引文格式:张良培,张乐飞,袁强强.遥感大模型:进展与前瞻[J].武汉大学学报(信息科学版),2023,48(10):1574-1581.DOI:10.13203/j.whugis20230341

Citation:ZHANG Liangpei,ZHANG Lefei,YUAN Qiangqiang.Large Remote Sensing Model: Progress and Prospects[J].Geomatics and Information Science of Wuhan University,2023,48(10):1574-1581.DOI:10.13203/j.whugis20230341

遥感大模型:进展与前瞻

张良培[1] 张乐飞[2] 袁强强[3]

1 武汉大学测绘遥感信息工程国家重点实验室,湖北 武汉,430079

2 武汉大学计算机学院,湖北 武汉,430072

3 武汉大学测绘学院,湖北 武汉,430079

摘 要:近年来,人工智能领域大语言模型和视觉基础模型的显著进展引发了学者们对遥感领域通用人工智能技术的关注,推动了遥感信息处理大模型研究的新范式。遥感大模型也称为遥感预训练基础模型,是一种利用大量的未标注遥感图像来训练大规模深度学习模型的方法,目的是提取遥感图像中的通用特征表示,进而提高遥感图像分析任务的性能、效率和通用性。遥感大模型的研究涉及3个关键因素:预训练数据集、模型参数量和预训练技术。其中,预训练数据集和模型参数量能够随着数据和计算资源的增加而灵活地扩大,预训练技术则是提升遥感大模型性能的关键因素。以遥感大模型的预训练技术为主线,归纳分析了现有的有监督单模态预训练遥感大模型、无监督单模态预训练遥感大模型和视觉-文本联合多模态预训练遥感大模型。最后,对遥感大模型在结合遥感领域知识与物理约束、提高数据泛化性、扩展应用场景以及降低数据成本4个方面,对遥感大模型进行了展望。

关键词:遥感大模型;预训练基础模型;多模态基础模型

中图分类号:P237 **文献标识码**:A **收稿日期**:2023-09-17

DOI:10.13203/j.whugis20230341 **文章编号**:1671-8860(2023)10-1574-08

Large Remote Sensing Model: Progress and Prospects

ZHANG Liangpei[1] *ZHANG Lefei*[2] *YUAN Qiangqiang*[3]

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

2 School of Computer Science, Wuhan University, Wuhan 430072, China

3 School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China

Abstract: In recent years, significant advancements in large language models and visual foundation models in the field of artificial intelligence have attracted scholars' attention to the potential of general artificial intelligence technology in remote sensing. These studies have propelled a new paradigm in the research of large models for remote sensing information processing. Large remote sensing models, also known as pre-trained foundation remote sensing models, are a kind of methodology that employs a vast amount of unlabeled remote sensing images to train large-scale deep learning models. The goal is to extract universal feature representations from remote sensing images, thereby enhancing the performance, efficiency, and versatility of remote sensing image analysis tasks. Research on large remote sensing models involves three key factors, including pre-training datasets, model parameters, and pre-training techniques. Among them, pre-training datasets and model parameters can be flexibly expanded with the increase in data and computational resources, while pre-training techniques are critical for improving the performance of large remote sensing models. This review focuses on the pre-training techniques of large remote sensing models and systematically analyzes the existing supervised single-modal pre-trained large remote sensing models, unsupervised single-modal pre-trained large remote sensing models, and visual-text joint multimodal pre-trained large remote sensing models. The conclusion section provides prospects for large remote sensing models in terms

基金项目:国家重点研发计划(2022YFB3903405)。

第一作者:张良培,博士,教授,研究方向为遥感信息处理与应用。zlp62@whu.edu.cn

of integrating domain knowledge and physical constraint, enhancing data generalization, expanding application scenarios, and reducing data costs.

Key words: large remote sensing model; pre-trained foundation model; multi-modal foundation model

随着遥感对地观测及人工智能技术的不断进步，当前的遥感图像解译技术已经广泛地应用于资源勘查、环境监测、精准农业和军事侦察等领域。在这些研究任务中，目标检测、语义分割、场景分类和变化检测等遥感视觉任务是实际应用的基本前提。自遥感领域发展进入深度学习时代至今[1-2]，上述遥感视觉任务的精度相较于传统机器学习算法已有了变革性的提升，但这些深度学习算法仍然局限于针对特定遥感视觉任务而设计特定模型，并利用相关数据集进行模型训练，得到的模型不能适用于相近任务甚至其他任务（如从场景分类到目标检测）。因此，要实现高效、准确的通用遥感图像理解仍然面临巨大挑战。

