E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP37　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)06-1744-48

论文引用格式:Feng M T, Shen J H, Wu Z J, Peng W X, Zhong H, Guo Y L, Shu X B, Zhang H, Dong W S and Wang Y N. 2025. Advancements in 3D vision understanding using multimodal large language models. Journal of Image and Graphics, 30(6):1744-1791(冯明涛, 沈军豪, 武子杰, 彭伟星, 钟杭, 郭裕兰, 舒祥波, 张辉, 董伟生, 王耀南. 2025. 多模态大模型驱动的三维视觉理解技术前沿进展. 中国图象图形学报, 30(6):1744-1791)[DOI:10.11834/jig.240588]

多模态大模型驱动的三维视觉理解技术前沿进展

冯明涛[1],沈军豪[1],武子杰[2],彭伟星[2],钟杭[2],郭裕兰[3],舒祥波[4],
张辉[2],董伟生[1],王耀南[2*]

1. 西安电子科技大学,西安　710072; 2. 湖南大学,长沙　410073;
3. 中山大学,广州　510006; 4. 南京理工大学,南京　210014

摘　要: 三维(3D)视觉感知和理解在机器人导航、自动驾驶以及智能人机交互等众多领域广泛应用,是计算机视觉领域备受瞩目的研究方向。随着多模态大模型的发展,它们与3D视觉数据的融合取得快速进展,为理解和与3D物理世界交互提供了前所未有的能力,并展现了独特优势,如上下文学习、逐步推理、开放词汇能力和丰富的世界知识。本文涵盖3D视觉数据基本表示,从点云到3D高斯泼溅;梳理主流多模态大模型的发展脉络;对联合多模态大模型的3D视觉数据表征方法进行归纳总结;梳理基于多模态大模型的3D理解任务,如3D生成与重建、3D目标检测、3D语义分割、3D场景描述、语言引导的3D目标定位和3D场景问答等;提炼基于多模态大模型的机器人具身智能系统中空间理解能力的提升策略;最后梳理了核心数据集和对未来前景的深刻讨论,以期促进该领域的深入研究与广泛应用。本文全面分析揭示了本领域的重大进展,强调利用多模态大模型进行3D视觉理解的潜力和必要性。因此,本综述目标是为未来的研究绘制一条路线,探索和扩展多模态大模型在理解和与复杂3D世界的互动能力,为空间智能领域的进一步发展铺平道路。

关键词: 三维视觉;多模态大模型;三维视觉表征;三维视觉生成;三维重建;机器人三维视觉;三维场景理解

Advancements in 3D vision understanding using multimodal large language models

Feng Mingtao[1], Shen Junhao[1], Wu Zijie[2], Peng Weixing[2], Zhong Hang[2], Guo Yulan[3],
Shu Xiangbo[4], Zhang Hui[2], Dong Weisheng[1], Wang Yaonan[2*]

1. Xidian University, Xi'an　710072, China; 2. Hunan University, Changsha　410073, China; 3. Sun Yat-sen University, Guangzhou　510006, China; 4. Nanjing University of Science and Technology, Nanjing　210014, China

Abstract: Three-dimensional (3D) visual perception and understanding are fundamental to numerous applications, including robotic navigation, autonomous driving, and intelligent human-computer interaction. As one of the most prominent research directions in computer vision, 3D vision has experienced rapid advancement, particularly with the rise of multimodal large language models (MLLMs). The integration of MLLMs with 3D visual data has unlocked unprecedented capabilities for understanding and interacting with the physical 3D world. These models bring distinct advantages, such as con-

收稿日期:2024-09-29;**修回日期**:2024-12-22;**预印本日期**:2024-12-29

*通信作者:王耀南　yaonan@hnu.edu.cn

基金项目:国家自然科学基金项目(62373293); 江西省自然科学基金青年项目(20242BAB20050)

Supported by: National Natural Science Foundation of China (62373293); Young Scholars Fund of the Natural Science Foundation of Jiangxi Province, China (20242BAB20050)

textual learning, step-by-step reasoning, open-vocabulary support, and rich world knowledge, making them transformative tools in the field of 3D vision. This paper provides a comprehensive overview of the latest progress in 3D vision understanding driven by MLLMs. It begins by addressing the foundational representations of 3D visual data. From point clouds to 3D Gaussian splatting, the review systematically examines mainstream data representation methods, which form the backbone for intelligent processing and analysis of 3D visual information. These representations enable the integration of semantic, spatial, and structural information, serving as a critical basis for downstream tasks in 3D vision. Following this, the paper traces the evolution of MLLMs, starting from the development of large language models and their extension into multimodal systems. It highlights the emergence of vision-language models that synergize text and visual data, offering significant potential for advancing 3D vision. The combination of multimodal pretrained priors and 3D data representations has opened new avenues for cross-modal understanding and interaction. The ability of MLLMs to align multimodal features while reasoning across modalities has proven crucial for overcoming traditional limitations in 3D vision, such as sparse data, occlusion, and noise. The review then focuses on the methods for representing 3D visual data using MLLMs, providing an in-depth synthesis of current strategies. It discusses how MLLMs leverage pretrained knowledge to interpret 3D information, facilitate multimodal feature alignment, and enhance the contextual understanding of complex 3D scenes. For instance, MLLMs enable the integration of 3D data with semantic priors, improving the efficiency and accuracy of tasks such as 3D object recognition, scene reconstruction, and spatial reasoning. In the context of specific 3D vision tasks, this paper explores various applications of MLLMs, including 3D generation and reconstruction, 3D object detection, semantic segmentation, scene description, language-guided 3D object localization, and 3D scene question answering. These tasks demonstrate the transformative influence of MLLMs on 3D vision, showcasing their ability to elevate task performance by incorporating multimodal capabilities. For example, 3D generation tasks benefit from the contextual knowledge of MLLMs, enabling the creation of semantically coherent and visually accurate 3D content. Similarly, 3D object detection and segmentation tasks leverage the reasoning capabilities of MLLMs to identify and classify objects in complex scenes more effectively. The role of MLLMs extends beyond traditional 3D vision tasks to applications in embodied robotic intelligence systems. Examples include robotic 3D grasping and 3D visual navigation, where MLLMs facilitate spatial understanding and decision making in dynamic environments. By integrating multimodal reasoning with 3D perception, MLLMs enable robots to perform language-guided manipulations, navigate cluttered spaces, and interact seamlessly with the physical world. The paper also provides a detailed examination of datasets that support research in this domain. It reviews fundamental 3D datasets, such as those designed for point cloud analysis, voxel-based representations, and mesh modeling, alongside multimodal 3D vision-language datasets. These datasets are analyzed in terms of their scale, diversity, and application scope, offering a robust foundation for training and evaluating MLLMs in various 3D vision tasks. However, the lack of diverse and representative datasets remains a significant challenge, limiting the generalizability and robustness of existing models. Building on these foundations, the paper addresses key challenges and future research priorities in MLLM-driven 3D vision understanding. Major challenges include unifying 3D representation formats, improving the spatial reasoning capabilities of MLLMs, enhancing the generalization of MLLMs for diverse 3D data processing, and addressing practical deployment mechanisms for multimodal models in real-world 3D vision tasks. In addition, the paper highlights the need for efficient model miniaturization for edge-side applications, such as autonomous robots and drones, and explores the potential of cloud-edge collaborative frameworks to enhance 3D vision tasks in resource-constrained environments. The future research directions proposed in this paper aim to address these challenges and further advance the field. Key priorities include developing scalable and efficient 3D data representations, designing domain-adaptive MLLMs, and creating comprehensive multimodal benchmarks that reflect real-world complexities. Moreover, the paper emphasizes the importance of fostering innovation in multimodal reasoning frameworks, enabling MLLMs to interpret, generate, and interact with 3D data seamlessly. Exploring the integration of real-time data streams with multimodal pretraining strategies can also provide valuable insights into dynamic 3D environments. By offering a comprehensive analysis of the progress, challenges, and future directions in this field, the paper underscores the transformative potential of MLLMs for 3D vision understanding. It highlights the necessity of leveraging these models to bridge the gap between perception and reasoning, enabling systems to interact effectively with the complex 3D world. Findings emphasize the importance of addressing existing limitations in 3D vision and pave the

way for the continued evolution of artificial intelligence in spatially complex and multimodal environments. This review aims to inspire further exploration and expansion of MLLMs in 3D vision understanding, providing a roadmap for future research in this domain. Through the synthesis of state-of-the-art developments, the paper lays a foundation for advancing spatial intelligence, fostering deeper integration between AI systems and the physical world. Ultimately, the analysis highlights the potential of MLLMs to revolutionize 3D vision tasks, promoting their broad application across industries and accelerating the progress of intelligent systems in complex, multimodal scenarios.

Key words: 3D vision; multimodal large model; 3D visual representation; 3D vision generation; 3D reconstruction; robot 3D vision; 3D scene understanding

0 引　言

三维(three dimensions, 3D)视觉是感知和理解世界三维结构的能力。它是一项基本的人类技能，理解3D空间环境对于许多现实世界的应用至关重要，也是许多应用的基本要求，包括自主机器人(自动驾驶车辆和无人驾驶飞机)、导航、增强/虚拟现实、工业自动化和城市规划。近年来，3D视觉领域取得了重大进展，从对3D环境和低层次视觉任务的主动/被动感知到高层次的语义理解和决策推理。

大型语言模型(large language model, LLM)的出现标志着自然语言处理的一个变革时代，使机器能够以前所未有的方式理解、生成和与人类语言交互。尽管LLM在大多数自然语言处理任务中表现出了令人惊讶的零/少样本推理性能，但本质上对视觉“视而不见”，因为它们只能理解离散文本。同时，大型视觉模型(large vision model, LVM)具备精细化视觉信息的解析能力，但推理能力通常会滞后。鉴于这种互补性，LLM和LVM相互发展，催生了多模态大模型的新领域，它基于LLM，具有接收、推理和输出多模态信息的能力。多模态大模型在图像字幕、问题回答、开放词汇识别和图像生成等广泛的图像处理问题上显著提高了性能。大量的外部模态信息被证明可以有效提高图像的理解能力，并具有良好的可移植性。多模态大模型与3D视觉数据的融合提供了一个独特的机会，可以增强计算模型对3D物理世界的理解和与物理世界的交互，从而引领各个子领域的创新。通过利用多模态大模型的固有优势，包括零样本学习、高级推理和丰富的知识，将多模态大模型与3D视觉数据集成在复杂的3D环境中进行解释、推理或规划。这促使3D视觉领域的研究人员探索大规模多模态模型，以改善3D视觉表示。如图1所示，在过去几年该领域发表论文数量(按arxiv提交/发表日期统计)稳步增长。为此，进行全面综述对于掌握多模态大模型在三维视觉理解领域的最新进展至关重要。

本综述对多模态大模型与3D视觉的交叉进行系统梳理，对该领域近3年来300余篇最新方法、应用和挑战进行概述，总体结构框架如图2所示，图中

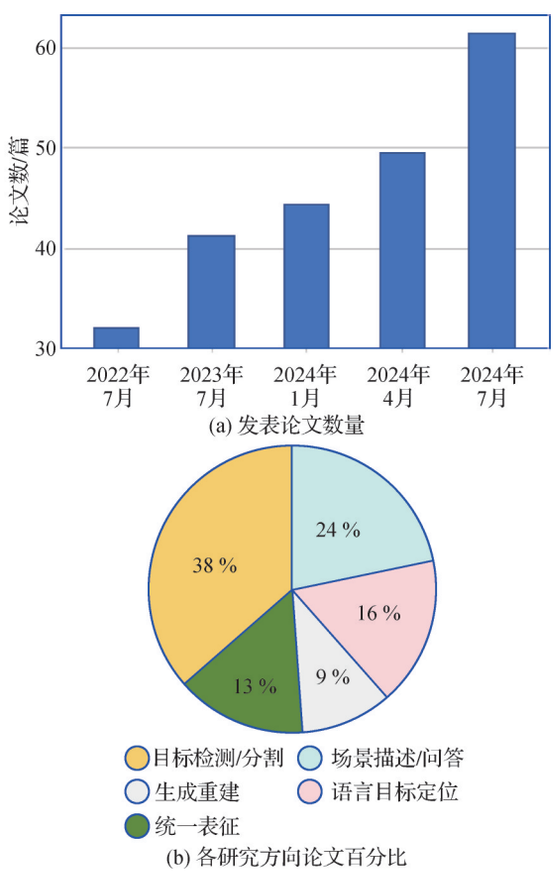

图1　基于多模态大模型的3D视觉论文数量及本综述中各研究方向论文百分比

Fig. 1　The number of 3D vision papers based on multimodal large models and the percentage of research directions represented by the papers in this review((a)the number of papers; (b)the percentage of research directions)

紫色箭头表示各节间逻辑顺序，灰色箭头表示数据集用于支撑训练和评估三维视觉应用。第 1 节介绍常见 3D 视觉表示的相关背景；第 2 节对多模态大模型发展脉络进行概述，重点分析与 3D 视觉结合紧密的模型；第 3 节介绍基于多模态大模型的 3D 视觉表征方法，这些表征方法通过对齐 3D 视觉特征与其他模态特征，挖掘多模态大模型的强大表征能力，为 3D 视觉理解下游任务赋能；第 4 节分析通过多模态大模型功能增强 3D 理解的下游任务，展示多模态大模型和 3D 视觉理解任务成功融合的各种领域，如将多模态大模型用来增强 3D 生成与重建、3D 目标检测、3D 语义分割、3D 场景描述、语言引导的 3D 目标定位和 3D 场景问答等任务；第 5 节介绍将多模态大模型与机器人 3D 视觉任务统一起来的任务，通过利用现有多模态大模型的知识，用于机器人 3D 抓取、3D 视觉导航等；第 4、5 节提供了多模态大模型赋能 3D 视觉理解任务方法的完整概述，以展示多模态大模型 + 3D 领域的完整图景；第 6 节梳理用于支撑训练和评估这些应用方向的数据集；第 7 节总结该领域的挑战和潜在的未来研究方向。

本综述主要贡献包括：1）梳理多模态大模型为 3D 视觉表示学习提供的智能解决方案，介绍利用多模态大模型的力量来学习 2D/3D 视觉、文本等多模态数据的联合表示，使多模态大模型能够更有效地推理空间关系、3D 对象属性和 3D 场景上下文。2）系统调研多模态大模型驱动的 3D 视觉理解任务，深入总结多模态 3D 视觉理解的最新研究进展，主要涵盖近 3 年内的最新研究成果，具有较强的时效性，并根据下游任务进行分类，方便读者获取。3）总结多模态 3D 视觉数据集，重点包含 3D、文本和图像 3 个模态的数据集及涵盖的任务方向，为不同领域任务发展提供支撑。4）对未来前景深刻讨论，在系统调研和综合性能比较的基础上，讨论未来研究方向，包括更高效的 3D 视觉表征、多模态大模型 3D 空间推理能力提升、机器人边端 3D 视觉任务场景下多模态大模型小型化和云边协同等。

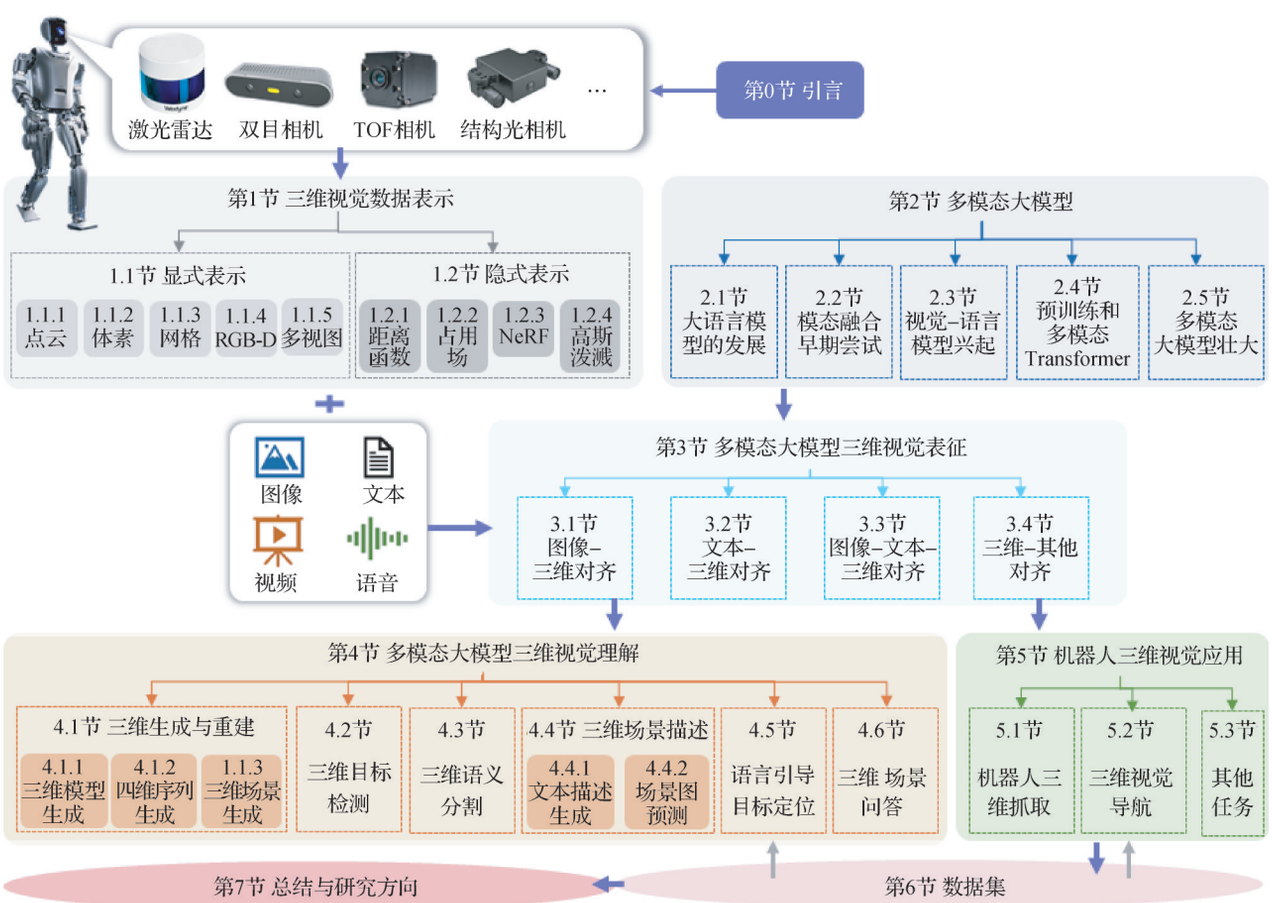

图 2　本综述总体架构

Fig. 2　The overall framework of this review

1 三维视觉数据的基本表示

机器人感知和理解真实三维世界的基础是获取三维数据，而三维数据表示了自然世界中物体和场景的三维几何信息。三维数据具有多种表示形式，如点云、体素、网格、符号距离函数、占用场和高斯泼溅等。每种表示涉及具体的三维视觉任务，都有其独特的优势和局限性。机器人利用各种传感器和三维成像技术来捕获三维数据，并根据任务需求选择最合适的数据表示形式。为了揭示不同表现方法所展现的三维形状特征和属性值，本文将三维数据的表示方法划分为显式表示和隐式表示，并采用对应表示方法重建3D物体，如图3和图4所示。

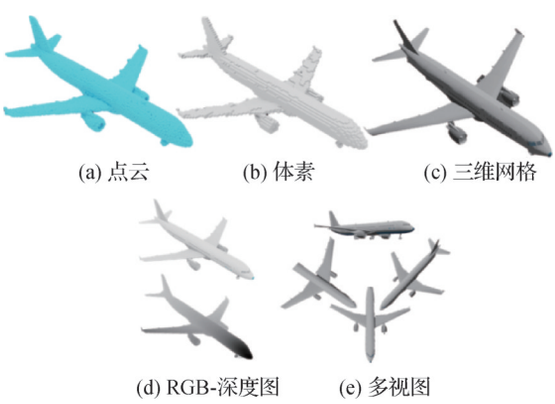

图3 三维视觉数据显式表示

Fig. 3 Explicit representation of 3D visual data
((a) point clouds; (b) voxels; (c) 3D mesh; (d) RGB-D images; (e) multi-view images)

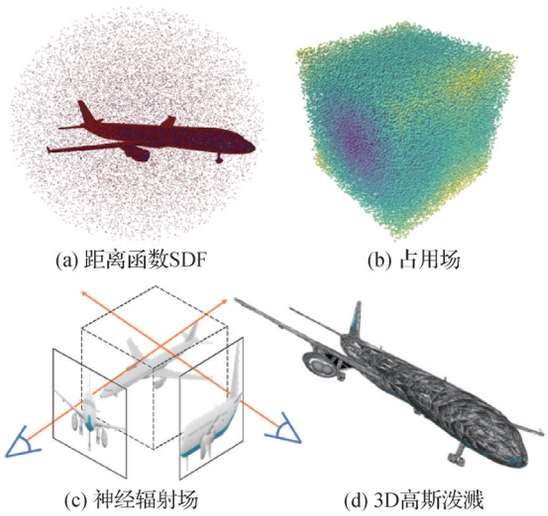

图4 三维视觉数据隐式表示

Fig. 4 Implicit representation of 3D visual data
((a) distance funciton SDF; (b) occupancy field; (c) neural radiance field; (d) 3D Gaussian splatting)

1.1 显式表示

1.1.1 三维点云

点云作为三维空间中大量离散点的集合，承载着三维空间坐标信息，还可以包含颜色、类别和强度等附加信息。机器人获取点云数据的主要途径包括使用激光雷达(light detection and ranging, LiDAR)设备和RGB-D相机等深度传感器。激光雷达通过发射激光脉冲并测量其反射回来的时间(time of flight, TOF)来获取距离信息。RGB-D相机则利用结构光或TOF技术来同时捕获场景的颜色和深度信息。此外，还可以采用TOF相机、结构光相机以及双目相机系统来捕获点云数据。深度学习技术近年来取得了显著进展，但由于点云的独特性质，如无序性，使其难以直接提取特征。因此，Qi等人(2017a)提出PointNet，利用多层感知机和最大池化来学习点云特征并聚合全局信息。为了更好地获取局部几何信息，Qi等人(2017b)进一步提出PointNet++，将点云分层并多尺度提取点云局部上下文和整体特征。然而，点云数据还具有旋转不变性，这给深度学习带来了新的挑战。为了解决这个问题，Hong等人(2023a)提出稳定网络输出的自适应闭环系统ACL-SPC框架，Yu等人(2023a)则提出基于Transformer的RoITr架构，两者都增强了点云旋转变化的鲁棒性。为了提升点云的建模能力，Guo等人(2021)提出点云Transformer(point cloud Transformer, PCT)架构。然后，一些研究工作(Wang等，2022a；Qian等，2022；Deng等，2023)基于PCT架构，继承并迭代了新的点云深度学习网络，以生成高维特征向量。同时，还有学者提出训练点云Transformer的新范式。Point-BERT(point bidirectional encoder representations from transformer)(Yu等，2022b)引入了启发于BERT的掩蔽点建模(masked point modeling, MPM)任务，而Point-MAE(point masked autoencoder)(Pang等，2022)则采用了掩蔽自编码器框架。两者通过学习未遮蔽点的潜在特征以预测被遮盖的点云部分，有效提高了点云Transformer的预训练效率。最近的研究还探索了应用于点云分析的新方法，Zhang等人(2023d)首次提出仅由不可学习组件构成的非参数网络Point-NN(point non-parametric network)，实现了

