遥感视觉语言模型：当前研究进展与未来趋势

Xiang Li†, *，*IEEE*会员，Congcong Wen†，*IEEE*会员，Yuan Hu，Zhenghang Yuan，朱小翔，*IEEE*会士。

arXiv:2305.05726v2 [cs.CV] 2 Apr 2024

摘要—**ChatGPT 和 GPT-4 取得的卓越成就，在通用人工智能（AGI）领域掀起了对大语言模型的广泛关注与研究热潮。这些模型能够提供接近人类思维的智能解决方案，使我们能够运用通用人工智能解决各种应用中的问题。然而，在遥感（RS）领域，关于通用人工智能实现方式的科学文献仍相对匮乏。现有的遥感人工智能相关研究主要聚焦于视觉理解任务，却忽视了对目标及其关系的语义理解。视觉语言模型正是在这方面表现出色，它们能够对图像及其相关文本描述进行推理，从而更深入地理解其潜在语义。视觉语言模型可以超越对遥感图像的视觉识别，对语义关系进行建模，并生成图像的自然语言描述。因此，它们更适用于需要视觉和文本理解的任务，例如图像描述和视觉问答。本文全面综述了遥感领域视觉语言模型的研究，总结最新进展，重点介绍面临的挑战，并指出潜在的研究机遇。具体而言，我们回顾了视觉语言模型在遥感主流任务中的应用，包括图像描述、基于文本的图像生成、基于文本的图像检索、视觉问答、场景分类、语义分割和目标检测。针对每项任务，我们分析了代表性工作并讨论了研究进展。我们还总结了常用的遥感视觉语言数据集、代码库和在线可访问资源。最后，我们概述了当前研究的局限性，并提出未来发展的潜在方向。本综述旨在全面介绍遥感领域视觉语言模型的当前研究进展，并激发这一充满前景且令人振奋的领域开展进一步研究。**

索引词—**遥感、视觉语言模型、通用人工智能、GPT、Transformer**

I. 引言

深度学习已经兴起，作为各种遥感（RS）应用的强大工具。早期的遥感研究主要搭配

Xiang Li 就职于沙特阿拉伯图瓦勒 23955 的阿卜杜拉国王科技大学（电子邮箱：xiangli92@ieee.org）。Congcong Wen 就职于阿联酋阿布扎比 129188 的纽约大学阿布扎比分校电气与计算机工程系（电子邮箱：wencc@nyu.edu）。Yuan Hu 就职于中国北京 100871 的北京大学遥感与地理信息系统研究所（电子邮箱：huyuan@pku.edu.cn）。Zhenghang Yuan 和朱小翔就职于德国慕尼黑 80333 的慕尼黑工业大学地球观测数据科学教席（电子邮箱：zhenghang.yuan@tum.de；xiaoxiang.zhu@tum.de）。朱小翔还就职于德国慕尼黑 80333 的慕尼黑机器学习中心。

* 通讯作者：Xiang Li。† 共同第一作者。

Z. Yuan 和 X. Zhu 的工作由德国联邦教育与研究部（BMBF）在慕尼黑机器学习中心国际未来人工智能实验室“AI4EO——地球观测人工智能：推理、不确定性、伦理及其他”（资助编号：01DD20001）框架下共同支持。

侧重于利用从图像中提取的视觉特征执行各种任务，例如目标检测、语义分割和变化检测。作为最常用的深度学习方法之一，卷积神经网络（卷积神经网络） [90] 能够自动学习遥感图像的层次化表示，使其能够捕捉局部和全局空间特征及模式。此外，注意力机制 [191] 已被纳入深度学习模型，通过使模型能够关注输入中的特定区域，提升其在遥感任务中的表现。得益于深度神经网络强大的特征学习能力，深度学习模型已被证明在各种遥感任务中有效，其性能优于传统机器学习方法，达到最先进水平 [246], [268]。然而，现有基于深度学习的研究大多侧重于视觉理解任务，却忽视了对目标及其关系的语义理解。例如，在进行土地覆盖分类时，如果建筑屋顶像素在视觉上与高速公路像素相似，仅依赖视觉信息的模型可能会将其分类为高速公路。这是因为模型缺乏高速公路不可能位于建筑屋顶内部这一常识。

近年来，大语言模型（LLMs）已成为自然语言处理（NLP）和计算机视觉领域的热门研究主题。这些模型构建了大规模 Transformer 网络，用于自然语言理解，并在语言建模、文本生成和问答等各种语言理解任务中取得了最先进的性能 [146], [24]。值得注意的是，ChatGPT [136]取得的惊人成就激发了人们对通用人工智能（AGI）领域的广泛兴趣和研究。在世界知识和先进推理能力的加持下，这些模型在语言理解方面展现出无与伦比的进步。大语言模型的巨大成功推动了视觉语言模型（VLMs）的众多研究，助力 AI2.0 时代的到来。

视觉语言模型通常被定义为一类人工智能模型，它结合计算机视觉和自然语言处理技术，以全面理解视觉信息和文本信息。凭借联合识别视觉与语义模式及其关系的能力，视觉语言模型不仅能识别图像中的目标，还能理解目标之间的关系，并生成图像的自然语言描述。因此，它们更适合需要同时理解视觉和文本信息的任务，例如图像

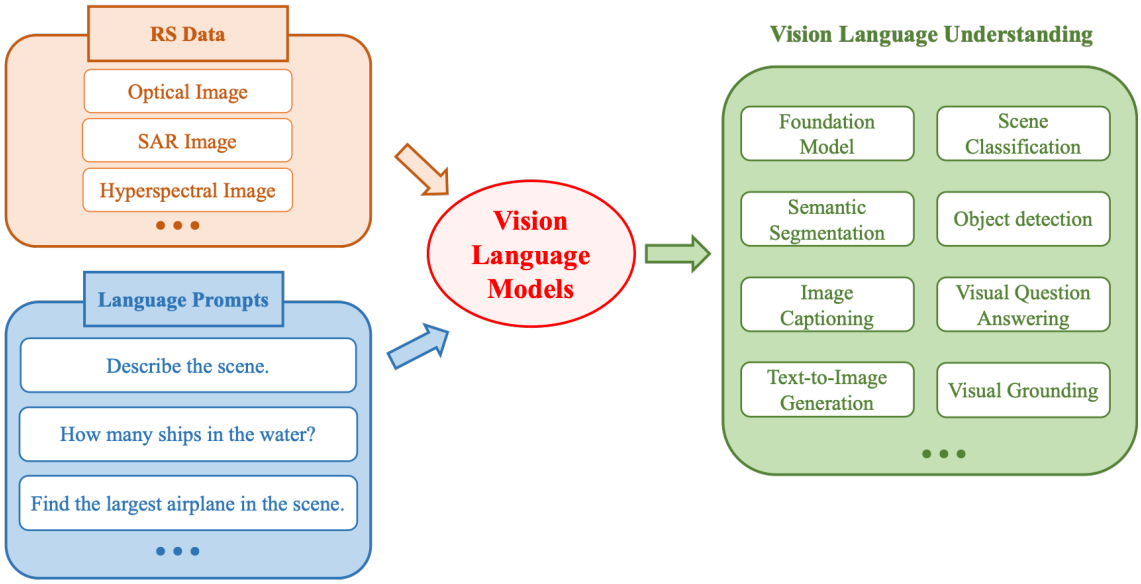

图 1. 遥感中的视觉语言模型。

图像描述、基于文本的图像检索和视觉问答。更重要的是，通过将视觉模型与大语言模型相结合，视觉语言模型能够以更加全面且接近人类的方式理解视觉内容。近年来，视觉语言模型在各种计算机视觉任务中展现出了令人瞩目的成果，包括图像理解 [23], [265], 视觉问答 [96], [95],文本到图像生成 [158],语义分割 [26], [252],目标检测 [248], [125], 等等。

在遥感领域，视觉语言模型的应用是一个相对较新且正在兴起的研究方向。视觉语言模型兼具视觉理解、人类般的世界知识和强大的推理能力，这些能力共同提升了遥感数据分析的智能化水平，使其更接近人类的分析方式。总体而言，与仅依赖视觉信息的以往方法相比，视觉语言模型具有多项优势。早期纯视觉模型主要侧重于监督学习，在处理分布外数据时面临挑战；而视觉语言模型则利用语义推理能力建立视觉概念之间的联系。因此，面对分布外数据时，视觉语言模型的表现显著提升。这一突破为零样本/开放词汇视觉理解任务铺平了道路，使其能够更好地适应现实应用，因为其中感兴趣的视觉目标可能属于此前未见的类别或概念。此外，视觉语言模型为探索将通用知识和专家知识融入遥感数据相关的视觉分析任务提供了契机。例如，视觉语言模型知道船舶更可能位于水上，而非陆地上。因此，基于视觉语言模型的目标检测模型往往不会在陆地上检测船舶，展现出提升遥感数据分析效果的潜力。此外，视觉语言模型还可支持执行需要视觉和语言理解能力的新任务，例如图像描述、视觉问答和语言引导编辑。这些任务同时需要视觉和语言理解能力，例如图像描述、视觉问答和语言引导编辑。设想使用基于视觉语言模型的聊天机器人，包括非遥感领域专家在内的用户都可以用日常语言进行对话，从而理解和分析遥感数据。此类系统的前景十分令人期待。近期研究 [138]探索了 Visual ChatGPT [213],的潜力。Visual ChatGPT 是一种基于 GPT架构的先进大语言模型，可用于构建能够处理多种遥感图像处理任务的聊天机器人；在该系统中，先进的 ChatGPT 模型充当高效的提示管理器，确保充分理解用户提示，并将其无缝映射到适当的工具或算法。

随着与遥感数据相关的文本元数据日益丰富，研究人员开始探索在这一领域应用视觉与语言模型的可能性 [189]。近年来，一些早期研究尝试将视觉语言模型用于各种遥感数据分析任务，包括遥感图像描述 [171], [127], [250], [251], [107], [202], [106], [257], [270], 基于文本的遥感图像生成 [14], [22], [256], [224], 基于文本的遥感图像检索 [1], [149], [240], [4], [31], [239], [238], [150], 视觉问答 [122], [259], [19], [3], [13],[235], 场景分类 [92], [173], [144], [197], [110], 语义分割 [26], [252],目标检测 [72], [248],[125], 等。值得注意的是，RS5M [255] 通过从公开数据集中仔细筛选遥感相关图像，构建了一个大规模遥感图像描述数据集，并利用 BLIP-2 模型 [95] 自动生成图像描述。开创性工作 RSGPT [62]构建了高质量的人工标注遥感图像描述数据集（RSICap），推动了遥感领域大型视觉语言模型的发展，并提出了一种用于遥感图像描述和视觉问答的基于GPT的模型。随着大规模遥感数据集日益丰富以及深度

学习技术不断进步，视觉语言模型有望在未来的遥感应用中发挥重要作用。

在这项工作中，我们全面回顾了遥感领域模型从视觉到语言再到视觉语言模型的发展历程。具体而言，我们对遥感领域视觉语言模型的最新进展进行了广泛的文献调研。此外，我们还为遥感应用中的视觉语言模型领域未来可能的研究方向提供了宝贵见解与建议。我们的工作有助于更好地了解视觉语言模型当前的前沿水平，并为该领域的研究人员提供指引，帮助他们探索这些模型在遥感任务中的潜力。

II. 从视觉中心模型到视觉语言模型

A. 以视觉为中心的模型

最常用的视觉中心模型是卷积神经网络（CNN）。卷积神经网络能够对输入图像执行卷积运算，随后进行池化和非线性激活函数处理，从而提取高级特征；因此，它们已成为最受欢迎、最成功的视觉模型之一。这些模型通常使用反向传播进行训练，反向传播是一种梯度下降方法，用于最小化预测输出与真实标签之间的误差。

卷积神经网络有着悠久的发展历史，其历史可以追溯到 20 世纪 80 年代。然而，直到 1998 年，卷积神经网络才首次应用于图像分类任务。在这项研究中，LeCun 等人 [90]提出了 LeNet-5 架构，该架构由多个卷积层和随后连接的全连接层组成。LeNet-5 架构在 MNIST 数据集上取得了当时最先进的成果，该数据集由手写数字识别样本构成。此后，卷积神经网络架构取得了多项进展。例如，2012 年提出的 AlexNet 架构 [81],在 ImageNet [40]数据集上取得了当时最先进的成果，该数据集包含超过一百万张图像，分为 1,000 个类别。AlexNet 采用了更深的网络、更小的滤波器以及更多的隐藏单元，使其能够捕获图像中更加复杂的特征。2014 年，VGG 架构 [172] 在 ImageNet 数据集上取得了与 AlexNet 相近的性能，同时使用了更少的参数。2015 年，GoogLeNet 架构 [180]，（也称为 Inception）被提出。该架构采用了一种名为 Inception 的模块，该模块并行使用多个尺寸不同的卷积滤波器，使网络能够捕获不同尺度的特征。2016 年，ResNet 架构 [53] 被提出，通过引入残差连接来解决深层网络中的梯度消失问题。ResNet 采用了跳跃连接，使网络能够学习残差函数，而不是直接学习底层映射，从而降低了深层网络的训练难度。ResNet 凭借一个包含 152 层的超深网络，在 ImageNet 数据集上取得了当时最先进的成果。2016 年，DenseNet 架构 [64]被提出，旨在引入层与层之间的密集连接，其中每一层都以一种前馈方式与所有后续层相连接。这种方法能够实现特征复用，并促进梯度流，从而以比传统深度神经网络更少的参数实现更优性能。2017 年，ResNeXt 架构 [217] 被提出，通过使用基数参数提升模型容量，同时避免显著增加计算复杂度。ResNeXt 证明，使用具有不同滤波器尺寸和通道数的并行路径，能够提升深度卷积神经网络的容量和准确率。2019 年，EfficientNet [183] 被提出，通过结合复合缩放、高效块结构和神经架构搜索，在保持较少参数和较低计算需求的同时，在图像识别任务上实现了最先进的性能。

近年来，最初为自然语言处理任务开发的基于Transformer的模型，已在众多计算机视觉任务中得到广泛探索。这些模型被称为视觉Transformer，通过自注意力机制从图像中提取特征，使其能够学习图像不同区域之间的全局依赖关系。自注意力机制的公式为

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{1}$$

其中，通过对查询 Q 与键 K执行点积运算来计算注意力权重，并应用缩放因子 $\sqrt{d_k}$ 和Softmax运算，对注意力权重进行归一化。然后，将得到的权重与相应的值特征 V 相乘，以生成最终的输出特征。

首个用于图像分类的基于Transformer架构是 [43]中提出的ViT。图2 展示了ViT模型的综述。人们提出了ViT架构的若干变体，包括DeiT [187], TNT [51], 和 PVT [205], , 通过融入蒸馏、令牌混合和金字塔视觉Transformer等技术进一步提升了性能。此外，考虑到ViT中粗略的图像块划分过程忽略了局部图像信息，因此提出了移位窗口Transformer（Swin） [121] ，利用沿空间维度移动的窗口来建模全局特征和边界特征。Twins [34]采用了空间可分离的自注意力机制，类似于逐通道卷积 [165], , 以建模局部-全局表征。ViL [247] 用一组局部嵌入替代单个类别令牌，对这些局部嵌入执行内部注意力，并与其对应的二维空间邻居进行交互。VOLO [234] 引入了展望注意力，类似于逐图像块的动态卷积，通过展开、线性权重注意力和折叠操作来关注更精细层级的特征。

B. 大语言模型

近年来，大语言模型（LLM）已成为自然语言处理（NLP）和计算机视觉领域的热门研究课题。这些模型构建了用于视觉和自然语言理解的大规模 Transformer 网络，并在语言建模、文本生成和问答等多种语言相关任务中取得了最先进的性能 [146], [42]。

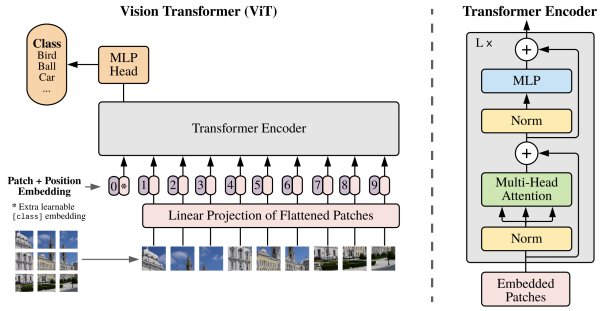

图 2. 视觉Transformer模型的架构综述 [43]

本节将概述大语言模型领域的一些重要进展。

一种开创性且广为人知的大语言模型是 GPT（生成式预训练Transformer）[146]模型，由 OpenAI 开发。GPT 模型使用海量网页文本语料库进行训练，在广泛的语言建模和文本生成任务中取得了令人瞩目的表现。图 3 展示了 GPT 方法的综述。给定一个词元序列 $t = \{t_1, ..., t_N\}$，GPT 将以下似然最大化：

$$\mathcal{L}(t) = \mathbb{E}_{t_i}[\log P(t_i|t_{i-k}, ..., t_{i-1}; \theta)], \tag{2}$$

其中， k 表示上下文窗口的长度， θ表示网络参数。

自GPT模型问世以来，GPT系列经历了多次迭代。GPT-2 [147] 拥有15亿个参数，在许多语言任务中取得了卓越表现，包括语言翻译、摘要生成、问答和文本补全。它能够生成高质量、连贯且流畅的文本，与人类撰写的文本难以区分，因此受到了广泛关注。GPT-3 [16] 表明，预训练大语言模型可以成为零样本学习者，并激发了人们对上下文学习的研究热情。InstructGPT [139],是ChatGPT采用的关键技术之一，它为提升大语言模型的可控性和灵活性引入了一种有前景的方法。该方法通过添加以自然语言短语或模板形式呈现的高层级指令来发挥作用。这些指令可用于引导生成过程，并确保输出文本符合特定约束或要求。最新版本GPT-4 [137] 展现了大规模语言模型在推动人工智能发展方面的巨大潜力，并为智能机器的新时代铺平了道路，使它们能够以更自然的方式理解我们并与我们交流。

另一种流行的大语言模型是谷歌开发的 BERT [42]，。与 GPT 模型不同，BERT 在预训练时采用了双向任务，这意味着它在预训练过程中能够同时考虑一个词的左侧和右侧上下文。这使 BERT 能够捕捉词语之间更细微的关系，并在各种自然语言处理任务中取得最先进的成果。具体而言，BERT 的预训练目标包含两个无监督任务：1）掩码语言模型（masked LM），根据附近的词元预测若干被掩码的词元；

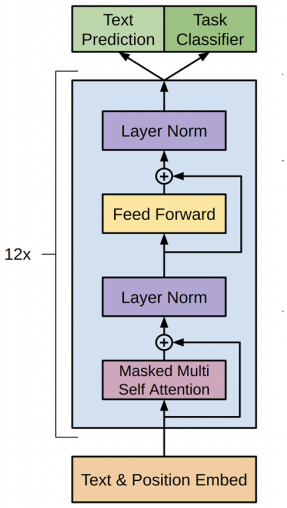

图 3. GPT 模型的网络架构和训练目标 [146]。

2）下一句预测，即根据前一句预测下一句，如图所示4。给定一个词元序列 $t = \{t_1, ..., t_N\}$，掩码语言模型的训练目标定义如下：

$$\mathcal{L}(t) = \mathbb{E}_{\mathcal{M}}[\sum_{i \in \mathcal{M}} \log P(t_i|t^{\mathcal{M}}; \theta)], \tag{3}$$

其中， $\mathcal{M}$ 表示随机掩码的位置， $t^{\mathcal{M}} = \{t_i : i \notin \mathcal{M}\}_{i=1}^{N}$ 表示根据 $\mathcal{M}$进行掩码处理后的序列， θ 表示网络参数。

BERT 模型有许多变体。例如，RoBERTa [120]，由 Facebook AI 开发，通过增加迭代次数和采用数据增强技术（包括动态掩码和加噪）改进了 BERT。ALBERT [85] 提出了三种技术，包括因式分解嵌入参数化、跨层参数共享和句子顺序预测任务，以减小 BERT 的模型规模并提高模型训练速度。MacBERT [36]提出用另一个同义词替换 [MASK]标记，而不是在掩码位置使用 [MASK] 标签。

除 GPT 和 BERT 外，近年来还出现了几种其他大语言模型。T5 [148] 通过在海量的多样化文本数据语料库上进行任务无关预训练，将单一统一架构应用于广泛的自然语言处理任务，旨在创建通用语言模型。CoT [211] 提出了一种实用技术，称为思维链提示，使大语言模型能够进行中间推理。

C. 视觉语言模型

鉴于预训练模型在计算机视觉和自然语言处理（NLP）领域取得的成功，研究人员尝试预训练能够整合这两种模态的大规模模型，这类模型称为视觉语言模型（VLM）。视觉语言模型可分为两种架构：融合编码器和双编码器模型。融合编码器模型使用多层跨模态 Transformer

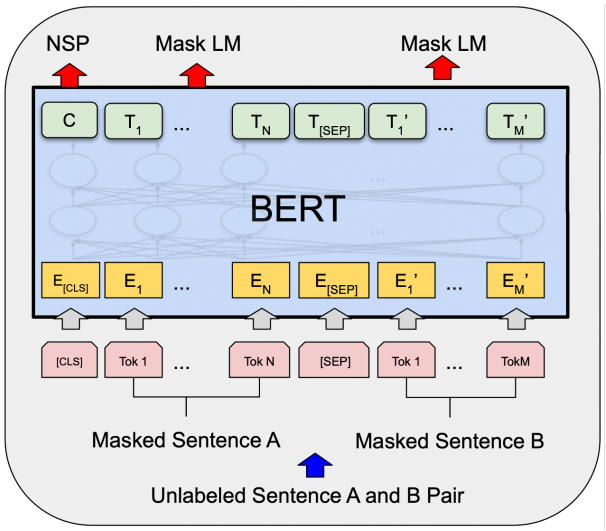

图 4. BERT 模型的预训练目标 [120]。

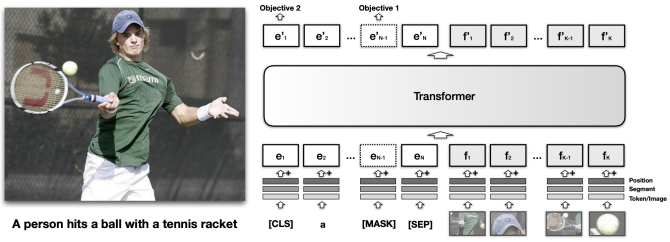

图 5. VisualBERT 架构示意图 [99]

编码器对图像与文本配对进行联合编码，并融合其视觉和文本表征。与此同时，双编码器模型分别对图像和文本进行编码，并使用点积或多层感知机来捕捉不同模态之间的交互。

1）融合编码器：融合编码器接收视觉特征和文本嵌入作为输入，并采用多种融合技术来捕捉视觉模态与文本模态之间的交互。经过自注意力或交叉注意力操作后，将最终层的潜在特征视为不同模态的融合表示。VisualBERT [99]是一项开创性工作，它利用自注意力机制隐式地将输入文本中的元素与相关输入图像中的区域对齐。它结合了 BERT [42] 来处理自然语言，并使用预训练的 Faster-RCNN [156] 生成目标提议。原始文本与从目标提议中提取的图像特征一起，被视为无序的输入词元，并输入 VisualBERT，通过多个 Transformer 层联合处理这些词元，以捕捉复杂的关联（见图 5）。随后，包括 Uniter [27],、OSCAR [104], 和 InterBert [112],在内的多个视觉语言模型，采用 BERT 作为文本编码器、Faster-RCNN 作为目标提议生成器，对视觉与语言之间的交互进行建模。

