

面向云遮挡与复杂地块场景下的跨模态注意力多源 遥感农作物分类

巫晨旭, 左浩龙, 李 刚*

(黑龙江工程学院测绘工程学院, 黑龙江哈尔滨 150050, 中国)

摘要: [目的/意义] 本研究旨在突破传统光学遥感在云遮天气与耕地破碎地形下的局限性, 通过构建一种具备优秀跨模态融合能力与泛化性能的深度网络模型来提升农业遥感分类精度。[方法] 提出一种基于注意力机制的3D卷积神经网络(3D Convolutional Neural Network Based on Attention Mechanism, Attention-3DCNN)模型: 其通过3D卷积+2D卷积结构处理时序哨兵二号多光谱影像, 在空间与时间维度上提取丰富特征; 同时以深度可分离卷积形式处理来自哨兵一号的合成孔径雷达(Synthetic Aperture Radar, SAR)数据, 实现对全天候可获取信息的高效抽取; 进一步, 模型引入“通道-时间-空间”三重注意力机制与残差连接策略, 对两个模态的特征进行动态加权与深度融合, 使得在光学数据缺失或遮挡严重的情形下, SAR数据能够有效补偿并维持分类性能。[结果和讨论] 为全面评价模型性能, 选取法国全景农业卫星时序数据集, 以及山东省沂水县实测数据集进行对比实验: 在法国数据集上模型达到97.5%的整体准确率, 在沂水县数据集上获得93%的准确率, 均显著优于对照基线模型; 同时, 通过对注意力分布的可视化分析可见, 模型聚焦的关键物候期与当地农业实地记录高度一致, 其高权重光谱波段亦符合农学机理, 这体现出模型在判别机制层面的可解释性。[结论] 综上, Attention-3DCNN模型在耕地破碎、云遮影响严重的山区条件下显著提升了作物分类精度, 具有良好的推广前景与应用价值。

关键词: 农作物分类; 深度学习; 3D卷积神经网络; 注意力机制; 遥感; 合成孔径雷达

中图分类号: S127; TP18

文献标识码: A

文章编号: SA202510010

引用格式: 巫晨旭, 左浩龙, 李刚. 面向云遮挡与复杂地块场景下的跨模态注意力多源遥感农作物分类[J]. 智慧农业(中英文), 2026, 8(2): 118-132. DOI: 10.12133/j.smartag.SA202510010

WU Chenxu, ZUO Haolong, LI Gang. Cross-Modal Attention for Multi-Source Remote Sensing Crop Classification under Cloud Occlusion and Complex Field Scenarios[J]. Smart Agriculture, 2026, 8(2): 118-132. DOI: 10.12133/j.smartag.SA202510010 (in Chinese with English abstract)

0 引言

作物制图作为评估农业系统变化、优化作物管理的核心技术, 是支撑粮食产量预估、农业资源动态监控的关键基础^[1]。然而, 在多云雨地区, 作物生长季的光学遥感观测常因云层遮挡而失效, 严重制约了农田监测的时效性与准确性^[2]。

为突破这一瓶颈, 合成孔径雷达(Synthetic Aperture Radar, SAR)等不受天气影响的遥感数据被视为有效的解决方案^[3]。现有研究已探索了多源数据融合^[4]、深度学习^[5]、深度学习与云平台结合^[6], 以及物候特征^[7]等在作物分类中的潜力。例如, 郭交等^[8]针对Pol-

SAR数据的农作物精细分类研究印证了SAR模态在复杂天气下作物分类中的独特价值。王佳玥等^[9]使用协同高分一号(GF-1), 高分二号(GF-2)和高分六号(GF-6)多源国产高分影像构建“光谱-时相-空间”三维特征影像来提升作物分类准确率; 林云浩等^[10]设计了耦合DeepLab与Transformer的双分支并行特征融合网络(Fusion DeepLab-Transformer Network, FDTNet)架构实现了高精度的作物分类; 张伟雄等^[11]通过多尺度时空全局注意力来挖掘遥感时序数据的价值。这些工作为多云区的作物分类奠定了良好基础。

近年来, 多源遥感数据的融合方法快速发展,

收稿日期: 2025-10-11

基金项目: 黑龙江省“双一流”学科协同创新成果项目(LJGXCG2025-P18)

作者简介: 巫晨旭, 硕士研究生, 研究方向为时空大数据。E-mail: 15589720597@163.com

*通信作者: 李 刚, 讲师, 研究方向为时空大数据。E-mail: 35987336@qq.com

copyright©2026 by the authors

(登录 www.smartag.net.cn 免费获取电子版全文)

典型策略包括早期融合、特征级（中期）融合和决策级融合，不同策略在鲁棒性、特征表达能力和模态交互深度方面具有显著差异^[12]。与此同时，注意力机制被大量引入多源遥感分类模型中，用于动态调节不同模态、不同区域和不同时段的重要性^[13,14]。部分研究利用通道注意力或空间注意力强化光学或SAR特征^[15]；也有研究采用并行注意力或Transformer结构实现多模态特征的深层交互，例如跨模态Transformer融合器^[16]、基于多通道的注意力融合器^[17]等方法在高光谱-SAR、光学-激光雷达（LightLaser Detection and Ranging, LiDAR）等任务中取得了优于传统融合策略的效果。然而，这些方法大多侧重于某一种或两种注意力机制，仍缺乏同时从“通道-空间-时间”多个维度对多源特征进行动态建模的能力，且在复杂作物-物候环境下的适应性仍有限。

在面对山东省沂水县这类兼具“高云量、碎地块、物候特异”的复杂场景时，现有方法仍存在明显局限：1）单一光学模型抗云能力弱；2）传统多源融合策略固定，无法根据作物与物候期动态调整数据权重；3）主流注意力机制多针对单模态设计，难以实现光学与SAR的跨模态协同。

本研究提出了一种动态自适应的多源融合模型，用于解决光学遥感在多云与耕地较为破碎的条件下的农作物分类精度较低的问题。该模型创新性地构建了“通道-时间-空间”三重注意力机制，实现了多源数据融合从静态权重叠加到动态智能协同的范式跨越，使模型能够依据实际观测条件自主决定“使用何种数据源、在何时段、关注哪些空间区域”。在方法层面，设计了面向跨模态数据的双分支特征提取与注意力融合架构，通过上述三重注意力机制实现特征的自适应加权与深度融合。在法国PASTIS公开数据集与山东沂水县实测数据集上进行实验，该模型具有优异的泛化能力和分类稳定性，性能显著优于基于固定权重的传统融合方法。此外，模型所学习到的注意力权重与作物关键物候阶段及遥感物理机理高度吻合，不仅提升了分类结果的可信度，也为实现可解释的农业遥感智能分析提供了切实可行的技术框架。

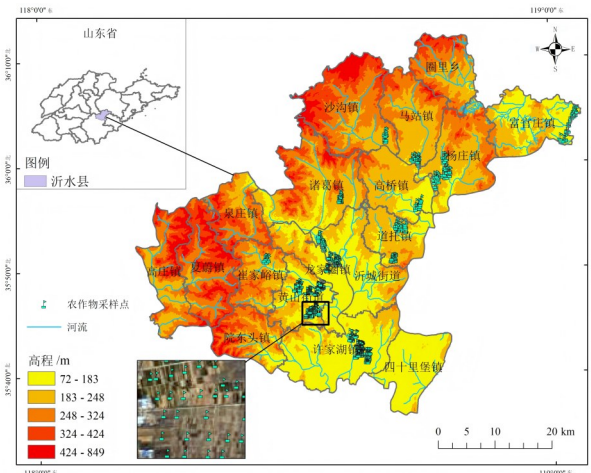
1 材料和方法

1.1 研究区域和数据集

1.1.1 研究区域

山东省东南部的沂水县（35°29′—35°56′N，

118°14′—118°48′E）地处鲁南丘陵向鲁西平原过渡带，低山丘陵和冲积平原交错分布。沂水县属于暖温带季风气候区，年平均气温12.3℃，年平均降水量784.8 mm，4月至9月是主要生长季，降水较多且云量达33%^[18]，使得遥感光学数据观测窗口受限。作物布局以冬软质小麦、夏玉米、大豆、果树为主，地块平均面积<0.5 hm²且边界不规则（碎片化特征显著），与法国PASTIS数据集的1.2 hm²连片地块形成鲜明对比，需模型具备更强的空间细节捕捉能力。作为临沂市农业经济核心区，当地农业生产链调控、粮食产量评估均依赖高精度作物分布数据，但复杂地形与高云量导致传统光学分类方法精度不足，亟须抗干扰的多源融合方案。研究区域如图1所示。



注：该图基于自然资源部标准地图服务网站下载的审图号为GS(2019)3333号标准地图制作，底图无修改。

图1 山东省沂水县研究区域示意图

Fig. 1 Location map of the study area in Yishui county, Shandong province

1.1.2 数据集

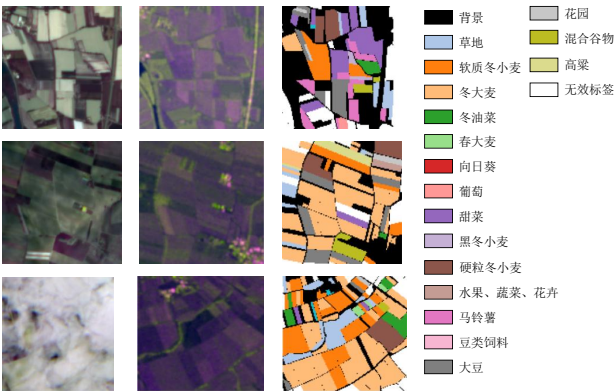
1.1.2.1 PASTIS数据集

PASTIS 由在法国四个不同地区采样的2 433个多光谱图像时间序列组成。每个图像块的像素尺寸为128×128，共计115 000张图像，保留10个10 m波段（排除3个60 m波段），未过滤云覆盖（约28%部分有云），标注18种作物及土地覆盖类别（含主要农作物以及自然景观）；并且包含哨兵1号（Sentinel-1）数据，经Orfeo Toolbox正射校正为后向散射系数（分贝），空间分辨率10 m，未做相干斑滤波以及辐射地形校正，每个观测是3通道（VV、VH、VV/VH）^[19]。PASTIS数据集的波段对比如表1所示，样本可视化如图2所示。

