doi:10.11887/j.issn.1001-2486.25050009 http://journal.nudt.edu.cn

雷达遥感基础模型研究:进展与展望

王欣超,陈思伟*
(国防科技大学 电子科学学院,湖南 长沙 410073)

摘 要:基础模型因其提供了一种通用、可泛化的解决方案,成为雷达遥感智能解译领域的关注重点。目前雷达遥感基础模型在理论与应用层面均取得了重要进展,及时总结现有研究进展具有重要意义。为了进一步推进雷达遥感基础模型相关科学问题的研究进展,阐述了基础模型的概念、关键技术和评价方法;在此基础上,介绍了目前雷达遥感基础模型的研究现状和应用效果,对典型的方法和基础模型实例进行了梳理和总结。最后从模型架构设计、可解释性研究、轻量化方法和安全性评估四个方面进行了讨论和展望。

关键词:基础模型;雷达遥感;自监督学习;知识迁移
中图分类号:TN957.52 **文献标志码**:A **文章编号**:1001-2486(2026)02-382-14

Research on foundation models for radar remote sensing: progress and prospects

*WANG Xinchao, CHEN Siwei**
(College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China)

Abstract: Foundation models have become a focus in radar remote sensing intelligent interpretation due to their provision of universal and generalizable solutions. Significant progress has been achieved in both theoretical and applied aspects of radar remote sensing foundation models, making it imperative to systematically summarize current research advancements. In order to further advance the research on radar remote sensing foundation models, the concept, key technologies, and evaluation methods of foundation models were expounded. Besides, current research progress and application performance were reviewed, with representative approaches and typical instances summarized. In conclusion, discussions and future directions were highlighted from four perspectives: model architecture design, interpretability research, lightweight methods, and security assessment.

Keywords: foundation model; radar remote sensing; self-supervised learning; knowledge transfer

雷达遥感对地观测技术飞速发展,在军事侦察、灾害评估、农林监测、海洋监视等关键领域获得了前所未有的广泛应用[1-3]。当前雷达遥感进入大数据时代,大规模、多样化、高分辨的雷达遥感影像数据为雷达遥感解译带来新的机遇和挑战。以深度学习为代表的人工智能技术应用,使目标检测、场景分类、语义分割、变化检测等解译任务的精度显著超越传统方法[4-6]。然而,面对日益激增的数据和更为复杂的解译任务,传统深度学习解译模型普遍面临理解和感知能力不足、泛化应用困难等问题。因此,构建通用且高效的雷达遥感解译模型是亟待攻克的难题。

近年来,不同类型的基础模型(如视觉、语言、多模态)展现出了卓越的通用性和理解感知能力,为雷达遥感智能解译带来了新的思路[7]。雷达遥感基础模型旨在构建一个可用于多种解译任务的通用模型,进一步通过知识迁移来提高下游解译任务的性能。如图1所示,研究人员构建了一系列的遥感解译基础模型。但在雷达遥感领域的研究中仍然面临众多的技术挑战。不同于自然图像可通过已有的多模态基础模型实现自动化标注,构建大规模雷达遥感图像数据集过程(预处理、数据清洗、标注)需要领域专家知识干预;由于成像机制及物理特性的差异,现有的预训练

收稿日期:2025-05-08
基金项目:国家自然科学基金资助项目(U24B20189,62122091);湖南省科技创新领军人才资助项目(2024RC1040)
第一作者:王欣超(1995—),男,新疆伊犁人,博士研究生,E-mail:wangxinchaol@163.com
***通信作者**:陈思伟(1984—),男,四川泸州人,教授,博士,博士生导师,E-mail:chenswnudt@163.com
引用格式:王欣超, 陈思伟. 雷达遥感基础模型研究:进展与展望[J]. 国防科技大学学报, 2026, 48(2): 382-395.
Citation: WANG X C, CHEN S W. Research on foundation models for radar remote sensing: progress and prospects[J]. Journal of National University of Defense Technology, 2026, 48(2): 385-395.

框架无法直接迁移至雷达遥感图像中,给模型的训练带来了困难。雷达遥感解译旨在为经济、军事侦察、灾害响应等重要领域提供精确可靠的态势感知,当前研究中缺乏有效的评估基准来对基础模型及其生成内容的可靠性进行评价。

当前,基础模型在遥感解译中得到了快速发展,Lu 等[8]和 Li 等[9]在综述中全面总结了遥感视觉基础模型和遥感视觉 - 语言基础模型的研究进展。然而,对于基础模型在雷达遥感解译领域中研究现状的讨论和总结并不充分。因此,系统性地总结该领域最新的进展,以及面临的主要挑战和未解决的难题,进一步探讨和挖掘具有研究潜力的方向是具有重要价值的。基于此,本文聚焦应用于雷达遥感的基础模型研究,旨在进一步填补该研究领域的空白。

首先,介绍了雷达遥感基础模型的基本内涵,系统阐释其技术体系框架、关键技术以及模型评估方法。其次,总结了典型模型的方法创新与技术突破。最后,探讨了雷达遥感基础模型未来的研究方向。

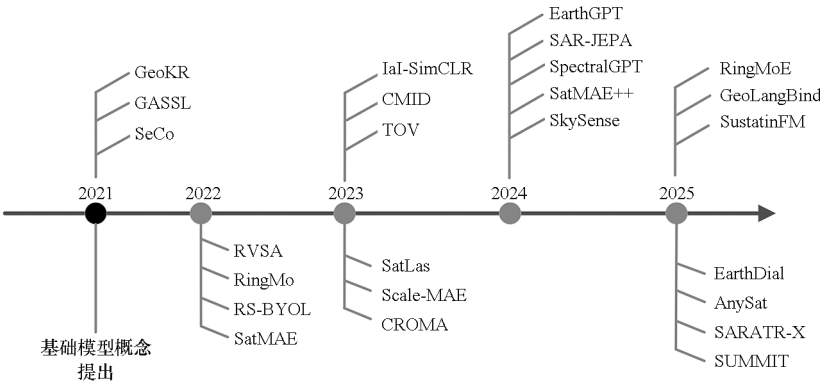

图 1　遥感基础模型发展历程

Fig. 1　Overview of remote sensing foundation models

1　雷达遥感基础模型概述

1.1　基本概念

预训练模型、基础模型、大模型等概念在遥感领域研究中被广泛提及,本小节重点总结了它们的关系和区别,如图 2 所示。

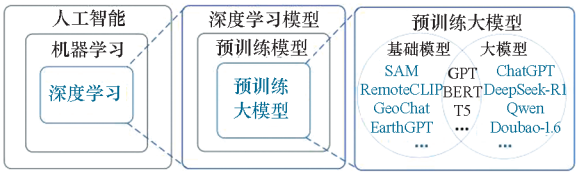

图 2　预训练模型、基础模型、大模型关系图

Fig. 2　Relationships of pre-trained model, foundation model and large model

预训练是机器学习领域中的重要方法,其核心是通过大量数据进行初步训练,捕捉数据中通用的特征和模式从而学习到广泛的知识[10]。这些学习到的特征和知识可以通过少量标注数据的微调,使得模型能够快速适配和提高特定任务上的性能。例如,在自然语言处理中,GPT[11]、BERT[12]等模型在大量文本数据上学习自然语言的基本结构与模式,经过微调用于文本分类、定位搜索、系统问答等任务;在计算机视觉领域中,ResNet[13]、VGG[14]、ViT(vision Transformer)[15]等模型在大规模图像数据上捕获图像的颜色、纹理、边缘等通用特征,适配于目标检测、场景分类、语义分割等视觉任务。

大模型通常指拥有巨大规模参数量的机器学习模型。这些模型遵循规模化法则[16](scaling law),通过扩大参数规模和训练数据量涌现小模型所不具备的解决更深层次问题的复杂能力,展现出类人的思维和智能。大模型强调模型参数和训练数据的规模,利用大模型涌现出的能力来实现复杂任务,如 ChatGPT、DeepSeek-R1[17]等利用人机对话完成信息检索、机器翻译、代码编写、文本总结等内容生成任务。

基础模型是指在多种任务中具有广泛适用性的大型预训练模型[18]。这些大型预训练模型被定义为基础模型有两方面因素:基础模型是作为通用的基础来构建和适配下游任务;使用“基础(foundation)”是为了强调模型架构稳定和安全的重要性。基础模型是特殊类型的预训练模型,与一般预训练模型相比参数规模更大、算力需求更高,更侧重多任务的泛化能力。

