引文格式:张云生,王旭颖,陈斯飏,等.点云智能理解:从点云中心至语言中心的转变[J].武汉大学学报(信息科学版),2026,51(6):1073-1087.DOI:10.13203/j.whugis20240497

Citation:ZHANG Yunsheng,WANG Xuying,CHEN Siyang,et al.Point Cloud Intelligent Understanding: Transition from Point-Centric to Language-Centric Paradigm[J].Geomatics and Information Science of Wuhan University,2026,51(6):1073-1087.DOI:10.13203/j.whugis20240497

点云智能理解:从点云中心至语言中心的转变

张云生[1,2,3] 王旭颖[1] 陈斯飏[1] 李海峰[1]

1 中南大学地球科学与信息物理学院,湖南 长沙,410075
2 实景三维建设与应用技术湖南省工程研究中心,湖南 长沙,410000
3 高速铁路建造技术国家工程研究中心,湖南 长沙,410075

摘 要:三维时空信息在实景三维中国建设的国家新型基础设施战略中占据重要地位,机载航空点云数据在大规模三维时空信息获取中发挥了关键作用。围绕点云数据的深度解译与语义信息提取需求,系统梳理了点云理解范式的发展脉络,从前深度学习时代的模型驱动与数据驱动范式,到后深度学习时代的自监督信号驱动范式,再到当前基于预训练基础大模型的大模型驱动范式。在此基础上,提出了以语言为中心的机载航空点云理解范式,系统分析了其面临的关键挑战及未来发展方向,旨在为点云智能理解的技术创新与研究拓展提供新的视角和思路。

关键词:机载航空点云;点云理解范式;大语言模型;视觉语言模型;以语言为中心

中图分类号:P237 **文献标识码**:A **收稿日期**:2025-04-16

DOI:10.13203/j.whugis20240497 **文章编号**:1671-8860(2026)06-1073-15

Point Cloud Intelligent Understanding: Transition from Point-Centric to Language-Centric Paradigm

ZHANG Yunsheng[1,2,3] *WANG Xuying*[1] *CHEN Siyang*[1] *LI Haifeng*[1]

1 School of Geosciences and Info-Physics, Central South University, Changsha 410075, China
2 Hunan Engineering Research Center of 3D Real Scene Construction and Application Technology, Changsha 410000, China
3 National Engineering Research Center of High-Speed Railway Construction Technology, Changsha 410075, China

Abstract: **Objectives**: In recent years, 3D spatiotemporal information has become a cornerstone of national new infrastructure construction, with airborne point clouds playing a critical role in large-scale 3D data acquisition. However, effectively interpreting these point clouds and extracting their semantic richness remain significant challenges. Traditional methods, including geometry-driven feature extraction and data-driven deep learning models, have achieved limited progress but struggle with the complexity and scale of airborne point clouds. Advances in large-scale foundation models (FM), particularly in natural language processing and vision-language integration, offer new opportunities for point cloud understanding. Large language models exhibit exceptional generalization and cross-modal semantic capabilities, enabling a language-centered paradigm. By leveraging language models to map point cloud data into a semantically enriched space, this approach addresses limitations of traditional methods. **Methods**: The evolution of point cloud understanding is examined across geometry- and data-driven approaches, self-supervised paradigms, and FM-driven methodologies. A language-centered framework for airborne point cloud understanding is proposed, tackling high-level semantic modeling, cross-modal alignment, and downstream task adaptation. Results demonstrate enhanced semantic representation, improved generalization, and significant advantages in com-

基金项目:国家自然科学基金(42171440,42271481);实景三维建设与应用技术湖南省工程研究中心开放课题(3DRS2024Z1);中国中铁股份有限公司科技研究开发计划重大专项(2021-专项-08);湖南省自然科学基金(2022JJ30698)。

第一作者:张云生,博士,教授,研究方向为基于人工智能/深度学习的点云解译。zhangys@csu.edu.cn

通信作者:李海峰,博士,教授。lihaifeng@csu.edu.cn

plex scenarios. **Results**: The findings provide new insights into point cloud understanding and establish a foundation for integrating large-scale models into 3D applications. **Conclusions**: These contributions offer innovative perspectives and technical solutions for advancing point cloud technologies in national infrastructure projects.

Key words: airborne point clouds; point cloud understanding paradigms; large language models; vision-language models; language-centered paradigm

实景三维技术作为一种能够真实、立体、动态地反映人类生产、生活和生态空间的时空信息手段,已成为国家新型基础设施建设的重要组成部分。为全面推动中国实景三维建设,自然资源部先后发布了《实景三维中国建设技术大纲(2021版)》《新型基础测绘与实景三维中国技术文件》以及《关于全面推进实景三维中国建设的通知》等政策文件及通知。在这一过程中,三维时空信息扮演着基础性且关键的角色,其中机载航空点云数据在获取三维时空信息中发挥了核心作用[1-4]。

机载激光扫描仪和倾斜航空摄影测量技术提供了大量高质量的机载点云数据。然而,要支持实景模型的重建和复杂地物关系的推理,需要对这些点云数据进行深度解译,以挖掘其蕴含的丰富语义信息[5-10]。点云解译的核心使命在于通过三维点云数据准确解析构成现实世界的复杂地物及其相互关系。然而,目前主流的点云理解任务大多局限于将点云映射为有限的固定目标类别,如S3DIS数据集的12类[11]、semantic3D数据集的8类[12]、WHU-Railway3D数据集的11类[13]、ISPRS Vaihingen数据集的9类[14]等。此类方法通过简化任务目标来平衡任务复杂度,重点在于对有限目标和场景的学习,从而推动了航空点云理解技术的初步发展。然而,这些方法在面对训练集之外的开放世界时,表现出较弱的泛化能力,尤其是在遇到未见过的场景和目标时,其语义理解能力仍存在显著不足。如图1所示,现有方法对点云的理解尚停留在中间阶段,即仅能识别有限、少数类别的物体,而缺失了对场景语义的深度挖掘,未能实现从类别识别迈向洞悉本质的深层语义理解。

大语言模型(large language model,LLM)被认为是人类历史上成功将语言知识以压缩形式编码的强大工具,其出现为多模态数据理解,尤其是视觉-语言模型的构建,提供了全新思路。在三维领域,LLM已被广泛应用于室内场景的点云数据等解译任务,显著增强了对超越单一类别

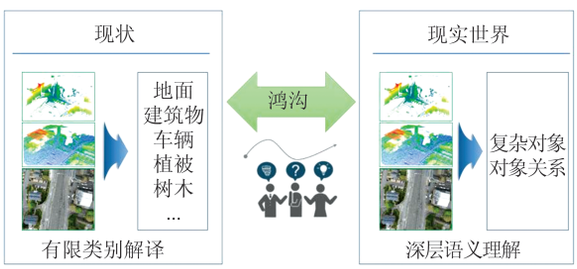


图1 点云理解现状与理想模式的鸿沟

Fig. 1 The Gap Between Status Quo and Ideal Model in Point Clouds Understanding

标签的对象间、对象与环境间以及对象与其他语义元素之间复杂关系的解析与建模能力。这一成功经验启发了研究者将其应用扩展至更具挑战性的室外场景,尤其是机载航空点云数据的理解中,为其语义解译和智能分析开辟了新的可能性。通过将点云映射到语言空间,LLM不仅能够为三维数据赋予丰富的对象级语义,还能推理场景中复杂的关系逻辑[15-18]。这一能力促使从根本上重新审视并重塑点云理解的任务内涵:从对有限固定类别的识别,转向基于全球知识体系(如地理常识、物理规律)的深度语义解译,实现三维空间中对象-关系-场景的层次化认知。

基于上述观点,本文首先回顾了点云智能理解范式的演进历程,重点分析了点云理解从以传统几何信息为中心向以语言知识为中心的新认知方式转变的趋势。该演进过程可划分为4个阶段:前深度学习时代的模型驱动范式和数据驱动范式,这两个阶段主要依赖传统算法和大规模数据输入以实现点云的基础理解;后深度学习时代的自监督信号驱动范式,标志着点云理解开始依赖智能算法实现自我学习和优化,从而减少对人工干预的依赖;当前的基础大模型时代,大模型驱动范式展现了点云理解向模拟人类认知和语言处理的显著转变。这4个阶段的理解范式、第一性原则、特征及主要任务总结如表1所示。此外,本文提出了以语言为中心的点云理解范式,系统分析了其面临的主要研究问题,并对未来的研究方向和技术创新进行了展望。