表 1 PASTIS数据集与 Sentinel-2 波段对应关系及分辨率
对比表

Table 1 Comparison of band correspondences and resolutions
between PASTIS dataset and Sentinel-2

PASTIS 波段	Sentinel-2 波段	原始分辨率/m	重采样后分辨率/m
1—4	B2—B4, B8	10	10
5—7	B5—B7	20	10
8	B8A	20	10
9—10	B11—B12	20	10



a. 光学影像 b. SAR数据 c. 分类标签
图 2 PASTIS数据集
Fig. 2 PASTIS dataset

1.1.2.2 沂水县主要农作物分类数据集

本研究构建的沂水县主要农作物分类数据集用于验证模型的区域泛化能力。首先，在 2024 年 4 月—9 月作物生长季期间，采用手持全球定位系统 (Global Positioning System, GPS) 设备采集沂水县典型种植区的小麦、玉米、大豆及果树样本点，共获得有效地面采样点 32 713 个。为确保采样信息的空间精度，对所有采样点进行坐标纠错与重复点检查，并剔除定位误差超过 5 m 的异常点。Sentinel-1 原始数据需经过一系列预处理步骤，包括数据导入、多视、配准、时间序列滤波、地理编码与辐射定标。首先，在多视处理中，通过沿距离向融合五个视窗的图像强度，以提升辐射分辨率并抑制散斑噪声，其参数由 Sentinel-1 头文件直接读取。随后，选取单一主影像，将所有从影像与之配准至同一几何基准，为时间序列滤波奠定基础。该滤波旨在解决由系统相干干涉与地物表面散射所引起的随机噪声，通过利用影像在时间维度上的关联性实现有效去噪。随后，将采样点与 Sentinel-1 (VV、VH 及额外衍生的比值极化通道) 和 Sentinel-2 (10 个光学波段) 影像进行时空配准。影像均采用同一投影坐标系 (WGS84/UTM)，并均通过轨道校正、大

气校正 (采用 Sen2Cor)、辐射定标、后向散射几何校正及多时相配准等步骤实现统一的几何与辐射基础。为显示出模型对云覆盖场景下的性能，本研究并没有对云遮挡进行特殊处理，数据集中保留大约三分之一云遮挡较为严重的数据。在预处理完成后，对预处理后的影像裁剪成多个 128×128 像素的影像块，结合多时相序列最终获得时空对齐的多源遥感数据立方体，共计 1 200 个有效样本。各样本均统一进行了尺度归一化 (z-score) 与 SAR 数据的 speckle 去噪 (Refined Lee 滤波)，同时对光学与 SAR 通道进行空间分辨率匹配 (以 10 m 为统一空间尺度，采用双三次/最近邻插值策略)。参考《沂水县 2023 年年鉴》确定核心分类体系为五类：小麦、玉米、大豆、果树及背景。背景包含裸地、乡村建设区及其他非农用地。沂水县数据集与 PASTIS 数据集在通道组成、时序长度与样本规模等方面的参数对比如表 2 所示。生成的地表覆盖标签与对应的多光谱与 SAR 数据示例如图 3 所示。

表 2 PASTIS 与沂水县数据集核心参数的对比

Table 2 Comparison of core parameters between PASTIS and the Yishui county dataset

参数	PASTIS(法国)	沂水县(中国)	对迁移的潜在影响
云覆盖率/%	28	33	沂水县需更强的抗云干扰能力
时相数量 (4—9月)	32 个时相	23 个时相	沂水县时序信息更稀疏, 考验时间注意力
平均地块面积/hm ²	1.2	<0.5	沂水县需更高的空间细节捕捉能力

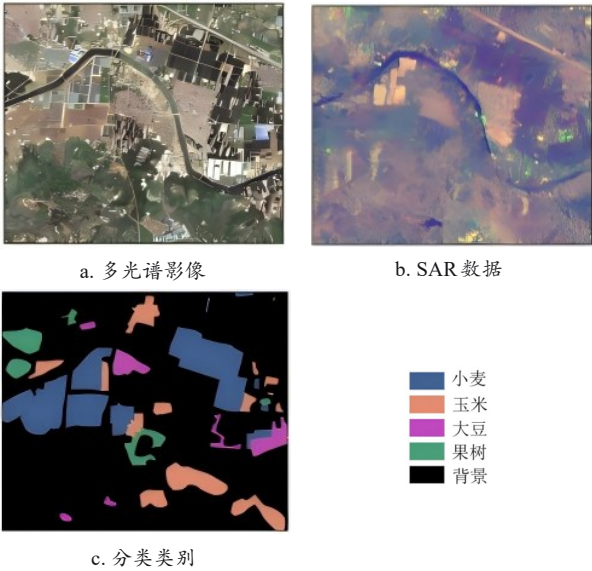


图 3 沂水县主要农作物分类数据集
Fig. 3 The dataset of main crops in Yishui county

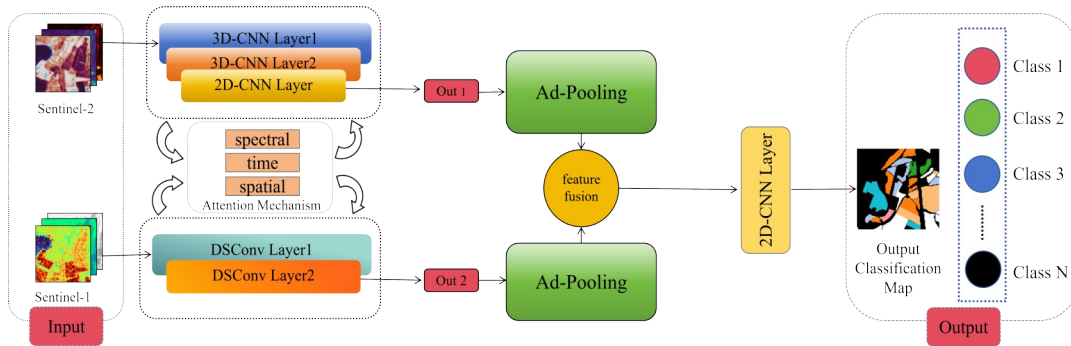
1.1.2.3 数据集对比

PASTIS 数据集与沂水县数据集在影像组成、作物类型、场景复杂度及云覆盖条件上均存在显著差异，为本研究验证基于注意力机制的3D卷积神经网络（3D Convolutional Neural Network Based on Attention Mechanism, Attention-3DCNN）模型的区域迁移能力提供了良好的实验基础。PASTIS数据集具有类别数量多、区域跨度大、时序密度高的特点，更适合作为模型的基础训练与验证平台；而沂水县数据集类别较少，但受云遮挡和耕地破碎化影响更为突出，贴合中国典型山区农业场景。两者均包含 Sentinel-1 与 Sentinel-2 数据，但 PASTIS 光学时序密度远高于沂水县，而沂水县数据集更贴近实际农业监测中的不规则时相与噪声条件。通过在两

个具有显著差异的数据集上的实验，可以全面评估模型在不同地区、不同作物结构与不同数据质量条件下的稳健性与迁移性能。

1.2 模型架构

针对沂水县“高云量，时序稀，地块碎”的痛点，模型采用“双分支特征提取→通道、时间、空间注意力机制→特征融合→分类输出”的四级架构（图4），实现多源数据的动态协同与精准分类。模型主要由两个分支组成，分别是多光谱影像分支、SAR分支。最后将两个分支的输出进行特征融合处理，随后通过一个 1×1 卷积层将融合后得到的特征图映射到类别数，再经 Softmax 函数得到逐像素的分类结果。



注：DSConv 为深度可分离卷积层；3D-CNN、2D-CNN 分别是 3D 卷积层和 2D 卷积层；Ad-Pooling 为自适应池化。

图4 Attention-3DCNN 模型总架构图

Fig. 4 Overall architecture diagram of the attention-3DCNN model

1.2.1 多光谱影像分支

近些年来 3DCNN 被广泛应用于视频、遥感和生理信号等三维数据的处理。在遥感分类领域，蔡玉林等^[20]将其与视觉转换器（Vision Transformer, ViT）结合，通过交叉注意力融合多模态数据，提升了树种分类精度；在行为识别领域，李泽慧等^[21]系统总结了其轻量化与注意力机制等改进策略；在情绪识别领域，杨朋辉等^[22]则利用其提取脑电时空特征，并结合注意力模块取得了领先性能。这些工作彰显了 3DCNN 在处理复杂序列数据方面的显著优势。

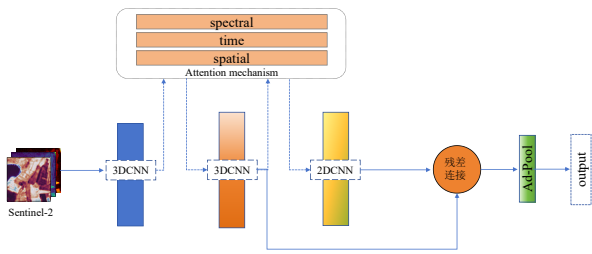
残差网络被广泛应用于图像识别，翻译，语音识别等领域^[23]。残差网络的核心原理在于将上一层的输出与后续层的输出直接相连，从而绕过了某些中间层，使得后一层既可以获得前一层的输出也可以获得原始的输入。这种设计可以让模型学习到残差映射——即目标输出与当前输出的差异。通过优化这些残差映射，训练过程会变得更加可控^[24]。

此外，残差网络还可以很大程度上缓解训练过程中的局部最小值问题^[25]。

基于以上理由，本研究在多光谱影像的特征提取中采用了 3DCNN 加 2DCNN 和残差网络的方式进行特征提取。具体而言，多光谱分支采用两个 3DCNN 卷积层和一个 2DCNN 卷积层，并进行了残差连接。该结构设计借鉴了 FARMONOV 等^[26]提出的 HypsLiDNet 框架中的多光谱影像特征提取的部分，与 HypsLiDNet 框架不同的是，为了适应沂水县地块较碎的特点，在注意力机制模块中增加了空间注意力机制。分支结构图如图 5 所示。

