遥感场景的图文跨模态理解研究进展

郑向涛[1]，赵郑营[2,3]，宋宝贵[1]，李浩[4]，卢孝强[1]

1. 福州大学 物理与信息工程学院, 福州 350108;
2. 西安理工大学 计算机科学与工程学院, 西安 710048;
3. 平顶山学院 软件学院, 平顶山 467000;
4. 空军预警学院, 武汉 430019

摘 要：随着遥感技术和人工智能的深度融合，人类对遥感数据的应用需求日益精细化。然而，单一模态数据在复杂场景解译中存在局限性，难以充分挖掘遥感图像中的深层信息。为此，多模态数据协同分析成为提升遥感解译能力的关键途径，并推动着遥感领域的进一步发展。图文跨模态理解通过文本描述建立遥感图像和人类认知的联系，借助文本语义信息增强视觉特征表征，实现跨模态信息互补，显著提升了遥感解译的性能。本文以遥感图文跨模态理解为主线，将遥感图文跨模态理解划分为遥感图像描述、文本生成图像、遥感图文对齐和遥感图像问答4个任务。首先概述了国内外图文跨模态研究的发展状况；然后对遥感图文跨模态理解常用的公开数据集和评价指标进行介绍；最后，总结了遥感图文跨模态理解面临的技术挑战，并对未来研究方向进行展望。

关键词：遥感图文跨模态，图像描述，文本生成图像，图文对齐，图像问答，遥感跨模态数据集

中图分类号：TP751/P2

引用格式：郑向涛,赵郑营,宋宝贵,李浩,卢孝强.2025.遥感场景的图文跨模态理解研究进展.遥感学报,29(6): 1566-1586

Zheng X T, Zhao Z Y, Song B G, Li H and Lu X Q. 2025. Advances in remote sensing image-text cross-modal understanding. National Remote Sensing Bulletin, 29(6):1566-1586[DOI:10.11834/jrs.20255125]

1 引 言

遥感图像是通过卫星或航空器等平台，获取地表反射或辐射信号的数字化影像。这些影像数据能够反映遥感场景的复杂信息，既包含颜色、纹理、光谱等像素级信息（唐振超 等，2023），又存在形状、轮廓、结构等目标级信息（张涛 等，2022），还包括地物组成、目标分布等场景信息（邓培芳 等，2021），展现出多层级的特性。然而，现有研究大多是对遥感图像进行像素、目标、场景的单层级信息解译。例如语义分割（Yuan等，2021；于纯妍 等，2024）对图像中地物进行逐个像素的识别，为图像中的每个像素分配地物类别标签，属于像素级解译。目标检测（Wang等，2022；徐丹青和吴一全，2024）通过算法自动发现图像中的感兴趣目标，获取目标的类别和位置信息，属于目标级解译。而场景分类（Cheng等，2020；李智 等，2024）侧重对遥感场景的整体认知，为每个场景图像分配不同的场景类别，属于场景级解译。现有单层级信息解译任务简化了图像内容的理解过程，仅提取感兴趣的信息，忽略了图像本身呈现的多层级信息特点。文本是人类记录和传递信息的主要方式，它通过定义特定的符号来描述认知的世界内容，不同文本描述可以勾勒出图像蕴含的各种内容，方便认知解析遥感场景中的多层级信息。因此，如何利用文本信息辅助图像对遥感场景进行理解变得尤为重要。

为弥补遥感图像单模态分析的不足，近年来学者们尝试联合遥感图像和文本模态进行遥感场景理解研究，探索遥感图文跨模态结合的新方向。遥感图文跨模态理解RS-ICU（Remote sensing image-text cross-modal understanding）是通过神经网络等人工智能技术识别遥感场景中的地物成分，建立遥感图像和语言文本的跨模态联系，填补遥

收稿日期：2025-04-08；**预印本**：2025-04-28

基金项目：国家自然科学基金(编号:62271484,61925112);陕西省重点研发计划(编号:2023-YBGY-225)

第一作者简介：郑向涛,研究方向为遥感图像处理。E-mail：zhengxiangtao@fzu.edu.cn

通信作者简介：卢孝强,研究方向为智能光学感知。E-mail：luxiaoqiang@fzu.edu.cn

感图像与人类认知之间的语义鸿沟，提高遥感场景解译的智能化水平。本文通过检索Web of Science（WOS）和中国知网，查询遥感图文跨模态理解相关论文的发表情况，检索关键词为“remote sensing caption”、“remote sensing image text retrieval”、“remote sensing text to image generation”、“remote sensing image text alignment”、“remote sensing visual question answering”、“遥感图像描述生成”、“遥感图文检索”、“遥感图像生成”、“遥感图文对齐”、“遥感图像问答”等，检索结果如图1所示。从近五年文献变化趋势来看，遥感图文跨模态理解文献数量虽然较少，但呈现出逐年增加的趋势，说明图文跨模态理解逐渐受到遥感领域相关学者的重视。

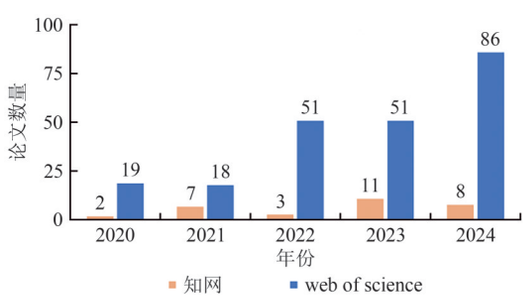


图1 近5年遥感图文跨模态理解文献数量统计

Fig. 1 Statistics of literatures on cross-modal understanding of remote sensing image-text in recent five years

国内外相关学者针对图文跨模态理解已展开大量研究。赵佳琦等（2021）提出了一种基于多尺度与注意力特征增强的遥感图像描述生成方法。Xia等（2024）全面概述了跨模态图文检索任务的研究进展，阐明图像和文本的多模态特性，并对广泛采用的图文检索技术和评价指标进行了比较分析。Lin等（2022）提出了一种融合上下文信息和统计知识的遥感影像场景图生成模型，基于可调Transformer结构来集成上下文信息并计算所有对象之间的注意力。Chen等（2024）提出一种多尺度显著遥感图像引导文本对齐方法，将文本与遥感图像对齐来学习显著信息，并设计一种损失函数增强不同模态之间的相似性，从而提高跨模态检索的效率。Yusuf等（2024）从图神经网络的角度对图像问答进行研究，介绍了图像问答的各种图神经网络模型，并对模型之间的性能进行比较，最后针对图像问答领域未来研究中仍存在的挑战提出一些建议。

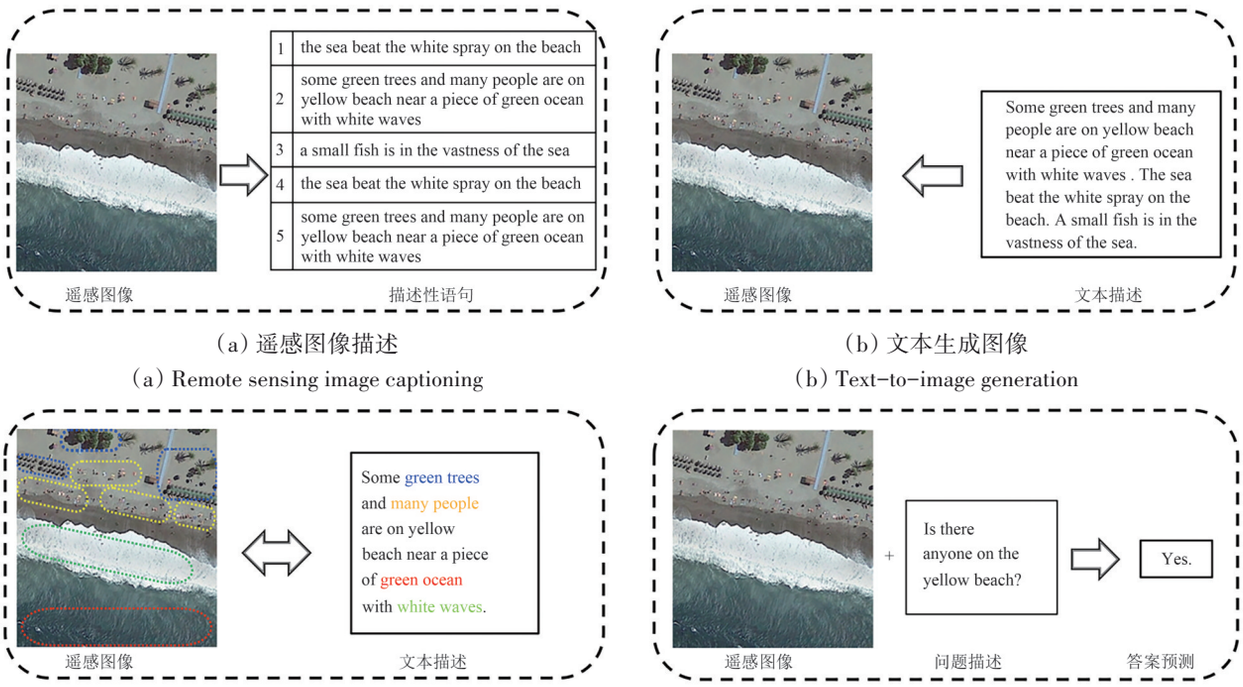
根据图像—文本的映射关系，本文从遥感图文跨模态理解角度进行概述，将现有的遥感图文跨模态研究划分为4个子任务：遥感图像描述（简称，图生文）、文本生成图像（简称，文生图）、遥感图文对齐（简称，图文对齐）和遥感图像问答（简称，图文对话）。如图2所示，从图文跨模态转化角度来看，遥感图文跨模态理解包含图生文和文生图两类任务，而在局部内容交互方面，遥感图文跨模态理解包含图文对齐与图文对话两类任务。首先，针对以上4类任务分别阐述相关研究进展，总结现有方法的解决思路，提出现有任务面临的挑战问题；然后，介绍遥感图文跨模态理解常用的公开数据集与评价指标；最后，总结遥感图文跨模态理解的未来发展趋势。

2 遥感图文跨模态理解

根据遥感领域图文跨模态研究的内容，本节概述了遥感图文跨模态理解的4个任务。已知遥感图像集合$I=\{i_1,i_2,\cdots,i_m\}$，i_j表示第j个遥感图像。文本集合$T=\{t_1,t_2,\cdots,t_n\}$，t_j表示第j个自然语言描述。图像问答中的问题集合记作$Q=\{q_1,q_2,\cdots,q_k\}$，q_j表示第j个问题，答案集合记作$A=\{a_1,a_2,\cdots,a_k\}$，a_j表示第j个问题对应的答案。通过定义图像I和文本$T/Q/A$之间的映射关系可以描述跨模态的四个任务，如表1所示。

2.1 遥感图像描述任务

遥感图像描述（图生文）任务通过根据遥感图像I生成反映遥感场景内容的描述性语句T，其映射关系为：$I\rightarrow T$。该描述语句力求简洁明了语法正确，能够反映遥感场景不同尺度下的地物成分，还能够描述这些地物之间的空间关系和属性。图生文任务可以帮助人类理解遥感图像中复杂地物信息，在灾害快速评估（如地震、洪涝等灾害的自动描述）、土地利用/覆盖监测（自动提取地物类型和分布）以及大规模遥感影像数据标注中具有重要应用价值。



（a）遥感图像描述

（a）Remote sensing image captioning

（b）文本生成图像

（b）Text-to-image generation

（c）遥感图文对齐

（c）Remote sensing image-text aligment

（d）遥感图像问答

（d）Remote sensing visual question answering

图2 遥感图文跨模态理解任务

Fig. 2 Cross-modal understanding task for remote sensing image-text

表1 遥感图文跨模态理解任务符号定义

Table 1 Symbol definition of remote sensing image-text cross-modal understanding tasks

任务类别	输入模态	输出模态	任务目标	映射关系
图生文	图像I	文本T	从图像I生成文本T	$I\rightarrow T$
文生图	文本T	图像I	根据文本T生成图像I	$T\rightarrow I$
图文对齐	图像I,文本T	对齐关系	判别图像I和文本T是否对齐	$(I,T)\rightarrow\{0,1\}$
图文对话	图像I,问题Q	答案A	结合图像I和问题Q生成答案A	$(I,Q)\rightarrow A$