近年来，在计算机视觉和自然语言处理领域，基于 Transformer 网络的视觉基础模型（如 CLIP[3]、Florence[4]和 BEiT[5]等）和大语言模型（如 GPT-3[6]、OPT[7]和 T5[8]等）在视觉和语言理解任务中表现出来的强大泛化性，引发了学者们对遥感预训练基础模型的极大关注，相关研究快速发展。相较于前期的深度学习模型，由于这些预训练基础模型的参数量非常巨大，因此也被称为遥感大模型。在遥感领域，已经有很多学者从扩大模型参数量的角度出发，设计了若干遥感预训练基础模型，并利用大规模数据对这些模型进行参数优化。随着遥感预训练基础模型的规模从 2 千万参数量的 JointSAREO 模型[9]发展到 3 亿参数量的 Scale-MAE 模型[10]，同时使用的训练数据量从数十万增加到百万级，遥感大模型在多个遥感视觉任务中的精度也逐渐提高。

然而，模型参数量和预训练数据量的增加终究有其上限，遥感预训练基础模型的本质是提取遥感图像中的通用特征表示，而非一味地扩大模型参数规模。因此，预训练技术成为了提升遥感大模型性能的关键因素。Chen 等[11]和 Risojević 等[12]对基于自然图像数据集预训练的权重和基于遥感图像数据集预训练的权重是否对遥感视觉任务产生影响进行了探讨。他们发现，遥感图像与自然图像在样本成像角度和地物分布上存在较大差异，因此，基于自然图像数据集的预训练权重限制了遥感预训练基础模型的性能。最近，Muhtar 等[13]进一步得出了基于遥感图像数据集预训练权重对遥感视觉任务至关重要的结论。因此，遥感大模型领域的最新研究不仅需要关注模型参数量和训练数据量，更需要注重设计适用于遥感视觉任务的预训练方法。此外，现有的遥感视觉任务主要关注如何从图像中提取更鲁棒、强健的视觉特征来执行各种任务，忽视了利用多模态遥感数据对于地物关系的语义理解。如在进行遥感图像语义分割任务时，如果建筑物屋顶的像素在形状、纹理和结构上与公路相似，仅基于视觉的模型可能会将建筑屋顶的像素分类为公路，这是因为视觉模型缺乏公路不能在建筑屋顶的通用知识。

为了全面地介绍现有遥感大模型的研究进展，本文首先分别从有监督单模态预训练遥感大模型、无监督单模态预训练遥感大模型和视觉-文本联合多模态预训练遥感大模型等方面对遥感预训练基础模型的进展进行回顾和总结；然后从遥感大模型在结合遥感领域知识与物理约束、提高数据泛化性、扩展应用场景和降低数据成本 4 个方面，对遥感大模型进行了展望。

1　遥感大模型预训练技术

为了使得遥感图像的高级抽象特征能在多种下游任务中最大限度地提升性能，许多学者提出了多种预训练技术。根据有无训练标签、有无文本信息，本文将遥感大模型预训练技术分为三大类：有监督单模态预训练、无监督单模态预训练和视觉-文本联合多模态预训练，其中，3 类方法的代表性算法包括 RSP[14]、GLCNet[15]、RVSA[16]、RemoteCLIP[17]和 RSGPT[18]等（如图 1 所示）。

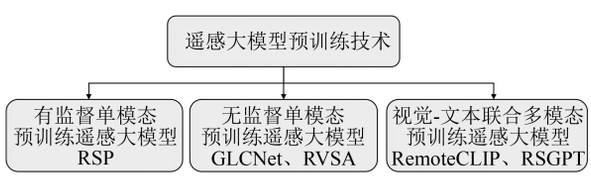


图 1　现有遥感大模型预训练技术的分类体系

Fig. 1　Existing Taxonomy of Remote Sensing Pre-trained Foundation Models

1.1　有监督单模态预训练遥感大模型

有监督单模态预训练是指在有标签的大规

模单一数据模态(如图像、文本、声音等)上进行预训练,以使模型在特定监督任务上具有更好的性能,其预训练框架如图2所示。Wang等[14]首次面向遥感领域,使用百万级遥感场景识别数据集(MillionAID)进行了有监督单模态预训练的探索,并在多个下游任务上验证了遥感视觉基础模型的性能。该团队利用获得的一系列卷积神经网络和视觉Transformer网络的骨干模型权重,对遥感预训练基础模型在场景识别、语义分割、目标检测和变化检测任务上的性能进行了评估。研究发现,当特定下游任务所需的表示粒度接近上游预训练任务时,遥感有监督预训练通常会导致更好的性能。虽然Fuller等[19]和Noman等[20]也开始对Transformer网络进行有监督单模态预训练,但是二者均是针对特定下游任务设计了新模型,并且仅在单一任务上进行性能验证。

