

# 融合分割与分类模型的遥感图像目标检测方法

童恒建, 陈昕浩, 赵 济

(中国地质大学(武汉) 计算机学院, 湖北 武汉 430010)

**摘要:** 为应对深度学习在遥感图像处理中面临的诸多挑战, 如标注过程耗时且复杂、数据多样性丰富、训练模型需要大量计算资源以及容易过拟合等问题, 提出一种融合 Segment Anything Model 分割模型和 Mobile-Friendly Vision Transformer 分类模型的遥感图像目标检测方法。实验结果表明, 相较于 YOLO-V2、Faster R-CNN 等模型, 该模型的参数量为 28.3 M, 平均训练时间为 32.5 s, 计算复杂度为 10.7 GFLOPs, 在数据使用效率、训练时间和分类准确率方面具有优势。

**关键词:** 遥感图像; 深度学习; 分割模型; 分类模型; 目标检测

DOI: 10.11907/rjdk.251067

中图分类号: TP391.41

文献标识码: A

文章编号: 1672-7800(2026)003-0188-07

扫描二维码阅读全文:



## Remote Sensing Image Object Detection Method Integrating Segmentation and Classification Models

TONG Hengjian, CHEN Xinhao, ZHAO Ji

(School of Computer Science, China University of Geosciences, Wuhan 430010, China)

**Abstract:** To address the various challenges that deep learning still faces in remote sensing image processing, such as the time-consuming and complex annotation process, high data diversity, significant computational resources required for model training, and susceptibility to overfitting, this paper proposes a remote sensing image object detection method that integrates the Segment Anything Model (SAM) segmentation model and the Mobile-Friendly Vision Transformer (MobileViT) classification model. Experimental results show that, compared to models such as YOLO-V2 and Faster R-CNN, this model has a parameter size of 28.3M, an average training time of 32.5 seconds, and a computational complexity of 10.7 GFLOPs, demonstrating advantages in data utilization efficiency, training time, and classification accuracy.

**Key Words:** remote sensing images; deep learning; segmentation model; classification model; object detection

### 0 引言

近年来,深度学习已成为图像识别领域的主流技术,凭借其强大的特征提取和非线性拟合能力,突破了传统手工设计特征的局限,让图像识别的精度和场景适配性有了质的飞跃。然而,在遥感图像处理中,深度学习仍面临诸多挑战,如图像标注耗时且依赖专业知识、数据多样性,以及遥感图像类别分布不平衡和数据量大导致计算资源需求较高等。为应对这些问题,本文提出融合分割模型与分类模型的遥感图像识别方法。通过 Segment Anything

Model (SAM) 进行实例提取,利用其强大的泛化能力和零样本学习能力解决目标尺度差异与图像的多样性问题,之后采用 MobileViT 模型对实例对象进行分类,以实现高效的遥感图像目标检测任务。该方法通过任务分解的方式降低了计算资源需求,同时提高了识别精度。相较于传统方法,该方法在训练效率、推理速度和准确率方面更具优势。

### 1 相关工作

在深度学习普及之前,遥感图像分类多依赖于传统的

收稿日期: 2025-02-02

基金项目: 国家自然科学基金面上项目(42171384)

**作者简介:** 童恒建(1970-),男,博士,中国地质大学(武汉)计算机学院教授,研究方向为面向地理对象的图像分析、遥感图像分割与分类、遥感图像处理与分析;陈昕浩(2000-),男,中国地质大学(武汉)计算机学院硕士研究生,研究方向为遥感图像分割与分类;赵济(1990-),男,中国地质大学(武汉)计算机学院副教授、博士生导师,研究方向为人工智能、计算机视觉和图像处理。