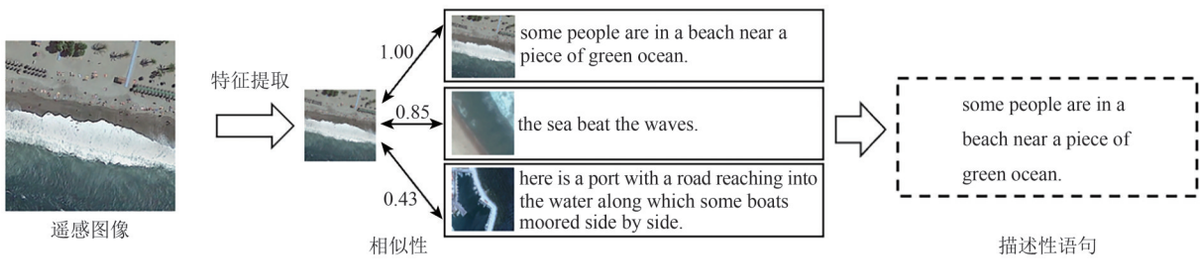
如图3所示，图生文任务主要包含3个步骤，首先，使用卷积神经网络CNN（Convolutional Neural Networks）或者Transformer从遥感图像中提取特征并将其编码为高维特征向量。然后，将图像特征转换为能够反映图像场景内容的文本特征。最后，将文本特征生成与图像相关的描述型语句。根据文本映射方式的不同，现有图生文方法主要分为以下3类：基于检索匹配的方法、基于语句模板的方法和基于编码—解码结构的方法。

2.1.1 基于检索匹配的方法

基于检索匹配的方法根据特征相似度度量搜索与图像特征最相似的图像，然后将检索到图像对应的描述分配给查询图像，生成文本描述。

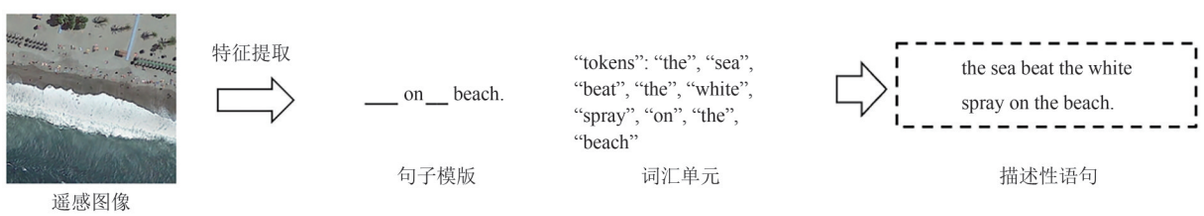
Ordonez等（2011）提出了一种基于大规模图像—文本数据集的非参数化检索方法，通过在共享语义空间中计算图像相似度，检索最匹配的图像及其文本描述，实现了自动图像描述生成。Wang等（2020）提出一种基于检索的递归记忆网络，通过建立主题库以记录训练数据集中的主题词，然后从主题库中检索图像的主题词实现语句描述的生成。Yuan等（2022b）提出一种轻量级的检索模型，通过知识提炼监督对方法进行优化，可以挖掘不同模态之间的关联，以提取最有价值的语义特征。

虽然基于检索匹配的方法生成的描述在语法上是正确的，但是这种方法仅限于根据现有示例生成描述，无法生成新的描述。因此，此类方法不能适应新的物体或者场景组合，甚至在某些特殊情况下，生成的描述可能与图像内容无关。



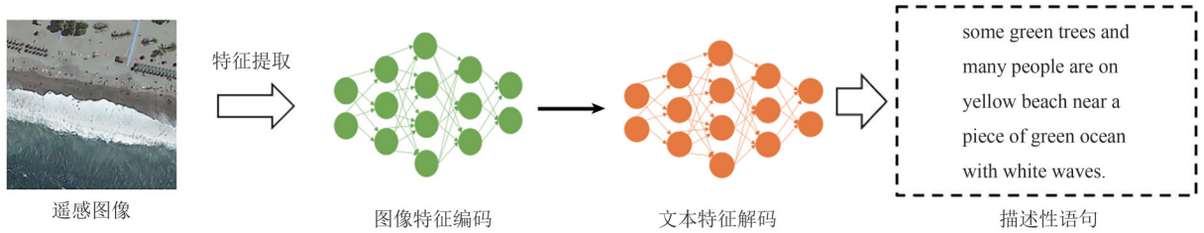
（a）基于检索匹配的方法

（a）Retrieval matching methods



（b）基于语句模版的方法

（b）Sentence template methods



（c）基于编码—解码结构的方法

（c）Encoder-decoder structure methods

图3 遥感图像描述的3种方法框架

Fig. 3 Three method frameworks for remote sensing image captioning

2.1.2 基于语句模版的方法

基于语句模板的方法是通过预定义句式来生成文本描述。根据提取的遥感图像特征，映射到预先设计的语句模板中，与模板中的占位符进行匹配，生成符合语法和语义的自然语言描述。

Mitchell等（2012）通过利用统计的语法信息驱动词汇的出现频率，过滤和约束来自视觉系统的噪声检测结果，从而生成类似人类的图像描述。Ivasic-Kos等（2016）提出一种基于模板的图像标注模型，通过目标级别和场景级别的注释，弥合图像特征和图像描述关键词之间的语义差距。Shi和Zou（2017）基于卷积神经网络提出一个遥感图像描述框架，该框架通过捕获对象、属性和关系的关键词，以较快速度生成鲁棒、全面的句子描述。

与基于检索匹配的方法相比，基于语句模板的方法生成的语句描述更符合图像内容，但是其高度依赖预定义的模板，往往会产生简单且重复的句子，使得此类方法生成的结果缺乏多样性和灵活性。

2.1.3 基于编码—解码结构的方法

基于编码—解码结构的方法由图像编码器和文本解码器两部分组成。图像编码器用于提取遥感图像的语义特征，而文本解码器通过解码图像特征生成语句描述。此类方法不仅可以提高生成描述的质量，而且可以增加生成结果的多样性，在图生文任务中被广泛使用。

Qu等（2016）首次将卷积神经网络和循环神经网络集成在遥感图像描述任务中，发布了UCM-Captions和Sydney-Captions两个遥感图文数据集，提高了遥感图像描述生成的多样性并验证了数据的合理性。Lu等（2018）在编码器—解码器架构中添加了额外的注意力机制，该架构通过解码器在生成文本单词时将感受野缩小到特定区域，进

一步提高生成描述的质量，并发布了一个大规模的图生文数据集RSICD。Du等（2023）基于语义分割提取的前景和背景，设计一种概念层次结构。该结构包括原始网格状特征图、背景、前景3层，每一层均有独特的编码方式，有效缓解了现有方法中遥感图像存在前景和背景信息容易混淆的问题。Kandala等（2022）提出一种基于Transformer的遥感图像描述模型，利用多标签分类作为补充任务，通过提供额外监督来缓解训练数据缺乏导致训练性能下降的问题。Hoxha和Melgani（2022）针对基于循环神经网络的框架需要大量带注释的样本引起高昂算力的问题，提出一种基于支持向量机的解码器，将图像信息直接解码为描述性语句。

基于编码—解码结构的方法可以弥补基于检索匹配和语句模板方法的不足，取得了较好的表现，已成为遥感图像描述任务的主流方法。

2.1.4 挑战与展望

随着编码—解码结构和注意力机制等深度学习技术的发展，遥感图像描述任务生成的语句质量得到了改善。然而，仍存在一些问题。

（1）图像描述的主观性：遥感图像覆盖范围广、信息量丰富，由于关注角度的不同，对同一幅图像的解读存在显著差异，从而使得图像描述具有较强的主观性。这种主观性不仅体现在解读者对目标区域的关注点和语义表达方式上，还对任务结果的多样性产生了深远影响。

（2）定制化描述生成：不同遥感应用领域对图像的需求和关注点存在显著差异。为满足用户的差异性需求，生成具有针对性和实用性的语句描述，可以引入遥感领域的先验知识和常识推理，提升模型的推理能力，使其能够根据具体应用场景生成定制化的描述。

（3）设计合理评价指标：如何评价生成描述性语句的准确性难度较大，现有的评价指标主要基于生成描述与参考描述之间的重叠度，忽略了遥感图像中蕴含的丰富语义信息，难以衡量语义相同但描述不同的语句。可以将遥感图像的语义信息考虑在内，提出新的图像描述评价指标，使得生成的语句描述更加贴近人类的理解。

2.2 文本生成图像任务

文本生成图像（文生图）任务根据给定的文本描述T，生成能够准确表示此描述的遥感图像I，其映射关系为：$T \rightarrow I$。该任务通过语义指导图像内容的生成，确保生成图像与文本描述高度一致。文生图可以解决数据获取困难和高成本问题，在城市规划模拟（生成未来城市景观图）、环境变化预测（模拟植被覆盖变化或污染扩散图）、数据增强（扩展遥感数据集用于模型训练）等方面具有潜在的应用价值。如果能够生成足够真实的遥感图像，就能以可控且低成本的方式构建遥感数据，这有助于挖掘深度学习在遥感领域的潜力。

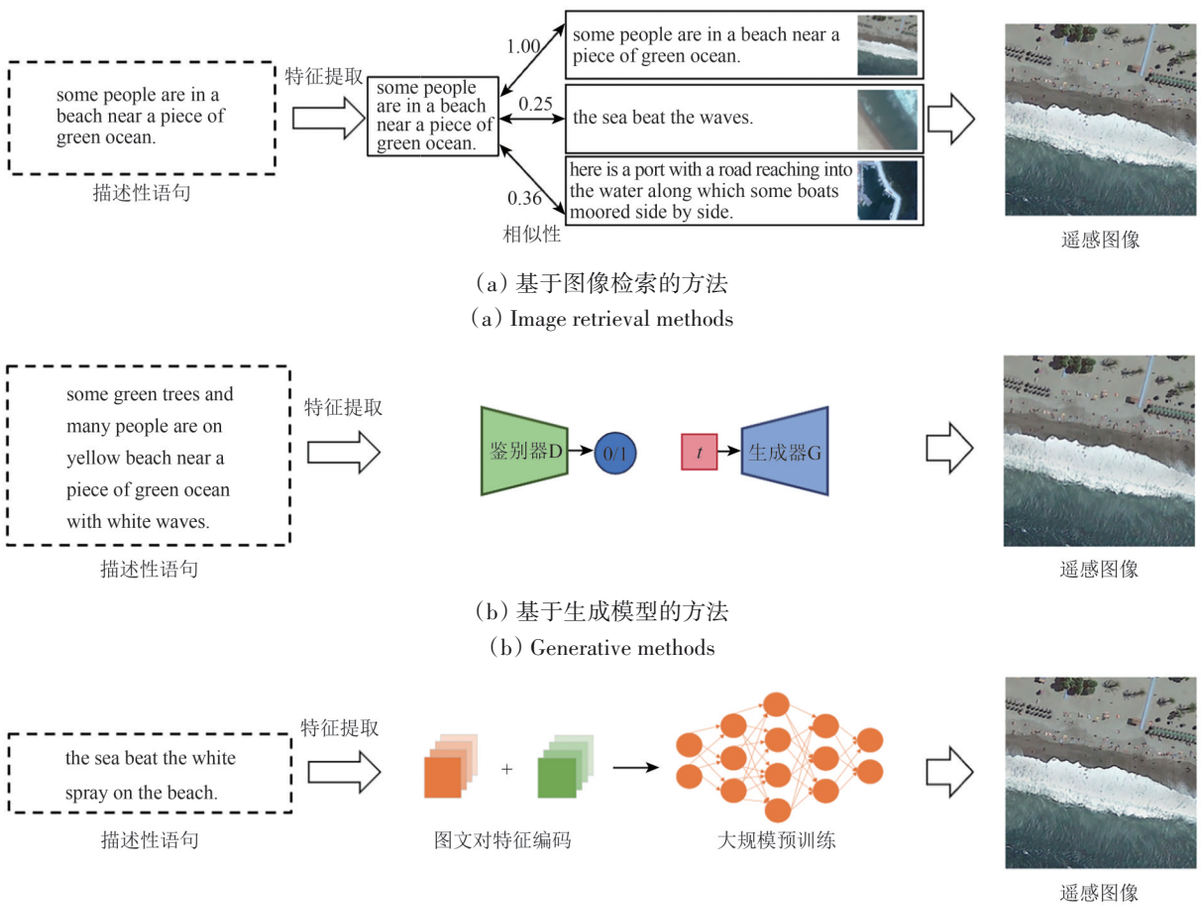
如图4所示，文生图任务主要包含3个步骤，首先，通过文本特征提取学习丰富的特征表示，以捕获文本中蕴含的属性、位置等信息。然后，将文本特征转换为等效的遥感图像特征。最后，将图像特征生成符合文本语义的遥感图像。根据遥感图像的映射过程不同，将现有文生图方法分为以下3类：基于图像检索的方法、基于生成模型的方法以及基于语言大模型的方法。

