中图法分类号:TP751;TP18　　文献标识码:A　　文章编号:1006-8961(2026)07-2381-27

论文引用格式:Li K Y, Cao X Y, Jiang Z X and Meng D Y. 2026. Advances in open vocabulary perception for remote sensing images. Journal of Image and Graphics, 31(7):2381-2407(李开宇, 曹相湧, 蒋梓轩, 孟德宇. 2026. 遥感图像开放词汇感知进展. 中国图象图形学报, 31(7):2381-2407)[DOI:10.11834/jig.260163]

遥感图像开放词汇感知进展

李开宇[1],曹相湧[2*],蒋梓轩[3],孟德宇[4]

1. 西安交通大学软件学院,西安　710049; 2. 西安交通大学计算机学院,西安　710049; 3. 西安交通大学人工智能学院,西安　710049; 4. 西安交通大学数学学院,西安　710049

摘　要: 传统的遥感图像智能解译技术大多建立在封闭集假设之上,高度依赖海量的人工标注数据,且在推理阶段仅能识别训练集中预先定义的固定类别。面对真实地球观测场景中复杂多变的地表环境、尺度剧烈变化的目标以及长尾分布的罕见地物,传统范式泛化能力受限,难以满足高度动态的开放世界解译需求。近年来,得益于视觉—语言基础模型的快速发展,开放词汇感知技术应运而生。该技术通过跨模态语义对齐打破了传统离散标签的束缚,在零样本与少样本场景下展现出强大的泛化潜力。然而,遥感影像独特的俯视成像视角、复杂的拓扑关联以及多源异构的物理模态,致使自然图像领域的通用大模型在向遥感垂直领域迁移时面临显著的领域鸿沟。为此,本文系统梳理并总结了遥感图像开放词汇感知领域的最新研究进展。从数据和方法两个维度,阐述了遥感视觉—语言预训练数据集的构建策略,以及预训练架构从基础域适配向异构数据感知与地理先验增强的演进脉络;全面剖析了开放词汇感知在零样本场景分类、跨模态检索、图像分割、目标检测与定位、变化检测以及三维点云理解等关键下游任务中的应用范式;深入探讨了当前该领域在高质量训练数据匮乏、细粒度评测基准缺失、多源异构模态深层对齐不足及模型可靠性等方面面临的核心挑战,并从多模态大语言模型驱动的生成式感知、全模态基础模型演进、时空因果推演及星地协同计算等方向对未来发展趋势进行了系统展望,以期为推动遥感智能解译迈向真实开放世界提供详实的理论参考。

关键词: 遥感图像;开放词汇感知;视觉—语言模型(VLM);零样本学习;智能解译

Advances in open vocabulary perception for remote sensing images

Li Kaiyu[1], Cao Xiangyong[2*], Jiang Zixuan[3], Meng Deyu[4]

1. School of Software Engineering, Xi'an Jiaotong University, Xi'an　710049, China; 2. School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an　710049, China; 3. College of Artificial Intelligence, Xi'an Jiaotong University, Xi'an　710049, China; 4. School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an　710049, China

Abstract: Remote sensing technology serves as the core mechanism for the observation of the Earth and the understanding of surface environments. It plays an irreplaceable role in critical fields, such as natural disaster monitoring, urban planning, resource exploration, and ecological protection. Driven by the rapid advancement of deep learning over the past

收稿日期:2026-03-30;**修回日期:**2026-04-22;**预印本日期:**2026-04-29

*通信作者:曹相湧　caoxiangyong@mail.xjtu.edu.cn

基金项目:教育部学科先导计划项目(JYB2025XDXM101);国家自然科学基金项目(62272375);国家自然科学基金数学天元基金项目(12426105)

Supported by: Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM101); National Natural Science Foundation of China (62272375); Tianyuan Fund for Mathematics of the National Natural Science Foundation of China (12426105)

decade, the intelligent interpretation of remote sensing images has achieved breakthrough progress in fundamental vision tasks. However, the traditional deep learning paradigm is intrinsically built upon a close-set assumption, i. e. , that models can only recognize a predefined and human-annotated set of fixed categories during the inference stage. When confronted with highly complex surface environments in real-world Earth observation scenarios, dynamic object morphology, and rare ground objects with long-tail distributions, this traditional paradigm not only incurs prohibitive costs for the construction of massive pixel-level annotated datasets but also easily falls into the trap of domain-specific overfitting. Consequently, the generalization and response capabilities of this paradigm are severely challenged by unseen categories or sudden events, making this paradigm inadequate for meeting the highly dynamic interpretation demands of the open world. In recent years, the rapid development of vision-language models has catalyzed a paradigm shift in artificial intelligence from task-specific models into general-purpose perception models. By mapping visual representations and natural language into a unified feature space through contrastive learning on massive image-text pairs, these models have broken the constraints of discrete labels, enabling a direct response to arbitrary natural language prompts. This capability is known as open vocabulary perception. Although this technology has demonstrated remarkable zero-shot generalization and cross-modal reasoning capabilities in the natural image domain, the direct application of these general vision-language models to the remote sensing domain encounters a severe domain gap. The uniqueness of remote sensing data poses multiple challenges to the adaptability of existing models. First, the distinct overhead imaging perspective causes drastic variations in object scale and complex background textures. Second, Earth observation tasks rely on multisource heterogeneous data from synthetic-aperture radar, multispectral or hyperspectral imaging, and thermal infrared sensors. The underlying physical mechanisms of these sensors exceed the inherent inductive biases of models that are pretrained solely on natural RGB images. Third, remote sensing objects often exhibit strong geospatial attributes and complex topological associations. To address these critical challenges, this study provides a comprehensive and systematic review of recent advancements in open vocabulary perception for remote sensing images. We first delve into the foundational aspect of this field: vision-language pretraining for remote sensing. We extensively review the evolution of construction strategies for large-scale datasets. We highlight the transition from limited, human-annotated image-text pairs into massive datasets generated via heuristic rules, the integration of geographic metadata, and advanced multimodal large language models, including innovative approaches that leverage OpenStreetMap and geographical coordinates to produce fine-grained, physics-aware descriptions across multiple modalities. Concurrently, we systematically summarize the progression of pretraining methodologies. Although early approaches have primarily focused on simple domain adaptation through continuous pretraining, recent state-of-the-art frameworks emphasize physics-aware encoding, fine-grained multilevel consistency learning, and geography-enhanced architectures. These frameworks better capture the intricate spatial relationships and modality diversities that are inherent in Earth observation data. Subsequently, this review conducts an in-depth analysis of the adaptation and optimization of open vocabulary perception techniques across a wide spectrum of crucial downstream tasks. For zero-shot scene classification and cross-modal retrieval, we discuss advanced strategies that are designed to mitigate the high intra-class similarity and complex interclass variances typical in remote sensing. We emphasize the shift toward fine-grained local-global alignment, hard negative mining, dynamic soft labeling, and prompt engineering. In the realm of open vocabulary image segmentation, we categorize the existing literature into training-based methods and training-free or annotation-free paradigms. Training-based methods leverage base categories to adapt models while preventing catastrophic forgetting through pseudo-label distillation and knowledge retention mechanisms. Training-free paradigms synergize foundational models, such as CLIP and the segment anything model, to extract structural masks and align semantics without the updating of network weights. For open vocabulary object detection and remote sensing visual grounding, we explore the approaches of researchers to deal with extreme scale variations, arbitrary orientations, and dense object distributions. These approaches include innovative frameworks for pseudo-label generation, multi-scale feature alignment, cross-modality context modeling, and interactive grounding mechanisms. Furthermore, we examine open vocabulary change detection, wherein recent studies employ either combinations of pretrained vision-language models or generative models to generate large-scale data. These approaches aim to identify arbitrary, text-specified surface transitions and simulate complex spatiotemporal changes without reliance on massive and costly bitemporal pixel-level annotations. We also briefly dis-

cuss emerging open vocabulary applications in 3D urban point clouds and cross-domain archeological remote sensing, illustrating the expanding horizon of this technology. Despite remarkable progress, the field of open vocabulary perception for remote sensing remains in a crucial developmental stage and faces several critical bottlenecks. This study critically identifies the limitations of current research, including the severe scarcity of high-quality and geographically balanced training data. This scarcity leads to geographic biases and performance degradation in data-poor regions. In addition, genuinely fine-grained and long-tailed open vocabulary evaluation benchmarks that can accurately reflect the performance of a model are prominently absent in extreme or unknown real-world scenarios. The inadequate physical understanding of heterogeneous modalities and the inherent black box unreliability of current large models in high-stake decision-making scenarios further constrain practical deployments. To chart the course for future research, we outline several promising and essential trajectories. First, we anticipate a paradigm shift toward generative perception that is driven by multimodal large language models. This shift unifies various spatial localization tasks into the direct generation of coordinate sequences or geometric property tokens to utilize fully the logical reasoning capabilities of foundational models. Second, we strongly advocate for the construction of rigorous, real-world, and fine-grained evaluation systems that incorporate complex spatiotemporal logic, diverse geographic conditions, and comprehensive evaluation metrics. Third, the development of omni-modal foundation models that explicitly integrate physical priors and deep learning is considered crucial for the achievement of all-weather and all-spectrum Earth observations, moving beyond pure data-driven approaches. Furthermore, we highlight the necessity to extend perception from static spatial analysis to dynamic spatiotemporal causal reasoning to decode the evolutionary processes of the Earth. Finally, addressing the severe conflict between the massive parameter scale of foundation models and the limited computing power of aerospace edge devices requires focused research into efficient, trustworthy, and safe edge-cloud collaborative computing architectures. By systematically synthesizing these advancements and challenges, this comprehensive review aims to serve as a foundational road map for researchers and practitioners. It accelerates the transition of the intelligent interpretation of remote sensing from isolated, close-set recognition toward artificial general intelligence that is capable of highly reliable, dynamic, and open world perception.

Key words: remote sensing image; open vocabulary perception; vision-language model(VLM); zero-shot learning; intelligent interpretation

0 引　言

遥感技术作为人类观测地球和认知地表环境的核心手段，在自然灾害监测、城市规划、资源勘查及生态保护等国家重大需求领域发挥着不可替代的作用(Yang等，2013；赵忠明 等，2019；Li等，2020)。过去十余年，得益于深度学习技术的飞速发展，遥感图像智能解译在场景分类、目标检测与图像分割等基础视觉任务上取得了突破性进展(Zhang等，2023b；Li等，2025f)。然而，传统深度学习范式本质上建立在封闭集(closed-set)假设之上，即模型在推理阶段仅能识别训练集中预先定义且人工标注过的固定类别。面对真实地球观测场景中高度复杂的地表环境、多变的目标形态以及长尾分布的罕见地物，传统范式不仅需要耗费巨大成本构建海量像素级标注数据集，且极易陷入特定领域的过拟合困境。一旦面对未知类别(unseen categories)或突发事件(如新型军事设施、罕见自然灾害)，封闭集模型的泛化面临严峻挑战，难以满足真实开放世界中高度动态的解译需求。

近年来，随着视觉—语言模型(vision-language model，VLM)的快速发展，人工智能领域的范式逐渐从面向特定任务的模型向通用感知模型转变(Radford等，2021；Cherti等，2023)。该范式通过在大规模图像—文本对上进行对比学习训练，将视觉表征与自然语言映射至统一的特征度量空间，其打破了传统离散标签的束缚，使模型具备直接响应任意自然语言提示的能力，即开放词汇感知(open vocabulary perception)。目前，基于视觉—语言模型的开放词汇感知已在自然图像领域的图像分类、目标检测(Li等，2022；Liu等，2024c；Cheng等，2024)，以及图像分割(Zhou 等，2022；Cho 等，2024；Kombol 等，2025)等任务中展现出强大的零样本泛化与跨模态推理能力。

然而，将通用自然图像领域的视觉—语言模型

直接应用于遥感领域，仍面临着严峻的领域鸿沟。遥感数据的独特性对现有VLM的适应性构成了多重挑战：首先，遥感影像独特的俯视成像视角导致了剧烈的目标尺度变化与复杂的背景纹理，这与自然图像以物为中心（object-centric）的表征分布存在显著差异；其次，真实的地球观测任务不仅依赖光学RGB图像，还需要深度融合合成孔径雷达（synthetic aperture radar，SAR）、多光谱/高光谱以及热红外等全天候、全谱段的多源异构数据，其内在的物理机理超出了以自然图像预训练的视觉—语言模型的固有归纳偏置（Zhu等，2017）；此外，遥感目标往往具有强烈的地理空间属性与复杂的拓扑关联，对细粒度属性的辨识要求高于常规自然图像。因此，如何克服上述挑战，构建契合遥感数据特性与物理机理的开放词汇感知框架，已成为当前地球视觉感知领域的重要命题。

鉴于该领域正处在快速发展的关键阶段，系统性地梳理其研究脉络与技术演进，对推动该领域的理论创新与应用落地具有重要的学术与实践价值。为此，本文全面回顾并总结了近年来遥感图像开放词汇感知领域的最新研究进展。本文将首先阐述遥感视觉—语言预训练的数据集构建与核心方法；其次，系统梳理该技术在分类、检测、分割以及变化检测等关键下游任务中的应用范式；最后，总结当前研究面临的核心挑战，并对未来发展方向进行展望。

1 遥感视觉—语言预训练

遥感视觉—语言预训练通过在大规模图文对数据上进行对比学习或生成式训练，将视觉表征与自然语言映射至统一的特征空间，从而赋予模型在无监督或零样本条件下处理开放世界概念的能力（支元杰 等，2026）。这一范式突破了传统遥感智能解译受限于封闭集预定义类别的瓶颈，已成为实现遥感开放词汇感知的基础。当前，遥感视觉—语言预训练研究正经历从通用模型适配向领域专用基础模型构建的转变。其发展逻辑在于化解两大根本矛盾：一是遥感领域高质量图像与文本配对数据的稀缺与基础模型对海量训练数据需求之间的矛盾；二是遥感多模态数据（如SAR、多光谱等）物理属性的复杂性与现有模型架构多局限于RGB模态之间的矛盾。针对上述问题，本节围绕预训练数据集的构建与预训练方法两个维度，阐述该领域的研究进展及面临的挑战。

1.1 遥感预训练数据集

数据规模、语义多样性与文本质量是决定视觉—语言预训练模型性能上限的核心要素。有别于自然图像领域动辄十亿级别的图文对数据，遥感图像受限于成像视角与专家知识壁垒，面临高质量配对数据匮乏的困境。如表1所示，早期遥感图文数据集主要依赖人工标注，规模普遍较小，代表性数据集如UCM-Captions、Sydney-Captions（Qu 等，2016）与RSICD（remote sensing image captioning dataset）（Lu等，2018）等，仅包含数千至数万幅图像。这类数据不仅规模受限，且描述多侧重于简单的场景分类或单一目标，缺乏对复杂空间关系和上下文的刻画。为打破数据量级的限制，研究者开始利用互联网公开数据与启发式规则构建数据集。例如，RS5M（Zhang等，2024b）通过关键词过滤通用数据集并结合BLIP-2（Li等，2023a）伪标签生成，构建了五百万量级的图文对。但由于数据源于网页和较简单的描述模型，文本存在明显的噪声与语义贫乏问题。随后，SkyScript（Wang等，2024c）利用地理坐标将谷歌地球引擎影像与OpenStreetMap（OSM）的语义标签强关联，构建了260万规模的数据集，虽然在一定程度上丰富了物体属性信息，但基于模板规则拼接的文本仍缺乏自然语言的流畅性与内在逻辑。

随着大语言模型（large language model，LLM）与多模态大语言模型（multi-modal large language model，MLLM）的快速演进，利用先进大模型重写或直接生成文本已成为构建高质量遥感数据集的主流范式，此类策略可显著提升文本的细粒度、词汇丰富度及物理真实性。在基于地理元数据的生成方面，RSTeller（Ge等，2025）提取了OSM中的精细矢量属性，并利用Mixtral模型生成和润色出高度复杂的场景描述；为满足多分辨率生成任务的需求，Git-10M（Liu等，2025a）利用GPT-4o结合OSM数据，显式地将空间分辨率与地理信息融入千万级文本描述中；RS-Landmarks（Barzilai等，2025）则将谷歌地图的地标信息与影像对齐，利用Gemini生成了包含1 800万个带有明确地标指向的高质量图文对。在基于现有视觉标签的语义转化方面，为充分利用现有的检测与分割数据集，RSM-ITD（remote sensing multisource image-text dataset）（He等，2024）设计了标注到描述

