聚焦视觉－语言对齐：遥感多模态大模型技术综述与前沿展望

王琼洁
中国电子信息产业发展研究院，北京 100036
DOI: 10.61369/TACS.2026050049

摘　　要： 在视觉大模型快速演进的背景下，测绘遥感领域正由以分类、检测和分割为核心的单任务解译模式，逐步走向以视觉—语言对齐、指令跟随和自然语言交互为特征的多模态理解范式。相较于自然图像场景，遥感图像具有俯视视角、尺度跨度大、目标密集、跨传感器差异显著以及专业知识依赖强等特点，使通用视觉大模型难以直接迁移。围绕这一问题，本文从遥感视觉—文本对齐的基础与难点出发，系统梳理了遥感多模态大模型的数据构建、训练范式与核心架构演进。首先，概述了以 CLIP 类模型为基础的遥感视觉—语言表征学习及其在领域持续预训练中的发展。其次，重点总结了遥感图文预训练数据、指令微调数据以及人机协同生成策略对模型能力形成的作用，并梳理了从全局图文理解到空间基础关联、再到多传感器统一建模的主要技术路线。最后，结合测绘遥感业务对空间定位精度、时空一致性、可靠性与工程可部署性的要求，分析了当前方法在遥感领域幻觉、细粒度定位、动态推理、评测体系和边缘部署等方面面临的主要问题。

关 键 词： 遥感多模态大模型；视觉－语言对齐；指令微调；空间基础关联；视觉问答；测绘遥感

Focusing on Vision-Language Alignment: Review and Frontier Prospect of Remote Sensing Multimodal Large Model Technology

Wang Qiongjie
China Center for Information Industry Development, Beijing 100036

Abstract: With the rapid evolution of visual large models, the field of surveying and remote sensing is shifting from single-task interpretation centered on classification, detection and segmentation to a multimodal understanding paradigm characterized by vision-language alignment, instruction following and natural language interaction. Different from natural images, remote sensing images feature top-down perspective, large scale span, dense targets, prominent cross-sensor differences and high reliance on professional knowledge, making general visual large models difficult to be directly transferred. Focusing on this issue, this paper systematically reviews the data construction, training paradigms and core architecture evolution of remote sensing multimodal large models based on the fundamentals and difficulties of remote sensing vision-language alignment. Firstly, it summarizes remote sensing vision-language representation learning based on CLIP and its development in domain continual pre-training. Secondly, it highlights the roles of remote sensing image-text pre-training data, instruction tuning data and human-machine collaborative generation strategies in model capability building, and sorts out the major technical routes from global image-text understanding to spatial correlation and multi-sensor unified modeling. Finally, in combination with the requirements of surveying and remote sensing services for spatial positioning accuracy, spatiotemporal consistency, reliability and engineering deployability, it analyzes the main challenges existing in current methods, including model hallucination, fine-grained localization, dynamic reasoning, evaluation system and edge deployment in the remote sensing field.

Keywords: remote sensing multimodal large model; vision-language alignment; instruction tuning; spatial correlation; visual question answering; surveying and remote sensing

引言

过去十余年，卷积神经网络和视觉 Transformer 在遥感场景分类、目标检测、变化检测与语义分割等任务中取得了显著进展，但这类方法大多遵循“特定任务—特定模型—特定数据集”的封闭式范式。模型通常只能输出固定类别标签、边界框或分割结果，难以支持

复杂语义查询、跨任务迁移以及与人类专家的自然语言交互[1][2][3]。随着通用多模态大模型的发展，遥感智能解译正在从“视觉识别”走向“视觉—语言联合理解”，这为地球观测信息服务方式的重构提供了新的技术契机[4][5]。然而，通用视觉大模型并不能直接等同于遥感解决方案。遥感图像多为俯视视角，目标尺度跨度大、密集度高，且不同传感器在成像机理、噪声形式和空间分辨率上差异显著。与此同时，遥感业务解释还高度依赖地理知识、场景上下文和时空关系，这使得通用模型在迁移到遥感领域时容易出现语义漂移、细粒度定位不稳以及专业知识幻觉等问题。因此，围绕视觉—语言对齐机制重构遥感大模型，已成为当前研究的核心议题。