2.2.1 基于图像检索的方法

基于图像检索的方法通过对文本进行特征提取，以关键词或短语特征作为查询条件，从数据库中检索与这些条件相关的文本，根据文本找到对应的图像。

Zhu等（2007）提出一种文本到图像的生成系统，通过集成自然语言处理、计算机视觉和机器学习，根据信息丰富的文本单元检索出可能性最大的图像，然后在文本的基础上对图片的分布进行优化。Lee等（2023）提出一种高效的文本生成图像方法，通过人类反馈改善图像与文本之间的对齐。该方法首先利用文本提示生成多样化的图像，然后通过训练一个奖励函数，以预测人类反馈。最后，通过优化和微调奖励函数，使得生成的图像能够呈现特定的颜色、数量和背景。

基于图像检索的方法主要关注输入文本描述与数据库中已有文本描述的语义相似性，并通过文本间匹配间接完成图文对齐。尽管生成的遥感图像通常具有较高的真实性，但由于过度依赖现有的数据库，其生成图像的多样性较低，且可能出现检索结果与文本不匹配的情况。



（a）基于图像检索的方法

（a）Image retrieval methods

（b）基于生成模型的方法

（b）Generative methods

（c）基于语言大模型的方法

（c）Large language methods

图4　文本生成图像的3种方法框架

Fig. 4　Three method frameworks for remote sensing text-to-image generation

2.2.2　基于生成模型的方法

基于生成模型的方法将提取的文本特征作为条件信息，对图像的生成过程进行指导，以学习一个从文本到图像的映射模型。根据映射模型的差异，可以将其划分为基于生成对抗网络和基于扩散模型两类方法。

（1）基于生成对抗网络的方法。基于生成对抗网络GAN（Generative Adversarial Network）的方法包含生成器和判别器两个部分。生成器用于生成与文本描述相关的遥感图像，判别器用于识别真实图像和生成图像的差异。Zhang等（2017）提出了堆叠生成对抗网络，通过堆叠多个不同尺度的GANs，实现文本到图像的生成。Xu等（2018）在GAN的基础上提出一种注意力生成网络，在合成图像时关注描述中的相关词汇和图像子区域，实现文本到图像的细粒度生成。然而，基于生成对抗网络的方法在遥感图像生成领域仍然面临巨大的挑战。Bejiga等（2019）提出基于GAN的文本到遥感图像生成方法，该方法可以实现从古代地理区域的文本描述到低空间分辨率的灰度遥感图像的生成。在随后的研究中，Bejiga等（2020，2021）通过预训练的Doc2Vec编码器改进文本编码，使得网络能够利用输入文本中不同层次的信息。但是，这些方法生成的灰度遥感图像空间分辨率较低，会造成图像细节信息的丢失。针对此问题，Zhao和Shi（2022）提出一种多阶段结构化生成对抗网络，利用无监督分割模块提取结构信息，使判别器以结构化的方式区分图像，最后利用多阶段框架生成分辨率逐渐提高的遥感图像。Xu等（2023）提出一种文本到遥感图像的生成网络，通过使用自回归方法获得预测的图像标记，然后将生成的图像标记输入到图像解码器中，以此生成更真实的遥感图像。

虽然基于生成对抗网络的方法可以生成较高质量的遥感图像，但是由于缺乏空间和语义的约束，此类方法生成的图像往往不可控，使得与原始图像的分布存在显著差异。

（2）基于扩散模型的方法。基于扩散模型的方法包含两个阶段：正向扩散和反向生成。在正向扩散阶段，通过在输入图像上逐步添加高斯噪声将其转化为无结构的噪声图像。在反向生成阶段，通过逐步学习逆扩散过程，将噪声图像恢复为原始输入图像或者生成新的图像。此类方法因生成图像质量高、表现多样而备受认可。

Rombach等（2022）通过在模型中引入交叉注意力，将扩散模型转变为强大且灵活的生成器，从而以卷积方式实现高分辨率图像的生成。Sebaq和ElHelw（2024）提出一种基于两阶段扩散模型的轻量级方法，通过文本描述逐步生成高分辨率的遥感图像。Tang等（2024）利用扩散模型的优势，引入一种条件控制机制，实现多尺度特征融合，从而增强控制条件的引导效果，具有更出色的遥感图像生成能力。

与基于生成对抗网络的方法相比，基于扩散模型的方法具有更好的可解释性、多样性和灵活性。然而，此类方法存在迭代采样速度慢且过于依赖噪声的问题，不仅降低了训练和预测效率，还使其容易受到噪声干扰。

2.2.3 基于语言大模型的方法

基于语言大模型的方法通过利用大规模的图像—文本对训练模型，实现文本到图像的转化。此类方法具有强大的生成能力，特别是在开放领域场景的零样本设置中，与基于生成对抗网络的方法中小规模数据形成鲜明对比。

OpenAI的DALL-E（Ramesh等，2021）中利用强大的Transformer模型和丰富的训练数据，将语义概念映射到像素级，生成高质量的图像。Yu等（2022）提出一种自回归模型，将文本到图像生成视为一个序列到序列的建模问题，通过扩大数据和模型规模，利用现有的大语言模型，合成真实且内容丰富的图像。

基于语言大模型的方法可以封装更复杂的语义，增强生成图像的真实感。然而大模型需要大量的数据支撑，这往往需要过多的人力和物力资源构建数据集。因此，将语言大模型应用于文本生成图像任务仍有待进一步探索。

2.2.4 挑战与展望

尽管文本生成图像任务在遥感领域取得显著进展，然而由于遥感图像具有覆盖范围广、场景信息复杂等特点，生成与给定文本描述一致的遥感图像仍面临诸多挑战，特别是在真实性和结构合理性方面。

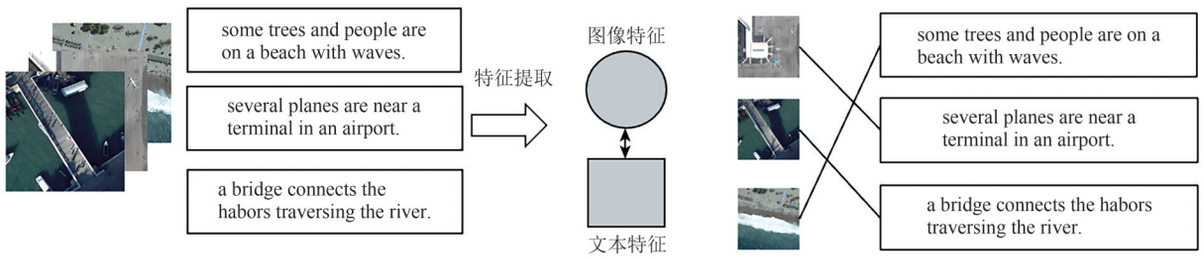
（1）生成图像的不确定性：遥感图像的场景分布差异较大，且对内容细节要求极高。根据文本描述生成遥感图像，需要同时满足语义表达的准确性和遥感场景的空间分布特性，这种双重约束显著增加了生成过程的不确定性。因此，如何在保证语义一致性的前提下，准确还原遥感图像特有的分布特征，是该领域面临的核心技术挑战。

（2）引入遥感先验知识：遥感图像的获取受到成像平台、成像载荷、大气扰动等条件的影响。为了提高生成遥感图像的可靠性，可以将遥感成像相关的物理知识（如传感器类型、地物反射特性、大气散射和吸收等）作为先验信息加以融合。从而有效约束生成过程，显著提升生成图像的真实性与可靠性，使生成结果更加贴合实际的遥感场景。

2.3 遥感图文对齐任务

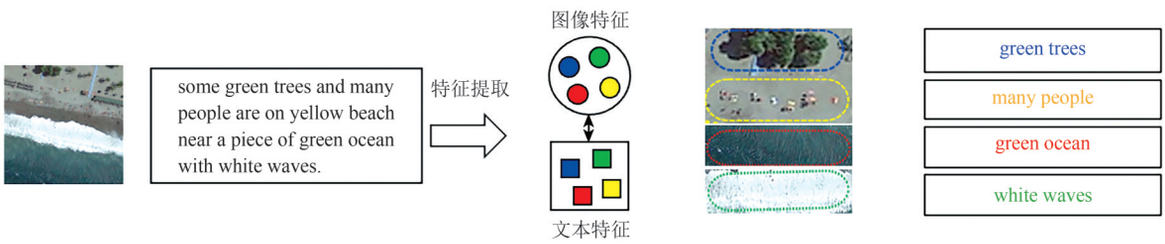
遥感图文对齐（图文对齐）根据图像I和文本描述T内容的不同，建立图像内容和单词语句的语义对应关系，其映射关系为：$(I, T) \rightarrow \{0, 1\}$。图文对齐任务通过特征提取将图像和文本映射到相同的特征空间进行相似性度量，让模型学习到不同模态间的相互表示，进而提高图文跨模态理解的性能，为遥感数据检索（基于语义描述快速搜索图像）、区域资源分布分析（例如矿产资源或农田识别）和自动环境报告生成（将遥感图像内容与文本报告进行匹配）等提供技术支持。

如图5所示，遥感图文对齐主要包含3个步骤，首先，通过特征提取器分别提取遥感图像和文本描述的特征。然后，将两种模态特征映射到共享特征空间计算相似性矩阵。最后，利用相似性结果构建遥感图像和文本描述的对齐关系。根据图文对齐内容的不同，现有图文对齐方法主要分以下3类：基于全局模态对齐的方法、基于局部目标对齐的方法和基于上下文关系对齐的方法。



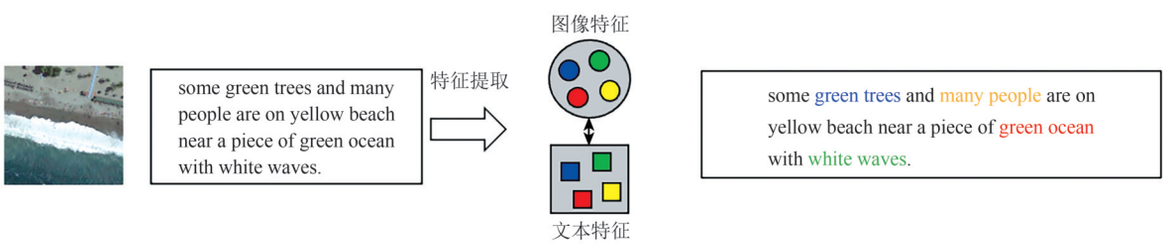
（a）基于全局模态对齐的方法

（a）Global modal alignment methods



（b）基于局部目标对齐的方法

（b）Local target alignment methods



（c）基于上下文关系对齐的方法

（c）Contextual relationship alignment methods

图5　遥感图文对齐的3种方法框架

Fig. 5　Three method frameworks for remote sensing image-text alignment

2.3.1　基于全局模态对齐的方法

基于全局模态对齐的方法从全局视角出发，将图像和文本两种跨模态实例作为整体对象进行对齐。

Faghri 等（2018）提取图像和文本的全局特征，并提出一种新的损失函数，通过显式引入硬负样本（hard negatives）优化嵌入学习，有效挖掘图像和文本之间的潜在语义关联。Zheng 等（2020）基于卷积神经网络提出一种双路径网络，并设计一个分类损失，将图像和文本嵌入到一个共享的视觉—文本空间中，从而挖掘不同模态间的细微差异。Cheng 等（2021）设计了一个语义对齐模块，通过使用注意力与门控机制筛选并优化数据特征，从而获得更具区分性的特征表示。

基于全局模态对齐的方法通常考虑整个图像和句子之间的对应关系，往往忽略了图像区域与文本内单词、句子之间的潜在关系。因此，对于涉及复杂场景中的真实情况，此类方法的跨模态对齐性能往往难以令人满意。

2.3.2　基于局部目标对齐的方法

基于局部目标对齐的方法通过提取图像的局部特征以及文本的单词特征，实现遥感图像局部区域和文本单词语义的对齐。

Lee 等（2018）首次将目标检测方法引入到图像—文本检索任务中，并揭示了图像区域与文本片段之间存在潜在的对齐关系，这一发现激发了大量关于图像与文本之间细粒度对齐关系的研究工作。Sun 等（2024）提出了一种基于 Transformer 的语义对齐算法，通过引入语义对齐约束，从细粒度的角度处理图像和文本信息，实现了更有效的图像与文本特征相似性捕捉。Abdullah 等（2020）提出了一种双向三元组深度网络，用于文本与遥感图像的对齐，采用平均融合策略来融合图像与