无需训练的3D点云分析。Liang等人(2024)则介绍了PointMamba,作为第一个用于点云分析的状态空间模型,通过采用线性复杂度算法显著降低了计算成本,同时保持了全局建模能力。紧接着,Han等人(2024c)提出Mamba3D,进一步引入局部规范池化和双向状态空间模型来提升局部特征提取能力,实现在多个点云分析任务的卓越性能。

点云以其精确的几何信息表征能力,为构建详尽的三维模型提供了数据支撑,增强了物体识别的准确性,并为机器人和自动驾驶车辆的环境感知、路径规划和障碍物检测提供了强有力的技术支持,从而提升了系统的整体性能和安全性。

1.1.2 体素

体素作为三维空间中规则排列的立方体网格单元,其构建三维数据的方式类似于二维图像的像素,主要用途是表征和存储三维形状信息,如几何占用等(Xu等,2022;Shim等,2023)。由于点云数据难以直接处理,研究者通常结合卷积神经网络(convolutional neural network,CNN)来体素化点云。例如,Maturana和Scherer(2015)提出的VoxNet,通过3种网格划分策略将点云转化为体素,并运用三维卷积提取特征以进行模型分类。基于VoxNet的工作,如3D ShapeNets(Wu等,2015)和3D-R2N2(Choy等,2016),通过占用网格体素化,实现了三维数据体素化的多种形式。然而,网格体素化增加了网格维度,可能导致分辨率和纹理信息的损失。为解决此类问题,研究工作(Zhang等,2022a;Xu等,2023c)引入了稀疏体素网格和动态体素网格优化等技术。此外,针对传统体素化表示在分辨率和占用率上的限制,诸多研究(Wang,2023;Zhou等2023)采纳了基于八叉树的表示方法进行三维形状建模。在整合不同技术优势方法上,PV-RCNN(Shi等,2020)及其增强版PV-RCNN++(Shi等,2023)结合了3D体素CNN在生成目标区域的能力与PointNet在提取关键点方面的专长,从而提升了目标检测的性能和效率。

体素因其规则的网格结构,适用于体积渲染与可视化。其与3D卷积网络的兼容性在3D对象的识别与分类任务中表现突出。此外,体素能与点云互相转化,为点云预处理开辟新途径,并在三维重建和语义分割领域发挥重要作用。

1.1.3 三维网格

网格是三维空间中由顶点、边和面构成的集合,以顶点相连的多边形作为基本单元。这些多边形网格通过点、边和面的连接勾勒出三维物体与场景的轮廓,既展现了元素的拓扑关系(Li等,2023d),也包含了几何属性的度量信息(Tang等,2023a)。由于网格明确表达了表面点间的连接,便于点间关系建模,且表示方式紧凑,许多研究(Lyu等,2023;Sella等,2023;Zhang等,2024a)将其应用于点云或体素数据的转换,以重建三维形状。然而,处理网格数据需兼顾顶点与边缘信息。Feng等人(2019)推出的MeshNet利用边面信息构建对称函数,并设计模块聚合面片信息。MeshCNN(Hanocka等,2019)以及MeshNet++(Singh等,2021)通过定义网格边缘的卷积层和池化层直接处理网格数据。另外,为了简化三维卷积操作,研究工作(Sella等,2023;Kim等,2023b;Li等,2024d)将网格数据转换为点云、体素或图等其他类型,其中Kim等人(2023b)创新地提出采用图神经网络生成非均匀网格的GMR-Net。此外,Gupta等人(2023)介绍了3DGen,该模型使用三平面变分自编码器(variational auto-encoder,VAE)提取纹理网格特征,经条件扩散模型生成面片特征。

三维网格以其对物体表面精确的细节描述,成为高质量渲染、机器人抓取和交互等应用的优选数据表示形式。此外,三维网格允许在复杂表面进行有效的纹理映射和形状分析,赋予其在物体位姿估计和增强现实领域的突出优势。

1.1.4 RGB-D数据

RGB-D图像是一种结合色彩与深度信息的结构化图像表示方式,其中每个像素包含红、绿、蓝3通道的颜色信息以及单通道的深度信息。RGB-D图像的生成得益于深度相机的发展和普及,例如,Kinect等深度相机(Tychola等,2022),通过结构光、TOF、LiDAR、立体相机系统或单目深度估计技术(Rajpal等,2023;Shao等,2023),实现对场景深度信息的快速获取和处理,从而适用于需要快速响应的应用场景。在RGB-D数据处理领域,Xing等人(2019)为了提取深度图像中的互补信息,使用HHA(horizontal disparity, height above ground, and the angle the pixel's local surface normal makes with the inferred gravity direction)编码处理RGB-D数据,将深度信息转换为三通道编码。然而,直接融合深度信息至RGB框架或简单整合双模态数据可能导致性能降低(Chen等,2021b)。为此,CANet(co-attention

network)(Zhou等,2020)通过共同注意力机制自适应融合颜色和深度特征,GLPNet(global-local propagation network)(Chen等,2021b)引入局部与全局上下文融合模块,实现深度与颜色信息的逐元融合。而Wu等人(2023a)设计的VirConvNet结合随机体素丢弃和抗噪声子流形卷积机制,高效融合了RGB图像与LiDAR数据。同时,Zheng和Gao(2024)针对RGB-D数据的大量内存占用问题进行探索,提出双分支RGB-D图像压缩框架,利用YUV域采样与模态内外注意力机制消除冗余,实现率失真优化。区别于其他3D数据表示方式,RGB-D图像引入深度信息,极大增强了计算机视觉系统的场景理解和分析能力,进而实现对人类立体视觉的模拟,尤其在低对比度和复杂背景的场景中提供了创新的检测方法(Sun等,2023a;Wang等,2024c)。随着注意力机制在深度学习领域的普及,许多研究(Goyal等,2023;Piekenbrinck等,2024)将其应用于RGB-D数据处理,为机器人三维视觉任务带来显著进展。

RGB-D数据通过结合颜色信息与深度信息,为机器人提供了全面的视觉感知能力,使其能有效进行实时三维场景重建、避障导航和路径规划等任务。此外,RGB-D相机连续捕获的深度信息增强了机器人在动态和复杂环境中的适应性和交互能力。

1.1.5 多视图几何

视图通过坐标变换将三维模型呈现于二维平面。基于视图的多视图几何从不同视点将3D对象渲染为2D图像,并使用2D卷积处理。这些视图数据可通过传统相机或模拟人眼视觉的双目相机采集,每个视图都表示特定视角下的二维特征。多视图几何不依赖复杂的三维特征,并且视图数据丰富,能有效利用先进网络架构。Su等人(2015)提出的多视图卷积神经网络(multi-view convolutional neural network,MVCNN)开创了三维模型多视图卷积神经网络,通过多视角投影生成视图,并采用最大池化融合特征,促进了该领域的快速发展。然而,最大池化仅保留最大值而忽略了其他信息。因此,Yu等人(2018)介绍了聚合局部卷积特征的MHBN,Han等人(2019)提出利用注意力机制提取全局特征的3DViewGraph。此外,研究表明(Wei等,2020;Li等,2023b),MVCNN忽视了视图间的特征联系。为此,有研究致力于挖掘视图间的相关性并从多尺度提取视图特征(Wei等,2023b),进而增强视图的旋转鲁棒性。例如,Xie等人(2023)推出的X-View通过平移原点,然后模拟不同视角来突破视角约束。在三维理解领域,Hong等人(2023b)提出一种从多视角图像学习三维概念并进行推理的方法。该方法使用神经场学习紧凑表示,并将其绑定到3D空间,通过神经推理运算符进行可视化推理。

多视图几何数据分析多个视角的图像信息,为三维场景提供了详尽的空间描述,在视觉导航、动态场景理解和增强现实等领域展现了独特优势。此外,作为SLAM技术的关键组件,多视图几何使机器人能够利用连续的图像序列实现环境地图的构建和自我定位。

1.2 隐式表示

1.2.1 距离函数

距离函数(signed distance field,SDF)是一种表征三维空间点与最近表面点间距的水平集表示法(Osher等,2004),分为有符号和无符号两种形式。有符号距离函数通过存储带符号的距离值来区分内外结构,擅于构建连续密闭表面,并常使用均匀体素网格实现离散化表示,进而能够描绘任意拓扑结构(Yang等,2023a;Yenamandra等,2024)。近年来,符号距离函数在深度学习形状分析领域受到广泛关注。例如,Park等人(2019)提出利用自编码器学习的DeepSDF,而Mittal等人(2022)推出了以VQ-VAE编码的AutoSDF。然后,Chou等人(2023)介绍了Diffusion-SDF,该模型将符号距离函数用于三维形状表征,并通过坐标多层感知机(multilayer perceptron,MLP)隐式编码物体表面。但由于符号距离函数仅限于表示封闭对象,因此许多研究(Chibane等,2020a;Long等,2023;Liu等,2024b)提出基于无符号距离函数的方法。Long等人(2023)提出的NeuralUDF(neural unsigned distance function)将无符号距离函数融入体积渲染,基于密度函数来处理任意拓扑结构。NeUDF(Liu等,2024b)则进一步优化了渲染权重函数和采样策略,并解决表面方向模糊问题,仅凭多视角监督即可重建复杂表面。

1.2.2 占用场

占用场是一种概率模型,用于描述三维空间中物体的分布。它通过记录每个点被曲面占用的概率来呈现物体分布,因此在处理不确定性和噪声方面具有鲜明优势。此外,占用场还能处理来自激光雷

达、摄像机以及超声波等多种传感器的数据融合问题（Chibane 等，2020b；Peng 等，2020；Chatzipantazis 等，2023）。Mescheder 等人（2019）的研究开创性地提出占用网络概念，其核心在于利用神经网络对连续的占用函数进行预测，以此编码物体的形状信息。该网络为空间内的每个点赋予一个占用概率值，进而采用多分辨率等值面提取算法（multiresolution isosurface extraction，MISE）将概率数据转换成三维网格形式。这一方法的应用例证之一是 IF-Nets（Chibane 等，2020b），它接受体素或点云数据作为输入，然后生成具有任意分辨率的表面模型。随后，Peng 等人（2020）进一步提出集成局部信息以预测占用概率的卷积占用网络，增强模型处理复杂几何形状的能力。Chatzipantazis 等人（2023）提出的 TF-ONet 模型引入了基于 SE（3）群的等变性注意力机制，提升模型处理 3D 形状的效率和准确性。为了优化稀疏数据的处理过程，Zhang 等人（2023f）最近提出的 OccFormer 模型，通过创新编码和注意力机制的应用，显著提高了静态及动态物体表面重建的保真度与效率。与此同时，为了满足现有 3D 表示的高计算量需求，Feng 等人（2022）介绍了傅里叶占用场（Fourier occupancy field，FOF），通过傅里叶级数将 3D 占用场分解为 2D 向量场，保留了 3D 拓扑和空间关系，并整合人体参数模型作为先验知识。FOF 基于拉普拉斯约束和自动机的不连续性匹配器扩展为 FOF-X，避免了由纹理和光照引起的性能下降，进一步提升了重建质量和鲁棒性。相比于其他表示方式，FOF 具有高效灵活性，在均衡高质量结果和快速处理方面效果优异。

1.2.3 神经辐射场

神经辐射场（neural radiance fields，NeRF）由 Mildenhall 等人（2020）提出，它是一种利用多层感知机进行视图合成和三维隐式空间建模的技术。在 NeRF 中，每个输入对应一个五维坐标，输出是该坐标的密度和幅度值。Zhang 等人（2020）深入分析 NeRF，提出 NeRF++，解决了形状—辐射模糊性、近场模糊性和无界场景参数化问题。为了降低 NeRF 场景校准所需的大量视图和计算时间，Yu 等人（2021）提出的 PixelNeRF 从少量图像合成三维场景，Lin 等人（2023b）则提出基于 NeRF 的视觉变换器（Vision Transformer，ViT），以减少对大量输入图像的依赖。同时，NeRF 为三维数据处理和渲染提供了新的方式。Chen 等人（2021a）提出的 MVSNeRF 利用三维卷积重建神经编码体积，Xu 等人（2023a）结合几何先验和 MLP，提出 NeRF-Det 来实现三维形状估计。此外，Barron 等人（2021）介绍的 MipNeRF 改进了 NeRF 的多尺度处理能力。Metzer 等人（2023）推出的 Latent-NeRF 结合扩散模型指导三维物体生成。其他工作研究使用深度学习技术和蒸馏算法来提升视图质量（Gu 等，2023a；Ye 等，2023）。

1.2.4 高斯泼溅

3D 高斯泼溅（3D Gaussian splatting，3DGS）是由 Kerbl 等人（2023）提出的一种基于辐射场的场景表示方法，它采用各向异性的三维球体和三维高斯分布建模辐射场。这种方法的表征能力和建模效率优异，但通常需要数百万参数且占用大量内存。为了解决内存效率问题，Lu 等人（2023c）提出受网格 NeRF 启发的内存高效模型 Scaffold-GS。另有一系列研究（Navaneet 等，2024；Girish 等，2024；Lee 等，2024）通过向量量化压缩参数，减少存储空间并提高渲染质量。3D 高斯泼溅的性能依赖于初始化稀疏点的数量和精度。为应对样本稀疏挑战，Chung 等人（2024）引入深度正则化和无监督几何平滑性约束，而 Zhu 等人（2024b）提出的 FSGS 方法通过扩展运动结构恢复（structure from motion，SFM）初始化的三维高斯分布来逐步填充场景。同时，Xiong 等人（2024）推出的 SparseGS 和 Li 等人（2024a）提出的 DNGaussian 均旨在从少量样本进行视图合成。针对高斯泼溅过程出现的锯齿和微影问题，许多研究进行了相关探讨（Yu 等，2023c；Zhang 等，2024d）。另有研究（Yang 等，2024c；Xie 等，2024）致力于将 3D 高斯泼溅从静态场景扩展到四维（four dimensions，4D）场景。

隐式 3D 表示方法彼此之间具有互补性，并提供不同的技术优势。符号距离函数能够精确地定义物体表面，常用于精细的表面重建任务。占用场则提供空间中某点被物体占用的概率信息，更适合于宏观层面的应用，如路径规划与避障。神经辐射场通过深度学习生成连续的 3D 场景表示，擅长于新视角合成，能够产生高质量的 3D 场景渲染。而高斯泼溅专注于提高 3D 形状表示和渲染的效率，尤其适用于实时渲染场景。这些方法根据其特性，广泛应用于机器人 3D 视觉任务中。

2 多模态大模型

2.1 大语言模型的发展

语言模型的研究可追溯至20世纪60年代，早期的研究工作（Bar-Hillel等，1960）主要依赖于基于规则的方法。然而，这种方法在处理语言的复杂性和多样性方面逐渐暴露出局限性。为了克服这些限制，研究者开始探索更为灵活和强大的方法。在随后的几十年里，统计模型逐渐成为主流，代表性的成果包括N-gram模型（Brown等，1992）和隐马尔可夫模型（Fine等，1998）。这两种统计模型在文本预测和生成方面展现出了鲜明的优势。进入21世纪，随着计算能力和算法的进步，研究者致力于开发能够更好地捕捉序列数据长期依赖关系的模型，例如Schuster和Paliwal（1997）提出的循环神经网络（recurrent neural network，RNN）。并且其变种长短期记忆网络（long short-term memory，LSTM）（Hochreiter和Schmidhuber，1997）的引入，有效解决了传统RNN在处理长序列时存在的梯度消失问题。此外，词嵌入技术如Word2Vec（Mikolov等，2013）的发展，使得机器能够以向量形式高效地表示词语，显著增强了模型对语义的理解能力。上述一系列技术进一步为Transformers架构（Vaswani等，2017）的诞生创造了条件。Transformers凭借其独特的自注意力机制，极大提高了处理自然语言任务的效率，使模型能高效处理大规模数据。Transformers及后续模型的规模和参数数量的显著增长，标志着大语言模型（large language model，LLM）的一个重要发展阶段。大语言模型是一类基于Transformer架构的语言模型，其参数规模通常达到数十亿甚至数百亿。LLM采用自注意力机制来捕捉输入序列中的长程依赖关系，并引入了位置编码、解码器和编码器技术。为了完成广泛的语言任务，LLM通常在大规模互联网数据集上通过文本生成任务进行训练。例如，GPT-3（generative pre-trained transformer-3）（Brown等，2020）、LLaMA（large language model meta AI）（Touvron等，2023）和GPT-4（OpenAI，2024）等模型以卓越的性能持续推动自然语言处理领域的前沿发展。LLM的一个显著特征是其涌现能力，这是随着语言模型规模和复杂性的增加而出现的新能力，如上下文学习（Brown等，2020）、指令跟随（Ouyang等，2022）和思维链推理（Wei等，2022），这些涌现能力在自然语言处理（natural language processing，NLP）领域已有广泛研究。同时LLM也通常用做视觉推理系统的控制器、决策器或语义修饰器（Yin等，2024a）。使用LLM开发具备人类智能的自主智能体已成为目前研究的热点领域（Liang等，2023b；Hong等，2024）。具身语言模型（Driess等，2023）更是结合现实世界的传感器数据与语言模型，直接建立文本和感知信息间的联系。

2.2 模态融合的早期尝试

在生物学领域，多模态表现为生物体利用多感官系统感知外部环境的能力。而在计算机科学领域，多模态涉及数据或信息的多种表达形式，如图像、文本、语音和视频等。单一模态数据因其信息维度有限，通常仅限于处理特定任务，尤其面对复杂的应用场景时表现不足。然而，多模态学习通过聚合多元数据源，能提供更准确、鲁棒的解决方案。Huang等人（2021）的研究表明，潜在的表征空间质量直接影响多模态学习模型的效能，且在数据充足时，模态多样性可提升表征估计的精确度。模态融合的早期工作可以追溯到1970年代提出的典型相关分析（canonical correlation analysis，CCA）、平行因子分解（parallel factor analysis，PARAFAC）以及其他张量分解方法。Slaney和Thiébaux（2001）提出根据文本指令重排彩色积木块的积木世界问题，尽管它并非基于深度学习技术，但这一早期尝试将视觉与语言结合，标志着视觉语言模型的启蒙。

2.3 视觉—语言模型的兴起

近年来，随着CNN和RNN等代表性深度学习技术的提出，多模态模型研究领域取得一系列突破性进展，并且推动视觉—语言任务处理方式的根本性变革。在过去的10年内，视觉理解与生成任务的处理主要采用早期融合和晚期融合两种策略。晚期融合利用联合嵌入技术，将不同的模态数据嵌入至一个共享的语义空间，然后测算其相似度，以促进跨模态的学习和理解（Mao等，2015；Antol等，2015；Vinyals等，2015）。早期融合策略则是在模型早期阶段将各种模态特征结合，并直接基于融合后的特征表示进行相似性分数预测（Yu等，2017；Wang等，2019；Xu等，2019）。除了视觉和语言模态，研究者还探索了涉及音频、语音和三维数据等其他模态的模型构建（Qi等，2016；Arandjelovic和Zisserman，

2017)。总而言之,研究者为了解决传统单一模态模型在处理多模态数据时理解和生成多模态内容的难题,提出整合图像和文本内容的视觉语言模型(visual language model, VLM)。VLM 的核心优势在于通过学习图像和文本之间的对应关系来捕获更丰富的语义信息。并通过多模态特征提取与融合、跨模态对比学习和多任务学习等方式促进不同模态的数据表征,这些方法具体表现于第3、4节,在图像字幕生成、视觉问答(visual question answering, VQA)以及图像检索等领域展现出了卓越能力。

2.4 预训练和多模态 Transformer

随着多模态模型领域的不断演进,预训练方法已成为一种重要的研究趋势。此方法首先在大规模数据集上预训练多模态模型,然后针对特定任务进行微调,已在多个应用领域证明了其提升模型性能的有效性。受诸如 BERT(Devlin 等, 2019)、T5(Raffel 等, 2020)和 GPT(Brown 等, 2020)等预训练 NLP 模型的启发,研究人员进一步开发了能够处理文本、图像、音频和点云等多模态输入的 Transformers 模型(Zhan 等, 2024; Wang 等, 2024d; Li 等, 2024e)。而 VLM 领域,Radford 等人(2021)提出对比语言—图像预训练模型(contrastive language-image pre-training, CLIP),其由两个核心组件构成:视觉编码器和文本编码器。CLIP 最大化给定图像的特征向量与相应的文本描述的特征向量之间的相似度来训练模型。且在推理阶段替换文本描述中的类别标签为潜在的候选标签,并搜索与输入图像特征最匹配的文本描述,以实现零样本迁移性。CLIP 后续更是作为基础模型,衍生出众多文生图大模型。Bao 等人(2022)受 BERT 模型的启发,提出 BEiT,一个以掩码图像模型为预训练任务的生成式自监督预训练大模型。但随着数据集的不断扩充,预训练模型面临着日益增长的复杂性和计算需求。为此,Wang 等人(2022b)提出 SimVLM,该模型采用大规模弱监督和单一的前缀语言建模目标进行端到端训练,证明了简洁的预训练框架也能高效处理视觉语言数据。Alayrac 等人(2022)推出了 Flamingo,它结合了两个互补的预训练模型,并引入了新的结构组件以链接这些模型在预训练期间积累的知识,从而实现跨任务的零样本或少样本迁移。此外,还有基于对比损失和描述损失训练的 CoCa(Yu 等,2022a),通过模块化接口调用已有大模型的 PaLI(Chen 等,2023f)等预训练模型。预训练过程不仅使模型能对齐不同模态信息,还增强了模型编码器的表征学习能力。这些模型旨在通过这种方式,创建一个跨任务的泛化系统,而无需特定任务的训练数据。总而言之,多模态模型的壮大为众多领域带来了前所未有的机遇。例如,DALL-E(Ramesh 等,2021)基于 GPT-3 架构,实现了从文本描述生成图像的成果;Stable Diffusion(Rombach 等,2022)和 ControlNet(Zhang 等,2023b)则利用基于 CLIP 和 UNet 的扩散模型,根据文本提示生成图像;BLIP(bootstrapping language-image pre-training)(Li 等,2022b)通过多模态混合编码器—解码器框架和文本描述生成与过滤方法,生成图像对应的文本描述。多模态模型在机器人技术(Wang 等,2024b)、医疗卫生(Thirunavukarasu 等,2023)和人文艺术(Ko 等,2023)等多个领域更是具有潜在应用价值。

