遥感基础模型发展综述与未来设想

付琨[1,2,3]，卢宛萱[1,2]，刘小煜[1,2]，邓楚博[1,2]，于泓峰[1,2]，孙显[1,2]

1. 中国科学院 网络信息体系技术重点实验室, 北京 100190;
2. 中国科学院空天信息创新研究院, 北京 100094;
3. 中国科学院大学, 北京 100101

摘 要：近年来，遥感智能解译技术快速发展，但大多为专用模型难以泛化到不同任务中，易造成资源浪费。基础模型是一种通用可泛化的解决方案，最近在遥感领域备受关注。尽管目前有大量工作已利用遥感单时相或多时相数据在感知识别和认知预测的部分任务上取得显著成果，但缺乏一个全面的综述给遥感基础模型提供系统概述。因此本文首先从数据、方法和应用角度对现有遥感基础模型的研究进展进行总结，然后通过分析现状存在的局限提出新一代遥感通用预测基础模型的设想，最后针对亟需研究的方向进行探讨与实验，为研究人员提供遥感基础模型过去成果与未来可能性之间的桥梁。

关键词：遥感智能解译，遥感基础模型，通用预测，多时相数据，多任务

中图分类号：TP701/P2

引用格式：付琨，卢宛萱，刘小煜，邓楚博，于泓峰，孙显.2024.遥感基础模型发展综述与未来设想.遥感学报，28(7)：1667-1680

Fu K，Lu W X，Liu X Y，Deng C B，Yu H F and Sun X. 2024. A comprehensive survey and assumption of remote sensing foundation modal. National Remote Sensing Bulletin，28(7)：1667-1680[DOI：10.11834/jrs.20233313]

1 引 言

近年来，卫星发射数量呈爆炸式增长，根据UCS（Union of Concerned Scientists）发布的卫星数据报告，截止至2023年5月1日，全球有超过1200颗地球观测卫星在轨运行（https：//www.ucsusa.org/resources/satellite-database[2023-07-23]）。随之带来的是获取大量遥感数据的能力，比如高分系列卫星每天可覆盖全球上亿平方公里区域，下传量达到百TB级。丰富全面的数据可支撑多场景（城市、乡村、山地、海洋等）、多要素（道路、植被、车辆、飞机等）、多时相（不同季节、不同气候等）任务。但在人工专家判读的模式下，遥感数据利用率不到获取量的5%，难以完成多样化任务。

近年来，越来越多研究人员从事遥感智能解译相关工作，针对不同平台、目标、任务单独设计专用模型（王威 等，2023；田壮壮 等，2023；李治 等，2023），如“十三五”高分支撑技术体系研制了近千个独立算法模型，提升了应用效益。但这种方式需要投入的成本大，模型无法泛化到其他任务中，在一定程度上造成了资源浪费。因此迫切需要寻找更通用、更泛化的解决途径。

基于海量数据的“基础模型+下游任务”模式最近在遥感领域备受关注，已成为一种可行的通用解决方案（Sun等，2023）。基础模型利用大规模无标签遥感数据进行训练，以获取数据中的通用泛化特征，再通过增量学习快速迁移，适应多种场景或任务。在多项工作中展示了该模式的有效性，如在场景分类、目标检测、要素分割、变化检测等遥感国际基准数据集中精度提升显著（Sun等，2023；Mañas等，2021；Li等，2022a），并在重点目标识别、海洋环境监测、国土资源分类、智慧城市建设、公共卫生管理等实际业务中刷新应用效果。但这些遥感基础模型主要着重于

收稿日期：2023-07-23；**预印本**：2023-11-29

基金项目：国家自然科学基金（编号：62201550，62171436）；中国科学院重点部署科研专项（编号：KGFZD-145-23-18）；科技创新2030—“新一代人工智能”重大项目（编号：2022ZD0118401）

第一作者简介：付琨，研究方向为空天遥感数据处理与应用。E-mail：fukun@aircas.ac.cn

通信作者简介：卢宛萱，研究方向为空天大数据智能分析。E-mail：luwx@aircas.ac.cn

分析目标环境中已发生或已具备的信息（感知识别）方面，一般采用的是单时相数据，较少利用遥感时序数据。

最近有一些工作也开始探索基于多时相数据的遥感基础模型（Yuan等，2022；Yuan和Lin，2021，Cong等，2022），它们通过学习多时相数据中的时间特征，提升土地覆盖、作物等时序分类精度，但它们依旧侧重于感知识别类方向，无法支撑气象预报、交通预测、生态演化等需要预测目标环境未来状态信息（认知预测）的任务。去年开始，部分研究人员在气象领域提出了气象预测基础模型（Bi等，2022；Chen等，2023），利用欧洲中期天气预报中心提供的ERA5数据挖掘一定时空范围内相关气象要素的变化规律，在风速预测、温度预测、热带气旋预测等方面取得超越传统物理方法的效果。然而这些模型都是针对气象相关任务的，无法适用于遥感领域其他预测方向（如森林退化预测、舰船轨迹预测、河道变迁预测等）。

针对以上问题，结合实际应用需求，本文提出新一代遥感通用预测基础模型的设想，通过学习天/临/空/地多源异构多时相数据规律，获取稳定泛化的时序通用特征，以支持多空间尺度、多时间尺度的认知预测任务。

图1展示了本文在谷歌学术上检索到与遥感基础模型相关的文章。从近4年来文章数量的变化趋势可以发现，遥感基础模型的文章逐年增加，已成为遥感领域的热点方向，但目前尚未有涵盖多类基础模型的总结分析工作。此外，相比基于单时相数据的基础模型，基于多时相数据的遥感时序基础模型较少，且目前没有关于通用预测基础模型的相关工作。

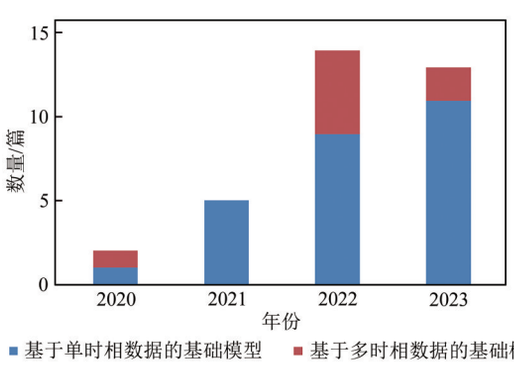

图1 遥感基础模型文章统计（此次检索于2023年6月进行）

Fig. 1 Statistics of related literatures of remote sensing foundation model (The search was conducted on June 2023)

本文希望可以为遥感领域做出以下3方面贡献：

（1）本文对遥感基础模型论文进行全面、及时的综述，并总结目前工作存在的局限。通过详尽的阐述，读者可以掌握遥感基础模型的大致情况；

（2）基于对现状的总结与分析，本文提出了新一代遥感通用预测基础模型的设想，对满足现实应用需求具有实际意义；

（3）在设想的基础上，本文进一步探讨了遥感通用预测基础模型亟需突破的技术和未来发展方向，并通过初步实验验证了设想的可行性。

2 遥感基础模型

本文按照使用的数据（单时相/多时相）和应用的任务类型（感知识别/认知预测），将现有遥感基础模型分为3类：基于单时相数据的感知识别基础模型、基于多时相数据的感知识别基础模型、基于多时相数据的认知预测基础模型。本节将依次综述每类遥感基础模型的现状，并总结分析现有方法的局限。

2.1 基于单时相数据的感知识别基础模型

与自然场景类似，遥感领域的基础模型大多采用自监督学习方式，基于大量未标记的遥感单时相数据，挖掘通用表征信息，并迁移到分类、检测、分割、变化检测等感知识别类下游任务中。本文根据采用的自监督学习方法的不同，将现有基于单时相数据的感知识别基础模型分为基于对比式学习的基础模型和基于生成式学习的基础模型，如图2所示，并在表1中总结了每个模型使用的方法、数据和任务。

