引文格式:钟燕飞,万瑜廷,马爱龙,等.高分辨率多模态遥感智能解译进展与展望:从遥感基础模型到遥感智能解译脑[J].武汉大学学报(信息科学版),2026,51(11):2329-2355.DOI:10.13203/j.whugis20260400

Citation: ZHONG Yanfei, WAN Yuting, MA Ailong, et al. Progress and Prospects in High-Resolution Multimodal Remote Sensing Intelligent Interpretation: From Remote Sensing Foundation Models to the Remote Sensing Intelligent Interpretation Brain[J]. Geomatics and Information Science of Wuhan University, 2026, 51(11): 2329-2355. DOI: 10.13203/j.whugis20260400

高分辨率多模态遥感智能解译进展与展望:从遥感基础模型到遥感智能解译脑

钟燕飞[1] 万瑜廷[1] 马爱龙[1] 刘寅贺[1] 李柯凡[1] 杨若雨[1] 陈梓航[1]

1 武汉大学测绘遥感信息工程全国重点实验室,湖北 武汉,430079

摘 要:高分辨率多模态遥感智能解译是实现复杂地学场景感知与理解的重要手段,在地表动态监测、资源环境调查和灾害应急等领域具有重要应用价值。近年来,随着深度学习与基础模型的发展,遥感智能解译在多模态特征感知、语义理解认知和复杂决策服务等方面取得了显著进展。系统梳理了遥感影像基础模型、遥感理解基础模型与遥感决策服务平台的研究进展,并在此基础上提出从遥感基础模型向遥感智能解译脑演进的发展思路。同时,分析了当前高分辨率多模态遥感智能解译面临的多模态统一表征不足、复杂地学知识推理可信性有限、感知认知与决策执行协同不足等关键挑战,并从空天地人一体化感知、多模态多尺度统一表征、知识引导的可信地学认知、线上实时自主决策、软硬一体化系统集成与应用等方面展望未来发展方向,以期为高分辨率多模态遥感智能解译的理论研究与技术发展提供参考。

关键词:高分辨率遥感;多模态遥感;智能解译;基础模型;类脑智能;感知-认知-决策

中图分类号:P237 **文献标识码**:A **收稿日期**:2026-09-28

DOI:10.13203/j.whugis20260400 **文章编号**:1671-8860(2026)11-2329-27

Progress and Prospects in High-Resolution Multimodal Remote Sensing Intelligent Interpretation: From Remote Sensing Foundation Models to the Remote Sensing Intelligent Interpretation Brain

ZHONG Yanfei[1] *WAN Yuting*[1] *MA Ailong*[1] *LIU Yinhe*[1] *LI Kefan*[1] *YANG Ruoyu*[1] *CHEN Zihang*[1]

1 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

Abstract: High-resolution multimodal remote sensing intelligent interpretation is an important means of perceiving and understanding complex geospatial scenes, with significant applications in dynamic land surface monitoring, natural resource and environmental surveys, and disaster emergency response. In recent years, advances in deep learning and foundation models have substantially improved remote sensing intelligent interpretation in multimodal feature perception, semantic understanding and cognition, and complex decision-making services. This paper systematically reviews the research progress in remote sensing image foundation models, remote sensing understanding foundation models, and remote sensing decision-making service platforms, and proposes an evolutionary pathway from remote sensing foundation models toward a remote sensing intelligent interpretation brain. Furthermore, it analyzes the key challenges facing high-resolution multimodal remote sensing intelligent interpretation, including insufficient unified multimodal representation, limited reliability of complex geoscientific knowledge reasoning, and inadequate coordination among perception, cognition, decision-making, and execution. Finally, future research directions are discussed in terms of integrated space-air-ground-human perception, unified multimodal and multiscale repre-

基金项目:国家自然科学基金(42301418, 42325105, 42671489, 42671491);湖北省科技创新人才计划(2024DJA020);湖北省自然科学基金(2025AFA077, 2026AFA013)。

第一作者:钟燕飞,博士,教授,研究方向为高分辨率遥感智能解译与应用。zhongyanfei@whu.edu.cn

通信作者:万瑜廷,博士,副教授。wanyuting@whu.edu.cn

sentation, knowledge-guided trustworthy geoscientific cognition, online real-time autonomous decision-making, and integrated software-hardware systems and applications. This review aims to provide a reference for theoretical research and technological development in high-resolution multimodal remote sensing intelligent interpretation.

Key words: high-resolution remote sensing; multimodal remote sensing; intelligent interpretation; foundation model; brain-inspired intelligence; perception-cognition-decision

随着对地观测技术的快速发展,遥感数据正朝着高空间分辨率、多传感器协同、多时相连续观测方向发展。光学、合成孔径雷达(synthetic aperture radar,SAR)、高光谱及红外等多模态遥感数据能够从不同角度刻画地表目标的光谱特征及空间结构,为精细地物识别、地表变化监测和复杂地学场景分析提供了丰富的信息基础[1-3]。然而,不同传感器在成像机理、数据分布和空间尺度等方面存在显著差异,并且高分辨率影像中地物尺度跨度大、场景语义复杂、地学关系隐含,如何实现异构数据的有效融合、复杂场景的准确理解及地学知识的深度挖掘,仍是遥感智能解译面临的重要挑战。

如图1所示,以机器学习、深度学习为代表的人工智能(artificial intelligence,AI)技术已广泛应用于遥感影像解译,推动地物分类、目标检测、语义分割和变化检测等任务由人工特征设计向数据驱动的自动特征学习转变[4-5],如早期的随机森林算法[6]和AlexNet[7]等。2022年ChatGPT的发布使基础模型受到更广泛的关注,指令学习与人类反馈进一步提升了预训练语言模型的任务适配能力[8]。基础模型通过大规模数据预训练获得通用表征能力,并能够通过微调、提示学习等方式适配多种下游任务[9-10]。同时,多模态基础模型逐步实现文本、图像等不同模态信息的统一表征与关联理解,为遥感影像分析拓展了自然语言问答、视觉语义理解及指令驱动的内容生成等新范式。在此背景下,遥感人工智能逐步迈向以基础模型为核心的发展阶段。以SatMAE(satellite masked autoencoder)[11]、RingMo[12]等为代表的研究探索了面向遥感影像的自监督预训练、多光谱信息建模及时序特征学习;视觉语言模型(vision-language model,VLM)的引入则进一步推动遥感影像与文本的跨模态对齐及语义交互,为开放词汇地物识别、遥感视觉问答(visual question answering,VQA)和指令驱动的影像理解等任务提供了新的技术路径[13]。

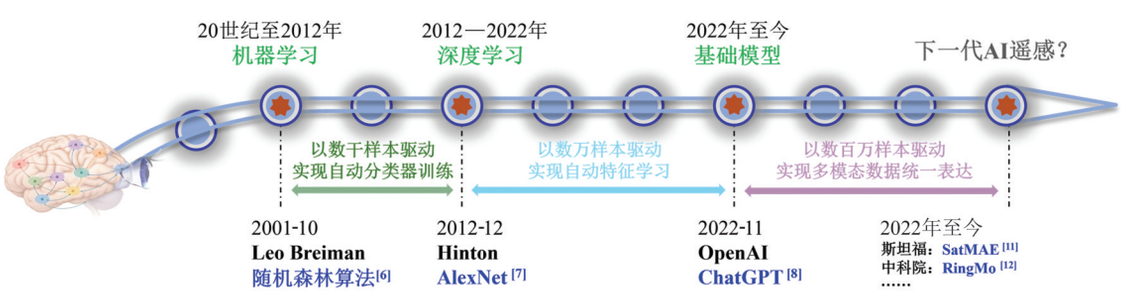


图1 人工智能技术发展历程与遥感基础模型时代

Fig. 1 The Development History of AI Technology and the Era of Remote Sensing Foundation Models

尽管遥感基础模型在通用特征提取和跨任务迁移方面取得了积极进展,但高分辨率多模态遥感智能解译仍存在从特征表征到语义认知、从单项任务处理到综合决策服务的能力鸿沟。一方面,现有模型需要进一步增强对多模态异构遥感数据、细粒度地物特征及跨尺度空间关系的统一表征能力;另一方面,遥感影像中蕴含的地理空间关系、领域知识和地学过程具有较强的专业性与复杂性,仅依赖视觉特征学习难以充分支撑高层次语义理解及知识推理。同时,实际应用不仅要求模型生成准确的解译结果,还需要综合利用多源数据、领域知识和软硬件平台,完成任务规划与决策支持。因此,遥感智能解译的研究对象有必要从单一模型与算法扩展至贯通感知、理解认知和决策服务的技术体系。

基于上述认识,本文从高分辨率多模态遥感智能解译的任务需求出发,将现有研究归纳为3个相互衔接的层次:遥感影像基础模型、遥感理解基础模型和遥感决策服务平台。其中,遥感影像基础模型侧重多源遥感数据的统一表征、跨模态特征学习与精细地物信息提取;遥感理解基础模型进一步面向高层次语义解析、地理空间关系建模及领域知识

推理;遥感决策服务平台则侧重整合遥感数据、模型算法与软硬件工具,支撑复杂任务的组织执行和决策服务。3个层次分别对应由数据表征感知、语义理解认知到快速应用决策的能力递进,共同构成遥感智能解译技术体系的重要基础。

然而,目前遥感感知、认知与决策过程仍相对分离,限制了地学应用的深度和广度。受人脑整合多模态感知、记忆、学习、推理和决策机制的启发,本文以遥感智能解译脑[14]的理论框架为基础,讨论从“遥感基础模型”迈向“遥感智能解译脑”的发展思路,面向复杂遥感任务构建融合多模态感知、地学知识、记忆推理和自主规划决策能力的“感知—认知—决策”全链路智能体系[14]。在此框架下,本文系统梳理遥感影像基础模型、遥感理解基础模型和遥感决策服务平台的研究进展,分析现有技术在多模态统一表征、复杂地学语义理解、可信推理和自主决策等方面的局限,并进一步探讨遥感智能解译脑的概念内涵、技术架构与发展方向,为高分辨率多模态遥感智能解译从通用表征学习迈向复杂地学认知与自主决策提供参考。人工智能与高分辨率多模态遥感大数据分析应用的主要任务见图2。

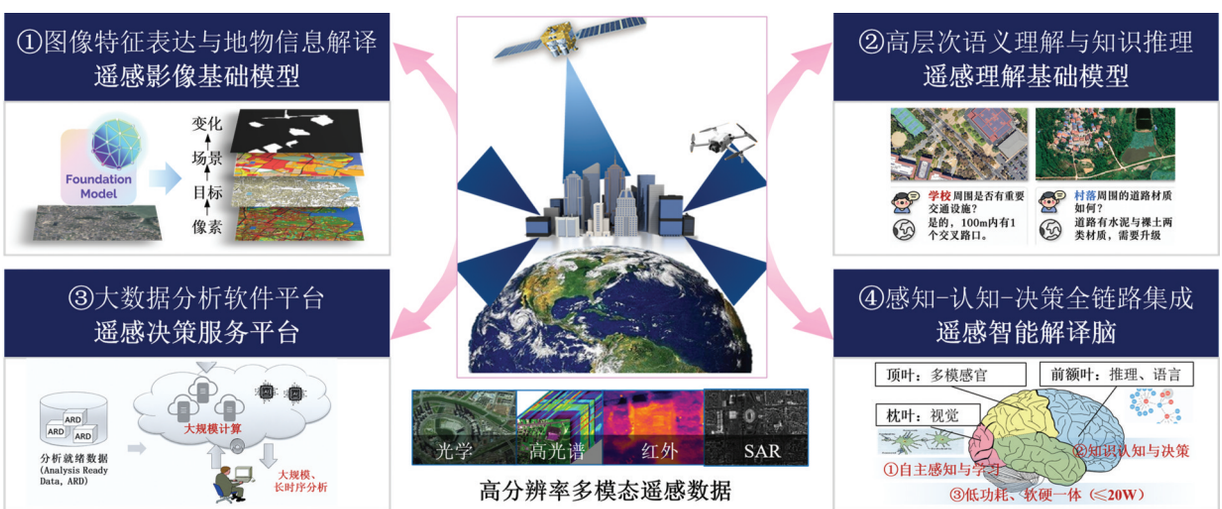


图2　人工智能和高分辨率多模态遥感大数据分析与应用的主要任务

Fig. 2　Main Tasks of AI and High-Resolution Multimodal Remote Sensing Big Data Analysis and Applications

1　遥感影像基础模型

遥感影像基础模型通过大规模遥感数据预训练学习可迁移的影像表征,为分类、检测、分割和变化检测等任务提供通用特征支撑。其利用影像的空间结构、光谱关系、时序关联和跨模态对应关系,降低下游任务对标注样本的依赖,并增强跨区域、跨传感器和跨时相应用的适应能力。根据观测数据特点与主要建模目标,相关研究可归纳为高分辨率、高光谱和多模态基础模型3类,典型模型如表1所示。3类模型分别侧重空间细节刻画、空间-光谱耦合和多源信息融合,彼此存在交叉,共同构成遥感通用感知的技术基础。

表1　代表性遥感影像基础模型对比表

Table 1　Comparison of Representative Remote Sensing Image Foundation Models

模型	年份	主要数据集	数据来源/模态	数据规模	训练方式	下游任务
RingMo[12]	2022	RingMo	GF-2	2 MB	MIM	IC, SS, OD, CD
GFM[15]	2023	GeoPile	NAIP/RGB	600 kB	MAE	IC, SS, CD, 影像重建
Scale-MAE[16]	2023	fMoW-RGB	RGB	363.6 kB	MAE	IC, SS
MAEST[17]	2023	Indian Pines、Pavia、Houston	高光谱	81 MB	MAE	IC
Prithvi[18]	2023	HLS	Landsat / S2	1 TB	MAE	SS
SatlasPretrain[19]	2023	Satlas	NAIP/S2	13 MB	监督多任务学习	IC, SS, 回归任务
USat[20]	2023	Satlas	NAIP/S2	13 MB	MAE	IC, SS
DECUR[21]	2023	SSL4EO	S1/S2	250 kB	Contrastive	IC
SkySense[22]	2023	SkySense	S1/S2/RGB	21.5 MB	Contrastive	IC, SS, OD, CD

