计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(7):84−94 [doi: 10.15888/j.cnki.csa.009167]

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

基于高分辨率图像的多尺度作物分类①

郭　金[1], 宋廷强[1,2], 巩传江[1], 孙媛媛[1,2], 马兴录[1], 范海生[3]

[1](青岛科技大学 信息科学技术学院, 青岛 266061)
[2](青岛科技大学 大数据学院, 青岛 266061)
[3](珠海市岭南大数据研究院 时空大数据研究室, 珠海 519080)
通信作者: 马兴录, E-mail: qdmxl@163.com

摘　要: 基于无人机平台获取的地面影像有着较高的空间分辨率, 但提供丰富的细节信息的同时, 也为农作物分类带来很多“干扰”, 尤其是在利用深度模型进行作物识别时, 存在边缘信息提取不充分及相似纹理作物误分, 导致分类效果欠佳等问题. 因此, 通过多尺度注意力特征提取的思路构建模型, 有效提取边缘信息, 提高作物分类精度. 所提出的多尺度注意力模型 (multi-scale attention network, MSAT) 通过多尺度块嵌入获取同一层级不同尺度的作物信息, 多尺度特征图被映射为多条序列独立地馈送到因子注意力模块中, 增强对农作物上下文信息的关注, 提高模型对地块边缘信息的提取, 因子注意力模块内置的卷积相对位置编码增强块内部局部信息的建模, 提高对相似纹理作物的区分能力, 最后通过融合局部特征与全局特征, 实现粗细双重信息的提取. 在水稻、甘蔗、玉米、香蕉和柑橘 5 种作物上的分类结果表明, MSAT 模型的 *MIoU* (mean intersection over union) 和 *OA* (overall accuracy) 指标达 0.816、98.10%, 验证了基于高分辨率图像的精细作物分类方法可行且设备成本低.

关键词: 无人机; 多尺度注意力; 作物分类; 因子注意力; 卷积相对位置编码

引用格式: 郭金,宋廷强,巩传江,孙媛媛,马兴录,范海生.基于高分辨率图像的多尺度作物分类.计算机系统应用,2023,32(7):84–94. http://www.c-s-a.org.cn/1003-3254/9167.html

Multi-scale Crop Classification Based on High-resolution Images

GUO Jin[1], SONG Ting-Qiang[1,2], GONG Chuan-Jiang[1], SUN Yuan-Yuan[1,2], MA Xing-Lu[1], FAN Hai-Sheng[3]

[1](College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China)
[2](School of Big Data, Qingdao University of Science and Technology, Qingdao 266061, China)
[3](Spatio-temporal Big Data Research Office, Zhuhai Big Data Research Institute, Zhuhai 519080, China)

Abstract: The ground images obtained by the unmanned aerial vehicle (UAV) platform have a high spatial resolution, but they also bring a lot of “interference” to crop classification while providing rich details. In particular, when depth models are used for crop recognition, there are problems such as insufficient edge information extraction and misclassification of similarly textured crops, which results in a poor classification effect. Therefore, a model is constructed by the idea of multi-scale attention feature extraction to effectively extract edge information and improve the accuracy of crop classification. The proposed multi-scale attention network (MSAT) obtains crop information on different scales at the same level through multi-scale block embedding. The multi-scale feature map is mapped into multiple sequences that are fed into the factor attention module independently, which enhances the attention to crop contexts and improves the model’s extraction ability of plot edge information. Moreover, the built-in convolutional relative position encoding of the factor attention module enhances the modeling of local information inside the module and the ability to distinguish similarly textured crops. Finally, the thickness information is extracted upon the fusion of local features and global features. The classification results of rice, sugarcane, corn, bananas, and oranges show that the mean intersection over union (*MIoU*)

① 基金项目: 山东省重点研发计划 (2019GGX101047); 山东省自然科学基金 (ZR2021QC120)
收稿时间: 2022-12-28; 修改时间: 2023-02-13; 采用时间: 2023-02-20; csa 在线出版时间: 2023-05-22
CNKI 网络首发时间: 2023-05-24

and overall accuracy (*OA*) of the MSAT model reach 0.816 and 98.10%, respectively, which verifies that the fine crop classification method based on high-resolution images is feasible, and the equipment cost is low.

Key words: unmanned aerial vehicle (UAV); multi-scale attention; crop classification; factorized attention; convolutional relative position encoding

精确的作物分类是精准农业的共同要求, 它对精准农业中的各个方向都有着重要意义, 包括作物长势分析[1]、作物产量估计[2]、作物种植管理[3]、农田信息提取[4] 等. 因此, 如何高效、精确地获取作物分布信息已成为当今精准农业的重要问题[5,6].

长期以来, 农田的作物种植信息主要通过综合统计或抽样调查获取的, 这种方法耗时耗力、主观性强且存在时滞问题[7]. 遥感技术的出现为农作物信息调查提供了便利, 其具有分辨率高、覆盖范围广的优势, 可以为相关部门提供更准确的农田信息[4,8,9], 而目前卫星遥感主要受天气和访问周期的限制, 因此低空、灵活的无人机成为作物信息调查的重要工具之一[10,11], 它具有成本低、不受云影响、更高的空间分辨率和可搭载特定传感器等优点[12–14].

