基于共享提示与Mamba适配器的遥感图像文本检索方法

杜文亮[1,2],许晓宇[1,2],赵佳琦[1,2],刘 兵[1,2],周 勇[1,2*]
(1. 中国矿业大学计算机科学与技术学院/人工智能学院,江苏徐州 221116;2. 矿山数字化教育部工程研究中心,江苏徐州 221116)

摘 要: 遥感图像文本检索旨在根据给定的图像或文本,从海量遥感图像文本数据库中快速、准确地检索出与之语义匹配的文本或图像. 随着对地观测技术的飞速发展,该技术在城市规划、灾害应急响应、环境监测等领域的应用价值日益凸显,已成为当前多模态信息处理领域的研究热点. 基于通用数据预训练的视觉语言预训练模型,通过实现图像与文本之间的高效语义对齐,为通用图像文本检索任务奠定了技术基础. 然而,通用数据与遥感数据之间存在显著的领域鸿沟,导致基于通用数据预训练的视觉语言预训练模型在直接应用于遥感任务时性能受限. 因此,需要通过微调使该视觉语言模型适应遥感领域独特的数据分布. 然而,现有微调方法应用到遥感领域时面临着两大核心挑战. 其一,跨模态对齐不足:现有微调方法缺乏显式的跨模态信息交互机制,难以充分建模图文之间的内在关联;其二,细粒度语义表征困难:现有方法往往难以捕捉遥感图像中目标尺度差异悬殊、地物类别间相似度高、空间拓扑关系复杂等精细化的语义信息. 尤其在处理小目标或由相似地物引发的语义混淆问题时性能受限,显著降低了检索准确性. 本文针对遥感图像文本检索任务中跨模态对齐不足与细粒度语义表征困难的问题,提出基于共享提示与Mamba适配器的微调方法. 该方法首先通过设计跨模态共享提示生成模块,建立图像与文本特征的显式交互机制;然后构建面向遥感场景的图像与文本的双分支Mamba适配器微调模块,分别实现图像与文本特征的细粒度表征;最后,采用对比损失与隶属损失,缓解由遥感图像小目标或相似地物引起的语义混淆问题. 实验结果表明,本方法在遥感图像描述数据集(Remote Sensing Image Captioning Dataset,RSICD)和遥感图像文本匹配数据集(Remote Sensing Image-Text Match Dataset,RSITMD)数据集上平均召回率分别达到37.3%和48.05%,相较于当前最优的适配器微调方法分别提升3.68%和1.52%. 此外,消融实验验证了共享提示生成模块与Mamba适配器的有效性.

关键词: 图像文本检索;遥感图像;Mamba适配器;视觉语言模型微调
基金项目: 国家自然科学基金(No.62272461,No.62276266,No.62002360)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2025)09-3358-13
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250326

A Remote Sensing Image Text Retrieval Method Based on the Shared Prompt and Mamba Adapter

DU Wen-liang[1,2], XU Xiao-yu[1,2], ZHAO Jia-qi[1,2], LIU Bing[1,2], ZHOU Yong[1,2*]
(1. *School of Computer Science and Technology / School of Artificial Intelligence, China University of Mining and Technology, Xuzhou, Jiangsu* 221116, *China*;
2. *Mine Digitization Engineering Research Center of the Ministry of Education, Xuzhou, Jiangsu* 221116, *China*)

Abstract: Remote sensing image-text retrieval aims to quickly and accurately retrieve semantically matching text or images from a massive remote sensing image-text database based on a given image or text. With the rapid development of Earth observation technology, the application value of this technology in fields such as urban planning, disaster emergency response, and environmental monitoring has become increasingly prominent, making it a research hotspot in the current field of multimodal information processing. Vision-language pre-training models, pre-trained on general-domain data, have laid the technical foundation for general image-text retrieval tasks by achieving efficient semantic alignment between images and text. However, a significant domain gap exists between general and remote sensing data, which limits the perfor-

收稿日期:2025-04-24;录用日期:2025-08-29;责任编辑:覃怀银
*通讯作者:周勇

mance of these pre-trained models when directly applied to remote sensing tasks. Therefore, fine-tuning is necessary to adapt the vision-language model to the unique data distribution of the remote sensing domain. However, existing fine-tuning methods face two core challenges when applied to the remote sensing domain. First, there is insufficient cross-modal alignment: current fine-tuning methods lack explicit cross-modal information interaction mechanisms, making it difficult to fully model the intrinsic correlation between images and text. Second, it is difficult to achieve fine-grained semantic representation: existing methods often struggle to capture fine-grained semantic information in remote sensing images, such as vast differences in target scales, high similarity between ground object classes, and complex spatial-topological relationships. Performance is particularly limited when dealing with small targets or semantic confusion caused by similar ground objects, which significantly reduces retrieval accuracy. This paper addresses the problems of insufficient cross-modal alignment and difficulty in fine-grained semantic representation in remote sensing image-text retrieval tasks by proposing a fine-tuning method based on a shared prompt and Mamba adapter. This method first establishes an explicit interaction mechanism for image and text features by designing a cross-modal shared prompt generation module. Then, it constructs a dual-branch Mamba adapter fine-tuning module for remote sensing scenarios to achieve fine-grained representation of image and text features, respectively. Finally, it uses contrastive loss and affiliation loss to alleviate the semantic confusion caused by small targets or similar ground objects in remote sensing images. Experimental results show that this method achieves mean average recall rates of 37.3% and 48.05% on the remote sensing image captioning dataset (RSICD) and remote sensing image-text match dataset (RSITMD) datasets, respectively, which are improvements of 3.68% and 1.52% compared to the current state-of-the-art adapter fine-tuning method. Furthermore, ablation studies have verified the effectiveness of the shared prompt generation module and the Mamba adapter.

Key words: image-text retrieval; remote sensing images; Mamba adapter; visual-language model fine-tuning

Foundation Item(s): National Natural Science Foundation of China (No.62272461, No.62276266, No.62002360)

1 引言

随着地球观测技术的快速发展[1],我国已积累了海量遥感数据,如何高效检索这些数据以充分发挥其应用价值是当前面临的重大挑战. 遥感图像文本检索(Remote Sensing Image-Text Retrieval,RSITR)技术可实现数据库中海量遥感图像与文本的快速检索与匹配,已成为城市规划、灾害应急、农业监测以及智能化作战等领域的重要技术支撑[2].

目前主流图像文本检索方法的重要技术基础为对比语言-图像预训练模型(Contrastive Language-Image Pre-training,CLIP)[3]. 该模型采用对比学习的多模态预训练框架,将语言文本特征与视觉图像特征映射至统一的潜在语义空间,实现了文本与图像的跨模态表征与对齐. 然而,由于CLIP模型的预训练数据主要源于通用领域的图像-文本对,缺乏对遥感领域图像与文本特征的学习,导致其直接应用在遥感图像文本检索任务时的效果欠佳.

微调技术能够将CLIP模型在通用领域的图像文本跨模态对齐能力迁移至遥感领域. 目前,围绕将CLIP模型微调至遥感领域的方法大致可以分为两类:全参数微调和适配器(Adapter)微调. 全参数微调技术使用大规模遥感图像文本训练数据对CLIP模型的全部参数进行微调,可实现CLIP模型针对遥感领域知识的全参数调整[4-6]. 但全微调方法通常需要耗费巨大人力构建大规模预训练数据集,并且需要消耗大量训练时间和计算资源,使得全微调方法难以推广[7].

适配器微调方法通过在CLIP模型中引入轻量级适配器模块,并在微调训练CLIP过程中仅优化轻量级适配器模块中的参数,以实现CLIP模型对遥感领域知识的迁移. 然而,现有适配器微调方法仍面临两方面关键挑战:一是当前方法在微调过程中缺乏显式的跨模态信息交互机制,使得图像与文本特征间的关联性建模不足,导致微调后的遥感图像-文本特征未能有效对齐[8,9];二是遥感图像中存在的不同类别的目标尺度差异显著以及目标之间的空间关系表征困难等问题,使得现有适配器微调方法难以实现对遥感场景的细粒度语义表征,从而限制了微调模型性能的进一步提升[10].