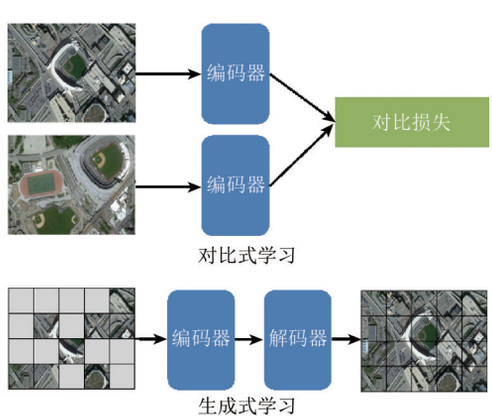

图2 对比式学习与生成式学习的对比

Fig. 2 A comparison of the contrastive learning and generative learning

表1 基于单时相数据的感知识别基础模型总结

Table 1 A gallery of the foundation model of perceptual recognition based on single-temporal data

类别	方法	数据	任务
对比式学习	Jung等(2022)：利用多个遥感视图求平均表示	ImageNet、xView	场景分类
	Ayush等(2021)：时序遥感图像作为正对，同时结合地理位置信息	fMoW、GeoImageNet	语义分割、目标检测、土地覆盖分类
	Jain等(2021)：将多光谱和SAR图像视为一张图像的不同增强视图	Sen12MS	场景分类
	Stojnić和Risojević(2021)：在遥感场景中应用CMC算法	ImageNet1000、BigEarthNet等	场景分类
	Zheng等(2021)：将MoCo、CLD与几何增强结合	KWD-Pre	野生动物识别
	SeCo(Mañas等，2021)：将对比学习方法与遥感时间信息相结合	SeCo	场景分类、变化检测
	GeoKR(Li等，2022a)：使用地理知识驱动遥感图像表示学习	Levir-KR	场景分类、目标检测、语义分割、云/雪检测
	GeCo(Li等，2022b)：利用全球土地覆盖产品进行预训练网络的自适应校正	Levir-KR	场景分类、语义分割、目标检测
	RS-BYOL(Jain等，2022)：将光谱信息和空间分辨率信息作为隐式增强学习不变特征嵌入	Sen12MS	场景分类、语义分割
	Bourcier等(2022)：设计扩展动量对比框架进行遥感预训练	fMoW	场景分类、车辆检测
	MATTER(Akiva等，2022)：利用多时相遥感图像学习光照与视角的不变性	Sentinel-2	场景分类、语义分割、变化检测
	CSP(Mai等，2023)：使用双编码器分别对图像和地理位置进行编码，并用对比损失学习一致性位置表示	iNat2018、fMoW	细粒度物种识别、场景分类
	DINO-MC(Wanyan等，2023)：不同尺寸作物的局部视图取代固定视图的方式进行训练	SeCo	场景分类、变化检测
	Patnala等(2023)：使用大气校正生成遥感图像的多个视图	SeCo	场景分类
	IaI-SimCLR(Prexl和Schmitt，2023)：利用遥感图像的地理信息提升负例难度	SEN12MS	场景分类、作物分类、生物数量估计
	CACo(Mall等，2023)：利用卫星图像的时序不变性特点设计新型对比损失	Sentinel-2	土地覆盖分类、语义分割、变化检测
	Heidler等(2023)：通过配对的遥感图像和音频数据间的对应关系提取遥感场景特性	SoundingEarth	场景分类、场景分割、视听场景分类、跨模态检索
	SSL4EO-L(Stewart等，2023)：通过三种不同的传感器和两种产品学习Landsat数据特征	SSL4EO-L	场景分类、语义分割
	TOV(Tao等，2023)：先学习自然场景数据集中的通用知识，然后学习遥感数据集中的领域相关专业知识完成模型训练	TOV-NI、TOV-RS	场景分类、目标检测、语义分割
生成式学习	Vincenzi等(2021)：利用高维光谱进行可见颜色的重建	BigEarthNet	场景分类、病毒检测
	RingMo(Sun等，2023)：并设计了针对复杂场景内小型密集物体的基础模型训练方法	RingMo	场景分类、目标检测、语义分割、变化检测
	Scale-MAE(Reed等，2023)：在预训练过程中明确学习不同已知尺度的数据间的关系	FMoW	土地利用分类、语义分割
	RVSA(Wang等，2022a)：设计旋转可变尺寸窗口注意力模块学习遥感特征	MillionAID	目标检测、场景分类、语义分割
	Cha等(2023)：扩增ViT参数量，并训练获取十亿级的遥感基础模型	MillionAID	旋转目标检测、语义分割
	CMID(Muhtar等，2023)：通过自蒸馏将对比学习与生成式学习结合训练	MillionAID	场景分类、语义分割、对象检测、变化检测
	GFM(Mendieta等，2023)：利用遥感数据时空结构的地理感知学习	GeoPile	场景分类、语义分割、目标检测

部分遥感基础模型基于对比学习方法，如MoCo（He等，2020；Chen等，2020c；Chen等，2021）、SimCLR（Chen等，2020a；Chen等，2020b）和CMC（Tian等，2020），通过数据增强产生多个样本，再利用对比损失学习通用特征。Jung等（2022）提出一种基于SimCLR框架的遥感平滑表示的自监督学习方法，输入多个图像并对其表示进行平均化操作。Zheng等（2021）结合了MoCo与几何增强等方法，提升了预训练模型的性能。部分研究人员发现可以利用遥感中同一个空间位置在不同时相上的数据属于同一类别的特点，因此设计了多时相视角的对比学习方法。Mañas等（2021）提出季节性对比损失进行遥感基础模型SeCo的训练，并收集大量遥感数据构建了同名数据集，在分类、变化检测等任务中取得显著效果。Mall等（2023）使用时序信息来对比具有长期和短期差异的图像，同时利用卫星图像不经常变化的特点设计了一种新的对比损失CACo Loss，和现有基础模型相比，提高了模型在土地覆盖分类、语义分割、变化检测等方面的准确率。在此基础上，一些研究人员还结合了遥感图像自带的地理信息提升基础模型性能。Li等（2022a）提出遥感地理知识驱动的基础模型训练方法GeoKR，将土地覆盖产品和地理位置视为地理知识，为模型训练提供自监督信息，同时构建了大规模数据集LevirKR支撑模型训练，减轻了场景分类、语义分割、目标检测等下游任务的标注负担。GeCo（Li等，2022b）方法利用地理先验知识指导并纠正表示学习过程，保证自适应校正过程的正确性，消除偏差影响，在场景分类、语义分割、目标检测等任务中取得了更好的效果。除此之外，还有一些工作引入了其他数据进行对比学习，如多光谱、SAR等多模态数据、音频数据等。Jain等（2021）将多光谱和SAR图像视为一张图像的不同增强视图来学习它们之间的相似性，以此获得更好的一致性表示。Heidler等（2023）使用配对的图像和音频数据进行训练，利用图像和音频数据间的对应关系，学习遥感场景中的关键属性，在航拍场景分类、航拍语义分割、视听场景分类、跨模态检索等任务中进行实验，证明方法的有效性。