总结而言,预训练模型是构建基础模型和大模型的基础。大模型侧重于扩展参数和训练数据

的规模来实现模型能力的提升;而基础模型则侧重于模型的通用性和适应性,构建能快速适用于多领域、多任务的基础底座。当前,基础模型和大模型的发展逐渐趋于同质化,概念之间的界限会越来越模糊。

1.2 关键技术

数据、模型架构和预训练技术是构建基础模型的基本要素。数据的质量以及数据集的构建方式是基础模型性能和泛化能力的基础。预训练算法的设计则引导模型从数据中学习到不同的特征表示。模型架构则决定对数据的建模方式,其特性决定了其捕获和理解现实世界信息的能力。本小节将从这三个要素依次进行展开介绍。

1.2.1 预训练数据集

雷达遥感解译的研究与数据集的发展有着密切联系,低质量的图像、文本描述、标注容易导致模型学习到错误的模式,使输出的结果出现偏差;Scaling Law[18]指出,数据的规模直接影响模型的性能,模型性能随训练数据量的增加呈幂律提升,但也存在边际递减效应;数据的多样性使模型学习到接近真实世界的特征表示,对新的场景具有更好的适应性。当前在雷达遥感领域,已有一系列研究致力于构建基础模型的预训练数据集,接下来根据数据模态的不同分别进行归纳和总结。表1和表2分别统计了用于雷达遥感视觉模型和视觉-语言模型训练的数据集。

表1 雷达遥感视觉数据集

Tab. 1 Radar remote sensing vision datasets

数据集	时间	样本量	波段	极化方式	分辨率/m	任务
MSTAR[19]	1996	9 466	X	单极化	0.3	分类
OpenSARShip-1.0[20]	2018	45 874	C	双极化	20	检测、分类
SAR-Ship-Dataset[21]	2019	43 819	C	单/双/全极化	1.7~25	检测
HRSID[22]	2020	5 604	C/X	双极化	0.5~3	检测、语义分割
FUSAR-Ship[23]	2020	16 144	C	全极化	1.1~1.7	分类
SSDD[24]	2021	2 456	C/X	单极化	1~15	检测、语义分割
AIR-PolSAR-Seg[25]	2022	2 000	C	全极化	8	分类
MSAR[26]	2022	60 396	C	单极化	1	检测
SAR-AIRcraft-1.0[27]	2023	4 368	C	单极化	1	检测
Kuro Siwo[28]	2024	1 600 000	C	双极化	10	洪水制图
SARDet-100K[29]	2024	116 598	C/X/Ka/Ku	单极化	0.1~25	检测、分类
RSAR[30]	2025	95 842	C/X/Ka/Ku	单极化	0.1~25	检测、分类
FAIR-CSAR-V1.0[31]	2025	29 965	C	单/双极化	1~5	检测、分类
NUDT4MSTAR[32]	2025	194 324	X/Ku	全极化	0.12~0.15	检测、分类
MuSID[33]	2025	569 635	C/X	全极化	0.1~25	预训练

表2 雷达遥感视觉-语言数据集

Tab. 2 Radar remote sensing vision-language datasets

数据集	时间	样本量	数据类型	属性描述
MMM-RS[34]	2024	约210万	光学、SAR、近红外、文本	图像生成
MMRS-1M[35]	2024	约100万	光学、SAR、红外、文本	分类、目标检测、图像字幕、视觉问答、视觉定位
SARChat-2M[36]	2025	约200万	SAR、文本	图像字幕、视觉问答、视觉定位、目标检测
SARLANG-1M[37]	2025	约100万	SAR、文本	图像字幕、视觉问答、视觉定位
EarthDial-Instruct[38]	2025	约1 111万	光学、SAR、近红外、红外、多光谱、文本	图像字幕、视觉问答、多标签分类、变化检测、灾害评估等44种任务
GeoLangBind-2M[39]	2025	205万	光学、SAR、多光谱、文本	分类、检索和语义分割
SustainFM[40]	2025	72.3万	光学、SAR、红外、文本	变化检测、语义分割、分类、回归
GeoPlex[41]	2025	—	光学、SAR、多光谱、文本	洪水分割、烧伤斑识别、农作物分类、树种识别、气候带划分和森林变化检测

(1)雷达遥感视觉数据集

早期雷达遥感数据集的构建以单极化、单波段为主,主要应用于雷达成像算法的验证。随着雷达遥感卫星技术的发展, OpenSARShip-1.0[20]和 SSDD[24]等数据集采集了多个卫星的雷达遥感影像,利用深度学习方法提取数据的丰富细节纹理信息,使得目标分类、检测等下游任务的性能得到了显著提高。但这类数据集中的目标均以切片为主,背景较为单一,所训练的模型泛化能力有限。SAR-Ship-Dataset[21]、HRSID[22]、MSAR[26]等数据集的构建则采集了机场、城区、岛礁、港口以及不同海况等多种场景下的多类目标,极大地丰富了雷达遥感数据集的多样性。然而,上述数据集的构建过程需要大量专家知识,标注成本较高,训练的模型用于特定任务,面对新的场景存在泛化性弱的问题。

随着基础模型在遥感领域得到了广泛的应用,数据集的构建朝着数量规模化和标注粒度精细化的方向发展。SARDet-100K[29]将多个公开数据集进行了统一标准化处理,构建了大规模多类别的合成孔径雷达目标检测数据集。其中包含了超过 11 万张 512 像素 ×512 像素的图像和 24 万个实例,涵盖了舰船、飞机等六类典型目标,其数据规模和场景覆盖度为雷达遥感基础模型的相关研究提供了训练和评估的数据支持。RSAR[30]数据集在此基础上对其进行了旋转框的标注,提升了该数据集标注的精细程度。MuSID[33]数据集则收集了 18 个公开的数据集,通过数据清洗和重采样构建了超过 56 万张 448 像素 ×448 像素图像的预训练数据集,进一步支撑了基础模型的训练。

不同于光学图像所见即所得的特性,雷达遥感图像解译的基础是对目标电磁散射特性的准确建模。现有的数据集多以单通道数字图片为主,为了将丰富电磁信息用于模型的特征表示和引导训练,FAIR-CSAR-V1.0[31]和 NUDT4MSTAR[32]数据集构建具有多种场景、类别和极化方式的复值图像数据集。其中,FAIR-CSAR-V1.0 数据集在标注方面由简单的目标框拓展到了属性关联,提供了细粒度的注释和丰富的特性信息(如星地相对方位角、强散射点分布);NUDT4MSTAR 数据集中详细地标注了目标信息以及成像条件(如目标尺寸、场景、擦地角、方位角)。

(2)雷达遥感视觉 - 语言数据集

视觉 - 语言基础模型将大语言模型与视觉编码器结合,通过在图像 - 文本对的数据上进行训练,利用自然语言交互的方式来处理视觉中典型的任务(场景分类、目标检测、语义分割)。近期涌现了多个用于支撑雷达遥感视觉 - 语言基础模型训练的大规模数据集,其中包含了多种传感器数据和大量的图像 - 文本对。

如表 2 所示,这些数据集的构建得益于不同遥感平台、不同体制成像技术的发展,通过整合多个已有公开数据集来拓展现有数据集的规模。然而,针对这种规模数据进行文本的标注和描述成本高,因此研究人员利用多模态大语言模型(如 ChatGPT-4o)来实现拓展文本信息标注和生成问答对话指令,如 EarthDial-Instruct[38]、GeoLangBind-2M[39]、GeoPlex[41]等数据集。

在这些数据集中,SARChat-2M[36]和 SARLANG-1M[37]是专门针对雷达图像的大规模多模态对话数据集。SARLANG-1M 数据集通过跨模态文本迁移的方式,利用光学图像生成高质量文本描述,再与其匹配的雷达图像进行对齐。此外,还通过已有的边界框标注来生成细粒度文本描述,这种结合了专家知识的构建方式提高了雷达图像 - 文本描述质量,并且降低了标注成本。SARChat-2M 数据集则利用原有标注的目标类别、几何属性、空间位置等信息通过跨模态学习进行图像 - 文本对的生成,实现了细粒度文本描述的自动化生成。SARChat-2M 和 SARLANG-1M 数据集的构建方式为雷达遥感的图像 - 文本数据集的构建提供了新的思路,减少了专家知识的干预并提高了自动化程度。

1.2.2　模型架构

骨干网络是基础模型的主要部分,负责输入数据的特征提取,不同的骨干网络架构决定了所构建模型的特性。现有骨干网络架构主要分为三类:卷积神经网络(convolutional neural networks, CNN)结构、Transformer 结构和 CNN-Transformer 混合结构。