无人机可见光图像作物分类还处于发展阶段, 使用植被指数和机器学习结合的分类方式对无人机可见光图像进行作物分类具有一定局限性. 原因是机载微型传感器采集的图像数据波段较少, 仅红、绿、蓝 3 个通道, 这导致光谱特征采集不足; 其次, 超高的空间分辨率带来了丰富的纹理信息, 对于具有相似纹理的作物更不易区分[15–17]. 深度学习的发展为充分利用高分辨率遥感图像中丰富的细节信息提供了保障[18], 近年来深度学习在计算机视觉领域得到广泛的应用, 其精度高, 鲁棒性好等优势为使用遥感图像进行作物识别提供新的思路[19,20]. 目前深度学习在计算机视觉领域的应用主要在自然场景图像上, 在图像分类、目标检测等领域相较于传统数字图像处理方法展示出明显优势. 基于遥感图像与自然场景图像有很高的相似性, 以及研究人员所做工作, 说明将深度学习技术应用在作物信息获取上是可行的, 已有研究人员对深度学习方法在无人机可见光图像上的作物分类进行了研究, 如 Chew 等人[21] 使用深度神经网络和迁移学习对卢旺达地区的无人机图像进行作物分类, 该模型在测试集上的 *F*1 指标达 0.86. 虽然深度神经网络可以很好应用于无人机图像的作物分类, 但在高分辨率图像中, 特征信息往往表现为多尺度, 单尺度在作物分类的适用性方面存在较大局限性, 因为小尺度往往会关注大面积特征, 而大尺度则会关注小面积特征[22–24]. 人类视觉系统遵循目标、环境、背景观察模式, 从高分辨率图像中提取信息应该具有类似的特征. 通常用于提取高分辨率可见光遥感图像中作物空间信息分布的卷积神经网络只能将遥感图像分割成局部图像块进行单独处理, 由于没有充分考虑不同作物目标的多尺度特征以及它们所处场景的上下文信息, 直接使用同一层级的单一尺度进行作物提取或识别的方法无法最大化作物间差距. 为了充分利用上下文信息, 需要对多尺度结构及其语义信息进行实际挖掘. Transformer[25] 的出现为上下文信息的提取提供了保障, 它通过多头注意力对全局依赖关系进行建模, 从而获取更丰富的全局信息, Dosovitskiy 等人[26] 提出的 ViT 模型首次将 Transformer 结构应用于图像分类领域, 很多研究人员也在此基础上将其应用于图像的下游任务中. Transformer 结构也进一步解决精准农业中精细作物分类问题, Reedha 等人[27] 将 ViT 模型应用于作物与杂草的识别, 解决了无人机图像中具有相似纹理特征的杂草和作物难以区分的问题. 也有研究人员将多尺度信息与 Transformer 进行结合, 提出了 MPViT 模型[28], 该模型探索了多尺度 patch embedding 和多路径结构在下游任务中的适用性, 在 ImageNet 数据集中表现出优异的结果. MPViT 模型的提出, 为多尺度多路径的 Transformer 在精细作物分类中的应用提供了研究基础.

针对使用单尺度卷积神经网络对作物误分及对地块边缘特征提取不充分问题, 本文提出了一种基于无人机图像的多尺度深度语义分割模型, 该模型对不同尺度的特征图进行同步的全局注意力操作, 以最大化不同尺度下的作物间差异, 提高模型对地块边缘特征的提取; 使用轻量的全局注意力降低并行结构的计算量, 提高模型的训练效率. 该模型进一步验证了使用高分辨率图像进行精细作物分类的可行性.

1 多尺度作物分类模型

1.1 模型结构

MSAT 模型通过深度可分离卷积获取多尺度特征图来获取同一网络层级下的不同尺度的作物信息, 这些信息被映射为一维序列独立的馈送到因子注意力模块中, 该模块通过建模远距离信息来获取作物的全局特征, 解决了对地块边缘信息提取不充分的问题, 内置的卷积相对位置编码弥补了对块 (patch) 内局部信息的建模, 提高了模块对作物细节信息的提取能力.

模型采用编码器-解码器结构, 编码器部分对图像特征进行逐级提取, 解码器部分还原低分辨率图像尺度并与浅层网络融合. MSAT 模型结构如图 1 所示, 编码器部分由骨干网络 Efficientnet-b1[29] 和多尺度因子注意力模块组成, 其中, Efficientnet-b1 网络对输入图像 $X \in \mathbb{R}^{H\times W\times C}$进行 4 次下采样, 逐步提取局部特征得到 $X_4 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C_4}$, 多尺度因子注意力模块首先对$X_4$使用卷积核分别为 1×1、3×3、5×5 和 7×7 的深度可分离卷积获取多尺度特征图, 再对其进行多尺度的 patch embedding 获取不同尺度下的特征序列X_4^1, X_4^2, X_4^3, X_4^4作为卷积注意力模块的输入; 卷积注意力模块对多尺度特征进行全局特征提取, 它相较于传统 Transformer 的优势是在获取全局信息的同时保证局部信息的关联性且计算效率更高, 传统的 Transformer 通过建模 patch 之间的关联性来获取上下文信息, 但无法构建 patch 内部的结构信息, 导致其对作物地块形状特征提取较为充分, 而忽略了作物的纹理特征, 从而出现对具有相似纹理的作物分类不准确的问题, 且其相似度矩阵的计算也会带来较大的参数量; 多尺度的卷积注意力机制相较于 DeepLabv3+[30] 的空间池化金字塔可提取更丰富的局部特征, DeepLabv3+中的空洞卷积可以捕获上下文信息, 但空洞率的增加也降低了其对局部信息的获取, 导致作物的分类不准确. 本文的卷积注意力模块使用因子注意力机制和卷积相对位置编码来降低模块的计算量并充分获取 patch的内部特征, 相对位置编码和卷积之间的内在联系使得可以使用类似卷积的操作来实现有效的自注意力. 最后将获取的全局特征$\{G^1, G^2, G^3, G^4\} \in \mathbb{R}^{N\times C}$与局部特征$L$融合得到$X_4^{\text{att}}$. 解码器部分将编码器部分的输出$X_4^{\text{att}}$与浅层网络特征 $X_2 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C_2}$融合, 实现全局信息、局部信息的获取和下采样过程中丢失的小目标信息的补充, 从而实现大、小作物地块的精准定位、分割.