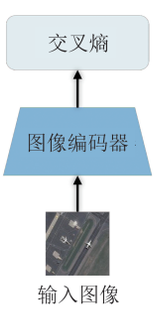


图2 有监督单模态预训练框架

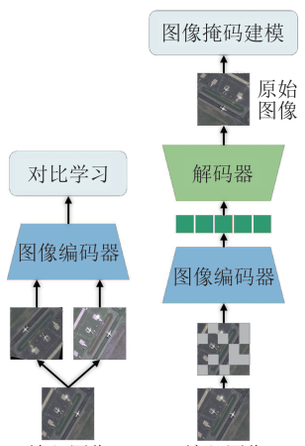
Fig. 2 Supervised Single-Modality Pre-training Framework

1.2 无监督单模态预训练遥感大模型

无监督单模态预训练是指在大规模单一数据模态上使用未标记的数据来预先训练模型,以学习数据的通用表示。自监督预训练是无监督单模态预训练的一种重要代表方法,其目标是使模型能够更好地捕捉数据中的结构和模式,从而在后续的任务中具有更好的性能。此外,无监督单模态预训练技术还包括稀疏编码[21]和自编码器[22]。在遥感任务中,Risojević等[12]首次注意到,自监督学习使用广泛且多样化的数据集训练的模型具有增强的稳健性。然而,他们缺少关键的实验验证。在这一启发下,遥感领域发展出了越来越多的自监督单模态预训练方法。根据自监督学习的框架和目标函数,可以将这些自监督单模态预训练方法分为图像对比自监督学习和图像掩码建模自监督学习两种,如图3所示。

1.2.1 图像对比自监督学习

遥感图像对比自监督学习是一种特征学习方法,它利用单张遥感图像来创建正样本和负样本对,通过将同类样本拉近,将不同类样本推远,以获得具有稳健性的特征表示。根据遥感图像正、负样本对的构建方式,可以进一步将遥感图像对比自监督学习分为实例级对比自监督学习和时间序列级对比自监督学习。



(a) 图像对比自监督学习 (b) 图像掩码建模自监督学习

图3 无监督单模态预训练框架

Fig. 3 Unsupervised Single-Modality Pre-training Framework

实例级对比自监督学习是遥感领域常见的自监督学习技术,它将相同类别图像的不同增强视角视为正样本,而不同类别图像的视角被视为负样本。在遥感领域,图像对比自监督学习主要以SimCLR方法作为遥感图像预训练的基准框架[23]。然而,SimCLR框架偏向于学习图像级表征,这会造成像素级信息丢失,从而不利于像素级遥感视觉任务。Li等[15]不仅利用风格特征来更好地学习图像级表征,还关注像素级特征匹配以消除图像级表征带来的信息损耗问题。尽管这些方法探索了对比自监督学习在遥感领域的潜力,但它们没有深入研究基于遥感图像的对比预训练模型在多种下游任务中的性能,时间序列级对比自监督学习进一步实现了对遥感预训练基础模型的促进作用,该类方法将在空间上相同但在时间上不同的样本视为正样本对。Ayush等[24]和Mañas等[25]都将不同时相而相同空间位置上的遥感图像引入对比自监督学习,并证明了自监督预训练方式在场景识别、语义分割、目标检测和变化检测任务中的促进作用。此外,Mañas等[25]深入研究了多时相遥感图像作为多种增强方式,进一步促进对比自监督预训练的表征能力。这两种方法证明了在遥感领域使用对比自监督预训练方式能够增强模型在多种下游任

务上的性能。然而，这两种方法存在和实例级对比自监督学习方法同样的问题，即没有对骨干模型的参数量进一步增大。

1.2.2　图像掩码建模自监督学习

图像掩码自监督学习框架首次由 He 等[26]提出，该框架使用视觉 Transformer 的编码器-解码器结构来重建随机掩码的图像块来学习鲁棒的图像表征，如图 3(b)所示。

