遥感跨模态图文检索：关键技术与挑战

王懿婧[1]，唐旭[1]，韩硕[1]，杜瑞琦[1]

1.西安电子科技大学 人工智能学院 西安 710126

摘 要:遥感跨模态图文检索作为连接自然语言与遥感影像的桥梁，旨在构建高效的双向语义关联，是遥感数据智能化分析的关键技术。本文全面概述了遥感跨模态图文检索领域的技术演进与研究现状。首先，细致分析了主流基准数据集在规模、场景类别及文本标注质量方面的特点，并介绍了通用的评价指标体系，为后续方法研究奠定基础。其次，回顾了文本特征表示从传统统计到深度学习，以及遥感影像特征表示从手工特征到深度神经网络的技术突破。再次，以是否采用跨模态预训练模型为划分，深入剖析了基于非跨模态预训练和基于跨模态预训练方法的原理及模型特点，并通过在三个主流数据集上的实验对比，揭示了跨模态预训练方法的性能优势及其不同微调策略的数据适配规律。最后，本文总结了当前研究面临的细粒度语义对齐、多源数据融合与跨域泛化、以及时序动态匹配机制缺失等核心挑战，并展望了未来在细粒度特征增强、多源异构数据协同建模和时序感知对齐机制和开发等方面的研究方向，以期推动遥感跨模态图文检索技术在实际应用中的深入发展。

关键词:遥感影像；跨模态检索；图文关系建模；深度学习；预训练模型；语义对齐；特征表示；跨模态预训练模型

中图分类号: TP701

引用格式：王懿婧，唐旭，韩硕，杜瑞琦. 遥感跨模态图文检索：关键技术与挑战.遥感学报.

收稿日期: 2025-10-15;

基金项目: 国家自然科学基金项目（面上项目，编号: 62571387）

第一作者简介: 王懿婧，1998 年生, 女, 博士研究生，研究方向为遥感跨模态图文检索的理论和应用。E-mail: yijingwang@stu.xidian.edu.cn

通信作者简介： 唐旭，1985 年生, 男, 教授，研究方向为遥感影像解译的理论和应用。E-mail: tangxu128@gmail.com

1 引 言

随着遥感技术的飞速发展与应用场景的持续拓展，海量遥感数据的高效检索与智能利用成为遥感信息处理领域的核心需求。遥感跨模态图文检索（Remote Sensing Cross-modal Image-text Retrieval，RS-CMITR）作为连接视觉感知与语言理解的关键技术，其任务定义为学习自然语言描述与遥感影像内容间的语义关联，借助于跨模态图文建模网络构建图文信息的映射桥梁，实现文本检索影像（Text to Image Retrieval，T2I）与影像检索文本（Image to Text Retrieval，I2T）的双向高效交互(Ma 等，2015; Xiao 等，2023)。具体如图 1 所示，T2I 任务要求系统依据自然语言查询从遥感影像数据中匹配语义相符的影像，I2T 任务则需根据给定影像检索最相关的文本描述，二者共同致力于打破模态壁垒，实现对遥感数据的精准语义定位。

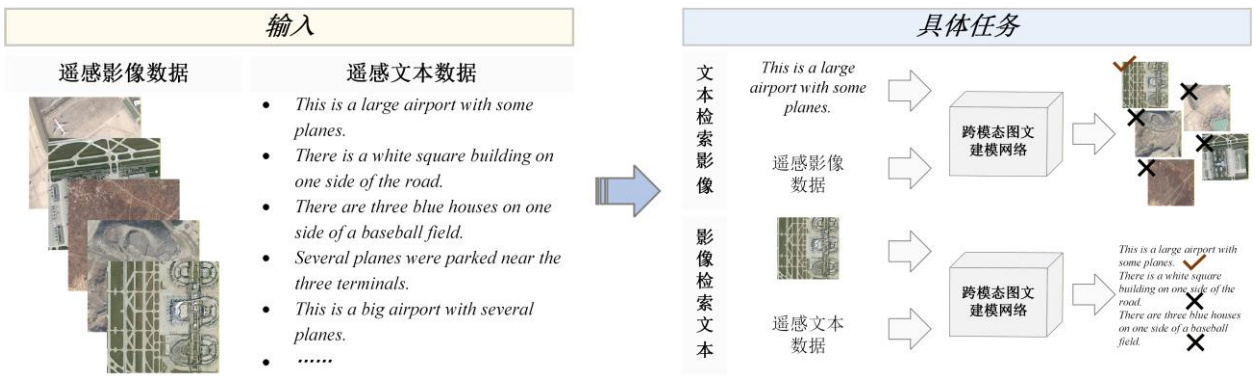

图 1 遥感跨模态图文检索示意图

Fig. 1 Remote sensing cross-modal image-text retrieval pipeline

RS-CMITR 技术的研究与突破具有重要的理论价值与现实意义(Wang 等，2024; Tang 等，2023)。在理论层面，其为解决跨模态语义对齐这一经典难题提供了遥感领域的独特视角，推动了计算机视觉与自然语言处理跨学科融合的深度发展(Tang 等，2024)；在应用层面，自然语言的灵活性与高层语义抽象能力显著提升了遥感数据的可访问性，为灾害检测中地震后房屋损毁区域的快速定位、环境监测里湿地生态系统变化的精准追踪、城市规划中商业区扩张范围的智能识别等场景提供了高效解决方案，成为海量遥感数据智能化分析的核心支撑(Yuan, Zhan, and Xiong 2023; Yuan 等，2021)。从技术发展脉络来看，RS-CMITR 经历了从传统方法到深度学习驱动的范式转型。早期研究依赖人工特征工程，文本特征表示学多采用词袋模型（Bag-of-Words, BoW）(Tsai 2012)、潜在语义分析（Latent Semantic Analysis, LSA）(Sumbul 等，2019)等统计方法，影像特征则依赖 SIFT(Wu 等，2013)、HOG(Pang 等，2011)等手工设计算子，但这类方法难以

收稿日期: 2025-10-15;
基金项目: 国家自然科学基金项目（面上项目，编号: 62571387）
第一作者简介: 王懿婧，1998 年生, 女, 博士研究生，研究方向为遥感跨模态图文检索的理论和应用。E-mail: yijingwang@stu.xidian.edu.cn
通信作者简介: 唐旭，1985 年生, 男, 教授，研究方向为遥感影像解译的理论和应用。E-mail: tangxu128@gmail.com

捕捉深层语义关联与复杂空间关系，检索精度受限。随着深度学习的兴起(Woo 等，2018; He 等，2016)，基于神经网络的表示学习成为主流：文本端从 Word2Vec 等静态词嵌入演进至 BERT(Koroteev 2021)等预训练语言模型，实现了语义信息的深度挖掘；影像端则通过卷积神经网络（Convolution Neural Network，CNN）(Li 等，2021)提取局部特征，再到视觉 Transformer（Vision Transformer，ViT）(Vaswani 等，2017)、Mamba(Liu, Tian, 等，2024)等架构突破局部性限制，完成全局语义建模。近年来，学习范式进一步分化，形成了基于非跨模态预训练模型(Wang 等，2024)（双编码器、融合编码器架构）与基于跨模态预训练模型（全学习、提示学习、适配器学习微调）(Yuan, Zhan, and Xiong 2023; Sun 等，2025)两大技术路线，后者凭借对大规模数据知识的有效利用，展现出更优的检索性能。

为系统梳理 RS-CMITR 领域的研究进展，本文将围绕核心技术与发展挑战展开全面论述。首先，阐述现有的典型遥感跨模态图文检索数据集 RSICD(Lu 等，2017)、RSITMD(Yuan, Zhang, Fu, 等，2022)、UCM-Captions(Qu 等，2016)与评价指标；再次，从基础表示学习入手，分别阐述遥感文本与影像的关键编码方法，对比传统方法与深度学习方法的技术特点及优劣；其次，以是否基于跨模态预训练为分类标准，深入解析两类学习范式的技术原理，结合典型模型案例展开细节论述，并基于三大常用数据集的实验结果，量化对比不同方法的检索性能，揭示技术发展规律；最后，总结当前研究面临的核心挑战，为未来研究方向提供参考。

2 遥感跨模态图文检索数据集与评价指标

RS-CMITR 作为连接遥感图像与自然语言理解的桥梁，旨在克服图像像素信息与人类高级语义之间的语义鸿沟。高质量的图文数据集是推动该领域发展的核心，其规模、场景多样性及文本标注的精细度直接影响模型性能。本文将重点介绍 RS-CMITR 领域的三个里程碑式数据集：UCM-Captions、RSICD 和 RSITMD。UCM-Captions 为早期探索提供了平台；RSICD 以其规模优势，推动了遥感场景描述的发展，但也暴露了文本多样性不足的局限；而 RSITMD 则通过高细粒度、多样的文本描述和关键词，成为高精度图文匹配算法的黄金标准。此外，本文还将概述 RS-CMITR 任务中常用的评价指标，包括召回率和平均召回率，以全面衡量模型的检索性能。

2.1 典型数据集介绍

RS-CMITR 是连接计算机视觉与自然语言处理的关键技术，其核心挑战在于弥合遥感图像的复杂像素

信息与人类高级语义理解之间的“语义鸿沟”。为了推动这一领域的发展，构建高质量的图文数据集至关重要。一个理想的数据集不仅需要具备大规模的图像覆盖和多样的场景类别，更关键的是其文本标注的质量——描述是否精确、细节是否丰富、内容是否多样。这些数据集是训练和验证深度学习模型的基础，其质量直接决定了模型能否准确理解并匹配复杂的图文对应关系。在众多数据集中，UCM-Captions、RSICD 和 RSITMD 是三个里程碑式的成果，其中 RSICD 和 RSITMD 分别代表了遥感图文数据集在“广度”和“深度”两个维度的探索，反映了该领域从通用标注向精细化检索的演进，对应数据集实例如图 2 所示。

