测绘学报　　2024 年　第 53 卷　第 10 期：1001—1595
Acta Geodaetica et Cartographica Sinica　　xb.chinasmp.com

多模态遥感基础大模型：研究现状与未来展望

张永军[1]，李彦胜[1]，党　博[1]，武　康[1]，郭　昕[2]，王　剑[2]，陈景东[2]，杨　铭[2]

1. 武汉大学遥感信息工程学院，湖北 武汉 430079；
2. 蚂蚁集团，浙江 杭州 310013

收稿日期：2024-01-12；修回日期：2024-09-08　　中图分类号：P237　　文献标识码：A
文章编号：1001-1595(2024)10-1942-13
基金项目：国家自然科学基金(42030102；42371321)
第一作者简介：张永军(1975—)，男，博士，教授，研究方向为航空航天摄影测量与遥感影像智能解译。
E-mail：zhangyj@whu.edu.cn
通信作者：李彦胜　　E-mail：yansheng.li@whu.edu.cn

摘　要：遥感对地观测能力的稳步提升为遥感基础大模型的涌现和发展奠定了数据基础。针对不同数据及任务类型，设计不同的深度网络骨架及优化方法必将浪费大量人力物力。为了解决上述问题，国内外研究学者转入遥感基础大模型研究，并提出了大量优秀统一模型。为提高遥感基础大模型的泛化性和可解释性，引入泛在的地学知识被认为是一项关键技术。目前，已有相关工作在遥感基础大模型的结构设计或预训练方法中挖掘或整合了地学知识，但尚无文献系统性阐述和总结地学知识引导的遥感基础大模型的研究现状。因此，本文首先对大规模遥感基础模型预训练数据集进行了归纳和总结，并分类回顾了遥感基础大模型的研究进展；然后，介绍了地学知识引导的遥感影像智能解译算法以及面向遥感基础大模型的地学知识挖掘与利用进展；最后，针对该领域仍然面临的挑战提出了几点未来研究展望，旨在为遥感基础大模型的未来研究提供探索方向参考。

关键词：预训练数据集；遥感智能解译；遥感基础大模型；地学知识

在遥感大数据时代，爆炸式增长的遥感影像数据为地球观测信息提取及知识发现带来了新的挑战和机遇[1]。目前，深度学习等先进人工智能技术能够从海量的多模态、多尺度、多时相遥感数据中自动学习特征表达与判别模型，进而提高遥感解译任务的效率和准确性。虽然众多任务特定的智能遥感解译算法已经被提出并在特定应用场景上取得了一定的进展[2-5]，但是任务之间的差异和任务特定解译模型的有限泛化能力使得每项任务都需要投入大量资源构建任务特定，甚至是场景特定的解译模型，导致算法解译效率低下和泛化应用困难。

近期，随着各类自然语言大模型、视觉基础大模型、多模态基础大模型的涌现和发展[6-8]，基础大模型在各个领域的探索成为研究热点。鉴于任务特定遥感解译模型的适用局限，许多学者开始探索针对地球观测任务的遥感基础大模型构建与应用。遥感基础大模型旨在利用大量未标注的遥感数据进行预训练，创建一个任务通用模型，即从大规模遥感数据中学习通用特征表达模型。进一步，通过迁移学习提高多种下游遥感解译任务的性能和效率[9-11]。然而，在遥感对地观测这个具有高度复杂性的领域中，仅依赖深度网络非线性映射模型难以全面理解地球的复杂特征，地学知识的挖掘与运用显得愈加关键。地学知识不仅包括丰富的时空信息、地形地貌等测绘地理信息数

引文格式：张永军，李彦胜，党博，等. 多模态遥感基础大模型：研究现状与未来展望[J]. 测绘学报，2024，53(10)：1942-1954. DOI：10.11947/j.AGCS.2024.20240019.
ZHANG Yongjun，LI Yansheng，DANG Bo，et al. Multi-modal remote sensing large foundation models：current research status and future prospect[J]. Acta Geodaetica et Cartographica Sinica，2024，53(10)：1942-1954. DOI：10.11947/j.AGCS.2024.20240019.

据，还涵盖了场景先验知识（如开放街道地图等）及领域专家知识（如领域常识等）。

目前，已经有一些遥感基础大模型开始尝试引入地学知识。具体来说，早期工作尝试利用时空信息（如成像时间和地理坐标）进行预训练算法建模[12-14]。后来，研究学者将地学产品嵌入基础模型预训练过程，利用公开获取的土地覆盖分类产品提供的地学知识优化基础模型[15-16]。结合地学参量约束模型参数更新也被验证是有效的[17]。最近，笔者所在团队提出的 SkySense[18] 通过对地理位置特定的大规模多模态时序遥感影像进行无监督学习，可以隐式挖掘时空敏感的地学知识，辅助提升解译精度。总体来说，上述方法涵盖了多样化地学知识整合方式，为提高模型性能和可解释性提供了有效途径。随着地学知识引导的强化，遥感基础大模型有望能够更好地适应不同地域、不同地貌、不同尺度、不同模态的智能遥感解译需求。

本文首先系统总结了当前用于遥感基础大模型预训练的大规模数据集情况；其次，回顾了遥感视觉基础大模型、遥感视觉-语言基础大模型、遥感视觉-地理位置基础大模型等 4 个方向的研究进展；然后，分析了当前面向遥感基础大模型的地学知识挖掘与利用的研究现状；最后，给出了遥感基础大模型发展面临的挑战与未来研究的几点展望。

1　大规模预训练数据集

大规模预训练数据是基础大模型的数据引擎。研究表明，在广泛而多样化的数据上进行预训练对于模型学习判别性通用特征表示具有显著促进作用[19-21]，有助于加速预训练模型在各种下游任务的微调收敛过程，减少对有标签数据的依赖，进而提升任务性能。这种任务通用的特征表示为模型在理解和处理不同场景数据时提供了坚实的基础，使其具备强大的泛化能力。在遥感领域，已有一系列相关研究致力于构建大规模预训练遥感数据集。根据数据模态的不同，接下来对大规模预训练数据集进行了归纳和总结。

1.1　遥感视觉预训练数据集

如表 1 所示，目前已经涌现出大量各具特色的遥感视觉预训练数据集。在这些数据集中，MillionAID[22] 和 SatlasPretrain[23] 包含了超高分辨率卫星影像，但仅涵盖可见光波段。通过这些数据集训练的遥感基础模型可能在依赖丰富光谱信息的任务（如农作物识别）等方面存在一定的缺陷。然而，超高分辨率影像所包含的细节纹理信息使得预训练模型在基于高分影像的实例分割、目标检测等下游任务上具有一定优势。相比之下，fMoW[24]、SeCo[12] 等数据集利用哨兵 2 号获得的中分辨率多光谱影像作为数据源。众所周知，遥感观测数据包括多种模态影像类型，这些数据具有独特的优势和相互补充的特性。如，光学图像提供了丰富的光谱信息和纹理细节，但容易受到天气及云层的影响。合成孔径雷达传感器能够在恶劣的天气条件下成像。为了满足更多需要依赖多种模态信息的下游任务，BigEarthNet-MM[25] 和 SSL4EO-S12[26] 数据集致力于构建成对的合成孔径雷达-多光谱影像数据集。这类数据集旨在提供更全面、多样化的信息，以支持多模态遥感基础大模型的训练和性能提升，有望促进多模态遥感技术的进步，使其在实际应用中更为灵活和有效。

