

视觉—语言协同微调 LLaVA 的遥感图像目标检测方法

张珈懿¹, 陈 琦¹, 屈 丹^{1,2}, 司念文¹, 于佳珺¹

(1. 信息工程大学, 河南 郑州 450001; 2. 先进计算与智能工程(国家级)实验室, 江苏 无锡 214083)

摘要: 针对遥感图像目标检测依赖大量的标注数据, 且现有通用参数高效微调技术在适配遥感任务时空感知能力不足的问题, 提出一种面向多模态大模型的视觉—语言协同高效微调方法。首先, 对现有数据集进行重构, 构建适用大型语言和视觉助手(LLaVA)微调的指令数据; 其次, 结合多认知视觉适配器(Mona)和量化低秩适配(QLoRA)方法, 在指令微调数据集上对 LLaVA 模型进行高效微调; 最后, 在推理阶段获取目标类别及坐标信息。在 NWPU VHR-10 数据集上的实验表明, 相较于预训练 LLaVA 模型, 协同微调模型在 $F1$ 值、准确率和 mIoU 上均有较大提高。在 NWPU VHR-10 和 DIOR 两个数据集上, 其 mAP 达到与传统检测模型可比的结果。该方法在低标注场景下仍表现优异, 为多模态大模型在遥感视觉任务中的高效应用提供新的技术途径。

关键词: 遥感图像; 目标检测; LLaVA 模型; 多认知视觉适配器; 指令数据集; 量化低秩适配

中图分类号: TP751

文献标识码: A

文章编号: 1671-0673(2026)03-0371-08

Remote Sensing Image Object Detection Method Based on Vision-Language Collaborative Fine-Tuning LLaVA

ZHANG Jiayi¹, CHEN Qi¹, QU Dan^{1,2}, SI Nianwen¹, YU Jiajun¹

(1. Information Engineering University, Zhengzhou 450001, China;

2. Laboratory for Advanced Computing and Intelligence Engineering, Wuxi 214083, China)

Abstract: To address the challenge that remote sensing image object detection heavily depends on extensive annotated data and that existing efficient fine-tuning methods for general parameters exhibit insufficient spatial perception when adapted to remote sensing tasks, a vision-language collaborative efficient fine-tuning method for multimodal large models is proposed. Firstly, the existing dataset is reconstructed to construct instruction data suitable for large language and vision assistant (LLaVA) fine-tuning. Secondly, combining the multi-cognitive visual adapter (Mona) and the quantized low-rank adaptation (QLoRA) method, the LLaVA model is efficiently fine-tuned on the instruction fine-tuning dataset. Finally, the target category and coordinate information are obtained in the inference stage. Experiments on the NWPU VHR-10 dataset show that compared with the pre-trained LLaVA model, the collaborative fine-tuning model has greatly improved in $F1$ score, accuracy, and mIoU. Its mAPs on both NWPU VHR-10 and DIOR datasets are comparable to those of traditional detection models. This method still performs well in low-annotation scenarios, providing a new technical approach for the efficient application of multimodal large models in remote sensing vision tasks.

Key words: remote sensing image; object detection; LLaVA model; Mona; instruction dataset; QLoRA

遥感图像目标检测是指从遥感图像中自动识别与定位特定目标,是地理信息科学和计算机视觉交叉领域的核心技术,在军事侦察、灾害评估和环境监测等领域具有重要的应用价值。

传统目标检测主要采用机器学习中的方向梯度直方图(Histogram of Oriented Gradient, HOG)^[1]、自适应提升(Adaptive Boosting, AdaBoost)^[2]算法框架和支持向量机(Support Vector Machine, SVM)^[3]等方法^[4]。当前,基于卷积神经网络(Convolutional Neural Network, CNN)的目标检测算法成为主流,其中主要分为以 R-CNN(Region-Based CNN)系列为代表的两阶段目标检测算法和以 SSD(Single Shot MultiBox Detector)^[5]、YOLO(You Only Look Once)^[6-7]系列为代表的一阶段目标检测算法。两阶段检测算法先通过区域生成网络(Region Proposal Network, RPN)产生大量候选区域,再将候选区域送入神经网络预测其目标的位置信息和类别概率,其目标检测的精确率较高,但检测速率较慢。一阶段目标检测算法不使用区域生成网络产生候选区域,直接在卷积神经网络中提取特征来预测物体分类和位置信息,其检测速率较快但精确率通常不如两阶段算法。随着深度学习技术的进步,基于 Transformer 架构的模型系列^[8]开始崭露头角,其利用 Transformer^[9]的强大序列建模能力,通过自注意力机制处理图像特征,理论上能够更好地捕捉全局上下文信息。研究人员通过增强骨干网络的特征提取能力^[10]、引入注意力机制等^[11]设计大量针对遥感图像目标检测任务的改进方法。