随着基于Transformer的生成式学习方法在计算机视觉领域基础模型方面取得巨大的成功，越来越多遥感领域的基础模型使用生成式学习，取得了较好的效果。Sun等（2023）率先提出RingMo遥感基础模型框架，构建了百万级大规模遥感数据集，并设计了针对复杂场景内小型密集物体的基础模型训练方法。在场景分类、目标检测、语义分割、变化检测等任务的国际标准数据集中性能提升显著。Scale-MAE（Reed等，2023）以已知比例掩码的输入图像来训练基础模型，在整个预训练过程中明确学习不同已知尺度的数据之间的关系，在8个遥感数据集上实现了下游任务的效果提升。Wang等（2022a）基于ViT提出针对多样遥感任务的基础模型RVSA，使用旋转可变尺寸窗口注意力来适应遥感图像的大尺寸和目标的任意方向，显著提高基础模型在分类、检测和分割等任务的准确率。Cha等（2023）构建了遥感领域十亿级基础模型，发现模型性能和数据效率随着参数数量的增加而提高，在旋转目标检测和语义分割等下游任务中实现了先进的性能。CMID模型（Muhtar等，2023）以自蒸馏的方式将对比式学习与生成式学习结合来学习全局和局部表示，CMID还可与CNN、ViT兼容，在多个下游任务中具有更好性能。Mendieta等（2023）以构建高效的遥感基础模型为立足点，首先建立了小型但多样化的数据集GeoPile，然后提出多目标持续预训练范式，兼顾蒸馏学习和生成式自监督学习，在利用最小资源的同时显著提升模型性能。

2.2 基于多时相数据的感知识别基础模型

近年来随着遥感技术发展，能够获取到时间间隔更短、空间分辨率更高的地球观测数据，针对同一地理区域连续获取的数据可转化为时序遥感数据（Gómez等，2016），此类数据包含丰富的地表状态和动态演变信息，常用来对局部或大范围的地表覆盖进行研究（Ienco等，2019）。

和基于单时相数据的感知识别基础模型类似，自监督学习方法常被用于时序遥感数据训练，并应用于土地覆盖、作物等细粒度分类任务中，表2对常见的基于多时相数据的感知识别基础模型使用的数据和任务进行了总结。SITS-Former（Yuan等，2022）基于Transformer利用自监督学习通过缺失数据补全任务在大量未标记的Sentinel-2多时相数据上进行训练。给定一个不完整的多时相数据，部分数据被随机掩码，模型被要求根据未掩码数据的信息恢复掩码数据，因此

模型可从数据中捕获高级的空间和时间依赖性，学习到判别性特征。SITS-Former将训练好的模型参数迁移到作物分类任务中获得显著的性能增益。

表2 基于多时相数据的感知识别基础模型总结

Table 2 A gallery of the foundation model of perceptual recognition based on time series data

类别	方法	数据	任务
单任务	SITS-Former(Yuan等,2022)：通过缺失数据补全任务在大量未标记的Sentinel-2多时相数据上自监督训练	Sentinel-2卫星数据	作物分类
多任务	SITS-BERT(Yuan和Lin,2021)：基于Transformer利用多时相数据固有的时间结构学习相关的通用时间特征	Sentinel-2卫星数据	作物分类、土地覆盖分类
	SatMAE(Cong等,2022)：基于MAE的多时相/多光谱卫星图像预训练框架	fMoW RGB、fMoW Sentinel数据集	土地覆盖分类、建筑物分割
	Presto(Tseng等,2024)：专为地球观测多时相数据设计的基于Transformer的轻量级基础模型	Sentinel-2卫星数据、ERA5气象数据、地形数据、土地覆盖数据	作物分割、燃料湿度回归、树木分类、土地覆盖分类

与此同时，基于多时相数据的感知识别基础模型不断扩展应用于其他下游任务。针对时序遥感数据标记少的问题，Yuan和Lin（2021）提出SITS-BERT模型，利用多时相数据固有的时间结构学习相关的通用时间特征，进行作物分类和土地覆盖研究，提高了模型的泛化性能并减少过拟合的风险。Cong等（2022）提出一种基于MAE（He等，2022）的多光谱多时相基础模型SatMAE，通过跨时间独立掩码方法充分利用时序信息，同时将多光谱数据编码为带有不同光谱的位置嵌入的频谱组，最后将训练好的基础模型参数迁移到土地覆盖分类和建筑物分割下游任务中，均取得了良好的结果。Tseng等（2024）提出一个专为地球观测多时相数据设计的基于Transformer的轻量级基础模型，通过自监督学习方法充分利用多传感器时间序列数据的结构，显著减少基础模型训练所需的参数量，并泛化于作物分割、燃料湿度回归、树木分类、土地覆盖分类等任务中。

2.3 基于多时相数据的认知预测基础模型

最近认知预测类任务开始受到研究人员的关注，因此也出现了一些基于多时相数据的认知预测基础模型。但目前，认知预测基础模型大多服务于气象预报应用，通过挖掘一定时空范围内相关气象要素的时空动态特征，了解气象要素变化规律，实现气象预测。当前已知的预测基础模型包括FourCastNet、盘古—气象、GraphCast、风乌—天气，表3对上述模型进行了详细介绍和说明。

表3 基于多时相数据的认知预测基础模型总结

Table 3 A gallery of the foundation model of cognitive prediction based on time series data

类别	方法	数据	任务
Transformer架构	FourCastNet(Pathak等,2022)：在ViT架构中结合傅里叶神经算子	ERA5	飓风预测、大气层河流预测、降水预测、温度预测、风速预测等
	盘古—气象(Bi等,2022)：将高度信息公式化为立方体数据并应用层次时间聚合减轻误差积累	ERA5	降水预测、温度预测、风速预测、热带气旋预测等
	风乌—天气(Chen等,2023)：基于多任务自动均衡权重与缓存回放策略,减少长时序自回归预测误差	ERA5	湿度预测、风速预测、温度预测等
图网络架构	GraphCast(Lam等,2023)：将原始经纬度网格映射到多网格,通过深度图网络实现有效信息传递	ERA5	湿度预测、风速预测、温度预测等

2022年2月，NVIDIA提出了FourCastNet（Pathak等，2022）天气预测模型，使用了64个Nvidia A100 GPU进行训练。模型结合了ViT（Dosovitskiy等，2021）和傅里叶神经网络（Guibas等，2022），模型的预报分辨率提升到了和物理模型相比拟的水平，速度与物理模型相比快了多个数量级，但在部分气象下游任务上精度仍低于传统物理模型。盘古—气象（Bi等，2022）基础模

型是2022年11月由华为提出的，使用了192个NVIDIA Tesla-V100 GPU进行训练，他们提出3D Transformer方法，输入和输出均为指定时间点的三维天气状态，并结合层次化时域聚合算法最小化迭代误差，他们的长期预报精度首次全面超过传统方法，并将时间效率提升至秒级。2022年12月，ECMWF提出一种基于图神经网络的自回归模型GraphCast（Lam等，2023），训练中模型使用了32台Cloud TPU v4，将原始经纬度网格映射到多网格上学习特征，通过深度图网络有效传递信息，实验结果证明，在中期天气预报中，模型性能优于欧洲气象中心的高精度预报。2023年4月，上海人工智能实验室提出风乌—天气（Chen等，2023）基础模型，采用多模态神经网络，结合多任务自动均衡权重策略，加强变量间协同优化作用，并提出了“缓存回放”策略，减少自回归预测误差，提高长期预测的性能，模型训练使用了32个Nvidia A100 GPU进行训练，在预报精度、预报时效和资源效率3方面均有了显著提升。

2.4 现有遥感基础模型的局限

近年来，遥感基础模型发展迅速，在众多应用任务中展示了显著的效果，但在认知预测方面还处于起步阶段，在数据、方法、任务上还存在一些局限：

（1）数据方面：当前遥感基础模型对于单时相数据的利用较为全面，不止从卫星、无人机等多个平台中获取数据，还涵盖了可见光、SAR、多光谱等多种模态。然而多时相数据的应用较为局限，目前只包含了从卫星平台获取的时序图像和气象数据，其他平台和模态的数据没有被充分利用。