图 2 遥感跨模态图文检索数据集实例

Fig. 2 Remote Sensing Cross-modal Image-Text Retrieval dataset examples

UCM-Captions 数据集：该数据集源自 Qu 等人提出的 UC Merced Land Use dataset，是一个专为遥感影像描述和跨模态检索任务设计的数据集，共包含 2,100 张航拍图像，覆盖了 21 个不同的地物类别，如表 1 所示。每张图像都配有 5 条描述性文字，总计 10500 个描述文本。图像的分辨率为 256×256 像素，空间分辨率为 1 英尺，这为 RS-CMITR 模型学习图像与文本之间的细粒度对应关系提供了统一且清晰的视觉信息。

RSICD 数据集：该数据集是为遥感图像标注任务设计的开创性大规模数据集。它包含了 10,921 张图像，覆盖了工业区、体育场、住宅区等 30 个核心场景类别，如表 2 所示。该数据集在提出时极大地丰富了领域内可用的数据资源，其主要目标是训练模型为遥感图像生成通顺、合理的自然语言描述。为此，数据集为每张图提供了 1 到 5 句描述，并通过复制等方式扩充至约 5.4 万句。然而，RSICD 的这一设计也带来了其核心局限性：文本描述的同质化严重，多样性不足（文本多样性分数量化为 1.67）(Yuan, Zhang, Fu, 等, 2022)。这意味着大量内容相似的图像，尤其是同一类别下的图像，往往被赋予了完全相同或高度相似的描述。这种“一对多”的模糊对应关系，虽然对生成通用描述影响不大，但对于需要精确区分相似场景的图文检索任务而言，则是一个致命缺陷，导致模型难以学习到细微的视觉差异。

表 1 UCM-Captions 数据集类别及影像数量

Table 1 UCM-Captions Dataset Categories and Counts

类别	数量	类别	数量	类别	数量
农田	100	森林	100	立交桥	100
飞机	100	高速公路	100	停车场	100
棒球场	100	高尔夫球场	100	河	100
海滩	100	港口	100	跑道	100
建筑物	100	十字路口	100	稀疏住宅区	100
灌木丛	100	中密度住宅区	100	储罐区	100
密集住宅区	100	移动式家庭公园	100	网球场	100
	总计			2,100	

表 2 RSICD 数据集类别及影像数量

Table 2 RSICD Dataset Categories and Counts

类别	数量	类别	数量	类别	数量
机场	420	农田	370	游乐场	1031
裸地	310	森林	250	池塘	420
棒球场	276	工业区	390	高架桥	420
海滩	400	草地	280	港口	389
桥梁	459	中密度住宅区	290	火车站	260
中心区	260	山脉	340	度假村	290
教堂	240	公园	350	河流	410
商业区	350	停车场	390	稀疏住宅区	300
密集住宅区	410	学校	300	储罐区	396
沙漠	300	广场	330	体育场	290
	总计			10,921	

RSITMD 数据集：为解决上述问题，RSITMD 应运而生，它是一个专为高精度 RS-CMITR 任务量身打造的高细粒度数据集。尽管其图像规模（4,743 张，32 类，如表 3 所示）小于 RSICD，但其核心优势在于无与伦比的文本质量。与 RSICD 形成鲜明对比，RSITMD 的文本多样性分数高达 4.60(Yuan, Zhang, Fu, 等，2022)。这一成就源于其严苛的标注流程：标注者被要求详细描述目标的具体属性（如颜色、尺寸、数量）和精确的空间邻接关系，并鼓励使用多样的词汇(Yuan, Zhang, Fu, 等，2022)。更重要的是，RSITMD 引入了创新的双重标注体系：每张图像不仅配有 5 句不同的完整描述句，还额外提供了 1 到 5 个高度精炼的关键词。这一设计使得 RSITMD 能够同时支持基于复杂场景描述的句子级检索和基于特定目标的关键词检索，极大地提升了其在精细化检索任务中的适用性和挑战性。

总而言之，RSICD、RSITMD 和 UCM-Captions 共同构成了遥感图文研究领域的重要基石，它们服务于不同的目标，体现了数据集构建理念的演进。UCM-Captions 作为早期结合图像与文本的数据集，为这一新兴方向提供了初步探索的平台；RSICD 凭借其规模优势，为 RS-CMITR 领域的发展铺平了道路，是理解跨模态遥感场景描述的宝贵资源；而 RSITMD 则聚焦于质量和精度，通过提供高细粒度、高多样性的文本描

105 述与关键词，精准地解决了现有数据集在检索任务上的短板，成为了评估和推动高精度图文匹配算法发展
106 的黄金标准。

107 **表 3 RSITMD 数据集类别及影像数量**

108 **Table 3 RSITMD Dataset Categories and Counts**

类别	数量	类别	数量	类别	数量	类别	数量
工业区	207	港口	178	海滩	154	沙漠	126
体育场	201	农田	177	商业区	148	森林	116
储罐区	201	度假村	166	中心区	146	火车站	112
广场	196	学校	165	停车场	143	山脉	109
游乐场	191	公园	160	机场	140	棒球场	105
河流	188	密集住宅区	157	教堂	136	十字路口	67
高架桥	188	稀疏住宅区	157	中密度住宅区	130	裸地	58
池塘	184	桥梁	155	草地	127	船只	55
总计					4,743		

109 ## 2.2 评价指标介绍

110 在 RS-CMITR 任务中，常用的评价指标是召回率(Recall，$\mathbf{R}@\mathrm{k}$)和平均召回率(Mean Recall, $m\mathbf{R}$)，这些
111 指标通常用于衡量检索结果中相关项被正确检索到的比例以及相关项的排序位置。RS-CMITR 通常涉及两
112 个检索方向：I2T 和 T2I。

113 $\mathbf{R}@\mathrm{k}$ 表示在检索结果的前 K 个中包含至少一个相关项的查询所占的比例。K 通常取 1、5、10，即 $\mathbf{R}@1$，
114 $\mathbf{R}@5$ 和 $\mathbf{R}@10$。对于一个查询，如果其对应的正确匹配项出现在检索结果的前 K 位，则该查询被认为是成
115 功的。

116 $$R@K = \frac{\text{在检索结果前}K\text{位找到相关匹配的查询数量}}{\text{总查询数量}}$$

117 因此，RS-CMITR 的评价指标包括：

118 ■ $\mathbf{R}@1$、$\mathbf{R}@5$ 和 $\mathbf{R}@10$ (I2T)：以图像检索文本时，正确文本出现在检索结果第一、五和十位的比例；

119 ■ $\mathbf{R}@1$、$\mathbf{R}@5$ 和 $\mathbf{R}@10$ (T2I)：以文本检索图像时，正确图像出现在检索结果第一、五和十位的比例；

120 平均召回率为以上的平均值。

121 $$m\mathbf{R} = \frac{\overbrace{\mathbf{R}@1+\mathbf{R}@5+\mathbf{R}@10}^{I2T} + \overbrace{\mathbf{R}@1+\mathbf{R}@5+\mathbf{R}@10}^{T2I}}{6}$$

122 通常，RS-CMITR 的综合性能会通过，$\mathbf{R}@1$，$\mathbf{R}@5$，$\mathbf{R}@10$ (I2T)，$\mathbf{R}@1$，$\mathbf{R}@5$，$\mathbf{R}@10$ (T2I)和 $m\mathbf{R}$
123 来共同衡量。

3 特征表示方法

RS-CMITR 的核心目标是建立自然语言描述与遥感影像之间的精准语义关联，实现二者的高效双向检索。这一目标的达成，高度依赖于对两种异构模态信息的有效特征提取与表示，为后续跨模态对齐与相似度计算奠定基础。本节将系统梳理 RS-CMITR 领域的核心表示技术：首先聚焦文本特征表示方法，对比传统统计方法与深度学习方法的技术原理、应用场景及局限性，阐述从手工词嵌入到预训练语言模型的技术进展；其次深入影像特征表示学技术，阐述 CNN 在局部特征提取中的核心作用与改进策略，以及 ViT、Mamba 等新型架构对全局语义建模的突破，同时探讨“CNN+全局模型”混合架构的技术优势；通过对两类模态表示方法的全面解析，揭示 RS-CMITR 从“特征匹配”到“语义对齐”的技术升级路径。

3.1 文本特征表示方法

遥感跨模态图文检索的核心目标是通过语义关联实现遥感影像与文本描述的高效互查，而文本特征表示学学习作为其中的关键技术，直接决定了文本语义特征的提取质量和跨模态对齐的精度。遥感领域文本数据通常包含地理信息、地物属性等专业信息，且与高分辨率遥感影像之间存在复杂的空间和语义关联，这对文本特征表示学方法提出了更高要求。传统方法依赖人工特征工程，难以捕捉文本的深层语义和上下文关联；而基于深度学习的方法通过端到端学习，能够自动挖掘文本中的隐含模式和多尺度语义信息，显著提升了跨模态检索的鲁棒性和泛化能力。尤其在处理遥感领域特有的长尾分布（如罕见地物类别）和多语言描述时，有效的文本特征表示学模型能够通过嵌入空间映射，弥合文本与影像之间的模态鸿沟。因此，文本特征表示学方法的演进不仅是自然语言处理技术的进步，更是推动遥感跨模态检索从理论走向应用的关键驱动力(Zhu 等，2024)。

