

基于多模态数据和多模型融合的大范围 土地覆盖制图

茹逸凡^{1,2}, 唐国良¹, 李春来²

1. 中国科学院大学 杭州高等研究院, 杭州 310024;

2. 中国科学院上海技术物理研究所, 上海 200083

摘要: 当前土地覆盖制图面临着单一数据源信息局限、数据异质性大、单一模型泛化能力不足等核心挑战。为此本研究基于多光谱影像 MSI (Multispectral Imagery) 的光谱特征与合成孔径雷达 SAR (Synthetic Aperture Radar) 数据的不同特性, 构建了多模态 (MSI+SAR) 数据集, 以弥补单模态 SAR 数据在复杂场景下的不足。此外, 在模型层面系统评估 7 种代表性模型后, 提出融合 3 类架构的多模型融合方法: 以 FCN 代表卷积神经网络 (CNN), 以 ConViT 代表视觉变换器 (ViT), 以 CoAtNet 代表 CNN-ViT 混合架构。实验结果表明: 多模态 (MSI+SAR) 数据较单模态 SAR 数据在总体精度 OA (Overall Accuracy) 和频率加权交并比 FWIoU (Frequency-Weighted Intersection over Union) 这两个代表性的语义分割评估指标上均实现了显著提升, 而所提出的多模型融合框架在 OA 和 FWIoU 两个指标上进一步取得了改善。本研究所提出的创新方法, 不仅增强了模型对复杂地物特征的提取能力, 还有效提高了在大范围土地覆盖制图中的分类效果, 验证了多模态数据与多模型融合在遥感制图中强大潜力。

关键词: 土地覆盖制图, 多光谱, 合成孔径雷达, 多模态, 多模型融合, 卷积神经网络, 视觉变换器

中图分类号: TP751/P2

引用格式: 茹逸凡, 唐国良, 李春来. 2026. 基于多模态数据和多模型融合的大范围土地覆盖制图. 遥感学报, 30(2): 458-471

Ru Y F, Tang G L and Li C L. 2026. Large-scale land cover mapping based on multimodal data and multimodel fusion. National Remote Sensing Bulletin, 30(2): 458-471 [DOI: 10.11834/jrs.20255006]

1 引言

土地覆盖制图是遥感科学的核心研究方向之一, 其通过对地表地物类型进行精确分类, 为全球环境监测、资源管理和可持续发展提供关键数据支撑 (Congalton 等, 2014; Wu 等, 2024)。土地覆盖制图应用范围极为广泛, 涵盖了水文监测 (Jacqueminet 等, 2013)、农业资源评估 (Oetter 等, 2001) 以及生态环境保护 (Zhang 等, 2021) 等众多领域 (García-Mora 等, 2012)。然而, 传统土地覆盖制图方法往往依赖单一传感器获取的单模态遥感数据 (Li 等, 2023), 尽管在特定场景下表现尚可, 但在面对复杂、异质地表环境时, 其分类精度和鲁棒性仍存在显著不足 (Joshi 等, 2016)。

多光谱影像 MSI (Multispectral Imagery) 作为当前主流的遥感数据源, 通过多个光谱通道记录地物的反射的光谱信息, 能够有效区分植被、水体、裸土等地物类型, 其主要优势在于提供丰富的光谱信息 (Thyagarajan 和 Vignesh, 2019)。然而, MSI 数据易受云层遮挡、大气散射及光照条件变化的影响, 在多云多雨地区或阴影覆盖区域, 其数据可用性及其分类精度均受到严重制约 (Haque 等, 2024)。此外, MSI 主要依赖光谱特征, 对具有相似光谱响应但不同结构的地物 (如不同作物类型、城市建筑材质等) 区分能力有限, 导致分类结果存在混淆误差。

相比之下, 合成孔径雷达 SAR (Synthetic Aperture Radar) 作为一种主动式微波遥感技术,

收稿日期: 2025-01-08; 预印本: 2025-05-07

基金项目: 浙江省科技计划项目 (编号: 2024C03032)

第一作者简介: 茹逸凡, 研究方向为机器学习与语义分割、遥感土地覆盖制图、图像处理、多光谱与高光谱。E-mail: ruyifan22@mails.ucas.ac.cn

通信作者简介: 李春来, 研究方向为光谱成像技术、数字图形处理、微弱信号处理技术。E-mail: Lichunlai@mail.sitp.ac.cn

其成像过程不依赖于外部光照条件，且对云层、雾霾等大气干扰具有较强的穿透能力，能够在各种复杂环境条件下持续稳定地获取地表信息。SAR影像通过雷达波反射信号反映地物的结构特征，广泛应用于地形变化监测、灾害评估和城市建设分析等领域。尽管SAR影像在提取地物结构特征方面具有明显优势，但其光谱信息有限，在某些任务中可能导致分类精度不及MSI影像（Brown和Porcello, 1969）。

因此，融合MSI与SAR数据的多模态方法已成为当前遥感图像处理领域的重要研究方向。通过有机结合MSI丰富的光谱特征与SAR独特的结构信息，这种融合策略能够有效弥补单一数据源的固有局限，显著提升地物分类的准确性。多模态（MSI+SAR）数据融合不仅增强了模型的表征学习能力，还大幅改善了复杂地物类型的识别效果，在土地覆盖制图应用中展现出显著的技术优势（Quan等, 2020）。具体而言，MSI数据提供的光谱特征可用于区分地物材质，而SAR数据提供的结构信息则有助于识别地物的三维特征。

近年来，随着人工智能技术，特别是深度学习方法的快速发展，土地覆盖制图技术取得了突破性进展（Ajibola和Cabral, 2024; Paheding等, 2024）。以Li等（2022）提出的L2HNet模型为例，该模型通过创新性地运用低分辨率的土地覆盖标签，成功实现了高分辨率土地覆盖图的生成，极大地提升了图像细节的表达能力。针对训练样本不足这一普遍性难题（Tong等, 2025），Li等（2024）进一步开发了弱监督框架“Paraformer”，该方法通过引入历史低分辨率数据作为监督信号，有效缓解了高质量标注数据稀缺的问题。

从全球范围来看，多个具有代表性的土地覆盖产品已相继发布（Grekousis等, 2015）。美国地质调查局（USGS）开发的国家土地覆盖数据库（NLCD）美国本土长期土地覆盖变化分析提供了重要的数据支撑（Wickham等, 2021）；国内研究团队研制的中国土地覆盖数据集（CLCD）则为区域尺度的土地覆盖监测提供了重要数据支持（Yang和Huang, 2021）。虽然现有土地覆盖产品已取得重要进展，但在复杂地表环境下的分类精度仍有待提升（Xu等, 2024）。

针对上述问题，本研究提出了一种创新性的技术方案：首先，构建了基于MSI影像与SAR数

据的多模态融合数据集，以克服单一SAR数据的局限性；其次，在多模态（MSI+SAR）数据基础上，融合了卷积神经网络（CNN）、视觉变换器（ViT）和CNN-ViT混合架构这3种先进模型（分别代表为FCN、ConViT和CoAtNet），从而显著提升了复杂地表场景的分类精度和模型的鲁棒性。

2 研究区概况与数据集

如图1所示，本研究选取的研究区域位于中国北京。实验所用的高分六号多光谱数据和高分三号SAR数据，来源于高分系列卫星，数据获取时段为2019年末至2020年初，以确保地表目标在不同影像之间的变化最小化。高分六号多光谱数据在ENVI 5.6平台上经过辐射校正、快速视线大气光谱超立方体分析（FLAASH）校正、正射纠正和波段选择等预处理步骤，最终将影像重采样至10 m的空间分辨率。波段选择操作用于去除水蒸气吸收、噪声和低质量波段，从而保持图像质量（Liu等, 2024）。

高分三号SAR数据在宽幅精细条带模式下采集，具有10 m的空间分辨率和100 km的幅宽。SAR数据通过使用ENVI SARscape Analytics工具箱中的去斑和地形校正功能进行预处理。选用Refined Lee滤波器，滑动窗口大小为5×5像素，用于去除斑点噪声；同时，使用GMTED2010全球数字高程模型（DEM）数据进行SAR数据的地形校正（Hong等, 2023）。此外，高分三号SAR数据的大小和分辨率与高分六号MSI影像相同（Liu等, 2024）。

实验所用MSI影像为三维数据立方体，尺寸为256×256×4通道。SAR影像同样也是三维数据立方体，尺寸为256×256×2通道的三维数据立方体。数据格式与结构详见图1（a）和图2。

