中图法分类号:TP751.1　　**文献标识码**:A　　**文章编号**:1006-8961(2026)04-0973-14

论文引用格式:Zhi Y J, Jiang Y W, Yang Z, Chen Y Z, Hao W K, Ma M Y, et al. 2026. Development status and prospects of pretrained foundation models for remote sensing imagery. Journal of Image and Graphics, 31(4):0973-0986(支元杰, 姜艺伟, 杨知, 陈奕州, 郝文魁, 马明阳, 魏江, 梅少辉. 2026. 面向遥感图像的预训练基础模型发展现状与展望. 中国图象图形学报, 31(4):0973-0986)[DOI:10.11834/jig.250424]

面向遥感图像的预训练基础模型发展现状与展望

支元杰[1],姜艺伟[1],杨知[2],陈奕州[1],郝文魁[2],马明阳[1],魏江[1],梅少辉[1*]

1. 西北工业大学电子信息学院,西安　710129; 2. 国网电力工程研究院有限公司,北京　100100

摘　要:随着训练数据体量和深度学习模型规模的不断提升,视觉基础模型(vision foundation model,VFM)和大语言模型(large language model,LLM)在多种类型下游任务中表现出优异的泛化能力,引发广大学者关注。在遥感(remote sensing,RS)领域,数据来源多样、模态复杂、地物类型丰富且分布异质,传统方法难以全面建模其中蕴含的语义与空间关系。围绕遥感多源数据特性和复杂地物关系设计预训练基础模型,对于提取通用、鲁棒的特征表示以及实现遥感影像智能解译具有重要意义。本文系统回顾了遥感预训练基础模型(remote sensing foundation model,RSFM)的研究进展,重点聚焦于单模态与多模态预训练策略的发展脉络和关键方法,梳理了当前主流遥感预训练数据集及其构建特性。在单模态方面,总结了典型的自监督对比学习(self-supervised contrastive learning,SSCL)与掩码生成预训练(masked generative pre-training)框架,并分析其在不同分辨率和多光谱影像中的应用效果;在多模态方面,重点回顾了图像—文本、图像—位置、图像—音频等多模态预训练策略及其特征对齐机制。进一步地,本文对遥感基础模型在跨场景适应、特征表征能力、预训练范式、数据质量与获取成本等方面所面临的主要挑战进行了分析,并从多模态融合、轻量化建模、跨域与跨时间泛化、模型透明度与可信性等角度,对未来遥感大模型的发展趋势与潜在研究方向进行了前瞻性探讨。本文旨在为遥感智能解译与大模型研究提供系统综述与理论参考。

关键词:遥感图像;遥感智能解译;预训练基础模型;多模态基础模型;通用预测;多任务

Development status and prospects of pretrained foundation models for remote sensing imagery

Zhi Yuanjie[1], Jiang Yiwei[1], Yang Zhi[2], Chen Yizhou[1], Hao Wenkui[2], Ma Mingyang[1], Wei Jiang[1], Mei Shaohui[1*]

1. School of Electronics and Information, Northwestern Polytechnical University, Xi'an　710129, China;

2. State Grid Electric Power Research Institute Co., Ltd., Beijing　100100, China

Abstract: Given the continuous expansion of training datasets and the rapid evolution of deep learning architectures, vision foundation models and large language models have demonstrated remarkable generalization and adaptability across diverse downstream tasks, thereby drawing increasing attention from the research community. Within the domain of remote sensing (RS), data exhibit significant heterogeneity across multiple sources, modalities, spatial scales, and temporal dimensions. Designing pretrained RS foundation models (RSFMs) capable of effectively capturing such complex geospatial dependencies is critical for robust feature representation and intelligent interpretation of RS imagery. This paper presents a

收稿日期:2025-09-08;**修回日期**:2025-11-03;**预印本日期**:2025-11-10

*通信作者:梅少辉　meish@nwpu.edu.cn

基金项目:国家自然科学基金项目(62171381, 62571442)

Supported by: National Natural Science Foundation of China(62171381, 62571442)

comprehensive review of the recent progress in pretraining strategies for RSFMs by emphasizing unimodal and multimodal learning paradigms. For unimodal models, representative frameworks based on self-supervised contrastive learning and masked image modeling are summarized. They leverage large-scale optical, hyperspectral, and radar imagery to learn transferable visual representations. These pretraining methods substantially enhance downstream performance in land cover classification, object detection, semantic segmentation, and change detection tasks. For multimodal models, we analyze the integration of image-text, image-location, and image-audio modalities through contrastive alignment strategies and cross-modal embedding learning, thereby effectively improving semantic coherence, generalizability, and interpretability in geospatial representation learning. Furthermore, widely adopted RS pretraining datasets, including their data sources, modality compositions, spatial resolutions, and annotation characteristics, are systematically summarized in this paper. Representative datasets, such as BigEarthNet, SEN12MS, and SkySenseGPT, are reviewed to demonstrate the diversity and scale of existing data resources. The importance of building open, standardized, and reproducible data repositories is emphasized, as these datasets serve as the foundation for training scalable and generalizable RSFMs. From a methodological perspective, this paper discusses the major pretraining paradigms that have shaped the current landscape of RSFMs, including contrastive self-supervised learning, generative self-supervised learning, and hybrid teacher-student distillation. These paradigms aim to maximize representational consistency between augmented views, reconstruct masked information, and align intermediate features between models, thereby enabling the extraction of semantically rich and transferable geospatial features. Despite these advances, several challenges remain unresolved in the development of RSFMs. Data-related issues, such as the scarcity of well-annotated multimodal datasets, geographic and temporal imbalance, and high acquisition costs, continue to hinder large-scale model training. Model scalability poses another limitation, as the billion-parameter-level architectures demand extensive computational resources and energy consumption during training and inference. Moreover, current RSFMs often suffer from limited cross-domain and cross-sensor generalization, thereby leading to performance degradation when applied to new regions or modalities. Transparency and interpretability also remain pressing concerns, as understanding the internal mechanisms of deep RSFMs and improving their robustness against adversarial perturbations are essential for reliable real-world deployment. Future research may address these challenges by focusing on developing scalable multimodal architectures that can jointly process optical, synthetic aperture radar, hyperspectral, and textual data, as well as by designing lightweight RSFMs through model compression, sparse training, and modular architecture optimization. Improving cross-domain and cross-temporal generalization by incorporating domain adaptation, meta-learning, and transfer learning techniques will further enhance model robustness under diverse acquisition conditions. In addition, integrating explainable artificial intelligence approaches, uncertainty quantification, and attention-based visualization can improve the interpretability and trustworthiness of RSFMs, thereby enabling their safe application in operational RS systems. Overall, this paper provides a systematic and forward-looking overview of the current development status, pretraining methodologies, benchmark datasets, and existing challenges of RSFMs. This work aims to offer a theoretical and methodological reference for the future construction of intelligent, scalable, and trustworthy foundation models in the RS domain by consolidating advances in unimodal and multimodal pretraining paradigms.

Key words: remote sensing images; remote sensing intelligent interpretation; pre-trained basic model; multimodal basic model; general prediction; multi-tasking

0 引　言

近年来,传感器技术与遥感载荷平台的快速发展,共同推动了高分辨率空间遥感技术的进步(Huang等,2025)。遥感影像通常由航空平台或地球卫星通过俯拍获取,具有覆盖地域广、分辨率高、空间位置信息丰富以及纹理结构复杂等特征(Dias等,2023)。在此背景下,依赖人工设计的传统方法在遥感影像信息提取中逐渐面临瓶颈:不仅需要消耗大量人力与时间成本,其效果与性能也难以保证(郭园方 等,2025)。作为机器学习的重要分支,深度学习通过构建多层神经网络模拟人脑的学习机制,能够自动实现影像特征的学习与提取(张帅豪和潘志刚 等,2025)。随着深度学习技术的迭代发展,神经网络的层数和神经元数量持续增加,模型规模

不断扩大，在图像分类、目标检测（Zhang 等，2024c）和图像分割（Chen 等，2024）等遥感视觉任务中展现出优异性能。基于上述优势，深度学习模型逐渐成为遥感影像解析与处理的核心技术手段（韦炎炎等，2025），在灾害监测（Tan 等，2023）、城市规划建设以及交通管理和国防安全等领域（燕琴 等，2024）发挥着关键作用。

目前，有许多学者从事遥感领域的模型设计工作，针对不同遥感影像处理任务，研发出了海量独立的深度学习模型（Li 等，2024a）。然而，这些模型的应用都局限于特定的平台、任务与数据，它们虽然在各自的任务中性能表现优异但却无法泛化迁移到其他解析任务中，利用率低下且浪费计算资源。与传统深度学习模型不同，基础模型可以在大量数据训练、强大模型架构和有效学习算法的共同作用下，同时处理多种任务，具有更强大的表达能力和泛化能力（黎宇哲 等，2025b）。