近年来,以 GPT-4 和大型语言和视觉助手(Large Language and Vision Assistant, LLaVA)^[12]为代表的多模态大语言模型(Large Language Model, LLM)在视觉—语言联合理解与推理方面展现出强大能力。这类模型通过在超大规模图像—文本对上进行预训练,构建通用的视觉概念与语义知识库,并展现出强大的零样本迁移能力。这为破解遥感目标检测的数据瓶颈提供了全新的技术路径。然而,如何将此类通用大模型高效适配至遥感视觉任务,仍面临核心挑战。现有的参数高效微调技术主要针对语言模型设计,通过低秩矩阵模拟参数更新^[13-15],或者固定预训练主干网络的原始参数,并添加额外可训练结构^[16-19],虽然能高效调整语义理解能力,却难以优化模型对图像空间结构、多尺度目标及复杂背景的理解能力,导致其直接应用于遥感图像目标检测时精度有限。因此,探

索一种能够同时增强视觉空间感知与语言语义理解的高效微调框架,成为推动多模态大模型在遥感领域落地的关键。

针对遥感图像目标检测依赖标注数据的难题,提出一种视觉—语言协同的高效微调方法。通过构建指令数据集,将检测任务转化为多轮对话学习,并提出 Mona-QLoRA 协同微调框架,其中多认知视觉适配器(Multi-cognitive Visual Adapter, Mona)^[20]增强视觉空间特征感知,量化低秩适配(Quantized Low-Rank Adaptation, QLoRA)^[21]优化语言理解能力,二者协同实现 LLaVA 模型的精准适配,并在推理阶段引入了后处理流程。实验结果表明,该方法实现了高精度目标检测。

1 基于微调 LLaVA 的遥感图像目标检测算法

1.1 视觉—语言微调数据集构建

为实现对多模态大模型 LLaVA 的有效微调,需将其训练数据格式与遥感目标检测任务要求对齐。本文基于 NWPU VHR-10 标准遥感检测数据集,提出一种结构化的指令数据构建方法。

NWPU VHR-10 数据集包含 800 幅图像,其中:正例图像 650 幅(包含目标);反例图像 150 幅(不包含目标)。首先,对数据集进行预处理,提取标注文件中的目标像素坐标与类别编号,通过图像宽高归一化处理,将像素坐标转换为 $[x_1, y_1, x_2, y_2]$ 格式,以图像左上角为原点, (x_1, y_1) 为目标边界框左上角坐标, (x_2, y_2) 为右下角坐标,坐标值为占图像长宽的比例。

坐标归一化具体计算流程为

$$x'_1 = \frac{x_1}{W}, y'_1 = \frac{y_1}{H}, x'_2 = \frac{x_2}{W}, y'_2 = \frac{y_2}{H} \quad (1)$$

式中: x_1, y_1, x_2, y_2 为标注文件中的原始像素坐标; W, H 分别为遥感图像的像素宽度和高度; x'_1, y'_1, x'_2, y'_2 为归一化后的相对坐标,取值范围均为 $(0, 1]$,确保数据格式统一。目标的类别编号与类别对应关系如表 1 所示。

表 1 目标的类别编号与类别对应关系

编号	类别	编号	类别
1	Airplanes	6	Basketball Courts
2	Ships	7	Ground Track Fields
3	Storage Tanks	8	Harbors
4	Baseball Diamonds	9	Bridges
5	Tennis Courts	10	Vehicles

为降低标注噪声对指令微调的影响,本文对标注数据进行了质量检查与清洗,主要包括:对存在越界的标注框(如 $x < 0$ 或 $x > W$)进行裁剪,使其落入有效图像范围;若出现 $x_1 > x_2$ 或 $y_1 > y_2$,则交换对应坐标,避免后续计算异常;过滤面积过小的框,以剔除可能的误标注或无效框。

传统LLaVA模型的训练数据是在设计问题模板后,通过提示GPT回答问题作为答案输出。由于数据集中图像的目标类别和坐标位置已经在txt文件中给出,故本文不再采用GPT的回答作为输出,直接将txt文件中所给的类别标识与边界框坐标进行整理,作为问题的答案输出。参考LLaVA模型微调阶段的多轮问答数据规范,设计3轮结构化提问:首先,判断图像中是否存在指定10类目标,仅输出“是”或“否”;其次,识别目标具体类别;最后,输出目标归一化坐标。最终构建的指令数据集以JSON文件存储,包含训练集650幅图像의id、路径及对话信息。指令数据集部分示例如图1所示。

```

{"id": "001",
 "image": "train001.jpg",
 "conversations": [
   {
     "from": "human",
     "value": "<image>\n It is now known that the ten type of targets are airplanes, ships, storage tanks, baseball diamonds, tennis courts, basketball courts, ground track fields, harbors, bridges, vehicles. Is there any of the above ten types of targets in this image? Only answer yes or no."
   },
   {
     "from": "gpt",
     "value": "Yes"
   },
   {
     "from": "human",
     "value": "If there are the above targets in the image, what are the types of these targets?"
   },
   {
     "from": "gpt",
     "value": "[ 'airplane ']"
   },
   {
     "from": "human",
     "value": "If there are the above targets in the image, use the top left corner of the image as the coordinate origin, provide the coordinates of the targets."
   },
   {
     "from": "gpt",
     "value": "[ [0.59,0.59,0.66,0.71]"
   }
 ]
}

