面向开放场景理解的遥感点云多模态融合与开放词汇技术综述

王琼洁
中国电子信息产业发展研究院，北京 100036
DOI: 10.61369/TACS.2026050048

摘　　要： 随着机载激光雷达、车载移动测量与无人机倾斜摄影等技术的快速发展，遥感场景中产生了海量高精度三维点云数据。传统遥感点云解译方法多建立在封闭类别监督范式之上，依赖大量点级标注，且在未知类别识别、跨场景迁移和复杂语义查询等方面存在明显局限。近年来，随着多模态预训练、开放词汇视觉理解以及大语言模型的发展，这些方式为遥感点云从“封闭集语义分割”走向“开放场景三维理解”提供了新的技术路径。基于此，本文围绕这一演进脉络，对相关研究进行系统梳理。首先，从遥感点云的大规模、稀疏性与弱纹理特征出发，概述了适用于大场景处理的点云表征骨干网络与自监督预训练范式。其次，本文重点总结了基于二维视觉－语言模型的2D-to-3D 跨模态蒸馏与特征反投影方法与面向点云直接输入的三维语言模型构建方法这两类关键路线。最后，结合测绘遥感任务对几何精度、跨尺度一致性与工程部署的特殊要求，分析了现有方法在空间精确对齐、纯几何语义表达、长尾类别识别、数据集构建与可信落地等方面的主要瓶颈。

关 键 词： 遥感点云；开放词汇分割；多模态融合；三维基础模型；3D-LLM；跨模态蒸馏

A Review of Multi-modal Fusion and Open Vocabulary Techniques for Remote Sensing Point Clouds in Open Scene Understanding

Wang Qiongjie
China Center for Information Industry Development, Beijing 100036

Abstract: With the rapid development of technologies such as airborne LiDAR, vehicle-mounted mobile measurement, and UAV oblique photogrammetry, massive amounts of high-precision 3D point cloud data have been generated in remote sensing scenes. Traditional remote sensing point cloud interpretation methods are mostly based on closed-category supervised paradigms, relying on a large number of point-level annotations, and have significant limitations in areas such as unknown category identification, cross-scene transfer, and complex semantic queries. In recent years, with the development of multimodal pre-training, open-vocabulary visual understanding, and large language models, these approaches have provided new technical paths for remote sensing point clouds to move from "closed-set semantic segmentation" to "open-scene 3D understanding." Based on this, this paper systematically reviews relevant research along this evolutionary path. First, starting from the large-scale, sparse, and weakly textured features of remote sensing point clouds, this paper outlines point cloud representation backbone networks and self-supervised pre-training paradigms suitable for large-scene processing. Second, this paper focuses on summarizing two key approaches: 2D-to-3D cross-modal distillation and feature back-projection methods based on 2D vision-language models, and 3D language model construction methods for direct point cloud input. Finally, and considering the specific requirements of remote sensing mapping tasks for geometric accuracy, cross-scale consistency, and engineering deployment, this paper analyzes the main bottlenecks of existing methods in areas such as precise spatial alignment, pure geometric semantic representation, long-tail category recognition, dataset construction, and reliable deployment.

Keywords: remote sensing point clouds; open-vocabulary segmentation; multimodal fusion; 3D foundation models; 3D-LLM; cross-modal distillation

