中图法分类号： 文献标识码： 文章编号：1006-8961(XXXX)XX-0001-21

论文引用格式：Zhang Ruonan, Geng Ying, Ma Kai, Zhou Aolin, Liu Libo. A survey on point cloud scene recognition: key challenges in feature representation, multi-modal association, and system adaptation[J/OL]. Journal of Image and Graphics, XXXX:1-21. DOI: 10.11834/jig.260212. (张若楠, 耿莹, 马凯, 周奥临, 刘立波. 点云场景识别综述:特征表达、多模态关联及系统适配的关键挑战[J/OL]. 中国图象图形学报, XXXX:1-21. DOI: 10.11834/jig.260212.)[DOI:10.11834/jig.260212]

点云场景识别综述:特征表达、多模态关联及系统适配的关键挑战

张若楠[1+],耿莹[2+],马凯[1+],周奥临[3+],刘立波[3*]

1. 宁夏大学信息与网络空间安全学院,宁夏中卫市 755000; 2. 宁夏大学旅游与健康学院,宁夏中卫市 755000; 3. 宁夏大学信息工程学院,宁夏银川市 750021

摘 要: 点云场景识别是三维感知领域的核心方向,本文聚焦于大规模三维点云检索驱动的地点识别与重定位任务,在自动驾驶、机器人导航等传统应用中发挥重要作用,并在低空经济等新兴场景展现出应用潜力。然而,受点云数据稀疏、无序、密度不均及低空复杂环境干扰影响,现有方法在真实动态场景中仍面临三大核心挑战:特征表达存在"语义盲区"、多模态关联存在"关系误区"、系统框架存在"适配限区"。本文从系统综述视角,将方法划分为特征表达、多模态关联及系统框架三类进行梳理,其中,特征表达包括点特征、体素特征与复值域特征表达,多模态关联涵盖点云内部多特征关联、外部输入源关联及多模态特征融合,系统框架分为单模态与多模态框架以适应不同应用场景。结合Oxford、KITTI、nuScenes等典型数据集,本文整理并对比了代表性方法的性能、适用条件及局限,揭示不同技术路线在识别精度、环境适应性、计算复杂度和部署可行性上的权衡。最后,总结研究瓶颈并展望鲁棒表征、多模态融合、通用化框架设计及统一评测体系等方向,为点云场景识别方法的系统理解与应用提供参考。

关键词: 点云场景识别;特征表达;多模态关联;复值域特征;系统框架

A survey on point cloud scene recognition: key challenges in feature representation, multi-modal association, and system adaptation

Zhang Ruonan[1+], Geng Ying[2+], Ma Kai[1+], Zhou Aolin[3+], Liu Libo[3*]

1. School of Information and Cybersecurity, Ningxia University, Zhongwei 755000, Ningxia Hui Autonomous Region, China; 2. School of Tourism and Health, Ningxia University, Zhongwei 755000, Ningxia Hui Autonomous Region, China; 3. School of Information Engineering, Ningxia University, Yinchuan 750021, Ningxia Hui Autonomous Region, China

Abstract: Point cloud scene recognition has emerged as a fundamental research pillar within the domain of three-dimensional (3D) perception. It plays an indispensable role in established autonomous systems, such as self-driving vehicles and robotic simultaneous localization and mapping (SLAM), while providing critical technical support for autonomous perception and scene understanding in increasingly complex low-altitude environments. With the rapid expansion of

收稿日期:2026-04-15;**修回日期:**2026-07-17

*通信作者:刘立波 liulib@163.com

基金项目:宁夏科技创新领军人才计划项目(No. 2022GKLRLX03);国家自然科学基金(No. 62506179, No. 62262053);宁夏自然科学基金(No. 2026AAC050038);银川市科技项目(No. 2025RC09)

Supported by: Ningxia Science and Technology Innovation Leading Talent Program Project (No. 2022GKLRLX03); National Natural Science Foundation of China (No. 62506179, No. 62262053); Natural Science Foundation of Ningxia (No. 2026AAC050038); Science and Technology Project of Yinchuan City (No. 2025RC09)

the "low-altitude economy," representative scenarios—including unmanned aerial vehicle (UAV) logistics, urban air traffic management, and precision infrastructure inspection—demand highly reliable spatial perception to ensure operational safety and intelligent decision-making under high-speed and variable-viewpoint conditions. Recent technological leaps in light detection and ranging (LiDAR), red-green-blue-depth (RGB-D) cameras, and four-dimensional (4D) millimeter-wave radar have significantly enhanced 3D data acquisition capabilities, pushing the field from traditional handcrafted geometric descriptor matching toward advanced data-driven deep learning paradigms. Despite this progress, achieving robust scene recognition in real-world dynamic environments remains an open and formidable challenge. Due to the intrinsic sparsity, irregularity, and uneven density of point clouds, coupled with the environmental interference inherent in low-altitude flight (such as motion blur and varying perspective scales), existing methods face three tightly coupled core limitations, defined in this review as the "semantic blind zone," the "relational misunderstanding," and the "adaptation limitation." Specifically, many real-valued embedding methods struggle to capture subtle geometric distinctions and latent semantic uncertainties, leading to a "semantic blind zone" where descriptors lose discriminative power under structural ambiguity. Furthermore, multi-modal association often falls into a "relational misunderstanding" by relying on shallow feature concatenation or simple alignment, which fails to exploit deep inter-modal synergies and cross-modal consistency. Finally, the "adaptation limitation" refers to the lack of scalability across heterogeneous sensors and varied computational platforms, particularly in resource-constrained low-altitude edge systems where the trade-off between latency and accuracy is critical. To provide a structured and comprehensive understanding of this research landscape, this review categorizes existing literature into three major dimensions: feature representation, multi-modal association, and system framework design. Regarding feature representation, we analyze point-based, voxel-based, and emerging complex-valued approaches. Point-based methods focus on local neighborhood modeling and fine-grained aggregation, while voxel-based methods utilize sparse convolutional operators to balance recognition accuracy and computational efficiency. Notably, complex-valued representation is explored as a promising path to model semantic uncertainty through amplitude-phase dual characteristics, providing a more cautious and mathematically robust way to handle noise compared to traditional real-valued frameworks. In terms of multi-modal association, the review distinguishes between internal feature association within point clouds (e. g. , global-local synergy), external input-source association (e. g. , incorporating RGB images, thermal data, or semantic priors), and explicit multi-modal feature fusion. We emphasize that non-linear interaction mechanisms, such as cross-attention transformers or interference-aware fusion modules, are essential for preserving semantic consistency across heterogeneous data sources. From a system framework perspective, we evaluate single-modal pipelines, which prioritize compact deployment and real-time response, against multi-modal frameworks that enhance robustness through informational redundancy at the cost of increased computational complexity. A detailed comparative analysis is conducted using representative public datasets, including the Oxford RobotCar, KITTI, and nuScenes benchmarks. These comparisons expose the inherent trade-offs between absolute retrieval accuracy, environmental adaptability (to weather and seasonal changes), and deployment feasibility. We observe that while voxel-based methods like MinkLoc3D demonstrate balanced performance across diverse urban datasets, point-based routes like PPT-Net often offer superior fine-grained details but may suffer from performance degradation when faced with significant density variations or occlusion. Furthermore, we introduce the evaluation of emerging low-altitude datasets such as UAVScenes and Pit30M, which highlight the specific challenges of drastic viewpoint changes, large-scale outdoor mapping, and cross-altitude matching that traditional ground-based methods fail to address adequately. Based on this comprehensive review, several critical research gaps and "bottlenecks" are identified: the sensitivity of descriptors to dynamic objects and environmental noise, the instability of fusion gains due to modal misalignment, and the lack of unified, standardized evaluation protocols for heterogeneous scenes. Looking forward, we propose that future research should prioritize several strategic directions: first, developing robust representation learning through self-supervised pre-training to reduce dependence on labeled data; second, exploring deeper and more adaptive multi-modal collaboration mechanisms to handle sensor failure or modal dropouts; third, investigating lightweight model compression and hardware-aware neural architecture search (NAS) for seamless integration into mobile low-altitude platforms; and fourth, establishing more generalized frameworks capable of zero-shot transfer across diverse sensing modalities. These advancements are particularly vital for the next generation of low-altitude applications, where safer and more intelligent

autonomous operations depend on the continued evolution of scene recognition technologies. In conclusion, this review aims to provide a challenge-oriented taxonomy and a strategic reference for researchers and engineers navigating the complexities of 3D point cloud perception, offering insights into building more resilient and adaptable autonomous systems.

Key words: point cloud scene recognition; feature representation; multi-modal association; complex-valued features; system framework

0 引　言

三维感知技术的快速发展为智能系统理解物理世界提供了重要支撑。作为三维环境感知中的关键任务，点云场景识别通过对点云数据进行表征、匹配与检索，实现环境场景的辨识与定位，已广泛应用于自动驾驶、机器人导航、无人系统感知等领域。与上述应用需求相对应，激光雷达、深度相机和四维毫米波雷达等传感器的推广应用，显著提升了点云数据获取能力，也推动场景识别研究从传统几何匹配向深度学习驱动的特征学习范式演进（Qi等，2017；Uy等，2018；Komorowski，2021；Gong等，2023）。深度学习方法能够从无序点云中提取丰富的几何和语义信息，为全局描述符构建、场景匹配与定位提供理论基础和工程可行性。

在特征表达方面，国际研究主要集中于点特征、体素特征和复值域特征三类方法。PointNetVLAD（Uy等，2018）首次将深度网络引入点云场景识别，开创深度点云检索先河。随后，PCAN（Zhang等，2019）引入注意力机制强化局部上下文关联，DAGC（Sun等，2020）结合动态图卷积与双向注意力提升拓扑表达能力，PPT-Net（Hui等，2021）采用金字塔Transformer实现多阶段语义编码。在体素特征方向，MinkLoc3D（Komorowski，2021）通过稀疏卷积实现轻量化实时推理，MinkUNeXt（Cabrera等，2025）通过参数规模扩展追求高精度上限。国内团队在复值域特征方面取得进展，ComPoint（Zhang等，2024）引入希尔伯特变换构建复值特征，利用虚部增强结构捕捉能力，PointWave（Li等，2023）以波函数隐喻驱动非线性交互，探索实值域表达局限突破的新范式。

多模态关联研究在国内外均取得显著成果。LPD-Net（Liu等，2019）通过并行融合手工几何特征与深度抽象特征验证特征互补性，MinkLoc++（Komorowski等，2021）采用后期融合策略整合点云与图像特征，QMMNet（Zhang等，2024）借鉴量子干涉概念设计非线性融合模块，实现点云与图像特征深度交互。

系统框架方面，单模态方法如PointNet++（Qi等，2017）和Point Transformer（Zhao等，2021）优化局部几何建模，多模态框架如CLIP2Point（Huang等，2023）尝试将视觉语言模型先验迁移至三维领域。国内研究也在轻量化特征建模和跨模态应用方面取得进展，SketchNet（Zhang等，2024）引入图像草图辅助提高跨场景鲁棒性，LFE-PointMamba（Zhou等，2025）实现全局序列特征高效提取。

尽管取得阶段性进展，现有方法在真实复杂动态场景中仍面临根本性局限。其一是点云场景特征层面的“语义盲区”，即在稀疏采样、结构重复或动态遮挡等复杂条件下，现有特征描述符的区分性不足，难以稳定捕捉场景中细微的几何差异与隐含的语义变化。其二是点云场景的多模态融合层面的“关系误区”，当前方法多基于浅层特征拼接，缺乏对跨模态信息多样性的有效建模，难以挖掘模态间的深层语义关联。其三是点云场景系统框架层面的“适配限区”，现有识别框架针对特定任务定制化设计，泛化与适配能力不足，难以在多样化场景和硬件条件下实现通用部署（Hao等，2024）。三重困境相互交织，共同构成当前领域亟待突破的瓶颈。