续表

模型	年份	主要数据集	数据来源/模态	数据规模	训练方式	下游任务
CROMA[23]	2023	SSL4EO	S1/S2	250 kB	Con-MAE	IC,SS
S2MAE[24]	2024	fMoW-S2, Big-EarthNet-S2	S2	1.5 MB	MAE	IC,CD
SatMAE++[25]	2024	fMoW-Sentinel, fMoW-RGB	S2/RGB	1 MB	MAE	IC
HyperFree[26]	2024	Hyper-Seg	高光谱	15 MB	提示分割预训练	IC, OD,异常检测,分割
MA3E[27]	2024	MillionAID	Google Earth	1 MB	MAE	IC, OD, SS
Cross-Scale MAE[28]	2024	fMoW-RGB	RGB	363.6 kB	MAE	IC
LeMeViT[29]	2024	MillionAID	Google Earth	1 MB	监督预训练	IC, OD, SS
DINO-MC[30]	2024	SeCo-100K		100 kB	Contrastive	SS, CD
UV-SAM[31]	2024	UV-SAM	RGB		提示适配与监督分割	城中村识别与分割
SARATR-X[32]	2024	ImageNet, 14 SAR datasets	SAR	0.2 MB	MAE	SAR目标识别
DOFA[33]	2024	GEO-Bench	NAIP+Gaofen+S1/S2+EnMAP		MAE	IC, SS, CD
OmniSat[34]	2024	OmniSat	NAIP/S1/S2	300 kB	Contrastive	IC, SS
SAR-JEPA[35]	2024	MSAR, SAR-Ship, SARSim, SAMPLE	SAR		联合嵌入预测	SAR目标识别
BIOSCAN-CLIP[36]	2024	Bioscan-CLIP	HR-RGB/Text/DNA	1 MB	Contrastive	分类学分类(从目到种),检索,零样本学习,未见物种识别,贝叶斯零样本分类
HyperSIGMA[37]	2025	HyperGlobal-450K	EO-1/GF-5	0.45 MB	MAE	HyperSpectral Tasks
SelectiveMAE[38]	2025	OpticalRS-13M		13 MB	MAE	IC, OD, SS
RingMo-Aerial[39]	2025	RingMo	RGB	5.4MB	Contrastive	IC, OD, SS, CD,三维重建,目标跟踪
DINOv3[40]	2025	DINOv3	Maxar	493 MB	Contrastive	IC, SS
ALPS[41]	2025	ALPS	RGB		伪标签监督预训练	SS
BioAnalyst[42]	2025	BioCube	环境/气象/物种		MAE	物种分布预测,气候变量线性探测,生物多样性动态预测
AlphaEarth[43]	2025	AlphaEarth	S2/Landsat/S1/LiDAR/环境	3 BB	Contrastive	专题制图(土地利用、土地覆盖、作物检测、作物类型识别及物种分布),生物物理变量估算(发射率、蒸散量),CD
AnySat[44]	2025	AnySat	NAIP/S1/S2		Contrastive	IC, SS
SeaMo[45]	2025	SSL4EO	S1/S2	250 kB	Contrastive	SS, OD, CD
OlmoEarth[46]	2025	OlmoEarth	S1/S2/Landsat	250 kB	Con-MAE	IC, SS
RingMoE[47]	2025	RingMo	RGB/S1/S2	400 MB	Con-MAE	IC, SS, OD, CD, 目标跟踪
TaxaBind[48]	2025	TaxaBind	RGB/Text/音频/环境	3 MB	Contrastive	零样本物种分类,跨模态检索,鸟类声音分类,地理特征回归,生态区分类,生物群系分类,植物物种预测,物种遭遇率预测,物种分布范围制图
Prithvi-EO-2.0[49]	2026	HLS	Landsat 8 and 9, S2	4.2 MB	MAE	IC, SS
TESSERA[50]	2026	TESSERA	S1/S2	120 kB	Contrastive	IC, SS, 回归分析
MaRS[51]	2026	MaRS-16M	RGB/UHR-SAR	16 MB	Con-MAE	多模态配准,模态缺失制图,跨模态图像翻译,建筑高度估计,道路提取,OD,CD,损毁评估等

表1中,fMoW为Functional Map of the World数据集,NAIP(National Agriculture Imagery Program)为美国国家农业影像计划,GF-2、S1、S2分别指高分二号、Sentinel-1、Sentinel-2卫星,RGB(red, green and blue)为红、绿、蓝三通道,HR(high resolution)为高分辨率,UHR(ultra-high resolution)为超高分辨率,LiDAR(light detection and ranging)为激光雷达,DNA(deoxyribonucleic acid)为脱氧核糖核酸,HLS(harmonized Landsat and Sentinel-2)为Landsat与Sentinel-2协调数据集,MIM(masked image modeling)为掩码图像建模,MAE(masked autoencoder)为掩码自编码,Con-MAE表示对比学习与掩码自编码的联合训练,Contrastive表示对比学习,IC(image classification)为影像分类,SS(semantic segmentation)为语义分割,OD(object detection)为目标检测,CD(change detection)为变化检测,CLIP(contrastive language-image pre-training)为对比语言-图像预训练模型,GFM(geospatial foundation model)为地理空间基础模型,MAEST(masked autoencoding spectral-spatial Transformer)为掩码自编码光谱-空间Transformer,DOFA(dynamic one-for-all)为动态通用模型。

1.1　高分辨率遥感基础模型

高分辨率遥感基础模型主要面向具有丰富纹理、清晰边界和复杂空间布局的光学遥感影像,通过大规模预训练学习局部目标与全局场景的通用表征。此类影像幅面大、目标尺度变化显著,同一场景内往往同时包含车辆等小型目标与道路、建筑群等大范围结构,因此模型需要兼顾细粒度特征和长距离空间关系。掩码图像建模是其重要技术路线:通过遮挡部分图像块并重建缺失内容,促使编码器学习纹理、边界与上下文联系。RingMo面向遥感场景开展大规模掩码预训练,为场景分类、目标检测和语义分割等任务提供可迁移特征,体现了利用未标注影像学习通用空间表征的潜力[12]。

针对不同传感器和观测尺度带来的差异,Scale-MAE将地理尺度信息引入掩码自编码过程,通过尺度感知表征与多尺度重建增强跨分辨率迁移能力[16];SatlasPretrain则利用大规模遥感影像及地理对象标注学习多任务共享的空间语义[19]。DINOv3采用教师-学生自蒸馏开展大规模自监督学习,并通过格拉姆矩阵锚定约束稳定图像块之间的特征关系,缓解长期训练中密集特征质量退化的问题[40]。其卫星影像版本面向高分辨率光学数据预训练,能够提供兼具局部细节与全局语义的密集表征,在冻结骨干网络的条件下结合轻量任务模块用于树冠高度估计等遥感任务,体现了自蒸馏路线在精细空间表征与高效迁移中的潜力。总体而言,高分辨率基础模型正由纹理与场景表征向尺度感知和细粒度对象理解发展,提升超大幅面、小目标及跨区域解译能力。

1.2　高光谱遥感基础模型

高光谱遥感基础模型需同时建模高维光谱信息与空间结构。由于光谱维度具有连续性和较强相关性,其建模更为复杂,相较于高分辨率遥感基础模型起步较晚。高光谱遥感基础模型面向具有连续窄波段信息的遥感影像,利用地物光谱响应及空间结构学习通用表征,服务于精细地物分类、矿物识别、植被分析和异常检测等任务。其核心是建立空间与光谱联合编码,通过空间掩码、光谱掩码及联合重建学习地物邻域关系、光谱连续性和跨波段依赖。高光谱数据维度高、波段冗余明显,且不同传感器的波段数量、中心波长和覆盖范围不一致,因此基础模型不仅需要保留具有材料辨识价值的光谱信息,还需要具备适应不同通道配置的能力。

SpectralGPT将光谱遥感数据作为三维空间-光谱结构进行建模,通过三维标记生成和多目标重建学习空间与光谱的耦合关系,并以渐进式训练适应不同来源的影像[52]。该研究主要基于大规模多光谱数据预训练,为高光谱基础模型提供了方法参考,但向连续窄波段数据迁移时仍需处理光谱采样及传感器响应差异。HyperFree进一步面向高光谱影像的通道适应和免调优应用,通过通道自适应嵌入将不同波段数量的输入转化为统一形式的特征标记,并结合提示信息完成下游解译[26]。这一思路缓解了固定输入通道对跨传感器应用的限制,推动高光谱解译由针对单一数据集训练的专用网络向可复用、可提示的基础模型发展。高光谱基础模型正在向通道自适应和少标注迁移发展,但仍受光谱冗余、数据质量及传感器差异制约。

1.3　多模态遥感基础模型

多模态遥感基础模型需要兼顾通用表征学习与跨传感器适配,其构建更具挑战性。此类模型联合利用光学、SAR、多光谱、高光谱和LiDAR等观测数据,在共享语义空间中学习不同成像机

制的互补信息。光学数据提供外观与纹理,SAR反映散射特性,LiDAR刻画三维结构,光谱数据则有助于区分地物材料。跨模态对比学习与联合掩码建模是其主要技术路线:前者通过对齐同一区域的跨模态特征建立一致的语义表征,后者通过恢复被遮挡内容学习模态内结构和模态间关联。CROMA(contrastive radar-optical masked autoencoders)将雷达-光学对比学习与掩码自编码结合,使模型同时利用跨模态一致性和影像自身结构[23]。在全球地表表征方面,AlphaEarth Foundations汇聚多源观测的空间、时间与测量信息,将其压缩为64维地表嵌入,为稀疏标注条件下的制图和监测提供可直接复用的特征[43]。TESSERA(temporal embeddings of surface spectra for earth representation and analysis)则面向Sentinel-1/2多模态时间序列开展像元级自监督学习,结合Barlow Twins目标与稀疏随机时间采样,使表征对有效观测的选取保持稳定,并保留与植被物候相关的时序信息,支撑地表分类、分割和树冠高度预测等任务[50]。

在通用解译方面,SkySense学习多源观测数据的空间与时序特征[22],SkySense++进一步通过像素、对象和场景之间的语义联系提高多任务迁移能力[53]。AnySat探索多分辨率、多尺度和多模态数据的统一建模[44];MaRS(a multi-modality very-high-resolution remote sensing foundation model)则面向甚高分辨率光学与SAR影像,通过跨粒度元模态学习协调两类数据的局部结构与空间表达[51]。AlphaEarth和TESSERA侧重可复用的像元级地表嵌入,SkySense、AnySat和MaRS等模型则更关注面向多任务解译的联合编码与特征适配,共同拓展了多模态基础模型的应用方式。多模态基础模型通过语义对齐与时空融合增强了多源观测的互补利用,但仍须进一步提高复杂观测条件下的迁移可靠性。

2 遥感理解基础模型

遥感影像基础模型侧重影像表征与信息提取,通常不直接输出自然语言解释或地学知识。遥感理解基础模型通过大规模遥感影像及相关语义数据预训练,学习具有跨任务迁移能力的视觉表征与视觉语言关联,实现对遥感场景、地物属性、空间关系及地理变化的综合理解,并支持图像描述、视觉问答、图文检索和目标定位等多种下游任务。如图3所示,遥感理解基础模型的技术发展由早期任务专用的遥感视觉语言模型(remote sensing vision-anguage model,RSVLM)逐步转向大规模图文对比式预训练、自回归生成式预训练与多模态指令学习,研究重点从基础语义识别进一步拓展至高分辨率精细感知、多时相变化理解及证据约束的复杂地理推理。在此基础上,文本条件影像生成与编辑技术通过构建语义可控的遥感场景,为理解模型的数据增强、语义对齐和能力评估提供补充支持。此外,支撑遥感理解基础模型发展的代表性数据集见表2,主要任务涉及图像描述、视觉问答、语言指代定位与分割、变化描述、变化视觉问答等。

2.1 早期遥感视觉语言任务模型

早期遥感视觉语言研究主要围绕图像描述、视觉问答和图文检索展开,通常采用任务专用的视觉编码、语言表示及跨模态融合结构,通过监督学习建立遥感影像与自然语言之间的语义关联。

图像描述旨在将遥感影像中的地物类别、空间分布及对象关系转化为自然语言。QU等[54]采用卷积神经网络(convolutional neural network,CNN)与循环神经网络(recurrent neural network,RNN)组合架构开展早期遥感影像描述研究;LU等[55]构建RSICD数据集,为后续方法提供了统一评测基础。随后,GAJBHIYE等[71]提出空间-通道注意力与记忆引导Transformer(spatial-channel attention-based memory-guided Transformer,SCAMET),通过该结构提升多尺度语义建模能力;ZHAO等[72]进一步引入区域特征与几何关系,增强对象级描述的准确性。视觉问答通过联合建模影像与自然语言问题,实现地物识别、数量统计和空间关系分析。LOBRY等[58]提出RSVQA任务与数据集,建立了基于视觉-语言特征融合的问答基线。WANG等[63]进一步提出EarthVQA数据集及语义对象感知框架(semantic object awareness,SOBA),通过对象感知和关系推理增强复杂地理场景的问答能力。图文检索通过跨模态语义匹配实现文本与遥感影像的双向检索。ABDULLAH等[73]提出TextRS双向三元组匹配网络,CHENG等[74]利用深度语义对齐增强跨模态特征匹配。随后,YUAN等[64]提出非对称多模态特征匹配网络(asymmetric multimodal feature matching network,AMFMN)及RSITMD数据集,推动多尺度细粒度检索研究;HE等[75]进一步融合全局显著特征与文本关系表

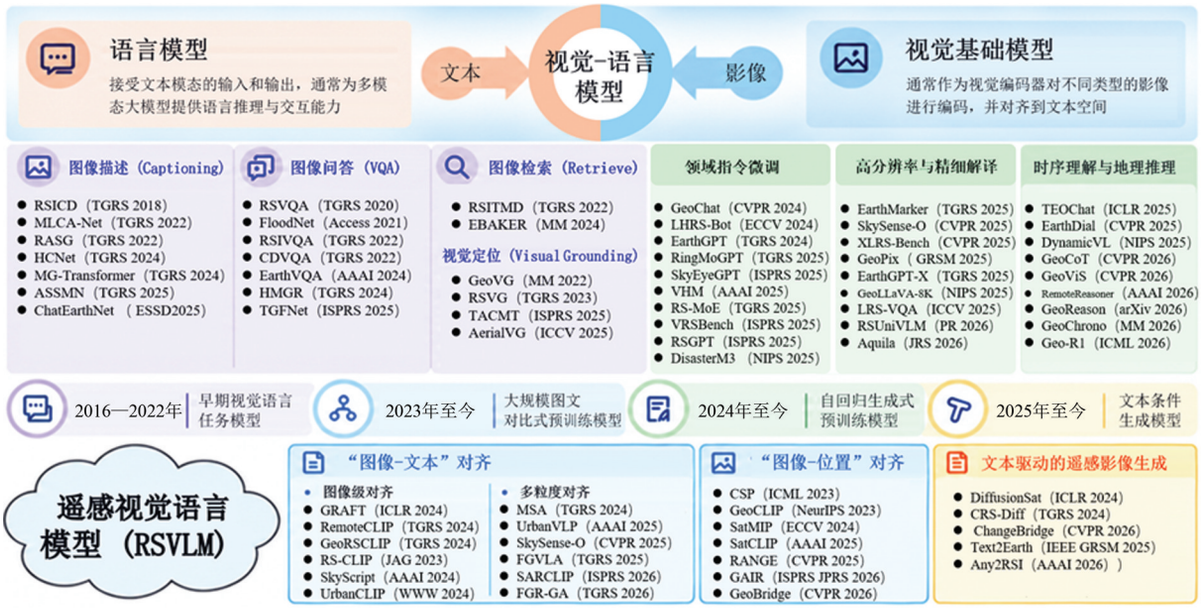


图 3　遥感理解基础模型发展脉络

Fig. 3　Evolution of Remote Sensing Understanding Foundation Models

表 2　遥感视觉语言任务的代表性数据集

Table 2　Representative Datasets for Remote Sensing Vision-Language Tasks

任务类型	数据集(年份)	影像规模	文本 / 标注规模	数据模态与主要特点
图像描述	UCM-Captions[54](2016)	2 100 幅	10 500 条描述	RGB / 文本;21 类场景,基础场景语义描述
图像描述	Sydney-Captions[54](2016)	613 幅	3 065 条描述	RGB / 文本;7 类城市及土地利用场景
图像描述	RSICD[55](2018)	10 921 幅	54 605 条描述	RGB / 文本;30 类场景,包含重复补齐的描述
图像描述	NWPU-Captions[56](2022)	31 500 幅	157 500 条描述	RGB / 文本;45 类场景,增强复杂场景覆盖
图像描述	RSICap[57](2025)	2 585 幅	2 585 条人工描述	RGB / 文本;细粒度地物属性、数量及空间关系
视觉问答	RSVQA-LR[58](2020)	772 幅	77 232 组问答	Sentinel-2 / 文本;地物存在性、数量及关系
视觉问答	RSVQA-HR[58](2020)	10 659 幅	1 066 316 组问答	高分辨率 RGB / 文本;建筑物、道路、面积及数量
视觉问答	RSVQA[59](2022)	37 264 幅	约 11.1 万组问答	RGB / 文本;分类、计数和属性识别
视觉问答	RSVQAxBEN[60](2021)	590 326 幅	14 758 150 组问答	Sentinel-2 多光谱 / 文本;土地覆盖、类别存在性
视觉问答	FloodNet[61](2021)	2 343 幅	训练集 4 511 组问答	无人机 RGB / 文本 / 掩膜;洪水灾害场景问答
视觉问答	HRVQA[62](2024)	53 512 幅	1 070 240 组问答	高分辨率航空 RGB / 文本;属性、位置与空间推理
视觉问答	EarthVQA[63](2024)	6 000 幅	208 593 组问答	RGB / 文本 / 语义掩膜;对象关系、计数及综合分析
图文检索	RSITMD[64](2022)	4 743 幅	23 715 条描述	RGB / 文本 / 关键词;细粒度双向图文检索
语言指代定位	DIOR-RSVG[65](2023)	17 402 幅	38 320 条指代表达	RGB / 文本 / 边界框;属性与空间关系定位
语言指代引导分割	RefSegRS[66](2024)	285 个场景	4 420 组图文掩膜三元组	航空 RGB / 文本 / 掩膜;语言指代分割
语言指代引导分割	RRSIS-D[67](2024)	17 402 组三元组	17 402 组指代与掩膜	RGB / 文本 / 掩膜;尺度、方向及关系表达
变化描述	LEVIR-CC[68](2022)	10 077 组双时相影像	50 385 条变化描述	双时相 RGB / 文本;地物新增、消失与变化描述
变化视觉问答	CDVQA[69](2022)	2 968 组双时相影像	超过 12.2 万组问答	双时相 RGB / 文本;变化类别、面积与比例问答
变化描述	SECOND-CC[70](2025)	6 041 组双时相影像	30 205 条变化描述	双时相 RGB / 文本 / 语义图;复杂变化场景