表 1　遥感视觉—语言预训练数据集

Table 1　Remote sensing vision-language pre-training datasets

数据集	年份	图像数量	图文对数量	分辨率	空间覆盖	图像模态	注释方法
UCM-Captions	2016	2 k	10 k	0.3 m	美国	光学 (RGB)	人工标注
Sydney-Captions	2016	0.6 k	3 k	0.5 m	悉尼	光学 (RGB)	人工标注
RSICD	2017	10 k	54 k	多种分辨率	–	光学 (RGB)	人工标注
RSTeller	2024	1.2 M	2.5 M	0.6 m	美国	光学 (RGB)	自动标注
RS5M	2024	5 M	5 M	多种分辨率	全球	光学 (RGB)	自动标注
SkyScript	2024	2.6 M	2.6 M	0.1～30 m	全球	光学 (RGB)	自动标注
LuojiaHOG	2024	94 k	94 k	多种分辨率	全球	光学 (RGB)	混合标注
RSM-ITD	2024	210 k	476 k	多种分辨率	全球	光学 (RGB)	自动标注
Git-10M	2025	10 M	10 M	0.5～128 m	全球	光学 (RGB)	自动标注
GAIA	2025	40 k	201 k	多种分辨率	全球	多模态视觉组合	自动标注
HQRS-IT-210K	2025	210 k	1.3 M	多种分辨率	全球	光学 (RGB 与 MSI)	自动标注
DGTRSD	2025	1 762 k	1 762 k	0.08～153 m	全球	光学 (RGB)	自动标注
MMSAR	2025	105 k	296 k	0.1～25 m	全球	SAR	自动标注
SARVLM-1M	2025	345 k	1.7 M	多种分辨率	全球	SAR	自动标注
SAR-TEXT	2025	136 k	136 k	多种分辨率	全球	SAR	自动标注
Llama3-SSL4EO-S12	2025	975 k	975 k	10 m	全球	光学 (MSI), SAR	自动标注
GeoLangBind-2M	2025	2 050 k	2 050 k	多种分辨率	全球	RGB, MSI, HSI, SAR, DEM, Infrared	自动标注
MGRS-200k	2025	200 k	400 k	多种分辨率	全球	光学 (RGB)	自动标注
RSFG-100k	2026	100 k	400 k	多种分辨率	全球	光学 (RGB)	自动标注

注：“–”表示原始数据中未提供相关信息。

与标注到指令算法，引导 Kosmos-2（Peng 等，2023）将边界框或掩码转化为自然语言；HQRS-IT-210K（He 等，2025b）提出多视角生成与融合的两阶段流水线，首阶段利用 Kosmos-2 与 LLaVA-1. 6 生成多视角初始描述，次阶段通过 LLaMA-3 进行融合与消除幻觉，大幅提升了信息密度；DGTRSD（dual-granularity remote sensing image-text dataset）（Chen 等，2025a）则利用 Qwen2. 5-VL 基于现有的短文本生成详细的长文本，构建了双粒度数据集，以解决长尾分布与注意力分配不均的问题。为解决通用大模型缺乏遥感物理先验的问题，GAIA（Zavras 等，2025b）结合专家知识，从权威遥感科学网站抓取带文本的图像，并利用 GPT-4o 生成严谨的合成描述，保障了地球观测任务所需的科学准确性。

除光学 RGB 图像外，遥感全天候、全谱段的特性催生了对 SAR 及多光谱数据的多模态对齐需求。然而，SAR 图像的相干斑噪声与非直观性使得文本标注极度困难。针对 SAR 模态，研究者探索了不同的转化策略。MMSAR（Wang 等，2025d）提出检测到描述算法，将异构的 SAR 目标检测框直接转化为多类别文本描述；SARVLM-1M（Ma 等，2025）基于领域知识和空间模板综合生成了 170 万对 SAR 图文数据。更进一步，SAR-TEXT（He 等，2025a）利用配对的光学与 SAR 图像作为中介，其首先由 MLLM 生成光学图像的高质量描述，随后通过精心设计的提示词与上下文学习机制，引导 DeepSeek-V3 模型将光学描述改写为符合 SAR 散射特征的文本，从而实现了零标注成本下的跨模态知识迁移。在多光谱与高维数据方面，Llama3-SSL4EO-S12（Marimo 等，2026）结合 Overture Maps 标签与 Llama3-LLaVA-Next 模型，为 Sentinel-2 多光谱图像生成了百万级文本标注；而 GeoLangBind-2M（Xiong 等，2025）将图文对齐进一步拓展至 RGB、SAR、多光谱、高光谱、红外及数字高程模型等 6 种异构模态，为全模态遥感基础模

型提供了数据基石。

一些研究反思了纯图文对范式的局限性，提出将地理坐标、地面图像或兴趣点（point of interest，PoI）作为全新的对齐桥梁。Mall等人（2023）收集带有地理位置的互联网地面图像作为卫星影像的中介匹配数据；Jain等人（2025b）则构建了Sentinel-2时间序列与带地理标签地面照片的跨视角数据集，用于支持免文本依赖的生态分类。在多源数据方面，Cambrin等人（2025）整合了结构化的土地覆盖与灾害标签，构建了包含64万样本的SAR与多光谱数据集。值得注意的是，如TorchSpatial（Wu等，2024）所指出的，现有数据集在地理分布上存在严重的地理偏差，未来的构建工作亟需重视全球空间分布的公平性，以避免预训练模型在数据稀缺地区出现性能坍塌。

1.2 遥感预训练方法

遥感视觉—语言预训练的核心目标是在高维特征空间中建立地球观测影像与自然语言描述之间的语义映射关系。由于遥感影像中复杂的空间关系和模态的多样性，直接迁移自然图像领域的对比语言—图像预训练（contrastive language-image pretraining，CLIP）（Radford等，2021）方法往往存在诸多问题。因此，当前的预训练方法正经历从简单的域适应微调向异构数据感知与地理增强的架构演进。早期的研究多沿用双塔架构，通过在遥感图文数据上进行持续预训练缓解领域漂移。RemoteCLIP（Liu等，2024a）、GeoRSCLIP（Zhang等，2024b）和SkyScript（Wang等，2024c）证实了海量遥感数据微调对零样本分类性能的提升。在此基础上，一些研究进一步对特征挖掘与语言容量进行了优化，例如DALIP（distribution alignment-based language-image pretraining）（Wu等，2025a）提出通过匹配图文特征的统计分布（而非单一[CLS]词元）以更好地捕获细粒度信息并泛化至遥感场景；RS-M-CLIP（Silva等人，2024）则结合局部—全局自蒸馏机制强化了视觉表征，并通过机器翻译增强探索了支持十余种语言的遥感多语言预训练。然而，受限于传统预训练模型较短的文本输入长度，此种简单迁移策略可能难以应对遥感图像大幅面、小目标及复杂场景描述的挑战。为此，DGTRS-CLIP（Chen等，2025a）提出了双粒度课程学习框架，通过知识保留扩展技术大幅提升文本编码器的输入容量，并采用从长文本到短文本动态调整权重的三阶段训练策略。该方法有效缓解了长文本带来的注意力弥散问题，在保留细粒度语义理解的同时提炼出了核心的判别特征。更进一步，GeoAlignCLIP（Yang等，2026b）提出了多粒度一致性学习框架，通过区域—短语对齐与难负样本挖掘，结合视觉内部一致性与层级文本一致性约束，有效应对了遥感影像中复杂的空间布局与高类间相似性。配合其构建的细粒度数据集，GeoAlignCLIP显著提升了模型在局部对象与具体属性上的视觉—语义对齐精度。

针对遥感领域丰富的异构模态（如SAR、多光谱、高光谱等），如何将非光学数据对齐至自然语言语义空间是当前研究的难点，主流解决方案分为跨模态知识迁移与物理感知架构设计。在知识迁移方面，核心思路是利用成熟的光学视觉模型引导异构模态特征学习。AlignEarth（Li等，2025e）利用无标注的光学与SAR配对数据，驱动SAR编码器的特征分布向冻结的光学特征空间对齐；SAR-RS-CLIP模型（He等人，2025a）则采用了渐进式微调策略，先在光学图文对上预训练，再迁移至SAR领域，有效缓解了模态鸿沟。在物理感知架构方面，DOFA-CLIP（dynamic-one-for-all CLIP）（Xiong等，2025）提出波长感知动态编码器，通过超网络动态生成适应不同传感器中心波长的卷积权重，并结合多教师模型蒸馏实现全模态支持；Llama3-MS-CLIP（Marimo等，2026）通过将新增的多光谱通道权重初始化为零，实现了从RGB到多光谱特征的平滑过渡；SARCLIP（Wang等，2025d）设计了噪声鲁棒编码与层次化提示学习，强制模型学习对相干斑等物理扰动不敏感的表征，并提出AnyReader架构以低分辨率编码器处理任意尺寸的高分辨率SAR图像。此外，CLOSP（contrastive language optical SAR pretraining）（Cambrin等，2025）提出了以文本为桥梁的统一空间，利用共享的文本编码器隐式拉近了非配对SAR与光学图像的特征距离，实现了多源传感器的融合检索。

遥感影像的强地理属性促使研究者广泛探索地理空间与跨视角知识的融合机制。为突破文本描述的局限，一些研究引入地面图像作为对齐锚点。GRAFT（ground remote alignment for training）（Mall等，2023）放弃了传统的文本监督，直接将卫星图像与同坐标的地面互联网图像的特征进行对比学习，传递性地赋予了模型文本理解能力。SenCLIP（Jain

等，2025a）亦采用类似思路，并引入注意力池化机制聚合同一位置多视角地面图像特征以增强表征的稳健性。TimeSenCLIP（Jain 等，2025b）进一步证明，无需大范围空间上下文，仅利用单像素的光谱时间序列与地面照片进行跨视角对齐，即可实现高精度的零样本生态与土地覆盖分类。另一类研究致力于显式融合地理坐标，如 SatCLIP（Klemmer 等，2025）利用球谐函数对经纬度进行位置编码并融入对比学习，VLM2GeoVec（Aimar 等，2025）则提出单编码器架构，将图像、文本、边界框与地理坐标交织为统一序列进行指令微调。AETHER（AlphaEarth-POI enriched representation learning）（Liu 等，2025c）进一步打破了纯地球观测特征的局限，通过引入兴趣点引导的对比学习，将 AlphaEarth（Brown 等，2025）等基础模型的物理形态特征与反映人类活动的社会经济文本语义进行多模态对齐。地理信息的引入还引发了关于特征干扰的探讨。Cambrin 等人（2025）的研究揭示了地理—语义权衡（geo-semantic trade-off）现象，即引入地理编码（如 GeoCLOSP）虽然能提升对特定地理事件（如灾害）的检索精度，却会稀释模型对一般性语义类别（如农田、森林）的视觉判别能力。

为应对遥感下游任务碎片化导致的模型适配成本过高问题，预训练方法正向多任务统一架构演进。传统的视觉—语言预训练模型多依赖于输出特征向量，需针对下游任务设计专用模型头，限制了其通用性。为此，基于 LLM 或 MLLM 的方法应运而生，并在细粒度感知与复杂逻辑推理等方面进行了一系列探索。在区域级感知层面，EarthMarker（Zhang 等，2025a）引入边界框与点作为视觉提示，通过将提示转化为伪图像并输入混合视觉专家模型构建共享编码机制。为强化语义关联与逻辑稳健性，SkySenseGPT（Luo 等，2024）通过构建覆盖场景图生成的指令集深化了模型对目标拓扑关系的理解；LHRS-Bot（Muhtar 等，2025）则利用志愿者地理信息（volunteered geographic information，VGI）实现语义平衡与课程学习对齐；针对常规微调易诱发模型产生事实性幻觉的缺陷，VHM（versatile and honest vision language model）（Pang 等，2025a）引入包含视觉欺骗性问题的数据集 HnstD，显著提升模型在复杂场景下的诚实度。TEOChat（Irvin 等，2025）首次将视觉指令拓展至时序观测序列，实现了交互式灾害评估；EarthDial（Soni 等，2025）通过构建千万级指令集实现了多源传感器数据的广泛覆盖，但在区分模糊纹理与密集重叠目标方面仍面临挑战。RSUniVLM（Liu 和 Lian，2024）利用语义描述符（Lan 等，2024）将像素级掩码强制转化为纯文本生成；Falcon（Yao 等，2025b）采用轻量级架构，通过坐标量化词元整合了 14 种异构任务，不仅极大地简化了下游微调流程，更通过多层级任务协同训练增强了模型对复杂指令的理解与对全面场景的感知能力。为摆脱专家知识驱动下的认知偏见，GeoZero（Wang 等，2025b）提出冷启动强化学习范式，不使用预定义思维链（chain of thought，CoT），采用答案锚定的组相对策略优化促使模型涌现地理空间推理能力。

综合来看，遥感视觉—语言预训练方法正处于从简单领域适配向底层逻辑重构的转型期。多粒度学习、物理感知编码、跨模态蒸馏及跨视角对齐等方法极大地拓宽了遥感专用模型的感知边界。未来，如何在统一架构内完美平衡物理机理的保持与语义理解的泛化，将是推动遥感基础模型迈向通用人工智能的重要课题。

2　遥感开放词汇感知任务

2.1　零样本场景分类和跨模态检索

零样本场景分类与跨模态检索是验证视觉—语言预训练模型特征对齐能力的基础任务。二者均基于计算图像与文本嵌入在联合特征空间中的相似度以实现预测。这种图文匹配范式使得模型可以无需额外监督微调，即可具备跨模态语义理解与开放词汇感知能力。因此，这两项任务不仅是遥感视觉—语言预训练模型的原生功能，亦是评估遥感基础模型泛化性及模态对齐精度的标准基准。

2.1.1　零样本遥感场景分类

零样本遥感场景分类（zero-shot remote sensing scene classification）旨在解决模型在面对训练阶段未曾见过的地物类别时的识别难题。其核心逻辑在于利用可见类别的样本进行训练，并引入属性列表、词向量或自然语言描述等辅助语义信息建立视觉空间与语义空间之间的映射关系，从而实现识别能力的跨类别迁移（Tan 等，2024b）。各遥感图像零样本场景分类方法对比如表 2 所示。

相较于传统监督学习，该任务摆脱了对大规模标注数据的依赖，为应对遥感领域类别动态更新、长

表2　遥感图像零样本场景分类方法对比

Table 2　Comparison of zero-shot scene classification methods for remote sensing images

/%

方法	年份	AID	EuroSAT	PatternNet	RESISC	方法	年份	AID	EuroSAT	PatternNet	RESISC
ViT-base:						ViT-large:					
CLIP	2021	69.60	32.10	64.10	65.70	CLIP	2021	69.30	41.90	71.40	66.70
RemoteCLIP	2024	87.10	30.70	56.10	67.90	RemoteCLIP	2023	70.90	27.80	61.90	74.30
GeoRSCLIP	2023	70.30	53.40	–	68.80	GeoRSCLIP	2023	74.40	59.90	77.40	73.80
SkyCLIP-50	2024	70.90	33.30	72.20	66.70	SkyCLIP-50	2024	71.70	51.30	80.90	70.90
RS-M-CLIP	2024	**88.90**	25.90	51.00	**92.60**	RSTeller	2024	66.80	56.40	78.90	69.60
RSTeller	2025	63.60	38.40	64.70	58.50	RS-TransCLIP	2025	**80.40**	**72.70**	**93.10**	**86.70**
RS-TransCLIP	2025	78.20	**69.00**	–	79.50	DOFA-CLIP	2025	75.50	59.00	80.20	73.20
DOFA-CLIP	2025	77.60	52.30	**76.70**	67.20	DGTRS-CLIP	2025	–	55.30	76.70	76.30
DGTRS-CLIP	2025	–	48.50	74.60	73.60	HQRS-CLIP	2025	73.90	60.60	–	78.40
DALIP	2025	75.60	64.70	–	72.90	VLM2GeoVec	2025	77.80	39.90	79.80	–
RSCLIP	2025	86.70	48.40	66.70	68.60	Barzilai等人(2025)	2025	72.00	–	–	72.30