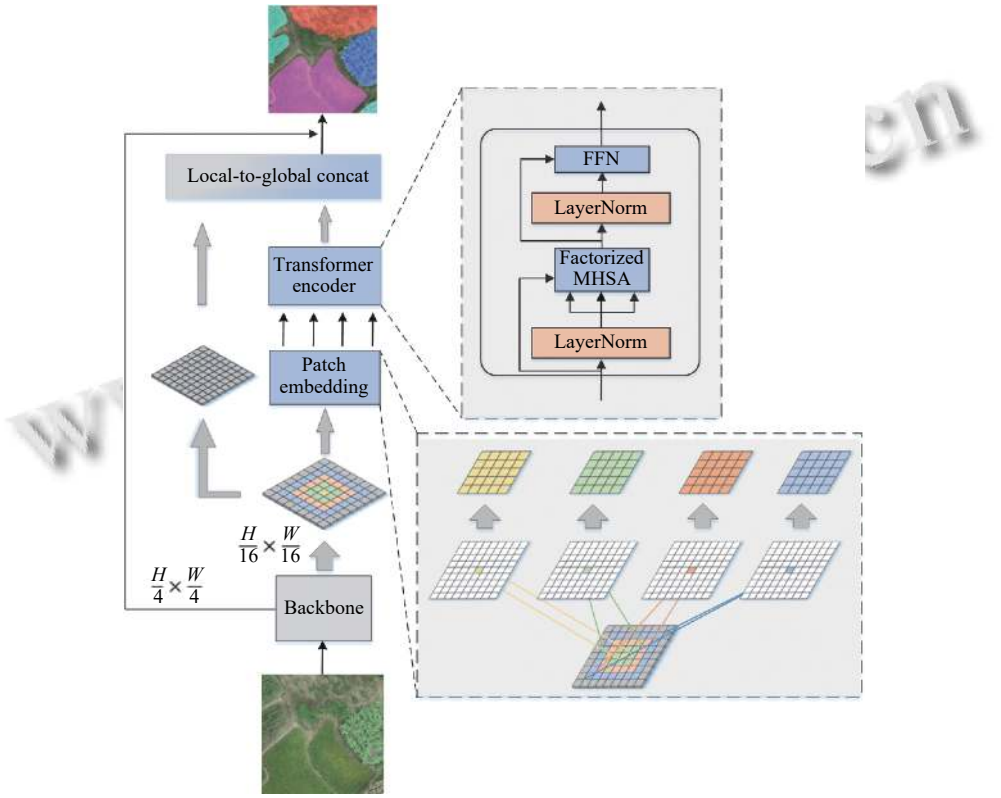


图 1 MSAT 模型结构

1.2 多尺度 patch embedding

本文设计了一个多尺度 patch embedding 层, 如图 2 所示, 该层通过对深度可分离卷积设置不同的填充值和卷积核大小来获取粗粒度和细粒度的作物特征图, 再使用重叠卷积对不同尺度的特征图进行 patch embedding, 最后对其进行 1D-Reshape 特征映射, 将 $X_4^j \in \mathbb{R}^{H_4 \times W_4 \times C_4}$ 映射为 $X_4^j \in \mathbb{R}^{N \times C}$, 其中, $j \in \{1,2,3,4\}$, N 和 C 分别表示 patch 的数量和嵌入维度.

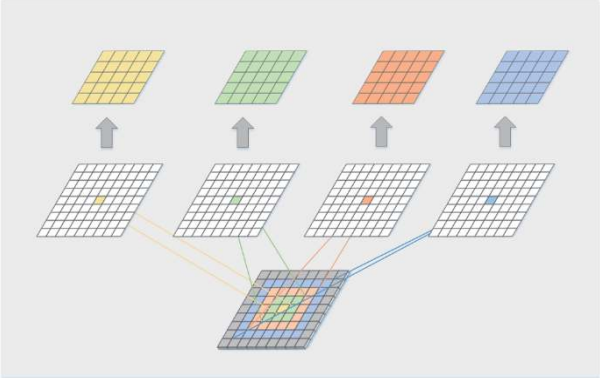


图 2 多尺度 patch embedding

DeepLabv3+模型中的空洞卷积通过设置不同的空洞率来获取作物的全局特征, 受像素间空洞的影响, 空洞卷积无法充分地建模相邻像素的依赖关系, 导致对作物的纹理特征关注较少, 若使用空洞卷积获取多尺度特征, 因子注意力模块在建模全局特征时就无法兼顾局部特征. 相反, 深度可分离卷积对相邻像素进行建模并通过填充来补充边缘特征, 相较于空洞卷积可以捕获更丰富的作物细节信息. 因此模型使用深度可分离卷积来获取多尺度特征, 具体而言, 对前一阶段的输出进行不同比率的填充, 补充对作物特征图边缘特征的提取, 深度可分离卷积可以通过调整填充比率来控制图像大小, 设置卷积核大小分别为 1×1、3×3、5×5 和 7×7, 以获取由细到粗的多尺度特征图. 将获取的多尺度特征图输入到重叠卷积中, 其中, 卷积核为 K (即 patch 的大小), 步长为 S, 填充为 P, 重叠卷积前后特征图尺寸变化如式 (1) 所示; 最后进行一维特征映射, 将 16×16 的 patch 映射为一维的特征序列, patch embedding 层可以通过定义输出的维度来控制序列的长度.

$$H' = \frac{H-K+2P}{S}+1,\ W' = \frac{W-K+2P}{S}+1 \tag{1}$$

1.3 卷积注意力机制

卷积注意力机制包括因子注意力机制、卷积相对位置编码和卷积位置编码 3 部分, 该模块通过在因子注意力模块中实现相对位置嵌入, 并采用高效的卷积块, 其生成的映射用于后续的前馈网络. 卷积注意力模块的结构如图 3 所示, 输入特征先进行卷积位置编码, 再通过嵌入相对位置编码的因子注意模块.

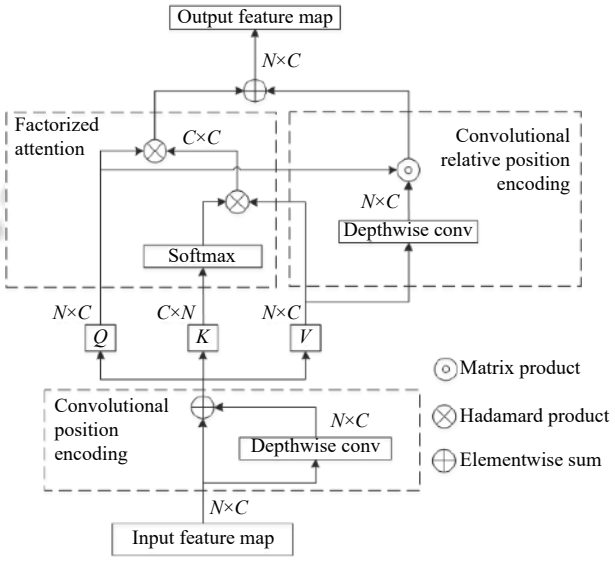


图 3 卷积注意力模块

卷积注意力的计算公式如式 (2) 所示, $\frac{Q}{\sqrt{C}}(Softmax(K)^{\mathrm{T}}V)$ 表示因子注意力机制, EV 为卷积相对位置编码.

$$ConvAtt(X) = \frac{Q}{\sqrt{C}}(Softmax(K)^{\mathrm{T}}V) + EV \tag{2}$$

(1) 因子注意力机制

Transformer 中的自注意力机制使用线性变换 W^Q, 和 W^V 将每个 $X_i \in \mathbb{R}^{N \times C}$ 投影到相应的查询、键和值向量中, 传统的注意力机制如式 (3) 所示. 原始的 Transformer 计算量巨大, 为减轻并行结构中计算的负担, 模型参照 CoAT[31] 使用了两个函数 $\phi(\cdot),\psi(\cdot): \mathbb{R}^{N \times C} \to \mathbb{R}^{N \times C'}$ 来近似 *Softmax* 注意力, 其中, $\phi(\cdot)$ 为单位函数, $\psi(\cdot)$ 为 *Softmax*, 从而降低高分辨率图像的计算复杂度, 因子注意力机制的计算公式, 如式 (4) 所示:

$$Att(X) = Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{C}}\right)V \tag{3}$$

$$\begin{aligned} FactorAtt(Q,K,V) &= \phi(Q)(\psi(K)^{\mathrm{T}}V) \\ &= \frac{Q}{\sqrt{C}}(Softmax(K)^{\mathrm{T}}V) \end{aligned} \tag{4}$$

其中, $Q,K,V \in \mathbb{R}^{N\times C}$是查询 (queries)、键 (keys) 和值 (values) 的线性映射.