注:LR(low resolution)为低分辨率;RSICD(remote sensing image captioning dataset)为遥感影像描述数据集;RSVQA(remote sensing visual question answering)为遥感视觉问答;RSITMD(remote sensing image-text match dataset)为遥感影像文本匹配数据集。

示,提升复杂场景中的语义对齐能力。指代分割则进一步建立自然语言表达与具体地物区域的像素级对应关系,实现基于类别、属性及空间关系描述的目标分割。YUAN等[66]提出指代遥感影像分割(referring remote sensing image segmentation,RRSIS)任务并构建RefSegRS数据集,通过语言引导的跨尺度特征增强模块(language-guided cross-scale enhancement,LGCE)改善小尺度及分散目标的分割效果。LIU等[67]进一步提出旋转多尺度交互网络(rotated multi-scale interaction network,RMSIN),构建RRSIS-D数据集,并通过多尺度特征交互与自适应旋转卷积增强不同尺度和方向地物的指代分割能力。

此类研究逐步由全局特征匹配发展至局部对象感知与细粒度跨模态交互,但模型仍主要围绕特定任务进行优化,跨任务迁移与开放式语义理解能力仍有限。这些研究为后续通用遥感视觉语言基础模型提供了任务定义、数据集与方法基础。

2.2 遥感大规模图文对比式预训练模型

遥感对比式视觉语言模型通过独立的视觉与文本编码器构建共享语义空间,利用图文对比学习实现跨模态特征对齐。遥感大规模图文对比式预训练模型的代表性研究方法如表3所示。

表3 遥感大规模图文对比式预训练模型的代表性方法

Table 3 Representative Methods for Remote Sensing Large-Scale Image-Text Contrastive Pre-training Models

模型(年份)	主要数据集	数据规模	数据模态	训练方式及主要任务
GeoRSCLIP[76](2024)	RS5M	约500万图文对	RGB/文本	领域对比预训练;零样本分类、图文检索
RemoteCLIP[77](2024)	RET-3、SEG-4、DET-10	165 745幅影像;828 725图文对	RGB/文本/检测与分割标注	标注转换、对比学习;分类、检索、计数
SkyCLIP[78](2024)	SkyScript	未筛选约520万图文对;筛选后训练子集约260万对	RGB/文本/地图标签	OpenStreetMap(OSM)语义监督、对比学习;开放词汇分类
GRAFT[79](2024)	共址卫星-地面影像	按跨视角影像对组织	卫星RGB/地面RGB	跨视角语义迁移;零样本分类、检索、分割
Sat2Cap[80](2024)	卫星-地面影像配对数据	约610万组跨视角影像对	卫星RGB/地面RGB	跨视角对比学习;自然语言地理制图
SenCLIP[81](2025)	Sen2LUCAS	LUCAS源数据约23.5万个地理位置;实际训练样本另计	Sentinel-2 RGB/地面RGB	地面语义迁移;零样本土地利用分类
Git-RSCLIP[82](2025)	Git-10M	约1 000万图文对	RGB/文本	大规模图文对比预训练;分类、检索
SAR-RS-CLIP[83](2025)	SAR-TEXT	超过13万SAR图文对	SAR/文本	渐进迁移、对比学习;SAR图文检索
SARCLIP[84](2026)	SARCAP	超过40万SAR图文对	SAR/文本	噪声鲁棒编码、层次化提示学习;检索、分类、计数
TimeSenCLIP[85](2026)	LUCAS、Sen4Map	Sen4Map约14万训练、3万验证、5万测试样本	Sentinel-2多光谱时序/地面影像语义	时序跨视角对比学习;土地覆盖、作物分类
SkySense-O[86](2025)	Sky-SA	183 375组局部图文对;1 763类标签	RGB/文本/局部空间标注	视觉自监督、局部对比学习;识别、定位、推理
SegEarth-OV[87](2025)	Million-AID及多个下游基准	约1.6万幅无标签影像训练特征上采样器	RGB/类别文本	CLIP特征恢复与局部校正;开放词汇分割
OVRS[88](2025)	DLRSD、iSAID、Potsdam、Vaihingen	4个遥感分割评测基准	RGB/类别文本/掩膜	旋转聚合、多尺度融合;开放词汇分割

RADFORD等[89]提出CLIP模型,通过大规模图文预训练学习可迁移的视觉语义表征,为零样本分类与图文检索提供了通用框架。在此基础上,遥感研究首先聚焦于领域适配,通过构建遥感图文数据提升模型对地理场景的理解能力。ZHANG等[76]构建包含约500万图文对的RS5M数据集,并提出GeoRSCLIP,提升零样本分类、图文检索和语义定位性能。针对遥感图文监督不足的问题,研究进一步拓展多源语义监督。LIU等[77]提出RemoteCLIP,通过边界框转描述(Box-to-Caption,B2C)和掩膜转边界框(Mask-to-Box,M2B)策略,将检测与分割标注转化为图文监督,增强跨任务迁移能力。WANG等[78]提出SkyScript,利用地理坐标关联OpenStreetMap

语义标签,构建包含260万图文对的数据集,并通过SkyCLIP实现开放词汇场景识别与跨模态检索。除直接利用遥感图文数据与地理标注外,跨视角语义迁移为降低文本标注依赖提供了新的途径。MALL等[79]提出GRAFT,通过共址地面与卫星影像对齐,将预训练CLIP的语言语义迁移至遥感视觉编码器,实现无需直接遥感图文标注的零样本理解。DHAKAL等[80]提出Sat2Cap,利用610万组卫星与地面影像建立跨视角语义关联,实现自然语言驱动的零样本地理信息制图。

上述研究推动遥感语义监督由直接图文配对向任务标注复用、地理知识关联及跨视角迁移拓展,但全局特征对齐仍难以充分表征密集小目标及局部空间关系,促使研究进一步转向细粒度语义建模。ZHONG等[90]提出RegionCLIP,将图文对比学习扩展至区域级匹配,为开放词汇目标检测提供了技术基础。针对遥感场景,ZHU等[86]提出SkySense-O,通过局部图文对比学习、视觉自监督及知识图谱增强开放类别语义建模,提升识别、定位和推理能力。在此基础上,对比式语义表征进一步向像素级开放词汇解译拓展。LI等[87]提出SegEarth-OV,通过特征上采样与局部校正缓解CLIP的全局语义偏置,在无需针对目标分割任务训练的条件下实现开放词汇分割;CAO等[88]提出OVRS,通过旋转聚合相似度计算与多尺度特征融合,改善不同方向和尺度地物的分割效果。总体而言,遥感对比式视觉语言模型逐步由领域图文预训练向多源语义监督、细粒度空间对齐及开放词汇解译发展,增强了表征的可迁移性与空间辨识能力,但复杂地理推理、长指令理解及开放式交互仍需结合语言生成模型进一步实现。

2.3　遥感自回归生成式预训练模型

自回归生成式视觉语言模型以遥感影像和自然语言指令为条件,通过逐词元预测生成场景描述、问题答案或结构化解译结果。典型架构由视觉编码器、跨模态连接模块和大语言模型(large language model,LLM)组成,通过视觉语言对齐与指令微调实现多任务统一建模。遥感自回归生成式预训练模型的代表性方法见表4。

表4　遥感自回归生成式视觉语言模型的代表性方法

Table 4　Representative Methods of Autoregressive Generative Vision-Language Models for Remote Sensing

模型(年份)	主要数据集	数据规模	数据模态	训练方式及主要任务
RS-LLaVA[91](2024)	RS-Instructions及通用指令数据	7 058条指令,训练集5 506条	RGB / 文本	LoRA、指令微调;图像描述、视觉问答
RSGPT[57](2025)	RSICap、RSIEval	2 585组图文对;936组评测问答	RGB / 文本	视觉语言对齐、领域适配;描述、问答
LHRS-Bot[92](2024)	LHRS-Align、LHRS-Instruct	约115万组图文对;约4万条生成式指令	RGB / 文本 / 地理信息	多层级对齐、课程学习;空间理解、问答
SkySenseGPT[93](2024)	FIT-RS	约180万条指令,其中141.5万条用于训练	RGB / 文本	细粒度指令微调;对象关系、场景图、问答
SkyEyeGPT[94](2025)	SkyEye-968k	约96.8万条指令	RGB / 文本	两阶段指令微调;描述、问答、分类、定位
VHM[95](2025)	VersaD、HnstD等	约140万组图文对;18万条SFT指令	RGB / 文本	场景描述多样化、可靠性增强微调;问答、定位
LHRS-Bot-Nova[96](2025)	LHRS-Align-Recap、LHRS-Instruct-Plus	约115万组图文对;新增32万条指令	RGB / 文本 / 地理信息	编码与连接模块优化;地理感知、空间认知
GeoChat[13](2024)	GeoChat-Instruct	约31.8万条指令	RGB / 文本 / 坐标	区域指令微调;对象问答、视觉定位
EarthMarker[97](2025)	RSVP	约365万组视觉提示-文本配对	RGB / 文本 / 点与区域提示	视觉提示学习;图像、区域、点级理解
GeoPix[98](2025)	GeoPixInstruct	65 463幅影像;140 412个对象实例	RGB / 文本 / 掩膜	分割词元与掩膜预测;像素级理解

续表

模型（年份）	主要数据集	数据规模	数据模态	训练方式及主要任务
GeoPixel[99]（2025）	GeoPixelD	超过5.3万条指代表达；约60万个对象标注	高分辨率RGB／文本／掩膜	高分辨率编码、掩膜交织生成；指代分割
SegEarth-R1[100]（2025）	EarthReason	5 434个掩膜；超过3万组问答	RGB／文本／掩膜	语言推理与掩膜解码；隐含语义分割
SegEarth-R2[101]（2026）	LaSeRS	约3.08万组三元组；超过4万个掩膜	RGB／文本／掩膜	空间注意力监督、分割查询；复杂指代分割
UniGeoSeg[102]（2026）	GeoSeg-1M、GeoSeg-Bench	约59万幅影像；110万组指令三元组	RGB／文本／掩膜	任务感知文本增强、渐进训练；指代、交互与推理分割
MIGRANT[103]（2026）	MIG-RS-Instruct、MIG-RS-Bench	50 114条训练指令；4 000组评测问答	多幅RGB／文本／坐标	两阶段指令微调；跨影像目标定位
EarthGPT[104]（2024）	MMRS-1M	超过100万组图文数据	光学／SAR／红外／文本	多传感器对齐、指令微调；问答、定位
EarthGPT-X[105]（2025）	M-RSVP	超过65万组影像-提示-文本三元组	光学／SAR／红外／文本／提示	多源空间提示融合；问答、定位、分割
EarthDial[106]（2025）	EarthDial-Instruct	超过1 111万条指令	RGB／SAR／多光谱／文本	多传感器统一指令学习；分类、检测、问答
EarthMind[107]（2025）	FusionEO、EarthMind-Bench	3万组训练样本；2 841组评测样本	光学／SAR／文本	空间注意力提示、跨传感器融合；问答、推理
Earth-OneVision[108]（2026）	MMRS-OneVision	约3 400万组问答对	光学／SAR／红外／多光谱／时序／视频／文本	全粒度对齐、渐进模态适配；9类理解任务
SAR-GPT[83]（2025）	SAR-TEXT及SAR问答数据	超过13万组SAR图文对	SAR／文本	SAR领域视觉指令微调；描述、问答
ChangeChat[109]（2025）	ChangeChat-87k	约8.7万条指令	双时相RGB／文本	变化指令微调；变化描述、量化、定位
TEOChat[110]（2025）	TEOChatlas	554 071条指令	多时相影像／文本	时序指令微调；变化问答、灾害分析
DVLChat[111]（2025）	DVL-Suite、DVL-Instruct	14 871幅时相影像；63 771组训练指令	多时相RGB／文本／变化标注	时序指令微调；变化量化、城市演化
GeoLLaVA-8K[112]（2025）	SuperRS-VQA、HighRS-VQA	81 367组高分辨率问答对	高分辨率RGB／文本	词元裁剪、锚点选择；8K影像问答
ZoomEarth[113]（2025）	LRS-GRO	SFT 1 011条；RL 2 500条问答	高分辨率RGB／文本／区域	SFT＋GRPO、主动感知；大幅面问答、定位
GeoVista[114]（2026）	APE-GRO	33 445条SFT轨迹；8 327条RL轨迹	超高分辨率RGB／文本／区域证据	多分支探索、证据状态管理；主动空间理解
GeoVLM-R1[115]（2025）	多任务遥感推理数据	约78.5万组GRPO任务问答	RGB／文本／空间标注	SFT＋GRPO、任务感知奖励；空间与变化推理
RSThinker[116]（2026）	Geo-CoT380k	约38万条结构化推理样本	RGB／文本／推理轨迹	感知约束推理链、GRPO；空间推理
GeoReason[117]（2026）	GeoReason-Bench	4 000条逻辑推理轨迹	RGB／文本／推理轨迹	逻辑一致性强化学习；地理推理、问答

注：LoRA（low-rank adaptation）为低秩适配；SFT（supervised fine-tuning）为监督微调；RL（reinforcement learning）为强化学习；GRPO（group relative policy optimization）为组相对策略优化；CoT（chain-of-thought）为思维链；VHM（versatile and honest vision language model）为通用且可靠的视觉语言模型。