具体而言，在模型的体量与特征表达上，基础模型可以更好地发掘复杂数据中的内在信息联系，学习更丰富更通用的语义和特征表示，从而提升模型的综合表现能力（Tao 等，2023a）；在训练数据与计算资源上，虽然基础模型需要更强大的计算设备和算力支持，但不同于传统深度学习需要数万乃至数十万精细数据训练（Osco 等，2023），基础模型结合自监督学习预训练，只需千百条精标数据即可实现模型微调，这在大规模数据集丰富但标注成本昂贵的遥感领域极其适用（Corley 等，2024）；在任务表现与可解释性上，虽然基础模型不利于对模型决策过程进行直观解释、分析和优化，但其能够适应不同数据分布并在许多下游任务上展现出相对较好的性能。

近年来，预训练基础模型在遥感领域备受关注，相关研究逐年增加，尤其在 2024 年与 2025 年，围绕通过挖掘遥感影像的地物关系与空间上下文分布提升遥感数据解译质量、解决多源数据融合与时空关联复杂问题的研究呈爆发式增长。

本文基于 Scopus 数据库统计了 2017 年 1 月至 2025 年 8 月间的相关文献，检索关键词为：TITLE-ABS-KEY（（“remote sensing” OR “earth observation”） AND （“foundation model” OR “pre-trained model” OR “pretrained model” OR “vision-language model” OR “multimodal mode” OR “large model”）），检索结果如图 1 所示。

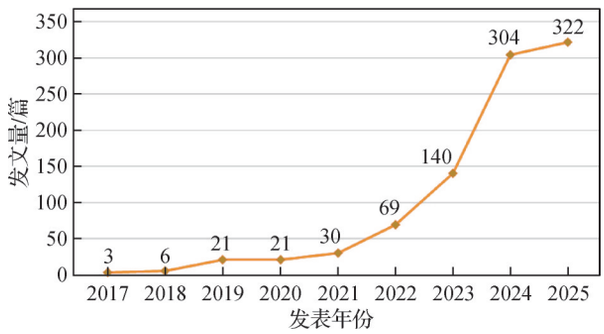

图 1　遥感预训练模型发文趋势
（2017 年 1 月至 2025 年 8 月）

Fig. 1　Publication trend of remote sensing pretraining models
(January 2017—August 2025)

为了全面介绍现有的遥感预训练基础模型的发展与训练方式、填补现有综述文献的时效性空缺，本文侧重选取近 3 年的预训练基础模型并进行综述分析。首先根据训练模态的类型，将遥感预训练模型分为单模态遥感预训练模型与多模态遥感预训练模型；进一步地，根据预训练方法的不同，将单模态模型分为自监督对比学习预训练与自监督掩码生成预训练，将多模态模型分为图像—文本模态预训练模型、图像—非文本模态预训练模型与遥感图像生成基础模型。随后，本文总结了当前用于遥感基础模型预训练的数据集及特点。最后，对遥感预训练基础模型发展提出了 4 点展望。

1　单模态遥感预训练模型

单模态遥感预训练模型专注于利用遥感图像的单一模态数据，通过自监督学习方法提取有用的特征表示，进而提升遥感图像的分析能力，在场景分类、目标检测和变化检测等任务中得到了广泛应用。本节根据预训练方法的不同将单模态遥感预训练模型分为对比式与生成式自监督学习预训练。

1.1　自监督对比学习预训练

对比学习通过对图像进行增强，生成多个视图，并通过对比损失来优化模型，使得模型能够学习到通用的、具有较好判别能力的特征，其预训练框架如图 2 所示（张良培 等，2023）。

针对高分系列卫星影像，Li 等人（2022）提出 GeoKR（geographical knowledge-driven representation），通过对齐现成地理知识表示和遥感图像表示来完成网络预训练。基于均值教师网络的高效预训

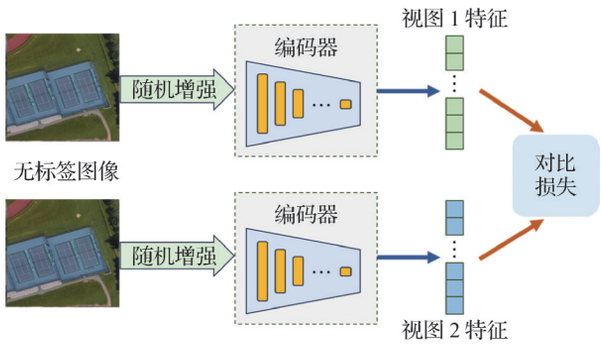

图 2　对比式自监督学习预训练框架

Fig. 2　Contrastive self-supervised pretraining framework

练框架减轻不同类型知识差异导致的噪声标签影响；该网络在场景分类、语义分割、目标检测和云/雪检测下游任务评估中，GeoKR 框架性能表现较好。针对中低分辨率多光谱卫星影像，Yao 等人（2021）提出 SeCo（sequence contrastive learning）模型，创新性地利用季节变化提供的信息学习可迁移的遥感特征表示，将对比自监督学习方法与卫星提供的时间信息相结合，学习泛化能力良好的时不变视觉表示，在 BigEarthNet 和 EuroSAT 土地覆盖分类、OSCD（one-shot conditional detection framework）变化检测下游任务性能表现优异。针对高分辨率遥感图像，Wang 等人（2023b）提出 GLCNet（global-local complementary network）模型，利用全局风格对比模块和局部匹配对比模块，将对比自监督学习方法与图像的全局风格特征和局部区域特征相结合，学习适用于语义分割的特征表示，以实现对高分辨率遥感图像的语义分割。在 ISPRS Potsdam（international society for photogrammetry and remote sensing potsdam）、DeepGlobe 等多个数据集的土地覆盖分类等下游语义分割任务中，该模型表现优异，尤其在仅使用少量原始数据集标签的情况下，性能提升显著。

1.2　自监督掩码生成预训练

生成学习通过学习图像的潜在分布来重建图像内容，通常使用编码器—解码器结构，通过生成模型重建图像，进而提取图像的鲁棒特征表示，生成式自监督学习预训练框架如图 3 所示（付琨 等，2024）。

面向 RGB 遥感图像，Wang 等人（2023a）设计约百万参数的 Transformer 基础模型，提出 RVSA（reweighted Volterra series algorithm），针对目标旋转角度多样的问题，在目标检测、场景分类和语义分割 3 种下游任务进行测试，相较普通 ViT（vision Transformer）性能更优越。Sun 等人（2023）提出 RingMo 模型，构建共包含 200 万幅遥感图像在内的预训练数据集，一定程度上解决了复杂遥感场景中容易忽略密集分布和小物体目标的问题。Cha 等人（2024）提出了一种针对遥感图像的基础模型 Billion-scale FM（billion-scale foundation model），其具备处理大规模遥感数据集的能力。该模型利用海量遥感数据进行预训练，学习了具有强大泛化能力的特征表示，可以应用于多种遥感任务。模型的设计充分考虑了遥感数据的多样性与复杂性，表现出在多个遥感影像任务（如分类、变化检测等）上的卓越性能，尤其在大规模数据处理方面显示了显著优势。Mendieta 等人（2023）提出的 GFM（geospatial foundation model）则采用蒸馏 + 自监督掩码生成的训练策略，即构建教师—学生架构并结合图像建模和特征蒸馏损失的训练方法，实验表明，这种训练方法得出的模型在变化检测、分类、语义分割和超分辨率等任务上均取得较大提升。

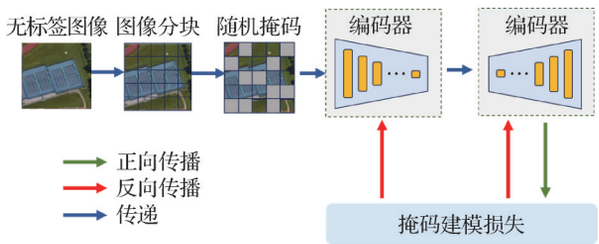

图 3　生成式自监督学习预训练框架

Fig. 3　Generative self-supervised pretraining framework

针对时间序列和多光谱卫星图像，Cong 等人（2022）提出了 SatMAE（satellite imagery based on masked autoencoder）模型，采用预训练 Transformer 架构。在训练时，运用自监督的掩码自编码器方法，对卫星图像的部分区域进行随机掩码，然后让模型学习恢复被掩码区域的信息，从而有效挖掘卫星图像的时间维度特征和多光谱特征，为后续的卫星图像分析任务奠定了良好的特征学习基础，在相关应用中取得了较好的性能表现。进一步地，Noman 等人（2024）提出的 SatMAE++ 在 SatMAE 模型训练方法上进行了改进和创新。通过优化预训练的策略和过程，更好地捕捉多光谱卫星图像中复杂的光谱信息和空间信息之间的关系，提升了 Transformer 模型在多光谱卫星图像任务中的特征提取能力和泛化能力，使得模型在多光谱卫星图像的分类、目标检测等