(2) 卷积相对位置编码

没有位置编码, Transformer 就仅由线性层和注意力层组成, 这样就限制了在每个 patch 内建模作物细节信息的能力, 无法计算其局部特征差异, 局部特征获取不充分, 则会导致相似纹理的作物被混淆, 为提高模型 patch 内部信息的建模能力, CoAT 集成了相对位置编码EV, 卷积相对位置编码的查询和值之间基于位置的局部关系, 可以建模相邻像素的依赖关系, 充分获取作物的纹理信息, 增大作物之间的特征差异. 使用深度卷积来计算EV, 卷积相对位置编码公式如式 (5) 所示:

$$EV = Q \circ DepthwiseConv1D(P,V) \tag{5}$$

其中, ∘表示 Hadamard 乘积.

将查询中的每个通道、位置编码和值向量视为内部头, 对于每个内部头l, 都有:

$$E_{ij}^{(l)} = M(i,j)q_i^{(l)}p_{j-i}^{(l)} \tag{6}$$

$$EV_i^{(l)} = \sum_j E_{ij}^{(l)}v_j^{(l)} \tag{7}$$

其中, $M(i,j)$表示指示函数, q_i为查询向量, v_j表示查询向量, p_{j-i}表示位置编码.

(3) 卷积位置编码

与绝对位置编码相似, 卷积位置编码的思想是将位置编码直接插入到输入图像特征中, 以丰富相对位置编码的效果. 首先, 将输入特征进行深度可分离卷积, 再按照标准绝对位置编码将产生的位置感知特征与输入特征进行融合, 从而完成对输入特征的卷积位置编码.

2 数据集介绍

2.1 卷积注意力机制

研究区位于广西壮族自治区贵港市的港南区和桂平市 (北纬 22°39′–24°2′, 东经 109°11′–110°39′). 该地属亚热带季风气候区, 年均气温 21.5 ℃, 年均降雨量 1 600 mm. 各季节气候特点是: 冬季偏暖, 降水偏少; 春季温度正常, 降水稍偏少; 夏季气温偏高, 降水正常; 秋季偏暖, 降水偏少. 该地区为水稻、甘蔗、玉米、香蕉和柑橘等作物提供了良好的种植条件.

研究选取了贵港市主要粮食作物水稻、玉米, 经济作物甘蔗、香蕉和柑橘作为分类对象. 各作物物候日历如图 4 所示. 水稻一般 4 月份和 8 月份种植, 分别于 7 月份和 11 月份收割; 玉米则在 3 月份和 7 月份种植, 6 月份和 10 月份收割, 生长周期短; 而甘蔗生长周期较长, 一般 3 月份播种, 12 月份收割; 香蕉和柑橘则一般成熟于 9–11 月份.

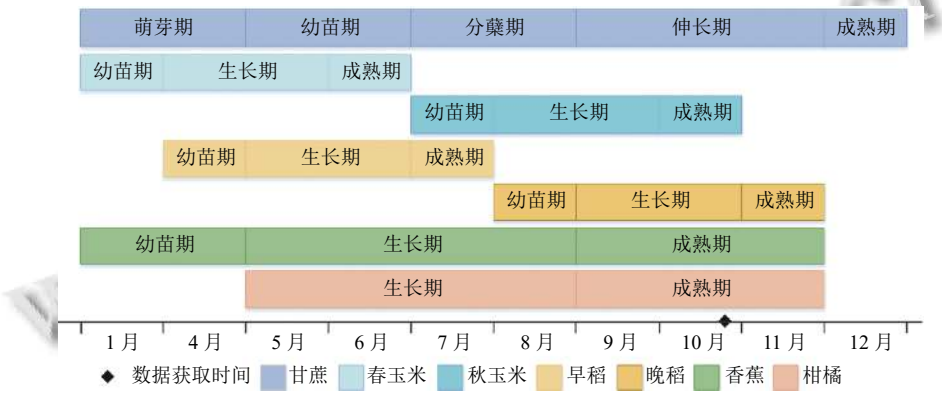


图 4 研究区主要作物物候日历

2.2 无人机图像采集

本文实验所用的无人机型是大疆 M300 RTK 无人机, 搭载精灵 4RTK 相机, 航拍时间为 2021 年 10 月 28 日, 拍摄当天天气晴朗、风轻, 飞行高度设置为 150 m, 地面分辨率设置为 6 cm. 采集的图像包含 3 个波段: 红色、绿色和蓝色. 我们根据研究区的实际情况采集 7 个作物信息丰富的区域进行实验, 总面积为 550 公顷.

从物候日历图 (图 4) 和作物采样图 (图 5) 可以看出, 10–11 月之间作物都处于成熟区、纹理特征差异较大. 该时期作物分别为晚稻、秋玉米、甘蔗、香蕉和柑橘. 对该时期的作物进行实地样本采集, 如图 5 所示, 可以看出, 水稻株苗密集、呈浅绿色; 甘蔗枝叶

错综复杂且种植密集; 玉米与甘蔗的叶面结构较为相似, 但玉米叶面发黄、种植有序且间距大; 香蕉相较于玉米和甘蔗叶片更大, 成放射状; 柑橘种植间隔大, 成点状排列.

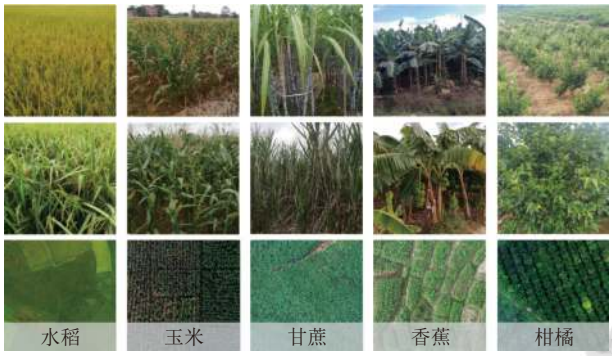


图 5 研究区作物采样图及正射影像

2.3 数据预处理

用于分类的 5 类研究区作物为水稻、玉米、甘蔗、香蕉和柑橘, 根据实地作物采样对研究区域进行田块矢量数据采集及点位采样, 从而根据田块矢量数据及点位数据对采集区域的正射影像进行目视解读, 并使用 ArcGIS 软件对各类作物进行真实值标注, 标注结果如表 1 所示.