早期遥感自回归生成式预训练模型研究主要关注通用模型的领域适配。LIU 等[118]提出大型语言与视觉助手(large language and vision assistant,LLaVA),利用视觉指令微调建立多模态交互框架,为遥感生成式模型提供了技术基础。HU 等[57]提出 RSGPT,构建 RSICap 数据集与 RSIEval评测基准,增强遥感场景描述能力;BAZI 等[91]提出 RS-LLaVA,通过低秩适配(low-rank adaptation,LoRA)与多任务指令训练实现图像描述和视觉问答的联合建模。为突破单一任务的局限,ZHAN 等[94]提出 SkyEyeGPT,构建约 96.8 万条多任务指令数据,通过两阶段微调统一图像描述、视觉问答、场景分类与视觉定位等任务。随着任务类型不断扩展,研究进一步关注地理知识融入与视觉语言表征优化。MUHTAR 等[92]提出 LHRS-Bot,利用志愿地理信息(volunteered geographic information,VGI)构建多层级视觉语言对齐数据,并通过指令学习增强地理场景理解;LI 等[96]提出 LHRS-Bot-Nova,通过改进视觉编码与跨模态连接模块,结合空间认知指令提升领域语义理解能力。然而,整图级视觉表征难以充分刻画密集小目标及复杂空间关系,推动研究由全局语义理解向细粒度空间解译拓展。KUCKREJA 等[13]提出 GeoChat,通过区域指令与坐标生成实现对象问答和目标定位;ZHANG 等[97]提出 EarthMarker,引入点和边界框等视觉提示,支持图像级、区域级和点级交互理解。在此基础上,研究进一步将语言指令与像素级预测相结合。OU 等[98]提出 GeoPix,通过语言驱动的掩膜预测机制实现细粒度地物分割;SHABBIR 等[99]提出 GeoPixel,结合高分辨率视觉编码与掩膜交织生成,支持最高 4 K 影像的单目标与多目标分割。除空间理解粒度的拓展外,多源遥感观测也推动生成式模型向多传感器联合理解发展。ZHANG 等[104]提出 EarthGPT,构建 MMRS-1M 数据集,通过统一指令微调融合光学、SAR 和红外影像,实现多传感器遥感解译;SONI 等[106]提出 EarthDial,进一步整合 RGB、SAR、多光谱及多时相影像,支持分类、检测、问答、定位与变化分析等任务。

遥感自回归生成式视觉语言模型逐步由领域指令适配向多任务统一建模、由全局场景理解向精细空间解译、由单一影像向多传感器联合理解拓展。然而,细粒度视觉信息损失、多模态语义冲突及语言先验引发的无依据生成仍制约解译可靠性,亟需加强视觉证据约束与空间语义一致性。

2.4　遥感文本条件生成模型

遥感影像理解与生成的协同建模旨在建立视觉内容与语言语义之间的双向映射,使模型既能够从影像中提取地理信息,又能够依据语义条件生成符合地物属性与空间关系的遥感场景。遥感文本条件影像生成与编辑的代表性模型见表 5。

表 5　遥感文本条件影像生成与编辑的代表性模型

Table 5　Representative Models for Text-Conditioned Remote Sensing Image Generation and Editing

代表模型(年份)	主要数据集	数据规模	数据模态	核心方法与任务
RSDiff[119](2024)	RSICD	10 921 幅影像	RGB / 文本	文本条件级联扩散;遥感影像生成
DiffusionSat[120](2024)	fMoW、Satlas、SpaceNet 等		RGB / 多光谱 / 文本 / 时空元数据	时空条件扩散;影像生成、超分辨与修复
Text2Earth[82](2025)	Git-10M	约 1 050 万图文对	RGB / 文本 / 地理信息	分辨率引导、动态条件适配;多尺度生成与编辑
RSEdit[121](2026)	遥感影像编辑数据	超过 6 万组双时相影像对	双时相 RGB / 文本指令	条件扩散编辑;局部地物修改、结构保持
GeoOne[122](2026)	GeoStructOmni	超过 47 万图文对	RGB / SAR / NIR / DEM / 地图 / 文本	时序语义解耦;影像生成、零样本描述

注:NIR(near infrared)为近红外;DEM(digital elevation model)为数字高程模型。

早期遥感文本条件生成模型研究主要关注文本条件影像生成,ZHAO 等[123]提出 StrucGAN,通过结构化生成对抗网络(generative adversarial network,GAN)学习增强文本与地物布局的一致性;SEBAQ 等[119]提出 RSDiff,利用文本条件与级联扩散生成遥感影像;TANG 等[124]

提出CRS-Diff,进一步融合文本、元数据及参考影像,提高生成过程的可控性。随着生成模型规模与语义建模能力的提升,LIU等[82]提出Text2Earth,基于大规模遥感图文数据构建生成基础模型,支持多分辨率影像生成与编辑。在此基础上,研究进一步由独立的条件生成向理解与生成统一建模发展。ZHANG等[122]提出GeoOne,通过时序语义解耦机制协调地理结构与语义条件,实现遥感影像生成和零样本描述。HOU等[125]提出UniEarth,基于Mamba-2架构构建共享表征,并通过跨模态语义锚定与难度感知优化联合学习影像生成、描述、问答和场景分类。进一步地,ZHANG等[126]提出Uni-RS,通过空间布局规划与空间感知查询监督,缓解影像理解与生成过程中的空间关系不一致问题,同时支持图像描述、视觉问答、目标定位及文本生成影像。相关研究呈现出由文本条件生成向视觉语言双向建模、由独立任务优化向理解与生成协同学习的发展趋势,为遥感场景语义一致性建模与可控数据构建提供了新的技术路径。

随着高分辨率多模态遥感视觉语言模型向复杂地理场景理解拓展,相关研究逐步聚焦于高分辨率感知、多时相变化理解与复杂地理推理。在高分辨率感知方面,GeoLLaVA-8K通过视觉词元筛选提升大幅面影像编码效率[112],ZoomEarth进一步引入主动感知机制,实现任务驱动的局部区域探索[113]。在时序理解方面,TEOChat通过多时相指令学习支持变化问答与灾害分析[110],DVLChat进一步拓展至变化量化和长期城市演化描述[111]。在地理推理方面,RSThinker通过感知约束的推理链增强视觉证据与推理过程的关联[116],GeoReason利用逻辑一致性强化学习提升推理与回答的一致性[117]。遥感理解基础模型正由静态场景识别向多尺度感知、动态过程理解及证据约束的复杂地理推理发展,但在跨尺度信息融合、时序变化解释及推理结果可靠性方面仍面临挑战。

3 遥感决策服务平台

在遥感影像基础模型与遥感理解基础模型等算法能力的支撑下,为实现从数据供给到智能决策的贯通,遥感平台的能力由数据共享逐步扩展至在线计算,再进一步延伸至面向任务的智能决策,可归纳为遥感数据服务平台、遥感计算服务平台与遥感软硬一体智能决策服务3类。其中,数据服务负责观测资料的整理、存储、检索与供给,输出可使用的数据及其质量信息;计算服务按既定流程运行算法与模型,输出解译结果或专题产品;决策服务则结合分析结果、业务目标与装备条件,确定任务优先级与后续行动。三者分别对应数据供给、分析处理与行动安排。一个平台可兼具多种能力,但应依据其实际提供的服务加以区分。

3.1 遥感数据服务平台

遥感数据服务平台以数据供给为主要职责,发展重点是将分散的影像档案组织为质量可查、时空关系明确的数据集合。澳大利亚地学数据立方体将长期卫星观测纳入统一时空网格,数字地球澳大利亚(digital earth Australia,DEA)在此基础上提供持续更新的数据产品[127-128]。Landsat分析就绪数据(analysis ready data,ARD)通过几何、辐射校正和质量标记,减少用户重复预处理的工作[129];HLS协调Landsat与Sentinel-2观测数据的差异,HLS 2.0进一步检验多传感器地表反射率的一致性[130-131]。这类平台将标准化处理前移至数据生产和发布环节,用户获取的既有影像像素数据,也有处理版本和质量信息。对于高分辨率多模态数据,统一网格只是组织条件,光学、SAR与高光谱的成像差异仍需在元数据和质量说明中保留。

数据供给从本地文件转向网络访问后,目录规范与存储布局成为平台建设的重要内容。STAC用统一元数据描述观测时间、覆盖范围和数据资产,于2025年成为开放地理空间联盟(open geospatial consortium,OGC)的社区标准[132];COG结合瓦片、概览和超文本传输协议(hypertext transfer protocol,HTTP)的范围请求,支持只读取影像中需要的区域[133],Zarr以分块数组组织多维数据[134]。这些方法使检索结果能够对应到具体的数据位置,并减少整景下载。软件中的索引与分块方式需要适应对象存储、缓存和网络条件:高分辨率影像可按空间窗口读取,多时相集合则需兼顾时间维度的访问。块过小会增加请求次数,过大又会读取无关数据,数据服务应根据常用访问方式配置存储,减少无效请求与多余读取。

数据服务还提供按需构建的数据集合和可复用的表征产品,地球系统数据立方体组织空间、时间和变量维度的数据[135],Cubo根据位置、分辨率与时段,从STAC目录和COG资产中生成面向学习任务的数据立方体[136]。Microsoft

Planetary Computer连接多源数据目录与访问接口[137],Rasteret通过可查询索引和按需像素读取减少云端COG访问的重复开销[138]。AlphaEarth Foundations提供由多源观测生成的年度地表嵌入,可作为少样本制图的共享表征[43]。

3.2　遥感计算服务平台

遥感计算服务平台接收数据和处理要求,完成算法执行、模型训练与推理,交付分析结果。Google Earth Engine把在线算子与分布式执行结合,减少本地下载和环境配置[139];开放地球引擎OGE通过数据、算法与算力的耦合支撑在线分析[140]。计算服务的发展由单机处理转向数据所在位置附近的并行执行。Gdalcubes利用延迟执行和分块处理组织计算[141],Xarray以带标签的多维数组保留维度含义[142],Pangeo连接云优化数据与按需分布式计算[143]。服务器的内存、存储及网络条件决定分块规模与并发上限,高分辨率任务还需处理跨块目标拼接和多模态配准。云到GPU吞吐研究表明,瓦片对齐读取与加载线程配置会影响训练数据供给[144],因此计算效率需同时考察运算、读取和传输开销。

遥感数据在线计算进一步要求处理流程能够跨平台复用,OpenEO用处理图描述数据加载、运算与输出[145],其联邦处理研究通过聚合器拆分跨云工作流,在不同后端之间传递中间结果[146]。OGC API-Processes提供处理任务与作业状态接口[147],Copernicus Data Space Ecosystem集成在线处理及数据访问接口[148]。基础模型的应用则使平台增加了训练、微调和模型服务:TorchGeo提供地理空间数据组合、采样与预处理组件[149],TerraTorch组织模型骨干、任务头和配置式微调[150],TerraStackAI连接数据准备、微调和推理服务器,并以弹性扩缩容支持批量及交互调用[151]。这类平台按照已确定的流程分配计算资源、管理依赖与模型版本。后端调度属于执行管理,其目标是完成计算作业,业务任务的优先次序和后续观测安排则由决策服务确定。

此外,计算能力向机载、星载设备延伸时,还需解决模型与资源受限边缘硬件的适配问题。SiamOHOT(Siamese network for onboard hyperspectral object tracking)通过空间-光谱蒸馏实现轻量高光谱跟踪[152],DisasterKD(disaster knowledge distillation)利用频率引导和跨解码器蒸馏保留灾害损毁结构[153],Prithvi与TerraMind轻量版探索基础模型在边缘设备上的复用[154],硬件感知架构搜索依据目标设备的实测时延与资源占用选择网络[155],Φ-Sat-1验证星上神经网络云筛查[156],WorldFloods研究低成本卫星洪水制图[157],后续试验展示地面重训练、模型上注与星上更新推理[158],RaVÆn以无监督变化检测筛选优先下传区域,并开展资源受限硬件验证[159],Φ-sat-2于2025年进入科学阶段,开展云检测、舰船检测、压缩与制图等应用[160]。SMFD(spiking multimodal fusion detection)融合微光和红外信息,结合脉冲计算与神经形态硬件验证夜间灾害受困人员实时应急搜救能力[161]。上述研究为近观测端处理提供了算法和设备基础,部署收益仍需按前处理、推理、数据搬运和产品输出的完整流程衡量。

3.3　遥感软硬一体智能决策服务

在遥感数据服务平台和计算服务平台建立的基础上,遥感软硬一体智能决策服务根据分析结果制定任务方案,并在执行中调整方案。其输入除遥感产品外,还包括业务目标、事件状态和可用装备;输出是观测、传输或核查安排,以及供业务人员判断的处置建议。GeoLLM-Engine提供多工具调用的地理空间任务环境[162],GeoLLM-Squad通过全局编排与专业智能体协作组织遥感工作流[163],Earth AI结合影像、人口和环境模型与推理智能体,研究跨模型、跨数据源分析[164],REMSA利用模型元数据,依据任务类型、输入数据和部署条件推荐基础模型[165]。

进一步地,遥感软硬一体决策服务的核心,是将遥感解译结果转化为可执行、可验证的任务安排,实现信息认知与实际行动的协同。平台综合云端分析与机载、星载筛查结果,结合数据来源、处理流程及质量信息,核验解译结论的可靠性,并依据任务目标与证据缺口,决定是否补充观测、调用更精细的模型或开展现场核查[166]。任务规划进一步统筹观测收益与设备约束,协同优化无人机航路、卫星任务调度及多平台信息获取顺序[167-170]。例如,针对变化检测发现的疑似灾损区域,系统可综合判断其重要程度、现有证据和设备剩余资源,确定补拍范围、观测时机与数据下传优先级。执行过程中,新获取的数据与现场核查结果持续反馈至解译和规划环节,动态修正灾情判断与后续任务。由此,数据获取、智能解译与任务执行不再是相互独立的处理环节,而是形成以任务目标为牵引、以软硬资源协同为保障、以反馈更新为驱动的闭环决策机制。

表6汇总了3类服务的代表性系统与方法。

表6 遥感数据、计算与软硬一体智能决策服务的代表性系统与方法

Table 6 Representative Systems and Methods for Three Types of Remote Sensing Platform Services

服务平台	系统或方法	年份	运行载体	关键技术	应用场景
遥感数据服务平台	澳大利亚地学数据立方体[127]	2017	地面/云端	数据一致化、数据立方体	时序监测
	HLS 2.0[131]	2025	地面/云端	数据一致化	时序监测、区域制图
	STAC 1.1.0[132]	2025	地面/云端	时空索引	数据检索
	COG 1.0[133]	2023	地面/云端	云原生存储	数据检索、区域制图
	cubo[136]	2024	地面/云端	时空索引、数据立方体	时序监测、区域制图
	AlphaEarth Foundations[43]	2025	云端集群	共享表征	区域制图
遥感计算服务平台	Google Earth Engine[139]	2017	云端集群	并行计算、工作流编排	时序监测、区域制图
	OGE[140]	2025	云端集群	数据立方体、工作流编排	时序监测、区域制图
	openEO[146]	2025	云端集群	工作流编排、并行计算	时序监测、区域制图
	Pangeo[1423	2021	云端集群	云原生存储、并行计算	时序监测
	TerraTorch[150]	2025	地面/云端	模型服务	区域制图、灾害评估
	TerraStackAI[151]	2026	地面/云端	工作流编排、模型服务	区域制图、灾害评估
	DisasterKD[153]	2026	机载平台	模型压缩、在线推理	灾害评估
	Prithvi/TerraMind轻量版[154]	2025	地面/云端	模型压缩、硬件适配	区域制图、灾害评估
	Φ-sat-2[160]	2025	星载平台	硬件适配、在线推理	区域制图、应急观测
	SMFD[161]	2026	机载平台	脉冲计算、硬件适配	应急观测
遥感软硬一体智能决策服务	GeoLLM-Squad[163]	2025	地面/云端	智能体编排、工作流编排	区域制图、灾害评估
	Earth AI[164]	2025	云端集群	智能体编排、模型服务	区域制图、灾害评估
	REMSA[165]	2026	地面/云端	智能体编排、模型服务	区域制图、灾害评估
	多无人机自适应路径规划[170]	2023	地面/云端	主动观测	时序监测、应急观测