（2）方法方面：针对时序认知预测的遥感基础模型大多采用Transformer架构，只有一个模型采用图网络。Transformer受限于输入长度的问题，难以扩展到遥感大图中。图网络虽然可以建模多尺度关系，但是训练一个大规模图网络难收敛，且容易过平滑（Ying等，2021）。遥感认知预测任务一般需要观察大场景中多个目标间的关联关系，且多时相数据体量大，模型训练困难。目前尚未有基础模型根据遥感特点突破常见模型架构。

（3）应用方面：地球上的要素不断变化，因此遥感时序认知预测应用很广泛，包含多样场景（城市、森林、河道、气象等）、多类任务（轨迹预测、要素演化、数值预测等），然而目前遥感预测基础模型局限在气象领域，只能泛化到风速、温度、湿度、热带气旋等方面，缺少通用预测基础模型，无法适应多样场景和多种任务。

3 面向多域异构多时相数据的新一代通用预测基础模型

基于第2节对现有遥感基础模型的总结与分析，本节提出新一代遥感通用预测基础模型的设想，并进一步讨论了在该设想下，亟需研究人员探索的未来方向。

3.1 核心思路

构建面向遥感多时相数据的新一代通用预测基础模型，共享学习多域异构多时相数据，支持多空间尺度、多时间尺度的预测任务，如图3所示。

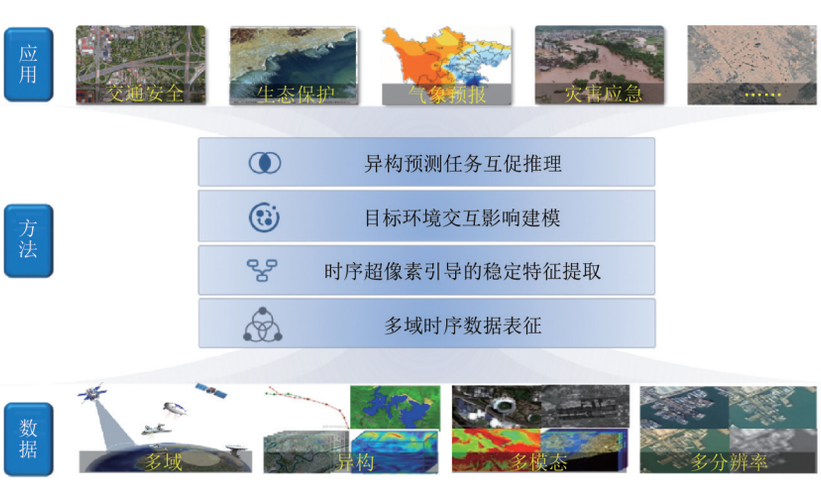

图3 面向多域异构时序数据的遥感通用预测基础模型

Fig. 3 Remote sensing prediction foundation model for multi-domain heterogeneous time-series data

具体来说：

（1）数据方面：涵盖天/临/空/地多平台、轨迹点/时序图像/视频/气象数据等多类型、可见光/SAR/多光谱等多模态、厘米级到百米级多分辨率的遥感多时相数据；

（2）方法方面：结合图网络和Transformer模型的优势，设计基础模型全新架构，具备对遥感大场景中多目标交互的长时序稳定预测能力，同时扩大模型容量，提升泛化效果；

（3）应用方面：通用预测基础模型可应用到多空间尺度（目标级、要素级、区域级）、多时间尺度（近实时、小时级、长时序）的多样化认知预测任务中。

3.2 探索性方向

新一代遥感通用预测基础模型的核心是打通多域异构多时相数据输入及多时间/空间尺度任务输出的信息通路，通过提取稳定泛化的时序超像素特征，实现对未来状态的精准认知预测。为了实现以上目的，本文提出多域时序数据表征、稳定规律特征提取、目标环境交互影响建模以及多任务互促推理四个探索性方向，以供从事遥感基础模型的研究人员参考。

（1）多域时序数据表征。为了在遥感领域实现通用认知预测的目的，需要兼容多域异构多时相数据，但这些数据在时间采样间隔、空间分辨率、数据维度等多方面均有显著差异。时间采样间隔上，普通视频的帧率为24帧/s，即每帧跨度约0.04 s；而大部分卫星受限于重访周期、云层干扰等原因，其获取清晰数据的时间跨度以小时/天为计量单位。在空间分辨率上，从厘米级的无人机数据到分辨率数米的卫星数据，其囊括的范围也从几百平方米跨越到全球尺度。在数据维度上，输入多时相数据的形态各异，如一维轨迹、二维气象数据、三维时序图像等不同维度的数据。因此，如何对多域异构多时相数据进行统一处理以实现多样特征的自动化提取是需要探索的方向。

针对以上问题，本文提供一些可能的解决思路以供研究人员考虑。针对时间采样间隔不一致的问题，可通过数据相邻帧的相似度衡量时序冗余度，并基于此为时序冗余度高的数据选取更高的掩码比率，使得模型能够处理不同时间间隔的数据，并具备对不同时间尺度数据的时空预测能力。针对空间分辨率不同的问题，可采用金字塔结构进行空间多尺度特征提取。金字塔是数据空间多尺度表达的一种，它实际上是一张图片在不同尺度下的集合。通过图像金字塔结构统一不同分辨率的数据特征到同一尺度，达到不同分辨率数据的统一处理。针对多时相数据维度不同问题，使用不同模态专家学习一维、二维、三维数据独有的特征，再使用“掩码预测”方式统一不同数据的训练策略，使得模型的训练过程更加简单高效。

（2）时序超像素引导的稳定特征提取。时序预测涉及在空间和时间维度上对未知系统状态的预测，需要对各种变量间的时空依赖进行建模。目标与环境的行为活动、变化过程遵守着显性或隐性的规律，以往传统模型只能基于显性规律人工建模物理方程，因此面临多重挑战。例如著名的三体问题，每一个物体在其他两个物体的万有引力作用下的运动方程可以表示成6个一阶的常微分方程。因此，一般三体问题的运动方程可以用18个微分方程描述。如果想要准确预测3个物体的未来状态，必须求解18个方程才能得到解析解。在这个例子中，通过观测而获得三体状态（位置、速度、加速度）的信息满足万有引力定律以及牛顿第二定律。上述例子是在理想环境下的建模，已被证明无法获得解析解，只能通过数值模拟进行预测。而遥感场景中的目标及环境变化更加复杂多变，其时序变化遵循多重规律，并且许多为隐性规律，难以通过显式的方程进行建模，因此传统预测模型难以从复杂多变的表象中挖掘隐藏在内部的隐性规律，获取遥感数据中的稳定特征。数据驱动的深度学习方法在一定程度上解决了复杂场景中高维、非线性规律、隐变量的拟合问题，然而基于单一数据源的预测基础模型只能实现单方面任务的认知，其所存储的稳定特征都是有限的，难以提取复杂场景下海量变化要素的规律特征，无法完成多样化预测任务。

针对以上问题，本文提出时序超像素概念。时序超像素定义为：在复杂变化场景中表现出稳定规律且具有关联约束的时序像素特征集合。虽然像素在变化，但变化规律是稳定的，时序超像素是可通过函数建模的时序像素集。通用预测基础模型需要从复杂、随机的变化中学习稳定和关

联的变化规律，获取时序超像素特征，并用其进行预测。通过多维大数据+通用基础模型的模式可提取稳定的超像素特征，大量时序超像素特征被提取和存储到基础模型的网络中，由此实现多样化任务的精准预测。同时，现有基础模型通常缺少预测的不确定性建模，可利用时序超像素整合数据驱动和物理模型驱动方法促进可解释性，降低预测误差，提高预测准确率。

（3）目标环境交互影响建模。在遥感数据中目标的行为活动规律预测是人们关注的重点之一，目标的行为活动和时序变化除了受到物理定律的约束之外，还会受到目标间、目标与环境间的交互影响。目标间在时间维度上存在相互依存的关系，使得模型在捕获长期依赖关系时面临与真实关联偏移逐渐扩大的情况，使得演化计算结果随时序的延长而误差逐渐增大。此外目标受到周围地形、天气等变化环境的影响，目标行为表现出突发性强、行动难追踪、交互变化快等特点，导致目标行为具有随机性，在此情况下长时序预测结果会产生较大误差。

