引 用 格 式:ZHANG Shuaihao,PAN Zhigang.Remote sensing large models: Review and future prospects[J].Remote Sensing Technology and Application,2025,40(1):1-13.[张帅豪,潘志刚.遥感大模型:综述与未来设想[J].遥感技术与应用,2025,40(1):1-13.]
DOI:10.11873/j.issn.1004-0323.2025.1.0001

遥感大模型:综述与未来设想

张帅豪[1,2],潘志刚[1]
(1.中国科学院空天信息创新研究院,北京 100190;
2.中国科学院大学,北京 100049)

摘要:深度学习极大地推动了遥感图像处理技术的发展,在精度和速度方面展现了显著优势。然而,深度学习模型在实际应用中通常需要大量人工标注的训练样本,且其泛化性能相对较弱。近年来,视觉基础模型和大语言模型的发展为遥感图像处理的大模型研究引入了新的范式。遥感大模型也称为遥感基础模型,基础模型因其在下游任务中的卓越迁移性能而备受瞩目,这些模型首先在大型数据集上进行与具体任务无关的预训练,然后通过微调适应各种下游应用。基础模型在语言和视觉及其他领域已经得到了广泛应用,其在遥感领域的潜力也正逐渐引起学术界的重视。然而,目前针对这些模型在遥感任务中的全面调查和性能比较仍然缺乏。由于自然图像与遥感图像之间存在固有差异,这些差异限制了基础模型的直接应用。在此背景下,本文从多个角度对常见的基础模型以及专门针对遥感领域的大模型进行了全面回顾,概述了最新进展,突出了面临的挑战,并探讨了未来发展的潜在方向。

关 键 词:遥感基础模型;微调;下游任务;预训练
中图分类号:TP75 **文献标志码**:A **文章编号**:1004-0323(2025)01-0001-13

1 引 言

随着遥感对地观测和人工智能技术的迅猛发展,我们正进入一个遥感大数据与人工智能相互融合的时代。高分辨率遥感图像在资源勘探、环境监测、精准农业和军事侦察等领域的应用已变得越来越广泛[1,2]。在这些应用中,目标检测、语义分割、场景分类和变化检测等遥感视觉任务构成了这些研究任务的基本前提和核心支撑。深度神经网络(DNNs)凭借丰富的标注数据和强大的GPU计算能力,显著提升了遥感视觉任务的准确性。然而,DNNs在满足遥感应用需求方面仍存在诸多不足。一方面,DNNs的性能通常受到耗时费力的标注过程和单一任务限制的影响(例如,每个视觉识别任务通常需要单独训练一个DNN);另一方面,特定传感器影像的标注数据非常稀缺。近年来,基础模型(Foundation Models, FMs)在视觉和语言理解任务中展现出卓越的通用性,代表性实例包括ChatGPT[3]、Gemini[4]和CLIP[5]。FMs在视觉任务中展现出巨大潜力的原因主要包括以下3个方面:数据和模型规模、学习策略和适应性。具体而言,大规模数据和模型使FMs能够捕捉数据中的复杂信息,从而生成更具代表性的特征。在学习策略方面,FMs通过自监督或半监督学习,利用多模态数据(包括文本、图像、音频和视频)进行训练,减少了对人工标注的依赖,显著减轻了标注工作的负担。FMs通常可以通过增加少量与任务相关的特定数据或参数,适应或微调以满足各种下游任务或领域的需求[6]。

收稿日期:2024-07-21;**修订日期**:2024-12-30
基金项目:国家重点研发计划项目(2017YFB0503001)。
作者简介:张帅豪,硕士研究生,主要从事SAR图像处理算法及应用系统设计。E-mail: zhangshuaihao23@mails.ucas.ac.cn
通信作者:潘志刚,研究员,主要从事智能化SAR图像处理与信息分析系统研究。E-mail: zgpan@mail.ie.ac.cn

基础模型已成为遥感领域的研究热点，并在遥感视觉任务中展现了巨大的应用潜力。然而，遥感图像与自然图像在本质上存在显著差异，这限制了FMs在遥感图像中的直接应用[7]。因此，亟需开发专门针对遥感领域的基础模型，并需充分考虑跨领域的通用技术以及遥感图像的固有特征。在此背景下，本文旨在对常见的基础模型和领域特定的遥感基础模型进行全面综述。首先，回顾基础模型的基本原理和常用技术，接着探讨基础模型在遥感领域的特定技术，最后总结这些模型的局限性，并展望未来的研究方向。

2 基础模型介绍

2.1 基础模型原理与关键技术

基础模型是一种预训练的深度神经网络，作为各种下游任务的骨干网络，通常在庞大且多样化的数据集上进行训练，以捕捉通用的视觉特征。基础模型的一个显著优势在于其强大的表示学习能力，能够在不同任务之间很好地泛化。

基础模型的关键技术包括模型结构、学习算法和微调。图1展示了FMs在通用视觉任务和遥感图像应用中的流程图，描述了其实际的应用过程。作为特征学习的核心骨干网络，基础模型通常可通过微调或使用适配器技术来适应不同的领域或下游任务。除了传统的视觉基础模型和视觉-语言基础模型之外，近年来一些研究还专注于开发基于视觉提示的大规模基础模型，例如Segment Anything Model(SAM)[8]。为了全面介绍遥感图像应用在视觉基础模型方面的最新进展，本文重点探讨了视觉基础模型、视觉—语言基础模型。

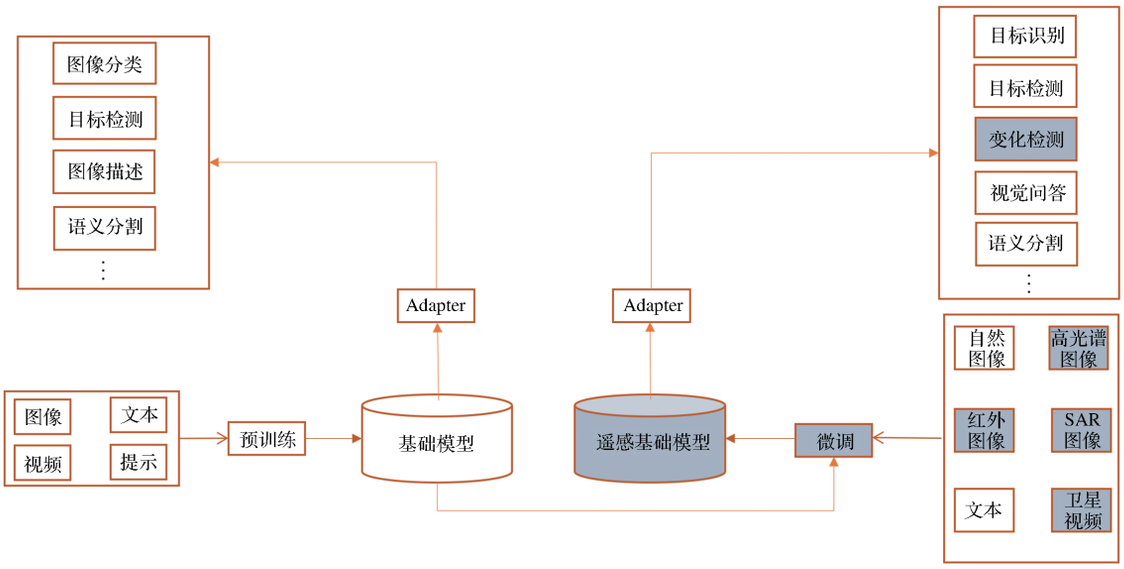

图1 遥感图像视觉任务的基础模型流程图

Fig.1 Flowchart of foundation models for vision tasks on remote sensing images

2.2 视觉基础模型

视觉模型的数据难以获取，不同的数据源有不同的结构，拥有同样规模和多样性的数据源非常困难。常见的视觉大模型包括将标准的Transformer结构直接应用于图像的基础视觉大模型，如ViT[9]、Swin-T[10]、CoAtNet[11]等，典型的视觉大模型如表2所示。