机器学习方法,如随机森林(Random Forest, RF)和支持向量机(Support Vector Machine, SVM)等<sup>[1-2]</sup>。这些方法主要依靠人工设计的特征提取策略,结合特定领域知识完成分类。然而,传统方法在处理高维数据时面临性能瓶颈,面对遥感图像丰富的光谱、空间纹理信息以及复杂场地场景下的纹理相似等问题,人工设计特征的表达能力有限,难以捕捉图像数据中的复杂关联。近年来,深度学习在遥感图像领域的应用逐渐成熟,并展现出强大的潜力。例如, Hu 等<sup>[3]</sup>最早尝试使用浅层卷积神经网络(Convolutional Neural Network, CNN)进行高光谱图像分类,通过降低网络复杂度来适应训练样本不足的情况。该方法减少了对训练样本量的需求,但未考虑到空间相关性,只考虑了光谱特征,并采用一种基于上下文分析的深度神经网络框架<sup>[4]</sup>。然而,这些方法都是浅层训练模型的折中方法。随后,深度网络逐步取代浅层网络成为主流方法。Chen 等<sup>[5]</sup>提出基于深度 CNN 的高光谱分类模型,通过 PCA(Principal Component Analysis)对图像进行降维和权重共享策略显著提升了模型性能。同时,注意力机制的引入进一步改善了模型的效率和适应性,例如, Xu 等<sup>[7]</sup>提出基于注意力的多尺度融合框架,采用基于注意力的多尺度融合、局部金字塔注意力以及自适应的深色空间金字塔池化3种机制来处理图像特征,能够处理复杂的空间分布多样性问题。但3种方法的计算复杂度都很高,处理高分辨率图片时计算资源消耗过大。

尽管基于深度学习的目标检测算法有了很大进步,但其在遥感图像领域的性能尚不够理想。针对高分辨率遥感图像的目标检测任务,其仍存在着数据稀缺、类别不均衡、目标多尺度且分布不均衡等问题。为解决这些问题,近年来对于目标检测在遥感图像方面的应用研究主要包括以下几个方向:数据预处理方法、骨干网络改进、损失函数改进<sup>[8-10]</sup>。其中,数据预处理方向包括数据清洗与数据增广。数据清洗是对数据重新进行审核校验,使其更加规范。数据增广是对原数据集进行随机处理,以产生相似且有效的训练样本<sup>[11]</sup>。而骨干网络的好坏以及使用是否合理将直接影响算法的性能<sup>[12]</sup>。同时, Zhang 等<sup>[13]</sup>基于 SAM 模型提出一种针对个性化场景分割的方案,适用于资源受限的场景。但其依赖于单个参考图像,存在尺度不确定性,且对于视频分割任务的适应性有限。除上述研究方向外,部分学者还针对遥感数据标注量少的问题进行了研究。宋云凯等<sup>[14]</sup>基于特征加权和特征融合的方式,利用特征学习模块增加感受野,有效提升了网络对多尺度遥感目标的感知能力,但在目标密集或背景复杂的遥感图像中仍可能存在误检的情况;曾宝秀等<sup>[15]</sup>提出一种结合空洞空间金字塔池化模块和通道、空间注意力模块的方法,能有效结合低级空间细节特征和高级语义信息,提供更全面的特征表示,但其对数据质量存在依赖,且计算复杂度高,参数调优复杂。而吕文琪等<sup>[16]</sup>提出的在编码与解码器跳跃连

接过程中加入一个注意力控制机制的方法,使网络将权重放在更值得注意的地方,有效提取了图像局部特征,降低了计算复杂度,具有较高的分割精度,但存在全局信息捕捉能力弱、边界优化不足、对数据规模有较强依赖等问题。

## 2 模型介绍

### 2.1 分割模型

Segment Anything Model(SAM)的设计目标是能够“分割任何东西”,即能够在各种图像上精确地分割出图像中的目标物体。其具有强大的灵活性和鲁棒性,这一特性使其能够处理具有多样性的遥感图像<sup>[17]</sup>。SAM 的核心优势在于其具备的零样本学习能力。不同于传统分割方法需要对每个新任务进行单独训练, SAM 在一个包含了数百万张图像和分割掩码的大规模数据集上进行了预训练。大规模预训练可帮助模型学习到更加丰富的视觉特征,从而提升其在各种任务中的表现。无论是何种场景或对象, SAM 能够在没有针对特定任务进行专门训练的情况下,处理全新的对象和场景。

SAM 采用一种基于 Transformer 的架构,具有强大的全局特征建模能力,使其能够处理分辨率高、目标尺度差异大的遥感图像。特别是在目标数量众多或类别分布复杂的图像中, SAM 能够精准分割每个目标实例。以下是 SAM 的工作流程:首先,对输入图像进行预处理,将遥感图像分割为多个小块,以减少计算复杂度,同时保留原始分辨率的信息。其次,对图像特征进行提取,通过 Transformer 编码器提取全局特征,并生成高维语义表示。再次,生成实例,结合注意力机制对目标实例进行区域划分和概率预测。最后,输出结果,以分割掩码的形式输出每个目标实例。