针对上述问题，本文考虑综合利用图网络和Transformer增强目标—目标、目标—环境间的多样化信息交互能力。利用图网络的谱性质设计算法对复杂网络中的关联关系进行较准确的挖掘，进而嵌入Transformer架构实现大规模网络训练，可考虑两方面的结合方式。一方面是在位置嵌入基础上，引入图网络的拓扑结构，以衡量节点间的关联关系。另一方面在Transformer的多头注意力基础上，加入图网络节点间特征和连接节点的边特征的结构信息修正注意力分数。在此基础上，本文还考虑对图网络与Transformer结合的方式进行拓展，提出超图Transformer结构。超图与简单图不同，对于一个简单图，其每条边均与两个顶点相关联，即每条边的度都被限制为2。而超图则允许每一条边的度为任何非负整数，因此超图可以处理多元关系和高阶关系。遥感场景中目标-目标、目标-环境间的相互关系是多元的，超图能够更好地建模这种多对多的关系，在目标环境交互预测中表现出更好的性能。

（4）异构预测任务互促推理。遥感时序预测应用包含了目标轨迹预测、要素演化预测、数值趋势预测等任务，各任务间差异大，具有不同时间尺度、不同空间尺度的特点。当前遥感预测基础模型都是针对气象预测单一应用场景的模型，无法支撑对跨场景多任务复杂情况进行分析与预测。此外，现有方法往往忽略对预测任务间特征关系的显式建模，任务间特征差异明显，梯度竞争严重，导致多任务处理能力受限，尤其是在序列化任务流中，存在严重的灾难性遗忘问题。

本文提出的面向时序多任务的通用预测基础模型需要增强任务间的关系，优化网络持续扩展策略，引导模型动态更新过程中的网络参数更新方向。通过对任务特征关系的建模，利用任务的互补信息实现相互促进。通过适当的训练，深度神经网络中由低到高的隐层充当着复杂程度不断增加的特征变换，这些变换共享不同任务中共有的隐藏特征。尽管对数域的线性分类器对不同的任务在必要时可以分开，但特征转换仍然可以在跨任务之间进行共享。具体来说，可先将数据特征映射到同一个向量空间创建一个联合的数据嵌入空间，使得多个任务可以共享特征空间中的特征，实现任务间共性和差异的显式建模，提高多任务推理预测精度。

3.3 初步实验进展

本团队已开展新一代遥感通用预测基础模型的技术攻关，利用3.2节提出的思路，采用简单直接的方式初步构建了通用预测基础模型的原型架构。首先利用不同模态数据专用的时序特征提取模型将输入数据转化为模型可计算的特征序列，然后对特征序列进行随机掩码后通过参数共享的多维信息交互Transformer基础模型学习稳定的超像素特征，进而利用之前时刻数据预测掩码的未来时刻数据，达到通用预测基础模型融合训练的目的。最后利用训练好的基础模型参数在多个预测任务中进行微调实验，取得一些实验进展。用于预训练的多域时序数据来自天/临/空/地多个平台，涵盖时序图像、视频、轨迹点等多个类型，共包含11359200帧数据。模型共训练200个epoch，优化器选用AdamW，初始学习率设置为0.0005，衰减策略为余弦衰减。通过以上训练设置得到的预测基础模型具有通用泛化的特点，适用于多种下游任务。

本文在多类遥感认知预测下游任务上进行实验，包括移动目标场景预测、降水即时预测、云图预测，选用的数据集分别为MOR-UAV、

HuaBei2021、CloudCast。对于每个数据集，根据数据集的常用设置划分训练集与测试集，MOR-UAV、HuaBei2021、CloudCast数据集的训练集比例分别为85%、90%、25%，SOTA方法与本文方法都是采用相同的训练集与测试集。其中移动目标场景预测采用MOR-UAV数据集，该数据集中包含30个无人机视频，移动目标包括小型汽车、重型车辆等，场景涵盖停车场、十字路口等，因此移动目标的运动会受到场景环境的影响。该任务是个短时预测任务，利用前面帧的视频数据预测未来帧。降水即时预测采用的是HuaBei2021数据集，该数据集包含2021年6月至8月华为地区的雷达回波数据。该任务根据雷达探测得到的回波数据来确定降水的变化情况，并预测数小时后雷达回波的状态。云图预测采用的是CloudCast数据集，该数据集共包含11种不同的云类型，在2017年—2018年期间每15 min记录一次。该任务通过给定过去一段时间的云图，学习当前时间段的时空动态预测未来一段时间同一区域的时序云图。具体实验结果如表4所示。

表4 认知预测任务定量精度对比

Table 4 Quantitative comparison of remote sensing cognitive prediction tasks

方法	移动目标场景预测			降水即时预测			云图预测	
	SSIM ↑	MSE ↓	MAE ↓	SSIM ↑	POD ↑	FAR ↓	SSIM ↑	PSNR ↑
SOTA	0.5301	3206.53	18731.12	0.9455	0.2101	0.3773	0.9633	41.23
本文方法	0.6665	1994.15	12516.83	0.9485	0.2280	0.3677	0.9640	41.30

注：“↑”表示数值越高越好，“↓”表示数值越低越好。

从表4可以看出，本文设计的遥感通用预测基础模型在3类认知预测下游任务中，无论是精度类指标还是误差类指标，都取得比当前最佳（SOTA）方法优异的性能。图4展示了移动目标场景预测的可视化图，尽管SOTA方法（Wang等，2022b）可以提取时空特征，但预测结果比较模糊，特别是遥感场景中的小型目标。相比之下，本文方法预测得到的预测模型结果更清晰，更接近真值。图5展示了降水即时预测的可视化图，图5中SOTA方法（Shi等，2015）输出的预测结果不仅模糊，而且与真值不一致，而本文方法给出了更清晰、更准确的结果。图6展示了云图预测的可视化图，可以看出本文方法的预测结果比SOTA方法（Wang等，2022b）更接近真实值，预测结果相对清晰。但目前训练出的预测基础模型还存在一些缺陷，一方面是部分任务的预测结果相对模糊，另一方面是目前模型的预测能力随着预测时间的推移而减弱，未来本团队将通过继续改进解决这些问题。