ViT结构如图2所示；对于ViT模型[9]，首先将图像分割成固定大小的patch，对每个patch进行线性嵌入，同时添加位置编码，然后将得到的向量序列送入Transformer Encoder，Transformer Encoder由多头自注意和MLP交替组成。在每个block之前进行应用层归一化，在每个区块之后应用残差连接。同时在序列中添加额外的可学习的嵌入向量。

Swin-T[10]针对Transformer在计算机视觉领域应用存在视觉实体的方差较大，图片分辨率高的两个问题，提出了一种基于滑动窗口机制、具有层级设计即下采样层的网络结构。通过结构的改进，Swin-T在视觉任务上取得了更优的性能，并成为了计算机视觉领域通用的骨干结构。

CoAtNet[11]针对Transformer具有更强的模型能力，但缺乏归纳偏置特性的现状，试图结合二者的长处。通过简单的相对注意力将深度卷积和自注意力自然统一，通过垂直摆放卷积层和注意力层

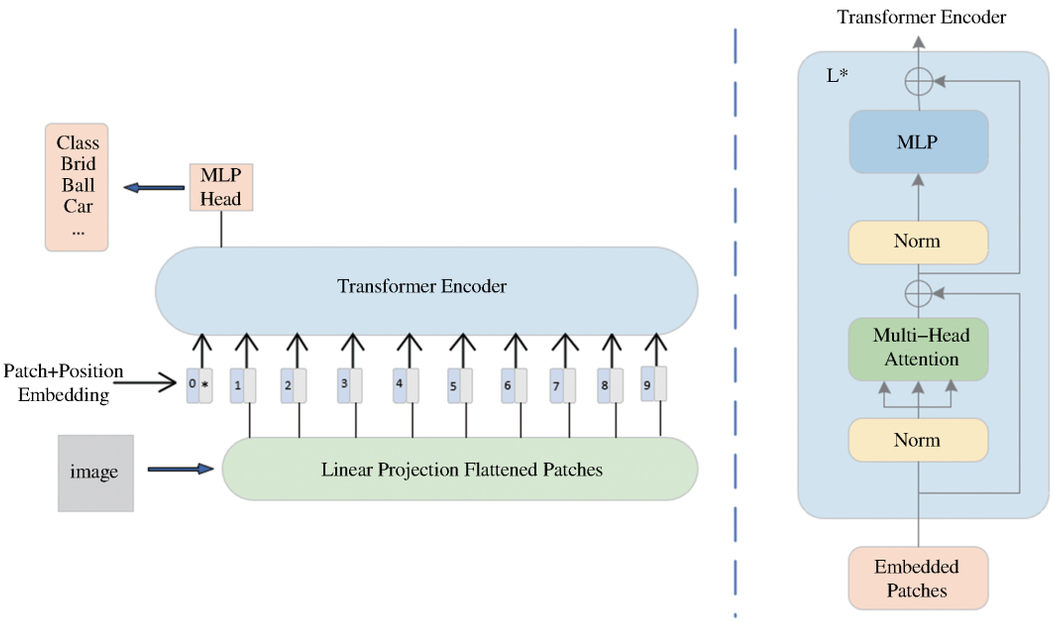

图 2　ViT 架构

Fig.2　ViT architecture

提升泛化性能和效率。CoAtNet解决了如何将卷积和自注意力结合在一个计算模块内,并将不同类别的计算模型垂直堆叠在一起,形成一个完整的网络的问题,模型在多个数据集上有最佳的表现。

2.3　视觉—语言基础模型

视觉—语言基础模型是视觉模型和自然语言模型的结合。视觉模型从图像中捕捉视觉特征,而语言模型从文本中获取编码信息。CLIP[5]模型被广泛应用视觉—语言基础模型当中,CLIP 模型如图3所示。给定N个图片—文件对,图片输入给Image Encoder得到特征 I_1、I_2、I_3...I_N,文本输入给 Text Encoder得到特征 T_1、T_2、T_3...T_N。定义(I_j,T_j)为正样本,其它属于副样本,通过训练使得最大化N个正样本的余弦相似度,最小化N个负样本的余弦相似度来做零样本迁移。通过充分学习多模态信息,视觉—语言大模型能够以“Zero-Shot”或“FewShot”的方式直接应用于图像—语言任务。常见的视觉—语言基础模型包括以下几种:CLIP[5]、mPLUG-owl[12]、ALBEF[13]、BLIP[14]、Flamingo[15]、LLaVA[16]、Kosmos1[17]等。其中,CLIP[5]通过Transformer表示文本特征,通过ResNets和ViT表示图片特征并在图片—文本对上训练,学习得到一个从视觉和语言输入到共同语义空间的映射,展现了强大的零样本泛化性能。ALBEF[13]首先使用对比学习对齐图像和文本,之后使用Transformer架构的变体融合特征,得益于交叉注意力机制的使用和视觉—文本对齐步骤,ALBEF适应于目标识别、图像描述等多种视觉—语言任务。BLIP[14]增加了文本编码器,实现了图像理解、多模态融合和语言生成,这使得BLIP特别适合执行图像描述任务。Flamingo[15]使用NormalizerFree ResNet学习视觉特征、Transformer解码器生成文本,通过交叉注意力机制实现深入的图文融合。Flamingo增加了感知器重采样模块,可以接受任意交错的图文数据作为输入,具有快速适应视觉问答等开放式任务的能力。LLaVA[16]利用CLIP视觉编码器和Vicuna大型语言模型,通过增强的多模态交叉注意力层,使模型能够在视觉和语言间进行深度信息交换,这种方式使其能够执行高级图像理解和语言生成任务,如图像描述、多模态问答和目标识别。Kosmos1[17]基于MAGNETO对齐感知和语言模型,融合多模态数据特征,利用xPos相对位置编码更好地进行长上下文建模,主要针对语言处理任务。表3总结了一些具有代表性的视觉—语言基础模型。

3　遥感基础模型

3.1　遥感基础模型应用框架

在基础模型中,预训练的权重通常包含了丰富的特征表示,可以适用于多个下游任务。然而,直接对整个模型进行微调可能会导致计算资源和训练时间的巨大开销。为了解决这个问题,目前主流的微调技术Adapter的方法被提出[18]。Adapter结

表 2 视觉基础模型总结

Table 2 Summary of Vision Foundation Models

模型	架构	参数量	训练数据	发布日期
ViT[9]	基于 Transformer	1.843 B	JFT-3B	2021/6/8
Swin-T[10]	基于 ViT 增加 shifted windows	3 B	ImageNet,CoCo	2021/11/8
CoAtNet[11]	基于 Transformer	24.4 B	ImageNet,JFT	2021/6/9
CoCa	结合了 contrastive loss 和 caption loss	21 B	JFT-3B, ALIGN	2022/6/14
DALL·E 3	它在 DALL·E2 的基础上进行了改进,提高了图像的保真度和质量	未公布	未公布	2023/10/19