```

图1 指令数据集部分示例

1.2 LLaVA模型微调框架

该方法基于多模态大模型LLaVA架构,通过构建指令数据集用于Mona-QLoRA协同微调,实现

端到端的遥感目标检测。与传统检测方法不同,本文方法不依赖显式的检测头,而是利用LLM的序列生成能力直接输出结构化检测结果。其结构如图2所示。视觉侧在CLIP-ViT特征提取后插入Mona以增强多尺度空间感知,语言侧在LLM线性层引入QLoRA以实现低秩、低显存微调,模型以序列生成方式输出目标类别与归一化坐标。

1.2.1 LLaVA模型

LLaVA是一种端到端训练的大型多模态模型,能够融合视觉和语言信息,提高理解和回应用户需求的能力,特别是在处理包含图像、视频等多模态输入的任务上。

LLaVA主要采用级联式的方法构建,由3个部分组成:预训练的大语言模型(Pretrained LLM)、预训练的视觉编码器和投影层。预训练的视觉编码器对输入图像进行编码,通常是CLIP中的视觉分支。连接在视觉编码器和语言模型之间的投影层为一个简单的线性层,作用是对图像编码得到的结果做线性映射。

1.2.2 微调方法

通过微调LLaVA模型,可以让模型学会专注于遥感图像中的特定目标和环境特征,从而提高目标检测的准确性。考虑计算资源、推理速度以及存储需求的限制,本文方法采用轻量化微调的方式对基于LLaVA的遥感图像目标检测系统进行改进。

传统的微调方法通常使用线性滤波器来处理输入特征^[22]。然而,视觉信号与语言信号在处理方式上存在很大差异。视觉任务需要处理的是图像的二维空间信息,线性滤波器不能很好地适应视觉信号的特性。Mona采用卷积滤波器,卷积层能够更有效地捕捉图像中的局部特征,提升特征的提取和传递效果。为最大限度减少新增参数量,Mona采用深度卷积(Depthwise Convolution, DW-Conv),上层特征在下投影(Down-Projection)后会经过 3×3 、 5×5 、 7×7 这3种深度可分卷积核形成多尺度滤波器组,以提供不同尺度的感受野。计算3个滤波器输出结果的平均值,再通过 1×1 卷积聚合特征。同时,在两种类型的卷积中添加跳跃连接。3个深度卷积权重为 $w_{DW}^i \in \mathbb{R}^{C_{in}^{DW} \times K_i \times K_i \times C_{out}^{DW}}$ ($i \in 1, 2, 3$),其中: C_{in}^{DW} 表示深度卷积后的输入通道数; K_i 表示第 i 个卷积核的尺寸; C_{out}^{DW} 表示输出通道数。此外,还使用1个点卷积(Pointwise Convolution, PWConv),权重为 $w_{PW} \in \mathbb{R}^{C_{in}^{PW} \times 1 \times 1 \times C_{out}^{PW}}$ 。