需要指出，上述三大挑战普遍存在于各类点云场景识别任务中，但在低空飞行平台与地面自动驾驶平台之间呈现不同的侧重。低空平台常以倾斜向下的视角观测场景，同一地物的点云形态随飞行高度和航向变化更为剧烈，对特征表达的跨视角不变性提出更高要求。

同时，低空平台在近地飞行与高空巡航之间点云密度波动显著，且载荷与能耗预算更为紧张，精度与效率的权衡更加苛刻。这些平台层面的差异在不同低空任务中进一步凝结为具体的技术约束。以城市低空物流配送为例，无人机需在建筑群间频繁起降并精确识别卸货点，视点高度与速度的剧烈变化导致点云密度快速切换，建筑物与广告牌造成局部

遮挡，要求特征表达具备较强的跨高度不变性与遮挡鲁棒性，系统框架则需支持在线的运动畸变补偿。电力与桥梁巡检任务聚焦于远距离细小目标的辨识，如杆塔绝缘子、销钉及桥梁裂缝等，此类目标在点云中仅占极少点数且受旋翼下洗气流噪声干扰，特征表达须强化细粒度几何保留与气流噪声抑制。应急救援与灾后评估场景环境高度非结构化，缺乏先验地图且常伴有烟尘雨雾，动态障碍物使场景持续变化，这对全天候感知能力、快速闭环检测及动态鲁棒性提出了极致要求。三类任务的约束各有侧重，但共同指向了特征表达的鲁棒性、多模态关联的容错性和系统框架的轻量化适配这三个核心维度，为后文方法讨论提供了具体的应用参照。

本文以系统综述为视角，围绕上述三大挑战，从特征表达、多模态关联、系统框架三个维度，对点云场景识别方法进行全面梳理与综合分析。通过对代表性方法、传感器模态及公开数据集的归纳比较，揭示不同技术路线在识别精度、计算复杂度与部署可行性之间的内在权衡，明确当前研究关键局限，并展望通用化框架及统一评测体系等未来方向，为该领域的理论深化和工程应用提供系统参考。

1 点云场景识别方法分类

在进入分类梳理之前，需明晰本文所涉核心任务的边界，并对其进行形式化描述。本文综述围绕点云地点识别展开，其核心目标是从三维点云中提取具有区分性的全局描述符，通过描述符匹配判断观测场景是否与数据库中某一已知场景属于同一地点。设数据库中含有N个已知场景的点云集合$M=\{P_1, P_2, \cdots, P_N\}$，给定查询点云$Q$，其数学形式为检索使得描述符距离最小化的候选点云，即：

$$P^* = arg\ min\ d\left(F\left(Q\right),F\left(P\right)\right)\#(1)$$

式中$F(\ )$为特征提取函数，将原始点云映射为紧凑的全局描述符，$d(\ ,\ )$为欧氏距离度量。场景检索侧重描述符构建完成后的高效索引与匹配，是实现地点识别的技术手段。重定位则是地点识别在同步定位与建图系统中的延伸应用，要求在已知地图中恢复精确的六自由度位姿。不同技术路线的差异主要体现在$F(.\ )$的具体设计。单模态方法将特征提取建模为从点云空间到描述符空间的映射：

$$d = F_{point}(P),P \in R^{N\times 3}\#(2)$$

式中$d\in R\hat{}D$为D维全局描述符，典型实现包括PointNetVLAD的PointNet加神经网络局部聚合描述符向量（network vector of locally aggregated descriptors，NetVLAD）聚合、MinkLoc3D的稀疏卷积加广义均值池化（generalized mean pooling，GeM）以及ComPoint的希尔伯特变换加复数卷积。多模态方法将映射扩展为描述符由点云与辅助模态共同决定的形式：

$$d = F_{fusion}(P,I)\#(3)$$

式中I表示图像、语义标签或草图等辅助模态，融合函数$F_fusion(\ )$的设计经历了从特征拼接到注意力交互再到非线性耦合的演进。

上述任务均围绕位置判定展开，而场景分类属于不同的问题范畴，目标是将场景划入预定义的语义类别，本文仅在讨论预训练范式或跨任务泛化时偶有涉及。场景理解作为一个层次更高的上位概念，涵盖语义分割、目标检测、场景分类等多类任务，本文在低空应用语境中将其作为地点识别技术的应用牵引加以讨论。后续各章涉及上述概念时，将严格保持此处界定的含义。

梳理现有文献可知，当前研究焦点与进展可围绕三个核心维度归纳：(1)场景点云特征表达、(2)场景多模态信息关联、(3)场景识别系统框架。这三个维度分别对应着前文所述的三大核心挑战。由于现有代表性方法和公开测评基准主要来源于自动驾驶、机器人导航及室内定位等研究场景，而低空飞行平台相关研究与标准化数据积累仍相对有限，本文以通用点云场景识别文献为主体，对代表性方法进行分类总结与评述，整体结构如图1所示。

1.1 场景点云特征表达相关方法

点云特征表达是点云场景识别的技术基石，核心目标是将原始、无序的三维点云转换为紧凑、信息丰富且具有区分性的特征表示，为后续场景匹配与识别提供基础。现有方法面临的核心困难是“语义盲区”，即实值域特征表达难以建模场景中几何结构歧义性与观测变化带来的匹配不确定性，导致在结构相似或外观剧变环境下特征描述产生本质混淆。根据数据表示形式和特征提取范式的差异，现有方法主要分为点特征表达、体素特征表达、复值域特征表达三类。点特征与体素特征方法虽在刻画精度与计算效率上各有侧重，但均受限于经典欧氏空间与

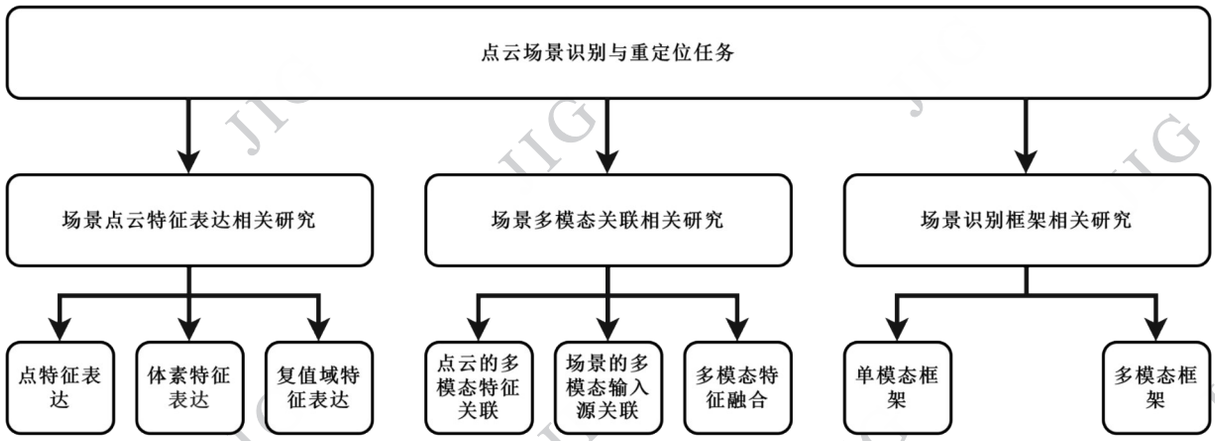

图1　点云地点识别与重定位方法分类体系

Fig. 1　Taxonomy of Point Cloud Place Recognition and Relocalization Methods

实值域框架，难以建模场景中普遍存在的语义不确定性，从而引发特征表达的“语义盲区”。复值域特征方法借助希尔伯特空间与复数表示，突破实值域在语义建模上的局限，这类方法不仅在理论上提出了新的特征表达范式，也已在实际点云场景识别任务中得到验证，显示出在处理语义不确定性和复杂环境下的有效性与适应性。三类方法在精度、效率与语义建模潜力上各有侧重。

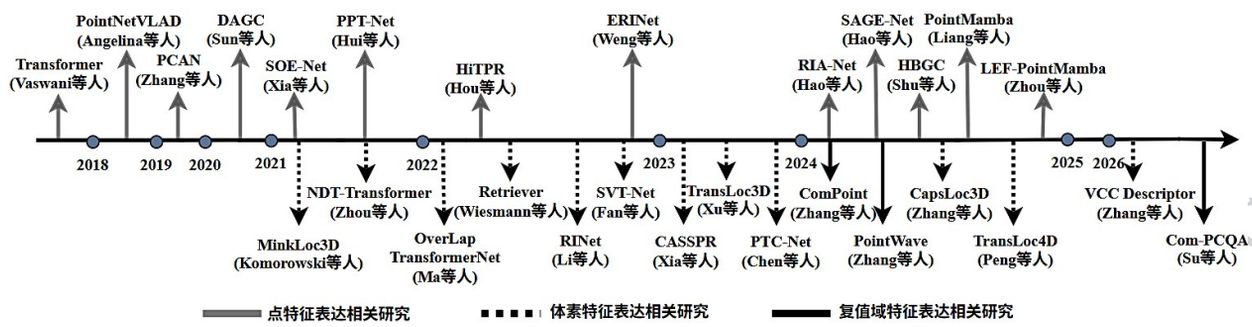

图2　点云特征表达相关研究时间轴

Fig. 2　Timeline of Research on Point Cloud Feature Representation

1.1.1　点特征表达

点特征表达以单个空间点为处理粒度，是早期经典技术思路。PointNetVLAD作为奠基性工作，首次将深度神经网络引入点云场景识别，其网络结构如图3所示。该方法以PointNet(Qi等，2017)为骨干网络，通过对称函数处理无序点云，并利用NetVLAD(Arandjelovic等，2016)，将局部特征软聚类为紧凑的全局描述符。其局限在于PointNet的感受野有限，难以捕捉远距离依赖。

后续研究围绕局部刻画与全局关联持续优化，PCAN引入PointNet++与注意力机制强化局部上下文关联；DAGC结合DGCNN(Wang等，2019)与双向注意力提升拓扑结构表达；SOE-Net(Xia等，2021)提出Shift操作子实现多尺度特征融合。PPT-Net代表了点特征路线的前沿水平，其创新在于金字塔Transformer(Vaswani等，2017)架构与分组自注意力机制，通过多阶段特征编码递进式挖掘深层语义，在精度上达到该路线当前最优。此外，ERINet(Weng等，2022)、RIA-Net(Hao等，2024)引入旋转不变性编码模块提升姿态变化适应性；SAGE-Net(Hao等，2024)、HBGC(Shu等，2023)融合注意力与卷积技术优化表达效果。上述方法主要基于实值域框架构建特征表示，注意力机制与多尺度编码策略在一定程度上增强了局部几何的判别力，但在结构高度相似或观测条件剧烈变化时，特征的区分性仍可能出现退化。

近年来，状态空间模型为点特征表达提供了新的架构选择。PointMamba（Liang 等，2024）首次将 Mamba 模型引入点云分析，利用空间填充曲线实现点云序列化，并采用纯 Mamba 编码器提取全局特征。该方法以线性计算复杂度实现全局建模，在物体分类任务上以更少参数量取得与 Transformer 相当的结果。目前 PointMamba 在点云场景识别任务的研究尚处于早期探索阶段，但其高效的序列化全局特征提取机制为后续地点识别研究提供了新的技术发展方向。在此基础上，LFE-PointMamba 针对局部几何捕捉不足，以及单向建模与点云无序性之间的矛盾进行改进，实现了高效的全局建模与空间拓扑建模，进一步验证了状态空间模型在点云分析中的可扩展性。