表 1 研究区作物类别及标注信息

作物类别	值	面积 (公顷)	样本量
水稻	1	83	3574
玉米	2	27	2882
甘蔗	3	24	2125
香蕉	4	21	1970
柑橘	5	48	3326

本实验根据已标注的无人机遥感图像构建样本数据库, 将 7 个区域的数据裁剪为 512×512 大小的图片, 如表 1 所示, 各样本的样本量比例为水稻:玉米:香蕉:甘蔗:柑橘=5:4:3:3:5, 样本较为均衡, 无长尾现象, 将所有数据划分为训练集、验证集和测试集, 选取其中 6 个区域的数据作为训练集和验证集, 包含 9810 张图片, 另一个区域作为测试集, 包含 3450 张图片.

3 实验结果及分析

3.1 实验设置

实验环境使用的 CPU 为 Intel Xeon W-2245, 128 GB 内存, 显卡型号为 NVIDIA GeForce RTX 3060Ti, CUDA 版本 11.3, 使用 PyTorch 框架搭建模型. 模型训练使用 Adam 优化器, 动量设置为 0.9, 权重衰减设置为 0.000 5; 学习率衰减策略使用指数衰减, 初始学习率设为 0.001; 批次大小设置为 8, 迭代次数设置为 100.

训练使用交叉熵损失函数和 Dice 损失函数的加权平均. 交叉熵损失函数表达式如式 (8) 所示, 其中, $p(x)$为真实值的概率分布, $q(x)$为预测值的概率分布, n表示类别数.

$$L_{\text{cross}} = -\sum_{i=1}^{n} p(x_i)\ln(q(x_i)) \tag{8}$$

Dice 损失的函数表达式如式 (9) 所示, 其中,$|x\cap y|$表示真实样本与预测集的交集, $|x|$和$|y|$分别表示真实值和预测值的个数.

$$L_{\text{Dice}} = 1 - \frac{2|x\cap y|}{|x|+|y|} \tag{9}$$

3.2 评估指标

为准确定量地验证模型在测试集中准确性, 本实验采用 *IoU* (intersection over union)、*MIoU* 和总体分类精度 (*OA*) 来评估分类结果. 其中, *IoU* 表示真实值与预测值的交并比, *MIoU* 为平均交并比, 该指标可以反映分割结果和真实结果之间的重合程度. 总体分类精度是所有正确预测数与总数的比值. $TP = p_{ii}$表示真正例, $FP = p_{ji}$表示假正例, 表示将其他类j预测为正确类i, $FN = p_{ij}$表示假负例, $TN = p_{jj}$表示真负例. *IoU*、*MIoU* 和 *OA* 的计算如式 (10)–式 (12) 所示:

$$IoU = \frac{TP}{TP+FN+FP} \tag{10}$$

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{11}$$

$$OA = \frac{TP+TN}{TP+FN+FP+TN} \tag{12}$$

3.3 对比实验

本实验对广西壮族自治区贵港市的水稻、玉米、甘蔗、香蕉和柑橘进行分类, 选用的数据集为贵港市 7 个区域的无人机图像数据, 图片大小为 512×512, 训练集包含 9810 张图片, 测试集包含 3450 张图片. 为评估 MSAT 模型在农作物分类中的有效性, 本实验将其与 UNet[32]、UNet++[33]、LinkNet[34]、DeepLabv3+模型进行了对比, 从定性和定量两方面分析评估结果, 在农作物数据集上的分类混淆矩阵和精度评估如图 6 和表 2 所示.