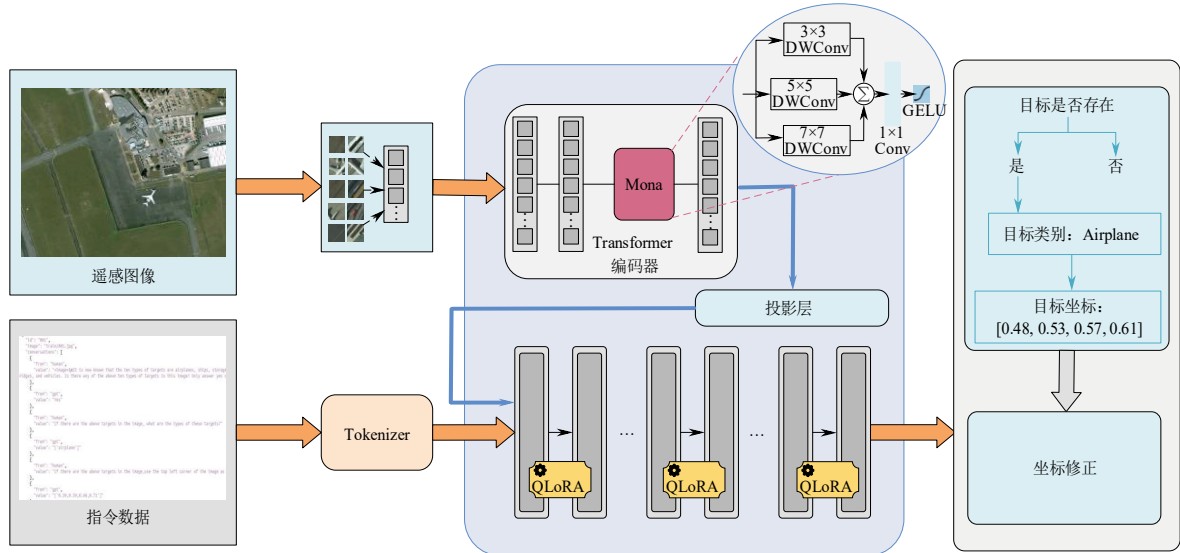


图2 Mona-QLoRA协同微调的整体框架与流程

上述两个卷积步骤可表示为如下形式:

$$f_{\text{DW}} = x + \text{avg} \left(\sum_{i=1}^3 w_{\text{DW}}^i \hat{\otimes} x \right) \quad (2)$$

$$f_{\text{PW}} = x + w_{\text{PW}} \bar{\otimes} x \quad (3)$$

式中, $\hat{\otimes}$ 和 $\bar{\otimes}$ 分别表示深度卷积和点卷积。

特征通过高斯误差线性单元 (Gaussian Error Linear Unit, GELU) 激活函数进行非线性变换, 并通过上投影 (Up-Projection) 恢复维度。Mona 的整体计算过程可表述如下:

$$x = x_0 + U^l \sigma \left(f_{\text{PW}} \left(f_{\text{DW}} \left(D^l(x_{\text{norm}}) \right) \right) \right) \quad (4)$$

式中: x_0 表示 Mona 的原始输入; D^l 和 U^l 分别表示第 l 个适配器的下投影和上投影; $\sigma(\cdot)$ 表示 GELU 激活函数。为使 Mona 能够调整输入分布以及来自固定层的输入占比, 在 Mona 的顶端添加 1 个归一化层和两个可学习权重, 用于调整输入分布。归一化操作有助于稳定前向传播的输入分布和反向传播的梯度。在 Mona 中采用层归一化, 可表示为

$$x_{\text{norm}} = s_1 \cdot \text{LayerNorm}(x_0) + s_2 \cdot x_0 \quad (5)$$

式中, s_1 和 s_2 表示可学习的权重。

将 Mona 分别插入 LLaVA 视觉编码器 CLIP-ViT 每个 Transformer 层的多头注意力和多层感知机输出, 确保多尺度空间信息被充分捕捉, 具体位置如图 3 所示。左侧为 Mona 微调, 每个 ViT 的多头注意力和多层感知机之后添加 Mona 模块。在训练时固定预训练层参数, 同时动态更新 Mona 参数。右侧为 Mona 模块细节, 其中 A 表示取平均。Mona 在下投影前采用缩放后的层归一化, 下投影后配置多认知卷积滤波器组和聚合滤波器。在 Mona 内部 4 个位置添加跳跃连接以增强自适应能力。

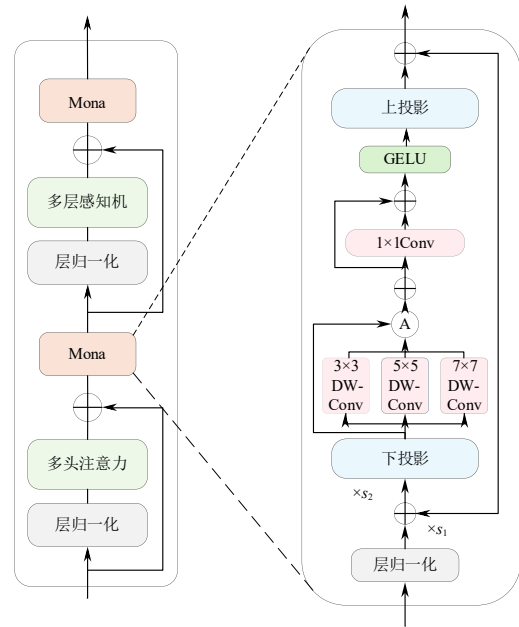


图3 Mona示意图

LoRA^[23] 是一种低资源微调大模型方法, 其核心思想是通过低秩分解来模拟参数的改变量, 从而以极小参数量实现大模型的间接训练。QLoRA 的核心是在 LoRA 基础上引入 4 位量化 (4-bit Quantization), 解决 LLM 微调时内存占用过高的问题, 同时保持模型性能接近全参数微调。

