 DOI: https://doi.org/10.12345/bdai.v6i4.30575

Intelligent NLP Retrieval for Remote Sensing Imagery based on Grid data Organization and Multimodal Deep Learning

Fuhu Ren[1,2] Lin Li[2,4] Zicong Du[5] Jinhua Dong[2] Yi Huang[3*]

1. School of Earth and Space Sciences, Peking University, Beijing, 100871, China
2. Institute of Earth and Space Technology, Peking University International S.&T. Innovation Center at Lin-gang Special Area, Shanghai, 201306, China
3. Fujian Big Data First level Development Co., Ltd., Fuzhou, Fujian, 350207, China
4. Fujian Mindu Innovation Laboratory, Fuzhou, Fujian, 350108, China
5. Zhongke Yunyao (Shenzhen) Technology Co., Ltd., Shenzhen, Guangdong, 518067, China

Abstract

To address the three major challenges in multi-temporal and multimodal retrieval of remote sensing imagery—namely, the semantic gap, operational complexity, and inefficient temporal analysis—this paper proposes an intelligent NLP retrieval framework for remote sensing imagery that integrates grid subdivision data organization with multimodal deep learning. The core contributions are: (1) Construction of a multi-scale spatiotemporal grid index based on the Discrete Global Grid System (DGGS), enabling efficient organization of image data; (2) Proposal of a remote sensing-specific multimodal deep learning model (Remote CLIP), which achieves a significant improvement in image-text matching accuracy on the RSICD & RSITMD dataset; (3) Design of a natural language instruction parsing engine that supports complex temporal queries and substantially enhances automated parsing accuracy.

Keywords

Remote sensing imagery retrieval; Multimodal deep learning; Grid subdivision; Remote CLIP model; Natural Language Processing

基于网格剖分组织与多模态深度学习的遥感影像自然语言智能检索

任伏虎[1,2] 李林[2,4] 杜子聪[5] 董锦华[2] 黄毅[3*]

1. 北京大学地球与空间科学学院，中国·北京 100871
2. 上海临港北京大学国际科技创新中心，中国·上海 201306
3. 福建大数据一级开发有限公司，中国·福建 福州 350207
4. 闽都创新实验室时空大数据中心，中国·福建 福州 350108
5. 中科云遥（深圳）科技有限公司，中国·广东 深圳 518067

摘　要

针对遥感影像多时相、多模态检索中存在的语义鸿沟、操作复杂及时序分析低效三大难题，本文提出一种融合网格剖分组织与多模态深度学习的遥感影像自然语言智能检索框架，主要包括：构建多尺度时空网格索引，通过全球离散网格系统实现影像数据高效组织；提出遥感专用多模态深度学习模型，在RSICD和RSITMD数据集的图文匹配准确率有显著提升；设计自然语言指令解析引擎，支持复杂时序查询，自动化解析准确率大幅提高。

关键词

遥感影像检索；多模态深度学习；网格剖分；Remote CLIP模型；自然语言处理

【课题项目】民用航天技术预先研究项目（项目编号：D040303）。

【作者简介】任伏虎（1962–），男，中国河北人，博士，教授，从事遥感与地理信息系统研究。

【通信作者】黄毅（1995–），男，中国福建人，硕士，从事数字经济与大数据研究。

1 引言

当前，全球对地观测系统日均产生超过 10TB 的遥感影像数据，传统的遥感数据检索方式面临如下多重挑战：（1）语义理解局限：依赖元数据（拍摄时间 / 传感器类型等）与关键词匹配，难以解析“城市扩张监测”等复杂语义；（2）多时相对比低效：用户需手动检索不同时相数据再计算差

异，耗时巨大，用户体验度差；（3）跨模态融合不足：现有方法如 AMFMN 仅支持图文单向检索，多光谱 -SAR 数据协同检索精度不足等。对遥感数据的智能检索特别是融合自然语言的智能检索成为行业共性需求之一。

2 问题的提出

2.1 研究现状与存在问题

现阶段，与遥感数据检索相关的方法和技术主要包括：

（1）基于元数据的遥感影像检索方案

通过建立空间索引和文本检索机制，根据用户输入的关键字段进行搜索。这类系统能够实现基本的检索功能并提供较快的空间过滤能力。然而，这种方法难以处理从多模态深度学习角度挖掘的潜在图像语义，并且无法支持基于自然语言的复杂描述，例如“polygon 范围内的农田减少”等。

