. 农业机械学报, 2024, 55(12): 314-321. DING Yongjun, YANG Wentao, ZHAO Yilong. Research on apple leaf spot segmentation model based on consistency semi-supervised learning[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(12): 314-321. (in Chinese)

[6] 王海燕, 江焯皓, 黎煊, 等. 基于弱监督数据集的猪只图像实例分割[J]. 农业机械学报, 2023, 54(10): 255-265. WANG Haiyan, JIANG Yehao, LI Xuan, et al. Instance segmentation of pig images based on a weakly supervised dataset[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(10): 255-265. (in Chinese)

[7] 高旺, 邓寒冰, 邢志鸿, 等. 温室番茄场景单目 RGB 模态向深度模态转换模型研究[J]. 农业机械学报, 2025, 56(6): 499-508, 574. GAO Wang, DENG Hanbing, XING Zhihong, et al. Research on monocular RGB-to-depth modality conversion model for greenhouse tomato scenes[J]. Transactions of the Chinese Society for Agricultural Machinery, 2025, 56(6): 499-508, 574. (in Chinese)

[8] 赵露露, 邓寒冰, 周云成, 等. 基于自生成标签的玉米苗期图像实例分割[J]. 农业工程学报, 2023, 39(11): 201-211. ZHAO Lulu, DENG Hanbing, ZHOU Yuncheng, et al. Instance segmentation of maize seedling images based on self-generated

- labels[J]. Transactions of the CSAE, 2023, 39(11): 201–211. (in Chinese)
- [9] 许华杰,肖毅烽. 基于多教师网络模型的半监督语义分割方法[J]. 计算机科学, 2023, 50(12): 279–284.
XU Huajie, XIAO Yifeng. Semi-supervised semantic segmentation method based on multi-teacher network model[J]. Computer Science, 2023, 50(12): 279–284. (in Chinese)
- [10] LIN X, LI C T, ADAMS S, et al. Self-supervised leaf segmentation under complex lighting conditions [J]. Pattern Recognition, 2023, 135: 109021.
- [11] 韦婷,李馨蕾,刘慧. 小样本困境下的图像语义分割综述[J]. 计算机工程与应用, 2023, 59(2): 1–11.
WEI Ting, LI Xinlei, LIU Hui. Survey on image semantic segmentation in dilemma of few-shot[J]. Computer Engineering and Applications, 2023, 59(2): 1–11. (in Chinese)
- [12] LU Z, HU S, ZHU X, et al. Simpler is better: few-shot semantic segmentation with classifier weight transformer [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 8741–8750.
- [13] AMAC M S, SENCAN A, BARAN B, et al. MaskSplit: self-supervised meta-learning for few-shot semantic segmentation [C] // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022: 1067–1077.
- [14] KARIMIJA FARBIGLOO S, AZAD R, MERHOF D. Self-supervised few-shot learning for semantic segmentation: an annotation-free approach [C] // International Workshop on Predictive Intelligence in Medicine. Cham: Springer Nature Switzerland, 2023: 159–171.
- [15] ZHOU T, WANG W, KONUKOGLU E, et al. Rethinking semantic segmentation: a prototype view [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 2582–2593.
- [16] WANG W, ZHOU T, YU F, et al. Exploring cross-image pixel contrast for semantic segmentation [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 7303–7313.
- [17] WANG K, LIEW J H, ZOU Y, et al. Panet: few-shot image semantic segmentation with prototype alignment [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 9197–9206.
- [18] ZHANG C, LIN G, LIU F, et al. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 9587–9595.
- [19] BOUDIAF M, KERVADEC H, MASUD Z I, et al. Few-shot segmentation without meta-learning: a good transductive inference is all you need? [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 13979–13988.
- [20] VAN GANSBEKE W, VANDENHENDE S, GEORGOULIS S, et al. Unsupervised semantic segmentation by contrasting object mask proposals [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10052–10062.
- [21] OUYANG C, BIFFI C, CHEN C, et al. Self-supervised learning for few-shot medical image segmentation [J]. IEEE Transactions on Medical Imaging, 2022, 41(7): 1837–1848.
- [22] 王炜航,张轶. MLDAC:多任务密集注意力计算自监督小样本分割方法[J]. 计算机工程与应用, 2025, 61(4): 211–221.
WANG Weihang, ZHANG Yi. MLDAC: multi-task dense attention computation self-supervised few-shot segmentation method [J]. Computer Engineering and Applications, 2025, 61(4): 211–221. (in Chinese)
- [23] 王森,杨克俭. 基于双线性插值的图像缩放算法的研究与实现[J]. 自动化技术与应用, 2008(7): 44–45, 35.
WANG Sen, YANG Kejian. Research and implementation of image scaling algorithm based on bilinear interpolation [J]. Automation Technology and Application, 2008(7): 44–45, 35. (in Chinese)
- [24] FELZENSZWALB P F, HUTTENLOCHER D P. Efficient graph-based image segmentation [J]. International Journal of Computer Vision, 2004, 59: 167–181.
- [25] RADHAKRISHNA A, APPU S, KEVIN S, et al. SLIC superpixels compared to state-of-the-art superpixel methods [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(11): 2274–2282.
- [26] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks [C] // Proceedings of the IEEE International Conference on Computer Vision, 2017: 764–773.
- [27] SHI X, WEI D, ZHANG Y, et al. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation [C] // European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 151–168.
- [28] MIN J, KANG D, CHO M. Hypercorrelation squeeze for few-shot segmentation [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 6941–6952.
- [29] HONG S, CHO S, NAM J, et al. Cost aggregation is all you need for few-shot segmentation [J]. arXiv Preprint, arXiv:2112.11685, 2021.