注:STAC(spatiotemporal asset catalog)为空间时相资产目录;COG(cloud optimized GeoTIFF)为云优化GeoTIFF;OGE(open geospatial engine)为开放地球引擎;SMFD(spiking multimodal fusion detector)为脉冲多模态融合检测器;REMSA是面向遥感基础模型选择的约束感知智能体。

4 遥感智能解译脑与多场景应用

尽管遥感影像基础模型、遥感理解基础模型与遥感决策服务平台已取得长足进展,为高分辨率多模态遥感数据的智能感知、地学认知与应用决策奠定了重要的技术基础,但现有遥感智能解译体系及地学应用服务中,感知、认知与决策环节仍相对独立,制约了高分辨率多模态遥感感知能力,削弱了认知结果的可靠性,并降低了应用决策的响应效率。因此,如何打破感知、认知与决策之间的技术壁垒,构建贯通多模态感知、地学知识认知、自主推理规划与应用决策执行的全链路智能解译体系,成为推动遥感智能解译由单一任务处理向复杂地学问题自主求解发展的关键。自李德仁院士提出“对地观测脑(earth observation brain,EOB)[171]”以来,遥感智能解译不断模拟人脑感知和认知过程,以期构建集测量、目标感知、认知为一体的实时智能对地观测系统。遥感智能解译脑(remote sensing intelligent interpretation brain,RSI2_Brain)理论框架[14]正是对这一发展方向的初步探索,旨在借鉴人脑感知、认知与决策的协同机制,统筹遥感基础模型、地学知识、智能体与软硬件平台,探索形成感知-认知-决策一体化的理论框架与技术体系,并服务于常态监测与灾害应急等典型应用。

4.1 遥感智能解译脑技术思想来源与定义

人脑整合视觉、听觉、嗅觉、味觉和触觉等多模态信息,并通过学习、记忆、认知与决策实现灵活的信息处理。这些功能启发了类脑智能对通用学习能力与低能耗运行方式的探索。当前,类脑计算研究主要涉及模型-框架-平台3个层面[172],通过混合计算架构和互补感知通路等设计,探索提高计算能效与感知效率的实现方式[173-174]。借鉴这一功能组织方式,遥感智能解译

脑的总体技术思想如图 4 所示:多模态对地观测提供感知输入,层次化信息组织连接像素、目标、场景和知识表达,软硬协同支撑信息传递及决策响应,应用需求则通过反馈影响后续对地观测。

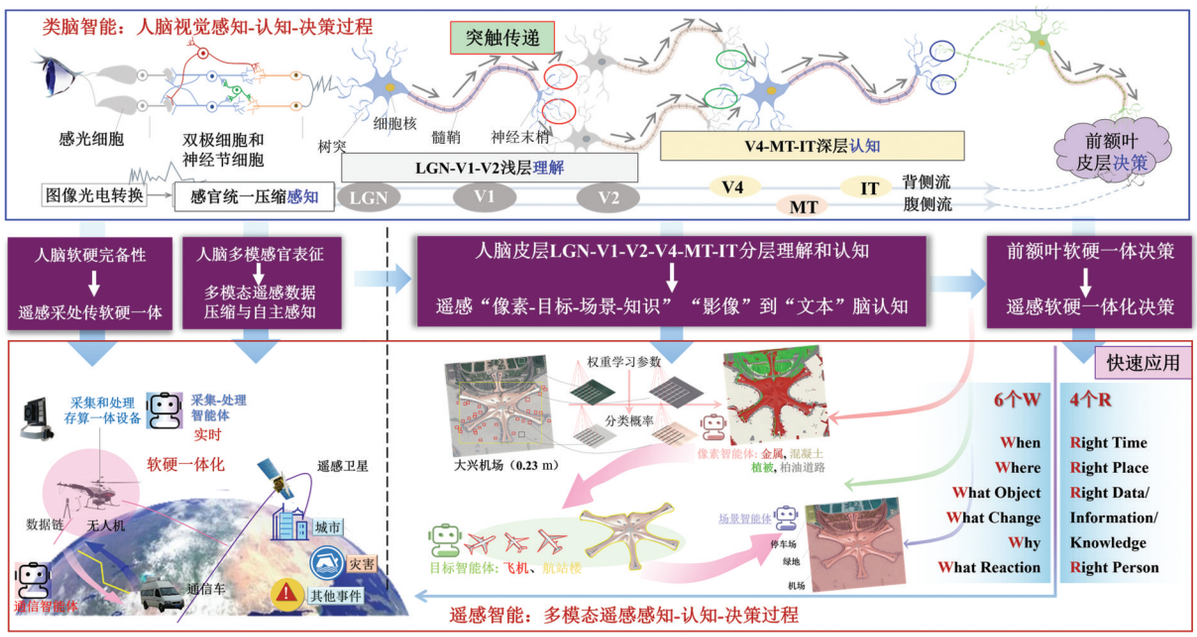


图 4　遥感智能解译脑总体技术思想来源[14](功能组织启发)

Fig. 4　Overall Conceptual Origins of the RSI2_Brain[14] (Inspired by Functional Organization)

具体而言,遥感智能解译脑借鉴人脑感知与认知系统的功能组织方式,构建采集—处理—传输软硬一体化系统。感知阶段参考多感官信息整合机制,对多模态遥感数据进行压缩与自主感知,识别地物及其状态;认知阶段借鉴外侧膝状体核(lateral geniculate nucleus,LGN)、初级视觉皮层(primary visual cortex,V1)、V2~V4 视觉区域(visual areas 2~4)、中颞视觉区(middle temporal area,MT)和下颞叶皮层(inferior temporal cortex,IT)的分层处理思想,结合空间关系、时间变化和领域知识,理解地物之间的联系及变化过程;决策阶段参考前额叶皮层的决策功能,依据认知结果与任务需求确定后续观测、传输、计算或行动方案。最终实现 6 个 W 和 4 个 R 的快速地学应用,6W 分别回答何时(When)、何地(Where)、何种对象(What Object)、发生何种变化(What Change)、为何变化(Why)及如何响应(What Reaction);4R 强调在恰当时间(Right Time)、恰当地点(Right Place),向合适人员(Right Person)提供所需数据、信息或知识(Right Data/Information/Knowledge)。感知、认知与决策构成连续的信息链,但并非简单串行流程。例如,决策端发现证据不足,可要求重新观测;认知端发现不同传感器的结果相互矛盾,可回到感知端检查配准与数据质量;感知端识别出异常分布,则应促使系统调整输出的可信度。因此,闭环需要将不确定性和资源状态与语义结果一起传递,为后续判断与行动提供依据。

由此,RSI2_Brain 可定义为一种具备多模遥感数据通用感知、模型自主学习可演化、像素-目标-场景-描述脑认知与推理、软硬一体化低功耗等能力,支撑快速决策应用的新型遥感解译体系[14]。如图 5 所示,首先,面向高分辨率多模态遥感数据的异构性与复杂性,通过多源传感器协同和多模态特征融合,实现遥感数据的通用感知与统一表征;其次,借鉴人脑的学习与记忆机制,结合模型自主学习和持续演化,实现遥感解译模型的动态适应与能力更新。在此基础上,面向像素、目标、场景与变化等多层次解译需求,融合遥感影像特征、领域知识与推理机制,推动遥感智能解译由信息提取向语义认知与逻辑推理发展。进一步地,通过遥感载荷、智能处理算法与嵌入式计算硬件的软硬件协同设计,实现数据采集、处理与传输的一体化协同,在降低系统能耗与响应时延的同时,支撑复杂场景下的快速决策与任务执行。最终推动遥感智能解译向多模态通用感知、自主学习演化、多层次认知推理与软硬件协同决策深度融合的新型智能体系发展。

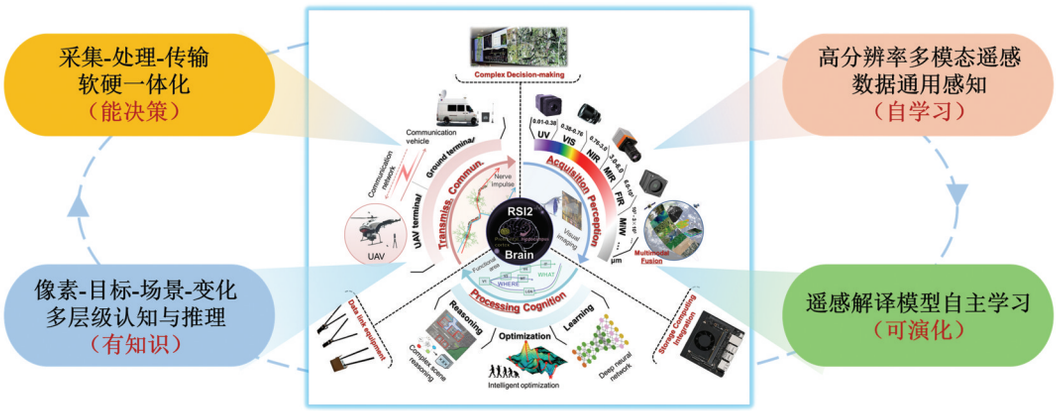
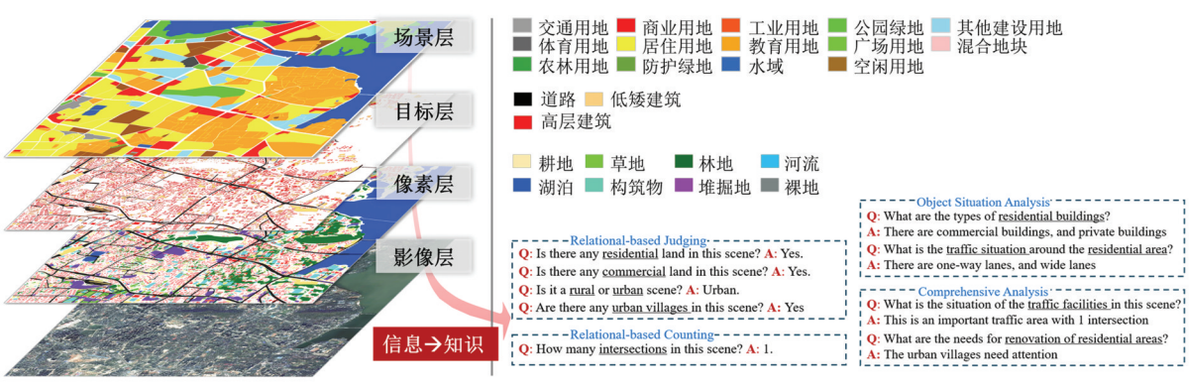


图5 遥感智能解译脑的定义（中间图来源于文献[14]）

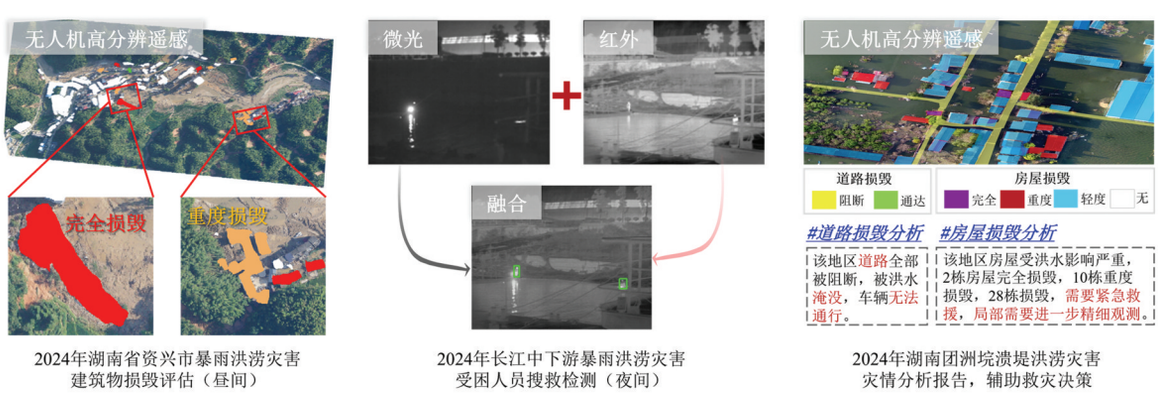
Fig. 5 Definition of RSI2_Brain (The Middle Figure is Sourced from Reference [14])

4.2 常态监测与灾害应急应用

遥感智能解译脑面向常态化对地观测与突发灾害应急等典型应用场景，融合多模态遥感感知、多层级地学认知、自主推理与任务决策能力，推动遥感解译由单一任务驱动的静态信息提取向感知-认知-决策闭环协同转变。在常态监测中，重点提升大范围地表信息的精细化解译与动态认知能力；在灾害应急中，则强调多源信息快速融合、灾情演变推理与救援任务自主规划，为复杂地学场景下的智能监测与快速响应提供支撑（图6）。



（a）对地常态监测示例：城市高分辨遥感像素—目标—场景—推理的多层级智能解译



2024年湖南省资兴市暴雨洪涝灾害建筑物损毁评估（昼间）

2024年长江中下游暴雨洪涝灾害受困人员搜救检测（夜间）

2024年湖南团洲垸溃堤洪涝灾害灾情分析报告，辅助救灾决策

（b）灾害应急响应示例：2024年长江中下游暴雨洪涝全天时实时灾害损毁评估、应急搜救与救灾决策

图6 对地常态监测与灾害应急响应应用示例

Fig. 6 Examples of Applications in Routine Earth Observation and Disaster Emergency Response

在对地常态监测方面，遥感智能解译脑通过像素-目标-场景-推理的多层级认知能力，支撑全国地表覆盖分类制图[175]、全球城市土地利用制图[176]与城市场景理解[177]等任务。例如，LIU等[175]利用低分辨率历史产品驱动深度学习模型训练，实现跨分辨率地表覆盖制图，在无需额外

标注的条件下完成全国尺度制图。此外,针对传统遥感分类以端到端静态处理为主、难以根据地物特征动态调整解译尺度的问题,相关研究引入强化学习智能体,将影像块作为决策单元,以分类精度提升为奖励信号,驱动智能体自主选择解译尺度[178]。结果表明,智能体的尺度调整行为与人类判读过程中对不同区域进行自适应缩放的方式具有一定相似性。这些研究推动遥感解译从固定尺度的静态分类向多层级认知与尺度自适应的动态解译拓展,为大范围、精细化地表监测提供了新的技术路径。

在灾害应急响应方面,遥感智能解译脑面向突发变化、信息不完整与通信受损等复杂条件,推动遥感应急模式由地面离线处理向采集、处理与传输一体化转变,为灾情评估与夜间搜救提供实时信息支持[14]。以洪涝灾害和夜间搜救为例,感知阶段综合利用卫星与无人机,以及SAR、可见光、微光和红外等多源数据,实现昼间道路、房屋等承灾体损毁评估[158,179]和夜间受困人员搜索与检测[166],并优先传输带有时空标记的关键事件信息;认知阶段结合遥感影像、文本资料和地学知识开展推理,生成暴雨洪涝灾害灾情分析报告,分析洪水演变、承灾体受损情况与交通可达性,为持续更新灾情判断和救灾决策提供依据;决策阶段依据灾情认知结果,制定重点区域补充观测、现场核查、救援力量配置及行动路线规划方案[167]。