与单流架构中采用的自注意力操作不同，双流架构利用交叉注意力机制来捕捉视觉和语言模态之间的交互。交叉注意力层通常由两个单向子层组成，一个处理从语言到视觉的信息，另一个处理从视觉到语言的信息。这些子层促进了两种模态之间的信息交换和语义对齐。一个著名的例子是ViLBERT [124], , 它分别处理视觉和文本输入，并采用协同注意力 Transformer层来实现模态之间的信息交换。图6 展示了每个流由Transformer层（TRM）和新型协同注意力 Transformer模块（Co-TRM）组成。此外，近期的研究工作，如LXMERT [182], Visual Parsing [226], ALBEF [97] 和Wen-Lan [68] 也在交叉注意力之前采用独立的 Transformer，以解耦模态内交互和跨模态交互。Chen 等人 [23] 提出了VisualGPT，通过利用一种新颖的自复活编码器-解码器注意力机制，使预训练语言模型能够适应少量领域内图文数据。

2）双编码器：双编码器使用两个独立的编码器，分别对每种模态进行编码，以独立编码视觉和文本信息。随后，通过注意力层或点积等操作，将相应编码器编码得到的图像和文本嵌入投影到共享语义潜在空间中，用于计算视觉与语言嵌入之间的相似度分数。与融合编码器机制相比，双编码器机制不使用 Transformer 中复杂的交叉注意力，因为它会预先计算并存储图像和文本嵌入，从而使视觉-语言交互建模更加高效。例如，对比语言-图像预训练（CLIP） [145], 如图 7所示，联合使用文本编码器和图像编码器，准确匹配（图像、文本）样本对。设 (x_i^I, x_i^T) 为一批 N 图像 - 文本对，且 z_i^I 和 z_j^T 分别表示 i_{th} 图像和 j_{th} 文本的归一化嵌入。CLIP 使用 InfoNCE [135] 计算损失，如公式 4所示。

$$
\begin{aligned}
L_{I2T} &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(sim(z_i^I, z_i^T)/\tau)}{\sum_{j=1}^{N}\exp(sim(z_i^I, z_j^T)/\tau)}, \\
L_{T2I} &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(sim(z_i^T, z_i^I)/\tau)}{\sum_{j=1}^{N}\exp(sim(z_i^T, z_j^I)/\tau)},
\end{aligned}
\tag{4}
$$

其中，sim(,)是使用点积计算的相似度函数，τ 是用于缩放 logits 的温度变量。该损失函数对于图像和文本编码器是对称的，最终损失是图像到文本匹配损失与文本到图像匹配损失的平均值。

此外，ALIGN [71] 提出了一种双编码器架构，通过对比损失在跨模态注意力之前对齐图像和文本表征。该损失会拉近匹配的图像—文本对，并推远不匹配的图像—文本对。为克服 CLIP 的局限性，Li 等人 [108]提出了一种名为数据高效 CLIP 的新型训练范式，提高了通用视觉特征的学习效率。该方法结合了模态内自监督和模态间多

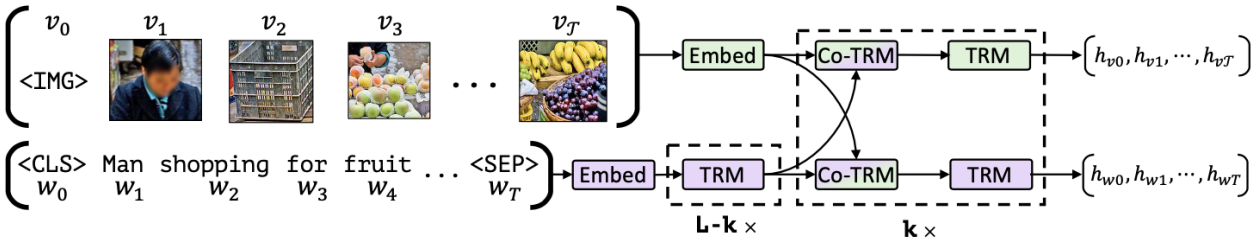

图 6. ViLBERT 架构示意图 [124]。

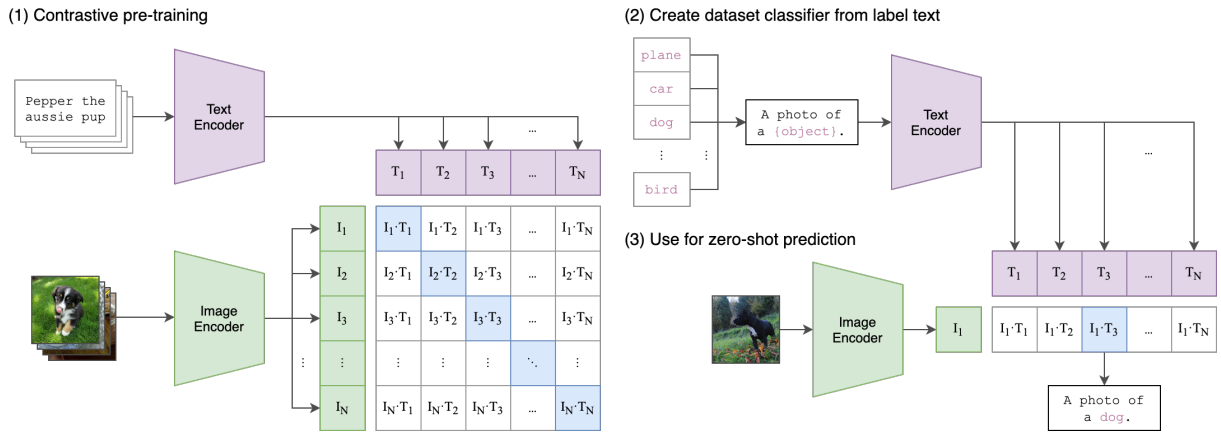

图 7. CLIP [145] 架构示意图。

视图监督，并进一步引入了一种监督信号，通过文本嵌入查找相似文本。Alayrac 等人提出了 Flamingo [5]，，这是一种仅使用少量输入/输出示例进行训练的模型，可接受交错的视觉数据和文本作为输入，并以开放式方式生成文本。具体而言，Flamingo 使用视觉编码器提取的时空特征，并初始化交叉注意力层，将这些视觉词元交错插入预训练语言模型的各层之间。类似地，Li 等人提出了一种通用且高效的视觉语言模型 BLIP-2 [95]，，该模型利用现成的冻结预训练图像编码器和冻结大语言模型，实现了视觉语言预训练的自举。他们首先利用冻结图像编码器自举视觉语言表征学习，随后利用冻结语言模型自举视觉到语言的生成式学习。近期，Kirillov 等人构建了基础分割模型 SAM [78]，可在零样本设置下应用于多种图像理解任务，例如语义分割、边缘检测、全景分割和实例分割。SAM 采用了一个大型图像编码器来获取图像嵌入，随后使用提示编码器根据输入提示生成查询，以输出多个目标掩码和置信度分数。Zhu 等人提出了 MiniGPT-4 [266]，通过训练单个线性投影层，将预训练 ViT 编码的视觉特征与预训练大语言模型对齐。这个仅含 13B 个参数的简化模型在多种跨模态任务上表现出色，例如图像描述、根据图像创作诗歌/故事，以及根据草稿构建网站图像。此外，InstructBLIP [38]，LLAVA [116],和 VisionLLM [204] 通过指令微调引入了通用视觉语言模型。Kosmos [65]，[141]介绍了一种从头开始在网络规模多模态语料库上训练的多模态大语言模型（MLLM）。该模型旨在感知各种模态、进行上下文学习（即以少样本方式学习）并遵循指令（即以零样本方式执行）。

III. 遥感视觉语言模型

A. 视觉中心基础模型

基础模型是指在海量数据上以任务无关方式训练的大型预训练深度学习神经网络。这些模型可以通过微调、少样本或零样本学习应用于各种下游任务。基础模型的示例包括 GPT-3 [16],BERT [42],和 T5 [148]。这些模型已在大量文本数据上进行预训练，可针对广泛的自然语言处理任务进行微调，例如语言翻译、问答和文本分类。在遥感（RS）领域，预训练对于提升分类、检测和分割任务的性能至关重要 [220]。以往的方法主要利用 ImageNet 数据集进行预训练。然而，由于自然图像与遥感图像之间存在显著差异，将 ImageNet 预训练模型迁移到遥感任务时会面临巨大的领域差距，而

*https://sites.google.com/site/michelevolpiresearch/data/zurich-dataset

方法	Year	Task	视觉编码器	文本编码器	数据集
RingMo [177]	2022	FM	视觉Transformer [43]、Swin Transformer [121]	-	从多个来源收集的两百万张遥感图像
Wang等人 [200]	2022	FM	视觉Transformer [43],ViTAE [225]	-	公开数据集和 GF-2
地理空间基础模型 [133]	2023	FM	Swin-B [121]	-	MillionAID [123]
Cha等人	2023	FM	视觉Transformer [43]	-	GeoPile [133]
Shi 等人 [171]	2017	RSIC	VGG-f [21]	-	MillionAID [123]
Lu 等人 [127]	2017	RSIC	AlexNet [81], VGG [172], GoogLeNet [180]	长短期记忆网络 [56]	Google Earth，GF-2
Zhang 等人 [250]	2017	RSIC	CaffeNet [81]	循环神经网络 [57]	RSICD [127]
VAA [254]	2019	RSIC	VGG16 [172]	长短期记忆网络 [56]	UCM [229]
Zhang 等人 [251]	2019	RSIC	VGG16 [172]	长短期记忆网络 [56]	UCM-Captions [143], Sydney-Captions [143]
Li 等人 [107]	2020	RSIC	ResNet-101 [53]	长短期记忆网络 [56]	UCM-Captions [143],Sydney-Captions [143], RSICD [127]
VRTMM [170]	2020	RSIC	VGG16 [172]	Transformer [191]	UCM-Captions [143],Sydney-Captions [143], RSICD [127]
Wang等人 [202]	2020	RSIC	AlexNet [81], VGG [172], ResNet [53]	Transformer [191]	西北工业大学-RESISC45 [28]
Li 等人 [106]	2020	RSIC	AlexNet [81], VGG [172], ResNet [53], 谷歌网络 [180]	长短期记忆网络 [56]	UCM-Captions [143], Sydney-Captions [143], RSICD [127]
Zhao等人 [257]	2021	RSIC	ResNet-50 [53]	长短期记忆网络 [56]	UCM-Captions [143],Sydney-Captions [143], RSICD [127]
Zia 等人 [270]	2022	RSIC	ResNet [53]	Transformer [191]	UCM-Captions [143],Sydney-Captions [143], RSICD [127]
VLCA。 [212]	2023	RSIC	CLIP [145]	GPT-2 [147]	DIOR-Captions [212]
Bejiga等人 [14]	2019	T2IG	生成对抗网络 [35]	-	中分辨率成像光谱仪
BTD-sGAN [22]	2021	T2IG	D-sGAN [130]	-	Oxford-102，GF
StrucGAN [256]	2021	T2IG	AttnGAN [223]	长短期记忆网络 [56]	RSICD [127]
Txt2Img-MHN [224]	2022	T2IG	向量量化变分自编码器 [190]，向量量化生成对抗网络 [44]	字节对编码 [169]	RSICD [127]
双线性深度网络 [1]	2020	T2IR	VGG [172], Inception V3 [181],ResNet-50 [53], EfficientNet [183]	长短期记忆网络 [56]	TextRS [1]
Rahhal等人 [149]	2020	T2IR	BiT [80]	双向长短期记忆网络 [49]	TextRS [1], UCM [229]
LW-MCR [239]	2021	T2IR	SqueezeNet [70]	组卷积神经网络	悉尼 [245],UCM [229], RSITMD [238], RSICD [127]
SAM [31]	2021	T2IR	Inception V3 [181]	双向门控循环单元 [168]	UCM-Captions [143],Sydney-Captions [143], RSICD [127], NWPU-RESISC45图像描述数据集
GaLR [240]	2022	T2IR	卷积神经网络 [240]	门控循环单元 [33]	RSICD [127], RSITMD [238]
Rahhal等人 [4]	2022	T2IR	视觉Transformer [43]	Transformer [191]	RSICD [127], RSITMD [238], UCM [229]
AMFMN [238]	2022	T2IR	ResNet-18 [53]	门控循环单元 [33]	悉尼 [245],UCM [229], RSITMD [238], RSICD [127]
Rahhal等人 [150]	2023	T2IR	Transformer [191]	Transformer [191]	TextRS [1], UCM [229],悉尼 [245], RSICD [127]
Lobry等 [122]	2020	VQA	ResNet-152 [53]	长短期记忆网络 [56]	RSVQA [122]
Zheng等 [259]	2021	VQA	VGG-16 [172]	门控循环单元 [33]	RSIVQA [259]
Chappuis 等人 [19]	2022	VQA	ResNet-152 [53]	BERT [42]	RSVQA [122]
Al 等人 [3]	2022	VQA	视觉Transformer	Transformer [191]	VQA-TextRS [3]
Bazi 等人 [13]	2022	VQA	视觉Transformer [43]	CLIP [145]	Sentinel-2 和航空影像
Yuan 等人 [235]	2022	VQA	卷积神经网络 [235]	循环神经网络 [57]	RSVQA [122], RSIVQA [259]
Yuan 等人 [236]	2022	VQA	视觉Transformer [43]	循环神经网络 [57]	变化检测视觉问答 [236]
Sun 等人 [179]	2022	VG	Darknetz[46]	BERT [42]	RSVG [179]
Zhan 等人 [242]	2023	VG	CNN	BERT [42]	DIOR-RSVG [242]
Li 等人 [92]	2017	SC	GoogLeNet [180]	Word2vec [134]	UCM [230],遥感场景分类7 [271],悉尼 900 [92]。
Sumbul 等人 [173]	2017	SC	卷积神经网络 [173]	Word2vec [134]	自行收集
Quan 等人 [144]	2018	SC	GoogLeNet [180]	Word2vec [134]	UCM [230],AID [214]。
Wang等人 [197]	2021	SC	ResNet-152 [53]	Word2vec [134]	UCM [230],AID [214],西北工业大学-45 [28]
Li 等人 [110]	2022	SC	ResNet-101 [53]	Word2vec [198], Fasttest [74], [15], GloVe [215], BERT [42]	UCM [230],AID-30 [214],西北工业大学-45 [28]
Zhang 等人 [248]	2023	OD	Faster R-CNN [156]	GloVe [215]	西北工业大学 VHR-10 [29], DIOR [98]
Lu 等人 [125]	2023	OD	Faster R-CNN [156]	门控循环单元 [33]	NWPU VHR-10数据集 [29], DIOR [98], FAIR1M数据集 [178]
Chen等人 [26]	2019	SS	ResNet [26]	-	Vaihingen [160], 苏黎世夏季数据集*
Jiang等 [72]	2021	SS	ResNet [53]	-	ISPRS Vaihingen [160]
Zheng等 [252]	2023	SS	三维卷积神经网络 [252]	GPT-2 [147]	休斯顿 [39], [89],帕维亚 [47], GID [186]

表 I 遥感领域知名视觉语言模型总结。“-”表示不适用。“FM”表示基础模型，“RSIC”表示遥感图像描述，“T2IG”表示基于文本的图像生成，“VQA”表示视觉问答，“VG”表示视觉定位，“SC”表示场景分类，“OD”表示目标检测，“SS”表示语义分割。

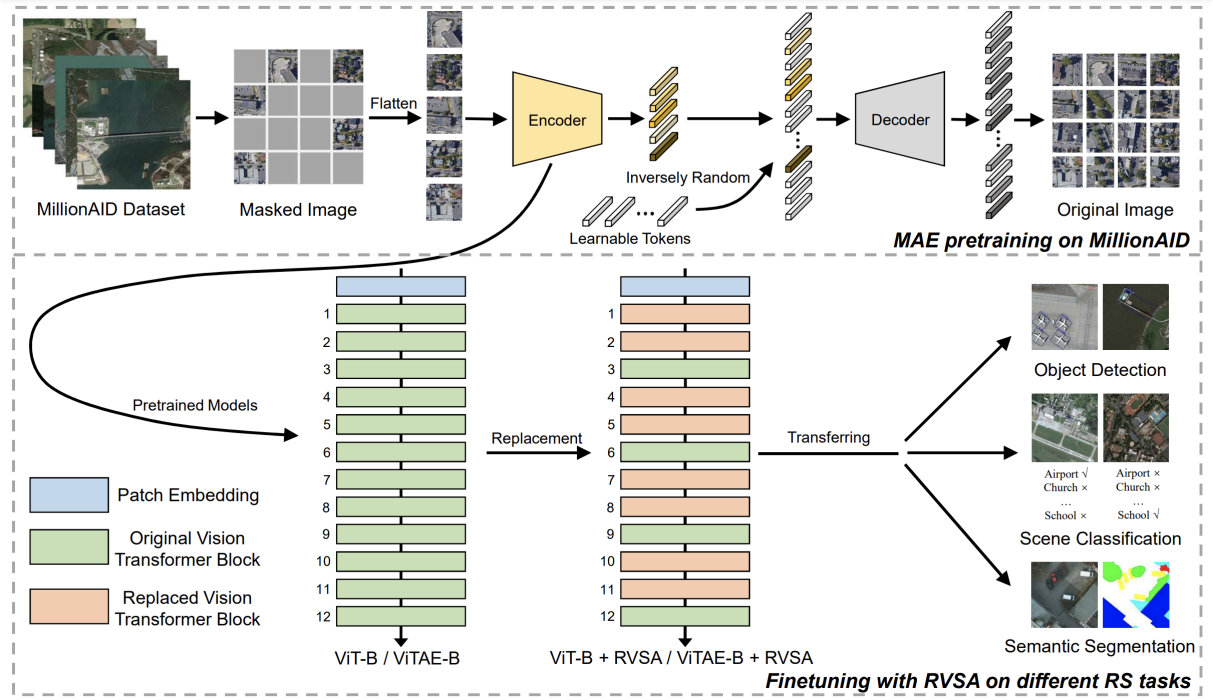

图 8. [200]中的遥感视觉变换器网络方法概览。

遥感图像。因此，设计一个专为遥感数据定制的基础模型是必要的。

研究人员通过两种方法实现这一目标：监督学习和自监督学习。在监督学习中，[199]研究人员在 MillionAID 数据集（一个大规模遥感数据集）上预训练深度神经网络，并提升了模型在其他下游遥感任务上的性能。然而，对大量标注数据的需求仍然是一大障碍，因为这可能会妨碍更大规模模型的训练。因此，自监督技术逐渐成为开发遥感基础模型的主要方法，因为这类技术能够利用大量未标注数据 [207]。早期研究 [2], [100], [6], [132], [54] 采用对比学习训练基础模型，并融入遥感特有的信息，例如地理数据、时间序列数据、音频数据等。近年来，掩码图像建模（MIM）在计算机视觉领域日益受到关注，例如 BEiT [10], MAE [52], SimMIM [218],，因为它无需标签、数据增强，也不必选择正负样本对，因此更容易利用海量数据。一些研究将 MIM 用于开发遥感基础模型。例如，[177]研究人员从卫星和航空平台收集了两百万张遥感图像，构建了一个大规模遥感数据集。基于该数据集，他们设计了首个生成式自监督遥感基础模型 RingMo。RingMo 在四种下游任务的八个数据集上取得了最先进的表现，这些任务包括变化检测、场景识别、目标检测和语义分割。[200] 研究人员首次尝试构建一种约含一亿个参数、面向遥感任务的大型视觉基础模型——纯视觉变换器。其方法概述如图8所示。他们还引入了旋转可变尺寸窗口注意力机制，以增强视觉Transformer适应遥感图像的能力。[133]研究人员发现，在构建地理空间基础模型时，不应忽视在多样化数据集（如 ImageNet-22k）上预训练的模型，因为这些模型的表征依然有效。因此，他们以可持续的方式构建了面向地理空间应用的基础模型。[18]研究人员开发了遥感领域首个十亿级基础模型，并证明了将模型规模从百万级扩大到十亿级的效果。关于遥感基础模型的全面文献综述可见 [73]。

B. 图像描述

遥感图像描述（RSIC）是一项复杂的任务，要求机器理解遥感（RS）图像的内容，并使用自然语言对其进行描述。这是一项具有挑战性的任务，因为生成的描述不仅必须捕捉不同尺度的地面元素，还要描述它们的属性以及相互关联的方式。不同于旨在预测单个标签或词语的其他任务，RSIC旨在生成全面的句子。为了生成简洁且有意义的句子描述，识别和理解不同层次的地面元素、分析其属性，并从高层次角度利用类别依赖关系和空间关系至关重要。

开创性工作 [171] 提出了一种解决方案，以应对遥感图像中相同地面元素在不同地理尺度下呈现不同语义这一挑战。他们引入了全卷积网络（FCN），为从 Google Earth 和高分二号获取的遥感图像生成全面且鲁棒的句子，同时确保理想的运行速度。他们首先处理了三个子任务，包括

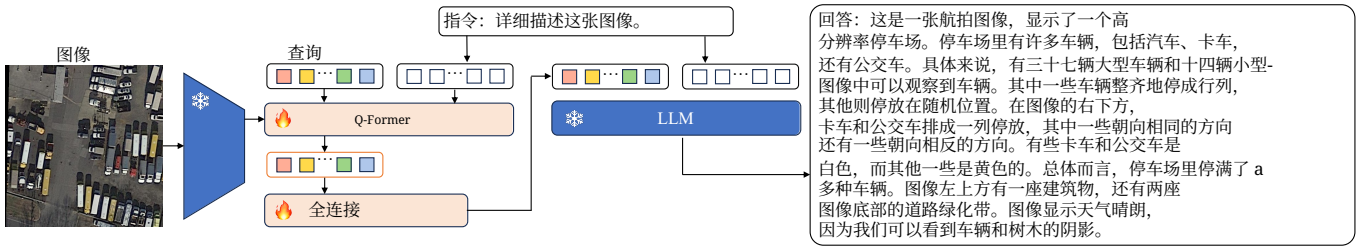

图 9. [62]提出的一种遥感图像描述方法综述。该模型由一个用于视觉特征提取的图像编码器、一个 Q-Former [38]（用于特征压缩）、一个线性投影层和一个冻结的大语言模型组成。

关键实例检测、环境分析和景观分析，然后整合各阶段的结果以生成语言描述。同样，Lu 等人 [127] 给出了全面描述遥感图像的指导，同时考虑尺度歧义、类别歧义和旋转歧义，并通过收集一万多张遥感图像，创建了一个大规模遥感图像描述数据集。相比之下，Zhang 等人 [250]采用卷积神经网络检测遥感图像中的主要目标，并使用循环神经网络语言模型生成对检测到的目标的自然语言描述。

近期研究还探索了利用注意力机制更好地对齐视觉特征与文本嵌入。例如，Zhang 等人 [254]提出了一种视觉对齐注意力模型，并采用精心设计的视觉对齐损失。该损失通过显式评估注意力关注的图像特征与相应词嵌入向量之间的特征相似度来确定。为解决训练注意力层时非视觉词语带来的影响，他们引入了视觉词汇表，在计算视觉对齐损失时从句子中剔除这类词语。此外，为弥合遥感图像中低级特征与高级语义之间的语义鸿沟，Zhang 等人 [251]利用全卷积网络（FCN）生成图像特征，并通过注意力机制获取中间向量，再将其作为 LSTM 解码器的输入，以生成遥感图像的描述。在此基础上，Li 等人 [107] 提出了一种多层次注意力模型，包含三种注意力结构：对图像不同区域的注意力、对不同词语的注意力，以及对视觉信息和语义信息的注意力。此外，他们还修正了现有数据集中的错误，包括词语错误、语法错误和不恰当的描述。另一项研究中，Shen 等人 [170]提出了一种基于变分自编码器（VAE）和强化学习的两阶段多任务学习模型，用于遥感图像描述。他们在第一阶段联合微调卷积神经网络与变分自编码器，随后利用 Transformer 和强化学习处理空间特征与语义特征，以生成文本描述。同样，Kandala 等人 [75]采用了基于 Transformer 的编码器—解码器网络来进行遥感图像描述。具体而言，为应对训练数据有限的问题，他们使用了一个辅助解码器，该解码器针对多标签场景分类进行训练，并利用其与图像描述的概念相似性以及突出语义类别的能力，辅助编码器完成训练。不同于使用

