用于零样本分类的遥感视觉语言模型：综述

檀晓萌[1,2]，席博博[1,3]，薛长斌[1]，李云松[3]，徐海涛[1]

1. 中国科学院国家空间科学中心 复杂航天系统电子信息技术重点实验室，北京 100190;

2. 中国科学院大学，北京 100190；

3. 西安电子科技大学 通信工程学院，陕西 西安 710071

摘 要：视觉语言模型在零-小样本分类、图文检索、图像字幕、视觉问答和视觉定位等多种多模态任务上取得了显著的进展。然而，大多数方法依赖于通用数据集的预训练，导致在如遥感、医学等特殊领域的泛化性能不佳。最近提出了许多遥感视觉语言模型，这些新模型通过构建大规模的遥感图像文本对数据来微调通用视觉语言模型，以实现具备地理感知能力的遥感域专用视觉语言模型。本文以零样本分类任务为主线，总结并分析了遥感视觉语言模型的最新发展，主要是利用遥感数据对通用视觉语言模型微调来构建遥感域专用模型，整体驱动方式依赖于高质量标注的遥感数据和高性能的算力资源。此外，当前模型的发展较为分散多样，这使得遥感视觉语言模型的统一基准评价难以建立。为解决上述问题，未来或许可结合模型架构设计、微调技术等进行算力优化，同时依据多样任务特点逐步完善评价体系。

关键词：遥感智能解译，视觉语言模型，遥感视觉语言模型，模型微调技术，多模态学习，图文对齐，零样本分类，遥感数据集构建

中图分类号：TP701 /P2　　**文献标志码：**A

1 引 言

截止至 2023 年 5 月 1 日，全球发射的卫星数量已达 7560 颗，其中大约 1290 颗卫星用于进行地球观测及相关任务（https：//www.ucsusa.org/resources/satellite-database）。获取和收集大量遥感数据已成为现实，遥感大数据时代随之而来。全面丰富的遥感数据包含着大量信息，特别是遥感图像中往往隐含了许多细节和精准信息。为充分利用这些图像，越来越多研究人员开展了遥感智能解译相关工作（黄媛等, 2024；李云飞等, 2022）。以遥感图像场景分类任务为例，不少学者针对不同平台、资源及任务设计专用模型（吴强强等, 2022；张涛等, 2022）。但这些方法投入成本高，支持的任

收稿日期：2024-09- 25;
基金项目：国家自然科学基金（编号:62401434），中国科学院国家空间科学数据中心青年开放课题（编号:NSSDC2302004）
第一作者简介：檀晓萌，1997 年生，女，博士研究生，研究方向为遥感图像处理、零样本学习。E-mail：tanxiaomeng22@mails.ucas.ac.cn
通信作者简介：薛长斌，1972 年生，男，研究员，研究方向为深空探测科学载荷全流程设计与仿真技术和深度学习。E-mail：xuechangbin@nssc.ac.cn

务类型单一。

近年来，随着计算机存储、算力等相关技术的发展，可获取的各个模态数据呈爆炸式增长。基于海量数据的“大模型+下游任务”的多任务模式率先在计算机视觉领域引发大量关注（付琨等, 2024），如大规模语言模型 ChatGPT（Alec 等, 2021）掀起的轰动，以及 2023 年大规模多模态模型 GPT-4（Waisberg 等, 2023）激起的强烈反响。这一系列的技术突破使得视觉语言模型的相关研究工作迅速增多。层出不穷且更新迅速的开源/闭源模型使得 2023 年成为大规模多模态模型的爆发年。然而，这些方法都是基于通用数据集训练得到的，其在遥感、医学等特殊领域的泛化性能不佳。

以零样本分类任务为例，分别采用通用视觉语言模型 CLIP 模型（Alec 等, 2021）作为基线，RemoteCLIP（Liu 等, 2024）、GeoRSCLIP（Zhang 等, 2023）、SkyCLIP（Wang 等, 2024）等三种遥感域专用视觉语言模型在常用的零样本遥感图像分类数据集 RSSDIVCS 上进行零样本分类，得到的结果如表 1 所示。基于通用数据集预训练得到的 CLIP 模型直接应用于零样本遥感图像场景分类任务的性能较差，而基于遥感域数据集预训练微调后的模型通常表现更好。这一方面说明了视觉语言模型的优势，另一方面也强调了数据对于特殊领域视觉语言模型的重要性。

表 1 遥感视觉语言模型在 RSSDIVCS 数据集上的零样本分类精度对比

Table 1 Comparison of zero-shot classification accuracy of remote sensing visual language models on the RSSDIVCS dataset

模型	总体精度（OA/%）	平均精度（AA/%）
CLIP	54.95	44.75
RemoteCLIP	52.64	42.86
GeoRSCLIP	62.72	51.07
SkyCLIP	56.35	45.89

鉴于视觉语言模型的有效性以及多任务属性，最近涌现了一批遥感图像文本对数据集以及基于其微调的遥感域专用视觉语言模型，即“自然图像预训练模型+遥感数据微调”的模式。这些方法按照训练数据的模式，可分为对比式和预训练式，表 2 列出了当前用于零样本分类的遥感视觉语言模型、发布日期及其特点。可以看出，这类模型首次出现是在 2023 年中，仅一年就已发布了 9 个模型，相关作品的数量还在快速增长。值得注意的是，当前这类模型的创新主要集中在遥感图像文本对数据的收集整理，对模型架构及微调技术的改进相对较少。大多数工作是基于 CLIP 模型、基于 LLaVA（Liu 等, 2024b） 和 MiniGPT-4（Zhu 等, 2023）框架进行数据驱动的模型微调。

随着遥感视觉语言模型的快速发展以及其在零样本分类任务上的优异表现，跟踪和比较用于零样本分类的遥感视觉语言模型的最新研究是值得的。本文的调查聚焦于零样本分类任务，总结用于零样本分类任务的遥感视觉语言模型的最新发展，可以为从事零样本和遥感方向的研究者提供最新了解。

表 2 用于零样本分类的遥感视觉语言模型相关信息

Table 2 The parameters settings of TS geometrical optics model in sparse distribution

模型	发布日期	特点
RemoteCLIP	2023 年 6 月 19 日	首个针对遥感领域的视觉-语言基础模型，开启了大规模遥感图文对数据集构建+预训练通用视觉语言模型微调的领域迁移方案，在遥感领域实现了视觉和语言的有效融合，提升了模型在多种任务中的表现。
GeoRSCLIP	2023 年 6 月 20 日	延续了大规模遥感图文对数据集构建+预训练通用视觉语言模型微调的领域迁移方案，在模型微调中，尝试了多种参数高效微调方法，进一步提升了模型在多种任务中的性能。
GeoChat	2023 年 11 月 24 日	首个专为遥感领域设计的多任务视觉语言模型，旨在处理高分辨率遥感图像，提供多任务对话能力。
GRAFT	2023 年 12 月 12 日	针对遥感图像的视觉-语言基础模型，旨在无需任何文本注释的情况下，通过地面图像与遥感图像的对齐，训练遥感图像的视觉-语言模型。
SkyCLIP	2023 年 12 月 20 日	构建了一个语义标签的更为全面的遥感视觉语言数据集，对未见类别的泛化性能更佳。
EarthGPT	2024 年 1 月 30 日	遥感领域的多模态视觉语言模型，统一处理遥感领域的多传感器图像理解任务。
LHRS-Bot	2024 年 2 月 4 日	采用了一种新颖的多层次视觉-语言对齐策略，结合课程学习方法，增强了模型在遥感图像理解中的表现。
H2RSVLM	2024 年 3 月 29 日	通过引入高质量的训练数据和增强模型的自我感知能力，该模型在提升遥感图像理解和减少错误生成方面展现了显著的优势。
SkySenseGPT	2024 年 6 月 14 日	引入了从关系推理到图像级场景图生成等复杂理解任务。

2 零样本分类

零样本分类问题是指训练类和测试类不相交时的图像分类问题。它的目标是让模型能够识别在训练阶段从未见过的类别。零样本分类模型需要利用与未见类别相关的辅助信息（如类别描述、属性或语义嵌入等）来实现对这些类别的归类识别。

根据训练阶段是否使用未见类别图像，零样本分类算法可分为归纳式和直推式两种范式（刘靖祎等, 2021）。归纳式算法是指模型在训练阶段只访问训练类别集合的样本和所有类别的属性，而在测试阶段需要对测试类别中的样本进行分类。它的目标是构建一个能够泛化到未见类别的模型。该模型尝试从已见类别中学习到足够的信息，以便能够推广到未见类别上。这类方法通常需要已见和未见类别之间存在一定的结构关系或共享的语义空间。直推式算法则在训练阶段也加入了测试类别图像样本，在测试阶段同样对测试类别样本做出分类。它的目标是利用未标记的测试类别样本提高对类别的分类性能。当前的视觉语言模型所具有的零样本性能，

通常是指归纳式零样本分类。

视觉语言模型中的零样本分类问题可以定义为：给定一个类别图像集合 C 和一个类别文本描述集合 S，其中 C 和 S 分别被分为训练集合 C_{train}、S_{train} 和测试集合 C_{test}、S_{test}。模型在训练阶段学习 C_{train} 和 S_{train} 之间的映射关系，在测试阶段，利用 S_{train} 和 S_{test} 共享的语义空间以及 C_{train} 和 S_{train} 之间的映射关系来实现从 C_{test} 到 S_{test} 的推理，实现图像分类。