表 1　大规模遥感视觉预训练数据集

Tab. 1　Large-scale remote sensing vision pre-training datasets

数据集	图像数量	图像大小/像素	空间分辨率/m	图像类型	图像数据源	覆盖地理位置
fMoW[24]	1 047 691	—	—	多光谱（4/8 波段）	Digital Globe	全球
SEN12MS[27]	180 662	256	10	合成孔径雷达-多光谱	哨兵 1 号、哨兵 2 号	全球
BigEarthNet-MM[25]	1 180 652	20～120	10～60	合成孔径雷达-多光谱	哨兵 1 号、哨兵 2 号	欧洲
MillionAID[22]	1 000 848	110～31 672	0.5～153	可见光	Google Earth	—
SeCo[12]	1 000 000	—	10	多光谱	哨兵 2 号	全球
fMoW-Sentinel[28]	882 779	45～60	10	多光谱（13 波段）	哨兵 2 号	全球
TOV-RS-Balanced[20]	500 000	600	1～20	可见光	Google Earth	—
SSL4EO-S12[26]	3 012 948	20～120	10～60	合成孔径雷达-多光谱	哨兵 1 号、哨兵 2 号	全球
SSL4EO-L[29]	5 000 000	264	30	多光谱	Landsat4-5，7-9	全球
SatlasPretrain[23]	856 000	512	0.5～2，10	可见光 & 多光谱	NAIP、哨兵 2 号	全球

1.2　遥感视觉-语言预训练数据集

目前，能够用于训练遥感视觉-语言基础大模型的数据集较少，其数据规模相对有限。如表 2 所示，多数预训练数据集集中于提供图像-文本描述。

表 2 大规模遥感视觉-语言预训练数据集

Tab. 2 Large-scale remote sensing vision-language pre-training datasets

数据集	数量	属性
RSICD[30]	24 333 个文本描述、10 921 张遥感影像	图像-文本描述
RSITMD[31]	23 715 个文本描述、4743 张遥感影像	图像-文本描述
RSVGD[32]	38 320 个语言表达、17 402 张遥感影像	视觉定位
RS5M[33]	500 万个图像文本对	图像-文本描述
RSICap[34]	2585 个图像文本对	图像-文本描述
文献[35]	828 725 个图像文本对	图像-文本描述
文献[36]	318 000 个图像指令提示对	图像-文本描述、定位描述、区域描述、复杂对话

具体来说，早期的遥感图像-文本描述数据集多为特定任务构建[30-32]，其中的文本描述较为简短，包含的有限语义信息不足以训练泛化性强的基础模型。RSICap[34]致力于创建高质量图像-文本描述信息，其中，每幅遥感影像带有场景、目标形状、目标绝对位置、相对位置、颜色和数量等细节信息的描述。文献[35]设计了“掩码转定位框”“定位框转文本描述”的转换流程，将遥感领域常用的 3 个图像检索数据集、10 个目标检测数据集、4 个语义分割数据集转换为图像-文本描述数据对，有效提升了遥感视觉-语言基础大模型的预训练数据多样性。相似地，文献[36]整合了一些遥感视觉问答、目标检测数据集，将其重构成图像-文本描述、定位描述和复杂对话等形式，以满足多功能对话智能体训练的需求。

2 遥感基础大模型

本文将遥感基础大模型归纳分为 4 类：遥感视觉基础大模型、遥感视觉-语言基础大模型、遥感视觉-地理位置基础大模型、遥感生成式基础大模型。图 1 展示了每种类型的遥感基础大模型所适应的典型下游任务。后续，本节将逐个类别回顾相关研究的前沿进展。

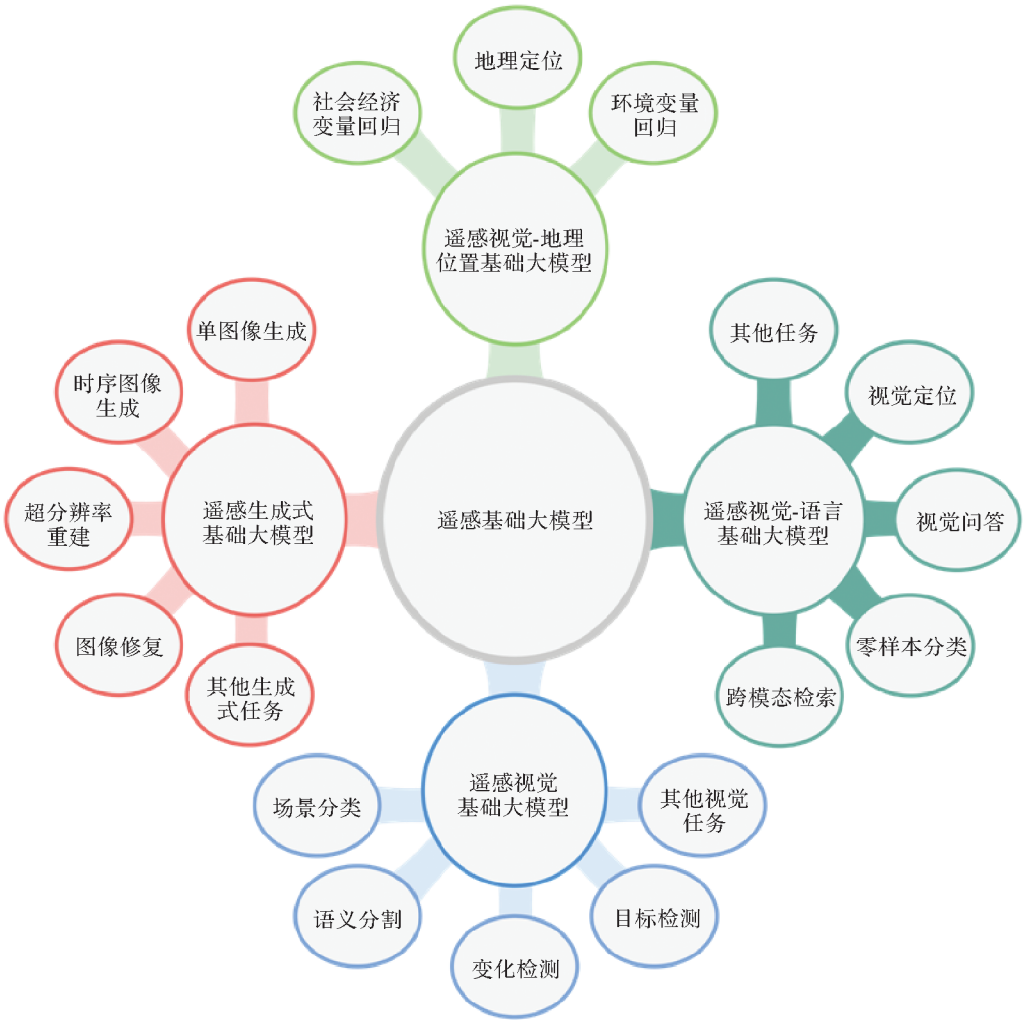


图 1 遥感基础大模型分类及典型适用的下游任务

Fig. 1 Classification of remote sensing foundation models and typical downstream tasks