（2）部分深度学习检索方案

这类研究利用深度学习模型对遥感影像进行特征提取，以实现场景分类或目标检测等快速检索任务。虽然在影像目标识别方面表现良好，但通常其只针对单一模态（如光学影像），或者需要用户输入精确的类别。此外，这类方案缺乏与自然语言描述的深度融合，难以满足多时相、多类别、组合条件的灵活检索需求。

（3）基于网格剖分的多时相影像对比方案

这些系统通过地球剖分网格对影像进行分割，从而支持局部区域的时序对比，具备多尺度、多时相的对比能力。但由于缺乏与自然语言理解的结合，其检索方式依赖 GIS 专业语句或脚本，操作门槛较高，难以满足用户对灵活检索的需求。

2.2 本文所要解决的问题

针对遥感数据检索现有技术解决方案存在的缺点，本文所要解决的技术问题是：如何通过深度学习多模态技术与遥感影像网格剖分组织技术相结合，实现多时相、多场景、多模态的遥感影像智能检索与差异对比，且能支持自然语言灵活描述的检索条件。

3 技术基础

3.1 全球离散网格（DGGS）技术

全球离散网格系统（DGGS）是一种基于离散几何网格对地球进行剖分和建模的空间数据管理方法。它通过将地球表面划分为一系列规则的网格单元，提供了一种标准化、层次化和高效的空间数据组织和分析方式。DGGS 的网格具体划分规则有多种。其中，北京大学研究团队提出的 GeoSOT 地球剖分框架具有较大影响力。研究团队出版了《空间信息剖分组织导论》[1] 等专著，推动发布了《地球空间网格编码规则》（GB/T 40087-2021）、《北斗网格位置码》（GB/T 39409-2020）、《北斗剖分时间码》（GB/T 42578-2023）等多项国家标准，并在各地进行了多个项目实践 [2]。

3.2 多模态深度学习技术

所谓多模态是指在同一个模型或系统中同时利用多种数据模式（如图像、文本、音频）进行学习或推断。CLIP (Contrastive Language-Image Pre-training)：即对比语言 - 图像预训练，是一种多模态深度学习模型 [3]。它通过在大规模的图像 - 文本对数据上进行对比学习，将图像和其对应的文本描述映射到同一个共享的特征空间（嵌入空间）中。在这个空间里，语义相似的图像和文本的向量表示会非常接近，而语义不相关的则会相互远离。

4 技术方案

本文提出了一种基于网格剖分组织与多模态深度学习的遥感影像自然语言智能检索方法。该方法将网格剖分数据组织、多模态深度学习 Embedding 以及自然语言处理三者进行结合，实现对多源多时相遥感影像数据的有效管理与智能检索。同时，与 AlphaEarth[4] 最大的区别在于，本方法通过剖分网格进行多维数据的组织索引，在后续的计算分析中能够发挥剖分网格编码计算的优势。

4.1 数据预处理与入库

（1）数据获取与预处理：数据来源可含多源遥感数据，包括光学影像、多光谱影像、SAR 影像等。对获得的遥感数据进行常规预处理：畸变校正、辐射校正、配准、融合。在此过程中形成统一的坐标参考系和时间标记，为后续切分与检索提供一致的数据格式。

（2）网格剖分与索引编码：

①网格剖分与多尺度索引：采用 DGGS 剖分方法，通过指定网格层级范围对大范围遥感影像进行分块，得到具有唯一编码的网格单元。

②遥感影像像元信息提取：获取遥感影像的拍摄时间、拍摄设备等有关元数据信息，并对像元的空间信息、辐射信息、位深度信息等属性信息进行提取。

③剖分网格影像数据存储：将每个基于剖分网格的影像切片数据及其元数据信息存储进主数据库（存储网格块及其各种属性信息）。

④将剖分网格按照编码推送至任务队列等待图像编码器进行高维度特征提取。

4.2 多模态深度学习（Remote CLIP）特征提取与存储：

（1）基于 CLIP 模型微调形成 Remote CLIP：

Remote CLIP 是一种借鉴了 CLIP 思路的多模态模型，将遥感影像同自然语言进行对齐与关联。CLIP 的关键在于它通过对比学习（在大量图像 - 文本对上预训练）将图像和文本映射到一个共享的嵌入空间 [5]。在这个空间中，语义上相似的图像和文本会有相近的向量表示。