表 1　各个阶段的点云理解范式、第一性原则、特征及主要任务

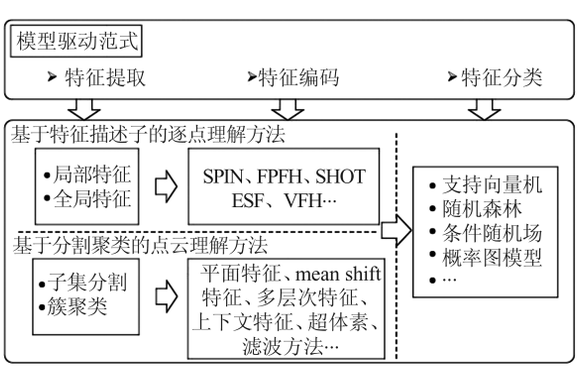
Table 1　Paradigm, Pirst Principles, Characteristics and Main Tasks of Point Cloud Understanding at Each Era

发展时代	理解范式	第一性原则	特征	主要任务
前深度学习时代	模型驱动范式（少标注-有限类别-单任务）	理论驱动	手工设计特征 自旋图(spin image,SPIN)、快速点特征直方图(fast point feature histograms,FPFH)、方向直方图签名(signature of histograms of orientations,SHOT)等	点云分类、目标识别、点云配准等单一任务
	数据驱动范式（多标注-有限类别-单任务）	通用逼近理论	架构 PointNet、点卷积神经网络(point convolutional neural network,PointCNN)、图卷积网络(graph convolutional network,GCN)、Transformer等	语义分割、目标检测等单一任务
后深度学习时代	自监督信号驱动的范式（弱标注-有限类别-单任务）	数据内在的特性作为数据独立、任务独立的通用表征	生成式学习 对比式学习	预训练通用任务-适配少数下游任务
基础大模型时代	大模型驱动的范式（强交互标注-无限类别-多任务）	数据压缩理论基础上的 Token 预测	多模态弱对齐(语言为中心)	包含世界关系的下游任务(语义分割、字幕描述、物体定位等)

1　点云智能理解范式研究进展

1.1　前深度学习时代——模型驱动范式

在前深度学习时代,点云理解以模型驱动范式为主,通过理论驱动和精心设计的特征编码方式,将点云几何信息映射到有限类别,完成分类、目标识别和配准等任务。研究者基于先验理论开发了适应不同场景的点云处理模型,在有限计算资源下为点云基础理解提供了解决方案。基于特征描述子的逐点理解法和基于分割聚类的点云理解方法是该范式下的两大主要策略,如图 2 所示。



注:ESF(ensemble of shape functions):形状函数集成;VFH(viewpoint feature histogram):视点特征直方图。

图 2　前深度学习时代模型驱动范式

Fig. 2　The Pre-Deep Learning Era: A Model-Driven Paradigm

1.1.1　基于特征描述子的逐点理解方法

基于特征描述子的逐点理解方法是该范式典型的点云理解方法之一,其核心在于为每个点提取描述子,捕捉局部几何特征及邻域信息,从而赋予点云数值化的几何特征。分类器利用这些特征实现逐点独立分类。手工设计的特征描述子通常根据关注的尺度分为两类:(1)局部特征,如 SPIN[19]、FPFH[20]、SHOT[21]。(2)全局特征,如 ESF 采用多种形状函数(如距离、角度、面积分布等)的组合描述三维对象的全局形态特征,尤其适用于大尺度地物分类[22],VFH 结合物体表面法线分布与观测视角信息,生成对视角变化鲁棒的全局特征,提升了跨视角场景下的识别能力[23]。WEINMANN 等[24]结合全局和局部的视角提取了 21 种特征,并使用随机森林分类器实现了逐点分类结果,成为这一领域的经典工作之一。此外,GUO 等[25]对常用的 10 种三维局部描述子进行了系统性的测试和评价,为分类工作的描述子选择提供了参考和指导。虽然逐点特征计算能够获得细粒度的表征,但由于忽略了点之间的上下文关系,其分类结果容易受到数据噪声和异常点的影响,进而导致结果的不平滑和不一致。为了改善上述问题,一些工作尝试在初始分类结果基础上进行后处理优化,如条件随机场[26]和基于图结构的优化方法[27]。

逐点方法依赖显式几何建模,无需大规模标注数据,同时特征表达具有较强的可解释性,使得点云理解过程更加透明。但其主要缺陷在于计算成本高、对特征描述子的依赖强以及特征提取过程中对先验知识和参数的敏感性。面临大规模复杂场景时,稳定性和泛化能力下降,使该类算法在不同场景中的表现难以保持一致。

1.1.2　基于分割聚类的点云理解方法

基于分割聚类的点云理解方法是模型驱动

范式中另一类常见的点云理解策略。该方法将空间邻近或特征相似的点聚合为多个子集或簇,每个簇通常对应一个区域或物体。通过简化点云复杂性,该方法为后续分类提供了高效的处理基础。聚类完成后,再对每个簇进行分类,以确定其所属的物体类别或场景类型。针对子集或簇的构建,有两种常见的思路:(1)先聚合再分类。例如,XU等[28]在城市场景中使用多种实体进行特征提取,其中包括平面和mean shift分割段。HUANG等[29]提出了一种语义标记框架,结合多层次(点-分段-对象)的特征和上下文信息,首先将非地面点进行多规则区域生长分割,然后计算相关特征并输入支持向量机进行分类。KANG等[30]先基于相似的几何和辐射属性生成超体素作为基本单元,然后构建多尺度视觉特征,并利用主题模型分析超体素内点的语义关联,最终通过概率图模型实现分类。(2)先分类再聚合。YANG等[31]通过设置超参数将点的分类结果进行聚合,并结合归一化切分方法控制过分割和欠分割问题,从而优化分类结果。此外,分类过程还可以与逐点理解方法相结合,先利用逐点特征完成初步点级分类,随后通过聚类算法将相似点归为同一类别。VERDOJA等[32]通过几何与颜色特征对点云进行粗分类,并基于此对超体素进行相似性聚类,实现高效且层次化的点云分割。

基于分割聚类的分类方法通过减少处理点的数量显著提升了点云理解模型的效率和鲁棒性。然而,该方法高度依赖先验知识,难以在复杂或未见场景中构建具有良好适应性的模型。此外,超体素分割难以确保类别单一性,影响了后续处理的可靠性。

1.2 前深度学习时代——数据驱动范式

在前深度学习时代的中后期,点云理解逐步从手工设计特征转向数据驱动范式。该范式以通用逼近理论为基础,通过利用海量标注数据训练深度神经网络,实现对训练数据与标注之间的映射关系的学习,摆脱对传统特征工程的依赖。如图3所示,基于深度学习的点云理解方法可分为间接法(间接理解)和直接法(直接理解)。间接法通过将点云转换为多视角投影图或体素网格,对转换后的模式进行解译,并将结果反投影到原始点云。相比之下,直接法直接对原始点云的每个点进行分类,无需额外转换。

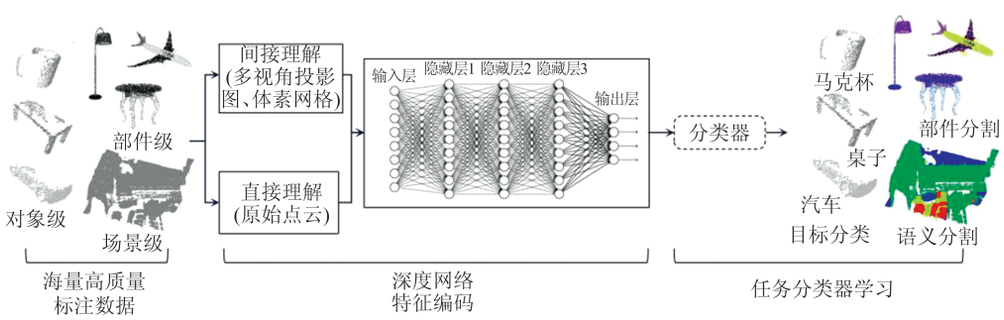


图3 前深度学习时代——数据驱动范式

Fig. 3 Pre-Deep Learning Era: A Data-Driven Paradigm

1.2.1 间接法

基于二维多视图投影和点云体素化的方法是两种常见的点云间接理解策略。基于二维多视图的方法通过将点云投影为不同视角的二维图像,利用成熟的二维图像理解模型对点云进行解译。例如,SU等[33]提出的多视图卷积神经网络(multi-view convolutional neural network,MVCNN)通过多视角图像实现三维物品分类,后续方法如多视角-MVCNN[34]、分组视角卷积神经网络[35]、池化约束点卷积神经网络[36]及多视角变换网络[37]分别从多分辨率特征提取、视角特征加权、池化操作优化、自适应视点选择等方面提升了性能。此类方法优势在于利用二维图像模型降低计算成本,但在处理密度不均或遮挡严重的机载点云时,可能丢失部分三维空间信息。基于体素的方法将三维点云划分为规则网格,并利用体素特征进行点云理解。VoxNet[38]是体素化方法的代表,但体素化方法计算内存开销较大。为此,研究者提出优化方案,如OctNet[39]和Kd-Net[40]分别利用八叉树和Kd-tree结构提升效率,MSNet[41]通过多尺度体素学习缓解边界问题,RPVNet[42]结合体素与点的优势,点-体素变换器[43]则引入稀疏注意机制提升体素化效率。近年来,稀疏卷积技术的引入进一步推动了体素化方法的革新。稀疏卷积仅对非空体素进行卷积操作,显著降低了计算和内存消耗。MinkowskiEngine[44]通过四维