因此,受近期Mamba模型在跨模态[11]及细粒度表征方面[12,13]表现的启发,本文提出基于共享提示与Mamba适配器的微调方法以应对上述挑战. 本研究的主要贡献如下:

(1)提出共享提示生成模块,通过设计图像和文本分支在微调过程中共同优化的共享参数矩阵,并将CLIP提取的模态标记(Token)与该矩阵融合为共享提示,建立跨模态信息的显式交互机制;与基于软提示方法仅通过损失函数设计间接约束跨模态对齐相比[14,15],共享提示直接优化图像与文本模态间的协同表征,有效提升CLIP模型对于遥感场景中图像特征相似但文本语义不同目标的表征能力.

(2)提出基于Mamba架构[16]适配器的微调模块,分别构建面向图像和文本特征提取的专用Mamba适配器,在图像适配器中引入树状状态空间框架,实现多尺度目标处理下的图像细粒度语义表征能力,在文本适配器中引入双向准可分矩阵,实现对复杂文本的细粒度语义解析和深层语义挖掘.

(3)提出融合共享提示与Mamba适配器的微调方法,将共享提示分别注入主干网络多层级Transformer结构对应的Mamba适配器中,有效实现CLIP模型对遥感图像与文本特征的多尺度跨模态对齐.相较于当前先进适配器微调方法,本文提出方法在RSICD和RSITMD两个遥感图像-文本检索数据集中取得更优的检索性能.

2 相关工作

2.1 遥感图像文本检索全参数微调

全参数微调方法通过大规模标注数据对预训练模型的所有参数进行优化.Liu等人[4]提出基于B2C(Box-to-Caption,B2C)与M2B(Mask-to-Box,M2B)的转换技术,将目标检测的边界框注释和语义分割掩码统一转换为自然语言描述,构建了规模达到当时公开数据集12倍的图像-文本检索预训练数据集,随后利用该数据集对改进的CLIP模型进行全参数微调,最终得到RemoteCLIP模型.Zhang等人[6]通过公开数据集过滤与标签数据描述生成方法,将现有遥感标注数据转化为自然语言描述,并结合网络公开图文数据,构建了首个规模达500万的遥感图文对数据集RS5M.同时,通过全参数微调在RS5M数据集上训练得到预训练模型GeoRSCLIP.

2.2 遥感图像文本检索适配器微调

适配器是一种轻量级的即插即用模块,由Houlsby等人[17]首次提出并将其嵌入在Transformer架构中,通过设计降维-激活-升维的三层结构实现跨模态特征的高效对齐.随着CLIP模型的出现,Gao等人[18]提出在CLIP末端添加双分支残差的适配器CLIP-Adapter,通过特征空间微调实现跨模态对齐.Chen等人[19]设计的AdaptFormer在视觉Transformer(Vision-Transformer,ViT)[20]中引入动态门控适配器,通过在前馈神经网络层中嵌入可学习分支,实现多尺度特征提取的自适应激活.Jiang等人[21]提出的交叉模态适配器结合时空注意力与条件门控机制,增强了文本-视频语义关联建模能力.Lu等人[22]提出的UniAdapter框架通过共享权重适配器与任务感知路由机制,实现多模态参数的动态分配.

在遥感领域适配器微调方面,Yuan等人[23]提出一种面向遥感图像-文本检索的参数高效迁移学习框架(Parameter-Efficient Remote Sensing Image-Text Retrieval,PE-RSITR),通过引入多模态适配器与混合对比损失,应对遥感数据类内相似性误匹配难题,在仅需微调0.16 M参数的情况下实现了性能的提升.Huang等人[11]进一步提出基于互信息门控适配器的HarMA框架,通过设计多模态适配器结合自适应三元组对比损失,应对跨模态对齐不足问题,在仅微调0.5 M模型参数情况下,其检索性能也超越了当时全微调模型的最优水平.

3 方法

本文提出的基于共享提示与Mamba适配器的遥感图像文本检索方法的整体框架如图1所示.该框架主要包括共享提示生成和Mamba适配器微调两个模块.

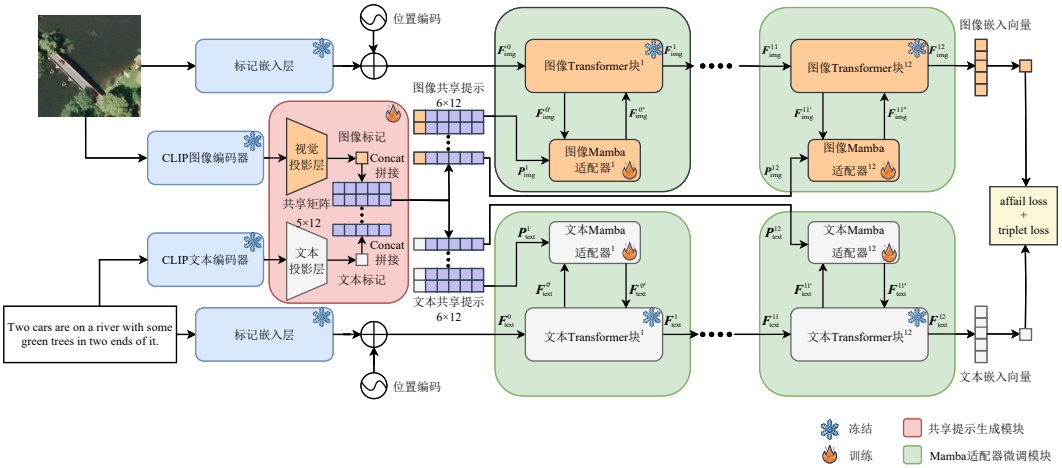


图1 本文提出方法的架构图

首先,将遥感图像-文本输入进共享提示生成模块.遥感图像-文本经过共享提示生成模块生成图像与文本提示.随后,将遥感图像-文本与提示输入进Mamba适配器微调模块.在该模块中,首先使用Trans-

former块进行初步的特征提取,之后将得到的浅层特征与提示输入进Mamba适配器中进行深层特征提取. 最后,将图像与文本特征嵌入到同一特征向量空间,采用对比损失与隶属损失进行训练. 接下来,将从共享提示生成模块、Mamba适配器微调模块、损失函数三个方面进行介绍.

3.1　共享提示生成模块

共享提示生成模块生成的提示包括图像共享提示和文本共享提示. 这两类提示均由预训练CLIP模型提取的标记及共享矩阵共同构成.

预训练的CLIP模型主要包括图像编码器和文本编码器两个部分. 图像编码器采用ViT架构,用于提取图像标记;文本编码器基于Transformer架构,用于提取文本标记. 为适配后续Mamba适配器的输入尺寸要求,设计图像投影层和文本投影层,分别将CLIP模型输出的512维的图像标记($\boldsymbol{T}_{\text{img}}$)与文本标记($\boldsymbol{T}_{\text{text}}$)分别映射至128维.

构建12组5×128维共享标记($\boldsymbol{W}^{i}_{\text{share}} \in \mathbf{R}^{5\times 128}$, $i=1,2,\cdots,12$),并将其拼接为可训练的共享矩阵:$\boldsymbol{W}_{\text{share}} \in \mathbf{R}^{12\times 5\times 128}$. 每组共享标记都分别与CLIP模型提取的图像标记和文本标记进行Concat拼接,生成图像共享提示($\boldsymbol{P}_{\text{img}} \in \mathbf{R}^{12\times 6\times 128}$)和文本共享提示($\boldsymbol{P}_{\text{text}} \in \mathbf{R}^{12\times 6\times 128}$). 其中,第$i$组图像共享提示和文本共享提示的生成如式(1)~(2)所示:

$$\boldsymbol{P}^{i}_{\text{img}} = \text{Concat}(\boldsymbol{W}^{i}_{\text{share}}, \boldsymbol{T}_{\text{img}}) \tag{1}$$

$$\boldsymbol{P}^{i}_{\text{text}} = \text{Concat}(\boldsymbol{W}^{i}_{\text{share}}, \boldsymbol{T}_{\text{text}}) \tag{2}$$

微调过程中,图像与文本分支的反向传播损失将协同优化共享矩阵的参数,形成图像与文本的协同表征. 文本模态提供的丰富语义信息引导模型解耦遥感场景中图像特征同质化的目标信息,增强对图像细粒度特征的语义判别能力.