表 3 视觉—语言基础模型总结

Table 3 Summary of Vision-Language Foundation Models

模型	架构	训练数据	发布日期
CLIP[5]	基于 ResNets、ViT 和 Transformer 图文双流对比预训练	WIT(WebImageText)	2021/2/26
ALBEF[13]	基于 BERT 和 ViT 图文双流先对齐后融合预训练,增加了动量蒸馏解决噪声问题	COCO 和 Visual Genome	2021/10/7
BLIP[14]	基于 BERT 和 ViT 多模态混合编码器-解码器,增加了 CapFilt 提高文本语料库质量	COCO、Visual Genome、Conceptual Captions	2022/2/15
Flamingo[15]	基于 NFNet 和 Transformer 图文交错多模态融合预训练,通过感知器重采样减少视觉标签	COCO、OKVQA、VQAv2、MSVDQA	2022/11/15
LLaVA[16]	基于 Vicuna、CLIP 视觉编码器进行特征对齐预训练,利用指令进行端到端的微调	ScienceQA、CC-595K、LLaVA-Instruct-158K	2023/12/11
Kosmos1[17]	基于 MAGNETO 对齐感知和语言模型,使用 xPos 相对位置编码更好地进行长上下文建模	The Pile、Common Crawl、LAION-2B、LAION-400M	2023/03/1

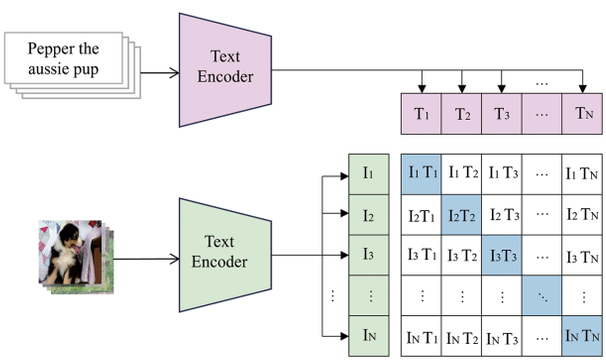

图 3 CLIP 预训练模型

Fig.3 CLIP pretrained model

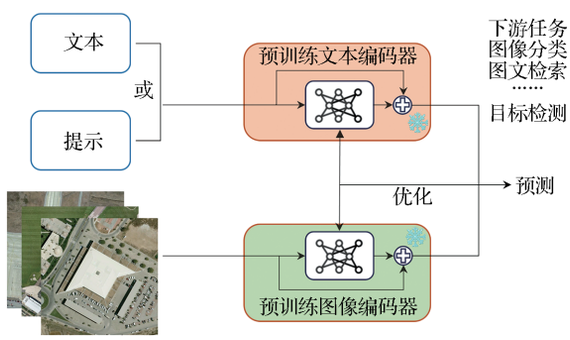

图 4 Adapter-tuning 结构图[19]

Fig.4 Adapter-Tuning architecture diagram

构如图 4 所示。对于不同的下游任务,如图像分类、图像描述、目标检测、变化检测、语义分割和文字生成等,Adapter 的方法可能有所不同。

在图像分类任务中,骨干网络选择一个在大规模图像分类任务上进行了预训练的模型作为基础模型。冻结其所有参数,即不对其进行任何调整。对于 Adapter 层,在模型顶部添加一个全连接层,输出类别预测。仅对 Adapter 层进行训练,通过微调适应特定的图像分类任务。

在目标检测任务中,骨干网络同样选取预训练的模型作为基础模型,并冻结其所有参数,在 Adapter 层替换或添加目标检测的头网络,如锚点机制或区域生成网络(Region Proposal Network, RPN),用于目标分类和边界框回归。仅对 Adapter 层以及与头网络相关的参数进行训练。可以使用目标检测任务专用的数据集进行微调。

在图像描述任务中,Adapter 的功能是将从骨干网络提取的视觉特征转换成适合序列模型处理的格式。通常,这涉及到特征向量的平展和嵌入,以便可以被如 LSTM 这类生成模型有效处理。Adapter 在这里桥接了视觉特征和语言输出的转换过程。

在变化检测任务中,Adapter 可能涉及对来自同一地点不同时间点的图像进行特征整合和比较。

它的主要任务是突出显示两个时间点之间的差异,帮助模型识别出变化的区域。这通常需要高度的特征匹配和融合技术。骨干网络选择一个适合用于语义特征提取的模型作为基础模型。Adapter层则添加卷积层和一个二元分类器,将特征图像映射到变化检测的二元标签,通过对Adapter层及相关参数进行训练,使模型适应变化检测任务的特定要求。

在语义分割任务中,骨干网络选择一个强大的图像语义特征提取模型作为基础模型,并冻结其所有参数,Adapter层通常处理来自骨干网络的特征图,在骨干网络的后面添加上采样或反卷积层,将特征图映射到与输入图像尺寸相同的语义分割掩码,训练Adapter层以及与语义分割相关的参数,以确保输出图像的像素级分类精度。例如,全卷积网络就直接在骨干网顶部增加上采样层来恢复图像原始尺寸,进行像素级分类。

在文字生成任务中,骨干网络选择一个强大的自然语言处理模型(如Transformer或循环神经网络)作为基础模型,并冻结其所有参数,Adapter层使用适合生成任务的模型结构,在骨干网络的输出上构建生成模型,用于生成文字序列,通过训练Adapter层和与生成模型相关的参数来微调模型,适应特定的文字生成任务。

3.2　大模型在遥感影像分析中的应用难点

理论上,由于其强大的零样本或少样本学习能力,上述预训练的基础模型可以直接应用于遥感图像。然而,考虑到自然图像和遥感图像之间的显著差异[20],在自然图像上预训练的基础模型需要经过仔细的微调和适应。具体而言,自然图像与遥感图像之间的差异主要体现在以下3个方面:

(1)遥感数据复杂性

遥感数据由于成像平台、成像机制以及成像条件的多样性,不同模态的图像如可见光、红外、高光谱以及SAR图像等在空间分辨率、光谱分辨率等方面存在显著的差异。这种高度的异质性对基础模型在遥感图像中的直接应用构成了挑战。例如,高光谱图像相较于传统的可见光图像,具有更多的通道,这就要求对模型进行额外的调整和优化,以适应其特有的数据特征。

(2)图像—文本稀缺性

大规模图像—文本对的可获得性对遥感基础模型的开发至关重要。然而,相较于自然图像领域,遥感领域发布的数据集在样本量上显得相对稀缺。此外,目前在自然图像领域关于深度理解和空间推理的描述,与遥感领域复杂的实际应用之间仍然存在显著差距。

(3)视觉任务难度

除了常见的挑战之外,遥感影像处理中的视觉任务与自然图像相比会更加复杂和具有挑战性。因为遥感影像具有一定的类间相似性与类内多样性的特点、不同传感器对应的同一目标图像会存在辐射特征、散射特征、光谱特征和几何特征的多样化,同时一张图中也会存在目标对象尺寸的巨大差异(例如飞机和机场)。

针对以上难点,研究人员通过开发适应遥感数据异质性的深度学习模型和改进数据生成方法,逐步解决了遥感图像处理中的关键挑战。这些努力为单模态预训练模型和视觉—文本联合训练模型的研究奠定了坚实基础。

3.3　单模态预训练模型

单模态预训练模型是指在遥感领域中未联合文本信息进行预训练的大型模型。单模态预训练模型的优势在于能够充分利用大规模遥感图像数据的信息,从而学习到更丰富、更强健的特征表达,提高模型对遥感图像的理解和处理能力。这种预训练方法还可以降低数据需求,因为预训练模型可以捕捉到遥感图像中的一般特征模态,从而在特定任务的训练样本较少时也能取得良好的效果。