2.5 多模态大模型的壮大

多模态预训练领域的研究已取得显著进展,整体发展流程如图5所示,并且这些模型持续不断地突破下游任务的性能界限。因此,多模态大语言模型(multimodal large language model, MLLM)已成为一个重要的研究领域,其经典框架如图6所示。输入的多模态数据首先通过模态编码器编码成统一的模态特征,然后经过特征投影、查询和融合转换器处理,将模态特征转换为适配 LLM 输入的形式,并将其和文本特征送入 LLM 进行语义理解和信息处理。最终,LLM 生成文本及利用生成器获得所需的多模态内容。例如,OpenAI 提出的 ChatGPT 和 GPT-4 系列模型,通过融合多模态数据和 LLM 的能力来执行跨越多种模态的任务。最近,GPT-4V 和 Gemini(Gemini Team Google, 2024)展示了令人印象深刻的多模态理解和生成能力,激发了众多学者对 MLLM 的研究热情。MLLM 的研究主要集中在多模态内容理解与文本生成,包括图像文本理解生成(Li 等,2023a; Zhu 等,2023b; Koh 等,2023)、视频文本理解生成(Rubenstein 等,2023)、音频文本理解生成(Chu 等,2023)以及三维数据文本理解(Hong 等,2023c; Li 等,2024e)。许多研究团队通过将 LLM 与外部工具结合,实现了接近于任意模态输入、输出的 MLLM(Shen 等,2023; Zhan 等,2024)。同时,为了减少级联系统中错误传播的问题,CoDi-2(Tang 等,2023b)和 ModaVerse(Wang 等,2024d)等模型开发了端到端的 MLLM。在 MLLM 中,关键技术和应用主要为多

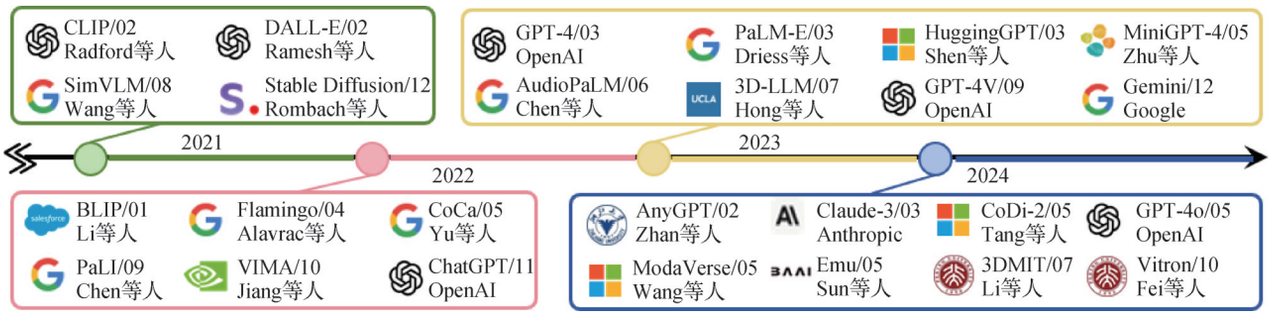

图5　代表性多模态大模型发展时间轴

Fig. 5　Timeline of representative MLLM

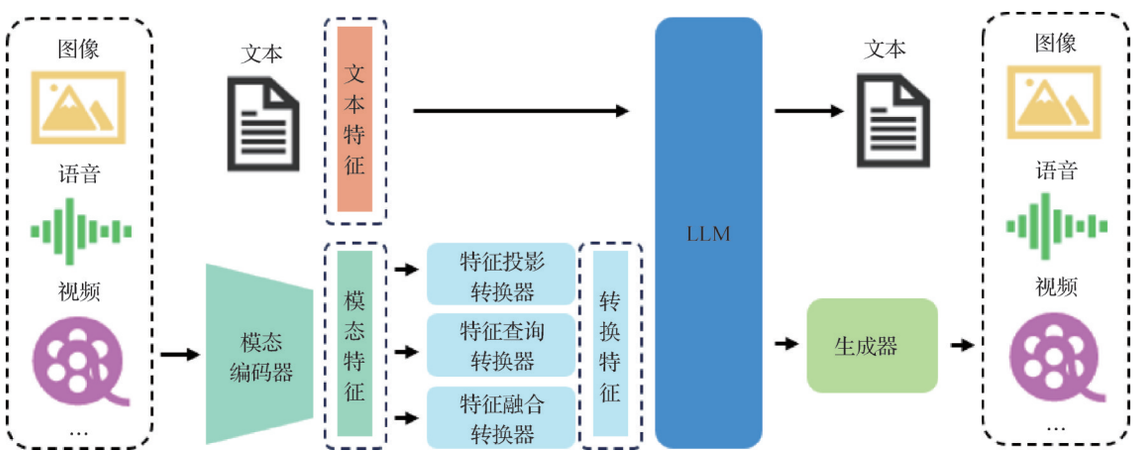

图6　多模态大模型的典型框架(Yin等,2024a)

Fig. 6　Typical framework of MLLM(Yin et al. ,2024a)

模态指令微调(Zhu等,2023b;Liu等,2023a)、多模态上下文学习(Yang等,2023d;Lu等,2023a)、多模态思维链(Ge等,2023;Stechly等,2024)以及LLM辅助的视觉推理(Shen等,2023;Lu等,2023a)。相较于传统的LLM,MLLM更符合人类感知世界的方式,如Mu等人(2023)提出的EmbodiedGPT,提供了更友好的用户界面并支持更广泛的下游任务。而OpenAI最新推出的GPT-4o代表了深度学习在实际应用方面的最新进展,它为诸多任务提供了灵活、高效和安全的多模态智能解决方案。同时,研究社区中的Fei等人(2024b)开发了名为Vitron的统一像素级通用视觉MLLM,能够处理从视觉理解到视觉生成和从低层次到高层次的一系列视觉任务。

根据第1节对3D数据基本形式的描述和第2节对MLLM的介绍,目前已经涌现出许多能处理3D模态输入的MLLM。这些模型不仅继承了通用MLLM的卓越性能,而且在3D数据处理方面展现出独特的适应性和高效性,为实现跨模态通用人工智能(artificial general intelligence,AGI)奠定了坚实的技术基础和理论支持。

3　多模态大模型三维视觉表征

本节将深入探讨图2所示的三维视觉数据表示如何与其他模态数据对齐,并利用LLM强大的语义理解能力来匹配和生成不同模态信息。通过这一过程,旨在展示如何整合不同模态的数据,对齐多模态大模型的三维视觉表征,以实现对三维视觉的全面理解。

3.1　三维模型—图像对齐表征

研究最初启发于CLIP,采用图像和文本编码器为每对图像—文本生成统一表征,然后对齐模态表征。由此,PointCLIP(Zhang等,2022b)首先对点云进行编码,将点云投影到多视图深度图中,端到端聚合多视图,利用CLIP进行零样本三维模型分类。在此基础上,PointCLIP V2(Zhu等,2023c)提出形状投

影模块生成更逼真的深度图，以缩小投影点云与自然图像之间的域差距。而其他研究更专注于对齐图像模态。CrossPoint（Afham 等，2022）利用第 2.3 节提及的跨模态对比学习方法，即在 3D 点云和 2D 图像间进行对比学习，减小特征空间中两种模态特征的距离。CLIP2Point（Huang 等，2023c）引入新的深度渲染方法将点云渲染为更易对齐的 RGB 图像和深度图像，整合了跨模态学习以增强深度特征，并采用内模态学习来提高深度聚合的不变性。然而，与对齐二维图像不同，三维数据的获取难度较大，加之大规模三维数据集和广泛类别的三维模型的数量均有限，这些因素都使得三维视觉预训练模型的训练充满挑战。此外，上述模型忽略了原始的 3D 结构信息，仅适用于分类任务。因此，CLIP-FO3D（Zhang 等，2023a）尝试改变投影方向，将多视图图像特征投影到点云上，并使用特征蒸馏训练 3D 模型，直接将 CLIP 的视觉特征空间转移到 3D 数据中，不需要任何形式的监督。EPCL（efficient point cloud learning）（Huang 等，2024e）更是提出高效的点云学习器，用于直接训练高质量的点云模型，通过在语义上对齐图像特征和点云特征来连接 2D 和 3D 模式，而无需配对 2D-3D 数据。针对现有基准模型对大规模注释 3D 点云数据的过度依赖问题，MvNet（Peng 等，2024）提出一种新的多视图视觉融合网络，将 3D 点云编码为多个不同视图的图像特征。并开发多视图提示融合模块，弥合了三维点云数据与二维预训练模型之间的差距。然后，导出一组 2D 图像提示符，以更好地描述用于少量 3D 点云分类的大规模预训练图像模型的先验知识。此外，Hess 等人（2024）介绍的 LidarCLIP，使用图像—激光雷达对，监督具有图像 CLIP 嵌入的点云编码器，有效地将文本和激光雷达数据以图像域为中介相关联，有针对性地解决不利的传感器条件下的场景检测。另有研究使用三模态预训练框架来集成更全面的视觉表征，如图 7 所示。

ULIP-2（Fei 等，2024a）输入点云数据，利用 Token 级 Transformer 自编码器，经过 Token 级掩码重建任务来获取离散 Token，进而提取点云特征。同时，初始点云也输入与 Token 级共享模态表征的点级 Transformer 自编码器，经过点级掩码重建任务和可微分渲染技术，生成深度图并提取特征。而最后一个分支，点云经 RGB 渲染器渲染成 RGB 图像，并嵌入特征空间，形成三模态的特征空间。最后，通过 3 种模态特征间的对比学习训练，实现了点云、RGB 图和深度图的无缝对齐。

通过将三维数据投影成 RGB 图像和深度图像，并以它们作为桥梁，连接视觉预训练模型，将 2D 特征知识转移到 3D 特征，并通过对比学习对齐多模态特征空间，以提升三维视觉任务的性能。这作为一个解决因大规模三维数据带来的 3D 预训练大模型能力较弱问题的有效方式，有着巨大的潜力。尽管如此，面对开放词汇和长尾分布数据集，三维模型与二维图像间的对齐精度仍需进一步提升。

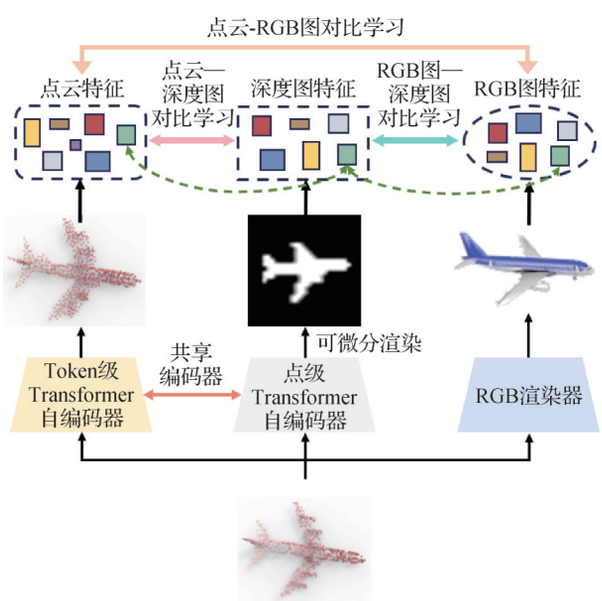

图 7　3D 点云数据映射并与 RGB 图像和深度图像对比学习框架（Fei 等，2024a）

Fig. 7　Projection and alignment of point cloud, RGB image, and depth image within a contrastive learning framework (Fei et al., 2024a)

3.2　三维模型—文本语言对齐表征

Point-BERT（Yu 等，2022b）开创了以自监督的方式，利用点云—文本预训练模型对齐点云和文本模态的方法。为了将 3D 特征融入 LLM，近期研究中，PointLLM（Xu 等，2024b）开辟了超越传统 2D 视觉数据的新路径，如图 8 所示。左侧图展示了 PointLLM 的模型架构，点云数据首先被点云编码器编码为点云词元，且用户提出的文本问题通过分词器被编码为文本词元。然后，点云词元和文本词元被共同输入映射器，映射器将词元转换为嵌入向量，送入 LLM 的特征空间，经由 LLM 处理以生成相应的

文本输出。右侧图展示了PointLLM在整合彩色点云和文本指令方面的具体应用。依据图中准确的对话示例，表明模型具备对三维物体特征的深入理解和常识推理能力。然而，PointLLM的模态对齐效果依赖于高质量的点云数据，为了进一步连接点云和语言模态，Huang等人（2024d）提出Text4Point模型。该模型采用二维图像作为跨模态融合的桥梁，通过对比学习在预训练阶段建立图像与点云的对应关系，并借助CLIP实现点云特征与文本嵌入的隐式对齐。此外，Text4Point引入了文本查询模块以整合语言信息至三维表示学习过程。在微调阶段，Text4Point能够在无2D图像标签的情况下，仅依赖语言指导来学习特定任务的三维表示。而ShapeLLM（Qi等，2024b）基于多视图图像蒸馏，设计并改进了3D编码器，增强几何形状的理解，在3D几何理解和语言统一的3D交互任务方面实现了最先进的性能。另一方面，Uni3DL（Li等，2023e）不依赖于多视图图像投影，设计了一个查询转换器用来学习与任务无关的语义，并基于3D视觉特征来屏蔽输出。Uni3DL通过统一网络结构的任务路由器，选择并生成针对不同任务需求的特定输出。然而，3D数据集的规模有限，存在如何使用低质量的点—文本特征来深入理解和生成3D对象的问题。为此，Qi等人（2023）构建了超过1 M个带注释样本的大型数据集，并提出GPT4Point模型，经此数据集训练，利用低质量的点—文本特征生成高质量的3D对象，同时保持几何形状和颜色的准确性。MiniGPT-3D（Tang等，2024）提出一种高效且功能强大的3D大语言模型（three dimensional-large language model，3D-LLM），仅需一台RTX 3090训练27 h即可实现多个SOTA（state-of-the-art）结果。MiniGPT-3D使用2D LLM的2D先验将3D点云与LLM对齐，利用2D和3D视觉信息之间的相似性，设计一种新的四阶段训练策略。它以级联的方式进行模态对齐，并引入了混合查询专家模块，以自适应聚合特征。

现有的3D—文本LLM主要在对象级别的数据集上训练，这限制了它们对复杂场景的深入理解。此外，这些模型更多关注静态3D对象的识别，而对动态对象的识别能力不足。期望通过结合3D表示处理方法与3D-LLM，充分发挥LLM的知识先验和泛化能力。

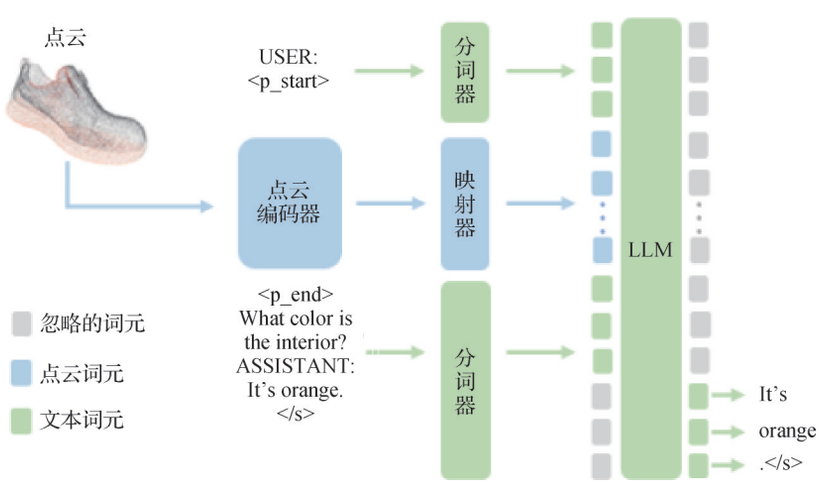

图8　理解彩色点云的多模态大语言模型（Xu等，2024b）

Fig. 8　MLLM for colored point cloud understanding（Xu et al.，2024b）

3.3　三维模型—图像—文本对齐表征

一种新兴的多模态表征学习范式是点云—图像—语言三模态表征（雷印杰 等，2024），它将3D点云、2D图像和自然语言融合在一起。ULIP（Xue等，2023）作为该领域的开创性工作，构建了包含3D点云、2D图像和自然语言描述的对象三元组。通过利用有限的自动合成三元组，学习与图像—文本空间对齐的3D表征空间，进行多模态预训练。在此基础上，Xue等人（2023）改进了ULIP对数据集元数据和类别名称的依赖性，提出简单而有效的三模态预训练框架ULIP-2，如图9所示。该框架通过从3D物体表面采样点云作为3D编码器输入。同时，从多个视

角渲染3D物体生成全方位图像，随后输入图像编码器来提取视觉特征。框架使用BLIP-2为每幅渲染图像自动生成描述性文本。为确保文本与视觉的特征匹配，还需CLIP对文本描述进行相似度排序。而在预对齐且冻结的视觉—语言特征空间，框架进一步对齐从3D编码器输出的点云特征，进而实现图像、文本和点云3种模态的一致性对齐。ULIP-2只需要3D数据作为输入，消除了手动3D注释的依赖，提供了无人工注释的可扩展多模态3D表示学习的新思路。随后，CG3D（CLIP goes 3D）（Hegde等，2023）提出一个基于对比损失的预训练框架。CG3D的训练数据源于ULIP，并进一步引入提示调整技术，将预训练视觉编码器的输入空间从计算机辅助设计（computer aided design，CAD）对象的渲染图像转移到自然图像，从而有效利用CLIP处理3D形状。

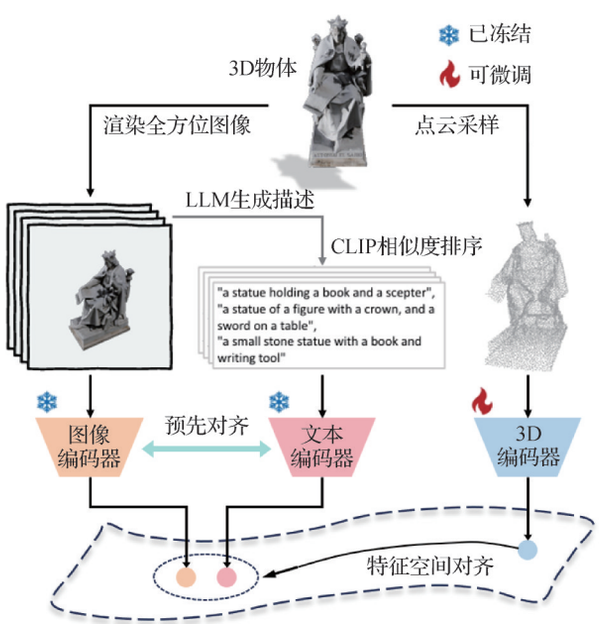

图9　对齐三维模型—图像—文本特征到同一空间（Xue等，2024）

Fig. 9　Alignment of 3D modal, image and text features into the unified space（Xue et al.，2024）

此外，OpenShape（Liu等，2023b）采用多模态对比学习框架用于学习文本、图像和点云的多模态联合表示。OpenShape的多模态嵌入通过编码丰富的视觉和语义信息，实现了细粒度的文本到3D、图像到3D交互。TAMM（triadapter multi-modal learning）（Zhang等，2024h）提出基于3个协同适配器的新型两阶段学习方法，以减少三模态表示学习过程中由于二维图像的域移位和每个模态的不同焦点带来的影响。CLIP图像适配器减轻了3D渲染图像和自然图像之间的域差距，双适配器将三维形状表示空间解耦为专注视觉属性和语义理解的两个互补的子空间，确保了全面有效的多模态预训练。Ji等人（2024）提出集成点云、文本和图像的综合方法JM3D。主要贡献包括结构化多模态组织器，通过多视图图像序列和分层文本树，增强视觉和语言模态的独立表示；联合多模态对齐，将语言理解与视觉表示相结合；通过高效的微调，将3D表示与LLM相结合，形成JM3D-LLM模型，在零样本3D分类和图像检索任务中取得出色成果。

上述模型通过三维模型—图像—文本特征空间对齐策略，有效提升三维特征的表征精度。然而，受限于样本数量不足和小范围预定义类别，现有模型对图像模态的利用不够深入，难以捕捉细粒度3D属性和特征。

3.4　三维模型—其他表征

音频、视频等模态信息与3D数据的表征对齐始于Guo等人（2023b）提出的Point-Bind，它是3D多模态表征对齐领域的一个创新工作，实现了点云、图像、文本、音频和视频5种模态的特征融合。如图10所示，3D编码器首先将3D点云编码为3D特征嵌入。同时，音频、文本和图像模态数据通过ImageBind（Girdhar等，2023）处理并嵌入到一个联合的特征空间。这些嵌入在特征空间通过对比损失函数与3D特征对齐，以提升3D视觉任务中的跨模态交互和精确理解。进一步地，Point-LLM采用参数高效的微调技术，将Point-Bind捕获的3D语义注入预训练的LLM中。这一过程不需要3D指令，但具有优越的3D和多模态问答能力，为3D点云在多模态应用中的扩展提供了新的视角。

OneLLM（Han等，2024b）提出统一的多模态编码器和渐进式多模态对齐策略，如图11所示。图像、音频和视频等8种不同模态数据通过通用编码器被转换为token序列，并利用可学习的模态切换器{model}，将不同长度的多模态输入token转化为固定长度。这些token随后被输入通用投影模块进行统一映射，最终输入LLM以实现多模态理解。OneLLM构建了一个能够处理广泛模态的统一且可扩展的编码器，为简化多模态编码做出了贡献。尽管如此，整个模态对齐过程采用渐进式的方式，这表

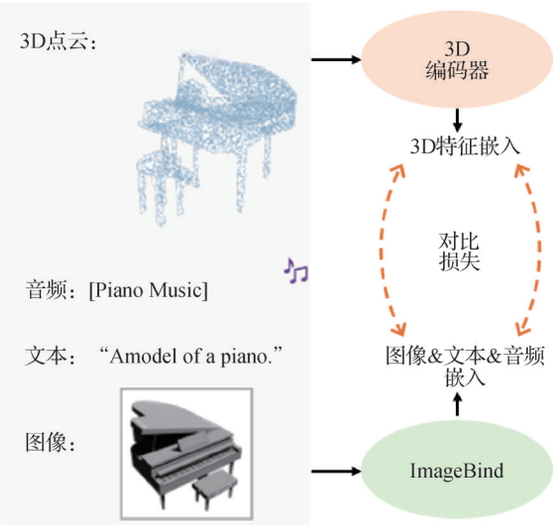

图 10　三维点云—图像—文本—语音多模态统一表征学习框架(Guo 等,2023b)

Fig. 10　Unified representation learning framework for 3D point clouds, images, text and speech(Guo et al. ,2023b)

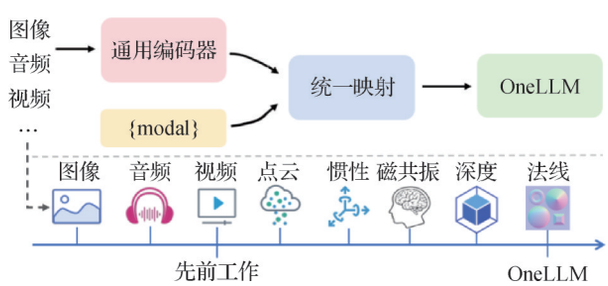

图 11　三维点云—图像—音频—视频—深度/法向图的统一表征学习框架(Han 等,2024b)

Fig. 11　3D unified representation learning framework for 3D point clouds, images, audio, video, and depth/normal maps(Han et al. ,2024b)