SAM 分割流程如图 1 所示。SAM 模型由 3 部分组成:图像编码器、提示编码器以及掩码解码器。图像编码器以及提示编码器分别对图像信息与点信息进行编码,因为需要提取全图的实例对象,所以在提示编码器部分输入稀疏点作为提示,每个点间隔 32 个像素。掩码解码器会提取图像实例对象及其位置信息,提取的对象实例将作为下一阶段的输入保存,其位置信息会单独保存。因为 SAM 模型可能会将图像中的背景、干扰物体提取为对象实例,所以在分类模型中也需要对这些干扰图片进行分类。同时,注意到提取的对象实例图像大小可能不同,需要在输入到分类模型之前将图像大小统一,以便于进行后续分类任务。

为了提升 SAM 模型在高分辨率遥感图像处理中的表现,本文进行了以下优化:将大尺寸遥感图像分割为更小的块来减轻显存压力,并让模型能够捕捉更多细节特征;针对模型输出的分割掩码进行后处理,去除多余的背景噪声,以增强目标区域边界的清晰度。

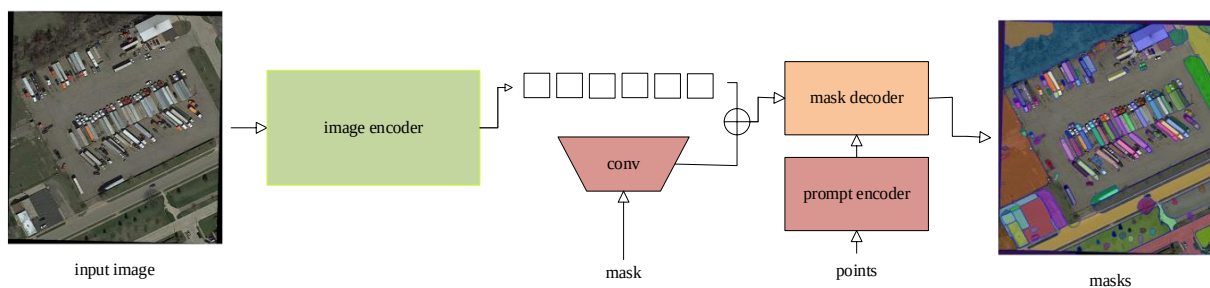


Fig. 1 SAM segmentation process

图1 SAM分割流程

## 2.2 分类模型

MobileViT 模型是一种视觉 Transformer (Vision Transformer, ViT) 架构,旨在结合 Transformer 和卷积神经网络的优点,通过卷积层提取局部特征来处理如边缘、纹理等低级别的视觉信息,以及通过 Transformer 层捕捉全局上下文信息来处理高层次语义信息的方式,提供一种高效的视觉模型<sup>[18]</sup>。该模型与 Transformer 模型最大的不同点在于全局表征层(global representation)。传统 Transformer 模型需先将特征图展平为一维序列后再开展后续计算,其自注意力机制会让序列中的每个元素与其余所有元素进行全局关联计算<sup>[19]</sup>。而 MobileViT 模型会先将输入的特征图划分为若干局部小块,再对所有小块中位置对应的元素进行注意力计算,既大幅降低了模型计算开销,又能在处理高分辨率图像时,取得与传统 Transformer 模型基本相当的建模效果。

由于每个对象实例生成的图片大小可能不同,为统一输入图片大小,本文通过缩小较大图片、使用空白背景填充较小图片的方式统一尺寸。MobileViT 模型在处理高光谱和多光谱数据时表现出色,轻量化的设计使其适合资源有限的场景。为进一步适应遥感图像的特性,本文在原有网络结构基础上增加了多尺度特征融合模块,以捕捉小目标和大面积区域特征,还在训练过程中使用数据增强技术来提升模型的鲁棒性。同时,由于硬件设备的实际计算能力有限,为实现高效的实时推理能力,对模型冗余结构进行剪枝操作,并对模型中参数进行量化处理。