稀疏张量表示与时空卷积,在点云分割与检测任务中实现了高效处理。SparseConvNet[45]则优化了并行计算能力,支持大规模稀疏点云的实时解析。此外,还有利用概率分布模型[46]、场检测滤波器[47]及变分自动编码器[48]优化体素化过程的方法。尽管体素化方法将点云转化为结构化形式便于处理,但面临高分辨率数据时计算开销大,如何在精度和计算资源间取得平衡是其面临的主要挑战。

1.2.2　直接法

基于邻域特征聚合的PointNet[49]是对直接点云数据本身进行理解的先驱,通过多层感知机与最大池化操作解决点云排列不变性问题,实现端到端点云处理。PointNet++[50]进一步结合最远点采样与多尺度特征融合,优化局部与全局信息整合,并通过Dropout抑制过拟合。随后,RandLA-Net[51]通过随机采样与高效特征聚合扩展感受野,提升了解译效率;GDANet[52]利用几何注意力划分边界和平面信息以增强特征聚合;PointNeXt[53]通过优化参数提升性能;PointMLP[54]采用轻量级几何仿射模块简化特征处理;DeLA[55]则解耦空间关系与局部聚合,进一步提升聚合效果。基于变种卷积神经网络的方法通过图卷积和3D卷积将卷积核扩展至3D模态,提取点云特征。动态图卷积神经网络(dynamic graph CNN,DGCNN)[56]通过边缘卷积保持排列不变性并提升局部特征捕捉能力,链DGCNN(linked DGCNN,LDGCNN)[57]结合动态卷积优化梯度计算,邻域图滤波卷积神经网络[58]和GridGCN[59]分别通过邻域图学习与优化查询模块提升速度与精度。PointCNN[60]利用连续体积卷积保持点云连续性,PointCNN采用X变换适应不规则点云[61],关系形状卷积神经网络[62]和核点卷积[63]通过改进卷积核与空间适应性进一步提升点云处理效果。近年来,基于注意力机制的Transformer架构在点云直接理解方面展现了巨大潜力。Point Transformer[64]通过位置编码增强局部特征建模,点云变换器[65]利用隐式偏移注意力满足排列不变性,Stratified Transformer[66]采用分层结构强化全局依赖建模能力。尽管Transformer对计算资源要求较高,但Point Transformer v2、v3[67-68]通过组向量注意力和多数据集训练显著提升了效率与精度,为点云智能理解提供了更强大的工具。

数据驱动的点云理解范式主要面临以下挑战:(1)高昂的标注成本是关键瓶颈,大规模、高质量的点云标注不仅耗时,还需投入大量资源。(2)标注过程依赖人工操作,不同标注者的主观判断可能导致标注不一致,进而使模型学习到错误信息,削弱其准确性和稳定性。为此,研究者正致力于开发减少标注数据依赖的新型学习范式,旨在提升模型的适应性和泛化能力,更有效应对多样化应用场景中的点云数据挑战。

1.3　后深度学习时代——自监督信号驱动范式

后深度学习时代的第一性原则是以数据内在特性为基础,构建任务独立、数据独立的通用表征,旨在通过通用表征提升模型的泛化能力,减少对特定任务和标注的依赖。自监督学习作为该阶段典型范式,通过代理任务挖掘数据内在结构,旨在生成不变性特征,为下游任务提供通用表征。点云理解中的自监督学习通常分为预训练-微调两阶段(如图4所示):预训练阶段设计代理任务,在大规模无标注数据集上训练模型获取初始参数;微调阶段利用少量标注数据训练,使模型适配特定下游任务。根据代理任务类型,自监督学习方法可分为基于对比式和基于生成式两类。

1.3.1　基于对比式自监督学习的点云理解方法

基于对比式自监督学习的点云理解方法通过构建正负样本对,在表征空间中拉近正样本、推远负样本,以学习点云数据的不变性特征。这类方法通常对点云进行刚体变换(如旋转、平移、尺度缩放),生成增强版本并编码为表征向量。PointContrast[69]是该类方法的早期代表,通过刚性变换生成点对(p,p')作为正负样本对,增强前后的点对为正样本,非对应点对为负样本。后续研究优化了正负样本选择策略、样本单元构建方法和引入多模态数据。例如,为解决正负样本选择中的语义矛盾,Proposal-Contrast[70]和Group-Contrast[71]利用表征聚类生成伪标注,以引导样本选择。JIANG等[72]通过投影头生成伪标注筛选正负样本。此外,难负样本策略[73]通过选择与正样本相似但非增强产物的点云块,提升了预训练效果。样本单元构建方面,从单点扩展到子块、体素、候选域及对象[74-75],以更高效地完成预训练任务。CrossPoint[74]引入影像数据,建立模态内(3D-3D)和模态间(2D-3D)对比学习框架,显著提升表征丰富性与分类精度。部分研究还结合多种样本单元进行混合使用[75],以增强多尺度表征学习效果。然而,对比式方法在样本构建和正负样本选择策略上仍需精心设计,以达到最

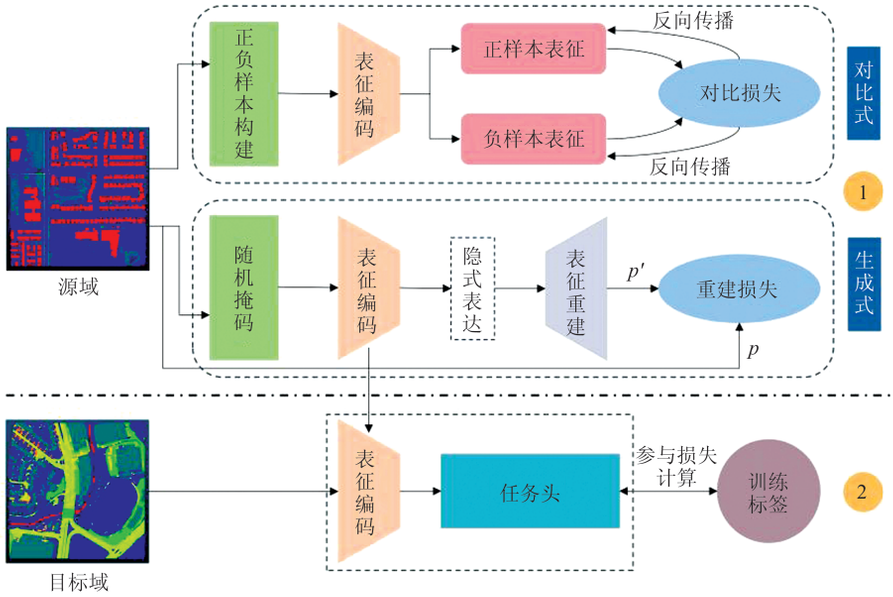


图4 后深度学习时代——自监督信号驱动范式

Fig. 4 Post-Deep Learning Era: A Self-Supervised Signal-Driven Paradigm

佳学习效果。

1.3.2 基于生成式自监督学习的点云理解方法

基于生成式自监督学习的点云理解方法通过设计代理任务,在恢复受扰动点云的过程中学习表征,主要分为掩码恢复任务和去噪恢复任务。掩码恢复任务是主流方向,受图像掩码任务的启发,研究者将类似策略引入三维点云。Point-BERT[76]将点云编码为离散词元(Token)并掩码后用BERT风格的Transformer恢复,同时引入对比学习和点块混合辅助,提升特征泛化能力;Point-MAE[77]改用标准Transformer,对60%~80%的点高比例掩码并轻量解码,提升了计算效率和位置保密性。用于层次化点云预训练的多尺度掩码自编码器[78]通过多尺度掩码增强了对局部结构和全局语义的理解。点云对比掩码自编码器[79]通过在点云生成式预训练框架中引入特征级对比约束,解决了对比学习辅助生成任务时的性能下降问题;Voxel-MAE[80]将方法拓展至大范围车载点云数据,结合体素化和高比例掩码实现高效表征学习。去噪恢复任务关注恢复受干扰点云,如密度变化、噪声和仿射变换。点云降噪自编码器[81]设计了通用去噪自编码器,并分析其在14种干扰类型下的表现。自监督点云上采样网络[82]通过自监督学习有效恢复稀疏点云细节信息,生成高分辨率的上采样点云数据。这类方法虽在特定任务中表现出色,但对代理任务设计依赖较高,不同类型扰动下的恢复效果差异较大,仍需进一步研究模型的通用性和适应性。