缺乏可解释性的编码器—解码器架构，Wang等人 [202] 提出了一种用于遥感图像描述的可解释词语—句子框架，该框架由两个网络组成：词语提取器和句子生成器。第一个网络提取给定遥感图像中的有用词语，第二个网络则将这些词语组织成连贯的句子。此外，Li 等人 [106] 提出了一种新颖的截断交叉熵损失，以解决遥感图像描述中由交叉熵损失导致的过拟合问题。RSGPT [62]微调了单个投影层，以对齐视觉特征与大语言模型，并在图像描述和视觉问答任务上均展现出出色的性能。图9 概述了 RSGPT 方法。

为了解决以往方法忽视领域知识的问题，Zhao 等人 [257] 提出了一种细粒度结构化注意力模型，以利用高分辨率遥感图像中语义内容的结构特征。此外，Zia 等人 [270]首次采用多尺度视觉特征编码器从遥感图像中提取细节信息，随后利用自适应多头注意力解码器，根据提取的多尺度特征优化描述生成。此外，他们还引入了主题敏感词嵌入，以生成更贴近人类表达且更具创新性的描述。最近，Wei 等人 [212] 提出了一种视觉语言对齐网络，用于联合表征视觉与语言，以实现遥感图像描述。该研究使用了 DIOR-Captions 数据集，这是一个新创建的数据集，通过人工标注的中文和英文内容，丰富了用于光学遥感图像目标检测的数据集。

我们总结了表 II 中各比较方法在 UCM-CAPTION [143]，悉尼-caption [143]，和 RSICD[127]数据集上的表现。我们报告了四项评估指标，包括 BLEU [140]，METEOR [8]，ROUGE-L [32],和 CIDEr-D [192]，，以验证比较模型的有效性。结果表明，现有视觉语言模型已取得不错的表现。具体而言，RS-GPT [62]优于所有比较方法，表现最佳。

然而，值得注意的是，前述三个常用于训练遥感图像描述视觉语言模型的数据集分别包含 2100、613 和 10921 张图像。这些数量远少于计算机视觉领域训练视觉语言模型时通常使用的数百万张图像。数据集规模上的这种差异在一定程度上

方法	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE L	CIDEr
UCM-CAPTION [143]							
VLAD + 循环神经网络 [127]	63.11	51.93	46.06	42.09	29.71	58.78	200.66
VLAD +长短期记忆网络 [127]	70.16	60.85	54.96	50.30	34.64	65.20	231.31
mRNN [143]	60.10	50.70	32.80	20.80	19.30	-	214.00
mLSTM [143]	63.50	53.20	37.50	21.30	20.30	-	222.50
mGRU [105]	42.56	29.99	22.91	17.98	19.41	37.97	124.82
mGRU 嵌入词 [105]	75.74	69.83	64.51	59.98	36.85	66.74	279.24
CSMLF [194]	37.71	14.85	7.63	5.05	9.44	29.86	13.51
SAA [126]	79.62	74.01	69.09	64.77	38.59	69.42	294.51
软注意力 [221]	74.54	65.45	58.55	52.50	38.86	72.37	261.24
硬注意力 [221]	81.57	73.12	67.02	61.82	**42.63**	76.98	299.47
SD-RSIC [175]	74.80	66.40	59.80	53.80	39.00	69.50	213.20
RTRMN（语义） [196]	55.26	45.15	39.62	35.87	25.98	55.38	180.25
RTRMN（统计） [196]	80.28	73.22	68.21	63.93	42.58	77.26	312.70
SVM-D 词袋 [59]	76.35	66.64	58.69	51.95	36.54	68.01	271.42
SVM-D CONC [59]	76.53	69.47	64.17	59.42	37.02	68.77	292.28
后处理 [60]	79.73	72.98	67.44	62.62	40.80	74.06	309.64
RSGPT [62]	**86.12**	**79.14**	**72.31**	**65.74**	42.21	**78.34**	**333.23**
Sydney-Captions [143]							
VLAD + 循环神经网络 [127]	56.58	45.14	38.07	32.79	26.72	52.71	93.72
VLAD +长短期记忆网络 [127]	49.13	34.12	27.60	23.14	19.30	42.01	91.64
mRNN [143]	51.30	37.50	20.40	19.30	18.50	-	161.00
mLSTM [143]	54.60	39.50	22.30	21.20	20.50	-	186.00
mGRU [105]	69.64	60.92	52.39	44.21	31.12	59.17	171.55
mGRU 嵌入词 [105]	68.85	60.03	51.81	44.29	30.36	57.47	168.94
CSMLF [194]	59.98	45.83	38.69	34.33	24.75	50.18	75.55
SAA [126]	68.82	60.73	52.94	45.39	30.49	58.20	170.52
软注意力 [221]	73.22	66.74	62.23	58.20	39.42	71.27	249.93
硬注意力 [221]	75.91	66.10	58.89	52.58	38.98	71.89	218.19
SD-RSIC [175]	72.40	62.10	53.20	45.10	34.20	63.60	139.50
SVM-D 词袋 [59]	77.87	68.35	60.23	53.05	37.97	69.92	227.22
SVM-D CONC [59]	75.47	67.11	59.70	53.08	36.43	67.46	222.22
后处理 [60]	78.37	69.85	63.22	57.17	39.49	71.06	255.53
RSGPT [62]	**82.26**	**75.28**	**68.57**	**62.23**	**41.37**	**74.77**	**273.08**
RSICD [127]							
VLAD + 循环神经网络 [127]	49.38	30.91	22.09	16.77	19.96	42.42	103.92
视觉语言注意力检测器（VLAD） +长短期记忆网络 [127]	50.04	31.95	23.19	17.78	20.46	43.34	118.01
mRNN [143]	45.58	28.25	18.09	12.13	15.69	31.26	19.15
mLSTM [143]	50.57	32.42	23.29	17.46	17.84	35.02	31.61
mGRU [105]	42.56	29.99	22.91	17.98	19.41	37.97	124.82
mGRU 嵌入词 [105]	60.94	46.24	36.80	29.81	26.14	48.20	**159.54**
CSMLF [194]	57.59	38.59	28.32	22.17	21.28	44.55	52.97
SAA [126]	59.35	45.11	35.29	28.08	26.11	49.57	132.35
软注意力 [221]	65.13	49.04	39.00	32.30	26.39	49.69	90.58
SD-RSIC [175]	64.50	47.10	36.40	29.40	24.90	51.90	77.50
RTRMN（语义） [196]	62.01	46.23	36.44	29.71	28.29	**55.39**	151.46
RTRMN（统计） [196]	61.02	45.14	35.35	28.59	27.51	54.52	148.20
SVM-D 词袋 [59]	61.12	42.77	31.53	24.11	23.03	45.88	68.25
SVM-D CONC [59]	59.99	43.47	33.55	26.89	22.99	45.57	68.54
MLAT [115]	66.90	51.13	41.14	34.21	27.31	50.57	94.27
后处理 [60]	62.90	45.99	35.68	28.68	25.30	47.34	75.56
RSGPT [62]	**70.32**	**54.23**	**44.02**	**36.83**	**30.10**	53.34	102.94

表 II UCM-CAPTION、SYDNEY-CAPTION 和 RSICD 数据集上的遥感图像描述性能。数据取自 [62]

限制了遥感图像描述视觉语言模型准确率大幅提升的潜力。更重要的是，除了 RSGPT [62]，中引入的 RSICap 数据集外，现有的遥感图像描述数据集仅包含简短描述。亟需构建包含详细描述的大规模数据集，以进一步推动遥感图像描述任务的发展。

C. 基于文本的图像生成

基于文本的图像生成是一个新兴研究领域，它结合自然语言处理与计算机视觉，根据文本描述生成逼真的图像。这项技术应用于遥感（RS）图像，在现实应用中具有巨大潜力。一个可能受益的领域是帮助城市规划师根据文本描述生成逼真的遥感图像，使他们能够评估设计的可行性，并做出更明智的决策。另一种潜在的

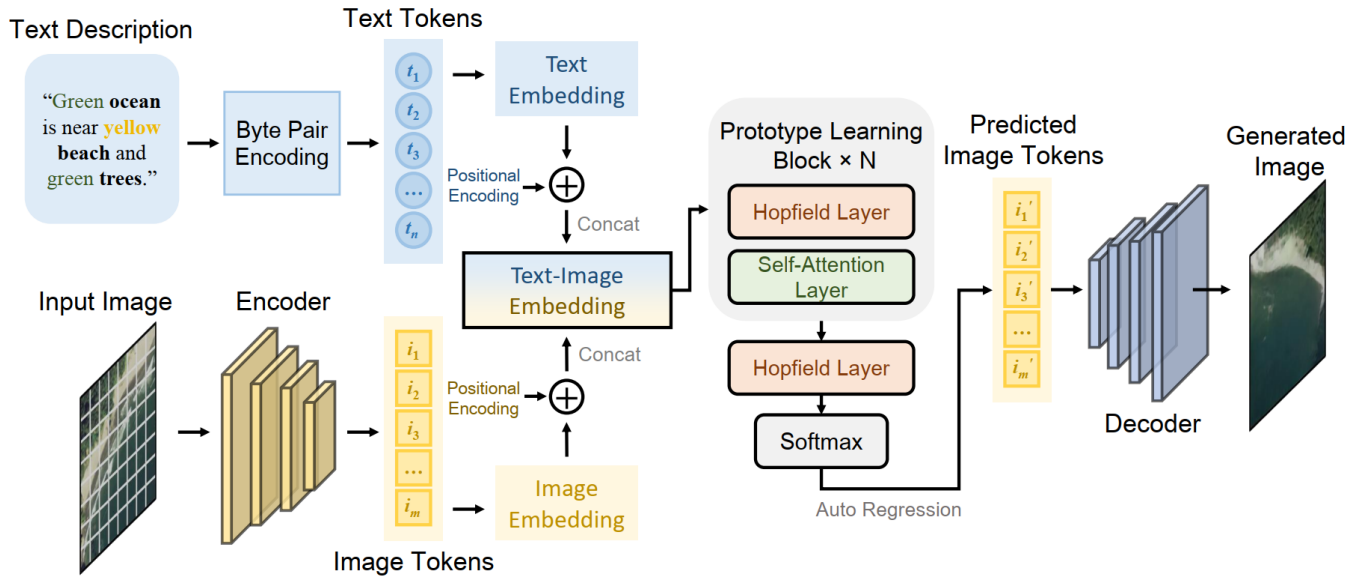

图 10. 所提出方法的示意图 [224] 用于基于文本的遥感图像生成。

应用场景是生成高质量的遥感图像标注数据集，这通常是一项颇具挑战且耗时的工作。基于文本的图像生成技术可用于根据文本描述创建合成遥感数据集，从而缓解标注样本短缺的问题。

已有多项研究探索了使用生成对抗网络（生成对抗网络）进行基于文本的遥感图像生成。例如，Bejiga等人 [14]率先开展了利用生成对抗网络根据文本描述合成遥感图像的研究。他们将文本表示转换为像素值，以捕捉尺寸和颜色等图像特征，并据此根据对地理区域的古代文本描述生成遥感图像。此外，Chen等人 [22] 提出了一种名为基于文本的深度监督生成对抗网络（BTD-sGAN）的增强型生成对抗网络，以解决遥感图像数量不足、质量较差的问题。BTD-sGAN采用双层UNet++ [264] 框架，并以柏林噪声、图像分割图和编码文本向量为输入来生成遥感图像。该方法在自然图像数据集Oxford-102上取得了3.66的Inception 分数，并在其自采集的遥感数据集上展现出良好的定性结果。此外，Zhao等人 [256]提出了一种结构化生成对抗网络，即StrucGAN，可在考虑结构信息的同时根据文本描述生成遥感图像。所提出的StrucGAN采用双向长短期记忆网络（长短期记忆网络 [56]）作为文本编码器，以捕捉语义特征，并采用StackGAN逐步生成从小到大不同尺度的图像。该模型还包含一个由区域建议模块（RPM）和判别器组成的分支，以确保生成结构合理的图像。该方法在广受欢迎的RSICD 数据集上取得了5.84的Inception 分数，前10名精确率达到16.20%。此外，Xu等人 [224] 提出了一种新颖的方法 Txt2Img-MHN，利用现代

方法	Inception 分数(↑)	FID 分数(↓)
Attn-GAN [223]	**11.71**	95.81
DAE-GAN [161]	7.71	93.15
DF-GAN [184]	9.51	109.41
Lafite [262]	10.7	**74.11**
DALL-E [152]	2.59	191.93
Txt2Img-MHN（向量量化变分自编码器） [153]	3.51	175.36
Txt2Img-MHN（向量量化生成对抗网络） [153]	5.99	102.44

表 III 基于文本的遥感图像生成在 RSICD 测试集上的结果。数值引自 [224]

Hopfield 网络 [153]。如图所示，图10，与以往侧重于学习联合文本图像表示的研究不同，Txt2Img-MHN 将 Hopfield 网络应用于文本图像嵌入，并以分层方式进行原型学习，从而确保学习步骤具有更好的粗粒度和细粒度表现，同时实现更丰富的语义编码。该方法在 RSICD 测试数据集上取得了显著的性能指标，包括 5.99 的 Inception 分数和 102.44 的 FID 分数。此外，我们在表 III 中总结了所有比较方法在 RSICD 测试数据集上的 Inception 分数 [163]和 FID 分数 [55]。

D. 基于文本的图像检索

如何高效地组织和管理海量遥感（RS）数据，长期以来一直是遥感社区面临的一项重大挑战。为应对这一挑战，基于文本的图像检索（TBIR）已成为一个重要的研究主题，旨在为遥感数据管理提供有效的解决方案。图像检索的主要目标是从大型数据集中提取特定图像，近年来受到了广泛关注。其基本思路是缩小目标图像的搜索范围，并检索与特定查询相匹配的图像。这项任务具有重要价值

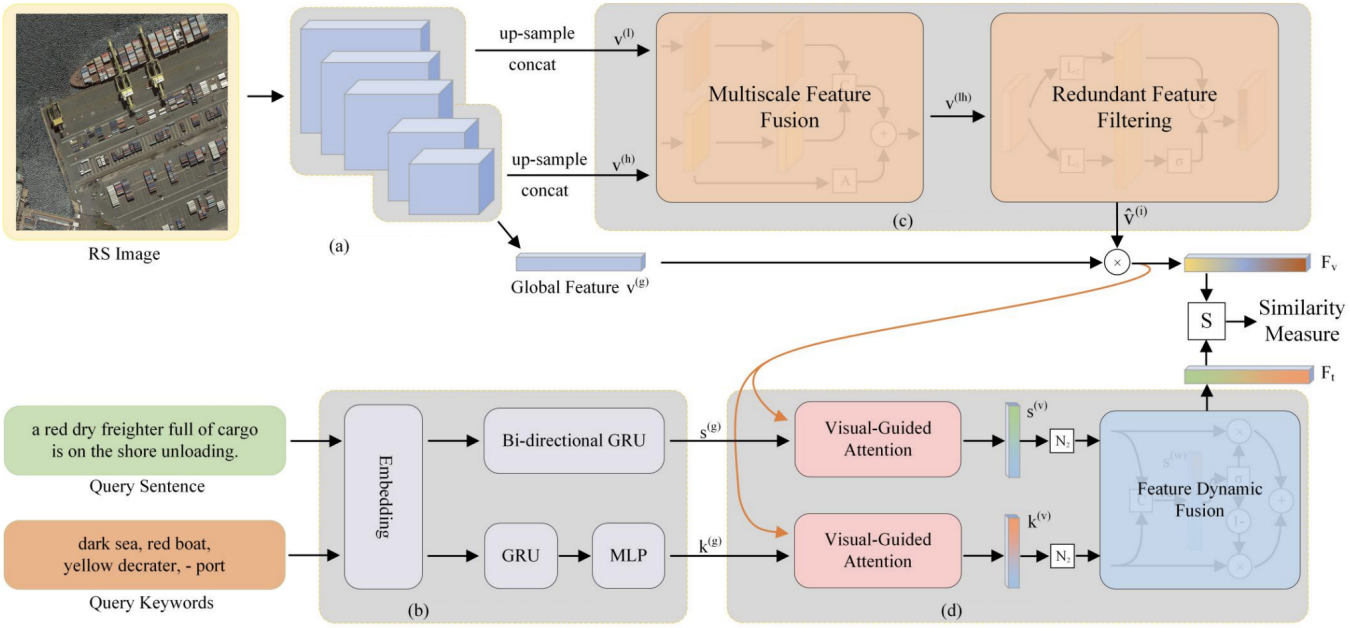

图11. [238] 用于基于文本的遥感图像检索的所提出方法示意图。

在森林砍伐检测、视觉导航和城市规划等实际应用中。

为实现这一目标，Abdullah 等人 [1] 构建了一个名为 TextRS 的新数据集，用于基于文本的图像检索任务。该数据集包含来自四个不同场景数据集的图像，每张图像均标注了五个对应的句子。为了检索与特定查询相匹配的图像，他们还提出了一种新颖的深度双向三元组网络，该网络由 LSTM [56] 和卷积神经网络组成，通过匹配文本与图像表示来实现遥感检索。实验结果表明，该模型在召回率@K方面表现出良好的性能，当 K = 1分别为 1、5 和 10 时，取得了 7.20%、51.39% 和 73.02% 的结果。类似地，Rahhal 等人 [149] 提出了一种无监督文本图像检索模型，用于遥感图像检索。该模型利用视觉 Big Transfer 模型学习图像表示，并采用双向 LSTM [56] 网络对文本描述进行编码。模型通过无监督嵌入损失进行优化，以确保图像的特征与其对应的文本描述最为接近，反之亦然，同时与其他图像的特征保持差异。该研究在 TextRS 数据集上实现了文本到图像检索，取得的召回率@1、召回率@5 和召回率@10 分别为 19.02%、55.25% 和 71.72%。此外，在 Merced 数据集上，召回率@1、召回率@5 和召回率@10 分别达到 21.86%、60.00% 和 75.58%。

Yuan 等人 [240] 提出了一种新颖的遥感文本-图像检索框架，该框架融合了局部和全局信息，并利用动态融合模块有效整合不同层级的特征。针对现有方法仅能处理以英语编写的查询这一问题，Rahhal等人 [4] 提出了一种多语言框架，该框架由语言编码器和视觉编码器组成，前者用于从文本描述中生成语言表示特征，后者用于提取视觉特征。作者使用四种不同语言（包括英语、阿拉伯语、法语和意大利语）的文本描述训练模型。此外，Chen等人 [31]提出了一种新颖的跨模态图文检索网络，该网络基于精心设计的语义对齐模块，通过注意力机制和门控机制建立遥感图像与其配对文本数据之间的直接关系，从而获得更具判别性的特征表示。

Yuan 等人 [239] 提出了一种简单而有效的跨模态模型，用于遥感图像检索，以应对遥感数据的多尺度和目标冗余特征所带来的挑战。该模型在遥感图像编码过程中融入多尺度信息，并动态过滤冗余特征，同时通过轻量级组卷积提取文本特征。此外，为了提升检索性能，他们开发了一种基于知识蒸馏的新型隐式监督优化方法。该方法使所提出的模型能够从教师网络的多层级和表示层中获取暗知识，从而显著提高轻量级模型的准确率。类似地，Yuan 等人 [238]采用多尺度自注意力模块来获取遥感图像的显著特征，并动态过滤冗余特征，如图 11所示。作者还构建了一个细粒度遥感图像-文本匹配数据集，使遥感图像检索能够分别基于关键词和句子进行。在悉尼、UCM、RSICD 和 RSITMD 这四个数据集上，上述方法取得的平均召回率分别约为 52%、46%、30% 和 16%。近期，Rahhal等人 [150] 提出了一种高效的文本-图像检索方法，利用视觉与语言 Transformer，通过两种对比损失将遥感图像的视觉表示与其对应的文本表示对齐，分别用于图像到文本和文本到图像分类

方法	RSICD 数据集							RSITMD 数据集						
	句子检索			图像检索			mR	句子检索			图像检索			mR
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
VSE++ [45]	3.38	9.51	17.46	2.82	11.32	18.10	10.43	10.38	27.65	39.60	7.79	24.87	38.67	24.83
SCAN-t2i [91]	4.39	10.90	17.64	3.91	16.20	26.49	13.25	10.18	28.53	38.49	10.10	28.98	43.53	26.64
SCAN-i2t [91]	**5.85**	12.89	19.84	3.71	16.40	26.73	14.23	11.06	25.88	39.38	9.82	29.38	42.12	26.28
CAMP-triplet [208]	5.12	12.89	21.12	4.15	15.23	27.81	14.39	**11.73**	26.99	38.05	8.27	27.79	44.34	26.20
CAMP-bce [208]	4.20	10.24	15.45	2.72	12.76	22.89	11.38	9.07	23.01	33.19	5.22	23.32	38.36	22.03
MTFN [203]	5.02	12.52	19.74	4.90	17.17	29.49	14.81	10.40	27.65	36.28	9.96	31.37	45.84	26.92
AMFMN-soft [238]	5.05	14.53	21.57	**5.05**	**19.74**	31.04	16.02	11.06	25.88	39.82	9.82	33.94	51.90	28.74
AMFMN-fusion [238]	5.39	**15.08**	**23.40**	4.90	18.28	**31.44**	**16.42**	11.06	**29.20**	38.72	9.96	34.03	52.96	29.32
AMFMN-sim	5.21	14.72	21.57	4.08	17.00	30.60	15.53	10.63	24.78	**41.81**	**11.51**	**34.69**	**54.87**	**29.72**
方法	UCM 数据集							悉尼数据集						
	句子检索			图像检索			mR	句子检索			图像检索			mR
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
VSE++ [45]	12.38	44.76	65.71	10.10	31.80	56.85	36.93	24.14	53.45	67.24	6.21	33.56	51.03	39.27
SCAN-t2i [91]	14.29	45.71	67.62	12.76	50.38	77.24	44.67	18.97	51.72	74.14	**17.59**	56.90	76.21	49.26
SCAN-i2t [91]	12.85	47.14	**69.52**	12.48	46.86	71.71	43.43	20.69	55.17	67.24	15.52	57.59	76.21	48.74
CAMP-triplet [208]	10.95	44.29	65.71	9.90	46.19	76.29	42.22	20.69	53.45	**75.86**	14.14	42.07	69.66	45.98
CAMP-bce [208]	14.76	46.19	67.62	11.71	47.24	76.00	43.92	15.52	51.72	72.41	11.38	51.72	76.21	46.49
MTFN [203]	10.47	47.62	64.29	**14.19**	52.38	78.95	44.65	20.69	51.72	68.97	13.79	55.51	77.59	48.05
AMFMN [238]	12.86	**51.90**	66.67	**14.19**	51.71	78.48	45.97	20.69	51.72	74.14	15.17	58.62	80.00	50.06
AMFMN-fusion [238]	**16.67**	45.71	68.57	12.86	**53.24**	**79.43**	**46.08**	24.14	51.72	**75.86**	14.83	56.55	77.89	50.17
AMFMN-sim [238]	14.76	49.52	68.10	13.43	51.81	76.48	45.68	**29.31**	**58.62**	67.24	13.45	**60.00**	**81.72**	**51.72**

表 IV RSICD、RSITMD、UCM 和悉尼数据集上的基于句子和基于图像的遥感检索结果。数据摘自 [238]。

。与此前的方法相比，该方法在各数据集上的检索性能略有提升。

此外，我们在表中总结了最先进的方法在 RSICD、RSITMD、UCM 和悉尼数据集上的表现IV。表中列出了 K 值为 1、5 和 10 时的 R@K 分数，以及平均召回率分数。这里，R@K 表示在前 K 个结果中找到真实标注的比例，而平均召回率表示 Huang 等人提出的 R@K 六种召回率的平均值。 [67]。