3 视觉语言模型

3.1 视觉语言模型简介

视觉语言模型是一种能够同时从图像和文本中学习的多模态模型，其输入为图像和文本，输出为

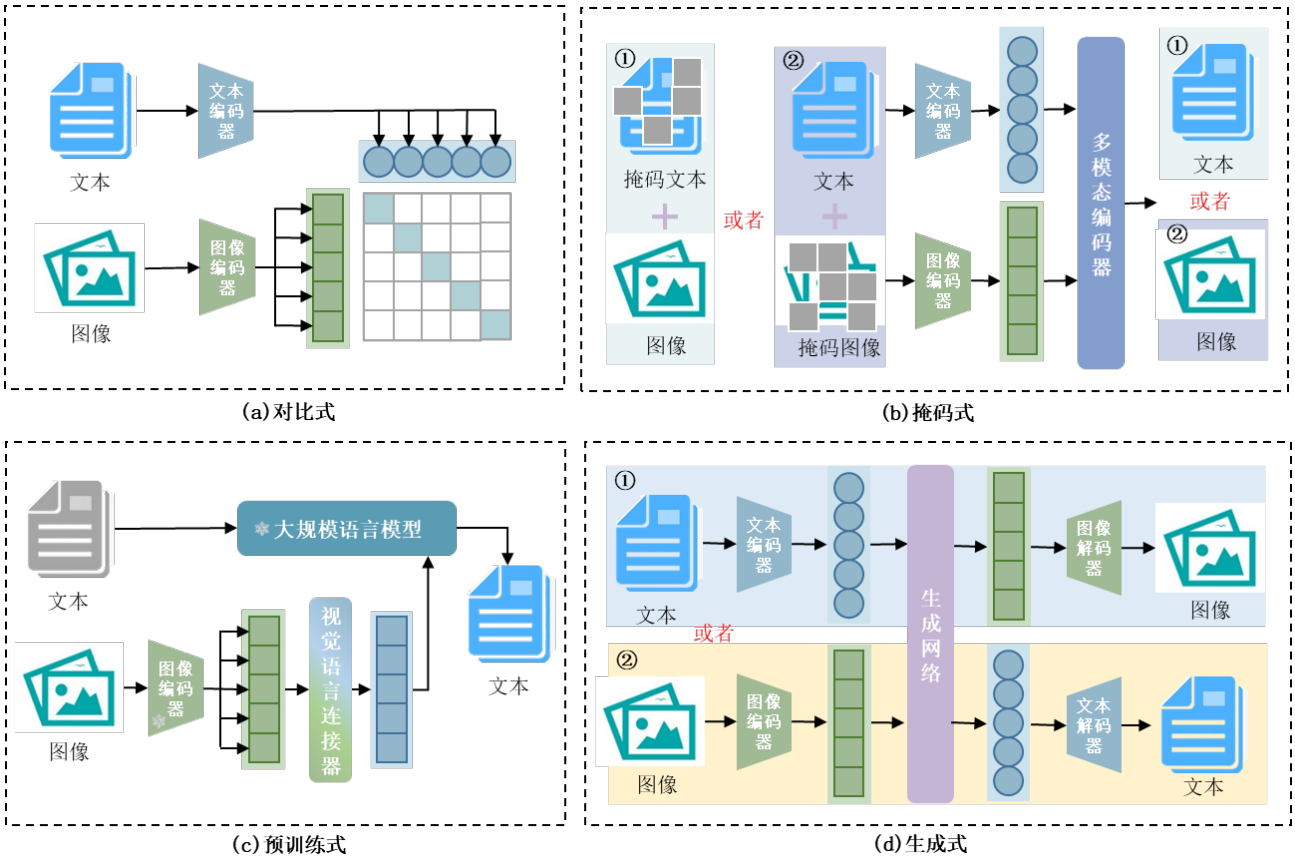

图 1 不同范式的视觉语言模型。（a）对比式（b）掩码式（c）预训练式（d）生成式

Fig.1 Visual language models with different paradigms. (a) contrastive (b) masked (c) pre-trained (d) generative

表 3 不同范式的视觉语言模型

Table 3 Visual language models with different paradigms

类型	核心组件	下游任务
对比式	图像编码器、文本编码器	图像分类、目标检测、语义分割等

掩码式	图像编码器、文本编码器、多模态编码器	图文检索、视觉问答、视觉推理、关系推理等
预训练式	图像编码器、文本编码器、视觉语言连接器	图像分类、视觉问答、图像描述、视觉定位等
生成式	文本编码器、生成网络、图像解码器/ 图像编码器、生成网络、文本解码器	图像分类、图文检索、视觉问答、视觉推理、图像描述等

文本。该模型具有良好的零样本学习能力，泛化能力强，应用广泛，支持的下游任务包括基于图像的会话，根据指令的图像识别、视觉问答、图像描述等。

当前的视觉语言模型融合了计算机视觉和自然语言处理领域的显著进展。根据模型训练模式的不同，视觉语言模型可分为以下四类，对比式、掩码式、预训练式和生成式，如图 1 所示。这四类范式的核心组件及主要支持的下游任务如表 3 所示。本节将依次简介各类视觉语言模型的训练思路、核心组件、网络架构和较为典型的方法。

（1） 对比式视觉语言模型

对比式视觉语言模型采用对比学习的思想，利用正负样本进行训练，通过预测正样本对的相似表示和负样本对的不同表示来进行训练。该范式的核心组件主要有两部分，分别是图像编码器和文本编码器，如图 1（a）所示。图像编码器采用卷积神经网络（Christian 等, 2015; Simonyan and Andrew 等, 2014; He 等, 2016）或者视觉 Transformer（Alexey 等, 2020）将图像编码为图像特征向量。并行的文本编码器通常采用 BERT 模型（Jacob 等, 2019）将文本描述处理为文本特征向量。随后，两种模态的特征向量将在共享的多维空间中对齐。最为典型的对比式视觉语言模型是 CLIP 模型（Alec 等, 2021），它利用 InfoNCE 损失函数（Oord 等, 2018）来对齐特征空间中的文本和图像。

（2） 掩码式视觉语言模型

掩码式视觉语言模型有两种训练思路，一种是通过给定部分未掩码文本来重建被掩码的图像块，即掩码图像建模。另一种是通过在标题中掩码词语，训练模型在给定未掩码图像的情况下重建这些词语，即掩码语言建模。这类模型的核心组件包括图像编码器、文本编码器和多模态编码器，如图 1（b）所示。与对比式模型类似，图像编码器通常也使用视觉 Transformer，并行的文本编码器通常也是基于 Transformer 的网络。除此之外，该范式独特的多模态编码器通常结合交叉注意力机制来整合视觉和文本信息。较为典型的方法有 FLAVA 模型（Singh 等, 2022）和 MaskVLM 模型（Kwon 等, 2022），它们均结合了掩码语言建模和掩码图像建模技术来学习文本和图像表示。

（3） 预训练式视觉语言模型

预训练式视觉语言模型通常利用开源的大型语言模型（Large Language Model，LLM）来学习图像编码器和 LLM 之间的映射。其核心组件包括预训练的图像编码器、预训练的 LLM 和视觉语言连接器，

如图 1（c）所示。该模型的学习过程更趋近于人类的感知过程。预训练的图像编码器类似人眼，接收和处理光信号。预训练的 LLM 的功能对比于人脑，理解文本信号并执行推理任务。视觉语言连接器负责将图像信息有效地投射到 LLM 可以理解的空间中以协调图像和文本信息，通常采用多层感知器（Multi-Layer Perceptron，MLP）。值得注意的是，这里的预训练图像编码器通常采用是图文对齐的对比式视觉语言模型中的图像编码器，最常用的来自于 CLIP 模型，此外，还有一些研究采用了 CLIP 模型的变体，如 EVA-CLIP 编码器。该范式的典型模型较多，例如 Frozen 模型（Tsimpoukelli 等, 2021）、Qwen 模型（Bai 等, 2023）、BLIP-2 模型（Li 等, 2023）、Mini-GPT 系列模型（Chen 等, 2023; Zheng 等, 2024; Zhu 等, 2023）和 LLaVA 系列模型（Liu 等, 2024）等。

（4） 生成式视觉语言模型

生成式视觉语言模型通过生成图像或者标题进行训练，即有两种模式，如图 1（d）所示。通过生成图像进行训练的生成式模型的核心组件包括文本编码器、生成网络和图像解码器。文本编码器和图像解码器与前两种范式类似。生成网络是文本-图像生成器，通常采用预训练的条件扩散模型。这类范式中较为典型的模型是 CoCa 模型（Yu 等, 2022），它采用多模态文本解码器作为图像-文本生成网络。通过标题进行训练的生成式模型的核心组件包括图像编码器、生成网络和文本解码器，它们的网络架构与前一模式的选择类似。这类范式中较为典型的模型是 Chameleon 模型（Chameleon 等, 2024），它采用基于 Transformer 架构的混合模态自回归语言模型作为图像-文本生成器。

3.2 遥感视觉语言模型简介

遥感视觉语言模型是一种结合了遥感技术和自然语言处理的多模态模型，旨在利用大规模、多模态遥感数据来实现具有更高效、准确的遥感图像分析。该模型的功能多样，应用广泛，包含遥感图像分类、遥感目标检测、遥感语义分割、遥感视觉定位、遥感视觉问答、遥感图像描述、遥感图像检索等多种任务。

当前遥感视觉语言模型的研究主要以数据为中心，采用“自然图像预训练模型+遥感数据微调”的模式实现了视觉语言模型的领域迁移，其模型架构、训练方式与计算机领域的视觉语言模型类似。如首个遥感视觉语言模型 RemoteCLIP 即为对比式模型，利用构建的大规模遥感数据来微调经典对比式通用视觉语言模型 CLIP 实现遥感域迁移。随后出现的 GeoChat 则是预训练式模型，通过构建多层感知器进行跨模态适应，将图像特征投影到 LLM 的语言空间，使 LLM 可以生成与图像对应的自然语言响应。

值得注意的是，不同的遥感视觉语言模型的预训练和测试数据选择差异较大，如表 3 所示。例如，RemoteCLIP、GeoRSCLIP、SkyCLIP 和 LHRS-Bot 模型将 EuroSAT 数据集设置为测试集，而 EarthGPT 选用它作为训练数据集。此外，从表中可看到常用

的数据集有 EuroSAT、RESISC45、WHU-RS19、SIRI-WHU、AID、RSI-CB256、PatternNet 和 FMoW。下面详细介绍这些数据集的来源、类别及分辨率等信息。