3.2　Mamba适配器微调模块

Mamba适配器微调模块采用双分支架构设计,包含图像特征微调分支和文本特征微调分支. 每个分支均包含12层特征微调模块,每层模块包含1个Transformer块和1个Mamba适配器. 其中,Transformer块由Attention模块和MLP模块构成,Mamba适配器嵌入在两模块之间. 各特征微调模块依次接收上层特征微调模块输出的模态特征以及共享提示生成模块提供的模态共享提示,在可训练的Mamba适配器与冻结参数的Transformer模块共同作用下,生成新的模态特征并传递至后续特征微调模块. Mamba适配器通过可训练的参数,将CLIP模型预训练的图像文本对齐能力迁移至遥感领域. 具体而言,Transformer块将前序特征微调模块输出的模态特征输入其Attention模块生成浅层特征,然后将浅层特征传递至Mamba适配器. Mamba适配器在共享提示的协助下将浅层特征提取为深层特征,并输出至Transformer块的MLP模块,MLP模块生成传递至后续特征微调模块的模态特征. 需注意的是,上述模态特征和模态共享提示在图像特征微调特征分支中对应图像特征和图像共享提示,在文本特征微调分支中则对应文本特征和文本共享提示.

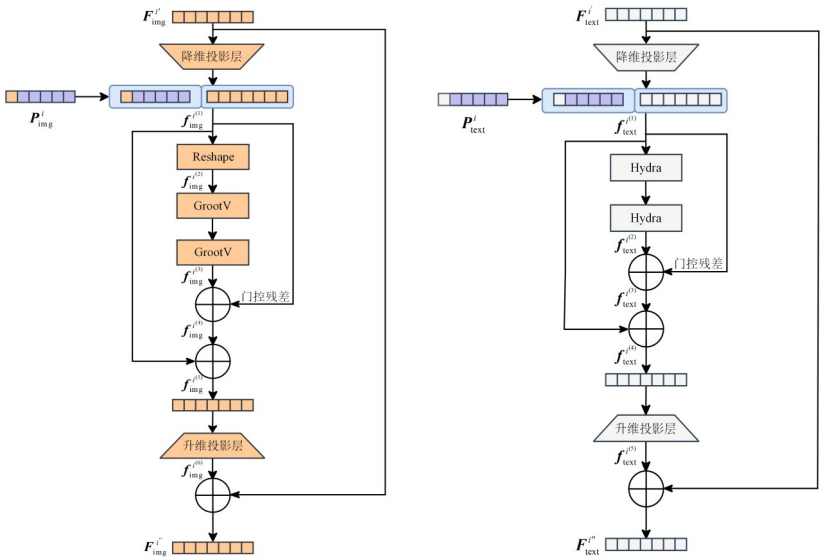
为提取最优模态特征,在图像特征微调分支与文本特征微调分支中分别设计不同的Transformer块与Mamba适配器. 图像特征微调分支中特征微调模块的Transformer块采用ViT架构. 图像Mamba适配器采用引入GrootV结构的Mamba模型[12],结构如图2(a)所示. GrootV结构继承自树状状态空间框架,其动态生成的拓扑网络能够自适应图像内容的空间分布,通过最小生成树结构有效捕捉像素间的局部连通性与全局层次关系,克服了传统状态空间模型因固定扫描顺序导致的二维视觉数据空间割裂问题. 这种自适应的树状信息传递机制使模型以更符合视觉规律的方式聚合图像特征,挖掘遥感图像中地物间的复杂空间关系,实现对遥感图像空间信息的细粒度表征. 文本特征微调分支中特征微调模块的Transformer块采用一般Transformer结构. 文本Mamba适配器采用引入Hydra结构[13]的Mamba模型,结构如图2(b)所示. 该结构通过引入双向准可分矩阵(Bidirectional Semiseparable Matrices,BSM)混合器,突破传统序列模型的单向性限制,在保留状态空间模型的高效计算特性的同时,建模上下文的前后依赖关系,实现对遥感描述中常见的属性信息(如颜色、大小等)和空间关系(如旁边、环绕等)等复杂文本描述的细粒度表征. 此外,Hydra模型的动态参数共享机制在减少冗余计算的同时,增强了模型对复杂语义逻辑的解析能力以及对语义特征的挖掘能力.

图像特征微调分支接收的初始图像特征由N个图像块,由分类标记嵌入(CLassification Token,CLS)和位置编码嵌入(Positional Embedding,PE)构建而成. 具体而言,输入图像$\boldsymbol{I} \in \mathbf{R}^{W\times H\times C}$通过卷积层后被划分为$N$个$P\times P$大小的图像块($N=H\times W/P^2$),每个图像块展平后通过可训练的线性层映射成1×512维的图像块嵌入向量$\boldsymbol{v}_j$($j=1,2,\cdots,N$). 随后在图像块嵌入向量前拼接可学习的分类嵌入向量$\boldsymbol{v}_{\text{cls}}$,并与位置编码$\boldsymbol{V}_{\text{pos}}$进行元素级求和,最终构建的初始图像特征$\boldsymbol{F}^{0}_{\text{img}}$,其构建过程如式(3)所示:

$$\boldsymbol{F}^{0}_{\text{img}} = (\boldsymbol{v}_{\text{cls}}, \boldsymbol{v}_1, \boldsymbol{v}_2 \cdots, \boldsymbol{v}_N) + \boldsymbol{V}_{\text{pos}} \tag{3}$$

在第i层的图像特征微调模块中,上层模块输出的图像特征$\boldsymbol{F}^{i-1}_{\text{img}}$输入至本层Transformer块的Attention模块中,得到浅层图像特征$\boldsymbol{F}^{i'}_{\text{img}}$浅层图像特征计算过程如式(4)所示:

$$\boldsymbol{F}^{i'}_{\text{img}} = \text{LS}\left(\text{MSA}\left(\text{LN}\left(\boldsymbol{F}^{i-1}_{\text{img}}\right)\right)\right) + \boldsymbol{F}^{i-1}_{\text{img}} \tag{4}$$



(*a*) 图像Mamba适配器 (*b*) 文本Mamba适配器

图2 Mamba适配器微调模块架构图

其中，LS、MSA和LN分别为尺度缩放层(Layer Scale，LS)、多头自注意力层(Multihead Self-Attention，MSA)和归一化层(Layer Norm，LN).

浅层图像特征 $\boldsymbol{F}_{\text{img}}^{i'}$ 与图像共享提示 $\boldsymbol{P}_{\text{img}}^{i}$ 共同输入至图像Mamba适配器，以获取深层图像特征. 具体而言，浅层图像特征 $\boldsymbol{F}_{\text{img}}^{i'}$ 经降维投影层降维至128维，并与图像共享提示 $\boldsymbol{P}_{\text{img}}^{i}$ 进行拼接，得到浅层图像共享特征 $\boldsymbol{f}_{\text{img}}^{i^{(0)}}$. 随后，将 $\boldsymbol{f}_{\text{img}}^{i^{(0)}}$ 在Reshape模块中由一维向量变换为二维矩阵 $\boldsymbol{f}_{\text{img}}^{i^{(2)}}$，以适配GrootV结构的输入形式. 然后，将 $\boldsymbol{f}_{\text{img}}^{i^{(2)}}$ 依次输入两个GrootV结构得到图像特征 $\boldsymbol{f}_{\text{img}}^{i^{(3)}}$.

图像Mamba适配器中采用动态门控双残差的融合机制实现特征增强. 其中，门控残差以非线性平滑的方式调控中间特征与原始输入的混合比例，形成初级特征校正，以避免单一残差路径可能引发的模态偏差. 直接残差则进一步强化梯度传播路径，以确保在复杂特征的底层关键信息在交互过程中得以保留. 图像特征 $\boldsymbol{f}_{\text{img}}^{i^{(0)}}$、$\boldsymbol{f}_{\text{img}}^{i^{(3)}}$ 和 $\boldsymbol{f}_{\text{img}}^{i^{(4)}}$ 经动态门控双残差得到图像特征 $\boldsymbol{f}_{\text{img}}^{i^{(5)}}$ 的计算过程如式(5)~(7)所示，其中gate表示可训练的门控参数.

$$\gamma=\text{sigmoid}(\text{gate}) \tag{5}$$

$$\boldsymbol{f}_{\text{img}}^{i^{(4)}}=\gamma\boldsymbol{f}_{\text{img}}^{i^{(0)}}+(1-\gamma)\boldsymbol{f}_{\text{img}}^{i^{(3)}} \tag{6}$$

$$\boldsymbol{f}_{\text{img}}^{i^{(5)}}=\boldsymbol{f}_{\text{img}}^{i^{(4)}}+\boldsymbol{f}_{\text{img}}^{i^{(0)}} \tag{7}$$