明模态对齐过程仍有优化空间。

4　多模态大模型三维视觉理解

第 3 节介绍了多模态表征对齐技术,这是优化 3D-LLM 的三维视觉理解能力的关键步骤。为了提升 3D-LLM 在三维视觉任务的性能,首先需要理解三维视觉任务。目前,3D-LLM 已经赋能了广泛的三维视觉任务,通过多任务学习方法,3D-LLM 能够在统一模型框架下同时学习多个相关任务来提高模型的泛化能力,这一点在第 2. 3 节有所提及。因此,本节介绍主要的三维视觉任务并展示如何利用 LLM 强大的语义理解能力来匹配和生成不同模态信息。同时,论述了 3D-LLM 在实际应用场景中的性能优势,为未来的研究和应用提供指导。

4. 1　三维视觉生成与重建

4. 1. 1　三维模型生成

三维建模是一个复杂且耗时的过程,通常依赖于手工设计的几何和外观特征,需要人工设定参数并借助特定设备处理。为了生成高质量和多样化的三维模型,需要面临两大挑战:1)图像或文本描述的空间与三维形状空间存在模态差异,难以获取准确的跨模态映射函数;2)三维模型的拓扑结构多样且复杂,难以转换成适合神经网络处理的表示形式。早期研究(Jain 等,2022;Sanghi 等,2022;Michel 等,2022)启发于 DALL-E 和 CLIP 等视觉—文本嵌入模型的文生图功能和零样本泛化能力,探索了不依赖文本—3D 数据集的预训练模型,仅利用文本自动生成 3D 模型。近期研究针对模态对齐问题,提出创新的对齐技术和多模态融合策略。ShapeGPT(Yin 等,2023)通过形状特定的 3D VQ-VAE 将三维形状离散化为 token,使得形状数据能与文本和图像共同构成多模态输入框架。GPT4Point(Qi 等,2023)则采用双流框架,通过 Point-QFormer 将点云几何特征与文本对齐,并输入到耦合的 LLM 和扩散路径。而 Michelangelo(Zhao 等,2024)提出一种 3D 形状生成前对齐方案,如图 12 所示。Michelangelo 包含形状图像文本对齐变分自编码器(shape image-text-aligned variational auto-encoder,SITA-VAE)和对齐形状潜在扩散模型(aligned shape latent diffusion model,ASLDM)两个模型。在 SITA-VAE 中,可学习的查询 token 引导 3D 形状编码器从输入点云中提取 3D 形状特征,并将特征映射到对齐的形状潜在空间。同时,图像和文本数据分别通过 CLIP 图像编码器和 CLIP 文本编码器提取特征,再经过特征对齐后,ASLDM 学习了从图像或文本特征到对齐形状潜在空间的概率映射函数。通过这种方式,3 种模态特征在形状潜在空间对齐,并基于 KL(Kullback-Leibler)散度正则化进行优化。最终,3D 形状解码器根据对齐空间的嵌入,以体素表征方式重建 3D 形状。上述模态对齐方法在特征空间对齐多模态特征,整合了语义信息,有效弥合不同模态间的领域差异,从而提高了 3D 形状生成的质量。而 Dream3D(Xu 等,2023b)将 3D 形状先验引入文本到图像扩散模型,强化了以文本为导向的 3D 重建。从层次特征提取角度,HyperSDFusion(Leng 等,2024)使用双分支扩散模型,引入双曲

文本—图像编码器来学习双曲空间中的文本顺序和多模态层次特征。另有研究利用分数蒸馏采样(score distillation sampling,SDS)的预训练扩散模型(Wang等,2023c),通过迭代优化过程,将文本或图像描述转换成高质量的3D模型,以实现跨模态转换。Direct2.5(Lu等,2024)更通过微调预训练的2D扩散模型构建2.5D扩散模型,不仅整合了3D扩散模型的数据处理和2D扩散模型的泛化性能优势,还填补了2D与3D扩散内容生成方法之间的空白。然而,基于SDS的扩散模型生成的3D模型可能有不真实与过饱和的效果,这促进了对三维模型的拓扑结构处理方法的研究。Fantasia3D(Chen等,2023a)和EXIM(Liu等,2023d)利用混合3D形状表示,拆分形状和颜色属性并独立处理。而Tsalicoglou等人(2024)聚焦于数据表示形式,提出以SDF为主干,扩展NeRF并微调网格纹理的Textmedh,以高度逼真的3D网格来模拟拓扑结构。值得注意的是,在应用创新方面,Yuan等人(2024b)通过程序合成来操纵LLM生成文本驱动的3D形状,Uni3d-llm(Liu等,2024a)在点云场景中利用LLM实现3D感知、生成和编辑任务。MLLM对三维重建领域的贡献不仅涵盖几何与外观,还扩展到纹理合成(Wei等,2023a;Lin等,2023a)、组合生成等任务(Jiang等,2023;Gao等,2024),开启了三维建模的新篇章。

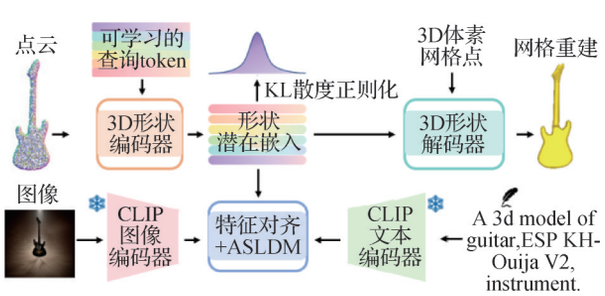

图12　三维模型生成(Zhao等,2024)

Fig. 12　3D model generation(Zhao et al.,2024)

除了3D场景物体的生成,3D数字人生成正逐渐成为三维视觉生成领域的一个热门方向。早期方法利用NeRF表示和可微分体积渲染从2D图像中学习3D形状(Gao等,2022;Weng等,2022),进而使用2D生成器进行3D人物生成。然而,这些方法计算成本高且生成分辨率有限。因此,涌现出了以EVA3d(Hong等,2022a)为代表的高分辨率生成模型,它们结合了2D人物生成和3D人物重建的先验知识。但由于2D人物数据集在不同姿势和观察视角的样本采集存在偏差,导致部分姿势和视角的数据缺失,难以生成全身拓扑几何。为了提高生成的鲁棒性,Chupa(Kim等,2023a)为人物生成两个法线图,通过双法线图优化来保持视图一致性。另一方面,随着扩散模型的迅猛发展,文本引导的3D人物生成成为一个新兴方向。为了解决文本到3D形状映射的复杂性以及文本—3D数据对的缺失问题,AvatarCLIP(Hong等,2022b)用SMPL(skinned multi-person linear)(Loper等,2015)初始化人物网格,通过CLIP计算文本标记与网格渲染图的匹配度来迭代优化3D人物。进一步地,DreamAvatar(Cao等,2024)利用参数化SMPL的姿势和形状作为先验来指导人物生成,HumanNorm(Huang等,2024f)通过学习扩散模型和引入多步骤SDS损失,克服了SMPL生成几何表面过于平滑且外观类似卡通风格的挑战。此外,Joint2Human(Zhang等,2024f)利用2D扩散模型直接生成具有全局结构和局部细节的多样化3D穿着人物,如图13所示。它首先利用FOF表征原始网格数据,经编码器编码,在潜在空间中训练扩散模型逐步添加噪声,生成特征向量。同时,利用姿态编码器编码人体关节姿势和3D关节球面嵌入,并利用图像编码器编码文本生成的2D人体图像。特征向量整合姿态特征和图像特征后,送入U-Net去噪。再通过低频解码器生成初步3D形状,然后由高频解码器增强高频信息。最终,通过多视图重雕刻策略生成高质量的3D数字人。多视图体雕刻在3D空间中对初始生成的人体网格进行旋转和FOF转换,并混合不同视角的占用场,以消除伪影。近期,CLAY(Zhang等,2024e)支持文本、图像和多种3D表示输入,通过结合大规模生成模型和多分辨率VAE,以及利用潜在的DiT(diffusion transformer)扩散模型,并融入LLM如GPT-4V来自动标注和处理3D数据,生成具有复杂几何结构和渲染纹理的3D数字人。

由于占用场可以保持拓扑信息,但需要转换为一个无缝隙的实体,这会不可避免地降低3D网格的原始质量,进而生成低分辨率3D物体。同时,生成3D模型的高质量3D数据相对稀缺,但通过设计有效的多模态模型对齐方法来增强3D生成模型,仍是一个充满希望的研究方向。

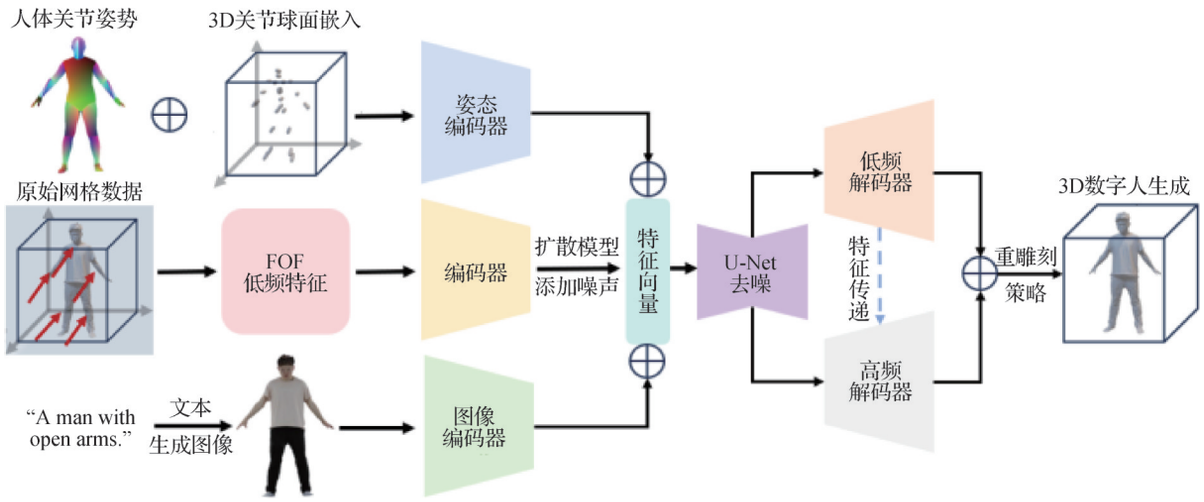

图 13　三维人体生成(Zhang 等,2024f)

Fig. 13　3D human generation(Zhang et al. ,2024f)

4.1.2　四维序列生成

四维序列生成是指创建包含三维空间信息以及随时间维度变化的动态场景,其核心在于模拟和渲染随时间进展发生外观和结构变化的场景。在构建4D序列时,首先必须明确其表示形式。早期研究尝试将4D物体表征为一个涵盖空间坐标和时间维度的函数(Xian 等,2021;Li 等,2022c),或融合静态3D高斯与变形场来生成4D物体(Chen 等,2023a)。其中,高斯泼溅因高效且高质量的重建能力而备受关注。例如,Yang 等人(2024d)采用时间切片方法,直接将3D高斯扩展到4D,为4D表示提供了新的视角。然而,这些4D表示方法通常需要大量的计算资源。为克服这一难题,研究者们采用4D平面分解(Cao 和 Johnson,2023)、哈希表征(Turki 等,2023)等多种方法,显著提升了4D重建的效率与质量。4D序列生成当前方法主要采用文本到视频扩散模型(Singer 等,2023),以此提取4D内容并合成相机轨迹,进而通过SDS方法优化渲染视频。以及结合预训练的扩散模型和动态3D泼溅的STAG4D(spatial-temporal anchored generative 4D Gaussians)(Zeng 等,2024)。最新的进展一方面集成多个扩散先验模型和利用LLM辅助来增强视图真实感,提供更有效的监督信号(Bahmani 等,2024);另一方面,Comp4D(Xu 等,2024a)利用LLM指导运动轨迹,创造出具有真实感和复杂交互的4D动态场景,如图14所示。首先,LLM根据文本输入将场景分解为多个对象的文本描述,再生成相应的多个独立3D对象。接着,LLM用于设计对象轨迹,得益于基于3D高斯分布的组合4D表示,Comp4D在组合得分蒸馏的每次迭代过程中,灵活地在以对象为中心的渲染和轨迹引导的渲染之间切换,以优化4D损失。通过这种方式,Comp4D生成一个动态的4D场景,其中每个对象根据其轨迹在时间维度上移动,最终实现文本到动态4D场景的高质量转换。

此外,对LLM进行微调还可以生成更精确的轨迹和更复杂的运动。其次,生成的四维序列仍受视频扩散模型能力的限制,例如真实性和长度方面,也将是未来研究的重点。

4.1.3　三维场景生成

3D场景生成是一项独特的挑战,也是机器人技术的重要研究方向,但3D场景生成仍拥有诸多难题。Text2NeRF(Zhang 等,2024c)结合扩散模型和单目深度估计技术,通过场景补全绘制出具有逼真视觉效果的图像,但受到3D一致性有限的困扰,即使微小的干扰也会导致场景视觉完整性的崩溃。因此促生了利用条件变分自编码器、生成对抗网络、NeRF与扩散模型的文本到3D对象生成模型(Huang 等,2023b;Zhang 等,2023c;Po 和 Wetzstein,2024),解决了生成场景与真实场景在物理、语义和视觉角度的不一致性,并使用局部条件扩散来增强场景的多样性。CompoNeRF(Bai 等,2024)则提出将文本解释为可编辑的、包含多个NeRF的3D布局,并对每个对象特定的NeRF进行组合、分解和重新排列,增强多对象场景的一致性。尽管这些模型将对象与环境合并以确保一致性,但它们通常产生质量较低的场,且只能容纳有限数量的对象。此外,场景生成还能

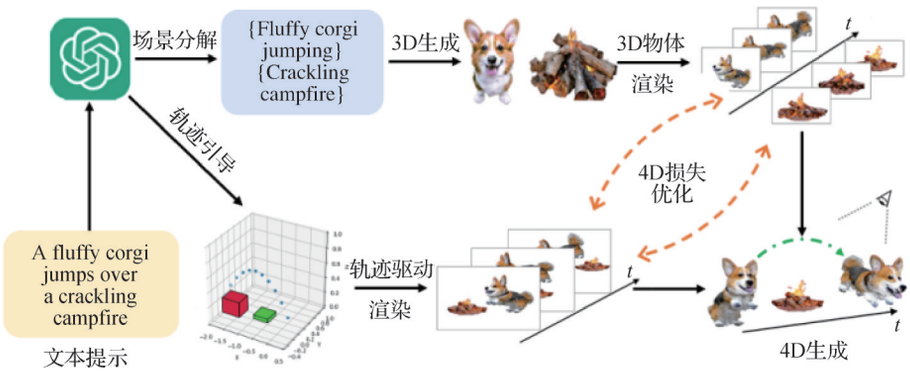

图 14　四维序列生成(Xu 等,2024a)

Fig. 14　4D sequence generation(Xu et al. ,2024a)

从以单个对象为中心的组件创建新场景(Wu 等,2022;Xu 等,2024a),却缺乏对象交互和空间位置的指导,难以构建包含多个对象且具有复杂交互的 3D 场景。还有研究引入手动设计或脚本代理的布局方式,采用 SDF 或 3D 高斯等信息表征手段来施加几何约束,以捕捉场景中多对象的交互关系(Gao 等,2024;Zhou 等,2024),如图 15 所示。LLM 首先根据文本输入创建粗略的场景布局,再采用 MVDream 的实例级扩散先验知识,结合组合优化策略,构建布局引导的高斯泼溅表示,然后融入自适应几何控制来规范化高斯分布的几何形状和空间分布。同时,利用 ControlNet 作为场景级扩散先验知识,细化了 LLM 生成的初始布局,实现对全局场景的一致性优化。最终,通过整合场景布局和实例高斯分布,GALA3D 生成了高交互性的 3D 场景。MLLM 引导了 3D 场景的布局生成和交互式编辑,但未能生成全面的场景。还有研究更注重于可编辑性,例如,LLplace(Yang 等,2024c)提出一种基于轻量级微调 Llama3 的新型 3D 室内场景布局架构,仅需用户指定房间类型和所需属性,减少了对场景对象空间关系的先验假设,增强对上下文信息的理解。Chu 等人(2024)利用 LLM 从基准真实的几何图形或辅助模型重建的 3D 场景中提取点云,实现了对象与场景的统一表示和架构重建。

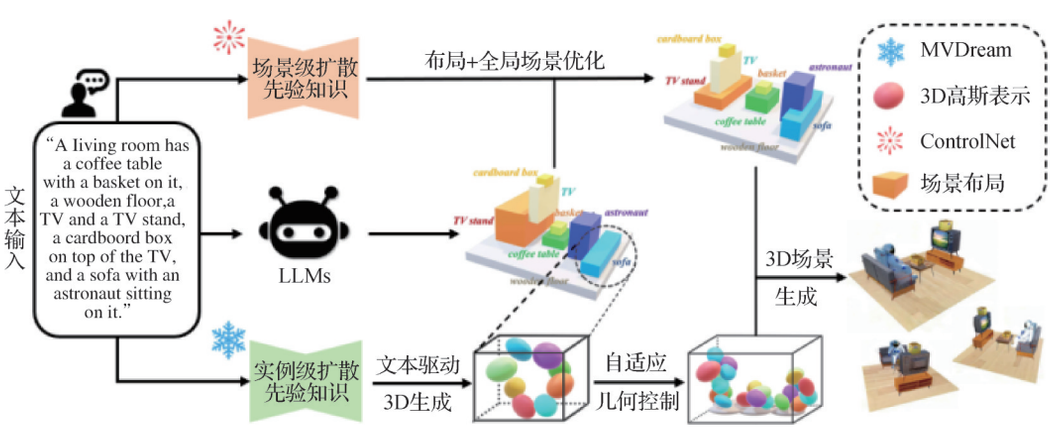

图 15　三维场景生成(Zhou 等,2024)

Fig. 15　3D scene generation(Zhou et al. ,2024)

质量、一致性和可编辑性阻碍了 3D 场景生成模型的广泛应用:低效的生成过程,常导致低质量的效果和较长的生成时间;不一致的 3D 场景视角,仅在有限视角下有良好的渲染结果;无法将物体与环境分离,进而无法对单个物体进行灵活编辑。

4.2　三维目标检测

近年来,自动驾驶因其减轻驾驶员负担、提高行车安全的潜力而受到越来越多的关注。3D 目标检测可以智能地预测自动驾驶车辆附近关键 3D 目标的位置、大小和类别,是感知系统的重要组成部分。

3D开放词汇目标检测的研究焦点在于不显著更改现有模型结构的前提下，增强模型的泛化能力，识别更多未知类别的新目标。其任务要求只提供少量样本或仅输入目标类别来实现（Zareian等，2021；Bangalath等，2022）。最初，由于2D目标检测领域存在大量的图像—文本对和区域—文本对，其发展较为迅速（Schuhmann等，2022），进而推动了利用大规模词汇—3D物体数据集来增强现有3D场景数据集，为开放词汇3D对象检测提供了基础数据集（Zhu等，2023a）。现有方法倾向于使用微弱的监督信号训练检测器（Yao等，2023，2024；Minderer等，2023），或采用蒸馏VLM对数据集进行预训练、微调或冻结等操作（Yu等，2023b）。为了提升目标检测的鲁棒性能，使用视觉文本嵌入替代视觉检测器的分类层（Shi等，2023；Cheng等，2024；Ma等，2023）。然而，开放词汇3D对象检测面临的关键挑战是新对象的定位和分类问题，新对象定位可通过增强3D几何先验知识来解决。Peng等人（2023）表明，标注3D边界框的高昂成本和时间耗费导致了有限规模的3D数据集，而使用未标记数据进行自监督预训练可以在有限标签的情况下提高新目标定位的准确性。因此，Khurana等人（2024）从未配对的RGB和LiDAR数据生成零样本3D边界框，并使用现成的图像基础模型将边界框作为伪标签训练。Coda（Cao等，2023b）则利用3D框几何先验和2D语义先验来生成新对象的伪框标签。为了分类新物体，Coda还开发了一个基于新物体边界框的跨模态对齐模块，以对齐三维点云与图像、文本的特征空间。此外，开放词汇3D目标检测的现有方法大多只围绕从单一基础模型中提取知识，而FM-OV3D（Zhang等，2024b）提出集成多个基础模型的跨模态知识混合方法，通过混合来自多个预训练基础模型的知识，提高3D模型的开放词汇定位和识别能力，实现了不受原始3D数据集约束的真正开放词汇。3D点云最易获取且属性信息丰富，但直接应用VLM等基础模型面临着获取大规模点云—文本对的挑战，以及图像和点云模态间的差异问题。因此，现有研究成果，如PointCLIP系列（Lu等，2023d；Zhu等，2023c；Etchegaray等，2024）使用CLIP处理从3D模态投影的多视图图像，如图16所示。首先，输入的3D点云通过CLIP模型的多视角投影模块被转换成深度图。接着，这些深度图被送入CLIP的视觉编码器中进行处理。同时，GPT模型接收3D指令提示，生成与3D点云对应的特定文本描述，这些文本描述随后被输入到CLIP的文本编码器中，通过CLIP模型，视觉特征和文本特征被对齐并结合，形成用于不同3D任务的特征表示。这些任务包括零样本3D部件分割、零样本/少样本3D分类以及零样本3D目标检测。PointCLIP V2作为一个3D开放世界学习的框架，实现了无需3D领域训练数据的开放世界理解能力。Guo等人（2023b）提出的Point-Bind&PointLLM，利用LLM和多模态语义实现零样本3D分析。