(1) CNN 结构

CNN 结构已经广泛应用于各类计算机视觉任务。典型的 CNN 由卷积层、池化层、全连接层和输出层组成,通过级联的方式将各层进行连接[42]。基于这种架构,CNN 能够学习到纹理、部件、对象等多层级特征。

CNN 无须手工设计特征,通过端到端的学习直接从雷达遥感原始数据提取多层次的特征,提高了检测、识别等任务的性能[4]。此外,通过数据增强技术(如尺度变换、噪声扰动等),使 CNN 模型学习到更鲁棒的特征,有效降低噪声带来的性能损失。但是,CNN 缺乏对全局信息的感知

力，导致其对长序列数据的处理性能不佳，易丢失数据空间位置信息。此外，CNN 较为依赖标注数据，且在雷达遥感图像应用中对成像参数（波段、俯仰角）变化敏感。

（2）Transformer 结构

Transformer 架构不仅支撑了大语言模型的性能突破，还为计算机视觉、多模态、科学计算等多个领域提供了相对统一的模型架构，如 ChatGPT、LLaMa[43]、CLIP[44]、AlphaFold[45]。在视觉基础模型中，通常将 Transformer 结构直接应用于图像数据处理，如 ViT、Swin-T（Swin Transformer）[46]等。

图 3 展示了基于 ViT 结构的雷达遥感图像目标检测模型。首先，将图像划分成固定大小的图像块，经向量化处理后通过可学习的线性层映射到高维隐空间。然后，通过位置编码器为图像块添加空间位置信息，以保留图像的空间结构并利用训练动态捕捉空间关系。最后，图像块经过多层 Transformer 编码器处理，聚合全局信息并进行任务导向的特征重组，输出检测、分割等任务的推理结果。

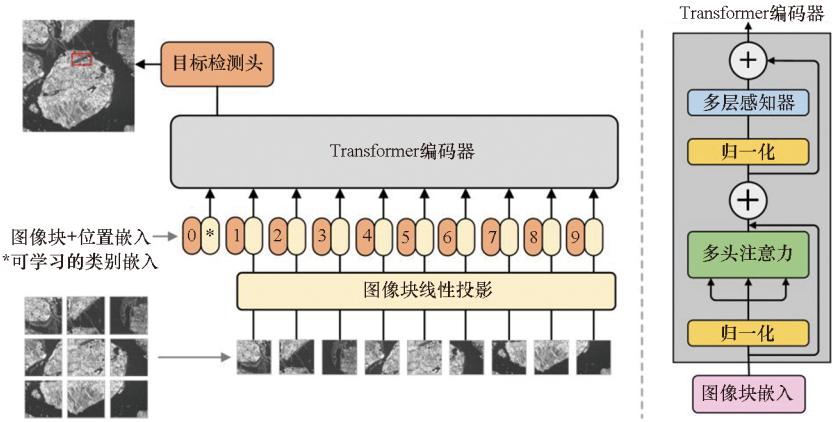

图 3 基于 ViT 的雷达遥感图像目标检测模型

Fig. 3 Radar remote sensing image object detection model based on ViT

Swin-T 针对 Transformer 在视觉应用中出现的多尺度目标实例、高分辨率图像计算资源消耗大的问题，提出了一种基于滑动窗口机制的多层级网络架构。通过在结构上的改进，Swin-T 能高效处理多分辨率图像，并降低了计算复杂度，在更细粒度的下游任务中取得了更优的性能，成为视觉基础模型常用的骨干结构。

Transformer 具有全局建模和跨模态交互的优势，在雷达遥感图像的应用中突破了传统 CNN 在长距离上下文建模和复杂场景理解的瓶颈。不仅在传统视觉任务中具有良好的性能，还通过跨模态交互的方式实现了多任务的统一。然而 Transformer 缺乏 CNN 的局部归纳偏置能力，在处理局部细节或小目标时易丢失高频信息。此外，Transformer 在训练时通常需要较大规模的数据，且处理大尺度数据需要较高的计算量，耗费的时间、硬件成本较高。

（3）CNN-Transformer 混合结构

CNN 具有良好的局部特征处理能力，能够提取图像局部高频的细节信息，但是处理全局信息的能力较弱。Transformer 具有良好的序列数据建模与生成能力，在处理全局信息方面具有优势，但是对于局部信息处理能力相对较弱。因此，研究人员通过将 CNN 结构和 Transformer 结构进行结合，可以有效地捕捉和处理全局和局部信息，如 CoaT[47]、ConvNeXt[48]等。

CoaT 引入了协同尺度机制，实现从精细到粗略、从粗略到精细以及跨尺度建模的能力，使模型能够在不同分辨率下既保留局部细节又能捕获全局上下文信息。此外，CoaT 还融合了 CNN 中的卷积和 Transformer 中的自注意力机制，设计了卷积自注意力机制，实现因子化注意力模块中卷积的相对位置嵌入，与 Transformer 中的普通自注意力层相比，计算效率显著提高。

ConvNeXt 在保持 CNN 结构的基础上，通过参考 Swin-T 的结构设计和训练优化方法，在 ImageNet 数据集上达到了和 Transformer 相近的性能。ConvNeXt V2[49]在此基础上，提出了全卷积掩码自编码器自监督学习框架，以此提高 ConvNeXt 的表示学习能力和扩展性。实验结果表明在同等参数量下 ConvNeXt V2 在 ImageNet-1k 数据集上性能超越了 ViT。

目前，CNN-Transformer 混合结构在雷达图像解译中得到了大量应用。GLNS[50]通过融合轻量级的 CNN 和 ViT 来捕捉局部和全局特征，这些特征随后被融合来执行雷达图像分类任务；

CRTransSar[26]结合了 CNN 的局部信息捕获能力和 Swin-T 的上下文学习能力,增强了雷达图像中目标特征属性的同时还提取到了更丰富的上下文特征信息。

总结而言,在未来雷达解译应用中,这种混合结构的局部 - 全局特征协同特点能提升目标识别、检测等任务的性能。此外,利用 CNN 构建辅助任务(去噪、超分辨、边缘提取)分别针对雷达图像的不同特性,通过联合训练的方式让模型学习更全面、更鲁棒的特征表示[33]。

1.2.3　预训练技术

当前,仅通过模型参数和预训练数据的规模扩展带来的收益存在边界效应。因此,预训练技术的创新成为提升基础模型性能的关键因素。

自监督学习技术通过挖掘数据本征结构来构建监督信号,打破了传统监督学习对人工标注的依赖。该技术利用图像的上下文关系(如空间结构、时序连续性或跨模态一致性)或图像之间的相似性设计代理任务,使模型从无标注数据中学习到对下游任务有价值的特征。

基础模型常用的预训练技术如图 4 所示。当前研究中预训练技术聚焦于对比学习和图像掩码建模:

1)对比学习(contrastive learning)。对比学习通过对比正负两个方面的实例来提取有意义的特征,将学习作为一项判别任务,让模型捕捉数据中的相关特征和相似性。常用的判别式对比学习损失函数表示为

$$L_i = -\ln \frac{\exp(\mathrm{sim}(x_i, x_i^+)/\tau)}{\exp(\mathrm{sim}(x_i, x_i^+)/\tau) + \sum_{j=1, j\neq i}^{N} \exp(\mathrm{sim}(x_i, x_j^-)/\tau)} \tag{1}$$

式中:x_i 是查询样本,x_i^+ 是与 x_i 相关的正样本,x_j^- 是与 x_i 不相关的负样本;$\mathrm{sim}(x_i, x_j)$ 是 x_i 与 x_j 的相似度,通常使用余弦相似度;τ 是温度参数,用于调整对比学习中相似度的尺度,避免梯度消失或爆炸问题。

2)图像掩码建模(masked image modeling, MIM)。MIM 方法将部分图像块进行掩码,用剩余可见的图像块来恢复被掩码的图像块,期望能同时兼顾局部细节(如边缘、纹理)与全局语义(如物体结构)[51]。MIM 方法中常用的均方误差(mean squared error, MSE)损失函数表示为

$$L_{\mathrm{MSE}} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

式中,y_i 表示原始图像的第 i 个像素值,$\hat{y}_i$ 表示重建图像对应位置的像素值,n 表示像素总数。

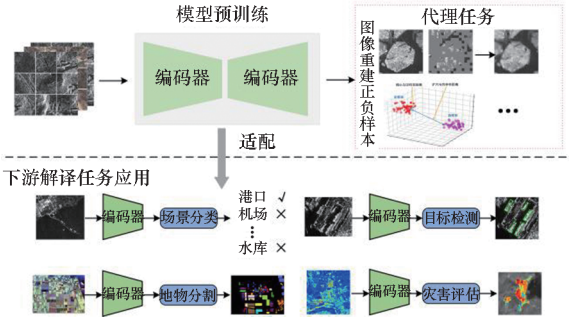

图 4　自监督预训练方法

Fig. 4　Self-supervised pre-trained method

在实际应用场景中,对比学习具有很强的特征判别能力,在图像增强(如超分辨率重建)、跨模态特征匹配等对区分精度要求较高的场景中表现突出。而图像掩码重建侧重于表征学习,在深度语义推理和内容生成等任务中展现出独特优势。