早期传统的文本特征表示学方法基于统计学实现。其中，词袋模型（Bag-of-Words, BoW）(Tsai 2012)通过统计文本中词汇的出现频率构建向量表示，结合文本融合加权策略突出关键词的重要性(Sumbul 等，2019)。LSA 通过奇异值分解将高维词频矩阵降维至潜在语义空间，解决了 BoW 的同义词和多义词问题。此外，主题模型如隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）(Blei, Ng, and Jordan 2003)提取隐含主题分布为主。然而，这些方法存在显著局限：其一，离散表示导致语义鸿沟，无法量化"河流宽度"与"水道直径"的相似性；其二，依赖人工构建的词典难以适应遥感领域新术语；其三，缺乏对词序和语法的建模，无法区分"农田环绕居民区"与"居民区环绕农田"的空间关系差异，制约了复杂查询场景下的检索精度。

随着深度学习的发展，基于神经网络的文本特征表示学方法在文本解译中展现出突破性进展。

Word2Vec(Church 2017)和 GloVe(Pennington, Socher, and Manning 2014)等静态词嵌入模型通过分布式表示将词汇映射到连续向量空间，显著提升了语义相关性计算能力。为进一步捕捉上下文动态特征，Bi-LSTM(Shahid, Zameer, and Muneeb 2020)和 Transformer(Vaswani 等，2017)架构被引入，显著提升文本信息解译性能(Cao 等，2022)。预训练语言模型如 BERT(Koroteev 2021)通过掩码语言建模和自注意力机制，实现了文本的深度语义理解。值得关注的是，多模态联合嵌入技术（如 CLIP 的遥感变体 RS-CLIP(Li 等，2023)通过对比学习将文本与影像嵌入到统一空间，实现了文本信息与遥感视觉信息的内容建模。

3.2 遥感影像特征表示方法

影像特征表示学习是遥感跨模态图文检索系统的另一个核心环节，其目标是将高维、异构的遥感影像转化为低维、语义可解释的特征向量，从而实现影像与文本模态的语义对齐。遥感影像具有空间分辨率高、光谱维度多、地物覆盖复杂等特点，传统基于手工设计特征（如 SIFT(Wu 等，2013)、HOG(Pang 等，2011)）的方法难以捕捉其深层语义信息。随着深度学习技术的发展，基于神经网络的表示学习方法能够通过分层特征提取机制，从像素级信息中挖掘局部细节和全局上下文关联。在跨模态检索任务中，有效的影像特征表示学需满足两方面需求：一方面需要保留局部地物特征（如建筑物边缘、道路纹理），另一方面需建模全局空间分布（如城市布局、农田分区）。

CNN(Li 等，2021)凭借其局部感受野和权值共享机制，成为提取遥感影像局部视觉特征的主流方法。经典模型如 ResNet(He 等，2016)、VGGNet(Vedaldi and Zisserman 2016)通过堆叠卷积层与池化层，能够逐级提取边缘、纹理、形状等低层特征到地物类别、空间组合等高层语义特征。针对遥感影像特点，研究者提出了多项改进策略：引入注意力机制增强关键区域响应，如通过通道和空间注意力动态调整特征权重(Woo 等，2018)；设计多尺度特征融合架构，如特征金字塔网络融合浅层特征与深层语义特征(Ghiasi, Lin, and Le 2019)；结合光谱特性改进卷积操作，如 3D-CNN(Alakwaa, Nassef, and Badr 2017)同时处理空间-光谱维度信息。然而，CNN 的局部归纳偏置也导致其难以建模长距离依赖关系，这对大范围地物（如河流走向、山脉延展）的表示构成挑战。

为突破 CNN 的局部性限制，视觉 ViT(Vaswani 等，2017)和 Mamba(Liu, Tian, 等，2024)等新型架构被引入遥感影像全局特征提取。ViT 将影像划分为序列化影像块，通过自注意力机制建立任意位置间的关联。针对遥感数据特点，GeoViT(Khirwar and Narang 2024)等改进模型通过加入地理坐标编码、多时相位置嵌入增强空间时序建模能力。而新兴的 Mamba 模型基于状态空间理论，采用选择性扫描机制实现线性复杂度下的长序列建模，在 512×512 像素的影像解译中相比 ViT 减少 37%的计算量。值得注意的是，全局模型常与

CNN 构成混合架构（如 ConvNeXt(Woo 等，2023)），通过 CNN 提取局部特征后输入 Transformer 进行全局关系推理，这类方法正逐步成为 RS-CMITR 影像特征表示学学习的新范式。

4 遥感跨模态图文检索方法与模型

上述的 RS-CMITR 公开数据集与文本/遥感影像的特征表示方法被广泛应用于基于深度学习的遥感跨模态图文检索研究领域，并为相关算法的性能验证与技术迭代提供了坚实的数据支撑。本节依据方法训练逻辑与架构设计的核心差异，从基于非跨模态预训练（细分为双编码器架构、融合编码器架构）和基于跨模态预训练（细分为全学习微调、提示学习、适配器学习）两大核心方向，对深度学习驱动的 RS-CMITR 算法展开系统总结与深入分析，同时结合多数据集上的实验结果（以 R@K、mR 为关键评估指标），揭示不同方法在检索精度、效率及场景适配性上的优势与特性。

现有的适用于 RS-CMITR 任务的方法主要分为两大类：基于非跨模态预训练的方法和基于跨模态预训练的方法。在非跨模态预训练方法中，遥感影像和文本分别送入已有的单模态编码器：影像编码器通常在 ImageNet 等自然影像数据集上预训练，文本编码器则在通用语料库上预训练。之后，再利用遥感图文配对数据，对两个编码器进行联合训练，使其在遥感场景下学习到一定的跨模态对齐能力；在跨模态预训练方法中，首先利用公开(Kim, Son, and Kim 2021)以及网络图文数据等大规模通用图文数据集(Radford 等，2021)对图像编码器和文本编码器进行跨模态预训练，使二者在大规模数据上学习到更强的视觉－语言对齐表征。随后，将得到的跨模态预训练模型迁移到遥感领域，即利用遥感影像和文本数据进行微调，以获得适应遥感场景的图文表示，其对比图与面向遥感设计的 RS-CMITR 任务方法发展流程如图 3 与图 4 所示。

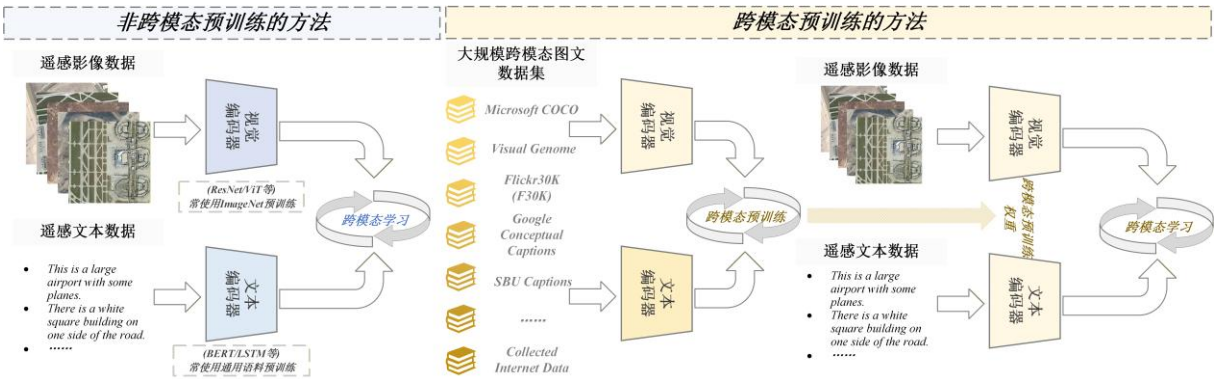

图 3 基于非跨模态预训练的方法和基于跨模态预训练的方法对比图

Fig. 3 Comparison between non-cross-modal pre-training-based methods and cross-modal pre-training-based methods.