自监督学习仍存在其局限性:首先,代理任务的设计对模型性能至关重要,在机载航空点云中,如何捕捉多场景下点云的多尺度几何特征和复杂空间结构是关键挑战。此外,无标注数据与下游任务语义目标难以直接对齐,限制了模型在高层次语义理解中的表现。随着研究的深入,点云理解正向统一且具有更强泛化能力的学习框架演进,通过构建统一知识空间,提升模型对多场景和多任务的适配能力。

1.4 基础大模型时代——大模型驱动范式

早期的深度学习模型聚焦于特定任务的定制化,需要研究人员针对每个新任务从头开始设计和训练模型。随着计算能力的提升和数据量的增长,大型预训练模型逐渐兴起。这些模型能够在广泛的数据集上学习丰富的先验知识和通用特征,GPT[83]在自然语言处理领域的成功,展示了通过大规模预训练捕获深层次通用特征的潜力。基础大模型时代的核心原则是利用数据压缩理论来预测Token,其优势在于预训练完成后,模型可以快速适应多种下游任务,无需大量标注数据和重复训练。大模型时代推动了点云理解范式的变革,从传统的模型微调转向基于视觉和语言大型预训练模型的统一知识空间。这一转变为研究人员提供了新的探索维度,使他们能够借助语言更深入地挖掘点云数据中的语义信息。基础大模型时代下的点云理解主要基于3种方法:基于图像介导的点云理解方法,基于自然语言直接训练的点云理解方法,以及基于图像-

文字-点云多模态数据协同的点云理解方法,其一般范式如图5所示。

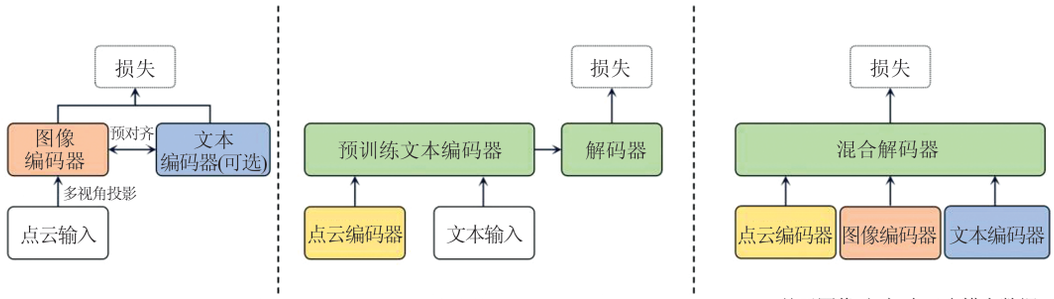


图 5　基础大模型时代——大模型驱动范式

Fig. 5　Foundation Model Era: A Foundation Models-Driven Paradigm

1.4.1　基于图像介导的点云理解方法

现有点云理解基础大模型面临的主要挑战是数据匮乏。相比二维图像数据,点云数据在多样性和覆盖范围上较为不足,限制了点云基础模型的训练。为此,研究者探索利用二维视觉基础模型和视觉语言基础模型(vision-language model, VLM)来弥补三维数据的不足。基于图像介导的点云理解方法通过将点云转换为二维图像,借助预训练的二维视觉基础模型进行三维数据理解。ZHANG 等[84]提出多视角投影法,将点云的结构信息与二维高级语义结合,显著提升了下游任务性能。YAN 等[85]通过二维知识迁移损失,引导三维特征与二维输出对齐。ZHOU 等[86]设计的 Uni3D 利用二维视觉预训练模型进行初始化,将三维模型扩展至十亿参数规模,在多项任务中刷新性能记录。KIRILLOV 等[87]提出的视觉基础模型(segment anything model, SAM)在 RGB 图像的零样本细粒度分割中表现出色。随后,研究人员将 SAM 扩展至三维领域[88-91],利用其对三维多视角投影图的预分割能力,将结果映射回三维空间,从而增强三维场景理解的表现。此外,OpenAI 的对比语言-图像预训练(contrastive language-image pre-training, CLIP)模型[92]在跨模态泛化中表现出色,为点云理解引入开放词汇理解提供了新方向。语言、图像与点云的统一表示学习(learning a unified representation of language, images, and point clouds, ULIP)[93]及其升级版 ULIP-2[94]将三维表征与 CLIP 的图像-文本空间对齐,实现了对象级三维分类的性能突破。OpenMask3D[95]和 OpenScene[96]则利用 CLIP 对三维场景中的对象属性进行实例分割。NAJIBI 等[97]结合车载点云与 VLM 的图像-文本知识,实现交通参与者的识别与追踪。尽管基于图像介导的方法展示了强大的潜力,但三维到二维的转换导致空间信息丢失,限制了点云中精确位置和结构细节的捕捉。此外,这些方法过于依赖二维中介,缺乏点云与自然语言的直接对齐能力,在复杂场景中的精度和语义理解深度仍有待提升。

1.4.2　基于自然语言直接训练的点云理解方法

基于自然语言直接训练的点云理解方法旨在直接对齐三维数据与自然语言,利用语言的开放词汇和语义泛化能力突破传统方法的局限。这类方法通过语言灵活描述三维对象属性(如材质、功能)及场景关系(如空间方位、动态交互),为点云理解提供了高效且通用的表达框架,不仅能精准识别已知类别,还可通过零样本学习扩展三维场景理解的边界。PointCLIP[98]是早期代表性工作,通过将点云投影为多视角深度图,与 CLIP 模型中的语言类别描述对齐,在数据稀缺条件下实现了有效的点云理解。ZHANG 等[99]进一步提出利用少量参数调整将三维点云分布与预训练语言模型对齐,激发语言模型的三维感知能力,从而提升多模态任务(如三维视觉定位)的表现。这些方法还支持多样化的语言理解点云任务,如点云描述生成、三维视觉问答、语言描述驱动的三维物体检索和三维场景解析。PointLLM[100]通过两阶段训练策略,先对潜在空间进行对齐,再对模型进行指令调优,展示了其对点云及常识的理解能力。ZHU 等[101]提出 3D-VisTA 模型,利用自注意力机制实现单模态建模与多模态融合,并通过掩码语言建模与场景-文本匹配预训练,在三维视觉语言任务上表现优异。QI 等[102]设计了 GPT4Point 模型,专注于三维对象的理解与生成,可高效处理多种点云语言任务,并在几何形状变更与颜色生成任务中展现出

高质量的结果。基于自然语言直接训练的方法在三维理解中仍面临诸多问题。首先,文本描述的局限性使其难以全面捕捉三维空间中的几何结构和精确的空间关系,无法充分展现三维场景的复杂性。其次,该方法对视觉模态中丰富的几何与外观信息利用不足,导致在场景细节的理解上存在缺陷。最后,在细粒度空间解析和多对象关系推理等复杂任务中,由于缺乏直观的视觉证据,模型表现不尽理想。

1.4.3 基于图像-文字-点云多模态数据协同的点云理解方法

为克服上述局限性,研究者们探索了基于图像-文字-点云多模态协同的方法,在混合解码器完成跨模态信息融合与任务推理,整合不同模态的编码特征,弥补单一模态信息的不足。LU等[103]提出了一种多模态零样本学习方法,结合点云和图像的互补信息,提高了视觉与语义的对齐精度,并在车载激光点云数据集上取得了SOTA效果。3D-LLM[104]则通过将二维视觉语言模型作为三维骨干网络,并引入三维定位机制,显著增强了模型对三维空间信息的捕捉能力,在二维-三维-语言任务中表现出色。MixCon3D[105]设计了一个创新的对比学习框架,融合多模态数据,有效提升了模型对三维对象及其文本描述关系的理解能力,在实际应用场景中展现了显著优势。Text4Point[106]利用密集对比学习对齐图像与点云表示,并结合CLIP对图像和文本特征的有效对齐,使点云特征与文本嵌入实现迭代对齐。SeCG[107]通过跨模态图注意力增强的语义3D视觉定位,结合语言模型读取点云整合语义知识与语言模型,进一步提升了三维视觉定位与实例分割的性能,在多关系定位任务中取得了显著改进。