每张MSI影像的4个通道分别对应波段范围如下：0.4—0.52 μm 的蓝谱段（B1）；0.52—0.60 μm 范围的绿谱段（B2）；0.63—0.69 μm 范围的红谱段（B3）；0.76—0.90 μm 范围的近红外谱段（B4）。每张SAR影像的两个通道分别对应HH和HV这两种不同极化方式。HH通道表示雷达发射和接收信号都是水平极化；HV通道表示雷达发射信号为水平极化，接收信号为垂直极化。同时每张SAR影像的两个通道均属于C波段（约5.3 GHz频率，波长大约为3.1 cm）。

本研究的土地覆盖制图任务，涉及14种地物类

别,需将影像中每个像素点精确归类。如图1底部所示,这些类别各有特定名称。每组实验数据含一幅 MSI 影像与一幅匹配的 SAR 影像,二者在空

间位置与分辨率上与真实土地的真实土地覆盖图作为参考标准进行对比。

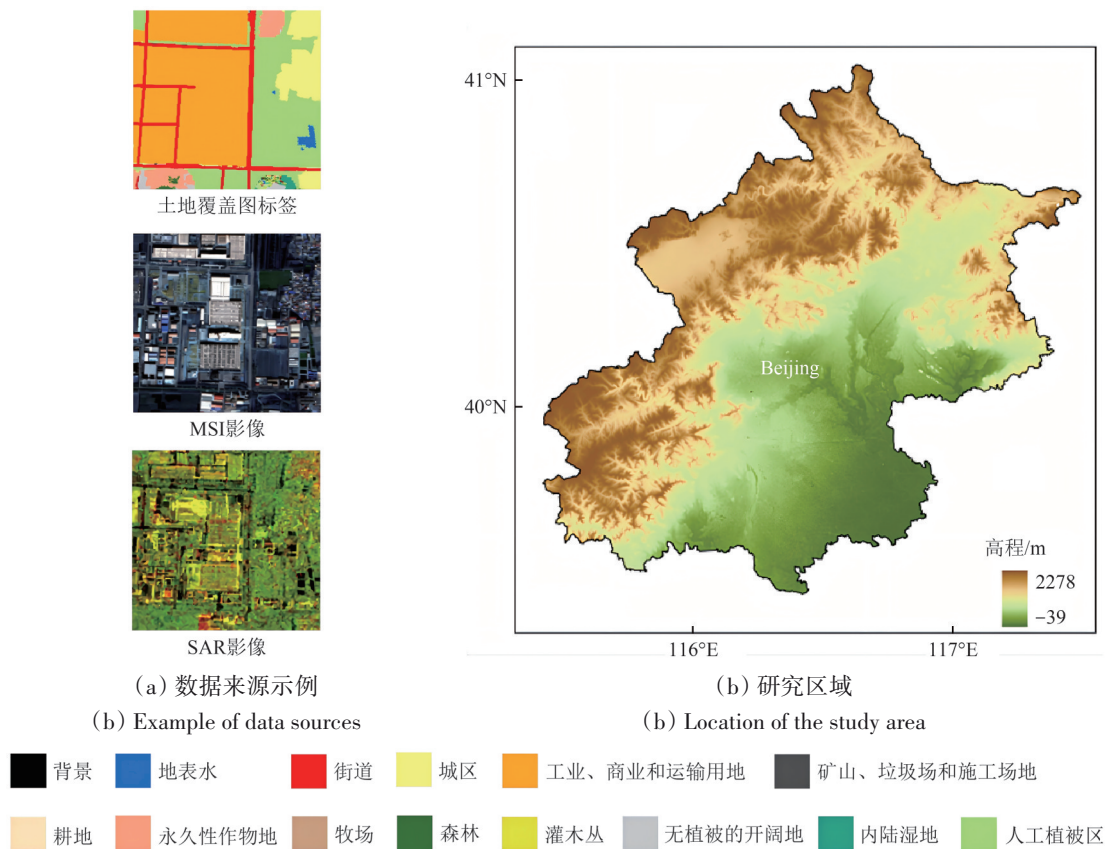


图1 研究区域和数据来源示意图

Fig. 1 Schematic of the study area and data sources

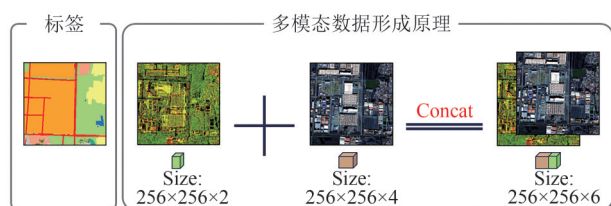


图2 单模态与多模态数据对比,从左到右依次是:基准土地覆盖图、单模态 SAR 数据、多光谱数据、多模态融合数据

Fig. 2 Comparison of single-modal and multi-modal data, from left to right: baseline land cover map, single-modal SAR data, multispectral data, multi-modal fusion data

3 研究方法

3.1 评估指标

本研究采用总体精度 OA (Overall Accuracy) 和频率加权交并比 FWIoU (Frequency-Weighted Intersection over Union) 作为核心评估指标指标

(Fitzgerald 和 Lees, 1994; LeCun 等, 1998), OA 和 FWIoU 数学表达式分别为

$$OA = \frac{\sum_{c=1}^C TP_c}{\sum_{c=1}^C n_c} \quad (1)$$

式中, C 表示类别总数; TP_c 表示模型正确地预测为某类别的像素数量; c 表示某一具体的类别; n_c 表示类别 c 的真实像素总数, 即 $n_c = TP_c + FN_c$, 其中 FN_c 表示未能正确预测为某类别的像素数量。

$$FWIoU = \frac{1}{\sum_{c=1}^C n_c} \cdot \sum_{c=1}^C \left(n_c \frac{TP_c}{TP_c + FP_c + FN_c} \right) \quad (2)$$

式中, n_c 同样表示类别 c 的真实像素总数, 即 $n_c = TP_c + FN_c$, 其中 FN_c 也表示未能正确预测为某类别的像素数量; TP_c 同样表示模型正确地预测为某类别的像素数量; c 表示某一具体的类别; FP_c 表示

模型错误地预测为某类别的像素数量； FN_c 表示未能正确预测为某类别的像素数量。

OA用以评定模型于整个数据集范畴内的分类精准度，在各类别分布处于相对均衡态势时，可提供具有效度的测度成果。FWIoU则兼顾类别分布的非平衡性，借由对交并比施以类别频次权重，进而更为全面地体现模型在类别分布不均衡情形下的性能水准。通常来说，此两项指标的数值越高，表征模型的分类精度与稳健性越为卓越。

3.2 多模态数据

本研究将 MSI 影像与 SAR 影像加以融合，生成多模态数据，并与由单一的 SAR 影像组成的单模态数据进行比对。随后，把这两类数据分别输入到 7 种具有代表性的语义分割机器学习模型进行对比，目的在于评估多模态数据相对单模态数据所具备的性能优势。

具体而言，借助监督学习方法，针对选定的 CoAtNet (Dai 等, 2021)、ConViT (D'Ascoli 等, 2022)、FCN (Shelhamer等, 2017)、HRNet (Wang等, 2021)、LinkNet (Chaurasia 和 Culurciello, 2017)、U-Net (Ronneberger 等, 2015) 这 6 个代表性的深度学习模型以及 Random Forest (Rodriguez-Galiano 等, 2012; Congalton 等, 2014) 这一种代表性的机器学习语义分割模型实施训练与测试。每个模型均分别采用基于单模态 SAR 数据和多模态 (MSI+SAR) 数据构建的两个数据集，以此实现对单模态与多模态数据在语义分割任务中性能差异的全面评价。

本研究所采用的数据可分为两类：第一，单模态数据与其真实对应的参考标签，分别对应图 2 里面的单模态 SAR 数据和基准土地覆盖图；第二，多模态数据与其真实对应的参考标签，分别对应图 2 里面的多模态融合数据和基准土地覆盖图。

详细而言，本研究搭建了上述两个数据集，各数据集均含有 7140 组训练样本。每组训练样本由单模态 SAR 数据、多模态 (MSI+SAR) 数据及其对应的真实参照标签共同组成，并且这三者在空间位置上呈现出绝对的一致性。在模型训练阶段，将第一个数据集中的 7140 组单模态 SAR 训练样本，与第二个数据集中的 7140 组多模态 (MSI+SAR) 训练样本，分别导入 CoAtNet、ConViT、FCN、HRNet、LinkNet、U-Net 以及 Random Forest 这 7 种

机器学习语义分割模型之中进行训练操作。随后，在模型测试阶段，将第 1 个数据集中的 1428 组测试样本和第 2 个数据集中与其对应的 1428 组测试样本，分别输入已完成训练的上述 7 种模型。最终，基于 OA 和 FWIoU 这两个关键指标，对模型在多模态 (MSI+SAR) 数据与单模态 SAR 数据场景下的性能，开展了综合性评估。