应用中能够发挥更出色的性能。

面向高光谱图像数据，Wang 等人（2025）提出 HyperSIGMA（hyperspectral sigma）模型，并且构建全局高光谱图像数据集 HyperGlobal-450K。HyperSIGMA 采用通道选择策略确保训练效率和提高数据丰富度，通过两个并行子网络提取空间和光谱特征，参数超过 10 亿，且可应对低层和高层图像解译任务，在高光谱分类、目标/变化检测等高层任务和光谱解混、图像去噪/超分辨低层任务性能领先。Hong 等人（2024）提出 SpectralGPT（spectral generative pretrained transformer），以渐进式训练方式充分利用不同尺寸、分辨率、时间序列和地理区域的光谱遥感大数据，在百万幅图像上训练生成参数超过 6 亿的基础模型，在单/多标签场景分类、语义分割和变化检测下游任务性能领先。

综上所述，单模态遥感预训练模型具有任务聚焦明确、数据处理流程简便等优点，因此在一些特定应用场景下，单模态研究有其独特的价值和意义。但是单模态遥感预训练模型在处理遥感数据时主要依赖单一类型的数据，导致其感知能力受限、泛化能力不足。正因如此，近年来遥感预训练模型研究正朝着多模态融合的方向发展。

2 多模态遥感预训练模型

如前所述，单一图像模态下的遥感视觉基础模型相对忽视对象间语义关系理解。随着人工智能技术的发展与数据集的日益完善，研究人员提出许多多模态预训练策略，多模态预训练有助于细粒度重识别典型目标并推断它们之间的关系，生成有关地物场景的自然语言描述，可服务于跨模态检索、目标计数、视觉描述与问答和图像字幕生成等下游任务。本节将多模态遥感预训练模型分为图像—文本、图像—非文本和生成基础模型 3 类。

2.1 图像—文本模态预训练模型

近年来，图像—文本模型得到了飞速发展，能够处理来自视觉和语言的多模态信息，用于图像描述、图像检索和视觉问答等任务。图像—文本模态预训练模型是通过联合学习图像和文本信息提升模型理解视觉与语言之间语义关系的能力，通用训练框架如图 4 所示。这类模型通过将图像数据和与之相关的自然语言描述进行对齐，从而使得模型能够生成更准确的图像描述、实现图像检索或回答与图像相关的问题。

在遥感图像语义理解与标签生成领域，诸多模型致力于结合视觉与语言信息，对遥感图像开展深入的语义层次分析，以实现精准的目标检测、物体分类和场景解析等关键任务。Bazi 等人（2024）提出的 RS-LLaVA（remote sensing-large language and vision assistant）专注于联合完成遥感图像的字幕生成与问答任务，通过视觉与语言模态的融合，深入理解图像语义，进而生成贴合图像内容的文本描述并准确回

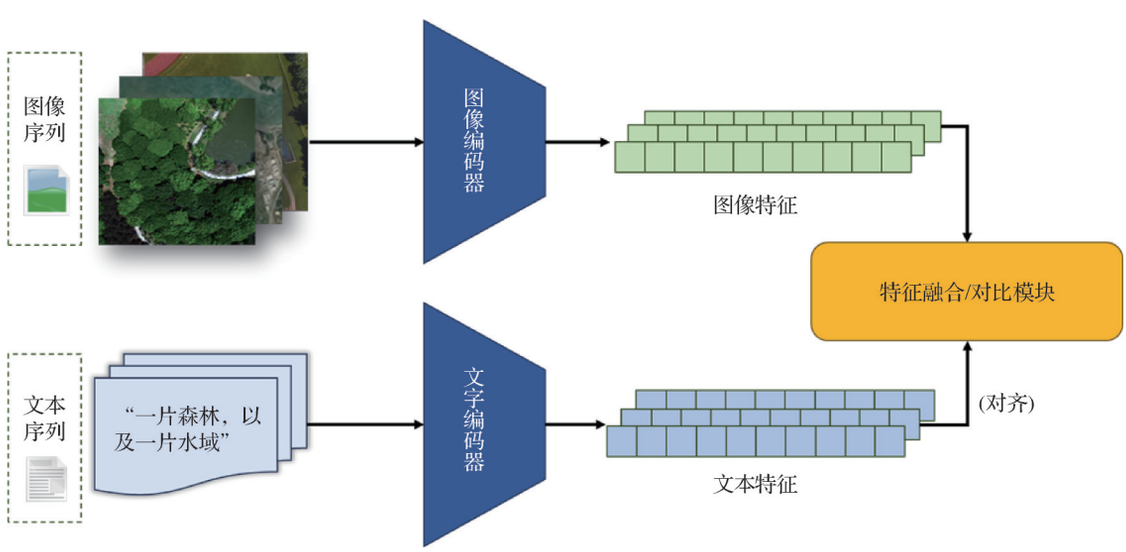

图 4 图像—文本预训练模型框架

Fig. 4 Image-text pretraining model framework

应相关问题。SkySenseGPT构建大规模指令调优数据集,以此为基础训练模型,显著提升对遥感视觉—语言的理解能力,在复杂的遥感视觉—语言任务中,凭借对图像语义的深度剖析,展现出卓越性能(Luo等,2024b)。Lu等人(2024)提出的Aquila模型则通过构建层次对齐的视觉—语言模型,全方位增强对遥感图像的理解,在结合视觉与语言信息的过程中,实现对图像更全面、细致的语义解读,为相关领域的研究与应用提供有力支持。这些模型的研究成果,极大地推动了遥感图像语义理解与标签生成技术的发展,为后续深入挖掘遥感图像信息奠定了坚实基础。

在地球观测数据处理领域,传统方法在处理多模态、多时相数据时面临诸多挑战(Zhou等,2024)。例如,难以在统一框架下高效处理不同类型视觉输入(Xiao等,2025),对时间序列数据的分析能力有限,且缺乏有效的交互式对话功能,无法满足日益增长的环境监测、城市发展分析等实际应用需求(Zhang等,2024a)。为应对这些问题,Soni等人(2025)提出的EarthDial通过引入大规模指令微调数据集,涵盖多种光谱模态,设计能支持多光谱、多时相和多分辨率图像的架构,将多传感器地球观测数据转化为交互式自然语言对话,有效解决了数据交互和多模态信息处理难题,实现了多种遥感任务。TEOChat(temporal earth observation chat)构建了包含丰富单图像和时间相关任务的TeoChatAtlas数据集,通过低秩适应技术微调大型语言模型,成功实现对地球观测数据时间序列的处理和对话功能,在时间场景分类、变化检测等任务中表现出色(Irvin等,2025)。Li等人(2024b)提出的UniRS(universal remote sensing)则采用统一的视觉表示方法,设计定制的变化提取模块和提示增强机制,并在混合数据集上联合微调,统一了多时态遥感任务处理,在视觉问答、变化描述等任务中展现出强大的通用性和有效性。

在图像精细理解领域,现有的遥感大模型缺乏图像精细理解的功能,具体而言,由于缺乏精细的、特定于遥感领域的数据,遥感大模型的对话能力的发展受到阻碍。为了突破这些局限,Liu和Lian(2024)提出一种统一的端到端RSUniVLM(remote sensing universal vision-language model)模型,提升了像素级任务的准确度。进一步地,Shabbir等人(2025)提出的GeoPixel通过在对话中生成交错掩码实现精细的视觉感知,在像素级理解方面表现出卓越的性能,在单目标和多目标分割任务中都取得了较大的提升;类似地,Ou等人(2025)为多模态大型语言模型配备掩码预测器,将视觉编码器的视觉特征转换为以大语言模型的分割标记嵌入为条件的掩码,提出了GeoPix,将图像理解功能扩展到像素级。