在 LLaVA 模型中, 视觉编码器负责提取图像的通用视觉特征, 其参数通常在微调中被冻结以保留预训练的视觉知识。多模态投影层作为连接视觉与语言模态的桥梁, 其结构较为简单, 对其进行低秩适配的收益有限。因此, 选择将 QLoRA 适配器主要应用于 LLM 的内部, 从而将模型有效地应用于遥感目标检测任务。

1.3 后处理流程

3轮提问输出的答案分别对应于目标是否存在、目标类别以及目标坐标的检测结果。生成式多模态大模型通过语言建模的方式输出坐标文本序列,而非传统检测网络中专门的坐标回归头,这种文本生成式输出本质是基于概率的逐字符预测,不具备天然的几何约束,因此在输出过程中可能出现坐标值超出归一化范围、坐标顺序颠倒等非规范情况。

为此,在推理阶段引入后处理流程,以提升检测的稳定性。具体而言,首先,对预测坐标进行合法性修正,将坐标裁剪至有效范围 $(0,1]$,并对不满足顺序约束的坐标对进行交换,以保证预测框满足 $x_1 < x_2$ 且 $y_1 < y_2$ 的基本条件。其次,对面积小于0.000 1的异常预测框进行过滤,以减少噪声输出的影响,获得最终检测结果,使模型输出更加符合目标检测任务的要求。

2 实验结果与分析

2.1 实验设置与评价指标

硬件平台配置在单张48 GB显存的NVIDIA GPU,运行在Ubuntu 20.04操作系统上。PyTorch版本为2.1.0,CUDA版本为12.1。

LLaVA预训练模型为llava-v1.5-7b,ViT模型为clip-vit-large-patch14-336。启用QLoRA适配器进行高效微调,QLoRA的设置与主流参数高效微调一致,秩为16,缩放因子 α 设为32,权重衰减正则化项的系数为0,梯度每累计8步更新1次权重;学习率为 3×10^{-5} ,学习率预热比例为0.03,同时使用余弦退火学习率调整策略。多模态投影器的学习率设为 2×10^{-5} ,比LLaVA稍低,多模态特征投影器的类型为带有GELU激活函数由两层组成的多层感知机,选择ViT模型倒数第2层的输出。总训练轮次为10轮,训练的批量大小为4。Mona的中间维度 n 设为64,学习率配置为 1×10^{-4} ,适配器通常用稍高学习率。

实验结果采用目标检测常用的评价指标:精确率(Precision)、召回率(Recall)、 $F1$ 值($F1$ Score)、准确率(Accuracy)、平均交并比(mean Intersection over Union, mIoU)和平均精度均值(mean Average Precision, mAP)。针对算法效率,采用训练参数量、训练参数占比、训练峰值显存和训练耗时作为评价指标。

2.2 实验分析

2.2.1 对比实验

按8:2比例划分为训练集640幅与测试集160幅,循环结束后,将检测结果与标准答案进行对照,分别计算3次提问得到的检测结果($F1$ 值、准确率和mIoU)。针对检测目标是否存在的问题,采用 $F1$ 值作为评价指标进行评价。针对检测目标类别的问题,采用准确率作为评价指标进行评价。对于检测出的目标,根据公式计算得到mIoU。为验证所提方法在遥感目标检测中的有效性,本实验与其他微调策略进行对比,其中:针对视觉部分进行Adapter微调,即在每个ViT的多头注意力和多层感知机之后添加标准Adapter;语言部分进行LoRA以及QLoRA微调。实验结果如表2所示。

表2 不同微调方式在NWPU VHR-10数据集上的性能对比

微调方式	$F1$ 值	准确率/%	mIoU/%
预训练LLaVA模型	0.861 9	47.40	32.58
Adapter微调(视觉)	0.929 7	74.46	57.38
LoRA微调(语言)	0.937 4	76.54	72.59
Adapter-LoRA协同微调	0.945 7	84.94	82.76
Mona-LoRA协同微调	0.949 5	85.00	83.33
Mona-QLoRA协同微调(本文)	0.959 0	85.47	85.28

视觉侧采用标准Adapter微调时,3项指标均为所有微调对比方案中最低,因为传统Adapter采用线性适配机制,未能针对遥感图像的二维空间特征进行优化。语言侧采用LoRA微调时,性能较标准Adapter微调有明显提升,由于未对视觉特征提取环节进行针对性适配,遥感图像中复杂背景与多尺度目标的特征感知不足问题仍未解决,导致其性能与协同微调方案存在差距。