## 2.3 组合模型

在计算机视觉领域,关于图像识别的任务有四大类:分类、定位、检测、分割。其中,分类与定位任务解决图像中的目标是什么,以及目标在图像中的哪里的问题,检测任务定位该目标的位置并知道目标是什么,而分割任务解决每个像素属于哪个目标或场景的问题。

将分割模型与分类模型结合使用的方式达到了目标检测的效果。因为目标检测的任务是定位并识别图像中的目标物体,其基本过程主要包含目标图像特征提取、候选区域生成,以及候选区域分类预测和候选区域位置与大小调整。针对本文提出的组合模型,先通过分割模型对图

像中不同目标物体的每个像素进行精确分割,再利用分类模型对实例对象进行特征提取与类别判定,最终实现目标检测任务。相比目标检测模型需要训练候选框以及对物体类别进行预测,组合模型的方式减少了训练任务。因为在分割模型对图像进行分割的过程中即完成了对目标的定位,所以不需要训练候选框。在目标检测过程中,损失函数如公式(1)所示。

$$Loss = \frac{L_{cls} + \lambda L_{reg} + L_{obj}}{N_{pos}} \quad (1)$$

其中, $L_{cls}$ 表示分类损失, $L_{reg}$ 表示定位损失, $L_{obj}$ 表示对象损失, $\lambda$ 表示定位损失的平衡系数, $N_{pos}$ 表示正样的锚点(Anchor Point)数。 $L_{cls}$ 和 $L_{obj}$ 采用二值交叉熵损失,而 $L_{obj}$ 采用 IoU 损失。

在本文的组合模型中,因为使用 SAM 直接分割图像得到了对象实例,所以只需要计算分类损失,如公式(2)所示。

$$Loss = L_{cls} \quad (2)$$

由于使用的 SAM 模型已通过大型数据集进行了训练,因此本文只需要训练分类模型,从而大大减少了训练样本量以及需要标注的数据量。

## 3 实验与结果分析

### 3.1 实验设置

#### 3.1.1 数据集

本文使用的数据集是 CLRS 以及 TGRS-HRRSD 数据集。其中,CLRS 数据集由 1.5 万张遥感图像组成,共分成 25 个场景类,如机场、桥梁、农田等,每个类有 600 张图片,图片大小为 256\*256,图像分辨率为 0.26~8.85 m。而 TGRS-HRRSD 数据集包含两万多张获取自 Google Earth 和 Baidu map 的图像,空间分辨率为 0.15~1.2 m,其中有 55 000 左右的对象实例。