3.3.1　有监督单模态预训练模型

有监督单模态预训练模型是在已知标签(即监督信息)的训练集上学习一个模型。通过该模型应用于新样本(或称为未知数据,即测试集中的样本),可以预测其对应的输出值。Fuller等[21]和Noman等[22]针对Transformer网络进行了有监督单模态预训练的研究。这两个研究都是为变化检测下游任务设计的新模型,并在单一任务上验证了其性能。Wang等[23]通过在大规模遥感(RS)数据集上训练3种骨干网络,包括传统的CNN、ViT模型和先进的ViTAE模型,研究了遥感数据预训练对模型的影响。他们进一步微调了上述模型,使用遥感或信息最大化的权重对4种下游任务进行初始化,包括图像分类、语义分割、目标检测和变化检测。共使用了9个数据集,并与其他方法进行了比较。研究结果显示,当特定下游任务所需的特征表示粒度接近上游预训练任务时,相比现有的SOTA方法,有监督单模态预训练模型ViTAE在几乎所有任务

上表现出最佳性能。

3.3.2 自监督单模预训练模型

自监督学习(SSL)是一种无监督学习形式,旨在从大量未标注数据中提取有用且可泛化的特征表示,以供下游任务使用[24-26]。自监督学习在训练FMs中起着至关重要的作用,许多最先进的基础模型在预训练阶段都使用了自监督学习。预训练使得基础模型能够获取丰富的特征和表示,之后再使用标注数据对特定的下游任务进行微调。遥感图像中的自监督学习是一种利用遥感数据中的未标记信息来学习有用特征的方法,目前已广泛应用于多光谱、超光谱和合成孔径雷达(SAR)图像。根据不同的前置任务方法,遥感图像的自监督学习方法可以分为3种:生成式、对比式和预测式[24, 27],如图5所示。

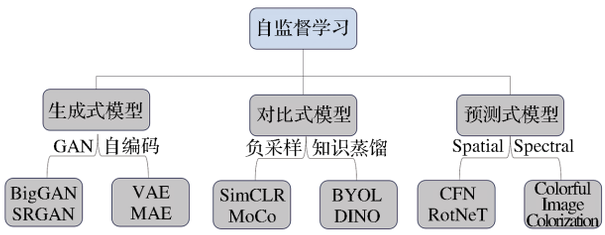

图5 自监督方法及经典模型

Fig.5 Summary of Self-Supervised Methods and Classical Models

遥感图像中的生成方法通常依赖于自编码器和生成对抗网络(GAN)等技术。然而,这些生成方法在遥感领域的应用不仅限于图像生成,还涵盖了一些特定的任务,例如城市洪水制图和高光谱解混。Peng等[28]提出了一个基于自监督学习的框架,用于通过多时相多光谱卫星图像进行块级城市洪水制图。该方法通过自监督中的自编码器学习,利用特征与块级变化向量分析相结合,生成能够突出潜在洪水影响区域的变化地图。AAENet等[29]针对无监督高光谱解混的新型网络,引入无绑定权重的自编码器、判别网络以及对抗过程,显著提升了模型的性能和鲁棒性。此外,这些生成方法还广泛用于图像生成和超分辨率任务中。例如,BigGAN[30]采用了大规模GAN架构,并引入了条件批量归一化、类别条件化和正交正则化等技术,使其能够生成高质量、多样化且类别特定的图像。而SRGAN[31]则专注于图像超分辨率任务,通过从大量低分辨率图像中学习恢复出高分辨率的细节纹理。这些技术的进展不仅提高了生成图像的质量,还扩展了遥感数据分析的能力。

遥感图像中的对比学习方法通过最大化相似样本的相似性和最小化不相似样本的相似性来学习数据的有效表示。其核心思想是通过对比目标函数,鼓励模型在特征空间中拉近相似样本,远离不相似样本。对比学习的经典方法可以分为4类:

(1)负采样:通过创建与锚样本不相似的负样本集来区分正负样本,学习有判别力的表示。典型模型包括SimCLR[32]和MoCo系列[33],使用卷积神经网络和动量更新策略,实现了在自监督学习任务中的优异表现。例如,MoCo[33]及其变体MoCov2[34]、MoCov3[35],通过对比学习表现出色。

(2)聚类:利用无监督的聚类算法将相似样本分组。代表性方法有DeepCluster[36]和SwAV[37],这些方法通过将图像的特征分组到簇中来鼓励相似表示的集合。

(3)知识蒸馏:通过教师模型向学生模型传递知识,学生模型在自监督学习中被训练。BYOL[38]和DINO[39]系列是该方法的典型例子,它们通过最大化不同视图的一致性来学习语义丰富的表示。

(4)冗余减少:减少学习表示中的冗余信息,确保表示捕捉到数据的基本特征。Barlow Twins模型[40]通过最大化互相关矩阵的非对角元素来学习有用的信息,同时最小化冗余。

遥感图像中的预测方法可以通过设计基于旋转预测的前置任务来增强对其他下游任务的支持。例如,Ji等[41]利用旋转预测识别输入图像的二维旋转,进而促进跨类别知识的迁移学习。他们将旋转预测与对比学习相结合,通过拉近正样本对并推开负样本对,增强了类内的一致性和类间的差异性。这些预训练任务与语义类别预测任务在端到端的框架中共同优化,最终提升了遥感图像场景分类的性能。此外,在文献[42]的研究中,旋转预测被用作预处理任务,以学习有效的特征表示,这些表示随后被转移至U-Net模型中,用于遥感城市场景的语义分割任务。上述方法各自通过不同的机制增强了模型的表示学习能力,在自监督学习任务中展现了强大的性能。

3.4 视觉—文本联合训练模型

视觉—文本基础模型可以利用其语言理解能力推断对象、属性及其周围环境之间的关系,从而从遥感图像中获取更具代表性的特征[43]。这些特性可应用于各种下游任务,如场景理解、目标检测和语义分割等,视觉—文本基础模型应用如图6所示。

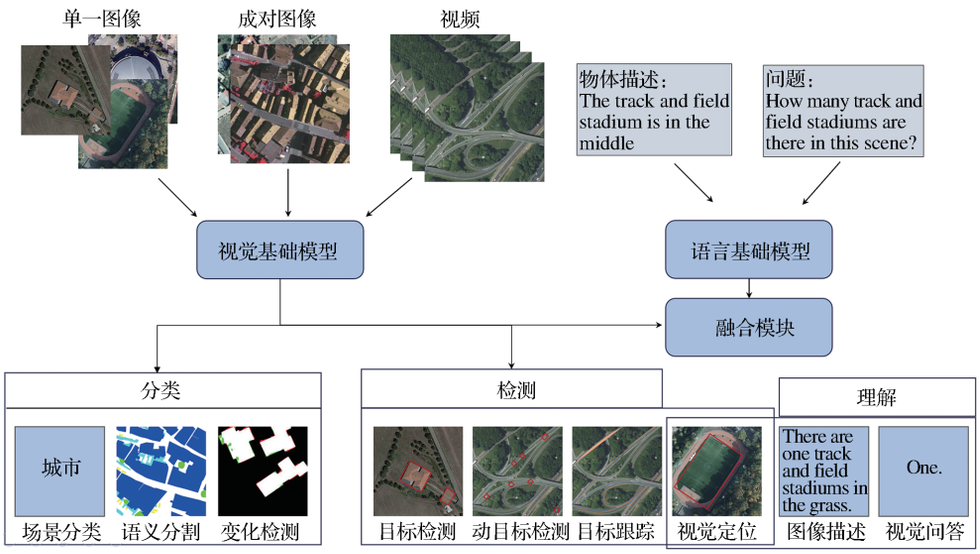

图 6　视觉—文本基础模型应用示意图

Fig.6　Diagram of Remote Sensing Foundation Model Applications

针对图像分类任务,Prompt-VQA[44]利用视觉模型提取上下文信息,将视觉模型上下文信息转化为文本特征并通过prompt与图像特征一同注入大模型,语言模型负责利用视觉上下文处理问题并提出有助于答案的特征。Mall等[45]使用真实影像作为遥感图像与文本对齐的媒介,使用大量配对的互联网和卫星图像训练遥感图像的图像编码器与CLIP编码器进行对齐。EarthPT[46]是一种用于地球观测的时间序列基础模型。该模型经过自监督训练,专为地球观测而设计,有效地预测了广泛光谱范围内像素级的表面反射率,并且其嵌入了有意义的信息,用于粒度动态土地利用分类等任务。SkySense[47]是一个基于精选的多模态遥感影像(RSI)数据集预训练的通用大规模模型,集成了一个分解的多模态时空编码器,该编码器以光学和SAR数据的时间序列为输入,通过多粒度对比学习进行预训练,学习不同模态和空间粒度上的表示。同时,引入了地理情境原型学习,以在RSI的多模态时空特征基础上学习区域感知的原型。受到计算机视觉和遥感领域基础模型的启发,FoMo-Bench[48]提出了统一的森林监测基准,由15个多样化的数据集组成,涵盖了卫星、航空和库存数据,涉及多种地理区域,包含了多光谱、可见光、合成孔径雷达和LiDAR数据,具有不同的时间、空间和光谱分辨率。