从点云智能理解范式的演进可以看出,尽管传统方法在简化机载航空点云任务和初步推动发展上发挥了一定作用,但在复杂场景的全面解析中表现乏力。语言不仅能够丰富三维空间的表达,还可以弥合视觉信息与复杂语义之间的差距,从而使模型在更广泛的应用场景中准确识别和理解对象。借助语言知识的可迁移性,模型还可进一步涌现出对未见地物类别与相关关系的泛化解译能力。因此,探索以语言为中心的点云理解方法,对于推动机载点云的深入研究和应用具有重要意义。

2 以语言为中心的航空点云理解范式主要研究问题和研究展望

以语言为中心的航空点云理解范式,是指直接利用自然语言作为先验知识或交互媒介,指导点云数据的解析与具体任务执行。该范式在三维信息解析任务中展现了巨大潜力,在某些应用场景(如自动驾驶、机器人导航、人机交互等)中的表现已引起广泛关注。然而,其在机载航空点云领域的挖掘仍有限,这一范式在推动该领域发展方面具有独特意义:(1)从数据层面看,机载航空点云作为国家基础建设的重要数据来源,规模大、覆盖广,但标注成本高昂。结合语言模型,可通过语义描述扩展数据理解,实现少样本甚至零样本学习,降低对大规模标注数据的依赖。(2)从理论与方法层面看,相较于传统点云理解方法,以语言为核心的模型能够灵活融合多模态信息,深度挖掘点云的内在语义特征,突破传统方法在复杂场景中的局限性。要实现以语言为核心驱动的航空点云智能理解框架,本文认为需重点围绕3个关键问题开展深入研究。

2.1 点云特征建模问题

以语言为中心的航空点云理解范式面临的首要挑战,是如何从点云数据中提取更具代表性的特征基元,以便在后续的处理环节中实现与语言信息的高效交互和融合。在自然语言处理中,Token是语言的最小单位,可以是单词、子词或字符[108]。在大语言模型(如GPT-4)中,Token是模型处理、理解和生成文本的核心构件。模型将文本分割成Token,以此为基础进行推理,并将生成的输出Token转化为可理解的语言。这一概念为点云理解提供了启发:在以语言为中心的机载点云理解中,如何有效建模三维数据特征,生成点云中最基本的处理单元,即点云Token。语言中的Token通常具备语义显著性和信息聚合性[109],即每个Token既能清晰表达特定意义,又能有效整合相似信息。对应地,点云Token应当具备能够代表三维空间的基本结构和语义单元的能力,同时需要考虑其在语义显著性、信息聚合能力等方面的特性。具体而言,点云Token的设计需回答以下核心问题:如何在机载航空点云中划分Token的范围?是否基于语义信息进行划分?如何确定合适的语义粒度,如砖块与建筑物在语义层次上的差异?是否采用固定的空间分辨率或特定几何特征进行划分?以及如何在

捕捉局部几何细节与全局结构信息之间实现平衡。针对这些问题的深入研究将为点云 Token 的设计提供理论支持，从而助力以语言为中心的点云智能理解范式的发展。

针对以上问题，本文提出从语义均一性和特征代表性两方面构建点云 Token。在语义均一性方面，虽然航空点云因飞行高度和视角变化可能呈现显著几何差异，但本质上，这些数据属于同一地物的多视图表达。因此，点云 Token 的构建需确保同一类别的 Token 在不同视角和采样条件下语义一致，无混淆，同时明确区分不同类别地物的语义边界，类似于语言中的词汇。现有基于欧氏距离的点云聚类方法虽在语义一致性上有所进展，但在航空点云中仍面临挑战。例如，城市区域中不同高度采集的建筑物点云可能因视角和密度差异，导致顶部与底部难以聚合，产生语义不一致问题。自然地形中，不同树种的点云在几何上相邻，其语义信息却截然不同，易造成聚类错误。因此，从语义信息聚合的角度优化点云 Token 的构建，规避传统物理距离聚类带来的语义混淆，是一个重要研究方向。在特征代表性方面，航空点云 Token 应具备高度凝练场景特征的能力，同时过滤冗余信息。航空点云通常包含大量冗余点，这些点对结构和语义信息贡献有限，却占用了大量计算资源。传统方法往往对所有点云全面处理，在大规模场景中，冗余数据严重拖累了计算效率。为此，可在语义均一性 Token 的基础上，通过相似度判断和剪枝机制，剔除冗余信息，仅保留核心代表性 Token。这种方法不仅能够显著降低数据处理量，还能将计算资源集中于关键特征，为以语言为中心的机载航空点云理解奠定高效可靠的基础。

2.2　跨模态信息映射问题

在点云特征建模中提取有效 Token 的基础上，以语言为中心的航空点云理解范式面临的下一个核心挑战，是如何实现点云与语言信息的跨模态映射，充分利用语言在描述开放世界方面的优势。机载点云是无序的三维数据，具有高度不规则性和复杂性，不同地物的空间分布和密度差异显著；而自然语言则通过简洁的词汇，以高度抽象的方式表达物体的外观、类别及相对空间关系。例如，位于街角高大的白色建筑在语言中能概括建筑特征，而点云数据则记录建筑的三维几何形状、表面特征及其空间位置。虽然“高大”可映射到点云的高度信息，但其具体数值取决于几何测量，而“白色”等颜色特征可能在点云中完全缺失。这种特征粒度的差异使点云与语言的对齐异常复杂。此外，航空点云因噪声、遮挡及采样密度不均等问题，进一步增加了跨模态对齐的难度。例如，自然环境下地形起伏和遮挡可能导致植被点云密度不均；城市场景中高楼或建筑物的遮挡可能引发地面点云数据缺失。这与语言描述假设对场景有完整感知的特点相矛盾，加剧了对齐的复杂性。因此，如何在非对齐的点云与语言数据间实现高效的信息映射，成为当前以语言为中心机载点云理解范式亟待解决的关键技术难题。

针对上述问题，本文提出一种可行思路：通过解耦低维数据与语义信息，在高层抽象语义层面实现对齐，从而缓解点云与语言在低层次细节上的差异对跨模态对齐的影响。核心方法是构建共享的高层语义空间，将点云和语言特征投射到统一表示中，以在抽象层面建立关联。文献[93]指出，利用预训练的共享知识空间可以有效对齐多模态特征。这启发人们思考一种可行的信息映射方法：利用大规模预训练模型生成高维嵌入空间，使点云和语言特征在该空间中实现一致的表示。语言描述被映射为高层语义特征，点云数据经过特征提取后投射到同一语义空间，从而在高层次实现对齐，增强跨模态对齐的鲁棒性和泛化能力。此外，在共享知识空间中利用多模态对比学习，通过拉近语义一致的点云与语言表征，同时推远语义不同的表征，进一步提升对齐效果。引入多尺度注意力机制可根据特征层次灵活调整对齐策略，重点处理高层语义特征，并简化对低层几何细节的处理，避免信息过载。同时，为应对航空点云中因采样不均或视角变化引发的噪声问题，设计自适应对齐机制尤为重要。该机制通过动态调节对齐层次和强度，能够根据数据质量和上下文灵活调整跨模态映射关系，即便在低质量点云条件下，也能依托高层语义信息实现稳定对齐。这种基于高层语义对齐的方法不仅有效缓解了低维数据与语义信息的不一致问题，还为多模态融合提供了统一的特征表示，为下游任务的高效实现奠定了基础。

2.3　下游任务适配问题

基于点云与语言信息的有效跨模态映射，获得以语言为中心的航空点云理解基础模型后，泛化适配下游任务成为将预训练模型的通用知识应用于特定任务的关键。然而，这一泛化过程通

常面临3个主要挑战:(1)模型预训练代理任务与下游任务之间存在不同,直接使用预训练特征通常难以实现最佳性能。(2)下游任务的数据分布与预训练数据往往存在较大差异,进一步削弱了模型的迁移能力。(3)许多特定任务的标注数据较为稀缺,难以支撑有效的微调训练,这可能使模型过度依赖少量标注数据,进而导致泛化能力下降,甚至在小规模任务中出现过拟合现象。

针对这些挑战,本文认为可行的解决方案是引入蒸馏学习[110]和提示学习[111]。蒸馏学习通过将预训练模型作为教师网络,利用其生成的高置信度伪标签指导学生网络学习,从而将教师网络中提取的通用特征传递给学生网络。这种方法能够适应不同任务的数据分布变化,尤其在数据不均衡的情况下,帮助模型维持泛化性能。而提示学习通过设计针对性的提示词,引导下游任务更好地适配预训练模型的特征空间。例如,建筑物检测任务可用建筑轮廓提示模型聚焦相关特征,植被分析任务则可用提示词强调高度与形状特征。提示学习不仅提升了模型对特定任务特征的敏感性,还增强了在数据稀缺条件下的适应能力。此外,这两种方法可结合应用。通过蒸馏学习利用语言生成的软标签约束模型学习,并通过用户提示指令激活语义相关的点云特征,实现多种类型知识的集成迁移,从而提高模型在多样化应用场景中的适应能力。这种方法为克服现有泛化适配问题提供了新的可能性,也为以语言为中心的航空点云智能理解在复杂场景中的应用提供了可能有效的解决路径。