将图像特征 $\boldsymbol{f}_{\text{img}}^{i^{(5)}}$ 的图像共享提示部分去除并通过升维投影层得到512维的图像特征 $\boldsymbol{f}_{\text{img}}^{i^{(6)}}$. 然后，将图像特征 $\boldsymbol{f}_{\text{img}}^{i^{(6)}}$ 与浅层图像特征 $\boldsymbol{F}_{\text{img}}^{i'}$ 进行残差连接，得到深层图像特征 $\boldsymbol{F}_{\text{img}}^{i''}$，计算过程如式(8)所示：

$$\boldsymbol{F}_{\text{img}}^{i''}=\boldsymbol{f}_{\text{img}}^{i^{(6)}}+\boldsymbol{F}_{\text{img}}^{i'} \tag{8}$$

最终，将深层图像特征 $\boldsymbol{F}_{\text{img}}^{i''}$ 输入至Transformer块的MLP模块中进行非线性变换，得到输入至下层特征微调模块的图像特征 $\boldsymbol{F}_{\text{img}}^{i}$，计算过程如式(9)所示：

$$\boldsymbol{F}_{\text{img}}^{i}=\text{LS}\left(\text{MLP}\left(\text{LN}\left(\boldsymbol{F}_{\text{img}}^{i''}\right)\right)\right)+\boldsymbol{F}_{\text{img}}^{i''} \tag{9}$$

其中，MLP表示由多层感知机(Multi-Layer Perceptron，MLP)实现的非线性变换操作.

文本特征微调分支接收的初始文本特征的构造方式如式(10)所示. 输入文本首先通过分词器进行分词处理，生成词向量序列. 随后，词向量经词向量嵌入层映射为文本向量 $\boldsymbol{t}_i$，再与位置嵌入 $\boldsymbol{T}_{\text{pos}}$ 进行元素级求和操作，最终得到初始文本特征 $\boldsymbol{F}_{\text{text}}^{0}$.

$$\boldsymbol{F}_{\text{text}}^{0}=(\boldsymbol{t}_1,\boldsymbol{t}_2\cdots,\boldsymbol{t}_N)+\boldsymbol{T}_{\text{pos}} \tag{10}$$

对比图2中的图像Mamba适配器和文本Mamba适配器可以看出，两适配器除中间的Mamba结构分别为两层GrootV和两层Hydra，其他特征处理过程完全一致，即对于第 i 层的文本特征微调模块，输入上层模块获得的文本特征 $\boldsymbol{F}_{\text{text}}^{i-1}$，将获得经过文本Mamba适配器微调后用于输入至下层模块的文本特征 $\boldsymbol{F}_{\text{text}}^{i}$.

输入图像与文本分别经过12层图像与文本特征微调模块微调后得到图像特征 $\boldsymbol{F}_{\text{img}}^{12}$ 和文本特征 $\boldsymbol{F}_{\text{text}}^{12}$. 使用线性层将 $\boldsymbol{F}_{\text{img}}^{12}$ 和 $\boldsymbol{F}_{\text{text}}^{12}$ 嵌入到同一向量空间，得到最终的图像特征 $\boldsymbol{F}_{\text{img}}$ 和文本特征 $\boldsymbol{F}_{\text{text}}$.

3.3 损失函数

本模型的损失函数为对比损失(Contrastive Loss，L_c)与隶属损失(Affiliation Loss，L_a)[24]的加权组合.

对比损失通过最大化跨模态正样本对的相似度并

抑制负样本对的匹配得分,建立了图像与文本的全局对齐关系. 该损失利用批内负样本构造和温度系数调节,避免了复杂采样策略,显著提升了跨模态检索的泛化性.

隶属损失基于类别先验信息,通过动态聚类中心约束同类样本在嵌入空间的语义聚合,强化了类内样本的紧凑性和类间样本的分离度,缓解了小目标或相似地物引起的语义混淆问题,适宜解决遥感图像中存在的模态内相似性高、主体不突出等问题. 具体损失函数定义如式(11)~(13)所示:

$$L_c=-\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp(s_{i,i}/\tau)}{\sum_{j=1}^{N}\exp(s_{i,j}/\tau)}+\sum_{i=1}^{N}\log\frac{\exp(s_{i,i}/\tau)}{\sum_{j=1}^{N}\exp(s_{j,i}/\tau)}\right] \tag{11}$$

$$L_a=-\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp(s^v_{i,i}/\tau)}{\sum_{j=1}^{N}\exp(s^v_{i,j}/\tau)}+\sum_{i=1}^{N}\log\frac{\exp(s^t_{i,i}/\tau)}{\sum_{j=1}^{N}\exp(s^t_{i,i}/\tau)}\right] \tag{12}$$

$$L=\lambda_1 L_c+\lambda_2 L_a \tag{13}$$

其中,$s_{i,j}=\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{t}_j$表示第$i$幅图像与第$j$个文本的相似度;$s^v_{i,j}=\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{t}_j^*$表示第$i$幅图像与第$j$个文本的聚类中心的相似度;$s^t_{j,i}=\boldsymbol{t}_j^{\mathrm{T}}\boldsymbol{v}_i^*$表示第$j$个文本与第$i$幅图像对应的聚类中心的相似度;$\lambda_1$和$\lambda_2$分别表示对比损失和隶属损失的权重;$\tau$为可训练温度参数;$\log(\cdot)$和$\exp(\cdot)$表示自然对数和自然指数函数.

4　实验结果及分析

本节将提出模型与8个当前最优适配器微调方法(Adapter[17]、CLIP-Adapter[18]、AdaptFormer[19]、Croos-Model Adapter[21]、UniAdapter[22]、R-Adapter[8]、PE-RSITR[23]和HarMA[11])和1个重参(Reparameterization)微调方法(PETR-RaMa[25])在RSICD[26]与RSITMD[27]两个遥感图像文本检索数据集上进行定量对比与分析,评估指标选择Recall@K和平均召回率. 同时,选取代表性实例的可视化结果进行定性对比与分析,并通过消融实验验证方法中各模块的作用.

4.1　实验设置

在本文实验中,提出方法与其他基于CLIP的对比方法均对预训练的CLIP-B/32模型进行微调. 所有实验的训练与推理均在拥有4张NVIDIA GeForce RTX 3090 GPU的服务器上完成. 提出方法在RSICD和RSITMD数据集上的训练轮次(epoch)数分别设为5和15,批次大小设置为128,优化器采用AdamW,学习率设置为4×10^{-4},权重衰减率设置为0.04,gate和τ的初始值分别为0.6和0.07.

4.2　数据集介绍

本文采用遥感图像文本检索任务广泛使用的数据集:RSICD和RSITMD. RSICD数据集包含10 921幅遥感图像及54 605条场景级文本描述,其样本具有类内多样性高和类间差异性低的特点,通常用于验证模型对全局语义的理解能力以及在不同场景的泛化能力. RSITMD数据集包含4 743张遥感图像和23 715条细粒度标注文本,其文本标注细粒度高于RSICD数据集,通常用于验证模型在目标级跨模态对齐中的局部敏感性. 两数据集分别从宏观场景适配与微观目标关联两个维度构建评估体系,可有效评估模型在复杂地物分布、多尺度语义关联及小目标干扰等挑战下的图像文本检索精度与鲁棒性. 本文将RSICD与RSITMD数据集均按照70%、20%、10%的比例划分为训练集、验证集与测试集.

4.3　评估指标

图像文本检索任务包括图像到文本检索(Image-Text Retrieval,ITR)和文本到图像检索(Text-Image Retrieval,TIR)两个任务. 为定量评估各方法在两个检索任务中的性能,本文实验使用R@K(Recall@K)和平均召回率mR(mean Recall,mR)作为评估指标. 其中,R@K指标表示以单张图像或单条文本作为查询输入,模型从候选库中检索出相关文本或图像,正确检索结果出现在前K个预测的概率. 在本文实验中,K分别取1、5、10.

4.4　对比实验和分析

4.4.1　定量实验结果与分析

各方法在RSICD和RSITMD数据集上微调后的定量实验结果分别如表1和表2所示. 相较于其他对比方法,本文提出方法在两个数据集的图像文本检索和文本图像检索两个任务中均取得了最优的R@K指标和平均召回率.