注：加粗字体表示每组各列最优结果。“–”表示未在该数据上评测。

尾分布及高昂标注成本等痛点提供了极具现实意义的解决方案。在视觉—语言大模型出现之前，该领域主要致力于构建稳健的视觉—语义对齐投影。早期研究多依赖预训练的语言模型（如Word2Vec、BERT）提取类别语义向量，并通过深度网络学习图像特征的映射函数。例如，ZSRSSC-LP（Li等，2017）利用标签传播算法挖掘同类遥感场景间的视觉相似性；DAN模型（Li等，2021）构建遥感知识图谱以提取强领域相关语义嵌入。然而，此类传统方法的语义空间通常固定且维度较低，加之视觉特征提取器多基于自然图像预训练，导致严重的领域漂移与语义鸿沟，难以精准刻画遥感图像复杂的属性。

随着CLIP等视觉—语言预训练模型的涌现，零样本遥感场景分类的研究范式迎来了向开放词汇分类的变革。这些模型凭借在大规模图文对上的对比学习，天然具备将图像与任意文本描述对齐的能力，使得零样本分类可直接转化为计算图像特征与类别提示词（prompt）之间的相似度，如图1所示。为使该范式适配遥感领域，GeoRSCLIP（Zhang等，2024b）与SkyScript（Wang等，2024c）等工作通过在海量遥感图文数据上进行持续预训练，显著增强了模型对地球观测场景的零样本表征能力。RSplitzero框架（Stacchio，2025）系统评估了多种视觉主干网络与生成式零样本学习方法，证实基于DINOv2（Oquab等，2024）的特征提取器结合生成模型能大幅提升对未知属性组合的泛化性能。在推理时的上下文利用方面，El Khoury等人（2025）指出传统归纳式推理忽略了测试集统计分布的缺陷，提出引入直推式推理策略，利用未标记测试样本的特征结构校准预测结果。该方法在不增加训练成本的前提下大幅提升了分类精度，充分证明了测试时优化策略在遥感零样本任务中的巨大潜力。

对于高光谱、SAR等非光学异构数据以及具备

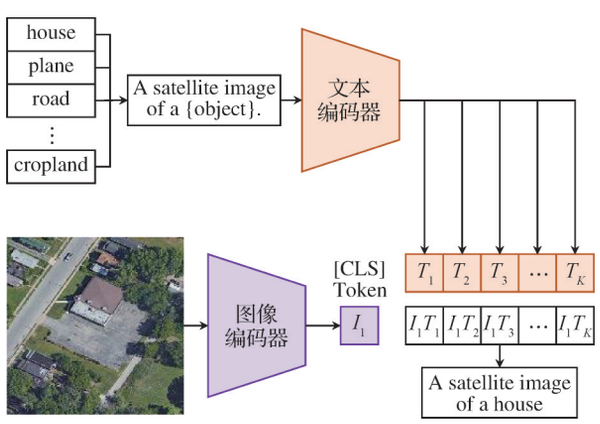


图1　遥感零样本场景分类推理流程

Fig. 1　Remote sensing zero-shot scene classification inference workflow

时间属性的观测序列，直接应用基于RGB预训练的CLIP模型面临着更为严峻的模态鸿沟。针对高光谱图像分类中从图像级理解到像素级精度的跨越难题，HZSCM框架（Huang等，2025a）设计了基于超像素的伪标签生成与光谱引导的标签校正机制，有效修正了视觉—语言模型生成的噪声标签；SPECIAL框架（Pang等，2025b）采用伪标签蒸馏策略，通过降维生成伪RGB图像获取初始预测，进而训练专用的光谱处理网络以规避常规视觉—语言模型无法直接处理多波段数据的问题；SpectralZero（Xia等，2026）则进一步引入大语言模型生成细粒度语义描述，并通过双分支网络实现空间与光谱特征的语义对齐。为进一步弥补多光谱等异构特征的模态鸿沟，Zavras等人（2025a）提出利用PAINT（patching with interpolation）权重插值技术微调CLIP，并通过跨模态知识蒸馏将多光谱特征直接映射到CLIP的共享嵌入空间中。在SAR图像方面，SARCLIP（Wang等，2025d）与SAR-TEXT（He等，2025a）通过构建大规模配对数据进行预训练或微调，成功赋予了模型理解相干斑噪声与雷达散射特性的零样本分类能力。进一步地，Wang等人（2024b）探索了利用视觉—语言模型提取光学遥感图像语义，结合扩散模型生成3D模型进行SAR仿真，并通过动态权重域适应技术缓解仿真与真实数据间的域偏移。在合成孔径声呐领域，Gerg（2026）证实了仅利用思维链文本提示描述目标的几何与阴影特征，视觉—语言模型即可在非消费级成像领域实现有效的零样本检测。此外，针对农作物分类等强时序依赖任务，Jain等人（2025b）证实仅利用单像素的光谱—时序特征与文本描述对齐即可实现高精度零样本分类，这揭示了在特定遥感任务中，时序演变信息相较于空间几何特征可能具有更强且更本质的零样本判别力。

2.1.2 遥感跨模态检索

遥感图像文本跨模态检索旨在使用户能够通过自然语言描述检索目标图像，或为遥感图像匹配精确的文本描述（Xu等，2025）。早期的遥感跨模态检索研究主要依赖于卷积神经网络与循环神经网络的组合，分别提取图像与文本的特征并将其映射至共享的度量空间中。然而，这种范式在面对遥感图像尺度多变、背景复杂且目标密集的场景时，通常仅能捕捉粗粒度的语义对应关系。随着视觉—语言模型的广泛应用，该领域的研究范式发生了变化。

针对预训练模型在遥感域迁移中面临的域偏移问题，一些研究开始关注遥感数据中普遍存在的类间相似性高的现象。由于遥感图像多为俯视视角，不同地理位置的场景在视觉纹理和文本描述上往往表现出较高的相似度。为了缓解弱相关样本对比学习的干扰，“先剔除后对齐”策略被提出（如EBAKER（eliminate before align strategy with keyword explicit reasoning framework）（Ji等，2024）及其改进版本iEBAKER（Zhang等，2026b）），其逻辑是在特征对齐前，直接在批次中剔除低全局相似度的正样本对。相关性引导的自适应学习机制（Chen等，2025b）则通过高斯混合模型对样本的语义相关性进行动态加权。然而，这类基于阈值截断或先验分布假设的机制存在一定的局限性：其在剔除噪声的过程中，可能误删视觉特征难以分辨但信息量丰富的困难样本（hard samples），进而限制模型在处理细微类内差异时的判别能力。为了保留困难样本并减少语义歧义，一些研究尝试引入外部先验或改进对比学习策略。例如，KTIR（Mi等，2024）和PriorCLIP（Pan等，2024）分别通过引入外部知识图谱和视觉场景分类先验来丰富语义表达。但是，通用知识库通常缺乏对遥感细粒度目标的领域内表示，使得知识注入的增益在复杂场景下受到限制。此外，针对模态内高相似度引发的困难样本难区分问题，DCCA（discriminative and consistent cross-modal alignment）（Song等，2026b）提出了具有负样本对扩展机制的全局对比学习策略，结合领域信息注入来强化辨别力；MIIA（masked interaction inferring and aligning）（Zhao等，2024）则采用掩码交互推理与动态对比学习，旨在无需额外监督的前提下增强样本间的区分度。为了进一步缓解高语义相似性给对比学习带来的二值标签噪声，DSD-RSITR（Cheng等，2025）提出了一种动态自蒸馏框架，利用动态更新的教师模型生成软标签，从而在低训练成本下引导模型学习更准确的全局语义关系。

除了全局特征优化，探索细粒度的局部语义对齐是提升检索精度的另一重要方向。遥感文本不仅描述全局场景，还常涉及具体的空间位置、几何拓扑和多目标关系。为此，一些研究提出了多种局部—全局协同对齐的架构。例如，CLGSA（context-aware local-global semantic alignment）（Chen等，2025c）和PR-CLIP（Guan等，2025）通过引入局部掩码重建机

制和跨模态位置信息重构，引导模型学习局部图像块与文本词元之间的细粒度映射；CPPMN（cross-modal progressive perspective matching network）（Zheng等，2025a）考虑了遥感图像的地理空间多视角感知特性，构建了渐进式的视角匹配网络；WSSCN（whole semantic sparse coding network）（Zheng等，2025b）利用稀疏编码技术解耦多重底层语义。FGR-GA（Yu等，2026）则通过特征解缠和组感知对齐，在宏观和微观层面同时约束表示空间。此外，RSITR-FFT（remote sensing image-text retrieval fine-grained fine-tuning）（Xiu等，2024）通过细粒度的词—区域对齐与一致性正则化在有限数据下实现了高效微调。尽管这些方法在公开数据集上取得了较高的评价指标，但其实用性仍面临计算复杂度的挑战。复杂的跨注意力计算、图神经网络等显著增加了模型的推理延迟和显存开销，这与大规模跨模态检索系统对检索效率的实际需求相悖。为缓解效率瓶颈，SGPD（sparse-guided partial dense）（Zhao等，2025c）提出了一种稀疏引导的局部稠密检索范式，旨在检索效率与精度之间取得平衡。

在面对真实应用中的复杂域偏移与多层级语义时，模型的泛化能力常常受限。UrbanCross（Zhong等，2024）探讨了不同国家城市景观间的地理迁移问题，指出常规模型在跨域场景下性能会显著下降，并引入了域适应机制进行修正。为了进一步提升模型对遥感特有数据特征的适应性，一些研究尝试结合提示工程（prompt engineering）与更具挑战性的数据集。Sun等人（2025）提出了强弱提示工程，通过注意力机制与预训练分类模型分别生成细粒度与全局语义提示，以增强模型对多尺度目标的特征捕获能力。同时，为了弥补传统数据集在地理空间感知与描述粒度上的不足，LuoJiaHOG（Zhao等，2025a）等数据集被提出，提供了层次化的地理标签与密集的文本描述，为验证细粒度对齐和跨域泛化方法构建了更加严谨的基准。

2.2 遥感开放词汇分割

遥感图像分割是地球观测领域的基础视觉任务，旨在为影像像素分配具备明确地理或物理意义的语义标签。各遥感图像开放词汇语义分割方法对比如表3所示。

表3 遥感图像开放词汇语义分割方法对比

Table 3 Comparison of open-vocabulary semantic segmentation methods for remote sensing images

/%

类别	方法	年份	数据集						
			LoveDA	iSAID	Postdam	Vaihingen	UAVid	VDD	UDD5
基于训练的方法	Cat-Seg	2023	36.9	21.7	47.1	29.1	–	–	–
	OVRS	2025	28.7	39.1	27.5	33.7	25.2	37.3	39.1
	GSNet	2025	29.3	42.0	26.5	35.2	25.4	38.1	40.1
	RSKT-Seg	2025	32.5	**44.0**	34.5	37.2	28.1	**41.2**	43.0
	SGSeg	2025	26.4	42.3	32.5	37.8	–	–	–
	FarSLIP	2025	34.2	18.7	47.5	22.8	**40.0**	40.8	**44.8**
	ROSS（Zhang等，2026）	2026	–	40.8	30.4	34.0	–	–	–
	SkySense-O	2025	**38.3**	43.9	**54.1**	**51.6**	–	–	–
免训练或免标注的方法	SegEarth-OV	2025	36.9	21.7	48.5	40.0	42.5	45.3	50.6
	RSCLIP	2025	38.0	–	47.4	28.9	–	–	50.8
	SegEarth-OV-3	2025	**47.4**	27.6	**57.8**	**60.8**	**54.7**	**64.5**	**71.7**
	AlignCLIP	2026	39.5	23.6	47.9	34.5	44.4	48.4	51.8
	Sosa等人（2026）	2026	38.2	21.9	50.2	40.6	44.3	46.8	53.8
	ReSeg-CLIP	2026	–	–	38.3	–	–	–	43.2
	SDCI	2026	43.7	**46.4**	47.5	50.1	–	–	–

注：加粗字体表示每组各列最优结果。“–”表示未在该数据上评测。

传统的遥感图像分割多建立在封闭集合假设之上，即推理阶段仅能识别训练集中预定义的类别。然而，真实地表环境具有高度动态性与复杂性，常涌现出未知的基础设施、罕见地貌或突发灾害特征。为突破预定义类别的局限，开放词汇分割（open-vocabulary segmentation）应运而生。该任务依托自然语言的无限表达能力，旨在根据任意文本提示对图像中已知（seen）与未知（unseen）类别进行像素级定位与分类，如图 2 所示。其底层逻辑在于利用视觉—语言基础模型（如 CLIP）在海量图文预训练中构建的跨模态对齐空间，将分割任务转化为像素与文本特征的相似度匹配问题。

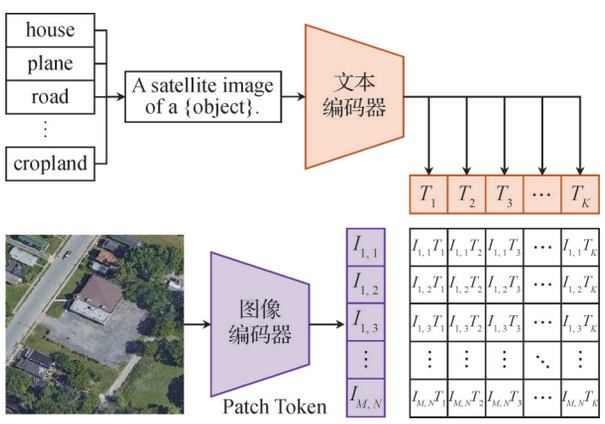


图 2　遥感开放词汇分割推理流程

Fig. 2　Remote sensing open-vocabulary segmentation inference workflow

根据现有研究，遥感开放词汇分割的研究方法可粗略分为两类：基于部分分割数据集训练（training-based）的方法与免训练或免标注（training-free/annotation-free）的方法，如图 3 所示。

基于训练的遥感开放词汇分割方法，其核心动机是利用包含基础类别的既有遥感分割数据集对通用视觉—语言模型进行微调或知识蒸馏，使其特征空间在维持对新颖类别零样本识别能力的同时，适配遥感图像的物理与几何特性。

针对遥感目标方向剧烈变化的问题，Cao 等人（2025）提出旋转聚合相似度计算模块，通过融合多方向特征并结合多尺度机制，构建了旋转与尺度不变的语义表示。

为弥合自然图像与遥感图像间的领域鸿沟，一些研究探索了双流架构以融合通用泛化性与遥感领域先验。例如，GSNet（Ye 等，2025）与 RSKT-Seg（Li

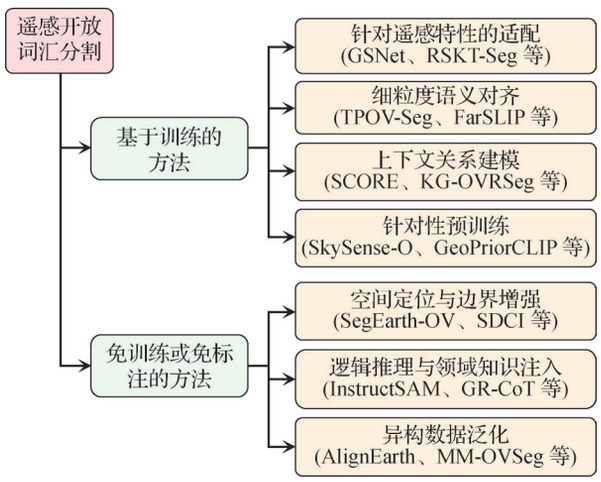


图 3　遥感开放词汇图像分割方法

Fig. 3　Remote sensing open-vocabulary image segmentation methods

等，2025a）均采用双流图像编码器，将通用 CLIP 与在遥感数据上自监督预训练的视觉模型并行提取特征，并通过查询引导或多方向成本图聚合策略实现多源特征融合；AerOSeg（Dutta 等，2025，2026）将 SAM（segment anything model）引入训练流程并设计语义反投影模块，以防止特征细化过程中的泛化能力流失。ZoRI（zero-shot remote sensing instance segmentation framework）（Huang 等，2025b）提出知识保持适应策略与先验注入机制，利用视觉原型缓存库缩小领域表征差异。

在细粒度语义对齐方面，TPOV-Seg（Zhang，2025b）引入特定文本模板生成器以丰富遥感目标的文本语义，并辅以文本感知的提示微调策略；FarSLIP（fine-grained aligned remote sensing language-image pretraining）（Li 等，2025j）则指出全局自蒸馏会破坏 CLIP 的语义连贯性，进而提出“Patch-to-Patch”局部蒸馏框架，显著增强了对微小目标的特征辨别力。此外，HG-RSOVSSeg（Huang 等，2026a）通过分层的视觉特征解码器和多模态特征聚合模块，进一步强化了模型在像素级对齐上的精确性。而 SGSeg（An 等，2025）则提出一种软引导策略，通过一个冻结的编码器补偿微调过程中可能出现的泛化能力损失，从而在适应下游任务的同时保留模型的通用性。