1.2.2 SAR 分支

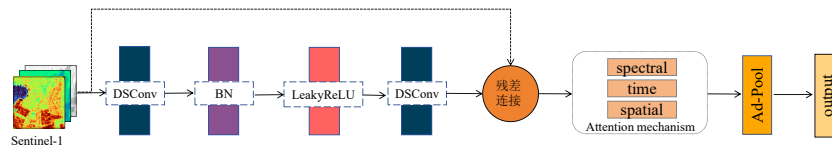
深度可分离卷积被广泛应用于图像分类、图像去噪和图像美学评估等领域^[27]，该技术主要包括两个独立操作：其一为深度方向卷积，使用较小核函数独立处理每个输入通道的空间特征，此方法可有效捕捉局部细节；其二是逐点卷积，用 1×1 卷积核将深度方向卷积输出的结果进行线性组合，以



注: Conv3D($3 \times 3 \times 3, s=1, p=1, C_{out}=64$), Conv2D($3 \times 3, s=1, p=1, C_{out}=128$), Ad-Pooling($1 \times 1, s=1, p=0, C_{out}=128$); s 为步长, p 为填充, C_{out} 为输出大小; spectral、time、spatial 分别代表通道、时间、空间注意力机制。

图5 多光谱分支模型图

Fig. 5 Multispectral branch model diagram.



注: BN 为批量归一化; LeakyReLU 为 LeakyReLU 激活函数 DSConv($3 \times 3, s=1, p=1, C_{out}=C_{in}$); Ad-Pooling($1 \times 1, s=1, p=0, C_{out}=128$); s 为步长; p 为填充; C_{out} 为输出大小; C_{in} 为输入大小。

图6 SAR分支模型架构图

Fig. 6 SAR branch model architecture diagram.

1.2.3 注意力模块

本研究引入了三种注意力机制: 通道注意力、时间注意力和空间注意力。流程如图7所示。

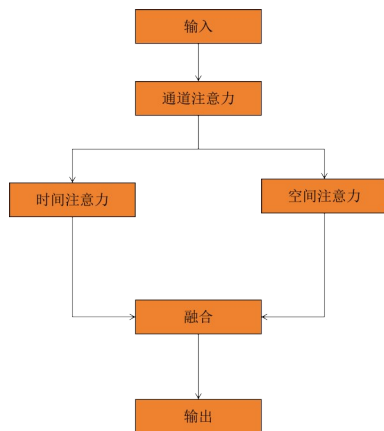


图7 注意力机制流程图

Fig. 7 Attention mechanism diagram

1.2.3.1 通道注意力

该注意力机制通过调整多通道影像中不同通道特征图的权重来自适应地选择重要特征。为避免引入大量参数, 参考 TONG 等^[29] 在通道注意力中的做法, 仅选择将密集块与过渡层融入通道注意力机制之内。其中, 过渡层由 1×1 卷积层和步长为2的平均池化层构成, 核心作用是缩减特征图的尺寸; 而这种与过渡层完成集成的通道注意力模块, 被定

及融合^[28]。

由于 Sentinel-1 的 SAR 数据几乎不受云量影响, 因此可以有效解决沂水县高云量的问题, 本研究采用了两个深度可分离卷积层, 并在两个深度可分离卷积层之间添加了批量归一化层 (Batch Normalization, BN) 来稳定训练过程, 加快收敛速度, 以及 LeakyReLU 激活函数 (LeakyReLU) 来引入非线性特征, 提高模型的表达能力。此外, 模型还将深度可分离卷积和残差网络进行结合, 大大提高了模型对于局部特征的捕捉能力, 并降低了模型的参数量和计算量。在模型的最后部分, 同样设立了注意力模块来接收残差连接后的输出。结构如图6所示。

义为自适应下采样。结构如图8所示。

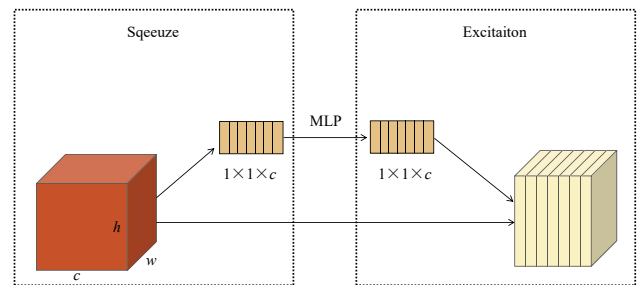


图8 通道注意力机制模型图

Fig. 8 Channel attention mechanism diagram

通道注意力采用挤压-激励 (Squeeze-Excitation, SE) 模块计算权重。

挤压是将输入特征压缩成一个长度为通道数的一维向量, 具体过程是使用全局池化将每一层特征压缩为一个值最终输出一个 $1 \times 1 \times C$ 的向量, 其中 C 是通道数。

在激励阶段, 各通道之间的依赖关系通过两个非线性全连接层的门控机制来捕捉。两个全连接层的维度分别是 $C/16$ 和 C 。设激励阶段的输出为 S_c , 如公式 (1) 所示。

$$S_c = \sigma(W_2 \text{ReLU}(W_1 z)) \quad (1)$$

式中: σ 是 Sigmoid 函数; W_1 、 W_2 为全连接层参数; z 为挤压操作的输出; ReLU 是非线性激活函

数。最后将原始特征映射与权重向量进行逐通道相乘，得到加权后的输出特征。

1.2.3.2 时空注意力

为捕获多光谱和SAR数据中的时空关系和物候信息，采用了时间和空间注意力机制，两种注意力机制根据编码特征块不同的展平形式独立提取空间和时间信息。将通道注意力的输出特征沿时间或空间尺度进行展平，设展平后的序列分别是 z_T 、 z_S ，如公式（2）~公式（5）所示。

$$z'_T = \text{MSA}(\text{LN}(z_T)) + z_T \quad (2)$$

$$z'_S = \text{MSA}(\text{LN}(z_S)) + z_S \quad (3)$$

$$\text{MST}(z) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^o \quad (4)$$

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right)V_i \quad (5)$$

式中： z'_T 、 z'_S 分别为经过时间和空间注意力机制强化后的特征；LN（Layer Normalization）表示层归一化操作；MSA（Multi-Head Self-Attention）表示多头自注意力，用于对输入序列进行多头注意力计算；MST表示对输入 z 进行多头时空注意力处理所得到的输出，即将所有注意力头（ head ）的输出拼接后通过线性变换得到最终特征； h 表示MSA中的 head 的数量， W^o 是用于特征空间变换的线性变换矩阵； Q_i 、 K_i 、 V_i 分别代表每个 head 的查询、键和值； d 为 Q_i 、 K_i 的维度； $\sqrt{d_k}$ 为缩放因子；Softmax函数对输入值进行归一化，输出注意力权重（取值范围 $[0, 1]$ ）；空间和时间维度的自注意力矩阵大小分别为 $[H_n W_n, H_n W_n]$ 和 $[T, T]$ ， H_n 代表第 n 层特征图的高度， W_n 代表第 n 层特征图的宽度， T 代表时间序列的长度。

在得到时空注意力提取的两个特征后，将两个特征拼接起来，采用一个 1×1 的卷积核进行融合降维从而得到注意力机制模块的最终输出。

1.2.4 分类输出模块

分类输出模块接收融合特征张量后，经一个 1×1 的卷积层及Softmax函数计算归一化概率分布，如公式（6）所示。

$$P_i = \frac{e^{x^T w_i}}{\sum_{l=1}^L e^{x^T w_l}} \quad (6)$$

式中： P_i 表示模型分配给第 i 类的似然值； w_i 表示最后一层的权重和偏置； x 是包含从图像中提取的深度特征的向量。通过反向传播计算并最小化真实目标 y_i 和网络输出 P_i 之间的交叉熵的期望值。

1.3 模型参数设置和硬件环境

模型运行过程中的步长为1；通道压缩率 $r=16$ ，以平衡表达能力与效率；优化器为Adam，初始学习率0.001，采用Cosine Annealing策略调至最低0.000 01；损失函数用交叉熵，结合早停机制防过拟合。

2 结果与分析

首先在国际公开基准PASTIS数据集上评估了Attention-3DCNN的基础分类能力。在数据集划分上遵循数据集的标准评估协议，运用五折交叉验证方案进行模型训练与测试。具体而言，使用其官方提供的固定五折划分，然后依次将其中四折用于训练，剩余一折用于测试，并报告五次实验的平均结果，以确保评估的稳健性。这种方法避免了因随机划分可能带来的空间泄露问题，确保评估结果的可靠性和可比性。随后，为检验模型在真实复杂场景下的泛化性与实用性，在沂水县数据集上进行了进一步验证。在划分沂水县训练集与测试集时（比例为7:3），充分考虑了空间分布独立性，确保来自同一空间区域或相邻像元的样本不会被同时分布于训练集和测试集中，从而杜绝空间泄露对模型性能评估的影响。同时，还控制了不同农作物类别（小麦、玉米、大豆、果树）在训练集与测试集中的分布比例，避免类别不均衡对评估结果造成偏差。对于受云量影响较大的区域，同样按照7:3的比例在训练集与测试集之间合理分配，确保二者在空间分布与云量条件上均具有代表性。本章将重点分析模型的分类准确性、异质区域适应性、注意力机制的可解释性，以及评估模型从公开基准（PASTIS）到真实复杂场景（沂水县2023年实测数据）的泛化能力与稳健性。

下文中用于对比的每一种模型均采用与本研究模型同样的训练集进行训练，并且采用相同的预处理策略、训练轮数、优化器（AdamW）、初始学习率、批大小和学习率调度策略。所有模型均在同一测试集上评估，并使用统一指标，训练时采用混合精度与梯度缩放技术以优化计算效率。损失函数综合了多尺度输出与类别加权策略，使用AdamW优化器配合余弦退火学习率调度进行参数更新，结果显示，模型在法国数据集上达到97.5%的整体准确率，在沂水县数据集上获得93%的准确率，两者均优于对比模型。