表 4 不同遥感视觉语言模型的数据选择

Table 4 Data selection of different remote sensing visual language models

模型	训练数据	测试数据
RemoteCLIP	自构建数据集	PatternNet, EuroSAT, OPTIMAL-31, RSC11, RSICB128, MLRSNet, RSI-CB256, RESISC45, WHU-earth, RS2800, WHU-RS19
GeoRSCLIP	fMoW, MillionAID, BEN	EuroSAT, AID, RESISC45
GeoChat	RESISC45	AID, UCM
GRAFT	自构建数据集 NAIP 和 Sentinel-2	EuroSAT, BEN, SAT-4, SAT-6
SkyCLIP	自构建数据集 SkyScript	PatternNet, EuroSAT, AID, fMoW, RESISC45, MillionAID, RSI-CB256
EarthGPT	EuroSAT, RESISC45, UCM, WHU-RS19, RSSCN7, FGSCR-42, DSCR	RESISC45, CLRS, NaSC-TG2
LHRS-Bot	fMoW, RSITMD, METER-ML, NWPU, UCM	AID, WHU-RS19, SIRI-WHU, EuroSAT, RESISC45, fMoW, METER-ML
H2RSVLM	fMoW, RSITMD, METER-ML, MillionAID, UCM, NWPU	NWPU, METER-ML, SIRI-WHU, AID, WHU-RS19
SkySenseGPT	RSITMD, UCM, NWPU	AID, WHU-RS19, SIRI-WHU

EuroSAT（Helber 等，2019）数据集来源于 Sentinel-2 卫星图像，涵盖了 13 个光谱带，由 10 个类别组成，总共 27000 个标记和地理参考图像。

RESISC45（Cheng 等, 2017）数据集，也被称为 NWPU45 数据集，是西北工业大学（NWPU）于 2017 年创建的遥感场景分类公共基准。它包含了 31,500 张图像，涵盖了 45 个场景类，每个类别有 700 张图像。

WHU-RS19（Dai and Yang 等, 2011）数据集来源于 Google Earth，由武汉大学发布。该数据集由 19 类航拍场景组成。每个类别有 50 个样本，图像大小为 600×600 像素。

SIRI-WHU（Zhao 等, 2016）数据集是从 Google Earth 上收集的。该数据集包含 12 个类别，共 2400 张图像，每张图像的尺寸为 200×200 像素。

AID（Xia 等, 2017）数据集也来自于 Google Earth，共 10000 张图像，由 30 种航拍场景类型组成。不同航空场景类型的样本图像数量差异较大，从 220 张到 420 张不等。

RSI-CB256（Li 等, 2017）数据集有 6 大类和 35 个子类别，总计约 24,000 张图像，平均每个类别约 690 张图像。图像大小为 256×256 像素。

PatternNet（Zhou 等, 2018）数据集由 38 个类别组成，每个类有 800 个图像，大小为 256×256 像素。

FMoW（Christie 等, 2018）数据集包含来自 200 多个国家/地区的超过 100 万张图像，共 63 个类别。

上述数据集均为分类数据集，其规模远小于遥感视觉语言模型预训练数据需要的规模（万张图像及以上）。

4 遥感视觉语言模型

4.1 遥感视觉语言模型与零样本分类

相较于一般的遥感零样本分类算法（如基于嵌入式模型（Li 等, 2017; Gencer 等, 2018; 李彦胜等, 2020; Li 等, 2021）、生成式模型（Wang 等, 2021; Ma 等, 2022）等的零样本分类方法），遥感视觉语言模型在零样本分类任务中展现出更强的语义理解、数据复杂性适应、开放集泛化等独特优势。

在语义理解方面，一般的遥感零样本分类方法依赖于预定义的类别/属性语义，难以全面表达类别特征，而遥感视觉语言模型可以利用开放式文本描述增强类别表达能力。例如，相较于传统方法利用数值属性（如“光谱反射率范围”）定义类别，遥感视觉语言模型可以利用“该地物通常出现在干旱地区，具有白色或浅色的光谱特征”来描述类别，从而提升模型的泛化能力。

遥感数据不仅包括光学图像，还涉及雷达图像（SAR）、多光谱、高光谱等数据。在数据复杂性适应方面，遥感视觉语言模型可以通过多模态融合（如结合文本、光学图像和 SAR 图像）来增强类别判别能力。而一般的遥感零样本分类方法难以有效利用多样的遥感数据。

在开放集泛化方面，一般的遥感零样本分类方法通常基于固定的类别嵌入（如 Word2Vec 训练好的向量），难以适应新任务。而遥感视觉语言模型允许灵活输入新类别描述，直接进行零样本推理。例如，传统方法中若类别库中没有“珊瑚礁”，则该类别无法识别。而在遥感视觉语言模型中，只要在推理中输入描述“海底生物结构，由钙质沉积形成”，即可识别珊瑚礁。

此外，通过预训练的视觉-语言模型，可以更好地适应不同遥感场景，减少域偏移问题，提升零样本分类的稳定性。例如，传统方法如果是在北美数据上训练的模型，那它在非洲地区可能失效。而遥感视觉语言模型可利用“这种地物常见于热带地区”的文本描述，增强跨区域适应性。

总的来看，遥感视觉语言模型突破了传统遥感零样本分类的局限，利用视觉-语言对齐技术，使得遥感图像分类更加灵活、泛化能力更强，适用于更广泛的遥感智能分析任务。

4.2 用于零样本分类的遥感视觉语言模型

RemoteCLIP（Liu 等, 2024）作为首个针对遥感的视觉语言模型，旨在通过调整模型的预训练数据进行模型参数微调来实现 CLIP 模型的遥感域迁移。为解决遥感域预训练数据稀缺的问题，该模型利用异构注释转换的方式进行了数据拓展，同时整合了无人机图像，得到了规模为 165k（现有数据集通常

为 10k 左右）的预训练数据集。在零样本分类任务中，该方法在 12 个下游数据集上的平均准确率超过了 CLIP 模型。

随后，遥感视觉语言模型基本延续了 RemoteCLIP 的研究思路，即“自然图像预训练模型+遥感数据微调”的模式，分为数据集构建和模型微调两部分主要内容。GeoRSCLIP（Zhang 等, 2023）在数据集构建方面与 RemoteCLIP 类似，都包含了整合现有数据集。不同是的，它还利用了 BLIP2（Li 等, 2023）为仅具有类别级别标签的 3 个大规模遥感数据集（RS5M-RS3）生成标题。在模型微调方面，除了完整微调，它还使用了 4 种不同的参数高效微调方法，即 Pfeiffer Adapter（Pfeiffer 等, 2021）、LoRA Adapter（Hu 等, 2021）、Prefix-tuning Adapter（Li and Liang 等, 2021）和 UniPELT Adapter（Mao 等, 2022）。与 CLIP 基线模型相比，GeoRSCLIP 模型的零样本分类性能在 AID（Xia 等, 2017）、RESISC45（Cheng 等, 2017）和 EuroSAT（Helber 等, 2019）数据集上改进了 3%至 20%。

与上述构建遥感图文数据集的方式不同，Wang 等人（Wang 等, 2024）采用从头收集整理的方式构建了一个包含 260 万个图文对，涵盖 2.9 万个不同语义标签的更为全面的遥感视觉语言数据集 SkyScript。随后，在该数据集上通过图文对比学习微调了 CLIP 模型，称为 SkyCLIP。该模型不仅显着提高了域内 SkyScript 分类数据集的性能，并且相较于基线 CLIP 模型能够更好地泛化到未见过的基准数据集。

鉴于上述方法产生的大量图像收集、文本注释等工作，GRAFT（Mall 等, 2024）利用地面拍摄的互联网图像为中介，实现遥感图像和文本的匹配，从而避免了文本注释，但需要互联网图像和遥感图像匹配的数据集。因此，它利用与地面图像相关联的地理标签，创建了大型的地面-卫星图像对数据集 NAIP（低分辨率）和 Sentinel-2（高分辨率）。该模型在分类任务上的性能的提升可达 20%。

除了图文对数据集构建和通用视觉语言模型微调之外，H2RSVLM（Pang 等, 2024）更注重提升自我感知能力，以应对模型产生的“幻觉”问题。它专门构建了 RSSA 数据集，通过在视觉问答任务中引入无法回答的问题，教会模型识别出这些问题并拒绝回答，以确保模型输出的真实性。

为处理多尺度、高分辨率的遥感图像，增加区域级推理和整体场景解释能力，Kuckreja 等人（Kartik 等, 2024）提出了 GeoChat。这是首个结合了大型语言模型（Large Language Model，LLM）的全能型遥感视觉语言模型。它通过构建多层感知器进行跨模态适应，将图像特征投影到 LLM 的语言空间，使得 LLM 可以生成与图像对应的自然语言响应。在 LLM 的微调方面，它采用了基于低秩自适应（Low Rank Adaptive，LoRA）的策略。GeoChat 在 AID 和 UCMerced 数据集上的零样本场景分类精度相较于通用视觉语言模型 Qwen-VL（Bai 等, 2023）、MiniGPTv2（Chen 等, 2023）和 LLaVA-1.5（Liu 等,

2024）表现出更好的性能。

此外，为统一集成多传感器遥感解释任务，Zhang 等人（Zhang 等, 2024）整合了现有 34 个多样化遥感数据集，构建了包含超过 1 兆个图文对的 MMRS 数据集，其中的遥感图像来自多传感器，例如光学雷达、合成孔径雷达和红外雷达。随后，EarthGPT 在数据集 MMRS 上进行了微调，并表现出了出色的粗粒度对话和细粒度本地化能力。

除构建信息更为丰富的大规模遥感图文数据集 LHRS-Align 外，LHRS-Bot（Dilxat 等, 2024）还提出了一种新颖的多级视觉语言对齐策略，采用三阶段课程学习，逐步解锁模型参数并引入越来越复杂的数据，使模型能够逐步吸收视觉知识并提高难度。

为增强遥感视觉语言模型的细粒度复杂理解能力，Luo 等人（Luo 等, 2024）构建了一个大规模的指示调整数据集 FIT-RS。除包括常见的解释任务，该数据集还创新地引入了几个难度不断升级的复杂理解任务，如关系推理和图像级场景图生成。SkySenseGPT 模型遵循主流的架构，由视觉编码器、多层感知器和一个 LLM 构成。该模型在 FIT-RS 数据集上训练，并通过 LoRA 技术对 LLM 进行了微调，获得了超越现有模型的出色性能。

4.3 讨论与分析

（1）关于数据集。

本节总结和讨论当前遥感图像分类数据集以及遥感视觉语言模型使用的遥感图文对数据集。遥感图像分类任务中推荐的数据集有 EuroSAT（Helber 等, 2019）、RESISC45（Cheng 等, 2017）、WHU-RS19（Dai and Yang 等, 2011）、SIRI-WHU（Zhao 等, 2016）和 AID（Xia 等, 2017）。除此之外，常用的图像场景分类数据集有 RSI-CB256（Li 等, 2017）、PatternNet（Zhou 等, 2018）和 FMoW（Christie 等, 2018）。