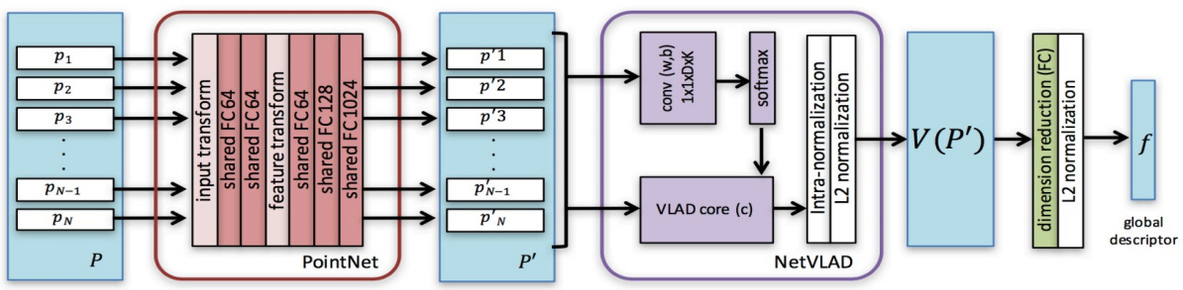

图3　PointNetVLAD 网络结构图（Uy 等，2018）

Fig. 3　Network Architecture of PointNetVLAD (Uy et al. , 2018)

1.1.2　体素特征表达

体素特征表达将无序点云映射至三维规则网格，并借助稀疏卷积平衡效率与精度。MinkLoc3D 奠定了体素特征路线的轻量化范式，在保持高计算效率的同时生成紧凑的全局描述符，其网络结构如图 4 所示。该方法以低参数量实现实时推理，但体素化过程中的离散化损失可能削弱几何细节刻画能力。

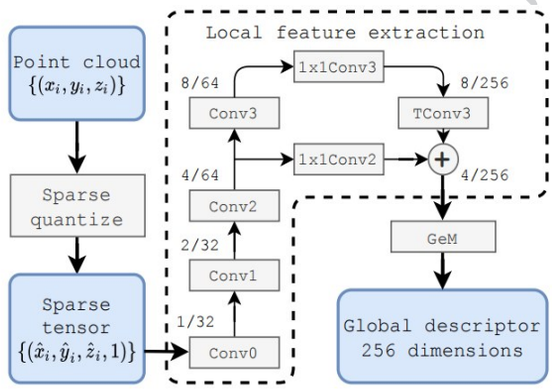

图4　MinkLoc3D 网络结构图（Komorowski，2021）

Fig. 4　Network Architecture of MinkLoc3D (Komorowski , 2021)

以此为基础，后续研究呈现出不同的技术演进取向。SVT-Net（Fan 等，2022）通过优化稀疏算子实现超轻量级部署，而 MinkUNeXt 则通过大幅提升参数规模实现高精度建模。与此同时，基于正态分布变换（normal distributions transform，NDT）的 Transformer 模型 NDT-Transformer（Zhou 等，2021）与 OverLapTransformer（Ma 等，2022）等工作引入 Transformer 架构，旨在利用自注意力机制强化多尺度特征的全局关联能力。此外，TransLoc4D（Peng 等，2024）采用体素化稀疏卷积处理 4D 雷达多模态数据，拓展了体素框架的应用边界。

在手工特征与体素表达融合方面，VCC Descriptor（Li 等，2026）提出了一种新型局部描述符。该方法从体素化点云中提取垂直特征并编码为圆弧直方图，在确保旋转不变性的同时显著提升了特征表示的紧凑性与提取效率，能够在毫秒级内完成回环检索，为体素特征表达在实时同步定位与建图（simultaneous localization and mapping，SLAM）系统中的应用提供了有效补充。体素方法通过上述路径的持续优化，已形成覆盖轻量部署到高精度建模的完整技术谱系。但离散化造成的信息损失与实值域表达局限仍是“语义盲区”的根本制约。

1.1.3　复值域特征表达

实值域框架将离散几何点映射为确定性的欧氏空间向量，在结构重复或几何歧义显著时，单纯依赖坐标值和邻域聚合的方式存在一定局限。近期研究转向希尔伯特空间的复值域表达，复数域的幅度-相位双重特性为语义建模提供新路径。

PointWave 提出波函数隐喻，将点特征建模为波

函数，即用一个复数来表达。核心思想是利用相位差动态调制特征交互，如图5所示，两个复值特征的叠加结果受其相位差调控，同相位增强、反相位削弱，实现更精细的非线性关联建模。

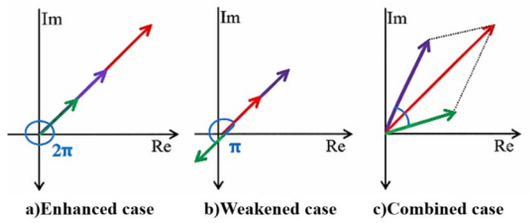

图5　基于波函数叠加的点特征交互原理(Li等,2023)

Fig. 5　Principle of Point Feature Interaction Based on Wave Function Superposition (Li et al. , 2023)

ComPoint作为前沿方法，创新引入希尔伯特变换构建初始复值表示，该转换过程及其效果直观展示于图6，虚部提供互补几何视角，再通过复数卷积块ComplexPointConv完成复值域特征提取与非线性融合，最后经软平衡聚合模块生成全局复值描述符，推进了基于复值域的点云结构捕捉与语义建模研究(Lan等,2026)。

上述两类方法通过复值域重构特征表达，为缓解“语义盲区”提供了一种具有启发性的路径。虽然目前仍面临复值算子优化程度不足、大规模基准验证有待丰富等挑战。但是实验数据表明，复值域方法在显著降低参数量与推理耗时的同时，依然保持了具有竞争力的识别精度。其展现出的轻量化表征能力与较高的推理时效，为点云特征表示的多维探索提供了有益参考。

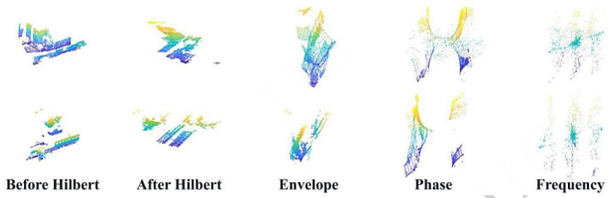

图6　希尔伯特变换构建点云复值表示示意图(Zhang等,2024)

Fig. 6　Schematic of Complex-Valued Point Cloud Representation via Hilbert Transform (Zhang et al. , 2024)

值得关注的是，复值表示在三维点云领域的探索已初步突破场景识别任务的边界。Com-PCQA(Su等,2026)将复值特征学习引入点云质量评估，借助希尔伯特变换构建复值解析信号，利用幅度与相位解耦的注意力机制预测点云感知质量。该工作表明复数域的幅度-相位双重特性在不同三维感知任务中均具备理论潜力，为复值特征表达方法从场景识别向更广泛的三维视觉任务推广提供了佐证。

1.2　场景多模态关联相关方法

多模态关联是弥补点云单一几何模态表达局限的关键技术方向，核心思想是利用多源异构信息的深度融合，强化场景语义与几何细节的互补描述，进而缓解“关系误区”带来的多模态关联不可靠问题。现有方法因缺乏对跨模态信息多样性的深度建模，多停留在浅层交互层面，导致几何信息与辅助模态的语义难以形成深度互补，在光照变化、传感器噪声或模态缺失等条件下识别性能显著下降且鲁棒性不足。依据多模态信息的来源差异，相关研究主要分为点云的多模态特征关联、场景的多模态输入源关联及多模态特征融合三类。

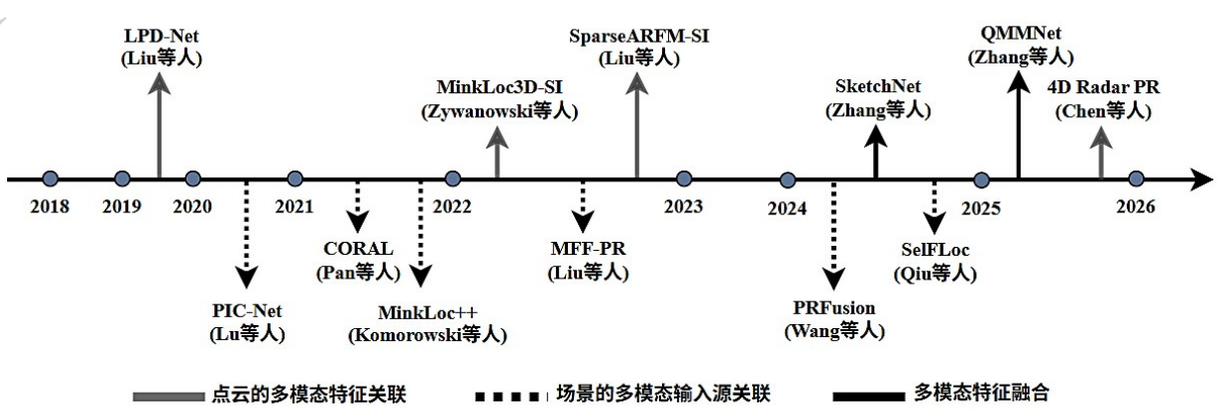

图7　多模态关联相关研究时间轴

Fig. 7　Timeline of Research on Multi-Modal Association

1.2.1　点云的多模态特征关联

聚焦单源点云数据内部的特征互补，核心是结合传统手工几何算子与深度学习抽象特征，兼顾特征可解释性与区分度。LPD-Net作为经典方案，其网络结构如图8所示，通过并联融合手工构造的几何特征与深度学习抽象特征优化场景识别性能，揭示了手工特征在局部结构刻画中的不可替代性，但融合策略停留在特征拼接层面，未能充分挖掘模态间的互补性。在传统的3D点云处理中，单源点云特征关联主要依赖于空间几何拓扑关系的挖掘。MinkLoc3D-SI（Zywanowski等，2021）针对点云空间密度异质性问题，利用球面坐标对空间分区，以分区点云密度为依据调整特征关联权重，提升了语义集中区域的特征优先级。SparseARFM-SI（Liu等，2022）作为前沿方法，在其基础上引入层级架构与自注意力机制，通过多分辨率特征交互强化模态关联性。

随着传感器技术从3D激光雷达（light detection and ranging，LiDAR）发展到4D radar，点云数据不仅包含三维空间坐标，还集成了瞬时多普勒速度与雷达散射截面（radar cross section，RCS）等物理属性，为特征关联提供了新的维度。4D radar PR（Chen等，2025）方法将RCS与空间坐标联合编码，在恶劣天气下提升了点云特征的鲁棒性与场景辨识能力。此外，TDFANet（Lu等，2025）面向4D雷达点云的稀疏性和动态环境挑战，结合时序多帧输入与速度信息，利用状态空间建模实现跨帧特征交互，从而增强单源点云内部多模态特征关联，提升动态环境下的识别鲁棒性。然而，上述方法均局限于单源数据内部融合，未引入外部语义信息，对“关系误区”的解决能力有限。

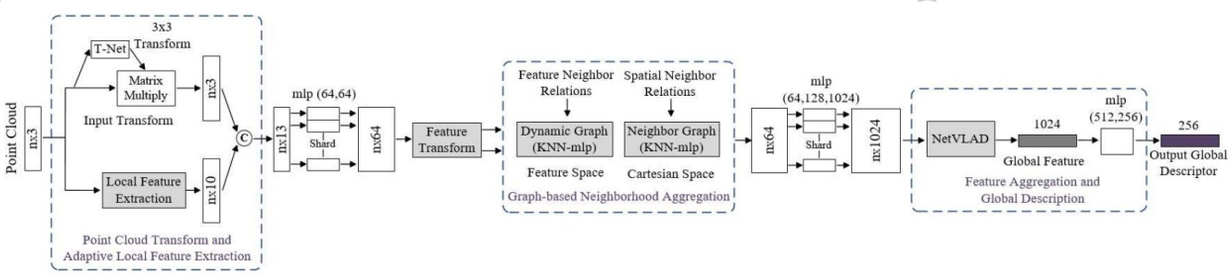

图8　LPD-Net网络结构图（Liu等，2019）

Fig. 8　Network Architecture of LPD-Net (Liu et al. , 2019)