然而，面对机器人交互、增强现实、自动驾驶等实时要求高、干扰性强的应用场景，3D开放词汇目标检测模型的推理速度和鲁棒性仍需提高。

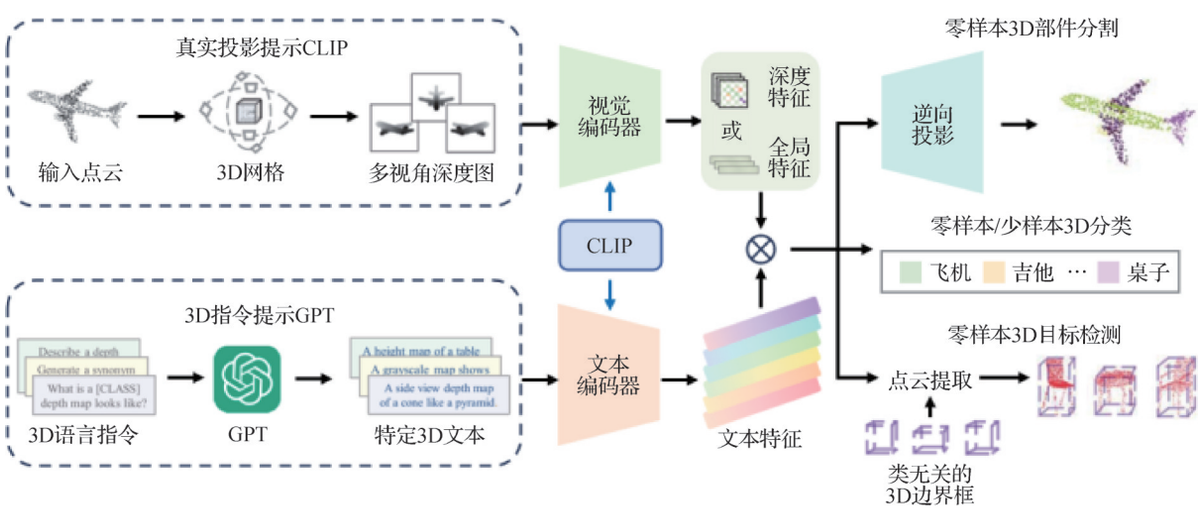

图16　三维语义分割和目标检测（Zhu等，2023c）

Fig. 16　3D semantic segmentation and object detection（Zhu et al.，2023c）

4.3 三维语义分割

开放词汇3D语义分割旨在识别和分割在训练集中未标注的新类别物体。在该领域，首先由Gu等人（2023b）和Peng等人（2023）利用VLM来辨识未知类别。其中，OpenScene（Peng等，2023）对齐了点特征与2D开放词汇语义分割模型（Luo等，2023a）的像素特征，ConceptGraphs（Gu等，2023b）通过多视图关联并将输出融合到3D空间，构建了涵盖对象属性的开放词汇场景图。ConceptGraphs进一步引申出3D语义分割的另一有效策略：通过使用多视图图像呈现3D场景并生成对应文本标注（Yang等，2024b）。此方法能直接对齐3D场景特征与文本特征，从而促进无缝的跨模态交互。然而，这种对齐主要在场景或区域层级进行，缺乏密集预测任务所需的细粒度指导。因此，OV3D（Jiang等，2024）通过整合VLM的开放场景知识，在3D点和文本实体描述之间建立细粒度的对应关系。在3D点云实例分割领域，点云—图像数据对相比点云—文本数据对更易获取，因而现行的研究方法（Chen等，2023b；Ding等，2023；Takmaz等，2023）通常借助图像作为桥梁，将点云特征与已经和文本特征校准的图像特征对齐，建立文本与3D模态的联系。OpenMask3D（Takmaz等，2023）首先使用3D点云生成类别不可知的实例掩码候选簇，并用于选择2D视图和掩码。区别于OpenMask3D，OVIR-3D（Lu等，2023b）整合了来自2D实例分割模型的2D掩码以生成3D候选簇。然而，它们都存在推理速度慢，且依赖于从2D模板生成新的3D候选模板的问题。故Open-YOLO 3D（Boudjoghra等，2025）采用一种高效的2D-3D联合推理策略，使用2D边界框预测来替代大计算量的分割模型，从而显著提升检测效率。为了更全面地捕获点云的细粒度几何特征，UniM-OV3D（He等，2024）设计了一种分层的点云提取器，用于获取局部和全局特征，并构建了分层的点—语义文本对，实现了3D点云与多模态数据的深度融合。Xiao等人（2024）融合了LiDAR特征和文本特征，并提出对象级和体素级蒸馏损失，首次实现了3D全景分割。最近的研究将当前MLLM中嵌入的知识从2D转移到3D，实现了零样本的3D场景理解（Tai等，2024；Ton等，2024）和3D推理分割（Chen等，2024c）。

目前用于大规模开放场景3D语义分割的数据集较少，通过开发更先进的自动化数据生成技术，以创建更大规模和多样化的三维视觉语言数据集来进一步提高模型性能。

4.4 三维场景描述

4.4.1 三维场景文本描述

三维场景文本描述任务旨在为整体3D场景和其中的实例对象生成详细且准确的文字描述。与2D视觉文本描述相比，这项任务更全面地呈现真实世界，但也因3D点云数据收集与处理的复杂性而颇具挑战。早期研究主要通过“先检测后描述”（Chen等，2021c；Yuan等，2022）和“集合到集合”（Cai等，2022；Chen等，2023e）两种范式来处理该任务。“先检测后描述”范式通过评估边界框间的关系进行建模，以级联方式执行对象定位和描述生成。例如，X-Trans2Cap（Yuan等，2022）首先利用检测器提取场景中的实例对象，然后使用师生框架提取输入的多模态信息，最终由解码器生成文本描述。亦或如Vote2Cap-DETR（Chen等，2023d）和Vote2Cap-DETR++（Chen等，2023e）采用的“集合到集合”范式，将3D场景文本描述视为一个集合到集合的问题，并使用基于Transformer的框架解耦对象定位与字幕生成任务，再分别进行独立处理。

但这些方法都只提取特定于任务的3D信息，未充分利用3D特征，使得模型的泛化性有限。最新研究引入了LLM技术，通过不同的模态对齐方式，转换3D数据以应用于多任务的大规模预训练任务。例如，PointLLM（Xu等，2024b）将3D场景的几何、外观信息融入点云编码器，利用LLM理解文本指令微调任务。LiDAR-LLM（Yang等，2023c）、LL3DA（Chen等，2023c）设计视图感知Transformer（view aware Transformer，VAT）桥接点云和LLM，从而将3D场景理解重述为语言建模问题，如图17所示。首先，激光雷达数据输入冻结的激光雷达编码器，以提取3D特征。其次，VAT接收嵌入视图位置信息的3D特征，并将它们映射到预训练的LLM的特征空间，进而对齐文本特征与3D特征。最终，LLM经Adapter模块微调，能实现多种3D下游任务，如图中展示的3D字幕生成、3D定位、3D问答和自动驾驶规划等。一些研究（Li等，2023e；Hong等，2023c）从多视图获取3D特征并以2D VLM为主干网络构建3D-LLM。值得注意的是，3DMIT（Li等，2024e）越过了视觉特征与文本嵌入对齐的阶段，直接将其作为每个3D多模态任务的提示，从而有效地微调3D-LLM。

然而3D场景文本描述仍面临以下挑战：首先，高质量的3D—文本对数据相对稀缺；其次，从稀疏的点云数据中提取复杂的几何和空间关系存在困难；最后，在开放和动态的3D环境中，难以理解和识别不同实例和对象之间的关系。

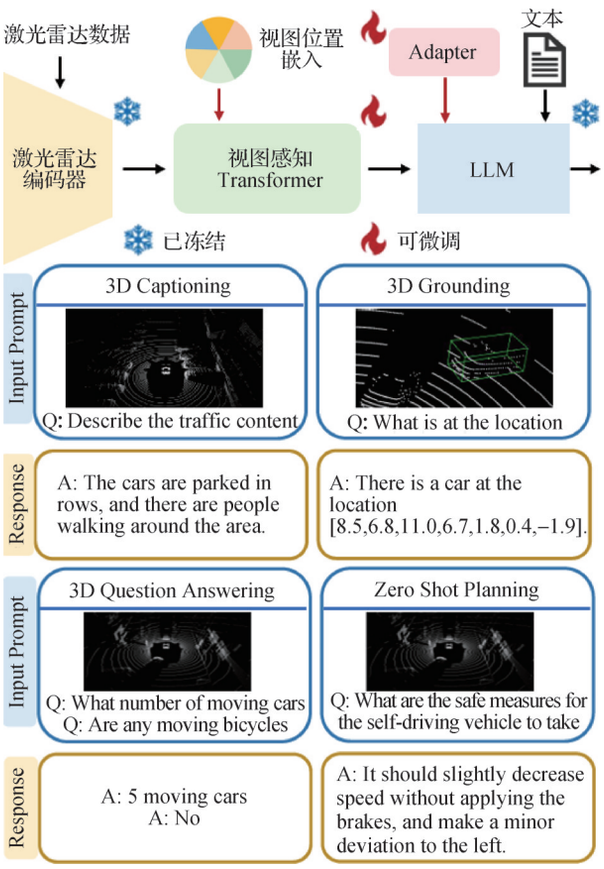

图17　三维场景描述、场景目标定位和场景问答（Yang等，2023c）

Fig. 17　3D scene captioning, scene grounding and scene question answering (Yang et al., 2023c)

4.4.2　三维场景图预测

3D场景图的概念最初由Armeni等人（2019）引入，3D场景图预测任务旨在构建一个有向图，其节点是3D实例的语义标签，边识别连接的3D实例之间的方向性语义或几何关系。自场景图提出以来，后续工作探索了大规模场景中分层3D场景图的构建（Wu等，2021；Hughes等，2022）。同时，一些方法更侧重于预测实例对象间的联系，并构建关系图（Feng等，2025；Wang等，2023b）。预测3D场景图的研究方法集中于基于视觉上下文和基于先验知识两个方向。基于视觉上下文的方法始于Wald等人（2020）提出的3DSSG数据集和SGPN（scene graph prediction network）模型，但无法预测细粒度实例和对象间的复杂关系。因此，Granular3D（Huang等，2024c）使用自适应实例包络法，构建具有全面上下文环境信息的子场景，实现了多粒度3D场景图预测。另一方面，基于先验知识的方法如类别分组嵌入（Qi等，2024a）和统计先验（Feng等，2025），将实例类别和对象间关系的常识嵌入3D表示，显著提高了3D场景图生成的预测准确性。VL-SAT（visual-linguistic semantics assisted training）（Wang等，2023b）通过训练多模态预测模型集成视觉语言知识，并利用CLIP来辅助3D模型，解决了子场景信息提取的缺陷问题。此外，为了实现零样本3D场景图预测，ConceptGraphs（Gu等，2023b）和Open3dsg（Koch等，2024）将3D场景嵌入VLM的特征空间，并利用LLM表征3D场景的语义信息。Chen等人（2024b）基于CLIP开发了一个强大的3D场景图（3D scene graph，3DSG）特征提取器，如图18所示。

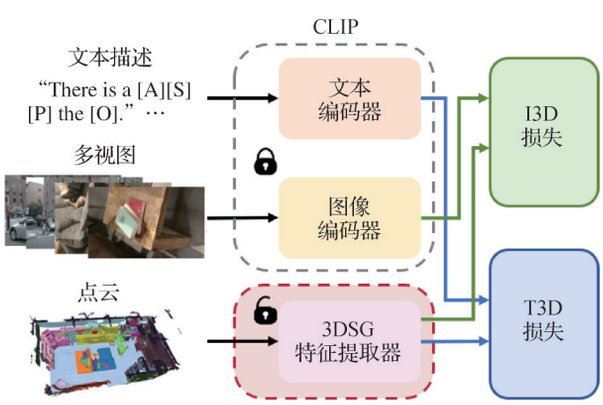

图18　三维场景图预测（Chen等，2024b）

Fig. 18　3D scene graph prediction (Chen et al., 2024b)

文本描述和多视图分别被冻结的CLIP文本编码器和图像编码器编码成文本特征和图像特征，而未标记的3D点云通过可微调的3DSG特征提取器提取3DSG特征。随后，分别经过文本—3D场景图对比（text-3DSG contrastive，T3D）损失和图像—3D场景图对比（image-3DSG contrastive，I3D）损失将文本、图像与3DSG特征对齐。因而，在推理阶段，经过良好对齐的3DSG特征能预测现实世界场景中的新3DSG类别，从而实现对新对象和谓词类别的识别。这些多模态大模型的应用不仅提高了3D场景图的预测精度，而且扩展其在一系列下游任务的应用，如复杂3D场景渲染、3D场景图任务规划（Rana等，2023；Dai等，2024），以及目标对象导航（Loo等，

2024）和移动操作的机器人交互（Honerkamp 等，2024）。

现有的3D场景图预测数据集较小，且存在数据长尾分布不平衡的问题。同时，3D场景预测任务的实例间关系更为复杂，并缺乏合适的预训练方法，与MLLM的结合较为浅显。

4.5 语言引导三维场景目标定位

语言引导的3D场景目标定位旨在以语言为导向，通过简单的文本描述定位复杂三维场景的特定物体。例如，ScanRefer（Chen 等，2020）和 ReferIt3D（Achlioptas 等，2020）作为该领域的开创性工作，最先提出3D视觉定位数据集及基准模型，为后续的语言监督训练奠定基础。语言监督训练通常有两种训练流程：单阶段和两阶段。更受关注的两阶段训练方法（Cai 等，2022；Jin 等，2023）首先使用预训练检测器生成物体边界框并提取特征，然后通过位置回归或属性分类来建模物体的位置及关系。而单阶段方法（Luo 等，2022；Wu 等，2023c）直接整合语言信息，通过逐点过滤在单一阶段内完成目标定位，极大简化了训练流程。然而，上述方法都依赖于文本监督，鉴于注释3D场景的成本非常高和LLM的2D开放词汇分割领域的进展（Li 等，2022a；Liang 等，2023a），为3D视觉定位任务的研究提供了新的思路（Yang 等，2023b；Yuan 等，2024a；Zhang 等，2024g）。这些方法大多以CLIP作为底层的视觉—语言交互桥梁，利用预训练编码器减少对文本标注的需求。但 Yuksekgonul 等人（2023）表明 CLIP 缺乏诸如关系、属性和文本或视觉顺序等组合理解能力，模型性能仅限于简单的名词短语文本引导。为应对该挑战，3D-CLR（Hong 等，2023b）使用文本蕴含的空间信息和3D数据来训练一个先于文本引导过程的、解析和组合语义查询的框架。CoT3DRef（Abdelrahman 等，2024）重述3D定位问题为序列到序列的Seq2Seq任务，通过预测一系列锚点来定位最终目标。Dora（Wu 等，2024）利用LLM从文本描述生成理想的参考顺序，使预测锚定对象无需确切位置也能逐步定位目标对象。然而，当前3D场景定位模型（Wu 等，2023c；Yin 等，2024b）在模型设计和损失设计上严重依赖于任务特定的知识或高级优化策略，并主要从2D VLM获取特征信息，未能充分捕捉到如物体空间关系等3D表征的关键信息。因此，GPS（Jia 等，2024）融合Transformer架构与LLM来实现语言和3D场景之间的多层次对齐。同时，多视图架构的ViewRefer（Guo 等，2023a）利用LLM扩展单文本至多个几何一致的描述，引入多头注意力机制以增强物体跨视图交互。此外，为了克服传统3D场景目标定位所需的大量文本注释和预定义词汇表的难题，许多研究（Yuan 等，2024a；Zhu 等，2024a；Zhang 等，2024g）利用LLM的先验知识和思维链推理能力，以零样本学习的方式实现3D场景目标定位及实例级3D视觉感知定位（Fei 等，2024c）。同时，LAMM（Yin 等，2024b）和 3DMIT（Li 等，2024e）基于 GPT 等MLLM进行文本引导的指令调优。LAC（Feng 等，2024）利用LLM提炼出词间语义关系的一般约束，并通过正则化将语言约束编码到神经符号概念学习器，显著提高自然监督下的目标定位准确性。

然而，以LLM作为核心推理代理需要巨大的计算成本，限制了模型在机器人等端侧设备有限环境的资源部署。且LLM模型的固有延迟给实时机器人领域的应用带来了挑战。MLLM在3D场景目标定位的研究仍需从这两个角度进行优化。

4.6 三维场景问答

3D场景理解的研究始于语言引导任务，例如3D场景目标定位。而随着定位任务相关探索的深入，学者们进一步提出3D视觉问答（3D visual question answering，3D-VQA）任务。该任务要求模型处理三维扫描数据，并回答关于场景的问题，旨在评估模型对于空间关系的理解和常识推理能力。为了实现3D-VQA任务，研究者构建了3D—文本问答对数据集。例如，ScanQA（Azuma 等，2022）基于ScanNet数据集（Dai 等，2017），构建了一个3D-VQA数据集，并设计一个融合模型来对齐三维对象和文本表征，以预测正确答案。Ma 等人（2023）介绍了SQA3D数据集，它专注于场景的具身化理解和问答，要求智能体根据文本描述确定其在三维空间坐标，并能根据周围环境进行问答。然而，面对三维场景中复杂的几何结构和物体间空间关系，Guo 等人（2023b）推出了第一个遵循3D多模态指令的三维语言模型，即Point-Bind LLM。经过参数微调，Point-Bind LLM将Point-Bind的多模态语义融入预训练的语言模型，展现出卓越的3D问答性能。对于零样本功能的实现，最近的研究工作（Parelli 等，2023；Chen 等，2024d）利用了LLM的2D先验知识引入场景对象的语义信息。Scene-LLM（Fu 等，2024）和 SIG3D（Man 等，

2024)结合文本引导的LLM和3D视觉编码器,以自身中心的视角获取场景信息,以应对3D视觉语言推理的态势感知难题。为了更高效地整合多模态特征,Tang等人(2024)提出一种四阶段级联训练策略来对齐多种模态空间,而Chen等人(2024d)则通过自举法和对比性语言—场景预训练,利用场景指代词作为特定的名词短语,实现了三维视觉与语言模型的整合。鉴于现有的3D-VQA模型性能受限于小规模数据集,SpatialVLM(Chen等,2024a)创新性地运用视觉MLLM自动生成和增强3D-VQA数据集,并凭此训练VLM以增强3D空间推理能力。此外,基于三维场景的问答研究进一步衍生出了具身问答任务(Das等,2018)。Singh等人(2024)与Patel等人(2024)分别探讨了室内和家居场景下,零样本学习的LLM在具身视觉问答中的应用,并且指出具身问答为3D场景问答提供更复杂的交互和认知场景。

对于包含多对象的上下文信息、需要复杂计算或多步骤逻辑推理,现有3D-VQA模型难以准确回答。同时,机器人具身问答在泛化到需要大量知识密集型推理的场景时,仍与人类表现有巨大差距。

5 多模态大模型机器人三维视觉应用

第3节介绍的多模态表征对齐技术除了在第4节的三维视觉任务中应用广泛,在机器人领域也做出了重大贡献。本节将多模态大模型赋能的机器人三维视觉任务分为机器人三维抓取、机器人三维视觉导航和其他任务,每个任务都采用特定的常用数据集和评估指标。并通过多模态大模型与传统和非大模型方法的对比,揭示3D-LLM在机器人三维视觉应用中的潜力,为未来的研究方向和应用实践提供参考。

5.1 机器人三维抓取

在机器人抓取领域中,传统算法通过分析物体与抓取装置的几何属性来生成与评估潜在的抓取姿势。然而,这些方法主要集中在形状匹配,通常依赖于详尽且精确的物体模型,缺乏对物体语义信息的理解。为降低对完整模型的依赖,研究者探索从三维数据中直接回归来识别潜在的抓取对象,既减少对物体先验知识的需求,又开辟了机器人3D抓取的新研究方向。此方法早期阶段,许多工作提出基于卷积神经网络的物体抓取方法,以点云(Lundell等,2020;Breyer等,2021)和视图(Morrison等,2018;Kumra等,2020)为主要的3D表示形式。而当前研究阶段,LLM已经在机器人领域得到了广泛的应用。为了使LLM能根据特定任务和3D场景上下文来生成相应动作,首先需要定义动作标记。LEO(Huang等,2024a)作为首个3D具身多任务多模态的通用智能体,在统一框架内依据视图、场景和指令设计了大量的特定标记来精确控制机器人运动。这类模型(Hong等,2024;Zhen等,2024)各自使用不同的动作标记,将指令映射为3D空间中的机器人动作,从而使机器人能执行复杂的指令引导任务。然而,它们忽视了对可操纵物体的语义理解(Li等,2024c),并且通常无法判断抓取装置的可操作性。为了解决这一问题,研究发现VLM具有通过视觉等感官输入与3D世界建立联系的能力。因此,催生出了结合VLM和LLM的3D具身MLLM。例如,Voxposer(Huang等,2023d)利用LLM从输入指令中提取可用性和约束条件,并整合VLM提供的视觉信息在观察空间中构建3D体素图,以指导机器人交互。而Singh等人(2024)凭借MLLM的常识库和鲁棒推理能力,使用图像和显隐式指令来推理生成最佳的抓取姿势。

现有的机器人三维抓取大多基于仿真环境,由于仿真环境在模拟现实世界的复杂光照和物体材质方面存在局限,模型仍需在真实世界进行评估和调试。同时,模型在某些特定类别物体抓取上仍有改进空间,未来研究可以探索如何结合3D视觉任务来适应开放场景中的新类别物体抓取。

5.2 机器人三维视觉导航

在机器人导航领域中,传统技术利用同步定位和映射(simultaneous localization and mapping, SLAM)方法来构建高精度的密集空间地图。近年来,由于机器人3D导航任务备受关注,许多研究(Chang等,2023;Singh等,2023;Liu等,2023c)采用3D场景图结构来解决大范围场景的空间表示问题。它们不仅提高了空间表示效率,还为LLM提供了有效的语义标记接口(Gu等,2023b;Werby等,2024),从而显著增强了机器人的理解和交互能力。例如,ConceptGraphs(Gu等,2023b)开发了一个以对象为中心的3D映射系统,如图19所示。该系统接收RGB-D序列作为输入,并通过开放词汇的目标检测/分割模型来识别和分割出候选对象。接着,这些对象通过几何和语义相似性度量来对齐特征,进而在

多个视角中关联，增量生成3D空间并构建物体集。然后，利用LVM为每个3D物体生成描述，再利用LLM推断相邻节点间的关系，这些关系在场景图中表现为连接物体的边。最终生成的3D场景图不仅具有开放词汇，能封装物体属性，还能应用于多种下游任务，如开放场景的机器人3D抓取和视觉导航等任务。而HOV-SG（Werby等，2024）提出一种分层结构的3D场景图，通过分解抽象查询，使用Voronoi图遍历检索场景目标以实现长视距机器人导航。尽管机器人3D视觉导航已经取得了一系列的突破，但依然面临众多挑战（Huang等，2023a；Wu等，2023c），例如对未知对象的导航、语言描述的精确解释以及依据对象属性的个性化导航等。针对上述挑战，许多研究工作进行了多种尝试与创新。VELMA（Schumann等，2024）通过文本提示查询过去的运动轨迹，然后利用LLM预测动作序列，有效应对了复杂城市环境下的导航问题。Vemprala等人（2024）通过设计提示词和创建高级函数库，促成了ChatGPT在机器人领域的应用。为了区分同一类别的不同实例对象，IVLMap（Huang等，2024b）将自然语言转换为具有实例和属性信息的导航目标，并基于自然语言命令完成零样本的端到端导航任务。此外，NaviLLM（Zheng等，2024）通过引入基于模式的指令，灵活地将各种任务转换为生成问题，使得LLM能适应具身导航的需求，并且显著提升了机器人在未知场景中的导航泛化能力。

当前的机器人三维视觉导航方法在未知的新环境中性能有所下降，且仅限于离散环境，仍需要探索如何扩展至大规模连续或复杂的场景。

5.3 其他任务

机器人领域的传统深度学习模型通常基于特定任务的小型数据集进行训练，这限制了它们在各种应用场景中的泛化性。为克服这一局限性，最近的研究方向（Song等，2023；Fu等，2024；Li等，2024e）趋向于将LLM和VLM等基础模型整合到机器人系统。这些模型不仅增强了机器人对环境的感知和交互能力，还促进了上下文感知机器人系统的发展。在三维视觉和机器人交叉领域的应用中，MLLM除了在机器人3D抓取和视觉导航任务贡献显著，还有许多相关的研究工作，如图19所示。在感知领域，VLM通过学习视觉和文本数据之间的关联来提供跨模态理解，从而辅助完成如3D分类（Hong等，2023c）、6D姿态估计（Cai等，2024）以及手术场景分割（刘敏 等，2024）等任务。机器人利用场景语义分割辨识未知对象，然后通过点云数据或物体形状的几何信息来估计目标坐标和姿态。同时，Foundationpose（Wen等人，2024）使用6D物体姿态跟踪，实现以时间线索的方式对视频序列进行更高效、流畅和准确的姿态预测。而Scene-LLM（Fu等，2024）进一步将VLM的上下文理解与3D现实世界对齐，通过将文本描述与3D环境中的特定对象、位置或动作相关联，机器人能够更好地理解和操作周围的环境。