2.1　遥感视觉基础大模型

在计算机视觉领域，视觉基础模型的研究重点已经从早期的利用大量标记数据的监督学习[37-38]（如在 ImageNet 数据集上进行预训练）发展到最近的对比学习范式[39-41]（在大规模未标记图像上开展无监督预训练）。随着自然语言处理领域中大语言模型的巨大成功[42]，掩码图像建模方法（如 MAE[43]、BEiT[44]等）受到广泛关注。研究指出[45]，基于对比学习的模型关注全局结构和形状等低频空间信息，而基于掩码图像建模的模型则更加侧重于挖掘高频空间信息（如局部结构和精细的纹理）。ibot、DINOv2[46-47]成功地结合了上述两种范式的优势，取得了先进的性能表现。

相较于自然图像，遥感影像往往附带时空地理元信息，并呈现出不同的空间尺度。遥感领域专家学者利用遥感数据的时空基准信息改造基础模型，将其扩展应对遥感数据分析。如，GASSL[48]利用地理位置预测作为 MoCo-v2 框架中的额外代理任务。SeCo[12]和 CACo[13]通过使用时间序列的时空结构来感知影像中地物的短期和长期变化。文献[20]使用自然图像和遥感图像作为初步和后续的预训练数据，构建正、负样本对进行对比学习，试验结果表明预训练数据的类别平衡性对于预训练模型学习有效通用表征是十分关键的。MATTER[49]对照明和视角不变性进行建模，以确保纹理在不变区域上的一致表示。DINO-MC[50]则利用不同大小的多个视图在 DINO 框架[51]内进行自监督学习。

此外，许多研究致力于改进基于掩码图像建模的框架，或者探索模型规模扩展[52]以及模型轻量化部署[53]。在可见光遥感影像为预训练数据的背景下，文献[54]提出了旋转可变大小窗口注意力方法处理遥感图像中大尺寸和任意方向的地物，并利用 MillionAID 设计了遥感亿级参数量的视觉大模型。RingMo[55]对 MAE 进行修改，更好地应对遥感影像密集目标检测任务。Scale-MAE[56]构建了一个带有尺度感知位置编码和拉普拉斯金字塔解码器的框架，实现了多尺度解码低频和高频特征。对于拥有更加丰富光谱信息的多光谱遥感影像数据，SpectralGPT[57]将多光谱图像作为 3D 张量数据进行掩码图像建模，提出多目标重建损失，有效捕捉空间光谱耦合特征和光谱顺序信息。考虑到卫星传感器能够以非规则和一定频率获取某一地点的时序多光谱影像，Prithiv[58]将常规的 2D 位置编码适应性改造为 3D 版本，由于其具有处理遥感时序数据的能力，该模型被成功应用于洪水检测、多时相农作物分割等场景。相似地，SatMAE[28]则利用时序多光谱数据来提高和验证基础模型处理时间序列的表现。为解决多光谱影像引起显存占用大的问题，现有遥感基础模型无法应对任意波段数据输入的缺陷，USat[59]首先对光学遥感影像的每个波段独立编码，然后使用光谱组池化操作聚合不同光谱波段的信息，同时保留不同空间分辨率的图像地理位置对齐位置编码。文献[17]借鉴掩码图像建模思想，提出特征引导的掩码自编码器，分别利用多光谱和合成孔径雷达影像重建人工特征描述符（如归一化指数、方向梯度直方图），结果表明相较于直接重建图像通过重建抽象特征可以获得更好的特征学习能力。

近期，CMID[21]、GFM[60]、Cross-Scale MAE[61]等研究将对比学习范式与掩码图像重建范式相结合，在场景分类、目标检测、语义分割、变化检测等众多图像级、对象级、像素级的典型遥感解译任务中展现出明显性能优势。类似地，CtxMIM[62]则在重建掩码图像损失的基础上增加上下文一致性约束，以提供额外的上下文信息。与大多数基础模型采用自监督预训练方法不同，SatLas[23]依托自建的具有丰富标注类型的大规模数据集 SatlasPretrain 进行有监督预训练，并将模型应用于热带雨林砍伐检测、可再生能源基础设施检测等任务。文献[63]面向遥感时空预测任务设计了包含空间、时间、时空建模 3 个分支的基础模型，并在雷达回波外推、卫星视频多目标跟踪和遥感视频预测等下游任务中取得了具有竞争力的结果。

除了仅依靠单模态图像预训练的工作外，CROMA[64]和 De-CUR[65]研究了使用静态影像进行单模态和多模态图像源的多模态预训练。Presto[66]同时利用时间和地理位置信息，联合多光谱、合成孔径雷达、高程等多模态信息训练了轻量级基础模型。遗憾的是，Presto 的预训练数据未包含高分辨率卫星图像，且缺乏在基于高分辨率影像的下游任务上广泛的测试以验证模型的泛化性。文献[67]则关注到跨模态协同解译中异构模态特征的空间相关性问题，采用不同的度量空间（即欧氏空间、复数空间和双曲空间）提取不同模态图像的特征，然后采用统一的编码器进行多模态特征融合。笔者所在团队则发展了目前参数量规模最大的多模态时序遥感基础大模型——SkySense[18]（20 亿参数量），通过时空解耦、时间感知嵌入等机制联合高分光学遥感影像、时序光学遥感影像、时序合成孔径雷达影像等多模态数据进行多粒度对

比学习。值得说明的是,灵活可插拔性和通用特征的强大泛化性使得 SkySense 在涵盖单模态图像级分类、目标级检测、像素级分割以及多模态农作物时序分类等 8 项任务(共计 16 个数据集)中均取得了最先进的水平。

2.2 遥感视觉-语言基础大模型

在自然语言处理领域,大型语言基础模型在自然语言理解、文本生成、智能问答等任务中取得了显著的成效[68]。特别是 ChatGPT 取得的巨大成功进一步推动了相关研究的发展。视觉-语言基础模型则集成了图像的视觉感知信息和语言的语义信息,旨在从视觉与语言的相互关系中学习通用特征,以更好地完成复杂场景的理解任务[11]。

在遥感领域,已有学者开始视觉-语言基础大模型相关研究工作。文献[69]专注于探索前沿的基础大模型(如 GPT-4V 等)在地理空间领域相关任务上的表现,为后续的研究提供基准参考。文献[34]利用构建的 RSICap 数据集微调了 InstructionBLIP 模型得到 RSGPT 模型,并在图像描述生成、视觉问答任务中显示出具有潜力的效果。RemoteCLIP[35]则采用对比语言-图像预训练(CLIP)方法在创建的视觉-语言数据集上进行了训练,获得的预训练模型在跨模态检索、零/少样本图像分类、目标计数等下游任务中进行了评估。GeoChat[36]致力于构建一个允许用户对给定的遥感影像视觉内容进行对话的多功能视觉-语言基础模型,能够完成图像级、区域级(指定图像中的特定区域)、定位式的对话任务。遗憾的是,目前 GeoChat 仅支持高分辨率的可见光影像,局限了其在众多下游场景的普适性。由于基于卫星影像的图像文本标注过程需要专家知识的干预,成本消耗巨大,目前已有的图像-文本描述数据相较于计算机视觉领域规模小很多。最近,GRAFT[70]考虑利用大规模带有地理位置信息的互联网数据作为数据中介,通过训练对齐相同地理位置的卫星影像和互联网图像的视觉特征,从卫星影像中抽取的视觉特征、互联网图像对应的视觉特征与已经训练好的文本语义特征共享至同一特征空间,从而在不需要文本标注的条件下实现影像编码与文本编码的关联。这大大降低了遥感视觉-语言模型训练的数据标注成本,为该方向提供了一个思路。此外,笔者所在团队创建了一个大规模遥感场景图数据集 STAR[71],并在此基础上延伸拓展出细粒度视觉-语言指令微调数据集 FIT-RS 及相应的视觉-语言基础模型 SkySenseGPT[72]。SkySenseGPT 具有对实例间关系的细粒度感知能力,能够基于用户指令完成复杂的图文交互任务。