1.2.2　场景的多模态输入源关联

该类方法通过引入图像、语义标签等外部模态，补充点云语义或几何细节。PIC-Net（Lu等，2020）与Coral（Pan等，2021）以图像为主，性能易受环境影响。MinkLoc++的输入包括点云和图像两种模态。在多模态输入处理中，它通过多损失函数训练平衡两类模态的贡献，从而在保持语义互补性的同时避免单一模态主导识别过程，后续特征融合步骤进一步提升整体表示能力。PRFusion（Wang等，2024）针对点云与图像模态在语义和几何表达上的异构性与互补性设计，通过注意力驱动的全局融合模块和滑动窗口式局部融合模块，分别捕捉场景级语义关联与局部几何细节匹配，从而有效提升了多模态关联鲁棒性，但仍未完全解决模态异构性带来的适配难题。此外，MFF-PR（Liu等，2022）与SelFloc（Qiu等，2024）虽提出通用特征融合逻辑，但未针对点云稀疏性、无序性进行适配设计，实际应用效果受限。

1.2.3　多模态特征融合

多模态特征融合策略直接决定了模态间信息交互的深度与有效性。早期工作如MinkLoc++分别采用独立分支来提取点云和图像的特征，再通过特征拼接或分值融合实现后期交互，虽简洁有效，但线性融合策略难以挖掘模态间的深层语义关联。

为突破线性融合的局限，近年来研究者引入非线性融合机制。QMMNet借鉴量子干涉概念设计非线性融合模块，其网络结构如图9所示，通过干涉操作实现点云与图像特征的深度交互，有效弥补点云语义表达的不足。SketchNet则创新性地采用图像草图作为辅助模态，利用草图拓扑结构受姿态扰动影响小的特点，通过交互式融合强化点云拓扑特征的鲁棒性，在跨场景识别中取得稳定性能增益。两类方法初步验证了非线性融合的可行性，但辅助模态质量波动与模态异构性仍是制约其泛化能力的关键因素，“关系误区”的根源问题仍未得到完全解决。

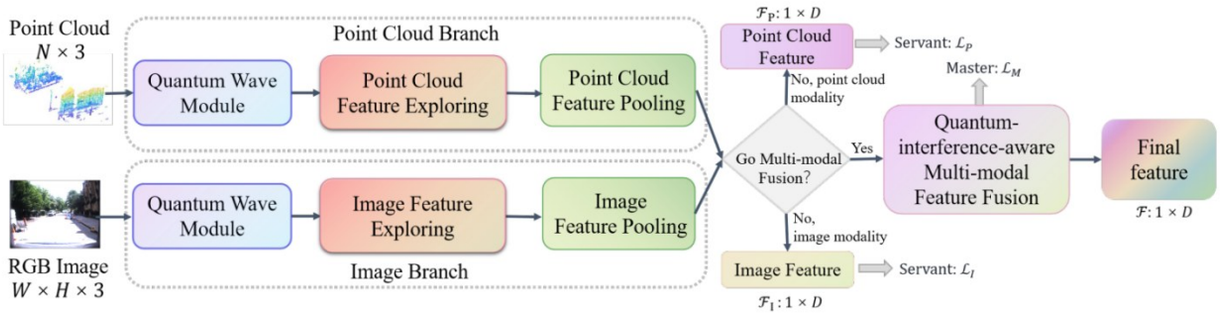

图9　QMMNet网络结构图(Zhang等,2024)

Fig. 9　Network Architecture of QMMNet (Zhang et al. , 2024)

1.3　场景识别框架相关方法

场景识别框架的设计需结合点云特征表达与多模态关联的技术思路,依据框架支持的输入模态种类,相关研究可划分为单模态框架与多模态框架两类,二者分别适配不同的场景识别需求。

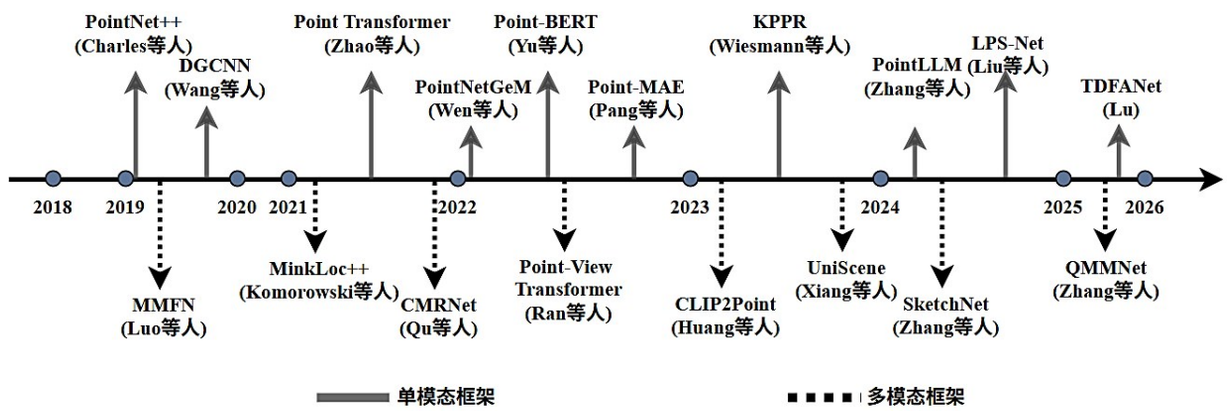

图10　场景识别框架相关研究时间轴

Fig. 10　Timeline of Research on Scene Recognition Frameworks

1.3.1　单模态框架

单模态框架以点云为唯一输入模态。PointNet-VLAD是深度学习范式下单模态框架的奠基性工作,构建了基于端到端深度学习的全局描述符提取范式。为后续由几何空间映射向深度语义表征的研究奠定了数学与工程基础。后续PointNet++、DGCNN、Point Transformer等模型通过层次化特征学习、图卷积与自注意力机制,提取局部几何特征并建模长程语义依赖等架构优化方法,增强复杂场景表征能力。对于单模态框架,4D雷达点云由于点云极度稀疏,其特征建模难度大,因此可采用专门的多帧/时序输入策略来增强场景表征。以TransLoc4D为代表的方法,针对4D雷达点云的极稀疏性与动态环境,采用多帧时序输入策略,在单模态框架下实现鲁棒的场景识别。预训练范式中,Point-BERT(Yu等,2022)、Point-UMAE(Pang等,2022)通过掩码建模在大规模无标签数据上预训练,学习通用场景先验以提升下游任务性能与泛化能力。前沿探索方面,PointLLM(Xu等,2024)将点云编码器与大语言模型对齐,探索开放场景理解与推理。

除了上述提升表征能力的核心方法外,另一部分研究的重心集中于框架的效率与轻量化优化。PointNetGeM(Wen等,2022)采用几何池化简化特征编码;LPS-Net(Liu等,2024)引入参数共享降低计算复杂度;KPPR(Wiesmann等,2022)结合预训练编码器与特征数据库加速训练与检索。此类设计通过压缩计算开销适配资源受限场景。

1.3.2　多模态框架

多模态框架支持点云与图像、草图等多模态数据输入,其核心思想是通过整体架构设计实现有效

的信息融合与协同。依据融合策略的发展,可划分为中期融合与后期融合两类。

中期融合策略在特征编码过程中实现模态间的深度交互。早期工作如MMFN(Li等,2019)确立了“双流编码+特征融合”的基线范式,分别提取图像与点云特征后,在特征层进行融合。该类方法有利于模态间的深层特征协同,但对传感器标定与时空对齐精度要求较高。

后期融合策略则先由各模态独立生成全局描述符,再在决策层进行融合,更具灵活性,在传感器异构或标定不精确的场景下表现出更强的鲁棒性。MinkLoc++采用分值融合方式,分别计算点云和图像描述符的相似度得分后进行加权求和,避免了特征层面的异构对齐难题。而QMMNet、SketchNet则在后期融合框架下引入非线性交互机制(参见1.2.3),进一步强化模态间的深层关联。后期融合方式既保留了各模态的独特信息,又能通过融合策略强化特征的互补性。

前沿探索中,CLIP2Point创新性地将视觉-语言大模型的先验知识迁移至3D领域,实现了开放词汇下的零样本场景识别;UniScene(Li等,2025)则通过场景分类与检索的多任务对比学习,旨在获得更具泛化性的通用场景表征。这些工作将多模态框架的边界从传统传感器融合拓展至语义知识与通用表征层面。

2 点云场景识别中的传感器模态

点云场景识别的性能不仅受算法结构影响,还在很大程度上取决于传感器的物理成像机制与观测条件。不同传感器在视场范围、点云密度、测量噪声、环境鲁棒性及附加属性等方面存在显著差异,这些差异不仅决定了场景可观测信息的完备性,也直接塑造了后续特征表达、多模态关联与系统框架的设计取向。因此,从传感器模态角度分析点云场景识别问题,是理解各类方法适用边界与性能差异的重要前提。

依据视场完备性与应用场景,主流点云传感器可分为激光雷达、4D毫米波雷达和红绿蓝深度(red green blue-depth, RGB-D)相机三类。图11示意了典型点云传感器的视场范围与点云形态。其中分别展示了旋转式LiDAR的360°全向扫描、固态LiDAR的有限扇形视场、4D雷达的稀疏点云及其极坐标表示以及RGB-D相机的彩色图像与深度图像的对齐关系。表1综合对比了各类传感器的物理特性、环境鲁棒性及其在低空任务中的适配策略。所示差异意味着,不同传感器对应的识别方法难以共享完全一致的建模策略。例如,全向但远距稀疏的旋转式LiDAR需侧重全局几何语义的挖掘,视场角(field of view, FoV)受限的固态LiDAR更依赖于时序补偿与子图构建,4D雷达则通常需要时序建模和跨模态辅助以弥补单帧区分性不足,而RGB-D更适合作为室内场景中的语义增强来源。

2.1 激光雷达

激光雷达通过发射激光束测量距离,是目前室外点云场景识别最常用的传感器。根据扫描机制,可进一步分为旋转式与固态/非重复扫描式两类。

旋转式激光雷达例如Velodyne VLP系列提供360°水平视场,点云呈环形分布,测量距离可达100至200米。其全向特性使得机器人在不同航向下对同一地点的观测具有天然的一致性,为设计视角不变的全局描述符如球面投影或柱面投影提供了理想

表1 典型传感器物理特性及其低空环境适配策略对比

Table 1 Physical Characteristics and Low-altitude Adaptation Strategies of Typical Sensors: A Comparative Analysis

传感器类型	视场角	点云密度	环境鲁棒性	优点	缺点	低空环境挑战	建议适配策略
旋转式LiDAR	360°全向	稀疏	不受光照及季节变化影响	视角不变性潜力大	远距离点云稀疏	远距稀疏退化,全局表征困难	强化视角不变性,挖掘全局几何语义
固态LiDAR	通常小于120°	高密	不受光照及季节变化影响	局部细节丰富	视角敏感,需多帧累积	FoV局限,高频姿态动态易失锁	异构时序联合补偿,滑动窗口子图构建
4D毫米波雷达	360°全向	极稀疏,噪声簇密集	全天候,抗雨雾雪	恶劣天气下稳定	单帧点云区分性差	噪声簇聚集,单帧表征区分度弱	时序能量流建模,径向速度特征关联
RGB-D相机	有限,约70°	稠密深度图	受强光及透明材质影响	室内语义丰富	室外性能显著下降	环境光敏感,有效量程受限	语义先验增强,辅助LiDAR精细化标注