已有许多工作[27-30]仅针对单一任务采用图像掩码自监督学习框架来增强遥感大模型的性能。Wang 等[16]发现，图像掩码自监督学习框架适用于大规模视觉模型，并且在加速模型训练的同时确保了模型的泛化性能。因此，针对遥感图像中大尺寸和任意方向的地物，他们设计了一种新的旋转多尺度窗口注意力(rotated varied-size attention，RVSA)，显著降低了计算成本和内存占用。RVSA 是全球首个面向遥感任务设计的亿级视觉 Transformer 基础模型，并在目标检测、语义分割、场景识别和变化检测遥感任务中取得了优异的性能。针对遥感小目标可能被随机掩码完全掩盖的问题，RingMo 遥感视觉基础模型认为在图像块中进行子随机掩码有助于保留遥感小目标信息，并在多个下游任务上验证了方案的有效性[31]。然而，在图像块中进行子随机掩码的操作无形中增加了遥感预训练基础模型的计算成本和内存占用。Reed 等[10]采用一种更为巧妙的地面采样距离位置编码来解决遥感图像地物尺度不一引起的遥感小目标问题。SatMAE[32]、Presto[33]和 ConSecutive PreTraining (CSPT)[34]是针对遥感图像的多时相特性和多源性构建的遥感预训练基础模型。其中，SatMAE 和 Presto 都采用了掩码建模的自监督学习方法，用于处理多时相的遥感图像。SatMAE 则引入了多时相掩码设计以提高模型在时间维度上的鲁棒性，而 Presto 则着眼于创建可大规模部署的通用遥感模型，这两种方法在各种下游任务中都表现出色。

随着视觉基础模型在遥感领域的潜力引起了越来越多的关注，上述预训练基础模型主要集中在预训练方法和数据集大小上，对模型参数数量的重视相对有限。Cha 等[35]通过将多头自注意力和前馈层以并行方式配置，构建了首个面向遥感任务的十亿级视觉 Transformer 基础模型。虽然该模型在参数量上大幅增加，但是它在下游任务上的性能提升幅度并不显著。因此，预训练基础模型不仅要关注预训练方法、数据集大小和模型参数量，还应该面向遥感图像多尺度、多粒度的特点进行更有效的结构设计。

1.3　视觉-文本联合多模态预训练遥感大模型

单模态遥感视觉基础模型主要关注视觉理解任务，而忽视了对象及其关系的语义理解。视觉语言模型不仅可以识别图像中的对象，还可以推断它们之间的关系，以及生成图像的自然语言描述，这使它们更适用于需要同时进行视觉和文本理解的任务，如图像字幕生成、基于文本的图像检索和视觉问题回答等任务。受此启发，遥感领域有关多模态通用基础模型的研究也越来越受重视。根据多模态预训练是否需要融合图像-文本信息，本文将视觉-文本联合多模态预训练分为图文双流对比预训练和图文双流融合预训练两种，如图 4 所示。

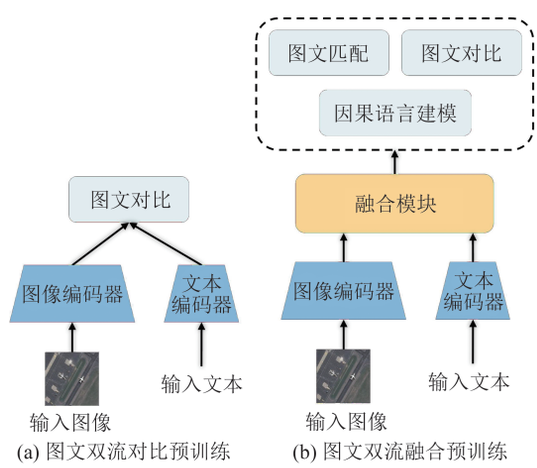


图 4　视觉-文本联合多模态预训练框架

Fig. 4　Multimodal Remote Sensing Pre-training with Visual-Text Integration Framework

1.3.1　图文双流对比预训练

图文双流对比预训练框架首次由 OpenAI 团队在对比语言-图像预训练(contrastive language-image pre-training，CLIP)模型[3]中提出，该框架利用图像文本对比目标函数度量图像和文本之间的相似性和差异性，并建立图像和文本之间的联系。然而，在遥感领域，与遥感图像相对应的精确文本描述往往难以获取，这导致遥感领域中与图文双流对比预训练相关的研究工作相对较少。Bazi 等[36]和 Mikriukov 等[37]以 CLIP 架构为基础，分别在遥感视觉问题回答任务和基于文本的图像检索任务中首次使用图像文本对比目标函数，以保持模态内和模态间的相似性。然而，这两个研究仅在相对较小的遥感图像文本对数据集上进行了对比预训练，而且分别只在两个单一的遥感任务上验证了模型的性能，因此并未充

分展示视觉-文本联合多模态预训练在其他潜在任务上的潜力。针对遥感图像文本对数据集规模较小的问题,RemoteCLIP将异构注释转换为基于Box-to-Caption和Mask-to-Box转换的统一图像-字幕数据格式,构建了一个大规模图像文本对预训练数据集[17]。RemoteCLIP在构建的大规模数据集上利用CLIP框架进行图像文本对比学习后,其性能在场景识别、少样本分类、图像-文本检索以及遥感图像目标计数任务中均有较大提升。