本文围绕遥感多模态大模型的关键研究链路展开综述，重点关注视觉—语言表征基础、数据与训练范式、核心架构演进以及应用中的现实挑战。本文聚焦遥感多模态大模型的关键链路，梳理其对齐基础、数据训练范式、核心架构演进，并剖析现实挑战与未来方向。

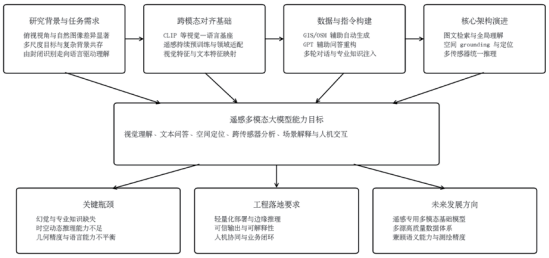

图 1 遥感视觉大模型研究的技术框架与演进逻辑

一、遥感视觉—语言对齐的基础与难点

为了更清晰地展示遥感视觉大模型研究从跨模态对齐基础、数据与指令构建到架构演进、关键挑战及未来方向的整体逻辑，本文对相关技术脉络进行了归纳，如图 1 所示。视觉—语言对齐的目标，是将图像与文本映射到共享语义空间，使模型既能理解视觉内容，又能根据语言指令完成检索、问答、定位与推理。CLIP 等模型通过大规模图文对的对比学习，为多模态理解提供了稳健基座，也成为遥感视觉大模型的重要起点[6][7]。在此基础上，RemoteCLIP[8] 和 GeoRSCLIP[9] 等工作通过持续预训练或领域化扩展，将遥感图像、航拍影像与配套文本纳入统一训练框架，提升了模型对遥感语义的敏感性和零样本迁移能力。这说明，视觉—语言对齐不仅是模型结构问题，更依赖于领域数据和领域语义空间的再塑造。

与自然图像相比，遥感视觉—语言对齐至少存在四方面特殊难点。一是俯视视角缺乏明显轮廓；二是高分图像内全局场景与微小目标并存；三是多传感器机理各异；四是场景解释高度依赖专业先验。正因为这些因素叠加，遥感视觉大模型的研究重点，不应只是“把通用模型搬过来”，而应是构建面向遥感场景的专门对齐机制。进一步看，遥感视觉—语言对齐所面对的并非单一语义映射问题，而是“视觉表征、空间关系与专业知识”三者之间的联合耦合。自然图像中的文本描述往往围绕物体外观、动作和场景氛围展开，而遥感语义表达更强调对象的地理功能、空间组合和上下位层级关系。例如，同样是一片建筑群，自然图像中的描述可能关注“住宅区”或“工厂”，而遥感判读中还需要进一步区分其是否临近交通干线、是否呈现规则化布局、是否具备特定的用地组织特征。这意味着，遥感视觉大模型不仅要学会把图像片段与词汇对应起来，还需要建立对象之间、对象与区域之间以及对象与业务知识之间的关系化表达能力。也正因如此，后续研究逐渐从单纯提升图文相似度，转向关注多尺度区域建模、空间基础关联和知识约束下的多模态推理。

二、数据基础与训练范式

（一）遥感图文预训练数据的构建

相较于自然图像领域可以直接从互联网获取大规模图文对，遥感多模态模型长期受制于高质量图文数据稀缺。为缓解这一问题，研究者开始借助地理坐标、GIS 矢量数据和开放地图语义信息构建遥感图文预训练语料。SkyScript[10] 通过地理坐标关联开放遥感图像与 OpenStreetMap 标签，形成大规模、语义多样的图文对数据集，为遥感视觉—语言预训练提供了重要数据基础。[10] 这一策略的价值在于，它不再仅依赖人工撰写描述，而是利用地理信息系统中已有的对象类别、道路结构和地物属性，自动扩展视觉—文本对应关系。