2.3 遥感视觉-地理位置基础大模型

区别于遥感视觉基础大模型以遥感影像为中心,遥感视觉-地理位置基础模型则以地理位置为核心,旨在从卫星影像中学习出对应于特定地理位置相关的通用特征表示。考虑到大量遥感数据包含了对应的地理位置信息,预训练后的位置编码器能够广泛应用于自然环境和社会经济等任务,如生物群落分类、人口密度回归等与地理位置相关的任务。

在计算机视觉领域中,一些学者采用了配对的自然图像和 GPS 数据训练位置编码器,以解决全球图像地理定位的挑战。如,GeoCLIP[73]设计了位置编码器,将 GPS 坐标映射为高维特征嵌入,并使用经过预训练的 CLIP 模型[6]作为图像编码器提取图像特征。随后,该研究将位置特征与图像特征映射到共享嵌入空间进行对比学习。不同地理位置的遥感影像的视觉特征受到与地理位置相关的气候、人口密度等自然环境和社会因素的密切影响。在这一背景下,CSP[74]采用多种方式构造正负样本对,并通过遥感数据集预训练后的图像编码器与提出的位置编码器进行对比学习。SatCLIP[14]则致力于捕捉全球不同地区的哨兵 2 号卫星影像的空间异质性,通过对比预训练的方式学习位置编码特征表示。相关试验证明,SatCLIP 模型的位置编码器成功学习到了与特定区域的社会经济与环境等因素高度相关的特征表示。上述技术为进一步深入分析地理位置与遥感影像之间的关联提供了有力支持。

2.4 遥感生成式基础大模型

遥感影像超分辨率重建、云去除等生成式解译方法能够帮助人类更完整、更细致地观察地表自然环境和人类活动的变化,吸引了众多学者的关注[4,75]。然而,先前的研究主要集中在为特定生成任务设计专用模型上,导致在实际应用中灵活性和通用性相对不足。稳定扩散模型(stable diffusion)在图像重建、视频生成等任务上取得显著进展,这使得诸多学者将其应用于多种遥感图像生成式任务,并取得了一定的进展。文献[76]采用文本描述、遥感影像以及附带的地理元信息(包括地理坐标、成像时间、空间分辨率等)训练了遥感生成式基础模型 DiffusionSat。该模型在单个遥感图像生成、多光谱图像超分辨率重建、时序图像生成和图像修复等多个下游任务上取得了先进的性能表现。文

献[77]则采用预训练扩散模型学习公开地图数据，可以生成视觉效果逼真、地物类别可控的合成卫星图像。该技术可以为数据缺失任务场景补充额外样本数据。尽管目前遥感生成式基础大模型仍处于初步发展阶段，研究成果相对较少，但其应用潜力巨大，预计将吸引更多学者深入研究。未来，我们可以期待这一领域的快速发展，为遥感生成式解译提供更为灵活、通用且性能卓越的模型。

3　地学知识引导的遥感基础大模型

地学知识主要包括地表人类活动与自然演变呈现的规律性时空先验信息和领域专家知识[78]。基于深度学习的智能遥感解译模型往往以数据驱动为主，解译模型的泛化性较低，同时缺乏足够的可解释性。为了弥补这一不足，引入地学知识成为提升解译模型性能的有效手段。本节首先回顾了地学知识引导的智能遥感解译技术，然后着重探讨了地学知识在提高智能遥感解译模型性能和可解释性等方面的潜在作用，最后对目前遥感基础大模型挖掘和利用地学知识的方法进行了分类阐述，旨在为未来相关研究提供参考和启示。

3.1　地学知识引导的遥感智能解译方法

近年来，面向遥感影像智能解译的地学知识引导技术受到国内外研究学者的广泛关注。在这一方向，笔者所在团队取得了若干研究进展[79-85]。

（1）利用自然语言嵌入模型或知识图谱表征模型引导的零样本遥感影像场景分类。如，文献[79]创建了遥感知识图谱 SR-RSKG 并开展知识图谱语义表征学习，进一步提出一种深度对齐网络在隐式空间中稳健地匹配视觉特征和语义特征，从而实现零样本遥感图像场景分类。SR-RSKG 包含丰富的显式关系信息（即“实体-关系-实体”或“实体-属性-属性值”），有助于更准确地描述复杂遥感场景。

（2）耦合知识图谱和深度网络的光学遥感影像语义分割。鉴于数据驱动的深度学习技术在可解释性方面存在不足，文献[82]借助遥感知识图谱的丰富语义关系建模与强大推理能力，引入高层次专家知识修正深度网络输出结果，并将知识推理输出用于进一步辅助深度学习模型的训练。此外，地物空间共生知识[85]也被用于提升遥感影像语义分割精度。

（3）经验知识引导的多模态遥感影像土地覆盖分类。通过融合光学、合成孔径雷达和高程等多模态信息，文献[81]提出了遥感指数等领域知识引导的深度协作融合网络（DKDFN）。该网络通过多头编码器协作融合多模态数据，利用多分支解码器创建多任务学习策略重建地学知识，显著提高了在土地覆盖分类任务上的精度和稳健性。

（4）多模态知识图谱推理驱动的合成孔径雷达影像溢油监测。文献[83]通过整合遥感影像、矢量、文本信息和大气-海洋模型信息等构建了海洋溢油监测知识图谱，结合规则推理和图神经网络方法可以在数据类别极不平衡的条件下得到优异的海洋溢油监测结果。通过构建多模态知识图谱，可以将与溢油监测相关的先验知识有效地组织在一起，从而克服传统方法存在的信息孤岛问题。在知识推理后，所有推理结果可以集成到知识图谱中，使知识图谱能够不断迭代演进，进而实现高精度溢油检测。

从上述的代表性地学知识引导的遥感影像解译算法可以看出，耦合地学知识的方式是多种多样的。由于结构化知识图谱具备可计算、可推理、可进化等优势，耦合地学知识图谱和深度学习有望成为新一代遥感智能解译范式[84]，为地学知识引导的遥感基础大模型研究提供有益的参考。