5　未来发展方向

高分辨率多模态遥感智能解译正由面向单一任务的模型驱动处理,向多源数据协同、地学知识引导与自主决策执行相结合的遥感智能解译脑系统发展。尽管遥感影像基础模型、遥感理解基础模型与决策服务平台已取得显著进展,但在多模态统一表征、时空演化推理、知识驱动的遥感理解以及采集-处理-传输协同等方面仍存在不足。未来,随着遥感数据获取能力、人工智能技术与星载、机载及地面计算平台的持续发展,遥感智能解译将进一步突破数据、模型、知识与决策之间的技术壁垒,推动遥感基础模型更进一步向遥感智能解译脑演进。以下从数据、信息、知识、决策、系统与应用5个方面讨论未来发展方向。

5.1　数据:从空天地协同观测走向空天地人一体化感知

未来可借鉴人脑整合多感官输入的机制,推动遥感数据由独立采集向协同感知、由被动获取向任务驱动的按需观测转变。遥感数据获取将由卫星、航空器与地面设备协同观测,进一步拓展至融合现场调查、专家经验与人类反馈的空天地人一体化感知。不同平台与传感器在观测范围、空间分辨率和时间频率上具有互补性,而现场核查与专业知识能够为遥感观测提供语义补充和真实性验证。未来需加强多平台、多传感器与人类参与的协同观测,构建覆盖宏观监测、局部精细观测和现场反馈的多层次数据体系。

5.2　信息:从单一任务解译走向多模态统一表征与模拟生成

以类脑分层对齐机制为启发[180],遥感基础模型将由面向分类、检测与变化识别等单一任务的特征学习,走向像素-目标-场景-变化的多层级统一表征与多尺度协同处理。未来可重点发展高光谱基础模型与多模态高分辨率基础模型,强化光谱、空间及时序信息的联合建模;结合智能体技术[181]实现解译尺度与处理策略的自适应选择,并利用模拟与生成式模型[120]拓展训练样本。这些方法有助于构建兼具多模态信息融合、多层级语义关联与跨任务迁移能力的统一表征体系,突破传统方法对单一数据源、固定解译尺度和特定任务的依赖,提升复杂场景下的信息提取能力,增强模型在跨区域、跨传感器及样本稀缺条件下的泛化适应性。

5.3　知识:从遥感信息提取走向知识引导的可信地学认知

遥感理解基础模型使遥感分析从信息提取进一步拓展至地物关系理解与地学知识获取。未来需融合领域知识图谱、知识嵌入、地学规则与多模态语义推理,将遥感观测与领域知识关联起来,增强模型对复杂地学问题的解释和推断能力。例如,可构建由地学思维链驱动的遥感认知与推测智能体,将解译-推理-推测的一致性与可验证性作为奖励信号,优化思维链规划,以期实现遥感地学知识的自动结构化输出。同时,结合多源证据验证、不确定性评估与推理过程追溯,提升解译结果的可解释性与可信度[182],为遥感智能解译脑的自主决策提供可靠的知识支撑。

5.4　决策:从线下离线分析走向线上实时自主决策

遥感决策将由数据回传后的集中处理与人工研判,逐步转向面向任务需求的在线推理、动

态规划与自主执行。决策系统能够依据实时解译结果、地学认知结论、证据不确定性及任务时限,自主判断是否需要补充观测、调用更精细的解译模型、调整数据传输优先级或开展现场核查,并综合考虑通信带宽、计算资源与能源消耗等约束,动态优化任务分配与执行方案。在此基础上,进一步融合智能体自主规划、工具调用与执行反馈能力,实现观测任务、解译模型和决策策略的协同调整,推动遥感决策从预设流程驱动向任务目标驱动、从被动响应向主动规划转变。未来应进一步建立"认知驱动行动、行动反馈认知"的闭环决策机制[183]。

5.5 系统与应用:从遥感智能体走向软硬一体化的遥感智能解译脑

数据、信息、知识与决策能力的深度融合,最终指向面向不同地学应用场景的遥感智能解译脑体系构建。当前遥感智能解译研究仍以感知、认知、决策等单项能力的技术集成为主,缺乏贯通任务理解、信息获取、知识推理、决策规划与行动执行的系统级闭环机制。系统方面,进一步借鉴生物神经系统的信息编码与事件驱动处理机制,发展遥感类脑脉冲计算系统,融合多模态遥感感知、脉冲神经网络与神经形态计算硬件[166],探索面向动态地学场景的低功耗实时感知、快速推理与边缘自主响应机制,为遥感智能解译脑提供高效、低功耗的底层计算支撑。应用方面,未来应推动遥感智能解译从单一功能智能体向场景驱动的遥感智能解译脑演进,面向城市治理、农业监测、生态评估与灾害应急等领域,构建具备领域知识、任务目标与自主协同能力的专用解译脑体系。其核心是以感知、认知、决策与行动智能体为功能单元,通过任务编排、记忆共享、信息反馈与策略迭代,形成多智能体闭环协同机制[183]。同时,协同优化遥感基础模型部署、异构算力调度、传感器观测规划与通信传输,实现软件智能与星载、机载及地面硬件资源的深度融合,推动遥感智能解译由技术模块拼接走向软硬一体、自主协同与持续演化的系统化应用。

综上所述,高分辨率多模态遥感智能解译将沿着"空天地人一体化感知—多模态多尺度统一表征—知识驱动的地学认知—线上实时自主决策—软硬一体化系统集成与应用"的主线演进。人脑机制为各环节协同提供了重要启发,而遥感智能解译脑则是对上述能力进行系统整合的理论探索。

6 结 语

本文系统梳理了高分辨率多模态遥感智能解译的研究进展与发展趋势,围绕遥感影像基础模型、遥感理解基础模型和遥感决策服务平台3类技术体系,分析了遥感智能解译从通用表征学习向地学语义理解与自主决策演进的发展路径,并进一步阐述遥感智能解译脑理论框架及其应用。研究表明,现有方法已在高分辨率与高光谱影像表征、多模态语义对齐、复杂场景理解、地学知识增强及智能决策服务等方面取得积极进展,但在多源异构数据的统一表征、复杂地学知识的可信推理、感知与决策的闭环协同,以及资源受限条件下的软硬件一体化运行等方面仍存在不足。

未来,高分辨率多模态遥感智能解译将沿着空天地人一体化感知、多模态多尺度统一表征、知识引导的可信地学认知、线上实时自主决策与软硬一体化系统集成的方向发展。通过融合类脑感知机制、多模态基础模型、地学知识与智能体技术,可推动遥感系统形成贯通信息获取、语义理解、知识推理、任务规划与反馈更新的全链路能力,促进遥感智能解译由面向单一任务的模型智能向具备持续学习、可信认知、自主决策与软硬协同能力的系统智能演进,为复杂地学问题分析与多场景遥感应用提供新的技术思路。

参 考 文 献

[1] REICHSTEIN M, CAMPS-VALLS G, STEVENS B, et al. Deep Learning and Process Understanding for Data-Driven Earth System Science[J]. *Nature*, 2019, 566(7743): 195-204.

[2] HANAN N P, ANCHANG J Y. Satellites Could soon Map Every Tree on Earth[J]. *Nature*, 2020, 587(7832): 42-43.

[3] VOIGT S, GIULIO-TONOLO F, LYONS J, et al. Global Trends in Satellite-Based Emergency Mapping[J]. *Science*, 2016, 353(6296): 247-252.

[4] ZHU X X, TUIA D, MOU L C, et al. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources[J]. *IEEE Geoscience and Remote Sensing Magazine*, 2017, 5(4): 8-36.

[5] ZHANG L P, ZHANG L F, DU B. Deep Learning for Remote Sensing Data: A Technical Tutorial on

the State of the Art[J]. *IEEE Geoscience and Remote Sensing Magazine*, 2016, 4(2): 22-40.

[6] BREIMAN L. Random Forests [J]. *Machine Learning*, 2001, 45(1): 5-32.

[7] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet Classification with Deep Convolutional Neural Networks[J]. *Communications of the ACM*, 2017, 60(6): 84-90.

[8] OUYANG L, WU J, XU J, et al. Training Language Models to Follow Instructions with Human Feedback [C]//International Conference on Neural Information Processing Systems, New Orleans, Louisiana, USA, 2022.

[9] BOMMASANI R, HUDSON D A, ADELI E, et al. On the Opportunities and Risks of Foundation Models[PP/OL]. V3. arXiv (2022-07-12)[2026-10-11]. https://doi. org/10.48550/arXiv. 2108.07258.

[10] BROWN T B, MANN B, RYDER N, et al. Language Models Are Few-Shot Learners[C]// International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2020.

[11] CONG Y Z, KHANNA S, MENG C L, et al. SatMAE: Pre-Training Transformers for Temporal and Multi-Spectral Satellite Imagery[C]//Advances in Neural Information Processing Systems 35, New Orleans, Louisiana, USA, 2022.

[12] SUN X, WANG P J, LU W X, et al. RingMo: A Remote Sensing Foundation Model With Masked Image Modeling[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, 61: 5612822.

[13] KUCKREJA K, DANISH M S, NASEER M, et al. GeoChat: Grounded Large Vision-Language Model for Remote Sensing[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024.

[14] WAN Y T, ZHANG C, MA A L, et al. Remote Sensing Intelligent Interpretation Brain: Real-Time Intelligent Understanding of the Earth[J]. *PNAS Nexus*, 2025, 4(6): pgaf182.

[15] MENDIETA M, HAN B R, SHI X J, et al. Towards Geospatial Foundation Models via Continual Pretraining[C]//IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023.

[16] REED C J, GUPTA R, LI S F, et al. Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning[C]// IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023.

[17] IBANEZ D, FERNANDEZ-BELTRAN R, PLA F, et al. Masked Auto-Encoding Spectral - Spatial Transformer for Hyperspectral Image Classification [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 1-14.

[18] JAKUBIK J, ROY S, PHILLIPS C E, et al. Foundation Models for Generalist Geospatial Artificial Intelligence [PP/OL]. V2. arXiv (2023-11-08)[2026-10-10]. https://doi. org/10.48550/arXiv. 2310.18660.

[19] BASTANI F, WOLTERS P, GUPTA R, et al. SatlasPretrain: A Large-Scale Dataset for Remote Sensing Image Understanding[C]//IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023.

[20] IRVIN J, TAO L, ZHOU J, et al. USat: A Unified Self-Supervised Encoder for Multi-Sensor Satellite Imagery [PP/OL]. V1. arXiv (2023-12-02)[2026-10-10]. https://doi. org/10.48550/arXiv. 2312.02199.

[21] WANG Y, ALBRECHT C M, ALI BRAHAM N A, et al. Decoupling Common and Unique Representations for Multimodal Self-Supervised Learning [C]// European Conference—ECCV 2024, Milan, Italy, 2025.

[22] GUO X, LAO J W, DANG B, et al. SkySense: A Multi-Modal Remote Sensing Foundation Model Towards Universal Interpretation for Earth Observation Imagery[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024.

[23] FULLER A, MILLARD K, GREEN J. CROMA: Remote Sensing Representations with Contrastive Radar-Optical Masked Autoencoders[C]//Advances in Neural Information Processing Systems 36, New Orleans, Louisiana, USA, 2023.

[24] LI X Y, HONG D F, CHANUSSOT J. S2MAE: A Spatial-Spectral Pretraining Foundation Model for Spectral Remote Sensing Data [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024.

[25] NOMAN M, NASEER M, CHOLAKKAL H, et al. Rethinking Transformers Pre-Training for Multi-spectral Satellite Imagery [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024.

[26] LI J T, LIU Y Y, WANG X Y, et al. HyperFree: A Channel-Adaptive and Tuning-Free Foundation Model for Hyperspectral Remote Sensing Imagery [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN,

USA, 2025.

[27] LI Z H, HOU B, MA S T, et al. Masked Angle-Aware Autoencoder for Remote Sensing Images[C]// The 18th European Conference—ECCV 2024, Milan, Italy, 2025.

[28] TANG M F, COZMA A, GEORGIOU K, et al. Cross-Scale MAE: A Tale of Multiscale Exploitation in Remote Sensing[C]//Advances in Neural Information Processing Systems 36, New Orleans, Louisiana, USA, 2023.

[29] JIANG W T, ZHANG J, WANG D, et al. LeMeViT: Efficient Vision Transformer with Learnable Meta Tokens for Remote Sensing Image Interpretation[C]//The 33rd International Joint Conference on Artificial Intelligence, Jeju, South Korea, 2024.

[30] WANYAN X, SENEVIRATNE S, SHEN S C, et al. Extending Global-Local View Alignment for Self-Supervised Learning with Remote Sensing Imagery [C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2024.

[31] ZHANG X, LIU Y, LIN Y M, et al. UV-SAM: Adapting Segment Anything Model for Urban Village Identification[C]//The AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2024.

[32] LI W J, YANG W, HOU Y N, et al. SARATR-X: Toward Building a Foundation Model for SAR Target Recognition[J]. *IEEE Transactions on Image Processing*, 2025, 34: 869-884.

[33] XIONG Z T, WANG Y, ZHANG F H, et al. Neural Plasticity-Inspired Multimodal Foundation Model for Earth Observation[PP/OL]. V3. arXiv (2025-10-15)[2026-10-10]. https://doi.org/10.48550/arXiv.2403.15356.

[34] ASTRUC G, GONTHIER N, MALLET C, et al. OmniSat: Self-Supervised Modality Fusion for Earth Observation[C]// European Conference—ECCV 2024, Milan, Italy, 2025.

[35] LI W J, YANG W, LIU T P, et al. Predicting Gradient Is Better: Exploring Self-Supervised Learning for SAR ATR with a Joint-Embedding Predictive Architecture[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2024, 218: 326-338.

[36] GONG Z M, WANG A T, HUO X L, et al. CLIBD: Bridging Vision and Genomics for Biodiversity Monitoring at Scale [PP/OL]. V5. arXiv (2025-12-08)[2026-10-11]. https://doi.org/10.48550/arXiv.2405.17537.

[37] WANG D, HU M Q, JIN Y, et al. HyperSIGMA: Hyperspectral Intelligence Comprehension Foundation Model[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, 47(8): 6427-6444.

[38] WANG F X, WANG H Z, WANG D, et al. Harnessing Massive Satellite Imagery with Efficient Masked Image Modeling[C]// IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 2025.

[39] DIAO W H, YU H C, KANG K Y, et al. RingMo-Aerial: An Aerial Remote Sensing Foundation Model With Affine Transformation Contrastive Learning [PP/OL]. V4. arXiv (2025-09-16)[2026-10-10]. https://doi.org/10.48550/arXiv.2409.13366.

[40] SIMÉONI O, VO H V, SEITZER M, et al. DINOv3[PP/OL]. V1. arXiv (2025-08-13)[2026-10-10]. https://doi.org/10.48550/arXiv.2508.10104.

[41] ZHANG S, WANG Q Z, LIU J Y, et al. ALPS: An Auto-Labeling and Pre-Training Scheme for Remote Sensing Segmentation With Segment Anything Model[J]. *IEEE Transactions on Image Processing*, 2025, 34: 2408-2420.