E. 视觉问答

视觉问答（VQA）是一项旨在为图像相关问题提供答案的任务。尽管这项任务在计算机视觉领域已广受欢迎，但在遥感领域仍处于起步阶段。遥感视觉问答系统使非专业用户能够以自然语言问题为查询，与遥感图像进行交互，从而以用户友好的方式帮助用户高层次地理解遥感图像。开创性工作 [122]构建了首个面向遥感图像的大规模视觉问答基准数据集，称为 RSVQA。研究人员从 OpenStreetMap 收集了低分辨率和高分辨率遥感图像，并为图像配上由人工生成的相关问题和答案。低分辨率数据集中共有 772 张图像和 77,232 个问答对；高分辨率数据集中共有 10,659 张图像和 1,066,316 个问答对。在 [122], 中，作者提出了一种基准方法，使用卷积神经网络（CNN）学习视觉特征，并采用 LSTM [56] 网络进行文本嵌入提取。此外，他们还设计了互注意力，以增强视觉特征与文本特征之间的对齐。所提出的方法取得了总体准确率在低分辨率和高分辨率数据集上，准确率均接近 80%。Zheng等人 [259] 提出了一个名为 RSIVQA 的遥感图像视觉问答数据集，其中包含约 3.7 万张图像。他们还开发了一种互注意力网络，以利用视觉特征与文本特征之间的语义对应关系，并采用双线性模块进行特征融合。所提出的方法在其 RSIVQA 数据集上取得了 77% 的总体准确率。Chappuis 等人 [19]提出使用大型语言变换器，例如 BERT [42],，来学习文本特征，并证明其性能优于循环神经网络。该方法在 RSVQA 数据集的低分辨率子集上取得了接近 84% 的总体准确率。最近，Yuan 等人 [237] 设计了 RSVQA 数据集的多语言版本，并利用增强数据集提高了 RSVQA 模型的鲁棒性。

不同于以往主要致力于解决封闭式场景下遥感视觉问答的方法，Al 等人 [3]提出了一个新数据集 VQA-TextRS，其中包含来自134个类别的2144幅遥感图像，问题和答案均通过人工标注生成。该数据集涵盖了多种多样的开放式问答对。为了解决开放式视觉问答任务，他们采用视觉与语言变换器网络，从图像及其配套问题中提取视觉特征和文本特征，随后利用变换器解码器，通过跨注意力机制融合这两种模态。因此，该方法在回答图像中是否存在目标的问题上达到了84.01%的准确率。随后，[13]有研究提出了类似思路，作者提出使用 CLIP [145] 模型将图像和问题嵌入为视觉与文本表征，随后通过注意力机制学习这些表征之间的相关性。图12 概述了所提出的遥感视觉问答方法。该方法达到了总体

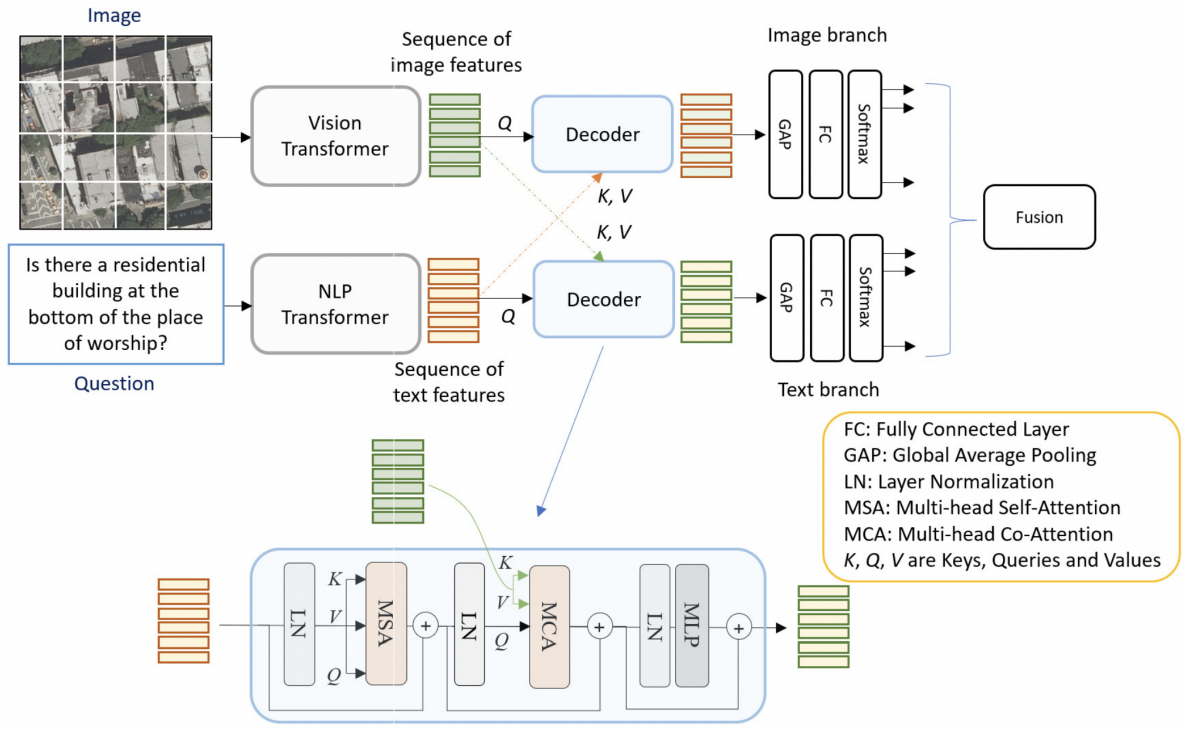

图 12. 文献 [13]提出的遥感视觉问答方法框架。该方法使用视觉语言变换器，将图像与问题对编码为文本和视觉特征表示。

准确率：在低分辨率 RSVQA 数据集上为85%，在高分辨率遥感视觉问答数据集的测试集1和测试集2上分别为85%和81%。不同于采用视觉编码器网络进行特征学习的传统方法，Chappuis 等人 [20]提出了一种新方法，将图像中的上下文信息转换为语言模型能够处理的文本提示。这样便可在统一的语言模型中同时处理问题和视觉上下文。Yuan 等人 [235] 提出了一种基于自步调课程学习（SPCL）的训练技术，以由易到难的方式训练 RSVQA 模型。该方法在低分辨率 RSVQA 数据集上取得了83%的总体准确率，在高分辨率遥感视觉问答数据集的测试集1和测试集2上分别取得了84.2%和79.3%的准确率，并在 RSIVQA 数据集上取得了79.7%的总体准确率。

近期，文献 [236] 提出了一种有趣的思路。作者借助视觉问答系统，对多时相航空影像进行变化检测。他们采用自动问答生成方法，构建了一个基于变化检测的视觉问答（CDVQA）数据集，其中包含多时相图像和问答对。该数据集包含 4,662 对带有语义变化图的图像。此外，他们还为 CDVQA 任务开发了一种基线方法，该方法包含四个部分：多时相特征编码、多时相融合、多模态融合和答案预测。该基线方法在 CDVQA 数据集的测试集1和测试集2上分别取得了 68.0% 和 63.4% 的总体准确率。

尽管现有 RSVQA 方法已展现出令人鼓舞的性能，但它们在各自的

RSVQA-HR 测试集 1 [122]			
方法	存在性	比较	平均准确率
RSVQA [122]	90.43	88.19	89.31
由易到难 [235]	91.39	89.75	90.57
双模态 [13]	92.03	91.83	91.93
SHRNet [253]	**92.45**	91.68	**92.07**
RSGPT [62]	91.86	**92.15**	92.00
RSVQA-HR 测试集 2 [122]			
RSVQA [122]	86.26	85.94	86.10
由易到难 [235]	87.97	87.68	87.83
双模态 [13]	89.37	89.62	89.50
SHRNet [253]	89.81	89.44	89.63
RSGPT [62]	**89.87**	**89.68**	**89.78**

表 V RSVQA-HR 测试集 1 和 RSVQA-HR 测试集 2 上的遥感视觉问答性能。数据引用自 [62]。

与目标存在性和分类任务相比，在目标计数方面的表现 [235]。不同于以往主要使用预定义问题输入的方法，近期的研究 [11]引入了视觉问题生成（VQG）方法，该方法能够针对遥感（RS）图像生成智能的自然语言问题。我们在表 [122] 和表 [122] 中分别总结了比较方法在 RSVQA-HR V 和 RSVQA-LR VI数据集上的性能。我们报告了每种问题类型的准确率，以及表中所有类型的平均准确率。从这些表格可以看出，RSGPT [62] 优于所有比较方法。

方法	存在性	比较	农村/城市	平均准确率
RSVQA [122]	87.46	81.50	90.00	86.32
由易到难 [235]	90.66	87.49	91.67	89.94
双模态 [13]	91.06	91.16	92.66	91.63
SHRNet [253]	91.03	90.48	94.00	91.84
RSGPT [62]	**91.17**	**91.70**	**94.00**	**92.29**

表 VI 遥感视觉问答在RSVQA-LR [122] 数据集上的性能。数据引用自 [62]。

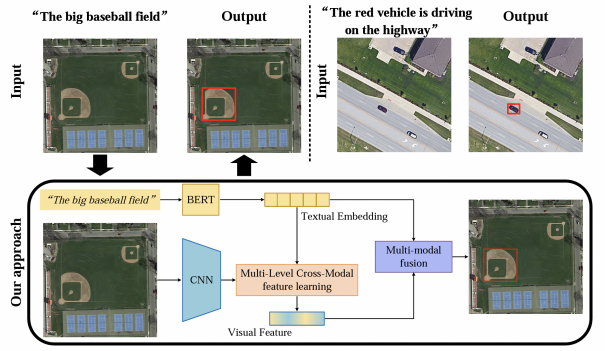

图 13. [242]提出的遥感视觉定位方法示意图。

F. 视觉定位

近年来，遥感数据（RS）背景下的视觉定位（RSVG）已成为一个新兴的研究领域。针对这一特定任务的研究目前仍处于起步阶段，相关成果相对较少。具体而言，RSVG利用一幅遥感图像及其关联的查询表达，为特定感兴趣目标提供边界框 [242]。通过使用自然语言引导在遥感场景中定位目标，RSVG能够实现目标级理解，并提高终端用户的可访问性。RSVG的潜在应用包括目标检测与识别、搜索与救援任务、城市规划等。

与自然图像中的查询表达相比，RSVG 中的表达往往涉及复杂的地理空间关系，而且感兴趣的目标通常并不显著。尽管自然图像视觉定位领域已开展了大量研究，但这一任务在遥感领域仍未得到充分探索。RSVG 首次发表于 [179], , 该研究不仅推出了一个新数据集，还提出了一种在遥感场景中实现视觉定位的新方法。RSVG 数据集包含 4,239 张 Google Earth 图像和 7,993 条指代表达，平均表达长度为 28.41。该研究提出的模型由三个部分组成，即语言编码器、图像编码器和融合模块。语言编码器用于构建地理空间关系图，其中节点表示地面目标，边则描述它们之间的地理空间关系。至于图像编码器，该方法利用自适应区域注意力模块，从大规模遥感场景中提取视觉特征。最后，融合模块用于将地理空间关系图融入视觉特征。所提出的方法取得了以下定位准确率：在交并比阈值分别为 0.5 和 0.25 时，准确率分别为 59.40% 和 64.95%。Zhan 等人最近也提出了一项研究。[242],在该研究中，他们构建了一个名为 DIOR-RSVG 的新数据集，并开发了一种新颖的基于 Transformer 的方法，如图所示。13。DIOR-RSVG 数据集包含 17,402 张遥感图像和 38,320 条指代表达，涵盖 20 个目标类别，平均表达长度为 7.47。所提出的方法利用多尺度视觉特征和多粒度语言特征学习判别性表征，以解决尺度变化问题。为应对杂乱背景，他们动态滤除无关噪声并增强显著特征。此外，该研究还基准评估了若干针对自然图像设计的最先进视觉定位方法在遥感数据上的表现。所提出的 MGVML 方法在 DIOR-RSVG 数据集测试集上的平均交并比最高，达到 68.04%。

表 VII 总结了在该数据集上各比较方法的性能。我们报告了不同阈值下的精确率、表中的平均 IoU 和累积交并比。有关这些评估指标，请参阅[242] 。尽管视觉语言模型在遥感视觉定位方面展现出令人瞩目的能力，但当前研究表明，与传统目标检测方法相比，现有 RSVG 方法的性能仍有提升空间。这一观察凸显了进一步开展研究、提升 RSVG 方法有效性和准确性的必要性。

G. 零样本场景分类。

零样本遥感场景分类旨在结合视觉特征以及语义类别之间的语义关系，识别未见场景概念。Li 等人 [92]提出了首个用于遥感场景分类的零样本学习方法。该方法采用在维基百科语料库上预训练的 Word2vec [198] 模型，为类别名称生成语义嵌入，随后构建语义图以捕捉类间关系。所提出的 ZSSC 方法在 UCM21 数据集（5 个未见类别）和 NWPU45 数据集（10 个未见类别）上的总体分类准确率分别达到 49.01% 和 47.00%。Quan 等人 [144]通过引入半监督 Sammon 嵌入算法 [164]来对齐语义原型和视觉原型，从而改进了该方法。所提出的 ZSSC 方法在 UCM21 和 NWPU45 数据集上的总体分类准确率分别达到 50.42% 和 48.40%。此外，Sumbul 等人 [173] 提出了一种用于细粒度遥感图像分类的零样本学习技术，该技术通过训练兼容性函数，建立图像特征与语义嵌入之间的关联，从而实现从已见类别到未见类别的知识迁移。在 Kodirov 等人 [79], 的研究基础上，Wang等人 [197] 提出了一种距离约束语义自编码器，用于对齐零样本遥感场景分类中的视觉特征和语义表征。在 [109], 中，Li 等人采用 BERT 等基于 Transformer 的大语言模型，从专家定义的各类别文本描述中提取语义嵌入。

Li 等人 [110]提出了一种基于生成对抗网络（GAN）的方法，而非使用判别模型进行遥感场景分类（RSSC），该方法——

方法	视觉编码器	语言编码器	Pr@0.5	Pr@0.6	Pr@0.7	精确率@0.8	精确率@0.9	平均交并比	累积交并比
单阶段：									
ZSGNet [162]	VGG	双向长短期记忆网络	48.12	43.79	36.82	25.04	6.62	40.11	46.11
ZSGNet [162]	ResNet-50	双向长短期记忆网络	51.67	48.13	42.30	30.11	10.15	44.12	51.15
FAOA-无空间 [232]	DarkNet-53	BERT	63.63	61.20	56.92	50.15	38.83	57.53	63.14
FAOA [232]	DarkNet-53	BERT	67.21	64.18	59.23	50.87	34.44	59.76	64.13
FAOA [232]	DarkNet-53	LSTM	70.86	67.37	62.04	53.19	36.44	60.74	66.20
ReSC [231]	DarkNet-53	BERT	72.71	68.92	63.01	53.70	33.37	64.24	68.10
LBYL-Net [63]	DarkNet-53	BERT	73.78	69.22	65.56	47.89	15.69	65.92	73.67
基于Transformer：									
TransVG [41]	ResNet-50	BERT	72.41	67.38	60.05	49.10	20.91	58.57	76.27
VLTG [228]	ResNet-50	BERT	69.41	65.16	58.44	45.56	24.17	57.27	71.78
VLTG [228]	ResNet-101	BERT	75.79	72.22	66.33	56.17	33.31	**69.11**	77.85
MGVLF [242]	ResNet-50	BERT	**76.78**	**72.68**	**66.74**	**56.42**	**35.07**	68.04	**78.41**

表 VII DIOR-RSVG 数据集上的遥感视觉定位性能。数据引自 [242]。

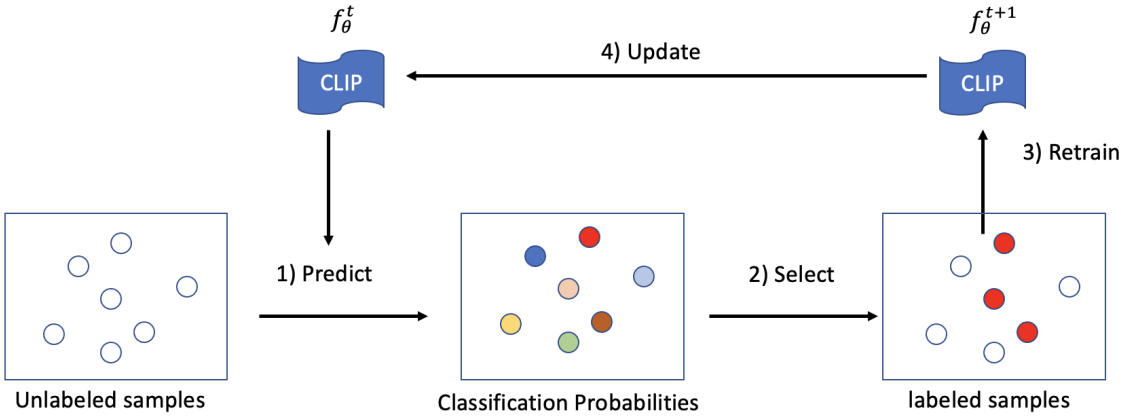

图 14. [103] 中提出的零样本遥感场景分类方法综述。

——用于零样本遥感场景分类，其中训练生成器网络，根据类别语义合成图像特征。该方法在 UCM21、AID30 和 NWPU45 数据集上的零样本分类准确率分别达到 62.66%、55.86% 和 50.66%。此外，作者研究了不同的语言处理模型，即 Word2vec [198]，Fasttest [74]，[15]，GloVe [215]，和 BERT [42]，，用于语义嵌入提取。与自然图像不同，遥感场景在结构和背景方面存在显著差异，这给模型学习用于场景理解的鲁棒视觉特征带来了巨大挑战。此外，遥感图像中不同场景类别之间的语义相关性比较弱于自然图像，这使得基于语义推理的零样本学习任务更具挑战性。

受大型视觉语言模型在（少样本/零样本）自然图像理解方面巨大成功的启发，Li 等人提出了一种名为 RS-CLIP 的大型视觉语言模型 [103]，，用于遥感场景分类。所提出的方法利用预训练的 CLIP 模型生成伪标签，并制定了课程学习策略，以逐步提升场景分类性能。图14 展示了所提出的零样本遥感场景分类方法的综述。表VIII 总结了其性能主流遥感场景分类方法。我们在表格中报告了新类别上的总体准确率。

方法	UCM (5/21)	NWPU (10/45)	AID (5/30)
稀疏自编码器 [79]	49.50	44.81	47.34
ZSP-LP [92]	49.01	47.00	46.77
ZSC-SA [144]	50.42	48.40	50.87
WDVSc [193]	55.91	50.68	52.61
RBGN [219]	57.93	44.60	51.99
DASE [197]	58.63	51.52	53.49
CSPWGAN [110]	62.66	51.52	55.86
RS-CLIP [103]	**95.54**	**96.95**	**93.34**

表 VIII
零样本-四个基准数据集上的遥感场景分类结果。括号中的数字表示新类别数和全部类别数。所有数值均引自 [110] 和 [103]。粗体数值表示最佳性能。

H. 少样本目标检测。

目标检测是遥感（RS）领域的一项重要任务，旨在通过确定目标实例的边界框和类别标签来检测目标实例。经过大量研究，近年来该领域取得了显著进展，包括

近年来的两阶段检测器，例如 Fast R-CNN [48] 和 Faster R-CNN [156], ，以及单阶段检测器，例如 SSD [118], YOLO [155], 和 RetinaNet [113],，以及近期提出的 DETR 变体 [17], [267]。遥感图像中的少样本目标检测（FSOD）是一个相对较新的研究领域，近年来受到广泛关注。该任务旨在仅使用少量标注样例，检测遥感图像中感兴趣的目标。Li 等人 [102] 的开创性工作，标志着专门针对遥感图像的少样本目标检测研究初步展开。现有的少样本目标检测方法大致可分为两类：1）基于元学习的方法 [102], [30], [263]和基于微调的方法 [258], [66], [249], [101]。基于元学习的方法包含一个元学习器，用于从大量采样任务中学习任务无关知识；每个任务（也称为任务片段）都包含一个支持集和一个查询集，并利用任务特定学习器在特定任务上执行检测。基于微调的方法首先在基类上训练检测模型，然后在新类别上微调网络权重。关于遥感图像少样本目标检测的综合文献综述见 [117]。

只有少数研究尝试将视觉语言模型用于遥感图像中的少样本目标检测。Zhang 等人 [248] 提出构建一个语料库，为每个区域提供语言描述，例如目标属性和关系，以编码相应的常识嵌入。所提出的 TSF-RGR 方法在 NWPU VHR-10 和 DIOR 数据集的 10 个新类别上，平均精度均值分别达到 0.77 和 0.49。其结果还表明，基于文本的语义关系能够显著提升少样本检测性能。Lu 等人 [125]提出将所有目标类别的文本描述用作额外特征，以缓解少样本目标检测中的分类混淆。图15 展示了所提出的遥感图像少样本目标检测方法的概览。所提出的 TEMO 方法在 NWPU VHR-10 和 DIOR 数据集的 10 个新类别上，平均精度均值分别达到 0.75 和 0.43。这些方法展现了融合视觉特征与语言特征以实现更优表征学习的潜力，从而提升遥感图像中的少样本目标检测性能。表IX总结了当前遥感图像少样本目标检测方法的性能。表中报告了各新类别的平均精度（AP）及平均 AP。PL、BD 和 TC 分别表示飞机、棒球场和网球场。

在计算机视觉领域，视觉语言模型在零样本目标检测方面展现出了巨大的潜力。多模态视觉语言预训练模型的应用使开放词汇目标检测（OVOD）成为一个活跃的研究领域，因为它能够考虑更加贴近现实的场景。与在固定且预定义的类别上进行训练和评估的传统目标检测方法不同，开放词汇目标检测涉及在标注数据集上进行训练，并将训练好的模型泛化到此前未见过的新类别。为了实现开放词汇目标检测，一种常见策略是通过匹配目标特征与类别文本嵌入来修改现有的目标检测头 [145], [50], [260], [151], [9]。在遥感领域，这种开放词汇设置为研究人员开辟了新的方向，使他们能够开发视觉语言模型，以在更具泛化能力和实用性的场景中检测未见类别中的目标。

I. 少样本/零样本语义分割

在语义分割领域，少样本学习方法能够利用数量有限的标注图像对新类别进行分割。近期的研究主要聚焦于两类方法，即基于参数匹配的方法和基于原型的方法。值得注意的是，开创性工作 PANet [201] 通过引入原型对齐模块，在少样本分割领域取得了突破。该模块为每个语义类别生成高度代表性的原型，并基于特征匹配对查询目标进行分割。Jiang等 [72] 提出了首个用于遥感图像分割的少样本学习方法，并展示了仅使用新类别的少量样本进行遥感图像语义分割的可行性。所提出的方法仅使用 5 个标注图像块，便在 ISPRS Vaihingen 和 Potsdam 数据集 [159] 上分别取得了 67.7% 和 53.2% 的总体分类准确率。

为克服基于深度学习的分割方法对数据的依赖，近期研究 [26], [252] 探索了自监督/半监督学习和弱监督学习，以减少对密集标注的需求。Chen等人 [26] 提出了一种基于对比学习的遥感图像小样本分割半监督方法。Zhang 等人 [252]提出了一种用于跨场景高光谱图像分类的网络，该网络利用语言引导实现域泛化。图16 概述了所提出的方法。Lang等人 [87]提出了一种基于全局校正和解耦配准的遥感小样本分割框架，并证明其在 iSAID-5i 数据集上的性能优于以往的 SOTA 方法 [209]。尽管这些方法展现了将视觉语言模型应用于遥感图像小样本语义分割的潜力，但视觉语言模型在遥感图像分割中的应用仍处于起步阶段。表X 总结了遥感图像小样本语义分割方法的性能。我们在表中报告类别平均 IoU。