针对目标检测下游任务,Wei等[49]提出一种RSIC的视觉语言对齐范式,从而共同表示视觉和语言,是一种跨模态注意力视觉语言对齐模型,可根据遥感图像生成准确、丰富的图像双语描述。韩国科技公司NAVER提出了ViLT[50],即多模态对齐的视觉-语言Transformer模型。探索了不同模态之间的对齐方式,将视觉和语言特征更好地融合。GLIP[51]旨在学习对象级别的、语言感知的、语义丰富的视觉表征。它将对象检测和短语定位任务相结合,通过预训练统一这两种任务。GLIP可充分利用检测和定位数据以改善这两个任务的表现,并通过对齐文本和图像来引导良好的定位结果。RemoteCLIP[52]通过学习丰富的视觉特征来实现目标检测,并通过预训练数据扩展技术将异构注释转化为统一的图像—文本对;弥补了预训练数据稀缺的问题,为下游任务提供无缝的文本嵌入对齐。

针对语义分割任务,Text2Seg[53]使用文本提示来指导遥感图像的语义分割任务,结合多个预训练的视觉—语言基础模型及其它在视觉领域取得显著进展的模型,包括SAM[54],Grounding DINO[55]和CLIP[5]等。RemoteCLIP[52]利用数据扩展,将异构注释转换为基于Box-to-Caption(B2C)和Mask-to-Box(M2B)转换的统一图像-文本数据格式。B2C方法将边界框注释转换为一组自然语言标题,而M2B方法将分割数据集统一为边界框注释,然后再转换为标题数据集。B2C生成方法基于规则,根据目标位置生成5个不同的标题。通过识别掩模图像中每个类别的连接区域的轮廓点,确定边界框的坐标。所有语义分割注释都通过M2B转换为边界框注释,然后执行B2C以获取相应的标题。

针对文字生成和视觉问答任务,RSVQA[56]使

用CNN进行视觉特征提取和RNN进行文本特征提取，为了增强视觉和文本特征之间的对齐，同时设计了相互注意机制。与以前主要侧重于在封闭场景中解决遥感VQA的方法不同，VQA-TextRS[57]通过人工注释创建问题和答案，其数据集包含了各种各样的开放式问答对。利用视觉和语言转换网络从图像和伴随的问题中提取视觉和文本特征，然后通过一个具有交叉注意机制的transformer解码器来整合这两种模态。Bazi等[58]提出了类似的想法，使用CLIP模型将图像和问题嵌入作为视觉和文本表示，然后通过注意力机制学习这些表示之间的相关性。PromptRSVQA[59]使用CNN进行图像特征提取，并将其作为提示集成到BERT中，将图像中的上下文信息转换为语言模型可以处理的文本提示。LiT-4-RSVQA[60-61]是一种简化的基于transformer的架构，经优化用于遥感领域中高效且精确的VQA。Mikriukov等[62]在图像检索任务中使用跨模式的图像—文本检索方式，将检索信息与图像在预训练过程中进行匹配，从而实现快速和高效的遥感图像检索。RSGPT[63]旨在使用大语言模型解决遥感图像字幕和视觉问答任务，然而其缺乏进行多任务对话的能力；RSGPT需要在不同的数据集上训练任务专用模型以独立解决任务，这限制了其开放式任务的能力。MLLM[64]是一种专为遥感视觉语言理解设计的统一多模态大型语言模型，该模型能够处理各种遥感视觉语言任务，无需额外的编码模块。它采用了简单有效的架构，包括视觉编码器、对齐层和基于LLM的解码器；该模型使用两阶段调优方法来处理遥感视觉特征和特定任务的指令。

针对变化检测任务，CDVQA是一个多时相航空影像变化检测视觉问答任务的网络，采用ViT作为视觉编码器，RNN作为文本编码器，设计了一种基本视觉问答任务框架，包含多时相特征编码、多时相融合、多模态融合和答案预测4个模块。SSCFNet[65]是一种空间—光谱交叉融合网络，组合了特征提取网络模块、增强模块和语义交叉融合模块。在语义交叉融合模块中，对新构建的语义特征块进行不同的卷积运算，得到不同层次的语义特征并进行交叉融合。SSCFNet在4个公开的遥感图像变化检测数据集上进行了验证，并在定量和定性评估上取得了良好的结果。文献[66]将SAM基础模型的潜在知识整合到变化检测中，可以有效地解决一般知识转移中的领域偏移问题，以及表达多时相图像中均匀和异质特征的挑战。RSBuilding[67]旨在增强跨场景的泛化能力和任务的通用性，为了统一任务表示并整合图像的时空线索，引入了一个带有任务提示的交叉注意力解码器；针对当前缺乏同时包含两个任务注释的数据集的问题，开发了一种联合训练策略，即使某些任务缺乏监督也能促进模型的平滑收敛，从而增强不同任务的互补性。实验结果证实，RSBuilding展示了强大的零样本泛化能力。

表4 遥感大模型汇总

Table 4 Summary of large models in remote sensing

模型	训练策略	数据集	基础网络	数据集模态
RingMo[68]	MAE+PIMask	多个数据集	Vit	单模态
RSPrompter[69]	冻结+微调	WHU，NWPU，SSDD	sam	单模态
SpectralGPT[70]	MAE+3DMask	fMoW/BigEarthNetS2	Vit	单模态
RemoteCLIP[52]	MAE	10个数据集	Vit-14	多模态
GeoChat[71]	LoRA微调	318k个指令对的RS数据集	LLaVA1.5	多模态
RSGPT[63]	文本监督	图像+文本描述+指令	InstructBLIP	多模态
EarthPT[46]	自回归	ClearSky	Unsupervised Multitask Learners	多模态
DINO-MM[72]	蒸馏+对比	多个数据集	Self-supervised Multitask Learners	多模态
SkySense[47]	冻结+微调	16个不同任务的数据集	Vit	多模态

关于融合光学图像和SAR图像的遥感基础模型，DINO-MM[72]是一种联合SAR图像和光学图像的表示学习方法，该模型采用蒸馏学习方法，以SAR图像、光学图像或联合的SAR和光学图像作为对比学习的输入，能够从单个或两个模态中同时提取图像特征。GeoChat[71]是第一个通用遥感VLM，提供高分辨率遥感图像的多任务会话功能。EarthGPT[73]统一集成了多种遥感解译任务，实现了遥感图像的通用解译，构造了一种视觉增强感知机制，提炼和整合粗尺度语义感知信息和细尺度详细感知信息；提出了一种统一的遥感领域多传感器多任务指令调整方法，统一包括场景分类、图像字幕、