[42] TRANTAS A, MENSIO M, STASINOS S, et al. BioAnalyst: A Foundation Model for Biodiversity[PP/OL]. V2. arXiv (2025-12-04)[2026-10-10]. https://doi.org/10.48550/arXiv.2507.09080.

[43] BROWN C F, KAZMIERSKI M R, PASQUARELLA V J, et al. AlphaEarth Foundations: An Embedding Field Model for Accurate and Efficient Global Mapping from Sparse Label Data [PP/OL]. V2. arXiv (2025-09-08)[2026-10-10]. https://doi.org/10.48550/arXiv.2507.22291.

[44] ASTRUC G, GONTHIER N, MALLET C, et al. AnySat: One Earth Observation Model for Many Resolutions, Scales, and Modalities[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2025.

[45] LI X Y, LI C Y, VIVONE G, et al. SeaMo: A Season-Aware Multimodal Foundation Model for Remote Sensing [J]. *Information Fusion*, 2026, 125: 103334.

[46] HERZOG H, BASTANI F, ZHANG Y W, et al. OlmoEarth: Stable Latent Image Modeling for Multimodal Earth Observation [PP/OL]. V1. arXiv (2025-11-17)[2026-10-10]. https://doi.org/10.48550/arXiv.2511.13655.

[47] BI H B, FENG Y C, TONG B Y, et al. RingMoE: Mixture-of-Modality-Experts Multi-Modal Foundation Models for Universal Remote Sensing

Image Interpretation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026, 48(4): 4388-4405.

[48] SASTRY S, KHANAL S, DHAKAL A, et al. TaxaBind: A Unified Embedding Space for Ecological Applications[C]//IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 2025.

[49] SZWARCMAN D, ROY S, FRACCARO P, et al. Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications[PP/OL]. V3. arXiv (2026-03-06)[2026-10-10]. https://doi. org/10.48550/arXiv. 2412.02732.

[50] FENG Z, ATZBERGER C, JAFFER S, et al. TESSERA: Temporal Embeddings of Surface Spectra for Earth Representation and Analysis[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Denver, Colorado, USA, 2026.

[51] YANG R, LIU Y, YAN H, et al. MaRS: A Multimodality Very-High-Resolution Remote Sensing Foundation Model with Cross-Granularity Meta-Modality Learning[C]//The AAAI Conference on Artificial Intelligence, Vancouver, British Columbia, Canada, 2026.

[52] HONG D F, ZHANG B, LI X Y, et al. SpectralGPT: Spectral Remote Sensing Foundation Model[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(8): 5227-5244.

[53] WU K, ZHANG Y Y, RU L X, et al. A Semantic-Enhanced Multi-Modal Remote Sensing Foundation Model for Earth Observation[J]. *Nature Machine Intelligence*, 2025, 7(8): 1235-1249.

[54] QU B, LI X L, TAO D C, et al. Deep Semantic Understanding of High Resolution Remote Sensing Image[C]//International Conference on Computer, Information and Telecommunication Systems (CITS), Kunming, China, 2016.

[55] LU X Q, WANG B Q, ZHENG X T, et al. Exploring Models and Data for Remote Sensing Image Caption Generation[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2018, 56(4): 2183-2195.

[56] CHENG Q M, HUANG H Y, XU Y, et al. NWPU-Captions Dataset and MLCA-Net for Remote Sensing Image Captioning[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 5629419.

[57] HU Y, YUAN J L, WEN C C, et al. RSGPT: A Remote Sensing Vision Language Model and Benchmark[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2025, 224: 272-286.

[58] LOBRY S, MARCOS D, MURRAY J, et al. RSVQA: Visual Question Answering for Remote Sensing Data[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2020, 58(12): 8555-8566.

[59] ZHENG X T, WANG B Q, DU X Q, et al. Mutual Attention Inception Network for Remote Sensing Visual Question Answering[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 5606514.

[60] LOBRY S, DEMIR B, TUIA D. RSVQA Meets Bigearthnet: A New, Large-Scale, Visual Question Answering Dataset for Remote Sensing[C]//IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 2021.

[61] RAHNEMOONFAR M, CHOWDHURY T, SARKAR A, et al. FloodNet: A High Resolution Aerial Imagery Dataset for Post Flood Scene Understanding[J]. *IEEE Access*, 2021, 9: 89644-89654.

[62] LI K, VOSSELMAN G, YANG M Y. HRVQA: A Visual Question Answering Benchmark for High-Resolution Aerial Images[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2024, 214: 65-81.

[63] WANG J J, ZHENG Z, CHEN Z H, et al. EarthVQA: Towards Queryable Earth *via* Relational Reasoning-Based Remote Sensing Visual Question Answering[C]//The AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2024.

[64] YUAN Z Q, ZHANG W K, FU K, et al. Exploring a Fine-Grained Multiscale Method for Cross-Modal Remote Sensing Image Retrieval[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 4404119.

[65] ZHAN Y, XIONG Z T, YUAN Y. RSVG: Exploring Data and Models for Visual Grounding on Remote Sensing Data[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, 61: 5604513.

[66] YUAN Z H, MOU L C, HUA Y S, et al. RRSIS: Referring Remote Sensing Image Segmentation[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 5613312.

[67] LIU S H, MA Y W, ZHANG X Q, et al. Rotated Multi-scale Interaction Network for Referring Remote Sensing Image Segmentation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024.

[68] LIU C Y, ZHAO R, CHEN H, et al. Remote Sensing Image Change Captioning With Dual-

Branch Transformers: A New Method and a Large Scale Dataset[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 5633520.

[69] YUAN Z H, MOU L C, XIONG Z T, et al. Change Detection Meets Visual Question Answering [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 5630613.

[70] KARACA A C, OZELBAS E, BERBER S, et al. Robust Change Captioning in Remote Sensing: SECOND-CC Dataset and MModalCC Framework [J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025, 18: 21494-21513.

[71] GAJBHIYE G O, NANDEDKAR A V. Generating the Captions for Remote Sensing Images: A Spatial-Channel Attention Based Memory-Guided Transformer Approach [J]. *Engineering Applications of Artificial Intelligence*, 2022, 114: 105076.

[72] ZHAO K, XIONG W. Region-Guided Transformer for Remote Sensing Image Captioning [J]. *International Journal of Digital Earth*, 2024, 17 (1): 2400988.

[73] ABDULLAH T, BAZI Y, AL RAHHAL M M, et al. TextRS: Deep Bidirectional Triplet Network for Matching Text to Remote Sensing Images [J]. *Remote Sensing*, 2020, 12(3): 405.

[74] CHENG Q M, ZHOU Y Z, FU P, et al. A Deep Semantic Alignment Network for the Cross-Modal Image-Text Retrieval in Remote Sensing[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021, 14: 4284-4297.

[75] HE Y P, XU X, CHEN H J, et al. Visual Global-Salient-Guided Network for Remote Sensing Image-Text Retrieval[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 5641814.

[76] ZHANG Z L, ZHAO T C, GUO Y L, et al. RS5M and GeoRSCLIP: A Large-Scale Vision-Language Dataset and a Large Vision-Language Model for Remote Sensing [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 5642123.

[77] LIU F, CHEN D L, GUAN Z, et al. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 5622216.

[78] WANG Z C, PRABHA R, HUANG T Y, et al. SkyScript: A Large and Semantically Diverse Vision-Language Dataset for Remote Sensing [C]// The AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2024.

[79] MALL U, PHOO C P, LIU M K, et al. Remote Sensing Vision-Language Foundation Models Without Annotations via Ground Remote Alignment[C]//International Conference on Learning Representations, Vienna, Austria, 2024.

[80] DHAKAL A, AHMAD A, KHANAL S, et al. Sat2Cap: Mapping Fine-Grained Textual Descriptions from Satellite Images [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2024.

[81] JAIN P, IENCO D, INTERDONATO R, et al. SenCLIP: Enhancing Zero-Shot Land-Use Mapping for Sentinel-2 with Ground-Level Prompting [C]//IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 2025.

[82] LIU C Y, CHEN K Y, ZHAO R, et al. Text2Earth: Unlocking Text-Driven Remote Sensing Image Generation with a Global-Scale Dataset and a Foundation Model [J]. *IEEE Geoscience and Remote Sensing Magazine*, 2025, 13(3): 238-259.

[83] HE Y G, CHENG X J, ZHU J J, et al. SAR-TEXT: A Large-Scale SAR Image-Text Dataset Built with SAR-Narrator and a Progressive Learning Strategy for Downstream Tasks[PP/OL]. V3. arXiv (2025-10-04)[2026-10-10]. https://doi.org/10.48550/arXiv.2507.18743.

[84] JIANG C W, WANG C, WU F, et al. SARCLIP: A Multimodal Foundation Framework for SAR Imagery via Contrastive Language-Image Pre-Training [J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2026, 231: 17-34.

[85] JAIN P, MARCOS D, IENCO D, et al. TimeSenCLIP: A Time Series Vision-Language Model for Remote Sensing [J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2026, 236: 99-119.

[86] ZHU Q, LAO J W, JI D Y, et al. SkySense-O: Towards Open-World Remote Sensing Interpretation with Vision-Centric Visual-Language Modeling [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2025.

[87] LI K Y, LIU R X, CAO X Y, et al. SegEarth-OV: Towards Training-Free Open-Vocabulary Segmentation for Remote Sensing Images [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2025.

[88] CAO Q L, CHEN Y T, MA C, et al. Open-Vocabulary High-Resolution Remote Sensing Image

Semantic Segmentation[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2025, 63: 5620614.

[89] RADFORD A, KIM J W, HALLACY C, et al. Learning Transferable Visual Models from Natural Language Supervision[C]//The 38th International Conference on Machine Learning, Online, 2021.

[90] ZHONG Y W, YANG J W, ZHANG P C, et al. RegionCLIP: Region-Based Language-Image Pretraining[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022.

[91] BAZI Y, BASHMAL L, AL RAHHAL M M, et al. RS-LLaVA: A Large Vision-Language Model for Joint Captioning and Question Answering in Remote Sensing Imagery[J]. *Remote Sensing*, 2024, 16(9): 1477.

[92] MUHTAR D, LI Z S, GU F, et al. LHRS-Bot: Empowering Remote Sensing With VGI-Enhanced Large Multimodal Language Model[C]//Computer Vision—ECCV 2024. 2025.

[93] LUO J W, PANG Z, ZHANG Y J, et al. SkySenseGPT: A Fine-Grained Instruction Tuning Dataset and Model for Remote Sensing Vision-Language Understanding[PP/OL]. V2. arXiv (2024-07-08)[2026-10-10]. https://doi.org/10.48550/arXiv.2406.10100.

[94] ZHAN Y, XIONG Z T, YUAN Y. SkyEyeGPT: Unifying Remote Sensing Vision-Language Tasks via Instruction Tuning with Large Language Model[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2025, 221: 64-77.

[95] PANG C, WENG X X, WU J, et al. VHM: Versatile and Honest Vision Language Model for Remote Sensing Image Analysis[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, 39(6): 6381-6388.

[96] LI Z S, MUHTAR D, GU F, et al. LHRS-Bot-Nova: Improved Multimodal Large Language Model for Remote Sensing Vision-Language Interpretation[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2025, 227: 539-550.

[97] ZHANG W, CAI M X, ZHANG T, et al. EarthMarker: A Visual Prompting Multimodal Large Language Model for Remote Sensing[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2025, 63: 5604219.

[98] OU R Z, HU Y, ZHANG F, et al. GeoPix: A Multimodal Large Language Model for Pixel-Level Image Understanding in Remote Sensing[J]. *IEEE Geoscience and Remote Sensing Magazine*, 2025, 13(3): 324-337.

[99] SHABBIR A, ZUMRI M, BENNAMOUN M, et al. GeoPixel: Pixel Grounding Large Multimodal Model in Remote Sensing[C]// International Conference on Machine Learning, Vancouver, Canada, 2025.

[100] LI K Y, XIN Z P, PANG L, et al. SegEarth-R1: Geospatial Pixel Reasoning *via* Large Language Model[PP/OL]. V1. arXiv (2025-04-13)[2026-10-10]. https://doi.org/10.48550/arXiv.2504.09644.

[101] XIN Z, LI K, CHEN L, et al. SegEarth-R2: Towards Comprehensive Language-Guided Segmentation for Remote Sensing Images[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, Tennessee, USA, 2025.

[102] NI S, WANG D, CHEN H, et al. UniGeoSeg: Towards Unified Open-World Segmentation for Geospatial Scenes[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition, Denver, Colorado, USA, 2026.

[103] AN X, ZHONG C, SUN J X, et al. Beyond Single and Earthbound: Exploring Multi-Image Grounding in Remote Sensing with Large Vision-Language Models[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2026, 238: 589-605.

[104] ZHANG W, CAI M X, 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: 5917820.

[105] ZHANG W, CAI M X, NING Y Q, et al. EarthGPT-X: A Spatial MLLM for Multilevel Multisource Remote Sensing Imagery Understanding With Visual Prompting[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2025, 63: 4709221.

[106] SONI S, DUDHANE A, DEBARY H, et al. EarthDial: Turning Multi-Sensory Earth Observations to Interactive Dialogues[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2025.

[107] SHU Y, REN B, XIONG Z T, et al. EarthMind: Leveraging Cross-Sensor Data for Advanced Earth Observation Interpretation with a Unified Multimodal LLM[PP/OL]. V2. arXiv (2025-09-28)[2026-10-10]. https://doi.org/10.48550/arXiv.2506.01667.

[108] CAI M X, WANG G Q, ZHANG W, et al. EarthOneVision: Extending Remote Sensing Multimodal

Large Language Models to More Sensor Modalities and Tasks[PP/OL]. arXiv (2026-06-09)[2026-10-10]. https://doi. org/10.48550/arXiv. 2606.10819.

[109]DENG P, ZHOU W Q, WU H L. ChangeChat: An Interactive Model for Remote Sensing Change Analysis *via* Multimodal Instruction Tuning[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025.

[110]IRVIN J A, LIU E R, CHEN J C, et al. TEOChat: A Large Vision-Language Assistant for Temporal Earth Observation Data [C]//International Conference on Learning Representations (ICLR), Singapore, 2025.

[111]XUAN W H, WANG J J, QI H L, et al. DynamicVL: Benchmarking Multimodal Large Language Models for Dynamic City Understanding[C]//Advances in Neural Information Processing Systems 38, San Diego, California, USA and Mexico City, Mexico, 2025.

[112]WANG F X, CHEN M S, LI Y Y, et al. GeoLLaVA-8K: Scaling Remote-Sensing Multimodal Large Language Models to 8K Resolution [PP/OL]. V2. arXiv (2025-11-04) [2026-10-10]. https://doi. org/10.48550/arXiv. 2505. 21375.

[113]LIU R X, FU B W, SONG J Y, et al. ZoomEarth: Active Perception for Ultra-High-Resolution Geospatial Vision-Language Tasks[PP/OL]. V1. arXiv (2025-11-15)[2026-10-10]. https://doi. org/10.48550/arXiv. 2511. 12267.

[114]ZHU J S, FU R H, HU J S, et al. GeoVista: Visually Grounded Active Perception for Vision-Language Understanding of Ultra-High-Resolution Remote Sensing Images[PP/OL]. V2. arXiv (2026-06-21)[2026-10-10]. https://doi. org/10.48550/arXiv. 2605. 14475.

[115]FIAZ M, DEBARY H, FRACCARO P, et al. GeoV-LM-R1: Reinforcement Fine-Tuning for Improved Remote Sensing Reasoning [PP/OL]. V3. arXiv (2025-10-14)[2026-10-10]. https://doi. org/10.48550/arXiv. 2509. 25026.