2.2 图像—非文本预训练模型

图像—文本任务尽管在遥感图像分析中有广泛的应用,但也存在一些局限性,如图5所示。这些局限性使得图像—非文本预训练成为一个更具吸引力的选择。具体而言,局限性可以概括为:1)文本生成的复杂性与不确定性。图像—文本任务依赖于将图像内容与文本描述对齐,而文本的生成往往带有一定的主观性和多样性(Xu等,2023)。尤其是在遥感图像的背景下,地物的描述可能因为不同的专业领域、语言差异或文化背景而有所不同,这增加了模型理解图像的难度(Mai等,2025)。此外,文本描述的质量直接影响任务的效果,生成不准确或不充分的文本描述会导致模型无法很好地理解图像内容。2)文本的稀缺性与标注成本。图像—文本任务通常依赖大量的标注文本数据,而这种数据的获取往往需要人工标注或依赖高质量的描述性数据集(Hoxha等,2024)。在遥感领域,生成准确的文本描述往往需要专业人员的参与,这对于大规模数据集来说是非常昂贵且耗时的。而且很多遥感图像数据集中可能缺少足够的文本描述,因此难以进行有效的训练。

针对以上问题,学者们提出了通用的图像—非文本模型训练框架,如图6所示。具体而言,Mai等人(2023)率先提出利用对比学习从图像中挖掘位置表示的模型CSP(contextual spatial pyramid),通过双编码器分别对图像及其对应的地理位置进行编码,利用对比目标从图像中学习有效的位置表示。

受CLIP(contrastive language-image pre-training)(Radford等人,2021)启发,Vivanco等人(2023)提出的GeoCLIP从全球地理定位的角度,通过构建随机傅里叶特征进行位置编码,利用位置编码器的层次化表示来解决地理定位问题,进一步拓展了图像—位置关系的应用场景。类似地,Klemmer等人(2025)提出的SatCLIP专注于利用卫星图像学习通用的位置嵌入,在更广泛的与位置相关任务中进行

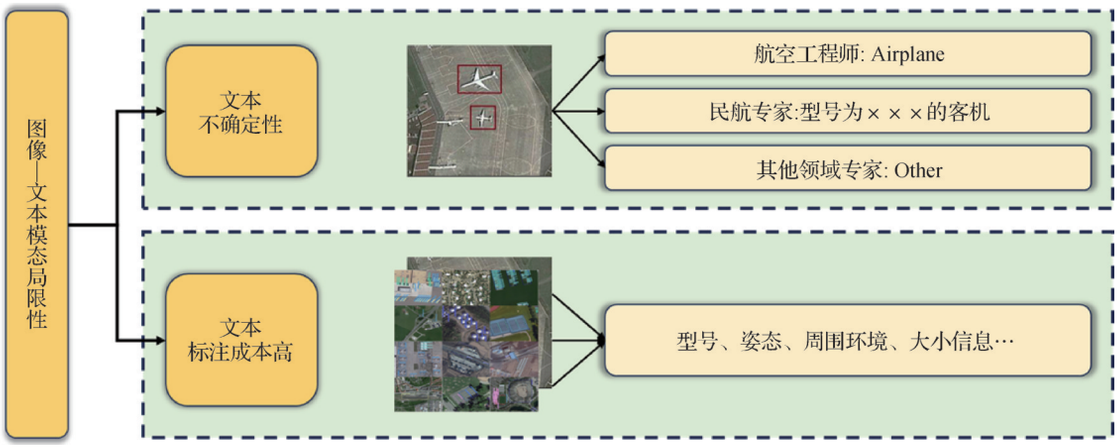

图5　图像—文本模态数据集局限性

Fig. 5　Limitations of image-text modal dataset

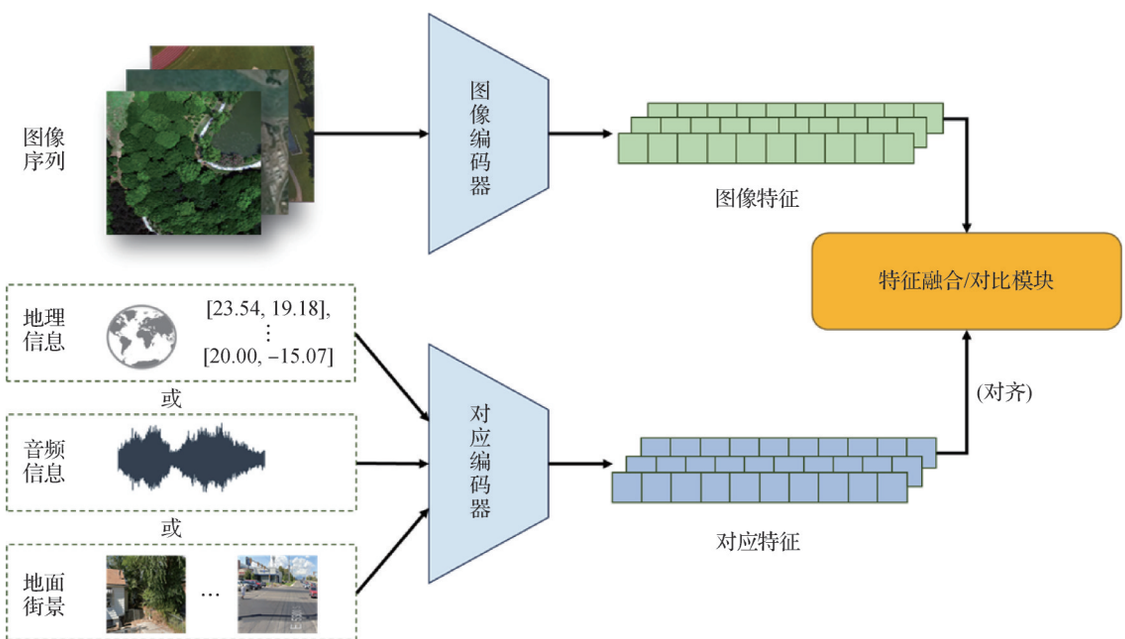

图6　图像—非文本模型训练框架

Fig. 6　Image-non-text model training framework

验证，强调位置表示的通用性和泛化性。而 Haas 等人（2023）提出的 StreetCLIP 模型，从新的元学习视角出发，解决开放域图像地理定位中的零样本学习难题，在已有图像—位置学习框架下，通过创新学习方式提升模型在无样本或少量样本场景下的地理定位能力。在此基础上，RANGE（retrieval augmented neural fields for multi-resolution geo-embeddings）模型（Dhakal 等，2025）的作者分析了图像与地理位置对比学习中存在的视觉信息损失问题，引入了信息论解释 + 检索增强机制，将单点的地理—图像对比学习扩展为多点视觉信息融合，这一跨领域思路在分类与回归任务均表现提升。

除了采用图像—位置数据策略外，Heidler 等人（2023）提出了 AV-ResNet（attention-vision residual network），这是一种自监督的视觉—音频表示学习方法，用于遥感数据的分析。该研究利用自监督学习方法，通过结合视觉和音频信息，辅助完成遥感图像和视频的分析任务。不同于传统的遥感数据处理方法，作者提出通过视听信号的协同学习，提升了模型对遥感图像中的多模态数据的理解与表征能力。自监督学习使得模型能够在无监督标注的条件下，自动挖掘图像和声音之间的内在联系，进而更有效

地从遥感数据中学习空间和时间的特征。这一方法不仅在遥感数据分析中展示了良好的性能，也为多模态学习和遥感图像处理提供了新的思路。受此启发，GAIR（geospatial artificial intelligence and remote sensing）（Liu等，2025b）模型进一步扩展了多模态遥感基础模型的边界。该方法不仅融合了遥感影像和地理位置信息，还首次引入街景影像，并通过设计隐式神经表示模块实现遥感与街景影像的空间对齐，从而在跨模态和跨尺度的特征建模方面展现出显著优势。通过联合对比学习，GAIR能够有效整合空中与地面视角信息，提升了模型在多样化地理任务中的泛化性和可迁移性。这一创新为构建更全面、更鲁棒的遥感多模态基础模型提供了新的方向。

2.3 遥感图像生成基础模型

2021年1月，OpenAI公布了其首个文本生成图像模型DALL·E，引发了业内对文本生成图像的关注，同时也引发了学者对于遥感图像生成基础模型研究的探索。生成式遥感基础模型在遥感领域中扮演着至关重要的角色。其主要的作用在于：在数据样本有限的情况下，通过生成高质量、多类型的遥感图像扩充训练数据集，推动遥感技术的发展和应用。近年来，多数遥感图像生成基础模型主要依靠U-Net架构的多轮稳定扩散模型实现图像生成，本文总结其通用结构，如图7所示。具体而言，Yu等人（2025）设计了一种新的噪声采样策略，用于去噪扩散模型，提出的MetaEarth模型实现了无界和任意大小的图像生成；类似地，Liu等人（2025a）利用体量更大的Git-10M数据集，提出了Text2Earth，支持生成RGB，SAR（synthetic aperture radar），PAN（panchromatic image）等模态的图像，并支持进行图像编辑与相互转化，在可控性和图像质量等多个任务中取得较大的提升；遗憾的是，该模型并未涉及高光谱图像。面向高光谱图像生成领域的空缺，Pang等人（2024）提出HSIGene（hyperspectral image generation），允许生成更精确和可靠的高光谱图像。在输入方式上，Tang等人（2024）通过引入一种新的条件控制机制，实现多尺度特征融合，实现了支持文本条件、元数据条件和图像条件控制输入的CRS-Diff（controllable remote sensing diffusion）模型，一定程度上解决了因输入形式单一而导致图像生成不准确的问题。