图 4 RS-CMITR 任务方法演进时间轴

Fig. 4 Evolution timeline of RS-CMITR task methods

4.1 基于非跨模态预训练的方法

基于非跨模态预训练的方法利用公开的 RS-CMITR 数据集的训练集训练模型，并利用对应的测试集评估模型的检索精度。根据模型架构，可将该方法分为两大类：基于双编码器架构和基于融合编码器架构的检索模型。采用双编码器架构的模型首先采用两个独立的编码器分别抽取遥感影像和文本特征，然后模型利用一些常规的相似度度量计算跨模态特征之间的相似度，以实现 RS-CMITR 任务。这类方法因其简单的思路和高效的检索过程而广泛应用于 RS-CMITR 任务中。Yuan 等人(Yuan 等，2021)提出了一种名为轻量化多尺度跨模态图文检索（Lightweight Multi-scale Crossmodal Text-image Retrieval，LW-MCR）模型。开发了一个双线性融合方案和一个知识蒸馏策略，以同时提高 RS-CMITR 性能并降低时间成本。考虑到遥感影像的内容复杂的特征，基于全局和局部信息的 RS-CMITR 模型（RS-CMITR Framework based on Global and Local information，GaLR）(Yuan, Zhang, Tian, 等，2022)通过增强视觉编码器表征遥感影像的能力的方式以确保其在 CMRSITR 任务上的性能。该模型首先采用图卷积网络（Graph Convolutional Network，GCN）(Kipf 2016)以探索遥感影像中的各种对象及其之间的关系；之后，该模型设计了多层级视觉自注意力模块（Multi-level Visual Self-attention，MVSA）(Yuan, Zhang, Fu, 等，2022)以充分的抽取遥感影像的多尺度信息；最后，通过整合错综复杂的对象级相关性和多尺度线索，GaLR 模型获得了较为完备的视觉特征，以更好的完成 RS-CMITR 任务。

采用基于融合编码器架构的模型通常涉及复杂的跨模态交互方案，以充分挖掘遥感影像和文本之间的关系，从而保证 RS-CMITR 任务的准确性。这类方法通常直接通过模型得到遥感影像和文本之间的匹配分数。语义对齐模型（Semantic Alignment Module，SAM）(Cheng 等，2021)采用传统的卷积神经网络（Convolutional Neural Network，CNN）和循环神经网络（Recurrent Neural Network，RNN）学习遥感影像和文本的特征；之后，SAM 引入语义对齐模块以探索跨模态关系，并通过互补方案增强特征。通过这些设计，SAM 能够获得较好的遥感影像-文本的检索能力。鉴于遥感影像的复杂性，除了跨模态关系外，研究人员注重对视觉特征的学习。例如，Yuan 等人(Yuan, Zhang, Fu, 等，2022)提出了适用于 RS-CMITR 任务的

非对称多模态特征匹配网络（Asymmetric Multimodal Feature Matching Network，AMFMN）。该模型设计了一个多尺度特征融合模块以探索遥感影像中的复杂内容；此外，其还开发了一个视觉引导的注意力模块以对齐跨模态语义并改进最终的检索性能。基于 AMFMN 模型的架构，跨模态遥感特征匹配网络（Cross-modal Remote Sensing Feature Matching Network，CMFM-Net）(Yu 等，2022)采用 GCN 网络以更好的建模遥感影像中对象之间的复杂关系，并利用 GCN 网络的架构增强不同模态的表征。为了进一步提取遥感影像的多尺度全局特征，增强模型跨模态融合能力，Tang 等人(Tang 等，2024)提出了基于历史经验的遥感视觉语言模型（Prior-experience-based Remote Sensing Vision-language Model，PERSVL）。该模型设计递归的多尺度补充架构，逐层提取高级语义特征缺失的尺度信息；此外，其设计一种双分支融合编码器利用跨模态信息丰富视觉和文本表征，并设计基于注意力机制的动态融合策略以融合跨模态特征；为了减轻数据噪声对模型训练的影响，该模型提出历史经验学习模块，根据模型训练的历史数据生成可靠伪标签，修正错误标签对模型的误导。

4.2 基于跨模态预训练的方法

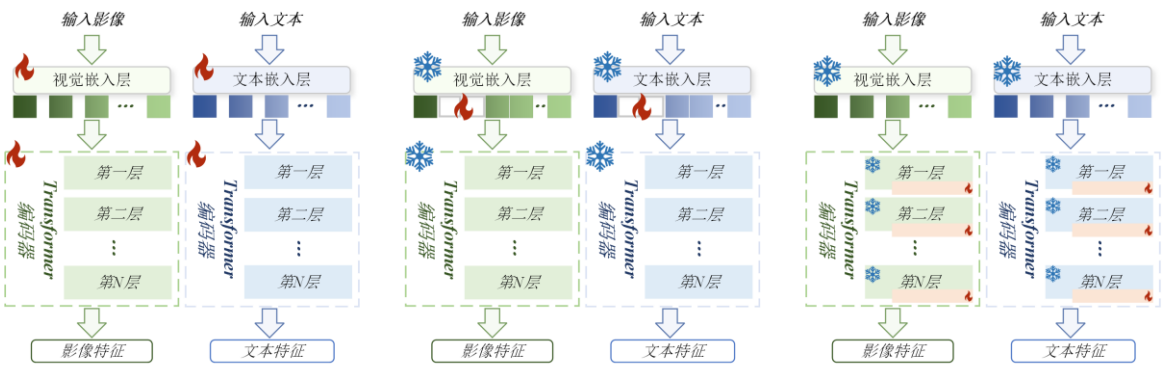

图 5 基于跨模态预训练模型的三种微调策略示意图

Fig. 5 Schematic diagram of three fine-tuning strategies for cross-modal pre-training models

基于跨模态预训练的方法首先利用大量的遥感图像-文本对来预训练模型，使其学习到丰富的跨模态关联知识。随后，模型加载这些预训练参数，并根据具体的 RS-CMITR 任务需求进行微调，以适应目标场景并提升性能。根据微调过程对预训练模型参数的修改程度和方式，基于跨模态预训练的方法可细分为三类策略，如图 5 所示：(a)全学习微调、(b)提示学习和(c)适配器学习。全学习微调的特点是更新预训练模型的所有参数，能够充分利用预训练模型的知识以实现对目标任务的最佳适配，但通常需要较大的计算资源和数据量；提示学习则在冻结大部分预训练模型参数的前提下，通过设计特定的“提示”模板来引导模型完成任务，无需进行大规模参数调整，从而提高样本效率并减少计算成本；而适配器学习的核心思想是在预训

练模型的主体参数被冻结的情况下，插入少量可训练的适配器模块来捕捉特定任务的语义信息，这种方法在保持模型通用性的同时提供了灵活性，具有参数高效、训练速度快的优点(Wang 等，2024)。

全学习微调是一种传统的微调方式，其核心思想是在目标任务上对整个预训练模型的参数进行更新。这种方法能够充分利用预训练模型的知识，并通过端到端的方式实现对目标任务的最佳适配。在遥感跨模态图文检索任务中，全学习微调已被广泛应用。特征交互增强 Transformer 模型（Interacting-enhancing Feature Transformer，IEFT）(Tang 等，2023)将自然多模态场景中预训练的权重迁移到 RS-CMITR 领域，并在全学习微调过程中设计通道级特征增强单元，以丰富视觉表征，改善模型在 RS-CMITR 上的检索精度。Al 等人提出一种多语言 Transformer（Multilanguage Transformer，MLT）模型，其用带有全局平均池化的对比语言-影像预训练（Contrastive Language-Image Pre-training，CLIP）(Radford 等，2021)编码器作为主干网络以完成 RS-CMITR 任务，并利用自然多模态场景预训练参数初始化 CLIP 主干网络。此外，他们在微调过程中提出一种多语言训练策略，以增强模型的泛化能力。与上述两种方法不同，Liu 等人(Liu, Chen, 等，2024)基于 CLIP 架构提出适用于遥感场景的 CLIP（RemoteCLIP）模型，利用大规模遥感多模态数据预训练该模型，以增强模型对遥感影像的特有性质的理解能力。在得到预训练权重后，该方法在 RS-CMITR 任务上微调模型，并取得了较好的检索精度。与 RemoteCLIP 类似，Zhang 等人(Zhang 等，2024)提出 GeoRSCLIP，利用大规模遥感多模态数据对 CLIP 模型进行预训练，并在 RS-CMITR 任务上微调。为了进一步增强模型对遥感图文对的感知能力，该方法对预训练数据集进行了筛选和增强，保障模型在 RS-CMITR 任务上的精度得到进一步的提升。

提示学习是一种新兴的微调方法，旨在通过设计特定的任务提示来引导预训练模型完成目标任务，而无需对模型参数进行大规模调整。具体而言，提示学习通过在输入数据中引入额外的提示模板，将遥感跨模态图文检索任务转化为预训练模型熟悉的下游任务形式。Wang 等人(Wang 等，2024)提出上下文和不确定性感知提示（Context and Uncertainty-aware Prompt，CUP）模型，该模型基于遥感影像上下文信息生成包含丰富遥感影像对象级语义信息的提示，有效促进自然场景预训练知识向遥感领域迁移。为了有效利用跨模态数据的有效信息，Sun 等人(Sun 等，2025)提出了强弱提示工程方法（Strong and Weak Prompt Engineering，SWPE）。该方法设计强和弱提示生成模块，通过注意力机制和预训练的分类模型生成细粒度的全局类别语义提示，之后利用提示引导式功能微调模块优化提示信息，增强细粒度细节和全局语义。

适配器学习是一种轻量级的微调方法，其核心思想是在预训练模型的基础上插入少量可训练的适配器模块。这些适配器模块通常由多层感知机或注意力机制组成，用于捕捉特定任务的语义信息。通过这种方