在计算机视觉中，预训练的视觉语言模型已被广泛探索用于开放词汇语义分割，这得益于其在开放词汇图像分类中的显著成功。在该任务中，模型无需额外使用该类别的标注图像，即可对任意类别进行分类。受 CLIP 的巨大成功启发，DenseCLIP [154]通过将每个像素与基于文本的类别描述进行匹配，解决了密集预测问题。MaskCLIP [261] 利用由 CLIP 生成的伪逐像素标签和自训练，实现了无需标注的分割。类似地， [241] 使用像素级伪标签进行密集监督。CLIPSeg [128] 基于 CLIP 模型构建，并根据任意提示生成图像分割结果。ZSSeg [222] 首先生成掩码候选区域，然后分两个阶段利用 CLIP 对生成的候选区域进行分类。LSeg [93] 利用图像编码器将像素嵌入与文本嵌入进行匹配，并利用文本编码器提供灵活的类别表示。OVSeg [111] 提出在 COCO Captions 数据集的掩码图像区域及其对应文本上对 CLIP 模型进行微调，从而获得更好的开放词汇分割性能。Fusioner [131]

Shot	方法	DIOR数据集						西北工业大学 VHR-10数据集			
		飞机	棒球场	网球场	火车站	风车	mAP	PL	BD	TC	mAP
3	Meta RCNN† [227]	9.1	17.8	33.5	3.8	9.1	14.7	6.8	7.1	5.7	6.5
	FsDetView† [216]	9.1	17.9	18.1	9.1	9.1	12.6	9.1	22.9	2.4	11.5
	TFA† [206]	11.9	22.6	**51.8**	9.7	8.1	20.8	9.1	16.2	14.5	13.2
	Meta YOLO [76]	-	-	-	-	-	-	13.0	12.0	11.0	12.0
	FSODM [102]	-	-	-	-	-	-	15.0	57.0	25.0	32.0
	PAMS-Det [258]	14.0	**54.0**	24.0	17.0	**31.0**	28.0	21.0	76.0	16.0	37.0
	FSCE† [176]	17.3	38.7	40.8	11.6	16.0	24.9	33.3	73.4	30.2	45.6
	DMLKI [101]	-	-	-	-	-	-	41.0	66.0	29.0	45.0
	TEMO† [125]	**18.4**	47.1	52.4	**17.7**	18.8	**30.9**	49.1	77.8	**45.8**	**57.6**
	TSF-RGR [248]	-	-	-	-	-	-	**57.0**	**79.0**	35.0	57.0
5	Meta RCNN† [227]	9.1	25.9	32.7	7.0	2.3	15.4	12.0	21.9	0.9	11.6
	FsDetView† [216]	9.1	22.3	26.3	5.0	1.1	12.8	16.5	28.4	7.8	17.6
	TFA† [206]	13.2	27.7	51.8	9.0	8.0	21.9	9.1	15.9	24.6	16.5
	Meta YOLO [76]	9.0	33.0	47.0	9.0	13.0	22.0	24.0	39.0	11.0	24.0
	RepMet [77]	9.0	19.0	11.0	1.0	1.0	8.0	20.0	36.0	14.0	23.0
	FSODM [102]	9.0	27.0	57.0	11.0	19.0	25.0	58.0	84.0	16.0	53.0
	PAMS-Det [258]	17.0	55.0	41.0	17.0	**34.0**	33.0	55.0	**88.0**	20.0	55.0
	FSCE† [176]	22.8	36.6	49.5	17.1	15.2	28.3	60.7	83.6	32.5	59.0
	DMLKI [101]	15.0	46.0	48.0	17.0	17.0	28.0	53.0	81.0	**49.0**	61.0
	TEMO† [125]	24.2	58.9	52.1	**25.0**	25.7	37.2	**71.4**	85.4	43.5	**66.8**
	TSF-RGR [248]	**58.0**	**79.0**	**58.0**	11.0	6.0	**42.0**	71.0	85.0	42.0	66.0
10	Meta RCNN† [227]	12.4	37.5	32.4	14.0	3.2	19.9	11.4	24.4	10.5	15.4
	FsDetView† [216]	9.1	31.7	25.8	11.1	6.1	16.7	9.1	20.0	12.4	13.8
	TFA† [206]	17.7	33.1	56.4	17.3	8.7	26.7	4.5	17.0	21.9	14.5
	Meta YOLO [76]	15.0	45.0	54.0	7.0	18.0	28.0	20.0	74.0	26.0	40.0
	RepMet [77]	13.0	33.0	24.0	1.0	1.0	14.0	22.0	39.0	18.0	26.0
	FSODM [102]	16.0	46.0	60.0	14.0	24.0	32.0	60.0	88.0	48.0	65.0
	PAMS-Det [258]	25.0	58.0	50.0	23.0	**36.0**	38.0	61.0	88.0	50.0	66.0
	FSCE† [176]	24.7	56.2	49.1	21.6	31.2	34.8	85.2	75.3	48.0	69.5
	DMLKI [101]	32.0	53.0	48.0	16.0	20.0	34.0	77.0	90.0	68.0	78.0
	TEMO† [125]	32.6	59.3	67.0	**23.9**	31.2	42.8	**86.4**	88.7	50.1	75.1
	TSF-RGR [248]	**72.0**	**79.0**	**67.0**	8.0	19.0	**49.0**	78.0	**94.0**	**60.0**	**77.0**

表 IX DIOR数据集和 NWPU VHR-10 数据集新类别上的少样本检测性能。† 表示 20 次随机实验的平均结果。粗体表示各列中的最佳性能。数据引用自 [125]。

结合不同的视觉与语言模型，利用跨模态融合模块实现开放词汇语义分割。SegCLIP [129]通过文本-图像对训练，将具有可学习中心的图像块聚合为语义区域，从而动态捕获语义组并生成分割结果。鉴于自然图像开放词汇语义分割的蓬勃发展，我们认为这可能是遥感领域一个很有前景的研究方向。

IV. 有用资源

本节总结了常用的遥感视觉语言数据集、代码库及其他资源。

A. 遥感视觉语言数据集

我们在表格中总结了遥感视觉语言数据集XI。尽管现有研究大多使用场景分类数据集来构建遥感基础模型，但所有列出的数据集都可能用于构建通用提案型遥感基础模型。请注意，我们仅列出了常用数据集，可能未能涵盖各项任务的所有相关数据集。关于遥感数据集的全面综述，请参见 [220]。

B. 开源代码库

我们进一步列出几个常用的开源代码库，为尝试开展遥感视觉语言模型研究的研究人员提供一个简易的起点。

- Huggingface*。Huggingface 是一个协作平台，供机器学习从业者共同开展模型、数据集和应用相关工作。该平台提供解决方案，以加快机器学习项目的进展，拥有超过 50,000 个组织用户。它通过 Transformers 库促进视觉语言模型的开发，简化了各种视觉语言模型的下载、运行和微调流程。

- MiniGPT-4。MiniGPT-4 [266] 是最早一批展现出卓越视觉语言能力的开源作品之一，例如根据手写文本生成网站，以及识别图像中的幽默元素。其增强版本 MiniGPT-v2 [24] 通过赋予模型空间理解和视觉定位能力，进一步提升了模型表现。该代码库可作为简单的入门工具包，帮助研究人员开发强大的视觉语言模型。

†https://sites.google.com/site/michelevolpiresearch/data/苏黎世-dataset *https://huggingface.co/

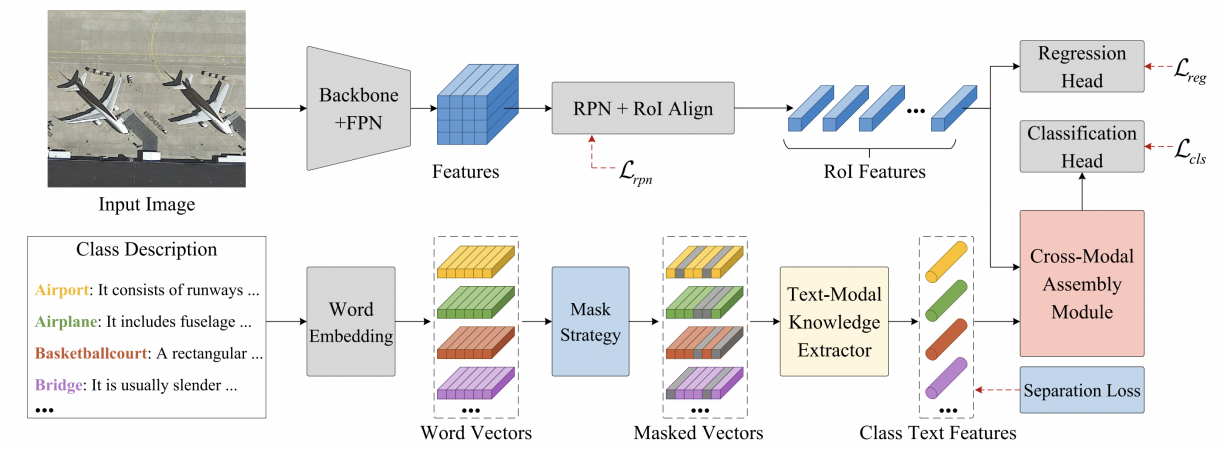

图 15. 架构 少样本目标检测方法的结构 [125]。类别描述用于目标分类 .

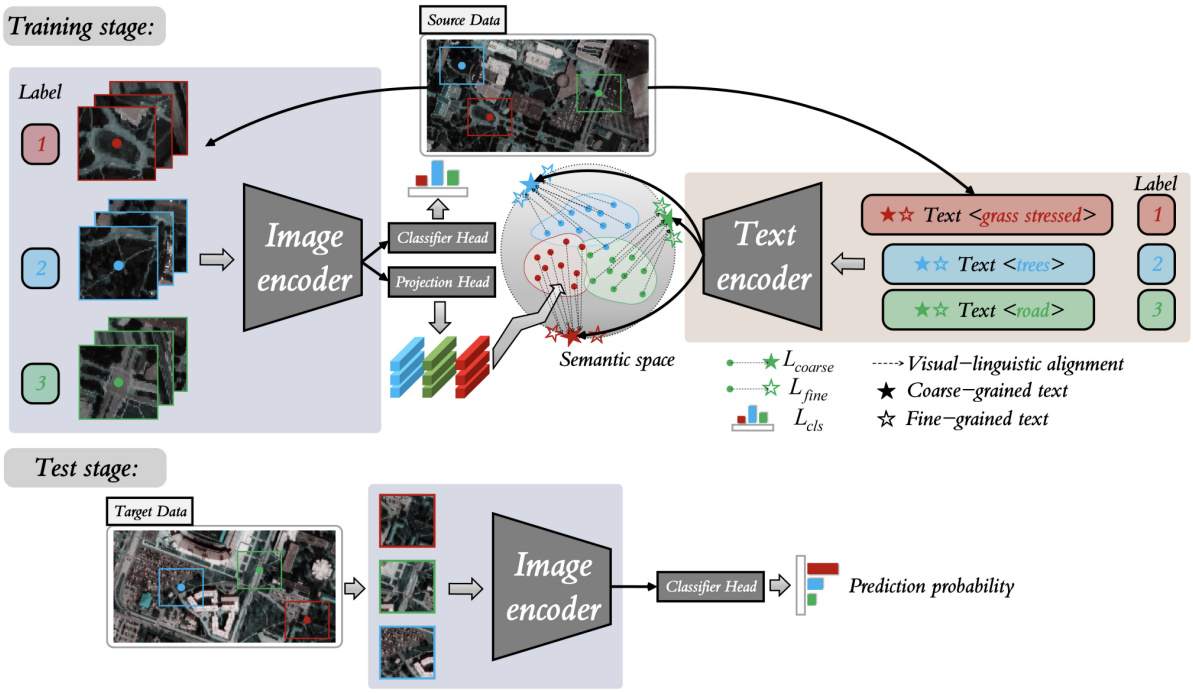

图 16. [252] 提出的跨场景高光谱图像分类方法综述。视觉特征和文本特征在语义空间中对齐。

•LLaVA。大型语言与视觉助手（LLaVA） [116]引入了一种大型多模态模型。该模型将视觉编码器与 Vicuna 融合，旨在实现通用的视觉与语言理解。它具备出色的聊天功能，让人联想到 GPT-4 的多模态特性，并在科学问答方面取得了无与伦比的准确率。多模态指令微调数据、模型和代码库均已公开。LLaVA-1.5 进一步改进了该模型，并在 11 项视觉语言基准测试中取得了最佳成绩。

• QWen-VL [7]。Qwen-VL由阿里云基于Qwen系列开发，是一种多模态模型，可处理图像、文本和边界框输入，并生成文本和边界框输出。与类似的开源大型视觉语言模型相比，它在多个英文基准上表现出色，例如零样本图像描述、视觉问答、文档视觉问答和视觉定位。该模型支持英语、中文和多语言交互，提升了图像中端到端双语文本识别的效果。

• Shikra [25]。Shikra 是一种多模态大语言模型，旨在通过自然语言展现其处理空间坐标输入与输出的能力，从而开启指代对话；整个过程无需额外词汇、位置编码器、检测前后处理措施或外部插件模型。

方法	Year	1-shot			5-shot		
		折叠-0	折叠-1	折叠-2	折叠-0	折叠-1	折叠-2
Panet [201]	2019	27.56	17.23	24.60	36.54	16.05	26.22
CANet [244]	2019	25.51	13.50	24.45	29.32	21.85	26.03
自采集 L [243]	2021	34.78	22.77	31.20	41.29	25.23	31.91
PFENet [185]	2022	35.84	23.35	27.20	44.42	25.34	33.59
NERTNet [119]	2022	34.93	**23.95**	28.56	**48.38**	26.73	37.19
DCP [88]	2022	37.83	22.86	28.92	41.52	**28.18**	33.43
BAM [86]	2022	39.43	21.69	28.64	29.32	27.92	38.62
DMML [195]	2022	28.45	21.02	23.46	30.61	23.85	24.08
SDM [233]	2022	27.96	21.99	27.82	28.50	25.23	31.07
DML [72]	2022	32.96	18.98	26.27	33.58	22.05	28.77
TBPNet [142]	2022	29.33	16.84	25.47	30.38	20.42	29.07
R^2Net [87]	2023	**41.22**	21.64	**35.28**	46.45	25.80	**39.84**

表 X 在 iSAID-5i 数据集上进行 1-shot 和 5-shot 设置时的小样本语义分割性能（以平均交并比计）。数值引自 [87]。所有对比方法均使用 ResNet-50 骨干网络。

C. 其他资源

此外，还有许多在线可访问的资源和工具，可用于构建遥感领域的视觉语言模型。

- OpenAI API†。尽管 OpenAI 未公开 GPT 系列的代码和模型，但它提供了便于用户调用其强大产品（如 GPT-4）的 API。OpenAI API 提供文本生成、图像生成、文本/图像嵌入及其他多模态工具（如音频转文本）的接口。它是研究人员准备自有数据集‡的实用工具。

- LAVIS。LAVIS [94] 是一个专为语言与视觉综合研究而设计的全面 Python 框架。它为图像描述、图像检索和视觉问答等多种任务提供统一平台。该库涵盖超过20个标准数据集，并支持超过30个预训练视觉语言模型。其目标是简化研究人员和开发人员多模态模型的开发流程。

- Midjourney§。Midjourney 目前是用于文本到图像生成的最令人印象深刻的工具之一。Midjourney 通过提供强大的多样化文本—图像对数据生成工具，有力地推动了视觉语言研究的发展。

V. 结论与未来趋势

通过融合视觉和语言模态，视觉语言模型有望彻底改变遥感数据的处理与解读方式。这些模型为研究人员和从业者提供了先进工具，使其能够从遥感影像中提取有意义的信息，从而更准确、更深入地分析地球表面及其动态变化。凭借弥合视觉感知与语言理解之间鸿沟的能力，视觉语言模型为开发创新的遥感应用和决策支持系统开辟了广阔前景。随着这一领域不断发展，视觉语言模型融入遥感工作流程预计将为遥感科学及其实践应用的进步作出重大贡献。

近年来，遥感（RS）领域的许多研究表明，在包括图像描述、基于文本的图像生成、基于文本的图像检索、视觉问答、场景分类、语义分割和目标检测在内的各种遥感任务中，视觉语言模型优于纯视觉模型。尽管这些早期尝试已经展示了将视觉语言模型应用于遥感的成功，但对于大多数研究人员而言，这仍然是一个新兴领域。因此，本文全面综述了视觉语言模型在遥感领域的应用，帮助其他研究人员深入了解这一快速发展的领域的背景和最新进展，同时也旨在鼓励对这一令人振奋且前景广阔的领域开展进一步研究。

回顾遥感领域视觉语言模型的相关文献后，我们发现当前研究存在若干局限。首先，用于训练视觉语言模型的遥感数据集数量有限，样本规模远小于计算机视觉领域数十亿级的图像数据集。其次，遥感领域现有的大多数视觉语言模型仍使用经典的卷积神经网络和循环神经网络作为图像和语言编码器，只有少数研究探索了预训练视觉 Transformer 和 GPT、BERT、Flan-T5 等大语言模型。这限制了这些模型的特征学习能力。此外，从头开始训练这些视觉语言模型需要大量计算资源，尤其是对于拥有数十亿参数的大型网络。因此，亟需有效的模型微调技术来训练遥感领域的大型视觉语言模型。此外，遥感数据可能会因光照条件、大气干扰和传感器噪声等因素而呈现高度变化。这种变化使视觉语言模型难以准确捕捉视觉信息与文本信息之间的关系。然而，现有研究鲜少考虑这一点。再者，视觉语言模型可能难以处理遥感数据较大的空间和时间尺度；这类数据覆盖范围广、时间跨度长，因而难以捕捉视觉信息与文本信息在时空维度上的关系。

鉴于现有遥感研究中视觉语言模型的局限性，我们列出了以下几个有前景的研究方向。

- **大规模图像-文本对数据集**。众所周知

†https://platform.openai.com/ ‡https://sharegpt.com/
§https://www.midjourney.com/

数据集	描述	任务
UCM [229]	加州大学默塞德分校土地利用数据集包含 2,100 张来自 21 个场景类别的航空影像，每个类别包含 100 张大小为 256×256 的图像。	场景分类，基础模型
悉尼数据集 [245]	来自 7 个不同场景类别的 613 张高分辨率遥感图像。	场景分类，基础模型
WHU-RS19 [37]	WHU-RS19 数据集包含来自 19 个场景类别的 1,013 张航空图像。总共有 1,013 张图像，每张包含 600×600 个像素。	场景分类，基础模型
NWPU-RESISC45 [28]	NWPU-RESISC45 数据集由来自 45 个场景类别的 31,500 张航空影像组成，且每个该类别有 700 张图像，尺寸为 256×256 像素。	场景分类，基础模型
AID [214]	AID 数据集包含来自 30 个场景类别的 10,000 张图像，每张图像的尺寸为 600×600 像素。	场景分类，基础模型
SEN12MS [166]	SEN12MS 是一个大规模多传感器遥感数据集，包含 180,662 组三元组双极化合成大小规律。合成孔径雷达（SAR）图像块、多光谱 Sentinel-2 图像块和中分辨率成像光谱仪土地覆盖 maps.	场景分类，基础模型
BigEarthNet [174]	BigEarthNet 是一个多模态、多标签遥感影像数据集，包含 590,326 对 Sentinel-1 和 Sentinel-2 图像切片。	场景分类，图像检索，基础模型。
MillionAID [123]	一个包含来自 51 个语义场景类别的百万张图像的大规模遥感影像数据集。	场景分类，基础模型
SatlasPretrain [12]	它整合了 Sentinel-2 和 NAIP 影像，包含来自 137 个类别和 7 种类型的 3.02 亿条标注。标签。	场景分类，基础模型
UCM-Captions [143]	基于 UCM 数据集的 2,100 对图像-描述对。每张图像均使用五个不同的句子进行描述。	图像描述、文本到图像检索。
悉尼描述 [143]	基于悉尼数据集的 613 个图像-描述对。每幅图像均使用五个不同的句子进行描述。	图像描述、文本-图像检索
RSICD [127]	RSICD 数据集包含 10,921 幅图像，来源于 Google Earth、百度地图等平台，MapABC 和天地图。每幅图像均已标准化为 224×224 像素的分辨率，并配有五条描述性语句。	图像描述，文本-到图像生成 / 检索
RS5M [255]	该数据集包含 500 万张带有机器生成的图像描述的遥感图像。该数据集构建于通过筛选公开可用的图文配对数据集，并专门为遥感数据集生成描述，使用一种预训练视觉语言模型。	图像描述
RSICap [62]	RSICap 数据集包含 2,585 条高质量的人工标注图像描述，信息丰富。	图像描述
TextRS [1]	TextRS 数据集汇集了来自四个不同场景数据集的 2,144 张图像：AID、Merced、PatternNet，以及西北工业大学数据集。每张图像均附有由五位不同标注员提供的描述性句子。	文本生成图像检索
RSVQA [122]	RSVQA-HR 包含 10,569 张高分辨率图像和 1,066,316 个问答对；RSVQA-LR 包含 772 张低分辨率图像和 77,232 个问答对。	视觉问答
RSIVQA [259]	RSIVQA 采用两种标注方法：人工标注和自动生成，由此构建了一个包含 110,000 多个视觉问答三元组的数据集。	视觉问答-回答
变化检测视觉问答 [236]	CDVQA 是一个变化检测视觉问答数据集，包含 2968 对航空影像。图像以及超过 122,000 组对应的问答对。	视觉问答-答
VQA-TextRS [3]	VQA-TextRS 是一个由来自四个源数据集的 2,144 张遥感图像组成的数据集，每张图像包含 2-3 个英文问答对，共计 6,245 对，并分为四个问题类别。	视觉问答答
RSIEval [62]	RSIEval 由 100 组高质量图像-描述对组成，每张图像对应一条描述，以及 936 个多样化的图像-问题-答案三元组，平均每张图像有 9 个问题。	视觉问答答
RSVG [179]	该数据集包含 4,239 张遥感图像，其中针对 5,994 个目标的文本标注共 7,933 条。每个每张图像平均包含 1.87 个指代表达，平均长度为 28.41 个词。	视觉定位
DIOR-RSVG [242]	DIOR-RSVG 数据集包含 38,320 条语言描述，涵盖 17,402 个遥感 DIOR数据集中的图像，涵盖20个不同的目标类别。表达式的平均值长度为 7.47 个词，词汇集包含 100 个术语。	视觉定位
西北工业大学甚高分辨率-10 [29]	该数据集包含 800 张航空影像，涵盖 10 个类别，空间分辨率为 0.5 至 2 米。650 这些图像中有一些标注目标，例如飞机、船舶、储罐和各种运动场。	目标检测
DIOR [98]	该数据集包含 23,463 张分辨率为 0.5 至 30 米的航空影像，其中有 192,472 个目标实例。来自20个类别，例如飞机、球场、桥梁、港口、车辆等。	目标检测
FAIR1M数据集 [178]	FAIR1M数据集包含超过40,000张图像中的100多万个实例，具有空间分辨率介于 0.3 米和 0.8 米之间。该数据集中的每个目标均有标注，分为 5 个主要类别，并进一步细分为 37 个子类别。	目标检测
ISPRS Vaihingen/Potsdam [160]	ISPRS Vaihingen（或 Potsdam）数据集分别包含 33 幅（或 38 幅）超高分辨率航拍图像，地面采样距离为 9 厘米（分别为 5 厘米）。图像数据集中的每个像素都被标注为六种类别之一。类别：建筑、不透水表面、低矮植被、树木、车辆和杂波。	语义分割-tion
苏黎世夏季数据集†	该数据集包含 20 个图像裁剪块，取自 QuickBird 于 2002年8月 对苏黎世的成像。The 图像已增强至全色分辨率，对应的地面采样距离约为 0.62 厘米。八个不同的城市及周围的城市类别已由人工标注。	语义分割-tion
休斯顿 [39], [89]	该数据集包含两幅图像：一幅高光谱图像和一幅基于激光雷达的数字表面模型，二者的空间分辨率均为 2.5 米。空间分辨率，拍摄于休斯敦大学及其周边地区。图像中的像素被标注为 16 个语义类别。	语义分割-tion
帕维亚 [47]	帕维亚中心和帕维亚大学数据采集自一次使用 ROSIS传感器对意大利北部进行的飞行任务。该帕维亚中心图像包含 102 个光谱波段和 1096×1096 像素，而帕维亚大学图像包含 103 个个光谱波段和 610×610 像素。每幅图像均划分为 9 个语义类别。	语义分割tion
GID [186]	GID 数据集包含两个部分：大尺度分类集和详细的土地覆盖分类集。大尺度分类集包含 150 幅像素级标注的高分二号影像，而详细的分类集由 30,000 个多尺度图像片段和 10 幅像素级标注的 GF-2 影像组成。	语义分割tion