上述数据集的详细信息已在 3.2 节介绍，为拓展这些数据集的规模，适应遥感视觉语言模型数据需求，大规模遥感图文对数据集的构建方式可分为两类，一类是下载原始遥感图像，通过网站或多模态模型获取其对应的注释；另一类是整合现有的遥感数据集资源，通过提示模板将原始注释转换为文本描述。目前已构建的遥感图文对数据集的相关信息如表 5 所示。下面详细介绍这些数据集的相关信息。RemoteCLIP 模型（Liu 等, 2024）使用的训练数据集是由语义分割数据集、目标检测数据集和图像分类数据集整合而成。它的异构注释过程是框-描述（box-to-caption，B2C）生成和掩码-框（mask-to-box，M2B）转换。B2C 生成是指基于边界框注释和标签生成目标检测数据集的文本描述。采用基于规则的方法来生成 5 个不同描述，分别包含图像中对象的位置和类别数量。M2B 转换是将语义分割数据集的注释方式转换为目标检测数据集的注释方式，从而与 B2C 生成方法衔接起来。通过这 3 种类型数据集整合，最终得到的训练数据集包含 165,745 张图像，每张图像都有五个相应的标题，总共 828,725 个训练图像-文本对。图像类别整合自多个数据集，较为

混杂。图像分辨率多样，从 224×224 到 1920×1080 不等。从图像分类、目标检测和语义分割三个维度整合了当前遥感常用的数据集，标注方式由原注释转换而来，相对降低了注释的主观性。

表 5 不同遥感训练数据集信息

Table 5 Different remote sensing training data sets information

模型/数据集	类型	图像类别	图像大小/像素	空间分辨率/m	注释方式	规模	支持的下游任务
RemoteCLIP	整合现有数据集	类型混杂	从 224×224 到 1920×1080 不等	--	使用文本模板将多种注释转换为图像标题	165k	场景分类、图文检索、目标计数等
RS5M-RS3	整合现有数据集	类型混杂	--	--	使用通用多模态模型 BLIP-2 生成图像标题	3M	场景分类、图文检索、语义定位等
GeoChat	整合现有数据集	类型混杂	从 256×256 到 3000×4000 不等	--	基于使用 SAM 技术生成的分割数据集构建遥感图文对数据集	318k	图像描述、图像特定区域描述、视觉问答、场景分类、视觉定位、目标检测、定位描述、多任务对话等
GRAFT	从头收集	来自 Flickr 上的收集结果	--	高分辨率（每像素 1 米）和低分辨率（每像素 10 米	将遥感图像与从社交媒体 Flickr 上收集的地面图像标题对齐	2M	场景分类、图文检索、视觉问答、语义分 割等
SkyScripts	从头收集	来自 OSM	--	从 0.1 米/像素到 30 米/像素的多分辨率	使用地理坐标自动连接开放、未标记的遥感图像与开源协作项目 OpenStreetMap2（OSM）来构建图像标题	5M	场景分类、细粒度场景分类、图文检索等
MMRS	整合现有数据集	类型混杂、多源图像	--	--	使用文本模板将多种注释转换为图像标题	1005k	图像描述、图像特定区域描述、视觉问答、场景分类、视觉定位、目标检测、多任务对话等
LHRS-Align	从头收集	来自 OSM	--	--	从 OSM 获取地理特征的标签，并使用纯语言模型 Vicuna-v1.5 基于这些标签生成图像标题	1.15 M	图像描述、视觉问答、场景分类、视觉定位、多任务对话等
HqDC-1.4M	整合现有数据集	类型混杂	512×512	--	使用视觉语言模型 Gemini 版本生成高质量细粒度描述数据集，包括类别、属性、固定和移动目标的未知描述等。	1.4M	图像描述、视觉问答、多标签场景分类、视觉定位等
FIT-RS	从头收集	基于 STAR 数据集	从 512×768 到 27860×31096 不等	--	用详细场景图标签手动注释，增加三元组和旋转边界框注释	1800.8k	图像描述、图像特定区域描述、视觉问答、多标签场景分类、目标检测、多任务对话、关系检测、关系推理、目标级推理、区域级场景图生成、图像级场景图生成等

RS5M 数据集（Zhang 等, 2023）除了收集公开可用的图文对数据集，从中筛选遥感类的可用数据外，

还使用 BLIP2（Li 等, 2023）为仅具有类级标签的 3 个大型遥感数据集（BigEarthNet（Sumbul 等, 2019）、FMoW（Christie 等, 2018）和 Million-AID（Long 等, 2021））生成描述。数据集中图像类别混杂，分辨率不等，存在世界某些地区的数据代表过多和代表性不足的问题。

GeoChat（Kartik 等, 2024）构建的遥感指令集合并了目标检测、场景分类和视觉问答 3 种不同类型的数据集。其中，目标检测数据集有 DOTA（Xia 等, 2018）、DIOR（Cheng 等, 2022）和 FAIRIM（Sun 等, 2022），它们构成了这个遥感指令集重要的子集 SAMRS 数据集。在这个子集上利用 ViTAE-RVSA 模型（Wang 等, 2023）添加了一些基本类别，如建筑物、道路和树木。场景分类数据集是 NWPURESISC-45 数据集，视觉问答数据集是 LRBEN 数据集（Lobry 等, 2020）和 Floodnet 视觉问答数据集（Rahnemoonfar 等, 2021）。值得注意的是，目标检测数据集中边界框的注释可为模型提供区域级信息，从而使得在其上训练的模型具备区域级推理能力。数据集中图像类别混杂，分辨率多样，从 256×256 到 3000×4000 不等。与简单的生成文本描述不同，本数据集构建的遥感指令集还需进行属性提取，如类别、颜色、相对大小、相对位置和目标相对关系。

与其他遥感视觉语言模型不同，GRAFT 模型（Mall 等, 2024）所需要的是地面-卫星图像对数据集。为同时支持图像级和像素级的图像识别任务，该数据集分别收集了高低两个分辨率的数据集，NAIP（高分辨率，每像素 1 米）和 Sentinel-2（低分辨率，每像素 10 米）。地面图像是从 Flick 上收集到的，为获取不同地区的代表性图像，位置是统一选择的，地理标签是准确的。卫星图像来源于 EarthEngine APIs2，它的中心设置在地面图像的地理标记处，地理标签属于该卫星图像的所有地面图像都与其进行匹配。最终构建的 NAIP 数据集包含 1020 万对地面-卫星图像对，Sentinel-2 数据集包含 870 万对。

SkyScripts 数据集（Wang 等, 2024）是通过将大规模未标记的遥感图像数据与来自 OpenStreetMap（OSM）的地理标记语义信息进行匹配得到的。其中的卫星和航空图像来源于 Google Earth Engine（GEE），形成了地面采样距离范围从 0.1 米/像素到 30 米/像素的多源、多分辨率的图像池。语义信息从 OSM 上收集。在 OSM 中，地图上的每个对象都由一个或多个标签来描述。每个标签由两个自由格式文本字段组成：键和值。根据定义，键用于描述主题、类别或特征类型（例如“地面”），而值则描述给定键的特定特征、属性或子类别（例如“沥青”）。这使得构建的数据集语义不仅包括大量的对象类别，还包括了细粒度的子类别和属性。

MMRS-1M 数据集（Zhang 等, 2024）是一个支持图文对话的丰富全面的指令跟踪遥感数据集。它涵盖了分类、检测、图像描述、视觉问答和视觉定位 5 个任务，以及光学、SAR 和红外 3 种模态。与

GeoChat 数据集类似，需要将类别、描述、问答等信息通过指定的模板转为遥感指令。

与 SkyScripts 数据集类似，LHRS-Align 数据集（Dilxat 等, 2024）是通过将来自 Google Earth3 的遥感图像与来自 OSM VGI 数据库的大量属性信息在地理上配对进行构建的，总共包含了 115 万个高质量遥感图文对。该数据集收集了来自世界各地的每个选定城市的 0.3°经度×0.28°纬度范围内的遥感图像和 OSM 特征。与 SkyScripts 数据集不同的是，LHRS-Align 数据集还使用了 LLM 模型 Vicuna-v1.5-13B（Chiang 等, 2023），通过汇总相应图像的键值标签来生成图像描述。该数据集的遥感图像遍及全球，可充分利用其中丰富的地理信息支持 LLM 释放在遥感应用方面的全部潜力。

HqDC-1.4M 数据集（Pang 等, 2024）利用“gemini-1.0-pro-vision”API 从多个公开遥感数据集中生成图像描述，从而获得了近 140 万个图像文本对数据。整合的数据集包括 Million-AID（Long 等, 2021）、CrowdAI（Mohanty 等, 2020）、FMoW（Christie 等, 2018）、CVUSA（Workman 等, 2015）、CVACT（Liu and Li 等, 2019）和 LoveDA（Wang 等, 2022）数据集，其中遥感图像的地面采样距离范围从 0.1 米到 100 米以上，包含广泛的空间分布和多样的陆地类型。图像分辨率统一裁剪为 512×512。生成的图像描述包括图像的类型、分辨率、属性、数量、颜色等信息，此外还有细粒度的图像中物体形状和空间位置信息。

为促进模型理解复杂遥感场景中实体之间丰富的语义关系，文献（Luo 等, 2024）首先构建了 STAR 数据集。该数据集包含了全球范围内的大尺寸超高分辨率遥感图像，每张图像都对应着详细的场景图标签。这些图像涵盖了多个复杂的语义场景（例如机场，大坝和港口），尺寸从 512×768 到 27,860×31,096 像素。由 STAR 数据集进行扩展得到了 FIT-RS 数据集。它包含超过 400,000 个三元组和超过 210,000 个具有 1273 个 RSI 中旋转边界框的对象。

（2）关于模型。

根据视觉语言模型的分类，可将 4.2 节中具备零样本分类功能的遥感视觉语言模型分为两类，对比式和预训练式。各模型的核心组件、微调技术及所需硬件资源如表 6 所示。