1.3　评价方法

对基础模型进行评价不仅可以判断模型的固有缺陷(如算法偏见、安全漏洞),还能评估其迁移至下游任务时可能引发的风险。此外,评价的结果可以反馈到模型设计和训练过程中,进一步指导改进的方向。

当前,基础模型的评价方法可分为内在评价和外在评价[18]。内在评价是对模型的内部机制、可解释性等方面的分析,聚焦于基础模型本身而不依赖于具体的下游任务。通常结合热力图、特征归因、因果推断等方法和工具[52-54],对模型内部决策机制、特征的交互、模型推理路径及认知对齐等方面进行分析[55-57]。外在评价则对基础模型的任务性能指标进行衡量,如视觉任务中常用的准确率、召回率、平均精度等[58]。

随着基础模型的广泛应用,对其评价的范围拓展到了稳健性、公平性、对抗鲁棒性以及计算效能等维度。如 SustainFM[40]以可持续发展目标为导向设计了地理空间基础模型评估基准,不仅涵盖了从资产财富预测到环境灾害检测的多样化任务,还将模型性能和微调的能源效率纳入评估范围,推动了地理空间基础模型向更加可持续、公平的方向发展。因此,构建一套标准化、多维度的评价体系值得深入研究。

2　雷达遥感基础模型研究进展

近期,研究人员设计了一系列用于雷达遥感解译任务的基础模型,如表 3 所示。本文将当前

应用于雷达遥感图像的基础模型分为三类:雷达遥感视觉基础模型、雷达遥感视觉 - 语言基础模型、物理知识引导的雷达遥感基础模型。后续,本节对逐个进行总结和分析。

表 3 雷达遥感基础模型总结

Tab. 3 Gallery of the foundation models for radar remote sensing

年 - 月	模型	架构	数据集	数据类型	任务应用
2022 - 08	SSLTransformer-RS[59]	Swin-T、UNet	SEN12MS	光学、SAR	图像分割
2023 - 11	RingMo-SAM[60]	ViT		光学、SAR	图像分割
2023 - 12	CROMA[61]	ViT	SSL4EO	光学、SAR	地物分类
2024 - 06	EarthGPT[35]	ViT、ConvNeXt	MMRS-1M	光学、SAR、文本	地物分类、图像字幕、视觉问答、目标检测
2024 - 09	SkySense[62]	ViT、Swin-T	WorldView-3/4 和 Sentinel-1/2 平台 2.15 亿张影像	光学、SAR	图像分割、目标检测、变化检测、地物分类
2024 - 09	SAR-JEPA[63]	ViT	MSAR、SARSim、SAMPLE、SAR-Ship	SAR	目标检测、目标识别
2024 - 09	SPT[64]	CLIP-ViT		SAR	目标识别
2024 - 11	FG-MAE[65]	ViT	SSL4EO-S12	光学、SAR	地物分类
2025 - 01	CWSAM[66]	ViT	FUSAR-Map1.0/2.0	光学、SAR	图像分割
2025 - 01	SARATR-X[67]	HiViT	SARDet-100K	SAR	目标识别、目标检测
2025 - 03	RingMoE[68]	Swin-T V2	RingMOSS	光学、SAR、多光谱	图像分类、目标检测、目标跟踪、图像分割、变化检测、深度估计
2025 - 03	GeoLangBind[39]	ViT、SigLIP	GeoLangBind-2M	光学、SAR、多光谱、红外、文本	图像分类、图像分割、跨模态检索
2025 - 04	SustainFM[40]	ViT、ResNet-50	Landsat、Sentinel-1/2 等 6 类平台的 72 万张影像	光学、SAR、多光谱、红外	回归、图像分割、变化检测、分类
2025 - 04	EarthDial[38]	ViT、InternLM2	EarthDial-Instruct	光学、SAR、文本	场景分类、目标检测、图像字幕、视觉问答
2025 - 05	AnySat[41]	ViT	GeoPlex	光学、SAR、高程数据	洪水分割、烧伤斑识别、农作物分类、树种识别、气候带划分等
2025 - 06	SUMMIT[33]	ViT	MuSID	SAR	图像分类、目标检测、图像分割

2.1 雷达遥感视觉基础模型

在视觉基础模型领域,训练方法的研究从早期利用大量有标注数据进行监督学习(如 ImageNet)发展到近期的对比学习方法(大规模数据自监督训练)。近期,受大语言模型成功的启发,图像掩码建模方法(如 MAE[69])被广泛关注,并成为当前视觉基础模型预训练的主要范式之一,如图 5 所示。

相较于自然图像,雷达遥感图像具有独特的成像机理,光学的预训练框架难以直接迁移。因此,研究人员设计了面向雷达遥感图像的视觉基础模型预训练框架,使其能有效学习和适应雷达遥感图像的特有物理属性(如散斑噪声、散射特征)。SAR-JEPA[63] 提出了基于联合嵌入预测架构的模型预训练和微调方法,通过局部掩码策略和多尺度梯度特征构建自监督任务,在特征空间构建了跨层级的自监督约束,有效地克服相干斑噪声带来的影响。值得注意的是,该研究进一步

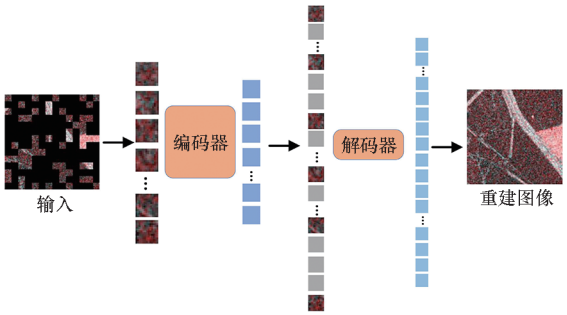

图 5　图像掩码建模方法[65]

Fig. 5　Masked image modeling method[65]

揭示了数据规模与模型泛化能力间的关系,从而构建了大规模预训练数据集 SARDet-100K[29] 和用于雷达图像目标识别的基础模型 SARATR-X[67],并且设计了双阶段预训练方法:首阶段通过 ImageNet 数据进行预训练,以获取更具多样化的初始模型权重;第二阶段结合雷达遥感图像的重要梯度特征表示来生成高质量的自监督信号,进一步抑制相干斑噪声对下游任务预测结果的影响。

基于单模态雷达遥感数据构建的基础模型在特定场景下展现出了强大的表征学习能力,其学习到的特征空间主要源于雷达成像的物理机制和几何特性。然而,开放世界的信息解译通常是多维度的,单一数据源不可避免地存在其固有局限。因此,通过多源数据融合不仅能够利用多传感器的优势互补有效缓解云雾遮挡、夜间观测等光学探测下的信息缺失问题,还能够通过建立跨域特征关联显著提升模型对复杂地物目标解译的鲁棒性。近期,研究人员利用基础模型结合多源数据来提高遥感解译任务的性能,对适应新体制观测体系和提供更精确的态势感知具有重要意义。

SSLTransformer-RS[59] 将同一位置不同传感器获取的卫星图像作为正样本对,将不同位置和传感器的图像作为负样本,通过对比学习构建多源数据下有效的结构表征。类似地,CROMA[61] 以光学图像为锚点,将与之在地理和时间上匹配的雷达遥感图像作为正样本,将同一批次中的其他雷达遥感图像作为负样本;反之,以雷达遥感图像为锚点时,光学图像的正负样本以相同规则构造。这种双向跨域对比机制使模型学习到传感器具有不变性的高阶语义特征,有效捕捉雷达和光学传感器数据之间共享的信息。AnySat[41] 提出一种对比自监督训练框架,将联合嵌入预测架构和跨模态对比学习结合,利用构建的 GeoPlex 大规模数据集进行预训练,并将模型应用于洪水分割、烧伤斑识别、农作物分类、树种识别、气候带划分和森林变化检测等多种任务。