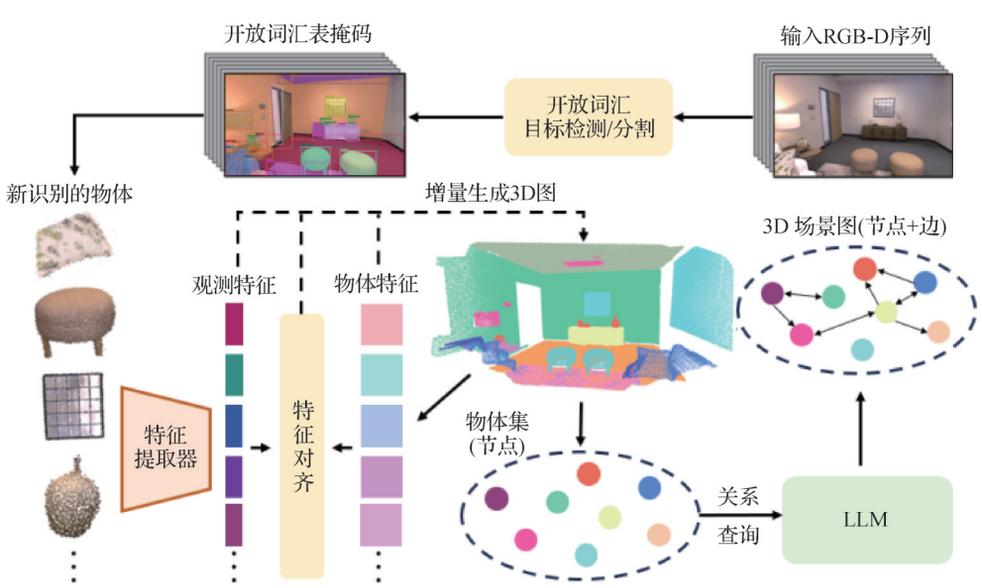

图19 机器人三维抓取、视觉导航（Gu等，2023b）

Fig. 19 Robotic 3D grasping, 3D vision-based navigation (Gu et al., 2023b)

在决策规划领域，LLM和VLM能协助机器人进行任务规范，以实现高层次规划（Rana等，2023；Song等，2023）。这些基于LLM的框架涵盖了广泛的推理层次，从初级的自回归代码策略生成（Liang等，2023b）到纯粹使用自然语言对任务计划进行推理（Mikami等，2024）。此外，研究人员还采用语言模型为策略学习技术提供反馈（Han等，2024a；Bartsch和Farimani，2024），提高机器人的学习效率和性能。

然而，由于缺乏大规模机器人3D数据和基础模型，限制了机器人感知的泛化能力。并且，一些方法涉及模态转换，将预训练的2D特征提升到3D空间，或将3D特征投影回2D空间以便2D模型输入。这些模态转换不可避免地导致空间信息的丢失，不能完全激发模型对3D空间关系的理解。

6 数据集

第4、5节的三维视觉应用驱动产生了3D数据集，并推动3D数据集不断发展。本节将介绍基本的3D数据集，它们是三维视觉研究的基础。然后，将详细介绍集成了多模态信息的3D视觉语言数据集。通过回顾这些数据集的发展历程及呈现图20的数据集示例图，直观展现了相关数据集的特点和内容。本节强调这些数据集在3D视觉理解领域发挥的关键作用，以及它们如何为研究者提供了宝贵的资源，从而促进了3D视觉语言数据集的全面理解。此外，将探讨这些数据集如何启发新数据集的构建，以满足未来研究的需求。

6.1 基本三维数据集

过去对2D图像中的对象进行语义理解和目标定位等视觉任务已经取得了巨大进展，但这些方法和数据集都仅限于2D图像。受限于RGB-D数据采集的高昂成本和3D数据集的低效注释，同期的3D数据集规模非常有限。为了获取大规模3D数据集以支持3D场景理解任务，学者们开放了多种工具并应用自动化技术，对3D数据集的收集和注释工作做出了重大贡献。在3D数据集的早期发展阶段，侧重于扩大数据集规模和专注于3D视觉信息，同时仅有一些探索将语言信息与3D视觉数据相结合的潜力。基于这些特点，将主要包含3D视觉信息而不涉及语言信息的数据集称为基本3D数据集。

ScanNet（Dai等，2017）包含1 513个室内场景的2.5 M幅RGB-D图像，每个场景都提供相机位置姿态、三维表面模型和语义信息。数据采集通过结合了自动化表面重建和众包语义标注的RGB-D捕获系统完成。该数据集适用于深度学习方法在3D场景理解领域的研究，在3D对象分类、语义体素标记和CAD模型检索都展现出了优异的性能。尽管ScanNet规模庞大且提供了丰富的注释信息，但数据采集依赖于众包，可能存在标注质量的不一致性。同时，ScanNet对硬件和处理流程的依赖限制了数据采集的便携性和实时性。

Matterport3D（Chang等，2017）包含90个建筑规模场景的194 k幅RGB-D图像，覆盖10.8 k个全景视图。采集过程利用三脚架安装的相机装置，通过旋转获取6个方向的高动态范围（high dynamic range，HDR）照片及连续深度数据，合成与彩色图像对齐的深度图。Matterport3D数据集以精确的全局对齐、多样的全景视角、高质量的深度信息为特点，适用于关键点匹配、视图重叠预测、表面法线估计、区域分类和语义体素标记等计算机视觉任务。Matterport3D具有多样性和高质量数据的优势，但也存在数据采集的高成本和隐私问题。

ShapeNet（Chang等，2015）从公共在线存储库的多个类别收集了包含超过3 M个3D CAD模型。数据采集过程包括自动化下载和人工验证，以确保模型的质量和准确性。ShapeNet通过WordNet的语义类别组织，支持多种任务，包括几何分析、对象识别和场景理解。ShapeNet的优点在于其广泛的覆盖范围和深度注释，但缺点是模型来源的偏差，倾向于人造物品，且对自然物体的覆盖较少。

3RScan（Wald等，2019）包含478个室内环境的1 482个RGB-D扫描序列，涵盖多次扫描以捕捉环境变化。数据集通过Google Tango应用采集，经相机姿态校准和纹理映射的后处理，以生成精确的3D重建模型。数据集的注释信息包括语义实例分割、对象对齐和对称性属性，因而适用于长期SLAM、场景变化检测和对象重定位等任务。尽管数据集在多样性和规模上具有优势，但动态室内环境的复杂性对其采集和标注过程提出挑战。

3DSSG（Wald等，2020）基于3RScan室内环境3D重建数据集，通过半自动方式生成，旨在提供丰富的语义场景图。3DSSG包含1 482个3D场景，每个场景都由一组场景图表示，这些图由节点和边组

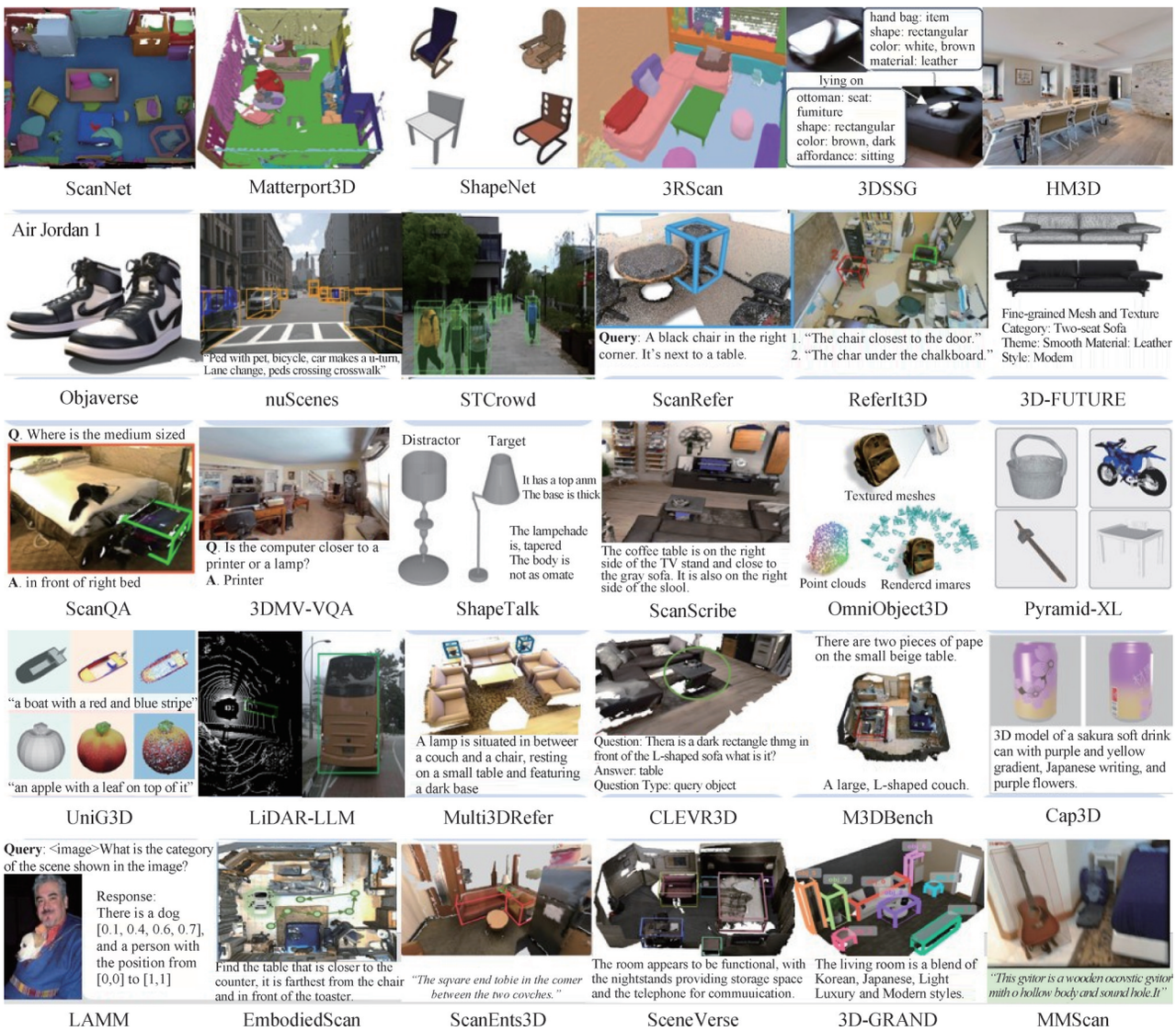

图 20　三维数据集示例图

Fig. 20　3D dataset example figure

成。节点代表 3D 扫描中的具体物体实例，而边则表示这些物体之间的关系。该数据集采集过程结合了语义标注、几何数据及人工验证，以确保数据的质量和准确性。3DSSG 适用于 3D 场景理解任务，如语义分割、对象检测和场景图生成等任务，尤其在跨域检索如 2D-3D 场景检索中表现突出。然而，3DSSG 处理真实世界扫描中的缺失数据、复杂对象关系和场景变化时仍存在挑战。

Structured3D（Zheng 等，2020）利用专业室内设计师的设计和先进的渲染引擎生成，包含 21 835 个房间的 3D 结构注释和超过 196 k 幅逼真的 2D 渲染图像。数据集使用自动化采集过程，从原始房屋设计文件中提取几何结构信息，通过行业领先的渲染引擎生成高质量图像。Structured3D 适用于结构化 3D 建模任务，如房间布局估计，通过结合真实图像训练深度网络，展示了在基准数据集上的改进性能。Structured3D 拥有丰富的注释和逼真渲染，但可能存在合成数据与真实世界数据的差异问题。

HM3D（Ramakrishnan 等，2021）由 Facebook AI Research 联合多所大学共同创建，包含来自全球 38 个国家的 1 000 个建筑规模的室内 3D 重建。这些数据通过 Matterport Pro2 设备扫描获取，并经过严格的筛选和质量控制流程，确保了数据的多样性和高质量。HM3D 目前主要支持几何任务，未来计划扩展至包含语义注释，以支持更高级的理解任务。HM3D 可用于具身智能代理的感知、导航和行动能力的研究，尤其适用于室内环境。尽管 HM3D 在视觉逼真度和重建完整性方面表现出色，但其数据来

源可能因扫描设备成本较高，主要集中于具备设备采购及数据上传能力的地区。

Objaverse（Deitke等，2023b）包含818 k个3D模型及其文本描述，覆盖21 k个不同类别，来源于Sketchfab平台并由超10万名艺术家设计。数据集的采集遵循Creative Commons许可，确保了使用的合法性与多样性。Objaverse通过提供描述性标题、标签和动画，支持广泛的3D应用，包括3D生成模型训练、长尾类别分割改进、开放词汇导航以及视觉模型鲁棒性分析。但由于Objaverse主要依赖于Sketchfab平台的内容，可能缺乏某些特定类别的全面覆盖。

Objaverse-xl（Deitke等，2023a）基于Objaverse数据集，包含超过10 M个从GitHub、Thingiverse、Sketchfab、Polycam和史密森尼学会等来源搜集的多样化3D对象。数据集的采集通过Python脚本自动化完成，并通过Blender软件进行渲染和去重处理。Objaverse-XL适用于3D视觉任务，如新视角合成、3D重建和生成，能够显著提升模型的零样本泛化能力。尽管数据集在规模和多样性上具有优势，但仍存在数据的潜在偏差和隐私问题，以及并非所有样本都对训练高性能模型至关重要。未来的工作可能会探索如何继续扩大3D数据集的规模，并考虑如何选择数据点进行训练。

nuScenes（Caesar等，2020）由nuTonomy公司采集，涵盖美国波士顿和新加坡两地，通过6个摄像头、5个雷达和1个激光雷达传感器，提供360度全方位视角。nuScenes包含1 000个场景，每个场景20 s长，详尽标注了23个类别的3D边界框和8个属性。采集过程注重场景多样性，包括不同天气和光照条件。nuScenes支持自动驾驶车辆的多种任务，如物体检测、跟踪和行为建模。nuScenes填补了缺乏全方位传感器数据的自动驾驶领域数据集的空白，但也存在数据的复杂性和类别间的长尾分布的挑战。

STCrowd（Cong等，2022）由上海科技大学等机构合作采集，通过车载128束激光雷达和单目摄像头在不同场景和天气条件下同步获取，包含219 k个行人实例，平均每帧20人，具有高密度和严重遮挡的特点。STCrowd经过两轮质量检查，确保了数据的准确性和一致性。它支持激光雷达、图像及多传感器融合的行人检测与跟踪任务，并提供了基准方法。STCrowd主要集中于校园环境，可能在其他场景下的泛化能力有待验证。然而，STCrowd的场景多样性相对有限，可能影响模型在复杂多变环境中的表现。

6.2 三维视觉语言数据集

由于LLM近年来取得显著进展，构建3D-LLM应用于3D场景理解已是一个不可避免的趋势。然而，先前的3D-LLM只能使用对象级别的数据集或有限的多模态场景数据集，因此受限于对象级别的理解能力和对象间空间关系的认识。与此相比，3D场景具有复杂的层次结构和丰富的上下文，这表明现有的3D-LLM性能尚未达到应有的期望。这促使学术界构建具有多模态注释且专注于语言基础的3D场景理解数据集，即3D视觉语言数据集，从而为训练和评估3D-LLM提供必要的数据支持。本节旨在介绍现有的赋能3D视觉理解任务的3D视觉语言数据集，且在表1中列出它们的应用任务、场景数以及模态类型。值得注意的是，这些3D视觉语言数据集大多基于现有的基本3D数据集，如ScanNet（Dai等，2017）和3RScan（Wald等，2019）。3D视觉语言数据集通过采用不同的注释策略和针对特定的3D视觉任务进行设计，从而展现出各自独特的特点。下面将详细介绍当前3D视觉语言数据集，包括它们的来源、数据结构和规模，以及它们支持的具体任务。

ScanRefer（Chen等，2020）基于ScanNet构建，涵盖了800个室内场景中的11 046个3D对象，并附有51 583条详细的人工文本注释。该数据集的构建过程包括数据收集、描述编写、验证和筛选等步骤。尽管ScanRefer提供了大规模且多样化的室内环境描述，但其质量可能受到3D对象检测准确性的限制，这直接影响了模型的定位能力。通过丰富的自然语言描述与3D场景的对应关系，ScanRefer数据集显著增强了3D视觉语言模型在复杂环境中的语言理解和空间定位性能，为3D视觉与语言交互的研究提供了重要资源。

ReferIt3D（Achlioptas等，2020）旨在提高3D对象在现实场景中的细粒度识别能力。该数据集基于ScanNet的室内场景构建，涵盖了76个细粒度的对象类别。通过自然语言描述构建的Nr3D数据集和基于空间关系的合成语言描述的Sr3D数据集，ReferIt3D为3D对象识别和空间关系识别任务提供了丰富的上下文信息。Nr3D通过众包平台收集文

本描述，而Sr3D则通过计算空间关系自动生成描述。ReferIt3D增强了对复杂场景的理解与识别能力，适用于3D目标识别和多模态交互任务，但同时也对模型的计算资源和算法设计提出更高要求。

3D-FUTURE（Fu等，2021b）由阿里巴巴集团创建，旨在填补高质量3D家具形状和纹理研究领域的空白。该数据集利用先进的3D渲染引擎，基于专业设计师构建的房间场景，生成了20 240幅高质感的室内图像，并提供了9 992个细节丰富的3D家具模型。这些模型已用于现代工业生产，具有高度精细的几何和纹理属性。3D-FUTURE提供的2D-3D精确对齐和6DoF姿态注释，为3D对象姿态估计、基于图像的3D形状检索、单视图3D重建和3D形状纹理恢复等研究任务提供了有力支持。然而，3D-FUTURE主要聚焦于家庭室内环境，限制了其在其他场景下的广泛应用。

ScanQA（Azuma等，2022）源自ScanNet，通过自动化生成和人工编辑过程构建，包含41 363个问答对，覆盖800个室内场景。数据采集涉及自动化问答对生成、问题过滤、人工编辑和答案收集。该数据集专为3D视觉问答任务设计，要求模型利用3D扫描信息和文本问题来定位场景中的对象并提供答案。ScanQA优势在于丰富的问答对和多样的问题类型，提高了任务的复杂性和实用性。然而，自动化生成的问题可能存在不准确或不相关的场景描述，需人工校正以确保数据质量。ScanQA对3D-LLM的

表1　三维视觉语言数据集

Table 1　3D visual language datasets

数据集	视觉问答	描述生成	目标定位	场景	模态		
					点云	文本	图像
ScanRefer（Chen等，2020）	×	√	√	0.7 k	√	√	×
ReferIt3D（Achlioptas等，2020）	×	√	√	0.7 k	√	√	×
3D-FUTURE（Fu等，2021b）	√	×	×	5 k	√	√	√
ScanQA（Azuma等，2022）	√	×	×	0.8 k	√	√	×
3DMV-VQA（Hong等，2023b）	√	×	×	5 k	√	√	√
ShapeTalk（Achlioptas等，2023）	×	√	×	–	√	√	×
ScanScribe（Zhu等，2023d）	√	√	√	1.2 k	√	√	×
OmniObject3D（Wu等，2023b）	×	√	√	–	√	√	√
Pyramid-XL（Qi等，2023）	√	√	×	–	√	√	√
UniG3D（Sun等，2023b）	√	√	×	–	√	√	√
LiDAR-LLM（Yang等，2023c）	×	√	√	–	√	√	√
Multi3DRefer（Zhang等，2023e）	×	√	√	0.7 k	√	√	×
CLEVR3D（Yan等，2023）	√	×	×	8.7 k	√	√	√
M3DBench（Li等，2023c）	√	√	√	–	√	√	√
Cap3D（Luo等，2023b）	×	√	×	–	√	√	√
LAMM（Yin等，2024b）	√	√	×	–	√	√	√
EmbodiedScan（Wang等，2023a）	×	×	√	5.2 k	√	√	√
ScanEnts3D（Abdelreheem等，2024）	×	√	√	0.7 k	√	√	×
SceneVerse（Jia等，2024）	×	√	√	68 k	√	√	√
3D-GRAND（Yang等，2024a）	√	√	√	40 k	√	√	√
MMScan（Lyu等，2024）	√	√	√	5.2 k	√	√	√

注：“√”表示数据集支持特定下游任务，或其包含对应模态的输入数据类型。“×”表示数据集不支持的下游任务，或其不包含对应模态的输入数据类型。

性能影响显著，其端到端的学习方法通过整合对象定位和问答模块，提升了模型对3D空间特征的理解和问答性能。

3DMV-VQA（Hong等，2023b）源自HM3DSem数据集，通过在模拟环境中操作具身代理来主动捕获RGB图像。该数据集包含约5 k个场景、600 k幅图像及50 k个问答对，涵盖概念、计数、关系和比较4种问题类型，旨在评估和提升模型在3D多视图视觉问答任务中的表现。数据集通过不同视角的图像采集，为3D场景理解提供丰富的视图信息。3DMV-VQA不仅规模庞大，而且通过场景图注释提供了精确的语义信息和对象间关系，为模型训练和评估提供了高质量的数据基础。然而，3DMV-VQA仍面临以下挑战：如何有效整合多视图信息以推断出紧凑的3D表示，并在此基础上进行准确的语义概念理解和复杂关系推理。

ShapeTalk（Achlioptas等，2023）整合ShapeNet、ModelNet和PartNet（Chang等，2015；Wu等，2015；Mo等，2019）等数据库资源，经过去重和质量筛选，最终提供36 k个3D形状和536 k条文本描述。数据采集使用人工注释方式，根据形状对比提供细粒度描述，确保语言的多样性和准确性。ShapeTalk涵盖30种物体类别，适用于语言驱动的3D形状编辑任务，促进了3D-LLM在语言引导的形状编辑任务中的表现。然而，ShapeTalk的局限性在于语言描述的模糊性，可能导致在实际应用中对形状变化的不精确理解。

ScanScribe（Zhu等，2023d）专为3D视觉语言任务设计的大规模3D场景文本配对数据集。它由ScanNet和3R-Scan室内场景的RGB-D扫描组成，并利用Objaverse数据集以增强场景中对象多样性。ScanScribe通过模板和GPT-3生成278 k个场景描述，涵盖1 185个独特室内场景的2 995个RGB-D扫描。ScanScribe的高质量文本和多样化场景表示为3D视觉语言领域的预训练模型带来显著的性能提升，尤其对于3D视觉定位、密集字幕生成、视觉回答和情境推理等任务。