引言

三维点云能够直接表达地物的几何形态、空间位置和结构关系，在地球空间信息科学与广义点云智能处理体系中，其已经成为数字孪生城市、自然资源调查、灾害监测和精细测图的重要数据基础[1]。与二维遥感影像相比，点云在描述建筑立面、植被垂直结构和地表高程变化方面具有天然优势；但同时，点云也具有数据规模大、采样不规则、密度变化强和纹理信息弱等特点，使其在特征学习与语义表达上显著区别于传统图像任务[2]。特别是在对地观测领域，航空点云作为一种离散的空间数据表征，能够精确刻画复杂场景的几何拓扑结构[3]。过去十余年，遥感点云解译主要围绕语义分割、目标检测和实例提取展开，形成了以稀疏卷积、点集学习和图结构建模为代表的一系列方法。这些方法在封闭类别设定下取得了较好效果，但通常需要大量精细标注，并且难以回答“未知类别是什么”、“文本描述对应哪些三维区域”以及“模型能否跨城市、跨平台迁移”等更开放的问题。与此同时，深度学习模型在实际应用中通常需要大量人工标注的训练样本，且其泛化性能相对较弱。近年来，视觉基础模型和大语言模型的发展为遥感图像处理的大模型研究引入了新的范式[4]。随着开放词汇视觉理解和多模态大模型的发展，研究重心正在从预定义类别识别逐步转向语言驱动的三维场景理解。对于遥感点云而言，这一变化不仅意味着类别集合的扩展，更意味着模型需要同时处理几何结构、传感器差异、跨尺度语义和测绘精度约束等问题。基于此，我们在保留通用三维开放词汇研究主线的同时，更强调其与遥感点云场景之间的适配关系。具体而言，我们并不将所有通用三维方法直接等同于遥感专用方法，而是从“可迁移技术路径”和“遥感应用约束”两个层面进行综述，以期为后续研究提供较为清晰的技术图谱。本文旨在梳理遥感点云从封闭集走向开放场景理解的技术脉络，重点对比跨模态蒸馏与三维语言建模路线，并剖析其现实挑战。

一、遥感点云开放场景理解的技术演进

遥感点云的开放场景理解并不是凭空出现的，它建立在大规模点云表征与三维预训练持续发展的基础之上。首先，从数据属性看，遥感点云来源复杂，既包括机载激光雷达形成的大范围地形地物点云，也包括车载移动测量、地面激光扫描和摄影测量重建点云。不同平台在视角、点密度、回波强度和噪声特征方面存在显著差异，因此同一模型在不同数据源间往往面临较强的域偏移。与室内场景点云相比，遥感点云常常覆盖范围更大、类别层级更复杂，且类别分布具有明显长尾特征，这决定了其开放词汇理解需求更强，也更难直接复用通用三维方法。为了更清晰地呈现这一发展脉络，本文梳理了该领域代表性技术的演进时间轴（如图1所示）。可以看出，整个技术生态经历了从早期的基础网络奠基与效率提升（2017–2020），到基于 Transformer 的自监督预训练（2021–2022），再到如今的开放词汇理解与多模态语言交互（2023至今）的深刻演变。

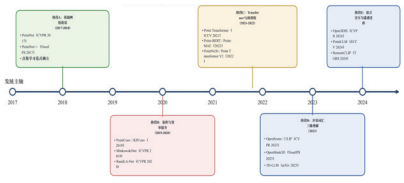

图1 遥感点云开放场景理解代表性技术发展时间轴

早期，基于结构化表征的点云语义分割方法都存在数据转换过程中空间信息丢失的问题，不可避免地限制其网络性能。而基于原始点云的深度学习方法直接把三维点云不做任何处理作为卷积神经网络的输入，这样保存了数据原有的空间信息。为了应对大场景点云的计算压力，研究者提出了两类具有代表性的底层表征路线。一类是直接点处理方法，典型如 RandLA-Net[5]，通过随机采样与局部特征聚合提升百万级点云的处理效率；另一类是基于稀疏卷积的体素化方法，典型如 MinkowskiNet[6]，通过仅在非空体素上进行卷积运算以控制计算复杂度。前者在保持原始点结构方面具有优势，后者在大规模场景建模和工程实现方面较为成熟。对遥感点云而言，二者并非简单替代关系，而是需要结合场景尺度、下游任务和硬件资源进行选择。

在此基础上，由于大多数语义分割方法工作在全监督的模式下，为数据标注带来了极大的压力，为了解决对于大规模点云标注数据的依赖问题[7]，三维预训练成为近年的重要方向。Point-BERT[8] 将掩码建模思想引入点云 Transformer，通过恢复被遮蔽的局部点块学习通用表征；Point-MAE[9] 进一步改进了掩码自编码策略，在分类、少样本学习和分割等任务中表现出较好的迁移能力。对于遥感点云，预训练的意义不仅在于降低标注依赖，更在于通过大规模无标签数据学习跨区域、跨平台的几何先验。不过，现有三维预训练研究仍以物体级或室内场景数据为主，如何使预训练目标更贴合遥感中的大尺度拓扑关系、地物层级结构和多源异构特征，仍是需要继续推进的问题。