式，适配器学习能够在保持预训练模型通用性的同时，为目标任务提供足够的灵活性。Yuan 等人(Yuan, Zhan, and Xiong 2023)设计了参数高效的迁移学习（Parameter-efficient Transfer Learning，PETL）方法，其设计一个轻量级的多模态遥感适配器（Multimodal Remote Sensing Adapter，MRS-Adapter），有效的将 CLIP 扩展到 RS-CMITR 任务中。为了进一步解决遥感影像与自然场景影像之间的语义鸿沟问题，Xu 等人(Xu 等，2025)提出一种字典驱动适配器（Dictionary-driven Adapter，DDA）。该方法通过引入共享字典和稀疏编码机制，聚合了遥感数据中的判别性特征，并将其融入到冻结的 CLIP 模型中，从而有效地弥合了跨模态语义差异，显著提升了遥感图文检索任务的性能。在此基础上，Li 等人(Li et al. 2025)基于适配器学习框架提出了细粒度视觉－语言对齐（Fine-Grained Visual-Language Alignment, FGVLA）方法。在冻结 CLIP 预训练参数的前提下，该方法在图像编码器和文本编码器的各个 Transformer 块中嵌入轻量级高效适配器，以实现从自然图像到遥感场景的自适应。在目标函数设计上，FGVLA 构建由粗粒度与细粒度项联合组成的混合损失；推理阶段则进一步引入与细粒度损失协同工作的推理机制，在全局特征相似度的基础上融合图像块－名词级相似度，从而在保证参数效率的同时显著提升遥感图文检索的检索性能。此外，针对遥感影像在频域中具有显著结构模式的特点，Wan 等人(Wan et al. 2025)提出频谱适配器（Frequency-domain-based RS image－text retrieval Adapter, FRS-Adapter）。该方法通过扩展频域感受野提取遥感影像特有的结构信息，并利用单模态滤波器组在多频带上进行频谱压缩，结合幅度与相位特征增强特征表示。同时，构建跨模态频谱互聚合模块，促进语言、空间与频谱信息的深度融合，引导保留与任务相关的频率成分、抑制无关成分，从而提高自然场景到遥感场景的迁移效果。与此不同，Yang 等人(Yang, Li, and Zhao 2025)从重参数化调优的角度出发，构建参数高效重参数化调优与排序－匹配框架（PERT-RaMa），提出基于 Kronecker 积的低秩适配模块，在不引入额外深层结构的前提下获得更高的内在秩表示。同时，作者将 RS-CMITR 任务形式化为“一对一匹配”和“一对多排序”问题，通过去除不必要的计算，显著加快了训练过程，进一步提升了 RS-CMITR 任务中的检索效率与整体性能。

4.3 实验对比与性能分析

本节在三个常用的 RS-CMITR 数据集，即 RSICD、RSITMD 以及 UCM-Captions 数据集，对提及的方法进行实验论证。本节对影像检索文本和文本检索影像两个基本任务均采用 **R**@k 以评估各个方法在每个任务上的检索结果，并采用 ***m*R** 指标衡量各个方法在两个任务上的综合表现。每个指标的最优性能在实验表中以粗体突出显示。

表 4 不同算法在 RSICD 数据集上的检索结果（%），最好的结果用加粗表示

Table 4 Retrieval result(%) of different methods on the RSICD dataset, the best result is indicated in bold.

方法类型		方法	影像检索文本			文本检索影像			mR
			R@1	R@5	R@10	R@1	R@5	R@10	
非跨模态预训练	双编码器	LW-MCR-b	4.57	13.71	20.11	4.02	16.47	28.23	14.52
		LW-MCR-d	3.29	12.52	19.93	4.66	17.51	30.02	14.66
		LW-MCR-u	4.39	13.35	20.29	4.30	18.85	32.34	15.59
		GaLR	6.59	19.85	31.04	4.69	19.48	32.13	18.96
	融合编码器	SAM t-i	5.91	18.59	30.63	4.82	18.10	33.44	18.58
		SAM i-t	6.41	20.10	32.89	5.91	18.08	29.73	18.85
		AMFMN-soft	5.05	14.53	21.57	5.05	19.74	31.04	16.16
		AMFMN-fusion	5.39	15.08	23.40	4.90	18.28	31.44	16.42
		AMFMN-sim	5.21	14.72	21.57	4.08	17.00	30.60	15.53
		CMFM-Net	5.40	18.66	28.55	5.31	18.57	30.03	17.75
		PERSVL	10.25	23.79	36.23	7.65	26.61	41.83	24.39
跨模态预训练	全学习	IEFT	6.26	24.00	39.98	8.11	27.45	43.25	24.84
		CLIP	13.54	30.83	43.46	11.55	33.14	49.83	30.39
		MLT	11.61	30.10	42.17	9.55	29.27	44.73	27.90
		RemoteCLIP	18.39	37.42	51.05	14.73	39.93	56.58	36.35
		GeoRSCLIP	21.13	41.72	55.63	15.59	41.19	57.99	38.87
	提示学习	CUP	9.86	31.60	47.15	12.05	32.66	48.70	30.34
		SWPE	18.66	39.52	53.61	15.33	40.86	57.73	37.62
	适配器学习	PETL	14.13	31.51	44.78	11.63	33.92	50.73	31.12
		GeoRSCLIP +DDA	**22.23**	**44.92**	**57.73**	**17.20**	**42.01**	**59.05**	**40.52**
		FGVLA	16.93	36.23	48.49	12.72	36.61	53.76	34.12
		FRS-Adapter	14.22	31.57	44.93	11.72	34.11	50.96	31.25
		PERT-RaMa	12.97	31.53	45.20	11.47	34.04	51.46	31.11

表 4 报告了多种不同算法在 RSICD 数据集上的检索结果。观察结果可知，采用跨模态预训练范式能够显著提升模型在 RS-CMITR 任务上的检索结果。以 ***m*R** 指标为例，在非跨模态预训练模型中，PERSVL 模型检索结果达到 24.39%，为该类型模型的最优性能。基于跨模态预训练的模型在该指标上的表现均优于 PERSVL。具体而言，IEFT 的性能相较于 PERSVL 提升了 0.45%，CLIP 提升了 6.0%，MLT 提升了 3.51%，RemoteCLIP 提升了 11.96%，GeoRSCLIP 提升了 14.48%，CUP 提升了 5.95%，SWPE 提升了 13.23%，PETL 提升了 6.73%，FGVLA 提升了 9.73%，FRS-Adapter 提升了 6.86%，PERT-RaMa 提升了 6.72%，以及 GeoRSCLIP+DDA 提升了 16.13%。该结果展现了跨模态预训练范式在 RS-CMITR 任务中的巨大潜力。对于跨模态预训练方法，在 GeoRSCLIP 的预训练模型基础上加入 DDA 的微调范式，在 RSICD 数据集上取得了最优的检索结果。该现象表明，合理的微调策略在 RS-CMITR 任务中发挥重要的作用。

表 5 报告了多种不同算法在 RSITMD 数据集上的检索结果。该表中呈现出与表 4 类似的趋势：绝大部分基于跨模态预训练的方法在该数据集上的检索结果显著优于非跨模态预训练的方法。与在 RSICD 数据集上的统治性的表现不同，GeoRSCLIP+DDA 的组合在影像检索文本和文本检索影像的 **R**@1以及 ***m*R** 三个指标上达到了最优值（34.51% / 25.97% / 52.63%）。相较而言，RemoteCLIP 在文本检索影像的 **R**@5上取得 59.51%的最高检索结果，以及 SWPE 在文本检索影像的 **R**@10上取得 75.27%的最高结果。RSITMD 数据集相较于 RSICD 数据集，其数据质量得到了进一步提升。对于 ***m*R** 指标，全学习最优模型 GeoRSCLIP 的性能达到 51.81%，提示学习最优模型 SWPE 的性能达到 50.54%，适配器学习最优模型 GeoRSCLIP+DDA 的性

能达到 52.63%。

表 5 不同算法在 RSITMD 数据集上的检索结果（%），最好的结果用加粗表示

Table 5 Retrieval result(%) of different methods on the RSITMD dataset, the best result is indicated in bold.

方法类型		方法	影像检索文本			文本检索影像			mR
			R@1	R@5	R@10	R@1	R@5	R@10	
非跨模态预训练	双编码器架构	LW-MCR-b	9.07	22.79	38.05	6.11	27.74	49.56	25.55
		LW-MCR-d	10.18	28.98	39.82	7.79	30.18	49.78	27.79
		LW-MCR-u	9.73	26.77	37.61	9.25	34.07	54.03	28.58
		GaLR	14.82	31.64	42.48	11.15	36.68	51.68	31.41
	融合编码器	AMFMN-soft	11.06	25.88	39.82	9.82	33.94	51.90	28.74
		AMFMN-fusion	11.06	29.20	38.72	9.96	34.03	52.96	29.32
		AMFMN-sim	10.63	24.78	41.81	11.51	34.69	54.87	29.72
		CMFM-Net	10.84	28.76	40.04	10.00	32.83	47.21	28.28
		PERSVL	20.58	42.29	54.65	14.87	46.06	65.75	40.70
跨模态预训练	全学习	IEFT	11.10	37.99	58.75	15.48	37.61	51.40	35.38
		CLIP	24.16	47.12	61.28	20.40	50.53	68.54	45.33
		MLT	22.56	43.80	57.30	19.29	51.37	68.18	43.75
		RemoteCLIP	28.76	52.43	63.94	23.76	**59.51**	74.73	50.52
		GeoRSCLIP	32.30	53.32	67.92	25.04	57.88	74.38	51.81
	提示学习	CUP	17.70	48.40	64.82	21.57	43.70	56.75	42.16
		SWPE	27.88	51.76	64.82	25.27	58.23	**75.27**	50.54
	适配器学习	PETL	23.67	44.07	60.36	20.10	50.63	67.97	44.47
		GeoRSCLIP+DDA	**34.51**	**53.32**	**69.47**	**25.97**	57.57	74.96	**52.63**
		FGVLA	26.33	51.99	64.38	21.77	54.12	72.92	48.58
		FRS-Adapter	23.81	44.53	60.47	20.44	50.70	68.25	44.70
		PERT-RaMa	20.66	44.16	58.94	17.95	50.81	69.08	43.60