不过，自动化构建的数据并不意味着可以直接满足高质量对齐需求。GIS 标签通常更偏向对象名称与属性记录，而缺乏自然语言中的关系表达、语义层次和任务导向。遥感大模型如果仅依赖此类弱监督语料，往往能够学习到图文相关性，却难以支撑复杂问答和细粒度推理。因此，预训练数据的关键不只是规模，更在于其是否覆盖多尺度场景、专业术语、对象关系以及不同传感器下的语义稳定性。

（二）指令微调与人机协同数据生成

在多模态大模型研究中，指令微调决定了模型能否将底层对齐能力转化为可交互、可问答、可推理的任务能力。遥感领域的难点在于，真实高质量的遥感对话数据非常有限。为此，研究者普遍采用“结构化标注 + 通用大模型重写”的方式，将目标框、分割掩码、类别标签和图像元信息组织为输入，再借助 GPT-4 类模型生成遥感问答、描述和推理指令。RSGPT[11] 即是这一方向的早期代表，它构建了遥感图像描述与问答基准，推动了面向遥感场景的视觉语言交互研究。

从训练范式角度看，遥感多模态数据构建还呈现出两个值得关注的趋势。其一，是由“静态描述型数据”向“任务驱动型数据”扩展。前者主要回答图像内容是什么，后者则更强调模型能否根据指令完成检索、比较、定位、判断和解释等复合任务。其

二，是由“单模态图像问答”向“多源协同问答”扩展，即在训练样本中同时引入影像、元数据、时间信息和传感器类型，使模型逐步形成对遥感业务场景的上下文感知能力。对综述研究而言，这一变化说明数据已不再只是模型训练的支撑材料，而是在很大程度上决定了模型能力边界：数据组织方式不同，模型最终形成的交互能力和推理深度也会显著不同。

从方法论上看，遥感指令微调数据的构建正在从单轮问答走向多层次、多任务和多粒度组织。一类数据关注图像级理解，如场景描述、土地利用判断和区域功能识别；另一类数据强调区域级或对象级理解，如指代表达、目标定位和属性描述；还有一类数据尝试融入跨时相、跨传感器和业务知识，提升模型的时空推理与专业问答能力。人机协同机制在此过程中尤为重要：通用大模型可以提供流畅语言，而遥感专家或规则系统则负责约束专业正确性。二者结合，才能避免训练语料“语言好看但知识不稳”的问题。

三、核心架构与技术演进

（一）从全局图文理解到空间基础关联

遥感视觉大模型的早期工作主要关注全局图文理解，即让模型回答“图像中有什么”“场景大致属于哪一类”这类问题。此时，视觉编码器负责提取整图特征，语言模型则在整体语义层面进行生成与推理。这一范式适合遥感图像描述、检索和粗粒度问答，但难以满足“目标在哪里”“某个设施位于何处”之类更接近业务需求的任务。GeoChat[12] 的代表性意义在于，它将高分辨率遥感图像、多任务对话和空间基础关联结合起来，使模型不仅能够回答图像内容，还能够根据文本指令输出对应区域或目标的位置，从而推动遥感多模态模型由“会描述”走向“会定位”。

空间基础关联之所以重要，是因为遥感应用普遍要求把语义结果落实到明确的地理位置上。无论是灾害应急中的受损设施定位，还是国土调查中的地物检索，都不能只停留在文本解释层面。与此同时，遥感图像中存在大量尺度差异显著、边界复杂甚至旋转分布明显的目标，这使空间基础关联比自然图像场景更具挑战。因此，这一阶段的方法设计重点，不仅包括视觉特征与文本特征的匹配，还包括多尺度区域建模、坐标离散化表达以及高分辨率输入下的局部推理机制。

值得注意的是，多传感器统一建模并不只是把多种影像简单拼接输入，而是要求模型能够理解不同传感器之间在信息表达上的互补性与差异性。光学数据更适合描述地物外观、纹理和颜色，SAR 数据则在穿透云雾、反映结构与粗糙度方面具有独特优势，红外数据又能够补充热特征与异常状态信息。若模型无法显式处理这些差异，就容易在统一训练中出现表征混淆，反而削弱各模态的判别能力。因此，当前多传感器路线的关键问题，一方面在于如何通过模态适配器、路由机制或分层对齐策略保留不同传感器的独特性，另一方面在于如何在统一语义空间中实现跨模态信息融合，使模型既能开展联合推理，又不丢失对单一传感器物理属性的敏感性。