字幕生成、VQA、视觉接地、目标检测等多种任务;构建了一个具有大规模多传感器多模态遥感指令跟踪特性的数据集MMRS-1M,该数据集包括基于34个现有不同遥感数据集的超过1M的图像-文本对,并包括光学、SAR和红外等多传感器图像。SkySense[47]是一种全面的遥感图像分析模型,填补了现有遥感图像分析模型在单一模态、静态输入和地理上下文等方面的空白。SkySense是迄今为止最大的MM-RSFM,采用模块化设计,能够处理从单模态到多模态、从静态到动态、从分类到本地化等多种任务。

视觉—语言基础模型在遥感中的主要挑战是需要高质量和描述性的图像数据。RemoteCLIP[52]和SkyCLIP[74]采用了创造性但不同的方法来解决这个问题。RemoteCLIP将现有的目标检测和语义分割标签转换为自然语言描述。例如,从一个飞机检测数据集中,包含中心附近的边界框的图像可以生成“图片中间有一架飞机”的描述。相比之下,SkyCLIP从OpenStreetMap的语义标签生成文本描述。

4　遥感大模型的挑战与未来展望

4.1　遥感大模型的挑战

(1)数据匮乏。由于大规模图像和文本对齐数据的应用,基础模型在视觉语言方面的能力得到了显著提升,增强了其理解和分析能力。然而,在遥感领域,由于遥感数据的专业性和封闭性,研究人员收集大量图像和文本配对数据并将其用于训练仍是一项挑战。

(2)可靠性。目前,基础模型仍存在诸多未解问题,如模型的不可靠性和理解能力不足[75]。以CLIP类型模型为例,尽管它们是使用最广泛的视觉模型之一,但在处理简单问题时也会出现失败现象。这是因为预训练的CLIP视觉编码器往往忽略了关键的视觉细节,无法有效地排序重要的视觉模式。此外,视觉上的不确定性放大了这些问题,导致更多的幻觉和错误识别。

(3)时空数据不一致。遥感数据种类繁多,不同的传感器提供了具有不同时间和空间参考以及不同格式的数据。由于成像机制的差异,不同传感器获取的图像在空间分辨率、光谱分辨率、辐射分辨率等方面存在较大差异;这种多样性使得收集和对齐高质量的多模态样本数据变得具有挑战性,从而限制了多模态RSFMs的发展。

(4)评估难度。对不同的基础模型进行公平统一的基准评估至关重要。然而,公平的模型评估却很困难。一方面,评估通常受到许多其他因素的影响。例如,仅仅在预训练中遵循图像调整和归一化的技巧,就可以在下游任务中显著提升性能[76],如在So2Sat随机拆分数据集上提高了32.28%的整体准确率,在EuroSAT数据集上提高了11.16%。在这种情况下,进行可靠的跨领域评估非常困难。

4.2　遥感大模型未来展望

视觉基础模型忽视了对对象及其邻域关系的语义理解。相比之下,视觉—语言基础模型可以利用它们的语言理解能力来推断对象、属性及其周围环境之间的关系,从而从遥感图像中获得更具代表性的特征。遥感大模型未来有以下5个发展趋势:

(1)持续发展大规模数据集:基于人工智能的遥感数据处理系统的准确性在很大程度上取决于训练数据集的规模和多样性。然而,在遥感领域中,现有最大的数据集,例如Million AID,与计算机视觉中使用的包含数十亿张图像的数据集相比仍然规模较小。为了满足对更大规模、更丰富数据集的需求,必须共同努力构建数据收集和共享机制。一个潜在的数据集是开源的LAION-5B[77],其中包含超过50亿个图像—文本对。因此,研究界需要合作构建更多样化的数据集,并结合语言描述,以促进大规模训练大型遥感模型。

(2)使用基于文本的图像生成的扩散模型:扩散模型已经在图像增强、图像生成、风格转换等任务上取得了惊人的效果[78]。然而,传统的神经网络的收敛通常需要大量数据进行训练,而数据的收集需要大量的人力和物力投入。相比之下,扩散模型能够生成更加详细、精细的高质量图像。通过在已有的文本描述基础上使用扩散模型生成新的图像,可以创建合成数据并能有效地扩展数据集的大小,以提高深度学习模型的稳健性和泛化能力。此外,结合风格转移或领域自适应等技术,可以生成更多样化、代表真实世界场景的合成图像,从而进一步丰富数据集并增强扩散模型的性能。

(3)开发特定领域的模型:多模态数据融合在遥感中起着至关重要的作用,未来的遥感大模型将更加关注不同数据模态的融合;包括可见光、激光雷达、热红外、SAR等多种模态的数据,使模型能够获得更全面、多维度的地物信息。传统模型已经针

对特定传感器或数据类型进行了定制，每种模型都有其独特的特点。这种专业化阻碍了利用多样化数据源的结合优势进行整体分析的潜力。因此，从自然图像中得出的表示并不总是与遥感图像的独特特性兼容，凸显了现有表示的局限性。在这种情况下，开发专门针对遥感领域的基础模型是必要的。

(4)评估策略和基准提升：评估挑战需要对更好的评估方法进行深入研究，以构建更稳健的基础模型。一种潜在的解决方案是开发基础模型代理，将性能评估建模为在不同任务和配置间传递的动态过程。通过这一方式，采用累计奖励而非静态准确率作为评估基准，可以在应对各种影响因素和预处理方法时提供更为可靠和稳健的评估结果。

(5)联合学习和迁移学习的应用：联合学习和迁移学习指在多个相关任务之间共享和转移学习知识。未来的遥感大模型将不仅仅关注单个任务，而是通过联合学习和迁移学习，将不同任务的知识紧密结合，提高模型的泛化能力和应用范围。未来的遥感大模型将会在数据集规模、扩散模型、多模态数据融合、解释性和可解释性以及联合学习和迁移学习等方面取得进一步发展。这些发展趋势将使遥感大模型在地球观测、资源管理、环境监测等领域发挥更大的作用，并提供更准确、高效的遥感数据分析和应用。

5 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。然而，遥感图像和自然图像之间的本质差异使得现有的基于自然图像的基础模型无法直接应用于遥感领域。本文通过整理基于基础模型的常用技术，为研究人员提供该领域的最新进展综述。在此基础上系统的总结、分析现有遥感基础模型的优缺点。最后针对当下遥感基础模型面临的挑战提出了未来的发展方向。

参考文献(References)：

[1] ZHANG L, ZHANG L. Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities [J]. IEEE Geoscience and Remote Sensing Magazine, 2022, 10(2): 270-294. DOI:10.1016/j.cageo.2022.105034

[2] SUN Z, SANDOVAL L, Crystal-Ornelas R, *et al*. A review of earth artificial intelligence[J]. Computers Geosciences,2022,159:105034. DOI:10.1016/j.cageo. 2022.105034

[3] ZHOU C,LI Q,LI C ,*et al*. A comprehensive survey on pretrained foundation models: A History from BERT to ChatGPT[J].arXiv Preprint 2023.DOI:10.48550/arXiv.2302.09419

[4] TEAM G,ANIL R,BORGEAUD S, *et al*. Gemini:A family of highly capable multimodal models[J].arXiv Preprint 2023. DOI:10.48550/ arXiv.2312.11805

[5] RADFORD A, KIM J W, HALLACY C, *et al*. Learning transferable visual models from natural language supervision [C]//International Conference on Machine Learning. PMLR, 2021: 8748-8763.