表 6 不同遥感视觉语言模型

Table 6 Different remote sensing visual language models

模型	类型	图像编码器	文本编码器	多模态编码器	微调	硬件资源
RemoteCLIP	对比	ViT-L（CLIP）	（CLIP）	—	完整微调	4*3090Ti（24G）
GeoRSCLIP	对比	ViT-H（CLIP）	（CLIP）	—	Pfeiffer、LoRA、Prefix-tuning、UniPELT、完整微调	1*A100（40G）+ 1*A100（80G）

GeoChat	预训练	ViT-L（CLIP）	Vicuna-v1.5-7B	a two-layer MLP	LoRA	3*A100（40G）
GRAFT	对比	ViT-B（CLIP）	（CLIP）	—	—	Unknown
SkyCLIP	对比	ViT-L（CLIP）	（CLIP）	—	—	4*A100（80G）
EarthGPT	预训练	ViT-L（DINOv2）+ ConvNext-L（CLIP）	LLaMA2-7B	a linear layer	—	16*A100（80G）
LHRS-Bot	预训练	ViT-L（CLIP）	LLaMA2-7B	a vision perceiver	LoRA	8*V100（32G）
H2RSVLM	对比	ViT-L（CLIP）	Vicuna-v1.5-7B	a two-layer MLP	—	2*A6000（48G）
SkySenseGPT	预训练	ViT-L（CLIP）	Vicuna-v1.5-7B	a two-layer MLP	LoRA	4*A100（40G）

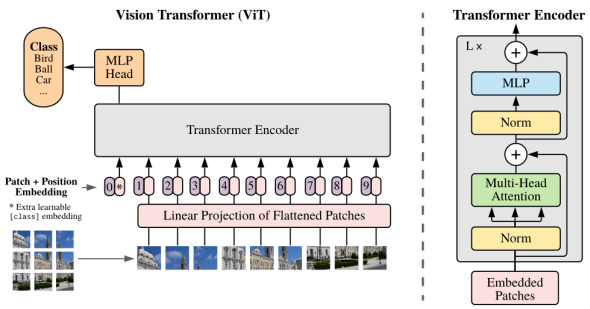

图 2 ViT 架构示意图（Alexey Dosovitskiy 等, 2020）

Fig.2 The framework of ViT（Alexey Dosovitskiy 等, 2020）

在图像编码器方面，多数模型采用了视觉转换器（Visio Transformer，ViT）（Alexey 等，2020），如图 2 所示。它通过将图像分割为固定大小的块，再对这些块进行线性嵌入和位置嵌入，生成图像级的向量序列。

在文本编码器方面，当前遥感视觉语言模型使用的主要有两种，一种是 CLIP 模型中采用的文本编码器，另一种是大型语言模型的开源版本。前者是一个具备 63M 参数、12 层和 8 个注意力头的 Transformer，最大文本序列长度上限设为 76。在该编码器中使用了掩码自注意力机制来保留使用预训练语言模型进行初始化或添加语言建模进行辅助的能力。后者常采用的开源版本有两种，分别是 LLaMA2（Touvron 等, 2023）和 Vicuna-v1.5（Chiang 等, 2023）。

LLaMA2 的架构由嵌入层、均方根标准化层、分组查询注意力层、前馈层和 Softmax 层组成，如图 3 所示。LLaMA2 有三点特殊的设计。一是均方根标准化层对每个 transformer 层的输入进行归一化；二是前馈层所采用的激活函数为 SwiGLU；三是分组查询注意力层将 Query 部分进行分组，每个组共享一组 KV，这样既避免了过多的性能损失，又能够加速推理。

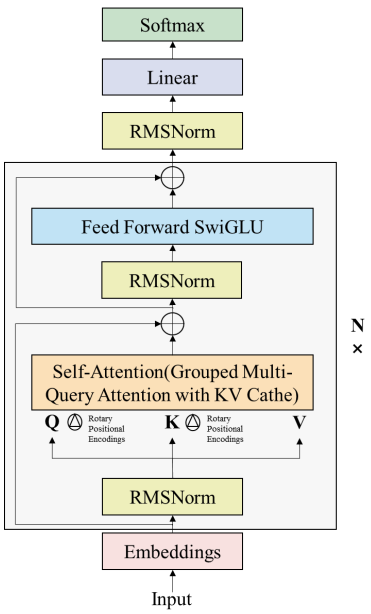

图 3 LLaMA2 架构示意图（Touvron 等, 2023）

Fig.3 The framework of LLaMA2（Touvron 等, 2023）

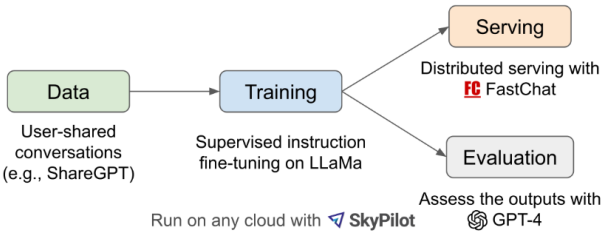

图 4 Vicuna 概述（Chiang 等, 2023）

Fig.4 The Workflow Overview of Vicuna（Chiang 等, 2023）

Vicuna-v1.5 的概述如图 4 所示，包括数据收集、监督训练、分布式服务和评估测试四部分。该模型的对话数据从 ShareGPT.com 上共收集了大约 7 万次。此外，该模型在 Stanford Alpaca 开源模型提供的训练脚本上进行了增强，以便于处理多轮对话和长序列。随后，它搭建了一个轻量级的分布式服务系统来提供对话演示。最后，模型的评估是通过 GPT-4 对模型的响应来判断的。

在模型微调技术方面，当前遥感视觉语言模型主要采用的是完整微调方法和基于适配器（Adapter）的方法（Houlsby 等, 2019）。完整微调方法是指对整个预训练模型的所有参数进行微调，可以获得更好的性能，但需要大量的计算资源和时间。Adapter 的出现是为了降低大规模预训练模型微调过程中需要改动的参数量，其主要思路是在模型网络结构中加入新定义的 Adapter 部分，在重新训练的过程中只更新 Adapter 部分的网络参数。经典 Adapter 先把输入的 d 维向量映射为一个小的 m 维向量，通过非线性层后，再从 m 维向量映射回 d 维向量，其结构如图 5 右所示。但该方法通常面临严重的灾难性遗忘和数据集平衡方面的困难。

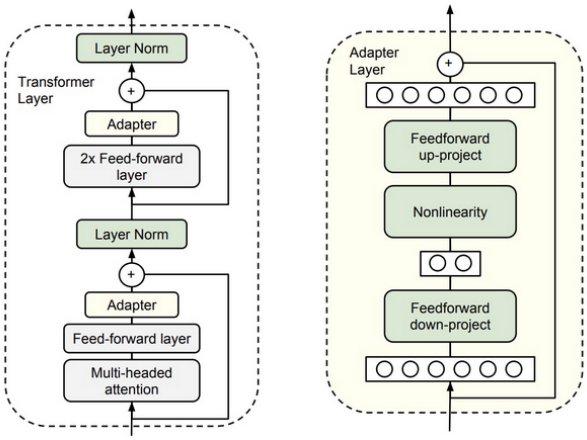

图 5 Adapter 概述（Houlsby 等, 2019）

Fig.6 The Workflow of Adapter（Houlsby 等, 2019）

当前用于遥感视觉语言模型的 Adapter 微调方法主要有 4 种变体，它们分别是 Pfeiffer Adapter（Pfeiffer 等, 2021）、LoRA Adapter（Hu 等, 2021）、Prefix-tuning Adapter（Li and Liang 等, 2021）和 UniPELT Adapter（Mao 等, 2022）。其中多数模型采用了 LoRA Adapter 方法进行微调。

Pfeiffer Adapter 为解决灾难性遗忘、任务间干扰和训练不稳定的问题，将学习过程分为两个阶段：知识提取阶段和知识组合阶段，如图 6 所示，可以利用来自多个任务的知识。知识提取阶段将训练 Adapter 模块学习下游任务的特定知识，将知识封装在 Adapter 模块参数中。知识组合阶段将预训练模型参数与特定于任务的 Adapter 参数固定，引入新参数组合多个 Adapter 中的知识，提高模型在目标任务中的表现。但 Adapter 模块的添加也导致模型整体参数量的增加，降低了模型推理时的性能。

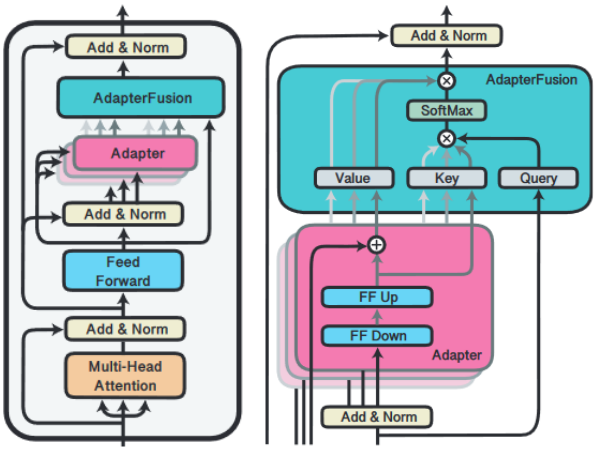

图 6 Pfeiffer Adapter 概述（Pfeiffer 等, 2021）

Fig.6 The Workflow of Pfeiffer Aapter（Pfeiffer 等, 2021）

LoRA adapter 为使微调更加高效，通过低秩分解的策略将权重更新矩阵表示为两个较小的新矩阵，如图 7 所示。这些新矩阵可以在适应新数据的同时保持整体变化数量较少地训练。该方法在微调的过程中，保持原始权重矩阵冻结，不需进一步调整。最终的参数更新矩阵由原始权重和适应后的权重组合得到。这有效降低了 Adapter 的训练参数量，减少了 GPU 显存占用，同时没有产生额外的推理延迟。