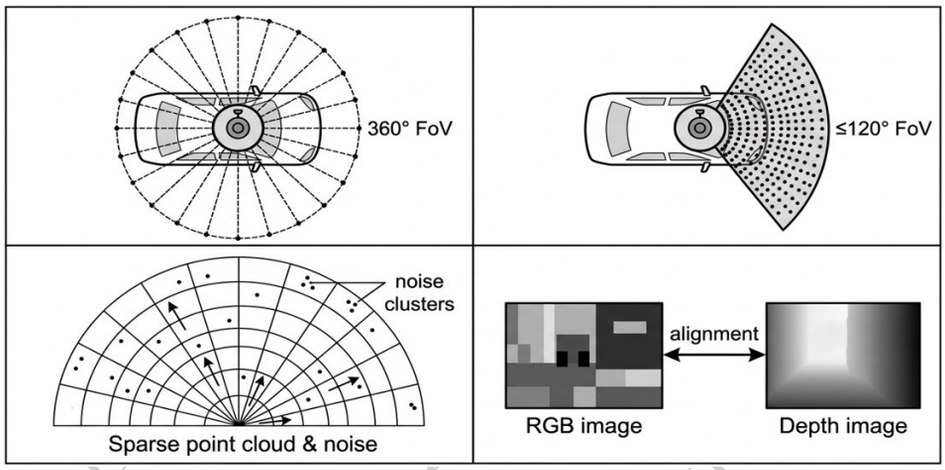

图11　典型点云传感器的视场范围与数据形态示意图

Fig. 11　Illustration of Field of View and Data Patterns of Typical Point Cloud Sensors

基础。然而，旋转式LiDAR的点云随距离增加急剧稀疏，远距离物体的几何结构不完整，容易导致误匹配。此外，旋转部件带来的机械磨损和成本也限制了其在某些轻量化平台上的应用。

固态/非重复扫描式激光雷达如Livox Mid系列采用棱镜或微机电系统（micro-electro-mechanical system，MEMS）振镜扫描，视场角通常小于120°，但点云密度显著高于旋转式，且无机械旋转部件，成本更低、寿命更长。高密度点云能够保留更细致的局部几何特征，有利于区分结构相似但细节不同的场景。其核心短板在于有限视场角，机器人从相反方向经过同一地点时，观测到的点云可能仅有少量重叠，导致识别失败。为缓解这一问题，现有方法常通过累积多帧点云形成局部子图，或与视觉及惯性导航融合来估计相对朝向，从而间接扩大有效观测范围。

2.2　4D毫米波雷达

4D毫米波雷达在传统距离、方位和多普勒速度之外增加了俯仰角信息，形成稀疏的四维点云。其最大优势在于对雨、雾、雪等恶劣天气以及光照变化的高鲁棒性，同时可测量目标的径向速度，为区分静态结构与动态物体提供补充信息。

然而，4D雷达点云具有极低的角分辨率，点云数量通常仅为激光雷达的十分之一至百分之一，且包含大量噪声簇和虚警。场景识别算法面临严峻挑战，单帧点云的区分性严重不足，易产生感知混淆；噪声簇会污染全局描述符，降低匹配精度。现有解决方案主要沿两个方向展开，一是引入时序信息，利用连续多帧点云构建序列描述符，通过运动一致性过滤噪声；二是跨模态辅助，将雷达点云与预建的高精度激光雷达地图或卫星图像进行匹配，借助外部几何先验提升可靠性。近年来，4D radar PR等工作开始联合编码雷达散射截面与空间坐标，增强单帧特征的物理可解释性；TDFANet则通过状态空间建模实现跨帧特征交互，初步缓解了稀疏观测下的描述符退化问题。

这表明，4D毫米波雷达场景识别通常难以依赖单帧点云完成稳定判别，而更需要借助时序信息积累与外部模态补偿来提升场景表示的可靠性。从系统框架的视角来看，4D雷达点云的极稀疏特性使其天然适合于与激光雷达或视觉传感器形成互补，在多模态融合框架下发挥全天候鲁棒感知的核心作用。未来，物理属性驱动的特征编码与事件驱动的时序建模有望为4D雷达场景识别开辟新的技术路径。

2.3　RGB-D相机

RGB-D相机通过结构光或飞行时间原理同时获取彩色图像和深度图。在室内环境中，RGB-D相机能够提供稠密的几何与纹理信息，语义内容丰富，非常适合服务机器人及仓储自动导引车等场景的定位与识别。彩色信息还可用于提取视觉词袋模型或深度学习特征，与几何特征形成互补。这种几何与纹理天然对齐的特性，使其在多模态关联研究中具有独特价值——红绿蓝（red green blue，RGB）图像和深度图的像素级对应关系为跨模态特征融合提供了理想的对齐基准，有助于缓解多模态关联中的“关

系误区”问题。

其局限性同样明显，深度测量受环境光照、物体表面材质和反射特性的影响，在强光下或对玻璃及黑色物体时深度值大量缺失；室外有效距离通常不足10米，且太阳光中的红外成分会严重干扰飞行时间测量。因此，RGB-D相机在自动驾驶及无人机等室外长距离场景识别中较少作为主传感器，更多出现在室内或跨模态辅助的角色中，利用其纹理与深度信息对几何表征进行补充和增强，例如为激光雷达点云提供语义标签。在低空飞行场景中，RGB-D相机在近地悬停状态下的稠密深度测量可为无人机自主起降和精确抓取提供关键的空间感知信息，但其作用高度受限于飞行高度和环境光照条件。

2.4 多传感器联合与跨模态

单一传感器通常难以在所有环境下保持稳定识别性能，多传感器融合已成为提升低空感知鲁棒性的必然趋势。其技术实现主要体现为联合感知与跨模态匹配两种核心形式。联合感知通过在同一平台上部署异构传感器，利用特征级或决策级融合综合各模态优势，例如利用图像语义弥补激光雷达的纹理缺失；跨模态匹配则侧重挖掘异源数据对同一地理结构的观测共性，旨在实现如机载点云与卫星影像、地面地图之间的直接对齐。

针对低空飞行平台特有的运动特性与应用需求，系统设计需在上述融合范式的基础上，从视角、运动及算力三个维度进一步深化适配策略。在视角鸿沟弥合方面，适配算法应重点强化特征描述符的旋转泛化与跨高度不变性，通过跨视角对比学习等技术路径降低姿态变换对匹配一致性的干扰。针对高速飞行导致的点云重叠度下降及动态扫描失锁问题，运动畸变补偿策略显得尤为重要，通过结合惯性或里程计数据进行时序补偿，并利用滑动窗口累积构建局部几何子图，能够有效弥补单帧观测的不完备性。此外，考虑到低空平台载荷及计算资源的严苛限制，能效与算力均衡策略也是系统适配的关键。算法选型应倾向于采用稀疏卷积或状态空间模型等轻量化架构，以实现在资源受限边缘端的高精度实时感知。通过上述策略的协同应用，系统方能有效突破异构环境下的感知瓶颈，提升低空场景识别的整体稳健性。多传感器联合与跨模态方法正是本文第1.2节所讨论的多模态关联相关方法的技术基础。

上述三大适配策略的逻辑源头正是引言所述低空物流、巡检与应急救援等典型任务对跨高度不变性、气流噪声抑制、动态鲁棒性及边缘端实时性提出的具体约束，策略与约束之间的对应关系为后续章节中的方法比较与选型提供了应用导向的参考依据。

3 公共数据集与训练策略

3.1 公共数据集

面向点云场景识别的研究高度依赖公开数据集进行算法验证与性能比较。根据传感器模态、场景规模与环境复杂度的不同，现有数据集可划分为单模态与多模态两大类，各自服务于不同的技术评估需求。

（1）Oxford数据集由牛津大学采集，专注于长期自动驾驶环境下的地点识别。该数据集包含在固定路线，为期一年重复采集的旋转式激光雷达点云序列，核心评测任务为纯点云的地点识别与重定位。其最大挑战在于场景外观随季节和光照发生剧烈变化，而算法仅能依赖几何结构信息实现稳健匹配。这一特性使其成为检验特征表达方法能否在“语义盲区”取得进展的关键基准，因为外观变化迫使模型必须挖掘更本质的几何语义。

（2）NUS（Inhouse）Datasets是室内点云地点识别的代表性基准，涵盖商业办公楼、住宅公寓和大学教学楼三种室内场景，数据通过手持激光雷达扫描采集。该数据集严格划分查询集与参考集，可用于评估室内场景下的点云地点检索与重定位性能。与室外道路场景相比，室内环境通常空间范围更受限，但布局更复杂，且墙体、家具、走廊等重复结构大量存在，容易引发感知混淆；同时，人员活动和局部遮挡也会进一步增加匹配难度。因此，该数据集更强调算法在复杂布局、结构相似与动态遮挡条件下的泛化能力，对特征表达的局部细节刻画能力和鲁棒判别能力提出了较高要求，是检验室内场景中精细化点云表征方法的重要基准。

（3）ScanNet是面向室内三维场景理解的大规模标注数据集，由RGB-D传感器重建为带颜色信息的三角网格模型，提供体素级语义标签与实例级目标分割标注，涵盖二十余类常见物体与结构，包含超过一千五百个扫描场景。尽管该数据集主要服务于

语义分割与场景分类任务，但其丰富的语义标注使其成为多模态关联研究的理想预训练来源。研究者可利用ScanNet的彩色图像与深度点云对齐数据，预先训练跨模态特征融合模块，从而缓解“关系误区”中模态异构性带来的适配困难。

（4）Oxford RobotCar Dataset是前述单模态Oxford数据集的完整多传感器扩展版本。该数据集由同一团队发布，面向长期自动驾驶研究。数据集同步采集激光雷达、相机等传感器同步数据，覆盖光照、天气等长期变化。主要呈现了真实动态环境中，感知系统为保障长期可靠运行所必须应对的极端外观与结构变化。其激光雷达点云子集已成为长期点云地点识别领域的权威评测基准。与ScanNet不同，该数据集侧重室外大尺度场景，对系统框架在动态环境下的鲁棒性和实时性提出了综合考验。因此，Oxford RobotCar Dataset不仅可用于长期点云地点识别评测，也为多模态关联方法和多传感器融合系统在真实动态场景中的性能验证提供了重要基准。

（5）KITTI数据集是自动驾驶感知领域的经典基准，采集了城市场景、郊区道路和高速公路的激光雷达（Velodyne HDL-64E）点云、双目相机图像及全球定位系统（global positioning system，GPS）/惯性测量单元（inertial measurement unit，IMU）信息。官方任务包括立体视觉、三维目标检测、跟踪等，为多模态感知和单模态特征学习提供评测基础。2019年发布的SemanticKITTI扩展在原KITTI Odometry序列基础上提供了28类逐点语义标注，将应用拓展至大规模户外点云语义分割。由于KITTI点云和图像按序列采集，其结构也便于评估基于时序信息的特征增强方法，同时可用于测试单帧或单模态场景识别性能。

（6）nuScenes数据集由Motional公司发布，旨在提升前代基准的数据规模与标注丰富度。数据采集于波士顿、新加坡的城市复杂道路，采用激光雷达、毫米波雷达和摄像头组成的完整传感器套件，包含1000个20秒驾驶场景片段，提供详细矢量地图及23类物体的三维边界框标注。相较于侧重单一模态或纯地点识别任务的数据集，nuScenes的优势在于其多传感器同步采集机制和丰富标注信息，能够较好反映真实城市交通环境中动态目标密集、遮挡频繁、道路结构复杂等特点。因此，该数据集不仅适用于多模态感知任务评测，也为分析点云与图像、雷达等异构信息之间的关联建模提供了重要基础。对于本文所讨论的“关系误区”问题，研究者可以在nuScenes上系统评估特征拼接、注意力融合、非线性交互等不同策略在多传感器协同条件下的实际增益，并进一步考察其在复杂动态环境中的鲁棒性与泛化能力。

（7）USyd Campus Dataset是面向跨模态地点识别的中等规模数据集，通过移动机器人平台在悉尼大学校园半结构化环境，同步采集激光雷达点云与高分辨率球形全景图像。数据集提供严格时空对齐的点云-图像对及采集轨迹信息，旨在为点云与视觉信息的关联互补研究提供对比数据，验证跨模态检索与融合定位算法，突出在几何结构清晰但视觉外观可能存在重复的环境中识别所面临的挑战。该数据集对于评估跨模态场景识别方法在半结构化环境下的鲁棒性和泛化能力尤其有价值，此类环境中几何结构可靠，但视觉外观可能存在重复或模糊。