OmniObject3D（Wu等，2023b）是一个大规模的3D物体数据集，包含6 k个来自日常生活的高质量3D扫描物体，涵盖190个类别，与ImageNet和LVIS等2D数据集共享类别，便于研究者探索通用的3D表示。OmniObject3D的每个3D物体通过2D和3D传感器捕获，提供了纹理网格、点云、多视角渲染图像和真实捕获视频，并进行注释。专业的扫描设备确保了高保真度的物体扫描，包括精确的形状和真实的外观。OmniObject3D支持包括鲁棒3D感知、新视角合成、神经表面重建和3D物体生成等多种研究任务。然而，OmniObject3D的复杂性要求算法能够处理高频率纹理和复杂几何形状，这对现有方法提出更高的要求。

Pyramid-XL（Qi等，2023）是一个自动化的点—文本注释引擎，通过3个层级逐步提升描述的质量和细节：Level 1利用单视图生成简短描述；Level 2合成多视图描述；Level 3利用高级视觉语言模型生成详细密集的描述和问答对。Pyramid-XL显著提升了3D-LLM在3D对象识别、文本推理任务中的性能，尤其在理解3D几何和颜色信息方面。然而，Pyramid-XL在生成高质量文本注释时成本较高，对计算资源的需求较大。

UniG3D（Sun等，2023b）是一个用于3D对象生成的数据集，通过通用数据转换管道，将Objaverse和ShapeNet数据集中的原始3D模型转化为包含文本、图像、点云和网格的多模态数据表示，并生成了1.6 M规模的UniG3D-ShapeNet和16 M规模的UniG3D-Objaverse数据集。此外，UniG3D还利用渲染引擎和多模态模型来丰富文本信息。该数据集适用于包括点云和SDF的多种3D对象生成任务，并通过Point-E和SDFusion等方法验证了其有效性。UniG3D的优点在于其通用性，可应用于任何3D数据集，且无需手动注释，展示了良好的可扩展性。

LiDAR-LLM（Yang等，2023c）采用MLLM和GPT4，为nuScenes数据集中的3D LiDAR场景图像增添文本描述，从而获取了包含420 k个LiDAR文本对的nu-Caption数据集。同时，该研究将3D边框融入文本问答对，创建了包含280 k个LiDAR字幕对的nu-Grounding数据集。这些数据集不仅覆盖了从简单到复杂的场景描述，还涉及了对象的详细描述和潜在风险识别，为3D字幕、3D定位和3D问答等任务提供了丰富的训练资源。通过精心设计的问题和答案对，LiDAR-LLM在理解3D场景和执行复杂空间推理方面展现出了优越的性能。

Multi3DRefer（Zhang等，2023e）旨在解决3D场景中自然语言描述与多个目标物体定位任务。Multi3DRefer扩展自ScanRefer数据集，包含61 926条

文本描述，覆盖11 609个物体实例，是首个支持零目标、单目标或多目标物体定位的3D视觉数据集。通过整合ChatGPT，Multi3DRefer在多样性和自然性方面得到增强，并利用人工审核确保了数据质量。该数据集适用于3D视觉定位、机器人导航及增强现实等任务，为多模态3D场景理解的研究提供了新的基准。尽管如此，Multi3DRefer的构建依赖于人工审核，可能会导致验证过程的主观性和时间成本增加。此外，对于模型而言，准确处理零目标和多目标场景的定位仍是一项挑战。

CLEVR3D（Yan等，2023）是专为3D点云上的视觉问答（VQA-3D）任务而设计的大规模数据集，包括了171 k个问题，覆盖8 771个3D场景，旨在提升模型对物体属性（如大小、颜色和材质）及其空间关系的深入理解。CLEVR3D利用结构化信息设计了问题模板，并填充模板变量，从而产生各种类型的问题。CLEVR3D通过真实场景和增强场景的组合场景操作策略构建，以减少模型对常规布局的依赖。CLEVR3D不仅适用于VQA-3D任务，还能显著提升3D场景图分析的性能。然而，CLEVR3D的局限性在于其依赖3D场景图的质量和多样性，对模型的泛化能力提出更高要求。

M3DBench（Li等，2023c）是一个大规模的3D多模态指令跟踪数据集，它通过结合文本、图像、3D对象等多模态提示构建。M3DBench构建过程利用了现有的3D数据集，并结合了自我指令生成方法，通过GPT系列模型生成与特定任务相关的指令数据，以确保数据的多样性和实用性。M3DBench适用于包括对象检测、视觉定位、密集描述、视觉问答、多区域推理、场景描述、多轮对话和机器人具身规划等3D视觉中心任务。其优点在于其大规模的指令—响应对，以及其支持的丰富多模态指令，显著增强了模型对3D场景的理解能力。但M3DBench在特定任务的性能评估和模型优化方面可能需要进一步的研究和开发。

Cap3D（Luo等，2023b）提出一种自动化3D文本描述生成方法，通过结合图像标题、图像文本对及LLMs，从多视图中整合文本描述，避免了耗时且成本高昂的手动注释过程。Cap3D应用于大规模3D数据集Objaverse，产生了660 k个3D-文本对。通过41 k人类注释的评估表明，Cap3D在质量、成本和速度方面均优于人工注释。Cap3D通过有效的提示工程，在ABO数据集上的几何描述生成方面与人类表现相当。此外，Cap3D收集的高质量3D-文本数据集适用于3D对象自动注释、几何描述生成，并作为文本到3D生成模型的基准测试。

LAMM（Yin等，2024b）介绍一个多模态指令调优数据集。LAMM的构建利用了公开可用的数据集，如COCO（common objects in context）、Visual Genome、3RScan等，并通过GPT-API生成196 k个3D指令响应对。LAMM数据采集过程包括日常对话、详细描述、事实知识对话和视觉任务对话，强调了细粒度信息和事实知识。LAMM适用于图像分类、目标检测和视觉问答等3D视觉任务，其优点在于提供了丰富的多模态信息和细粒度描述，但可能存在生成数据的偏差和不准确性。LAMM提升了3D-LLM对视觉和语言任务的理解和执行能力，尤其经过特定任务微调。然而，LAMM在应用于需要精确边界框预测的任务中仍存在挑战。

EmbodiedScan（Wang等，2023a）是为促进具身智能而构建的多模态3D数据集。该数据集基于ScanNet、3RScan和Matterport3D等现有大规模3D室内场景扫描数据，通过重新注释和处理，以适应从第一人称视角的连续场景感知变化。EmbodiedScan包含超过5 k个扫描场景，近1 M个第一人称视角的RGB-D视图，以及丰富的多模态注释，如160 k个3D边界框，涵盖超过760个类别以及1 M个对象间空间关系的描述。这些注释通过精细的标注工具生成，为3D场景理解提供了丰富的上下文信息。Embodiedscan不仅提供了丰富的多模态数据和详尽的注释，还通过引入语言描述，为3D场景理解提供了新的视角，适用于3D检测、语义占用预测和基于语言的3D场景理解等任务。

ScanEnts3D（Abdelreheem等，2024）通过语言关联3D场景对象，并运用Web界面和专业标注团队注释来扩充ReferIt3D和ScanRefer数据集。它涵盖了705个真实世界场景，以及84 k个自然引用语句和369 k个对象间的显式对应关系。ScanEnts3D无需额外推理注释，通过训练时的辅助损失函数增强模型性能。同时，该数据集有丰富的3D视觉语言对应关系，适用于3D场景密集文本描述和目标物体定位任务。

SceneVerse（Jia等，2024）作为首个百万规模的3D视觉语言数据集，通过人工注释和自动生成的方

式构建，包含68 k个室内场景、1.5 M个3D对象以及2.5 M个场景语言对。数据采集过程包括房间分割、点云归一化和语义标签对齐等预处理步骤，确保数据的一致性和质量。SceneVerse还引入基于3D场景图和LLM的自动化生成流程，以提升数据多样性和规模。SceneVerse适用于3D视觉定位、场景理解和多模态交互等任务，通过大规模预训练，能够显著提升3D视觉—语言模型的性能。

3D-GRAND（Yang等，2024a）从3DFRONT（Fu等，2021a）和Structured3D（Zheng等，2020）的合成数据中整合了40 k个家庭场景，并使用GPT-4生成6.2 M个配对的场景语言指令。3D-GRAND适用于多种3D文本任务，包括对象引用、场景描述和问答等。其优点在于提供了大规模、密集关联的数据，有助于提升3D-LLM的性能，特别是在模拟环境到现实世界的迁移学习能力上展现出潜力。然而，3D-GRAND主要基于合成场景，可能与真实场景存在偏差，对模型在现实世界的泛化性能产生影响。

MMScan（Lyu等，2024）是由上海人工智能实验室联合多所大学共同构建的大规模多模态3D场景数据集，它基于EmbodiedScan，通过精心设计的语言提示和人工校正，生成了层次化的语言注释。该数据集包含1.4 M个文本描述，涵盖109 k个对象和7.7 k个区域，为3D视觉定位和问答任务提供了超过3.04 M的样本。MMScan的采集过程利用了自上而下的逻辑，从区域到对象级别，从单一目标到目标间关系，全面覆盖了空间和属性理解的各个方面。MMScan的高质量和多样性使其成为训练和评估3D-LLM的重要资源，特别是在提升模型在复杂空间和属性理解方面的表现上具有显著影响。

7 结语与展望

7.1 挑战和研究展望

尽管多模态大模型与3D视觉数据集成技术已经取得初步进展，但是表示载体的统一、模型应用推进以及落地后的纠正完善等挑战仍然严重限制了多模态大模型驱动的三维视觉进一步推广应用，因此综合分析当前的技术现状，多模态大模型在3D视觉领域的集成技术发展需要创新的解决方案以解决以下挑战：

1）3D表征载体的统一是3D视觉语言模型通用性的扩展关键。3D视觉的不同表示在不同应用场景具有不同的侧重性，例如点云主要用于表示室内（例如网格的顶点）和室外（例如LiDAR点云）环境，因为它们简单高效且易与大规模神经网络兼容。然而，点云空间表示很难捕捉到准确、丰富的空间模型细节，同时空间结构表征能力不强。而体素等结构化表示本征体现了3D数据在空间上结构上下文关系的同时，往往带来繁复的计算量以及有限分辨率的表征细粒度。

因此，开发能够更有效地弥合空间信息和语言之间差距的新型3D场景表示可以开启全新的理解与交互模式，整合表示体系优势。通过寻找在3D表示中编码语言和语义信息的创新方法，例如使用精简语言和语义嵌入，可以帮助弥合空间信息和语言模式之间的差距。

2）多模态大模型的空间推理能力提升是赋能3D视觉理解任务的关键。对空间关系的理解和推理是3D视觉理解任务的基本能力。虽然目前多模态大模型在某些视觉问答基准测试中表现出色，是功能强大的通用模型，适用于广泛的任务，但大多数最先进的VLM仍然在空间推理方面存在显著挑战，即需要理解物体在3D空间中的位置、识别物理对象的数量关系、距离或大小差异或它们之间的空间关系的任务。因此，研究如何解决多模态大模型训练数据中缺乏三维空间知识，并试图通过使用互联网规模的空间推理数据来训练，提升多模态大模型的空间推理能力，并进一步赋能3D视觉任务，十分关键。

3）多模态大模型驱动的3D视觉数据处理普适性受限难题仍然阻碍着技术的推广。随着3D环境的复杂度和语言模型尺度的增加，大模型技术仍然严重制约于个人用户硬件设备配置。当前多模态大模型需要一整套计算机集群进行高性能计算，大模型的训练与使用被少数能负担大型计算代价的研究者以及公司主导，不利于整体计算机社区参与到技术更新迭代中。因此，为强适应性和高计算效率而设计的多模态大模型架构，能够降低模型技术使用门槛，此项技术的进步可以大大拓宽其应用范围。

此外，通过仿真引擎获取的3D视觉数据虽然能够在一定程度上支撑多模态大模型的训练，仿真数据和真实场景之间的偏差往往会造成下游任务性能的下降。因此，如何通过新的泛化性和知识迁移方

法研究，在适应真实数据的同时保持模型的强泛化性和良好性能，非常重要。

4）3D 视觉任务下的多模态大模型落地完善机制至关重要。最后，当前多模态大模型与 3D 视觉集成技术的构建设计到应用落地是一个持续自我完善的过程，亟须提升完善机制纠正过程中的不足，使得多模态大模型技术更良好地理解 3D 任务目标，实现高效精准作业。

在模型性能提升方面：当前 3D 视觉感知理解任务基准的粒度过于简单，限制了对复杂 3D 环境理解的洞察。特别是在 3D 推理方面，基准的规模以及完整性阻碍了对空间推理技能的评估以及 3D 决策/交互系统的开发。此外，目前使用的基准数量不足以支撑多模态大模型在 3D 环境中的全部能力发挥。

在模型安全与道德方面：当前的模型设计仍旧处于技术实现阶段，没有充分考虑到模型落地应用之后的安全性以及道德性应用问题。多模态大模型经常不可预测且难以解释的方式输出不准确、不安全的信息，导致关键 3D 应用任务失效，容易造成国民经济的重大损失。同时，模型训练数据的社会偏见切实影响着 3D 场景预测的决策优先度，导致某些群体处于不利地位。因此，在 3D 环境任务作业过程中，构建多模态大模型完善机制，采用策略促使模型自我迭代补全模型缺陷，用于偏差检测和纠正的有力升级框架，确保多模态大模型在 3D 视觉领域的广泛落地，不断自我升级，输出有效且公正的结果。

5）面向边端侧场景的多模态大模型小型化和云边协同增强机器人 3D 视觉任务。面向机器人 3D 视觉感知等边端侧低算力场景，对多模态大模型提出更高的时延和性能要求。传统 3D 视觉任务小模型方案面临 3D 跨场景泛化性不足、任务效能低、重复开发交付成本高等问题。基于多模态大模型技术的 3D 场景泛化性优势，探索模型量化、剪枝和蒸馏等模型压缩技术，降低大模型应用门槛。此外，针对云边端多源异构算力等实际 3D 场景，结合多模态大模型通用能力和 3D 视觉任务小模型专用能力，实现云—边结合模式下的推理框架，研究大、小模型在协同规划、协同推理以及数据隔离等方面的技术实现，将是未来一个重要的研究方向。

7.2 总结

大型语言模型时代的 3D 视觉技术正愈发成为感知与理解世界结构的关键。本文立足于多模态大模型与 3D 视觉数据的集成技术，系统性回顾了多模态大模型在感知、理解和生成 3D 视觉数据方面的体系、架构以及改良应用，专注于充分发挥多模态大模型在一系列 3D 任务中的驱动能力与变革潜力，是实现当前语言模态模型智能推介至 3D 空间智能的基础。

其中，本文确认了多模态大模型相较于传统技术的独特优势，包括但不限于零样本学习、高级推理和世界知识认知领域，通过提炼并对齐大型语言模型中的文本信息与 3D 空间解释以弥合模态间的偏差，增强 3D 环境中的空间理解和交互，展示了多模态大模型与 3D 视觉数据集成技术在广泛任务上的成功应用案例，推动空间 AI 系统的发展。

本文还强调了统一表示、模型推广和落地完善等方向上面临的重大挑战，克服这些阻碍是保障多模态大模型切实落地的前置必要条件。总之，本文不仅全面概述了使用多模态大模型的 3D 视觉任务现状，还呼吁共同努力探索和扩展多模态大模型在复杂 3D 世界理解与交互方面的能力，为空间智能领域的持续性技术革新铺平道路。

参考文献（References）

Abdelrahman E，Ayman M，Ahmed M，Slim H and Elhoseiny M. 2024. CoT3DRef：chain-of-thoughts data-efficient 3D visual grounding［EB/OL］.［2024-09-20］. https://arxiv.org/pdf/2310.06214.pdf

Abdelreheem A，Olszewski K，Lee H Y，Wonka P and Achlioptas P. 2024. ScanEnts3D：exploiting phrase-to-3D-object correspondences for improved visio-linguistic models in 3D scenes//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa，USA：IEEE：3512-3522［DOI：10.1109/WACV57701.2024.00349］

Achlioptas P，Abdelreheem A，Xia F，Elhoseiny M and Guibas L. 2020. ReferIt3D：neural listeners for fine-grained 3D object identification in real-world scenes//Proceedings of the 16th European Conference on Computer Vision. Glasgow，UK：Springer：422-440［DOI：10.1007/978-3-030-58452-8_25］

Achlioptas P，Huang I，Sung M，Tulyakov S and Guibas L. 2023. ShapeTalk：a language dataset and framework for 3D shape edits and deformations//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：12685-12694［DOI：10.1109/CVPR52729.2023.01220］

Afham M，Dissanayake I，Dissanayake D，Dharmasiri A，Thilakarathna K and Rodrigo R. 2022. Crosspoint：self-supervised cross-modal contrastive learning for 3D point cloud understanding//Pro-

ceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 9892-9902 [DOI: 10.1109/CVPR52688.2022.00967]

Alayrac J B, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Millican K, Reynolds M, Ring R, Rutherford E, Cabi S, Han T D, Gong Z T, Samangooei S, Monteiro M, Menick J, Borgeaud S, Brock A, Nematzadeh A, Sharifzadeh S, Binkowski M, Barreira R, Vinyals O, Zisserman A and Simonyan K. 2022. Flamingo: a visual language model for few-shot learning//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 23716-23736

Antol S, Agrawal A, Lu J S, Mitchell M, Batra D, Zitnick C L and Parikh D. 2015. VQA: visual question answering//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 2425-2433 [DOI: 10.1109/ICCV.2015.279]

Arandjelovic R and Zisserman A. 2017. Look, listen and learn//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 609-617 [DOI: 10.1109/ICCV.2017.73]

Armeni I, He Z Y, Zamir A, Gwak J, Malik J, Fischer M and Savarese S. 2019. 3D scene graph: a structure for unified semantics, 3D space, and camera//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 5663-5672 [DOI: 10.1109/ICCV.2019.00576]

Azuma D, Miyanishi T, Kurita S and Kawanabe M. 2022. ScanQA: 3D question answering for spatial scene understanding//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 19107-19117 [DOI: 10.1109/CVPR52688.2022.01854]

Bahmani S, Skorokhodov I, Rong V, Wetzstein G, Guibas L, Wonka P, Tulyakov S, Park J J, Tagliasacchi A and Lindell D B. 2024. 4D-fy: text-to-4D generation using hybrid score distillation sampling [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2311.17984.pdf

Bai H T, Lyu Y, Jiang L T, Li S J, Lu H N, Lin X D and Wang L. 2024. CompoNeRF: text-guided multi-object compositional NeRF with editable 3D scene layout [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2303.13843.pdf

Bangalath H, Maaz M, Khattak M U, Khan S and Khan F S. 2022. Bridging the gap between object and image-level representations for open-vocabulary detection//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 33781-33794

Bao H B, Dong L, Piao S H and Wei F R. 2022. BEIT: BERT pre-training of image Transformers [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2106.08254.pdf

Bar-Hillel Y. 1960. The present status of automatic translation of languages. Advances in Computers, 1: 91-163 [DOI: 10.1016/S0065-2458(08)60607-5]

Barron J T, Mildenhall B, Tancik M, Hedman P, Martin-Brualla R and Srinivasan P P. 2021. Mip-NeRF: a multiscale representation for anti-aliasing neural radiance fields//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 5835-5844 [DOI: 10.1109/ICCV48922.2021.00580]

Bartsch A and Farimani A B. 2024. LLM-Craft: robotic crafting of elasto-plastic objects with large language models [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2406.08648.pdf

Boudjoghra M E A, Dai A, Lahoud J, Cholakkal H, Anwer R M, Khan S and Khan F S. 2025. Open-YOLO 3D: towards fast and accurate open-vocabulary 3D instance segmentation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2406.02548.pdf

Breyer M, Chung J J, Ott L, Siegwart R and Nieto J. 2021. Volumetric grasping network: real-time 6 DOF grasp detection in clutter [EB/OL]. [2021-01-04]. https://arxiv.org/pdf/2101.01132.pdf

Brown P F, deSouza P V, Mercer R L, Della Pietra V J and Lai J C. 1992. Class-based *n*-gram models of natural language. Computational Linguistics, 18(4): 467-479

Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D M, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I and Amodei D. 2020. Language models are few-shot learners//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 1877-1901

Caesar H, Bankiti V, Lang A H, Vora S, Liong V E, Xu Q, Krishnan A, Pan Y, Baldan G and Beijbom O. 2020. nuScenes: a multimodal dataset for autonomous driving//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11618-11628 [DOI: 10.1109/CVPR42600.2020.01164]

Cai D G, Zhao L C, Zhang J, Sheng L and Xu D. 2022. 3DJCG: a unified framework for joint dense captioning and visual grounding on 3D point clouds//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 16443-16452 [DOI: 10.1109/CVPR52688.2022.01597]

Cai J H, He Y S, Yuan W H, Zhu S Y, Dong Z L, Bo L F and Chen Q F. 2024. OV9D: category-levelopen-vocabulary 9D object pose and size estimation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2403.12396.pdf

Cao A and Johnson J. 2023. HexPlane: a fast representation for dynamic scenes//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 130-141 [DOI: 10.1109/CVPR52729.2023.00021]

Cao Y, Zeng Y H, Xu H and Xu D. 2023b. CoDA: collaborative novel box discovery and cross-modal alignment for open-vocabulary 3D object detection//Proceedings of the 37th International Conference

on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 71862-71873

Cao Y K, Cao Y P, Han K, Shan Y and Wong K Y K. 2024. Dreamavatar: text-and-shape guided 3D human avatar generation via diffusion models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 958-968 [DOI: 10.1109/CVPR52733.2024.00097]

Chang A, Dai A, Funkhouser T, Halber M, Nießner M, Savva M, Song S R, Zeng A and Zhang Y D. 2017. Matterport3D: learning from RGB-D data in indoor environments [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/1709.06158.pdf

Chang A X, Funkhouser T, Guibas L, Hanrahan P, Huang Q X, Li Z M, Savarese S, Savva M, Song S R, Su H, Xiao J X, Yi L and Yu F. 2015. ShapeNet: an information-rich 3D model repository [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/1512.03012.pdf

Chang Y, Ballotta L and Carlone L. 2023. D-Lite: navigation-oriented compression of 3D scene graphs for multi-robot collaboration. IEEE Robotics and Automation Letters, 8(11): 7527-7534 [DOI: 10.1109/LRA.2023.3320011]

Chatzipantazis E, Pertigkiozoglou S, Dobriban E and Daniilidis K. 2023. SE(3)-equivariant attention networks for shape reconstruction in function space [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2204.02394.pdf

Chen A P, Xu Z X, Zhao F Q, Zhang X S, Xiang F B, Yu J Y and Su H. 2021a. MVSNeRF: fast generalizable radiance field reconstruction from multi-view stereo//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 14104-14113 [DOI: 10.1109/ICCV48922.2021.01386]

Chen B Y, Xu Z, Kirmani S, Driess D, Florence P, Ichter B, Sadigh D, Guibas L and Xia F. 2024a. SpatialVLM: endowing vision-language models with spatial reasoning capabilities [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2401.12168.pdf