#### 3.1.2 实验过程

本文设置了两个实验,实验 1 对组合模型的性能进行比较,实验 2 分析使用 SAM 模型提取的实例对象与传统分类图片的差异。实验 1 的流程如图 2 所示。

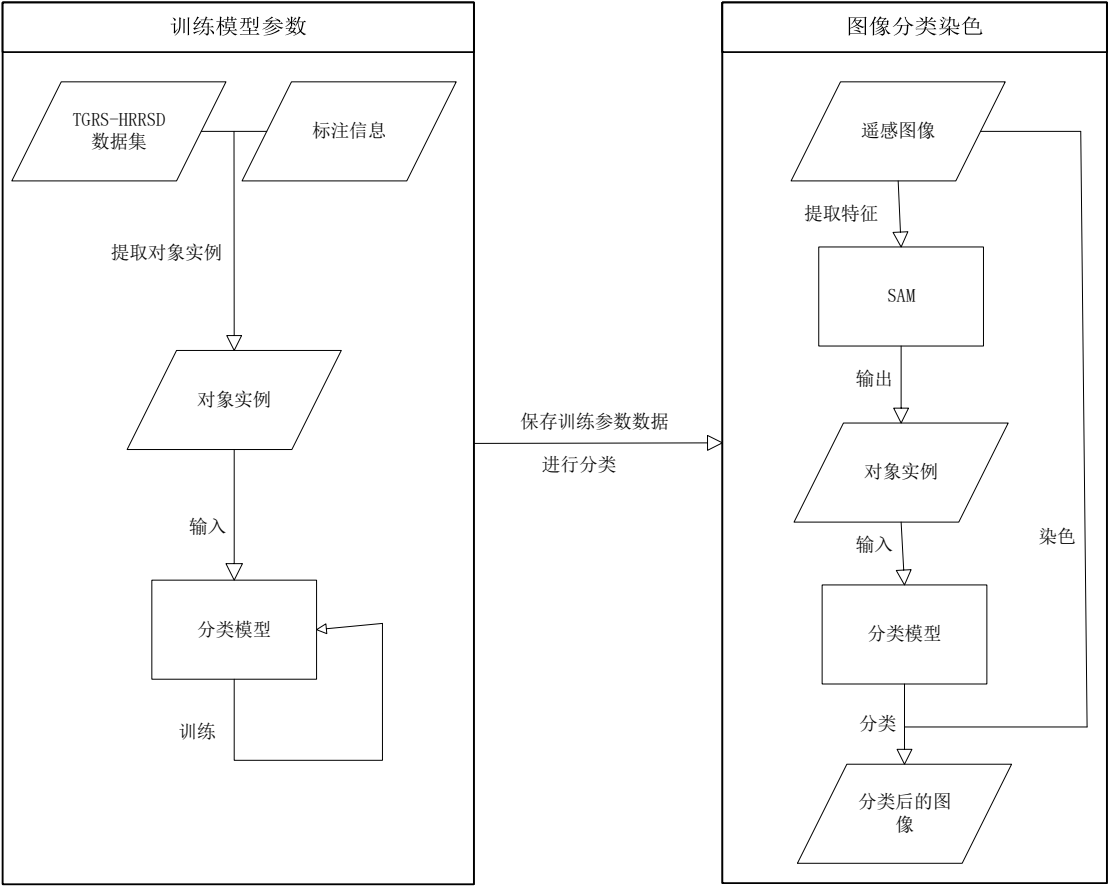


图 2 实验 1 流程

本文将 TGRS-HRRSD 数据集中标注的对象实例作为输入对 MobileViT 模型进行训练。训练完成后将两个模型组合起来,令遥感图像通过 SAM 模型生成对象实例,然后使用 MobileViT 模型对这些对象实例进行分类染色,实现对原图像进行目标检测的目的。在实验过程中,本文只需训练分类模型。使用已训练过的参数,通过迁移学习的方式训练模型,可提高训练效率,并减少需要使用的标注数据量。

在实验 2 中,本文将遥感图像通过 SAM 模型生成对象实例,如图 3 所示。由于得到的对象实例数量过少,因此需要对这些对象实例进行数据增强操作,再将这些对象实例与 CLRS 数据集生成混合数据集,并与未混合过的原 CLRS 数据集进行对比实验。需要注意的是,SAM 模型提取的不只是对象实例,可能还会得到一些干扰对象,所以在混合数据集前需要将这些干扰对象排除。

3.1.3 基准模型

本文使用目标检测模型作为基准模型,选取 YOLO-v2、Fast R-CNN 和 Faster R-CNN 等传统目标检测模型作为对比基准,分别计算每个模型对不同类别的分类正确率以及平均准确率。

3.2 性能比较

在实验 1 中,本文从 TGRS-HRRSD 数据集的 21 761 张

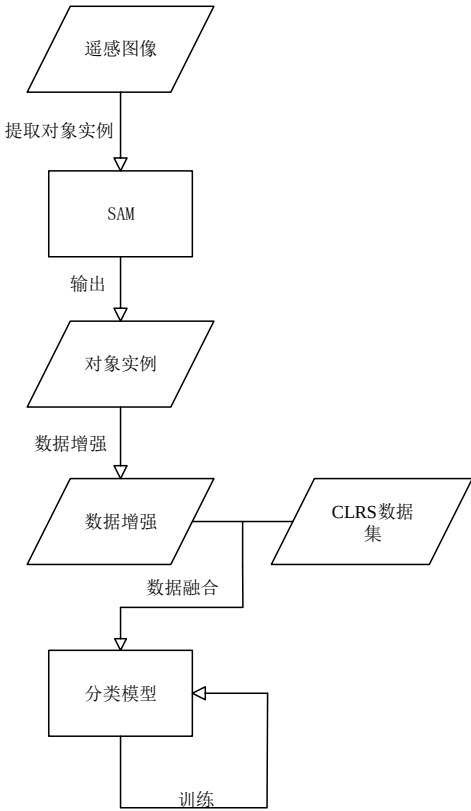


图 3 实验 2 流程

图像中提取 50 004 个实例对象作为训练数据训练分类模型,得到结果如表 1 所示。组合模型的平均准确率为 88.7%,而且对其中 10 个类的分类中,有 6 个类的准确率是最高的。实验结果表明,本文所提出的组合模型应用于遥感图像检测任务时,在准确率方面高于目标检测模型。相