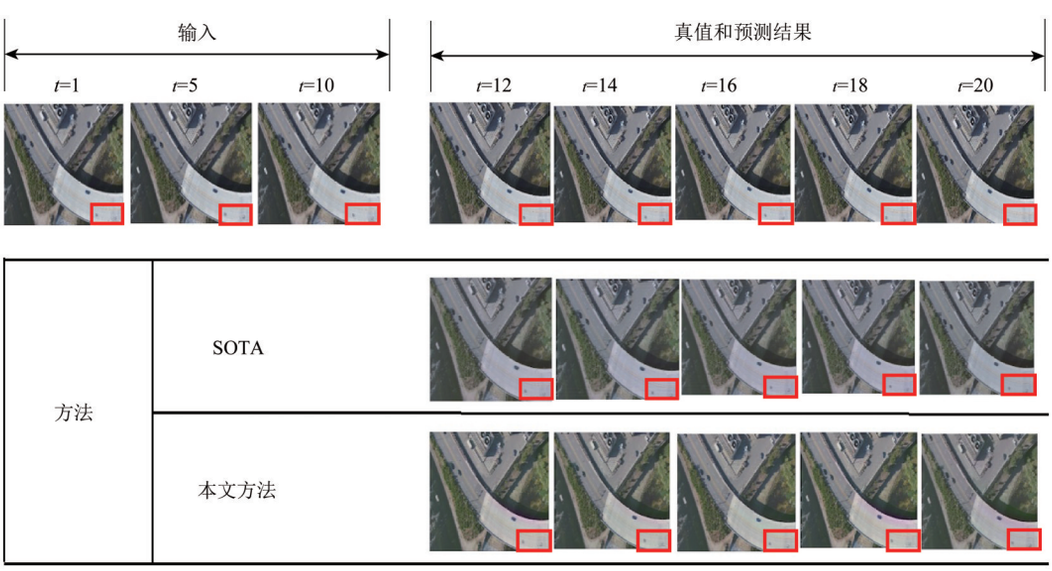

图4 移动目标场景预测可视化结果

Fig. 4 Qualitative visual comparison of moving object scenarios prediction tasks

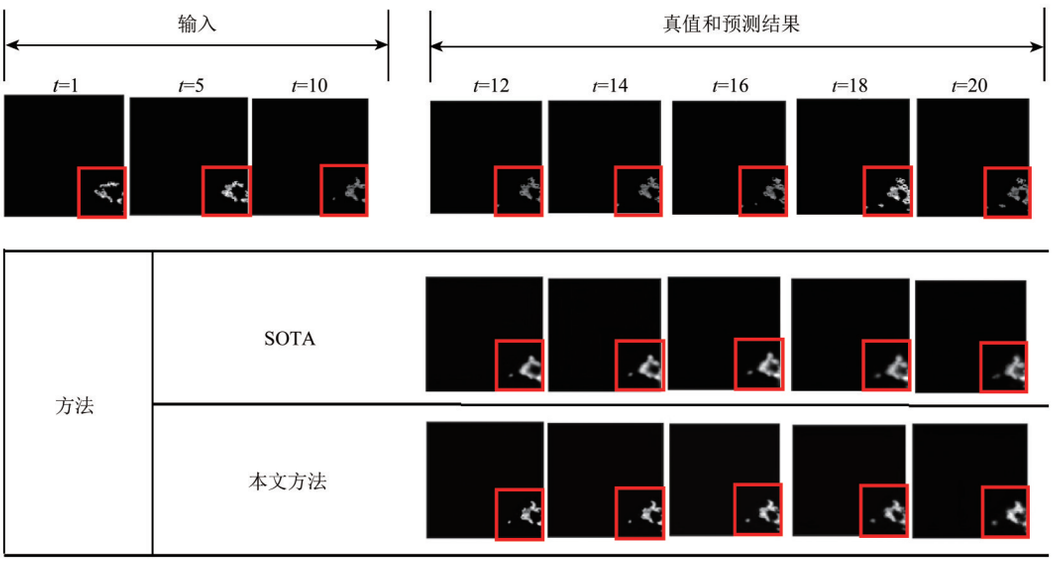

图5 降水即时预测可视化结果

Fig. 5 Qualitative visual comparison of radar echo extrapolation tasks

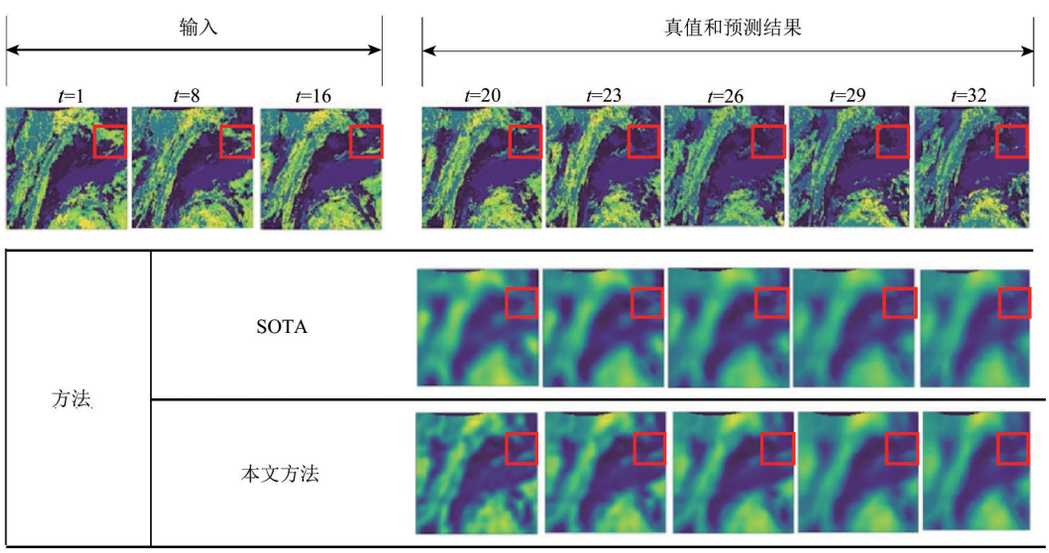

图6 云图预测可视化结果

Fig. 6 Qualitative visual comparison of cloud forecasting tasks

以上展示了目前在新一代遥感通用预测基础模型设想下，进行部分技术攻关后取得的初步进展，未来本团队会继续突破相关技术，在更全面的认知预测任务中获得明显能力增益。

4 结 论

具有通用泛化能力的基础模型对于遥感智能解译的进一步发展至关重要。本文通过整理基于单时相数据的感知识别基础模型、基于多时相数据的感知识别基础模型、基于多时相数据的认知预测的基础模型的研究现状，为研究人员提供该领域的最新进展综述。在此基础上，通过分析当前遥感基础模型在数据、方法、应用上存在的局限，提出新一代遥感通用预测基础模型的设想，并进一步明确该设想下亟需突破的4个探索性方向并进行初步实验。后续工作将在多域多时序数据表征、稳定规律特征提取、目标环境交互影响建模以及多任务互促推理方面进行针对性的关键技术突破，同时继续探索更为通用的遥感基础模型，将感知识别与认知预测整合到一个架构中。

参考文献(References)

Akiva P, Purri M and Leotta M. 2022. Self-supervised material and

texture representation learning for remote sensing tasks//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE: 8193-8205 [DOI: 10.1109/CVPR52688.2022.00803]

Ayush K, Uzkent B, Meng C L, Tanmay K, Burke M, Lobell D and Ermon S. 2021. Geography-aware self-supervised learning//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 10161-10170 [DOI: 10.1109/ICCV48922.2021.01002]

Bi K F, Xie L X, Zhang H H, Chen X, Gu X T and Tian Q. 2022. Pangu-weather: a 3D high-resolution model for fast and accurate global weather forecast. arXiv preprint arXiv: 2211.02556 [DOI: 10.48550/arXiv.2211.02556]

Bourcier J, Floquet T, Dashyan G, Ceillier T, Alahari K and Chanussot J. 2022. Self-supervised pretraining on satellite imagery: a case study on label-efficient vehicle detection. arXiv preprint arXiv: 2210.11815 [DOI: 10.48550/arXiv.2210.11815]

Cha K, Seo J and Lee T. 2023. A billion-scale foundation model for remote sensing images. arXiv preprint arXiv: 2304.05215 [DOI: 10.48550/arXiv.2304.05215]

Chen K, Han T, Gong J C, Bai L, Ling F H, Luo J J, Chen X, Ma L M, Zhang T N, Su R, Ci Y Z, Li B, Yang X K and Ouyang W L. 2023. FengWu: pushing the skillful global medium-range weather forecast beyond 10 days lead. arXiv preprint arXiv: 2304.02948 [DOI: 10.48550/arXiv.2304.02948]

Chen T, Kornblith S, Norouzi M and Hinton G. 2020a. A simple framework for contrastive learning of visual representations//Proceedings of the 37th International Conference on Machine Learning. Virtual: JMLR.org: 1597-1607

Chen T, Kornblith S, Swersky K, Norouzi M and Hinton G. 2020b. Big self-supervised models are strong semi-supervised learners//Proceedings of the 34th International Conference on neural Information Processing Systems. Vancouver: Curran Associates Inc.: 22243-22255 [DOI: 10.48550/arXiv.2006.10029]

Chen X L, Fan H Q, Girshick R and He K M. 2020c. Improved baselines with momentum contrastive learning. arXiv preprint arXiv: 2003.04297 [DOI: 10.48550/arXiv.2003.04297]

Chen X, Xie S, He K. An empirical study of training self-supervised vision transformers[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 9640-9649.