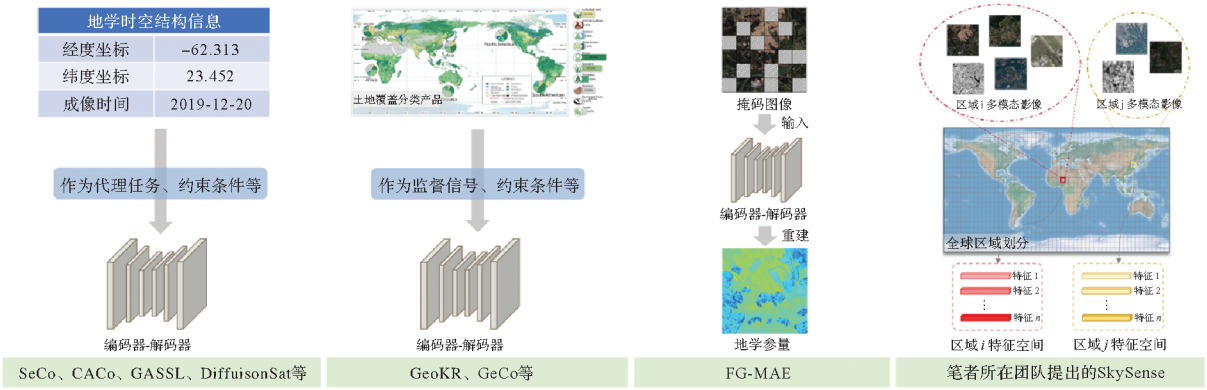
3.2　面向遥感基础大模型的地学知识挖掘与利用

目前，一些遥感基础大模型的预训练或推理已经开始探索地学知识的挖掘与利用。总体来说，遥感基础大模型的地学知识挖掘与利用方法可以大致分为以下 4 种类型（图 2）。

（1）时空结构信息挖掘与利用。遥感影像附带成像时间、经纬度坐标等元信息，这些地学时空信息能够有效改善遥感基础模型预训练性能。如，拍摄自同一地点但不同成像时间的遥感影像可用于对比预训练[12-13]；地理坐标编码可作为预训练的代理任务[48]；地理坐标、成像时间等时空信息可用作预训练约束条件[76]；结合视觉信息学习的地理位置编码器[14]可进行特定区域的变量回归等任务。

（2）土地覆盖分类产品嵌入学习。土地覆盖分类产品（如 GlobeLand30[86]、FROM_GLC10[87] 等）蕴含着丰富的地学先验知识。这些地学先验信息的嵌入建模正成为遥感基础大模型研究热点。GeoKR[15]通过对齐视觉特征与公开地学产品提取出的知识特征促进骨干网络学习，以缓解遥感影像和地理知识之间的时间与空间分辨率差异的影响。GeCo[16]根据地学产品中“时序变化小”“空间聚合性高”的先验信息定义可学习的纠正矩阵，以学习地学产品中的类别分布特点。此外，利用地学先验信息干预参与预训练的遥感数据的类别平衡，能够在一定程度上改善基础模型学习到的通用特征的

有效性[20]。结合地学先验知识和生成式基础模型，文献[77]将开放街道图(OSM)提供的道路、建筑物等地物目标信息作为输入条件，基于ControlNet[88]生成内容可控的遥感合成影像，有望应用于众多下游任务的有监督数据扩展。



(a) 地学时空结构信息挖掘与利用 (b) 地学产品先验信息嵌入 (c) 地学参量约束 (d) 隐式地学知识挖掘与融合

图 2 面向遥感基础大模型的地学知识挖掘与利用的 4 种方式

Fig. 2 Four ways of mining and utilizing geoscience knowledge for remote sensing foundation model

(3) 地学参量约束。定量遥感旨在将多源遥感观测数据定量反演或推算为地学目标参量，形成时空遥感数据产品[89]。相关地学参量(如归一化指数等)通过物理机理、成像光谱信息反映地表的属性信息，FG-MAE[17]结合经典的掩码图像建模算法重建相关地学参量，从而约束大模型参数更新。

(4) 隐式地学知识挖掘与融合。地理景观的形成是气候、地质、水文、生物多样性和人类活动等多种因素的错综复杂相互作用[90]。这些因素共同促使地理区域呈现出特定的地理特征，即不同地区的遥感影像往往呈现出明显的地理异质性。笔者所在团队提出的SkySense[18]发展了地理空间敏感的上下文学习范式，旨在从遥感大数据中隐式挖掘与融合地学知识。具体而言，将全球划分为众多子区域，通过对地理位置特定的大规模多模态时序遥感影像进行无监督学习，以隐式挖掘时空敏感的聚类特征，这些聚类特征一定程度上可以较好地反映不同区域的语义先验。在推理阶段，可以通过注意力机制融合视觉特征和语义先验来改善遥感影像的解译性能。

4 面临挑战与未来展望

如前文所述，目前各个方向的遥感基础大模型均取得了一定的进展和突破，但仍然面临着诸多挑战。本节从预训练数据集、评估基准、基础模型架构、地学知识的嵌入和挖掘及大规模应用等方面对遥感基础大模型面临的挑战进行梳理，并提出了几点展望。

4.1 面临挑战

(1) 多模态预训练数据稀缺与评估基准不足。在自然语言处理和计算机视觉等领域，大量卓越基础模型的成功案例均揭示预训练数据集的规模和质量是影响模型泛化性的重要因素[47]。尽管遥感领域逐渐涌现出规模较大的预训练数据集(详见第1节)，但仍然缺乏不同卫星源、不同波段组合、不同空间分辨率、不同成像模式的多模态预训练数据集，无法支撑多模态遥感基础大模型的充分训练。此外，全面、统一、可靠的评估基准能够帮助全面衡量遥感基础模型的能力。早期遥感基础模型评估所选用的数据集、下游任务、评估方式各不相同，未形成系统全面的评估数据集及指标体系。笔者所在团队提出的SkySense[18]在单模态图像级分类、目标级识别、像素级分割、多模态时序分类等众多数据集上建立了统一的评估基准结果，以方便后续方法进行对比。未来还应该不断补充更多任务类型。此外，在弱监督下游任务的条件下评估预训练模型的泛化性更加符合实际应用场景的需求，如评估在少样本、含大量噪声标签等下游数据条件下遥感基础大模型的稳健性。

(2) 缺少灵活支持多模态、多时序输入的统一预训练框架。在遥感领域，遥感影像数据往往呈现出不同分辨率、光谱信息、成像模式、时间序列长度等特性。每种模态数据的成像机理和物理性质各不相同，时序影像包含的时序信息有助于改善时间

敏感的下游任务性能。多模态时序数据联合解译有利于获得更加全面、准确的特征表达。尽管目前一些遥感基础模型开始探索多模态、多时序数据联合的预训练，但仍然缺少能够灵活支持波段任意的影像、文本甚至是音频的统一预训练框架。

（3）缺乏地学知识挖掘与嵌入。在提升遥感基础模型的可解释性和稳定性方面，引入地学知识被认为是一项关键的改进手段。尽管已有一些遥感基础模型试图从多个角度隐式或显式地整合地学知识，但对地学知识的应用和挖掘有待进一步加强。

4.2　未来展望

（1）多模态预训练数据的丰富与评估标准的完善。为了增强遥感基础大模型在不同数据源和任务上的泛化性能，未来的研究需要创建大规模、多样化的多模态遥感预训练数据集。此外，在现有评估基准的基础上增加更多遥感定量反演、时序预测与生成任务（如定量遥感分析、地物要素矢量生成、遥感影像时间序列修复等）的评估对比。