3.3 多模型融合策略

多模型融合策略的具体方式呈现于图 3。该策略有效整合 CNN 提取局部细节特征的能力与 ViT 感知全局信息的能力，以此提升模型的特征表达与分类性能 (Chua, 1997; Dosovitskiy 等, 2021)。对模型分类性能的评定，亦采用 OA 和 FWIoU 这两项指标予以评估。

如图 3 所示，左侧为多模型融合方法的输入多模态 (MSI+SAR) 数据。在右侧，红色虚线框内包含 3 个分支，分别为视觉变换器 (ViT) 分支、卷积神经网络 (CNN) —视觉变换器 (ViT) 混合分支和卷积神经网络 (CNN) 分支。每个分支均输出一个尺寸为 $256 \times 256 \times 14$ 三维张量，分别对应 3 个类别概率图 P^{ViT} 、 $P^{CNN-ViT}$ 、 P^{CNN} 。每个类别概率图里的每个通道对应一个尺寸为 256×256 的二维矩阵，代表图 1 中 14 个类别之一的二维预测概率分布图。

具体而言，ViT 分支采用 ConViT 模型，以探索基于自注意力机制的特征提取效果；CNN-ViT 混合分支采用 CoAtNet 模型，结合卷积与变换器优势实现特征融合；CNN 分支采用 FCN 模型，侧重于传统卷积网络的特征提取能力。

P_{fused} 可以以用下面的两个数学公式表示：

$$P_{fused}^{(c)} = w_1 \times P_{:::c}^{ViT} + w_2 \times P_{:::c}^{CNN-ViT} + w_3 \times P_{:::c}^{CNN} \quad (3)$$

$$P_{fused} = [P_{fused}^{(0)}, P_{fused}^{(1)}, \dots, P_{fused}^{(13)}] \quad (4)$$

式中， $P_{:::c}^{ViT}$ 表示从视觉变换器 (ViT) 分支获得的与第 c 类相对应的 256 像素 $\times$ 256 像素的二维预测概率矩阵。 $P_{:::c}^{CNN-ViT}$ 表示从卷积神经网络 (CNN) 与视觉变换器 (ViT) 混合分支获得的与第 c 类相对应的 256 像素 $\times$ 256 像素的二维预测概率矩阵。 $P_{:::c}^{CNN}$ 表示从卷积神经网络 (CNN) 分支获得的与第 c 类相对应的 256 像素 $\times$ 256 像素的二维预测概率矩阵。

w_1 、 w_2 、 w_3 分别表示图 3 当中的权重 weight 1、weight 2、weight 3。代表不同模型的权重系数。

$P_{fused}^{(c)}$ 表示 3 个第 c 个类别相对应的 256 像素 $\times$ 256

像素的二维预测概率矩阵 $P_{\text{fused}}^{\text{ViT}}$ 、 $P_{\text{fused}}^{\text{CNN-ViT}}$ 、 $P_{\text{fused}}^{\text{CNN}}$ 经过融合后得到的新的二维预测概率矩阵。 P_{fused} 表示在通道维度上将二维预测概率矩阵 $P_{\text{fused}}^{(c)}$ 从第一个相对应的 $256 \text{ 像素} \times 256 \text{ 像素}$ 的二维预测概率矩阵 $P_{\text{fused}}^{(0)}$ 到最后一个的 256×256 的二维预测概率矩阵 $P_{\text{fused}}^{(13)}$ 合并, 形成新的尺寸大小为 $256 \text{ 像素} \times 256 \text{ 像素} \times 14$ 通道的三维的类别概率图 P_{fused} 。

通过式 (5) 进行转换, 最终输出为尺寸为 $256 \text{ 像素} \times 256 \text{ 像素}$ 的二维矩阵 P_{result} , 该输出结果对应了经过模型后生成的土地覆盖图。

$$P_{\text{result}}(a, b) = c[\max_{c \in [0, 13]} P_{\text{fused}}(a, b, c)] \quad (5)$$

式 (5) 描述了二维矩阵 P_{result} 中每个元素与三维张量 P_{fused} 对应元素之间的关系。其中, c 的取值范围为 0 到 13 的整数 (包括 0 和 13), 对应 14 个类

别索引, 每个特定的 c 值表示一个类别索引, 并与 P_{fused} 中的某一通道索引相对应。具体类别与索引的对应关系可参考图 1 中的类别标注。在公式中, a 表示 P_{result} 和 P_{fused} 中相对应的任意行索引。 b 表示 P_{result} 和 P_{fused} 中相对应的任意列索引。

式 (5) 说明了二维矩阵 P_{result} 中任意元素 $P_{\text{result}}(a, b)$ 的值与三维张量 $P_{\text{fused}}(a, b, c)$ 的关系。当 c 在 0 至 13 的整数范围内 (包括 0 和 13), 且 $P_{\text{fused}}(a, b, c)$ 达到最大值时, 最大值对应的 c 的值即为 $P_{\text{result}}(a, b)$ 的值。这一 c 值与图 1 中的类别标注对应。

最后, 通过对 P_{result} 与图 3 左侧的输入数据集所对应的真实土地覆盖图标签进行对比, 测算 OA 与 FWIoU 这两个指标, 从而对该多模型融合策略的性能表现作出评估。

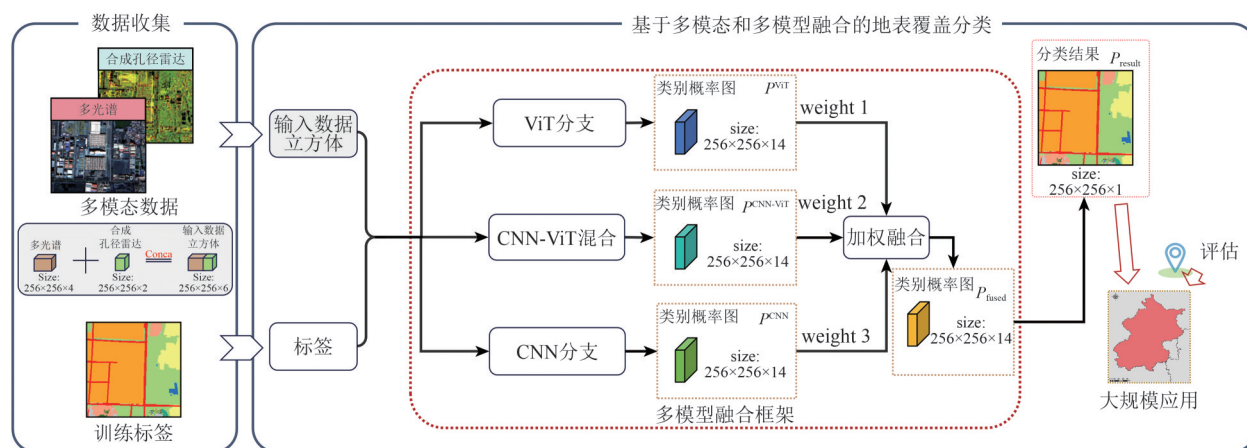


图3 多模型融合的框架示意图

Fig. 3 Schematic diagram of the multi-model fusion framework

4 实验设置与结果分析

4.1 实验设置

本实验中硬件环境采用英伟达公司提供的 RTX 4090 GPU。软件环境方面, 使用 Python 3.9 编程语言和 PyTorch 2.3 框架。

在深度学习模型的训练过程中, 参数设置如下: 训练轮次 (epoch) 为 600, 每轮包含 892 次迭代; 批量大小 (batch size) 为 32。

训练轮次的选择具有其科学性, 通过对训练过程中的性能变化进行观察与分析, 本文确定了 600 轮训练的合理性。具体原因如下:

(1) 训练初期性能的快速提升: 如图 4 和图 5 所示, 模型在前 100 次迭代中, 性能 (如 FWIoU 和 OA 指标) 显著提升。之后, 随着训练轮次的增

加, 性能提升的速率逐渐减缓, 趋于平稳。因此, 600 轮训练能够保证模型充分学习并达到稳定的性能。

(2) 实验观察与性能稳定性: 我们通过多次实验发现, 600 轮训练次数能够使模型达到较高的稳定性, 且进一步增加训练轮次对性能的提升已经非常有限。由图 4 和图 5 可见, 模型性能在前期迭代中的显著提升之后, 进入了稳定阶段, 因此 600 轮训练是一个合理的选择。

(3) 避免过拟合与模型稳定性: 考虑到训练过程中模型性能的逐渐稳定, 选择 600 轮训练不仅避免了过拟合的风险, 还能确保模型的性能在较长时间内维持一致性。

除此以外, 输入数据的空间尺寸为 $256 \text{ 像素} \times 256 \text{ 像素}$ 的图块, 在训练时随机裁剪为 4 个尺寸为

224 像素×224 像素的子图块，并相应划分标签以匹配子图块。优化器选用 AdamW，初始学习率设为 0.001；损失函数为交叉熵损失函数（Cross-Entropy Loss）；学习率调度器采用 ReduceLROnPlateau。