二、开放词汇三维理解的主要技术路线

（一）基于二维视觉－语言模型的2D-to-3D 蒸馏

在缺乏大规模“三维点云－文本”配对数据的背景下，利用成熟的二维视觉－语言模型为三维网络提供语言语义，是当前开放词汇三维理解最具代表性的路线之一。以 OpenScene[10] 为代表的方法，通过多视角渲染、二维密集特征提取和三维反投影，将 CLIP[11] 等二维大模型的语义空间迁移到三维点云特征上，从而支持零样本的开放词汇查询。此类方法的优势在于能够借助二维大规模图文对预训练成果，较快建立点云与自然语言的联系；但其性能高度依赖视图选择、相机配准和遮挡处理，对于遥感点云中

常见的大尺度场景、低纹理区域和多平台几何误差尤为敏感。

（二）面向开放查询的实例级与区域级建模

除了逐点语义蒸馏，OpenMask3D[12] 等工作进一步将开放词汇能力拓展到三维实例级分割。其基本思想是先生成类别无关的三维实例候选，再借助多视角图像特征与语言嵌入完成语义匹配。相较于逐点特征方法，这类方法更适合处理以“对象”为中心的文本查询，如特定设施、道路附属物或局部构件。Open3DIS[13] 则构建了三维类别不可知提议与二维基础大模型的深度协同机制，不仅能精准提取三维空间中的细粒度目标，还能有效处理极其长尾的复杂实例级语言查询。然而，遥感场景中的地物边界常常不规则、尺度差异极大，许多地物并不天然具备清晰的实例定义，因此实例级开放词汇方法在遥感中的适用性仍有待进一步验证。

（三）面向遥感点云的适配问题

通用开放词汇三维方法为遥感点云研究提供了重要启发，但其并不能直接等同于解决方案。一方面，遥感点云往往缺乏稳定的 RGB 纹理，甚至仅包含 XYZ 与强度信息，导致许多依赖图像外观语义的方法难以保持效果；另一方面，测绘任务更关注边界位置、拓扑关系和几何精度，而不仅是语义标签是否匹配。因此，遥感点云的开放词汇理解需要在语义泛化和几何可靠性之间建立新的平衡。

三、三维语言模型及其对遥感点云的启示

（一）从跨模态蒸馏到三维语言建模

随着大语言模型的发展，研究开始探索将三维场景直接纳入语言模型的输入空间。视觉 - 文本会话式遥感大模型利用大语言模型强大的语义理解能力，通过向大语言模型附加特定于任务的组件来支持密集预测任务 [14]。在此框架下，3D-LLM[15] 尝试通过三维特征提取、视觉语言桥接和指令数据构建，使模型能够完成三维问答、描述和定位等任务。这表明三维场景理解正在从“语言辅助分类”走向“语言驱动交互”。PointLLM[16] 则进一步聚焦点云输入，通过点云编码器与大语言模型之间的映射，实现对点云对象的语言理解与响应生成。尽管这些工作多数仍以对象级或室内点云为主，但它们说明：三维点云不再只是传统意义上的几何输入，而可以逐步演化为可参与对话、检索和推理的语义载体。3D-LLM 的潜力在于：支持自然语言检索罕见设施、辅助交互式标注，并将几何、影像与文本知识统一到同一框架下。然而，这类潜力目前仍主要停留在通用三维基准层面，距离遥感应用中的高精度定位、复杂场景鲁棒推理和工程可部署性仍有较大差距。