（2）模态任意、波段任意、时序任意的遥感基础模型框架设计。考虑到遥感影像数据的多源、多模态等特性，亟须发展一种支持多样性输入的可插拔模型框架，以此实现灵活支持任意模态、任意光谱波段/极化方式、任意时序长度输入，满足不同遥感任务的需求。通过这一可扩展的框架设计，模型能够更好地适应不同应用场景下的遥感数据特性，提高模型的通用性和适用性。

（3）高效低成本下游微调算法研究。在大模型和海量遥感数据的背景下，下游任务全参数微调需要消耗大量的时间和计算成本，因此，亟须发展参数量可控的高效下游微调方法，以达到甚至超过全参数量微调的效果。此外，遥感影像标注成本较高，下游具体应用场景中可获取的有标签样本有限，因此需要开发低标注样本量的下游任务微调算法，更好地服务于实际应用场景。

（4）地学知识图谱构建与引导。在具体的遥感智能解译任务中引入地学知识以提升深度网络的性能和可解释性已经受到许多学者关注。地学知识图谱的构建和利用也被认为是未来遥感解译的发展趋势之一[84,91]。通过将源自文本语料库、时空信息、地形地貌、场景先验与专家知识等的地学知识整合，以知识图谱的统一形式进行重构并融入遥感基础模型的训练和推理过程是提升基础模型的性能和可解释性的重要方向之一。地学知识图谱的构建与融入不仅可以提升遥感基础大模型的实际性能表现，还有望为遥感下游应用提供更为全面和深度的结果溯源解释。

（5）全球尺度大规模复杂场景应用。目前许多遥感基础模型已经在大量的下游任务数据集上评估了效果，但基础模型在大规模复杂场景应用上的适应性还需更多探索研究。遥感基础模型的通用表征能力使其在全球尺度大规模复杂场景制图应用等方面表现出较大潜力。由于地表呈现出不同景观格局，需要验证和优化基础大模型在应对复杂场景的高效性和稳定性。因此，对遥感基础模型在大规模复杂场景应用中的适用性和性能进行深入研究，将有助于填补当前研究的空白，有助于解决人道主义救援、农业监测和粮食安全评估、可持续发展评估等全球性问题。

5　结　语

遥感基础大模型为遥感影像智能解译带来了新的机遇。通过充分整合地学知识，可以有效辅助遥感基础大模型感知地表的复杂时空特征与语义信息。本文首先回顾了大规模遥感预训练数据集；其次，讨论了遥感视觉基础大模型、遥感视觉-语言基础大模型、遥感视觉-地理位置基础大模型和遥感生成式基础大模型；然后，总结了地学知识引导的遥感基础大模型的研究现状；最后，分析了目前研究面临的挑战，并围绕数据、算法、知识建模与引导等方面作出了几点未来的研究展望供学者们参考。

参考文献

[1]　付琨，卢宛萱，刘小煜，等. 遥感基础模型发展综述与未来设想[J]. 遥感学报，2024，28(7)：1667-1680.
FU Kun，LU Wanxuan，LIU Xiaoyu，et al. A comprehensive survey and assumption of remote sensing foundation modal[J]. National Remote Sensing Bulletin，2024，28(7)：1667-1680.

[2]　LI Yansheng，CHEN Wei，HUANG Xin，et al. MFVNet：a deep adaptive fusion network with multiple field-of-views for remote sensing image semantic segmentation[J]. Science China Information Sciences，2023，66(4)：140305.

[3]　LI Yansheng，DANG Bo，ZHANG Yongjun，et al. Water body classification from high-resolution optical remote sensing imagery：achievements and perspectives[J]. ISPRS Journal of Photogrammetry and Remote Sensing，2022，187：306-327.

[4] LI Yansheng, WEI Fanyi, ZHANG Yongjun, et al. HS2P: Hierarchical spectral and structure-preserving fusion network for multimodal remote sensing image cloud and shadow removal[J]. Information Fusion, 2023, 94: 215-228.

[5] PENG Daifeng, ZHAI Chenchen, ZHANG Yongjun, et al. High-resolution optical remote sensing image change detection based on dense connection and attention feature fusion network[J]. The Photogrammetric Record, 2023, 38(184): 498-519.

[6] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[EB/OL]. [2024-01-05]. https://arxiv.org/pdf/2103.00020.

[7] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[EB/OL]. [2024-01-05]. https://arxiv.org/pdf/2304.02643.

[8] YANG Zhengyuan, LI Linjie, LIN K, et al. The dawn of LMMs: preliminary explorations with GPT-4V (ision)[EB/OL]. [2024-01-05]. https://arxiv.org/pdf/2309.17421.

[9] 张良培, 张乐飞, 袁强强. 遥感大模型: 进展与前瞻[J]. 武汉大学学报(信息科学版), 2023,48(10): 1574-1581.
ZHANG Liangpei, ZHANG Lefei, YUAN Qiangqiang. Large remote sensing model: progress and prospects[J]. Geomatics and Information Science of Wuhan University, 2023, 48(10): 1574-1581.

[10] JIAO Licheng, HUANG Zhongjian, LU Xiaoqiang, et al. Brain-inspired remote sensing foundation models and open problems: a comprehensive survey[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16: 10084-10120.

[11] LI Xiang, WEN Congcong, HU Yuan, et al. Vision-language models in remote sensing: current progress and future trends[EB/OL]. [EB/OL]. [2024-01-05]. https://arxiv.org/abs/2305.05726.

[12] MANAS O, LACOSTE A, GIRO-I-NIETO X, et al. Seasonal contrast: unsupervised pre-training from uncurated remote sensing data [C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 9414-9423.

[13] MALL U, HARIHARAN B, BALA K. Change-aware sampling and contrastive learning for satellite images[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 5261-5270.

[14] KLEMMER K, ROLF E, ROBINSON C, et al. SatCLIP: global, general-purpose location embeddings with satellite imagery[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2311.17179.

[15] LI Wenyuan, CHEN Keyan, CHEN Hao, et al. Geographical knowledge-driven representation learning for remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5405516.

[16] LI Wenyuan, CHEN Keyan, SHI Zhenwei. Geographical supervision correction for remote sensing representation learning[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5411520.

[17] WANG Yi, HERNÁNDEZ H H, ALBRECHT C, et al. Feature guided masked autoencoder for self-supervised learning in remote sensing [EB/OL]. [2024-01-05]. https://arxiv.org/pdf/2310.18653.

[18] GUO Xin, LAO Jiangwei, DANG Bo, et al. SkySense: a multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 27672-27683.

[19] TAO Chao, QI Ji, GUO Mingning, et al. Self-supervised remote sensing feature learning: learning paradigms, challenges, and future works[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5610426.

[20] TAO Chao, QI Ji, ZHANG Guo, et al. TOV: The original vision model for optical remote sensing image understanding via self-supervised learning[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16: 4916-4930.

[21] MUHTAR D, ZHANG Xueliang, XIAO Pengfeng, et al. CMID: a unified self-supervised learning framework for remote sensing image understanding[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5607817.

[22] LONG Yang, XIA Guisong, LI Shengyang, et al. On creating benchmark dataset for aerial image interpretation: reviews, guidances, and million-AID[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 4205-4230.

[23] BASTANI F, WOLTERS P, GUPTA R, et al. SatlasPretrain: a large-scale dataset for remote sensing image understanding[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2211.15660v3.