为解决遥感密集预测中普遍存在的类别歧义与上下文依赖问题，一些研究进一步引入了复杂的逻辑结构与知识约束。SCORE（Huang 等，2025c）强调

了场景上下文在实例分割中的关键作用，通过提取全局与区域上下文动态增强类别嵌入，有效区分了视觉相似但空间分布各异的目标；TACOSS（text as supervision for contrastive semantic segmentation）（Zermatten等，2025）借助文本数据增强策略提升了跨数据集标签体系的迁移能力；KG-OVRSeg（Huang等，2026b）则引入知识图谱，通过建模同义词与上下位词关系生成结构化类别嵌入，缓解了孤立词汇匹配引发的语义混淆。

面对高质量像素级标注成本高昂的问题，T2ASeg（Wang等，2025g）与LandSegmenter（Liu等，2025b）探索了基于文本或廉价含噪土地覆盖产品的弱监督训练与置信度引导融合策略。

在整体架构概念的拓展上，FreeMix（Wu等，2025b）进一步将开放词汇分割问题细化至开放词汇领域泛化（open-vocabulary domain generalization，OVDG）；RemoteSAM（Yao等，2025c）将多类视觉任务统一于指代表达分割范式，极大地增强了复杂指令理解能力。此外，为推动原生遥感分割大模型的发展，SkySense-O（Zhu等，2025a）构建了全像素精细标注的Sky-SA数据集，并引入以视觉为中心的建模原则与图谱正则化来纠正图文对齐歧义；GeoPrior-CLIP（Liang等，2026）则通过地理感知跨模态注意力模块将地图拓扑先验直接注入视觉编码器，显著增强了模型对地物边界的空间感知。

尽管基于训练的方法在多项基准测试中取得了显著的性能增益，但从方法论层面审视，其实际应用仍受限于若干固有缺陷。首先，此类范式本质上仍依赖源域基础类别的像素级标注，削弱了开放词汇分割旨在降低标注成本的目标；其次，利用闭集数据微调基础模型极易引发灾难性遗忘，即模型在拟合遥感特定特征时容易向高频基础类别发生模型坍塌，从而抑制了对长尾或未知类别的泛化响应。此外，当面对与训练集特征分布差异巨大的极端场景时，特定微调模型易陷入过拟合困境，其泛化稳定性的上限被微调数据的多样性严格约束。

为彻底摆脱像素级标注依赖并最大化保留视觉—语言模型原生开放语义空间，免训练或免标注范式成为遥感开放词汇分割的另一重要演进分支。该范式主张冻结原始的权重以直接利用其预训练的图文匹配能力，并引入其他视觉基础模型（如SAM或DINO）提供的几何与结构先验，以弥补CLIP等模型在密集预测及局部定位上的短板。

针对CLIP特征图分辨率低且存在全局偏置导致的边界粗糙问题，SegEarth-OV（Li等，2025g）引入无需后训练的上采样器恢复空间细节，并设计全局偏差缓解操作以提升局部语义保真度。这一思想也启发了后续的工作，RSCLIP（Wang等，2025f）通过邻域感知块和多头多尺度注意力机制，进一步增强了对CLIP内部特征的操控能力；而AlignCLIP（Liao等，2026）则通过自引导对齐和文本特定的视觉原型缓解跨模态不匹配问题。

为了进一步强化空间边界，利用SAM等模型提供高质量类无关掩码进行引导成为一种高效途径。ReSeg-CLIP（Heidarianbaei等，2026）在CLIP视觉编码器的多个层级中引入SAM生成的分层注意力掩码，从而在不同尺度上硬性约束了注意力的发散。Text2Seg（Zhang等，2024a）系统性地比较了Grounding DINO（Liu等，2024c）与SAM等基础模型的多种串联方式，为免训练流水线的设计提供了经验指导。SDCI（Wang和Ni，2026）则提出了一种双向交叉感知机制，在特征提取阶段让CLIP的语义与DINO的结构进行注意力交互，并在推理时利用超像素结构进行凸优化，实现了高度几何精确性的边界锐化。此外，随着MLLM的发展，复杂指令理解被引入分割流程。InstructSAM（Zheng等，2025c）摒弃了传统的置信度阈值过滤，转而利用MLLM预测图像中目标类别的存在性与数量，并将其作为二分类整数规划的硬约束条件进行掩码—标签匹配。SegEarth-OV3（Li等，2025h）则直接探索了最新一代的SAM 3（Carion等，2026），利用其解耦的存在性检测头过滤场景中不存在的类别，从而减少了在大词汇表检索时假阳性错误。

对于免训练方法而言，其核心挑战在于面对遥感细粒度类别时，通用大模型的原生识别能力极其有限。由于缺乏遥感视角的微调，模型很难直接区分“不透水层”、“裸土”或具有特定属性的农田等地物。如何在完全不进行参数更新的前提下，将专业的领域知识精准嵌入，是当前免训练范式面临的重要挑战。针对这一问题，GR-CoT（geospatial reasoning chain of thought）（Zhou等，2026）引入MLLM的思维链推理，通过宏观场景锚定与视觉特征解耦动态生成图像自适应词汇表，利用逻辑先验缓解光谱相似地物间的分类歧义。此外，另一个挑战在于向异

构数据模态的泛化与拓展。真实的地球观测任务依赖SAR、红外以及多光谱等多维度数据，而现存的强泛化性的视觉—语言基础模型大多基于自然场景的RGB图像进行预训练，它们在面对异构数据时往往发生退化甚至失效。针对这一问题，现有研究已展开初步探索。例如，AlignEarth（Li等，2025e）提出了一种跨模态知识蒸馏机制，在无需文本标注的前提下，将光学视觉—语言模型强大的开放语义空间有效迁移至SAR特征编码器中；MovSeg（Ji等，2026）则针对多光谱数据的物理独特性，引入多光谱输入适配模块，依托参数高效微调策略将近红外频段信息融入开放词汇感知框架。进一步地，MM-OVSeg（Wei等，2026b）构建了光学与SAR多模态融合的开放词汇分割框架，其通过跨模态统一过程将SAR特征与通用视觉模型的RGB表征空间对齐，并借助双编码器融合模块整合多源密集特征与全局文本语义，实现了复杂天气环境下的鲁棒感知。尽管上述工作提供了启发性的解题思路，但要构建具备全天候、全谱段感知能力的遥感开放词汇系统，仍需进行更为深远的探索。

2.3 遥感开放词汇检测和定位

遥感开放词汇目标检测（open-vocabulary object detection）与遥感视觉定位（visual grounding）在技术基础与应用目标上具有高度相关性。前者旨在借助文本提示（通常为类别名称）识别并定位未知类别目标，侧重于类别级语义与候选区域之间的对齐；后者则要求模型根据包含空间关系、属性特征或功能描述的自然语言表达，精确定位特定目标实例，更强调面向实例级对象的细粒度语言理解与定位。Zhou等人（2025）指出，在遥感与无人机场景中，实际应用不仅需要应对开放环境中的未知类别，还需要结合多样化语言描述完成目标识别与定位，这表明两类任务的协同建模具有明确的现实需求。与此同时，GLIP（grounded language-image pre-training）（Li等，2022）和Grounding DINO（Liu等，2024c）等工作表明，目标检测可以被重构为视觉定位任务，从而在预训练阶段将检测与定位纳入统一的视觉—语言对齐框架之中，如图4所示。这一思路也为遥感场景下开放词汇检测与视觉定位的融合研究提供了重要启发。

早期相关研究主要依赖预训练视觉—语言模型开展知识蒸馏与伪标签生成，以在尽量减少人工标

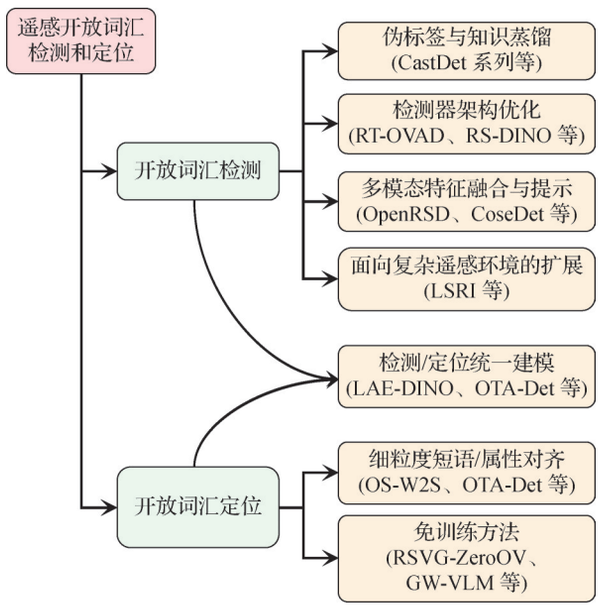


图4 遥感开放词汇目标检测和视觉定位方法

Fig. 4 Remote sensing open-vocabulary object detection and visual grounding methods

注的前提下扩展检测器词汇表。该类方法通常将通用大型视觉—语言模型作为教师模型，以指导学生检测器学习新类别。Li等人（2025i）提出CastDet框架，利用定位教师和外部分类教师构成多教师自学习机制，以生成高质量伪标签；进一步地，该方法通过动态标签队列对伪标签进行维护与采样，从而提升模型对新类别的学习能力。在后续工作中，Li等人（2026b）进一步针对遥感目标方向任意的特点，引入尺度抖动与角度抖动方差等伪框筛选策略，并结合旋转边界框表示，将该自学习框架扩展到面向旋转目标的开放词汇检测任务。针对伪标签生成过程中对新旧类别严格分离这一理想化假设，Wang等人（2025e）提出基于选择性掩码的教师—学生框架，通过遮挡基类区域，使模型能够在单阶段训练中直接利用包含混合类别的部分标注数据，从而提升部分标注样本的利用效率。针对遥感目标尺度变化带来的泛化挑战，Li（2026）将开放词汇检测的泛化过程解耦为视觉空间转换与词汇扩展两个阶段，并通过级联知识蒸馏逐步缓解视觉空间与语义空间之间的对齐困难。与此同时，Yao等人（2025a）提出VK-Det（visual knowledge-guided open-vocabulary object detection），引入原型感知的伪标签生成机制，通过特征聚类构建类别原型并进行匹配，从而减少对外部文本监督的依赖，并缓解由文本引入的语义偏差。

总体来看，基于伪标签与蒸馏的方法有效推动了遥感开放词汇检测的发展，但其性能仍在较大程度上依赖启发式置信度阈值与教师模型质量，容易放大早期预测偏差，并对低显著性目标的召回性能造成不利影响。

为降低对伪标签质量的依赖，后续研究开始更多地从检测器结构设计出发，通过增强跨模态特征融合与多模态上下文提示来提升开放词汇检测的稳健性。针对无人机视角下小目标密集、背景复杂的挑战，Ju等人（2026）提出RS-DINO。该方法在视觉编码阶段引入多尺度大核注意力机制，以同时增强全局与局部信息建模能力；在多模态解码阶段，则结合卷积门控前馈网络，以提升密集场景中的特征表达能力与跨模态融合稳定性。另一方面，Weng等人（2025）提出轻量级CAGE（cross-attention gated enhancement）模块，通过交叉注意力门控与全局调制在极低计算开销下实现了图文对齐。RT-OVAD（Wei等，2024）提出轻量化的图像—文本协同建模框架，由图文协作编码器和文本引导解码器构成，通过跨模态交互同时增强视觉与文本表征，并引导查询聚焦于类别相关特征。从多模态上下文提示角度，OpenRSD（Huang等，2025d）和RS-MPOD（Yang等，2026a）通过同时引入图像提示与文本提示实现开放式检测，其中图像提示能够在缺乏精确类别描述时提供视觉先验，从而减轻用户对专业遥感类别知识的依赖。Hwang和Woo（2025）提出FASE（feature-aligned scene encoding）框架，通过特征对齐场景编码，将领域特定场景上下文与通用文本特征进行深度融合。Gu等人（2025）提出CoseDet，通过对周边区域进行显式建模来捕获遥感场景中的关键上下文依赖，并利用伪词机制实现区域特征与文本语义空间的对齐。LLaMA-Unidetector（Wang等，2025c）采用解耦式设计，将类不可知定位与MLLM分类过程分离，从而更充分地利用大模型的零样本推理能力。

除检测器结构优化外，针对遥感领域特有挑战，一些研究融合领域先验与细粒度特征，提升模型在遥感开放环境中的适应性。针对大幅视角变化带来的挑战，Kini等人（2025）提出跨视角对齐框架，通过多实例语义关联实现跨视角特征迁移。在多传感器数据融合方面，Wang等人（2025h）构建了首个对齐的RGB—红外低空开放词汇检测数据集，并提出LSRI模型，利用热辐射信息与纹理信息的互补性提升小目标检测性能。Li等人（2025b）则通过构建ShipSem-VL数据集，引入细粒度语义描述，以增强模型对目标属性的理解能力。尽管这些研究从架构与数据两个层面拓展了开放词汇遥感感知的研究边界，但一些系统评估结果表明，现有视觉—语言模型在细粒度类别区分任务中的性能仍会显著下降，而上下文信息在不同场景下也可能同时带来正面与负面的影响（Ouerghemi等，2026）。进一步地，Yang等人（2026a）指出，遥感场景中的文本提示容易受到语义歧义与类别定义差异的影响；尽管多模态提示可以提升模型鲁棒性，但其性能仍会受到跨数据集分布偏移的制约。可以发现，模型性能的下降并非完全源于视觉表征不足，还与语言提示歧义、上下文依赖失衡以及跨域分布偏移密切相关。这些现象表明，当前方法中可能存在语义伪对齐，即模型更倾向于依赖全局语义或主导词汇进行匹配，而未能充分建模目标的细粒度属性特征。

针对上述问题，一些研究进一步关注细粒度视觉—语言对齐以及统一数据与任务框架的构建。Yang等人（2026b）提出GeoAlignCLIP，通过多粒度对比学习和多视图一致性学习，在区域—短语层面建立更为精确的视觉—语言对应关系。为支撑统一建模，Wei等人（2025）提出OS-W2S自动标注引擎，并构建MI-OAD数据集。该数据集覆盖词、短语和句子三级语言描述，共包含16万幅图像和200万个图文对。在此基础上，Wei等人（2026a）提出OTA-Det（open-text aerial detection），将开放词汇目标检测与遥感视觉定位纳入统一框架，并通过图像级注释聚合与属性级数据分解实现更密集的细粒度语义对齐，从而缓解语义伪对齐问题。类似地，Pan等人（2025）提出LAE-Label数据引擎，并结合动态词汇构建和视觉引导文本提示学习策略，训练得到LAE-DINO这一开放词汇基础检测模型。与前一阶段以单任务优化为主的研究不同，这类工作更加强调从数据构建、语义粒度控制与任务统一建模3个层面系统提升模型的开放环境感知能力，为遥感开放词汇检测与视觉定位的一体化研究奠定了基础。

近期，免训练范式也开始被引入开放词汇遥感感知任务。Hu等人（2026）提出KGCS，通过场景描述模块构建判别性语义描述字典，利用结构感知的双路径目标提议策略缓解复杂结构目标的分割碎片

化问题，结合图文相似度模块完成自适应筛选，从而在严格零样本设定下取得了较好的检测效果。Zhu等人（2026）提出GW-VLM，通过多尺度视觉—语言搜索将类别不可知区域的多尺度视觉搜索结果软对齐为文本片段，并作为上下文提示使视觉—语言模型与大语言模型以“Guess What”式交互完成开放词汇检测推理。Li等人（2026a）提出RSVG-ZeroOV，结合视觉—语言模型的交叉注意力图与扩散模型的自注意力图，并通过总览—聚焦—演化策略实现零样本开放词汇遥感视觉指代表达定位。不过，这类免训练方法的能力边界在很大程度上仍取决于其所依赖基础模型的泛化能力与细粒度感知能力，同时，多模型串联的推理流程通常也会带来额外的计算开销与时延。因此，尽管该方向在弱监督甚至无训练设定下展现出较强潜力，其实际应用价值仍有待在更复杂遥感场景中进一步验证。