较于 YOLO-v2、Fast R-CNN 和 Faster R-CNN 等传统目标检测方法,本文方法仅需训练分类模型,无需对复杂的目标检测模型进行完整训练,不仅大幅降低了模型训练难度,而且有效减少了数据标注的工作量。同时,在训练速度方面也优于目标检测方法。

Table 1 Model performance on the TGRS-HRRSD dataset

表 1 TGRS-HRRSD 数据集上的模型性能

(%)

| 方法           | data       | bridge | crossroad | ship | airplane | car  | house | forest | playground | land | river | mAP  |
|--------------|------------|--------|-----------|------|----------|------|-------|--------|------------|------|-------|------|
| YOLO-v2      | TGRS-HRRSD | 79.0   | 43.4      | 78.5 | 84.6     | 74.4 | 65.1  | 64.2   | 94.4       | 62.2 | 71.1  | 71.7 |
| Fast R-CNN   |            | 75.1   | 67.1      | 75.0 | 83.3     | 76.0 | 66.5  | 66.9   | 90.0       | 82.6 | 73.6  | 75.6 |
| Faster R-CNN |            | 85.5   | 88.6      | 88.5 | 90.8     | 78.4 | 81.5  | 79.6   | 90.6       | 86.9 | 82.7  | 85.3 |
| 本文方法         |            | 94.2   | 94.5      | 85.8 | 91.1     | 95.3 | 80.4  | 73.8   | 91.9       | 95.0 | 85.1  | 88.7 |

将组合模型的参数量、计算复杂度以及训练时间与 YOLO-v2、Faster R-CNN 模型进行比较,结果如表 2 所示。其中,GFLOPs 表示每秒 10 亿次的浮点数运算。由表 2 可知,本文方法的训练时间最少、计算复杂度最低,参数只有 28.3 M。

Table 2 Parameters, average training time and GFLOPs of different models

表 2 不同模型参数、平均训练时间与 GFLOPs

| 模型           | 参数/M | 平均训练时间/s | GFLOPs |
|--------------|------|----------|--------|
| YOLO-v2      | 50.1 | 78.4     | 39.5   |
| Faster R-CNN | 60.2 | 152.3    | 90.3   |
| 本文方法         | 28.3 | 32.5     | 10.7   |

Table 3 Comparison of dataset impact on model performance

表 3 数据集对模型性能的影响比较

(%)

| 方法        | 数据            | bridge | forest | river | storage-tank | commercial | playground | airplane | car  | mAP  |
|-----------|---------------|--------|--------|-------|--------------|------------|------------|----------|------|------|
| MobileViT | CLRS          | 87.2   | 78.6   | 79.1  | 80.0         | 77.4       | 84.3       | 89.1     | 90.7 | 83.3 |
| MobileViT | CLRS+SAM 分割实例 | 88.0   | 76.2   | 81.9  | 84.2         | 77.6       | 79.7       | 92.3     | 94.1 | 84.3 |

### 3.3 消融实验

为了评估 SAM 和 MobileViT 的独立贡献,实验分别测试了只使用单一模型的效果。仅使用 SAM 模型虽然在分割任务中表现优异,但其缺乏对目标类别的准确判断,所以并不对只使用 SAM 模型的方法进行测试。

仅使用 MobileViT 模型的测试方案为:利用目标检测数据集中的标注信息提取图像中的实例对象,并开展测试。实验结果显示,该方法在分类任务中可达到较高精

度,但在目标检测场景下的性能表现欠佳。这一结果表明,两种模型的结合能够有效弥补各自的不足,实现更好的综合性能。

### 3.4 实验结果分析

图 4—图 6 为训练模型的效果展示。其中,图 4 是原图,图 5 是使用 SAM 提取的实例对象,染色部分表示该实例被 SAM 识别并提取,而图 6 表示实例对象经过本文模型分类后的目标检测结果。