文本的相关特征，实现图像和文本的细粒度对齐。Zia等（2022）为了应对遥感图像因尺度和视角变化带来的巨大差异，通过提取多尺度视觉特征来描述不同尺度下的物体或场景，然后利用自适应多头注意力动态衡量多尺度图像特征对下一个单词预测的贡献，以生成细粒度描述。

基于局部目标对齐的方法通过学习遥感图像和文本描述之间更精细的关联性，往往能够取得优于全局模态方法的对齐效果。然而，此类方法忽视了图像和文本上下文背景信息，可能会引起不理想的对齐结果。

2.3.3 基于上下文关系对齐的方法

除了上述两种对齐方式外，许多研究者考虑了两者的结合，既考虑了整体的语义关系，又处理不同文本单词特征与图像局部特征之间的关系，本文将这种图像—文本上下文联合语义对齐的方法称为基于上下文关系对齐的方法。

Wang等（2024a）通过视觉和语言的全局表示以及跨模态交叉注意力模块实现图文全局和局部级别的提取，然后利用多粒度特征融合模块将局部信息融入全局特征，实现全局与局部层面的对齐。Yang等（2024）提出一种隐式—显式关系推理模型，首次在遥感图文跨模态检索中挖掘文本和图像特征的隐式和显式关系。通过学习局部图像—文本之间的关系，在不需要额外先验监督的情况下实现图文的一致性对齐。Gao等（2021）提出了一种跨尺度对齐方法，包括不同尺度间的自适应对齐。该方法通过自适应地从图像和文本中提取多粒度的视觉和语义特征，从而增强文本的上下文关联。Yuan和Shi（2024）提出了一种多粒度视觉语义交互融合模型，该模型包括多粒度特征提取和多粒度特征融合两个网络。前者用于从图像文本中提取多粒度的视觉和语义特征，后者用于对齐来自不同模态的特征，以实现深度的内部和外部融合。Ma等（2024）设计了一种基于方向导向的视觉—语义嵌入模型，通过区域导向注意模块自适应地调整语义空间中图像和文本嵌入的距离。然后，利用轻量级模块扩展文本标识的可处理范围，实现遥感图像和文本对齐准确度的提高。

基于上下文关系对齐的方法充分考虑了图像和文本描述中物体与非物体内容的语义关联，有效减少了冗余信息的干扰，从而实现了更精准的图文对齐。然而，此类方法可能会导致模型的计算成本增高，降低了对齐效率，因此仍需进一步的改进。

2.3.4 挑战与展望

虽然现有的遥感图文对齐任务从不同角度实现遥感图像和文本的对齐，但仍存在一些问题。

（1）遥感图文多层级对齐：遥感图像信息的呈现具有多层级的特性，涵盖场景级、目标级和像素级3个层次。然而，现有的对齐算法主要是基于目标级对齐，忽视了场景级全局语义的宏观描述和像素级细节的微观特征。同时，文本描述中的介词、形容词等语义细节也未被充分利用。因此，将场景、目标、像素级信息与文本语义结合起来，充分挖掘层级间的互补性，是突破遥感图文对齐的关键。

（2）遥感图文轻量级对齐：在遥感图文对齐任务中，特别是面向资源受限的设备或系统，传统算法往往因计算复杂度高、模型规模大而难以满足实时性和轻量化需求。因此，如何在保证对齐准确性的同时，优化算法的计算复杂度和模型规模，使其能够高效运行于资源受限环境，并适配多种下游任务，是遥感图文对齐亟待解决的关键问题。

2.4 遥感图像问答任务

遥感图像问答任务（图文对话）是指根据给定的遥感图像I和与之相关的问题文本Q，利用问题的语义信息在图像中提取相关内容，生成准确的答案A，其映射关系为：$(I, Q) \rightarrow A$。这一任务以对遥感图像及问题的深度理解为基础，要求模型能够根据文本语义对图像内容进行推理，为遥感领域的场景分类、目标检测及关系推理等研究提供了新的视角。同时，它还适用于灾害监测与应急响应（回答受灾区域受损情况）、设施定位与评估（如建筑物状态查询）以及遥感数据解译自动化（快速回答特定地物信息）。

如图6所示，遥感图文对话主要包含3个步骤，首先，使用图像信息编码器和文本信息编码器分别对遥感图像的空间信息以及问题描述的语义信息进行编码。然后，将编码得到的图像和文本特征放在共享特征空间中使用不同的对齐方法生成多模态表示特征。最后，通过分类器或者生

成器，对与图像和问题相关的答案进行预测。根据图文对齐方法的不同，现有图文对话方法主要分为以下3类：基于联合嵌入的方法、基于场景推理的方法和基于学习策略的方法。

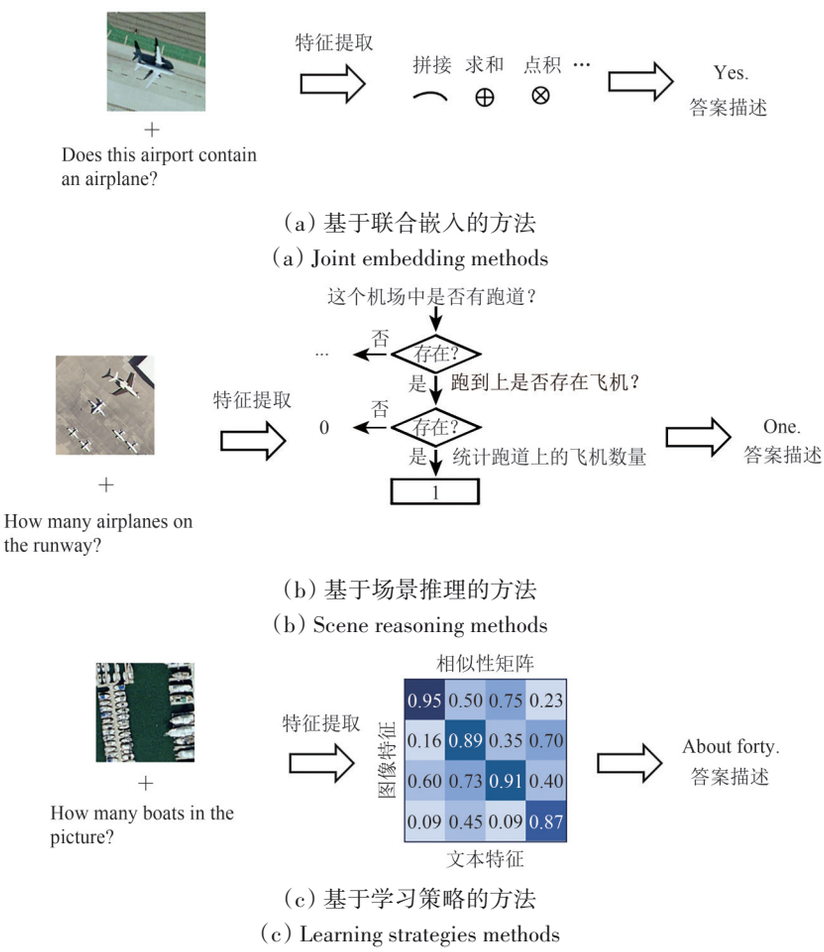


图6 遥感图像问答的3种方法框架

Fig. 6 Three method frameworks for remote sensing visual question answering

2.4.1 基于联合嵌入的方法

基于联合嵌入的方法首先对图像和文本进行特征提取，然后利用点积、加权求和或拼接等方式将两者进行对齐，最后将对齐后的特征送入分类器或者生成器输出答案。此类方法需要将图像和文本映射到一个共同的特征空间，进而实现图像和文本描述的交互。

Bazi等（2022）提出一种基于Transformer的视觉语言模型，使用CLIP预训练网络将图像块和问题词嵌入到一系列图像和文本表示序列中。然后，利用点积来捕捉这些表征内部和之间的关系。最后，通过对两个分类器上的预测结果求平均值生成最终答案。Yuan等（2022a）构建了一个基于变化检测的图像问答框架，通过多时相特征编码、多时相融合、多模态融合以及答案预测来实现与图像相关变化信息的生成。对于多个特征图的对齐，作者比较和分析了元素减法、乘法、求和以及串联方法的效果。Feng等（2023）提出了一种融合上下文和交替引导的注意力网络，使用跨模态交替引导注意力模块映射图像和文本特征，结合学习到的权重，使用加权求和的方法进行模态对齐。同时使用一种改进的多类别损失函数，弥补样本不平衡引起的模型偏差。

基于联合嵌入的方法采用模态直接融合的方式生成相应的答案，是当前大多数遥感图像问答方法的基础。然而，此类方法在联合嵌入的过程中会引入大量无用信息，导致图像区域与问题关键词对齐困难。此外，由于缺乏对问题的深入推理，答案生成的质量还有很大的提升空间。

2.4.2 基于场景推理的方法

基于场景推理的方法通过捕捉场景中物体间的语义和空间关系，利用问题的语义线索引导图像内容进行推理，最后生成相应的答案。此类方法不仅需要对图像中物体进行检测和识别，而且需要更高层次的视觉理解，提高生成答案的准确性。

Zhang等（2023b）提出一种空间层次推理网络，使用基于哈希算法的空间多尺度视觉表示模块对多尺度视觉特征进行编码，嵌入位置信息。在文本信息的指导下，进行空间层次推理，最后采用图像问答交互模块进行答案预测。Zhang等（2023a）提出了一个端到端的多步问题驱动遥感图像问答方法，采用多步注意力机制进行交互式推理，标记与问题最相关的区域。该方法构建了一个基于问题驱动的模块，对问题类型和关键词进行分类，实现对不同尺度特征图的有效指导。

基于场景推理的方法能够理解对象之间的语义关系，尤其是在包含复杂地形结构或时间序列的遥感图像中，依然可以表现出优异的图文对话性能。

2.4.3 基于学习策略的方法

基于学习策略的方法将对比学习、迁移学习、强化学习等技术引入到遥感图像问答任务中，以适应不同的应用场景和问题类型，生成稳健准确的答案。

Yuan等（2023）提出一种多层次视觉特征学习方法，该方法利用跨模态全局注意力、跨模态Transformer以及基于自适应学习的策略，共同提取语言引导的整体和区域图像特征，提高了图文对话模型的可靠性和鲁棒性。Li等（2024a）针对遥感图像问答任务提出一种跨模态迁移学习模型，该模型通过在大规模遥感数据上预训练的权重进行初始化。在跨模态融合模块中，利用注意力机制融合图像和文本的上下文表示，从而促进图像和文本的跨模态表示学习。

基于学习策略的方法能够在数据不平衡、数据噪声等情况下有效提升模型的学习能力。然而，这类方法也存在一些缺点，如训练过程复杂、模型收敛速度较慢，以及调参难度较大。

2.4.4 挑战与展望

尽管现有研究能够生成较为准确的答案，但由于遥感图像具有复杂的空间结构和丰富的语义信息，遥感图像问答任务仍面临以下挑战。

（1）遥感问答的图像多源性：遥感图像通常来自不同的传感器类型，例如光学图像、雷达图像和红外图像，这些图像在分辨率和光谱特征等方面均存在差异。为了准确回答与图像相关的问题，需要集成并分析这些异构数据，以提高语义信息的综合提取能力。然而，多源数据间可能存在空间覆盖、成像时间和光谱特性的差异。因此，如何在保持多源数据完整性的同时，利用它们的互补性，成为遥感问答技术面临的一个复杂挑战。

（2）遥感问答的问题多样性：在遥感问答中，问题类型的多样性对系统提出了较高要求。这些问题涵盖从具体指标的查询（如某地区的植被覆盖率）到复杂事件的分析（如灾害评估和城市扩张）。同时，不同用户可能以专业术语或日常语言的形式提问，这要求系统能够理解和解析自然语言中包含的多种表达方式和意图，从而在复杂的语境中定位准确的信息。此外，通过基于已有答案的反馈实施递进式提问，不仅可以逐步深化对遥感场景的细节挖掘，还能提升对复杂问题的综合理解能力。

（3）遥感问答的答案开放性：由于遥感问答涉及多样化的问题，许多问题难以用简单的标签或单句文本回答。答案的表达形式需要适应具体的任务要求和用户需求，既可能是简单的数值，也可能需要图表、摘要或详细报告。这种开放性要求问答系统不仅需要具备强大的语义推理能力，能够准确理解问题，还能根据图像和背景知识生成科学、严谨且适合目标用户需求的答案。因此，如何有效实现答案的开放性，关系到遥感问答系统的高效应用与实际价值。