表 6 报告了多种不同算法在 UCM-Captions 数据集上的检索结果。同样的，基于跨模态预训练的方法显著优于非跨模态预训练的方法。该现象表明，对于数据量较小的 UCM-Captions 数据集，预训练阶段获得的知识对于模型的结果和泛化性的提升至关重要。然而，与 RSICD 和 RSITMD 数据集不同的是，在该数据集上，可学习参数量最少的提示学习的表现显著优于全学习和适配器学习。以 ***m*R** 指标为例，提示学习中表现最优的模型 SWPE 取得 60.41%的结果，高于全学习最优模型 RemoteCLIP 和适配器学习的最优模型 PETL。该结果表明，在小数据上动用过多参数微调，不仅破坏从预训练中获得的知识，也容易陷入过拟合的情况。

为了公平对比不同跨模态预训练范式的效果差异，本文选取了多个基于 CLIP ViT-B/32 预训练模型的代表性方法进行对比分析。具体而言，CLIP 方法代表全训练学习范式，CUP 方法代表提示学习范式，PETL、FGVLA、FRS-Adapter 和 PERT-RaMa 方法代表适配器学习范式。所有方法均，以相同的 CLIP ViT-B/32 预训练权重(Radford 等，2021)作为初始化，并在统一的遥感图文数据集上进行微调从而有效隔离模型架构与预训练数据异质性对实验结果的潜在影响，使不同微调策略的本质效能得以客观评估。实验结果表明，基于 CLIP ViT-B/32 的适配器学习方法在三个数据集上整体优于提示学习方法与全训练学习范式，且在数据质量更高的 RSITMD 数据集上优势更为明显。在适配器学习方法内部，FGVLA 展现出最优的检索性能，显著优于其他三种适配器方法，这表明细粒度视觉-语言对齐机制能够更有效地捕捉遥感影像与文本描述之间的语义关联。值得注意的是，FRS-Adapter、PERT-RaMa 和 PETL 在 RSICD 数据集上的性能较为接近，但在

RSITMD 数据集上性能分化更为明显，这表明在数据质量更高、语义更复杂的数据集上，不同适配器设计策略的有效性差异会被进一步放大。相较于提示学习方法，适配器学习通过引入额外的可学习参数模块，能够在保留预训练知识的同时更灵活地学习任务特定的特征表示，从而在遥感图文检索任务中取得更优的

表 6 不同算法在 UCM-Captions 数据集上的检索结果（%），最好的结果用加粗表示

Table 6 Retrieval result(%) of different methods on the UCM-Captions dataset, the best result is indicated in bold.

方法类型		方法	影像检索文本			文本检索影像			mR
			R@1	R@5	R@10	R@1	R@5	R@10	
基于非跨模态预训练	双编码器架构	LW-MCR-b	12.38	43.81	59.52	12.00	46.38	72.48	41.10
		LW-MCR-d	15.24	51.90	62.86	11.90	50.95	75.24	44.68
		LW-MCR-u	18.10	47.14	63.81	13.14	50.38	79.52	45.35
		GaLR	11.56	46.19	72.25	10.95	42.29	65.62	41.48
	融合编码器	SAM t-i	9.50	47.60	78.60	13.80	55.20	94.80	49.91
		SAM i-t	11.90	47.10	76.20	10.50	47.60	93.80	47.85
		AMFMN-soft	12.86	51.90	66.67	14.19	51.71	78.48	45.97
		AMFMN-fusion	16.67	45.71	68.57	12.86	53.24	79.43	46.08
		AMFMN-sim	14.76	49.52	68.10	13.43	51.81	76.48	45.68
		CMFM-Net	11.90	44.29	66.19	12.19	47.81	76.86	43.21
		PERSVL	19.05	50.95	75.24	18.51	53.62	82.67	50.01
跨模态预训练	全学习	IEFT	15.75	60.56	**91.41**	16.50	57.94	82.54	54.11
		CLIP	17.14	55.24	79.52	13.90	56.95	91.81	52.43
		MLT	19.52	55.23	81.42	17.71	**67.47**	93.52	55.81
		RemoteCLIP	20.48	59.85	83.33	18.67	61.52	94.29	56.36
	提示学习	CUP	15.06	56.98	91.14	15.24	54.76	79.37	52.09
		SWPE	**26.19**	**68.10**	86.19	**22.29**	66.48	93.24	**60.41**
	适配器学习	PETL	22.71	55.81	80.33	18.82	62.84	93.72	55.71
		FGVLA	24.29	64.76	86.67	19.71	64.67	**96.00**	59.35
		FRS-Adapter	22.83	58.90	80.72	19.03	63.07	93.84	56.89
		PERT-RaMa	19.05	59.05	84.38	18.82	64.90	94.82	56.83

性能。为直观展示不同跨模态预训练范式的性能差异，图 6 展示了 PETL 和 CUP 两种方法在 T2I 和 I2T 检索任务上的典型结果对比。在 T2I 任务中，黄色边框标注了与查询文本最匹配的正确图像；在 I2T 任务中，绿色文字标注了与查询图像最匹配的正确文本描述。从 T2I 检索结果来看，两种方法在不同场景下表现各有优劣。在"There are many white planes parked beside the blue house."这一细粒度 T2I 检索中，CUP 方法在首位即检索到正确图像(黄色边框)，而 PETL 方法的正确结果排在第 2 位，这表明提示学习在该特定场景下的匹配结果更高。在"it is a large stadium with numerous bleachers and a soccer field where players are playing on half of it."场景中，两种方法均在首位检索到正确结果(黄色边框)，表现相当。从 I2T 检索结果来看，CUP 方法展现出更稳定的优势。针对学校场景的遥感图像，CUP 在首位检索到的文本描述(绿色标注)准确捕捉了核心语义，而 PETL 的正确描述排在第 2 位。针对住宅区场景，两种方法均在首位检索到正确文本(绿色标注)，表现相当。综合来看，这些可视化结果表明，适配器学习方法和提示学习方法各有适用场景，这种差异可能源于两种方法的学习机制不同。

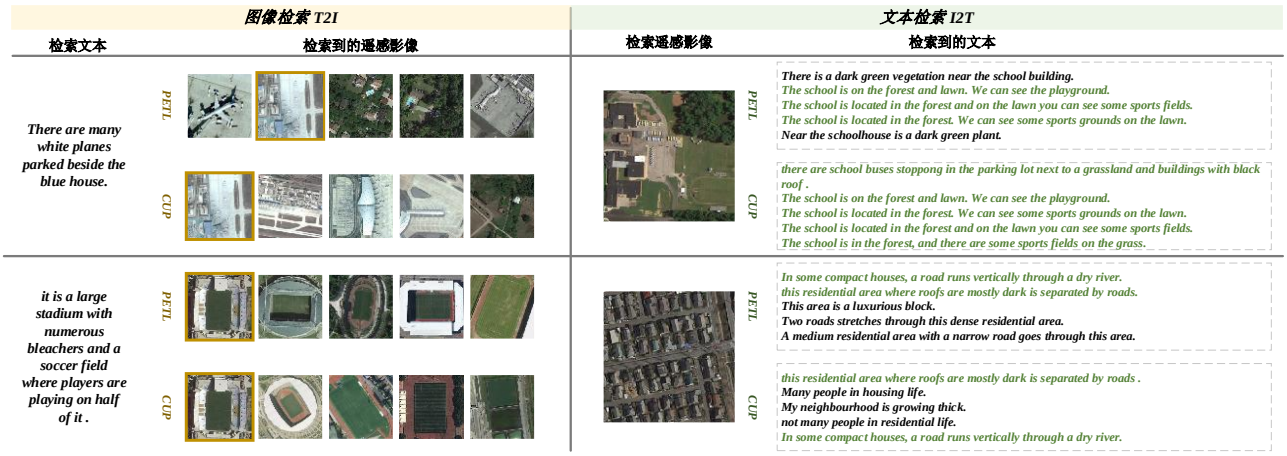

图 6 检索结果对比图

Fig. 6. Comparative illustration of retrieval results.

5 挑战与展望

近些年 RS-CMITR 任务发展迅速，但精度与泛化能力受模型层面、训练策略与数据制约，核心挑战集中于三方面：一是细粒度语义对齐难，影像细节密集、地物组合复杂而文本聚焦核心地物，且现有 CNN、ViT 等难捕捉相似地物细微差异；二是多源数据融合与跨域泛化弱，多源数据异构且专用标注少、融合方式简单，模型跨区域/跨传感器任务性能下滑；三是时序动态匹配机制缺失，现有研究聚焦静态影像，无法将文本时间描述与影像时序特征精准对齐。具体如下：

（1）细粒度语义对齐问题。遥感影像具有像素级细节复杂，地物组合多样的特点，常包含多类地物的繁杂组合，且细节信息十分密集；而文本描述往往聚焦核心地物，易出现影像细节冗余与文本信息缺失的匹配偏差。同时，对于相似地物的细微差异，现有特征表示方法，像 CNN 的局部特征提取、ViT 的全局注意力机制，难以有效捕捉区分性信息，导致细粒度场景下检索精度受限。如何让模型精准捕捉遥感影像与文本间的细粒度语义关联，是亟待解决的关键问题。

（2）多源数据融合与跨域泛化问题。实际遥感应用中，光学、SAR、高光谱、红外等多源数据能提供互补信息（如 SAR 的全天候成像能力、高光谱的精细光谱特征），但多源数据存在异构特性，现有文本-影像对齐方法难以直接适配其特征表示；且领域内缺乏针对多源数据的专用文本标注，现有研究多采用简单特征拼接的融合方式，未充分挖掘数据源间的互补价值，多源 CM-RSIT 研究仍处初步阶段。此外，现有模型在同域数据集中能维持一定性能，但在跨区域检索任务中，性能大幅下滑。这源于不同区域地物分布的显著差异的视觉特征鸿沟，且现有预训练模型多基于单一数据源与特定区域数据构建，缺乏跨域知识的