在雷达遥感视觉基础模型构建中,图像掩码建模和对比学习等方法能够有效地挖掘大规模数据的内在结构和规律,提高了通用特征提取的能力并降低对标注数据的依赖。这种非监督模式通过自动学习的方式能更好地适应不同场景、不同数据的特性,具有更强的适应性和通用性。然而,这些方法也面临共同的问题和挑战。非监督模式通过数据自身构造监督信号来进行训练(如图像掩码建模的重建代理任务,对比学习正负样本的构造),在训练过程中相较于有监督模式的标签信息,需要根据数据特性设计更合适的自监督信号。

2.2　雷达遥感视觉 – 语言基础模型

视觉 – 语言基础模型旨在统一图像和自然语言的信息表征,通过文本交互的方式完成对图像的解译。在雷达遥感视觉 – 语言基础模型的研究中通常采用预训练 – 微调的范式,其核心是在大规模图像 – 文本数据中进行预训练,使模型通过文本交互方式完成多解译任务的统一。

EarthGPT[35] 在构建的 MMRS-1M 数据集上采用了视觉增强感知、跨模态理解、统一指令微调等策略,在分类任务、图像描述、视觉问答、零样本推理等应用中展示出了巨大潜力。类似地,GeoLangBind[39] 构建了更大规模数据集 GeoLangBind-2M,通过最大化相似样本在嵌入空间中的相似度,最小化不相似样本的相似度,来学习统一的视觉 – 语言表示,并在光学、雷达等图像上以文本交互方式实现了多种视觉任务的统一。

尽管雷达图像已经在上述研究中被广泛使用,但由于其标注和文本描述需要大量专家知识干预,自动化标注和文本生成较为困难。因此,针对雷达遥感视觉 – 语言基础模型的研究有很大的提升空间。近期,SARChat-Bench-2M[36] 构建了针对雷达图像的大规模对话数据集,同时选取了 InternVL2.5[70]、DeepSeek-VL[71]、GLM-Edge-V 和 mPLUG-Owl[72] 等 16 种不同参数规模的视觉 – 语言基础模型对 6 类视觉和多模态任务进行评估。同样地,SARLANG-1M 构建了一个大规模雷达图像细粒度描述的图像 – 文本数据集,选取 DeepSeek-VL、Qwen2.5-VL[73] 等 10 种不同视觉 – 语言基础模型进行微调,微调的模型对图像描述、视觉问答、视觉定位、实例计数以及参数反演等下游任务进行评估。此外,研究还表明,对雷达图像进行预处理(去噪、图像增强)能使雷达遥感视

觉－语言基础模型更有效地关注目标和区域。

当前涌现了一批可应用于雷达遥感解译的视觉－语言基础模型，这些模型通过视觉和语言统一建模的方式融合了多源数据，以文本交互的方式实现了视觉任务和理解任务的统一，提高了雷达遥感解译中处理复杂问题的能力。特别是，SARChat-Bench-2M 和 SARLANG-1M 为文本标注困难的雷达图像领域提供了数据收集、标注、预训练和评估的方法和流程，为其他垂直领域提供了思路。然而，雷达遥感领域的基础模型旨在为民生经济、军事侦察等领域提供更有效、可靠的信息，当前对视觉－语言基础模型的研究缺乏常识和物理约束，生成的信息存在幻觉和对抗攻击等安全风险，因此还需要构建对模型可靠性和安全性的基准测试。

2.3 物理知识引导的雷达遥感基础模型

雷达遥感数据所包含的信息是地物目标对雷达波束的反映，不同的波段、极化方式、入射角都会影响雷达遥感图像所蕴含的信息[74]。因此，研究人员结合雷达遥感图像的物理机理与几何特性提出了一些方法。

如1.2节所述，非监督模式的预训练方法需要通过自身构建监督信号。近期，部分研究人员通过利用雷达遥感图像的物理特性来构建自监督信号。FG-MAE[65]利用方向梯度直方图等特征描述子提取图像边缘梯度、方向等信息构建自监督信号，实验表明该方法相较于直接使用 MAE 方法生成的特征表示具有更强的地物区分能力。类似地，SUMMIT[33]利用 Canny 边缘检测算法和 Harris 角点检测算法提取边缘图和散射点图构建辅助自监督信号，融合了边缘和散射点信息的同时还增加了自监督去噪的预处理分支，有效地抑制了噪声带来的影响。此外，RingMoE[68]结合了全极化数据的四个极化通道，利用极化功率特征来进行图像的掩码重建，以此增强模型在预训练阶段的可解释性。

除了在预训练阶段利用物理特性构建自监督信号，还有研究关注任务的微调阶段。RingMo-SAM[60]在 SAM 视觉基础模型上，分析了遥感图像中目标密集的特点和雷达数据的极化散射特性，通过提示学习机制嵌入了雷达成像机理与地物特性，实现了对雷达遥感数据的多要素语义分割，并具备在新场景数据上的零样本泛化能力。CWSAM[66]同样在 SAM 视觉基础模型上，分析了雷达信号的频域特性并设计了处理低频信息的输入模块，通过快速傅里叶变换来提取土地覆盖特征的语义信息，引导模型增强对细粒度语义分割的能力。SPT[64]提出了一种基于散射提示的微调方法，将雷达遥感图像中的散射信息通过文本编码器转换为文本描述，并将其作为提示信息与视觉图像描述一起处理。

上述方法均基于预训练－微调范式，在不同阶段融合了雷达遥感图像的几何特性和散射信息来改进模型训练范式，利用自监督学习和物理知识的约束驱动模型学习雷达遥感图像通用的特征表示，提高了雷达遥感基础模型的可解释性。但是，当前对于物理知识的运用依然局限于原有的光学预训练框架，对于所添加物理约束和生成的结果在一致性方面并未进行深入的研究。

3 雷达遥感基础模型应用

基础模型是为雷达遥感图像解译任务构建通用的基座，1.3节介绍了现有雷达遥感基础模型的评价方法。在研究中，评价模型性能最直接的方法依然是结合具体应用进行分析。本节将介绍雷达遥感基础模型在实际任务中的应用效果。

3.1 图像分割

雷达遥感图像由于其散射特性成像与光学图像视觉差异大，包含相干斑噪声、复杂散射特征等特性。CWSAM[66]为了克服雷达遥感图像特性给 SAM 基础模型带来的挑战，引入了端到端的轻量适配器，并结合地物特征的语义信息进行地物分割。该方法对6个省份8个区域的高分3号雷达遥感图像进行分割，结果如图6所示。实验表明，在 mIOU、OA、Precision 指标上相较于最高对比方法提升了0.59%、1.81%、2.19%，并且在大多数类别中均取得了最佳性能，进一步表明了基础模型具有更高的模型鲁棒性和泛化能力。

除了地物分割应用，基础模型在洪水分割任务中也得到了广泛应用。Kuro Siwo 构建了一个覆盖330亿 m^2 的全球洪水事件数据集，实验中测试了多种架构，包括 ResNet、ConvNeXt、ViT、Swin-T，结果表明 ViT 和 Swin-T 的性能与大多数模型相近，并没有展现出显著的优势。经过分析发现，复杂的地形和不同的土地覆盖类型会影响雷达信号的相互作用，风或植被的存在会使水面粗糙度改变，进而影响后向散射，难以准确识别洪水。

3.2 目标检测

对高价值军事目标的检测是雷达遥感解译的核心任务之一。SARATR-X[67]设计了基于自监督学习的雷达目标特征表示学习方法，该方法在分

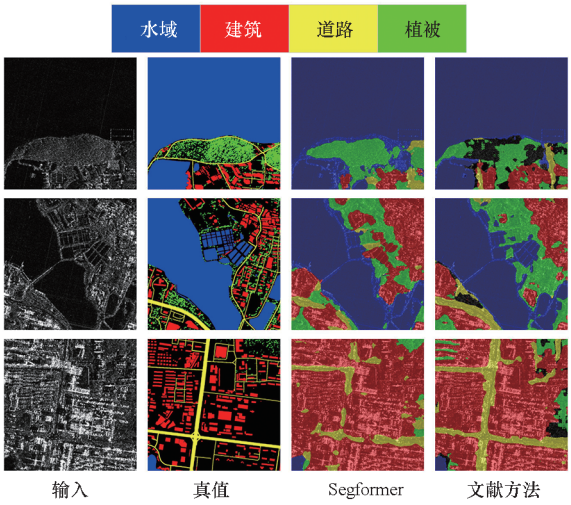

图 6　FUSAR-Map2.0 数据集地物分类定性结果[73]

Fig. 6　FUSAR-Map2.0 datasets qualitative results of landcover classification[73]