3 数据集与评价指标

本节讨论了遥感图文跨模态理解中4类任务常用的数据集及评价指标，包括遥感图像描述，文本生成图像，遥感图文对齐和遥感图像问答。

3.1 遥感图像描述数据集及评价指标

3.1.1 遥感图像描述数据集

遥感图像描述任务常使用的数据集如表2所示，本节对3个数据集UCM-Caption（Qu等，2016），Sydney-Captions（Qu等，2016）和NWPU-Captions（Cheng等，2022）进行介绍。

表2 遥感图文跨模态理解数据集

Table 2 Remote sensing image-text cross-modal understanding datasets

任务	数据集	下载地址
图生文	UCM-Captions（Qu等，2016）	https://pan.baidu.com/s/1mjPToHq
	Sydney-Captions（Qu等，2016）	https://pan.baidu.com/s/1hujEmcG
	NWPU-Captions（Cheng等，2022）	https://github.com/HaiyanHuang98/NWPU-Captions/tree/main
	RSICD（Lu等，2018）	https://github.com/XiangtaoZheng/RSICD
	RSITMD（Yuan等，2022b）	https://github.com/xiaoyuan1996/AMFMN
	SkyScript（Wang等，2024b）	https://github.com/wangzhecheng/SkyScript
	MMM-RS（Luo等，2024）	https://anonymous.4open.science/r/MMM-RS-C73A
	RS5M（Zhang等，2024）	https://huggingface.co/datasets/Zilun/RS5M
	VRSBench（Li等，2024b）	https://huggingface.co/datasets/xiang709/VRSBench
文生图	RSICD（Lu等，2018）	https://github.com/XiangtaoZheng/RSICD
	SkyScript（Wang等，2024b）	https://github.com/wangzhecheng/SkyScript
	MMM-RS（Luo等，2024）	https://anonymous.4open.science/r/MMM-RS-C73A
图文对齐	UCM-Captions（Qu等，2016）	https://pan.baidu.com/s/1mjPToHq#list/path=%2F
	Sydney-Captions（Qu等，2016）	https://pan.baidu.com/s/1hujEmcG#list/path=%2F
	NWPU-Captions（Cheng等，2022）	https://github.com/HaiyanHuang98/NWPU-Captions/tree/main
	RSICD（Lu等，2018）	https://github.com/XiangtaoZheng/RSICD
	SkyScript（Wang等，2024b）	https://github.com/wangzhecheng/SkyScript
	RSITMD（Yuan等，2022b）	https://github.com/xiaoyuan1996/AMFMN
	SeeTRUE（Yarom等，2023）	https://huggingface.co/datasets/yonatanbitton/SeeTRUE
图文对话	RSVQA（Lobry等，2020）	https://zenodo.org/records/6344334
	RSIVQA（Zheng等，2022）	https://github.com/XiangtaoZheng/RSIVQA
	RSVQAxBEN（Lobry等，2021）	https://zenodo.org/records/5084904
	RescueNet-VQA（Sarkar和Rahnemoonfar，2023）	https://github.com/BinaLab/RescueNet-VQA
	TextRS-VQA（Al Rahhal等，2022）	https://textvqa.org/
	VRSBench（Li等，2024b）	https://huggingface.co/datasets/xiang709/VRSBench

（1）UCM-Captions数据集：UCM-Captions数据集是从美国地质调查局国家地图城市区域影像集合中提取的，包含农业、飞机、棒球场、海滩、建筑物、灌木丛等21类场景类别，每个类别有100幅256×256大小的图像，每幅图像包含5个不同的描述，共有2100幅图像和10500个描述。其中，训练集包括1680幅图像，验证集和测试集均包括210幅图像。与其他数据集相比，UCM-Captions数据集语句较为简单，可用于图生文和图文对齐任务。

（2）Sydney-Captions数据集：Sydney-Captions数据集来自谷歌地球的悉尼场景分类数据集，包括住宅、机场、草地、河流、海洋、工业和跑道7个类别的数据，是一个小规模的数据集。该数据集空间分辨率为0.5 m，经过裁剪和选择后，总共包含613幅图像，每幅图像大小为500×500，并配有5个图像描述。其中，训练集包含497幅图像，验证集和测试集均包括58幅图像。与其他数据集相比，Sydney-Captions数据集规模较小，但相比于UCM-Captions，其描述语句更长，信息量更大，可用于图生文和图文对齐任务。

（3）NWPU-Captions数据集：NWPU-Captions数据集是在场景分类数据集的基础上进行扩展。该数据集包含45个场景类别，每个类别包含700幅图像。每幅图像都用5个不同的句子进行描述，每个句子不少于6个单词，总共包含31500幅大小为256×256的图像和157500个句子。其中，25200幅图像用于训练，验证集和测试集各有3150幅图像。NWPU-Captions数据集样本量较大，种类数量较多，可用于图生文和图文对齐任务。

3.1.2 遥感图像描述评价指标

遥感图像描述任务常使用BLEU（Papineni等，2002），METEOR（Denkowski和Lavie，2014），ROUGE（Lin，2004），CIDEr（Vedantam等，2015）和SPICE（Anderson等，2016）评价生成描述的质量。

（1）BLEU通过计算生成描述与参考描述之间的n-gram重叠率来评估生成描述的质量。其中，n-gram是包含n个单词组成的序列，n通常设置为1，2，3或者4。BLEU的计算公式如下：

$$\mathrm{BLEU} = \mathrm{BP} \cdot exp\left(\sum_{i=1}^{N} \boldsymbol{w}_n \log \boldsymbol{P}_i\right) \tag{1}$$

$$\mathrm{BP} = \begin{cases} 1, & c > r \\ \mathrm{e}^{(1-r/c)}, & c \leqslant r \end{cases} \tag{2}$$

式中，$\boldsymbol{w}_n$是n-gram的权重因子，满足$\sum_{i=1}^{N} \boldsymbol{w}_n = 1$；BP表示短句惩罚因子；$P$表示精确率，即生成描述中与参考描述匹配的元素（n-gram）占生成描述中总元素数量的比例；P_i表示基于n-gram单词的精确率；N表示n-gram的最大长度，通常设为4。由于生成的较短句子往往获得更高的分数，BLEU引入一个长度惩罚因子BP，用于约束生成句子比参考句子短的情况，r表示参考描述的长度，c表示生成描述的长度。

（2）METEOR解决了BLEU的一部分局限性，特别是在处理同义词、词序和部分匹配方面。它首先计算参考描述与生成描述之间的BLEU-1分数，然后结合精确率和召回率来评估匹配结果。此外，METEOR还引入了显式的排序组件，用于考虑生成描述与参考描述之间n-gram的顺序，从而使得评价指标比BLEU更具鲁棒性。METEOR分数通过计算精确率和召回率的调和平均数得到，具体公式如下：

$$\mathrm{METEOR} = \left(\frac{10 \cdot P \cdot R}{R + 9P}\right)(1 - \mathrm{BP}_1) \tag{3}$$

$$\mathrm{BP}_1 = 0.5 \cdot \left(\frac{c}{M_u}\right) \tag{4}$$

式中，P表示精确率，即生成描述与参考描述匹配的元素占生成描述总元素数量的比例；R表示召回率，即生成描述与参考描述匹配的元素占参考描述总元素数量的比例；BP_1表示惩罚因子，用于惩罚匹配元素中不连贯的序列；M_u表示匹配的单词总数；c表示参考描述与生成描述中连续且排序相同的元素数。

（3）ROUGE最初用于评估文本摘要的质量，包括ROUGE-1、ROUGE-2、ROUGE-W和ROUGE-L等指标，每种指标针对不同的任务设计，其中ROUGE-L常用于文本生成任务。ROUGE-L主要通过计算生成描述与参考描述之间的词序匹配程度，以评估生成描述的质量。与BLEU不同，ROUGE-L基于精确率P和召回率R的调和平均计算评分，综合评价生成描述与参考描述的匹配程度。ROUGE-L的计算公式如下：

$$\mathrm{ROUGE-L} = \frac{R \cdot P(1 + (P/R)^2)}{R + P \cdot (P/R)^2} \tag{5}$$

（4）CIDEr是一种专门用于评估生成描述质量的指标，它通过计算生成描述与参考描述之间的余弦相似度，并结合TF-IDF（Term Frequency-Inverse Document Frequency）权重来评估描述中每个n-gram的重要性。CIDEr特别关注与描述语义相关的关键短语，从而对生成描述的语义质量进行综合评价。CIDEr的具体公式如下：

$$\mathrm{CIDEr} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{CIDEr}_n(\boldsymbol{c}_i, \boldsymbol{s}_i) \tag{6}$$

$$\mathrm{CIDEr}_n = \frac{1}{M}\sum_{i=1}^{M} \frac{\boldsymbol{g}^n(\boldsymbol{c}_i) \cdot \boldsymbol{g}^n(\boldsymbol{s}_{ij})}{\left\|\boldsymbol{g}^n(\boldsymbol{c}_i)\right\| \cdot \left\|\boldsymbol{g}^n(\boldsymbol{s}_{ij})\right\|} \tag{7}$$

式中，$\boldsymbol{c}_i$表示第i个图像的生成描述；$\boldsymbol{s}_{ij}$表示第i个图像生成的第j个参考描述；N为测试图像的总数；$\boldsymbol{g}^n(\boldsymbol{x})$表示$x$中n-gram的TF-IDF权重向量，$M$表示每幅图像参考描述的数量。

（5）SPICE是一种基于语义图的评价指标，它通过将参考描述和生成描述解析为语义图，计算二者在对象、属性和关系上的匹配程度。SPICE的核心思想是从语义层面评估描述的质量，而不是单纯依赖于词汇或句法匹配。计算公式如下：

$$\mathrm{SPICE}(\boldsymbol{C}, \boldsymbol{S}) = \frac{2P \cdot R}{P + R} \tag{8}$$

式中，$\boldsymbol{C}$表示生成描述的语义图，是所有的$\boldsymbol{C}_i$经过语义解析后得到的结构化表示，其中$\boldsymbol{C}_i$表示第i个图像的生成描述；$\boldsymbol{S}$表示参考描述的语义图，是所有的$\boldsymbol{S}_{ij}$经过语义解析后得到的结构化表示，其中$\boldsymbol{S}_{ij}$表示第i个图像生成的第j个参考描述；P表示精确率；R表示召回率。

3.2 文本生成图像数据集及评价指标

3.2.1 文本生成图像数据集

文本生成图像任务常使用的数据集如表2所示，

本小节对RSICD（Lu等，2018），SkyScripts（Wang等，2024b），MMM-RS（Luo等，2024）3个数据集进行介绍。

（1）RSICD数据集：RSICD数据集是一个大规模的遥感文本—图像数据集，包括机场、荒地、棒球场、海滩、桥梁等30个场景类别。数据集总共含有10921幅大小为224×224高分辨率遥感图像和24333个文本描述，每幅图像都有5个描述，对于少于5个描述的图像，描述会被重复填充。其中训练集包含8734幅图像，验证集和测试集分别包含1094和1093副图像。相比其他数据集，它包含的数据量较大，对象类型多样。该数据集适用场景较多，可用于图生文、文生图以及图文对齐任务。

（2）SkyScript数据集：SkyScript数据集由Google Earth Engine（GEE）平台获取，并结合OpenStreetMap（OSM）数据，以丰富遥感图像的语义信息。该数据集包括260万个图文数据对，涵盖29000个不同的标签，这些标签形成了10万个单目标文本和120万个多目标文本，其中图像的空间分辨率为0.1—30 m。SkyScript数据集可用于图生文、文生图以及图文对齐任务。

（3）MMM-RS数据集：MMM-RS数据集是一种多模态、多分辨率、多场景的遥感数据集，涵盖了9个公开的的遥感数据集。通过将图像裁剪或缩放至512×512的大小，得到约210万个文本—图像数据对，其中图像模态包含可见光、合成孔径雷达和近红外3种，分别有1806889、289384和7000个文本—图像对。该数据集可用于图生文和文生图任务。

3.2.2 文本生成图像评价指标

常用的文本生成图像评价指标主要有IS（Inception Score）（Salimans等，2016）和FID（Fréchet Inception Distance）（Heusel等，2017）。