未来生成式遥感模型的应用潜力仍然巨大（Zhang等，2024b）。首先，随着数据的多样化和复杂化，生成式模型将在遥感数据的增强和样本扩充中发挥越来越重要的作用，尤其是在数据稀缺或难以获得的地区。其次，随着多模态输入的不断发展，我们期待生成式遥感模型能够整合不同类型的数据，如光学影像、SAR数据、地理元数据和文本描述，进一步提升生成图像的准确性与适用性。

3 遥感预训练数据集

随着近年来遥感大模型研究的不断深入，遥感领域的数据集逐渐趋于完善，涵盖了更为丰富的多样化数据源（Tao等，2025）。这些数据集不仅包括传统的光学图像，还逐步融入了合成孔径雷达、高光谱、红外及其他遥感数据类型，从而为大模型的训练

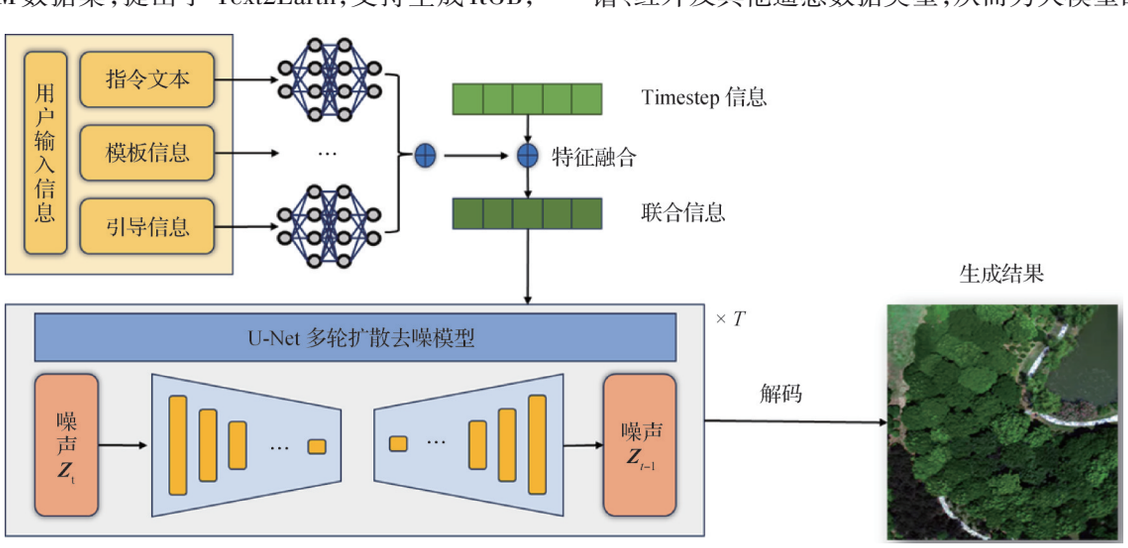

图7 生成基础模型训练框架

Fig. 7 Generative base model training framework

提供了更为全面的样本和信息支持。在数据属性的分类上，遥感数据集可以分为单模态与多模态两类。

3.1 单模态数据集

单模态数据集通常由同一种类的遥感图像构成，如光学影像或SAR图像，适用于针对特定数据类型的任务。为系统展示当前典型单模态遥感预训练数据集的组成与特征，表1对相关数据集的来源、规模、类型及主要应用方向进行了汇总与说明。

3.2 联合模态数据集

随着遥感数据融合技术的发展，联合模态数据集越来越受到重视，尤其是近两年来，优秀的联合模态遥感大模型预训练数据集大量涌现。为进一步概述多源、多模态融合背景下典型遥感预训练数据集的构建方式与应用范围，表2对近年来具有代表性的联合模态数据集进行了整理与说明。

多模态数据集集成了来自不同传感器和数据源的信息，结合了图像和文本数据，形成蕴含更多信息的图像—文本对数据集，为遥感大模型提供了更多维度的输入。这些多模态数据集能够提升模型对复杂场景的理解和生成能力，推动了遥感技术在更广泛应用领域的落地与发展。

4 展 望

4.1 遥感多维信息深度发掘与知识约束

1）高分辨率时空光谱视觉表示提取。针对基础模型涌现出的强大泛化能力与高信息容量，建立高分辨率遥感场景下的空间—时间—光谱多维度视觉表示体系，追求图像特征鲁棒性的同时探索跨模态数据新型利用方式。

2）地物关系通用知识与语义理解。基于不同传感器获取的多源异构数据和丰富信息模态类型，构建语义理解模型，将地物目标的视觉信息与场景本身的语义信息有机关联，结合数学物理建模与地理约束先验知识发掘模型学习通用本征特征的潜在能力。

表1 单模态数据集的特点及描述

Table 1 Characteristics and descriptions of single-modal datasets

数据集名称	数据类型	发布年份	特点及描述	下载地址
BigEarthNet_1.0（Sumbul等，2019）	多光谱	2019	提供丰富的光谱信息，用于土地覆盖分类、遥感图像场景分类等遥感图像理解任务。	https://github.com/ljl5261/MMM-RS
ImageNet-22k（Russakovsky等，2015）	RGB图像	2015	具有高分辨率和丰富细节，都经过专业人工标注，标注信息包括图像中物体的类别、位置等。	https://image-net.org/index.php
SSL4EO-L（Stewart等，2023）	多光谱	2023	专注Landsat图像的自监督学习数据集，推动遥感图像的无监督学习和大规模训练能力。	https://github.com/torchgeo/torchgeo
HYPERGLOBAL-450K（Wang等，2025）	高光谱	2024	覆盖全球多个地区，空间分辨率为30 m，涵盖森林、草原、裸地和农作物等多种地表类型。	https://github.com/WHU-Sigma/HyperSIGMA
SatlasPretrain（Bastani等，2023）	可见光	2023	结合了Sentinel-2和NAIP图像，具有137个类别和7种标签类型的302 M标签。	https://github.com/allenai/satlas
TOV-RS-Balanced（Tao等，2023b）	可见光	2023	专注遥感图像的自监督学习任务，利用视觉模型对光学遥感图像进行理解。	https://github.com/GeoX-Lab/G-RSIM/tree/main/TOV_v1
SEN12MS（Schmitt等，2019）	SAR/多光谱	2019	结合来自Sentinel-1/2卫星的多光谱图像数据，专用遥感图像的深度学习和分类任务。	https://opendatalab.org.cn/OpenDataLab/SEN12MS