2.4 遥感开放词汇变化检测

遥感图像变化检测（change detection）旨在通过对比同一地理区域在不同时期获取的影像，精准识别地表的动态演变。遥感开放词汇变化检测方法如图5所示。

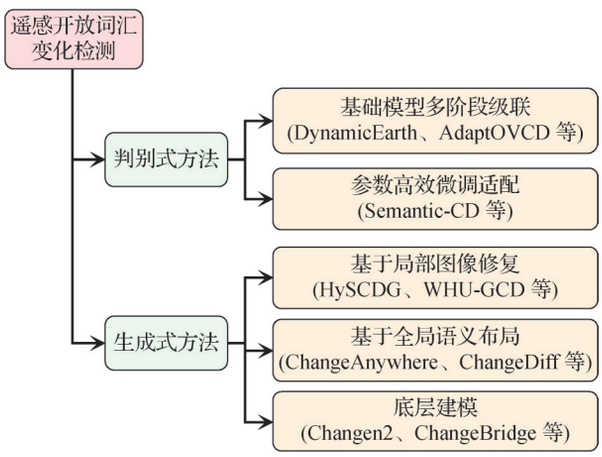


图5 遥感开放词汇变化检测方法

Fig. 5 Remote sensing open-vocabulary change detection methods

传统数据驱动型变化检测模型高度依赖预定义的类别体系与大量像素级标注，仅能在闭集场景下运作（Li等，2024a，2025d；Liu等，2024b）。然而，真实世界的地表变化具备高度的多样性与不可预见性。为突破固定类别设定的局限，开放词汇变化检测（open-vocabulary change detection）应运而生。该任务利用自然语言描述作为引导，要求模型在无需特定类别标注的前提下，灵活定位并识别任意文本指定的变化类型，如图6所示。当前，围绕开放词汇变化检测任务的研究主要沿两条相辅相成的逻辑脉络演进：其一是直接利用视觉基础模型与视觉—语言模型开展免训练或参数高效的开放词汇推理；其二是借助生成式大模型合成海量且多样的变化检测数据，从数据引擎层面增强下游深层检测器的泛化能力。

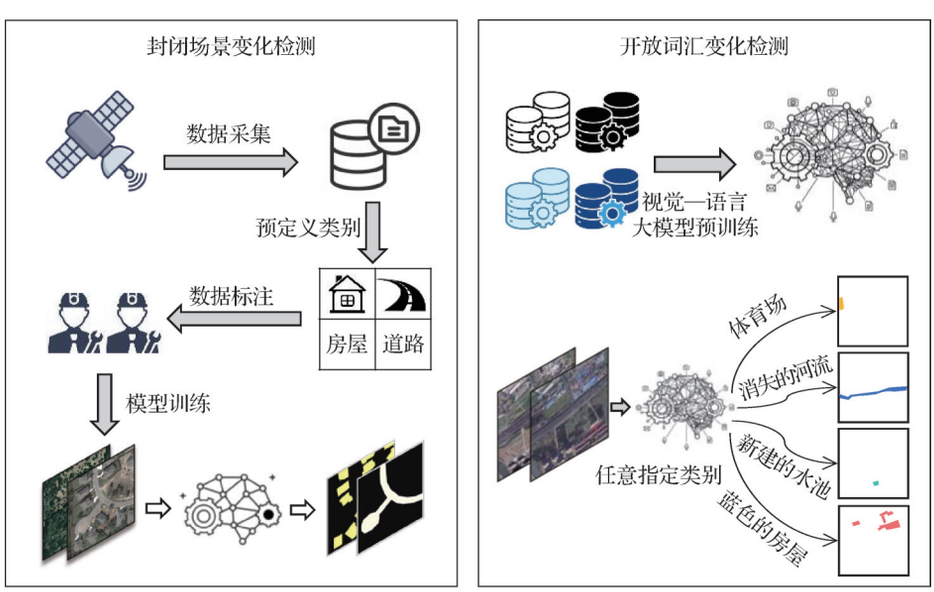


图6 封闭场景变化检测和开放词汇变化检测的对比

Fig. 6 Comparison of closed-scene change detection and open-vocabulary change detection

在直接开展开放词汇变化检测的路径中，早期探索主要聚焦于通用视觉基础模型的多阶段级联。AnyChange（Zheng等，2024）率先探索了零样本变化检测，利用SAM在潜在特征空间内进行双时相匹配，实现了类别无关的变化定位。为进一步引入语义识别能力，SCM（Tan等，2024a）将FastSAM（Zhao等，2023）与CLIP予以结合，提出分段语义注意力机制，从而在无标注条件下利用文本提示过滤伪变化并精准识别特定目标的动态变化。随着研究不断深入，DynamicEarth（Li等，2025c）首次对开放词汇变化检测任务进行了形式化定义，抽象出掩码—比较—识别（M-C-I）与识别—掩码—比较（I-M-C）两大通用推理框架，并系统评估了多种异构模型的组合表现。然而，这种基于模块拼接的级联范式存在一定的局限性，即各子模型在特征空间上未充分对齐，且预训练数据缺乏遥感视角的领域感知，极易在多阶段推理过程中引发误差的级联累积。

为缓解异构模型级联导致的误差传播，后续研究逐步向多层次特征对齐与统一融合架构演进。AdaptOVCD（Dou等，2026）提出了双维度多级信息融合架构，在数据级引入自适应辐射对齐，在特征级结合边缘先验动态划分变化边界，并在决策级执行置信度校准，在无需训练的前提下实现了异构模型的协同。UniVCD（unified open-vocabulary change detection）（Zhu和Yang，2025）则构建了轻量级特征对齐模块，将SAM2（Ravi等，2025）的高分辨率空间细节与CLIP的语义先验桥接至统一表示空间，有效保留了地物的边界细节并注入了丰富的开放词汇语义。为进一步简化推理流程，OmniOVCD（Zhang等，2026a）借助SAM3的提示概念分割能力，使用协同融合与实例解耦策略，在生成高质量的土地覆盖掩码后，基于拓扑连接性将其解耦为独立实例进行双时相匹配。此外，部分研究侧重于参数高效微调适配，如Semantic-CD（Zhu等，2025b）将任务解耦为二值变化与语义变化检测，在适配CLIP视觉编码器的基础上引入开放语义提示器以构建精细的语义代价空间。

鉴于高质量双时相语义标注获取成本极高，利用生成模型合成变化数据以反哺下游检测器，逐渐成为开放词汇变化检测领域中一个有潜力的方向。早期的合成方法多依赖于三维图形学或基础的生成模型。例如，SyntheWorld（Song等，2024）结合3D程序化建模与扩散模型生成的纹理，构建了全合成的土地覆盖与变化数据集。随后，基于稳定扩散模型（stable diffusion）的局部图像修复（inpainting）技术广泛应用于单时相图像的编辑，以生成伪双时相数据。PRISM（Cho等，2025）通过多任务预训练模型（Wang等，2024a）提取单时相图像中的建筑物伪标签，结合形态学操作扩展修复区域，并利用文本提示驱动扩散模型，在无真实双时相数据的情况下重绘建筑物变化。HySCDG（Benidir等，2025）则利用扩散模型与ControlNet（Zhang等，2023a），在真实土地覆盖语义图的引导下，对选定的地理实例执行精确的局部重绘，从而构建了大规模的混合数据集。然而，纯合成或单纯局部重绘的数据往往存在背景高度一致的缺陷，易导致下游模型对光照、季节等非语义环境差异产生过拟合。为此，WHU-GCD（Zan等，2025）利用CLIP等模型对扩散模型进行两阶段微调，并在生成策略中引入了真实环境中未发生变化的短时距双时相图像作为负样本；同时结合SAM与局部修复模型，在真实双时相图像的同质区域内直接重绘新目标。该策略不仅实现了背景的复杂化，还显著增强了检测器在真实光照与季节差异等干扰下的鲁棒性。

尽管基于局部重绘的方法取得了一些进展，但其本质仍局限于对既有实例的替换，难以准确刻画真实地表演变中全局环境的协同变化。为突破局部编辑的空间局限性，一些研究转向基于全局语义布局的条件扩散生成机制。Changen（Zheng等，2023）提出了生成式概率变化模型（generative probabilistic change model），将复杂的随机变化过程解耦为语义级别的变化事件模拟与图像级别的语义合成。ChangeAnywhere（Tang等，2024）进一步提出了基于语义潜在扩散模型的生成流程，通过对单时相语义掩码执行异或操作模拟目标的出现与消失，实现了从单时相图像至多时相变化数据集的全局转化。为了实现对复杂变化场景的精细化控制，ChangeDiff（Zang等，2025）构建了文本到布局与布局到图像的两阶段架构，并设计了多类别分布引导文本提示，通过类别分布优化损失有效泛化了扩散模型，实现了依据文本对复杂时序变化图像的精确比例控制。此外，针对公开图像分割数据集与实际遥感应用域之间存在的分布鸿沟问题，DreamCD（Tang等，2026）利用预训练分割器生成的伪掩码作为弱条件训练语义

扩散模型，并设计内容—语义—风格合成机制，利用自适应实例归一化将真实后时相图像的风格特征注入生成过程，有效提升了合成数据在变化检测任务中的域适应能力。

为了追求更高的物理真实感与时空连贯性，一些研究直接深入扩散模型的隐空间与噪声空间，对时序演变过程本身进行底层数学建模。Noise2-Change（Liu 等，2026）摒弃了启发式规则与显式文本引导，通过在离散扩散模型中直接操纵初始采样噪声的语义成分，使模型能够在扩散先验的引导下自动演化出结构连贯、形态自然的连续变化，有效解决了生成图像间空间对齐困难的问题。进一步地，ChangeBridge（Zhao 等，2025b）引入了扩散桥理论（diffusion bridge），打破了传统扩散模型从纯噪声到图像的范式，提出了支持多模态条件控制的异步漂移时空桥接模型。该方法构建了像素级的漂移映射图，对发生突变的前景事件（高漂移）与渐变背景（低漂移）赋予差异化的演化强度，从而精准模拟了遥感场景中异质化的时空演变特征。在基础模型层面上，Changen2（Zheng 等，2025d）将生成式概率变化模型与可扩展分辨率的扩散 Transformer（RS-DiT）深度融合，借助 SAM 提取的目标轮廓作为条件，利用海量无标注单时相数据完成了自监督预训练。该模型不仅生成了包含极多变化类型的百万级数据集，其自身亦演进为面向遥感变化定制的生成式基础模型，在零样本变化预测与任务级表征迁移中展现出了极强的潜力。

总的来说，当前针对开放词汇变化检测的研究已在不同维度取得显著进展，直接的判别式方法不断优化跨模态特征的对齐机制与统一架构，而生成式方法则为解决数据稀缺难题提供了强劲动力。这两大技术路径的深度融合，即利用生成模型产生的高保真、多模态变化数据，精细化调优具备统一架构的视觉语言基础模型，可能成为推动遥感地表动态监测迈向真实开放世界的重要途径。

2.5 其他遥感开放词汇感知任务

随着遥感开放词汇感知在二维图像分类、检测与分割等基础任务上渐趋成熟，该领域的研究边界正加速向更复杂的数据维度与更深度的专家任务拓展。在三维点云的开放词汇感知任务中，鉴于纯三维模态语义标注匮乏，Wang 等人（2025a）提出了一种免标注框架 OpenUrban3D，通过设计多视角多粒度投影模块生成对齐的虚拟图像，并结合 LLM 解析复杂的空间指令。该方法利用样本平衡特征融合策略将二维语义知识蒸馏至三维主干网络，从而有效解决了大尺度城市场景下目标尺度变异与类别不平衡导致的学习困难。Alami 和 Remondino（2026）同样聚焦于大规模点云，构建了一种免训练的开放词汇分割框架，其依托 Grounding DINO 和 Sa2VA（Yuan 等，2025）等二维检测与分割模型，结合自适应阈值策略，在二维图像中提取语义后将其投影至三维空间，并利用多视角投票与局部几何特征细化来消除投影噪声与误分类。Xu 等人（2025）指出利用二维图像作为中介去桥接三维与文本关系的次优性，并提出直接对齐点云与文本的城市级三维基础模型 CitySeg。针对多源三维数据集间点云分布不均与标注标准冲突的问题，该模型设计了局部—全局交叉注意力网络以增强对稀疏全局上下文的捕捉，并提出层次化分类图策略以统一不同数据源和粒度的标签。该工作证明了在不依赖任何二维视觉图像输入的前提下，也能实现城市级点云场景的零样本泛化，为大尺度空间感知的技术路线选择提供了不同的视角。

此外，开放词汇感知也可以为面临极端数据稀缺的交叉遥感应用提供新的解决思路。在城市能源建模领域，由于建筑元数据常因隐私保护或更新滞后难以获取，HeatPrompt（Thota 等，2026）将 MLLM 作为“municipal heat planner”，从卫星图像中提取屋顶老化度、建筑密度等语义特征作为物理元数据的先验，大幅提升了连续热负荷回归预测的准确性与可解释性。在考古遥感中，受限于已发现的隐蔽遗址数量极少，研究人员难以构建足够规模的训练集来微调深度学习模型。Landauer 和 Klassen（2025）对此进行了深入探讨，在无需任何特定领域微调的设定下，评估了视觉基础模型在考古特征检测任务中的可行性。该研究涵盖了从卫星影像、航空 LiDAR 阴影图到无人机视频等多种模态。其表明，即便未经专业考古数据微调，这些模型也能达到与人类初级专家或传统自动化方法相近的检测水平。

3 结　语

遥感图像开放词汇感知技术的快速演进，标志着地球观测领域的视觉智能解译正经历一场深刻的

范式跃迁:从高度依赖专家标注的预定义封闭集模式,迈向基于自然语言驱动的开放世界通用感知模式。通过视觉—语言预训练模型的跨模态语义对齐,现有的图像分类、跨模态检索、目标检测、图像分割和变化检测等基础任务,在零样本或少样本场景下均展现出了令人瞩目的泛化潜力。然而,遥感数据本身所具备的尺度多变性、背景复杂性、多模态异构性以及强烈的地理空间属性,使得现有基于自然图像域发展而来的视觉—语言模型在向遥感垂直领域迁移时面临重重阻碍。要真正实现遥感领域的开放感知,当前研究在评测基准构建、底层模型架构设计、多模态深度融合以及工程化边缘落地等方面,仍有诸多亟待突破的瓶颈。

1)高质量训练数据的匮乏与地理偏差。尽管大模型辅助的数据生成引擎极大地扩充了遥感图文对的规模,但对遥感特定的密集目标群、复杂的背景干扰与抽象的语义属性仍难以被自然语言精准、无歧义地描述。高质量的图像—包围框和图像—掩码数据的批量构造也依然是个难题。此外,现有的开源预训练数据在地理分布上存在严重的区域不平衡,导致模型在处理数据稀缺地区(如欠发达国家和地区)的影像时,易发生表征退化与性能坍塌。

2)缺乏细粒度开放词汇验证基准。现有的遥感开放词汇模型评估大多依赖于传统的封闭集数据集(如LoveDA、DOTA等),或仅在有限的常见类别测试集上进行验证。这种在已知或高频类别上测试的评价方式,无法真实反映模型对长尾地物(如罕见军用设施、特定生长期的农作物)、精细属性以及复杂空间关系的泛化能力。由于缺乏一个真正包含海量类别且呈现真实世界长尾分布的验证集,当前部分模型的开放词汇能力仍停留在“伪开放”阶段,难以准确衡量其在极端或未知场景下的真实表现。

3)多源异构模态的开放感知。真实的地球观测依赖于全天候、全谱段的传感器网络,但当前开放感知研究过度依赖光学RGB图像与自然语言的对齐。对于SAR图像的相干斑噪声与微波散射机制、多光谱/高光谱的高维连续性以及热红外的辐射特性等,现有视觉—语言基础模型缺乏内在的物理规律认知。如何打破光学视角的局限,构建涵盖多源异构数据的跨模态语义表示空间,仍是巨大的挑战。

4)模型可靠性与可解释性瓶颈。遥感解译常被应用于国家安全、灾害应急和环境监测等高风险决策场景,这对模型的可靠性提出了极高要求。当前的视觉—语言模型本质上仍是黑箱,其决策过程缺乏透明度。模型可能会出现高置信度的虚假预测,或在面对域外样本时发生无提示的性能骤降。这种内在的不确定性与不可解释性,制约了开放词汇感知技术在关键任务中的应用落地,使其难以成为自动化决策闭环中的可信赖环节。