Fig. 4 Original images

图 4 原图像

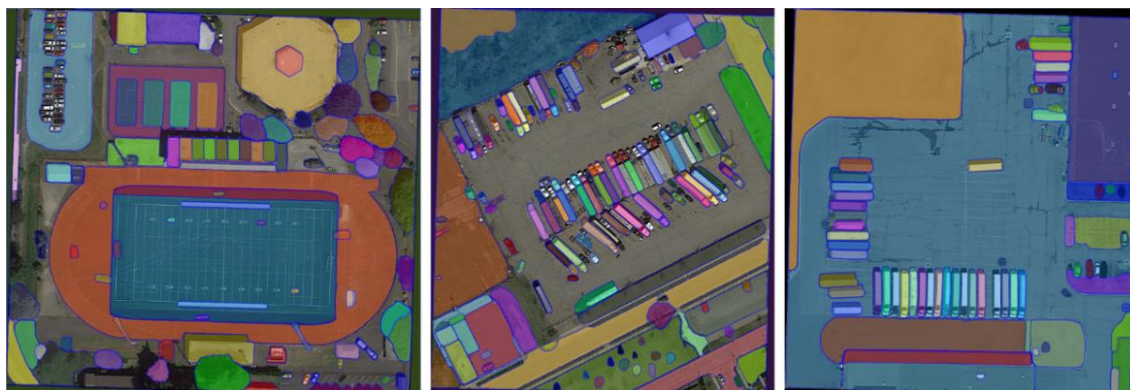
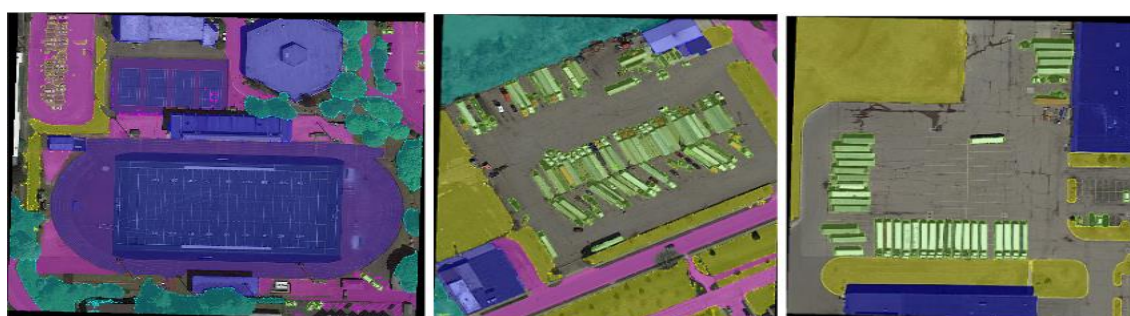


Fig. 5 Instance objects extracted by SAM

图5 SAM提取的实例对象



注：红色表示桥梁；绿色表示车辆；蓝色表示建筑；黄色表示土地；品红表示道路；青色表示森林

Fig. 6 Object detection results

图6 目标检测结果

通过对比图4和图5发现,SAM模型对遥感图像中的大部分实例对象进行了提取,虽然仍存在一些实例对象未被识别,但绝大多数实例都被正确提取。而观察图6可以发现,本文模型可以比较准确地对SAM模型提取的图片实例进行分类,最终达到目标检测的目的。

## 4 结语

本文提出结合SAM分割与MobileViT分类的遥感图像目标检测方法,通过SAM模型精确分割目标区域,并结合MobileViT进行高效的分类,成功提高了目标检测精度。实验结果表明,组合模型在计算资源消耗和分类准确率方面表现更优。

未来的研究将在以下几个方向进行深入探索:首先,通过SAM分割技术生成更多遥感图像类别实例,进一步改善遥感图像标注难度大、数据不足的问题,提升模型对多样化数据的适应性。其次,可尝试将SAM模型以及MobileViT模型替换为SAM模型的改进模型<sup>[20-21]</sup>或其他分类模型,寻找最有效率的组合方案。此外,探索更先进的深度学习架构,如结合自监督学习和元学习等新兴技术,以提升模型在小样本学习和实时处理任务中的表现,也将是未来研究的重要方向。