[6] BOMMASANI R, HUDSON D A, ADELI E, *et al*. On the opportunities and risks of foundation models[J].arXiv Preprint 2021. DOI:10.48550/arXiv.2108.07258

[7] ROLF E, KLEMMER K, ROBINSON C, *et al*. Mission critical-satellite data is a distinct modality in machine learning [J].arXiv Preprint 2024. DOI:10.48550/arXiv. 2402. 01444

[8] KIRILLOV A, MINTUN E, RAVI N, *et al*. Segment anything[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 4015-4026.

[9] DOSOVITSKIY A. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv Preprint 2020. DOI:10.48550/ arXiv.2010.11929

[10] LIU Z, LIN Y, CAO Y, *et al*. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[11] DAI Z, LIU H, LE Q V, *et al*. Coatnet: Marrying convolution and attention for all data sizes[J]. Advances in Neural Information Processing Systems, 2021, 34: 3965-3977.

[12] YE Q, XU H, XU G, *et al*. mplug-owl: Modularization empowers large language models with multimodality[J]. arXiv Preprint,2023. DOI:10.48550/arXiv.2304.14178

[13] LI J, SELVARAJU R, GOTMARE A ,*et al*. Align before fuse: Vision and language representation learning with momentum distillation[J]. Advances in Neural Information Processing Systems, 2021, 34: 9694-9705.

[14] LI J, LI D, XIONG C ,*et al*. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation [C] //International Conference on Machine Learning. PMLR, 2022: 12888-12900.

[15] ALAYRAC J-B,DONAHUE J,LUC P,*et al*.Flamingo:A visual language model for few-shot learning[J].Advances in Neural Information Processing Systems,2022,35:23716-23736.

[16] LIU H,LI C,WU Q,*et al*. Visual instruction tuning[J].arXiv Preprint,2023. DOI:10.48550/arXiv.2304.08485

[17] PENG Z,WANG W,DONG L ,*et al*. Kosmos-2: Grounding multimodal large language models to the world[J]. arXiv Preprint,2023. DOI:10.48550/arXiv.2306.14824

[18] REBUFFI S, BILEN H,VEDALDI A.Learning multiple visual domains with residual adapters[J]. arXiv Preprint,2017. DOI:10.48550/arXiv.1705.08045

[19] XING J,LIU J,WANG J ,*et al*. A survey of efficient fine-tuning methods for Vision-Language Models — Prompt and Adapter[J].Computers & Graphics,2024,119:103885. DOI:

10.1016/j.cag.2024.01.012

[20] ROLF E,KLEMMER K,ROBINSON C ,*et al*. Mission critical-Satellite data is a distinct modality in machine learning[J]. arXiv Preprint,2024. DOI:10.48550/arXiv.2402.01444

[21] FULLER A, MILLARD K, GREEN JRJAPA. Transfer learning with pretrained remote sensing transformers[J].arXiv Preprint 2022. DOI:10.48550/arXiv.2209.14969

[22] NOMAN M,FIAZ M,CHOLAKKAL H,*et al*. Remote sensing change detection with transformers trained from scratch[J]. IEEE Transactions on Geoscience and Remote Sensing,2024, 62:1-14, DOI:10.1109/TGRS. 2024.3383800

[23] WANG D,ZHANG J,DU B ,*et al*. An empirical study of remote sensing pretraining[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-20. DOI: 10.1109/TGRS.2022.3176603

[24] JING L, TIAN Y. Self-Supervised Visual Feature Learning With Deep Neural Networks:A survey[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2020,43(11): 4037-4058.

[25] LIU X,ZHANG F,HOU Z ,*et al*. Self-supervised learning: Generative or contrastive[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(1): 857-876.

[26] STOJNIC V,RISOJEVIC V.Self-supervised learning of remote sensing scene representations using contrastive multiview coding[C]//In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 1182-1191.

[27] WANG Y,ALBRECHT C M,BRAHAM N A A,*et al*. Self-supervised learning in remote sensing:A review[J].IEEE Geoscience and Remote Sensing Magazine,2022,10(4):213-247.

[28] PENG B, HUANG Q, VONGKUSOLKIT J, *et al*. Urban flood mapping with bitemporal multispectral imagery via a self-supervised learning framework[J].IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021,14:2001-2016.DOI:10.1109/JSTARS. 2020. 3047677

[29] JIN Q,MA Y,FAN F,*et al*.Adversarial autoencoder network for hyperspectral unmixing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 34(8): 4555-4569.

[30] BROCK A,DONAHUE J,SIMONYAN K J A E-P . Large scale GAN training for high fidelity natural image synthesis[J]. arXiv preprint,2018. DOI:10.48550/arXiv.1809.11096

[31] LEDIG C,THEIS L,HUSZAR F, *et al*. Photo-realistic single image super-resolution using a generative adversarial network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4681-4690.

[32] CHEN T, KORNBLITH S, NOROUZI M, *et al*. A simple framework for contrastive learning of visual representations [C]//International Conference on Machine Learning. PMLR, 2020:1597-1607.

[33] HE K,FAN H,WU Y ,*et al*. Momentum contrast for unsupervised Visual Representation Learning [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9729-9738.

[34] CHEN X,FAN H,GIRSHICK R ,*et al*. Improved baselines with momentum contrastive learning[J].arXiv Preprint,2020. DOI:10.48550/ arXiv.2003.04297

[35] CHEN X,XIE S, HE K J A E-P. An empirical study of training self-supervised visual transformers[J].arXiv Preprint,2021. DOI:10.48550/ arXiv.2104.02057

[36] CARON M,BOJANOWSKI P,JOULIN A,*et al*. Deep clustering for unsupervised learning of visual features [J]. arXiv Preprint,2018. DOI:10.48550/arXiv.1807.05520

[37] CARON M, MISRA I, MAIRAL J , *et al*. Unsupervised learning of visual features by contrasting cluster assignments [J]. arXiv Preprint,2020. DOI:10.48550/arXiv.2006.09882

[38] GRILL J-B ,STRUB F,ALTCHé F ,*et al*. Bootstrap Your Own Latent: A new approach to self-supervised learning[J]. arXiv Preprint,2020. DOI:10.48550/arXiv.2006.07733

[39] CARON M,TOUVRON H,MISRA I ,*et al*. Emerging properties in Self-Supervised Vision Transformers[J]. arXiv Preprint,2021. DOI:10.48550/arXiv.2104.14294

[40] LONG Y,XIA G S,LI S, *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, DOI:10.1109/JSTARS.2021.3070368

[41] JI H,GAO Z,ZHANG Y ,*et al*. Few-Shot scene classification of optical remote sensing images leveraging calibrated pretext tasks[J]. IEEE Transactions on Geoscience and Remote Sensing,2022,60:1-13. DOI:10.1109/TGRS.2022.3184080

[42] GONZáLEZ-SANTIAGO J, SCHENKEL F MIDDEL-MANN W. Self-supervised image colorization for semantic segmentation of urban land cover[C]//2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS.IEEE,2021: 3468-3471.

[43] LI X, WEN C, HU Y ,*et al*. Vision-language models in remote sensing: Current progress and future trends [J]. IEEE Geoscience and Remote Sensing Magazine, 2024, 12: 32-66. DOI:10.1109/ MGRS.2024.3383473

[44] CHAPPUIS C,ZERMATTEN V,LOBRY S,*et al*. Prompt-RSVQA: Prompting visual context to a language model for remote sensing visual question answering[J].IEEE/CVF Conference on Computer Vision and Pattern Recognition,2022,19 (7):1372-1381.

[45] MALL U,PHOO C P,LIU M K,*et al*. Remote sensing vision-language foundation models without annotations via ground remote alignment[J]. arXiv Preprint, 2023. DOI: 10.48550/arXiv.2312.06960

[46] SMITH M J,FLEMING LGEACH JEJAE P. EarthPT:A time series foundation model for earth observation [J]. arXiv Preprint,2023. DOI:10.48550/arXiv.2309.07207

[47] GUO X,LAO J,DANG B ,*et al*. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery[J].IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, 16(7):

27672-27683.