表2 联合模态数据集的特点及描述

Table 2 Characteristics and descriptions of joint modal datasets

数据集名称	数据类型	发布年份	特点及描述	下载地址
RSICap & RSIEval	图像—文本对	2023	推动遥感视觉语言模型的研究，专注图像与语言的交互与理解。	https://github.com/Lavender105/RSGPT
SATIN（Roberts等，2023）	图像—文本对	2023	基于多任务学习的遥感图像分类数据集，结合了多种类型的卫星图像与文本描述。	https://hf-mirror.com/datasets/jonathan-roberts1/SATIN
SkyScript（Wang等，2024）	图像—文本对	2024	语义丰富的遥感图像与文本配对数据集，专注遥感任务中图像和语言的深度融合。	https://github.com/wangzhecheng/SkyScript
ChatEarthNet（Yuan等，2025）	图像—文本对	2024	全球范围内的遥感图像和文本数据集，专注遥感图像的深度学习任务。	https://github.com/zhu-xlab/ChatEarthNet
LuoJiaHOG（Zhao等，2025）	图像—文本对	2024	聚焦地理感知的图像和文本数据集，增强遥感图像的检索和文本理解能力。	https://arxiv.org/abs/2403.10887v1
FIT-RS（Luo等，2024b）	图像—文本对	2024	包含1 800 851个高质量指令样本，专注复杂远程感知场景中理解对象间语义关系。	https://github.com/Luo-Z13/SkySenseGPT
RS-GPT4V（Xu等，2024）	图像—文本对	2024	提供用于多模态指令任务的遥感图像和文本数据集，推动遥感领域的多模态学习。	https://github. com/GeoX-Lab/RS-GPT4V? tab=readme-ov-file
MMM-RS（Luo等，2024a）	图像—文本对	2024	多模态、多个分辨率的遥感数据集，侧重文本到图像的生成任务。	https://github.com/ljl5261/MMM-RS
DDFAV（Li等，2024a）	图像—文本对	2024	大规模遥感视觉语言数据集，专门用于遥感图像和文本的评估基准。	https://github.com/HaodongLi2024/rspope
VALOR-1M（Li等，2025a）	图像—文本—听觉	2025	大规模、高质量的三模态数据集，包含100万个带有人工注释的视听字幕的可听视频。	https://github.com/TXH-mercury/VALOR
GeoPixInstruct（Ou等，2025）	图像—文本对	2025	包括65 463幅图像和140 412个实例，实例具有文本描述、边界框和掩码进行注释。	https://github.com/Norman-Ou/GeoPix
GeoPixelD（Shabbir等，2025）	图像—文本对	2025	包含5 427个经过验证的参考表达—掩码对和61 384个标注对象，每个对象都配有平均647个字符的详细描述。	https://github.com/mbzuai-oryx/GeoPixel
REO-Instruct（Xue等，2024）	图像—文本对	2024	160万个多模态EO影像和语言对。	https://github. com/REO-VLM-anonymous/REO-VLM
SoundingEarth（Heidler等，2023）	图像—音频样本	2023	由全球范围内共定位的航空影像和音频样本组成，实现了视听数据的结合。	https://github.com/khdlr/SoundingEarth

4.2 遥感基础模型训练数据的规模化构建与标注

1）低成本大规模遥感数据集合。基于先进星载/机载遥测设备和生成式模型等技术提高训练数据体量，同时建立公共开源的数据平台和统一规范的格式标准，有助于缓解基础数据匮乏问题，促进学术界和工业界在遥感预训练模型发展上的合作与交流。然而，目前遥感数据存在获取成本高、隐私安全限制严和跨区域分布不均等难点，这对构建真正大规模、高覆盖的数据集提出了挑战。如何在保证数据多样性的同时降低构建与存储成本，将成为未来

的关键趋势。

2)高质量文本—影像跨模态标注。创建公开规范的标注平台并明确物体识别、场景描述等各项标注任务，设计具备准确性、一致性和完整性的标注评估准则，邀请具有相关领域知识的人员定期进行质量检查和审核，从数据源角度提高遥感基础模型的训练效果。

4.3 遥感基础模型架构创新与训练范式转变

1)模型架构与核心原理研究创新。在当前有监督和自监督训练范式、对比学习和生成式学习框架、Transformer 和 Mamba 网络结构的基础上进一步创新，提高遥感基础模型在单/多标签分类等图像级、检测分割等像素级下游任务上的综合性能和效率。

2)增强模型透明度和可解释性。通过蒸馏或重设计等方式优化模型网络结构，有助于阐释编解码特征解译表达、评估模型预测结果时的全面性能和分析模型受到对抗性样本攻击时的稳定性，进一步推动遥感基础模型成为通用方法。然而，大模型的“黑箱”特性依然严重制约其在高风险应用中的推广，如何实现可解释性与性能之间的平衡，是未来研究的重要方向。

3)特定场景领域模型定制化设计。针对具体下游任务和数据模式设计特定的遥感基础模型，有助于降低模型对于海量优质数据的依赖、训练推理过程中的算力和显存成本开销，并提高该垂直领域直接关切指标的表现能力。但这类定制化设计往往牺牲通用性，如何实现“通用模型”与“专用模型”的协同发展，将成为遥感基础模型演进中的关键问题。

4.4 遥感基础模型技术落地与新质应用

1)构建遥感应用生态与统一评估体系。针对农业、城市规划和环境监测等实际应用领域，从定向数据获取与模型优化、客户端需求反馈等角度构建生态，推广开发并不断完善遥感预训练模型应用，通过校企合作积极探索新型应用场景。

2)云端平台搭建与端侧加速部署。充分发挥遥感基础模型的实际价值，以独立应用程序或软件插件接口等形式提供检索查找、定位跟踪和探测评价等多方面服务，提升自然资源管理、人员搜救以及农业估产等遥感细分应用任务的精确程度和分析效率。

3)数据泄露风险规避与隐私安全保障。结合遥感基础模型数据获取与处理、训练测试与部署、产品应用与维护 3 个不同阶段的生命周期，设计具备前瞻性的安全框架，进一步缓解数据安全、隐私泄露以及抗攻击能力提升等现存问题。

5 结 语

遥感基础大模型的兴起为遥感智能处理提供了新范式，也为遥感解译在多源异构数据场景下的泛化能力提升带来了新机遇。本文首先梳理了单模态遥感预训练模型与多模态遥感预训练模型的研究现状，并分析了训练方法；然后，总结了目前常用于遥感预训练的单模态与多模态数据集，并分析其特点；最后，对未来遥感预训练模型的发展做出了研究展望，供学者们参考。

参考文献(References)

Bastani F, Wolters P, Gupta R, Ferdinando J and Kembhavi A. 2023. SatlasPretrain: a large-scale dataset for remote sensing image understanding//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 16726-16736 [DOI: 10.1109/ICCV51070.2023.01538]

Bazi Y, Bashmal L, Al Rahhal M M, Ricci R and Melgani F. 2024. RS-LLaVA: a large vision-language model for joint captioning and question answering in remote sensing imagery. Remote Sensing, 16(9): #1477 [DOI: 10.3390/rs16091477]

Cha K, Seo J and Lee T. 2024. A billion-scale foundation model for remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing: 1-17 [DOI: 10.1109/JSTARS.2024.3401772]

Chen K Y, Liu C Y, Chen H, Zhang H T, Li W Y, Zou Z X, et al. 2024. RSPrompter: learning to prompt for remote sensing instance segmentation based on visual foundation model. IEEE Transactions on Geoscience and Remote Sensing, 62: #4701117 [DOI: 10.1109/TGRS.2024.3356074]

Cong Y Z, Khanna S, Meng C L, Liu P, Rozi E, He Y T, et al. 2022. SatMAE: pre-training transformers for temporal and multi-spectral satellite imagery//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: ACM: #15

Corley I, Robinson C, Dodhia R, Lavista Ferres J M and Najafirad P. 2024. Revisiting pre-trained remote sensing model benchmarks: resizing and normalization matters//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3162-3172 [DOI: 10.1109/CVPRW63382.2024.00322]

Dhakal A, Sastry S, Khanal S, Ahmad A, Xing E and Jacobs N. 2025. RANGE: retrieval augmented neural fields for multi-resolution geo-embeddings//Proceedings of 2025 Computer Vision and Pattern

Recognition Conference. Nashville, USA: IEEE: 24680-24689 [DOI: 10.1109/CVPR52734.2025.02298]

Dias P, Potnis A, Guggilam S, Yang L X, Tsaris A, Medeiros H, et al. 2023. An agenda for multimodal foundation models for earth observation//2023 IEEE International Geoscience and Remote Sensing Symposium. Pasadena, USA: IEEE: 1237-1240 [DOI: 10.1109/IGARSS52108.2023.10282966]

Fu K, Lu W X, Liu X Y, Deng C B, Yu H F and Sun X. 2024. A comprehensive survey and assumption of remote sensing foundation modal. National Remote Sensing Bulletin, 28(7): 1667-1680 (付琨, 卢宛萱, 刘小煜, 邓楚博, 于泓峰, 孙显. 2024. 遥感基础模型发展综述与未来设想. 遥感学报, 28(7): 1667-1680) [DOI: 10.11834/jrs.20233313]

Guo Y F, Yu Z T, Liu A S, Zhou W B, Qiao T, Li B, et al. 2025. Recent progress of the security research for multimodal large models. Journal of Image and Graphics, 30(6): 2051-2081 (郭园方, 余梓彤, 刘艾杉, 周文柏, 乔通, 李斌, 等. 2025. 多模态大模型安全研究进展. 中国图象图形学报, 30(6): 2051-2081) [DOI: 10.11834/jig.250067]

Haas L, Alberti S and Skreta M. 2023. Learning generalized zero-shot learners for open-domain image geolocalization [EB/OL]. [2025-09-08]. https://arxiv.org/pdf/2302.00275.pdf