3 结 语

三维时空信息在国家新型基础设施建设中发挥着关键作用,机载航空点云数据为大规模三维信息的获取提供了重要支撑。然而,如何高效解译和利用这些数据仍面临诸多挑战。本文系统回顾了点云智能理解范式的发展历程:从前深度学习时代的模型驱动与数据驱动范式,到后深度学习时代的自监督信号驱动范式,再到当前基于预训练基础大模型驱动的范式,揭示了点云理解从依赖传统算法与大数据输入,逐步转向自我学习与优化,并最终迈向模拟人类认知与语言处理的演进趋势。基于上述分析,本文提出以语言为中心的航空点云理解范式,通过语言对点云语义信息的深度建模与映射,显著提升点云语义解译能力。在大规模复杂场景中,该范式不仅扩展了点云理解的广度与深度,还赋予了点云数据丰富语义。针对这一新范式,本文探讨了三大核心挑战:点云特征建模、跨模态信息映射,以及下游任务适配问题,并提出了相应的解决思路和实施方案。

参 考 文 献

[1] 唐小川,涂子涵,任绪清,等. 一种识别植被覆盖滑坡的多模态深度神经网络模型[J]. 武汉大学学报(信息科学版),2024,49(9):1566-1573.
TANG Xiaochuan, TU Zihan, REN Xuqing, et al. A Multi-modal Deep Neural Network Model for Forested Landslide Detection[J]. *Geomatics and Information Science of Wuhan University*, 2024, 49(9): 1566-1573.

[2] 邓博,张会,柏君,等. 利用机载LiDAR的深圳市斜坡类地质灾害危险性评价[J]. 武汉大学学报(信息科学版),2024,49(8):1377-1391.
DENG Bo, ZHANG Hui, BAI Jun, et al. Hazard Evaluation of the Slope Based on Airborne LiDAR Data in Shenzhen, China[J]. *Geomatics and Information Science of Wuhan University*, 2024, 49(8): 1377-1391.

[3] 刘小莎,董秀军,钱济人,等. 高植被山区泥石流物源LiDAR遥感精细识别方法研究[J]. 武汉大学学报(信息科学版),2024,49(3):400-410.
LIU Xiaosha, DONG Xiujun, QIAN Jiren, et al. Airborne LiDAR-Based Debris Flow Material Sources Remote Sensing Recognition in Lush Mountainous Area[J]. *Geomatics and Information Science of Wuhan University*, 2024, 49(3): 400-410.

[4] 董秀军,邓博,袁飞云,等. 航空遥感在地质灾害领域的应用:现状与展望[J]. 武汉大学学报(信息科学版),2023,48(12):1897-1913.
DONG Xiujun, DENG Bo, YUAN Feiyun, et al. Application of Aerial Remote Sensing in Geological Hazards: Current Situation and Prospects[J]. *Geomatics and Information Science of Wuhan University*, 2023, 48(12): 1897-1913.

[5] LIAN R B, WANG W X, MUSTAFA N, et al. Road Extraction Methods in High-Resolution Remote Sensing Images: A Comprehensive Review[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2020, 13: 5489-5507.

[6] 詹杰,徐帆,时逸舟,等. 深度学习三维点云上采样算法综述[J]. 中国图象图形学报,2026,31(3):686-718.
ZHAN Jie, XU Fan, SHI Yizhou, et al. Review of

Deep Learning-Based 3D Point Cloud Upsampling Algorithms[J]. *Journal of Image and Graphics*, 2026, 31(3): 686-718.

[7] GUO Y L, WANG H Y, HU Q Y, et al. Deep Learning for 3D Point Clouds: A Survey[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 43(12): 4338-4364.

[8] 陈军, 刘建军, 田海波. 实景三维中国建设的基本定位与技术路径[J]. 武汉大学学报(信息科学版), 2022, 47(10): 1568-1575.
CHEN Jun, LIU Jianjun, TIAN Haibo. Basic Directions and Technological Path for Building 3D Realistic Geospatial Scene in China[J]. *Geomatics and Information Science of Wuhan University*, 2022, 47(10): 1568-1575.

[9] 龚健雅, 季顺平. 从摄影测量到计算机视觉[J]. 武汉大学学报(信息科学版), 2017, 42(11): 1518-1522.
GONG Jianya, JI Shunping. From Photogrammetry to Computer Vision[J]. *Geomatics and Information Science of Wuhan University*, 2017, 42(11): 1518-1522.

[10] 张立强, 李洋, 侯正阳, 等. 深度学习与遥感数据分析[J]. 武汉大学学报(信息科学版), 2020, 45(12): 1857-1864.
ZHANG Liqiang, LI Yang, HOU Zhengyang, et al. Deep Learning and Remote Sensing Data Analysis[J]. *Geomatics and Information Science of Wuhan University*, 2020, 45(12): 1857-1864.

[11] ARMENI I, SENER O, ZAMIR A R, et al. 3D Semantic Parsing of Large-Scale Indoor Spaces[C]//IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016.

[12] HACKEL T, SAVINOV N, LADICKY L, et al. Semantic3D. net: A New Large-Scale Point Cloud Classification Benchmark[PP/OL]. V1. arXiv (2017-04-12)[2024-10-19]. https://arxiv.org/abs/1704.03847.

[13] QIU B, ZHOU Y Z, DAI L, et al. WHU-Railway3D: A Diverse Dataset and Benchmark for Railway Point Cloud Semantic Segmentation[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2024, 25(12): 20900-20916.

[14] NIEMEYER J, ROTTENSTEINER F, SOERGEL U. Contextual Classification of LiDAR Data and Building Object Detection in Urban Areas[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2014, 87: 152-165.

[15] 杨必胜, 陈一平, 邹勤. 从大模型看测绘时空信息智能处理的机遇和挑战[J]. 武汉大学学报(信息科学版), 2023, 48(11): 1756-1768.
YANG Bisheng, CHEN Yiping, ZOU Qin. Opportunities and Challenges of Spatiotemporal Information Intelligent Processing of Surveying and Mapping in the Era of Large Models[J]. *Geomatics and Information Science of Wuhan University*, 2023, 48(11): 1756-1768.

[16] BAI Y T, GENG X Y, MANGALAM K, et al. Sequential Modeling Enables Scalable Learning for Large Vision Models[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024.

[17] ZHANG A S, LI S, WU J, et al. Exploring Semantic Information Extraction from Different Data Forms in 3D Point Cloud Semantic Segmentation[J]. *IEEE Access*, 2023, 11: 61929-61949.

[18] 丁博, 张立宝, 秦健, 等. 语义增强图像-文本预训练模型的零样本三维模型分类[J]. 电子与信息学报, 2024, 46(8): 3314-3323.
DING Bo, ZHANG Libao, QIN Jian, et al. Zero-Shot 3D Shape Classification Based on Semantic-Enhanced Language-Image Pre-Training Model[J]. *Journal of Electronics & Information Technology*, 2024, 46(8): 3314-3323.

[19] JOHNSON A E, HEBERT M. Surface Matching for Object Recognition in Complex Three-Dimensional Scenes[J]. *Image and Vision Computing*, 1998, 16(9/10): 635-651.

[20] RUSU R B, BLODOW N, BEETZ M. Fast Point Feature Histograms (FPFH) for 3D Registration[C]//IEEE International Conference on Robotics and Automation, Kobe, Japan, 2009.

[21] SALTI S, TOMBARI F, DI STEFANO L. SHOT: Unique Signatures of Histograms for Surface and Texture Description[J]. *Computer Vision and Image Understanding*, 2014, 125: 251-264.

[22] WOHLKINGER W, VINCZE M. Ensemble of Shape Functions for 3D Object Classification[C]//IEEE International Conference on Robotics and Biomimetics, Karon Beach, Thailand, 2012.

[23] RUSU R B, BRADSKI G, THIBAUX R, et al. Fast 3D Recognition and Pose Using the Viewpoint Feature Histogram[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, China, 2010.

[24] WEINMANN M, JUTZI B, HINZ S, et al. Semantic Point Cloud Interpretation Based on Optimal Neighborhoods, Relevant Features and Efficient Classifiers[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2015, 105: 286-304.

[25] GUO Y L, BENNAMOUN M, SOHEL F, et al. A Comprehensive Performance Evaluation of 3D Local Feature Descriptors[J]. *International Journal of Computer Vision*, 2016, 116(1): 66-89.