Adapter-LoRA、Mona-LoRA协同微调的性能均显著优于单一微调方法,当视觉侧将标准Adapter替换为本文设计的Mona适配器后,性能进一步优化,这一结果验证了Mona适配器的设计优越性。本文提出的Mona-QLoRA协同微调算法在所有对比策略中表现最佳, $F1$ 值、准确率和mIoU分别达到0.959 0、85.47%和85.28%。QLoRA的低秩适配与量化技术结合,通过精细化的参数更新的方式优化语言模型处理多模态信息的关键路径,使得视觉特征与语言语义的对齐更精准,进而降低类别误判与坐标漂移问题,实现了检测性能的提升。

为验证所提方法的轻量化,从可训练参数规模与训练资源开销两个维度进行评估,实验结果如表3所示。

表3 不同微调方式的效率对比

微调方式	可训练参数量/ 10^6	可训练参数占比/%	训练峰值显存/GB	训练耗时/h
Adapter 微调(视觉)	12.38	0.16	28.50	1.50
LoRA 微调(语言)	30.04	0.39	33.80	1.70
Adapter-LoRA 协同微调	43.44	0.57	38.60	2.05
Mona-LoRA 协同微调	48.38	0.63	39.20	2.15
Mona-QLoRA 协同微调(本文)	48.20	0.63	27.40	2.25

Mona-LoRA 与 Mona-QLoRA 参数量相近, Mona 适配器为强化遥感空间特征捕捉增加少量适配参数, 且 Mona-QLoRA 通过量化实现参数微幅优化, 保持轻量化特性, 训练峰值显存为所有方式最低。Mona-QLoRA 协同微调算法实现轻量化、低显存、高性能的协同优化, 参数量与占比保持低位, 显存为全方案最低。本文方法训练耗时略长于其他方法, 这是由于引入多尺度卷积增加了前向计算量, 但换来了最优的检测性能, 在显存效率和检测精度之间取得最佳平衡, 为多模态大模型在遥感目标检测领域的工程化部署提供关键效率支撑。

采用 mAP 评价指标分别在 NWPU VHR-10 和 DIOR 两个数据集上进行实验, DIOR 数据集的预处理、坐标归一化、标注清洗等过程与 NWPU VHR-10 数据集一致。实验结果如表 4 所示。

表4 不同算法在两个数据集上的 mAP 单位: %			
模型类型	模型名称	mAP	
		NWPU VHR-10	DIOR
一阶段	YOLOv8 ^[24]	79.3	61.6
两阶段	Faster RCNN	80.2	62.7
Transformer	DETR	84.7	56.4
多模态大模型	Mona-QLoRA 协同微调(本文)	80.5	63.2

在 NWPU VHR-10 数据集中, 本文多模态大模型的 mAP 达到 80.5%, 与两阶段代表模型 Faster RCNN 精度基本持平, 略低于 Transformer 架构的 DETR。而在含 20 个类别、场景更复杂的 DIOR 数据集中, 该模型表现突出, mAP 达到 63.2%。这一结果表明, 传统检测模型在类别数量增加、背景复杂度提升时容易出现精度下降, 而本文方法通过视觉—语言协同微调框架, 借助 LLaVA 预训练模型的通用知识迁移能力与 Mona-QLoRA 的任务适配能力, 有效缓解了复杂场景下的特征学习不足问题, 展现出更强的环境适应性。而且本文方法同时保留自然语言交互能力, 具备传统方法所缺乏的交互式检测能力和多任务统一处理能力, 用户可通过自然语言指令灵活定制检测需求, 拓展

了遥感目标检测的应用场景。

2.2.2 低标注比例实验

为验证所提方法在低标注条件下的性能表现, 在 NWPU VHR-10 训练集中按比例随机抽取标注样本构建子集进行微调, 其余样本不参与训练。测试集保持不变, 每个标注比例随机采样 3 次取平均, 实验结果如表 5 所示。

表5 Mona-QLoRA 协同微调算法在不同标注比例下的性能对比

标注比例/%	F1 值	准确率/%	mIoU/%
10	0.842 7	75.23	70.90
30	0.929 7	80.22	78.28
50	0.937 4	83.36	82.33
100	0.959 0	85.47	85.28

10% 低标注下, 模型已实现较好检测效果, 预训练通用视觉—语言知识与任务适配性微调, 无需大量标注即可完成遥感目标的基础识别与定位; 标注比例提升至 30%, 模型 3 项指标大幅提升, 少量新增标注让模型学习遥感目标核心特征, 优化了特征语义对齐精度, 显著提升检测性能; 50% 标注比例下即可实现接近全标注的检测性能, 实现了高效的知识迁移, 减少了对大量标注数据的依赖。