2.1 消融实验

采用控制变量的方式进行消融实验，将本研究模型分别与标准 3D-CNN（仅 Sentinel-2）、光学-SAR 简单融合模型（特征堆叠）、仅 SAR 3D-CNN（Sentinel-1）、无注意力双分支融合（特征连接），以及单一注意力模型（通道注意力双分支融合模型、时间注意力双分支融合模型、空间注意力双分支融合模型）进行对比。这些模型覆盖了“单一数据源→传统融合→无注意力融合→单一注意力融合”的技术谱系。为评估模型性能，分别在 PASTIS 和沂水县数据集的训练集上训练模型，并分别

在各自的测试集上评估模型性能，评估模型性能所采用的指标为总体精度（Overall Accuracy, OA），宏平均 F_1 分数（Macro- F_1 ），Kappa 系数，所有模型的初始学习率为 0.001，批量大小为 32，最多训练 100 轮，10 轮无改进则终止。实验表明，本研究提出的 Attention-3DCNN 模型无论是在 PASTIS 数据集上还是在沂水县主要农作物分类数据集上均取得了优异的分类效果，直接验证了引入跨模态注意力机制的有效性，证明了其自适应加权策略优于简单的特征堆叠或连接方法。模型表现见表 3、图 9 和图 10。

表 3 Attention-3DCNN 模型与其他模型在不同数据集上的表现
Table 3 Performance of the Attention-3DCNN and other models on different datasets

模型名称	OA(PASTIS)/%	宏平均 F_1 分数(PASTIS)	Kappa(PASTIS)	OA(沂水县)/%	宏平均 F_1 分数(沂水县)	Kappa(沂水县)
标准 3D-CNN	94.2	0.932	0.915	87.2	0.845	0.822
光学-SAR 简单融合模型	95.3	0.945	0.928	89.5	0.876	0.858
仅 SAR 3D-CNN 模型	91.6	0.902	0.889	88.3	0.861	0.842
注意力双分支融合模型	95.1	0.938	0.935	89.9	0.882	0.870
通道注意力双分支融合模型	96.5	0.953	0.942	90.0	0.898	0.890
时间注意力双分支融合模型	95.8	0.945	0.942	89.5	0.894	0.887
空间注意力双分支融合模型	96.9	0.959	0.956	91.0	0.906	0.900
Attention-3DCNN	97.5	0.970	0.965	93.0	0.920	0.910

2.2 对比实验与模型性能分析

为突出模型的特点，还选择了当前主流模型以同样的方式和流程进行实验。选择的模型为 3D-ConvSTAR^[30]、Self-Attention 3D^[26]、UNet++^[31]、基于卷积神经网络和长短时记忆网络的双流模型^[32]（Dual-Stream Model Based on Convolutional Neural Network and Long Short-Term Memory Network, CNN-LSTM-DS），以及基于 Transformer 与 Gist 卷积神经网络的融合网络^[33]（Fusion Network Based on Transformer and Gist Convolutional Neural Network, TGF-Net）。这些基线模型的选择基于以下三个维度：

- 1) 传统时序/三维卷积模型（如 3D-ConvSTAR），代表基于卷积的时序建模与局部特征融合能力；
- 2) 注意力/Transformer 增强的时序模型（如 Self-Attention 3D、TGF-Net），代表近年来在多模态时序任务中表现优异的自注意力机制；
- 3) 经典分割/编码器-解码器模型（如 UNet++）与双流时序模型（如 CNN-长短时记忆网络（Long

Short-Term Memory, LSTM）），分别代表侧重多尺度特征提取与分流时空特征的设计范式。

上述选择原则旨在覆盖主要的方法学谱系（卷积、注意力、编码器-解码器、双流融合），并涵盖遥感作物分类领域中被广泛认可或代表前沿思路的工作，从而能够公平地评估本研究方法在不同设计目标下的优势与局限。

其中，3D-ConvSTAR 模型结合三维卷积与卷积循环神经网络，侧重于时序建模；Self-Attention 3D 模型融合形态学特征与注意力机制，在高光谱与激光雷达数据间执行选择性特征增强；UNet++ 模型借助嵌套跳跃连接实现多尺度语义信息融合；CNN-LSTM-DS 模型则通过双流架构分别提取空间与时序特征并进行后期融合，TGF-Net 模型结合了 Transformer 与 CNN 架构，用于融合高光谱（Hyperspectral Imaging, HSI）和 SAR 数据的特征，设计了基于 Transformer 的光谱特征提取模块和空间特征提取模块，并在两个 HSI+SAR 数据集上验证了其像素级分类效果。

实验中的所有模型的初始学习率为 0.001，批

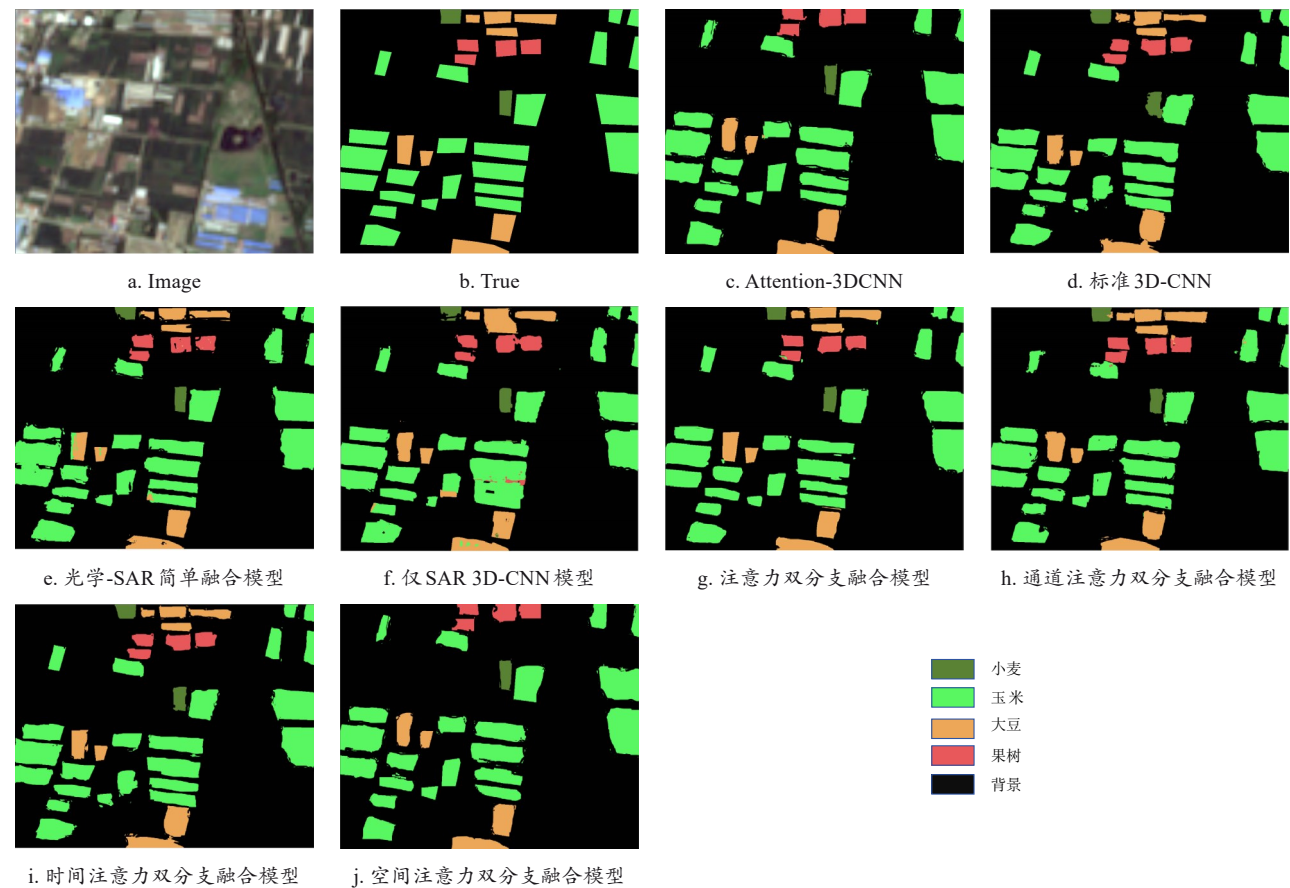


图9 在消融实验中 Attention-3DCNN 与其他模型在沂水县数据集上的结果

Fig. 9 Results of the Attention-3DCNN and other models in the ablation study on the Yishui county dataset

量大小为 32，最多训练 100 轮，10 轮无改进则终止，实验结果如表 4 和图 11 所示。表 4 展示了各模型在沂水县数据集上的定量对比结果，除分类精度外，本研究还进一步报告了参数量、计算量 GFLOPs 与单景平均推理时间，以全面评估方法的性能-效率均衡性。结果表明，尽管上述模型在各自设计目标中均取得了一定成效，但在应对本研究关注的“高云量、时序稀疏、地块破碎”等多重挑战时仍存在明显局限。具体而言，它们的融合机制依然多依赖静态融合权重或单一注意力模块，难以

在通道、时间和空间三个维度上实现自适应特征校准，导致光学影像缺失或复杂地形条件下分类性能下降。此外，TGF-Net 模型等基于 Transformer 的方法虽具有较强的表征能力，但其参数量与计算开销显著增大，不利于实际部署。相比之下，本研究提出的 Attention-3DCNN 通过引入“通道-时间-空间”三重注意力机制，能够依据不同区域和作物生长期的遥感响应特性动态调整多源数据融合权重，显著提升了在异质环境下的判别能力。最终，本模型在山东沂水县数据集上均取得了最优分类精度。