1.3.2 图文双流融合预训练

图文双流融合预训练通过交叉注意力构建的融合模块来深度关联图像和文本特征,并使用多种目标函数度量图像和文本间的相似性,如图像文本对比损失、图像文本匹配损失、图像引导的文本建模损失等。

在遥感领域中,Prompt-RSVQA首次利用图文双流融合预训练方法来提升遥感视觉问题回答任务的性能[38]。Liu等[39]首次在遥感图像变化描述任务中引入了图像双流融合与训练方法,通过融合模块来充分利用冻结的图像编码器和冻结的大语言模型的稳健表征能力以及推理能力,这个思想继承自视觉-文本联合多模态预训练模型——BLIP-2[40]。与Liu等[39]的工作类似,Wei等[41]提出了双语遥感图像描述任务,并借鉴了BLIP-2的设计思想来构建模型,但是该方法在性能上并没有大幅领先传统遥感图像描述方法。Hu等[18]和Zhang等[42]以BLIP-2为基础预训练模型,并分别构建了两个新的遥感图像文本对预训练数据集——RSICap和RS5M。两种方法均在遥感图像字幕生成、基于文本的遥感图像检索和遥感视觉问题回答任务上进行了测试,并取得了优异的结果。尽管图文双流融合预训练模型在提升遥感图像文本理解任务和图像文本生成任务方面表现出了一定的性能,但是该模型尚未进行充分的性能验证以适应下游的纯视觉任务。因此,视觉-文本联合遥感通用基础模型在单模态遥感任务和多模态遥感任务中的有效性还有待更加深入的研究。

2 遥感大模型未来展望

如前所述,随着遥感对地观测技术和通用人工智能的发展,研究人员在遥感大模型方面已经取得了若干研究成果,无论是预训练数据集规模、预训练大模型性能还是预训练自适应程度都已取得了重要进展。但目前的遥感大模型面对数据源的多样性、预训练大模型的高效性、泛化性与可靠性等难题依然存在挑战。本文对遥感大模型在结合遥感领域知识与物理约束、提高数据泛化性、扩展应用场景以及降低数据成本4个方面进行展望。

2.1 结合遥感领域知识与物理约束

目前遥感大模型在表征能力和推理能力上表现出强大性能,然而,这些模型在高效部署及可解释性等方面仍存在研究空白。由于数据驱动模型的黑匣子特性,研究人员不仅难以合理解释遥感大模型的决策,也不能通过可解释性原则设计更高效、鲁棒的遥感大模型[43]。在计算机视觉与自然语言处理领域,已有学者提出多领域知识的大、小模型协同表达理论框架,该框架能够同时发挥大、小模型的通用性和专用性优势,使得预测结果具备更强的鲁棒性和可解释性[44]。大、小模型协同表达通常利用知识图谱等方式构建特定领域的结构化知识来提升大模型的能力和可解释性[45]。通过将地物的形状、纹理等人工特征以及地物光谱特性等遥感特定先验信息构建多重知识图谱,进一步约束遥感领域小模型的自由度,是实现遥感大模型在垂直领域高效部署、提升遥感大模型可解释性的重要方向之一。

2.2 提高遥感大模型的数据泛化性

尽管最近遥感大模型展现出了强大的泛化能力,但它们在实际应用部署中仍然受限于训练-测试图像分布的差异。遥感领域的训练-测试图像分布的差异主要来源于成像传感器参数、成像气候条件以及地形地貌的多样性等多方面。在计算机视觉领域中,提示学习方法和特征自适应方法已经被探索用于提高大模型在自然图像上的数据泛化性,同时减少了大规模微调的需求。视觉-语言大模型提示学习通过找到最优的提示而不是微调整个视觉-语言大模型,以使视觉-语言大模型高效地适应下游任务场景[46],现有的相关的研究可以分为文本提示学习、视觉提示学习和文本-视觉提示学习3类[47-49]。除此以外,特征自适应方法也被广泛探索,它的核心思想是通过使用一个额外的轻量级特征适配器来微调大模型[50]。由于提示学习方法和特征自适应方法简单有效的特性,将它们应用在遥感领域来提高遥感预训练大模型的泛化性将是有前景的方向。