Cong Y Z, Khanna S, Meng C L, Liu P, Rozi E, He Y T, Burke M, Lobell D B and Ermon S. 2022. SatMAE: pre-training transformers for temporal and multi-spectral satellite imagery//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc.: 197-211

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2020. An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929 [DOI: 10.48550/arXiv.2010.11929]

Gómez C, White J C and Wulder M A. 2016. Optical remotely sensed time series data for land cover classification: a review. ISPRS Journal of Photogrammetry and Remote Sensing, 116: 55-72 [DOI: 10.1016/j.isprsjprs.2016.03.008]

Guibas J, Mardani M, Li Z Y, Tao A, Anandkumar A and Catanzaro B. 2021. Adaptive fourier neural operators: efficient token mixers for transformers. arXiv preprint arXiv: 2111.13587 [DOI: 10.48550/arXiv.2111.13587]

He K M, Chen X L, Xie S N, Li Y H, Dollár P and Girshick R. 2022. Masked autoencoders are scalable vision learners//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE: 15979-15988 [DOI: 10.1109/CVPR52688.2022.01553]

He K M, Fan H Q, Wu Y X, Xie S N and Girshick R. 2020. Momentum contrast for unsupervised visual representation learning//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 9726-9735 [DOI: 10.1109/CVPR42600.2020.00975]

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

Ienco D, Interdonato R, Gaetano R and Minh D H T. 2019. Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for land cover mapping via a multi-source deep learning architecture. ISPRS Journal of Photogrammetry and Remote Sensing, 158: 11-22 [DOI: 10.1016/j.isprsjprs.2019.09.016]

Jain P, Schoen-Phelan B and Ross R. 2021. Multi-modal self-supervised representation learning for earth observation//2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. Brussels: IEEE: 3241-3244 [DOI: 10.1109/IGARSS47720.2021.9553741]

Jain P, Schoen-Phelan B and Ross R. 2022. Self-supervised learning for invariant representations from multi-spectral and SAR images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15: 7797-7808 [DOI: 10.1109/JSTARS.2022.3204888]

Jung H, Oh Y, Jeong S, Lee C and Jeon T. 2022. Contrastive self-supervised learning with smoothed representation for remote sensing. IEEE Geoscience and Remote Sensing Letters, 19: 8010105 [DOI: 10.1109/LGRS.2021.3069799]

Lam R, Sanchez-Gonzalez A, Willson M, Wirnsberger P, Fortunato M, Alet F, Ravuri S, Ewalds T, Eaton-Rosen Z, Hu W H, Merose A, Hoyer S, Holland G, Vinyals O, Stott J, Pritzel A, Mohamed S and Battaglia P. 2022. GraphCast: learning skillful medium-range global weather forecasting. arXiv preprint arXiv: 2212.12794 [DOI: 10.48550/arXiv.2212.12794]

Li W Y, Chen K Y, Chen H and Shi Z W. 2022a. Geographical knowledge-driven representation learning for remote sensing

images. IEEE Transactions on Geoscience and Remote Sensing, 60: 5405516 [DOI: 10.1109/TGRS.2021.3115569]

Li W Y, Chen K Y and Shi Z W. 2022b. Geographical supervision correction for remote sensing representation learning. IEEE Transactions on Geoscience and Remote Sensing, 60: 5411520 [DOI: 10.1109/TGRS.2022.3202499]

Li Z, Sui Z W, Fu Q Y, Zheng J J and Bu T. 2023. High-resolution remote sensing extraction of urban buildings based on morphological sequences and multi-source a priori information. National Remote Sensing Bulletin, 27(4): 998-1008 (李治, 隋正伟, 傅俏燕, 郑璀璀, 卜桐. 2023. 基于形态学序列和多源先验信息的城市建筑物高分遥感提取. 遥感学报, 27(4): 998-1008) [DOI: 10.11834/jrs.20221077]

Mai G C, Lao N, He Y T, Song J M and Ermon S. 2023. CSP: self-supervised contrastive spatial pre-training for geospatial-visual representations. //International Conference on Machine Learning. PMLR, 2023: 23498-23515 [DOI: 10.48550/arXiv.2305.01118]

Mall U, Hariharan B and Bala K. 2023. Change-aware sampling and contrastive learning for satellite images//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE: 5261-5270 [DOI: 10.1109/CVPR52729.2023.00509]

Mañas O, Lacoste A, Giró-i-Nieto X, Vazquez D and Rodríguez P. 2021. Seasonal contrast: unsupervised pre-training from uncurated remote sensing data//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 9394-9403 [DOI: 10.1109/ICCV48922.2021.00928]

Mendieta M, Han B R, Shi X J, Zhu Y and Chen C. 2023. GFM: building geospatial foundation models via continual pretraining. arXiv preprint arXiv: 2302.04476

Muhtar D, Zhang X L, Xiao P F, Li Z S and Gu F. 2023. CMID: a unified self-supervised learning framework for remote sensing image understanding. IEEE Transactions on Geoscience and Remote Sensing, 61: 5607817 [DOI: 10.1109/TGRS.2023.3268232]

Pathak J, Subramanian S, Harrington P, Raja S, Chattopadhyay A, Mardani M, Kurth T, Hall D, Li Z Y, Azizzadenesheli K, Hassanzadeh P, Kashinath K and Anandkumar A. 2022. FourCastNet: a global data-driven high-resolution weather model using adaptive fourier neural operators. arXiv preprint arXiv: 2202.11214 [DOI: 10.48550/arXiv.2202.11214]

Patnala A, Stadtler S, Schultz M G and Gall J. 2023. Generating views using atmospheric correction for contrastive self-supervised learning of multispectral images. IEEE Geoscience and Remote Sensing Letters, 20: 2502305 [DOI: 10.1109/LGRS.2023.3274493]

Prexl J and Schmitt M. 2023. Multi-modal multi-objective contrastive learning for Sentinel-1/2 imagery//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver: IEEE: 2136-2144 [DOI: 10.1109/CVPRW59228.2023.00207]

Reed C J, Gupta R, Li S F, Brockman S, Funk C, Clipp B, Keutzer K, Candido S, Uyttendaele M and Darrell T. 2023. Scale-MAE: a scale-aware masked autoencoder for multiscale geospatial representation learning. arXiv preprint arXiv: 2212.14532 [DOI: 10.48550/arXiv.2212.14532]

Shi X J, Chen Z R, Wang H, Yeung D Y, Wong W K and Woo W C. 2015. Convolutional LSTM network: a machine learning approach for precipitation nowcasting//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press: 802-810

Stewart A J, Lehmann N, Corley I A, Wang Y, Chang Y C, Braham N A A, Sehgal S, Robinson C and Banerjee A. 2023. SSL4EO-L: datasets and foundation models for landsat imagery. arXiv preprint arXiv: 2306.09424 [DOI: 10.48550/arXiv.2306.09424]

Stojnić V and Risojević V. 2021. Self-supervised learning of remote sensing scene representations using contrastive multiview coding//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville: IEEE: 1182-1191 [DOI: 10.1109/CVPRW53098.2021.00129]

Sun X, Wang P J, Lu W X, Zhu Z C, Lu X N, He Q B, Li J X, Rong X E, Yang Z J, Chang H, He Q L, Yang G, Wang R P, Lu J W and Fu K. 2023. RingMo: a remote sensing foundation model with masked image modeling. IEEE Transactions on Geoscience and Remote Sensing, 61: 5612822 [DOI: 10.1109/TGRS.2022.3194732]