针对上述挑战,本文对未来的几个研究课题进行展望。

1)MLLM驱动的生成式感知范式。传统基于视觉与文本特征相似度匹配的对比学习范式在应对遥感复杂任务时逐渐显露瓶颈,未来研究将加速向以大语言模型为核心的自回归生成式感知范式演进。这种范式通过将目标检测、精细分割和多边形提取等空间定位任务统一建模为坐标序列或几何属性词元的直接生成,能够深度复用大模型在海量文本预训练中积累的逻辑推断与零样本泛化能力。针对离散词元表示引发的定位精度损失以及密集场景下的高复杂度问题,可引入更加合理的定位形式(Lan等,2024;Jiang等,2025;Tang等,2025;Song等,2026a),使模型在连续物理空间与离散符号空间之间建立精准映射。这种从特征度量向语言序列预测的底层架构重构,不仅能显著提升模型对复杂指令的响应精度,更有望从根本上解决遥感解译中结构化输出与深层语义理解脱节的难题,实现感知与推理的有机统一。

2)真实开放世界细粒度评测体系构建。学术界亟需突破现有封闭式数据集的桎梏,联合构建涵盖海量地理类别、极度长尾分布以及复杂时空逻辑的遥感大规模开放词汇验证基准。该体系的构建应深度集成测绘地理信息专业知识图谱与多源地理数据,确保评测样本在地理纬度、气候条件、成像载体以及光照环境等方面具备极高的多样性。在评价准则层面,应打破单一依赖交并比或平均精度的局限,建立一套包含语义逻辑一致性、多尺度定位精确度、跨模态常识推演能力以及环境干扰鲁棒性在内的综合评估方式。通过引入针对具体地物属性和细粒度场景的深度评测指标,该体系将能够揭示模型在面对罕见地物、突发事件及专业领域需求时的真实泛化水平,为遥感开放词汇感知的实用化进程提供严谨且客观的衡量尺度。

3)向全模态基础模型的演进。地球观测领域具

有极强的传感器异构性，而现有视觉大模型普遍缺乏对非光学成像物理规律的底层认知。未来的跨模态感知研究应从纯粹的数据驱动转向物理先验与深度学习的结合，系统性地探索将SAR、高光谱和热红外等数据的物理规律作为先验，显式地注入深层神经网络的特征提取与语义对齐环节。通过构建物理机理与自然语言共监督的多模态联合特征表示空间，可以有效消解传感器观测值波动与地物本质语义之间的关联偏差。这一发展路径将促使遥感基础模型真正具备处理全天候、全谱段异构数据的能力，实现从浅层视觉表征匹配向深层物理机理理解的跨越，从而显著提升模型在多源协同观测任务下的解译可靠性。

4）具备时空因果推演能力的动态演变解译。地球表面是一个高度动态且复杂的交互系统，遥感感知的边界正从二维静态图像分析向三维立体空间以及融合时间维度的四维动态演变过程全面延伸。未来的开放词汇感知模型需引入更高效的长期记忆机制与时序因果建模模块，使其不仅能精准理解大规模城市三维点云或高程模型中的空间几何关系，更能在跨时相的观测序列中识别地表的演化规律。通过赋予模型对自然灾害演进过程、城市扩展模式或生态植被周期等事件的逻辑分析能力，遥感开放词汇感知模型将不再局限于单一时间点的目标识别，而是演进为对地表动态演化过程的深度因果推断。这种向四维时空感知的拓展，将极大地增强模型在监测复杂地理过程、预警突发灾害事件以及模拟未来地貌演变等深层次应用中的科学支撑作用。

5）面向星地协同的高效轻量化与可信安全计算。考虑到遥感基础模型庞大的参数规模与航空航天边缘设备（如在轨卫星、无人机载荷）极其受限的算力与功耗之间的剧烈矛盾，模型的高效压缩与分布式协同计算技术将成为落地的关键。未来需深入探索参数高效微调、低位宽量化以及跨模态知识蒸馏等技术，构建云端大模型持续自主演化与边缘端模型敏捷协同推理的星地一体化计算架构。与此同时，在面向军事侦察、应急减灾等高风险应用场景时，遥感感知系统的安全性与可解释性不可忽视。通过集成思维链推理以提供透明的逻辑化推演过程，并结合系统性的不确定性量化方法，将显著提升开放感知模型在面对对抗攻击或极端工况下的鲁棒性，确保开放词汇感知技术在满足高性能需求的同时，具备极高的决策可靠性与公信力。

综上所述，遥感图像开放词汇感知正在向通用化迈进。视觉—语言预训练模型及其他多模态大模型的引入，打破了传统封闭集视觉任务的类别壁垒，为复杂多变的地球观测场景提供了前所未有的灵活解译能力。尽管当前在评测基准的完备性、多源异构数据的解译等方面仍面临诸多挑战，但随着通用领域和遥感领域大模型的不断演进，这些技术瓶颈有望被逐步攻克。可预见的是，遥感开放感知技术将加速向全天候、全谱段、四维时空及星地协同计算等方向发展，最终为构建具备强大因果推理与深度解译能力的通用地球视觉智能奠定坚实基础，全面赋能灾害预警、资源探测、生态保护与全球变化监测等重大科学与工程应用。

参考文献（References）

Aimar E S, Zhambulova G, Khan F S, Xu Y H and Felsberg M. 2025. VLM2GeoVec: toward universal multimodal embeddings for remote sensing [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2512.11490.pdf

Alami A and Remondino F. 2026. Open-vocabulary segmentation of aerial point clouds. Remote Sensing, 18(4): #572 [DOI: 10.3390/rs18040572]

An K, Wang Y P and Chen L. 2025. Soft-guided open-vocabulary semantic segmentation of remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 63: #5652216 [DOI: 10.1109/TGRS.2025.3628336]

Barzilai A, Gigi Y, Helmy A, Silverman V, Refael Y, Jaber B, et al. 2025. A recipe for improving remote sensing VLM zero shot generalization [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2503.08722.pdf

Benidir Y, Gonthier N and Mallet C. 2025. The change you want to detect: Semantic change detection in earth observation with hybrid data generation//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 2204-2214 [DOI: 10.1109/CVPR52734.2025.00211]

Brown C F, Kazmierski M R, Pasquarella V J, Rucklidge W J, Samsikova M, Zhang C H, et al. 2025. AlphaEarth foundations: an embedding field model for accurate and efficient global mapping from sparse label data [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2507.22291.pdf

Cambrin D R, Vaiani L, Gallipoli G, Cagliero L and Garza P. 2025. CLOSP: a unified semantic space for SAR, MSI, and text in remote sensing [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2507.10403.pdf

Cao Q L, Chen Y T, Ma C and Yang X K. 2025. Open-vocabulary high-resolution remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 63: #5620614 [DOI: 10.1109/TGRS.2025.3559557]

Carion N, Gustafson L, Hu Y T, Debnath S, Hu R H, Suris D, et al. 2026. SAM 3: segment anything with concepts [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2511.16719.pdf

Chen W Z, Deng Y P, Jin W, Chen J B, Chen J S, Feng Y M, et al. 2025a. DGTRSD and DGTRSCLIP: a dual-granularity remote sensing image-text dataset and vision-language foundation model for alignment. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18: 29113-29130 [DOI: 10.1109/JSTARS.2025.3625958]

Chen X M, Zheng X T and Lu X Q. 2025b. Relevance-guided adaptive learning for remote sensing image-text retrieval. IEEE Transactions on Geoscience and Remote Sensing, 63: #5632713 [DOI: 10.1109/TGRS.2025.3587097]

Chen X M, Zheng X T and Lu X Q. 2025c. Context-aware local-global semantic alignment for remote sensing image-text retrieval. IEEE Transactions on Geoscience and Remote Sensing, 63: #5617312 [DOI: 10.1109/TGRS.2025.3552304]

Cheng T H, Song L, Ge Y X, Liu W Y, Wang X G and Shan Y. 2024. YOLO-world: real-time open-vocabulary object detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 16901-16911 [DOI: 10.1109/CVPR52733.2024.01599]

Cheng Q M, Zhou Z L, Yuan L F and Du Y J. 2025. Efficient yet effective:a dynamic self-distillation framework for remote sensing image-text retrieval. IEEE Geoscience and Remote Sensing Letters,22:1-5 [DOI:10. 1109/LGRS. 2025. 3586775]

Cherti M, Beaumont R, Wightman R, Wortsman M, Ilharco G, Gordon C, et al. 2023. Reproducible scaling laws for contrastive language-image learning//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 2818-2829 [DOI: 10.1109/CVPR52729.2023.00276]

Cho E, Won S, Choo O S and Kim S T. 2025. PRISM: pseudo-labeling and region-based inpainting for synthetic change detection modeling. IEEE Geoscience and Remote Sensing Letters, 22: #6009205 [DOI: 10.1109/LGRS.2025.3569426]

Cho S, Shin H, Hong S, Arnab A, Seo P H and Kim S R. 2024. CAT-seg: cost aggregation for open-vocabulary semantic segmentation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 4113-4123 [DOI: 10.1109/CVPR52733.2024.00394]

Dou M Y, Qiu S, Hu M, Chen Y F, Ye H P, Liao X H, et al. 2026. AdaptOVCD: training-free open-vocabulary remote sensing change detection via adaptive information fusion [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2602.06529.pdf

Dutta S, Vasim A, Gole S, Rezatofighi H and Banerjee B. 2025. AerO-Seg: harnessing SAM for open-vocabulary segmentation in remote sensing images//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville, USA: IEEE: 2245-2255 [DOI: 10.1109/CVPRW67362.2025.00212]

Dutta S, Vasim A, Rezatofighi H and Banerjee B. 2026. AerOSeg++: scale-aware and texture-guided open-vocabulary segmentation with SAM features for remote sensing images. ACM Transactions on Multimedia Computing, Communications and Applications, 22(5): 1-21. [DOI: 10.1145/3787522]

El Khoury K, Zanella M, Gérin B, Godelaine T, Macq B, Mahmoudi S, et al. 2025. Enhancing remote sensing vision-language models for zero-shot scene classification//Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad, India: IEEE: 1-5 [DOI: 10.1109/ICASSP49660.2025.10888395]

Ge J Y, Zhang X, Zheng Y, Guo K T and Liang J M. 2025. RSTeller: scaling up visual language modeling in remote sensing with rich linguistic semantics from openly available data and large language models. ISPRS Journal of Photogrammetry and Remote Sensing, 226: 146-163 [DOI: 10.1016/j.isprsjprs.2025.05.002]

Gerg I D. 2026. Prompted, not trained: on zero-shot classification of synthetic aperture imagery with vision-language models [EB/OL]. [2026-03-20]. https://www.researchgate.net/profile/Isaac-Gerg/publication/399392672_Prompted_Not_Trained_On_Zero-Shot_Classification_of_Synthetic_Aperture_Imagery_with_Vision-Language_Models/links/695bf47b0c98040d4827af3a/Prompted-Not-Trained-On-Zero-Shot-Classification-of-Synthetic-Aperture-Imagery-with-Vision-Language-Models.pdf

Gu J Z, Fan L L, Zhao J and Cao X H. 2025. CoseDet: open-vocabulary remote sensing object detection with contextual semantic information. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18: 26863-26875 [DOI: 10.1109/JSTARS.2025.3622239]

Guan J H, Shu Y L, Li W G, Song Z H and Zhang Y C. 2025. PR-CLIP: cross-modal positional reconstruction for remote sensing image-text retrieval. Remote Sensing, 17(13): #2117 [DOI: 10.3390/rs17132117]

He Y G, Cheng X J, Zhu J J, Qiu C P, Wang J, Zhang X C, et al. 2025a. SAR-TEXT: a large-scale SAR image-text dataset built with SAR-narrator and a progressive learning strategy for downstream tasks [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2507.18743.pdf

He Y G, Zhu J J, Li Y Y, Huang Q J, Wang Z Y and Yang K. 2024. Rethinking remote sensing CLIP: leveraging multimodal large language models for high-quality vision-language dataset//Proceedings of the 31st International Conference on Neural Information Processing. Auckland, New Zealand: Springer: 417-431 [DOI: 10.1007/

978-981-96-6972-1_29]

He Y G, Zhu J J, Li Y Y, Zhang X Y, Qiu C P, Wang J, et al. 2025b. Enhancing remote sensing vision-language models through MLLM and LLM-based high-quality image-text dataset generation [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2507.16716.pdf

Heidarianbaei M, Dorozynski M, Kanyamahanga H, Mehltretter M and Rottensteiner F. 2026. Open-vocabulary semantic segmentation in remote sensing via hierarchical attention masking and model composition [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2602.23869.pdf

Hu W, Hu S H, Ma F, Zhao Q H and Zhang F. 2026. KGCS: zero-annotation expert-knowledge injection for object detection in aerial images. IEEE Transactions on Geoscience and Remote Sensing, 64: #5611716 [DOI: 10.1109/TGRS.2026.3670221]

Huang L B, Chen Y S, Li Z K, Ghamisi P and Du Q. 2025a. HZSCM: hyperspectral image zero-shot classification via vision-language models. IEEE Transactions on Geoscience and Remote Sensing, 63: #5528420 [DOI: 10.1109/TGRS.2025.3618636]

Huang S Q, He S T and Wen B H. 2025b. ZoRI: towards discriminative zero-shot remote sensing instance segmentation//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press: 3724-3732 [DOI: 10.1609/aaai.v39i4.32388]

Huang S Q, He S T, Qin H Y and Wen B H. 2025c. SCORE: scene context matters in open-vocabulary remote sensing instance segmentation//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision (ICCV). Honolulu, USA: IEEE: 12559-12569

Huang W B, Deng F, Li H C and Yang J. 2026a. HG-RSOVSSeg: hierarchical guidance open-vocabulary semantic segmentation framework of high-resolution remote sensing images. Remote Sensing, 18(2): #213 [DOI: 10.3390/rs18020213]

Huang W B, Li H C, Zhang S and Deng F. 2026b. Reducing semantic ambiguity in open-vocabulary remote sensing image segmentation via knowledge graph-enhanced class representations. ISPRS Journal of Photogrammetry and Remote Sensing, 231: 837-853 [DOI: 10.1016/j.isprsjprs.2025.11.029]

Huang Z Y, Feng Y C, Yang S, Liu Z Q, Liu Q J and Wang Y H. 2025d. OpenRSD: towards open-prompts for object detection in remote sensing images [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2503.06146.pdf

Hwang H and Woo S S. 2025. FASE: feature-aligned scene encoding for open-vocabulary object detection in remote sensing//Proceedings of the 34th ACM International Conference on Information and Knowledge Management. Seoul, Korea(South): Association for Computing Machinery: 4822-4826 [DOI: 10.1145/3746252.3760838]

Irvin J A, Liu E R, Chen J C, Dormoy I, Kim J, Khanna S, et al. 2025. TEOChat: a large vision-language assistant for temporal earth observation data//Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: ICLR: 1-29

Jain P, Ienco D, Interdonato R, Berchoux T and Marcos D. 2025a. SenCLIP: enhancing zero-shot land-use mapping for sentinel-2 with ground-level prompting//Proceedings of 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Tucson, USA: IEEE: 5656-5665 [DOI: 10.1109/WACV61041.2025.00552]

Jain P, Marcos D, Ienco D, Interdonato R and Berchoux T. 2025b. TimeSenCLIP: a time series vision-language model for remote sensing using single-pixel [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2508.11919.pdf

Ji Y R, Wang C H, Chen J S, Chen J B, Yue A Z, Meng Y, et al. 2026. MovSeg: efficient adaptation of vision-language models for multispectral open-vocabulary segmentation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 19: 8044-8055 [DOI: 10.1109/JSTARS.2026.3658442]

Ji Z, Meng C X, Zhang Y, Wang H R, Pang Y W and Han J G. 2024. Eliminate before align: a remote sensing image-text retrieval framework with keyword explicit reasoning//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: Association for Computing Machinery: 1662-1671 [DOI: 10.1145/3664647.3681270]

Jiang Q, Huo J A, Chen X Y, Xiong Y D, Zeng Z Y, Chen Y H, et al. 2025. Detect anything via next point prediction [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2510.12798.pdf

Ju M H, Feng Y C, Diao W H and Liu C B. 2026. Addressing dense small-object detection in remote sensing: an open-vocabulary object detection framework. Remote Sensing, 18(6): #851 [DOI: 10.3390/rs18060851]