（2）特征提取流程：

①从任务队列中并行取出待处理的剖分网格影像；

②将按照网格剖分组织形成的遥感影像块输入 Remote CLIP 的图像编码器，获得其高维图像特征向量；

③提取的特征向量存储进特征数据库中（如向量数据库 FAISS、Milvus、Qdrant 等）；

④最终所有剖分网格影像提取特征后，特征距离相近的切分网格其语义也相近。

4.3 自然语言的理解与解析

（1）当用户输入查询指令 Query，例如 Query(“从 2015 年到 2017 年间 Polygon 范围内农田减少的影像”)、Query（“查询包含高楼建筑的影像”）时，系统需要通过命名实体识别与解析规则，提取出关键信息（时间范围、空间范围、地物类型“农田”等）。

（2）提取出结构化信息可以是由 JSON 描述的信息。(代码略)

4.4 检索执行器解析执行：

（1）当目标类型为“查询包含高楼建筑的影像”或“查询 2015 年包含高楼建筑的影像”等不包含时间变化状态 (change_type 为 false) 的查询指令时，具体执行步骤如下：

①记忆总体查询指令：Query（“查询包含高楼建筑的影像”）；

②将用户输入指令 Query 经过 Remote CLIP 进行句子嵌入的特征提取后，得到 text_query_embedding；

③检索到相关的切片网格影像，在执行该步骤时，需要协同时间范围和空间范围进行数据过滤，即（时间范围内 & 空间范围内 & 相似范围内）查询条件进行查询过滤；

④最终检索到时间范围内指定空间和语义相关的切片网格影像或网格码集合。

（2）当目标类型为“农田减少”等带有时间变化状态的查询指令时（change_type 为 true），具体执行步骤如下：

①记忆总体查询指令：Query(“从 2015 年到 2017 年间 Polygon 范围内农田减少的影像”)；

②通过 LLM 拆分总体查询指令为更多子查询指令列表：QueryList [Query1(“2015 年 Polygon 范围内农田影像”)，Query2(“2017 年 Polygon 范围内的裸地或少量农田影像”)]；

③将子查询指令列表通过不带有时间变化状态的查询指令进行查询；

④得到 grid1，也得到 grid2；

⑤ grid1 和 grid2“自 2015 年到 2017 年农田减少部分”的网格集。

4.5 结果可视化

为提高检索结果的直观性和实用性，可整合一套多维度的可视化展示方案，包括以下内容：①差异区域高亮显示；②多时相影像叠加；③动态统计分析等。

5 讨论

选用 RSICD 和 RSITMD 数据集对本方案进行了初步验证，结果表明具备技术可行性，且在图文匹配准确率、召回率等指标方面较传统遥感影像识别方法有不同程度的提升。对于本技术方案的优点与创新总结如下：

（1）具备自然语言与图像多模态融合能力

本文通过引入 CLIP 多模态深度学习模型，将遥感影像的视觉特征与自然语言描述统一映射到同一语义空间中，解决了传统技术中依赖元数据或固定关键词检索的局限性。与现有技术相比，本技术方案能够灵活支持用户通过自然语言描述复杂的检索需求。

（2）建立“多时相 - 空间 - 语义”的三重索引。

本技术方案将空间网格剖分、时间戳以及特征向量三者结合，使检索能够在三重维度上综合筛选，可大幅提升检索的效率与准确度。同时，良好的系统设计可以提升网格遥感影像检索的智能化程度。

（3）自动化差异检测

在传统方式中，多时相影像对比分析通常需要用户手动执行多个步骤，操作复杂且容易出错。本技术方案通过自然语言解析，实现了多时相差异检测的自动化，能够显著降低用户操作的复杂性，使非专业用户也能轻松完成复杂检索任务，同时确保结果的准确性和一致性。

6 结论与展望

本文提出“网格剖分组织 + 多模态深度学习”融合的遥感自然语言智能检索框架，对于将人工智能大模型技术引入遥感领域，破解当前海量遥感数据管理中语义鸿沟、时序分析低效、操作复杂等共性痛点，具有一定创新性。发展“预测式检索”新范式，如基于历史数据的洪涝风险区域推演等；建立开源技术生态，通过模块化设计开放 DGGS 剖分引擎、预训练权重及查询解析算法等，推动在不同场景的应用。

参考文献

[1] 程承旗,任伏虎 等著. 空间信息剖分组织导论[M]. 科学出版社.2012

[2] 李林,程承旗 等. 北斗网格码:数字孪生城市CIM时空网格框架[J]. 信息通信技术与政策,2021,30(11):1-5.

[3] 张程皓. 基于CLIP预训练模型的跨模态哈希检索在遥感图像上的研究及应用[D]. 重庆师范大学,2023

[4] Brown F.C. et al. AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data. arXiv:2507.22291 [cs.CV] [Google Scholar]

[5] Devlin, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv 2018, arXiv:1810.04805. [Google Scholar]