2.3 扩展遥感大模型的应用场景

目前,遥感多模态大型模型主要采用了CLIP或者BLIP架构进行训练,这些模型使用了

数十亿规模的遥感图像文本对进行训练,在学习遥感图像的视觉语义特征的同时也学习了与这些视觉特征良好对齐的文本嵌入。然而,由于视觉-语言多模态大模型通常是为图像级别表示设计的架构进行预训练的,对于一些小尺度的任务(如目标检测、像素分割)并不适用[51]。在计算机视觉领域中,知识蒸馏技术最近被用来提高视觉-语言大模型在细粒度下游任务中的应用场景。这项技术主要聚焦于将图像级别的知识转移到区域或像素级别的任务,让大模型在任意开放环境的目标检测和语义分割等任务上也可以发挥较好的效果[52-53]。这类方法通常将预训练的视觉-语言模型作为教师模型,把预训练的视觉-语言模型学习到的知识蒸馏到检测或分割模型中,上述过程可以通过提示学习、知识对齐和伪标签使用等技术实现[54]。通过知识蒸馏技术高效挖掘遥感大型模型的内在知识,对于进一步扩展遥感大模型的应用场景便显得尤为迫切。

2.4　降低遥感大模型的数据成本

当前遥感大模型的训练过程需要大量的训练数据支撑,因此,更加高效经济的方式来采集或生成高质量的多模态遥感训练数据便成为一个重要的研究方向。近两年,扩散模型因其能够生成更真实的高质量图像而在计算机视觉领域受到了大量关注。与变分自动编码器和生成对抗网络等生成网络不同,扩散模型在前向阶段对图像逐步施加噪声,直至图像被破坏变成完全的高斯噪声,然后在逆向阶段学习从高斯噪声还原为原始图像过程[55]。扩散模型已在各类图像增强、图像生成、风格转换等任务上都取得了令人惊叹的效果[56-57]。因此,一个自然的想法是通过使用扩散模型基于现有文本描述生成遥感图像,这样就能够利用生成的合成数据有效地扩大数据集的规模,帮助提高遥感大模型的鲁棒性和泛化能力。此外,通过将风格转移或域适应等技术与遥感图像的特点相结合,还可以进一步将扩散模型进行适应性改进,生成更加多样化且贴近真实应用需求的合成遥感图像,支持遥感大模型的训练过程。

参　考　文　献

[1] Zhang L P, Zhang L F, Du B. Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art[J]. *IEEE Geoscience and Remote Sensing Magazine*, 2016, 4(2): 22-40.

[2] Chen Y S, Lin Z H, Zhao X, et al. Deep Learning-based Classification of Hyperspectral Data [J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2014, 7(6): 2094-2107.

[3] Radford A, Kim J W, Hallacy C, et al. Learning Transferable Visual Models from Natural Language Supervision[C]//International Conference on Machine Learning, Salt Lake City, USA, 2021.

[4] Yuan L, Chen D D, Chen Y L, et al. Florence: A New Foundation Model for Computer Vision[EB/OL]. (2021-09-25)[2023-05-23]. https://arxiv.org/abs/2111.11432.

[5] Bao H, Dong L, Piao S, et al. BEiT: BERT Pre-Training of Image Transformers[C]//The Tenth International Conference on Learning Representations, Vienna, Austria, 2022.

[6] Brown T B, Mann B, Ryder N, et al. Language Models are Few-shot Learners[C]//The 34th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2020.

[7] Zhang S S, Roller S, Goyal N, et al. OPT: Open Pre-trained Transformer Language Models [J]. *CoRR*, 2022, abs/2205.01068.

[8] Raffel C, Shazeer N, Roberts A, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer[J]. *Journal of Machine Learning Research*, 2020, 21(140):1-67.

[9] Wang Y, Albrecht C M, Zhu X X. Self-supervised Vision Transformers for Joint SAR-optical Representation Learning[C]//IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 2022.

[10] Reed C J, Gupta R, Li S F, et al. Scale-MAE: A Scale-aware Masked Autoencoder for Multiscale Geospatial Representation Learning [EB/OL]. (2022-12-30) [2023-05-23]. https://arxiv. org/abs/2212.14532.

[11] Chen Z L, Wang Y Y, Han W, et al. An Improved Pre-training Strategy-based Scene Classification with Deep Learning[J]. *IEEE Geoscience and Remote Sensing Letters*, 2020, 17(5): 844-848.

[12] Risojević V, Stojnić V. Do We Still Need ImageNet Pre-training in Remote Sensing Scene Classification? [EB/OL]. (2021-11-05)[2023-05-23]. https://arxiv.org/abs/2111.03690.

[13] Muhtar D, Zhang X L, Xiao P F. Index Your Position: A Novel Self-Supervised Learning Method for Remote Sensing Images Semantic Segmentation[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 1-11.