表 4 Attention-3DCNN 模型与其他模型的对比

Table 4 Comparison of Attention-3DCNN with other models

模型名称	核心方法简述	OA(沂水县)/%	Kappa(沂水县)	参数量/M	计算量/ GFLOPs	推理时间/(ms/景)
3D-ConvSTAR	3DCNN 多源融合(固定权重)	89.5	0.858	45.2	128.3	156
Self-Attention 3D	自注意力机制+3DCNN	90.5	0.872	52.7	145.6	183
UNet++	编码器-解码器结构,多尺度特征融合	88.0	0.835	68.9	212.4	245
CNN-LSTM-DS	光学影像时序+纹理特征融合	86.5	0.815	43.0	98.7	132
TGF-Net	基于 Transformer 与卷积(CNN)架构	90.5	0.900	105.3	285.1	312
Attention-3DCNN	跨模态三重注意力(通道-时间-空间)	93.5	0.910	41.3	97.2	133

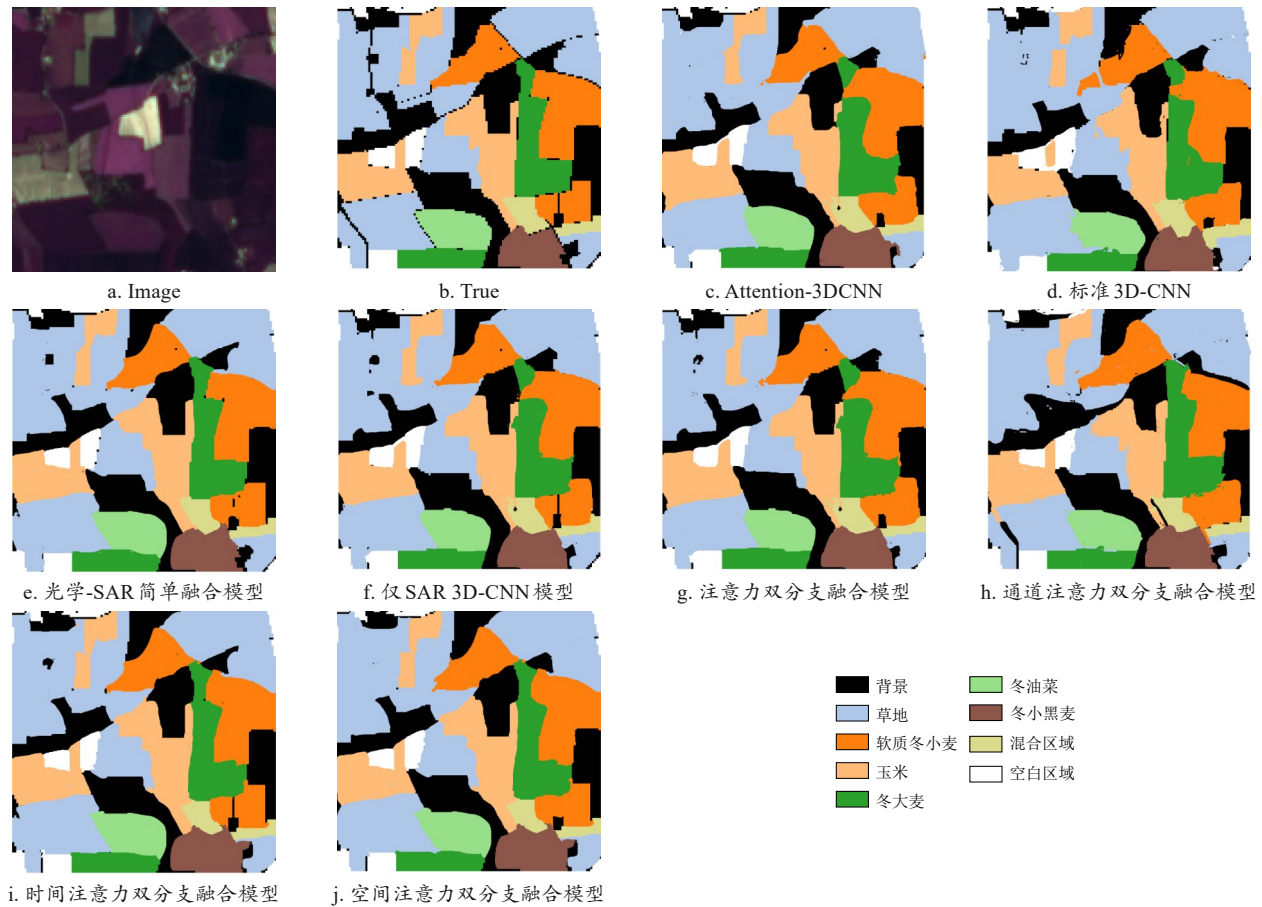


图 10 在消融实验中 Attention-3DCNN 模型与其他模型在 PASTIS 数据集上的结果

Fig. 10 Results of Attention-3DCNN and other models in the ablation study on the PASTIS dataset

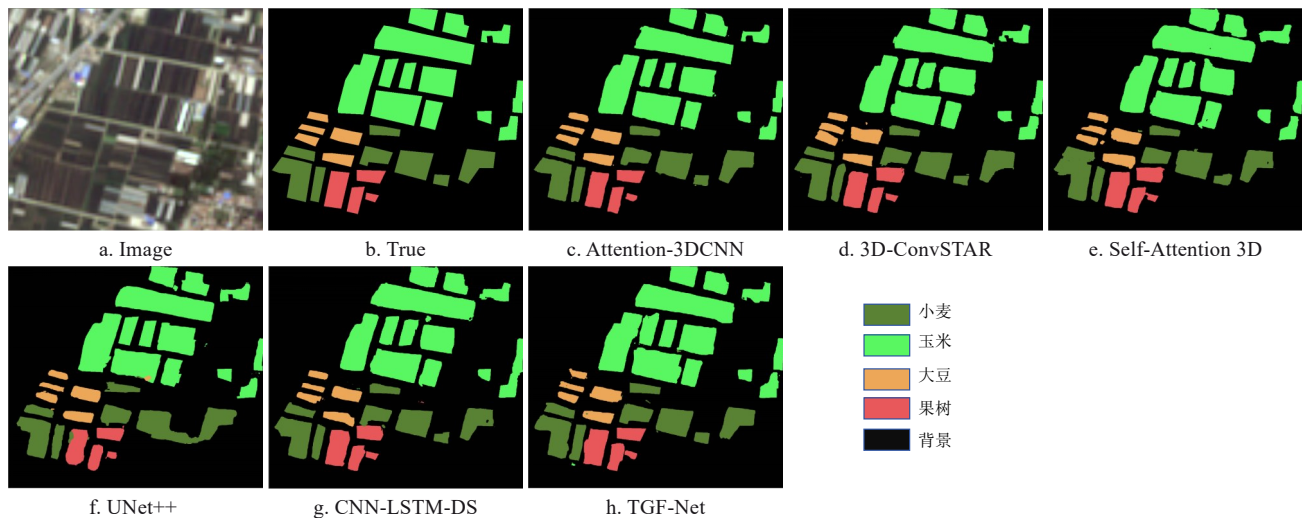


图 11 在对比实验中 Attention-3DCNN 模型与其他模型在沂水县数据集上的对比结果

Fig. 11 Results of the Attention-3DCNN and other models in the comparative experiments on the Yishui county dataset

2.3 云量较大条件下的模型测试

为进一步评估模型在光学信息严重受限情形下的鲁棒性, 本研究从沂水县测试集中筛选云覆盖程

度较高的样本构建高云量子测试集。该子集包含约 120 个样本, 覆盖不同作物类型及地形条件, 能够代表典型多云山区农业遥感场景。鉴于高云量子测试集中各类别样本数量进一步减少, 且云遮对不同

作物类别的影响程度存在差异，本节在常规整体准确率与Kappa系数之外，额外引入宏平均 F_1 分数，以更全面地评估模型在类别层面的分类稳定性。采用2.2节中使用的模型进行实验，模型编号、初始学习率、批量大小等不变。该实验旨在模拟多云雨山区农业遥感监测中的典型应用场景，重点考察模型在光学观测质量下降时对多源信息的利用能力。实验结果如表5、表6和图12所示。

表5 在高云量的条件下 Attention-3DCNN 模型与其他模型在沂水县数据集的对比

Table 5 Comparison of Attention-3DCNN and other models under high cloud coverage conditions on the Yishui county dataset

模型名称	主要依赖信息	OA/%	宏平均 F_1	Kappa
3D-ConvSTAR	光学时序为主	83.6	0.802	0.775
Self-Attention 3D	光学 + 时序注意力	85.1	0.821	0.796
UNet++	光学空间特征	82.4	0.789	0.761
CNN-LSTM-DS	光学时序 + 手工特征	81.9	0.781	0.754
TGF-Net	Transformer + CNN 融合	86.3	0.836	0.812
Attention-3DCNN	跨模态三重注意力 (S2+S1)	89.4	0.872	0.846

表5展示了不同模型在高云量条件下的整体分类性能。可以观察到，所有模型在该子测试集上的整体准确率、宏平均 F_1 分数及Kappa系数均较常规条件有所下降，表明云层遮挡对作物光谱与时序特征的提取具有普遍影响。其中，主要依赖光学影像的模型（如 UNet++与 3D-ConvSTAR）性能下降最

表6 多源遥感农作物分类研究中不同模型在高云量条件下的下降幅度

Table 6 The decline of different models under high cloud cover in multi-source remote sensing crop classification studies

模型	常规 OA/%	高云量 OA/%	下降幅度/百分点
3D-ConvSTAR	89.5	83.6	5.9
Self-Attention 3D	90.5	85.1	5.4
UNet++	88.0	82.4	5.6
CNN-LSTM-DS	86.5	81.9	4.6
TGF-Net	90.0	86.3	3.7
Attention-3DCNN	93.0	89.4	3.6

为明显，其整体准确率降至82%~84%区间，反映出在光学信息严重受限时，此类模型难以维持稳定判别能力。

相比之下，引入时序建模或注意力机制的模型在一定程度上缓解了云遮影响，但其融合策略仍以光学特征为主，缺乏对不同模态贡献度的动态调节能力，因此在高云量条件下仍表现出较为明显的性能衰减。基于 Transformer 的 TGF-Net 在该场景下保持了相对较高的分类精度，表明全局建模与特征重加权机制有助于提升模型对不完整观测数据的适应性。