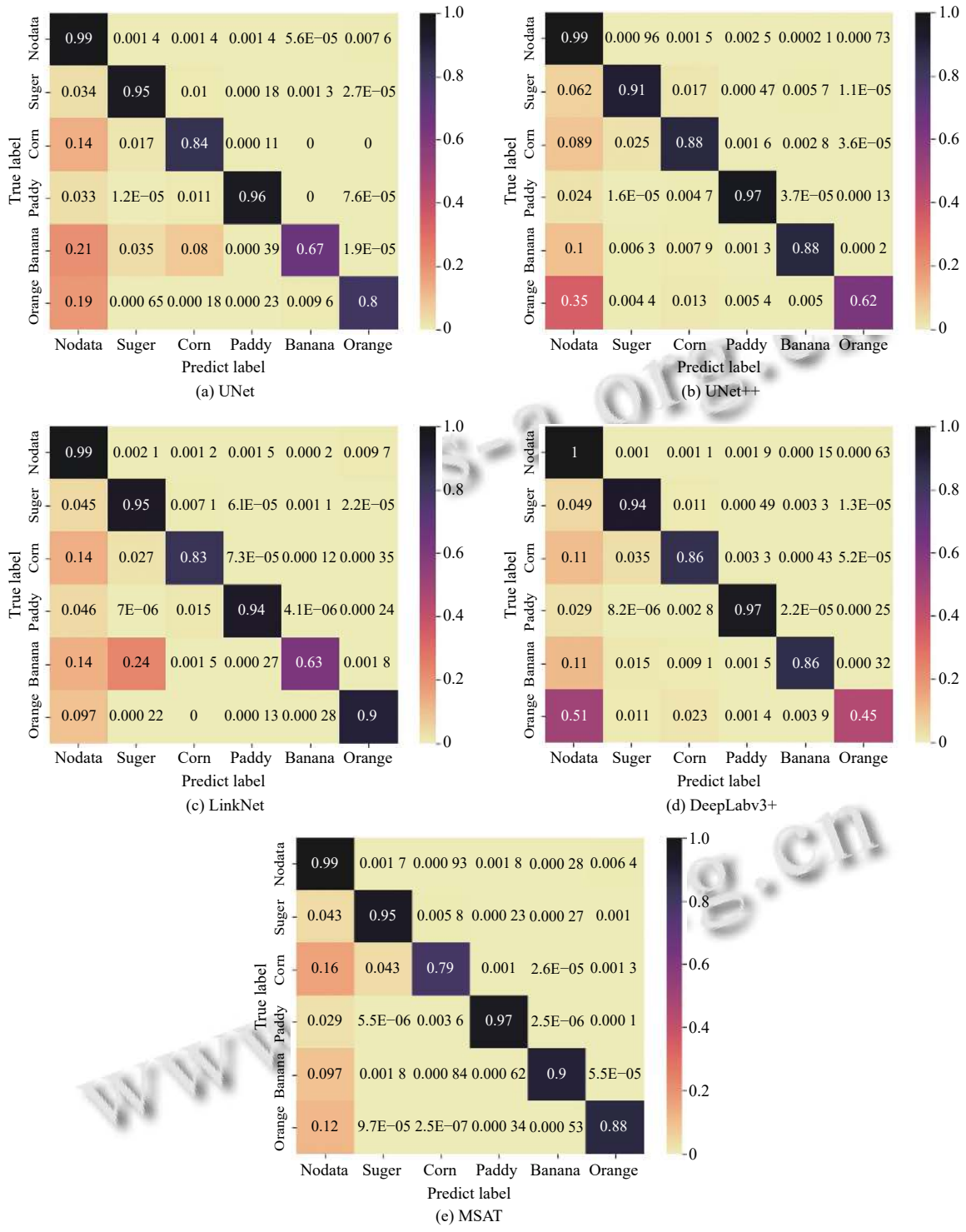


图 6 各模型的分类结果可视化

由图 6 和表 2 可以看出 MSAT 模型相较于其他模型在农作物数据集上有更高的分割精度. 具体而言, 如表 2 所示, 与 DeepLabv3+模型相比, 本文提出的多尺度因子注意力模型 MSAT 在 *MIoU* 和 *OA* 指标上分别提高了 8.6%, 1.32%. 与 UNet 模型、UNet++模型和 LinkNet 模型相比, *MIoU* 和 *OA* 分别提高了 6.8%、5.2%、7.5% 和 0.53%、0.66%、0.42%. 从图 6 的 5 种模型的混淆矩阵可以看出, 柑橘被误分为背景类, 原因

是柑橘种植区内有其他呈点状种植的树种与柑橘纹理较相似; 图像的分辨率较高, 且玉米、甘蔗和香蕉 3 种作物的叶面结构较为相似, 从而增加了模型对 3 者的区分难度, 导致了这 3 类作物之间不同程度的误分. 而 MSAT 模型的多尺度全局注意力机制使其不止关注作物纹理特征, 更可以结合作物的种植背景和种植结构来区分, 如玉米植株的排列较为整齐, 排与排之间有空隙, 地块边缘整齐; 甘蔗种植紧密, 植株之间无明显空隙, 地块边缘不规则, 呈弯曲状; 香蕉植株之间距离较大, 排列无序且稀疏, 地块边缘无不规则. 如果仅依据 3 种作物的叶面纹理特征进行区分, 则无法突出玉米、甘蔗和香蕉 3 类作物之间的差异, MSAT 充分获取上下文信息, 可以从叶面纹理结构和种植背景、种植结构 3 方面对 3 类作物进行分类, 同时也增强了模型对柑橘和其他树种的区分能力. MSAT 与其余 4 类模型相比, 在水稻、玉米、香蕉、柑橘 4 类作物上都有提升, 但甘蔗的分割精度相较于 UNet 下降了 0.021, 结合混淆矩阵, 分析原因可能是与玉米存在一定的误分, 该问题可通过增加玉米和甘蔗的样本量来提升模型对二者的区分能力.

表 2 各模型在农作物数据集上的分割精度

模型	*IoU*					*MIoU*	*OA* (%)
	水稻	玉米	甘蔗	香蕉	柑橘		
UNet	0.916	0.666	0.855	0.611	0.693	0.748	97.57
UNet++	0.895	0.683	0.834	0.797	0.613	0.764	97.44
LinkNet	0.897	0.693	0.763	0.599	0.754	0.741	97.68
DeepLabv3+	0.913	0.662	0.826	0.802	0.449	0.73	96.78
MSAT	0.918	0.701	0.834	0.849	0.779	0.816	98.10

图 7 为 5 类模型对水稻、玉米、甘蔗、香蕉和柑橘 5 种作物分割结果, 从左到右依次为真实值标注, MSAT、UNet、UNet++、LinkNet 和 DeepLabv3+模型的分割结果. 结合表 2 与图 7, 分析可知基于卷积网络的 UNet、UNet++和 LinkNet 对玉米、甘蔗和香蕉的分割结果较差, 且对地块边缘的分割不规则, 香蕉大多被误分为玉米和甘蔗. 原因是 UNet、UNet++和 LinkNet 这 3 个模型主要通过建模相邻像素信息来获取局部特征, 对上下文信息获取不充分, 导致玉米、甘蔗和香蕉这 3 类有相似纹理的作物无法很好地区分, 分割边缘不平滑. DeepLabv3+模型具有多尺度卷积结构, 通过空洞卷积扩大感受野以获取更充分的上下文信息, 但该结构对局部信息的构建能力较弱, 导致作物也呈现一定程度的误分, 而 MSAT 模型基于多尺度的因子注意力机制, 多尺度结构可以充分获取全局信息, 因子注意力机制可以补充局部信息, 二者的结合具有强大的全局和局部特征表达能力, 提高作物的识别能力.

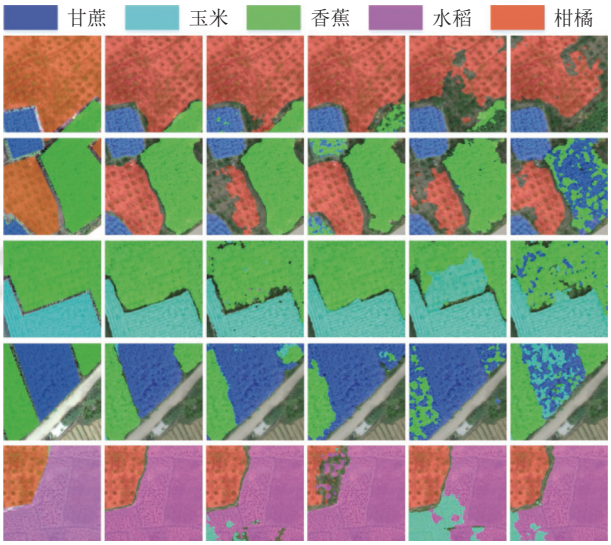
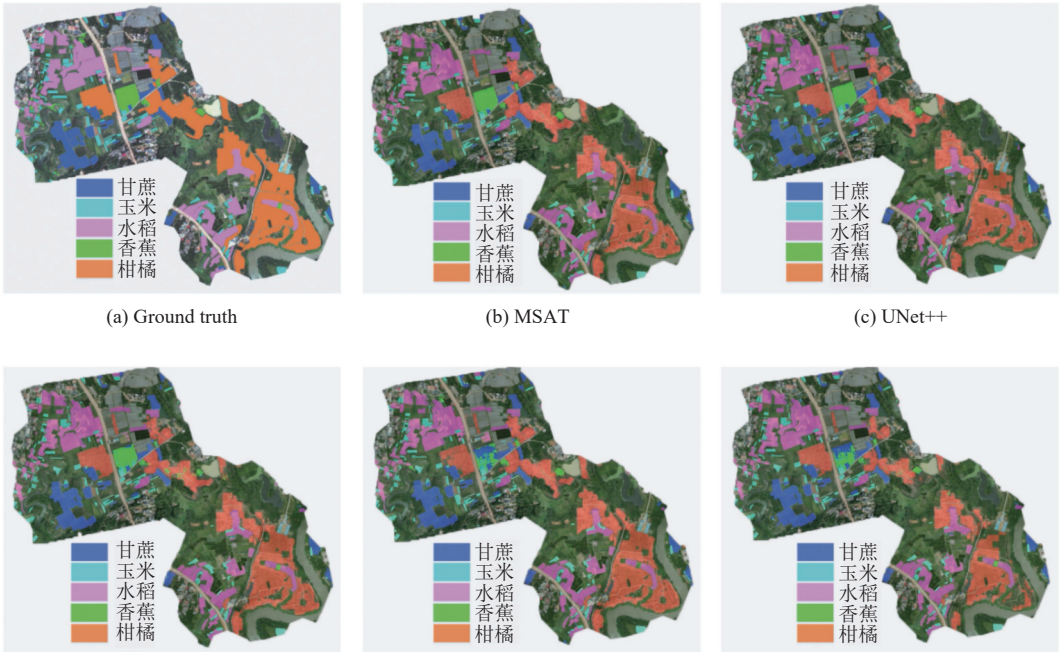