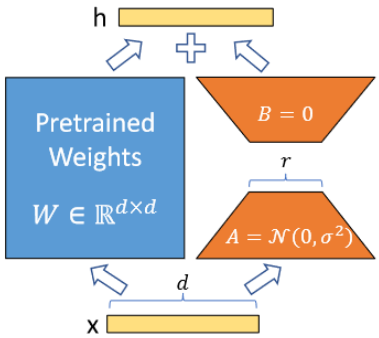

图 7 LoRA Adapter 概述（Hu 等, 2021）

Fig.7 The Workflow of LoRA Adapter（Hu 等, 2021）

Prefix-tuning Adapter 在输入令牌（token）之前构造一段与任务相关的虚拟令牌（virtual tokens）作为前缀（Prefix），然后训练的时候只更新 Prefix 部分的参数，而模型中的其他部分参数固定，如图 8 所示。值得注意的是，该方法针对不同的模型结构，需要构造不同的 Prefix。具体而言，针对自回归架构模型，需在句子前面添加 Prefix。针对编码器-解码器结构模型，需要在编码器和解码器中都增加 Prefix。

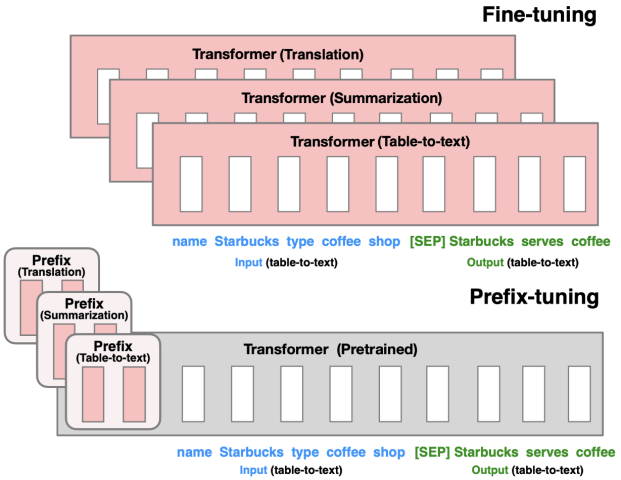

图 8 Prefix-tuning Adapter 概述（Li and Liang 等, 2021）

Fig.8 The Workflow of Prefix-tuning Adapter（Li and Liang 等, 2021）

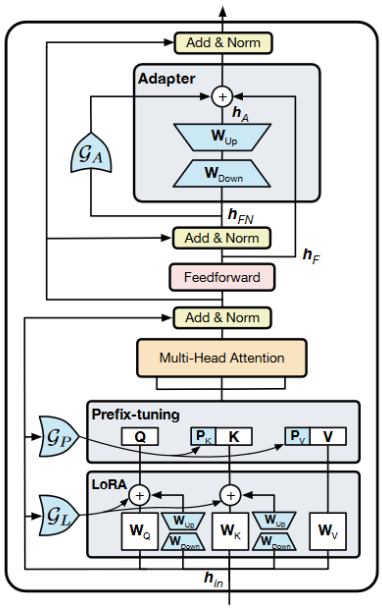

图 9 UniPELT Adapter 概述（Mao 等, 2022）

Fig.9 The Workflow of UniPELT Adapter（Mao 等, 2022）

UniPELT Adapter 是一个将上述的微调方法进行

统一的架构，如图 9 所示。它通过门控机制学习激活最适合当前数据或任务设置的方法，具体来说，这个架构为每个 Transformer 层中的每个子模块添加一个可训练门。对于每个模块，门控由线性层实现，并通过门参数控制 Prefix-tuning、LoRA 和 Adapter 方法的开关。

在遥感视觉语言模型中，对比式模型更适合使用完整微调方法，这样可获得更好的性能，但需要大量的计算资源和时间。虽然也有学者尝试使用基于 Adapter 的方法微调模型以减少计算资源和时间，但所获模型性能明显不如完整微调（Zhang 等, 2023）。对于预训练式模型而言，预训练模型的参数规模过大，完整微调耗时太长，当前最常用的模型微调技术是 LoRA 方法，该方法较为经典，操作简单高效。

综上所述，目前的遥感视觉语言模型，无论是图像还是文本，编码器还是解码器都是以 Transformer 架构为基础。常用的图像编码器是基于 CLIP 模型图文对齐的 ViT 网络。常用的文本编码器除了 CLIP 模型的文本编码器外，就是基于 LLM 微调的开源模型。对于预训练范式的遥感视觉语言模型中的多模态编码器，则通常采用多层感知器。总的来说，当前的遥感视觉语言模型在模型架构和微调方法等方面都有所探索，还构建了一批高质量、丰富多样的遥感图文对数据集，为将通用视觉语言模型转变为遥感领域专用模型奠定了坚实的基础。

5 存在的问题及未来发展

目前用于零样本分类的遥感视觉语言模型仍处于快速发展阶段，其前景十分可观，但现有方法中仍然存在一些问题。下面就这些问题进行分析并对其未来发展进行讨论。

5.1 存在的问题

当前的遥感视觉语言模型取得了一系列进展，同时也暴露出了几点问题。首先是严重的数据依赖。目前大多数模型都依赖于大规模遥感图文数据集的收集整理，是以数据为中心的研究工作，在模型架构及微调方法上的改进较少。其次是高额的训练成本。如表 5 所示，遥感视觉语言模型对于硬件资源要求较高，大部分工作使用 A100GPU 完成。然而，高性能的算力资源造成了研究障碍，不可避免地阻碍了这一方向的发展。最后是欠缺的基准评价。当前对于遥感视觉语言模型的评价体系并不完备，这主要体现在现有的基准测试没有覆盖遥感视觉语言模型的各种能力，难以全面衡量其性能。另一方面是各个模型对于数据集的选择较为多样，如表 3 所示，难以得到统一的评价结果。

5.2 未来的发展

作为遥感领域的一个热门方向，用于零样本分类的遥感视觉语言模型衍生于计算机视觉领域的视觉语言模型，与深度学习紧密相连。目前来看，用于零样本分类的遥感视觉语言模型在未来的研究中

有以下几个潜在的研究方向：

首先是模型架构的设计改进。纵观基于深度学习的图像处理方法的发展脉络，当模型技术突破到一定阶段后会出现大量数据爆发的阶段，但随后又回归到对于模型架构技术的深入研究中。当前架构以 Transformer 为主，其二次复杂性限制了该模型处理长序列的能力。结合遥感任务需求，降低模型复杂性或为架构改进方向之一。

其次是数据训练的优化调整。高算力的研究障碍是一个亟需解决的问题，利用低算力场景实现遥感视觉语言模型的训练微调将是未来研究人员想要的实现方式。或可将模型微调技术和遥感模型范式、数据特点结合，针对性地优化模型微调技术，以降低模型微调所需算力。

最后是基准评价的统一完善。随着这一方向的发展，基准评价体系的统一和完善将有利于促进这一方向技术的深入发展。当前遥感视觉语言模型的发展分散多样，下游任务类型层出不穷，结合任务类型和数据特点逐步完善评价体系较为可行。

6 结　论

用于零样本分类的遥感视觉语言模型是近两年新兴的热门方向，当前已经取得了一系列的成果。这一方向与传统零样本分类的区别在于预训练数据的规模更大，模型的参数量和复杂度更高，分类的精度和泛化性能更好，这在遥感数据爆发的时代具有十分重要的价值和意义以及相当可观的研究前景。本文首先介绍了视觉语言模型的基础概念，随后以零样本分类为主线，对比分析了遥感视觉语言模型的发展现状，其构建方式以“通用预训练模式+遥感数据微调”为主。这使得当前模型发展以大规模数据为驱动，高性能算力为支撑，且模型多样性和复杂性明显，统一的模型评价体系尚未出现。因此，遥感语言模型未来的发展应结合遥感域数据特点对模型架构进行改进设计，或改进模型微调技术以优化数据微调的算力需求，同时逐步完善模型的评价体系。

参考文献（References）

Huang Y，He X G and Wan Y L. 2024. Hyperspectral-image classification method combining superpixel dimension reduction with post-processing optimization. National Remote Sensing Bulletin，28（2）：494-510. (黄媛，贺新光，万义良.2024.联合超像素降维和后处理优化的高光谱图像分类方法.遥感学报，28（2）： 494-510)

Li Y F，Li J and He L. 2022. Convolutional neural network based single image pair method for spatiotemporal fusion. National Remote Sensing Bulletin， 26（8）：1614-1623. (李云飞，李军，贺霖.2022.单样本对卷积神经网络遥感图像时空融合.遥感学报，26（8）：1614-1623)

Wu Q Q，Wang S，Wang B and Wu Y L. 2022. Road

extraction method of high-resolution remote sensing image on the basis of the spatial information perception semantic segmentation model. National Remote Sensing Bulletin，26（9）： 1872-1885. (吴强强，王帅，王彪，吴艳兰.2022.空间信息感知语义分割模型的高分辨率遥感影像道路提取.遥感学报，26（9）： 1872-1885)

Zhang T，Yang X G，Lu X Q，Lu R T and Zhang S X. 2022. Ship detection in remote sensing image based on dense RFB and LSTM. National Remote Sensing Bulletin，26（9）：1859-1871. (张涛，杨小冈，卢孝强，卢瑞涛，张胜修.2022.Dense RFB 和 LSTM 遥感图像舰船目标检测.遥感学报，26（9）： 1859-1871)

Fu K，Lu W X， Liu X Y， Deng C B， Yu H F and Sun X. 2024. A comprehensive survey and assumption of remote sensing foundation modal. National Remote Sensing Bulletin，28（7）：1667-1680. (付琨，卢宛萱，刘小煜，邓楚博，于泓峰，孙显.2024.遥感基础模型发展综述与未来设想.遥感学报，28（7）：1667-1680)

Alec R, Jong W K, Chris H, Aditya R, Gabriel G, Sandhini A, Girish S, Amanda A, Pamela M, Jack C, Gretchen K, Ilya S, 2021. Learning Transferable Visual Models From Natural Language Supervision, in: International Conference on Machine Learning（2021）. Presented at the International Conference on Machine Learning, 1–48.

Alexey D, Lucas B, Alexander K, Dirk W, Zhai X H, Thomas U, Mostafa D, Matthias M, Georg H, Sylvain G, Jakob U, Neil H, 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv: 2010.11929. [DOI:10.48550/arXiv.2010.11929]

Bai J, Bai S, Yang S, Wang S, Tan S, Wang P, Lin J, Zhou C, Zhou J, 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv preprint arXiv: 2308.12966. [DOI:10.48550/arXiv.2308.12966]

Zhao B, Zhong Y F, Xia G S, Zhang L P, 2016. Dirichlet-Derived Multiple Topic Scene Classification Model for High Spatial Resolution Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing 54, 2108–2123. [DOI:10.1109/TGRS.2015.2496185]

Chameleon T, 2024. Chameleon: Mixed-Modal Early-Fusion Foundation Models. arXiv preprint arXiv: 2405.09818. [DOI:10.48550/arXiv.2405.09818]

Pang C, Wu J, Li J Y, Liu Y, Sun J X, Li W J, Weng X X, Wang S, Feng L T, Xia G S , He C H, 2024. H2RSVLM: Towards Helpful and Honest Remote Sensing Large Vision Language Model. arXiv preprint arXiv: 2403.20213.