类和检测任务中均获得了显著提升,如图 7 所示。在 MSTAR 数据集中,它的目标分类性能优于现有的自监督学习方法,提升了 4.5%,在多类别的目标检测中平均提升约 4%,在舰船和飞机目标检测中提升了 2.6% 和 5.7%。

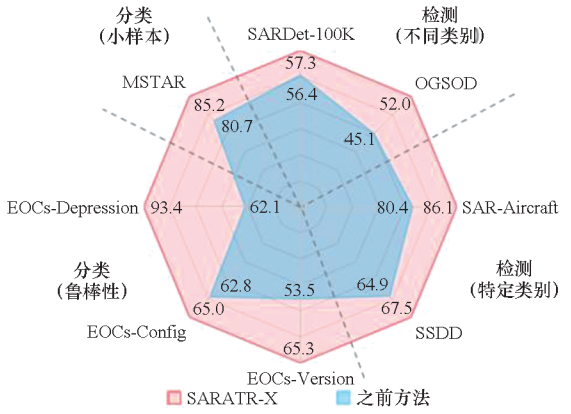

图 7　SARATR-X 目标识别性能[74]

Fig. 7　Target recognition performance of SARATR-X[74]

SUMMIT[33]提出了针对雷达图像特性的自监督预训练框架,设计了包含自监督去噪、空间散射特征增强等辅助任务,在不同阶段参与模型的训练。所构建的模型在 SSDD 数据集和 SAR-AIRcraft-1.0 数据集上相比于仅在 SAR 图像上进行预训练的模型 mAP 指标提高了 2.4% 和 3.9%,在 SARDet-100K 数据集上 mAP、mAP50 和 mAP75 等指标提升均超过 7%。实验还进一步表明了在基础模型构建中,图像预处理和特征增强对性能的提升起到重要作用。

基础模型在军事目标识别领域相较于传统深度学习模型具有更强的适应性和泛化能力。但是,实际战场具有复杂的电磁对抗环境和多类型的人造假目标,当前基础模型的研究缺乏相应的应对措施。

3.3　地物分类

对于雷达遥感图像地物分类的难点主要在于图像中存在的相干斑噪声,这种噪声表现为颗粒状干扰,通常被建模为乘性噪声,这限制了基础模型中自监督学习算法的性能。FG-MAE[65]在基于掩码图像建模的自监督学习方法基础上,结合方向梯度直方图特征作为自监督信号来增强模型对空间信息的学习并抑制相干斑噪声的影响。实验结果表明,在 EuroSAT-SAR 地物分类数据集上,使用改进的预训练技术后模型对 10 类地物的平均精度提升 5.9%,个别地物提升达 11.1%,改进前后方法对比效果如图 8 所示。

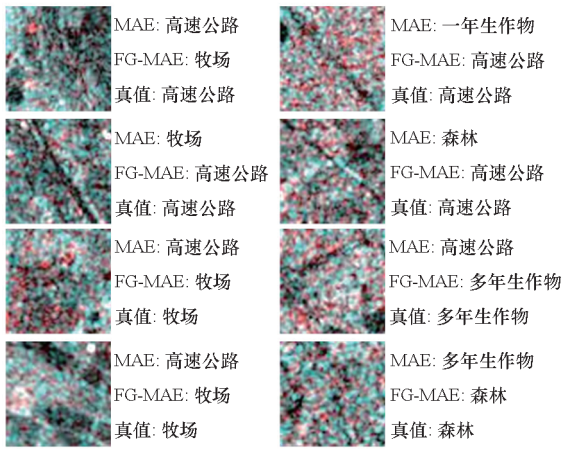

图 8　FG-MAE 地物分类结果对比[65]

Fig. 8　Comparison of the landcover classification results by FG-MAE[65]

FG-MAE 在预训练框架中加入物理特性约束,在雷达图像地物分类任务中提升显著。但是,使用的特征注意力机制依赖预定义的特征图(边缘特征),容易受噪声影响导致提取到伪特征,影响预测结果。此外,较高的掩码率容易丢失局部特征和散射特征,导致重建的细节模糊。图像掩码建模是当前基础模型自监督训练中主要的方法之一,挖掘能全面刻画雷达图像的自监督信号是未来对其优化的重点方向。

3.4　多模态学习任务

语言 – 视觉基础模型在自然图像领域取得了显著成功,然而在遥感领域的发展仍处于起步阶段。基于此,研究人员设计了多任务通用遥感语言 – 视觉基础模型 EarthGPT[35],包含雷达、红外等多种传感器的图像解译。模型利用自然语言指令的形式完成对雷达遥感图像的场景分类、图像

字幕、目标检测等多个任务,效果如图 9 所示。

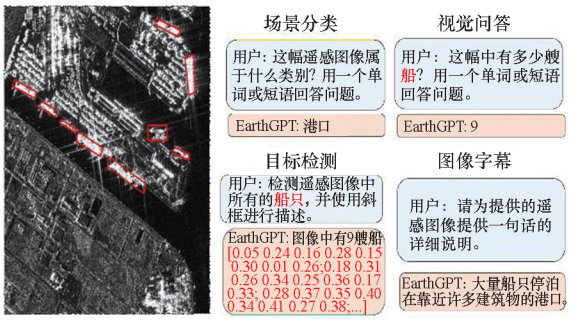

图 9 EarthGPT 多任务视觉问答[36]

Fig. 9 Multitask visual question answering of EarthGPT[36]

为了进一步对比分析视觉 - 语言基础模型和人类在雷达图像解译方面的能力,SARLANG-1M[37]在 SARLANG-1M-VQA 数据集上对比了多种模型和解译专家的实验评分。其中,经过雷达图像数据微调的 QWEN2.5-VL 模型准确率为 63.33%,超过了解译专家 57.76% 的准确率。此外,还分析了雷达图像预处理的必要性。在图像描述任务中,QWEN2.5-VL-3B 模型通过预处理在 BLEU_1、ROUGE_L 和 CIDEr 指标上分别取得了 5.39%、5.63% 和 0.46% 的提升,在视觉问答任务中整体准确率提升了 2.81%。

视觉 - 语言基础模型在雷达遥感上的应用推动了雷达遥感智能解译的发展,非行业人员不需要特定的专家知识就能对雷达遥感图像进行解译应用。目前,视觉 - 语言基础模型在雷达遥感领域的应用仍然存在挑战。视觉和文本特征融合过程中常出现对齐偏差(低质量数据、错误描述),导致生成的信息与图像语义脱节。此外,多模态的信息融合增加了对抗攻击的维度,例如可通过视觉、文本两个维度注入对抗样本。

4 研究展望

如前文所述,目前研究人员在数据集构建、预训练方法、任务应用等各个方面都取得了一定的进展和突破。本节针对模型架构、可解释性、轻量化以及安全性四个方面进行展望。

4.1 机理融合的基础模型架构

雷达遥感基础模型与机理融合时面临多重挑战。虽然已有研究将极化信息、入射角、波段等参数纳入模型设计,但这种半定制的物理化架构未能充分挖掘雷达遥感图像数据蕴含的特性,又因过度依赖先验假设而削弱了数据驱动本有的特征挖掘能力。因此,设计通用且灵活的机理融合基础模型架构对物理机理与数据驱动能力的融合具有重要意义。

4.2 基础模型可解释性

基础模型的构建通常基于深度神经网络,其内部工作原理不透明。目前主流的可解释工具与方法依赖于局部扰动生成的近似解释,无法完整还原复杂模型的真实决策逻辑,例如有研究表明注意力机制热力图与模型实际推理路径的关联性较弱[75]。此外,在理论基础与验证上依然存在缺陷,有研究尝试用符号化方法解释神经网络,但仍缺乏对 Transformer 等基础模型架构的内在逻辑的严格数学证明[76]。以往的可解释性方法主要针对特定任务的模型设计,而基础模型由于能够适用于广泛的领域,并可能展现出未预见的特性,因此对现有解释框架提出了新的挑战。

4.3 基础模型轻量化方法

雷达遥感基础模型的部署和应用是面向军事、航天、国家基础设施智能化的重大需求,而实时响应差、应用可信度低和决策不透明等是要解决的首要难题。轻量化方法的研究在平衡基础模型小型化的部署成本与性能之间起到积极作用。在成本上可降低硬件计算资源消耗,达到实时响应的目的;在性能上减小轻量化后的损失,提升泛化能力。因此,模型通用轻量化方法的研究对军事侦察、海洋监测等领域的智能化具有重要的应用价值。