[26] YAN X, ZHENG C D, LI Z, et al. PointASNL: Robust Point Clouds Processing Using Nonlocal Neural Networks with Adaptive Sampling [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020.

[27] LANDRIEU L, RAGUET H, VALLET B, et al. A Structured Regularization Framework for Spatially Smoothing Semantic Labelings of 3D Point Clouds [J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2017, 132: 102-118.

[28] XU S, VOSSELMAN G, OUDE ELBERINK S. Multiple-Entity Based Classification of Airborne Laser Scanning Data in Urban Areas[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2014, 88: 1-15.

[29] HUANG R, HONG D F, XU Y S, et al. Multi-Scale Local Context Embedding for LiDAR Point Cloud Classification[J]. *IEEE Geoscience and Remote Sensing Letters*, 2020, 17(4): 721-725.

[30] KANG Z Z, YANG J T. A Probabilistic Graphical Model for the Classification of Mobile LiDAR Point Clouds[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018, 143: 108-123.

[31] YANG B S, DONG Z. A Shape-Based Segmentation Method for Mobile Laser Scanning Point Clouds [J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2013, 81: 19-30.

[32] VERDOJA F, THOMAS D, SUGIMOTO A. Fast 3D Point Cloud Segmentation Using Supervoxels with Geometry and Color for 3D Scene Understanding [C]//IEEE International Conference on Multimedia and Expo, Hong Kong, China, 2017.

[33] SU H, MAJI S, KALOGERAKIS E, et al. Multi-view Convolutional Neural Networks for 3D Shape Recognition[C]//IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015.

[34] QI C R, SU H, NIEßNER M, et al. Volumetric and Multi-view CNNs for Object Classification on 3D Data [C]//IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016.

[35] FENG Y F, ZHANG Z Z, ZHAO X B, et al. GVCNN: Group-View Convolutional Neural Networks for 3D Shape Recognition[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018.

[36] WANG C, PELILLO M, SIDDIQI K. Dominant Set Clustering and Pooling for Multi-view 3D Object Recognition[PP/OL]. V1. arXiv (2019-06-04) [2024-01-06]. https://arxiv.org/abs/1906.01592.

[37] HAMDI A, GIANCOLA S, GHANEM B. MVTN: Multi-view Transformation Network for 3D Shape Recognition[C]//IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2022.

[38] MATURANA D, SCHERER S. VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems, Hamburg, Germany, 2015.

[39] RIEGLER G, ULUSOY A O, GEIGER A. OctNet: Learning Deep 3D Representations at High Resolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017.

[40] KLOKOV R, LEMPITSKY V. Escape from Cells: Deep Kd-Networks for the Recognition of 3D Point Cloud Models[C]//IEEE International Conference on Computer Vision, Venice, Italy, 2017.

[41] WANG L, HUANG Y C, SHAN J, et al. MS-Net: Multi-scale Convolutional Network for Point Cloud Classification[J]. *Remote Sensing*, 2018, 10(4): 612.

[42] XU J Y, ZHANG R X, DOU J, et al. RPVNet: A Deep and Efficient Range-Point-Voxel Fusion Network for LiDAR Point Cloud Segmentation[C]// IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2022.

[43] ZHANG C, WAN H C, SHEN X Y, et al. PVT: Point-Voxel Transformer for Point Cloud Learning [J]. *International Journal of Intelligent Systems*, 2022, 37(12): 11985-12008.

[44] CHOY C, GWAK J, SAVARESE S. 4D Spatio-temporal ConvNets: Minkowski Convolutional Neural Networks[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2020.

[45] TANG H T, LIU Z J, ZHAO S Y, et al. Searching Efficient 3D Architectures with Sparse Point-Voxel Convolution[C]//Computer Vision-ECCV, Glasgow, UK, 2020.

[46] WU Z R, SONG S R, KHOSLA A, et al. 3D ShapeNets: A Deep Representation for Volumetric Shapes[C]//IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015.

[47] LI Y Y, PIRK S, SU H, et al. FPNN: Field Probing Neural Networks for 3D Data[PP/OL]. V3. arXiv

(2016-10-25)[2024-08-17]. https://arxiv.org/abs/1605.06240.

[48] MENG H Y, GAO L, LAI Y K, et al. VV-Net: Voxel VAE Net with Group Convolutions for Point Cloud Segmentation[C]//IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 2020.

[49] CHARLES R Q, HAO S, MO K C, et al. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017.

[50] QI C R, YI L, SU H, et al. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space[PP/OL]. V1. arXiv (2017-06-07)[2024-09-21]. https://arxiv.org/abs/1706.02413.

[51] HU Q Y, YANG B, XIE L H, et al. RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020.

[52] XU M T, ZHANG J H, ZHOU Z P, et al. Learning Geometry-Disentangled Representation for Complementary Understanding of 3D Object Point Cloud[C]//AAAI Conference on Artificial Intelligence, Online, 2021.

[53] QIAN G C, LI Y C, PENG H W, et al. PointNeXt: Revisiting PointNet++ with Improved Training and Scaling Strategies[C]//Conference on Neural Information Processing Systems, New Orleans, USA, 2022.

[54] MA X, QIN C, YOU H X, et al. Rethinking Network Design and Local Geometry in Point Cloud: A Simple Residual MLP Framework[PP/OL]. V2. arXiv (2022-11-29)[2024-11-14]. https://arxiv.org/abs/2202.07123.

[55] CHEN B J, XIA Y Z, ZANG Y, et al. Decoupled Local Aggregation for Point Cloud Learning[PP/OL]. V1. arXiv (2023-08-31)[2024-07-17]. https://arxiv.org/abs/2308.16532.

[56] WANG Y, SUN Y B, LIU Z W, et al. Dynamic Graph CNN for Learning on Point Clouds[J]. *ACM Transactions on Graphics*, 2019, 38(5): 1-12.

[57] ZHANG K G, HAO M, WANG J, et al. Linked Dynamic Graph CNN: Learning on Point Cloud via Linking Hierarchical Features[PP/OL]. V2. arXiv (2019-08-06)[2024-11-15]. https://arxiv.org/abs/1904.10014.

[58] LU Q, CHEN C, XIE W J, et al. PointNGCNN: Deep Convolutional Networks on 3D Point Clouds with Neighborhood Graph Filters[J]. *Computers & Graphics*, 2020, 86: 42-51.

[59] XU Q G, SUN X D, WU C Y, et al. Grid-GCN for Fast and Scalable Point Cloud Learning[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020.

[60] ATZMON M, MARON H, LIPMAN Y. Point Convolutional Neural Networks by Extension Operators[PP/OL]. V1. arXiv (2018-05-27)[2024-12-05]. https://arxiv.org/abs/1803.10091.

[61] LI Y Y, BU R, SUN M C, et al. PointCNN: Convolution on X-Transformed Points[C]//Neural Information Processing Systems, Montreal, Canada, 2018.

[62] LIU Y C, FAN B, XIANG S M, et al. Relation-Shape Convolutional Neural Network for Point Cloud Analysis[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2020.

[63] THOMAS H, QI C R, DESCHAUD J E, et al. KPConv: Flexible and Deformable Convolution for Point Clouds[C]//IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 2019.

[64] ZHAO H S, JIANG L, JIA J Y, et al. Point Transformer[C]//IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2022.

[65] GUO M H, CAI J X, LIU Z N, et al. PCT: Point Cloud Transformer[J]. *Computational Visual Media*, 2021, 7(2): 187-199.

[66] LAI X, LIU J H, JIANG L, et al. Stratified Transformer for 3D Point Cloud Segmentation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022.

[67] JIANG L, LAO Y X, LIU X H, et al. Point Transformer V2: Grouped Vector Attention and Partition-Based Pooling[C]//Conference on Neural Information Processing Systems, New Orleans, Louisiana, USA, 2022.

[68] WU X Y, JIANG L, WANG P S, et al. Point Transformer V3: Simpler, Faster, Stronger[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024.

[69] XIE S N, GU J T, GUO D M, et al. PointContrast: Unsupervised Pre-Training for 3D Point Cloud Understanding[C]//European Conference on Computer Vision, Glasgow, UK, 2020.

[70] YIN J B, ZHOU D F, ZHANG L J, et al. ProposalContrast: Unsupervised Pre-Training for LiDAR-

Based 3D Object Detection[C]//European Conference on Computer Vision, Tel Aviv, Israel, 2022.

[71] WANG C Y, JIANG L, WU X Y, et al. GroupContrast: Semantic-Aware Self-Supervised Representation Learning for 3D Understanding [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024.

[72] JIANG L, SHI S S, TIAN Z T, et al. Guided Point Contrastive Learning for Semi-Supervised Point Cloud Semantic Segmentation [C]//IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2022.