[24] CHRISTIE G, FENDLEY N, WILSON J, et al. Functional map of the world[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 6172-6180.

[25] SUMBUL G, DE WALL A, KREUZIGER T, et al. BigEarthNet-MM: a large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval[software and data sets][J]. IEEE Geoscience and Remote Sensing Magazine, 2021, 9(3): 174-180.

[26] WANG Yi, ALI BRAHAM N A, XIONG Zhitong, et al. SSL4EO-S12: a large-scale multimodal, multitemporal dataset for self-supervised learning in Earth observation[software and data sets][J]. IEEE Geoscience and Remote Sensing Magazine, 2023, 11(3): 98-106.

[27] SCHMITT M, HUGHES L H, QIU C, et al. SEN12MS—a curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion[J]. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2019, 4: 153-160.

[28] CONG Yezhen, KHANNA S, MENG Chenlin, et al. SatMAE: pre-training transformers for temporal and multi-spectral satellite imagery[J]. Advances in Neural Information Processing Systems, 2022, 35: 197-211.

[29] STEWART A J, LEHMANN N, CORLEY I A, et al. SSL4EO-L: datasets and foundation models for landsat imagery[EB/OL].

[2024-01-05]. https://arxiv.org/abs/2306.09424.

[30] LU Xiaoqiang, WANG Binqiang, ZHENG Xiangtao, 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.

[31] YUAN Zhiqiang, ZHANG Wenkai, FU Kun, 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.

[32] ZHAN Yang, XIONG Zhitong, YUAN Yuan. RSVG: exploring data and models for visual grounding on remote sensing data[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5604513.

[33] ZHANG Zilun, ZHAO Tiancheng, GUO Yulong, et al. RS5M: a large scale vision-language dataset for remote sensing vision-language foundation model[EB/OL]. [2024-01-05]. http://export.arxiv.org/abs/2306.11300v4.

[34] HU Yuan, YUAN Jianlong, WEN Congcong, et al. RSGPT: a remote sensing vision language model and benchmark[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2307.15266v1.

[35] LIU Fan, CHEN Delong, GUAN Z, et al. RemoteCLIP: a vision language foundation model for remote sensing[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2306.11029v4.

[36] KUCKREJA K, DANISH M S, NASEER M, et al. GeoChat: grounded large vision-language model for remote sensing[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2311.15826v1.

[37] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770-778.

[38] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 6000-6010.

[39] CARON M, TOUVRON H, MISRA I, et al. Emerging properties in self-supervised vision transformers[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 9650-9660.

[40] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2002.05709v2.

[41] HE Kaiming, FAN Haoqi, WU Yuxin, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 9729-9738.

[42] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: open and efficient foundation language models[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2302.13971v1.

[43] HE Kaiming, CHEN Xinlei, XIE Saining, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 16000-16009.

[44] BAO H, DONG L, PIAO S, et al. BEiT: BERT pre-training of image transformers[EB/OL]. [2024-01-05]. https://arxiv.org/pdf/2106.08254.

[45] PARK N, KIM W, HEO B, et al. What do self-supervised vision transformers learn? [EB/OL]. [2024-01-05]. https://arxiv.org/pdf/2305.00729.

[46] ZHOU Jinghao, WEI Chen, WANG Huiyu, et al. iBOT: image BERT pre-training with online tokenizer[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2111.07832v3.

[47] OQUAB M, DARCET T, MOUTAKANNI T, et al. DINOv2: learning robust visual features without supervision[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2304.07193v2.

[48] AYUSH K, UZKENT B, MENG Chenlin, et al. Geography-aware self-supervised learning[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 10181-10190.

[49] AKIVA P, PURRI M, LEOTTA M. Self-supervised material and texture representation learning for remote sensing tasks[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 8203-8215.

[50] WANYAN Xinye, SENEVIRATNE S, SHEN Shuchang, et al. DINO-MC: self-supervised contrastive learning for remote sensing imagery with multi-sized local crops[EB/OL]. [2024-01-05]. https://ar5iv.labs.arxiv.org/html/2303.06670.

[51] ZHANG Hao, LI Feng, LIU Shilong, et al. DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2203.03605v4.

[52] CHA K, SEO J, LEE T. A billion-scale foundation model for remote sensing images[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2304.05215v4.

[53] WANG Yuelei, ZHANG Ting, ZHAO Liangjin, et al. RingMo-lite: a remote sensing multi-task lightweight network with CNN-transformer hybrid framework[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2309.09003v1.

[54] WANG Di, ZHANG Qiming, XU Yufei, et al. Advancing plain vision transformer toward remote sensing foundation model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5607315.

[55] SUN Xian, WANG Peijin, LU Wanxuan, et al. RingMo: a remote sensing foundation model with masked image modeling[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 3194732.

[56] REED C J, GUPTA R, LI S, et al. Scale-MAE: a scale-aware masked autoencoder for multiscale geospatial representation learning[C]//

Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 4088-4099.

[57] HONG D, ZHANG B, LI X, et al. SpectralGPT: spectral foundation modele[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2311.07113.

[58] JAKUBIK J, ROY S, PHILLIPS C E, et al. Foundation models for generalist geospatial artificial intelligence[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2310.18660v2.

[59] IRVIN J, TAO L, ZHOU J, et al. USat: a unified self-supervised encoder for multi-sensor satellite imagery[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2312.02199v1.

[60] MENDIETA M, HAN Boran, SHI Xingjian, et al. Towards geospatial foundation models via continual pretraining[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 16806-16816.

[61] TANG M, COZMA A L, GEORGIOU K, et al. Cross-scale MAE: a tale of multiscale exploitation in remote sensing[C]//Proceedings of the 37th Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc.,2023.

[62] ZHANG Mingming, LIU Qingjie, WANG Yunhong. CtxMIM: context-enhanced masked image modeling for remote sensing image understanding[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2310.00022v4.

[63] YAO Fanglong, LU Wanxuan, YANG Heming, et al. RingMo-sense: remote sensing foundation model for spatiotemporal prediction via spatiotemporal evolution disentangling[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 3316166.

[64] FULLER A, MILLARD K, GREEN J R. CROMA: remote sensing representations with contrastive radar-optical masked autoencoders [EB/OL]. [2024-01-05]. https://arxiv.org/abs/2311.00566.

[65] WANG Yi, ALBRECHT C M, ALI BRAHAM N A, et al. Decoupling common and unique representations for multimodal self-supervised learning[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2309.05300v3.

[66] TSENG G, CARTUYVELS R, ZVONKOV I, et al. Lightweight, pre-trained transformers for remote sensing timeseries[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2304.14065v4.

[67] FENG Yingchao, WANG Peijin, DIAO Wenhui, et al. A self-supervised cross-modal remote sensing foundation model with multi-domain representation and cross-domain fusion[C]//Proceedings of 2023 IEEE International Geoscience and Remote Sensing Symposium. Pasadena: IEEE, 2023: 2239-2242.

[68] ZHAO W X, ZHOU Kun, LI Junyi, et al. A survey of large language models[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2303.18223v14.

[69] ROBERTS J, LÜDDECKE T, SHEIKH R, et al. Charting new territories: exploring the geographic and geospatial capabilities of multimodal LLMs[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2311.14656v3.

[70] MALL U, PHOO C P, LIU M K, et al. Remote sensing vision-language foundation models without annotations via ground remote alignment[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2312.06960v1.