（1）IS是一种用于评估生成图像质量和多样性的指标。该指标使用Inception-v3模型提取生成图像的特征，计算生成图像预测类别分布$p(y|\boldsymbol{x})$与总体类别分布$p(y)$之间的KL散度，并将其指数化。计算公式如下：

$$\mathrm{IS}=exp\left(\mathrm{E}_{\boldsymbol{x}}\left(D_{\mathrm{KL}}(p(y|\boldsymbol{x})\|p(y))\right)\right) \tag{9}$$

$$p(y)=\frac{1}{N}\sum_{i=1}^{N}p(y|\boldsymbol{x}_i) \tag{10}$$

$$D_{\mathrm{KL}}(p\|q)=\sum_{y}p(y)\log\frac{p(y)}{q(y)} \tag{11}$$

式中，$\boldsymbol{x}$表示生成的图像；$\boldsymbol{x}_i$表示生成的第i个图像；y表示Inception-v3模型的预测标签；$p(y|\boldsymbol{x})$表示给定生成图像$\boldsymbol{x}$的预测类别分布；$p(y)$是所有生成图像的$p(y|\boldsymbol{x})$汇总得到的边界类别分布；N为生成图像的总数量；D_{KL}表示KL散度，用于衡量两个概率分布之间的差异。

（2）FID作为对现有指标IS的改进，FID不仅考虑生成图像的质量，还同时评估其多样性。该指标使用Inception系列神经网络对真实图像和生成图像提取高维特征，通过计算这些特征的均值和协方差，将其建模为两个多元高斯分布。随后，计算这两个分布之间的Fréchet距离（欧氏距离的平方）得到最终的FID分数。较低的FID分数表示生成图像的特征分布与真实图像的特征分布更接近，即生成图像的质量和多样性更好。由于FID能够有效捕捉生成图像的质量和多样性，同时与人类对图像质量的主观感知具有较强相关性，因此被广泛用于评估生成图像的质量。具体计算公式如下：

$$\begin{aligned}\mathrm{FID}=&\left\|\mu_r-\mu_g\right\|^2+\\&\mathrm{Tr}(\sum\boldsymbol{r}+\sum\boldsymbol{g}-2(\sum\boldsymbol{r}\sum\boldsymbol{g})^{1/2})\end{aligned} \tag{12}$$

式中，μ_r表示真实图像特征分布的均值；$\sum\boldsymbol{r}$表示真实图像特征分布的协方差矩阵；μ_g表示生成图像特征分布的均值；$\sum\boldsymbol{g}$表示生成图像特征分布的协方差矩阵；$\|\mu_r-\mu_g\|^2$表示两个均值欧几里得距离的平方，用于衡量生成图像与真实图像在特征均值上的差异；Tr表示矩阵迹运算，用于衡量协方差矩阵之间的差异。

3.3 遥感图文对齐数据集及评价指标

3.3.1 遥感图文对齐数据集

遥感图文对齐任务常用数据集如表2所示，本节对RSITMD（Yuan等，2022b）和SeeTRUE（Yarom等，2023）两个数据集进行介绍。

（1）RSITMD数据集：RSITMD数据集是从RSICD数据集和谷歌地图收集得到，场景类别与RSICD数据集相同。该数据集包含4743幅256×256大小的图像，其中4291幅图像用于训练，用于验证和测试的数据相同，均包含452幅图像。与RSICD数据集相比，RSITMD数据集数据量更少，

但文本描述种类更多，常用于图生文和图文对齐任务。

（2）SeeTRUE数据集：SeeTRUE数据集是一个综合的图文对齐数据集，基于多个已有数据集，包括SNLI-VE（Saharia等，2022）、Winoground（Thrush等，2022）、DrawBench（Wang等，2023）、EditBench（Xie等，2019）、MS COCO（Lin等，2014）和Pick-a-Pick（Kirstain等，2023）。在这些数据集的基础上进行人工标注，标注内容包括图文是否在语义上对齐。该数据集包含真实图像与合成图像以及真实文本与合成文本多种组合的31,855对图文实例，适用于图文对齐、文生图等任务。

3.3.2 遥感图文对齐评价指标

常用的遥感图文对齐评价指标主要有Recall@K，RUSM和问答正确率。

（1）遥感图文对齐使用Recall@K（R@K）作为评价指标，Recall为召回率，是一个用于衡量正确检索到的数量与给定查询的所有相关数量之比的指标，计算公式如下：

$$\mathrm{R@K} = \frac{TP@\mathrm{K}}{TP@\mathrm{K} + FN@\mathrm{K}} \tag{13}$$

$$\begin{aligned}\mathrm{RSUM} = &(\mathrm{R@1} + \mathrm{R@5} + \mathrm{R@10})_{i-t} + \\ &(\mathrm{R@1} + \mathrm{R@5} + \mathrm{R@10})_{t-i}\end{aligned} \tag{14}$$

式中，TP表示真阳样本数量，FN表示假阴样本数量，K可取1、5和10。在图对齐文的任务中使用$(\mathrm{R@K})_{i-t}$表示图到文的对齐衡量；文对齐图任务中使用$(\mathrm{R@K})_{t-i}$表示文到图的对齐衡量。此外，为了显示整体对齐性能，在图像到文本和文本到图像方向上将所有R@K求和为RSUM，具体计算如公式（14）所示。

（2）Yarom等（2023）和Miyamoto等（2024）针对图文对齐任务的评价指标，提出使用问答正确率衡量。首先，使用GPT等模型根据文本描述生成问答对，然后，在问答模型中计算问题回答的正确率，如式（16）。然而，这种方法没有根据图像生成对应的问答对，且评价过程较为繁琐。

俗话说“一图解千文”，图像和文字各有其独特的表达力量。一幅图可以比千言万语更直观、更清晰地传达信息，而一篇精炼的文字描述也能够详细解释图像的含义。两者相辅相成，共同传递信息。目前，遥感图文对齐任务中使用的评价指标基于图文检索和图像问答。要精确评价图像与文字之间的对齐性与不对齐性，我们认为应在现有评价体系的基础上进一步拓展更加细致、有效的评价指标。

3.4 遥感图像问答数据集及评价指标

3.4.1 遥感图像问答数据集

遥感图像描述任务常使用的数据集如表2所示，本节对RSVQA（Lobry等，2020），RSIVQA（Zheng等，2022）和RSVQAxBEN（Lobry等，2021）3个数据集进行介绍。

（1）RSVQA数据集：RSVQA数据集包含低分辨率数据集和高分辨率数据集两个部分。低分辨率数据集是在荷兰区域内获取得到，空间分辨率为10 m，包含772幅大小为256×256的图像和77232对问答。其中训练集、验证集和测试集分别包含572幅、100幅、100幅图像。高分辨率数据集来自美国地质调查局中空间分辨率为15 cm的正射影像，由10659幅大小为512×512的图像和1066316对问答组成。其中，训练集占61.5%、验证集占11.2%、测试集包含两组，分别占20.5%和6.8%。

（2）RSIVQA数据集：RSIVQA数据集包含3个遥感图像分类数据集UCM（Yang和Newsam，2010）、Sydney（Zhang等，2015）和AID（Xia等，2017）以及两个目标检测数据集HRRSD（Zhang等，2019）和DOTA（Xia等，2018）。该数据集大多数问题是通过场景和物体级别的注释自动生成，有一少部分是由人工注释。总共包含图像—问题—答案三元组111134个，答案分为3种类型：存在型、计数型和其他。其中，80%的数据用于训练，10%用于验证，其余的10%用于测试。

（3）RSVQAxBEN数据集：该数据集包含590326张空间分辨率为10 m的图像，利用CLC标签的随机算法生成了14758150个图像—问题—问答三元组。数据集包括两种类型的问题，分别是土地覆盖类型存在与否的是/否问题，和答案为一个或多个土地覆盖类型名称的问题。

3.4.2 遥感图像问答评价指标

为了评估遥感图像问答模型的稳健性，使用平均准确率AA（Average Accuracy）和整体准确率OA（Overall Accuracy）作为遥感图像问答的评价

指标。其中，AA和OA的计算公式如下：

$$\mathrm{AA}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{A}_i \tag{15}$$

$$\mathrm{OA}=\frac{B}{Q} \tag{16}$$

式中，A_i表示问答中每一类问题的准确率，N表示问题类别数量，B表示回答正确的问题数量，Q表示总问题数量。

4 结 语

4.1 遥感图文跨模态理解挑战性难题

遥感图文跨模态理解能够综合图像和文本两种模态的信息，结合计算机视觉和自然语言处理技术，促进模态之间的转化与交互，为多种遥感应用场景提供技术支撑。本文分别阐述了4类不同任务的相关研究进展，总结现有方法的解决思路，提出现有任务面临的挑战问题。然后，介绍遥感图文跨模态理解常用的公开数据集及评价指标。对现有遥感图文跨模态理解面临的挑战性难题总结如下：

（1）模态对齐问题：遥感图像和文本在数据形式和信息表达上具有显著差异。例如，遥感图像为低维连续数据，承载丰富且复杂的空间分布信息。而文本则是离散的符号表达，倾向于概括和简化描述。这种模态差异使得精准对齐遥感图像与文本成为一项技术难题，需要研究有效的对齐方法来实现两者之间的交互。

（2）跨模态可解释性：跨模态可解释性是提升遥感图像解读质量的关键。由于主观因素的影响，不同人对同一遥感图像的解读存在显著差异。同时遥感图像的多层级特性进一步加剧了这种差异，导致文本描述的多样性和主观性增加。因此，提升跨模态理解的可解释性，有助于用户更好地理解决策依据，改善用户体验，并确保跨模态理解在实际应用中的有效性。

（3）跨模态推理：跨模态理解不仅要求图像和文本有直接的对应关系，更需要深入解析两者之间的内在逻辑。遥感图像通常包含多个层级的信息，层级间既存在紧密的相关性，也具有明显的独立性。如何在这些层级信息之间建立有效的关联，挖掘图像与文本之间的深层逻辑关系，构建具有较好泛化性与精准性的跨模态推理机制，是当前跨模态理解面临的一大挑战。

4.2 遥感图文跨模态理解未来发展趋势

通过分析遥感图文跨模态理解的现有研究，遥感图文跨模态领域未来的发展趋势如下：

（1）跨模态信息深度挖掘：随着科技的进步，遥感图文跨模态理解的需求已不仅仅停留在简单的交互层面，逐渐开始转向高层次的需求。如按照人类的要求，完成指定目标的分割、识别和视觉定位等任务。通过深入挖掘遥感图像中的多层次信息，结合文本的语义内容，可以进一步提升图文跨模态理解的精度和应用广度。

（2）地球科学知识图谱构建：地球科学知识图谱是涵盖地球科学领域所有实体及其相互关系的结构化知识库。为了增强遥感图文跨模态理解任务的专业性和领域适应性，利用知识图谱中蕴含的地理、气象等专业知识，能够帮助模型进行更精确的理解和推理。因此构建一个全面的地球科学知识图谱，对推动遥感图文跨模态理解的发展至关重要。

（3）人机交互：遥感图文跨模态理解涉及对物理世界的抽象与解读，体现了主观世界与物理世界的互动。为了进一步提升图文跨模态理解能力，有必要将人类的反馈与机器的互动联合到遥感图文跨模态任务中，有助于开发符合实际应用的模型。因此，如何将人机互动有效融入到遥感图文跨模态理解任务中仍需进一步探索。

（4）遥感图文大模型：遥感图文跨模态任务的性能可通过预训练的视觉语言模型得到显著提升。通过在大规模遥感数据集上进行预训练，并结合迁移学习等技术，可以为遥感图文任务提供更强大的支持。然而，如何训练一个通用的遥感领域视觉语言大模型，以适应多种跨模态任务，仍然是一个亟待解决的难题。

（5）语言多元化：当前的遥感图文跨模态理解主要以英文描述为主。中华文化博大精深，传统诗词意境优美，民族文化底蕴深厚。如果能通过多国混合语言生成遥感图像的理解文本，包括汉语、少数民族语言及其他各国语言，不仅可以丰富遥感图文数据的多样性，还能大幅提升遥感图文跨模态理解的实用性和文化包容性。

（6）遥感多源数据：当前的遥感图文跨模态理解研究大多集中在光学图像数据上，这在一定

程度上限制了其在更广泛场景中的应用。因此，需要探索其他类型的遥感数据，如多光谱数据、激光雷达、近红外等，以拓展遥感图像跨模态理解任务的适用范围，这是未来值得深入研究的发展方向。

参考文献(References)

Abdullah T, Bazi Y, Al Rahhal M M, Mekhalfi M L, Rangarajan L and Zuair M. 2020. TextRS: deep bidirectional triplet network for matching text to remote sensing images. Remote Sensing, 12(3): 405 [DOI: 10.3390/rs12030405]

Al Rahhal M M, Bazi Y, Alsaleh S O, Al-Razgan M, Mekhalfi M L, Al Zuair M and Alajlan N. 2022. Open-ended remote sensing visual question answering with transformers. International Journal of Remote Sensing, 43(18): 6809-6823 [DOI: 10.1080/01431161.2022.2145583]