[14] Wang D, Zhang J, Du B, et al. An Empirical Study

of Remote Sensing Pre-training[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, 61: 1-20.

[15] Li H F, Li Y, Zhang G, et al. Global and Local Contrastive Self-Supervised Learning for Semantic Segmentation of HR Remote Sensing Images[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 1-11.

[16] Wang D, Zhang Q M, Xu Y F, et al. Advancing Plain Vision Transformer Toward Remote Sensing Foundation Model[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, 61: 1-15.

[17] Liu F, Chen D L, Guan Z, et al. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing [EB/OL]. (2023-06-19)[2023-09-15]. https://arxiv.org/abs/2306.11029.

[18] Hu Y, Yuan J L, Wen C C, et al. RSGPT: A Remote Sensing Vision Language Model and Benchmark [EB/OL]. (2023-07-28) [2023-09-10]. https://arxiv.org/abs/2307.15266.

[19] Fuller A, Millard K, Green J R. Transfer Learning with Pretrained Remote Sensing Transformers[EB/OL]. (2022-09-28)[2023-08-20]. https://arxiv.org/abs/2209.14969.

[20] Noman M, Fiaz M, Cholakkal H, et al. Remote Sensing Change Detection with Transformers Trained from Scratch [EB/OL]. (2023-03-13) [2023-08-20]. https://arxiv.org/abs/2304.06710.

[21] Aharon M, Elad M, Bruckstein AM. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation[J]. *IEEE Transactions on Signal Processing*, 2006, 54(11): 4311-4322.

[22] Hinton G E, Salakhutdinov R R. Reducing the Dimensionality of Data with Neural Networks[J]. *Science*, 2006, 313(5786): 504-507.

[23] Stojnić V, Risojević V. Self-supervised Learning of Remote Sensing Scene Representations Using Contrastive Multiview Coding[EB/OL]. (2023-03-14) [2023-05-23]. https://arxiv.org/abs/2104.07070.

[24] Ayush K, Uzkent B, Meng C L, et al. Geography-Aware Self-supervised Learning[EB/OL]. (2020-11-19)[2023-05-23]. https://arxiv.org/abs/2011.09980.

[25] Mañas O, Lacoste A, Giro-i-Nieto X, et al. Seasonal Contrast: Unsupervised Pre-training from Uncurated Remote Sensing Data[EB/OL]. (2021-04-30) [2023-05-23]. https://arxiv.org/abs/2103.16607.

[26] He K M, Chen X L, Xie S N, et al. Masked Autoencoders are Scalable Vision Learners[EB/OL]. (2022-12-30) [2023-05-23]. https://arxiv.org/abs/2111.06377.

[27] Muhtar D, Zhang X L, Xiao P F, et al. CMID: A Unified Self-Supervised Learning Framework for Remote Sensing Image Understanding[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, 61: 1-17.

[28] Zhang Y X, Zhao Y, Dong Y N, et al. Self-Supervised Pre-training via Multimodality Images with Transformer for Change Detection[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 1024, 61: 1-11.

[29] Li Y Y, Alkhalifah T, Huang J P, et al. Self-supervised Pre-training Vision Transformer with Masked Autoencoders for Building Subsurface Model[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, DOI: 10.1109/TGRS.2023.3308999.

[30] Zhang T, Zhuang Y, Chen H, et al. Object-centric Masked Image Modeling-based Self-supervised Pre-training for Remote Sensing Object Detection[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023, 16: 5013-5025.

[31] Sun X, Wang P J, Lu W X, et al. RingMo: A Remote Sensing Foundation Model with Masked Image Modeling[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 61: 1-22.

[32] Cong Y Z, Khanna S, Meng C L, et al. SatMAE: Pre-training Transformers for Temporal and Multi-spectral Satellite Imagery[EB/OL]. (2022-12-30) [2023-05-23]. https://arxiv.org/abs/2207.08051.

[33] Tseng G, Zvonkov I, Purohit M, et al. Lightweight, Pre-trained Transformers for Remote Sensing Timeseries [EB/OL]. (2022-12-30) [2023-05-23]. https://arxiv.org/abs/2304.14065.

[34] Zhang T, Gao P, Dong H, et al. Consecutive Pre-Training: A Knowledge Transfer Learning Strategy with Relevant Unlabeled Data for Remote Sensing Domain[J]. *Remote Sensing*, 2022, 14(22): 5675.

[35] Cha K, Seo J, Lee T. A Billion-scale Foundation Model for Remote Sensing Images [EB/OL]. (2022-12-30) [2023-05-23]. https://arxiv.org/abs/2304.05215.