根据模型参数规模,可将本文对比的方法分为两类:模型参数量较小的方法和模型参数量较大的方法. 其中,Adapter、Adapter Former、Croos-Model Adapter和PE-RSITR采用单Adapter架构设计,因此模型参数量较小;其他方法为参数量较大的方法. Adapter和Adapter Former仅能将适配器部署在图像或文本的单一特征微调分支,无法实现图像与文本特征的同步微调,导致图像与文本特征的信息表征存在不对称性,从而限制了这两种方法的检索性能. Croos-Model Adapter和PE-RSITR的图像与文本特征微调分支中的特征微调模块共用一个适配器,实现了两个分支的同步微调. 但参数规模有限,导致其共享特征的表征能力较弱. 此外,这两种方法的共享特征模块可能在融合过程中覆盖原始模型提取

的部分特征信息,导致模态特征信息的丢失. 然而PE-RSITR凭借动态参数共享机制与多模态特征融合优化策略,使得其检索性能优于参数量更大的CLIP-Adapter、UniAdapter、R-Adapter和PERT-RaMa. 与之相对的是,本文提出方法在图像特征微调分支与文本特征微调分支的特征微调模块中分别设置了独立的适配器,并针对图像与文本两个模态分别设计了不同结构的Mamba适配器. 同时,通过采用拼接共享提示的方法,实现了在保留原始特征完整性的前提下共享图像与文本的特征. 实验结果显示,相较于参数量较小模型中性能最优的PE-RSITR,本文提出方法在RSICD和RSITMD数据集上的平均召回率分别提升了6.18%和3.58%.

表1 各方法在RSICD数据集的定量实验结果 单位:%

方法	参数量/M	图像-文本检索			文本-图像检索			平均召回率
		R@1	R@5	R@10	R@1	R@5	R@10	
CLIP	—	6.77	15.37	23.15	5.01	15.75	24.21	15.04
Adapter	0.17	8.73	24.73	37.81	8.43	26.02	43.33	24.84
CLIP-Adapter	0.52	7.11	19.48	31.01	7.67	24.87	39.73	21.65
AdaptFormer	0.17	12.46	28.49	41.86	9.09	29.89	46.81	28.10
Croos-Model Adapter	0.16	11.18	27.31	40.62	9.57	30.74	48.36	27.96
UniAdapter	0.55	12.65	30.81	42.74	9.61	30.06	47.16	28.84
R-Adapter	1.70	10.45	32.48	48.20	14.00	31.29	43.64	30.01
PERT-RaMa	0.53	12.97	31.53	45.20	11.47	34.04	51.46	31.11
PE-RSITR	0.16	14.13	31.51	44.78	11.63	33.92	50.73	31.12
HarMA	0.50	16.36	34.48	47.74	12.92	37.17	53.07	33.62
Ours	0.55	**20.77**	**40.16**	**52.97**	**14.75**	**38.76**	**56.4**	**37.3**

注:加粗表示最优结果,下划线表示次优结果.

表2 各方法在RSITMD数据集的定量实验结果 单位:%

方法	参数量/M	图像-文本检索			文本-图像检索			平均召回率
		R@1	R@5	R@10	R@1	R@5	R@10	
CLIP	—	9.29	26.33	37.39	7.79	23.67	38.89	23.89
Adapter	0.17	13.75	27.64	39.96	12.89	40.09	59.91	32.37
CLIP-Adapter	0.52	12.83	28.84	39.05	13.30	40.20	60.06	32.38
AdaptFormer	0.17	16.71	30.16	42.91	14.27	41.53	61.46	34.81
Croos-Model Adapter	0.16	18.16	36.08	48.72	16.31	44.33	64.75	38.06
UniAdapter	0.55	19.86	36.32	51.28	17.54	44.89	56.46	39.23
R-Adapter	1.70	16.55	46.81	**64.69**	21.24	42.70	53.98	41.00
PERT-RaMa	0.53	20.66	44.16	58.94	17.95	50.81	69.08	43.60
PE-RSITR	0.16	23.67	44.07	60.36	20.10	50.63	67.97	44.47
HarMA	0.50	25.81	48.37	60.61	19.92	53.27	71.21	46.53
Ours	0.55	**28.76**	**49.78**	59.51	**22.52**	**55.53**	**72.17**	**48.05**

注:加粗表示最优结果,下划线表示次优结果.

对于参数量较大的模型,CLIP-Adapter仅在CLIP模型后加入一个瓶颈层,没有在CLIP模型的各特征提取层中进行知识迁移,导致其未能在RSICD和RSITMD数据集上学习到足够的遥感特征. UniAdapter虽然在各特征提取层中进行知识迁移,但其网络结构设计过于复杂,难以有效提取图像与文本中目标的特征. R-Adapter模型仅通过构建多正样本边际噪声对比估计损失来约束图像与文本的特征对齐,导致其在两数据集上的文本-图像检索效果均不理想. PERT-RaMa模型通过构建基于Kronecker积的低秩适配(Low-Rank Adaptation,LRA)缓解过拟合问题,但该低秩适配的内在秩以及超参数的组合依赖人工设置,难以适应遥感图像固有的类分布不均衡和场景复杂性等问题. HarMA采用了相对简单的网络结构设计并引入了共享机制,但其共享方式仅为结构层面的隐性共享. 与之相对的是,本文提出方法不仅在CLIP模型的各特征提取层中进行知识迁移,而且适配器的网络结构设计也较为简单. 同时,基于Mamba架构的适配器能够有效提取图像与文本中目标的特征,显著提升了检索任务的准确度. 实验结果显示,相较于参数量较大模型中性能最优

的HarMA,本文提出方法在RSICD和RSITMD数据集上的平均召回率分别提升了3.68%和1.52%.

4.4.2 特征微调可视化定性结果与分析

本节使用Grad-CAM[28]可视化方法对不同方法的神经网络激活图进行可视化展示与分析,以验证本文提出方法对于遥感特征的微调有效性. 对比实验涵盖未微调的预训练基础模型CLIP、前文定量评估中表现次优的HarMA方法以及本文提出方法,具体选取较大目标类别的"beach"和"lake"、较小目标类别的"cars"以及包含细粒度属性的目标类别"large house with brown roof"作为特征提取目标,展示三种方法对这四个类别进行特征表示过程中的神经网络激活情况,可视化结果如图3所示. 其中,输入文本以红色标注,模型图像分支对输入文本的激活程度以颜色梯度表示:颜色越红表示激活程度越高,颜色越蓝表示激活程度越低.

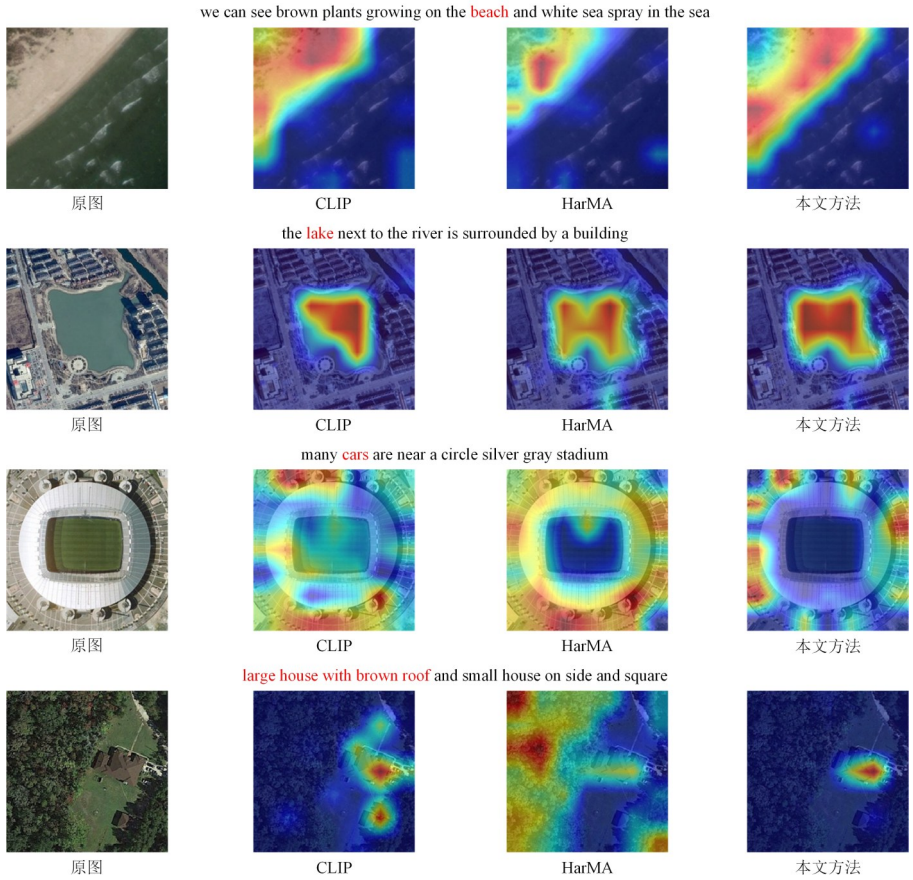


图3　特征微调可视化结果

从图3第一行和第二行的可视化结果可以看出,原始CLIP在提取"beach"目标特征的同时激活了非目标类别"plants"的特征. 对于同样是大目标的"lake"类别,CLIP获取的特征表示不完全. 微调的两种方法对于两个大目标类别的特征表示效果优于原始CLIP. 相较于HarMA方法,本文提出方法对于目标类别的激活集中程度更高,而HarMA方法激活了少部分非目标类别"plants"和"buildings"的特征,表明本文提出方法对于两个大目标类别获取了更好的特征表示.