Kini J, Gupta R and Shah M. 2025. Cross-view open-vocabulary object detection in aerial imagery[EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2510.03858.pdf

Klemmer K, Rolf E, Robinson C, Mackey L and Rußwurm M. 2025. SatCLIP: global, general-purpose location embeddings with satellite imagery//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press: 4347-4355 [DOI: 10.1609/aaai.v39i4.32457]

Kombol N, Martinović I and Šegvić S. 2025. A survey on training-free open-vocabulary semantic segmentation [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2505.22209.pdf

Lan M C, Chen C F, Zhou Y, Xu J X, Ke Y P, Wang X J, et al. 2024. Text4Seg: reimagining image segmentation as text generation//Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: ICLR: 1-28

Landauer J and Klassen S. 2025. Visual foundation models for archaeological remote sensing: a zero-shot approach. Geomatics, 5(4): #52 [DOI: 10.3390/geomatics5040052]

Li A X, Lu Z W, Wang L W, Xiang T and Wen J R. 2017. Zero-shot scene classification for high spatial resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing,

55(7): 4157-4167 [DOI: 10.1109/TGRS.2017.2689071]

Li B Y, Dong H C, Zhang D, Zhao Z Y, Sun H and Gao J Y. 2025a. Exploring efficient open-vocabulary segmentation in the remote sensing//Proceedings of the 40th AAAI Conference on Artificial Intelligence. Singapore, Singapore: AAAI Press: 5982-5991 [DOI: 10.1609/aaai.v40i8.37521]

Li H D. 2026. Open-vocabulary object detection for high-resolution remote sensing images. Computer Vision and Image Understanding, 263: #104566 [DOI: 10.1016/j.cviu.2025.104566]

Li J H, Pan J C, Sun Y Z and Huang X M. 2025b. Semantic-aware ship detection with vision-language integration//Proceedings of 2025 IGARSS IEEE International Geoscience and Remote Sensing Symposium. Brisbane, Australia: IEEE: 6903-6907 [DOI: 10.1109/IGARSS55030.2025.11243976]

Li J, Pei Y Q, Zhao S H, Xiao R L, Sang X and Zhang C Y. 2020. A review of remote sensing for environmental monitoring in China. Remote Sensing, 12(7): #1130 [DOI: 10.3390/rs12071130]

Li J N, Li D X, Savarese S and Hoi S. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: #814 [DOI: 10.5555/3618408.3619222]

Li K Y, Cao X Y, Deng Y P, Pang C, Xin Z P, Qiao H, et al. 2025c. DynamicEarth: how far are we from open-vocabulary change detection?//Proceedings of the 40th AAAI Conference on Artificial Intelligence. Singapore, Singapore: AAAI Press: 6279-6287 [DOI: 10.1609/aaai.v40i8.37554]

Li K Y, Cao X Y, Deng Y P, Song J Y, Liu J M, Meng D Y, et al. 2025d. SemiCD-VL: visual-language model guidance makes better semi-supervised change detector. IEEE Transactions on Geoscience and Remote Sensing, 63: #5601613 [DOI: 10.1109/TGRS.2024.3512548]

Li K Y, Cao X Y, Liu R X, Wang S H, Jiang Z X, Wang Z, et al. 2025e. Annotation-free open-vocabulary segmentation for remote-sensing images [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2508.18067.pdf

Li K Y, Cao X Y and Meng D Y. 2024a. A new learning paradigm for foundation model-based remote-sensing change detection. IEEE Transactions on Geoscience and Remote Sensing, 62: #5610112 [DOI: 10.1109/TGRS.2024.3365825]

Li K Y, Jiang Z X, Cao X Y, Wang J Y, Xiao Y C, Meng D Y, et al. 2025f. DescribeEarth: describe anything for remote sensing images [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2509.25654.pdf

Li K Y, Liu R X, Cao X Y, Bai X R, Zhou F, Meng D Y, et al. 2025g. SegEarth-OV: towards training-free open-vocabulary segmentation for remote sensing images//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 10545-10556 [DOI: 10.1109/CVPR52734.2025.00986]

Li K Y, Zhang S Q, Deng Y P, Wang Z, Meng D Y and Cao X Y. 2025h. SegEarth-OV3: exploring SAM 3 for open-vocabulary semantic segmentation in remote sensing images [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2512.08730.pdf

Li K, Wang D, Wang T, Dong F Y, Zhang Y M, Zhang L Y, et al. 2026a. RSVG-ZeroOV: exploring a training-free framework for zero-shot open-vocabulary visual grounding in remote sensing images//Proceedings of the 40th AAAI Conference on Artificial Intelligence. Singapore, Singapore: AAAI Press: 6288-6296 [DOI: 10.1609/aaai.v40i8.37555]

Li L H, Zhang P C, Zhang H T, Yang J W, Li C Y, Zhong Y W, et al. 2022. Grounded language-image pre-training//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 10955-10965 [DOI: 10.1109/CVPR52688.2022.01069]

Li Y, Guo W W, Yang X, Liao N, He D Y, Zhou J Q, et al. 2025i. Toward open vocabulary aerial object detection with CLIP-activated student-teacher learning//Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer: 431-448 [DOI: 10.1007/978-3-031-73016-0_25]

Li Y, Guo W W, Yang X, Liao N, Zhang S F, Yu Y, et al. 2026b. Exploiting unlabeled data with multiple expert teachers for open vocabulary aerial object detection and its orientation adaptation. International Journal of Computer Vision, 134(4): #156 [DOI: 10.1007/s11263-026-02733-2]

Li Y S, Kong D Y, Zhang Y J, Chen R X and Chen J D. 2021. Representation learning of remote sensing knowledge graph for zero-shot remote sensing image scene classification//Proceedings of 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. Brussels, Belgium: IEEE: 1351-1354 [DOI: 10.1109/IGARSS47720.2021.9553667]

Li Z S, Yu W K, Muhtar D, Zhang X L, Xiao P F, Ghamisi P, et al. 2025j. FarSLIP: discovering effective CLIP adaptation for fine-grained remote sensing understanding [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2511.14901.pdf

Liang A K, Xiao X, Hu X Y, Ke T, Wang T Y, Xiong Y B, et al. 2026. GeoPriorclip: a foundational remote sensing vision-language model enhanced with cascaded geographic information priors. Geospatial Information Science: 1-24 [DOI: 10.1080/10095020.2026.2619233]

Liao W B, Gao Y C, Zhu H, Ma Y K and Dai F. 2026. AlignCLIP: self-guided alignment for remote sensing open-vocabulary semantic segmentation//Proceedings of the 14th International Conference on Learning Representations. Rio de Janeiro, Brazil: 1-27

Liu C Y, Chen K Y, Zhao R, Zou Z X and Shi Z W. 2025a. Text2Earth: unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model. IEEE Geoscience and Remote Sensing Magazine, 13(3): 238-259 [DOI: 10.1109/MGRS.2025.3560455]

Liu C Y, Huang W and Zhu X X. 2025b. LandSegmenter: towards a flexible foundation model for land use and land cover mapping [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2511.08156.pdf

Liu F, Chen D L, Guan Z Q Y, Zhou X C, Zhu J L, Ye Q L, et al. 2024a. RemoteCLIP: a vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 62: #5622216 [DOI: 10.1109/TGRS.2024.3390838]

Liu J Y, Qin Q, Dong G S, Wang X L, Feng J Z, Zeng Z C, et al. 2025c. Beyond AlphaEarth: toward human-centered geospatial foundation models via POI-guided contrastive learning [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2510.09894.pdf

Liu Q, Kuang Y, Yue J, Ghamisi P, Xie W Y and Fang L Y. 2026. Generating any changes in the noise domain. IEEE Transactions on Pattern Analysis and Machine Intelligence, 48(3): 3698-3713 [DOI: 10.1109/TPAMI.2025.3643733]

Liu R X, Li K Y, Song J Y, Sun D W and Cao X Y. 2024b. MV-CC: mask enhanced video model for remote sensing change caption [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2410.23946.pdf

Liu S L, Zeng Z Y, Ren T H, Li F, Zhang H, Yang J, et al. 2024c. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 38-55 [DOI: 10.1007/978-3-031-72970-6_3]

Liu X and Lian Z H. 2024. RSUniVLM: a unified vision language model for remote sensing via granularity-oriented mixture of experts [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2412.05679.pdf

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 W, Pang Z, Zhang Y J, Wang T Z, Wang L L, Dang B, et al. 2024. SkySenseGPT: a fine-grained instruction tuning dataset and model for remote sensing vision-language understanding [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2406.10100.pdf

Ma Q W, Wang Z Y, Liu W, Lu X K, Deng B, Duan P H, et al. 2025. SARVLM: a vision language foundation model for semantic understanding and target recognition in SAR imagery [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2510.22665.pdf

Mall U, Phoo C P, Liu M K, Vondrick C, Hariharan B and Bala K. 2023. Remote sensing vision-language foundation models without annotations via ground remote alignment [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2312.06960.pdf

Marimo C T, Blumenstiel B, Nitsche M, Jakubik J and Brunschwiler T. 2026. Beyond the visible: multispectral vision-language learning for earth observation//Ribeiro R P, Pfahringer B, Japkowicz N, Larrañaga P, Jorge A M, Soares C, et al, eds. Machine Learning and Knowledge Discovery in Databases. Research Track. Cham, Germany: Springer: 359-375 [DOI: 10.1007/978-3-032-06106-5_21]

Mi L, Dai X J, Castillo-Navarro J and Tuia D. 2024. Knowledge-aware text-image retrieval for remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 62: #5646813 [DOI: 10.1109/TGRS.2024.3486977]

Muhtar D, Li Z S, Gu F, Zhang X L and Xiao P F. 2025. LHRS-bot: empowering remote sensing with VGI-enhanced large multimodal language model//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 440-457 [DOI: 10.1007/978-3-031-72904-1_26]

Oquab M, Darcet T, Moutakanni T, Vo H V, Szafraniec M, Khalidov V, et al. 2024. DINOv2: learning robust visual features without supervision [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2304.07193.pdf

Ouerghemi N, Tomoiagă C and Detyniecki M. 2026. Planes, not A380: how prompting, context, and granularity shape VLM performance in aerial imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 19: 5009-5020 [DOI: 10.1109/JSTARS.2025.3649701]

Pan J C, Liu Y X, Fu Y Q, Ma M Y, Li J H, Paudel D P, et al. 2025. Locate anything on earth: advancing open-vocabulary object detection for remote sensing community//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI: 6281-6289 [DOI: 10.1609/aaai.v39i6.32672]

Pan J C, Ma M Y, Ma Q, Bai C and Chen S Y. 2024. PriorCLIP: visual prior guided vision-language model for remote sensing image-text retrieval [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2405.10160.pdf

Pang C, Weng X X, Wu J, Li J Y, Liu Y, Sun J X, et al. 2025a. VHM: versatile and honest vision language model for remote sensing image analysis//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press: 6381-6388 [DOI: 10.1609/aaai.v39i6.32683]

Pang L, Yao J, Li K Y, Zhou J, Meng D Y and Cao X Y. 2025b. SPECIAL: zero-shot hyperspectral image classification with CLIP [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2501.16222.pdf

Peng Z L, Wang W H, Dong L, Hao Y R, Huang S H, Ma S M, et al. 2023. Kosmos-2: grounding multimodal large language models to the world [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2306.14824.pdf

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

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Virtual: PMLR: 8748-8763

Ravi N, Gabeur V, Hu Y T, Hu R H, Ryali C, Ma T Y, et al. 2025.

SAM 2: segment anything in images and videos//Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: ICLR: 1-44

Silva J D, Magalhães J, Tuia D and Martins B. 2024. Multilingual vision-language pre-training for the remote sensing domain//Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems. Atlanta, USA: ACM: 220-232 [DOI: 10.1145/3678717.3691318]

Song J, Chen H R X and Yokoya N. 2024. SyntheWorld: a large-scale synthetic dataset for land cover mapping and building change detection//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE: 8272-8281 [DOI: 10.1109/WACV57701.2024.00810]

Song T H, Lu H Y, Yang H, Sui L, Wu H N, Zhou Z D, et al. 2026a. Towards pixel-level VLM perception via simple points prediction [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2601.19228.pdf

Song Z H, Shu Y L, Li W G, Guan J H and Zhang Y C. 2026b. Towards discriminative and consistent cross-modal alignment for remote sensing image-text retrieval. Remote Sensing, 18(4): #662 [DOI: 10.3390/rs18040662]

Soni S, Dudhane A, Debary H, Fiaz M, Munir M A, Danish M S, et al. 2025. EarthDial: turning multi-sensory earth observations to interactive dialogues//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 14303-14313 [DOI: 10.1109/CVPR52734.2025.01334]

Sosa J, Rukhovich D, Kacem A and Aouada D. 2026. Enabling training-free text-based remote sensing segmentation [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2602.17799.pdf

Stacchio L, Nepi L, Paolanti M and Pierdicca R. 2025. RSplitzero: generalized zero-shot learning in remote sensing across attribute splits with single and multi-modal representations. International Journal of Digital Earth, 18(2): #2551869 [DOI: 10.1080/17538947.2025.2551869]

Sun T C, Zheng C Y, Li X, Gao Y L, Nie J, Huang L, et al. 2025. Strong and weak prompt engineering for remote sensing image-text cross-modal retrieval. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18: 6968-6980 [DOI: 10.1109/JSTARS.2025.3534474]

Tan X L, Chen G Z, Wang T, Wang J Q and Zhang X D. 2024a. Segment change model (SCM) for unsupervised change detection in VHR remote sensing images: a case study of buildings//Proceedings of 2024 IGARSS IEEE International Geoscience and Remote Sensing Symposium. Athens, Greece: IEEE: 8577-8580 [DOI: 10.1109/IGARSS53475.2024.10642429]

Tan X M, Xi B B, Li J J, Zheng T, Li Y S, Xue C B, et al. 2024b. Review of zero-shot remote sensing image scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17: 11274-11289 [DOI: 10.1109/JSTARS.2024.3410995]

Tang H, Xie C W, Wang H Y, Bao X Y, Weng T Y, Li P D, et al. 2025. UFO: a unified approach to fine-grained visual perception via open-ended language interface//Proceedings of the 39th Annual Conference on Neural Information Processing Systems. Mexico City, Mexico: [s.n.]