表 XI 遥感视觉语言任务常用数据集汇总。

人工智能系统的准确性在很大程度上依赖于训练数据集的规模和多样性。然而，在遥感领域，现有规模最大的数据集（包括 Million-AID）与计算机视觉领域所采用的网络规模数据集相比仍显不足，后者包含数十亿张图像。例如，LAION-5B [167] 数据集是一个开源数据集，包含超过 50 亿个图像-文本对。为满足对更丰富数据集日益迫切的需求，从而支持大规模训练遥感领域的大型视觉语言模型，必须协同努力建立数据收集与共享机制。因此，研究群体有必要开展合作，构建足够多样化且配有语言描述的数据集。

- **统一视觉语言模型**。继 MiniGPT-4 [266] 和 LLaVA [116],取得显著成功之后，计算机视觉领域对利用大语言模型（大语言模型）开展视觉语言任务的研究兴趣日益浓厚。借助大语言模型强大的推理能力，近期的多模态大语言模型已被设计用于处理来自多种模态的输入，包括图像、视频、音频和点云。在这一研究趋势的推动下，RSGPT [62] 和 GeoChat [84] 等近期工作探索了开发统一视觉语言模型的可行性，以处理广泛的遥感任务，包括但不限于图像描述、视觉问答、视觉定位和区域描述。对更先进、更通用的遥感视觉语言模型的探索，为增强对遥感影像的理解提供了一条前景广阔的研究路径。

- **使用扩散模型进行基于文本的图像生成**。现有神经网络通常需要大量数据才能训练至收敛，但数据采集需要耗费大量人力和物力。另一方面，扩散模型 [158],近期因其生成细节丰富、保真度高的高质量图像的能力而备受关注。通过利用扩散模型根据现有文本描述生成新图像，我们可以创建合成数据，有效扩充数据集规模，从而提升深度学习模型的鲁棒性和泛化能力。此外，通过融入风格迁移或领域自适应等技术，我们可以生成更多样化、更能代表真实场景的合成图像，进一步提升利用扩散模型进行数据增强的效果。

- **少样本/零样本学习**。得益于大语言模型强大的推理能力，视觉语言模型能够根据视觉数据中词语与概念之间的关系识别未见过的目标或模式，展现出在数据高效学习方面的巨大潜力。这使其尤其适用于少样本/零样本学习场景，即训练时可用的标注数据有限。尽管此前已有研究使用规模较小的视觉模型和语言模型，探索在少样本/零样本设置下对遥感图像的理解，但这些模型缺乏理解和识别未见过的目标或模式所必需的推理能力。随着我们迈向通用人工智能时代，必须设计新的技术，以便将大语言模型更好地融入遥感图像理解任务，例如目标检测、语义分割和变化检测，尤其是在少样本/零样本设置下。

- **遥感数据上的高效微调**。 现有的大语言模型通常包含数十亿个参数（例如，GPT-3 拥有 1750 亿个参数），因此，对整个模型进行微调以适配遥感数据并不现实。因此，需要采用高效的模型微调技术，使大语言模型（例如 LLaMA [188]）能够适应遥感图像分析任务。目前有三种潜在的解决方案：1）提示微调 [16] ，即设计可学习的提示，并在新领域中对其进行微调；2）适配器网络 [58],[114] ，即在深度神经网络的现有层之间插入适配器层；3）低秩适配，即将可训练的秩分解矩阵注入 Transformer 架构的每一层。例如，近期提出的 LoRA [61]能够将可训练参数的数量减少 10,000 倍，并将 GPU 内存占用减少三倍，因此有望成为在遥感任务中微调大型模型的一种有效解决方案。

- **将遥感专家知识融入大语言模型**。为了更好地利用大语言模型分析遥感数据，一个重要步骤是将遥感专家知识恰当地融入大语言模型——这需要为大语言模型注入有关遥感图像的领域知识，例如传感器成像理论、空间相关性以及地面目标的光谱特征。近期研究开发了一种名为指令微调的新技术 [210],，用于提升大语言模型遵循指令时的性能。该指令在若干全样本任务上进行微调，随后评估其在特定任务上的零样本泛化能力。在遥感领域，应用指令微调有望使基于知识的指令能够生成和理解遥感图像。

- **通过地理定位将文本信息与遥感相联系**。大语言模型可用于分析与地理位置相关的文本数据，例如社交媒体文本消息[269],报纸等，以提取语言特征，甚至地理信息，随后可将这些信息进一步与遥感数据融合。这为广泛的应用开辟了新的视角，例如建筑物的语义理解 [69],灾害响应 [83],以及具有地理感知能力的社会动态分析 [82], [157], ，并为利用与遥感数据互补的非常规地理数据源提供了新的可能性。

- **气候变化适应与减缓**视觉语言模型（VLMs）提出了一种全面的方法，用于支持和加强气候变化适应与减缓举措，这些举措是联合国可持续发展目标（SDGs）的重要组成部分。例如，视觉语言模型可以在监测森林覆盖变化和加强森林保护工作方面发挥重要作用。视觉语言模型能够无缝整合卫星影像、森林相关文本数据及科学报告。

这种整合有助于自动识别和分类关键森林地标及多种植被物种，从而实现对森林覆盖变化、火灾易发区域和植被转变的实时监测。通过利用视觉语言模型，我们能够有效遏制温室气体排放，维护生态平衡，并以积极主动的方式应对气候变化带来的挑战。

参考文献

[1] T. Abdullah、Y. Bazi、M. M. Al Rahhal、M. L. Mekhalfi、L. Rangarajan 和 M. Zuair。Textrs：用于将文本与遥感图像匹配的深度双向三元组网络。《遥感》，12(3):405，2020。2，[2] 7，12，21 P. Akiva、M. Purri 和 M. Leotta。面向遥感任务的自监督材质与纹理表征学习。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 8203–8215，2022 8 – 页。M. M. Al Rahhal、Y. Bazi、S. O. [3] Alsaleh、M. Al-Razgan、M. L. Mekhalfi、M. Al Zuair 和 N. Alajlan。基于Transformer的开放式遥感视觉问答。《国际遥感杂志》，43 18 6809 6823，2022 2，7，13，21 ( ): –。M. M. Al [4] Rahhal、Y. Bazi、N. A. Alsharif、L. Bashmal、N. Alajlan 和 F. Melgani。用于改进文本到遥感图像检索的多语言Transformer。《IEEE应用地球观测与遥感精选主题期刊》，15:9115–9126，2022。2，7，[5] 12 J.-B. Alayrac、J. Donahue、P. Luc、A. Miech、I. Barr、Y. Hasson、K. Lenc、A. Mensch、K. Millican、M. Reynolds 等。Flamingo：用于少样本学习的视觉语言模型。《神经信息处理系统进展》，35 23716 23736，2022 6 : –。K. Ayush、[6] B. Uzkent、C. Meng、K. Tanmay、M. Burke、D. Lobell 和 S. Ermon。地理感知自监督学习。载于《IEEE/CVF国际计算机视觉会议论文集》，第 10181 10190，2021 8 – 页。J. Bai、S. Bai、S. Yang、S. Wang、S. Tan、[7] P. Wang、J. Lin、C. Zhou 和 J. Zhou。Qwen-vl：具备多种能力的前沿大型视觉语言模型。arXiv预印本 arXiv: 2308 12966，2023 19 。. S. [8] Banerjee 和 A. Lavie。METEOR：一种自动机器翻译评估指标，与人工判断的相关性更高。载于ACL机器翻译和/或摘要生成内在与外在评估指标研讨会论文集，第65 72页，2005 –。[9] 9 A. Bansal、K. Sikka、G. Sharma、R. Chellappa 和 A. Divakaran。零样本目标检测。载于《欧洲计算机视觉会议》（ECCV）论文集，第 384 400，2018 17 – 页。H. Bao、L. Dong、S. Piao 和 F. Wei。BEiT：图像Transformer的BERT预训练。载于《国际学习表征会议》，2021 8。L. Bashmal、Y. [10] Bazi、F. Melgani、R. Ricci、M. M. Al Rahhal 和 M. Zuair。基于遥感图像的视觉问题生成。《IEEE应用地球观测与遥感精选主题期刊》，16 3279 3293，2023 14 : –。F. Bastani、P. Wolters、[11] R. Gupta、J. Ferdinando 和 A. Kembhavi。SatlasPretrain：用于遥感图像理解的大规模数据集。载于《IEEE/CVF国际计算机视觉会议论文集》，第 16772 16782，2023 21 – 页。Y. Bazi、M. M. [12] Al Rahhal、M. L. Mekhalfi、M. A. Al Zuair 和 F. Melgani。用于遥感影像视觉问答的双模态基于Transformer的方法。《IEEE地球科学与遥感汇刊》，60 1 11，2022 2，7，13，14，15 : –。M. B. [13] Bejiga、F. Melgani 和 A. Vascotto。复古遥感：从古代文本生成图像。《IEEE应用地球观测与遥感精选主题期刊》，12 960，2019 3 950 ( ): –。2，7，11 [14] P. Bojanowski、E. Grave、A. Joulin 和 T. Mikolov。利用子词信息丰富词向量。《计算语言学协会汇刊》，5 135 146，2017 7，16 : –。T. Brown、B. [15] Mann、N. Ryder、M. Subbiah、J. D. Kaplan、P. Dhariwal、A. Neelakantan、P. Shyam、G. Sastry、A. Askell 等。语言模型是少样本学习者。《神经信息处理系统进展》，33:1877–1901，2020。4，6，22 [16] N. Carion、F. Massa、G. Synnaeve、N. Usunier、A. Kirillov 和 S. Zagoruyko。利用Transformer进行端到端目标检测。载于《计算机视觉—ECCV 2020：第16届欧洲会议，英国格拉斯哥，2020年8月23日至28日，论文集》，第I部分第16卷，第213–229页。施普林格，2020 17 。K. Cha、J. Seo 和 T. Lee。面向遥感图像的十亿级基础模型。arXiv预印本 arXiv: 2304 05215，2023 8 。。C. Chappuis、V. [17] Mendez、E. Walt、S. Lobry、B. Le Saux 和 D. Tuia。用于遥感视觉问答的语言Transformer。载于 IGARSS 2022-2022 IEEE国际地球科学与遥感研讨会，第 4855 4858 2022 2，7，13 –页。IEEE，。C. [18] Chappuis、V. Zermatten、S. Lobry、B. Le Saux 和 D. Tuia。Prompt-rsvqa：通过提示视觉上下文引导语言模型进行遥感视觉问答。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第1372– 1381，2022 14 页。K. Chatfield、K. [19] Simonyan、A. Vedaldi 和 A. Zisserman。细节中的魔鬼归来：深入探索卷积网络。载于《2014年英国机器视觉会议论文集》。英国机器视觉协会，2014 7 。C. Chen、H. Ma、G. Yao、N. Lv、H. [20] Yang、C. Li 和 S. Wan。基于文本描述的遥感图像增强，用于水岸变化检测。《遥感》，13 10 1894，2021 2，7，11 ( ): 。J. [21] Chen、H. Guo、K. Yi、B. Li 和 M. Elhoseiny。VisualGPT：以数据高效的方式适配预训练语言模型进行图像描述。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 18030 18040，2022 2，5 –页。J. Chen、D. Zhu、X. Shen、X. Li、Z. Liu、P. Zhang、R. Krishnamoorthi、V. Chandra、[22] Y. Xiong 和 M. Elhoseiny。MiniGPT-v2：将大语言模型用作视觉语言多任务学习的统一接口。arXiv: 2310 09478，2023 1，18 。。K. Chen、Z. Zhang、W. Zeng、R. Zhang、[23] F. Zhu 和 R. Zhao。Shikra：释放多模态大语言模型指代对话的魅力。arXiv预印本 arXiv: 2306 15195，2023 19 。。Y. Chen、C. Wei、D. Wang、C. Ji 和 B. Li。[24] 用于遥感图像少样本分割的半监督对比学习。《遥感》，14 17 4254，2022 2，7，17 ( ): 。Y.-C. Chen、L. Li、L. Yu、[25] A. El Kholy、F. Ahmed、Z. Gan、Y. Cheng 和 J. Liu。Uniter：通用图文表征学习。载于欧洲计算机视觉会议，第 104 120 2020 5 –页。施普林格，。G. Cheng、J. Han 和 X. Lu。遥感图像场景分类：基准与最新技术水平。[26] 《IEEE会刊》，105 10 1865 1883，2017 7，21 ( ): –。G. Cheng、J. Han、P. Zhou 和 L. Guo。[27] 基于部件检测器集合的多类地理空间目标检测与地理图像分类。《ISPRS摄影测量与遥感学报》，98 119 132，2014 7，21 : –。G. Cheng、B. Yan、P. Shi、K. Li、X. Yao、L. Guo 和 J. Han。[28] 用于遥感图像少样本目标检测的原型CNN。《IEEE地球科学与遥感汇刊》，60 1 10，2021 17 : –。Q. [29] Cheng、Y. Zhou、P. Fu、Y. Xu 和 L. Zhang。用于遥感跨模态图文检索的深度语义对齐网络。《IEEE应用地球观测与遥感精选主题期刊》，14 4284 4297，2021 2，7，12 : –。L. [30] Chin-Yew。Rouge：一种用于自动评估摘要的工具包。载于《文本摘要拓展研讨会论文集》，2004，2004 9 。K. Cho、B. van Merrienboer、C。[31] Gulcehre、D. Bahdanau、F. Bougares、H. Schwenk 和 Y. Bengio。利用RNN编码器-解码器学习用于统计机器翻译的短语表示。载于《2014年自然语言处理实证方法会议（EMNLP）论文集》，第 1724 1734，2014 7 –页。X. Chu、Z. Tian、Y. [32] Wang、B. Zhang、H. Ren、X. Wei、H. Xia 和 C. Shen。Twins：重新审视视觉Transformer中的空间注意力设计。《神经信息处理系统进展》，34 9355 9366，2021 3 : 。A. Creswell、T. White、V. Dumoulin、K. [33] Arulkumaran、B. Sengupta 和 A. A. Bharath。生成对抗网络：综述。《IEEE信号处理杂志》，35 1 53 65，2018 7 ( ): –。[34] Y. Cui、W. Che、T. Liu、B. Qin、S. Wang 和 G. Hu。重新审视中文自然语言处理的预训练模型。arXiv预印本 arXiv: 2004 13922，2020 4 。。D. Dai 和 W. Yang。[35] 利用双层稀疏编码和偏置图像表示进行卫星图像分类。《IEEE地球科学与遥感汇刊》，8 1 173 176，2011 21 ( ): –。W. Dai、J. Li、D. Li、[36] A. M. H. Tiong、J. Zhao、W. Wang、B. Li、P. N. Fung 和 S. Hoi。InstructBLIP：迈向具备指令微调能力的通用视觉语言模型。《神经信息处理系统进展》，第36卷，2024年。6、9 [37] [38]

[39] C. Debes、A. Merentitis、R. Heremans、J. Hahn、N. Frangiadakis、T. van Kasteren、W. Liao、R. Bellens、A. Pi˘zurica、S. Gautama 等。高光谱与激光雷达数据融合：2013年GRSS数据融合竞赛的成果。《IEEE应用地球观测与遥感精选主题期刊》，7 6 2405 2418, 2014 7, 21 ( ): –。[40] J. Deng、W. Dong、R. Socher、L.-J. Li、K. Li 和 L. Fei-Fei。ImageNet：大型分层图像数据库。载于《2009年IEEE计算机视觉与模式识别会议》，第248 255 2009 3 –页。IEEE，。[41] J. Deng、Z. Yang、T. Chen、W. Zhou 和 H. Li。TransVG：利用Transformer进行端到端视觉定位。载于《IEEE/CVF国际计算机视觉会议论文集》，第1769–1779页，2021年。[42] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova。BERT：用于语言理解的深度双向Transformer预训练。arXiv预印本 arXiv:1810 04805, 2018 3, 4, 5, 6, 7, 13, 16 。。A. Dosovitskiy、L. [43] Beyer、A. Kolesnikov、D. Weissenborn、X. Zhai、T. Unterthiner、M. Dehghani、M. Minderer、G. Heigold、S. Gelly 等。一张图像胜过16×16个词：用于大规模图像识别的Transformer。载于《国际学习表征会议》，2020年。3, 4, 7 [44] P. Esser、R. Rombach 和 B. Ommer。驾驭Transformer，实现高分辨率图像合成。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第12873–12883, 2021 7 页。F. Faghri、D. J. Fleet、J. R. Kiros 和 S. Fidler。VSE++：[45] 利用困难负样本改进视觉语义嵌入。载于《英国机器视觉会议论文集》2018, 2018 13 。A. Farhadi 和 J. Redmon。[46] YOLOv：渐进式改进。载于《计算机视觉与模式识别》，第1804卷，第1–6页。施普林格，德国柏林/海德堡，2018 7 。P. 3 Gamba。用于城市区域表征的数据集合。载于 IGARSS 2004，即2004年IEEE国际地球科学与遥感研讨会，第1 2004 7, 21 卷。[47] IEEE，。R. Girshick。Fast R-CNN。载于《IEEE国际计算机视觉会议论文集》，第1440 1448, 2015 17 –页。A. Graves 和 J. [48] Schmidhuber。利用双向长短期记忆网络及其他神经网络架构进行逐帧音素分类。《神经网络》，18 5 6 602 610, 2005 7 (-): –。X. Gu、T. Lin、[49] W. Kuo 和 Y. Cui。通过视觉与语言知识蒸馏实现零样本检测。corr abs/2104 13921 2021 17 。( )。K. Han、[50] A. Xiao、E. Wu、J. Guo、C. Xu 和 Y. Wang。Transformer中的Transformer。《神经信息处理系统进展》，34 15908 15919, 2021 3 : –。K. He、X. Chen、[51] S. Xie、Y. Li、P. Doll´ar 和 R. Girshick。掩码自编码器是可扩展的视觉学习器。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第16000 16009, 2022 8 –页。K. He、X. Zhang、S. Ren 和 J. Sun。用于图[52] 像识别的深度残差学习。载于《IEEE计算机视觉与模式识别会议论文集》，第770 778, 2016 3, 7 –页。K. Heidler、L. Mou、[53] D. Hu、P. Jin、G. Li、C. Gan、J.-R. Wen 和 X. X. Zhu。用于遥感数据的自监督视听表征学习。《国际应用地球观测与地理信息学杂志》，116 103130, 2023 8 : 。M. Heusel、H. Ramsauer、T. [54] Unterthiner、B. Nessler 和 S. Hochreiter。通过双时间尺度更新规则训练的生成对抗网络收敛至局部纳什均衡。《神经信息处理系统进展》，第30卷，2017年。11 [55] S. Hochreiter 和 J. Schmidhuber。长短期记忆。《神经计算》，9 8 1735 1780, 1997 7, 11, 12, 13 ( ): –。J. J. Hopfield 和 D. [56] W. Tank。7 优化问题的“神经”决策计算。《生物控制论》，52(3)：141–152，1985年。[57] N. Houlsby、A. Giurgiu、S. Jastrzebski、B. Morrone、Q. De [58] Laroussilhe、A. Gesmundo、M. Attariyan 和 S. Gelly。面向自然语言处理的参数高效迁移学习。载于《国际机器学习会议》，第2790 2799 2019 22 –页。PMLR，。G. Hoxha 和 F. Melgani。[59] 用于遥感图像描述的新型基于支持向量机的解码器。《IEEE地球科学与遥感汇刊》，60 1 14, 2021 10 : –。G. Hoxha、G. Scuccato 和 F. [60] Melgani。利用后处理策略改进图像描述系统。《IEEE地球科学与遥感汇刊》，2023 10 。E. J. Hu、P. Wallis、Z. [61] Allen-Zhu、Y. Li、S. Wang、L. Wang、W. Chen 等。LoRA：大语言模型的低秩适配。载于《国际学习表征会议》，2021 22 。Y. Hu、J. [62] Yuan、C. Wen、X. Lu 和 X. Li。RSGPT：一种遥感视觉语言模型及其基准。arXiv预印本 arXiv:2307 15266，2023年，2、9、10、14、15、21、22。。

[63] B. Huang、D. Lian、W. Luo 和 S. Gao。未雨绸缪：为单阶段视觉定位学习地标特征。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第16888 16897, 2021 16 – 页。G. Huang、Z. Liu、L. [64] Van Der Maaten 和 K. Q. Weinberger。密集连接卷积网络。载于《IEEE计算机视觉与模式识别会议论文集》，第4700–4708, 2017 3 页。S. Huang、L. Dong、W. Wang、Y. Hao、S. Singhal、S. Ma、T. Lv、L. Cui、O. K. Mohammed、B. [65] Patra 等。语言并非你所需的一切：使感知与语言模型对齐。《神经信息处理系统进展》，36, 2024 6 。X. Huang、B. He、M. Tong、D. Wang 和 C. He。通过共享注意力模块和平衡微调策略实现遥感图像上的[66] 少样本目标检测。《遥感》，13 19 3816, 2021 17 ( ): 。Y. [67] Huang、Q. Wu、C. Song 和 L. Wang。学习图像与句子匹配中的语义概念及顺序。载于《IEEE计算机视觉与模式识别会议论文集》，第6163 6171, 2018 13 – 页。Y. Huo、M. Zhang、G. Liu、H. Lu、Y. Gao、G. Yang、J. Wen、H. Zhang、B. [68] Xu、W. Zheng 等。Wenlan：通过大规模多模态预训练连接视觉与语言。arXiv预印本 arXiv:2103 06561，2021。[69] 5 ¨ M. Haberle、E. J. Hoffmann 和 X. X. Zhu。从地理标记推文中提取的语言特征能否帮助进行遥感中的建筑功能分类？《ISPRS摄影测量与遥感学报》，188 255 268, 2022 22 : – 页。F. N. [70] Iandola、S. Han、M. W. Moskewicz、K. Ashraf、W. J. Dally 和 K. Keutzer。SqueezeNet：以少50倍的参数实现AlexNet级别的精度，并且模型大小为¡ 0 07360，2016 5 1602。arXiv预印本 arXiv: 。[71] 7 C. Jia、Y. Yang、Y. Xia、Y.-T. Chen、Z. Parekh、H. Pham、Q. Le、Y.-H. Sung、Z. Li 和 T. Duerig。利用噪声文本监督扩展视觉和视觉语言表征学习。载于《国际机器学习会议》，第4904 4916 2021 5 – 页。PMLR，。[72] X. Jiang、N. Zhou 和 X. Li。使用深度度量学习进行遥感图像少样本分割。《IEEE地球科学与遥感快报》，19 1 5, 2022 2, 7, 17, 20 : – 页。L. [73] Jiao、Z. Huang、X. Lu、X. Liu、Y. Yang、J. Zhao、J. Zhang、B. Hou、S. Yang、F. Liu 等。受脑启发的遥感基础模型及开放问题：综合综述。《IEEE应用地球观测与遥感精选主题期刊》，2023 8 。A. Joulin、[74] E. Grave 和 P. B. T. Mikolov。高效文本分类的技巧汇编。EACL 2017, 427, 2017 7, 16 , 页。H. Kandala、S. [75] Saha、B. Banerjee 和 X. X. Zhu。探索用于遥感图像描述的Transformer与多标签分类。《IEEE地球科学与遥感快报》，19 5，2022 1: –。[76] 9 B. Kang、Z. Liu、X. Wang、F. Yu、J. Feng 和 T. Darrell。通过特征重加权实现少样本目标检测。载于《IEEE/CVF国际计算机视觉会议论文集》，第8420–8429, 2019 18 页。[77] L. Karlinsky、J. Shtok、S. Harary、E. Schwartz、A. Aides、R. Feris、R. Giryes 和 A. M. Bronstein。RepMet：基于代表元的分类与少样本目标检测度量学习。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第5197 5206, 2019 18 – 页。A. Kirillov、E. Mintun、[78] N. Ravi、H. Mao、C. Rolland、L. Gustafson、T. Xiao、S. Whitehead、A. C. Berg、W.-Y. Lo 等。Segment Anything。载于《IEEE/CVF国际计算机视觉会议论文集》，第4015 4026, 2023 6 – 页。E. Kodirov、[79] T. Xiang 和 S. Gong。用于零样本学习的语义自编码器。载于《计算机视觉与模式识别会议论文集》，第3174 3183, 2017 15, 16 – 页。A. Kolesnikov、[80] L. Beyer、X. Zhai、J. Puigcerver、J. Yung、S. Gelly 和 N. Houlsby。Big Transfer（BiT）：通用视觉表征学习。载于《计算机视觉—ECCV 2020：第16届欧洲计算机视觉会议，英国格拉斯哥，2020年8月23日至28日，论文集，第V部分16》，第491–507页。施普林格，2020 7 。A. Krizhevsky、I. Sutskever 和[81] G. E. Hinton。使用深度卷积神经网络进行 ImageNet 分类。《 Co mmunications of the ACM》，60 6 84 90, 2017 3, 7 ( ): –。A. [82] Kruspe、M. Haeberle、I. Kuhn 和 X. X. Zhu。新冠肺炎疫情期间欧洲Twitter消息的跨语言情感分析。载于ACL 2020自然语言处理与COVID-19 2020 22 (NLP-COVID)研讨会，。A. Kruspe、J. [83] Kersten 和 F. Klan。危机事件中可采取行动推文的检测。《Natural Hazards and Earth System Sciences》，21(6):1825–1845，2021。22