在此基础上，表6进一步给出了各模型在高云量条件下相对于常规测试条件的性能下降幅度。结果显示，传统光学主导模型的整体准确率下降幅度普遍超过5%，而引入跨模态融合的性能衰减相对较小。本研究提出的 Attention-3DCNN 模型在

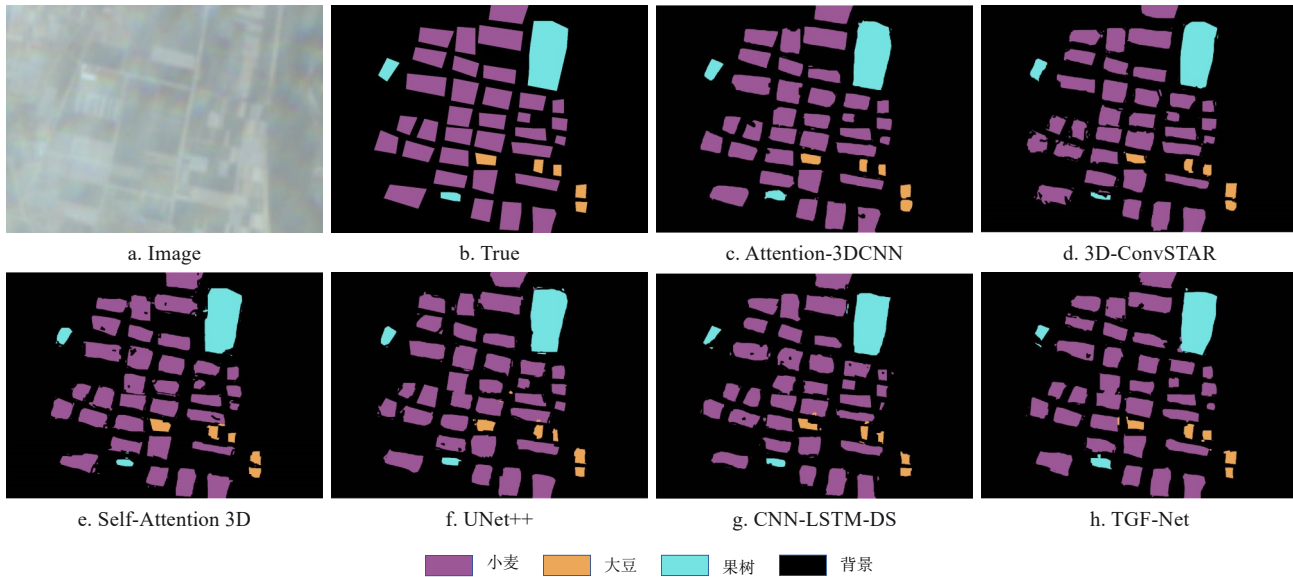


图12 在高云量的条件下 Attention-3DCNN 与其他模型在沂水县数据集的结果图

Fig. 12 Results of Attention and other models on the Yishui county dataset under conditions of high cloud coverage

高云量条件下的整体准确率仅下降3.6个百分点,为所有对比模型中最低,显示出最强的稳定性。

该结果表明, Attention-3DCNN 不仅在绝对分类精度上具有优势,其在复杂气象干扰下的性能保持能力同样优于现有方法。这主要得益于模型中通道—时间—空间三重注意力机制的协同作用,使模型能够在光学信息质量下降时自适应提升SAR特征权重,并在不同时间尺度与空间区域上动态调整多源特征的融合策略,从而有效抑制云遮对分类性能的负面影响。

综上,高云量条件下的性能衰减分析从“绝对精度”和“相对稳定性”两个层面验证了本方法的鲁棒性优势,进一步说明所提出的跨模态注意力机制在多云多雨、山区农业遥感场景中具有显著的应用潜力。

2.4 地块破碎程度较高条件下的模型测试

除云量干扰外,耕地破碎化也是制约山区农业遥感分类精度的重要因素之一。小尺度、边界不规则的农田地块容易造成混合像元比例升高与空间结构信息破碎,从而对模型的空间特征提取能力和边界判别能力提出更高要求。为评估不同模型在高度破碎耕地条件下的适应能力,本研究从沂水县测试集中进一步筛选地块面积较小、边界形态复杂的样本,构建地块破碎程度较高的子测试集,并在该子集上开展对比实验,该子集有230个样本,实验模型采用2.2节中的模型,模型编号、初始学习率等不变。实验结果如表7、表8和图13所示。

表7给出了不同模型在地块破碎条件下的分类性能对比结果。整体来看,相较于常规测试条件,

表7 地块破碎程度较高条件下 Attention-3DCNN 与其他模型在沂水县数据集的对比

Table 7 Comparison of Attention-3DCNN and other models on the Yishui county dataset under the condition of high degree of land fragmentation					
模型名称	主要结构特点	OA/%	宏平均 F_1	Kappa	
3D-ConvSTAR	3D 卷积,局部时空建模	84.9	0.816	0.792	
Self-Attention 3D	3DCNN + 自注意力	86.2	0.829	0.806	
UNet++	编码器-解码器,多尺度跳连	83.5	0.801	0.776	
CNN-LSTM-DS	空间 CNN+时序 LSTM	82.7	0.793	0.768	
TGF-Net	Transformer + CNN 融合	87.1	0.842	0.818	
Attention-3DCNN	跨模态三重注意力 (含空间注意力)	90.2	0.881	0.856	

表8 多源遥感农作物分类研究中地块破碎程度较高条件下不同模型在沂水县数据集上的性能下降幅度

Table 8 The performance degradation of different models on the Yishui county dataset under the condition of high degree of land fragmentation			
模型名称	常规测试 OA/%	破碎地块 OA/%	下降幅度/百分点
3D-ConvSTAR	89.5	84.9	4.6
Self-Attention 3D	90.5	86.2	4.3
UNet++	88.0	83.5	4.5
CNN-LSTM-DS	86.5	82.7	3.8
TGF-Net	90.0	87.1	2.9
Attention-3DCNN	93.0	90.2	2.8

各模型在该子集上的整体准确率、宏平均 F_1 分数及 Kappa 系数均出现不同程度下降,表明地块破碎对作物分类任务具有显著干扰作用。其中,依赖连续空间语义结构的模型(如 UNet++) 在破碎地块条件下性能下降较为明显,其分类结果在地块边界区域易出现类别混淆,反映出编码器-解码器结构在应对小尺度、不规则农田斑块时的局限性。

相比之下,基于时序建模或注意力机制的模型在一定程度上缓解了空间结构破碎带来的不利影响。其中, TGF-Net 借助 Transformer 的全局建模能力,在复杂空间结构下保持了相对稳定的性能,说明引入全局上下文信息有助于提升模型对碎片化地块的判别能力。然而,该类方法在空间细节捕捉与多尺度局部建模方面仍存在不足,其性能提升幅度有限。

表8从性能下降幅度的角度量化了不同模型在地块破碎条件下的稳定性差异。结果表明,传统卷积模型及编码器-解码器结构的整体准确率下降幅度普遍超过4个百分点,而引入跨模态融合与注意力机制的模型性能衰减相对较小。本研究提出的 Attention-3DCNN 在破碎地块子测试集上的整体准确率仅下降2.8个百分点,为所有对比模型中最低,表现出最强的稳定性。

这一优势主要源于模型中引入的空间注意力机制与多尺度特征建模策略,使模型能够在地块边界复杂、空间连续性较差的条件下,动态聚焦于具有判别意义的局部区域特征,并有效抑制混合像元与边界噪声的干扰。同时,跨模态融合框架使得 SAR 数据在空间纹理与结构信息表达上的优势得以充分发挥,进一步增强了模型在破碎耕地场景下的分类鲁棒性。

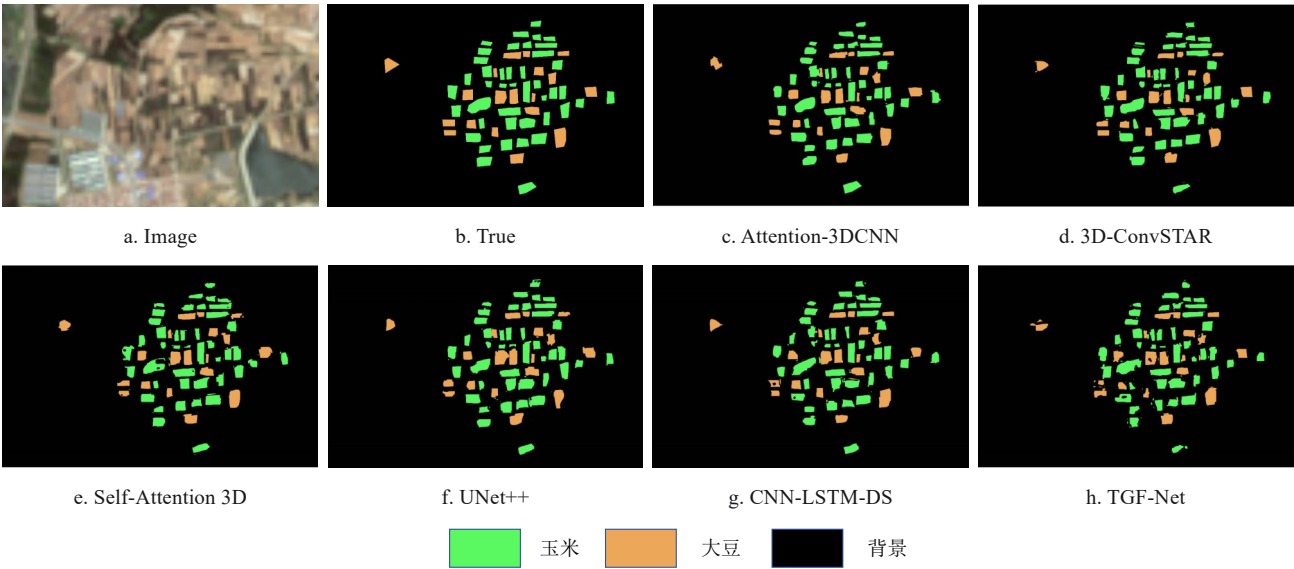


图 13 在地块破碎的条件下 Attention-3DCNN 模型与其他模型在沂水县数据集的结果图
Fig. 13 Results of Attention-3DCNN and other models on the Yishui county dataset under fragmented land conditions