[48] BOUNTOS N, OUAKNINE A, ROLNICK D. Fomo-bench: A multi-modal, multi-scale and multitask forest monitoring benchmark for remote sensing foundation models[J]. arXiv Preprint, 2023. DOI: 10.48550/ arXiv.2312.10114

[49] WEI T, YUAN W, LUO J, *et al*. Vlca: vision-language aligning model with cross-modal attention for bilingual remote sensing image captioning[J]. Journal of Systems Engineering and Electronics, 2023, 34: 9-18, DOI: 10.23919/JSEE.2023.000035

[50] KIM W, SON B, KIM I. Vilt: Visionand-language transformer without convolution or region supervision[J]. arXiv Preprint, 2021. DOI: 10.48550/ arXiv.2102.03334

[51] LI LH, ZHANG P, ZHANG H, *et al*. Grounded languageimage pre-training[J]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 18(7): 10965-10975.

[52] LIU F, CHEN D, GUAN Z, *et al*. Remoteclip: A vision language foundation model for remote sensing[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-16, DOI: 10.1109/TGRS.2024.3390838

[53] ZHANG J, ZHOU Z, MAI G, *et al*. Text2seg: Remote sensing image semantic segmentation via text-guided visual foundation models[J]. arXiv Preprint, 2023. DOI: 10.48550/arXiv.2304.10597

[54] MAZUROWSKI M, DONG H, GU H, *et al*. Segment anything model for medical image analysis: An experimental study [J]. Medical Image Analysis, 2023, 89: 102918. DOI: 10.1016/ j.media.2023.102918

[55] LIU S, ZENG Z, REN T, *et al*. Grounding dino: Marrying dino with grounded pre-training for open-set object detection [C]//European Conference on Computer Vision. Springer, Cham, 2025: 38-55.

[56] 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.

[57] AL RAHHAL M M, BAZI Y, ALSALEH S O, *et al*. Open-ended remote sensing visual question answering with transformers[J]. International Journal of Remote Sensing, 2022, 43(18): 6809-6823.

[58] BAZI Y, AL RAHHAL M M, MEKHALFI M L, *et al*. Bimodal transformer-based approach for visual question answering in remote sensing imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-11.

[59] CHRISTEL C, VALERIE Z, SYLVAIN L, *et al*. Prompt-RSVQA: Prompting visual context to a language model for remote sensing visual question answering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 1372-1381.

[60] YUAN Z, MOU L, WANG Q, *et al*. From easy to hard: Learning language-guided curriculum for visual question answering on remote sensing data[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-11.

[61] LOBRY S, TUIA D. Visual question answering on remote sensing images[M]. Advances in Machine Learning and Image Analysis for GeoAI. Elsevier, 2024: 237-254.

[62] MIKRIUKOV G, RAVANBAKHSH M, DEMIR BJAPA. Deep unsupervised contrastive hashing for largescale cross-modal text-image retrieval in remote sensing[J]. arXiv Preprint, 2022. DOI: 10.48550/ arXiv.2201.08125

[63] HU Y, YUAN J, WEN C, *et al*. Rsgpt: A remote sensing vision language model and benchmark[J]. arXiv Preprint, 2023. DOI: 10.48550/arXiv.230 7.15266

[64] ZHAN Y, XIONG Z, YUAN YJAPA. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model[J]. arXiv Preprint, 2024. DOI: 10.48550/arXiv.2401.09712

[65] WANG J, LIU F, JIAO L, *et al*. Sscfnet: A spatialspectral cross fusion network for remote sensing change detection[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16: 4000-4012.

[66] LI K, CAO X, MENG DJITOG, *et al*. A new learning paradigm for foundation model-based remote-sensing change detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-12.

[67] WANG M, CHEN K, SU L, *et al*. RSBuilding: Towards general remote sensing image building extraction and change detection with Foundation Model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-17. DOI: 10.1109/ TGRS.2024.3439395

[68] SUN X, WANG P, LU W, *et al*. RingMo: A remote sensing foundation model with masked image modeling [J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-22, DOI: 10.1109/TGRS.20 22.3194732

[69] CHEN K, LIU C, CHEN H, *et al*. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model[J]. 2024, 62: 3356074. DOI: 10.1109/ tgrs.2024.3356074

[70] HONG D, ZHANG B, LI X, *et al*. SpectralGPT: Spectral remote sensing foundation model[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46: 5227-5244, DOI: 10.1109/TPAMI.2024. 3362475

[71] KUCKREJA K, DANISH M, NASEER M, *et al*. Geochat: Grounded large vision-language model for remote sensing[J]. arXiv Preprint, 2023. DOI: 10.48550/arXiv.2311.15826

[72] WANG Y, ALBRECHT C, ZHU X. Self-supervised vision transformers for joint SAR-optical representation learning[J]. arXiv Preprint, 2022. DOI: 10.48550/arXiv.2204.05381

[73] ZHANG W, CAI M, ZHANG T, *et al*. EarthGPT: A universal multimodal large language model for multisensor image comprehension in remote sensing domain[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-20, DOI: 10.1109/ TGRS.2024.3409624

[74] WANG Z, PRABHA R, HUANG T, *et al*. SkyScript: A large and semantically diverse vision-language dataset for remote

sensing[J].arXiv Preprint,2023. DOI:10.48550/arXiv. 2312.12856.

[75] LEVKOVICH IELYOSEPH Z.Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: Vignette study [J]. JMIR Mental Health, 2023, 10: e51232. DOI: 10.2196/51232

[76] CORLEY I, ROBINSON C, DODHIA R , *et al*. Revisiting pre-trained remote sensing model benchmarks: Resizing and normalization matters[J]. arXiv Preprint, 2023. DOI: 10.48550/arXiv.2305.13456

[77] DONG R, YUAN S, LUO B , *et al*. Building bridges across spatial and temporal resolutions: Reference-based super-resolution via change priors and conditional diffusion model [J]. arXiv Preprint, 2024. DOI: 10.48550/arXiv.2403.17460.

[78] ROMBACH R, BLATTMANN A, LORENZ D, *et al*. High-resolution image synthesis with latent diffusion models[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 10684-10695.

Remote Sensing Large Models: Review and Future Prospects

ZHANG Shuaihao[1,2], PAN Zhigang[1]

(1.*Aerospace Information Research Institute*, *Chinese Academy of Sciences*, *Beijing* 100190, *China*;
2.*University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

Abstract: Deep learning has significantly advanced remote sensing image processing technology, demonstrating notable improvements in both accuracy and speed. However, deep learning models typically require large amounts of manually labeled training samples in practical applications, and their generalization performance is relatively weak. In recent years, the development of visual foundation models and large language models has introduced a new paradigm for research on large models in remote sensing image processing. Remote sensing large models, also known as remote sensing foundation models, have garnered attention for their outstanding transfer performance in downstream tasks. These models are first pretrained on large datasets unrelated to specific tasks and are then fine-tuned to adapt to various downstream applications. Foundation models have already been widely applied in language, vision, and other fields, and their potential in the field of remote sensing is increasingly gaining attention from the academic community. However, there is still a lack of comprehensive surveys and performance comparisons of these models in remote sensing tasks. Due to the inherent differences between natural images and remote sensing images, these differences limit the direct application of foundation models. Against this backdrop, this paper provides a comprehensive review of common foundation models and large models specifically designed for the field of remote sensing from multiple perspectives. It outlines the latest advancements, highlights the challenges faced, and explores potential future directions for development.

Key words: Remote Sensing Foundation Model; Fine-Tuning; Downstream Tasks; Pretraining