（二）需要谨慎看待的现实限制

首先，现有三维语言模型所依赖的数据体系与遥感场景差异较大，训练语料中对遥感专业对象、地理过程和测绘知识的覆盖有限。其次，许多模型通过离散 Token 或区域特征压缩三维输入，这虽然有助于降低计算量，却可能削弱厘米级或分米级几何细节的保持能力。再次，语言模型固有的幻觉问题在遥感场景中风险更高，因为错误的三维定位或错误的地物描述可能直接影响测图、巡检和灾害研判。因此，将 3D-LLM 概念迁移到遥感点云时，应坚持“能力提升”和“可靠约束”并重的技术路线。

四、结论

总体来看，遥感点云理解正在经历由封闭集监督识别向开放场景多模态理解的深刻转变。稀疏卷积、点云预训练、二维到三维跨模态蒸馏以及三维语言模型等技术，共同构成了这一转变的主要推动力量。需要指出的是，现阶段真正面向遥感点云的开放词汇研究仍处于发展阶段，很多代表性方法来自通用三维视觉领域，其对遥感场景的适配仍需结合测绘精度要求、数据源差异和实际业务目标进行再设计。从综述视角看，这一领域的研究评价也不应仅停留于网络精度比较，而应同时关注数据组织方式、语义空间构建、几何约束机制以及业务场景下的可解释性与可落地性。未来，随着遥感专用三维基础模型、高质量多模态数据体系和更可靠的评测标准不断完善，开放词汇与语言驱动的三维理解有望成为遥感智能解译的重要发展方向。

参考文献

[1] 杨必胜，董震 . 点云智能研究进展与趋势 [J]. 测绘学报 ,2019,48(12):1575-1585.

[2] 胡伏原，李晨露，周涛，等 . 面向深度学习的三维点云补全算法综述 [J]. 中国图象图形学报 ,2025,30(2):309-333. DOI:10.11834/jig.240124.

[3] 潘洁晨，邢帅，曹家印，等 . 基于深度学习的航空点云语义分割研究进展 [J]. 地球信息科学学报，2025,27(9):1999-2020. DOI:10.12082/dqxxkx.2025.250151.

[4] 张帅豪，潘志刚 . 遥感大模型：综述与未来设想 [J]. 遥感技术与应用 ,2025,40(1):1-13. DOI:10.11873/j.issn.1004-0323.2025.1.0001.

[5]Hu Q, Yang B, Xie L, et al. Randla-net: Efficient semantic segmentation of large-scale point clouds[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11108-11117.

[6]Choy C, Gwak J Y, Savarese S. 4d spatio-temporal convnets: Minkowski convolutional neural networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3075-3084.

[7] 郑智鸿，宋海川 . 基于组对比学习的弱监督三维点云语义分割方法 [J]. 华东师范大学学报（自然科学版），2024(2):108-118. DOI:10.3969/j.issn.1000-5641.2024.02.012.

[8]Yu X, Tang L, Rao Y, et al. Point-bert: Pre-training 3d point cloud transformers with masked point modeling[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 19313-19322.

[9]Pang Y, Tay E H F, Yuan L, et al. Masked autoencoders for 3d point cloud self-supervised learning[J]. World Scientific Annual Review of Artificial Intelligence, 2023, 1: 2440001.

[10]Peng S, Genova K, Jiang C, et al. Openscene: 3d scene understanding with open vocabularies[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 815-824.

[11]Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PmLR, 2021: 8748-8763.

[12]Takmaz A, Fedele E, Sumner R W, et al. Openmask3d: Open-vocabulary 3d instance segmentation[J]. arXiv preprint arXiv:2306.13631, 2023.

[13]guyen P, Ngo T D, Kalogerakis E, et al. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 4018-4028.

[14] 付琨，卢宛萱，刘小煜，等 . 遥感基础模型发展综述与未来设想 [J]. 遥感学报，2024, 28(7).DOI:10.11834/jrs.20233313.

[15]Hong Y, Zhen H, Chen P, et al. 3d-llm: Injecting the 3d world into large language models[J]. Advances in Neural Information Processing Systems, 2023, 36: 20482-20494.

[16]Xu R, Wang X, Wang T, et al. Pointllm: Empowering large language models to understand point clouds[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 131-147.