[36] 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.

[37] Mikriukov G, Ravanbakhsh M, Demir B. Deep Unsupervised Contrastive Hashing for Large-scale Cross-modal Text-image Retrieval in Remote Sensing [EB/OL]. (2022-12-30) [2023-05-23]. https://arxiv.org/abs/2201.08125.

[38] Chappuis C, Zermatten V, Lobry S, et al. Prompt-RSVQA: Prompting Visual Context to a Language Model for Remote Sensing Visual Question Answering [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, USA, 2022.

[39] Liu C Y, Zhao R, Chen J Q, et al. A Decoupling Paradigm with Prompt Learning for Remote Sensing Image Change Captioning [J]. *Geoscience*, 2023, DOI: 10.1109/TGRS.2022.3232784.

[40] Li J N, Li D X, Savarese S, et al. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models [EB/OL]. (2022-12-30) [2023-05-23]. https://arxiv.org/abs/2301.12597.

[41] Wei T T, Yuan W L, Luo J R, 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(1): 9-18.

[42] Zhang Z L, Zhao T C, Guo Y L, et al. RS5M: A Large Scale Vision-language Dataset for Remote Sensing Vision-language Foundation Model [EB/OL]. (2023-06-20) [2023-08-23]. https://arxiv.org/abs/2306.11300.

[43] Yang Y, Zhuang Y T, Pan Y H. Multiple Knowledge Representation for Big Data Artificial Intelligence: Framework, Applications, and Case Studies [J]. *Frontiers of Information Technology & Electronic Engineering*, 2021, 22(12): 1551-1558.

[44] Yang Z X, Yang Y. Decoupling Features in Hierarchical Propagation for Video Object Segmentation [EB/OL]. (2022-10-18) [2023-08-23]. https://arxiv.org/abs/2210.09782.

[45] Zhang X M, Wu C Y, Zhang Y, et al. Knowledge-enhanced Visual-Language Pre-training on Chest Radiology Images [J]. *Nature Communications*, 2023, 14: 4542.

[46] Zhou K Y, Yang J K, Loy C C, et al. Learning to Prompt for Vision-Language Models [J]. *International Journal of Computer Vision*, 2022, 130(9): 2337-2348.

[47] Zhong Y W, Yang J W, Zhang P C, et al. RegionCLIP: Region-based Language-Image Pre-training [EB/OL]. (2021-12-16) [2023-05-23]. https://arxiv.org/abs/2112.09106.

[48] Rao Y M, Zhao W L, Chen G Y, et al. DenseCLIP: Language-guided Dense Prediction with Context-Aware Prompting [EB/OL]. (202-12-02) [2023-05-23]. https://arxiv.org/abs/2112.01518.

[49] Jia M, Tang L, Chen B C, et al. Visual Prompt Tuning [C]//17th European Conference on Computer Vision, Tel Aviv, Israel, 2022.

[50] Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-Efficient Transfer Learning for NLP [EB/OL]. (2019-02-02) [2023-05-23]. https://arxiv.org/abs/1902.00751.

[51] Liu Y, Zhu G, Zhu B, et al. TaiSu: A 166M Large-scale High-Quality Dataset for Chinese Vision-Language Pre-training [C]//Annual Conference on Neural Information Processing Systems, New Orleans, USA, 2022.

[52] Rasheed H, Maaz M, Khattak M U, et al. Bridging the Gap Between Object and Image-Level Representations for Open-Vocabulary Detection [EB/OL]. (2022-07-07) [2023-05-23]. https://arxiv.org/abs/2207.03482.

[53] Mal Z, Luo G, Gao J, et al. Open-Vocabulary One-stage Detection with Hierarchical Visual-Language Knowledge Distillation [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, USA, 2022.

[54] Xie J, Zheng S. ZSD-YOLO: Zero-Shot YOLO Detection Using Vision-Language Knowledge Distillation [EB/OL]. (2021-09-24) [2023-05-23]. https://arxiv.org/abs/2109.12066.

[55] Rombach R, Blattmann A, Lorenz D, et al. High-Resolution Image Synthesis with Latent Diffusion Models [C]//IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022.

[56] Wang W L, Bao J M, Zhou W G, et al. Semantic Image Synthesis via Diffusion Models [EB/OL]. (2022-06-30) [2023-05-23]. https://arxiv.org/abs/2207.00050.

[57] Dhariwal P, Nichol A Q. Diffusion Models Beat GANs on Image Synthesis [C]//Annual Conference on Neural Information Processing Systems, Nice, France, 2021.