（8）低空无人机视角数据集。上述数据集主要面向地面自动驾驶与室内导航场景，其点云分布、观测视角及运动平台特性与低空飞行平台存在差异。近年来，研究开始关注无人机载多模态感知数据的系统化构建。UAVScenes基于MARS-LVIG扩展形成，提供超过12万帧图像与LiDAR点云的语义标注，数据包含惯性测量单元（inertial measurement unit，IMU）与全球导航卫星系统（global navigation satellite system，GNSS）信息，并支持地点识别、六自由度（six degrees of freedom，6-DoF）定位等任务，可为低空多模态场景理解提供新的评测基础。与此同时，CS-Campus3D以及近期发布的CS-Wild-Places进一步面向空地跨源点云地点识别，重点评估地面与机载LiDAR之间的跨视角匹配能力。此类数据集的共同特点在于视点高度变化显著，且跨源点云在覆盖范围、密度分布与可见结构上存在系统差异，因此对特征表达的旋转泛化能力、跨高度不变性及跨模态鲁棒性提出了更高要求。

总体而言，现有主流基准仍以地面平台为主，其在观测视角、运动模式和环境条件方面与低空场景存在系统性差异。UAVScenes等新兴数据集为验证方法在低空条件下的迁移能力提供了初步依据，后文第4.5节在跨数据集一致性分析中对此有所涉及，但低空场景下的大规模标准化评测仍有待进一步建设。

表2　面向点云场景识别的主要公共数据集概览

Table 2　Overview of Major Public Datasets for Point Cloud Scene Recognition

分类	数据集	年份	数据模态	场景	规模	主要用途	低空视角	跨季节/天气	多模态同步	场景识别
单模态	Oxford	2013	激光点云	城市道路	长期采集	室外地点识别	否	是	否	是
	NUS(Inhouse)	2014	激光点云	室内扫描	多遍历轨迹	室内地点识别	否	否	否	是
	ScanNet	2017	RGB-D点云	室内扫描	21类物体标注	实例分割、场景分类	否	否	是	可
多模态	Oxford RobotCar	2016	多模态同步	城市道路	超1000小时数据	多传感器融合、长期定位	否	是	是	是
	KITTI/SemanticKITTI	2012/2019	LiDAR点云+双目图像	城郊与高速公路	28类物体标注	3D检测、跟踪、语义分割	否	否	是	是
	nuScenes	2019	激光雷达+毫米波雷达+相机	城市街道	1000个片段,140万3D框	3D检测跟踪、多模态融合、轨迹预测	否	否	是	可
	USyd Campus	2016	LiDAR点云+全景图像	大学校园	多序列同步数据对	跨模态地点识别	否	否	是	是
	UAVScenes	2025	LiDAR点云+双目图像+IMU+GNSS	航空模型机场、岛屿、山谷	21个序列,120k+标注帧	无人机地点识别、多模态感知、场景理解	是	部分	是	是

3.2　训练策略

点云场景识别方法的训练目标是学习具有区分性的全局描述符,使同一地点的点云在描述符空间中彼此靠近,不同地点的点云相互远离。为实现这一目标,领域内广泛采用基于度量学习的训练策略,其核心是三元组损失及其变体。

给定一个锚点样本X_a、一个与锚点属于同一地点的正样本X_p以及一组来自不同地点的负样本X_n,以$G(.)$表示全局描述符提取函数。锚点与正样本的欧氏距离:

$$\rho_p = d\left(G\left(X_a\right),G\left(X_p\right)\right)\#(4)$$

锚点与第j个负样本的距离:

$$\rho_{n_j} = d\left(G\left(X_a\right),G\left(X_{n_j}\right)\right)\#(5)$$

三元组损失的基本形式为铰链函数约束下锚点与正样本距离和锚点与负样本距离之差的最小化:

$$\mathcal{L}_T = \sum_j max\left(0,\rho_p - \rho_{n_j} + \beta\right)\#(6)$$

式中β为边界参数,通常设为0.5,其作用在于控制正负样本对之间应保持的最小距离裕量。在此基础上,最大三元组损失将求和操作替换为取最大值,在每次迭代中仅针对最难区分的负样本进行优化,有助于学习更具判别力的特征表示。

为进一步提升大规模场景下的检索性能,四元组损失在三元组的基础上引入一个与所有样本结构均不相似的随机假样本X_n。最大四元组损失结合难样本挖掘与四元组约束,其形式为:

$$\mathcal{L}_{maxQ} = \max_j \left(0,\rho_p - \rho_{n_j} + \beta\right) + \max_j \left(0,\rho_p - \rho_n + \gamma\right)\#(7)$$

式中γ为第二个边界参数,通常设为0.2,用于控制锚点与假样本之间的距离裕量。与三元组损失相比,最大四元组损失通过额外的假样本约束有效避免模型陷入局部最优,在Oxford等大规模基准上通常表现更优。此外,软间隔损失采用求和操作对正负样本距离的对数损失进行平滑平均,训练过程更稳定但忽略了对难样本的针对性处理。上述训练策略在PointNetVLAD、LPD-Net、PPT-Net、ComPoint等各类方法中被广泛采用,具体选择通常取决于数据集规模和任务复杂度。

4　方法综合比较与分析

基于前文对点云场景识别方法的分类梳理,本

节通过量化数据对比与质性特性分析，揭示不同技术路线的优势、局限及内在权衡。表3-表6中汇总的性能数据均引自原始文献，各方法在训练数据集划分、点云下采样策略、输入点数量、优化器及学习率衰减方案等方面存在差异，部分方法还使用了不同的数据增强策略。因此，跨方法的数值对比仅可反映大致性能趋势，不作为严格精度排序的依据。各方法的详细实验设置请参见对应的原始文献。

4.1 评价指标

点云场景识别方法的性能评估依赖若干定量指标。这些指标从识别精度、检索效率和排序质量三个维度刻画算法特性。为便于理解后续各表的实验数据，本节对常用指标的定义与适用情形进行说明。

在识别精度方面，召回率是最主要的度量。R@1（Recall@1）指检索结果中排名第一的场景与查询场景属于同一地点的比例。该指标直接反映算法的最高匹配准确度，是衡量重定位性能的核心。R@1%（Recall@1%）将考察范围放宽至检索结果的前百分之一。若该候选池中包含至少一个正确匹配，则视为命中。该指标侧重评估算法将正确结果纳入候选池的能力，对于后续需进行几何验证的系统尤为重要。Recall@N进一步将范围扩展至前N个候选项。通过改变N的取值可绘制召回率曲线，观察算法在不同宽松程度下的性能变化。精度指标刻画了描述符的匹配质量，而效率指标则决定了方法在实际系统中的可用性。

检索延迟是衡量实时性的关键指标，通常以毫秒为单位。它记录从输入单帧点云到输出全局描述符的完整推理耗时，直接反映算法的部署可行性。与之配合的静态效率指标包括参数量和浮点运算量，前者衡量模型规模，后者衡量计算复杂度。

在排序质量方面，部分研究引入了F1分数和均值平均精度。F1分数兼顾精确率与召回率，适合评估正负样本不平衡条件下的整体判别能力。均值平均精度通过计算不同召回水平下精确率的均值，综合反映检索结果的排序质量，适用于多类别或细粒度场景匹配任务。

上述指标各有侧重。召回率类指标聚焦匹配准确率，效率类指标关乎部署成本，排序质量类指标则提供更全面的性能视角。在同一评测基准下综合考察这些指标，有助于更客观地判断方法的综合性能与适用边界。

4.2 特征表达方法性能对比

特征表达的核心目标是突破前文所述的“语义盲区”，实现点云几何结构与隐含语义的完备刻画，精度与效率的平衡是方法设计的核心矛盾。由于多模态方法在不同数据集上评测标准不一，本节以Oxford数据集上的R@1%、R@1为主要参考指标，该数据集对特征表达的区分性要求较高。表3汇总了代表性方法在Oxford数据集上的性能与效率指标，不同路线呈现出鲜明的分化趋势。

以PPT-Net为代表的点特征表达方法，凭借金字塔Transformer与分组自注意力机制，在细粒度特征与全局语义关联上优势显著，达到当前该路线的最佳精度，但参数量相对较高，适合精度优先的场景。

体素特征表达内部已形成完整性能频谱，MinkLoc3D奠定了该路线的轻量化范式并适配嵌入式部署；SVT-Net在保持超轻量级的同时，通过稀疏体素Transformer实现精度与速度的更优平衡；而MinkUNeXt则以较大参数量取得最高精度，展现了体素方法在极致精度上的潜力。

以ComPoint和PointWave为代表的复值域探索方法，ComPoint创新性地引入希尔伯特变换构建初始复值表示并利用虚部提供的互补几何视角强化结构捕捉能力。实验结果表明，复值表示具有极强的即插即用灵活性，当ComPoint作为增强模块集成至PPT-Net等先进实值域模型时，其在Oxford数据集上的平均召回率R@1%可达到97.9%且R@1指标达到93.2%。尽管涉及复数域运算，其独立架构的推理耗时仅为11毫秒。PointWave进一步提出波函数隐喻并利用复值特征的相位差动态调制特征交互。此类方法通过复数域的幅度与相位双重特性，旨在探索一种不同于传统实值加权的非线性交互机制，为实现更具判别力的特征建模提供了新的研究视角。

复值域表征利用幅度-相位双重特性建模语义不确定性，在显著降低参数量与推理耗时的同时保持了具有竞争力的识别精度，但其硬件算子加速适配尚不成熟。综合来看，点特征、体素特征与复值域特征三类方法在精度、效率与语义建模能力上形成了互补的技术谱系，三类方法分别适配高精度需求、轻量化部署与创新探索场景，研究者需结合实际应用约束灵活抉择。

表3　不同特征表达方法在Oxford数据集上的性能与效率对比

Table 3　Performance and Efficiency Comparison of Feature Representation Methods on the Oxford Dataset

方法	核心类别	Oxford		参数量(M)	耗时(ms)
		R@1%	R@1		
PointNetVLAD	点特征表达	80.3%	63.3%	19.8	15
PCAN	点特征表达	83.8%	70.7%	20.4	55
LPD-Net	点特征表达	94.9%	86.4%	19.8	26
PPT-Net	点特征表达	98.1%	93.5%	13.1	22
MinkLoc3D	体素特征表达	97.9%	93.2%	1.1	8
SVT-Net	体素特征表达	97.8%	93.7%	0.9	11
MinkUNeXt	体素特征表达	99.3%	97.7%	43.5	–
ComPoint	复值域探索	97.9%	93.2%	14.4	–
PointWave	复值域探索	97.5%	92.4%	6.6	–

注：各方法实验设置存在差异，跨方法对比仅供参考，旨在呈现总体趋势。表中ComPoint性能数据采用其作为插件集成于PPT-Net后的SOTA表现。

4.3　多模态关联方法特性分析

多模态关联旨在突破“关系误区”，即弥补单一模态在语义与几何表达上的不足。以PointNetVLAD在Oxford数据集上的63.3%为基准，表4数据显示，LPD-Net通过融合手工几何特征实现23.1%的增益，反衬出纯数据驱动的点云特征在局部几何结构刻画上表达能力不足，手工先验起到了有效的补充作用。以RGB图像为辅助模态的方法中，PRFusion引入注意力驱动的全局与局部融合模块，通过动态重标定模态贡献获得最高增益；QMMNet借鉴量子干涉概念设计非线性融合模块，以更低的参数量和计算开销取得了与注意力融合相当的效果。两者在解决“关系误区”上路径不同但效果相当，表明深度非线性交互是挖掘模态互补价值的核心路径，而简单特征拼接的增益相对有限。SketchNet以图像草图作为辅助模态，利用草图拓扑结构受姿态扰动影响小的特点，在跨场景识别中取得稳定增益，进一步验证了辅助模态选择的多样性对缓解“关系误区”的重要性。

综合上述方法，手工特征关联增强了局部可解释性，缺乏深层语义；特征拼接灵活性高，但增益有限；非线性融合增益显著，但算力需求高；草图辅助在跨场景拓扑鲁棒性上具独特优势，但依赖草图可得性。多模态融合需在增益幅度、系统复杂度与辅助模态可靠性之间综合权衡。