（二）多传感器统一建模与任务协同

随着模型能力的提升，研究开始从单一光学图像理解扩展到多传感器统一建模。EarthGPT[13] 试图在统一框架下处理光学、SAR 与红外等多种遥感数据，并通过多任务指令微调提升跨传感器理解能力。[13] 这一趋势说明，遥感视觉大模型的发展不再局限于“把图像和文本对齐”，而是进一步迈向“把不同传感器、不同任务和不同表达粒度纳入同一语义空间”。对于真实遥感业务而言，这种统一建模具有明显优势：光学影像受云雾影响时，可借助 SAR 弥补感知盲区；图像级判断与区域级定位也可以在一个对话框架中协同完成。

四、结论

总体来看，遥感视觉大模型的研究正在推动地球观测智能解译由单一视觉识别向视觉—语言联合理解转变。以 RemoteCLIP 为代表的领域视觉—语言基座、以 SkyScript 为代表的数据构建策略、以 RSGPT 为代表的指令微调实践，以及以 GeoChat 和 EarthGPT 为代表的空间基础关联与多传感器统一建模方法，共同构成了当前遥感多模态大模型的主要发展脉络。从综述视角看，这一方向的关键不在于简单复制通用视觉大模型框架，而在于面向遥感场景重建数据、对齐、推理与评测体系。未来，随着高质量遥感多模态数据的持续积累、几何与物理约束下的对齐机制不断完善，以及更可靠的行业知识注入方式逐步成熟，遥感视觉大模型有望成为测绘遥感智能解译的重要基础设施。

参考文献

[1] 龚健雅，季顺平．摄影测量与深度学习 [J]. 测绘学报，2018，47(6)：693-704.

[2] 李德仁，张良培，夏桂松．遥感大数据自动分析与数据挖掘 [J]. 测绘学报，2014(12)：1211-1216. DOI：10.13485/jc.nki1.1-20892.0140.187.

[3] 燕琴，顾海燕，杨懿，等．智能遥感大模型研究进展与发展方向 [J]. 测绘学报，2024，53(10)：1967-1980. DOI：10.11947/j.AGCS.2024.20240053.3..

[4] 张良培，张乐飞，袁强强．遥感大模型：进展与前瞻 [J]. 武汉大学学报（信息科学版），2023，48(10)：1574-1581. DOI：10.13203/j.whugis20230341.

[5] 张永军，李彦胜，党博，等．多模态遥感基础大模型：研究现状与未来展望 [J]. 测绘学报，2024，53(10)：1942-1954. DOI：10.11947/j.AGCS.2024.20240019.

[6]Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PmLR, 2021: 8748-8763.

[7]Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//International conference on machine learning. PMLR, 2023: 19730-19742.

[8]Liu F, Chen D, Guan Z, et al. Remoteclip: A vision language foundation model for remote sensing[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-16.

[9]Zhang Z, Zhao T, Guo Y, 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: 1-23.

[10]Wang Z, Prabha R, Huang T, et al. SkyScript: A large and semantically diverse vision-language dataset for remote sensing[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(6): 5805-5813.

[11]Hu Y, Yuan J, Wen C, et al. Rsgpt: A remote sensing vision language model and benchmark[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2025, 224: 272-286.

[12]Kuckreja K, Danish M S, Naseer M, et al. Geochat: Grounded large vision-language model for remote sensing[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 27831-27840.

[13]Zhang W, Cai M, Zhang T, et al. EarthGPT: A universal multimodal large language model for multisensor image comprehension in remote sensing domain[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-20.

[14]Li X, Wen C, Hu Y, et al. Vision-language models in remote sensing: Current progress and future trends[J]. IEEE Geoscience and Remote Sensing Magazine, 2024, 12(2): 32-66.

[15]Cha K, Yu D, Seo J. Pushing the limits of vision-language models in remote sensing without human annotations[J]. arXiv preprint arXiv:2409.07048, 2024.