上述配置旨在优化模型的训练效果，并确保实验的科学性。

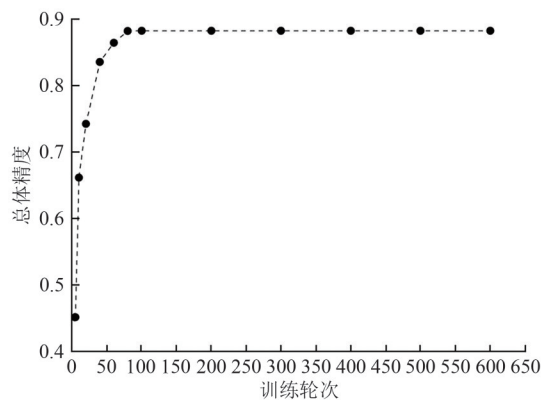


图4 多模态(MSI+SAR)数据下,ConViT模型的输出结果的OA平均值与训练轮次的关系

Fig. 4 The relationship between the average FWIoU of the ConViT model's output results and the number of training iterations under multimodal (MSI+SAR) data

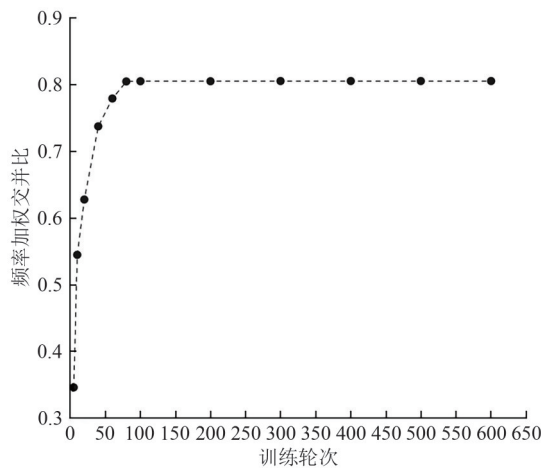


图5 多模态(MSI+SAR)数据下,ConViT模型的输出结果的FWIoU平均值与训练轮次的关系

Fig. 5 The relationship between the average FWIoU of the ConViT model's output results and the number of training iterations under multimodal (MSI+SAR) data

4.2 多模态数据与单模态数据的对比实验

表1记录了单模态SAR数据在CoAtNet、ConViT、FCN、HRNet、LinkNet、U-Net、Random Forest等6个深度学习语义分割方法和一个经典机器学习像素分类模型中的性能表现。在性能评估方面，采用总体精度和频率加权交并比这两项指标。表格中

的数据，是各模型在测试时，基于同样的1428组数据所得到OA和FWIoU的均值。

表2呈现了多模态（MSI+SAR）数据在上述7种模型中的性能表现情况，评估指标同样设定为OA与FWIoU。表中数据源于与表1相同的1428组测试数据输出结果的OA与FWIoU的平均值。

表1 不同的模型在单模态数据上的性能

Table 1 The performance of different models on Single-modality data

模型	OA	FWIoU
CoAtNet	<u>59.04</u>	<u>50.44</u>
ConViT	80.00	68.98
FCN	52.22	41.88
HRNet	53.27	44.74
LinkNet	51.53	43.08
U-Net	48.99	41.13
Random Forest	41.52	34.46

注：加粗为最优模型,加下划线为次优模型。

表2 不同的模型在多模态数据上的性能

Table 2 The performance of different models on Multi-modality data

Models	OA	FWIoU
CoAtNet	64.36	55.34
ConViT	88.21	80.48
FCN	<u>66.23</u>	<u>56.35</u>
HRNet	58.37	49.63
LinkNet	58.87	50.71
U-Net	54.20	45.99
Random Forest	44.66	37.08

注：加粗为最优模型,加下划线为次优模型。

借助双轴折线图，对各异模型于单模态SAR数据和多模态（MSI+SAR）数据上的性能表现进行了对比，详见图6。

结合表1—2还有图6，可以得出以下结论：在这7个模型中，相较于由单一SAR数据构成的单模态数据，融合MSI数据的多模态数据在平均OA和平均FWIoU方面均实现了显著提升。这充分说明多模态数据的引入可有效强化模型对地表特征的识别能力；此外，就单模态SAR数据和多模态（MSI+SAR）数据而言，ConViT模型在上述两项指标上的表现均优于其余模型，体现出更为强大的特征提取与语义分割能力。

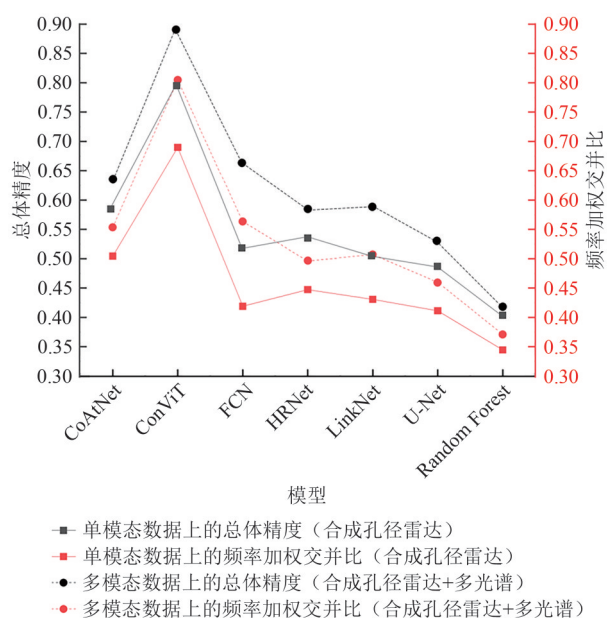


图6 不同模型在多模态数据和单模态数据下的性能表现

Fig. 6 Performance evaluation of different models on multi-modal SAR+MSI data and single-modal SAR data

4.3 多模型融合实验

为更有效地提升模型性能，在多模态数据的支撑下，本研究创新性地推出一种多模型融合策略。该策略融合了3种性能出色的深度学习语义分割模型的优势之处：以ViT为核心结构的ConViT模型；CNN与ViT相混合的CoAtNet模型；以及CNN为主要结构的FCN模型。

如表3所示，根据图3中定义的3个分支权重weight1 (w_1)、weight2 (w_2)、weight3 (w_3)进行了7组不同参数组合的对比实验。这些权重组合的选择依据主要考虑了以下4个方面：

(1) 基于表2“不同的模型在多模态数据上的性能”的结果，我们发现CoAtNet与FCN的精度相近，故赋予两者相等权重有助于在多模型融合中平衡两者的贡献，即 w_2w_3 。这样的权重设置确保了二者能够互补并共同增强最终模型的性能，避免任何一个分支的过度主导。

(2) 由于ConViT模型在OA和FWIoU上表现出色，相对于其他模型（如CoAtNet和FCN），它能够提供更高质量的特征表示。因此通过逐步增加，可以测试ConViT模型在多模型融合中的影响，尤其是当ConViT的权重逐渐增强时，是否能带来性能上的明显提升。

(3) 关于权重组合 $w_1 = \frac{1}{3}$, $w_2 = \frac{1}{3}$, $w_3 = \frac{1}{3}$ ，是为了在没有偏向某一模型的情况下，验证多个模型的特征如何在融合中共同工作，并观察是否能获得较好的性能。

(4) 除此以外， w_1 步长的选择上，通过使用0.1的步长来调整，能够在合理的实验范围内有效捕捉到不同权重配置对最终性能的影响较大的步长使得实验的每一次变化都能反映出显著的性能差异，从而有助于快速识别最佳的权重组合。相比之下，过小的步长（如0.05）会增加实验次数，可能导致细微的变化对最终结果的影响不显著，同时也增加了计算和分析的负担。

在此过程中，以多模态（MSI+SAR）数据作为输入，先对ConViT模型、CoAtNet模型、FCN模型进行训练，所得模型作为后续实验的基础。针对7组不同权重组合的融合模型，在测试时，依据表3中的OA和FWIoU2个指标，对各组合进行评估。每组权重组合的测试结果取自与“4.2多模态数据与单模态数据的对比实验”一致的1428组数据的输出结果的平均值。