图 7 各模型预测结果对比图

图 8 展示了 5 类模型在整个测试区域的农作物分割结果, 依次为真实值标注、MSAT、UNet++、UNet、LinkNet 和 DeepLabv3+模型. 其中, LinkNet 和 DeepLabv3+对香蕉作物有不同程度的误分, LinkNet 模型对香蕉的分类精度不高, *IoU* 只有 0.599, DeepLabv3+模型将部分柑橘误分为背景, 使得模型对柑橘的分类结果较低, *IoU* 为 0.449, 图 6 的混淆矩阵也显示了 LinkNet 将香蕉误分为甘蔗的概率和 DeepLabv3+将柑橘误分为背景的概率较大. UNet++模型和 UNet 模型分别在柑橘和香蕉的分类上有一定程度的误分. 相比之下, MSAT 模型对作物的分类效果更好, 无明显误分现象, 且对作物地块的提取也更完整. 实验选取的测试区域各作物种植较分散, 存在多种作物交叉种植情况, 如图 8 中的 ground truth 所示, 左上角和右下角区域交叉分布较为密集, 主要容易产生交叉种植的作物为住房周围的玉米、水稻及河道边的香蕉, 对比图 5 的 ground truth 与 MSAT, 交叉种植的作物无明显误分现象, 中间区域存在甘蔗和柑橘误分的情况, 对比原始图像发现原因为地块边缘的水稻存在小部分倒伏现象, 纹理较为混

乱. 根据图 6(e) 混淆矩阵分析各类作物的分类情况, 无明显误分现象, 存在少量玉米、香蕉、甘蔗预测为背景值的情况, 分析原因可能是存在部分收割后的玉米、甘蔗残株, 及种植较为零散的香蕉.



(a) Ground truth (b) MSAT (c) UNet++

(d) UNet (e) LinkNet (f) DeepLabv3+

图 8 各模型在测试集上的全域预测结果

3.4 消融实验

为进一步验证多尺度因子注意力模块的有效性, 本文设置了一组消融实验. 实验将本文提出的多尺度因子注意力模型 MSAT 与原模型 DeepLabv3+、多尺度卷积为空洞卷积的因子注意力模型 MSAT-AC 进行比较, 通过对比三者在农作物数据集上的分割精度来讨论多尺度因子注意力结构的有效性. 如表 3 所示, 本文提出的基于深度可分离卷积的因子注意力模型 MSAT 分割结果优于基于空洞卷积的因子注意力模型 MSAT-AC 和 DeepLabv3+模型. 对比 DeepLabv3+模型与 MSAT-AC 模型, 直接在 DeepLabv3+模型上加入多尺度的因子注意力模块分割精度不增反减, 原因是因子注意力机制中的卷积相对位置编码可以增强多尺度特征图中对局部信息的建模能力, 而空洞卷积后的特征图相邻像素之间无直接联系, 这导致 CoAT 无法有效建模 patch 内的局部信息且在建模上下文信息时也会漏掉一些信息, 从而导致直接在 DeepLabv3+模型上应用因子注意力机制后模型性能降低. 为解决这一问题, 使用深度可分离卷积替换空洞卷积, 对比 MSAT 模型和 MSAT-AC 模型, *MIoU* 和 *OA* 分别提升了 11.1%、1.95%. 原因是深度可分离卷积弥补了空洞卷积相邻像素相关性弱的不足, 利于 CoAT 对多尺度特征图局部信息和上下文信息的建模.

表 3 验证多尺度注意力模块在农作物数据集上的有效性

模型	DWConv	AtrousConv	CoAT	*MIoU*	*OA* (%)
DeepLabv3+	—	√	—	0.73	96.78
MSAT-AC	—	√	√	0.705	96.15
MSAT	√	—	√	0.816	98.10

4 结论与展望

本文提出了一种基于多尺度因子注意力的农作物分类模型 MSAT. 该模型对无人机图像进行逐级特征提取, 并对第 4 阶段的特征图进行多尺度嵌入, 通过一个多尺度注意力机制, 该机制解决了常规 Transformer 无法充分获取图像局部信息的问题, 增加了纹理相似作物之间的差距, 增强了对边缘信息的提取能力, 提高了分类结果的准确度. 在农作物数据集上的对比实验表明, 该模型对农作物的总体分类精度达到了 98.1%, 平均交并比达到 0.861, 相较于对比模型都有一定提升, 尤其对玉米、甘蔗和柑橘的分类结果更佳. 为进一步

验证 MSAT 模型的性能, 设计了一组消融实验, 探索基于深度可分离卷积的多尺度注意力机制的有效性. 多尺度模型要求并行的 Transformer 结构需要足够轻量, 较大的注意力结构导致较慢的训练速度, 但降低模型规模也会一定程度影响模型分类准确度, 针对该问题, 后续将继续探索兼顾训练速度和分类精度的多尺度模型. 本研究为多尺度注意力机制在农作物分类领域的探索提供了研究基础, 并进一步验证了无人机图像在精准农业中的适用性.