4.4 基础模型安全性

对抗攻击给雷达遥感基础模型在军事侦察等关键领域带来了极大的风险和挑战[77]。尽管基础模型具有大的模型容量、广泛的知识和复杂的推理机制,但是实际会面临更为复杂的攻击环境。对于多平台、多波段、多极化等特性的数据处理增加了对抗攻击的范围,使得攻击形式更加复杂多变。多源数据的应用虽然弥补了单一类型数据的局限,但是给对抗攻击增加了攻击维度,容易受到认知偏差的影响,例如提示注入(prompt injection)、越狱技术(jailbreak techniques)对多模态基础模型进行攻击[78-80]。因此,开展雷达遥感领域基础模型的对抗攻击研究将推动雷达遥感智能解译从被动防御转向主动免疫,对高可靠场景的部署应用具有重要现实意义。

5 总结

具有通用泛化能力的基础模型对于雷达遥感智能解译的发展具有重要意义。本文首先梳理了基础模型的基本概念与特质、模型构建的关键技

术和目前常用的评价方法。其次总结了雷达遥感视觉基础模型、雷达遥感视觉 - 语言基础模型、物理知识引导的雷达遥感基础模型研究与应用现状。在此基础上,分析了雷达遥感基础模型在构建和应用上的挑战,并对其模型架构、可解释性、轻量化、安全性四个方面进行了展望。总结而言,基础模型在雷达遥感领域的应用是遥感智能解译的一项重要进展,给遥感数据的智能解译与应用能力带来了显著提升,为实现人工智能赋能遥感领域迈出重要一步。

参考文献(References)

[1] CHEN S W, WANG X S, XIAO S P, et al. Target scattering mechanism in polarimetric synthetic aperture radar-interpretation and application [M]. Singapore: Springer, 2018.

[2] CHEN S W. Characterization and application of electromagnetic scattering in polarimetric imaging radar [D]. Sendai, Japan: Tohoku University, 2012.

[3] LEE J S, POTTIER E. Polarimetric radar imaging: from basics to applications [M]. Boca Raton, US: CRC Press, 2009.

[4] 徐丰, 王海鹏, 金亚秋. 深度学习在 SAR 目标识别与地物分类中的应用[J]. 雷达学报, 2017, 6(2): 136 - 148.
XU F, WANG H P, JIN Y Q. Deep learning as applied in SAR target recognition and terrain classification[J]. Journal of Radars, 2017, 6(2): 136 - 148. (in Chinese)

[5] 金亚秋. 多模式遥感智能信息与目标识别: 微波视觉的物理智能[J]. 雷达学报, 2019, 8(6): 710 - 716.
JIN Y Q. Multimode remote sensing intelligent information and target recognition: physical intelligence of microwave vision[J]. Journal of Radars, 2019, 8(6): 710 - 716. (in Chinese)

[6] LIAO L Y, DU L, CHEN J, et al. EMI-Net: an end-to-end mechanism-driven interpretable network for SAR target recognition under EOCs [J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 5205118.

[7] 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 K, LU W X, LIU X Y, et al. A comprehensive survey and assumption of remote sensing foundation model [J]. National Remote Sensing Bulletin, 2024, 28(7): 1667 - 1680. (in Chinese)

[8] LU S Q, GUO J L, ZIMMER-DAUPHINEE J R, et al. Vision foundation models in remote sensing: a survey[J]. IEEE Geoscience and Remote Sensing Magazine, 2025: 2 - 27.

[9] LI X, WEN C 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(2): 32 - 66.

[10] HAN X, ZHANG Z Y, DING N, et al. Pre-trained models: past, present and future[J]. AI Open, 2021, 2: 225 - 250.

[11] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners [C]//Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020: 1877 - 1901.

[12] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171 - 4186.

[13] HE K M, ZHANG X, REN S Q, et al. Deep residual learning for image recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770 - 778.

[14] TAMMINA S. Transfer learning using VGG-16 with deep convolutional neural network for classifying images [J]. International Journal of Scientific and Research Publications (IJSRP), 2019, 9(10): 143 - 150.

[15] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16 × 16 words: transformers for image recognition at scale [EB/OL]. (2021 - 06 - 03)[2025 - 01 - 22]. https://arxiv.org/pdf/2010.11929.pdf.

[16] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models [EB/OL]. (2020 - 01 - 23)[2025 - 04 - 07]. https://arxiv.org/pdf/2001.08361.pdf.

[17] GUO D Y, YANG D J, ZHANG H W, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning [EB/OL]. (2025 - 01 - 22)[2025 - 04 - 07]. https://arxiv.org/pdf/2501.12948.pdf.

[18] BOMMASANI R, HUDSON D A, ADELI E, et al. On the opportunities and risks of foundation models [EB/OL]. (2022 - 07 - 12)[2025 - 04 - 07]. https://arxiv.org/pdf/2108.07258.pdf.

[19] KEYDEL E R, LEE S W, MOORE J T. MSTAR extended operating conditions: a tutorial [J]. Algorithms for Synthetic Aperture Radar Imagery Ⅲ, 1996, 2757: 228 - 242.

[20] HUANG L Q, LIU B, LI B Y, et al. OpenSARShip: a dataset dedicated to sentinel-1 ship interpretation[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2018, 11(1): 195 - 208.

[21] WANG Y Y, WANG C, ZHANG H, et al. A SAR dataset of ship detection for deep learning under complex backgrounds[J]. Remote Sensing, 2019, 11(7): 765.

[22] WEI S J, ZENG X F, QU Q Z, et al. HRSID: a high-resolution SAR images dataset for ship detection and instance segmentation [J]. IEEE Access, 2020, 8: 120234 - 120254.

[23] HOU X Y, AO W, SONG Q, et al. FUSAR-Ship: building a high-resolution SAR-AIS matchup dataset of Gaofen-3 for ship detection and recognition [J]. Science China Information Sciences, 2020, 63(4): 140303.

[24] ZHANG T W, ZHANG X L, LI J W, et al. SAR ship detection dataset (SSDD): official release and comprehensive data analysis[J]. Remote Sensing, 2021, 13(18): 3690.

[25] WANG Z R, ZENG X, YAN Z Y, et al. AIR-PolSAR-Seg: a large-scale data set for terrain segmentation in complex-scene PolSAR images [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 3830 - 3841.

[26] XIA R F, CHEN J, HUANG Z X, et al. CRTransSar: a visual transformer based on contextual joint representation learning for SAR ship detection[J]. Remote Sensing, 2022, 14(6): 1488.

[27] 王智睿, 康玉卓, 曾璇, 等. SAR-AIRcraft-1.0:高分辨率SAR飞机检测识别数据集[J]. 雷达学报, 2023,12(4): 906 – 922.
WANG Z R, KANG Y Z, ZENG X, et al. SAR-AIRcraft-1.0: high-resolution SAR aircraft detection and recognition dataset [J]. Journal of Radars, 2023, 12(4): 906 – 922. (in Chinese)

[28] BOUNTOS N I, SDRAKA M, ZAVRAS A, et al. Kuro Siwo: 33 billion m^2 under the water. A global multi-temporal satellite dataset for rapid flood mapping [C]//Proceedings of the 38th Conference on Neural Information Processing Systems, 2024.

[29] LI Y X, LI X, LI W J, et al. SARDet-100K: towards open-source benchmark and ToolKit for large-scale SAR object detection [C]//Proceedings of the 38th Conference on Neural Information Processing Systems, 2024.

[30] ZHANG X, YANG X, LI Y X, et al. RSAR: restricted state angle resolver and rotated SAR benchmark[C]//Proceedings of the IEEE/CVFConference on Computer Vision and Pattern Recognition (CVPR), 2025: 7416 – 7426.

[31] WU Y M, SUO Y X, MENG Q B, et al. FAIR-CSAR: a benchmark dataset for fine-grained object detection and recognition based on single-look complex SAR images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2025, 63: 5201022.

[32] LIU Y X, LI W J, LIU L, et al. ATRNet-STAR: a large dataset and benchmark towards remote sensing object recognition in the wild [EB/OL]. (2025 – 03 – 13)[2025 – 04 – 07]. https://arxiv.org/pdf/2501.13354.pdf.

[33] DU Y T, CHEN Y S, HUANG L B, et al. SUMMIT: a SAR foundation model with multiple auxiliary tasks enhanced intrinsic characteristics[J]. International Journal of Applied Earth Observation and Geoinformation, 2025, 141: 104624.

[34] LUO J L, WANG Y Z, GU Z Q, et al. MMM-RS: a multi-modal, multi-GSD, multi-scene remote sensing dataset and benchmark for text-to-image generation[C]//Proceedings of 38th Conference on Neural Information Processing Systems, 2024.