表3 在测试阶段不同权重组合下的结果

Table 3 The results of different weights in the test phase

权重	OA/%	FWIoU/%
$w_1=0.6, w_2=0.2, w_3=0.2$ (baseline)	89.28	81.89
$w_1 = \frac{1}{3}, w_2 = \frac{1}{3}, w_3 = \frac{1}{3}$	84.93	76.35
$w_1=0.4, w_2=0.3, w_3=0.3$	87.16	79.07
$w_1=0.5, w_2=0.25, w_3=0.25$	<u>89.09</u>	<u>81.65</u>
$w_1=0.7, w_2=0.15, w_3=0.15$	89.09	81.64
$w_1=0.8, w_2=0.1, w_3=0.1$	88.81	81.25
$w_1=0.9, w_2=0.05, w_3=0.05$	88.51	80.86

注：加粗为最优模型，加下划线为次优模型。

实验结果表明，在当 $w_1=0.6$ 、 $w_2=0.2$ 、 $w_3=0.2$ 的时候，多模态融合策略的整体性能优于其他组合，平均总体精度达到89.28%，平均加权交并比达到81.89%。该权重组合的性能显著优于其余6组组合，表现出更优的融合效果。

除此以外，在 $w_1=0.6$ 、 $w_2=0.2$ 、 $w_3=0.2$ ； $w_1=0.5$ 、 $w_2=0.25$ 、 $w_3=0.25$ ； $w_1=0.7$ 、 $w_2=0.15$ 、 $w_3=0.15$ ； $w_1=0.8$ 、 $w_2=0.1$ 、 $w_3=0.1$ ； $w_1=0.9$ 、 $w_2=0.05$ 、 $w_3=0.05$ 这5组权重组合下，该多模型融合策略的平均总体精

度和平均频率加权交并比均优于以多模态数据为输入时性能表现最好的单模型 ConViT。在多模态数据输入的条件下, ConViT 的平均总体精度和平均频率加权交并比分别为 88.21% 和 80.48%, 而多模型融合策略在这两项指标上表现出更优异的性能。

4.4 可视化结果展示

图 7 展示了单模态 SAR 数据下不同深度学习模型的输出结果对比; 图 8 展示了多模态 (SAR+

MSI) 下不同深度学习模型的输出结果对比。

通过对比图 7 (b) 和图 8 (b) 的分类结果与对应空间位置的真实标签可见: 单模态 SAR 数据下的分类结果, 如图 7 (b) 所示, 与真实标签的吻合度较低; 而引入 MSI 数据构成多模态输入后, 如图 8 (b) 所示, 分类结果的准确性显著提升。这一现象验证了多模态 (MSI+SAR) 数据在土地覆盖分类中的优越性, 其通过互补性特征增强了模型的判别能力。

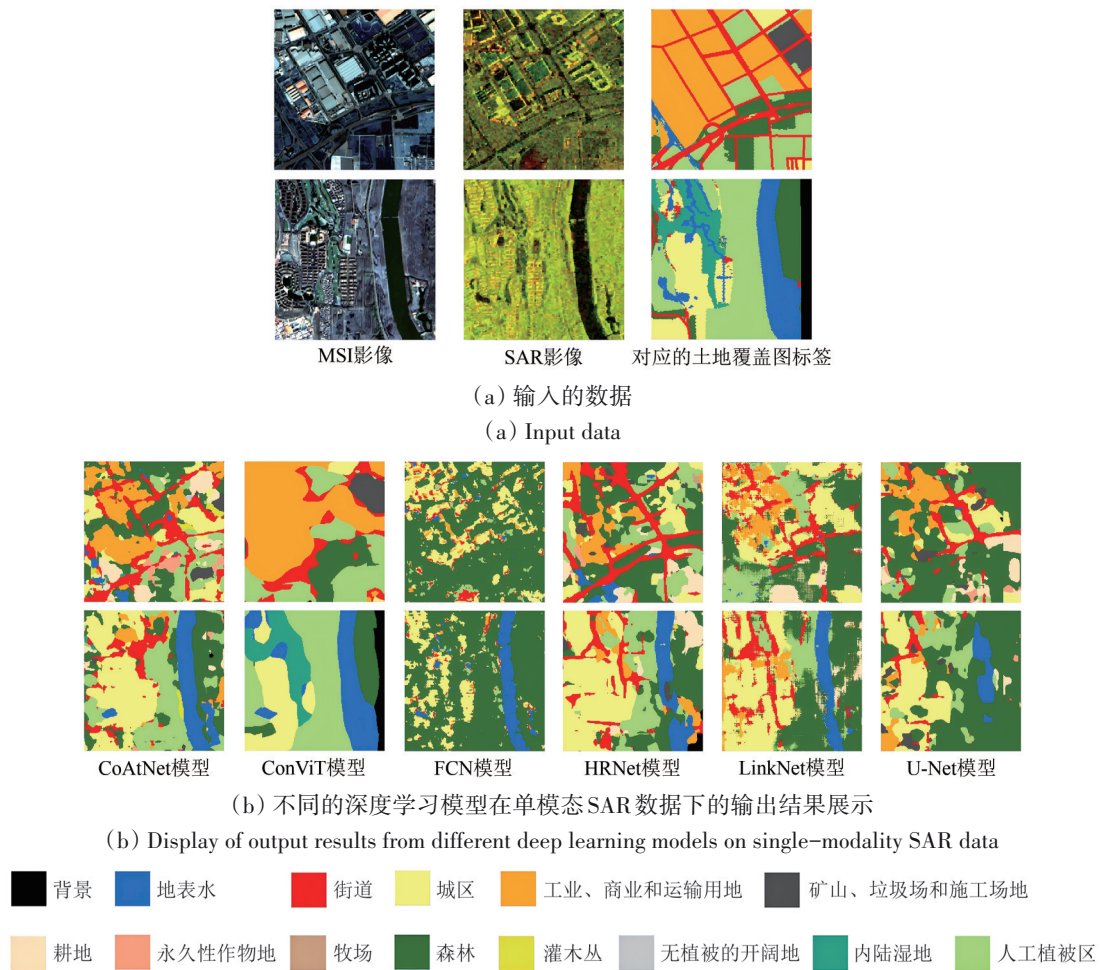


图 7 不同深度学习模型在单模态 SAR 数据下的输出结果对比

Fig. 7 The comparison of output results from different deep learning models on single-modal SAR data

图 9 (b) 展示了在多模态数据作为输入条件下, 不同权重组合的多模型融合 (MSI+SAR) 数据融合实验结果。与图 9 相比, 图 8 (b) 中的可视化结果在不同程度上仍存在一定局限性。例如, CoAtNet、HRNet 和 LinkNet 模型在细节信息捕捉方面能力不足, 导致分类结果中出现较为明显的裂

缝; 而 FCN 模型则由于过于侧重局部特征提取, 忽视了整体空间结构一致性。上述对比结果充分表明, 通过采用合理的权重组合进行多模型融合, 能够在多模态 (MSI+SAR) 数据基础上进一步优化土地覆盖制图的分类效果。