[DOI:10.48550/arXiv.2403.20213]

Chen J, Zhu D, Shen X, Li X, Liu Z, Zhang P, Krishnamoorthi R, Chandra V, Xiong Y, Elhoseiny M, 2023. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv: 2310.09478. [DOI:10.48550/arXiv.2310.09478]

Cheng G, Wang J, Li K, Xie X, Lang C, Yao Y, Han J, 2022. Anchor-Free Oriented Proposal Generator for Object Detection. IEEE Transactions on Geoscience and Remote Sensing 60, 1–11. [DOI:10.1109/TGRS.2022.3183022]

Chiang W L, Li Z, Lin Z, Sheng Y, Wu Z, Zhang H, Zheng L, Zhuang S, Zhuang Y, Gonzalez J E, Stoica I, Xing E P, 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. [https://lmsys.org/blog/2023-03-30-vicuna].

Christian S, Liu W, Jia Y Q, Pierre S, Scott R, Dragomir A, Dumitru E, Vincent V, Andrew R, 2015. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition （CVPR）IEEE, Boston, MA, USA: 1–9. [DOI:10.1109/CVPR.2015.7298594]

Christie G, Fendley N, Wilson J, Mukherjee R, 2018. Functional Map of the World, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition （CVPR） IEEE, Salt Lake City, UT: 6172–6180. [DOI:10.1109/CVPR.2018.00646]

Dai D X, Yang W, 2011. Satellite Image Classification via Two-Layer Sparse Coding With Biased Image Representation. IEEE Geoscience and Remote Sensing Letters 8, 173–176. [DOI:10.1109/LGRS.2010.2055033]

Dilxat M, Li Z S, Gu F, Zhang X L, Xiao P F, 2024. LHRS-Bot: Empowering Remote Sensing with VGI-Enhanced Large Multimodal Language Model. arXiv preprint arXiv: 2402.02544. [DOI:10.48550/arXiv.2402.02544]

Cheng G, Han J W, Lu X Q, 2017. Remote Sensing Image Scene Classification: Benchmark and State of the Art. Proc. IEEE 105, 1865–1883. [DOI:10.1109/JPROC.2017.2675998]

Xia G S, Hu J W, Hu F, Shi B G, Bai X, Zhong Y F, Zhang L P, Lu X Q, 2017. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. IEEE Transactions on Geoscience and Remote Sensing 55, 3965–3981. [DOI:10.1109/TGRS.2017.2685945]

Helber P, Bischke B, Dengel A, Borth D, 2019. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover

Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12, 2217-2226. [DOI: 10.1109/JSTARS.2019.2918242]

Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, Laroussilhe Q.D, Gesmundo A, Attariyan M, Gelly S, 2019. Parameter-Efficient Transfer Learning for NLP, in: Proceedings of the 36th International Conference on Machine Learning. Presented at the International Conference on Machine Learning PMLR: 2790–2799.

Hu E J, Shen Y, Wallis P, Allen Zhu Z, Li Y, Wang S, Wang L, Chen W, 2021. LoRA: Low-Rank Adaptation of Large Language Models. [DOI:10.48550/arXiv.2106.09685]

Jacob D, Chang M W, Kenton L, Kristina T, 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies

K. Simonyan, Andrew Zisserman, 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 (2014).

He K M, Zhang X Y, Ren S Q, Sun J, 2016. Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition （CVPR）. Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition （CVPR） IEEE, Las Vegas, NV, USA: 770–778. [DOI:10.1109/CVPR.2016.90]

Kartik K, Muhammad S D, Muzammal N, Abhijit D, Salman K, Fahad S K, 2024. GeoChat: Grounded Large Vision-Language Model for Remote Sensing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition（2024）. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 27831–27840.

Kwon G, Cai Z, Ravichandran A, Bas E, Bhotika R, Soatto S, 2022. Masked Vision and Language Modeling for Multi-modal Representation Learning, in: The Eleventh International Conference on Learning Representations （ 2022 ） . Presented at the Eleventh International Conference on Learning Representations, 1–18.

Li H, Tao C, Wu Z, Chen J, Gong J, Deng M, 2017. RSI-CB: A Large Scale Remote Sensing Image Classification Benchmark via Crowdsource Data. arXiv preprint arXiv:1705.10450. [DOI:/10.48550/arXiv.1705.10450]

Li J, Li D, Savarese S, Hoi S, 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large

Language Models. arXiv preprint arXiv: 2301.12597. [DOI:10.48550/ arXiv.2301.12597]

Li X L, Liang P, 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 1: 4582–4597. [DOI:10.18653/v1/2021.acl-long.353]

Li A X, Lu Z W, Wang L W, Xiang T, Wen J R, 2017. Zero-Shot Scene Classification for High Spatial Resolution Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing 55, 4157–4167. [DOI:10.1109/TGRS.2017.2689071]

Wang C, Peng G H, Bernard D B, 2021. A Distance-Constrained Semantic Autoencoder for Zero-Shot Remote Sensing Scene Classification. IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 14, 12545–12556. [DOI:10.1109/JSTARS.2021.3132189]

Gencer S, Ramazan G C, Selim A, 2018. Fine-Grained Object Recognition and Zero-Shot Learning in Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing 56, 770–779. [DOI:10.1109/TGRS.2017.2754648]

Ma S Q, Liu C, Li Z, Yang W, 2022. Integrating Adversarial Generative Network with Variational Autoencoders towards Cross-Modal Alignment for Zero-Shot Remote Sensing Image Scene Classification. Remote Sensing 14, 4533. [DOI:10.3390/rs14184533]

Li Y S, Kong D Y, Zhang Y J, Tan Y H, Chen L, 2021. Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification. ISPRS Journal of Photogrammetry and Remote Sensing 179, 145–158. [DOI:10.1016/j.isprsjprs.2021.08.001]

Li Y S, Zhu Z H, Yu J G, Zhang Y J, 2021. Learning Deep Cross-Modal Embedding Networks for Zero-Shot Remote Sensing Image Scene Classification. IEEE Transactions on Geoscience and Remote Sensing 59, 10590–10603. [DOI:10.1109/TGRS.2020.3047447]

Li Y S, Kong D Y, Zhang Y J, Ji Z, Xiao R, 2020. Zero-shot remote sensing image scene classification based on robust cross-domain mapping and gradual refinement of semantic space[J]. Acta Geodaetica et Cartographica Sinica, 2020, 49(12): 1564-4574.（李彦胜，孔德宇，张永军，季铮，肖锐, 2020. 联合稳健跨域映射和渐进语义基准修正的零样本遥感影像场景分类. 测绘学报 49，1564－1574）

Liu F, Chen D, Guan Z, Zhou X, Zhu J, Ye Q, Fu L, Zhou J, 2024. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. IEEE Transactions on Geoscience and Remote Sensing 62, 1–16. [DOI:10.1109/TGRS.2024.3390838]

Liu H, Li C, Li Y, Lee Y J, 2024a. Improved Baselines with Visual Instruction Tuning. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA: 26296–26306. [DOI:10.1109/CVPR52733.2024.02484.]

Liu H, Li C, Wu Q, Lee Y J, 2024b. Visual instruction tuning, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS'23. Curran Associates Inc, Red Hook, NY, USA: 34892–34916. [DOI:10.5555/3666122.3667638]

Liu L, Li H, 2019. Lending Orientation to Neural Networks for Cross-View Geo-Localization, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition （CVPR）. Presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition （CVPR）IEEE, Long Beach, CA, USA: 5617–5626. [DOI:10.1109/CVPR.2019.00577]

Lobry S, Marcos D, Murray J, Tuia D, 2020. RSVQA: Visual Question Answering for Remote Sensing Data. IEEE Transactions on Geoscience and Remote Sensing 58, 8555–8566. [DOI:10.1109/TGRS.2020.2988782]

Long Y, Xia G S, Li S, Yang W, Yang M Y, Zhu X X, Zhang L, Li D, 2021. On Creating Benchmark Dataset for Aerial Image Interpretation: Reviews, Guidances, and Million-AID. IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 14, 4205–4230. [DOI:10.1109/JSTARS.2021.3070368]

Luo J, Pang Z, Zhang Y, Wang T, Wang L, Dang B, Lao J, Wang J, Chen J, Tan Y, Li Y, 2024. SkySenseGPT: A Fine-Grained Instruction Tuning Dataset and Model for Remote Sensing Vision-Language Understanding. arXiv preprint arXiv: 2406.10100. [DOI:10.48550/arXiv.2406.10100]

Mall U, Phoo C P, Liu M K, Vondrick C, Hariharan B, Bala K, 2024. Remote Sensing Vision-Language Foundation Models without Annotations via Ground Remote Alignment, in: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024.