综上所述，地块破碎条件下的对比实验从绝对分类性能与相对稳定性两个层面验证了 Attention-3DCNN 在复杂空间结构场景中的适应能力。该结果与 2.3 节云量较大条件下的实验结论相互印证，表明本方法在多云雨、耕地破碎等典型山区农业遥感场景中均具有良好的鲁棒性与推广潜力。

2.5 注意力权重的区域适应性分析

本模型的性能优势主要得益于其注意力机制的

自适应调整能力。如表 9 所示，面对两地差异，模型的注意力权重发生了显著且合理的变化。该实验有力地证明了 Attention-3DCNN 不仅是一个高性能的分类器，更具备较强的跨区域泛化能力。其内在的注意力机制能根据新区域的数据特征动态调整聚焦策略，从而克服地域差异带来的挑战，为模型在全国乃至全球同类地区的推广应用提供了坚实依据。

表 9 法国 PASTIS 与中国沂水县各类别关键特征及模型注意力调整对比

Table 9 Comparison of key characteristics and model attention adjustment in various categories between France PASTIS and Yishui county of China				
分类类别	关键特征	法国 PASTIS 注意力权重(均值)	沂水县注意力权重(均值)	调整分析
小麦	抽穗期	时间: 0.76; 通道(B5): 0.21	时间: 0.79; 通道(B5): 0.23	物候提前,时间权重前移;红边波段重要性增强
大豆	开花结荚期	通道(VH): 0.35	通道(VH): 0.62	光学数据受限,SAR通道权重显著提升以补偿
背景	非农用地	空间: 0.58	空间: 0.71	地块破碎,模型更关注局部空间结构

2.6 注意力机制的可解释性分析与农学验证

不同作物对遥感信息源的依赖存在显著差异，小麦和大豆在近红外波段都具有较高重要性值；果树分类的最佳光谱波段是 S2 的绿色（B3）和 SWIR_2（B10），而 S1 则是垂直水平（VH）极化波段^[34]；背景识别主要依赖的是 Sentinel-2 可见光波段（B2-B4）及空间注意力机制。研究证实，通道注意力机制是能够根据作物生物学特性动态调整关注重点，实现关键信息源的自适应聚焦，进而突破传统统一加权模式的局限性。该发现为多源遥感

数据融合策略的作物特异性优化提供的是理论依据。

时间注意力识别的关键物候期与沂水县农业农村局 2023 年物候观测数据高度吻合，具体数据如表 10 所示。这种时间窗口的精准对标排除“模型自定物候期”的主观性，证实其是能够从遥感数据中自动提取与农学观测一致的关键生长阶段，为分类精度提升提供可靠的物候学基础。

上述分析表明，通道注意力能够针对不同作物的生物学特性动态分配波段权重，时间注意力识别的关键物候期与实地观测高度吻合，两者共同构成

表 10 模型识别的关键物候期与沂水县农业局观测数据对比

Table 10 Comparison of key phenological periods identified by the model with the observation data from the agricultural bureau of

Yishui county				
作物类型	模型识别关键物候期(第 X—X 天)	对应公历时间	沂水县农业农村局观测记录	匹配度/%
小麦	120—150 天	4 月中下旬—5 月上旬	4 月 20 日—5 月 10 日抽穗	100
玉米	180—210 天	7 月上旬—8 月上旬	7 月 5 日—8 月 5 日灌浆	100
大豆	160—190 天	6 月下旬—7 月下旬	6 月 25 日—7 月 25 日结荚	100
果树	70—90 天	2 月下旬—3 月中旬	2 月 28 日—3 月 15 日萌芽	100

了模型“自适应聚焦关键信息”的实证基础，为分类精度的提升提供了可解释的机制支撑。

3 结论与展望

3.1 结论

本研究针对高云量、碎地块区域作物分类精度低的难题，成功构建并验证了基于跨模态注意力机制的 Attention-3DCNN 模型，用于融合光学与 SAR 遥感数据进行农作物分类。该模型通过“通道-时间-空间”三重注意力的协同作用，实现了多源遥感数据的自适应融合，在国际公开基准（PASTIS）和山东沂水县实测场景下均取得了领先的分类精度（OA 分别达 97.5% 和 93%），展现了出色的鲁棒性和泛化能力。

本研究证实了自适应注意力加权策略在多源遥感融合中的核心价值，为突破多云雨地区农业遥感监测瓶颈提供了一种创新且有效的技术途径。

3.2 展望

尽管本研究取得了上述成果，但仍存在一些局限性，未来将围绕以下方向进一步深化研究，以推动遥感农作物分类方法向更高精度、更强鲁棒性与更高效的应用方向发展：

（1）发展高精度地块边界优化方法。未来研究将探索融合多尺度上下文信息与边缘感知的优化技术，重点提升复杂破碎地块及混合像元区域的边界分割完整性，实现从像素级分类到对象级精细化识别的跨越。

（2）构建鲁棒的时序缺失数据重建与融合模型。针对长时间、大范围云覆盖导致的数据质量问题，将重点发展基于生成式模型与跨模态学习的缺失信息重建方法，深化光学与 SAR 等多源数据的时序协同利用机制，增强模型在低质量数据条件下的稳定分类能力。

（3）致力于模型轻量化与业务化部署研究。为支撑大区域、高频次业务化运行需求，后续工作将

系统开展模型剪枝、量化与知识蒸馏等压缩技术研究，研发适配边缘计算及云端高效推理的轻量化模型，推动技术在农业监测实际业务环境中的规模化应用。

后续研究将在全国更多典型农区开展验证，并逐步增加作物类型，进一步检验和拓展模型的普适性，最终为智慧农业和国家粮食安全提供可靠的遥感技术支撑。

利益冲突声明：本研究不存在研究者以及与公开研究成果有关的利益冲突。

参考文献：

[1] 刘伊,张彦军. ReluformerN:轻量化高低频增强高光谱农业地物分类方法[J]. 智慧农业(中英文), 2024, 6 (5): 74-87.
LIU Y, ZHANG Y J. ReluformerN: Lightweight high-low frequency enhanced for hyperspectral agricultural landcover classification[J]. Smart Agriculture, 2024, 6(5): 74-87.

[2] SUN Y W, LI Z L, LUO J C, et al. Farmland parcel-based crop classification in cloudy/rainy mountains using Sentinel-1 and Sentinel-2 based deep learning[J]. International Journal of Remote Sensing, 2022, 43(3): 1054-1073.

[3] YE Y X, ZHANG J C, ZHOU L, et al. Optical and SAR image fusion based on complementary feature decomposition and visual saliency features[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 5205315.

[4] LI J J, ZHANG J C, YANG C, et al. Comparative analysis of pixel-level fusion algorithms and a new high-resolution dataset for SAR and optical image fusion[J]. Remote Sensing, 2023, 15(23): 5514.

[5] RAČIČ M, OŠTIR K, ZUPANČ A, et al. Multi-year time series transfer learning: Application of early crop classification[J]. Remote Sensing, 2024, 16(2): 270.

[6] VIZZARI M, LESTI G, ACHARKI S. Crop classification in Google Earth Engine: Leveraging Sentinel-1, Sentinel-2, European CAP data, and object-based machine-learning approaches[J]. Geo-Spatial Information Science, 2025, 28(3): 815-830.

[7] ZHAO F, YANG G J, YANG X D, et al. Determination of key phenological phases of winter wheat based on the time-weighted dynamic time warping algorithm and MODIS time-series data[J]. Remote Sensing, 2021, 13(9): 1836.

[8] 郭交,王鹤颖,项诗雨,等. 基于多特征优化的 PolSAR 数据农作物精细分类方法[J]. 农业机械学报, 2024, 55(9): 275-285.
GUO J, WANG H Y, XIANG S Y, et al. Crop classification based on PolSAR data using multiple feature optimization[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55 (9): 275-285.

[9] 王佳玥,蔡志文,王文静,等. 协同多源国产高分影像和面向对象方法的南方农作物遥感识别[J]. 中国农业科学, 2023, 56 (13): 2474-2490.