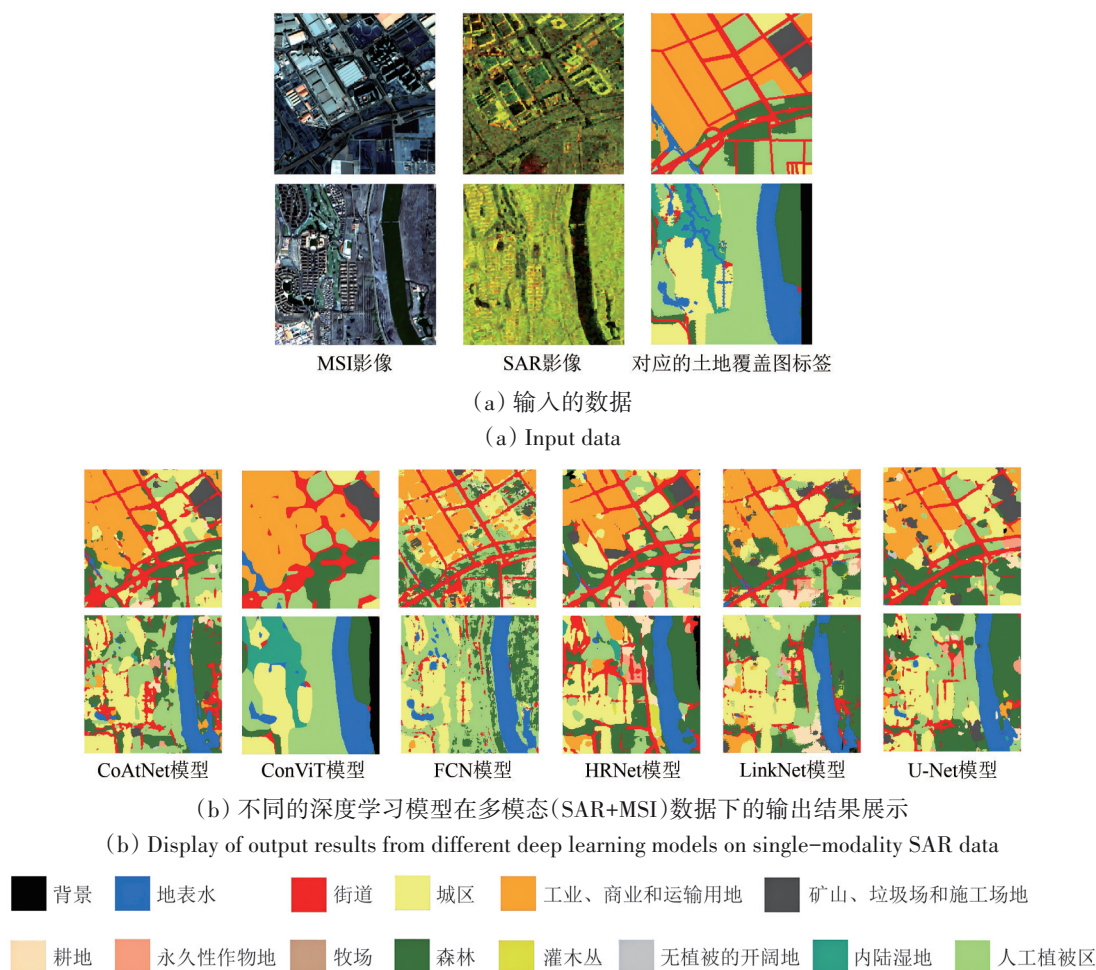


图8 不同深度学习模型在多模态(SAR+MSI)数据下的输出结果对比

Fig.8 The comparison of output results from different deep learning models on multimodal (SAR+MSI) data

5 讨论

在本研究中,我们采用了总体精度(OA)和频率加权交并比(FWIOU)作为性能评估的主要指标,并通过可视化结果对比,展示了所提出的多模型融合方法在土地覆盖制图任务中的有效性。尽管通过OA和FWIOU指标可以定量地评估模型的性能,但仅通过这些指标和视觉结果并不能完全揭示其背后的原因。因此,本节将结合遥感技术的背景,深入探讨所提出方法的优势及其背后的科学原因。

5.1 遥感数据的多样性及复杂性

在土地覆盖制图任务中,遥感影像的应用面临着地表异质性强、地物类型复杂等关键挑战(Lu等,2010)。地形地貌的多样性以及地物光谱

特征的时空变异性,使得基于单一数据源的分类方法往往难以实现理想的精度。具体而言,遥感数据的多波段特性虽然提供了丰富的光谱信息,但同时也带来了同物异谱(同一地物在不同波段呈现不同光谱响应)和异物同谱(不同地物在特定波段表现相似)等固有难题(Sharma等,2018; Van Der Werff和Van Der Meer,2008; Mickaelian和Sargsyan,2010)。例如,在同物异谱问题中,某些地物类别(如森林与耕地)可能在某些波段上具有相似的光谱响应,但由于生长环境及其物理特性不同,这些地物在不同波段的表现存在显著差异。而在异物同谱问题中,不同地物类别(如水体与湿地)在某些光谱波段的反射特征高度重叠,从而导致单模态遥感影像在此情境下的分类精度显著下降。

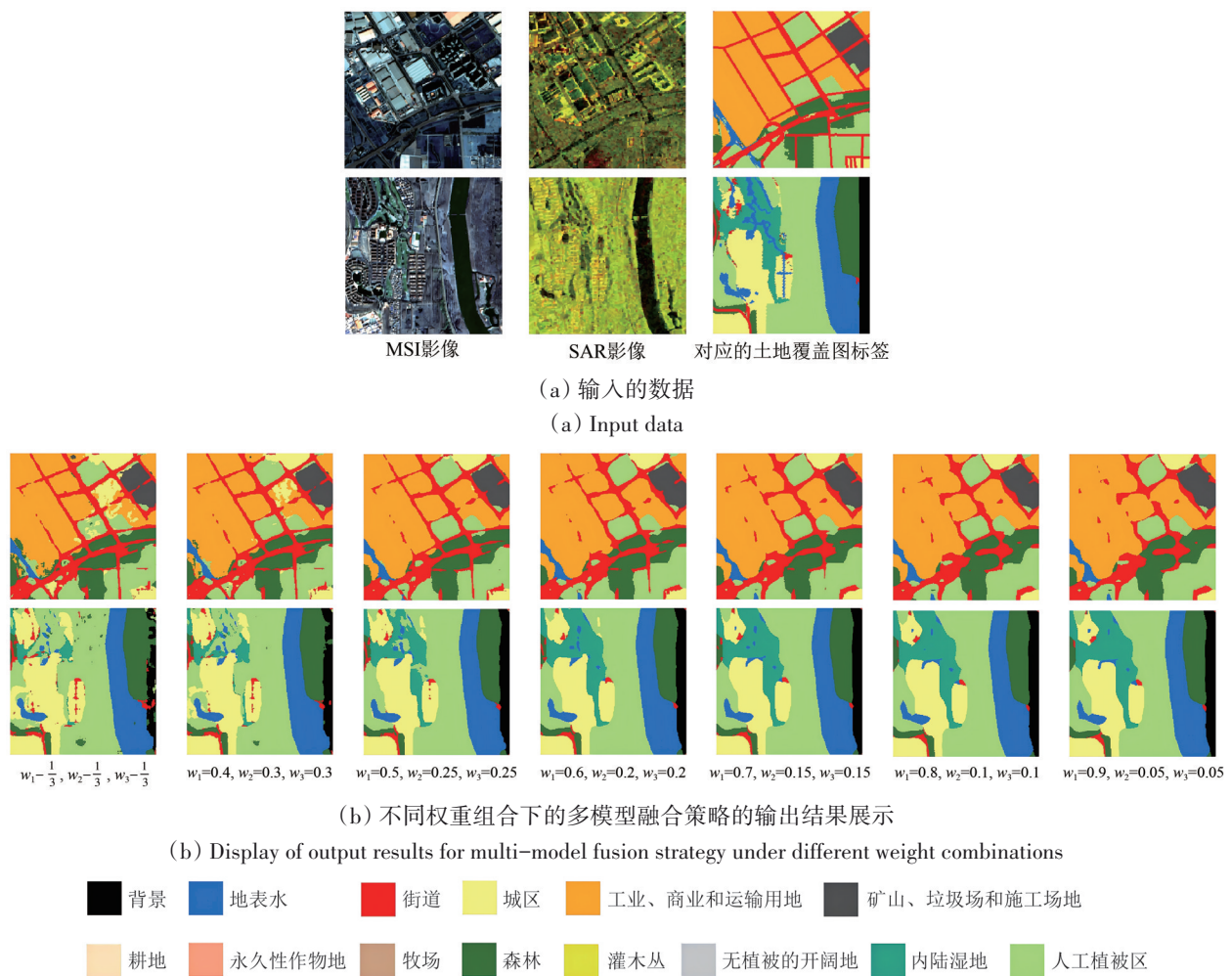


图9 基于多模态(SAR+MSI)数据的多模型融合策略在不同权重组下的输出效果

Fig. 9 The output results of the multimodel fusion strategy based on multimodal (SAR+MSI) data under different weight combinations

这一问题在山区、湿地等复杂场景中尤为突出。复杂的地形起伏与密集的地物分布使得不同类别的地物光谱特征呈现出更高的相似性，加之噪声干扰、环境变化及光照条件等因素的影响，进一步降低了传统单模态遥感影像的分类可靠性(Rehman等, 2021)。针对这一挑战，多模态数据融合策略展现出显著优势。其中，MSI影像通过多波段光谱信息能够有效捕捉地物的光谱差异，在同物异谱问题的解决上具有独特价值；而SAR数据则通过地表结构特征的提取，在异物同谱场景中表现出较强的区分能力。

实验结果表明，MSI与SAR数据的协同融合显著提升了模型的分类性能(表1和表2)。这种多模态融合策略不仅通过光谱—结构特征的互补增强了模型的判别能力，更在复杂地形区域表现出优异的适应性。

5.2 多模型融合的优势

由表1—3可见，虽然ConViT模型在OA和FWIoU指标上优于CoAtNet和FCN等单模型，但通过融合这3种模型，分类性能得到了进一步提升。这种性能提升主要源于各模型在特征提取和分类决策层面的互补性协同效应(Arpitha等, 2023)。具体而言，ConViT模型凭借其全局注意力机制，在长程依赖建模方面表现出色；CoAtNet则通过结合CNN的局部特征提取能力和ViT的全局建模优势，在处理复杂空间结构地物时展现出更强的鲁棒性，特别是在地物边界模糊、地形变化剧烈的区域；而FCN模型则因其全卷积结构，在细节特征捕捉和精细分类任务中具有独特优势。因此，通过多模型融合策略，不仅能够保留各个模型的优势，还能够弥补单一模型在特定地物类别上的弱点，从而提升了整体分类精度。