2.2.3 消融实验

为研究 Mona 和 QLoRA 两种微调方式协同的必要性, 分别对 LLaVA 模型进行视觉和语言部分的微调。实验结果如表 6 所示。

表6 不同算法在 NWPU VHR-10 数据集上的性能对比

微调方式	F1 值	准确率/%	mIoU/%
预训练 LLaVA 模型	0.861 9	47.40	32.58
QLoRA 微调	0.942 3	84.66	81.78
Mona 微调	0.925 0	75.45	60.32
Mona-QLoRA 协同微调(本文)	0.959 0	85.47	85.28

微调显著提升预训练 LLaVA 模型在遥感目标检测任务上的性能。相较于预训练 LLaVA 模型, 3 种微调方法在 F1 值、准确率和 mIoU 上均取得显著

提升。其中:与预训练 LLaVA 模型相比,QLoRA 微调后 $F1$ 值从 0.861 9 提升至 0.942 3,大大降低了漏检率,准确率从 47.40% 提升至 84.66%,mIoU 从 32.58% 提升至 81.78%,提升幅度最为显著;Mona 微调在 3 个方面也均有较大提升,说明视觉侧适配器能够有效缓解冻结视觉编码器在遥感域上的特征不匹配,提升对多尺度目标的感知与定位质量;Mona-QLoRA 协同微调算法取得了最佳的综合性性能,在 $F1$ 值、准确率和 mIoU 上分别达到 0.959 0、85.47% 和 85.28%。与单一微调方法相比,协同微调在 mIoU 上表现最为突出,同时准确率、 $F1$ 值保持最优水平,表明边界框定位需要更精细的参数调整,协同微调通过多层次适配机制,在空间坐标回归任务上表现出优势。

为验证所提方法中 Mona 卷积核组合设计的有效性,在 NWPU VHR-10 数据集上分别对 3×3 、 5×5 、 7×7 单个卷积核,以及 $(3\times3+5\times5+7\times7)$ 多尺度卷积核组合进行实验。实验结果如表 7 所示。

表 7 不同卷积核在 NWPU VHR-10 数据集上的性能对比

不同卷积核	$F1$ 值	准确率/%	mIoU/%
3×3	0.942 0	82.15	79.87
5×5	0.945 2	82.62	80.66
7×7	0.932 0	80.40	77.23
$3\times3+5\times5+7\times7$	0.959 0	85.47	85.28

从表 7 可以看出,3 个尺度组合取得最佳结果,说明遥感目标检测需要同时兼顾细粒度纹理与大尺度上下文信息。多尺度组合通过跨尺度特征融合,提高了复杂背景下的目标可分性与定位稳定性,从而在 mIoU 上表现最为突出,验证了多尺度卷积核组合设计的合理性。

3 结束语

围绕遥感图像目标检测任务中标注数据稀缺、现有方法泛化能力有限且现有通用参数高效微调技术在遥感领域不适配的核心挑战,探索多模态大模型在该领域的应用。首先,构建了面向遥感检测的多模态指令数据集与评估体系。其次,提出了 Mona-QLoRA 协同高效微调框架,Mona 通过多尺度深度卷积结构,专门增强模型对遥感图像空间特征的感知能力,有效捕捉了多尺度目标与复杂上下文信息,将其与优化语言理解的 QLoRA 协同工作,形成了针对多模态大模型视觉与语言组件的协同微调策略。实验结果表明,相

较于单一的 QLoRA 或 Mona 微调,视觉与语言共同微调取得了更优的性能。

本文方法在遥感目标检测任务中取得较好效果,但仍存在若干局限性。本文方法依赖一定规模的目标框标注来构建指令数据集,属于监督微调范式,距离完全零标注以及开放词汇检测仍有差距;尽管采用 QLoRA 与 Mona 显著降低训练成本,但推理端仍受 7B 级多模态大模型规模限制,与专用检测网络相比部署成本仍偏高。未来工作将围绕零标注迁移以及模型轻量化部署展开。

4 参考文献:

- [1] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C]//Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2005:886-893.
- [2] FREUND Y, SCHAPIRE R E. Experiments with a new boosting algorithm [C]//Proceedings of the Thirteenth International Conference on International Conference on Machine Learning. New York, USA: ACM, 1996:148-156.
- [3] VAPNIK V N. The nature of statistical learning theory [M]. New York, USA: Springer, 1995:1-203.
- [4] 谢俊章,彭辉,唐健峰,等.改进 YOLOv4 的密集遥感目标检测[J].计算机工程与应用,2021,57(22):247-256.
- [5] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector [M]//LEIBE B, MATAS J, SEBE N, et al. Computer vision-ECCV 2016. Cham, Switzerland: Springer, 2016:21-37.
- [6] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2016:779-788.
- [7] 刘春友,唐志斌,刘智国,等.面向智能监控的轻量 YOLOv10 目标检测算法[J].信息工程大学学报,2025,26(4):444-449.
- [8] ZHAO Y A, LV W Y, XU S L, et al. DETRs beat YOLOs on real-time object detection [C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2024:16965-16974.
- [9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017,30:5998-6008.
- [10] XU D Q, WU Y Q. FE-YOLO: a feature enhancement network for remote sensing target detection [J]. Remote Sensing, 2021,13(7):No.1311.
- [11] LI Y S, LUO J W, ZHANG Y J, et al. Learning to holisti-