Mao Y, Mathias L, Hou R, Almahairi A, Ma H, Han J, Yih S, Khabsa M, 2022. UniPELT: A Unified Framework for Parameter-Efficient Language Model Tuning, in: Muresan, S, Nakov, P,

Villavicencio, A. （Eds.）, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics （Volume 1: Long Papers ） . Presented at the ACL 2022 Association for Computational Linguistics, Dublin, Ireland: 6253–6264. [DOI:10.18653/v1/2022.acl-long.433]

Mohanty S P, Czakon J, Kaczmarek K A, Pyskir A, Tarasiewicz P, Kunwar S, Rohrbach J, Luo D, Prasad M, Fleer S, Göpfert J P, Tandon A, Mollard G, Rayaprolu N, Salathe M, Schilling M, 2020. Deep Learning for Understanding Satellite Imagery: An Experimental Survey. Frontiers in Artificial Intelligence 3

Oord A. van den, Li Y, Vinyals O, 2018. Representation Learning with Contrastive Predictive Coding. arXiv preprint arXiv: 1807.03748. [DOI:10.48550/ arXiv.1807.03748]

Pfeiffer J, Kamath A, Rücklé A, Cho K, Gurevych I, 2021. AdapterFusion: Non-Destructive Task Composition for Transfer Learning. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume 487–503. [DOI:10.18653/v1/2021.eacl-main.39]

Rahnemoonfar M, Chowdhury T, Sarkar A, Varshney D, Yari M, Murphy R.R, 2021. FloodNet: A High Resolution Aerial Imagery Dataset for Post Flood Scene Understanding. IEEE Access 9, 89644–89654. [DOI:10.1109/ACCESS.2021.3090981]

Singh A, Hu R, Goswami V, Couairon G, Galuba W, Rohrbach M, Kiela D, 2022. FLAVA: A Foundational Language And Vision Alignment Model, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition （CVPR）. Presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition （CVPR）IEEE, New Orleans, LA, USA: 15617–15629. [DOI:10.1109/CVPR52688.2022.01519]

Sumbul G, Charfuelan M, Demir B, Markl V, 2019. Bigearthnet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding, in: IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium. Presented at the IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing SymposiumIEEE, Yokohama, Japan: 5901–5904. [DOI:10.1109/IGARSS.2019.8900532]

Sun X, Wang P, Yan Z, Xu F, Wang R, Diao W, Chen J, Li J, Feng Y, Xu T, Weinmann M, Hinz S, Wang C, Fu K, 2022. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. ISPRS Journal of Photogrammetry and

Remote Sensing 184, 116–130. [DOI:10.1016/j.isprsjprs.2021.12.004]

Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D, Blecher L, Ferrer C.C, Chen M, Cucurull G, Esiobu D, Fernandes J, Fu J, Fu W, Fuller B, Gao C, Goswami V, Goyal N, Hartshorn A, Hosseini S, Hou R, Inan H, Kardas M, Kerkez V, Khabsa M, Kloumann I, Korenev A, Koura P.S, Lachaux M.-A, Lavril T, Lee J, Liskovich D, Lu Y, Mao Y, Martinet X, Mihaylov T, Mishra P, Molybog I, Nie Y, Poulton A, Reizenstein J, Rungta R, Saladi K, Schelten A, Silva R, Smith E.M, Subramanian R, Tan X.E, Tang B, Taylor R, Williams A, Kuan J.X, Xu P, Yan Z, Zarov I, Zhang Y, Fan A, Kambadur M, Narang S, Rodriguez A, Stojnic R, Edunov S, Scialom T, 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv preprint arXiv: 2307.09288. [DOI:10.48550/arXiv.2307.09288]

Tsimpoukelli M, Menick J.L, Cabi S, Eslami S.M.A, Vinyals O, Hill F, 2021. Multimodal Few-Shot Learning with Frozen Language Models, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.200–212

Waisberg E, Ong J, Masalkhi M, Kamran S.A, Zaman N, Sarker P, Lee A.G, Tavakkoli A, 2023. GPT-4: a new era of artificial intelligence in medicine. Ir J Med Sci 192, 3197–3200. [DOI:10.1007/s11845-023-03377-8]

Wang D, Zhang Q, Xu Y, Zhang J, Du B, Tao D, Zhang L, 2023. Advancing Plain Vision Transformer Toward Remote Sensing Foundation Model. IEEE Transactions on Geoscience and Remote Sensing 61, 1–15. [DOI:10.1109/TGRS.2022.3222818]

Wang J, Zheng Z, Ma A, Lu X, Zhong Y, 2022. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. arXiv preprint arXiv: 2110.08733. [DOI:10.48550/arXiv.2110.08733]

Wang Z, Prabha R, Huang T, Wu J, Rajagopal R, 2024. SkyScript: A Large and Semantically Diverse Vision-Language Dataset for Remote Sensing, in: Wooldridge, M.J, Dy, J.G, Natarajan, S.（Eds.）, Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada. AAAI Press5805–5813. [DOI:10.1609/AAAI.V38I6.28393]

Zhou W X, Shawn N, Li C M, Shao Z F, 2018. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS Journal of

Photogrammetry and Remote Sensing, Deep Learning RS Data 145, 197–209. [DOI:10.1016/j.isprsjprs.2018.01.004]

Workman S, Souvenir R, Jacobs N, 2015. Wide-Area Image Geolocalization with Aerial Reference Imagery, in: 2015 IEEE International Conference on Computer Vision （ICCV）. Presented at the 2015 IEEE International Conference on Computer Vision （ICCV） IEEE, Santiago, Chile: 3961–3969. [DOI:10.1109/ICCV.2015.451]

Xia G S, Bai X, Ding J, Zhu Z, Belongie S, Luo J, Datcu M, Pelillo M, Zhang L, 2018. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition （CVPR）IEEE, Salt Lake City, UT: 3974–3983. [DOI:10.1109/CVPR.2018.00418]

Li Y S, Zhu Z H, Yu J G, Zhang Y J, 2021. Learning Deep Cross-Modal Embedding Networks for Zero-Shot Remote Sensing Image Scene Classification. IEEE Transactions on Geoscience and Remote Sensing 59, 10590–10603. [DOI:10.1109/TGRS.2020.3047447]

Yu J, Wang Z, Vasudevan V, Yeung L, Seyedhosseini M, Wu Y, 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. arXiv preprint arXiv: 2205.01917. [DOI:10.48550/arXiv.2205.01917]

Zhang W, Cai M, Zhang T, Zhuang Y, Mao X, 2024. EarthGPT: A Universal Multimodal Large Language Model for Multisensor Image Comprehension in Remote Sensing Domain. IEEE Transactions on Geoscience and Remote Sensing 62, 1–20. [DOI:10.1109/TGRS.2024.3409624]

Zheng K, He X, Wang X E, 2024. MiniGPT-5: Interleaved Vision-and-Language Generation via Generative Vokens. [DOI:10.48550/arXiv.2310.02239]

Zhu D, Chen J, Shen X, Li X, Elhoseiny M, 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv preprint arXiv: 2304.10592. [DOI:10.48550/arXiv.2304.10592]

Zilun Zhang, Tiancheng Zhao, Yulong Guo, Jianwei Yin, 2023. RS5M and GeoRSCLIP: A Large Scale Vision-Language Dataset and A Large Vision-Language Model for Remote Sensing, IEEE Transactions on Geoscience and Remote Sensing 62, 1-23. [DOI:10.1109/TGRS.2024.3449154]

Liu J Y, Shi C J, Tu D J and Liu S. 2021. Survey of Zero-Shot Image Classification. Journal of Frontiers of Computer Science and

Technology, 15, 812–824.（刘靖祎，史彩娟，涂冬景，刘帅，2021. 零样本图像分类综述. 计算机科学与探索 15, 812–824）

Remote Sensing Visual Language Models for Zero-Shot Classification: A Survey

TAN Xiaomeng [1,2], XI Bobo [1,3], XUE Changbin [1], LI Yunsong [3], XU Haitao[1]

1. *National Space Science Center, Key Laboratory of Electronic Information Technology for Complex Aerospace Systems, Beijing 100190, China;*

2. *University of Chinese Academy of Sciences, Beijing 100190, China*

3. *School of Cyber Engineering, Xi'dian University, Xi'an 710071, China*

Abstract: Visual language models (VLMs) have achieved remarkable success across a range of multimodal tasks, including zero-shot classification, image-text retrieval, image captioning, visual question answering, and visual localization. However, most existing methods are pre-trained on general-purpose datasets, which often leads to suboptimal generalization performance in specialized domains such as remote sensing and medical imaging. To address this limitation, a growing number of domain-specific remote sensing VLMs have recently been proposed. These models aim to incorporate geo-awareness by fine-tuning general VLMs using large-scale remote sensing image-text pair datasets. In this paper, we review and analyze the latest advancements in remote sensing VLMs, focusing on zero-shot classification as a central task.

Current VLMs can be broadly categorized into four types based on their training paradigms: contrastive, masked, generative, and pre-trained models. Contrastive models typically consist of an image encoder and a text encoder that are jointly trained to align visual and textual representations. Masked models build upon this architecture by incorporating a multimodal encoder to facilitate cross-modal understanding through masked prediction tasks. Pre-trained models further extend this framework by introducing a visual-language connector to bridge the gap between modalities. In contrast, generative models differ substantially in structure and functionality. Depending on the output modality—text or image—generative models can be divided into two subtypes, each employing distinct network architectures: one uses a text encoder, a generative network, and an image decoder; the other utilizes an image encoder, a generative network, and a text decoder.

Based on the aforementioned classification, this paper systematically categorizes existing remote sensing VLMs (RSVLMs) and provides a detailed analysis of their core components, fine-tuning

strategies, and associated datasets. Building on this analysis, we highlight several key challenges faced by current methods, including strong dependency on large-scale annotated data, high computational costs, and the absence of standardized benchmark evaluations. To address these issues, we propose several promising research directions.

RSVLMs for zero-shot classification have emerged as a prominent research direction in recent years, yielding a series of notable advancements. Compared with traditional zero-shot classification methods, these models are characterized by their reliance on significantly larger-scale pre-training datasets, increased model complexity and parameter counts, and superior classification accuracy and generalization capabilities. These features are particularly valuable in the context of the rapid expansion of remote sensing data, offering both substantial theoretical significance and practical applicability, as well as promising avenues for future research. This paper provides a comparative analysis of the current development of RSVLMs for zero-shot classification tasks, which are primarily constructed following the paradigm of general pre-training followed by fine-tuning on remote sensing data. This approach has led to a development pattern that is driven by large-scale datasets and supported by high-performance computing resources, resulting in considerable model diversity and complexity. However, a unified evaluation framework has yet to be established. Therefore, future development of RSVLMs should focus on designing model architectures that are better aligned with the unique characteristics of remote sensing data, and on improving fine-tuning techniques to reduce computational demands. At the same time, efforts should be made to progressively establish a standardized and comprehensive evaluation system for these models.

Key words: remote sensing intelligent interpretation, visual language model, remote sensing visual language model, model fine-tuning, multi-modal learning, image-text alignment, zero-shot classification, remote sensing dataset construction

Supported by National Natural Science Foundation of China (No. 62401434), the Youth Open Project of National Space Science Data Center (No. NSSDC2302004).