[84] K. Kuckreja、M. S. Danish、M. Naseer、A. Das、S. Khan 和 F. S. Khan。Geochat：基于地理信息的视觉语言大模型，用于遥感。arXiv预印本 arXiv: 2311.15826, 2023 22 。。Z. Lan、M. Chen、S. Goodman、K. Gimpel、P. Sharma 和 R. Soricut。ALBERT：用于语言表征自监督学习的轻量级BERT。载于《国际学习表征会议》，2019年。[85] 4 [86] C. Lang、G. Cheng、B. Tu 和 J. Han。学习不分割什么：一种少样本分割的新视角。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 8057–8067, 2022 20 – 页。C. Lang、G. Cheng、B. Tu 和 J. Han。遥感影像少样本分割的全局校正与解耦配准。《IEEE地球科学与遥感汇刊》，[87] 17, 20 2023年。[88] C. Lang、B. Tu、G. Cheng 和 J. Han。超越原型：用于少样本分割的分治代理。arXiv预印本 arXiv: 2204.09903, 2022 20 。。B. Le Saux、N. Yokoya、R. H¨ansch 和 S. Prasad。2018年IEEE GRSS数据融合竞赛：多模态土地利用分类 [技术委员会] 2018 [89] 6 1 52–54, 。《IEEE地球科学与遥感杂志》，(): –。[90] 7, 21 Y. LeCun、L. Bottou、Y. Bengio 和 P. Haffner。应用于文档识别的基于梯度的学习。《IEEE会刊》，86 11 2278–2324, 1998 1, 3 (): –。K.-H. Lee、X. Chen、G. Hua、H. Hu 和 X. He。用于图文匹配的堆叠交叉注意力。载于《欧洲计算机视觉会议（ECCV）论文集》，第 201–216, 2018 13 – 页。A. Li、Z. Lu、L. Wang、T. Xiang 和 J.-R. Wen。面向高空间分辨率遥感图像的零样本场景分类。《IEEE地球科学与遥感汇刊》，55 4167，2017 7 4157(): –。[91] [92] 2, 7, 15, 16 B. Li、K. Q. Weinberger、S. Belongie、V. Koltun 和 R. Ranftl。语言驱动的语义分割。载于《国际学习表征会议》，2021 17 。D. Li、J. Li、H. Le、G. Wang、S. Savarese 和 S. C. Hoi。LAVIS：面向语言视觉智能的一站式库。载于《计算语言学协会第61届年会论文集》（第 3 卷：系统演示），31–41, 2023 20 页。[93] [94] J. Li、D. Li、S. Savarese 和 S. Hoi。BLIP-2：利用冻结的图像编码器和大语言模型引导语言图像预训练。载于《国际机器学习会议》，第19730–19742 页。PMLR，2023 2, 6 。。J. Li、D. Li、C. Xiong 和 S. Hoi。BLIP：为统一视觉语言理解与生成引导语言图像预训练。载于《国际机器学习会议》，第12888–12900页。PMLR，2022 2 。J. Li、R. Selvaraju、A. Gotmare、S. Joty、C. Xiong 和 S. C. H. Hoi。在融合前对齐：通过动量蒸馏进行视觉语言表征学习。NeurIPS，2021 5 。K. Li、G. Wan、G. Cheng、L. Meng 和 J. Han。光学遥感图像中的目标检测：综述与新基准。《ISPRS摄影测量与遥感学报》，159:296–307，2020。7，[95] [96] [97] [98] 21 L. H. Li、M. Yatskar、D. Yin、C.-J. Hsieh 和 K.-W. Chang。VisualBERT：一种简洁且高性能的视觉与语言基线。arXiv预印本 arXiv: 1908.03557, 2019 5 。。W. Li、K. Chen、H. Chen 和 Z. Shi。面向遥感图像的地理知识驱动的表示学习。《IEEE地球科学与遥感汇刊》，[99] [100] 60 1 16, 2021 8 : –。W.-z. Li、J.-w. Zhou、X. Li、Y. Cao 和 G. Jin。通过深度度量学习与知识继承实现航空影像少样本目标检测。《国际应用地球观测与地理信息学杂志》，[101] 122 103397, 2023 17, 18 :。X. Li、J. Deng 和 Y. Fang。遥感图像上的少样本目标检测。《IEEE地球科学与遥感汇刊》，[102] 60 1 14, 2021 17, 18 : –。X. Li、C. Wen、Y. Hu 和 N. Zhou。RS-CLIP：通过对比视觉语言监督实现零样本遥感场景分类。《国际应用地球观测与地理信息学杂志》，[103] 124 103497, 2023 16 :。X. Li、X. Yin、C. Li、P. Zhang、X. Hu、L. Zhang、L. Wang、H. Hu、L. Dong、F. Wei 等。OSCAR：面向视觉语言任务的对象语义对齐预训练。载于《欧洲计算机视觉会议》，第121–137页。施普林格，2020年。[105] [104] 3 X. Li、A. Yuan 和 X. Lu。用于图像描述的多模态门控循环单元。《 Multimedia Tools and Applications》，77:29847–29869, 2018 10 。X. Li、X. Zhang、W. Huang 和 Q. Wang。截断交叉熵损失 [106] 用于遥感图像描述生成。《IEEE地球科学与遥感汇刊》，59 6 5246–5257, 2020 2, 7, 9 (): –。Y. Li、S. Fang、L. Jiao、R. Liu 和 R. Shang。用于遥感图像描述的多层次注意力模型。《遥感》，12 939，2020 6(): 。[107] [108] 2, 7, 9 Y. Li、F. Liang、L. Zhao、Y. Cui、W. Ouyang、J. Shao、F. Yu 和 J. Yan。监督无处不在：一种数据高效的对比式语言—图像预训练范式。载于《国际学习表征会议》，2021 5 。Y. Li、Z. Zhu、J.-G. Yu 和 Y. Zhang。面向零样本遥感图像场景分类的深度跨模态嵌入网络学习。《IEEE地球科学与遥感汇刊》，59 12 10590–10603, 2021 15 (): –。[109] Z. Li、D. Zhang、Y. Wang、D. Lin 和 J. Zhang。用于零样本遥感场景分类的生成对抗网络。《应用科学》，12 8 3760, 2022 2, 7, 15, 16 (): 。F. Liang、B. Wu、X. Dai、K. Li、Y. Zhao、H. Zhang、P. Zhang、P. Vajda 和 D. Marculescu。使用掩码适配 CLIP 的开放词汇语义分割。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 7061–7070, 2023 17 – 页。J. Lin、A. Yang、Y. Zhang、J. Liu、J. Zhou 和 H. Yang。InterBERT：用于多模态预训练的视觉与语言交互。arXiv 预印本 arXiv: 2003.13198, 2020 5 。T.-Y. Lin、P. Goyal、R. Girshick、K. He 和 P. Doll´ar。用于密集目标检测的焦点损失。载于《IEEE国际计算机视觉会议论文集》，第 2980–2988, 2017 17 – 页。[110] [111] [112] [113] [114] Z. Lin、A. Madotto 和 P. Fung。通过参数高效迁移学习探索多功能生成式语言模型。arXiv 预印本 arXiv: 2004.03829, 2020 22 。C. Liu、R. Zhao 和 Z. Shi。基于多层聚合变换器的遥感图像描述生成。《IEEE地球科学与遥感快报》，19 1–5, 2022 10 : –。H. Liu、C. Li、Q. Wu 和 Y. J. Lee。视觉指令微调。《神经信息处理系统进展》，36, 2024 6, 19, 22 。S. Liu、Y. You、H. Su、G. Meng、W. Yang 和 F. Liu。遥感图像解译中的少样本目标检测：机遇与挑战。《遥感》，14 18 4435, 2022 17 (): 。W. Liu、D. Anguelov、D. Erhan、C. Szegedy、S. Reed、C.-Y. Fu 和 A. C. Berg。SSD：单发多框检测器。载于《计算机视觉—ECCV 2016：第14届欧洲计算机视觉会议》，荷兰阿姆斯特丹，2016年10月11日至14日，论文集，第一部分14，第21–37 页。施普林格，2016 17 。Y. Liu、N. Liu、Q. Cao、X. Yao、J. Han 和 L. Shao。为小样本语义分割学习非目标知识。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 11573–11582, 2022 20 – 页。Y. Liu、M. Ott、N. Goyal、J. Du、M. Joshi、D. Chen、O. Levy、M. Lewis、L. Zettlemoyer 和 V. Stoyanov。RoBERTa：一种稳健优化的 BERT 预训练方法。arXiv 预印本 arXiv: 1907.11692, 2019 4, 5 。Z. Liu、Y. Lin、Y. Cao、H. Hu、Y. Wei、Z. Zhang、S. Lin 和 B. Guo。Swin Transformer：使用移位窗口的分层视觉变换器。载于《IEEE/CVF国际计算机视觉会议论文集》，第 10012–10022, 2021 3, 7 – 页。S. Lobry、D. Marcos、J. Murray 和 D. Tuia。RSVQA：面向遥感数据的视觉问答。《IEEE地球科学与遥感汇刊》，[115] [116] [117] [118] [119] [120] [121] [122] 58 12 8555–8566, 2020 2, 7, 13, 14, 15, 21 (): –。Y. Long、G.-S. Xia、S. Li、W. Yang、M. Y. Yang、X. X. Zhu、L. Zhang 和 D. Li。关于创建航空图像解译基准数据集：综述、指南与 Million-AID。《IEEE应用地球观测与遥感精选主题期刊》，14 4230，2021 4205: –。[123] [124] 7, 21 J. Lu、D. Batra、D. Parikh 和 S. Lee。ViLBERT：为视觉与语言任务预训练任务无关的视觉语言表征。《神经信息处理系统进展》，32, 2019 5, 6 。X. Lu、X. Sun、W. Diao、Y. Mao、J. Li、Y. Zhang、P. Wang 和 K. Fu。在文本模态知识引导下进行航空影像少样本目标检测。《IEEE地球科学与遥感汇刊》，2023 2, 7, 17, 18, 19 。X. Lu、B. Wang 和 X. Zheng。用于遥感图像描述生成的声音主动注意力框架。《IEEE地球科学与遥感汇刊》，[125] [126] 58 3 1985–2000, 2019 10 (): –。X. Lu、B. Wang、X. Zheng 和 X. Li。探索遥感图像描述生成的模型与数据。《IEEE地球科学与遥感汇刊》，56(4):2183–2195，2017。2, 7, 9, 10,[127] [128] 21 T. L¨uddecke 和 A. Ecker。使用文本和图像提示进行图像分割。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 7086–7096, 2022 17 – 页。H. Luo、J. Bao、Y. Wu、X. He 和 T. Li。SegCLIP：图块聚合 [129]

通过可学习中心实现开放词汇语义分割。载于《国际机器学习会议》，第23033–23044页。PMLR，2023 18。N. Lv、H. Ma、C. Chen、Q. Pei、Y. Zhou、F. Xiao 和 J. Li。通过对抗训练进行遥感数据增强。《IEEE应用地球观测与遥感精选主题期刊》，14 9318 9333, 2021 7 : – 。C. Ma、Y. Yang、Y. Wang、Y. Zhang 和 W. Xie。使用冻结的视觉语言模型进行开放词汇语义分割。载于《2022年英国机器视觉会议论文集》，2022年。[132] O. Manas、A. Lacoste、X. Gir´o-i Nieto、D. Vazquez 和 P. Rodriguez。季节性对比：从未经筛选的遥感数据进行无监督预训练。载于《IEEE/CVF国际计算机视觉会议论文集》，第9414 9423, 2021 8 – 页。M. Mendieta、B. Han、X. Shi、Y. Zhu、C. Chen 和 M. Li。GFM：通过持续预训练构建地理空间基础模型。arXiv 预印本 arXiv:2302 04476, 2023 7, 8 。T. Mikolov、K. Chen、G. Corrado 和 J. Dean。高效估计词语在向量空间中的表征。arXiv 预印本 arXiv:1301 3781，2013。[135] A. v. d. Oord、Y. Li 和 O. Vinyals。基于对比预测编码的表征学习。arXiv 预印本 arXiv:1807.03748，2018。[136] OpenAI。ChatGPT。OpenAI API，2023 1 。OpenAI。GPT 技术报告，。[138] L. P. Osco、E. L. d. Lemos、W. N. Gonc¸alves、A. P. M. Ramos 和 J. Marcato Junior。Visual ChatGPT 在遥感领域的潜力。《遥感》，15 13 3232, 2023 2 ( ): 。L. Ouyang、J. Wu、X. Jiang、D. Almeida、C. Wainwright、P. Mishkin、C. Zhang、S. Agarwal、K. Slama、A. Ray 等。训练语言模型，使其遵循人类反馈中的指令。《神经信息处理系统进展》，35 27730 27744, 2022 4 : – 。K. Papineni、S. Roukos、T. Ward 和 W.-J. Zhu。BLEU：一种机器翻译自动评估方法。载于《计算语言学协会第40届年会论文集》，第311 318, 2002 9 – 页。Z. Peng、W. Wang、L. Dong、Y. Hao、S. Huang、S. Ma 和 F. Wei。Kosmos-2：将多模态大语言模型视觉定位到真实世界。arXiv 预印本 arXiv:2306 14824, 2023 6 。G. Puthumanaillam 和 U. Verma。基于纹理的原型网络，用于森林覆盖的小样本语义分割：推广至不同地理区域。《神经计算》，538:126201，2023。[143] B. Qu、X. Li、D. Tao 和 X. Lu。高分辨率遥感图像的深度语义理解。载于《2016年计算机、信息与电信系统国际会议（CITS）》，第1 5 2016 7, 9, 10, 21 – 页。IEEE，。J. Quan、C. Wu、H. Wang 和 Z. Wang。基于结构对齐的遥感场景零样本分类。载于《2018 年 IEEE 电子与通信工程国际会议（ICECE）》，第17 21 2018 2, 7, 15, 16 – 页。IEEE，。A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark 等。通过自然语言监督学习可迁移的视觉模型。载于《国际机器学习会议》，第8748–8763页。PMLR，2021。5,[146] A. Radford、K. Narasimhan、T. Salimans、I. Sutskever 等。通过生成式预训练改进语言理解。2018 1, 3, 4 。A. Radford、J. Wu、R. Child、D. Luan、D. Amodei、I. Sutskever 等。语言模型是无监督的多任务学习者。OpenAI 博客，1 8 9, 2019 4, 7 ( ): 。C. Raffel、N. Shazeer、A. Roberts、K. Lee、S. Narang、M. Matena、Y. Zhou、W. Li 和 P. J. Liu。利用统一的文本到文本 Transformer 探索迁移学习的极限。《机器学习研究杂志》，21 1 5485 5551, 2020 4, 6 ( ): – 。M. M. A. Rahhal、Y. Bazi、T. Abdullah、M. L. Mekhalfi 和 M. Zuair。利用文本线索进行遥感图像检索的深度无监督嵌入。《应用科学》，10 24 8931, 2020 2, 7, 12 ( ): 。M. M. A. Rahhal、M. A. Bencherif、Y. Bazi、A. Alharbi 和 M. L. Mekhalfi。多模态遥感图像检索的双 Transformer 架构对比。《应用科学》，13(1):282，2023。2, 7, 12 [151] S. Rahman、S. Khan 和 N. Barnes。改进零样本目标检测的视觉—语义对齐。载于《AAAI人工智能会议论文集》，第34卷，11939，2020 11932页 – 。[152] A. Ramesh、M. Pavlov、G. Goh、S. Gray、C. Voss、A. Radford、M. Chen 和 I. Sutskever。零样本文本到图像生成。载于《国际机器学习会议，页码8821 8831 2021 11 – 。PMLR，。[153] H. Ramsauer、B. Schafl、J. Lehner、P. Seidl、M. Widrich、L. Gruber、M. Holzleitner、T. Adler、D. Kreil、M. K. Kopp 等。霍普菲尔德网络就是你所需要的一切。载于国际学习表征会议，2020 11 。Y. Rao、W. Zhao、G. Chen、Y. Tang、Z. Zhu、G. Huang、J. Zhou 和 J. Lu。DenseCLIP：通过上下文感知提示实现语言引导的密集预测。载于 IEEE/CVF计算机视觉与模式识别会议论文集，页码18082 18091, 2022 17 – 。J. Redmon、S. Divvala、R. Girshick 和 A. Farhadi。你只需看一次：统一的实时目标检测。载于 IEEE计算机视觉与模式识别会议论文集，页码 779 788，2016 – 。[156] S. Ren、K. He、R. Girshick 和 J. Sun。Faster R-CNN：借助区域提议网络实现实时目标检测。神经信息处理系统进展，28, 2015 5, 7, 17 。S. Rode-Hasinger、A. Kruspe 和 X. X. Zhu。是真是假？利用图神经网络检测社交媒体上的虚假信息。载于第八届噪声用户生成文本研讨会论文集（W-NUT 2022 222 229, 2022 22 ），页码 – 。R. Rombach、A. Blattmann、D. Lorenz、P. Esser 和 B. Ommer。利用潜在扩散模型进行高分辨率图像合成。载于 IEEE/CVF计算机视觉与模式识别会议论文集，页码10684 10695, 2022 2, 22 – 。F. Rottensteiner、G. Sohn、M. Gerke 和 J. D. Wegner。ISPRS语义标注竞赛。ISPRS：德国利奥波尔茨赫厄，（）：。F. Rottensteiner、G. Sohn、J. Jung、M. Gerke、C. Baillard、S. Benitez 和 U. Breitkopf。ISPRS城市目标分类与三维建筑重建基准。ISPRS摄影测量、遥感与空间信息科学年刊 I-3 1，1 2012 1 293 298，2012（），第（）：– 。S. Ruan、Y. Zhang、K. Zhang、Y. Fan、F. Tang、Q. Liu 和 E. Chen。DAE-GAN：用于文本生成图像的动态方面感知生成对抗网络。载于 IEEE/CVF国际计算机视觉会议论文集，页码13960 13969, 2021 11 – 。A. Sadhu、K. Chen 和 R. Nevatia。基于自然语言查询的目标零样本定位。载于 IEEE/CVF国际计算机视觉会议论文集，页码 4694–4703，2019。[163] T. Salimans、I. Goodfellow、W. Zaremba、V. Cheung、A. Radford 和 X. Chen。改进生成对抗网络训练的技术。神经信息处理系统进展，29, 2016 11 。J. W. Sammon。用于数据结构分析的非线性映射。IEEE计算机汇刊，100 5 401 409, 1969 15 ( ): – 。M. Sandler、A. Howard、M. Zhu、A. Zhmoginov 和 L.-C. Chen。MobileNetV2：倒置残差与线性瓶颈。载于 IEEE计算机视觉与模式识别会议论文集，页码4510 4520, 2018 3 – 。M. Schmitt、L. H. Hughes、C. Qiu 和 X. X. Zhu。SEN12MS——用于深度学习与数据融合的地理参考 Sentinel-1/2 多光谱影像精选数据集。载于 ISPRS 摄影测量、遥感与空间信息科学年刊，第 IV-2 7, 153 160, 2019 21 /W 卷，页码 – 。C. Schuhmann、R. Beaumont、R. Vencu、C. Gordon、R. Wightman、M. Cherti、T. Coombes、A. Katta、C. Mullis、M. Wortsman 等。LAION-5B：用于训练新一代图像-文本模型的开放大规模数据集。神经信息处理系统进展，35 25278 25294, 2022 22 : – 。M. Schuster 和 K. K. Paliwal。双向循环神经网络。IEEE信号处理汇刊，45 11 2673 2681, 1997 7 ( ): – 。R. Sennrich、B. Haddow 和 A. Birch。利用子词单元实现低频词的神经机器翻译。载于计算语言学协会第54届年会论文集（第一卷：长篇论文），页码1715 1725, 2016 7 – 。X. Shen、B. Liu、Y. Zhou、J. Zhao 和 M. Liu。通过变分自编码器和强化学习实现遥感图像描述。知识型系统，203 105920, 2020 7, 9 : 。Z. Shi 和 Z. Zou。机器能否为遥感图像生成类人语言描述？IEEE地球科学与遥感汇刊，55 6 3623 3634, 2017 2, 7, 8 ( ): – 。K. Simonyan 和 A. Zisserman。用于大规模图像识别的超深卷积网络。载于第三届国际学习表征会议（ICLR 2015）。计算与生物学习学会，2015 3, 7 。G. Sumbul、R. G. Cinbis 和 S. Aksoy。遥感影像中的细粒度目标识别与零样本学习。IEEE地球科学与遥感汇刊，56 2 770 779, 2017 2, 7, 15 ( ): – 。G. Sumbul、A. De Wall、T. Kreuziger、F. Marcelino、H. Costa、P. Bene- vides、M. Caetano、B. Demir 和 V. Markl。BigEarthNet-MM：用于遥感图像分类与检索的大规模、多模态、多标签基准档案

图像分类与检索 [软件和数据集]。IEEE地球科学与遥感杂志，9 3 174 180, 2021 21 ( ): – 。

[175] G. Sumbul、S. Nayak 和 B. Demir。SD-RSIC：由摘要驱动的深度遥感图像描述。IEEE地球科学与遥感汇刊，59 8 6922 6934, 2020 10 ( ): – 。

[176] B. Sun、B. Li、S. Cai、Y. Yuan 和 C. Zhang。FSCE：通过对比式提议编码实现少样本目标检测。载于 IEEE/CVF计算机视觉与模式识别会议论文集，页码 7352 7362, 2021 18 – 。