Anderson P, Fernando B, Johnson M and Gould S. 2016. SPICE: semantic propositional image caption evaluation//14th European Conference on Computer Vision. Amsterdam: Springer: 382-398 [DOI: 10.1007/978-3-319-46454-1_24]

Bazi Y, Rahhal M M A, Mekhalfi M L, Zuair M A A and Melgani F. 2022. Bi-modal transformer-based approach for visual question answering in remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 60: 4708011 [DOI: 10.1109/TGRS.2022.3192460]

Bejiga M B, Hoxha G and Melgani F. 2020. Retro-remote sensing with Doc2Vec encoding//2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS). Tunis: IEEE: 89-92 [DOI: 10.1109/M2GARSS47143.2020.9105139]

Bejiga M B, Hoxha G and Melgani F. 2021. Improving text encoding for retro-remote sensing. IEEE Geoscience and Remote Sensing Letters, 18(4): 622-626 [DOI: 10.1109/LGRS.2020.2983851]

Bejiga M B, Melgani F and Vascotto A. 2019. Retro-remote sensing: generating images from ancient texts. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(3): 950-960 [DOI: 10.1109/JSTARS.2019.2895693]

Chen Y X, Huang J H, Li X Y, Xiong S W and Lu X Q. 2024. Multi-scale salient alignment learning for remote-sensing image–text retrieval. IEEE Transactions on Geoscience and Remote Sensing, 62: 4700413 [DOI: 10.1109/TGRS.2023.3340870]

Cheng G, Xie X X, Han J W, Guo L and Xia G S. 2020. Remote sensing image scene classification meets deep learning: challenges, methods, benchmarks, and opportunities. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13: 3735-3756 [DOI: 10.1109/JSTARS.2020.3005403]

Cheng Q M, Huang H Y, Xu Y, Zhou Y Z, Li H Y and Wang Z Y. 2022. NWPU-captions dataset and MLCA-Net for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing, 60: 5629419 [DOI: 10.1109/TGRS.2022.3201474]

Cheng Q M, Zhou Y Z, Fu P, Xu Y and Zhang L. 2021. A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14: 4284-4297 [DOI: 10.1109/JSTARS.2021.3070872]

Deng P F, Xu K J and Huang H. 2021. CNN-GCN-based dual-stream network for scene classification of remote sensing images. National Remote Sensing Bulletin, 25(11): 2270-2282 (邓培芳, 徐科杰, 黄鸿. 2021. 基于CNN-GCN双流网络的高分辨率遥感影像场景分类. 遥感学报, 25(11): 2270-2282) [DOI: 10.11834/jrs.20210587]

Denkowski M and Lavie A. 2014. Meteor universal: language specific translation evaluation for any target language//Proceedings of the Ninth Workshop on Statistical Machine Translation. Baltimore: ACL: 376-380 [DOI: 10.3115/v1/W14-3348]

Du R Y, Cao W, Zhang W K, Zhi G, Sun X, Li S K and Li J H. 2023. From plane to hierarchy: deformable transformer for remote sensing image captioning. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16: 7704-7717 [DOI: 10.1109/JSTARS.2023.3305889]

Faghri F, Fleet D J, Kiros J R and Fidler S. 2018. VSE++: improving visual-semantic embeddings with hard negatives//British Machine Vision Conference 2018. Newcastle: BMVA Press

Feng J F, Tang E T, Zeng M M, Gu Z J, Kou P L and Zheng W. 2023. Improving visual question answering for remote sensing via alternate-guided attention and combined loss. International Journal of Applied Earth Observation and Geoinformation, 122: 103427 [DOI: 10.1016/j.jag.2023.103427]

Gao C Y, Cai G Y, Jiang X Y, Zheng F, Zhang J, Gong Y F, Peng P, Guo X W and Sun X. 2021. Contextual non-local alignment over full-scale representation for text-based person search. arXiv preprint arXiv: 2101.03036

Heusel M, Ramsauer H, Unterthiner T, Nessler B and Hochreiter S. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc.: 6629-6640

Hoxha G and Melgani F. 2022. A novel SVM-based decoder for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing, 60: 5404514 [DOI: 10.1109/TGRS. 2021.3105004]

Ivasic-Kos M, Pobar M and Ribaric S. 2016. Two-tier image annotation model based on a multi-label classifier and fuzzy-knowledge representation scheme. Pattern Recognition, 52: 287-305 [DOI: 10.1016/j.patcog.2015.10.017]

Kandala H, Saha S, Banerjee B and Zhu X X. 2022. Exploring transformer and multilabel classification for remote sensing image captioning. IEEE Geoscience and Remote Sensing Letters, 19: 6514905 [DOI: 10.1109/LGRS.2022.3198234]

Kirstain Y, Polyak A, Singer U, Matiana S, Penna J and Levy O. 2023. Pick-a-Pic: an open dataset of user preferences for text-to-image generation//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc.: 1594

Lee K, Liu H, Ryu M, Watkins O, Du Y, Boutilier C, Abbeel P, Ghavamzadeh M and Gu S S. 2023. Aligning text-to-image models using human feedback. arXiv preprint arXiv: 2302.12192

Lee K H, Chen X, Hua G, Hu H D and He X D. 2018. Stacked cross attention for image-text matching//15th European Conference on Computer Vision. Munich: Springer: 212-228 [DOI: 10.1007/978-3-030-01225-0_13]

Li P F, He J L, Liu G and Zhong S J. 2024a. PECR: parameter-efficient transfer learning with cross-modal representation learning for remote sensing visual question answering//2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul: IEEE: 6740-6744 [DOI: 10.1109/ICASSP48485.2024.10446146]

Li X, Ding J and Elhoseiny M. 2024b. VRSBench: a versatile vision-language benchmark dataset for remote sensing image understanding//38th Conference on Neural Information Processing Systems. Vancouver: [s.n]

Li Z, Gao L R, Zheng K and Ni L. 2024. Research progress of high-resolution remote sensing image scene classification. National Remote Sensing Bulletin, 28(11): 2739-2760 (李智, 高连如, 郑珂, 倪丽 . 2024. 高分辨率遥感图像场景分类研究进展 . 遥感学报, 28(11): 2739-2760) [DOI: 10.11834/jrs.20243519]

Lin C Y. 2004. ROUGE: a package for automatic evaluation of summaries//Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004). Barcelona: ACL: 74-81

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//13th European Conference on Computer Vision. Zurich: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Lin Z Y, Zhu F, Wang Q, Kong Y Z, Wang J Y, Huang L and Hao Y M. 2022. RSSGG_CS: remote sensing image scene graph generation by fusing contextual information and statistical knowledge. Remote Sensing, 14(13): 3118 [DOI: 10.3390/rs14133118]

Lobry S, Demir B and Tuia D. 2021. RSVQA meets bigearthnet: a new, large-scale, visual question answering dataset for remote sensing//2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. Brussels: IEEE: 1218-1221 [DOI: 10.1109/IGARSS47720.2021.9553307]

Lobry S, Marcos D, Murray J and Tuia D. 2020. RSVQA: visual question answering for remote sensing data. IEEE Transactions on Geoscience and Remote Sensing, 58(12): 8555-8566 [DOI: 10.1109/TGRS.2020.2988782]

Lu X Q, Wang B Q, Zheng X T and Li X L. 2018. Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing, 56(4): 2183-2195 [DOI: 10.1109/TGRS.2017.2776321]

Luo J L, Wang Y Z, Gu Z Q, Qiu Y D, Yao S Z, Wang F Y, Xu C Y, Zhang W H, Wang D and Cui Z. 2024. MMM-RS: a multi-modal, multi-GSD, multi-scene remote sensing dataset and benchmark for text-to-image generation//38th Conference on Neural Information Processing Systems. Vancouver: [s.n]

Ma Q, Pan J C and Bai C. 2024. Direction-oriented visual–semantic embedding model for remote sensing image–text retrieval. IEEE Transactions on Geoscience and Remote Sensing, 62: 4704014 [DOI: 10.1109/TGRS.2024.3392779]

Mitchell M, Dodge J, Goyal A, Yamaguchi K, Stratos K, Han X F, Mensch A, Berg A, Berg T and Daumé III H. 2012. Midge: generating image descriptions from computer vision detections//Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon: ACL: 747-756

Miyamoto M, Morita R and Zhou J J. 2024. Visual question answering based evaluation metrics for text-to-image generation//2024 IEEE International Symposium on Circuits and Systems (ISCAS). Singapore: IEEE: 1-5 [DOI: 10.1109/ISCAS58744.2024.10558259]

Ordonez V, Kulkarni G and Berg T L. 2011. Im2Text: describing images using 1 million captioned photographs//Proceedings of the 25th International Conference on Neural Information Processing Systems. Granada: Curran Associates Inc.: 1143-1151

Papineni K, Roukos S, Ward T and Zhu W J. 2002. BLEU: a method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia: ACL: 311-318 [DOI: 10.3115/1073083.1073135]

Qu B, Li X L, Tao D C and Lu X Q. 2016. Deep semantic understanding of high resolution remote sensing image//2016 International Conference on Computer, Information and Telecommunication Systems (CITS). Kunming: IEEE: 1-5 [DOI: 10.1109/CITS.2016.7546397]

Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M and Sutskever I. 2021. Zero-shot text-to-image generation//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 8821-8831

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE: 10674-10685 [DOI: 10.1109/CVPR52688.2022.01042]

Saharia C, Chan W, Saxena S, Lit L, Whang J, Denton E, Ghasemipour S K S, Ayan B K, Mahdavi S S, Gontijo-Lopes R, Salimans T, Ho J, Fleet D J and Norouzi M. 2022. Photorealistic text-to-image diffusion models with deep language understanding//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc.: 2643

Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A and Chen X. 2016. Improved techniques for training GANs//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona: Curran Associates Inc.: 2234-2242

Sarkar A and Rahnemoonfar M. 2023. RESCUENet-VQA: a large-scale visual question answering benchmark for damage assessment//2023 IEEE International Geoscience and Remote Sensing Symposium. Pasadena: IEEE: 1150-1153 [DOI: 10.1109/IGARSS52108.2023.10281747]

Sebaq A and ElHelw M. 2024. RSDiff: remote sensing image genera-

tion from text using diffusion model. Neural Computing and Applications, 36(36): 23103-23111 [DOI: 10.1007/s00521-024-10363-3]

Shi Z W and Zou Z X. 2017. Can a machine generate humanlike language descriptions for a remote sensing image? IEEE Transactions on Geoscience and Remote Sensing, 55(6): 3623-3634 [DOI: 10.1109/TGRS.2017.2677464]

Sun Y J, Wang M and Ma Y. 2024. Semantic-alignment transformer and adversary hashing for cross-modal retrieval. Applied Intelligence, 54(17): 7581-7602 [DOI: 10.1007/s10489-024-05501-2]

Tang D T, Cao X Y, Hou X S, Jiang Z Y, Liu J M and Meng D Y. 2024. CRS-Diff: controllable remote sensing image generation with diffusion model. IEEE Transactions on Geoscience and Remote Sensing, 62: 5638714 [DOI: 10.1109/TGRS.2024.3453414]

Tang Z C, Wei W, Luo W R, Hu J and Zhang D Y. 2023. Remote sensing image semantic segmentation method combining cosine annealing with atrous convolution. National Remote Sensing Bulletin, 27(11): 2579-2592 (唐振超, 韦蔚, 罗蔚然, 胡洁, 张东映. 2023. 融合余弦退火与空洞卷积的遥感影像语义分割. 遥感学报, 27(11): 2579-2592) [DOI: 10.11834/jrs.20211038]

Thrush T, Jiang R, Bartolo M, Singh A, Williams A, Kiela D and Ross C. 2022. Winoground: probing vision and language models for visio-linguistic compositionality//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE: 5228-5238 [DOI: 10.1109/CVPR52688.2022.00517]

Vedantam R, Zitnick C L and Parikh D. 2015. CIDEr: consensus-based image description evaluation//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE: 4566-4575 [DOI: 10.1109/CVPR.2015.7299087]

Wang B Q, Zheng X T, Qu B and Lu X Q. 2020. Retrieval topic recurrent memory network for remote sensing image captioning. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13: 256-270 [DOI: 10.1109/JSTARS. 2019. 2959208]

Wang S, Saharia C, Montgomery C, Pont-Tuset J, Noy S, Pellegrini S, Onoe Y, Laszlo S, Fleet D J, Soricut R, Baldridge J, Norouzi M, Anderson P and Chan W. 2023. Imagen editor and editbench: advancing and evaluating text-guided image inpainting//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE: 18359-18369 [DOI: 10.1109/CVPR52729.2023.01761]