对于较小目标类别"cars",原始CLIP对其在图中的表示较为分散,并且对于某些明显有"cars"目标的位置未能实现有效的特征表示. 经过微调的HarMA方法特征表示更加集中,但对于"stadium"出现了错误的激活. 相较而言,本文提出方法对于"cars"的特征表示更加集中且准确.

图3最后一行展示了三种方法对于包含细粒度属性的目标类别"large house with brown roof"的特征表示效果. 原始CLIP不仅激活了"large house with brown roof"的目标特征并且错误地激活了"small house"的目标特征,表明CLIP仅实现了对"house"类别的特征表示,没有表示"large""brown"和"small"的细粒度属性特征. 本例中,HarMA的特征表示结果甚至不如未微调的原始CLIP,说明HarMA对于包含细粒度属性的类别特征没能实现有效微调. 相较而言,本文提出方法对包含细粒度属性的目标类别实现了准确的特征表示,正确激活了带有"large"和"brown"细粒度属性特征的"house"类别.

综上所述，本文提出方法不仅实现了多尺度目标特征表示，而且对于细粒度属性特征也实现了准确表示，验证了本方法中共享提示模块对于图文特征信息交互能力的加强.

4.4.3 检索结果实例定性结果与分析

本文从RSITMD数据集中随机展示两组图像-文本检索和文本-图像检索的测试实例，定性分析本文提出方法的检索性能. 数据集中每图像关联5个文本，因此以图像为查询时至多可获得5个正确文本检索，而以文本为查询时则至多只能获得1个正确图像检索. 对比实验同样涵盖未微调的预训练基础模型CLIP、前文定量评估中表现次优的HarMA方法以及本文提出方法，检索结果实例如图4所示.

在图像-文本检索任务中，第一个查询图像展示了由工业厂房构成的工业区场景，图像右上方包含1个较小的河流目标. 本文提出方法成功将该图像关联的5个正确文本全部纳入前5检索结果. 相比之下，HarMA方法检索结果均为错误文本，尽管其能够识别图像中“building”和“road”的目标特征，但未能有效表征“industrial”场景属性，表明HarMA方法对于场景属性的特征表示能力弱于本文提出方法.

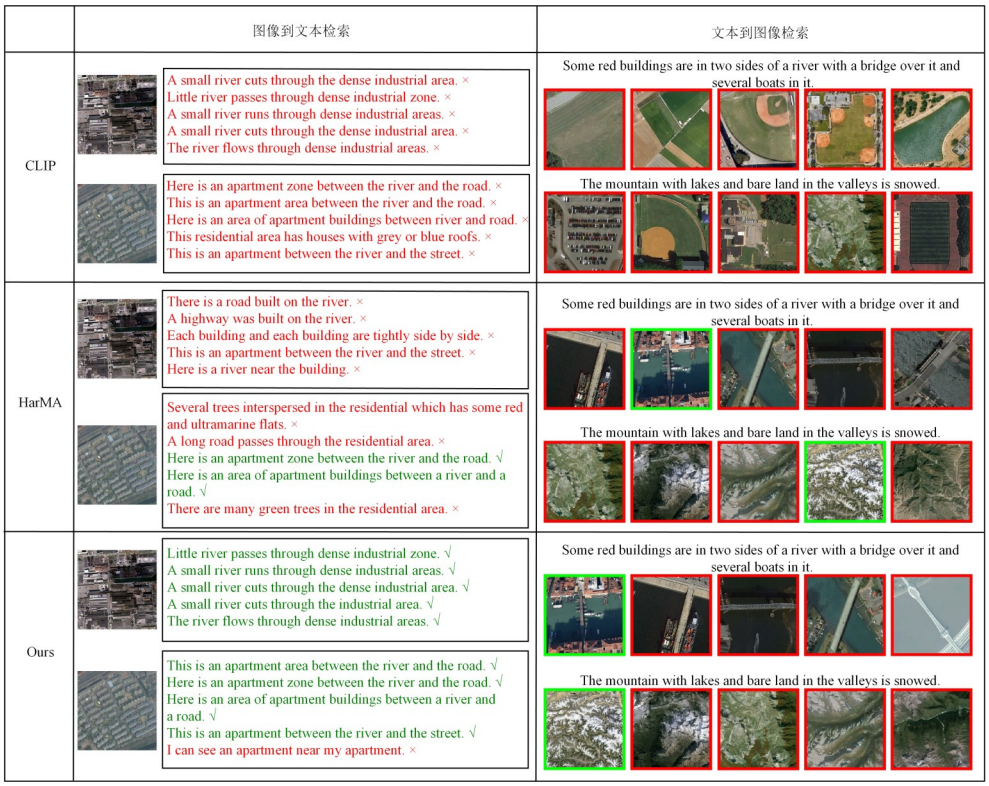


图4 检索结果实例

第二个查询图像主要描述了河流与道路间分布的公寓楼场景. 本文提出方法前4个检索结果均为正确文本，R@5指标达到80%. 而HarMA方法仅在第3、4位出现正确文本，R@5指标为40%. 进一步分析显示，HarMA方法错误检索的文本虽包含“road”和“residential”目标特征，但未能正确表征目标类别的空间关系，表明HarMA方法对于空间关系的特征表示能力弱于本文提出方法.

在上述两个图像-文本检索示例中，未微调的CLIP模型均未能在前5个检索中获取正确检索结果，表明了微调对于遥感图像-文本检索任务的必要性.

在文本-图像检索任务中，第一个查询文本为“Some red buildings are in two sides of a river with a bridge over it and several boats in it”，本文提出算法检索概率最高的结果即为正确图像. 相比之下，HarMA方法的正确图像仅位于其检索概率第二位，其概率最高的结果虽包含更显著“river”目标类别，但缺少“red buildings”目标类别，表明HarMA对于细粒度特征的表示能力弱于本文提出方法.

第二个查询文本是“The mountain with lakes and bare land in the valleys is snowed”，本文提出算法检索概率最高的结果同样为正确图像. 而HarMA方法的正确图像位于其检索概率第四位，其前三项检索图像虽包含“valleys”目标类别，但“snowed”属性的特征表示都不显著，表明HarMA方法对于属性特征的表示能力弱于本文提出方法.

在上述两个文本-图像检索示例中,未微调CLIP模型的检索图像中均未出现“river”“mountain”和“lakes”等查询文本中核心目标类别,表明了微调对于遥感文本-图像检索任务的必要性.

4.5　消融实验

4.5.1　模块消融实验

本文提出方法设计了共享提示生成模块和Mamba适配器微调模块,并在对比损失的基础上引入了隶属损失. 本节设计消融实验以验证两个模块以及隶属损失在图像文本检索任务的有效性和必要性. 本消融实验在RSITMD数据集上完成,实验结果如表3所示. 其中,Full表示提出方法的完整模型,SA表示共享提示生成模块,MA表示Mamba适配器微调模块中的Mamba适配器,AL表示隶属损失,w/o表示缺少即with out. 由于SA模块生成的提示需要在MA模块中使用,因此没有设置实验w/o|MA和w/o|MA+AL.