## 参考文献:

- [1] XIA J S, CHANUSSOT J, DU P J, et al. Rotation-based support vector machine ensemble in classification of hyperspectral data with limited training samples [J]. IEEE Transactions on Geoscience & Remote Sensing, 2016, 54(3):1519-1531.
- [2] CHEN Y S, ZHAO X, LIN Z H. Optimizing subspace SVM ensemble for hyperspectral imagery classification [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2014, 7(4): 1295-1305.
- [3] HU W, HUANG Y Y, WEI L, et al. Deep convolutional neural networks for hyperspectral image classification [J]. Journal of Sensors, 2015 (1): 258619.
- [4] MA X R, GENG J, WANG H Y. Hyperspectral image classification via contextual deep learning [EB/OL]. <https://link.springer.com/article/10.1186/s13640-015-0071-8>.
- [5] CHEN Y S, LIN Z H, ZHAO X, et al. Deep learning-based classification of hyperspectral data [J]. IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing, 2014, 7(6):2094-2107.
- [6] XU F L, ZHANG G, SONG C, et al. Multiscale and cross-level attention learning for hyperspectral image classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5501615.
- [7] XU Q S, YUAN X, OUYANG C J, et al. Attention-based pyramid network for segmentation and classification of high-resolution and hyperspectral remote sensing images [J]. Remote Sensing, 2020, 12(21): 3501.
- [8] ZHANG L, ZHANG Y S, YU Y, et al. Survey on object detection in tilt-

- ing box for remote sensing images[J]. National Remote Sensing Bulletin, 2023, 26(9): 1723-1743.
- [9] QIAN X L, LIN S, CHENG G, et al. Object detection in remote sensing images based on improved bounding box regression and multi-level features fusion[J]. Remote Sensing, 2020, 12(1): 143.
- [10] ZAND M, ETEMAD A, GREENSPAN M. Oriented bounding boxes for small and freely rotated objects[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 4701715.
- [11] LI X Y, LONG S P, ZHU J. Survey of few-shot learning based on deep neural network[J]. Application Research of Computers, 2020, 37(8): 2241-2247.
- 李新叶,龙慎鹏,朱婧. 基于深度神经网络的少样本学习综述[J]. 计算机应用研究, 2020, 37(8):2241-2247.
- [12] ZHAO Y Q, RAO Y, DONG S P, et al. Survey on deep learning object detection[J]. Journal of Image and Graphics, 2020, 25(4): 629-654.
- 赵永强,饶元,董世鹏,等. 深度学习目标检测方法综述[J]. 中国图象图形学报, 2020, 25(4):629-654.
- [13] ZHANG R R, JIANG Z K, GUO Z Y, et al. Personalize segment anything model with one shot[DB/OL]. <https://arxiv.org/pdf/2305.03048>.
- [14] SONG Y K, WU Y X, YE Y Y, et al. Few-shot object detection on remote sensing images based on feature weighting and fusion[J]. Software Guide, 2024, 23(4): 150-156.
- 宋云凯,吴原项,叶蕴瑶,等. 基于特征加权与融合的小样本遥感目标检测[J]. 软件导刊, 2024, 23(4):150-156.
- [15] ZENG B X, DONG C X. Water body detection based on multi-scale feature fusion for remote sensing images[J]. Software Guide, 2024, 23(5): 168-174.
- 曾宝秀,董传祥. 基于多尺度特征融合的遥感影像水体检测[J]. 软件导刊, 2024, 23(5):168-174.
- [16] LYU W Q, MA X, JIAN Y M, et al. Research on lightweight remote sensing image semantic segmentation method based on multilayer perceptron[J]. Software Guide, 2024, 23(1): 173-181.
- 吕文琪,马骁,简夜明,等. 基于多层感知机的轻量级遥感影像语义分割方法研究[J]. 软件导刊, 2024, 23(1):173-181.
- [17] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 4015-4026.
- [18] MEHTA S, RASTEGARI M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer[DB/OL]. <https://arxiv.org/pdf/2110.02178>.
- [19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6000-6010.
- [20] MA Z H, HONG X P, SHANGGUAN Q N. Can SAM count anything? an empirical study on sam counting[DB/OL]. <https://arxiv.org/pdf/2304.10817>.
- [21] CHEN T R, ZHU L Y, DING C T, et al. SAM fails to segment anything? --SAM-adapter: adapting SAM in underperformed scenes: camouflage, shadow, medical image segmentation, and more[DB/OL]. <https://arxiv.org/abs/2304.09148>.

(责任编辑:黄健)