Wang W Z, Di X G, Liu M Z and Gao F. 2024a. Multi-level symmetric semantic alignment network for image–text matching. Neurocomputing, 599: 128082 [DOI: 10.1016/j.neucom.2024.128082]

Wang Y, Bashir S M A, Khan M, Ullah Q, Wang R, Song Y L, Guo Z and Niu Y L. 2022. Remote sensing image super-resolution and object detection: benchmark and state of the art. Expert Systems with Applications, 197: 116793 [DOI: 10.1016/j. eswa. 2022. 116793]

Wang Z C, Prabha R, Huang T Y, Wu J J and Rajagopal R. 2024b. SkyScript: a large and semantically diverse vision-language dataset for remote sensing//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver: AAAI: 5805-5813 [DOI: 10.1609/aaai.v38i6.28393]

Xia B, Yang R N, Ge Y X and Yin J B. 2024. A review of cross-modal retrieval for image-text//Fifteenth International Conference on Graphics and Image Processing (ICGIP 2023). Suzhou: SPIE: 1308915 [DOI: 10.1117/12.3021468]

Xia G S, Bai X, Ding J, Zhu Z, Belongie S, Luo J B, Datcu M, Pelillo M and Zhang L P. 2018. DOTA: a large-scale dataset for object detection in aerial images//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 3974-3983 [DOI: 10.1109/CVPR.2018.00418]

Xia G S, Hu J W, Hu F, Shi B G, Bai X, Zhong Y F, Zhang L P and Lu X Q. 2017. AID: a benchmark data set for performance evaluation of aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing, 55(7): 3965-3981 [DOI: 10.1109/TGRS. 2017.2685945]

Xie N, Lai F, Doran D and Kadav A. 2019. Visual entailment: a novel task for fine-grained image understanding. arXiv preprint arXiv: 1901.06706

Xu D Q and Wu Y Q. 2024. Progress of research on deep learning algorithms for object detection in optical remote sensing images. National Remote Sensing Bulletin, 28(12): 3045-3073 (徐丹青, 吴一全. 2024. 光学遥感图像目标检测的深度学习算法研究进展. 遥感学报, 28(12): 3045-3073) [DOI: 10.11834/jrs.20243166]

Xu T, Zhang P C, Huang Q Y, Zhang H, Gan Z, Huang X L and He X D. 2018. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 1316-1324 [DOI: 10.1109/CVPR.2018.00143]

Xu Y H, Yu W K, Ghamisi P, Kopp M and Hochreiter S. 2023. Txt2Img-MHN: remote sensing image generation from text using modern hopfield networks. IEEE Transactions on Image Processing, 32: 5737-5750 [DOI: 10.1109/TIP.2023.3323799]

Yang L L, Zhou T Q, Ma W T, Du M Z, Liu L, Li F, Zhao S and Wang Y W. 2024. Remote sensing image-text retrieval with implicit-explicit relation reasoning. IEEE Transactions on Geoscience and Remote Sensing, 62: 4511411 [DOI: 10.1109/TGRS. 2024. 3466909]

Yang Y and Newsam S. 2010. Bag-of-visual-words and spatial extensions for land-use classification//Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. San Jose: ACM: 270-279 [DOI: 10.1145/1869790.1869829]

Yarom M, Bitton Y, Changpinyo S, Aharoni R, Herzig J, Lang O, Ofek E and Szpektor I. 2023. What you see is what you read? Improving text-image alignment evaluation//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc.: 79

Yu C Y, Li D L, Song M P, Yu H Y and Chang C I. 2024. Edge-perception enhanced segmentation method for high-resolution remote sensing image. National Remote Sensing Bulletin, 28(12): 3249-3260 (于纯妍, 李东霖, 宋梅萍, 于浩洋, Chang C I. 2024. 边缘感

知强化的高分辨遥感影像语义分割. 遥感学报, 28(12): 3249-3260) [DOI: 10.11834/jrs.20233098]

Yu J H, Xu Y Z, Koh J Y, Luong T, Baid G, Wang Z R, Vasudevan V, Ku A, Yang Y F, Ayan B K, Hutchinson B, Han W, Parekh Z, Li X, Zhang H, Baldridge J and Wu Y H. 2022. Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv: 2206.10789

Yuan X H, Shi J F and Gu L C. 2021. A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Systems with Applications, 169: 114417 [DOI: 10.1016/j. eswa. 2020.114417]

Yuan Z H, Mou L C, Xiong Z T and Zhu X X. 2022a. Change detection meets visual question answering. IEEE Transactions on Geoscience and Remote Sensing, 60: 5630613 [DOI: 10.1109/TGRS. 2022.3203314]

Yuan Z H, Mou L C and Zhu X X. 2023. Multilingual augmentation for robust visual question answering in remote sensing images// 2023 Joint Urban Remote Sensing Event (JURSE). Heraklion: IEEE: 1-4 [DOI: 10.1109/JURSE57346.2023.10144189]

Yuan Z Q, Zhang W K, Fu K, Li X, Deng C B, Wang H Q and Sun X. 2022b. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. IEEE Transactions on Geoscience and Remote Sensing, 60: 4404119 [DOI: 10.1109/TGRS. 2021.3078451]

Yuan Z Y and Shi C. 2024. MGN-Net: multigranularity graph fusion network in multimodal for scene text spotting. IEEE Internet of Things Journal, 11(14): 25088-25098 [DOI: 10.1109/JIOT. 2024. 3390943]

Yusuf A A, Feng C, Mao X L, Ally Duma R, Abood M S and Chukkol A H A. 2024. Graph neural networks for visual question answering: a systematic review. Multimedia Tools and Applications, 83 (18): 55471-55508 [DOI: 10.1007/s11042-023-17594-x]

Zhang F, Du B and Zhang L P. 2015. Saliency-guided unsupervised feature learning for scene classification. IEEE Transactions on Geoscience and Remote Sensing, 53(4): 2175-2184 [DOI: 10. 1109/TGRS.2014.2357078]

Zhang H, Xu T, Li H S, Zhang S T, Wang X G, Huang X L and Metaxas D N. 2017. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks//2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE: 5908-5916 [DOI: 10.1109/ICCV.2017.629]

Zhang M M, Chen F and Li B. 2023a. Multistep question-driven visual question answering for remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 61: 4704912 [DOI: 10.1109/ TGRS.2023.3312479]

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) [DOI: 10.11834/jrs. 20211042]

Zhang Y L, Yuan Y, Feng Y C and Lu X Q. 2019. Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. IEEE Transactions on Geoscience and Remote Sensing, 57(8): 5535-5548 [DOI: 10.1109/TGRS. 2019. 2900302]

Zhang Z L, Zhao T C, Guo Y L and Yin J W. 2024. 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: 5642123 [DOI: 10.1109/ TGRS.2024.3449154]

Zhang Z X, Jiao L C, Li L L, Liu X, Chen P H, Liu F, Li Y X and Guo Z C. 2023b. A spatial hierarchical reasoning network for remote sensing visual question answering. IEEE Transactions on Geoscience and Remote Sensing, 61: 4400815 [DOI: 10.1109/TGRS. 2023.3237606]

Zhao J Q, Wang H Z, Zhou Y, Zhang D and Zhou Z Y. 2021. Remote sensing image description generation method based on attention and multi-scale feature enhancement. Computer Science, 48(1): 190-196 (赵佳琦, 王瀚正, 周勇, 张迪, 周子渊. 2021. 基于多尺度与注意力特征增强的遥感图像描述生成方法. 计算机科学, 48(1): 190-196) [DOI: 10.11896/jsjkx.200600076]

Zhao R and Shi Z W. 2022. Text-to-remote-sensing-image generation with structured generative adversarial networks. IEEE Geoscience and Remote Sensing Letters, 19: 8010005 [DOI: 10.1109/LGRS. 2021.3068391]

Zheng X T, Wang B Q, Du X Q and Lu X Q. 2022. Mutual attention inception network for remote sensing visual question answering. IEEE Transactions on Geoscience and Remote Sensing, 60: 5606514 [DOI: 10.1109/TGRS.2021.3079918]

Zheng Z D, Zheng L, Garrett M, Yang Y, Xu M L and Shen Y D. 2020. Dual-path convolutional image-text embeddings with instance loss. ACM Transactions on Multimedia Computing, Communications, and Applications, 16(2): 51 [DOI: 10.1145/3383184]

Zhu X J, Goldberg A B, Eldawy M, Dyer C R and Strock B. 2007. A text-to-picture synthesis system for augmenting communication// Proceedings of the 22nd National Conference on Artificial Intelligence. Vancouver: AAAI Press: 1590-1595

Zia U, Mohsin Riaz M and Ghafoor A. 2022. Transforming remote sensing images to textual descriptions. International Journal of Applied Earth Observation and Geoinformation, 108: 102741 [DOI: 10.1016/j.jag.2022.102741]

Advances in remote sensing image-text cross-modal understanding

ZHENG Xiangtao[1], ZHAO Zhengying[2,3], SONG Baogui[1], LI Hao[4], LU Xiaoqiang[1]

1.College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China;
2.School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China;
3.College of Software, Pingdingshan University, Pingdingshan 467000, China;
4.Department of Intelligence, Air Force Early Warning Academy, Wuhan 430019, China

Abstract: With the deep integration of remote sensing technology and artificial intelligence, the human demand for the application of remote sensing data has become increasingly refined. However, single-modal data has limitations in the interpretation of complex scenes. Optical imagery, while rich in spatial information, suffers from weather dependencies; Synthetic Aperture Radar (SAR) data provides all-weather capability but lacks intuitive interpretability; and hyperspectral data, despite its detailed spectral signatures, presents challenges in data redundancy and computational complexity. Therefore, single-modal data is difficult to fully exploit the deeper information in remote sensing images. These limitations underscore the critical need for more advanced analytical approaches that can overcome the constraints of single-modal data interpretation.

For this reason, the collaborative analysis of multi-modal data has become a key way to enhance the interpretation capability of remote sensing and is driving the further development of the remote sensing field. By integrating complementary data sources and leveraging their synergistic relationships, multi-modal approaches enable more robust and comprehensive scene interpretation. Among various multi-modal strategies, remote sensing image-text cross-modal understanding has gained particular prominence as it establishes a vital connection between remote sensing image features and human semantic cognition. This framework enhances visual feature representations with the help of text semantic information and achieves cross-modal information complementarity, which significantly improves the performance of remote sensing interpretation. This paper provides a comprehensive examination of remote sensing image-text cross-modal understanding, which is categorized into four main types of tasks: remote sensing image captioning (referred to as image-to-text), text-to-image generation (referred to as text-to-image), remote sensing image-text alignment (referred to as image-text alignment), and remote sensing visual question answering (referred to as image-text dialogue). From the perspective of cross-modal transformation of images and text, remote sensing image-text cross-modal understanding includes image-to-text and text-to-image. In terms of local content interaction, remote sensing image-text cross-modal understanding includes image-text alignment and image-text dialogue.

This paper begins with a thorough review of the historical development and current state of image-text cross-modal research. The research progress at home and abroad of four tasks, namely image-to-text, text-to-image, image-text alignment, and image-text dialogue, is reviewed. The key technological breakthroughs in the fields of image-text cross-modal are mainly introduced. On this basis, the in-depth analysis is conducted on the technical difficulties faced by the four tasks. Then, it presents a detailed analysis of commonly used public datasets and evaluation metrics for remote sensing image-text cross-modal understanding. Finally, it summarizes the technical challenges in this field, which mainly include three aspects: modal alignment of remote sensing images and text, cross-modal interpretability, and cross-modal reasoning. Based on the proposed problem of remote sensing image-text cross-modal, it looks forward to the future research directions. (1) In-depth mining of cross-modal information (2) Construction of earth science knowledge graph. (3) Human-computer interaction. (4) Large-scale remote sensing image-text model. (5) Language diversity. (6) Remote sensing multi-source data.

This comprehensive investigation not only synthesizes current research but also provides a clear roadmap for future research in remote sensing image-text cross-modal understanding, with potential implications for numerous practical applications in environmental monitoring, urban planning, and disaster management.

Key words: remote sensing image-text cross-modal, image captioning, text-to-image generation, image-text alignment, visual question answering, remote sensing cross-modal datasets

Supported by National Natural Science Foundation of China (No. 62271484); National Science Fund for Distinguished Young Scholars (No. 61925112); Key Research and Development Program of Shanxi (No. 2023-YBGY-225)