5.3 地物细节的分类

本研究涵盖14类具有显著异质性的地物类型,包括自然与人工环境中的典型地表覆盖物。具体而言,研究中的14类地物类型包括:背景,通常表示为不确定或无关部分,常被去除以突出重要内容;地表水,包括河流、湖泊和水库等水域;街道,包括城市或农村的道路系统;城区,涵盖高度建设的城市区域,包括住宅与商业区;工业、商业和运输用地,包括工业设施、商业场所及交通基础设施,如机场和火车站;矿山、垃圾场和施工场地,包括地表矿、废物管理区及建筑工地;人工植被区,指人工栽培的草坪、绿地和花园等植被;耕地,用于农业生产的农田;永久性作物地,种植果树、葡萄和橄榄等长期作物的区域;牧场,草地及其他用于放牧的区域;森林,包括天然生长和人工栽培的针叶林与阔叶林;灌木丛,以灌木植物为主的覆盖区域;无植被的开阔地,指裸土、沙地、岩石区及其他无植被覆盖的土地;内陆湿地,包括非沿海区域,如沼泽和泥炭地。

5.4 可视化结果分析

从图7—图9中的可视化结果可以直观地看到,融合模型在不同地物类型的识别上具有显著优势。尤其是在地形复杂、地物类型繁多的区域,融合模型在细节上的表现更加突出,能够较好地地区分相似地物类别,减少分类裂缝和误分类现象。此外,图9中的不同权重组合下的融合结果也表明,合理的权重设置能够显著提升最终的分类效果,优化模型的性能表现。

具体而言,图9(b)展示了不同权重组合的多模型融合策略在处理地形地貌杂、地物细节丰富的区域时的优势。通过合理设置各模型的权重,能够优化模型对不同地物类型的响应,提升城市、湿地及森林等复杂地物类别的识别能力。对比图8和图7(b),在多模态(MSI+SAR)数据输入下的模型分类结果相比单模态SAR数据,显示出更加细致的地物分布,尤其在市区和水体区域,分类精度有了显著提升。

6 结 论

本研究创新性地提出了一种面向土地覆盖制图的多模态数据融合与多模型协同优化框架。在数据层面,通过MSI影像与SAR数据的通道级融

合,构建了具有互补特征表达的多模态数据集。实验验证表明,该数据融合策略较单一SAR数据显著提升了模型的分类精度和泛化能力。

在模型架构方面,本研究提出的多模型融合机制通过协同优化CNN分支、ViT分支、CNN和ViT相结合的混合分支,实现了局部特征提取与全局上下文建模的优势互补。定量实验与可视化分析均证实,该融合策略能有效提升复杂场景下的地物识别精度和鲁棒性。

本研究为基于深度学习的土地覆盖制图提供了新的技术思路,其多模态数据处理框架和多模型融合策略不仅为土地覆盖制图领域提供了重要的理论支持,同时也为未来研究和实际应用提供了新思路 and 方向。

志 谢 本研究所使用的C2Seg-BW数据集来源于文献“Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks”,我们对此表示由衷的感谢。该数据集已公开并免费提供,可通过以下链接获取:<https://github.com/danfenghong/Outcome-of-the-2023-IEEE-WHISPERS-C2Seg-Challenge>。

参考文献(References)

- Ajibola S and Cabral P. 2024. A systematic literature review and bibliometric analysis of semantic segmentation models in land cover mapping. *Remote Sensing*, 16(12): 2222 [DOI: 10.3390/rs16122222]
- Brown W M and Porcello L J. 1969. An introduction to synthetic-aperture radar. *IEEE Spectrum*, 6(9): 52-62 [DOI: 10.1109/MSPEC.1969.5213674]
- Chaurasia A and Culurciello E. 2017. LinkNet: exploiting encoder representations for efficient semantic segmentation//2017 IEEE Visual Communications and Image Processing (VCIP). St. Petersburg: IEEE: 1-4 [DOI: 10.1109/vcip.2017.8305148]
- Chua L O. 1997. CNN: a vision of complexity. *International Journal of Bifurcation and Chaos*, 7(10): 2219-2425 [DOI: 10.1142/S0218127497001618]
- Congalton R G, Gu J Y, Yadav K, Thenkabail P and Ozdogan M. 2014. Global land cover mapping: a review and uncertainty analysis. *Remote Sensing*, 6(12): 12070-12093 [DOI: 10.3390/rs61212070]
- Dai Z H, Liu H X, Le Q V and Tan M X. 2021. CoAtNet: marrying convolution and attention for all data sizes//Proceedings of the 35th International Conference on Neural Information Processing

- Systems. Virtual Event: Curran Associates Inc.: 3965-3977
- D'Ascoli S, Touvron H, Leavitt M L, Morcos A S, Biroli G and Sagun L. 2022. ConViT: improving vision transformers with soft convolutional inductive biases. *Journal of Statistical Mechanics: Theory and Experiment*, 2022: 114005 [DOI: 10.1088/1742-5468/ac9830]
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Housley N. 2021. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929* [DOI: 10.48550/arXiv.2010.11929]
- Fitzgerald R W and Lees B G. 1994. Assessing the classification accuracy of multisource remote sensing data. *Remote sensing of Environment*, 47(3): 362-368 [DOI: 10.1016/0034-4257(94)90103-1]
- García-Mora T J, Mas J F and Hinkley E A. 2012. Land cover mapping applications with MODIS: a literature review. *International Journal of Digital Earth*, 5(1): 63-87 [DOI: 10.1080/17538947.2011.565080]
- Grekousis G, Mountrakis G and Kavouras M. 2015. An overview of 21 global and 43 regional land-cover mapping products. *International Journal of Remote Sensing*, 36(21): 5309-5335 [DOI: 10.1080/01431161.2015.1093195]
- Haque M A, Reza M N, Ali M, Karim M R, Ahmed S, Lee K D, Khang Y H and Chung S O. 2024. Effects of environmental conditions on vegetation indices from multispectral images: a review. *Korean Journal of Remote Sensing*, 40(4): 319-341 [DOI: 10.7780/kjrs.2024.40.4.1]
- Jacqueminet C, Kermadi S, Michel K, Béal D, Gagnage M, Branger F, Jankowsky S and Braud I. 2013. Land cover mapping using aerial and VHR satellite images for distributed hydrological modeling of periurban catchments: application to the Yzeron catchment (Lyon, France). *Journal of Hydrology*, 485: 68-83 [DOI: 10.1016/j.jhydrol.2013.01.028]
- Joshi N, Baumann M, Ehammer A, Fensholt R, Grogan K, Hostert P, Jepsen M R, Kuemmerle T, Meyfroidt P, Mitchard E T A, Reiche J, Ryan C M and Waske B. 2016. A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring. *Remote Sensing*, 8(1): 70 [DOI: 10.3390/rs8010070]
- LeCun Y, Bottou L, Bengio Y and Haffner P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278-2324 [DOI: 10.1109/5.726791]
- Li Z H, He W, Cheng M F, Hu J X, Yang G Y and Zhang H Y. 2023. SinoLC-1: the first 1 m resolution national-scale land-cover map of China created with a deep learning framework and open-access data. *Earth System Science Data*, 15(11): 4749-4780 [DOI: 10.5194/essd-15-4749-2023]
- Li Z H, He W, Li J P, Lu F X and Zhang H Y. 2024. Learning without exact guidance: updating large-scale high-resolution land cover maps from low-resolution historical labels//*Proceedings of the 2014 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle: IEEE: 27717-27727 [DOI: 10.1109/cvpr52733.2024.02618]
- Li Z H, Zhang H Y, Lu F X, Xue R Y, Yang G Y and Zhang L P. 2022. Breaking the resolution barrier: a low-to-high network for large-scale high-resolution land-cover mapping using low-resolution labels. *ISPRS Journal of Photogrammetry and Remote Sensing*, 192: 244-267 [DOI: 10.1016/j.isprsjprs.2022.08.008]
- Liu Y H, Wang Y, Zhang Y F, Mei S H, Zou J Q, Li Z H, Lu F X, He W, Zhang H Y, Zhao H L, Chen C, Xia C, Li H, Vivone G, Hänsch R, Taskin G, Yao J, Qin A K, Zhang B, Chanussot J and Hong D F. 2024. Cross-City Semantic Segmentation (C2Seg) in multimodal remote sensing: outcome of the 2023 IEEE WHIS-PERS C2Seg Challenge. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17: 8851-8862 [DOI: 10.1109/jstars.2024.3388464]
- Lu D S, Hetrick S and Moran E. 2010. Land cover classification in a complex urban-rural landscape with QuickBird imagery. *Photogrammetric Engineering and Remote Sensing*, 76(10): 1159-1168 [DOI: 10.14358/PERS.76.10.1159]
- Arpitha M, S A Ahmed and Harishnaika N. 2023. Land use and land cover classification using machine learning algorithms in Google earth engine. *Earth Science Informatics*, 16(4): 3057-3073 [DOI: 10.1007/s12145-023-01073-w]
- Mickaelian A M and Sargsyan L A. 2010. Spectral observations of big objects. *Astrophysics*, 53(4): 483-489 [DOI: 10.1007/s10511-010-9139-4]
- Oetter D R, Cohen W B, Berterretche M, Maersperger T K and Kennedy R E. 2001. Land cover mapping in an agricultural setting using multiseasonal Thematic Mapper data. *Remote Sensing of Environment*, 76(2): 139-155 [DOI: 10.1016/s0034-4257(00)00202-9]
- Paheding S, Saleem A, Siddiqui M F H, Rawashdeh N, Essa A and Reyes A A. 2024. Advancing horizons in remote sensing: a comprehensive survey of deep learning models and applications in image classification and beyond. *Neural Computing and Applications*, 36(27): 16727-16767 [DOI: 10.1007/s00521-024-10165-7]
- Quan Y H, Tong Y P, Feng W, Dauphin G, Huang W J and Xing M D. 2020. A novel image fusion method of multi-spectral and SAR images for land cover classification. *Remote Sensing*, 12(22): 3801 [DOI: 10.3390/rs12223801]
- Rehman A U, Ullah S, Shafique M, Khan M S, Badshah M T and Liu Q J. 2021. Combining Landsat-8 spectral bands with ancillary variables for land cover classification in mountainous terrains of northern Pakistan. *Journal of Mountain Science*, 18(9): 2388-2401 [DOI: 10.1007/s11629-020-6548-7]
- Rodriguez-Galiano V F, Ghimire B, Rogan J, Chica-Olmo M and Rigol-Sanchez J P. 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67: 93-104 [DOI: 10.1016/j.isprsjprs.2011.11.002]