[116]LIU J, SUN L, FU R, et al. Towards Faithful Reasoning in Remote Sensing: A Perceptually-Grounded GeoSpatial Chain-of-Thought for Vision-Language Models[C]//International Conference on Learning Representations, Sydney, Australia, 2026.

[117]LI W S, XIANG X T, WEN Z X, et al. GeoReason: Aligning Thinking and Answering in Remote Sensing Vision-Language Models *via* Logical Consistency Reinforcement Learning [PP/OL]. V2. arXiv (2026-01-08) [2026-10-10]. https://doi. org/10.48550/arXiv. 2601. 04118.

[118]LIU H T, LI C Y, WU Q Y, et al. Visual Instruction Tuning[C]//Advances in Neural Information Processing Systems 36, New Orleans, Louisiana, USA, 2023.

[119]SEBAQ A, ELHELW M. RSDiff: Remote Sensing Image Generation from Text Using Diffusion Model [J]. *Neural Computing and Applications*, 2024, 36(36): 23103-23111.

[120]KHANNA S, LIU P, ZHOU L Q, et al. DiffusionSat: A Generative Foundation Model for Satellite Imagery[C]//International Conference on Learning Representations (ICLR), Vienna, Austria, 2024.

[121]CHEN Z Y, ZHANG Z C, ZHANG F. RSEdit: Text-Guided Image Editing for Remote Sensing [PP/OL]. V2. arXiv (2026-05-18)[2026-10-10]. https://doi. org/10.48550/arXiv. 2603. 13708.

[122]ZHANG W Z, CHEN Z L, ZHANG M X, et al. GeoOne: Unifying Structural and Semantic Control in Diffusion Models *via* Temporal Decoupling[J]. *Knowledge-Based Systems*, 2026, 343: 116023.

[123]ZHAO R, SHI Z W. Text-to-Remote-Sensing-Image Generation With Structured Generative Adversarial Networks[J]. *IEEE Geoscience and Remote Sensing Letters*, 2022, 19: 8010005.

[124]TANG D T, CAO X Y, HOU X S, et al. CRS-Diff: Controllable Remote Sensing Image Generation With Diffusion Model[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 5638714.

[125]HOU Y Y, WEI Z H, REN C, et al. UniEarth: A Unified Foundation Model for Remote Sensing Image Generation and Understanding[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2026, 64: 4707619.

[126]ZHANG W Y, HU Y, LI Y, et al. Uni-RS: A Spatially Faithful Unified Understanding and Generation Model for Remote Sensing[PP/OL]. V1. arXiv (2026-01-25)[2026-10-10]. https://doi. org/10.48550/arXiv. 2601. 17673.

[127]LEWIS A, OLIVER S, LYMBURNER L, et al. The Australian Geoscience Data Cube—Foundations and Lessons Learned[J]. *Remote Sensing of Environment*, 2017, 202: 276-292.

[128]DHU T, DUNN B, LEWIS B, et al. Digital Earth Australia-Unlocking New Value from Earth Observation Data[J]. *Big Earth Data*, 2017, 1(1/2): 64-74.

[129]DWYER J L, ROY D P, SAUER B, et al. Analy-

sis Ready Data: Enabling Analysis of the Landsat Archive[J]. *Remote Sensing*, 2018, 10(9): 1363.

[130]CLAVERIE M, JU J C, MASEK J G, et al. The Harmonized Landsat and Sentinel-2 Surface Reflectance Data Set[J]. *Remote Sensing of Environment*, 2018, 219: 145-161.

[131]JU J C, ZHOU Q, FREITAG B, et al. The Harmonized Landsat and Sentinel-2 Version 2.0 Surface Reflectance Dataset[J]. *Remote Sensing of Environment*, 2025, 324: 114723.

[132]Open Geospatial Consortium. SpatioTemporal Asset Catalog (STAC) Community Standard 1.1.0: OGC 25-004[S/OL]. 2025[2026-10-10]. https://www.ogc.org/standards/stac/.

[133]Open Geospatial Consortium. OGC Cloud Optimized GeoTIFF Standard 1.0: OGC 21-026[S/OL]. (2023-07-14)[2026-10-10]. https://docs.ogc.org/is/21-026/21-026.html.

[134]Zarr Developers. Zarr Core Specification Version 3[S/OL]. 2023[2026-10-10]. https://zarr-specs.readthedocs.io/en/latest/v3/core/index.html.

[135]MAHECHA M D, GANS F, BRANDT G, et al. Earth System Data Cubes Unravel Global Multivariate Dynamics[J]. *Earth System Dynamics*, 2020, 11(1): 201-234.

[136]MONTERO D, AYBAR C, JI C N, et al. On-Demand Earth System Data Cubes[C]// IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 2024.

[137]Microsoft. Planetary Computer Documentation[EB/OL]. [2026-10-10]. https://planetarycomputer.microsoft.com/docs/changelogs/history.

[138]Terrafloww. Rasteret 0.3.5[EB/OL]. (2026-03-13)[2026-10-10]. https://github.com/terrafloww/rasteret/releases/tag/v0.3.5.

[139]GORELICK N, HANCHER M, DIXON M, et al. Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone[J]. *Remote Sensing of Environment*, 2017, 202: 18-27.

[140]龚健雅, 乐鹏, 向隆刚, 等. 开放地球引擎OGE设计与研制[J]. 测绘学报, 2025, 54(4): 587-602.
GONG Jianya, LE Peng, XIANG Longgang, et al. The Design and Implementation of the Open Geospatial Engine(OGE)[J]. *Acta Geodaetica et Cartographica Sinica*, 2025, 54(4): 587-602.

[141]APPEL M, PEBESMA E. On-Demand Processing of Data Cubes from Satellite Image Collections with the Gdalcubes Library[J]. *Data*, 2019, 4(3): 92.

[142]HOYER S, HAMMAN J. Xarray: N-D Labeled Arrays and Datasets in Python[J]. *Journal of Open Research Software*, 2017, 5(1): 10.

[143]ABERNATHEY R P, AUGSPURGER T, BANIHIRWE A, et al. Cloud-Native Repositories for Big Scientific Data[J]. *Computing in Science & Engineering*, 2021, 23(2): 26-35.

[144]ZAYTAR A, ROBINSON C, TADESSE G A, et al. Optimizing Cloud-to-GPU Throughput for Deep Learning With Earth Observation Data[PP/OL]. V1. arXiv (2025-06-06)[2026-10-10]. https://doi.org/10.48550/arXiv.2506.06235.

[145]SCHRAMM M, PEBESMA E, MILENKOVIĆ M, et al. The openEO API - Harmonising the Use of Earth Observation Cloud Services Using Virtual Data Cube Functionalities[J]. *Remote Sensing*, 2021, 13(6): 1125.

[146]MOHR M, PEBESMA E, DRIES J, et al. Federated and Reusable Processing of Earth Observation Data[J]. *Scientific Data*, 2025, 12: 194.

[147]Open Geospatial Consortium. OGC API-Processes Part 1: Core 1.0.0: OGC 18-062r2[S/OL]. (2021-12-20)[2026-10-10]. https://docs.ogc.org/is/18-062r2/18-062r2.html.

[148]Copernicus Data Space Ecosystem. APIs-Documentation[EB/OL]. [2026-10-10]. https://documentation.dataspace.copernicus.eu/APIs.html.

[149]STEWART A J, ROBINSON C, CORLEY I A, et al. TorchGeo: Deep Learning with Geospatial Data[C]//The 30th International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, 2022.

[150]GOMES C, BLUMENSTIEL B, DE SOUSA ALMEIDA J L, et al. TerraTorch: The Geospatial Foundation Models Toolkit[C]// IEEE International Geoscience and Remote Sensing Symposium, Brisbane, Australia, 2025.

[151]Watson C, Wachira C, Lickorish R, et al. TerraStackAI: Bringing Earth and Space AI to Red Hat and the World[EB/OL]. (2026-03-03)[2026-10-10]. https://research.ibm.com/blog/terrastackai.

[152]SUN C, WANG X Y, LIU Z Q, et al. SiamOHOT: A Lightweight Dual Siamese Network for Onboard Hyperspectral Object Tracking via Joint Spatial-Spectral Knowledge Distillation[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, 61: 5521112.

[153]GUO J C, WAN Y T, WANG J J, et al. Disaster-KD: Frequency-Guided Cross-Decoder Knowledge Distillation for UAV Real-Time Disaster Damage Assessment[J]. *Pattern Recognition*, 2026, 174: 112968.

[154]IBM Research. AI-Powered Satellites Will Upend How We Observe Our Changing Planet[EB/OL]. (2025-10-10)[2026-10-10]. https://research.ibm.com/blog/terramind-prithvi-tiny-small-models-geospatial.

[155]DEL PRETE R, THIND P K, MAZZEO A, et al. Optimizing Deep Learning Models for On-Orbit Deployment Through Neural Architecture Search[J]. *Scientific Reports*, 2025, 15: 37783.

[156]GIUFFRIDA G, FANUCCI L, MEONI G, et al. The Φ-Sat-1 Mission: The First On-Board Deep Neural Network Demonstrator for Satellite Earth Observation[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 5517414.

[157]MATEO-GARCIA G, VEITCH-MICHAELIS J, SMITH L, et al. Towards Global Flood Mapping Onboard Low Cost Satellites with Machine Learning[J]. *Scientific Reports*, 2021, 11: 7249.

[158]MATEO-GARCIA G, VEITCH-MICHAELIS J, PURCELL C, et al. In-Orbit Demonstration of a Re-Trainable Machine Learning Payload for Processing Optical Imagery[J]. *Scientific Reports*, 2023, 13: 10391.

[159]RŮŽIČKA V, VAUGHAN A, DE MARTINI D, et al. RaVÆn: Unsupervised Change Detection of Extreme Events Using ML On-Board Satellites[J]. *Scientific Reports*, 2022, 12: 16939.

[160]ESA. Φ-sat-2 Begins Science Phase for AI Earth Images[EB/OL]. (2025-07-15)[2026-10-10]. https://www.esa.int/Applications/Observing_the_Earth/FutureEO/Phsat-2/Phsat-2_begins_science_phase_for_AI_Earth_images.

[161]LIU S Y, WAN Y T, LI K F, et al. A Brain-Inspired Low-Light and Infrared Multimodal Remote Sensing Detection Network for Nighttime Personnel Search and Rescue[J]. *Information Fusion*, 2026, 136: 104518.

[162]SINGH S, FORE M, STAMOULIS D. GeoLLM-Engine: A Realistic Environment for Building Geospatial Copilots[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2024.

[163]LEE C, PARAMANAYAKAM V, KARATZAS A, et al. Multi-Agent Geospatial Copilots for Remote Sensing Workflows[C]// IEEE International Geoscience and Remote Sensing Symposium, Brisbane, Australia, 2025.

[164]BELL A, AIDES A, HELMY A, et al. Earth AI: Unlocking Geospatial Insights with Foundation Models and Cross-Modal Reasoning[PP/OL]. V4. arXiv (2026-02-18)[2026-10-10]. https://doi.org/10.48550/arXiv.2510.18318.

[165]CHEN B E, BÖK T E, RASTI B, et al. REMSA: Foundation Model Selection for Remote Sensing via a Constraint-Aware Agent[PP/OL]. V3. arXiv (2026-06-29)[2026-10-10]. https://doi.org/10.48550/arXiv.2511.17442.

[166]MUNIR M A, SHEIKH M U, SHABBIR A, et al. Agentic AI for Remote Sensing: Technical Challenges and Research Directions[PP/OL]. V3. arXiv (2026-06-01)[2026-10-10]. https://doi.org/10.48550/arXiv.2604.24919.

[167]WAN Y T, ZHONG Y F, MA A L, et al. An Accurate UAV 3-D Path Planning Method for Disaster Emergency Response Based on an Improved Multiobjective Swarm Intelligence Algorithm[J]. *IEEE Transactions on Cybernetics*, 2023, 53(4): 2658-2671.

[168]SONG Y J, OU J W, PEDRYCZ W, et al. Generalized Model and Deep Reinforcement Learning-Based Evolutionary Method for Multitype Satellite Observation Scheduling[J]. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024, 54(4): 2576-2589.

[169]POPOVIĆ M, VIDAL-CALLEJA T, HITZ G, et al. An Informative Path Planning Framework for UAV-Based Terrain Monitoring[J]. *Autonomous Robots*, 2020, 44(6): 889-911.

[170]WESTHEIDER J, RÜCKIN J, POPOVIĆ M. Multi-UAV Adaptive Path Planning Using Deep Reinforcement Learning[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, USA, 2023.

[171]LI D R, WANG M, DONG Z P, et al. Earth Observation Brain (EOB): An Intelligent Earth Observation System[J]. *Geo-spatial Information Science*, 2017, 20(2): 134-140.

[172]ZHANG W H, MA S C, JI X L, et al. The Development of General-Purpose Brain-Inspired Computing[J]. *Nature Electronics*, 2024, 7(11): 954-965.

[173]PEI J, DENG L, SONG S, et al. Towards Artificial General Intelligence with Hybrid Tianjic Chip Architecture[J]. *Nature*, 2019, 572(7767): 106-111.

[174]YANG Z Y, WANG T Y, LIN Y H, et al. A Vision Chip with Complementary Pathways for Open-World Sensing[J]. *Nature*, 2024, 629(8014): 1027-1033.

[175]LIU Y H, ZHONG Y F, MA A L, et al. Cross-

Resolution National-Scale Land-Cover Mapping Based on Noisy Label Learning: A Case Study of China[J]. *International Journal of Applied Earth Observation and Geoinformation*, 2023, 118: 103265.

[176]ZHONG Y F, YAN B W, YI J J, et al. Global Urban High-Resolution Land-Use Mapping: From Benchmarks to Multi-Megacity Applications[J]. *Remote Sensing of Environment*, 2023, 298: 113758.

[177]WANG J J, ZHONG Y F, CHEN Z H, et al. EarthVL: A Progressive Earth Vision-Language Understanding and Generation Framework [PP/OL]. V1. arXiv (2026-01-06) [2026-10-10]. https://doi.org/10.48550/arXiv.2601.02783.

[178]LIU Y H, ZHONG Y F, SHI S N, et al. Scale-Aware Deep Reinforcement Learning for High Resolution Remote Sensing Imagery Classification[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2024, 209: 296-311.

[179]ZHENG Z, ZHONG Y F, WANG J J, et al. Building Damage Assessment for Rapid Disaster Response with a Deep Object-Based Semantic Change Detection Framework: From Natural Disasters to Man-Made Disasters[J]. *Remote Sensing of Environment*, 2021, 265: 112636.

[180]MUTTENTHALER L, GREFF K, BORN F, et al. Aligning Machine and Human Visual Representations Across Abstraction Levels[J]. *Nature*, 2025, 647(8089): 349-355.

[181]XU W J, YU Z J, MU B Y, et al. RS-Agent: Automating Remote Sensing Tasks Through Intelligent Agent [J]. *Science China Information Sciences*, 2026, 69(8): 180302.

[182]WANG S, HAN W, HUANG X H, et al. Trustworthy Remote Sensing Interpretation: Concepts, Technologies, and Applications[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2024, 209: 150-172.

[183]CHEN Z H, ASADI SHAMSABADI E, JIANG S, et al. Integration of Large Vision Language Models for Efficient Post-Disaster Damage Assessment and Reporting[J]. *Nature Communications*, 2026, 17: 1481.