- cally detect bridges from large-size VHR remote sensing imagery [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(12):11507-11523.
- [12] LIU H T, LI C Y, WU Q Y, et al. Visual instruction tuning [J]. *Advances in Neural Information Processing Systems*, 2023, 36:34892-34916.
- [13] ZHONG H Q, SHEN S C, CAI K, et al. Serial low-rank adaptation of vision transformer [C] // *Proceedings of the 2025 IEEE International Conference on Multimedia and Expo*. Piscataway, USA: IEEE, 2025. DOI: 10.1109/ICME59968.2025.11209194.
- [14] LEI C, LI A, YAO H, et al. Rethinking token reduction with parameter-efficient fine-tuning in ViT for pixel-level tasks [C] // *Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway, USA: IEEE, 2025:14954-14964.
- [15] LIU S Y, WANG C Y, YIN H X, et al. DoRA: weight-decomposed low-rank adaptation [DB/OL]. (2024-07-09) [2026-01-25]. <https://arxiv.org/abs/2402.09353>.
- [16] ZHU B E, NIU Y L, HAN Y C, et al. Prompt-aligned gradient for prompt tuning [C] // *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision*. Piscataway, USA: IEEE, 2023:15613-15623.
- [17] HE X H, LI C Y, ZHANG P C, et al. Parameter-efficient model adaptation for vision transformers [DB/OL]. (2023-07-13) [2026-01-25]. <https://arxiv.org/abs/2203.16329>.
- [18] YIN D S, YANG Y R, WANG Z C, et al. 1% VS 100%: parameter-efficient low rank adapter for dense predictions [C] // *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway, USA: IEEE, 2023:20116-20126.
- [19] GONG M, DENG Y N, QIN A, et al. Structure-learnable adapter fine-tuning for parameter-efficient large language models [C] // *Proceedings of the International Conference on Artificial Intelligence and Computational Engineering 2025*. London, UK: IET, 2025:225-230.
- [20] YIN D S, HU L Y, LI B, et al. 5%>100%: breaking performance shackles of full fine-tuning on visual recognition tasks [C] // *Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway, USA: IEEE, 2025:20071-20081.
- [21] DETTMERS T, HOLTZMAN A, PAGNONI A, et al. QLoRA: efficient finetuning of quantized LLMs [J]. *Advances in Neural Information Processing Systems*, 2023, 36:10088-10115.
- [22] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP [C] // *Proceedings of the 36th International Conference on Machine Learning*. New York, USA: PMLR, 2019:2790-2799.
- [23] HU E J, SHEN Y L, WALLIS P, et al. LoRA: low-rank adaptation of large language models [DB/OL]. (2021-10-16) [2026-01-25]. <https://arxiv.org/abs/2106.09685>.
- [24] 席阳丽, 屈丹, 王芳芳, 等. 基于 FEW-YOLOv8 遥感图像目标检测算法 [J]. *郑州大学学报(工学版)*, 2025, 46(4):62-69.

(编辑: 李志豪)

(上接第 370 页)

- [18] DAW A, KARPATNE A, WATKINS W D, et al. Physics-guided neural networks (PGNN): an application in lake temperature modeling [M] // *Knowledge-guided machine learning*. Boca Raton, USA: Chapman and Hall/CRC, 2022:353-372.
- [19] WILLMOTT C J, MATSUURA K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance [J]. *Climate Research*, 2005, 30:79-82.
- [20] BISHOP C M, NASRABADI N M. *Pattern recognition and machine learning* [M]. New York, USA: Springer, 2006:192-195.
- [21] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 42(2):318-327.
- [22] CIPOLLA R, GAL Y, KENDALL A. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics [C] // *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway, USA: IEEE, 2018:7482-7491.
- [23] HYNDMAN R J, KOEHLER A B. Another look at measures of forecast accuracy [J]. *International Journal of Forecasting*, 2006, 22(4):679-688.
- [24] DAVIS J, GOADRICH M. The relationship between precision-recall and ROC curves [C] // *Proceedings of the 23rd International Conference on Machine Learning*. New York, USA: ACM, 2006:233-240.

(编辑: 李志豪)