[177] X. Sun、P. Wang、W. Lu、Z. Zhu、X. Lu、Q. He、J. Li、X. Rong、Z. Yang、H. Chang 等。RingMo：采用掩码图像建模的遥感基础模型。IEEE地球科学与遥感汇刊，2022 7, 8 。

[178] X. Sun、P. Wang、Z. Yan、F. Xu、R. Wang、W. Diao、J. Chen、J. Li、Y. Feng、T. Xu 等。FAIR1M：用于高分辨率遥感影像细粒度目标识别的基准数据集。ISPRS摄影测量与遥感学报，184:116–130，2022。

[179] Y. Sun、S. Feng、X. Li、Y. Ye、J. Kang 和 X. Huang。遥感图像中的视觉定位。载于第30届 ACM 多媒体国际会议论文集，页码 404–412，2022。

[180] C. Szegedy、W. Liu、Y. Jia、P. Sermanet、S. Reed、D. Anguelov、D. Erhan、V. Vanhoucke 和 A. Rabinovich。通过卷积网络不断深入。载于 IEEE计算机视觉与模式识别会议论文集，页码 1 9, 2015 3, 7 – 。

[181] C. Szegedy、V. Vanhoucke、S. Ioffe、J. Shlens 和 Z. Wojna。重新思考面向计算机视觉的 Inception 架构。载于 IEEE计算机视觉与模式识别会议论文集，页码 2818 2826, 2016 7 – 。

[182] H. Tan 和 M. Bansal。LXMERT：从 Transformer 中学习跨模态编码器表征。载于2019年自然语言处理经验方法会议暨第九届自然语言处理国际联合会议论文集（EMNLP-IJCNLP），页码 5100 5111, 2019 5 – 。

[183] M. Tan 和 Q. Le。EfficientNet：重新思考卷积神经网络的模型缩放。载于国际机器学习会议，页码 6105 6114 2019 3, 7 – 。PMLR，。

[184] M. Tao、H. Tang、F. Wu、X.-Y. Jing、B.-K. Bao 和 C. Xu。DF-GAN：一种简单有效的文本生成图像基线方法。载于 IEEE/CVF计算机视觉与模式识别会议论文集，页码 16515 16525, 2022 11 – 。

[185] Z. Tian、H. Zhao、M. Shu、Z. Yang、R. Li 和 J. Jia。用于少样本分割的先验引导特征增强网络。IEEE模式分析与机器智能汇刊，44(2):1050– 1065, 2020 20 。

[186] X.-Y. Tong、G.-S. Xia、Q. Lu、H. Shen、S. Li、S. You 和 L. Zhang。利用可迁移的深度模型对高分辨率遥感图像进行土地覆盖分类。《遥感环境》，237 111322, 2020 7, 21 : 。

[187] H. Touvron、M. Cord、M. Douze、F. Massa、A. Sablayrolles 和 H. J´egou。通过注意力机制训练高数据效率的图像 Transformer 并进行蒸馏。载于国际机器学习会议，页码 10347 10357 2021 3 – 。PMLR，。

[188] H. Touvron、T. Lavril、G. Izacard、X. Martinet、M.-A. Lachaux、T. Lacroix、B. Rozi`ere、N. Goyal、E. Hambro、F. Azhar 等。Llama：开放高效的基础语言模型。arXiv 预印本 arXiv: 2302 13971, 2023 22 。。

[189] D. Tuia、R. Roscher、J. D. Wegner、N. Jacobs、X. Zhu 和 G. Camps-Valls。推动形成地球科学数据分析人工智能的集体议程。IEEE地球科学与遥感杂志，9(2):88–104，2021。

[190] A. Van Den Oord、O. Vinyals 等。神经离散表征学习。神经信息处理系统进展，30，2017。

[191] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin。注意力即你所需的一切。神经信息处理系统进展，30, 2017 1, 7 。

[192] R. Vedantam、C. Lawrence Zitnick 和 D. Parikh。Cider：基于共识的图像描述评估。载于 IEEE计算机视觉与模式识别会议论文集，页码 4566–4575, 2015 9 。

[193] Z. Wan、D. Chen、Y. Li、X. Yan、J. Zhang、Y. Yu 和 J. Liao。结合视觉结构约束的传导式零样本学习。神经信息处理系统进展，32, 2019 16 。

[194] B. Wang、X. Lu、X. Zheng 和 X. Li。高分辨率遥感图像的语义描述。IEEE地球科学与遥感快报，16 8 1274 1278, 2019 10 ( ): – 。

[195] B. Wang、Z. Wang、X. Sun、H. Wang 和 K. Fu。DMML-Net：深度用于遥感影像少样本地理目标分割的元度量学习。《IEEE地球科学与遥感汇刊》，60 1 18, 2021 20 : –。

[196] B. Wang、X. Zheng、B. Qu 和 X. Lu。用于遥感图像描述的检索主题循环记忆网络。《IEEE应用地球观测与遥感精选主题期刊》，13 256 270, 2020 10 : –。

[197] C. Wang、G. Peng 和 B. De Baets。用于零样本遥感场景分类的距离约束语义自编码器。《IEEE应用地球观测与遥感精选主题期刊》，14 12545 12556, 2021 2, 7, 15, 16 : –。

[198] D. Wang、Y. Li、Y. Lin 和 Y. Zhuang。零样本学习的关系知识迁移。载于《第三十届AAAI人工智能会议》，2016 7, 15, 16 。

[199] D. Wang、J. Zhang、B. Du、G.-S. Xia 和 D. Tao。遥感预训练的实证研究。《IEEE地球科学与遥感汇刊》，2022 8 。

[200] D. Wang、Q. Zhang、Y. Xu、J. Zhang、B. Du、D. Tao 和 L. Zhang。推动纯视觉Transformer发展为遥感基础模型。《IEEE地球科学与遥感汇刊》，2022年。

[201] K. Wang、J. H. Liew、Y. Zou、D. Zhou 和 J. Feng。Panet：通过原型对齐实现少样本图像语义分割。载于《IEEE/CVF国际计算机视觉会议论文集》，第 9197 9206, 2019 17, 20 –页。

[202] Q. Wang、W. Huang、X. Zhang 和 X. Li。用于遥感图像描述的词语—句子框架。《IEEE地球科学与遥感汇刊》，59 12 10532 10543, 2020 2, 7, 9 ( ): –。

[203] T. Wang、X. Xu、Y. Yang、A. Hanjalic、H. T. Shen 和 J. Song。利用多模态张量融合和重排序实现图像与文本匹配。载于《第27届ACM多媒体国际会议论文集》，第 12 20, 2019 13 –页。

[204] W. Wang、Z. Chen、X. Chen、J. Wu、X. Zhu、G. Zeng、P. Luo、T. Lu、J. Zhou、Y. Qiao 等。VisionLLM：大语言模型也是面向视觉中心任务的开放式解码器。《神经信息处理系统进展》，36, 2024 6 。

[205] W. Wang、E. Xie、X. Li、D.-P. Fan、K. Song、D. Liang、T. Lu、P. Luo 和 L. Shao。金字塔视觉Transformer：一种无需卷积的密集预测通用骨干网络。载于《IEEE/CVF国际计算机视觉会议论文集》，第568– 578, 2021 3 页。

[206] X. Wang、T. E. Huang、T. Darrell、J. E. Gonzalez 和 F. Yu。简单得令人沮丧的少样本目标检测。载于《第37届机器学习国际会议论文集》，第9928–页，2020年。

[207] Y. Wang、C. M. Albrecht、N. A. A. Braham、L. Mou 和 X. X. Zhu。遥感中的自监督学习：综述。《IEEE地球科学与遥感杂志》，10 4 213 247, 2022 8 ( ): –。

[208] Z. Wang、X. Liu、H. Li、L. Sheng、J. Yan、X. Wang 和 J. Shao。Camp：用于文本图像检索的跨模态自适应消息传递。载于《IEEE/CVF国际计算机视觉会议论文集》，第 5764 5773, 2019 13 –页。

[209] S. Waqas Zamir、A. Arora、A. Gupta、S. Khan、G. Sun、F. Shah-baz Khan、F. Zhu、L. Shao、G.-S. Xia 和 X. Bai。iSAID：用于航空图像实例分割的大规模数据集。载于《IEEE/CVF计算机视觉与模式识别会议研讨会论文集》，第 28 37, 2019 17 –页。

[210] J. Wei、M. Bosma、V. Zhao、K. Guu、A. W. Yu、B. Lester、N. Du、A. M. Dai 和 Q. V. Le。微调语言模型是零样本学习者。载于《国际学习表征会议》，2021 22 。

[211] J. Wei、X. Wang、D. Schuurmans、M. Bosma、F. Xia、E. Chi、Q. V. Le、D. Zhou 等。思维链提示激发大语言模型的推理能力。《神经信息处理系统进展》，35 24824 24837, 2022 4 : –。

[212] T. Wei、W. Yuan、J. Luo、W. Zhang 和 L. Lu。VLCA：采用跨模态注意力的视觉语言对齐模型，用于双语遥感图像描述。《系统工程与电子技术》，34 1 9 18, 2023 7, 9 ( ): –。

[213] C. Wu、S. Yin、W. Qi、X. Wang、Z. Tang 和 N. Duan。Visual ChatGPT：利用视觉基础模型进行对话、绘图和编辑。arXiv预印本 arXiv: 2303 04671, 2023 2 。。

[214] G.-S. Xia、J. Hu、F. Hu、B. Shi、X. Bai、Y. Zhong、L. Zhang 和 X. Lu。AID：用于航空场景分类性能评估的基准数据集。《IEEE地球科学与遥感汇刊》，55 7 3965 3981, 2017 7, 21 ( ): –。

[215] Y. Xian、Z. Akata、G. Sharma、Q. Nguyen、M. Hein 和 B. Schiele。用于零样本分类的潜在嵌入。载于《IEEE计算机视觉与模式识别会议论文集》，第69–77页，2016年。7，16

[216] Y. Xiao、V. Lepetit 和 R. Marlet。针对野外目标的少样本目标检测和视点估计。《IEEE模式分析与机器智能汇刊》，45 3 3090 3106, 2022 18 ( ): –。S.

[217] Xie、R. Girshick、P. Dollar、Z. Tu 和 K. He。用于深度神经网络的聚合残差变换。载于《IEEE计算机视觉与模式识别会议论文集》，第1492– 1500, 2017 3 页。

[218] Z. Xie、Z. Zhang、Y. Cao、Y. Lin、J. Bao、Z. Yao、Q. Dai 和 H. Hu。SimMIM：一种用于掩码图像建模的简单框架。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 9653 9663, 2022 8 –页。Y. Xing、S.

[219] Huang、L. Huangfu、F. Chen 和 Y. Ge。用于广义零样本学习的鲁棒双向生成网络。载于ICME，第 1 6 2020 16 –页。IEEE，。Z. Xiong、F.

[220] Zhang、Y. Wang、Y. Shi 和 X. X. Zhu。Earthnets：赋能地球观测领域的人工智能。arXiv预印本 arXiv:2210 04936，2022。。[221] K. Xu、J. Ba、R.

Kiros、K. Cho、A. Courville、R. Salakhudinov、R. Zemel 和 Y. Bengio。展示、关注与讲述：结合视觉注意力的神经图像描述生成。载于《国际机器学习会议》，第 2048 2057 2015 10 –页。PMLR，。M. Xu、

[222] Z. Zhang、F. Wei、Y. Lin、Y. Cao、H. Hu 和 X. Bai。采用预训练视觉语言模型实现开放词汇语义分割的简单基线。载于《计算机视觉—ECCV 2022：第17届欧洲计算机视觉会议，以色列特拉维夫，2022年10月23日至27日，论文集，第XXIX部分》，第 736 753 2022 17 –页。施普林格，。

[223] T. Xu、P. Zhang、Q. Huang、H. Zhang、Z. Gan、X. Huang 和 X. He。AttnGAN：利用注意力生成对抗网络进行细粒度文本到图像生成。载于《IEEE计算机视觉与模式识别会议论文集》，第1316–1324页，2018年。[224] Y. Xu、

W. Yu、P. Ghamisi、M. Kopp 和 S. Hochreiter。Txt2Img-MHN：利用现代霍普菲尔德网络从文本生成遥感图像。《IEEE图像处理汇刊》，2023 2, 7, 11 。Y. Xu、Q.

[225] Zhang、J. Zhang 和 D. Tao。ViTAE：通过探索内在归纳偏置改进视觉 Transformer。《神经信息处理系统进展》，34 28522 28535, 2021 7 : –。H. Xue、Y.

[226] Huang、B. Liu、H. Peng、J. Fu、H. Li 和 J. Luo。探索模态间关系：利用自注意力进行视觉解析，以实现视觉语言预训练。NeurIPS，2021 5 。X. Yan、Z. Chen、A. Xu、X.

[227] Wang、X. Liang 和 L. Lin。Meta R-CNN：面向实例级低样本学习的通用求解器。载于《IEEE/CVF国际计算机视觉会议论文集》，第 9577 9586, 2019 18 –页。L. Yang、Y. Xu、C. Yuan、W. Liu、

[228] B. Li 和 W. Hu。利用视觉语言验证和迭代推理改进视觉定位。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 9499 9508, 2022 16 –页。Y. Yang 和 S. Newsam。

[229] 用于土地利用分类的视觉词袋及其空间扩展。载于《第18届SIGSPATIAL地理信息系统进展国际会议论文集》，第 270 279, 2010 7, 21 –页。Y. Yang 和 S. Newsam。用于图像分类的空间金字塔共生。载于《2011年国际计算机视觉会议》，第 1465 1472 2011 7 –页。IEEE，。Z. Yang、

[230] T. Chen、L. Wang 和 J. Luo。通过递归子查询构造改进单阶段视觉定位。载于《计算机视觉—ECCV 2020：第16届欧洲计算机视觉会议，英国格拉斯哥，2020年8月23日至28日，论文集，第XIV部分16》，第387–404页。施普林格， 2020 16 。

[231] Z. Yang、B. Gong、L. Wang、W. Huang、D. Yu 和 J. Luo。一种快速准确的单阶段视觉定位方法。载于《IEEE/CVF国际计算机视觉会议论文集》，第 4683 4693, 2019 16 –页。X. Yao、Q. Cao、X. Feng、G. Cheng 和 J. Han。用

[232] 于少样本航空图像语义分割的尺度感知精细匹配。《IEEE地球科学与遥感汇刊》，60 1 11, 2021 20 : –。L.

[233] Yuan、Q. Hou、Z. Jiang、J. Feng 和 S. Yan。VOLO：用于视觉识别的视觉观察器。《IEEE模式分析与机器智能汇刊》，2022 3 。Z. Yuan、

[234] L. Mou、Q. Wang 和 X. X. Zhu。从易到难：学习语言引导课程，以在遥感数据上进行视觉问答。《IEEE地球科学与遥感汇刊》，60 1 11, 2022 2, 7, 14, 15 : –。

[235] Z. Yuan、L. Mou、Z. Xiong 和 X. Zhu。变化检测与视觉问答相结合。《IEEE地球科学与遥感汇刊》，60 1 13, 2022 7, 14, 21 : –。Z. Yuan、

[236] L. Mou 和 X. X. Zhu。通过多语言增强提升遥感图像视觉问答的鲁棒性。载于《2023年联合城市遥感会议（JURSE），第 1 4 2023 13 –页。IEEE，。

[237] Z. Yuan、W. Zhang、K. Fu、X. Li、C. Deng、H. Wang 和 X. Sun。探索用于跨模态遥感图像检索的细粒度多尺度方法。《IEEE地球科学与遥感汇刊》，60 3078451, 2022 2, 7, 12, 13 : 。Z. Yuan、W. Zhang、X. Rong、

[238] X. Li、J. Chen、H. Wang、K. Fu 和 X. Sun。遥感领域中的轻量级多尺度跨模态文本图像检索方法。《IEEE地球科学与遥感汇刊》，60 1 19, 2021 2, 7, 12 : –。Z. Yuan、W. Zhang、C. Tian、

[239] X. Rong、Z. Zhang、H. Wang、K. Fu 和 X. Sun。基于全局与局部信息的遥感跨模态文本图像检索。《IEEE地球科学与遥感汇刊》，60 1 16, 2022 2, 7, 12 : –。N. Zabari 和 Y. Hoshen。不看任

[240] 何分割示例，在真实场景中进行语义分割。arXiv预印本 arXiv:2112 03185，2021。。

[241] Y. Zhan、Z. Xiong 和 Y. Yuan。RSVG：探索用于遥感数据视觉定位的数据与模型。《IEEE地球科学与遥感汇刊》，61 1 13, 2023 7, 15, 16, 21 : –。B. Zhang、

[242] J. Xiao 和 T. Qin。用于少样本分割的自引导与交叉引导学习。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第8312– 8321, 2021 20 页。C. Zhang、G.

[243] Lin、F. Liu、R. Yao 和 C. Shen。CANet：结合迭代细化与注意力少样本学习的类别无关分割网络。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 5217 5226, 2019 20 –页。F. Zhang、B.

[244] Du 和 L. Zhang。用于场景分类的显著性引导无监督特征学习。《IEEE地球科学与遥感汇刊》，53 4 2175 2184, 2014 7, 21 （）: –。L. Zhang、

[245] L. Zhang 和 B. Du。遥感数据的深度学习：关于最新技术水平的技术教程。《IEEE地球科学与遥感杂志》，4 2 22 40, 2016 1 （）: –。

[246] P. Zhang、X. Dai、J. Yang、B. Xiao、L. Yuan、L. Zhang 和 J. Gao。多尺度视觉Longformer：一种用于高分辨率图像编码的新型视觉Transformer。载于《IEEE/CVF国际计算机视觉会议论文集》，第2998–3008页，2021年。[248] S.

[247] Zhang、F. Song、X. Liu、X. Hao、Y. Liu、T. Lei 和 P. Jiang。用于遥感图像少样本目标检测的文本语义融合关系图推理。《遥感》，15（5）：1187，2023年。2, 7, [249] T. Zhang、X. Zhang、P. Zhu、X. Jia、X. Tang 和 L. Jiao。遥感图像中的广义少样本目标检测。《ISPRS摄影测量与遥感学报》，195 353 364, 2023 17 : –。X. Zhang、

[250] X. Li、J. An、L. Gao、B. Hou 和 C. Li。基于深度学习的遥感图像自然语言描述。载于《2017年IEEE国际地球科学与遥感研讨会（IGARSS）》，第 4798 4801 2017 2, 7, 9 –页。IEEE，。X. Zhang、X.

[251] Wang、X. Tang、H. Zhou 和 C. Li。基于属性注意力机制的遥感图像描述生成。《遥感》，11 6 612, 2019 2, 7, 9 （）: 。Y. Zhang、M.

[252] Zhang、W. Li、S. Wang 和 R. Tao。用于跨场景高光谱图像分类的语言感知域泛化网络。《IEEE地球科学与遥感汇刊》，2023 2, 7, 17, 19 。Z. Zhang、L. Ji

[253] ao、L. Li、X. Liu、P. Chen、F. Liu、Y. Li 和 Z. Guo。用于遥感视觉问答的空间层次推理网络。《IEEE地球科学与遥感汇刊》，61 1 15, 2023 14, 15 : –。

[254] Z. Zhang、W. Zhang、W. Diao、M. Yan、X. Gao 和 X. Sun。VAA：用于遥感图像描述的视觉对齐注意力模型。《IEEE Access》，7 137355 137364, 2019 7, 9 : –。Z. Zhang、T. Zhao、Y. Guo

[255] 和 J. Yin。RS5M：用于遥感视觉语言基础模型的大规模视觉语言数据集。arXiv预印本 arXiv: 2306 11300, 2023 2, 21 。。R. Zhao 和 Z. Shi。

[256] 使用结构化生成对抗网络进行文本生成遥感图像。《IEEE地球科学与遥感快报》，19 1 5, 2021 2, 7, 11 : –。R. Zhao、Z. Shi 和 Z. Zou。

[257] 基于结构化注意力的高分辨率遥感图像描述生成。《IEEE地球科学与遥感汇刊》，60 1 14, 2021 2, 7, 9 : –。Z. Zhao、P. Tang、

[258] L. Zhao 和 Z. Zhang。通过两阶段微调实现遥感图像少样本目标检测。《IEEE地球科学与遥感快报》，19 1 5, 2021 17, 18 : –。X.

[259] Zheng、B. Wang、X. Du 和 X. Lu。用于遥感视觉问答的互注意力Inception网络。《IEEE地球科学与遥感汇刊》，60 1 14, 2021 2, 7, 13, 21 : –。Y. Zheng、

[260] R. Huang、C. Han、X. Huang 和 L. Cui。背景

用于零样本目标检测的可学习级联。在《亚洲计算机视觉会议论文集》中，2020 17 。C. Zhou、C. C. Loy 和 B. Dai。从CLIP中提取自由密集标签。载于《计算机视觉——ECCV 2022：第17届欧洲会议，特拉维夫，以色列，2022年10月23日至27日，论文集，第XXVIII部分》，第696– 712 2022 17 页。施普林格，。Y. Zhou、R. Zhang、C. Chen、C. Li、C. Tensmeyer、T. Yu、J. Gu、J. Xu 和 T. Sun。迈向文本到图像生成的无语言训练。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第17907 17917, 2022 11 –页。Z. Zhou、J. Chen、Z. Huang、H. Wan、P. Chang、Z. Li、B. Yao、B. Wu、L. Sun 和 M. Xing。Fsods：一种用于合成孔径雷达图像少样本目标检测的轻量级元学习方法。《IEEE地球科学与遥感汇刊》，60 1 17, 2022 17 ： –页。Z. Zhou、M. M. Rahman Siddiquee、N. Tajbakhsh 和 J. Liang。Unet++：一种用于医学图像分割的嵌套U-Net架构。载于《医学图像分析中的深度学习与临床决策支持的多模态学习：第4届国际研讨会，DLMIA 2018，以及第8届国际研讨会，ML-CDS 2018，与MICCAI 2018联合举办，20，2018，4，3 11 2018格拉纳达，西班牙，9月，论文集，第–页。施普林格，。[265] D. Zhu、J. Chen、K. Haydarov、X. Shen、W. Zhang 和 M. Elhoseiny。ChatGPT提问，BLIP-2作答：通过自动提问丰富视觉描述。arXiv预印本 arXiv:2303 06594, 2023 2 。。D. Zhu、J. Chen、X. Shen、X. Li 和 M. Elhoseiny。MiniGPT-4：利用先进的大型语言模型增强视觉语言理解。载于《第十二届国际学习表征会议》，2023 6, 18, 22 。X. Zhu、W. Su、L. Lu、B. Li、X. Wang 和 J. Dai。可变形DETR：用于端到端目标检测的可变形Transformer。载于《国际学习表征会议》，2020 17 。[268] X. X. Zhu、D. Tuia、L. Mou、G.-S. Xia、L. Zhang、F. Xu 和 F. Fraundorfer。遥感中的深度学习：全面综述与资源列表。《IEEE地球科学与遥感杂志》，5 4 8 36, 2017 1 （）： –。X. X. Zhu、Y. Wang、M. Kochupillai、M. Werner、M. H¨aberle、E. J. Hoffmann、H. Taubenb¨ock、D. Tuia、A. Levering、N. Jacobs、A. Kruspe 和 K. Abdulahhad。从社交媒体数据中采集地理信息：一种社区遥感方法。《IEEE地球科学与遥感杂志》，10 4 150 180, 2022 22 （）： –。U. Zia、M. M. Riaz 和 A. Ghafoor。将遥感图像转换为文本描述。《国际应用地球观测与地理信息学杂志》，108 102741, 2022 2, 7, 9 ： 。Q. Zou、L. Ni、T. Zhang 和 Q. Wang。基于深度学习的遥感场景分类特征选择。《IEEE地球科学与遥感快报》，12（11）：2321–2325，2015年。7