[73] DU B A, GAO X, HU W, et al. Self-Contrastive Learning with Hard Negative Sampling for Self-Supervised Point Cloud Learning[C]//ACM International Conference on Multimedia, Chengdu, China, 2021.

[74] AFHAM M, DISSANAYAKE I, DISSANAYAKE D, et al. CrossPoint: Self-Supervised Cross-Modal Contrastive Learning for 3D Point Cloud Understanding[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022.

[75] LIU M Z, ZHOU Q, ZHAO H S, et al. Prototype-Voxel Contrastive Learning for LiDAR Point Cloud Panoptic Segmentation[C]//International Conference on Robotics and Automation, Philadelphia, USA, 2022.

[76] YU X M, TANG L L, RAO Y M, et al. Point-BERT: Pre-Training 3D Point Cloud Transformers with Masked Point Modeling[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022.

[77] PANG Y T, WANG W X, TAY F E H, et al. Masked Autoencoders for Point Cloud Self-Supervised Learning[C]//European Conference on Computer Vision, Tel Aviv, Israel, 2022.

[78] FANG R Y, GAO P, GUO Z Y, et al. Point-M2AE: Multi-scale Masked Autoencoders for Hierarchical Point Cloud Pre-Training[C]//Conference on Neural Information Processing Systems, New Orleans, Louisiana, USA, 2022.

[79] REN B, MEI G F, PANI PAUDEL D, et al. Bringing Masked Autoencoders Explicit Contrastive Properties for Point Cloud Self-Supervised Learning [C]//Asian Conference on Computer Vision, Hanoi, Vietnam, 2024.

[80] HESS G, JAXING J, SVENSSON E, et al. Masked Autoencoder for Self-Supervised Pre-Training on LiDAR Point Clouds[C]//IEEE/CVF Winter Conference on Applications of Computer Vision Workshops , Waikoloa, USA, 2023.

[81] ZHANG Y B, LIN J H, LI R H, et al. Point-DAE: Denoising Autoencoders for Self-Supervised Point Cloud Learning[PP/OL]. V4. arXiv (2024-07-26) [2024-12-01]. https://arxiv. org/abs/2211.06841.

[82] LIU X H, LIU X C, LIU Y S, et al. SPU-Net: Self-Supervised Point Cloud Upsampling by Coarse-to-Fine Reconstruction with Self-Projection Optimization[J]. *IEEE Transactions on Image Processing*, 2022, 31: 4213-4226.

[83] OPENAI, ACHIAM J, ADLER S, et al. GPT-4 Technical Report[PP/OL]. V6. arXiv (2024-03-04) [2024-11-28]. https://arxiv. org/abs/2303.08774.

[84] ZHANG R R, WANG L H, QIAO Y, et al. Learning 3D Representations from 2D Pre-Trained Models via Image-to-Point Masked Autoencoders [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023.

[85] YAN S M, SONG C, KONG Y K, et al. Multi-view Representation Is What You Need for Point-Cloud Pre-Training [C]//International Conference on Learning Representations, Kigali, Rwanda, 2023.

[86] ZHOU J S, WANG J S, MA B R, et al. Uni3D: Exploring Unified 3D Representation at Scale[PP/OL]. V1. arXiv (2023-10-10) [2024-10-09]. https://arxiv. org/abs/2310.06773.

[87] KIRILLOV A, MINTUN E, RAVI N, et al. Segment Anything[C]//IEEE/CVF International Conference on Computer Vision, Paris, France, 2024.

[88] YANG Y H, WU X Y, HE T, et al. SAM3D: Segment Anything in 3D Scenes[PP/OL]. V1. arXiv (2023-06-06)[2024-11-21]. https://arxiv. org/abs/2306.03908.

[89] PENG X, CHEN R, QIAO F, et al. Learning to Adapt SAM for Segmenting Cross-domain Point Clouds[PP/OL]. V4. arXiv (2024-09-22)[2024-12-05]. https://arxiv. org/abs/2310.08820.

[90] XU M T, YIN X, QIU L T, et al. SAMPro3D: Locating SAM Prompts in 3D for Zero-Shot Instance Segmentation[PP/OL]. V2. arXiv (2025-02-04) [2025-02-21]. https://arxiv. org/abs/2311.17707.

[91] CEN J, CHEN K, CHEN R N, et al. Segment any Point Cloud Sequences by Distilling Vision Foundation Models[C]//Conference on Neural Information Processing Systems, New Orleans, Louisiana, USA, 2023.

[92] RADFORD A, KIM J W, HALLACY C, et al.

Learning Transferable Visual Models from Natural Language Supervision[PP/OL]. V1. arXiv (2021-02-26) [2025-02-01]. https://arxiv. org/abs/2103.00020.

[93] XUE L, GAO M F, XING C, et al. ULIP: Learning a Unified Representation of Language, Images, and Point Clouds for 3D Understanding [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023.

[94] XUE L, YU N, ZHANG S, et al. ULIP-2: Towards Scalable Multimodal Pre-Training for 3D Understanding [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024.

[95] ENGELMANN F, FEDELE E, POLLEFEYS M, et al. OpenMask3D: Open-Vocabulary 3D Instance Segmentation[C]//Conference on Neural Information Processing Systems, New Orleans, Louisiana, USA, 2023.

[96] PENG S Y, GENOVA K, JIANG C Y, et al. OpenScene: 3D Scene Understanding with Open Vocabularies[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023.

[97] NAJIBI M, JI J W, ZHOU Y, et al. Unsupervised 3D Perception with 2D Vision-Language Distillation for Autonomous Driving [C]//IEEE/CVF International Conference on Computer Vision, Paris, France, 2024.

[98] ZHANG R R, GUO Z Y, ZHANG W, et al. PointCLIP: Point Cloud Understanding by CLIP [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022.

[99] ZHANG Y, GONG K, KANG Y, et al. From Language to 3D Worlds: Adapting Language Models for Point Cloud Perception [EB/OL]. (2023-09-19) [2024-05-17]. https://openreview. net/forum? id=H49g8rRIiF.

[100]XU R S, WANG X L, WANG T, et al. PointLLM: Empowering Large Language Models to Understand Point Clouds[C]//European Conference on Computer Vision, Milan, Italy, 2024.

[101]ZHU Z Y, MA X J, CHEN Y X, et al. 3D-VisTA: Pre-Trained Transformer for 3D Vision and Text Alignment [C]//IEEE/CVF International Conference on Computer Vision, Paris, France, 2024.

[102]QI Z Y, FANG Y, SUN Z Y, et al. GPT4Point: A Unified Framework for Point-Language Understanding and Generation [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024.

[103]LU Y H, JIANG Q, CHEN R N, et al. See more and Know More: Zero-Shot Point Cloud Segmentation via Multi-modal Visual Data[C]//IEEE/CVF International Conference on Computer Vision, Paris, France, 2024.

[104]CHEN P H, CHEN Z F, DU Y L, et al. 3D-LLM: Injecting the 3D World into Large Language Models[C]//Conference on Neural Information Processing Systems, New Orleans, USA, 2023.

[105]GAO Y, WANG Z, ZHENG W S, et al. Sculpting Holistic 3D Representation in Contrastive Language-Image-3D Pre-training[PP/OL]. V2. arXiv (2024-04-05)[2024-12-05]. https://arxiv. org/abs/2311.01734.

[106]HUANG R, PAN X R, ZHENG H, et al. Joint Representation Learning for Text and 3D Point Cloud [J]. *Pattern Recognition*, 2024, 147: 110086.

[107]XIAO F, XU H B, WU Q X, et al. SeCG: Semantic-Enhanced 3D Visual Grounding via Cross-Modal Graph Attention [PP/OL]. V1. arXiv (2024-03-13) [2024-12-16]. https://arxiv. org/abs/2403.08182.

[108]COLLOBERT R, WESTON J, BOTTOU L, et al. Natural Language Processing (Almost) from Scratch[J]. *Journal of Machine Learning Research*, 2011, 12: 2493-2537.

[109]SHAO R, ZHANG Z Y, TAO C, et al. Homogeneous Tokenizer Matters: Homogeneous Visual Tokenizer for Remote Sensing Image Understanding [J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2024, 218: 294-310.

[110]HINTON G, VINYALS O, DEAN J. Distilling the Knowledge in a Neural Network[PP/OL]. V1. arXiv (2015-03-09)[2025-01-14]. https://arxiv. org/abs/1503.02531.

[111]LESTER B, AL-RFOU R, CONSTANT N. The Power of Scale for Parameter-Efficient Prompt Tuning [PP/OL]. V2. arXiv (2021-09-02)[2025-01-15]. https://arxiv. org/abs/2104.08691.