迁移与适配能力。

（3）时序遥感影像解译与跨模态动态匹配问题。时序遥感影像能展现地物随时间的动态变化（如植被生长衰退、城市建筑的新建拆除、水域的涨落等），为 RS-CMITR 提供了更丰富的语义维度。然而，现有研究多聚焦于静态遥感影像与文本的跨模态匹配，尚未充分利用时序影像的动态特性。一方面，时序影像的多时间步特征关联难以被有效捕捉，现有特征表示方法主要针对单张影像的空间特征提取，缺乏对时序维度上“变化模式”（如植被的季相变化规律、土地利用类型的渐变过程）的建模能力，导致模型无法理解文本中“随时间变化”的描述（如“夏季的河流”“近年扩张的工业区”）；另一方面，跨模态动态匹配机制的缺失，使得文本与时序影像序列的对齐存在困难，例如文本描述的是某一时间段内的地物变化过程，而模型只能基于单张影像的特征进行匹配，无法将文本的“时间维度语义”与影像的“时序变化特征”精准关联。未来需结合时序分析方法（如循环神经网络、时间注意力机制），设计时序感知的跨模态动态对齐模块，并引入地物动态变化规律的先验知识（如植被物候模型、城市发展时序模式），提升时序 RS-CMITR 的精度与实用性。

需要指出的是，尽管细粒度语义对齐、跨域泛化、时序匹配等问题在通用跨模态检索中也存在，但遥感场景因其独特的数据特性使这些挑战更为复杂。具体而言，遥感任务的特殊性主要体现在两个方面：其一，跨尺度语义鸿沟问题。遥感影像涵盖从厘米级到千米级的多尺度信息，而文本描述往往是抽象概念的高度凝练表达，如"城市中心的商业区"可能对应影像中数千个像素点构成的复合地物组合（包含建筑、道路、绿地、停车场等的空间配置），这种"文本抽象概念-影像多尺度细节"的跨层级匹配要求模型同时理解宏观场景布局与微观地物特征，远超自然图像中"一段文本描述一个物体"的简单对应关系；其二，类内差异大、类间差异小的矛盾特性。同一类别地物因成像条件、区域差异等因素在影像中呈现截然不同的视觉特征，而不同类别地物又常因光谱、纹理相似而难以区分（如稀疏林地与灌木丛、沥青路面与裸土），这种"相似但类别不同"与"不同但类别一致"的细微差异是遥感领域特有的识别难题。

针对上述遥感任务的独特挑战，未来研究需从以下方向寻求突破：一是构建多尺度分层对齐机制，设计"全局场景-局部地物-像素细节"的层次化特征提取框架，结合文本语义层级构建跨尺度对齐损失，使模型能根据文本粒度动态调整影像特征聚合尺度；二是发展判别性特征学习策略，针对"类内差异大、类间差异小"的特性设计困难样本挖掘方法，通过对比学习与度量学习拉大类间距离、压缩类内分布，保障大幅提升细粒度检索精度；三是融合遥感先验知识，通过知识图谱或规则约束增强模型语义理解能力；四是推进遥感专用的大规模预训练模型研究，在预训练阶段引入多光谱/高光谱波段等遥感特有信息，使模型学习领域

内的复合语义表达；五是探索时序感知的动态对齐机制，结合循环神经网络或 Transformer 建模地物变化模式，并融入植被物候模型、城市发展时序规律等先验知识，实现文本时间语义与影像时序特征的精准匹配；此外，多源异构数据融合方面需突破简单特征拼接的局限，设计针对 SAR、光学、高光谱等不同模态的深度对齐网络，并构建大规模多源遥感-文本标注数据集；跨域泛化方面可借鉴元学习、域自适应等方法，提升模型在少样本场景下的迁移能力。通过上述多维度的技术创新与领域知识融合，有望显著提升 RS-CMITR 任务在复杂场景下的检索精度与实用性。

6 结论

在 RS-CMITR 领域，技术发展与实际应用需求之间的适配性、方法选择的科学性一直是待厘清的关键问题。本文围绕这些核心需求展开工作，一方面梳理了 RS-CMITR 从传统人工特征工程到深度学习驱动的完整技术脉络，明确了文本特征表示方法从词袋模型、潜在语义分析到 Word2Vec、BERT 预训练模型，影像特征表示方法从 SIFT、HOG 手工特征到 CNN 局部提取、ViT 与 Mamba 全局建模的发展逻辑；另一方面，系统解析了非跨模态预训练（双编码器、融合编码器架构）与跨模态预训练（全学习、提示学习、适配器学习）两类技术路线的差异，并通过多种算法在三大数据集上的实验，量化得出跨模态预训练方法的性能优势，以及不同微调策略适配不同数据场景的规律，为实际应用中数据集与模型的选择提供了明确依据。这项工作的意义在于，理论上搭建起 RS-CMITR 领域“数据-特征-模型”的关联框架，推动计算机视觉与自然语言处理技术在遥感领域的深度融合。同时，本文也指出了当前领域存在的细粒度语义对齐难、多源数据融合不足、时序动态匹配缺失等问题，为后续研究突破瓶颈、拓展 RS-CMITR 的应用范围奠定了基础。

参考文献(**References**)

Alakwaa, Wafaa, Mohammad Nassef, and Amr Badr. 2017. "Lung cancer detection and classification with 3D convolutional neural network (3D-CNN)." *International Journal of Advanced Computer Science and Applications* 8 (8). doi: [10.14569/IJACSA.2017.080853].

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent dirichlet allocation." *Journal of machine Learning research* 3 (Jan):993–1022. doi: [10.1162/jmlr.2003.3.4-5.993].

Cao, Min, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. 2022. "Image-text retrieval: A survey on recent research and development." *arXiv preprint arXiv:2203.14713*. doi: [10.48550/arXiv.2203.14713].

Cheng, Qimin, Yuzhuo Zhou, Peng Fu, Yuan Xu, and Liang Zhang. 2021. "A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14:4284–97. doi: [10.1109/JSTARS.2021.3070872].

Church, Kenneth Ward. 2017. "Word2Vec." *Natural Language Engineering* 23 (1):155–62. doi: [10.1017/S1351324916000334].

Ghiasi, Golnaz, Tsung-Yi Lin, and Quoc V Le. 2019. Nas-fpn: Learning scalable feature pyramid architecture for object detection. Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Khirwar, Madhav, and Ankur Narang. 2024. GeoViT: versatile vision transformer architecture for geospatial image analysis. Paper presented at the 2024 International Conference on Machine Intelligence for GeoAnalytics and Remote Sensing (MIGARS).

Kim, Wonjae, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. Paper presented at the International conference on machine learning.

Kipf, TN. 2016. "Semi-supervised classification with graph convolutional networks." *arXiv preprint arXiv:1609.02907*. doi:

[10.48550/arXiv.1609.02907].
Koroteev, Mikhail V. 2021. "BERT: a review of applications in natural language processing and understanding." *arXiv preprint arXiv:2103.11943*. doi: [10.48550/arXiv.2103.11943].
Li, Xiang, Congcong Wen, Yuan Hu, and Nan Zhou. 2023. "RS-CLIP: Zero shot remote sensing scene classification via contrastive vision-language supervision." *International Journal of Applied Earth Observation and Geoinformation* 124:103497. doi: [10.1016/j.jag.2023.103497].
Li, Zewen, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. 2021. "A survey of convolutional neural networks: analysis, applications, and prospects." *IEEE transactions on neural networks and learning systems* 33 (12):6999–7019. doi: [10.1109/TNNLS.2021.3084827].
Li, Shuo, Haoyang Ji, Fang Liu, Licheng Jiao, Xutong Min, Xinyan Huang, Jiahao Wang, Long Sun, Lingling Li, and Xu Liu. 2025. "Fine-Grained Visual-Language Alignment for Remote Sensing Image-Text Retrieval." *IEEE Transactions on Geoscience and Remote Sensing. doi: 10.1109/TGRS.2025.3596333.*
Liu, Fan, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. 2024. "Remoteclip: A vision language foundation model for remote sensing." *IEEE Transactions on Geoscience and Remote Sensing* 62:1–16. doi: [10.1109/TGRS.2024.3390838].
Liu, Yue, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. 2024. "Vmamba: Visual state space model." *Advances in neural information processing systems* 37:103031–63.
Lu, Xiaoqiang, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. 2017. "Exploring models and data for remote sensing image caption generation." *IEEE Transactions on Geoscience and Remote Sensing* 56 (4):2183–95. doi: [10.1109/TGRS.2017.2776321].
Ma, Yan, Haiping Wu, Lizhe Wang, Bormin Huang, Rajiv Ranjan, Albert Zomaya, and Wei Jie. 2015. "Remote sensing big data computing: Challenges and opportunities." *Future Generation Computer Systems* 51:47–60. doi: [10.1016/j.future.2014.10.029].
Pang, Yanwei, Yuan Yuan, Xuelong Li, and Jing Pan. 2011. "Efficient HOG human detection." *Signal processing* 91 (4):773–81. doi: [10.1016/j.sigpro.2010.08.010].
Pennington, Jeffrey, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. Paper presented at the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).
Qu, Bo, Xuelong Li, Dacheng Tao, and Xiaoqiang Lu. 2016. Deep semantic understanding of high resolution remote sensing image. Paper presented at the 2016 International conference on computer, information and telecommunication systems (Cits).
Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. 2021. Learning transferable visual models from natural language supervision. Paper presented at the International conference on machine learning.
Shahid, Farah, Aneela Zameer, and Muhammad Muneeb. 2020. "Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM." *Chaos, Solitons & Fractals* 140:110212. doi: [10.1016/j.chaos.2020.110212].
Sumbul, Gencer, Marcela Charfuelan, Begüm Demir, and Volker Markl. 2019. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. Paper presented at the IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium.
Sun, Tianci, Chengyu Zheng, Xiu Li, Yanli Gao, Jie Nie, Lei Huang, and Zhiqiang Wei. 2025. "Strong and weak prompt engineering for remote sensing image-text cross-modal retrieval." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. doi: [10.1109/JSTARS.2025.3534474].
Tang, Xu, Dabiao Huang, Jingjing Ma, Xiangrong Zhang, Fang Liu, and Licheng Jiao. 2024. "Prior-experience-based vision-language model for remote sensing image-text retrieval." *IEEE Transactions on Geoscience and Remote Sensing*. doi: [10.1109/TGRS.2024.3464468].
Tang, Xu, Yijing Wang, Jingjing Ma, Xiangrong Zhang, Fang Liu, and Licheng Jiao. 2023. "Interacting-enhancing feature transformer for cross-modal remote-sensing image and text retrieval." *IEEE Transactions on Geoscience and Remote Sensing* 61:1–15. doi: [10.1109/TGRS.2023.3280546].
Tsai, Chih-Fong. 2012. "Bag‐of‐words representation in image annotation: a review." *International Scholarly Research Notices* 2012 (1):376804. doi: [10.5402/2012/376804].
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need." *Advances in neural information processing systems* 30.
Vedaldi, Andrea, and Andrew Zisserman. 2016. "Vgg convolutional neural networks practical." *Department of Engineering Science, University of Oxford* 66.
Wan, Ziyi, Enyuan Zhao, Jie Nie, Ze Zhang, Zhiqiang Wei, Nan Zheng, and Yuting Zhao. 2025. "Frequency Spectrum Adaptor for Remote Sensing Image-Text Retrieval." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. doi: 10.1109/JSTARS.2025.3589786.*
Wang, Yijing, Xu Tang, Jingjing Ma, Xiangrong Zhang, Fang Liu, and Licheng Jiao. 2024. "Cross-modal remote sensing image–text retrieval via context and uncertainty-aware prompt." *IEEE transactions on neural networks and learning systems*. doi: [10.1109/TNNLS.2024.3458898].
Woo, Sanghyun, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.
Woo, Sanghyun, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. Paper presented at the Proceedings of the European conference on computer vision (ECCV).
Wu, Jian, Zhiming Cui, Victor S Sheng, Pengpeng Zhao, Dongliang Su, and Shengrong Gong. 2013. "A Comparative Study of SIFT and its Variants." *Measurement science review* 13 (3):122. doi: [10.2478/msr-2013-0021].
Xiao, Yi, Qiangqiang Yuan, Kui Jiang, Jiang He, Yuan Wang, and Liangpei Zhang. 2023. "From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution." *Information Fusion* 96:297–311. doi: [10.1016/j.inffus.2023.03.021].
Xu, Junwei, Tao Huang, Zhenyu Wang, Weisheng Dong, and Xin Li. 2025. Bridging Task Boundaries: Remote Sensing Image-Text Retrieval via Dictionary-Driven Adaptation. Paper presented at the ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
Yang, Jian, Shengyang Li, and Manqi Zhao. 2025. "Parameter-Efficient Reparameterization Tuning for Remote Sensing Image-Text Retrieval." *IEEE Transactions on Geoscience and Remote Sensing*. doi: 10.1109/TGRS.2025.3558246.
Yu, Hongfeng, Fanglong Yao, Wanxuan Lu, Nayu Liu, Peiguang Li, Hongjian You, and Xian Sun. 2022. "Text-image matching for cross-modal remote sensing image retrieval via graph neural network." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16:812–24. doi: [10.1109/JSTARS.2022.3231851].
Yuan, Yuan, Yang Zhan, and Zhitong Xiong. 2023. "Parameter-efficient transfer learning for remote sensing image–text retrieval." *IEEE Transactions on Geoscience and Remote Sensing* 61:1–14. doi: [10.1109/TGRS.2023.3308969].

Yuan, Zhiqiang, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. 2022. "Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval." *arXiv preprint arXiv:2204.09868*. doi: [10.1109/TGRS.2021.3078451].
Yuan, Zhiqiang, Wenkai Zhang, Xuee Rong, Xuan Li, Jialiang Chen, Hongqi Wang, Kun Fu, and Xian Sun. 2021. "A lightweight multi-scale crossmodal text-image retrieval method in remote sensing." *IEEE Transactions on Geoscience and Remote Sensing* 60:1–19. doi: [10.1109/TGRS.2021.3124252].
Yuan, Zhiqiang, Wenkai Zhang, Changyuan Tian, Xuee Rong, Zhengyuan Zhang, Hongqi Wang, Kun Fu, and Xian Sun. 2022. "Remote sensing cross-modal text-image retrieval based on global and local information." *IEEE Transactions on Geoscience and Remote Sensing* 60:1–16. doi: [10.1109/TGRS.2022.3163706].
Zhang, Zilun, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. 2024. "Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing." *IEEE Transactions on Geoscience and Remote Sensing*. doi: [DOI10.1109/TGRS.2024.3449154].
Zhu, Lilu, Yang Wang, Yanfeng Hu, Xiaolu Su, and Kun Fu. 2024. "Cross-modal contrastive learning with spatiotemporal context for correlation-aware multiscale remote sensing image retrieval." *IEEE Transactions on Geoscience and Remote Sensing* 62:1–21. doi: [10.1109/TGRS.2024.3417421].

Remote sensing cross-modal image-text retrieval: key technologies and challenges

WANG Yijing[1], TANG Xu[1], HAN Shuo[1], DU Ruiqi[1]

1. *School of Artificial Intelligence, Xidian University, Xi'an, 710126, China*

Abstract: With the rapid development of remote sensing technology and the continuous expansion of application scenarios, efficient retrieval and intelligent utilization of massive remote sensing data have become core demands in the field of remote sensing information processing. Remote sensing cross-modal image-text retrieval (RS-CMITR), as a key technology connecting visual perception and language understanding, establishes semantic associations between natural language descriptions and remote sensing image content to achieve efficient bidirectional interaction between text-to-image retrieval and image-to-text retrieval. This technology breaks down modality barriers and enables precise semantic localization of remote sensing data, providing intelligent solutions for disaster detection, environmental monitoring, and urban planning. This paper aims to systematically review the technological evolution, core methodologies, and major challenges in the RS-CMITR field, providing comprehensive reference for in-depth research in this domain.

This review systematically analyzes RS-CMITR technology from datasets, feature representation, model architectures, to learning paradigms. First, we introduce three mainstream benchmark datasets—UCM-Captions, RSICD, and RSITMD—analyzing their characteristics in terms of scale, scene diversity, and text annotation quality, and explain the evaluation metrics including Recall@K and

mean Recall. Second, we review the evolution of text feature representation methods from traditional statistical approaches to deep learning methods, as well as the development of remote sensing image feature representation from hand-crafted features to deep neural networks. Third, using cross-modal pre-training as the classification criterion, we categorize existing RS-CMITR methods into two major classes: methods based on non-cross-modal pre-training (including dual-encoder architecture and fusion encoder architecture) and methods based on cross-modal pre-training (including full fine-tuning, prompt learning, and adapter learning). We provide in-depth analysis of the technical principles, model characteristics, advantages, and conduct comprehensive performance comparisons of representative algorithms across the benchmark datasets.

Experimental comparisons reveal several important findings regarding RS-CMITR methods. Cross-modal pre-training methods consistently demonstrate superior retrieval performance compared to non-cross-modal pre-training methods across all benchmark datasets. Different fine-tuning strategies exhibit distinct data adaptability patterns: full fine-tuning and adapter learning excel on large-scale datasets, while prompt learning shows advantages on small-scale datasets, highlighting the effectiveness of parameter-efficient fine-tuning. Dataset quality, particularly text diversity, significantly influences model performance. The review demonstrates that RS-CMITR has evolved from traditional feature engineering to deep learning-driven intelligent retrieval paradigms, with cross-modal pre-training combined with parameter-efficient fine-tuning emerging as the mainstream technical approach.

Despite significant progress in RS-CMITR technology, three core challenges remain: (1) Fine-grained semantic alignment is difficult, as existing methods struggle to capture subtle differences between similar land cover types and establish precise image-text correspondences; (2) Multi-source data fusion and cross-domain generalization capabilities are insufficient, with models showing significant performance degradation in cross-domain and cross-sensor tasks; (3) Temporal dynamic matching mechanisms are absent, as current research focuses on static images and cannot effectively model temporal changes in land cover. Future research should focus on enhancing fine-grained feature representation, collaborative modeling of multi-source heterogeneous data, and construction of temporal-aware dynamic alignment mechanisms to advance RS-CMITR technology from theory to practical applications.

Key words: remote sensing images; cross-modal retrieval; image-text relationship modeling; deep learning; pre-trained models; semantic alignment; feature representation; pre-trained vision-language models

579 **Supported by** The National Natural Science Foundation of China (General Program, No. 62571387)