- WANG J Y, CAI Z W, WANG W J, et al. Integrating multi-source Gaofen images and object-based methods for crop type identification in South China[J]. *Scientia Agricultura Sinica*, 2023, 56(13): 2474-2490.
- [10] 林云浩, 王艳军, 李少春, 等. 一种耦合 DeepLab 与 Transformer 的农作物种植类型遥感精细分类方法[J]. *测绘学报*, 2024, 53(2): 353-366.
- LIN Y H, WANG Y J, LI S C, et al. A coupled DeepLab and Transformer approach for fine classification of crop cultivation types in remote sensing[J]. *Acta Geodaetica et Cartographica Sinica*, 2024, 53(2): 353-366.
- [11] 张伟雄, 唐婷, 孟瑜, 等. 基于多尺度时空全局注意力的遥感影像时间序列农作物分类[J]. *遥感学报*, 2024, 28(11): 2865-2877.
- ZHANG WX, TANG P, MENG Y, et al. Crop type classification of remote sensing image time series based on multi-scale spatial-temporal global attention model[J]. *National Remote Sensing Bulletin*, 2024, 28(11): 2865-2877.
- [12] ZHANG L P, ZHANG L F, DU B. Deep learning for remote sensing data: A technical tutorial on the state of the art[J]. *IEEE Geoscience and Remote Sensing Magazine*, 2016, 4(2): 22-40.
- [13] HUANG X X, ZHANG X J, WANG L B, et al. MMA-net: A semantic segmentation network for high-resolution remote sensing images based on multimodal fusion and multi-scale multi-attention mechanisms[J]. *Remote Sensing*, 2025, 17(21): 3572.
- [14] MIAO J M, GAO J, WANG L, et al. Deep learning application of fruit planting classification based on multi-source remote sensing images[J]. *Applied Sciences*, 2025, 15(20): 10995.
- [15] LIU X, ZOU H J, WANG S X, et al. Joint network combining dual-attention fusion modality and two specific modalities for land cover classification using optical and SAR images[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024, 17: 3236-3250.
- [16] MEI Y, FAN J, FAN X, et al. CSTC: Visual transformer network with multimodal dual fusion for hyperspectral and LiDAR image classification[J]. *Remote Sensing*, 2025, 17(18): 3158.
- [17] RAMATHILAGAM A B, NATARAJAN S, KUMAR A. TransCropNet: A multichannel transformer with feature-level fusion for crop classification in agricultural smallholdings using Sentinel images[J]. *Journal of Applied Remote Sensing*, 2023, 17(2): 024501.
- [18] YU L S, ZHANG F G, ZANG K, et al. Potential ecological risk assessment of heavy metals in cultivated land based on soil geochemical zoning: Yishui county, North China case study[J]. *Water*, 2021, 13(23): 3322.
- [19] SAINTE FARE GARNOT V, LANDRIEU L, CHEHATA N. Multi-modal temporal attention models for crop mapping from satellite time series[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022, 187: 294-305.
- [20] 蔡玉林, 王兴路, 高洪振, 等. 融合 3DCNN 和 Vision Transformer 的多模态遥感数据树种分类方法[J]. *激光与光电子学进展*, 2025, 62(20): 2028001.
- CAI Y L, WANG X L, GAO H Z, et al. Tree species classification method based on multi-modal remote sensing data combined with 3DCNN and vision transformer[J]. *Laser & Optoelectronics Progress*, 2025, 62(20): 2028001.
- [21] 李泽慧, 张琳, 山显英. 三维卷积神经网络方法改进及其应用综述[J]. *计算机工程与应用*, 2025, 61(3): 48-61.
- LI Z H, ZHANG L, SHAN X Y. Review on improvement and application of 3D convolutional neural networks[J]. *Computer Engineering and Applications*, 2025, 61(3): 48-61.
- [22] 杨朋辉, 杨长青, 刘静, 等. 基于 2D-3D 卷积神经网络的情绪识别模型[J]. *燕山大学学报*, 2025, 49(1): 66-73.
- YANG P H, YANG C Q, LIU J, et al. Emotion recognition model based on 2D-3D convolutional neural network[J]. *Journal of Yan-shan University*, 2025, 49(1): 66-73.
- [23] SHAFIQ M, GU Z Q. Deep residual learning for image recognition: A survey[J]. *Applied Sciences*, 2022, 12(18): 8972.
- [24] WANG H Y, MIAO F. Building extraction from remote sensing images using deep residual U-Net[J]. *European Journal of Remote Sensing*, 2022, 55(1): 71-85.
- [25] LI H, XU Z, TAYLOR G, et al. Visualizing the loss landscape of neural nets[C]// *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. New York, USA: ACM, 2018: 6391-6401.
- [26] FARMONOV N, ESMAEILI M, ABBASI-MOGHADAM D, et al. HypsLiDNet: 3-D - 2-D CNN model and spatial-spectral morphological attention for crop classification with *DESI*S and LiDAR data[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024, 17: 11969-11996.
- [27] KHAN Z Y, NIU Z D. CNN with depthwise separable convolutions and combined kernels for rating prediction[J]. *Expert Systems with Applications*, 2021, 170: 114528.
- [28] LIU F C, XU H, QI M, et al. Depth-wise separable convolution attention module for garbage image classification[J]. *Sustainability*, 2022, 14(5): 3099.
- [29] TONG W, CHEN W T, HAN W, et al. Channel-attention-based DenseNet network for remote sensing image scene classification[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2020, 13: 4121-4132.
- [30] LIU N T, ZHAO Q S, WILLIAMS R, et al. Enhanced crop classification through integrated optical and SAR data: A deep learning approach for multi-source image fusion[J]. *International Journal of Remote Sensing*, 2024, 45(19/20): 7605-7633.
- [31] WANG L J, WANG J Y, LIU Z Z, et al. Evaluation of a deep-learning model for multispectral remote sensing of land use and crop classification[J]. *The Crop Journal*, 2022, 10(5): 1435-1451.
- [32] 史洁宁, 吴田军, 黄启厅, 等. 耦合 NDVI 与纹理时序特征的地块作物遥感分类[J]. *南方农业学报*, 2025, 56(1): 29-40.
- SHI J N, WU T J, HUANG Q T, et al. Land parcel crop remote sensing classification via coupling with time series features of NDVI and texture[J]. *Journal of Southern Agriculture*, 2025, 56(1): 29-40.
- [33] WANG H Q, WANG H J, WU L F. TGF-Net: Transformer and gist CNN fusion network for multi-modal remote sensing image classification[J]. *PLoS One*, 2025, 20(2): e0316900.
- [34] CHABALALA Y, ADAM E, ALI K A. Machine learning classification of fused Sentinel-1 and Sentinel-2 image data towards mapping fruit plantations in highly heterogeneous landscapes[J]. *Remote Sensing*, 2022, 14(11): 2621.

Cross-Modal Attention for Multi-Source Remote Sensing Crop Classification under Cloud Occlusion and Complex Field Scenarios

WU Chenxu, ZUO Haolong, LI Gang*

(School of Geomatics Engineering, Heilongjiang Institute of Technology, Harbin 150050, China)

Abstract: [Objective] Accurate and timely crop mapping is fundamental for agricultural management, yield forecasting, and food security assessment. However, in mountainous and hilly regions characterized by frequent cloud cover and highly fragmented farmland,

crop classification methods relying solely on optical remote sensing data are severely constrained. Persistent cloud contamination introduces data gaps and temporal inconsistencies in optical image time series, significantly degrading classification accuracy and robustness. To address these limitations, a robust and adaptive deep learning framework is developed capable of effectively integrating multi-modal remote sensing data. The primary objective is to enhance crop classification accuracy and stability under complex conditions where optical observations are scarce or unreliable, thereby supporting reliable agricultural monitoring in cloudy and fragmented landscapes. **[Methods]** A novel deep neural network architecture named 3D convolutional neural network based on attention mechanism (Attention-3DCNN) was proposed, designed to jointly exploit multi-temporal optical and synthetic aperture radar (SAR) observations. The model integrated Sentinel-2 multispectral time-series imagery with weather-insensitive Sentinel-1 SAR data through a dedicated cross-modal fusion strategy driven by a triple-attention mechanism. The network adopted a dual-branch feature extraction architecture. For the Sentinel-2 data, a hybrid module combining three-dimensional and two-dimensional convolutional neural networks (3D-CNN and 2D-CNN) was employed to capture discriminative spatiotemporal features and crop phenological dynamics across the growing season. This design enabled effective modeling of the spectral-temporal interactions inherent in crop development. For the Sentinel-1 SAR data, depthwise separable convolutions were utilized to efficiently extract spatial and textural features related to crop structure and surface scattering characteristics while reducing computational complexity. Features extracted from both modalities were subsequently integrated using a custom-designed attention-based fusion module. This module consisted of three complementary attention mechanisms: channel attention, temporal attention, and spatial attention. Residual connections were incorporated throughout the network to facilitate stable training and effective gradient propagation. The proposed model was evaluated on two datasets to assess both its performance and generalizability. The first was the publicly available panoptic agricultural satellite time series (PASTIS) benchmark dataset from France, which contained dense time-series observations and multiple crop classes. The second was a real-world dataset constructed for Yishui county, Shandong province, China, which was characterized by high cloud frequency (approximately 33%), highly fragmented farmland (average parcel size < 0.5 hm²), and a relatively simple crop rotation system. Comparative experiments were conducted against several state-of-the-art models, including 3D-ConvSTAR, UNet++, Self-Attention 3D, CNN-LSTM dual-stream network, and TGF-Net. Ablation studies were also performed to quantify the contribution of each attention component. **[Results and Discussions]** Experimental results demonstrated that Attention-3DCNN consistently outperformed all baseline methods on both datasets. On the PASTIS benchmark, the model achieved an overall accuracy (OA) of 97.5%, confirming its strong classification capability under favorable observation conditions. On the more challenging Yishui county dataset, Attention-3DCNN attained an OA of 93%, outperforming the other comparison models. Ablation experiments confirmed the effectiveness of the proposed triple-attention mechanism, as removing any attention component resulted in a clear reduction in classification performance. Under heavy cloud coverage, Attention-3DCNN exhibited the smallest accuracy degradation, with an OA drop of only 3.6 percentage points, indicating its ability to adaptively rely on SAR information when optical data quality deteriorated. In regions with highly fragmented farmland, the proposed model also maintained the highest accuracy and the smallest performance decline (2.8 percentage points), benefiting from the spatial attention mechanism. Moreover, attention visualization provided meaningful interpretability. Temporal attention peaks aligned with key crop phenological stages, while channel attention highlighted spectrally and physically informative optical bands and SAR polarizations, which was consistent with established agronomic and remote sensing knowledge. **[Conclusions]** This study presents the Attention-3DCNN model for accurate and robust crop classification in regions affected by persistent cloud cover and fragmented agricultural landscapes. By fusing Sentinel-2 optical and Sentinel-1 SAR time-series data through a channel-temporal-spatial triple-attention mechanism, the proposed framework enables adaptive integration of complementary multi-modal information. The model achieves outstanding performance on both benchmark and real-world datasets, demonstrates strong robustness under adverse conditions, and offers enhanced interpretability. Overall, the proposed approach provides a reliable and practical solution for crop mapping in complex agricultural environments.

Keywords: crop classification; deep learning; 3D convolutional neural network; attention mechanism; remote sensing; SAR

Foundation items: Heilongjiang Province Double First-Class Discipline Coordinated Innovation Achievement Project (LJGXCG 2025-P18)

First author: WU Chenxu, E-mail: 15589720597@163.com

***Corresponding author:** LI Gang, E-mail: 35987336@qq.com