参考文献

1 Nasirzadehdizaji R, Cakir Z, Sanli FB, *et al*. Sentinel-1 interferometric coherence and backscattering analysis for crop monitoring. Computers and Electronics in Agriculture, 2021, 185: 106118. [doi: 10.1016/j.compag.2021.106118]

2 周亮, 慕号伟, 马海姣, 等. 基于卷积神经网络的中国北方冬小麦遥感估产. 农业工程学报, 2019, 35(15): 119–128. [doi: 10.11975/j.issn.1002-6819.2019.15.016]

3 Mubin NA, Nadarajoo E, Shafri HZM, *et al*. Young and mature oil palm tree detection and counting using convolutional neural network deep learning method. International Journal of Remote Sensing, 2019, 40(19): 7500–7515. [doi: 10.1080/01431161.2019.1569282]

4 Yang GF, He Y, Yang Y, *et al*. Fine-grained image classification for crop disease based on attention mechanism. Frontiers in Plant Science, 2020, 11: 600854. [doi: 10.3389/fpls.2020.600854]

5 Tatsumi K, Yamashiki Y, Morante AKM, *et al*. Pixel-based crop classification in Peru from Landsat 7 ETM+ images using a random forest model. Journal of Agricultural Meteorology, 2016, 72(1): 1–11. [doi: 10.2480/agrmet.D-15-00010]

6 Kumar P, Prasad R, Choudhary A, *et al*. A statistical significance of differences in classification accuracy of crop types using different classification algorithms. Geocarto International, 2017, 32(2): 206–224.

7 Wu MQ, Yang LC, Yu B, *et al*. Mapping crops acreages based on remote sensing and sampling investigation by multivariate probability proportional to size. Transactions of the Chinese Society of Agricultural Engineering, 2014, 30(2): 146–152.

8 Du XD, Cai YH, Wang S, *et al*. Overview of deep learning. Proceedings of the 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC). Wuhan: IEEE, 2016. 159–164.

9 Liang J, Zheng ZW, Xia ST, *et al*. Crop recognition and evaluation using red edge features of GF-6 satellite. Journal of Remote Sensing, 2020, 24(10): 1168–1179.

10 刘帅兵, 杨贵军, 周成全, 等. 基于无人机遥感影像的玉米苗期株数信息提取. 农业工程学报, 2018, 34(22): 69–77. [doi: 10.11975/j.issn.1002-6819.2018.22.009]

11 Yang MD, Tseng HH, HsuY C, *et al*. Semantic segmentation using deep learning with vegetation indices for rice lodging identification in multi-date UAV visible images. Remote Sensing, 2020, 12(4): 633. [doi: 10.3390/rs12040633]

12 韩文霆, 张立元, 牛亚晓, 等. 无人机遥感技术在精量灌溉中应用的研究进展. 农业机械学报, 2020, 51(2): 1–14. [doi: 10.6041/j.issn.1000-1298.2020.02.001]

13 Gómez-Candón D, Virlet N, Labbé S, *et al*. Field phenotyping of water stress at tree scale by UAV-sensed imagery: New insights for thermal acquisition and calibration. Precision Agriculture, 2016, 17(6): 786–800. [doi: 10.1007/s11119-016-9449-6]

14 Stöcker C, Bennett R, Nex F, *et al*. Review of the current state of UAV regulations. Remote Sensing, 2017, 9(5): 459. [doi: 10.3390/rs9050459]

15 Ozdarici-Ok A, Ok AO, Schindler K. Mapping of agricultural crops from single high-resolution multispectral images—Data-driven smoothing vs parcel-based smoothing. Remote Sensing, 2015, 7(5): 5611–5638. [doi: 10.3390/rs70505611]

16 Zhao WZ, Du SH. Learning multiscale and deep representations for classifying remotely sensed imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 113: 155–165. [doi: 10.1016/j.isprsjprs.2016.01.004]

17 Kwak GH, Park NW. Impact of texture information on crop classification with machine learning and UAV images. Applied Sciences, 2019, 9(4): 643. [doi: 10.3390/app9040643]

18 Zhu XX, Tuia D, Mou LC, *et al*. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine, 2017, 5(4): 8–36. [doi: 10.1109/MGRS.2017.2762307]

19 Rebetez J, Satizábal HF, Mota M, *et al*. Augmenting a convolutional neural network with local histograms—A case study in crop classification from high-resolution UAV imagery. Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, 2016. 27–29.

20 汪传建, 赵庆展, 马永建, 等. 基于卷积神经网络的无人机遥感农作物分类. 农业机械学报, 2019, 50(11): 161–168.

[doi: 10.6041/j.issn.1000-1298.2019.11.018]

21 Chew R, Rineer J, Beach R, *et al*. Deep neural networks and transfer learning for food crop identification in UAV images. Drones, 2020, 4(1): 7. [doi: 10.3390/drones4010007]

22 Han ZM, Dian YY, Xia H, *et al*. Comparing fully deep convolutional neural networks for land cover classification with high-spatial-resolution Gaofen-2 images. ISPRS International Journal of Geo-Information, 2020, 9(8): 478. [doi: 10.3390/ijgi9080478]

23 Yang YD, Zhuang Y, Bi FK, *et al*. M-FCN: Effective fully convolutional network-based airplane detection framework. IEEE Geoscience and Remote Sensing Letters, 2017, 14(8): 1293–1297. [doi: 10.1109/LGRS.2017.2708722]

24 Karim F, Majumdar S, Darabi H. Insights into LSTM fully convolutional networks for time series classification. IEEE Access, 2019, 7: 67718–67725. [doi: 10.1109/ACCESS.2019.2916828]

25 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

26 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929, 2020.

27 Reedha R, Dericquebourg E, Canals R, *et al*. Transformer neural network for weed and crop classification of high resolution UAV images. Remote Sensing, 2022, 14(3): 592. [doi: 10.3390/rs14030592]

28 Lee Y, Kim J, Willette J, *et al*. MPViT: Multi-path vision transformer for dense prediction. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 7277–7286.

29 Tan MX, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 6105–6114.

30 Chen LC, Zhu YK, Papandreou G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 833–851.

31 Xu WJ, Xu YF, Chang T, *et al*. Co-scale conv-attentional image transformers. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 9981–9990.

32 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241.

33 Zhou ZW, Rahman Siddiquee M, Tajbakhsh N, *et al*. UNet++: A nested U-Net architecture for medical image segmentation. Proceedings of the 4th Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Granada: Springer, 2018. 3–11.

34 Chaurasia A, Culurciello E. LinkNet: Exploiting encoder representations for efficient semantic segmentation. Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP). St. Petersburg: IEEE, 2017. 1–4.

(校对责编: 牛欣悦)