(1)共享提示生成模块. 表3的第四行所示为仅移除共享提示生成模块后的实验结果. 实验结果表明移除共享提示模块后,模型的平均召回率下降了0.69%. 其中,图像-文本检索任务的R@1、R@5和R@10指标分别下降1.55%、0.89%和0.22%,文本-图像检索任务的R@1、R@5和R@10指标分别下降1.06%、0.6%和0.8%. 这表明共享提示生成模块对图像-文本检索性能具有显著提升作用,尤其在图像-文本检索任务的R@1指标中提升尤为显著.

(2)Mamba适配器微调模块. 表3的第三行所示为仅在Mamba适配器微调模块中加入了Mamba适配器的实验结果. 与第六行所有模块都不添加的实验结果相比,添加了Mamba适配器后模型的平均召回率提升了0.8%. 其中,文本-图像检索任务的R@1、R@5和R@10指标分别提升了2.15%、3.78%和0.62%,图像-文本检索任务的R@1和R@5指标分别提升了1.99%和3.54%,但R@10指标下降了1.11%,说明添加了Mamba适配器后,正确检索结果更容易集中在第1个和前5个.

(3)隶属损失. 表3的第五行所示为仅移除隶属损失的实验结果. 实验结果表明移除隶属损失后,模型的平均召回率下降了0.77%. 其中,图像-文本检索任务的R@1和R@5指标分别下降了1.99%和0.66%,文本-图像检索任务的R@1和R@5指标分别下降了1.15%和1.55%. 但两任务中的R@10指标都有不同程度上升,说明隶属损失对于正确检索结果出现在前10的影响不大.

综上所述,每增加一个模块(损失函数)或增加两个模块(损失函数)的组合,模型的图像文本检索性能都会有不同程度的提升,验证了提出的两个模块和引入损失函数的有效性. 同时,最优性能出现在提出方法的完整模型,验证了提出两个模块和引入损失函数的必要性.

表3　模块消融实验结果　　单位:%

模型	MA	SP	AL	图像-文本检索			文本-图像检索			平均召回率
				R@1	R@5	R@10	R@1	R@5	R@10	
Full	√	√	√	**28.76**	<u>49.78</u>	59.51	**22.52**	<u>55.53</u>	72.17	**48.05**
w/o|MA+SP			√	<u>27.43</u>	48.45	59.96	21.37	54.91	71.77	47.32
w/o|SP+AL	√			25.88	**50.22**	<u>60.84</u>	20.58	53.98	<u>72.35</u>	47.31
w/o|SP	√		√	27.21	48.89	59.29	<u>21.46</u>	**55.93**	71.37	<u>47.36</u>
w/o|AL	√	√		26.77	49.12	59.73	21.37	53.98	**72.7**	47.28
w/o|MA+SP+AL				23.89	46.68	**61.95**	20.4	54.42	71.73	46.51

注:加粗表示最优结果,下划线表示次优结果.

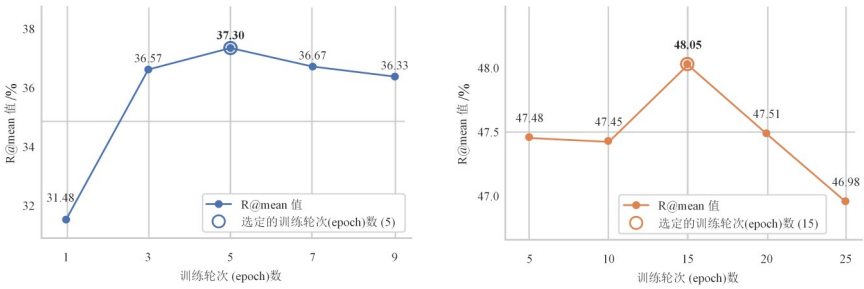
4.5.2　训练轮次数消融实验

本节在RSICD和RSITMD数据集上设置不同的训练轮次(epoch)数,分析不同训练轮次数对提出模型在两个数据集上的影响,并说明本文训练轮次数的设置原因. 实验结果如图5所示,图5(a)和图5(b)分别代表RSICD和RSITMD数据集上结果,两个数据集训练轮次数分别设置为1、3、5、7、9和5、10、15、20、25. 实验结果表明,随着训练轮次数的增加,提出模型在两个数据集上的性能均呈先上升后下降的趋势,轮次数分别为5(RSICD)和15(RSITMD)时达到性能最优. 因此,本文提出方法在RSICD和RSITMD数据集上的训练轮次数分别设置为5和15.

4.5.3　损失函数权重消融实验

本文提出方法包含对比损失和隶属损失两个损失函数,对应的权重分别为α和β. 本节在RSITMD上测试不同的权重比例设置对提出模型性能的影响,α:β的比例分别选取为1∶1、0.5∶1、1∶0.5、0.5∶0.5,实验结果如表4所示.

实验结果表明,当对比损失与隶属损失比例设置为1∶1时,提出方法在图像-文本检索及文本-图像检索任务中性能显著优于其他权重比例. 当对比损失权重减半时(权重比例0.5∶1),提出方法平均召回率下降



(*a*) RSICD数据集 (*b*) RSITMD数据集

图5 不同数据集上训练轮次(epoch)数的消融实验结果

表4 损失函数不同权重占比下的实验结果 单位:%

α:*β*	图像-文本检索			文本-图像检索			平均召回率
	R@1	R@5	R@10	R@1	R@5	R@10	
1∶1	**28.76**	**49.78**	59.51	**22.52**	**55.53**	**72.17**	**48.05**
0.5∶1	26.99	47.57	**63.5**	21.55	54.65	71.46	47.62
1∶0.5	25.00	48.01	62.61	21.99	55.04	70.66	47.22
0.5∶0.5	26.55	48.89	60.84	21.77	53.54	70.58	47.03

注:加粗表示最优结果,下划线表示次优结果.

0.43%,而隶属损失权重减半(权重比例1∶0.5)导致平均召回率显著下降0.83%,这表明隶属损失对提出方法检索性能的贡献更为关键.

因此,本文提出方法将对比损失和隶属损失权重比例设置为1∶1,以实现最优的检索性能.

4.5.4 Mamba适配器输入维度消融实验

本节在RSITMD数据集上对提出模型的Mamba适配器设置了不同的输入维度,分析输入维度对于模型参数量以及模型性能的影响,并说明本文输入维度的设置原因. 实验结果如表5所示. 其中,Mamba适配器输入维度设置为64、128、256. 实验结果表明,随着输入维度的提升,模型的参数量也随之提升,但模型性能却在输入维度为128而不是256时达到最优. 这说明过高的Mamba适配器输入维度和参数量,可能会影响CLIP模型对图像和文本的表征能力. 因此,本文将Mamba适配器输入维度设置为128,以取得最优的检索性能以及合理的模型参数量.

4.5.5 共享提示中标记维度比例消融实验

本文共享提示生成模块分别生成图像共享提示和文本共享提示. 其中,图像共享提示由图像标记和共享标记拼接而成,文本共享提示由文本标记和共享标记拼接而成. 本文中的模态标记,由1个1×128的1维特征向量构成,共享标记由1个5×128的5维特征向量构成. 本节在RSITMD上测试模态标记与共享标记在不同维度比例情况下对提出方法检索性能的影响. 为了限制共享提示的参数量,将其总维度限制为6. 本消融实验将模态标记维度与共享标记维度的比例(*M*∶*S*)设置为:1∶5、2∶4、3∶3、5∶1,实验结果如表6所示. 其中,模态标记为图像标记或文本标记.

表5 Mamba适配器不同输入维度下的的实验结果 单位:%

维度	参数量/M	图像-文本检索			文本-图像检索			平均召回率
		R@1	R@5	R@10	R@1	R@5	R@10	
64	**0.23**	27.65	46.46	**61.28**	21.73	53.63	70.58	46.89
128	0.55	**28.76**	**49.78**	59.51	**22.52**	55.53	**72.17**	**48.05**
256	1.52	27.88	46.68	60.84	20.49	**55.62**	71.86	47.23

注:加粗表示最优结果,下划线表示次优结果.

表6 模态标记与共享标记的维度比例对提出方法检索性能的影响 单位:%

M:*S*	参数量/M	图像-文本检索			文本-图像检索			平均召回率
		R@1	R@5	R@10	R@1	R@5	R@10	
1:5	**0.55**	**28.76**	**49.78**	59.51	22.52	55.53	72.17	**48.05**
2:4	0.56	28.54	48.67	59.07	20.84	54.60	**72.79**	47.42
3:3	0.58	**28.76**	46.46	57.74	21.59	**55.84**	72.57	47.16
4:2	0.59	28.32	47.35	59.29	21.64	55.49	71.55	47.27
5:1	0.60	26.77	48.01	**59.96**	**22.83**	54.16	71.42	47.19

注:加粗表示最优结果,下划线表示次优结果.