Heidler K, Mou L C, Hu D, Jin P, Li G Y, Gan C, et al. 2023. Self-supervised audiovisual representation learning for remote sensing data. International Journal of Applied Earth Observation and Geoinformation, 116: #103130 [DOI: 10.1016/j.jag.2022.103130]

Hong D F, Zhang B, Li X Y, Li Y X, Li C Y, Yao J, et al. 2024. SpectralGPT: spectral remote sensing foundation model. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(8): 5227-5244 [DOI: 10.1109/TPAMI.2024.3362475]

Hoxha G, Sumbul G, Henkel J, Möllenbrok L and Demir B. 2024. Annotation cost-efficient active learning for deep metric learning-driven remote sensing image retrieval. IEEE Transactions on Geoscience and Remote Sensing, 62: #5636211 [DOI: 10.1109/TGRS.2024.3438430]

Huang Z Y, Yan H X, Zhan Q Q, Yang S, Zhang M M, Zhang C K, et al. 2025. A survey on remote sensing foundation models: from vision to multimodality [EB/OL]. [2025-09-08]. https://arxiv.org/pdf/2503.22081.pdf

Irvin J A, Liu E R, Chen J C, Dormoy I, Kim J, Khanna S, et al. 2025. TEOChat: a large vision-language assistant for temporal earth observation data//Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: OpenReview.net

Klemmer K, Rolf E, Robinson C, Mackey L and Rußwurm M. 2025. SatCLIP: global, general-purpose location embeddings with satellite imagery//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Pennsylvania, USA: AAAI: 4347-4355 [DOI: 10.1609/aaai.v39i4.32457]

Li H D, Zhang X F and Qu H C. 2025a. DDFAV: remote sensing large vision language models dataset and evaluation benchmark. Remote Sensing, 17(4): #719 [DOI: 10.3390/rs17040719]

Li W Y, Chen K Y, Chen H and Shi Z W. 2022. Geographical knowledge-driven representation learning for remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 60: #5405516 [DOI: 10.1109/TGRS.2021.3115569]

Li X, Wen C C, Hu Y, Yuan Z H and Zhu X X. 2024a. Vision-language models in remote sensing: current progress and future trends. IEEE Geoscience and Remote Sensing Magazine, 12(2): 32-66 [DOI: 10.1109/MGRS.2024.3383473]

Li Y J, Xu W J, Li G Z, Yu Z J, Wei Z W, Wang J N, et al. 2024b. UniRS: unifying multi-temporal remote sensing tasks through vision language models [EB/OL]. [2025-09-08]. https://arxiv.org/pdf/2412.20742.pdf

Li Y Z, Fu L, Zhu L H, Luo Q D and Tu L. 2025b. Multimodal large model-based method for generating visual Q&A data for electronic document images. Journal of Image and Graphics, 30(9): 3083-3096 (黎宇哲, 伏凌, 朱泠皞, 罗琪峘, 涂来. 2025b. 多模态大模型面向电子文档视觉问答的数据生成. 中国图象图形学报, 30(9): 3083-3096) [DOI: 10.11834/jig.240610]

Liu C Y, Chen K Y, Zhao R, Zou Z X and Shi Z W. 2025a. Text2Earth: unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model. IEEE Geoscience and Remote Sensing Magazine, 13(3): 238-259 [DOI: 10.1109/MGRS.2025.3560455]

Liu J, Chen S H, He X J, Guo L T, Zhu X X, Wang W N, et al. 2025b. VALOR: vision-audio-language Omni-perception pretraining model and dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(2): 708-724 [DOI: 10.1109/TPAMI.2024.3479776]

Liu X and Lian Z H. 2024. RSUniVLM: a unified vision language model for remote sensing via granularity-oriented mixture of experts [EB/OL]. [2025-09-08]. https://arxiv.org/pdf/2412.05679.pdf

Lu K X, Zhang R Q, Huang X and Xie Y X. 2024. Aquila: a hierarchically aligned visual-language model for enhanced remote sensing image comprehension [EB/OL]. [2025-09-08]. https://arxiv.org/pdf/2411.06074.pdf

Luo J L, Wang Y Z, Gu Z Q, Qiu Y D, Yao S Z, Wang F Y, et al. 2024a. MMM-RS: a multi-modal, multi-GSD, multi-scene remote sensing dataset and benchmark for text-to-image generation//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: ACM: #388

Luo J W, Pang Z, Zhang Y J, Wang T Z, Wang L L, Dang B, et al. 2024b. SkySenseGPT: a fine-grained instruction tuning dataset and model for remote sensing vision-language understanding [EB/OL]. [2025-09-08]. https://arxiv.org/pdf/2406.10100.pdf

Mai G, Xie Y Q, Jia X W, Lao N, Rao J M, Zhu Q, et al. 2025. Towards the next generation of geospatial artificial intelligence. International Journal of Applied Earth Observation and Geoinformation, 136: #104368 [DOI: 10.1016/j.jag.2025.104368]

Mai G C, Lao N, He Y T, Song J and Ermon S. 2023. CSP: self-supervised contrastive spatial pre-training for geospatial-visual rep-

resentations//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: ACM: #981

Mendieta M, Han B R, Shi X J, Zhu Y and Chen C. 2023. Towards geospatial foundation models via continual pretraining//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 16760-16770 [DOI: 10.1109/ICCV51070.2023.01541]

Noman M, Naseer M, Cholakkal H, Anwar R M, Khan S and Khan F S. 2024. Rethinking transformers pre-training for multi-spectral satellite imagery//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 27811-27819 [DOI: 10.1109/CVPR52733.2024.02627]

Osco L P, de Lemos E L, Gonçalves W N, Ramos A P M and Marcato Junior J. 2023. The potential of visual chatGPT for remote sensing. Remote Sensing, 15(13): #3232 [DOI: 10.3390/rs15133232]

Ou R Z, Hu Y, Zhang F, Chen J X and Liu Y. 2025. GeoPix: a multimodal large language model for pixel-level image understanding in remote sensing. IEEE Geoscience and Remote Sensing Magazine, 13(3): 324-337 [DOI: 10.1109/MGRS.2025.3560293]

Pang L, Cao X Y, Tang D T, Xu S, Bai X R, Zhou F, et al. 2024. HSIGene: a foundation model for hyperspectral image generation [EB/OL]. [2025-09-08]. https://arxiv.org/pdf/2409.12470.pdf

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 8748-8763

Roberts J, Han K and Albanie S. 2023. SATIN: a multi-task metadataset for classifying satellite imagery using vision-language models [EB/OL]. [2025-09-08]. https://arxiv.org/pdf/2304.11619.pdf

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S A, et al. 2015. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115(3): 211-252 [DOI: 10.1007/s11263-015-0816-y]

Schmitt M, Hughes L H, Qiu C P and Zhu X X. 2019. SEN12MS — A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion [EB/OL]. [2025-09-08]. https://arxiv.org/pdf/1906.07789.pdf

Shabbir A, Zumri M, Bennamoun M, Rahaman M S, Zhou J and Zhang L. 2025. GeoPixel: pixel grounding large multimodal model in remote sensing//Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: OpenReview.net

Soni S, Dudhane A, Debary H, Fiaz M, Munir M A, Danish M S, et al. 2025. EarthDial: turning multi-sensory earth observations to interactive dialogues//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14303-14313 [DOI: 10.1109/CVPR52734.2025.01334]

Stewart A J, Lehmann N, Corley I A, Wang Y, Chang Y C, Ait N, et al. 2023. SSL4EO-L: datasets and foundation models for landsat imagery//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: ACM: #2612

Sumbul G, Charfuelan M, Demir B and Markl V. 2019. Bigearthnet: a large-scale benchmark archive for remote sensing image understanding//Proceedings of 2019 IEEE International Geoscience and Remote Sensing Symposium. Yokohama, Japan: IEEE: 5901-5904 [DOI: 10.1109/IGARSS.2019.8900532]

Sun X, Wang P J, Lu W X, Zhu Z C, Lu X N, He Q B, et al. 2023. RingMo: a remote sensing foundation model with masked image modeling. IEEE Transactions on Geoscience and Remote Sensing, 61: #5612822 [DOI: 10.1109/TGRS.2022.3194732]

Tan C J, Cao Q, Li Y W, Zhang J L, Yang X, Zhao H Q, et al. 2023. On the promises and challenges of multimodal foundation models for geographical, environmental, agricultural, and urban planning applications [EB/OL]. [2025-09-08]. https://arxiv.org/pdf/2312.17016.pdf

Tang D T, Cao X Y, Hou X S, Jiang Z Y, Liu J M and Meng D Y. 2024. CRS-diff: controllable remote sensing image generation with diffusion model. IEEE Transactions on Geoscience and Remote Sensing, 62: #5638714 [DOI: 10.1109/TGRS.2024.3453414]