[71] LI Yansheng, WANG Linlin, WANG Tingzhu, et al. STAR: a first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2406.09410v3.

[72] LUO Junwei, PANG Zhen, ZHANG Yongjun, et al. SkySenseGPT: a fine-grained instruction tuning dataset and model for remote sensing vision-language understanding[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2406.10100v2.

[73] CEPEDA V V, NAYAK G K, SHAH M. GeoCLIP: clip-inspired alignment between locations and images for effective worldwide geolocalization[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2309.16020.

[74] MAI Gengchen, LAO Ni, HE Yutong, et al. CSP: self-supervised contrastive spatial pre-training for geospatial-visual representations [EB/OL]. [2024-01-05]. https://arxiv.org/abs/2305.01118v2.

[75] WOLTERS P, BASTANI F, KEMBHAVI A. Zooming out on zooming in: advancing super-resolution for remote sensing[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2311.18082v1.

[76] KHANNA S, LIU P, ZHOU Linqi, et al. DiffusionSat: a generative foundation model for satellite imagery[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2312.03606v2.

[77] ESPINOSA M, CROWLEY E J. Generate your own Scotland: satellite image generation conditioned on maps[EB/OL]. [2024-01-05]. https://arxiv.org/abs/2308.16648v1.

[78] 李彦胜, 吴敏郎, 张永军. 知识图谱约束深度网络的高分辨率遥感影像场景分类[J]. 测绘学报, 2024,53(4): 677-688. DOI: 10.11947/j.AGCS.2024.20230125.
LI Yansheng, WU Minlang, ZHANG Yongjun. Knowledge graph-guided deep network for high-resolution remote sensing image scene classification[J]. Acta Geodaetica et Cartographica Sinica, 2024, 53(4): 677-688. DOI: 10.11947/j.AGCS.2024.20230125.

[79] LI Yansheng, KONG Deyu, ZHANG Yongjun, et al. Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 179: 145-158.

[80] LI Yansheng, ZHU Zhihui, YU Jingang, et al. Learning deep cross-modal embedding networks for zero-shot remote sensing image scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(12): 10590-10603.

[81] LI Yansheng, ZHOU Yuhan, ZHANG Yongjun, et al. DKDFN: domain knowledge-guided deep collaborative fusion network for multi-

modal unitemporal remote sensing land cover classification[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 186: 170-189.

[82] LI Yansheng, OUYANG Song, ZHANG Yongjun. Combining deep learning and ontology reasoning for remote sensing image semantic segmentation[J]. Knowledge-Based Systems, 2022, 243: 108469.

[83] LIU Xiaojian, ZHANG Yongjun, ZOU Huimin, et al. Multi-source knowledge graph reasoning for ocean oil spill detection from satellite SAR images[J]. International Journal of Applied Earth Observation and Geoinformation, 2023, 116: 103153.

[84] 李彦胜, 张永军. 耦合知识图谱和深度学习的新一代遥感影像解译范式[J]. 武汉大学学报(信息科学版), 2022,47(8): 1176-1190.
LI Yansheng, ZHANG Yongjun. A new paradigm of remote sensing image interpretation by coupling knowledge graph and deep learning [J]. Geomatics and Information Science of Wuhan University, 2022, 47(8): 1176-1190.

[85] 李彦胜, 武康, 欧阳松, 等. 地学知识图谱引导的遥感影像语义分割[J]. 遥感学报, 2024, 28(2): 455-469.
LI Yansheng, WU Kang, OUYANG Song, et al. Geographic knowledge graph-guided remote sensing image semantic segmentation[J]. National Remote Sensing Bulletin, 2024, 28(2): 455-469.

[86] CHEN Jun, CHEN Lijun, CHEN Fei, et al. Collaborative validation of GlobeLand30: methodology and practices[J]. Geo-spatial Information Science, 2021, 24(1): 134-144.

[87] GONG Peng, LIU Han, ZHANG Meinan, et al. Stable classification with limited sample: transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017[J]. Science Bulletin, 2019, 64(6): 370-373.

[88] ZHANG Lümin, RAO Anyi, AGRAWALA M. Adding conditional control to text-to-image diffusion models[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 3836-3847.

[89] 龚健雅, 李彦胜. 定量遥感与机器学习能够融合吗? [J]. 地球科学, 2022, 47(10): 3911-3912.
GONG Jianya, LI Yansheng. Can quantitative remote sensing and machine learning be integrated? [J]. Earth Science, 2022, 47(10): 3911-3912.

[90] GOODCHILD M F. The validity and usefulness of laws in geographic information science and geography[J]. Annals of the Association of American Geographers, 2004, 94(2): 300-303.

[91] 张兵, 杨晓梅, 高连如, 等. 遥感大数据智能解译的地理学认知模型与方法[J]. 测绘学报,2022,51(7):1398-1415. DOI: 10.11947/j.AGCS.2022.20220279.
ZHANG Bing, YANG Xiaomei, GAO Lianru, et al. Geo-cognitive models and methods for intelligent interpretation of remotely sensed big data[J]. Acta Geodaetica et Cartographica Sinica, 2022, 51(7): 1398-1415. DOI: 10.11947/j.AGCS.2022.20220279.

(责任编辑:张艳玲)

Multi-modal remote sensing large foundation models: current research status and future prospect

ZHANG Yongjun[1], LI Yansheng[1], DANG Bo[1], WU Kang[1], GUO Xin[2], WANG Jian[2], CHEN Jingdong[2], YANG Ming[2]

1. *School of Remote Sensing and Information Engineering, Wuhan University, Wuhan* 430079, *China*;
2. *Ant Group, Hangzhou* 310013, *China*

Abstract: The increasing remote sensing capabilities for Earth observation have eased the access to abundant data and enabled the emergence and development of remote sensing foundation models (RSFMs). Designing distinct deep neural networks and optimizing for different data and task types require substantial development efforts and prohibitively high computational resources. In order to address these issues, researchers in the remote sensing field have shifted their focus to the study of RSFMs and presented many dedicated designed unified models. To enhance the generalizability and interpretability of RSFMs, the integration of extensive geographic knowledge has been recognized as a pivotal/key approach. While existing works have explored or incorporated geographic knowledge into the architecture design or pre-training methods of RSFMs, there lacks of a comprehensive survey to review the current status of geographic knowledge-guided RSFMs. Therefore, this paper starts with summarizing and categorizing large-scale pre-training datasets and then provides an overview of the research progress in this field. Subsequently, we introduce intelligent interpretation algorithms for remote sensing imagery guided by geographic knowledge, along with advancements in the exploration and utilization of geographic knowledge specifically tailored for RSFMs. Finally, several future research prospects are outlined to tackle the persisting challenges in this field, aiming to shed light on future investigations into RSFMs.

Key words: pre-training dataset; remote sensing intelligent interpretation; remote sensing foundation models; geographic knowledge

Foundation support: The National Natural Science Foundation of China (Nos. 42030102;42371321)

First author: ZHANG Yongjun (1975—), male, PhD, professor, majors in aerospace photogrammetry and remote sensing intelligent interpretation.
E-mail: zhangyj@whu.edu.cn
Corresponding author: LI Yansheng
E-mail: yansheng.li@whu.edu.cn