Tang K and Chen J. 2024. ChangeAnywhere: sample generation for remote sensing change detection via semantic latent diffusion model [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2404.08892.pdf

Tang K, Zheng Z, Chen H R X, Chen X J and Chen J. 2026. DreamCD: a change-label-free framework for change detection via a weakly conditional semantic diffusion model in optical VHR imagery. International Journal of Applied Earth Observation and Geoinformation, 146: #105125 [DOI: 10.1016/j.jag.2026.105125]

Thota K, Mu X H, Schlachter T and Hagenmeyer V. 2026. HeatPrompt: zero-shot vision-language modeling of urban heat demand from satellite images [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2602.20066.pdf

Wang C Y, Jing K L, Zhu J H and Wang D. 2025a. OpenUrban3D: annotation-free open-vocabulary semantic segmentation of large-scale urban point clouds [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2509.10842.pdf

Wang D, Liu S Y, Jiang W T, Wang F X, Liu Y, Qin X L, et al. 2025b. GeoZero: incentivizing reasoning from scratch on geospatial scenes [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2511.22645.pdf

Wang D, Zhang J, Xu M Q, Liu L, Wang D S, Gao E Z, et al. 2024a. MTP: advancing remote sensing foundation model via multitask pretraining. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17: 11632-11654 [DOI: 10.1109/JSTARS.2024.3408154]

Wang G Q, Xie J L, Zhang T, Sun Y K, Chen H, Zhuang Y, et al. 2025c. LLaMA-unidetector: an LLaMA-based universal framework for open-vocabulary object detection in remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 63: #4409318 [DOI: 10.1109/TGRS.2025.3564332]

Wang J Y, Sun H, Tang T, Sun Y L, He Q S, Lei L, et al. 2024b. Leveraging visual language model and generative diffusion model for zero-shot SAR target recognition. Remote Sensing, 16(16): #2927 [DOI: 10.3390/rs16162927]

Wang J Z and Ni H. 2026. Bidirectional cross-perception for open-vocabulary semantic segmentation in remote sensing imagery [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2601.21159.pdf

Wang P F, Lu Z H, Li Y J, Ding B G and Zhang D. 2025d. SARCLIP: the first vision-language foundation model for SAR image. IEEE Transactions on Geoscience and Remote Sensing, 63: #5223211 [DOI: 10.1109/TGRS.2025.3630131]

Wang S J, Song Y, Xiang J J, Chen Y Y, Zhong P and Fu R G. 2025e. Mask-guided teacher-student learning for open-vocabulary object

detection in remote sensing images. Remote Sensing, 17(19): #3385 [DOI: 10.3390/rs17193385]

Wang S W, Han J G and Zhu X X. 2025f. RSCLIP for training-free open-vocabulary remote sensing image semantic segmentation [EB/OL]. [2026-03-20]. https://www.techrxiv.org/doi/10.36227/techrxiv.175790902.28615776/v1

Wang W Z, Xiao A R, He W, Zhu H Y and Xiao L. 2025g. Text-to-image activation for open-vocabulary semantic segmentation in remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 63: #4419917 [DOI: 10.1109/TGRS.2025.3619504]

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

Wang Z Q, Zhang Z K, Chen Q H, Xiong M M and Zhang J L. 2025h. Open-vocabulary object detection for low-altitude scenarios using RGB-infrared data: a benchmark and a new method//Proceedings of the 14th International Conference on Learning Representations. Rio de Janeiro, Brazil: 1-12

Wei G T, Liu Y, Yuan X, Xue X Z, Guo L L, Yang Y F, et al. 2025. OS-W2S: an automatic labeling engine for language-guided open-set aerial object detection [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2505.03334.pdf

Wei G T, Yuan X, Liu Y, Shang Z H, Xue X Z, Wang P, et al. 2024. RT-OVAD: real-time open-vocabulary aerial object detection via image-text collaboration [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2408.12246.pdf

Wei G T, Yuan X, Zhou Y, Jing H Z, Liu Y, Qi X B, et al. 2026a. Open-text aerial detection: a unified framework for aerial visual grounding and detection [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2602.07827.pdf

Wei Y M, Xiao A R, Chen H R X, Xia J S and Yokoya N. 2026b. MM-OVSeg: multimodal optical-SAR fusion for open-vocabulary segmentation in remote sensing [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2603.17528.pdf

Weng Z H, Li X J, Wu C, He W J, Lyu J F, Zhou D, et al. 2025. Light-weight cross-modal enhancement method with benchmark construction for UAV-based open-vocabulary object detection [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2509.06011.pdf

Wu J J, Xie J T, Zhang Z L, Wang Q L, Hu Q H, Li P H, et al. 2025a. DALIP: distribution alignment-based language-image pre-training for domain-specific data//Proceedings of 2015 IEEE/CVF International Conference on Computer Vision (ICCV). Honolulu, USA: IEEE/CVF: 2099-2109

Wu J Y, Shi J Y, Zhao Z Y, Liu Z Y and Zhi R C. 2025b. FreeMix: open-vocabulary domain generalization of remote-sensing images for semantic segmentation. Remote Sensing, 17(8): #1357 [DOI: 10.3390/rs17081357]

Wu N M, Cao Q, Wang Z Y, Liu Z P, Qi Y L, Zhang J L, et al. 2024. TorchSpatial: a location encoding framework and benchmark for spatial representation learning//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #2588 [DOI: 10.5555/3737916.3740504]

Xia Y, Yue X, Liu X Z, Chen N, Liu H, Yue J, et al. 2026. SpectralZero: text-driven spectral-spatial alignment for zero-shot hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 64: #5507016 [DOI: 10.1109/TGRS. 2026. 3669516]

Xiong Z T, Wang Y, Yu W K, Stewart A J, Zhao J, Lehmann N, et al. 2025. DOFA-CLIP: multimodal vision-language foundation models for earth observation [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2503.06312.pdf

Xiu D, Ji L Y, Geng X R and Wu Y R. 2024. RSITR-FFT: efficient fine-grained fine-tuning framework with consistency regularization for remote sensing image-text retrieval. IEEE Geoscience and Remote Sensing Letters, 21: #6016805 [DOI: 10.1109/LGRS. 2024.3478176]

Xu J L, Wei Z Z, You W K, Li L Y and Sun W J. 2025. CitySeg:a 3D open vocabulary semantic segmentation foundation model in city-scale scenarios [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2508.09470.pdf

Xu L X, Wang L Y, Zhang J Z, Ha D and Zhang H S. 2025. A review of cross-modal image-text retrieval in remote sensing. Remote Sensing, 17(24): #3995 [DOI: 10.3390/rs17243995]

Yang J, Gong P, Fu R, Zhang M H, Chen J M, Liang S L, et al. 2013. The role of satellite remote sensing in climate change studies. Nature Climate Change, 3(10): 875-883 [DOI: 10.1038/nclimate1908]

Yang S, Huang Z Y, Chen J X, Liu Q J and Wang Y H. 2026a. Beyond open vocabulary: multimodal prompting for object detection in remote sensing images [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2602.01954.pdf

Yang X, Fu R H, Duan Z R, Lin Z W, Liu X Y and Yang B. 2026b. GeoAlignCLIP: enhancing fine-grained vision-language alignment in remote sensing via multi-granular consistency learning [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2603.09566.pdf

Yao J H, Zheng Y B, Lu S Q, Xu W Y and Sun P. 2025a. VK-det: visual knowledge guided prototype learning for open-vocabulary aerial object detection//Proceedings of the 40th AAAI Conference on Artificial Intelligence. Singapore, Singapore: AAAI Press: 11875-11882 [DOI: 10.1609/aaai.v40i14.38174]

Yao K L, Xu N, Yang R, Xu Y Y, Gao Z Y, Kitrungrotsakul T, et al. 2025b. Falcon: a remote sensing vision-language foundation model (Technical Report) [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2503.11070.pdf

Yao L, Liu F, Chen D L, Zhang C Y, Wang Y J, Chen Z Y, et al.

2025c. RemoteSAM: towards segment anything for earth observation//Proceedings of the 33rd ACM International Conference on Multimedia. Dublin, Ireland: ACM: 3027-3036 [DOI: 10.1145/3746027.3754950]

Yu C Y, Zheng Y, Cai L K, Xiang T and Gao C Q. 2026. FGR-GA: integrating fine-grained representation refinement and group-aware alignment for remote sensing image-text retrieval. IEEE Transactions on Geoscience and Remote Sensing, 64: #5402715 [DOI: 10.1109/TGRS.2026.3670002]

Ye C Y, Zhuge Y Z and Zhang P P. 2025. Towards open-vocabulary remote sensing image semantic segmentation//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press: 9436-9444 [DOI: 10.1609/aaai.v39i9.33022]

Yuan H B, Li X T, Zhang T, Sun Y Y, Huang Z L, Xu S L, et al. 2025. Sa2VA: marrying SAM2 with MLLM for dense grounded understanding of images and videos [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2501.04001.pdf

Zang Q, Yang J Y, Wang S, Zhao D, Yi W J and Zhong Z. 2025. ChangeDiff: a multi-temporal change detection data generator with flexible text prompts via diffusion model//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press: 9763-9771 [DOI: 10.1609/aaai.v39i9.33058]

Zavras A, Michail D, Demir B and Papoutsis I. 2025a. Mind the modality gap: towards a remote sensing vision-language model via cross-modal alignment. ISPRS Journal of Photogrammetry and Remote Sensing, 228: 270-287 [DOI: 10.1016/j.isprsjprs.2025.06.019]

Zavras A, Michail D, Zhu X X, Demir B and Papoutsis I. 2025b. GAIA: a global, multi-modal, multi-scale vision-language dataset for remote sensing image analysis [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2502.09598.pdf

Zermatten V, Castillo-Navarro J, Marcos D and Tuia D. 2025. Learning transferable land cover semantics for open vocabulary interactions with remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 220: 621-636 [DOI: 10.1016/j. isprsjprs. 2025. 01.006]

Zhang D, Zeng M and Li X. 2026. Towards open-vocabulary semantic segmentation for remote sensing images. Pattern Recognition, 175: #113120 [DOI:10. 1016/j. patcog. 2026. 113120]

Zhang J L, Zhou Z L, Mai G C, Hu M X, Guan Z H, Li S, et al. 2024a. Text2Seg: zero-shot remote sensing image semantic segmentation via text-guided visual foundation models//Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery. Atlanta, USA: ACM: 63-66 [DOI: 10.1145/3687123.3698287]

Zhang L M, Rao A Y and Agrawala M. 2023a. Adding conditional control to text-to-image diffusion models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 3813-3824 [DOI: 10.1109/ICCV51070.2023.00355]

Zhang W, Cai M X, Zhang T, Zhuang Y, Li J and Mao X R. 2025a. EarthMarker: a visual prompting multimodal large language model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 63: #5604219 [DOI: 10.1109/TGRS.2024.3523505]

Zhang X, Li D Y, Xia Y J, Dong X H, Yu H L, Wang J Y, et al. 2026a. OmniOVCD: streamlining open-vocabulary change detection with SAM 3 [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2601.13895.pdf

Zhang X K, Zhou C F, Huang J Z and Zhang L F. 2025b. TPOV-Seg: textually enhanced prompt tuning of vision-language models for open-vocabulary remote sensing semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 63: #4420117 [DOI: 10.1109/TGRS.2025.3624767]

Zhang X R, Zhang T Y, Wang G C, Zhu P, Tang X, Jia X P, et al. 2023b. Remote sensing object detection meets deep learning: a metareview of challenges and advances. IEEE Geoscience and Remote Sensing Magazine, 11(4): 8-44 [DOI: 10.1109/MGRS.2023.3312347]

Zhang Y, Ji Z, Meng C X and Pang Y W. 2026b. iEBAKER: improved remote sensing image-text retrieval framework via eliminate before align and keyword explicit reasoning. Expert Systems with Applications, 296: #128968 [DOI: 10.1016/j.eswa.2025.128968]

Zhang Z L, Zhao T C, Guo Y L and Yin J W. 2024b. 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]

Zhao X, Ding W C, An Y Q, Du Y L, Yu T, Li M, et al. 2023. Fast segment anything [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2306.12156.pdf

Zhao Y X, Zhang M, Yang B N, Zhang Z, Kang J J and Gong J Y. 2025a. LuoJiaHOG: a hierarchy oriented geo-aware image caption dataset for remote sensing image-text retrieval. ISPRS Journal of Photogrammetry and Remote Sensing, 222: 130-151 [DOI: 10.1016/j.isprsjprs.2025.02.009]

Zhao Z H, Wu C, Cao X Y, Wang D, Chen H R X, Tang D T, et al. 2025b. ChangeBridge: spatiotemporal image generation with multimodal controls for remote sensing [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2507.04678.pdf

Zhao Z M, Gao L R, Chen D, Yue A Z, Chen J B, Liu D S, et al. 2019. Development of satellite remote sensing and image processing platform. Journal of Image and Graphics, 24(12): 2098-2110 (赵忠明, 高连如, 陈东, 岳安志, 陈静波, 刘东升, 等. 2019. 卫星遥感及图像处理平台发展. 中国图象图形学报, 24(12): 2098-2110) [DOI: 10.11834/jig.190450]

Zhao Z P, Miao X R, He C, Hu J F, Min B B, Gao Y M, et al. 2024. Masking-based cross-modal remote sensing image-text retrieval via dynamic contrastive learning. IEEE Transactions on Geoscience and Remote Sensing, 62: #5626215 [DOI: 10.1109/TGRS.2024.3406897]

Zhao Z P, Miao X R, Liu L, Xu X Z, Liu Y, Hu J F, et al. 2025c. Sparse-guided partial dense for cross-modal remote sensing image-text retrieval. IEEE Transactions on Geoscience and Remote Sensing, 63: #5617213 [DOI: 10.1109/TGRS.2025.3555956]

Zheng C Y, Li X, Liang X Y, Huang L, Du S, Nie J, et al. 2025a. Cross-modal progressive perspective matching network for remote sensing image-text retrieval. IEEE Transactions on Multimedia, 27: 3966-3978 [DOI: 10.1109/TMM.2025.3535365]

Zheng C Y, Wen Q, Li X, Yang C X, Nie J, Guo Y Y, et al. 2025b. Whole semantic sparse coding network for remote sensing image-text retrieval. IEEE Transactions on Geoscience and Remote Sensing, 63: #5406913 [DOI: 10.1109/TGRS.2025.3604386]

Zheng Y J, Wu W J, Li Q Y, Wang X H, Zhou X, Ren A A, et al. 2025c. InstructSAM: a training-free framework for instruction-oriented remote sensing object recognition//Proceedings of the 39th Annual Conference on Neural Information Processing Systems. Mexico City, Mexico:[s.n.]

Zheng Z, Ermon S, Kim D, Zhang L P and Zhong Y F. 2025d. Changen2: multi-temporal remote sensing generative change foundation model. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(2): 725-741 [DOI: 10.1109/TPAMI.2024.3475824]

Zheng Z, Tian S Q, Ma A L, Zhang L P and Zhong Y F. 2023. Scalable multi-temporal remote sensing change data generation via simulating stochastic change process//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 21761-21770 [DOI: 10.1109/ICCV51070.2023.01994]

Zheng Z, Zhong Y F, Zhang L P and Ermon S. 2024. Segment any change//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #2581

Zhi Y J, Jiang Y W, Yang Z, Chen Y Z, Hao W K, Ma M Y, et al. 2026. Development status and prospects of pretrained foundation models for remote sensing imagery. Journal of Image and Graphics, 31(4): 973-986 (支元杰, 姜艺伟, 杨知, 陈奕州, 郝文魁, 马明阳, 等. 2026. 面向遥感图像的预训练基础模型发展现状与展望. 中国图象图形学报, 31(4): 973-986) [DOI: 10.11834/jig.250424]

Zhong S R, Hao X X, Yan Y B, Zhang Y, Song Y Q and Liang Y X. 2024. UrbanCross: enhancing satellite image-text retrieval with cross-domain adaptation//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM: 6307-6315 [DOI: 10.1145/3664647.3680604]

Zhou C, Loy C C and Dai B. 2022. Extract free dense labels from CLIP//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 696-712 [DOI: 10.1007/978-3-031-19815-1_40]

Zhou C F, Wang J, Liu X Y and Zhang X K. 2026. Geospatial-reasoning-driven vocabulary-agnostic remote sensing semantic segmentation [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2602.08206.pdf

Zhou Y, Li J J, Ou C Y, Yan D W, Zhang H K and Xue X Z. 2025. Open-vocabulary object detection in UAV imagery: a review and future perspectives [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2507.13359.pdf

Zhu G Y, Yang B W, Zhuang Y, Zhang T, Wang G Q, Che Z H, et al. 2026. A training-free guess what vision language model from snippets to open-vocabulary object detection [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2601.11910.pdf

Zhu Q, Lao J W, Ji D Y, Luo J W, Wu K, Zhang Y Y, et al. 2025a. SkySense-O: towards open-world remote sensing interpretation with vision-centric visual-language modeling//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 14733-14744 [DOI: 10.1109/CVPR52734.2025.01373]

Zhu X X, Tuia D, Mou L C, Xia G S, Zhang L P, Xu F, et al. 2017. Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine, 5(4): 8-36 [DOI: 10.1109/MGRS.2017.2762307]

Zhu Y S, Li L, Chen K Y, Liu C Y, Zhou F G and Shi Z W. 2025b. Semantic-CD: remote sensing image semantic change detection towards open-vocabulary setting//Proceedings of 2025 IGARSS IEEE International Geoscience and Remote Sensing Symposium. Brisbane, Australia: IEEE: 6388-6392 [DOI: 10.1109/IGARSS55030.2025.11243524]

Zhu Z Q and Yang B W. 2025. UniVCD: a new method for unsupervised change detection in the open-vocabulary era [EB/OL]. [2026-03-20]. https://arxiv.org/pdf/2512.13089.pdf

作者简介

李开宇，男，博士研究生，主要研究方向为计算机视觉和遥感图像解译。E-mail: likyoo.ai@gmail.com

曹相湧，通信作者，男，副教授，主要研究方向为机器学习和计算机视觉。E-mail: caoxiangyong@mail.xjtu.edu.cn

蒋梓轩，男，本科生，主要研究方向为遥感多模态大模型。E-mail: andrewjiang@stu.xjtu.edu.cn

孟德宇，男，教授，主要研究方向为机器学习和计算机视觉。E-mail: dymeng@mail.xjtu.edu.cn