Tao C, Qi J, Guo M N, Zhu Q and Li H F. 2023a. Self-supervised remote sensing feature learning: learning paradigms, challenges, and future works. IEEE Transactions on Geoscience and Remote Sensing, 61: #5610426 [DOI: 10.1109/TGRS.2023.3276853]

Tao C, Qi J, Zhang G, Zhu Q, Lu W P and Li H F. 2023b. TOV: the original vision model for optical remote sensing image understanding via self-supervised learning. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16: 4916-4930 [DOI: 10.1109/JSTARS.2023.3271312]

Tao L J, Zhang H K, Jing H Z, Liu Y, Yan D W, Wei G T, et al. 2025. Advancements in vision-language models for remote sensing: datasets, capabilities, and enhancement techniques. Remote Sensing, 17(1): #162 [DOI: 10.3390/rs17010162]

Vivanco Cepeda V, Nayak G K and Shah M. 2023. GeoCLIP: clip-inspired alignment between locations and images for effective worldwide geo-localization//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: ACM: #379

Wang D, Hu M Q, Jin Y, Miao Y C, Yang J Q, Xu Y C, et al. 2025. HyperSIGMA: hyperspectral intelligence comprehension foundation model. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(8): 6427-6444 [DOI: 10.1109/TPAMI.2025.3557581]

Wang D, Zhang Q M, Xu Y F, Zhang J, Du B, Tao D C, et al. 2023a. Advancing plain vision transformer toward remote sensing foundation model. IEEE Transactions on Geoscience and Remote Sensing, 61: #5607315 [DOI: 10.1109/TGRS.2022.3222818]

Wang X F, Cui Q Z, Xu L X, Liu H F, He L X, Luo B, et al. 2023b. GLCNet: global-local complementary network for 3D shape recognition//Proceedings of 2023 International Joint Conference on Neural Networks (IJCNN). Gold Coast, Australia: IEEE: 1-8 [DOI: 10.1109/IJCNN54540.2023.10191731]

Wang Z C, Prabha R, Huang T Y, Wu J J and Rajagopal R. 2024. Sky-

Script: a large and semantically diverse vision-language dataset for remote sensing//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 5805-5813 [DOI: 10.1609/aaai.v38i6.28393]

Wei Y Y, Mao T Y, Li B A, Wang F, Li F, Zhang Z, et al. 2025. Visual and large multimodal models promote image restoration and enhancement: research progress. Journal of Image and Graphics, 30(5): 1197-1219 (韦炎炎, 毛天一, 李柏昂, 王飞, 李锋, 张召, 等. 2025. 视觉模型及多模态大模型推进图像复原增强研究进展. 中国图象图形学报, 30(5): 1197-1219) [DOI: 10.11834/jig.240436]

Xiao A R, Xuan W H, Wang J J, Huang J X, Tao D C, Lu S J, et al. 2025. Foundation models for remote sensing and Earth observation: a survey. IEEE Geoscience and Remote Sensing Magazine, 13(4): 2-29 [DOI: 10.1109/MGRS.2025.3576766]

Xu L R, Zhao L, Guo W, Li Q J, Long K W, Zou K Q, et al. 2024. RS-GPT4V: a unified multimodal instruction-following dataset for remote sensing image understanding [EB/OL]. [2025-09-08]. https://arxiv.org/pdf/2312.17016.pdf

Xu Y H, Yu W K, Ghamisi P, Kopp M and Hochreiter S. 2023. Txt2Img-MHN: remote sensing image generation from text using modern Hopfield networks. IEEE Transactions on Image Processing, 32: 5737-5750 [DOI: 10.1109/TIP.2023.3323799]

Xue X Z, Wei G T, Chen H, Zhang H K, Lin F, Shen C H, et al. 2024. REO-VLM: transforming VLM to meet regression challenges in earth observation [EB/OL]. [2025-09-08]. https://arxiv.org/pdf/2412.16583.pdf

Yan Q, Gu H Y, Yang Y, Li H T, Shen H T and Liu S Q. 2024. Research progress and trend of intelligent remote sensing large model. Acta Geodaetica et Cartographica Sinica, 53(10): 1967-1980 (燕琴, 顾海燕, 杨懿, 李海涛, 沈恒通, 刘世琦. 2024. 智能遥感大模型研究进展与发展方向. 测绘学报, 53(10): 1967-1980) [DOI: 10.11947/j.AGCS.2024.20240053]

Yao T, Zhang Y H, Qiu Z F, Pan Y W and Mei T. 2021. SeCo: exploring sequence supervision for unsupervised representation learning//Proceedings of the 35th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 10656-10664 [DOI: 10.1609/aaai.v35i12.17274]

Yu Z P, Liu C Y, Liu L Q, Shi Z W and Zou Z X. 2025. MetaEarth: a generative foundation model for global-scale remote sensing image generation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(3): 1764-1781 [DOI: 10.1109/TPAMI.2024.3507010]

Yuan Z H, Xiong Z T, Mou L C and Zhu X X. 2025. ChatEarthNet: a global-scale image-text dataset empowering vision-language geo-foundation models. Earth System Science Data, 17(3): 1245-1263 [DOI: 10.5194/essd-17-1245-2025]

Zhang H, Xu J J, Cui H W, Li L, Yang Y W, Tang C S, et al. 2024a. When geoscience meets foundation models: Toward a general geoscience artificial intelligence system. IEEE Geoscience and Remote Sensing Magazine, 13(4): 79-118 [DOI: 10.1109/MGRS.2024.3496478]

Zhang L P, Zhang L F and Yuan Q Q. 2023. Large remote sensing model: progress and prospects. Geomatics and Information Science of Wuhan University, 48(10): 1574-1581 (张良培, 张乐飞, 袁强强. 2023. 遥感大模型: 进展与前瞻. 武汉大学学报(信息科学版), 48(10): 1574-1581) [DOI: 10.13203/j.whugis20230341]

Zhang M, Yang B N, Hu X Y, Gong J Y and Zhang Z X. 2024b. Foundation model for generalist remote sensing intelligence: potentials and prospects. Science Bulletin, 69(23): 3652-3656 [DOI: 10.1016/j.scib.2024.09.017]

Zhang S H and Pan Z G. 2025. Remote sensing large models: review and future prospects. Remote Sensing Technology and Application, 40(1): 1-13 (张帅豪, 潘志刚. 2025. 遥感大模型: 综述与未来设想. 遥感技术与应用, 40(1): 1-13) [DOI: 10.11873/j.issn.1004-0323.2025.1.0001]

Zhang Y, Ye M, Zhu G Y, Liu Y, Guo P Y and Yan J H. 2024c. FFCA-YOLO for small object detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 62: #5611215 [DOI: 10.1109/TGRS.2024.3363057]

Zhao Y X, Zhang M, Yang B N, Zhang Z, Kang J J and Gong J Y. 2025. LuoJiaHOG: a hierarchy oriented geo-aware image caption dataset for remote sensing image-text retrieval. ISPRS Journal of Photogrammetry and Remote Sensing, 222: 130-151 [DOI: 10.1016/j.isprsjprs.2025.02.009]

Zhou Y, Feng L T, Ke Y P, Jiang X, Yan J C, Yang X, et al. 2024. Towards vision-language geo-foundation model: a survey [EB/OL]. [2025-09-08]. https://arxiv.org/pdf/2406.09385.pdf

作者简介

支元杰,男,副教授,主要研究方向为多模态数据融合。
E-mail: zyjyuyan@nwpu.edu.cn
梅少辉,通信作者,男,教授,主要研究方向为遥感图像处理、视频处理、模式识别与机器学习、航天图像获取与处理、光学/电磁图像探测与对抗。E-mail:meish@nwpu.edu.cn
姜艺伟,男,硕士研究生,主要研究方向为遥感图像处理。
E-mail: jiangyiwei@mail.nwpu.edu.cn
杨知,男,高级工程师,主要研究方向为电力遥感应用、智能巡检与防灾减灾。E-mail: yangzhi@ei.sgcc.com.cn
陈奕州,男,硕士研究生,主要研究方向为遥感图像处理。
E-mail: yzchen@mail.nwpu.edu.cn
郝文魁,男,高级工程师,主要研究方向为数据处理与分析。
E-mail: hwk198516@163.com
马明阳,男,副教授,主要研究方向为视频图像处理和遥感图像处理。E-mail: mamingyang@nwpu.edu.cn
魏江,男,副教授,主要研究方向为图像处理、数字信号处理及应用、虚拟现实。E-mail: joyway@nwpu.edu.cn