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

Tian Y L, Krishnan D and Isola P. 2020. Contrastive multiview coding//16th European Conference on Computer Vision. Glasgow: Springer: 776-794 [DOI: 10.1007/978-3-030-58621-8_45]

Tian Z Z, Zhang H W, Wang K, Liu S Q, Zou Q J, Zhao Z and Chen Y B. 2023. Application of an improved CenterNet in remote sensing images object detection. National Remote Sensing Bulletin, 27(12): 2706-2715 (田壮壮, 张恒伟, 王坤, 刘盛启, 邹前进, 赵镇, 陈育斌. 2023. 改进CenterNet在遥感图像目标检测中的应用. 遥感学报, 27(12): 2706-2715) [DOI: 10.11834/jrs.20231638]

Tseng G, Cartuyvels R, Zvonkov I, Purohit M, Rolnick D and Kerner H. 2024. Lightweight, pre-trained transformers for remote sensing timeseries. arXiv preprint arXiv: 2304.14065 [DOI: 10.48550/arXiv.2304.14065]

Vincenzi S, Porrello A, Buzzega P, Cipriano M, Fronte P, Cuccu R, Ippoliti C, Conte A and Calderara S. 2021. The color out of space: learning self-supervised representations for earth observation imagery//2020 25th International Conference on Pattern Recognition (ICPR). Milan: IEEE: 3034-3041 [DOI: 10.1109/ICPR48806.2021.9413112]

Wang D, Zhang Q M, Xu Y F, Zhang J, Du B, Tao D C and Zhang L P. 2022a. Advancing plain vision transformer toward remote sensing foundation model. IEEE Transactions on Geoscience and Remote Sensing, 61: 5607315 [DOI: 10.1109/TGRS.2022.3222818]

Wang W, Li X J and Wang X. ADC-CPANet:A Remote Sensing Image Classification Method Based on Local-Global Feature Fusion. National Remote Sensing Bulletin, (王威, 李希杰, 王新 . ADC-CPANet: 一种局部—全局特征融合的遥感图像分类方法 . 遥感学报) [DOI: 10.11834/jrs.20232658]

Wang Y B, Wu H X, Zhang J J, Gao Z F, Wang J M, Yu P S and Long M S. 2022b. PredRNN: a recurrent neural network for spatiotemporal predictive learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2): 2208-2225 [DOI: 10.1109/TPAMI.2022.3165153]

Wanyan X Y, Seneviratne S, Shen S C and Kirley M. 2023. DINO-MC: self-supervised contrastive learning for remote sensing imagery with multi-sized local crops. arXiv preprint arXiv: 2303.06670 [DOI: 10.48550/arXiv.2303.06670]

Ying C X, Cai T L, Luo S J, Zheng S X, Ke G L, He D, Shen Y M and Liu T Y. 2021. Do transformers really perform bad for graph representation?. arXiv:2106.05234 [DOI: 10.48550/arXiv.2106.05234]

Yuan Y and Lin L. 2021. Self-supervised pretraining of transformers for satellite image time series classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14: 474-487 [DOI: 10.1109/JSTARS.2020.3036602]

Yuan Y, Lin L, Liu Q S, Hang R L and Zhou Z G. 2022. SITS-Former: a pre-trained spatio-spectral-temporal representation model for Sentinel-2 time series classification. International Journal of Applied Earth Observation and Geoinformation, 106: 102651 [DOI: 10.1016/j.jag.2021.102651]

Zheng X C, Kellenberger B, Gong R, Hajnsek I and Tuia D. 2021. Self-supervised pretraining and controlled augmentation improve rare wildlife recognition in UAV images//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal: IEEE: 732-741 [DOI: 10.1109/ICCVW54120.2021.00087]

A comprehensive survey and assumption of remote sensing foundation modal

FU Kun[1,2,3], LU Wanxuan[1,2], LIU Xiaoyu[1,2], DENG Chubo[1,2], YU Hongfeng[1,2], SUN Xian[1,2]

1.Key Laboratory of Network Information System Technology (NIST), Chinese Academy of Sciences, Beijing 100190, China;
2.Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China;
3.University of Chinese Academy of Sciences, Beijing 100101, China

Abstract: In recent years, remote sensing intelligent interpretation technologies have advanced rapidly, but most established models are task oriented. Therefore, generalizing them to different tasks is difficult, and considerable amounts of resources are wasted. The foundation model is a straightforward approach that has recently attracted considerable interest in the field of remote sensing. Although many works have achieved remarkable results in some tasks for perception recognition and cognitive prediction by using remote sensing single-temporal or multitemporal data, a comprehensive review that provides a systematic overview of the remote sensing foundation model is lacking. Thus, this paper begins by summarizing developments in research on existing remote sensing foundation models from the perspectives of data, methods, and applications. Then, after analyzing the current situation's limits, we proposed a novel general predictive foundation model. Finally, some essential research areas were highlighted, and past achievements were linked with the future possibilities of remote sensing foundation model.

Existing remote sensing foundation models were categorized into three groups according to the data types used (single-temporal/multitemporal) and the tasks involved (perceptual recognition/cognitive prediction): the foundation model of perceptual recognition based on single-temporal data, the foundation model of perceptual recognition based on multitemporal data, and the foundation model of cognitive prediction based on multitemporal data. According to the different self-supervised learning methods adopted, we divided the existing foundation models of perceptual recognition based on single-temporal data into those based on contrastive learning and those based on generative learning. According to the number of tasks, the foundation model of perceptual recognition based on multitemporal data was divided into a single-task-oriented foundation model and a multitask-oriented foundation model. According to different model architectures, the cognitive prediction foundation models based on multitemporal data were divided into transformer-based and graph network-based foundation models. In accordance with the aforementioned categorization, we described the current state of each type of remote sensing foundation models and summarized their data, methods, and application restrictions.

Based on the summary and analysis of the existing remote sensing foundation models, a novel general predictive foundation model assumption was proposed. The information pipeline for multidomain or temporal data input and multitime or spatial scale task output can be opened up by extracting stable and generalized time-series hyper-pixel features. This approach enabled the accurate cognitive prediction of

the future state. Tens of millions of multiplatform, multitype, multimodal, and multitemporal data were included. By combining the benefits of the transformer model and the graph network, a new foundation model architecture was created, which increased the model's capacity and enhanced generalization while predicting multitarget interactions in large remote sensing scenes over the long term. In terms of application, the general predictive foundation model can be applied to diverse cognitive prediction tasks with multiple spatial and time scales. Under this assumption, we proposed four exploratory directions: multidomain time series data representation, stable feature extraction, object-environment interaction modeling, and multitask interaction reasoning, aiming to provide a reference for researchers exploring remote sensing foundation models.

In general, foundation models with generalization ability are crucial to development of remote sensing intelligent interpretation. We provided an overview of current advances in this field by collating the current state of research on remote sensing foundation models. By analyzing the limitations of current remote sensing foundation models in terms of data, methods, and applications, we proposed a novel general predictive foundation model assumption and further clarified four exploratory directions that urgently need breakthroughs under this idea. The follow-up work will make specific and important technological breakthroughs in multidomain time series data representation, stable feature extraction, object–environment interaction modeling, and multitask interaction reasoning. We explored a general remote sensing foundation model integrating perception recognition and cognitive prediction into a single architecture.

Key words: remote sensing intelligent interpretation, remote sensing foundation models, general prediction, multi temporal data, multi-task

Supported by National Natural Science Foundation of China (No. 62201550, 62171436); Key Deployment Program of the Chinese Academy of Sciences (No. KGFZD-145-23-18); National Key Research and Development Program of China (No. 2022ZD0118401)