实验结果表明,当模态标记维度与共享标记维度的比例设置为1∶5时,提出方法表现出了最优的检索性能以及最小的模型参数量. 相较而言,其他维度比例没有展示出其在参数量或在检索性能上的优势. 因此,本文将模态标记维度与共享标记维度的比例设置为1∶5.

5 结束语

本文提出一种基于共享提示与Mamba适配器的遥感图像文本检索微调方法,将Mamba架构引入遥感图像-文本检索任务,构建的共享提示与双分支适配器有效解决了现有方法跨模态交互薄弱和细粒度表征不足的问题. 通过自适应树状特征聚合与双向语义建模机制,实现了遥感图像空间关系与文本语义逻辑的深度对齐与细粒度表征. 实验验证了提出方法在多个检索指标上的优越性,可视化分析进一步揭示了其对多尺度目标及属性特征的精准捕捉,未来将探索更高效、更快速的参数共享策略,并拓展至遥感图像描述领域.

参考文献

[1] 罗忠涛, 龚彦如, 黎雾萱, 等. 天波超视距雷达地海杂波图像增强与检测器设计[J]. 电子学报, 2024, 52(12): 4037-4047.
LUO Z T, GONG Y R, LI J X, et al. Land-sea clutter image enhancement and detector design for sky-wave over-the-horizon radar[J]. Acta Electronica Sinica, 2024, 52(12): 4037-4047. (in Chinese)

[2] 张若愚, 聂婕, 宋宁, 等. 基于布局化-语义联合表征遥感图文检索方法[J]. 北京航空航天大学学报, 2024, 50(2): 671-683.
ZHANG R Y, NIE J, SONG N, et al. Remote sensing image-text retrieval based on layout semantic joint representation[J]. Journal of Beijing University of Aeronautics and Astronautics, 2024, 50(2): 671-683. (in Chinese)

[3] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. Cambridge: PMLR, 2021: 8748-8763.

[4] LIU F, CHEN D L, GUAN Z, et al. RemoteCLIP: A vision language foundation model for remote sensing[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 5622216.

[5] WANG Z C, PRABHA R, HUANG T Y, et al. SkyScript: A large and semantically diverse vision-language dataset for remote sensing[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(6): 5805-5813.

[6] ZHANG Z L, ZHAO T C, GUO Y L, et al. RS5M and GeoRSCLIP: A large-scale vision- language dataset and a large vision-language model for remote sensing[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 5642123.

[7] WANG Y D, GHAMISI P. RSAdapter: Adapting multimodal models for remote sensing visual question answering[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 5628313.

[8] KIM S, JEONG B, KIM D, et al. Efficient and versatile robust fine-tuning of zero-shot models[M]//Computer Vision-ECCV 2024. Cham: Springer Nature Switzerland, 2024: 440-458.

[9] LI X, LIAN D Z, LU Z H, et al. GraphAdapter: Tuning vision-language models with dual knowledge graph[C]//Proceedings of the 37th Conference on Neural Information Processing Systems. San Diego: NeurIPS, 2023: 13448-13466.

[10] MOUGHNIEH H, CHALHOUB M, NASRALLAH H, et al. Efficient adaptation for remote sensing visual grounding[C]//Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. Piscataway: IEEE, 2025: 1-5.

[11] HUANG T. Efficient remote sensing with harmonized transfer learning and modality alignment[C]//Proceedings of the International Conference on Learning Representations (Workshop). Washington: ICLR, 2024: 1-14.

[12] HWANG S, LAHOTI A S, PUDUPPULLY R, et al. Hydra: Bidirectional state space models through generalized matrix mixers[C]//Advances in Neural Information Processing Systems 37. San Diego: NeurIPS, 2024: 110876-110908.

[13] XIAO Y, SONG L, HUANG S, et al. Mambatree: Tree topology is all you need in state space model[C]//Proceedings of the 38th International Conference on Neural Information Processing Systems 37. San Diego: NeurIPS, 2024: 75329-75354.

[14] ZHOU K Y, YANG J K, LOY C C, et al. Learning to prompt for vision-language models[J]. International Journal of Computer Vision, 2022, 130(9): 2337-2348.

[15] JIA M L, TANG L M, CHEN B C, et al. Visual prompt tuning[M]//Computer Vision - ECCV 2022. Cham: Springer Nature Switzerland, 2022: 709-727.

[16] GU A, DAO T. Mamba: Linear-time sequence modeling with selective state spaces[EB/OL]. (2024-05-31) [2025-04-18]. https://arXiv.org/abs/2312.00752.

[17] HOULSBY N, GIURGIU A, JASTRZEBSKIS, et al. Parameter-efficient transfer learning for NLP[C]//International Conference on Machine Learning. Cambridge: PMLR, 2019: 2790-2799.

[18] GAO P, GENG S J, ZHANG R R, et al. CLIP-adapter:

Better vision-language models with feature adapters[J]. International Journal of Computer Vision, 2024, 132(2): 581-595.

[19] CHEN S F, GE C J, TONG Z, et al. Adaptformer: Adapting vision transformers for scalable visual recognition[C]// Proceedings of the 36th International Conference on Neural Information Processing Systems. Cambridge: PMLR, 2022: 16664-16678.

[20] DOSOVITSKIY A, BEYER L, KOLESNIKO-V A, et al. An image is worth 16x16 words:Transformers for image recognition at scale[EB/OL]. (2021-06-03) [2025-04-18]. https://arxiv.org/abs/2010.11929.

[21] JIANG H J, ZHANG J K, HUANG R, et al. Cross-modal adapter for vision-language retrieval[J]. Pattern Recognition, 2025, 159: 111144.

[22] LU H, HUO Y, YANG G, et al. Uniadapter: Unified parameter-efficient transfer learning for cross-modal modeling[EB/OL]. (2023-05-21) [2025-04-18]. https://doi.org/10.48550/arXiv.23-02.06605.

[23] YUAN Y, ZHAN Y, XIONG Z T. Parameter-efficient transfer learning for remote sensing image-text retrieval[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5619014.

[24] PAN J C, MA Q, BAI C. A prior instruction representation framework for remote sensing image-text retrieval[C]// Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 611-620.

[25] YANG J, LI S Y, ZHAO M Q. Parameter-efficient reparameterization tuning for remote sensing image-text retrieval[J]. IEEE Transactions on Geoscience and Remote Sensing, 2025, 63: 4702315.

[26] LU X Q, WANG B Q, ZHENG X T, et al. Exploring models and data for remote sensing image caption generation[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(4): 2183-2195.

[27] YUAN Z Q, ZHANG W K, FU K, et al. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 4404119.

[28] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: Visual explanations from deepnetworks via gradient-based localization[C]//Proeedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 618-626.

作者简介

杜文亮 男,1989年2月出生于江苏省徐州市. 中国矿业大学计算机科学与技术学院/人工智能学院讲师. 主要研究方向为遥感图像处理、人工智能、机器学习、模式识别、医学图像处理等方面理论与应用.

E-mail: wldu@cumt.edu.cn

许晓宇 男,2001年11月出生于北京市. 中国矿业大学计算机科学与技术学院/人工智能学院硕士研究生. 主要研究方向为多模态学习和计算机视觉.

E-mail: xiaoyuxu@cumt.edu.cn

赵佳琦 男,1988年5月出生于江苏省徐州市. 中国矿业大学计算机科学与技术学院/人工智能学院副教授、硕士生导师. 主要研究方向为人工智能、计算机视觉、无人驾驶等方面理论与应用.

E-mail: jiaqizhao@cumt.edu.cn

刘 兵 男,1981年8月出生于河南省永城市. 中国矿业大学计算机科学与技术学院/人工智能学院副教授. 主要研究方向为人工智能、图像处理和矿山智能化等领域的理论与应用.

E-mail: liubing@cumt.edu.cn

周 勇 男,1974年9月出生于江苏省徐州市. 中国矿业大学计算机科学与技术学院/人工智能学院教授、博士生导师. 主要研究方向为机器学习、人工智能、数据科学与工程等方面的理论与应用.

E-mail: yzhou@cumt.edu.cn