- Ronneberger O, Fischer P and Brox T. 2015. U-net: convolutional networks for biomedical image segmentation//18th International Conference on Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. Munich: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]
- Sharma A, Liu X W and Yang X J. 2018. Land cover classification from multi-temporal, multi-spectral remotely sensed imagery using patch-based recurrent neural networks. *Neural Networks*, 105: 346-355 [DOI: 10.1016/j.neunet.2018.05.019]
- Shelhamer E, Long J and Darrell T. 2017. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4): 640-651 [DOI: 10.1109/TPAMI.2016.2572683]
- Thyagarajan K K and Vignesh T. 2019. Soft computing techniques for land use and land cover monitoring with multispectral remote sensing images: a review. *Archives of Computational Methods in Engineering*, 26(2): 275-301 [DOI: 10.1007/s11831-017-9239-y]
- Tong X Y, Dong R M and Zhu X X. 2025. Global high categorical resolution land cover mapping via weak supervision. *ISPRS Journal of Photogrammetry and Remote Sensing*, 220: 535-549 [DOI: 10.1016/j.isprsjprs.2024.12.017]
- Van Der Werff H M A and Van Der Meer F D. 2008. Shape-based classification of spectrally identical objects. *ISPRS Journal of Photogrammetry and Remote Sensing*, 63(2): 251-258 [DOI: 10.1016/j.isprsjprs.2007.09.007]
- Wang J D, Sun K, Cheng T H, Jiang B R, Deng C R, Zhao Y, Liu D, Mu Y D, Tan M K, Wang X G, Liu W Y and Xiao B. 2021. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10): 3349-3364 [DOI: 10.1109/TPAMI.2020.2983686]
- Wickham J, Stehman S V, Sorenson D G, Gass L and Dewitz J A. 2021. Thematic accuracy assessment of the NLCD 2016 land cover for the conterminous United States. *Remote Sensing of Environment*, 257: 112357 [DOI: 10.1016/j.rse.2021.112357]
- Wu H, Li Y, Lin A Q, Fan H C, Fan K X, Xie J Y and Luo W T. 2024. A review of crowdsourced geographic information for land-use and land-cover mapping: current progress and challenges. *International Journal of Geographical Information Science*, 38(11): 2183-2215 [DOI: 10.1080/13658816.2024.2379468]
- Xu P P, Tsendbazar N E, Herold M, De Bruin S, Koopmans M, Birch T, Carter S, Fritz S, Lesiv M, Mazur E, Pickens A, Potapov P, Stolle F, Tyukavina A, Van De Kerchove R and Zanaga D. 2024. Comparative validation of recent 10 m-resolution global land cover maps. *Remote Sensing of Environment*, 311: 114316 [DOI: 10.1016/j.rse.2024.114316]
- Yang J and Huang X. 2021. The 30 m annual land cover dataset and its dynamics in China from 1990 to 2019. *Earth System Science Data Discussions*, 13(8): 3907-3925 [DOI: 10.5194/essd-13-3907-2021]
- Zhang T T, Du Z R, Yang J Y, Yao X C, Ou C, Niu B W and Yan S. 2021. Land cover mapping and ecological risk assessment in the context of recent ecological migration. *Remote Sensing*, 13(7): 1381 [DOI: 10.3390/rs13071381]
- Hong D, Zhang B, Li H, Li Y, Yao J, Li C, Werner M, Chanussot J, Zipf A, Zhu X. 2023. Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. *Remote Sensing of Environment*, 299: 113856. [DOI: 10.1016/j.rse.2023.113856]

Large-scale land cover mapping based on multimodal data and multimodel fusion

RU Yifan^{1,2}, TANG Guoliang¹, LI Chunlai²

1. Hangzhou Institute for Advanced Study, School of Physics and Optoelectronic Engineering, University of Chinese Academy of Sciences, Hangzhou 310024, China;

2. Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China

Abstract: Land cover mapping is a vital task in Earth observation; it provides fine-scale details of landscape and support many downstream applications like ecology, hydrology, and resource management. However, current land cover mapping faces key challenges such as limited information from single-source data, substantial data heterogeneity, and insufficient generalization capability of individual models. To solve these issues, this study addresses key challenges in land cover mapping by proposing an innovative framework that integrates multimodal remote sensing data with a multimodel deep learning-based framework for collaborative decision-making. This work aims to provide a novel pathway for large-scale high-resolution land cover mapping and support many other downstream applications.

In this study, leveraging the spectral characteristics of multispectral imagery (MSI) and the distinct properties of Synthetic Aperture Radar (SAR) data, a complementary multimodal (i. e., MSI+SAR) dataset is constructed as feature input, effectively overcoming the limitations of using a single SAR modality in complex Earth observation scenarios. Moreover, at the model architecture level, a systematic

evaluation of the performance differences among seven representative machine learning-based models is conducted. Based on these foundations, a multimodel fusion strategy is further proposed, combining convolutional neural networks (CNNs), vision transformers (ViTs), and a hybrid CNN/ViT architecture (represented by FCN, ConViT, and CoAtNet, respectively). These three models have demonstrated outstanding performance in previous comparative experiments between multimodal (i.e., SAR+MSI) data and single-modal SAR data.

In the experimental section, we conduct a comprehensive evaluation in the entire Beijing City. Remote sensing data are collected from GF-3 (SAR) and GF-6 (MSI). On the basis of these multimodal data, we compare six advanced semantic segmentation models and one pixel-based classification model. Results show that compared with using single SAR data, the utilization of multimodal data significantly improves two key evaluation metrics used in semantic segmentation task or land cover mapping task—Overall Accuracy (OA) and Frequency-Weighted Intersection over Union (FWIoU)—in Beijing City. The proposed multimodel fusion framework further enhances performance in OA and FWIoU, validating the effectiveness of the method in semantic segmentation tasks with multimodal remote sensing data.

This paper presents an innovative approach that enhances a model's capability to extract complex land cover features by effectively integrating multimodal remote sensing data (MSI and SAR) with a multimodel fusion framework. The proposed method demonstrates superior performance in large-scale land cover mapping, achieving significant improvements in classification accuracy and robustness compared with conventional single-source or -model approaches. The success of this framework highlights the powerful potential of multimodal data fusion and collaborative deep learning in overcoming challenges such as spectral ambiguity, cloud interference, and limited labeled samples.

Key words: land-cover mapping, multi-spectral, synthetic aperture radar, multi-modal, multi-model fusion, convolutional neural networks, vision transformers

Supported by Zhejiang Provincial Science and Technology Program (No. 2024C03032)