表4　多模态关联方法性能增益与特性

Table 4　Performance Gain and Characteristics of Multi-Modal Association Methods

方法	辅助模态	融合策略	融合后精度(R@1)	性能增益(R@1提升)
LPD-Net	手工特征	特征拼接	86.4%	+23.1%
MinkLoc++	RGB图像	特征拼接	96.7%	+33.4%
PRFusion	RGB图像	注意力融合	98.2%	+34.9%
QMMNet	RGB图像	干涉融合	96.7%	+33.4%
SketchNet	图像草图	交互式融合	95.9%	+32.6%

注：基准模型为PointNetVLAD在Oxford数据集的R@1(63.3%)。各方法实验设置存在差异，跨方法对比仅供参考。

4.4 系统框架效率与泛化能力探讨

系统框架设计需突破“适配限区”，核心在于平衡识别性能、资源消耗与场景适配性。单模态框架以点云为唯一输入，呈现专精化适配优势，能够针对特定传感器特性进行深度优化。多模态框架支持点云与图像、草图等多源输入，以系统复杂度换取更全面的环境感知能力。表5归纳了不同框架类型在输入模态、参数量及典型适用场景方面的对比。

综合来看，单模态框架泛化性依赖特征表达的通用性，其性能上限受限于单一模态所能承载的环境信息量；多模态框架如QMMNet和SketchNet的泛化能力则受限于辅助模态的可获得性。在无视觉光照的夜间或恶劣天气环境下，依赖图像的多模态框架部署困难程度增加，此时可回归单模态或采用雷达等全天候传感器作为补充方案。理想的多模态框架应具备根据传感器状态和环境条件动态调整各模态贡献的能力，而非在某一模态失效时整体性能显著下降。

模块化可扩展架构设计与跨场景预训练范式，仍是突破“适配限区”、实现框架在多样化场景与硬件条件下通用部署的关键方向。在此基础上，轻量化多模态融合、模态缺失下的鲁棒推理以及硬件感知的模型压缩等技术，正推动框架从专精化走向通用化。近年来兴起的动态网络推理范式也为框架适配提供了新的思路，根据输入复杂度和可用资源自适应调整网络深度或通道宽度，有望在同一架构内实现精度与效率的灵活平衡。系统框架的选型需根据任务需求、环境条件和资源约束进行综合权衡，未来的设计方向应向着按需配置、动态适配的目标持续演进。

表5 系统框架效率与适用场景对比

Table 5 Efficiency and Applicability Comparison of System Frameworks

框架类型	方法	输入模态	参数量(M)	使用场景
单模态	PointNet++	点云	1.7	层次化局部特征提取
	DGCNN	点云	1.7	动态图卷积，局部拓扑表达
	MinkLoc3D	点云	1.1	轻量级部署
多模态	QMMNet	点云、RGB图像	3.5	大规模多模态
	SketchNet	点云、图像草图	3.6	跨场景点云

注：单模态框架响应实时、系统紧凑，在无光等视觉受限环境下稳定性优，但易受点云稀疏退化制约；多模态框架通过信息冗余提升鲁棒性，但系统复杂度较高。

4.5 跨数据集性能一致性与鲁棒性评估

本节进一步评估代表性方法在多样化场景中的普适性能。现有的性能评估多局限于单一数据集，难以全面反映算法在不同传感器配置与环境动态下的泛化能力。为此，本文选取各技术路线中的代表性模型，在涵盖长期外观变化以及结构异质性、大规模城市交通干扰的多个异构数据集上进行一致性对比分析。表6汇总了主流方法在三大数据集上的召回性能与鲁棒性特质。

通过跨数据集的量化对比可以发现，体素方法在密度变化的场景间表现出较强的一致性平衡，如MinkLoc3D在Oxford与U. S. 之间仅波动4个百分点。相比之下，点特征方法PPT-Net虽在Oxford上精度领先，但在迁移至KITTI时骤降至39. 3%，推测与其邻域搜索策略对点云密度变化敏感有关。上述结果表明，不同技术路线在跨场景泛化能力上存在显著差异，单纯依赖单一基准的精度排名可能掩盖模型在真实环境中的鲁棒性问题。

值得注意，In-house U. S. 数据集上的性能差异进一步印证了体素方法与点方法在环境适应性上的结构性分野。In-house U. S. 为室内结构化环境，具有较规则的几何布局。MinkLoc3D的稀疏卷积在规则网格上操作，在室内规整环境中能更稳定地提取特征，因而取得97. 2%的R@1；而PPT-Net的邻域搜索在狭窄通道和密集家具环境中易受遮挡和局部点云缺失影响，性能相对受限(90. 1%)。这一现象表明，方法选型需充分考虑目标场景的结构特征。

此外，SketchNet由于依赖图像草图作为辅助模

表6　代表性场景识别方法在多源异构数据集下的性能一致性对比

Table 6　Performance consistency comparison of representative scene recognition methods across heterogeneous multi-source datasets

方法类别	代表性模型	Oxford 数据集(R@1)	KITTI 数据集(R@1)	In-house U. S.(R@1)	环境适应性与鲁棒性简析
基准方法	PointNetVLAD	63.3%	46.4%	86.1%	基准性能受限,跨场景衰减显著
点特征表达	PPT-Net	93.5%	39.3%	90.1%	精度上限较高,对密度异质性敏感
体素特征表达	MinkLoc3D	93.2%	58.3%	97.2%	环境适应性较强,性能表现均衡
复值域探索	ComPoint	93.2%	45.2%	92.3%	跨场景波动较小,特征表征较稳健
多模态关联	SketchNet	95.9%	79.8%	-	跨场景鲁棒性极强,泛化优势显著

注:各方法实验设置存在差异,跨方法对比仅供参考。In-house U. S. 为室内结构化数据集,体素方法(MinkLoc3D)因规则网格划分在规整环境中更具优势,点方法(PPT-Net)因邻域搜索在狭窄空间易受遮挡干扰。SketchNet依赖图像草图模态,In-house U. S. 数据集为纯激光点云,无法进行该实验。

态,而 In-house U. S. 数据集仅提供纯激光点云,缺少对应的图像输入,因此在该数据集上无法进行评测,表中以"-"标注。这也从侧面说明,多模态方法虽然泛化优势显著,但对辅助模态的可用性存在依赖。

复值域方法 ComPoint 展现了相位信息在缓解匹配歧义方面的独特价值,其在跨数据集评估中表现出比传统实值点特征更稳健的趋势。而融入了非线性交互的多模态框架如 SketchNet 则展现出卓越的泛化能力,尤其在结构差异巨大的 KITTI 场景中,通过引入草图关联有效补全了点云缺失的纹理结构信息,实现了显著的泛化优势。本小节通过对这些波动规律的解析,总结出能够兼顾识别精度与环境普适性的关键优化路径,从而为突破前文所述的适配限区提供实证支撑。

综上所述,模型在跨数据集任务中的表现波动,揭示了特定路线在应对环境迁移时的脆弱性。在低空飞行等对感知可靠性要求极高的应用中,通过建立动态补偿与多模态协同的适配体系,方能有效应对环境异质性带来的识别挑战。

5　结　语

本文以点云地点识别与重定位为核心任务,对特征表达、多模态关联与系统框架三条技术主线相关研究进行了系统梳理与综合分析。通过对代表性方法、传感器模态及公开数据集的归纳比较可知,现有研究已在局部几何建模、全局描述符构建、跨模态信息互补和识别框架优化等方面取得持续进展,推动点云场景识别由单一特征驱动逐步发展为面向复杂环境的多层次协同建模体系。综合全文分析结果可见,特征表达、多模态关联与系统设计分别对应复杂场景下语义表达不足、模态协同不稳和部署适配受限等核心问题,不同技术路线的差异本质上体现为识别精度、环境鲁棒性、计算复杂度与工程可实现性之间的权衡。总体而言,现有方法已显著提升了点云场景识别性能,但对复杂动态环境的稳定适应能力仍有不足,引言提出的关键问题尚未得到统一解决。

进一步分析表明,当前研究仍存在若干共性局限:(1)点云特征表达对稀疏采样、结构重复、远距离退化和动态遮挡等因素仍较敏感,描述符的区分性与稳定性有待提升。实值域方法虽可通过注意力机制、对比学习或大规模预训练等策略增强语义建模能力,但在极端条件下性能退化仍较明显。复值域等新兴表征路线尽管在特定基准上展现了潜力,但在算子优化程度、工业硬件兼容性及长效稳定性方面仍面临显著瓶颈,其在实际生产环境中的可靠性尚待大规模背书。(2)多模态方法易受模态差异、对齐误差与质量波动影响,融合收益尚不稳定;(3)现有框架多面向特定传感器配置和任务条件设计,跨场景、跨平台与资源受限条件下的适配能力仍显不足;(4)不同研究在输入形式、训练策略与评价指标等方面缺乏统一规范,制约了方法间的客观比较与规律总结。

未来研究应围绕上述问题推进协同突破。在特

征表达方面，应加强局部几何、全局上下文与场景语义的联合建模，并结合旋转不变表示、时序信息利用、自监督预训练与轻量化设计，提升模型鲁棒性与泛化能力。在多模态关联方面，应由浅层融合转向稳定的深层交互与自适应协同机制，提高模型对模态缺失、弱对齐和质量波动的容忍能力。在系统框架方面，应推动模型向模块化、通用化方向发展，增强其对传感器条件、环境变化与任务需求的动态适配能力。同时，应完善覆盖多传感器、多场景和多时间跨度的统一评测体系，加强对鲁棒性、实时性、泛化性与资源开销的综合评估，以促进点云场景识别技术向真实复杂环境应用持续推进。

参考文献

Arandjelovic R, Gronat P, Torii A, Pajdla T and Sivic J. 2016. NetVLAD: CNN architecture for weakly supervised place recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 5297-5307 [DOI: 10.1109/CVPR.2016.572]

Cabrera J J, Santo A, Gil A, Viegas C and Payá L. 2025. MinkUNeXt: point cloud-based large-scale place recognition using 3D sparse convolutions. Array, 25: 100569 [DOI: 10.1016/j.array.2025.100569]

Chen Y, Zhuang Y, Wang B and Huai J. 2025. 4D RadarPR: context-aware 4D radar place recognition in harsh scenarios. ISPRS Journal of Photogrammetry and Remote Sensing, 221: 210-223 [DOI: 10.1016/j.isprsjprs.2025.01.033]

Fan Z X, Song Z, Liu H Y, Lu Z, He J and Du X. 2022. SVT-Net: super light-weight sparse voxel transformer for large scale place recognition//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual: AAAI Press: 551-560 [DOI: 10.1609/aaai.v36i1.19971]

Gong J Y, Lou Y J, Liu F Q, Zhang Z W, Chen H M, Zhang Z Z, Tan X, Xie Y and Ma L Z. 2023. Scene point cloud understanding and reconstruction technologies in 3D space. Journal of Image and Graphics, 28(06): 1741-1766 (龚靖渝，楼雨京，柳奉奇，张志伟，陈豪明，张志忠，谭鑫，谢源，马利庄. 2023. 三维场景点云理解与重建技术. 中国图象图形学报，28(06)：1741-1766) [DOI: 10.11834/jig.230004]

Hao W C, Zhang W F and Jin H Y. 2024. SAGE-Net: employing spatial attention and geometric encoding for point cloud based place recognition. IEEE Robotics and Automation Letters, 9(6): 4958-4965 [DOI: 10.1109/LRA.2024.3387112]

Hao W C, Zhang W F and Su H. 2024. RIA-Net: rotation invariant aware 3D point cloud for large-scale place recognition. IEEE Robotics and Automation Letters, 9(6): 5014-5021 [DOI: 10.1109/LRA.2024.3386597]

Huang T Y, Dong B W, Yang Y H, Huang X, Lau RW, Ouyang W, et al. 2023. CLIP2Point: transfer CLIP to point cloud classification with image-depth pre-training//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22157-22167 [DOI: 10.1109/ICCV51070.2023.02157]