[35] ZHANG W, CAI M X, ZHANG T, et al. EarthGPT: a universal multimodal large language model for multisensor image comprehension in remote sensing domain[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 5917820.

[36] MA Z M, XIAO X Y, DONG S H, et al. SARChat-Bench-2M: a multi-task vision-language benchmark for SAR image interpretation[EB/OL]. (2025 – 03 – 04)[2025 – 06 – 27]. https://arxiv.org/pdf/2502.08168.pdf.

[37] WEI Y M, XIAO A R, REN Y X, et al. SARLANG-1M: a benchmark for vision-language modeling in SAR image understanding [EB/OL]. (2025 – 04 – 04)[2025 – 06 – 27]. https://arxiv.org/pdf/2504.03254.pdf.

[38] SONI S, DUDHANE A, DEBARY H, et al. EarthDial: turning multi-sensory earth observations to interactive dialogues[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025: 14303 – 14313.

[39] XIONG Z T, WANG Y, YU W K, et al. GeoLangBind: unifying earth observation with agglomerative vision-language foundation models[EB/OL]. (2025 – 03 – 08)[2025 – 04 – 07]. https://arxiv.org/pdf/2503.06312.pdf.

[40] GHAMISI P, YU W K, ZHANG X K, et al. Geospatial foundation models to enable progress on sustainable development goals[EB/OL]. (2025 – 05 – 30)[2025 – 06 – 27]. https://arxiv.org/pdf/2505.24528.pdf.

[41] ASTRUC G, GONTHIER N, MALLET C, et al. AnySat: one earth observation model for many resolutions, scales, and modalities[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025: 19530 – 19540.

[42] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278 – 2324.

[43] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMa: open and efficient foundation language models [EB/OL]. (2023 – 02 – 27)[2025 – 04 – 07]. https://arxiv.org/pdf/2302.13971.pdf.

[44] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]// Proceedings of the 38th International Conference on Machine Learning, 2021.

[45] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold [J]. Nature, 2021, 596(7873): 583 – 589.

[46] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: hierarchical vision Transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 9992 – 10002.

[47] XU W J, XU Y F, CHANG T, et al. Co-scale conv-attentional image transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 9961 – 9970.

[48] LIU Z, MAO H Z, WU C Y, et al. A ConvNet for the 2020s[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 11966 – 11976.

[49] WOO S, DEBNATH S, HU R H, et al. ConvNeXt V2: co-designing and scaling ConvNets with masked autoencoders[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 16133 – 16142.

[50] LIU X Y, WU Y, LIANG W K, et al. High resolution SAR image classification using global-local network structure based on vision transformer and CNN[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 4505405.

[51] BAO H B, DONG L, PIAO S H, et al. BEiT: BERT pre-training of image transformers [EB/OL]. (2022 – 09 – 03)[2025 – 04 – 07]. https://arxiv.org/pdf/2106.08254.pdf.

[52] ZHOU B L, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 2921 – 2929.

[53] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017: 618 – 626.

[54] MA J, BAI Y L, ZHONG B N, et al. Visualizing and understanding patch interactions in vision transformer[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(10): 13671 – 13680.

[55] YEH C K, HSIEH C Y, SUGGALA A, et al. On the (in) fidelity and sensitivity of explanations[C]//Proceedings of the

33rd Conference on Neural Information Processing Systems, 2019.

[56] HINTERSDORF D, STRUPPEK L, KERSTING K, et al. Finding NeMo: localizing neurons responsible for memorization in diffusion models[C]//Proceedings of the 38th Conference on Neural Information Processing Systems, 2024.

[57] DRAVID A, GANDELSMAN Y, EFROS A A, et al. Rosetta neurons: mining the common units in a model zoo[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023: 1934 – 1943.

[58] RASCHKA S. Model evaluation, model selection, and algorithm selection in machine learning[EB/OL]. (2020 – 11 – 11)[2025 – 04 – 07]. https://arxiv.org/pdf/1811.12808.pdf.

[59] SCHEIBENREIF L, HANNA J, MOMMERT M, et al. Self-supervised vision transformers for land-cover segmentation and classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022: 1421 – 1430.

[60] YAN Z Y, LI J X, LI X X, et al. RingMo-SAM: a foundation model for segment anything in multimodal remote-sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5625716.

[61] FULLER A, MILLARD K, GREEN J. CROMA: remote sensing representations with contrastive radar-optical masked autoencoders[C]//Proceedings of 37th Conference on Neural Information Processing Systems, 2023.

[62] GUO X, LAO J W, DANG B, et al. SkySense: a multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024: 27662 – 27673.

[63] LI W J, YANG W, LIU T P, et al. Predicting gradient is better: exploring self-supervised learning for SAR ATR with a joint-embedding predictive architecture[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2024, 218: 326 – 338.

[64] GUO W L, LI S Y, YANG J. Scattering prompt tuning: a fine-tuned foundation model for SAR object recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2024: 3056 – 3065.

[65] WANG Y, HERNÁNDEZ H H, ALBRECHT C M, et al. Feature guided masked autoencoder for self-supervised learning in remote sensing[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024, 18: 321 – 336.

[66] PU X Y, JIA H C, ZHENG L H, et al. ClassWise-SAM-adapter: parameter-efficient fine-tuning adapts segment anything to SAR domain for semantic segmentation[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2025, 18: 4791 – 4804.

[67] LI W J, YANG W, HOU Y N, et al. SARATR-X: towards building a foundation model for SAR target recognition[J]. IEEE Transactions on Image Processing, 2025, 34: 869 – 884.

[68] BI H B, FENG Y C, TONG B Y, et al. RingMoE: mixture-of-modality-experts multi-modal foundation models for universal remote sensing image interpretation[EB/OL]. (2025 – 04 – 04)[2025 – 05 – 01]. https://arxiv.org/abs/2504.03166.

[69] HE K M, CHEN X L, XIE S N, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 15979 – 15988.

[70] CHEN Z, WU J N, WANG W H, et al. Intern VL: scaling up vision foundation models and aligning for genericvisual-linguistic tasks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024: 24185 – 24198.

[71] LU H Y, LIU W, ZHANG B, et al. DeepSeek-VL: towards real-world vision-language understanding[EB/OL]. (2024 – 03 – 11)[2024 – 05 – 01]. https://arxiv.org/pdf/2403.05525.pdf.

[72] YE Q H, XU H Y, XU G H, et al. mPLUG-Owl: modularization empowers large language models with multimodality[EB/OL]. (2024 – 03 – 29)[2024 – 05 – 01]. https://arxiv.org/pdf/2304.14178.pdf.

[73] BAI S, CHEN K Q, LIU X J, et al. Qwen2.5-VL technical report[EB/OL]. (2025 – 02 – 19)[2025 – 05 – 01]. https://arxiv.org/pdf/2502.13923.pdf.

[74] 陈思伟. 成像雷达极化旋转域解译: 理论与应用[M]. 北京: 科学出版社, 2024.
CHEN S W. Imaging radar polarimetric rotation domain interpretation theory and application[M]. Beijing: Science Press, 2024. (in Chinese)

[75] LI Y C, SUN X G, CHEN H X, et al. Attention is not the only choice: counterfactual reasoning for path-based explainable recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(9): 4458 – 4471.

[76] REN Q H, GAO J Y, SHEN W, et al. Where we have arrived in proving the emergence of sparse symbolic concepts in AI models[EB/OL]. (2024 – 09 – 13)[2025 – 05 – 01]. https://arxiv.org/abs/2305.01939.

[77] 阮航, 崔家豪, 毛秀华, 等. SAR 目标识别对抗攻击综述: 从数字域迈向物理域[J]. 雷达学报, 2024, 13(6): 1298 – 1326.
RUAN H, CUI J H, MAO X H, et al. A survey of adversarial attacks on SAR target recognition: from digital domain to physical domain[J]. Journal of Radars, 2024, 13(6): 1298 – 1326. (in Chinese)

[78] ZANG Y, YUN T, TAN H, et al. Pre-trainedvision-language models learn discoverable visual concepts[EB/OL]. (2025 – 01 – 13)[2025 – 04 – 07]. https://arxiv.org/pdf/2404.12652.pdf.

[79] LENG S C, ZHANG H, CHEN G Z, et al. Mitigating object hallucinations in large vision-language models through visual contrastive decoding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024: 13872 – 13882.

[80] ZHANG J M, YI Q, SANG J T. Towards adversarial attack on vision-language pre-training models[C]//Proceedings of the 30th ACM International Conference on Multimedia, 2022: 5005 – 5013.