Hui L, Yang H, Cheng M M, Xie J and Yang J. 2021. Pyramid point cloud transformer for large-scale place recognition//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 6098-6107 [DOI: 10.1109/ICCV48922.2021.00604]

Komorowski J. 2021. MinkLoc3D: point cloud based large-scale place recognition//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 1790-1799 [DOI: 10.1109/WACV48630.2021.00183]

Komorowski J, Wysoczańska M and Trzcinski T. 2021. MinkLoc++: LiDAR and monocular image fusion for place recognition//2021 International Joint Conference on Neural Networks. Shenzhen, China: IEEE: 1-8 [DOI: 10.1109/IJCNN52387.2021.9533373]

Lan Tiansheng, Zhang Ling, Liu Libo, Zhang Ruonan. Wave-particle duality-driven feature representation paradigm for point cloud place recognition [J/OL]. Journal of Image and Graphics, 2026, 1-17. (兰天升，张玲，刘立波，张若楠. 波粒二象性驱动的点云场景识别特征表达范式[J/OL]. 中国图象图形学报，2026，1-17 [DOI: 10.11834/jig.250631]

Li B W, Guo J Q, Liu H B, Zou Y, Ding Y, Chen X, et al. 2025. Uniscene: unified occupancy-centric driving scene generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 11971-11981 [DOI: 10.1109/CVPR52734.2025.01118]

Li G and Zhang R. 2023. A point is a wave: point-wave network for place recognition//ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island, Greece: IEEE: 1-5 [DOI: 10.1109/ICASSP49357.2023.10094873]

Li W, Ma Y, Ren J, Ou J, Zhou J and Huang P. 2026. VCC: vertical feature and circle combined descriptor for 3D place recognition. Sensors, 26(4): 1185 [DOI: 10.3390/s26041185]

Li Z K, Huang L and He J R. 2019. A multiscale deep middle-level feature fusion network for hyperspectral classification. Remote Sensing, 11(6): 695 [DOI: 10.3390/rs11060695]

Liang D, Zhou X, Xu W, Zhu X, Zou Z, Ye X, et al. 2024. PointMamba: a simple state space model for point cloud analysis. Advances in Neural Information Processing Systems, 37: 32653-32677 [DOI: 10.5555/3666122.3666169]

Liu C F, Chen G and Song R. 2024. LPS-Net: lightweight parameter-shared network for point cloud-based place recognition//2024 IEEE International Conference on Robotics and Automation. Yokohama,

Japan: IEEE: 448-454 [DOI: 10.1109/ICRA57147.2024.10610758]

Liu W, Fei J L and Zhu Z T. 2022. MFF-PR: point cloud and image multi-modal feature fusion for place recognition//2022 IEEE International Symposium on Mixed and Augmented Reality. Singapore: IEEE: 647-655 [DOI: 10.1109/ISMAR55827.2022.00082]

Liu X, Zhang R, Wang J and Li G. 2022. SparseARFM-SI: rotary point cloud place recognition based on multi-resolution and attention mechanism//2022 IEEE International Conference on Visual Communications and Image Processing. Suzhou, China: IEEE: 1-5 [DOI: 10.1109/VCIP56404.2022.10008851]

Liu Z, Zhou S, Suo C, Yin P, Chen W, Wang H, et al. 2019. LPD-Net: 3D point cloud learning for large-scale place recognition and environment analysis//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE: 2831-2840 [DOI: 10.1109/ICCV.2019.00292]

Lu S, Zhuo G, Wang H, Zhou Q, Zhou H, Huang R, et al. 2025. TDFANet: encoding sequential 4D radar point clouds using trajectory-guided deformable feature aggregation for place recognition//2025 IEEE International Conference on Robotics and Automation. Atlanta, USA: IEEE: 1-7 [DOI: 10.1109/ICRA55743.2025.11127648]

Lu Y, Yang F, Chen F and Xie D. 2020. PIC-Net: point cloud and image collaboration network for large-scale place recognition. arXiv preprint: arXiv:2008.00658 [DOI: 10.48550/arXiv.2008.00658]

Ma J, Zhang J, Xu J, Ai R, Gu W and Chen X. 2022. OverlapTransformer: an efficient and yaw-angle-invariant transformer network for LiDAR-based place recognition. IEEE Robotics and Automation Letters, 7(3): 6958-6965 [DOI: 10.1109/LRA.2022.3178797]

Pan Y, Xu X, Li W, Cui Y, Wang Y and Xiong R. 2021. Coral: colored structural representation for bi-modal place recognition//2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague, Czech Republic: IEEE: 2084-2091 [DOI: 10.1109/IROS51168.2021.9635839]

Pang Y, Wang W, Tay F E H, Liu W, Tian Y and Yuan L. 2022. Masked autoencoders for point cloud self-supervised learning. //Proceedings of the European Conference on Computer Vision (ECCV). Cham, Switzerland: Springer: 604-621. [DOI: 10.1007/978-3-031-20086-1_35]

Peng G, Li H, Zhao Y, Zhang J, Wu Z, Zheng P, et al. 2024. TransLoc4D: transformer-based 4D radar place recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 17595-17605 [DOI: 10.1109/CVPR52733.2024.01666]

Qi CR, Su H, Mo K and Guibas LJ. 2017. PointNet: deep learning on point sets for 3D classification and segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 652-660 [DOI: 10.1109/CVPR.2017.16]

Qi CR, Yi L, Su H and Guibas LJ. 2017. PointNet++: deep hierarchical feature learning on point sets in a metric space. Advances in Neural Information Processing Systems, 30: 5099-5108 [DOI: 10.5555/3294771.3294858]

Qiu Q, Wang W, Ying H, Liang D, Gao H and He X. 2024. SelFLoc: selective feature fusion for large-scale point cloud-based place recognition. Knowledge-Based Systems, 295: 111794 [DOI: 10.1016/j.knosys.2024.111794]

Shu D W and Kwon J. 2023. Hierarchical bidirected graph convolutions for large-scale 3-D point cloud place recognition. IEEE Transactions on Neural Networks and Learning Systems, 35(7): 9651-9662 [DOI: 10.1109/TNNLS.2023.3265115]

Su J, Li G, Wang S, Su H, Lin W and Gao W. 2026. Com-PCQA: no-reference point cloud quality assessment via complex-valued feature learning. IEEE Transactions on Image Processing, 35: 3339-3353. [DOI: 10.1109/TIP.2026.3676290]

Sun Q, Liu H, He J, Fan Z and Du X. 2020. DAGC: employing dual attention and graph convolution for point cloud based place recognition//Proceedings of the 2020 International Conference on Multimedia Retrieval. Dublin, Ireland: ACM: 224-232 [DOI: 10.1145/3372278.3390693]

Uy M A and Lee G H. 2018. PointNetVLAD: deep point cloud based retrieval for large-scale place recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4470-4479 [DOI: 10.1109/CVPR.2018.00470]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. 2017. Attention is all you need. Advances in Neural Information Processing Systems, 30: 5998-6008 [DOI: 10.5555/3294771.3294781]

Wang S, Kang Q, She R, Zhao K, Song Y and Tay WP. 2024. PRFusion: toward effective and robust multi-modal place recognition with image and point cloud fusion. IEEE Transactions on Intelligent Transportation Systems, 25(12): 20523-20534 [DOI: 10.1109/TITS.2024.3465830]

Wang Y, Sun Y, Liu Z, Sarma SE, Bronstein MM and Solomon JM. 2019. Dynamic graph CNN for learning on point clouds. ACM Transactions on Graphics, 38(5): 1-12 [DOI: 10.1145/3326362]

Wen K, Zhang R and Li G. 2022. PointNetGeM: simple and efficient point cloud based network for place recognition//2022 IEEE International Conference on Visual Communications and Image Processing. Suzhou, China: IEEE: 1-5 [DOI: 10.1109/VCIP56404.2022.10008869]

Weng S, Zhang R and Li G. 2022. Erinet: effective rotation invariant network for point cloud based place recognition//2022 IEEE International Conference on Visual Communications and Image Processing. Suzhou, China: IEEE: 1-5 [DOI: 10.1109/VCIP56404.2022.10008830]

Wiesmann L, Nunes L, Behley J and Stachniss C. 2023. KPPR: exploiting momentum contrast for point cloud-based place recognition. IEEE Robotics and Automation Letters, 8(2): 592-599 [DOI: 10.

1109/LRA.2022.3228174]

Xia Y, Xu Y, Li S, Wang R, Du J, Cremers D, et al. 2021. SOE-Net: a self-attention and orientation encoding network for point cloud based place recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 11348-11357 [DOI: 10.1109/CVPR46437.2021.01119]

Xu R, Wang X, Wang T, Chen Y, Pang J and Lin D. 2024. PointLLM: empowering large language models to understand point clouds// European Conference on Computer Vision. Cham: Springer Nature Switzerland: 131-147 [DOI: 10.1007/978-3-031-72698-9_8]

Yu X, Tang L, Rao Y, Huang T, Zhou J and Lu J. 2022. Point-BERT: pre-training 3D point cloud transformers with masked point modeling//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 19313-19322 [DOI: 10.1109/CVPR52688.2022.01871]

Zhang R, Li G, Gao W and Li TH. 2024. ComPoint: can complex-valued representation benefit point cloud place recognition? . IEEE Transactions on Intelligent Transportation Systems, 25(7): 7494-7507 [DOI: 10.1109/TITS.2024.3351215]

Zhang R, Liu X, Li G, Li TH and Zhao P. 2024. Sketch-aided interactive fusion point cloud place recognition//Proceedings of the 2024 International Conference on Multimedia Retrieval. Phuket, Thailand: ACM: 1115-1119 [DOI: 10.1145/3652583.3657613]

Zhang R, Li G, Gao W and Liu S. 2025. A quantum-inspired framework in leader-servant mode for large-scale multi-modal place recognition. IEEE Transactions on Intelligent Transportation Systems, 26(2): 2027-2039 [DOI: 10.1109/TITS.2024.3497574]

Zhang W X and Xiao C X. 2019. PCAN: 3D attention map learning using contextual information for point cloud based retrieval//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 12436-12445 [DOI: 10.1109/CVPR.2019.01272]

Zhao H, Jiang L, Jia J, Torr P and Koltun V. 2021. Point transformer// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 16259-16268 [DOI: 10.1109/ICCV48922.2021.01595]

Zhou B, Zhan L and Jiang J. 2025. LFE-PointMamba: point cloud learning via local feature enhancement and state space model. Knowledge-Based Systems, 315: 114350 [DOI: 10.1016/j.knosys.2025.114350]

Zhou Z, Zhao C, Adolfsson D, Su S, Gao Y and Duckett T. 2021. NDT-Transformer: large-scale 3D point cloud localisation using the normal distribution transform representation//2021 IEEE International Conference on Robotics and Automation. Xi'an, China: IEEE: 5654-5660 [DOI: 10.1109/ICRA48506.2021.9560932]

Zywanowski K, Banaszczyk A, Nowicki M R and Komorowsko J. 2022. MinkLoc3D-SI: 3D LiDAR place recognition with sparse convolutions, spherical coordinates, and intensity. IEEE Robotics and Automation Letters, 7(2): 1079-1086 [DOI: 10.1109/LRA.2021.3136863]

作者简介

张若楠：女，博士，准聘副教授，硕士生导师，主要研究领域为点云场景识别方法研究。E-mail：zhangrn@nxu.edu.cn.

刘立波（通讯作者）：女，博士，教授，博士生导师，主要研究领域为智能信息处理、计算机视觉。E-mail：liulib@163.com

耿莹：女，管理学学士在读，主要研究领域为计算机视觉、深度学习。E-mail：gy01162026@163.com.

马凯：男，工学学士在读，主要研究领域为深度学习、自然语言处理和情感分析。E-mail：KaiMa1231@outlook.com.

周奥临：男，工科硕士在读，主要研究领域为点云场景识别方法研究。E-mail：12025130987@stu.nxu.edu.cn