Chen D Z, Chang A X and Nießner M. 2020. ScanRefer: 3D object localization in RGB-D scans using natural language//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 202-221 [DOI: 10.1007/978-3-030-58565-5_13]

Chen D Z, Gholami A, Nießner M and Chang A X. 2021c. Scan2Cap: context-aware dense captioning in RGB-D scans//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 3192-3202 [DOI: 10.1109/CVPR46437.2021.00321]

Chen L G X, Wang X J, Lu J L, Lin S H, Wang C B and He G Q. 2024b. CLIP-driven open-vocabulary 3D scene graph generation via cross-modality contrastive learning//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 27863-27873 [DOI: 10.1109/CVPR52733.2024.02632]

Chen R, Chen Y W, Jiao N X and Jia K. 2023a. Fantasia3D: disentangling geometry and appearance for high-quality text-to-3D content creation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22189-22199 [DOI: 10.1109/ICCV51070.2023.02033]

Chen R N, Liu Y Q, Kong L D, Zhu X G, Ma Y X, Li Y K, Hou Y N, Qiao Y and Wang W P. 2023b. CLIP2Scene: towards label-efficient 3D scene understanding by CLIP//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 7020-7030 [DOI: 10.1109/CVPR52729.2023.00678]

Chen S H, Zhu X X, Liu W, He X J and Liu J. 2021b. Global-local propagation network for RGB-D semantic segmentation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2101.10801.pdf

Chen S J, Chen X, Zhang C, Li M S, Yu G, Fei H, Zhu H Y, Fan J Y and Chen T. 2023c. LL3DA: visual interactive instruction tuning for Omni-3D understanding, reasoning, and planning [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2311.18651.pdf

Chen S J, Zhu H Y, Chen X, Lei Y J, Yu G and Chen T. 2023d. End-to-end 3D dense captioning with Vote2Cap-DETR//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 11124-11133 [DOI: 10.1109/CVPR52729.2023.01070]

Chen S J, Zhu H Y, Li M S, Chen X, Guo P, Lei Y J, Yu G, Li T H and Chen T. 2023e. Vote2Cap-DETR++: decoupling localization and describing for end-to-end 3D dense captioning [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2309.02999.pdf

Chen T R, Yu C, Li J, Zhang J Q, Zhu L Y, Ji D Y, Zhang Y, Zang Y, Li Z J and Sun L Y. 2024c. Reasoning3D — grounding and reasoning in 3D: fine-grained zero-shot open-vocabulary 3D reasoning part segmentation via large vision-language models [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2405.19326.pdf

Chen X, Wang X, Changpinyo S, Piergiovanni A J, Padlewski P, Salz D, Goodman S, Grycner A, Mustafa B, Beyer L, Kolesnikov A, Puigcerver J, Ding N, Rong K R, Akbari H, Mishra G, Xue L T, Thapliyal A, Bradbury J, Kuo W C, Seyedhosseini M, Jia C, Ayan B K, Riquelme C, Steiner A, Angelova A, Zhai X H, Houlsby N and Soricut R. 2023f. PaLI: a jointly-scaled multilingual language-image model [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2209.06794.pdf

Chen Y L, Yang S, Huang H F, Wang T, Xu R S, Lyu R Y, Lin D H and Peng J M. 2024d. Grounded 3D-LLM with referent tokens [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2405.10370.pdf

Cheng T H, Song L, Ge Y X, Liu W Y, Wang X G and Shan Y. 2024. YOLO-world: real-time open-vocabulary object detection [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2401.17270.pdf

Chibane J, Alldieck T and Pons-Moll G. 2020b. Implicit functions in feature space for 3D shape reconstruction and completion//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6968-6979 [DOI: 10.1109/CVPR42600.2020.00700]

Chibane J, Mir A and Pons-Moll G. 2020a. Neural unsigned distance fields for implicit function learning//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 21638-21652

Chou G N, Bahat Y and Heide F. 2023. Diffusion-SDF: conditional generative modeling of signed distance functions//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 2262-2272 [DOI: 10.1109/ICCV51070.2023.00215]

Choy C B, Xu D F, Gwak J, Chen K and Savarese S. 2016. 3D-R2N2: a unified approach for single and multi-view 3D object reconstruction//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 628-644 [DOI: 10.1007/978-3-319-46484-8_38]

Chu T, Zhang P, Dong X Y, Zang Y H, Liu Q and Wang J Q. 2024. Unified scene representation and reconstruction for 3D large language models [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2404.13044.pdf

Chu Y F, Xu J, Zhou X H, Yang Q, Zhang S L, Yan Z J, Yan Z, Zhou C and Zhou J R. 2023. Qwen-Audio: advancing universal audio understanding via unified large-scale audio-language models [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2311.07919.pdf

Chung J, Oh J and Lee K M. 2024. Depth-regularized optimization for 3D Gaussian splatting in few-shot images [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2311.13398.pdf

Cong P S, Zhu X G, Qiao F, Ren Y M, Peng X D, Hou Y N, Xu L, Yang R G, Manocha D and Ma Y X. 2022. STCrowd: a multimodal dataset for pedestrian perception in crowded scenes//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 19576-19585 [DOI: 10.1109/CVPR52688.2022.01899]

Dai A, Chang A X, Savva M, Halber M, Funkhouser T and Nießner M. 2017. ScanNet: richly-annotated 3D reconstructions of indoor scenes//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2432-2443 [DOI: 10.1109/CVPR.2017.261]

Dai Z R, Asgharivaskasi A, Duong T, Lin S S, Tzes M E, Pappas G and Atanasov N. 2024. Optimal scene graph planning with large language model guidance [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2309.09182.pdf

Das A, Datta S, Gkioxari G, Lee S, Parikh D and Batra D. 2018. Embodied question answering//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1-10 [DOI: 10.1109/CVPR.2018.00008]

Deitke M, Liu R S, Wallingford M, Ngo H, Michel O, Kusupati A, Fan A, Laforte C, Voleti V, Gadre S Y, VanderBilt E, Kembhavi A, Vondrick C, Gkioxari G, Ehsani K, Schmidt L and Farhadi A. 2023a. Objaverse-XL: a universe of 10M+ 3D objects//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 35799-35813

Deitke M, Schwenk D, Salvador J, Weihs L, Michel O, VanderBilt E, Schmidt L, Ehsanit K, Kembhavi A and Farhadi A. 2023b. Objaverse: a universe of annotated 3D objects//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 13142-13153 [DOI: 10.1109/CVPR52729.2023.01263]

Deng X, Zhang W Y, Ding Q and Zhang X M. 2023. PointVector: a vector representation in point cloud analysis//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 9455-9465 [DOI: 10.1109/CVPR52729.2023.00912]

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional Transformers for language understanding [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/1810.04805.pdf

Ding R Y, Yang J H, Xue C H, Zhang W Q, Bai S and Qi X J. 2023. PLA: language-driven open-vocabulary 3D scene understanding//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 7010-7019 [DOI: 10.1109/CVPR52729.2023.00677]

Driess D, Xia F, Sajjadi M S, Lynch C, Chowdhery A, Ichter B, Wahid A, Tompson J, Vuong Q, Yu T H, Huang W L, Chebotar Y, Sermanet P, Duckworth D, Levine S, Vanhoucke V, Hausman K, Toussaint M, Greff K, Zeng A, Mordatch I and Florence P. 2023. PaLM-E: an embodied multimodal language model [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2303.03378.pdf

Etchegaray D, Huang Z, Harada T and Luo Y D. 2024. Find n' Propagate: open-vocabulary 3D object detection in urban environments [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2403.13556.pdf

Fei B, Li Y X, Yang W D, Ma L P and He Y. 2024a. Towards unified representation of multi-modal pre-training for 3D understanding via differentiable rendering [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2404.13619.pdf

Fei H, Wu S, Zhang H, Chua T S and Yan S. 2024b. Vitron: a unified pixel-level vision llm for understanding, generating, segmenting, editing [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2412.19806.pdf

Fei J J, Ahmed M, Ding J, Bakr E M and Elhoseiny M. 2024c. Kestrel: point grounding multimodal LLM for part-aware 3D vision-language understanding [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2405.18937.pdf

Feng C, Hsu J, Liu W Y and Wu J J. 2024. Naturally supervised 3D visual grounding with language-regularized concept learners [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2404.19696.pdf

Feng M T, Hou H R, Zhang L, Guo Y L, Yu H S, Wang Y N and Mian A. 2025. Exploring hierarchical spatial layout cues for 3D point cloud based scene graph prediction. IEEE Transactions on Multimedia, 27: 731-743 [DOI: 10.1109/TMM.2023.3277736]

Feng Q, Liu Y B, Lai Y K, Yang J Y and Li K. 2022. FOF: learning fourier occupancy field for monocular real-time human reconstruction//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 7397-7409

Feng Y T, Feng Y F, You H X, Zhao X B and Gao Y. 2019. MeshNet: mesh neural network for 3D shape representation//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI: 8279-8286 [DOI: 10.1609/aaai.v33i01.33018279]

Fine S, Singer Y and Tishby N. 1998. The hierarchical hidden Markov model: analysis and applications. Machine Learning, 32(1): 41-62 [DOI: 10.1023/A:1007469218079]

Fu H, Cai B W, Gao L, Zhang L X, Wang J M, Li C, Zeng Q X, Sun C Y, Jia R F, Zhao B Q and Zhang H. 2021a. 3D-FRONT: 3D furnished rooms with layouts and semantics//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 10913-10922 [DOI: 10.1109/ICCV48922.2021.01075]

Fu H, Jia R F, Gao L, Gong M M, Zhao B Q, Maybank S and Tao D C. 2021b. 3D-Future: 3D furniture shape with texture. International Journal of Computer Vision, 129(12): 3313-3337 [DOI: 10.1007/s11263-021-01534-z]

Fu R, Liu J Y, Chen X L, Nie Y X and Xiong W H. 2024. Scene-LLM: extending language model for 3D visual understanding and reasoning [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2403.11401.pdf

Gao G G, Liu W Y, Chen A P, Geiger A and Schölkopf B. 2024. GraphDreamer: compositional 3D scene synthesis from scene graphs [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2312.00093.pdf

Gao J, Shen T C, Wang Z A, Chen W Z, Yin K X, Li D Q, Litany O, Gojcic Z and Fidler S. 2022. GET3D: a generative model of high quality 3D textured shapes learned from images//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 31841-31854

Ge J X, Luo H Y, Qian S Y, Gan Y L, Fu J and Zhang S H. 2023. Chain of thought prompt tuning for vision-language models [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2304.07919.pdf

Gemini Team Google. 2024. Gemini: a family of highly capable multimodal models [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2312.11805.pdf

Girdhar R, El-Nouby A, Liu Z, Singh M, Alwala K V, Joulin A and Misra I. 2023. ImageBind: one embedding space to bind them all//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 15180-15190 [DOI: 10.1109/CVPR52729.2023.01457]

Girish S, Gupta K and Shrivastava A. 2024. Eagles: efficient accelerated 3D Gaussians with lightweight encodings [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2312.04564.pdf

Goyal A, Xu J, Guo Y J, Blukis V, Chao Y W and Fox D. 2023. RVT: robotic view transformer for 3D object manipulation//Proceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR: 694-710 [DOI: 10.48550/arXiv.2306.14896]

Gu J T, Trevithick A, Lin K E, Susskind J M, Theobalt C, Liu L J and Ramamoorthi R. 2023a. NerfDiff: single-image view synthesis with nerf-guided distillation from 3D-aware diffusion//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR: 11808-11826

Gu Q, Kuwajerwala A, Morin S, Jatavallabhula K M, Sen B, Agarwal A, Rivera C, Paul W, Ellis K, Chellappa R, Gan C, de Melo C M, Tenenbaum J B, Torralba A, Shkurti F and Paull L. 2023b. ConceptGraphs: open-vocabulary 3D scene graphs for perception and planning [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2309.16650.pdf

Guo M H, Cai J X, Liu Z N, Mu T J, Martin R R and Hu S M. 2021. PCT: point cloud transformer. Computational Visual Media, 7(2): 187-199 [DOI: 10.1007/s41095-021-0229-5]

Guo Z, Tang Y W, Zhang R, Wang D, Wang Z G, Zhao B and Li X L. 2023a. ViewRefer: grasp the multi-view knowledge for 3D visual grounding//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 15326-15337 [DOI: 10.1109/ICCV51070.2023.01410]

Guo Z Y, Zhang R R, Zhu X Y, Tang Y W, Ma X Z, Han J M, Chen K X, Gao P, Li X Z, Li H S and Heng P A. 2023b. Point-Bind and Point-LLM: aligning point cloud with multi-modality for 3D understanding, generation, and instruction following [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2309.00615.pdf

Gupta A, Xiong W H, Nie Y X, Jones I and Oğuz B. 2023. 3DGen: triplane latent diffusion for textured mesh generation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2303.05371.pdf

Han D G, McInroe T, Jelley A, Albrecht S V, Bell P and Storkey A. 2024a. LLM-Personalize: aligning LLM planners with human preferences via reinforced self-training for housekeeping robots [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2404.14285.pdf

Han J M, Gong K X, Zhang Y Y, Wang J Q, Zhang K P, Lin D H, Qiao Y, Gao P and Yue X Y. 2024b. OneLLM: one framework to align all modalities with language [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2312.03700.pdf

Han X, Tang Y, Wang Z X and Li X Z. 2024c. Mamba3D: enhancing local features for 3D point cloud analysis via state space model [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2404.14966.pdf

Han Z Z, Wang X Y, Vong C M, Liu Y S, Zwicker M and Chen C L P. 2019. 3DViewGraph: learning global features for 3D shapes from a graph of unordered views with attention [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/1905.07503.pdf

Hanocka R, Hertz A, Fish N, Giryes R, Fleishman S and Cohen-Or D. 2019. MeshCNN: a network with an edge. ACM Transactions on Graphics, 38(4): #90 [DOI: 10.1145/3306346.3322959]

He Q D, Peng J L, Jiang Z K, Wu K, Ji X Z, Zhang J N, Wang Y B,

Wang C J, Chen M G and Wu Y S. 2024. UniM-OV3D: unimodality open-vocabulary 3D scene understanding with fine-grained feature representation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2401.11395.pdf

Hegde D, Valanarasu J M J and Patel V M. 2023. CLIP goes 3D: leveraging prompt tuning for language grounded 3D recognition//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision Workshops. Paris, France: IEEE: 2020-2030 [DOI: 10.1109/ICCVW60793.2023.00217]

Hess G, Tonderski A, Petersson C, Åström K and Svensson L. 2024. LidarCLIP or: how I learned to talk to point clouds//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 7423-7432 [DOI: 10.1109/WACV57701.2024.00727]

Hochreiter S and Schmidhuber J. 1997. Long short-term memory. Neural Computation, 9(8): 1735-1780 [DOI: 10.1162/neco.1997.9.8.1735]

Honerkamp D, Büchner M, Despinoy F, Welschehold T and Valada A. 2024. Language-grounded dynamic scene graphs for interactive object search with mobile manipulation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2403.08605.pdf

Hong F Z, Chen Z X, Lan Y S, Pan L and Liu Z W. 2022a. EVA3D: compositional 3D human generation from 2D image collections [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2210.04888.pdf

Hong F Z, Zhang M Y, Pan L, Cai Z G, Yang L and Liu Z W. 2022b. AvatarCLIP: zero-shot text-driven generation and animation of 3D avatars. ACM Transactions on Graphics, 41(4): #161 [DOI: 10.1145/3528223.3530094]

Hong S M, Yavartanoo M, Neshatavar R and Lee K M. 2023a. ACL-SPC: adaptive closed-loop system for self-supervised point cloud completion //Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 9435-9444 [DOI: 10.1109/CVPR52729.2023.00910]

Hong Y N, Lin C R, Du Y L, Chen Z F, Tenenbaum J B and Gan C. 2023b. 3D concept learning and reasoning from multi-view images//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 9202-9212 [DOI: 10.1109/CVPR52729.2023.00888]

Hong Y N, Zhen H Y, Chen P H, Zheng S H, Du Y L, Chen Z F and Gan C. 2023c. 3D-LLM: injecting the 3D world into large language models//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 20482-20494

Hong Y N, Zheng Z S, Chen P H, Wang Y A, Li J Y and Gan C. 2024. MultiPLY: a multisensory object-centric embodied large language model in 3D world [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2401.08577.pdf

Huang C G, Mees O, Zeng A and Burgard W. 2023a. Visual language maps for robot navigation//Proceedings of 2023 IEEE International Conference on Robotics and Automation. London, UK: IEEE: 10608-10615 [DOI: 10.1109/ICRA48891.2023.10160969]

Huang J Y, Yong S L, Ma X J, Linghu X K, Li P H, Wang Y, Li Q, Zhu S C, Jia B X and Huang S Y. 2024a. An embodied generalist agent in 3D world [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2311.12871.pdf

Huang J C, Zhang H T, Zhao M B and Wu Z. 2024b. IVLMap: instance-aware visual language grounding for consumer robot navigation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2403.19336.pdf

Huang K X, Yang J R, Wang J, He S F, Wang Z, He H Y, Zhang Q F and Lu G D. 2024c. Granular3D: delving into multi-granularity 3D scene graph prediction. Pattern Recognition, 153: #110562 [DOI: 10.1016/j.patcog.2024.110562]

Huang R, Pan X R, Zheng H, Jiang H J, Xie Z F, Wu C, Song S J and Huang G. 2024d. Joint representation learning for text and 3D point cloud. Pattern Recognition, 147: #110086 [DOI: 10.1016/j.patcog.2023.110086]

Huang S Y, Wang Z, Li P H, Jia B X, Liu T Y, Zhu Y X, Liang W and Zhu S C. 2023b. Diffusion-based generation, optimization, and planning in 3D scenes//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 16750-16761 [DOI: 10.1109/CVPR52729.2023.01607]

Huang T Y, Dong B W, Yang Y H, Huang X S, Lau R W H, Ouyang W L and Zuo W M. 2023c. CLIP2Point: transfer CLIP to point cloud classification with image-depth pre-training//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22100-22110 [DOI: 10.1109/ICCV51070.2023.02025]

Huang W L, Wang C, Zhang R H, Li Y Z, Wu J J and Li F F. 2023d. VoxPoser: composable 3D value maps for robotic manipulation with language models [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2307.05973.pdf

Huang X, Shao R Z, Zhang Q, Zhang H W, Feng Y, Liu Y B and Wang Q. 2024f. HumanNorm: learning normal diffusion model for high-quality and realistic 3D human generation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4568-4577 [DOI: 10.1109/CVPR52733.2024.00437]

Huang X S, Huang Z, Li S, Qu W T, He T, Hou Y N, Zuo Y F and Ouyang W L. 2024e. Frozen CLIP transformer is an efficient point cloud encoder//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 2382-2390 [DOI: 10.1609/aaai.v38i3.28013]

Huang Y, Du C Z, Xue Z H, Chen X Y, Zhao H and Huang L B. 2021. What makes multi-modal learning better than single (provably)//Proceedings of the 35th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 10944-10956

Hughes N, Chang Y and Carlone L. 2022. Hydra: a real-time spatial

perception system for 3D scene graph construction and optimization [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2201.13360.pdf

Jain A, Mildenhall B, Barron J T, Abbeel P and Poole B. 2022. Zero-shot text-guided object generation with dream fields//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 857-866 [DOI: 10.1109/CVPR52688.2022.00094]

Ji J Y, Wang H W, Wu C L, Ma Y W, Sun X S and Ji R R. 2024. JM3D and JM3D-LLM: elevating 3D understanding with joint multi-modal cues [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2310.09503.pdf

Jia B X, Chen Y X, Yu H Y, Wang Y, Niu X S, Liu T Y, Li Q and Huang S Y. 2024. SceneVerse: scaling 3D vision-language learning for grounded scene understanding [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2401.09340.pdf

Jiang L, Shi S S and Schiele B. 2024. Open-vocabulary 3D semantic segmentation with foundation models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 21284-21294 [DOI: 10.1109/CVPR52733.2024.02011]

Jiang R X, Wang C, Zhang J B, Chai M L, He M M, Chen D D and Liao J. 2023. Avatarcraft: transforming text into neural human avatars with parameterized shape and pose control//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 14325-14336 [DOI: 10.1109/ICCV51070.2023.01322]

Jin Z, Hayat M, Yang Y W, Guo Y L and Lei Y J. 2023. Context-aware alignment and mutual masking for 3D-language pre-training//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 10984-10994 [DOI: 10.1109/CVPR52729.2023.01057]

Kerbl B, Kopanas G, Leimkühler T and Drettakis G. 2023. 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 42(4): #139 [DOI: 10.1145/3592433]

Khurana M, Peri N, Hays J and Ramanan D. 2024. Shelf-supervised cross-modal pre-training for 3D object detection [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2406.10115.pdf

Kim B, Kwon P, Lee K, Lee M, Han S, Kim D and Joo H. 2023a. Chupa: carving 3D clothed humans from skinned shape priors using 2D diffusion probabilistic models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 15919-15930 [DOI: 10.1109/ICCV51070.2023.01463]

Kim M, Lee J and Kim J. 2023b. GMR-Net: GCN-based mesh refinement framework for elliptic pde problems. Engineering with Computers, 39(5): 3721-3737 [DOI: 10.1007/s00366-023-01811-0]

Ko H K, Park G, Jeon H, Jo J, Kim J and Seo J. 2023. Large-scale text-to-image generation models for visual artists' creative works [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2210.08477.pdf

Koch S, Vaskevicius N, Colosi M, Hermosilla P and Ropinski T. 2024. Open3DSG: open-vocabulary 3D scene graphs from point clouds with queryable objects and open-set relationships [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2402.12259.pdf

Koh J Y, Fried D and Salakhutdinov R. 2023. Generating images with multimodal language models//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 21487-21506

Kumra S, Joshi S and Sahin F. 2020. Antipodal robotic grasping using generative residual convolutional neural network//Proceedings of 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas, USA: IEEE: 9626-9633 [DOI: 10.1109/IROS45743.2020.9340777]

Lee J C, Rho D, Sun X Y, Ko J H and Park E. 2024. Compact 3D Gaussian representation for radiance field [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2311.13681.pdf

Lei Y J, Xu K, Guo Y L, Yang X, Wu Y W, Hu W, Yang J Q and Wang H Y. 2024. Comprehensive survey on 3D visual-language understanding techniques. Journal of Image and Graphics, 29(6): 1747-1764 (雷印杰, 徐凯, 郭裕兰, 杨鑫, 武玉伟, 胡玮, 杨佳琪, 汪汉云. 2024. "三维视觉—语言"推理技术的前沿研究与最新趋势. 中国图象图形学报, 29(6): 1747-1764) [DOI: 10.11834/jig.240029]

Leng Z Y, Birdal T, Liang X H and Tombari F. 2024. HyperSDFusion: bridging hierarchical structures in language and geometry for enhanced 3D Text2Shape generation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2403.00372.pdf

Li B Y, Weinberger K Q, Belongie S, Koltun V and Ranftl R. 2022a. Language-driven semantic segmentation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2201.03546.pdf

Li J, Liu Z, Li L, Lin J Q, Yao J and Tu J M. 2023b. Multi-view convolutional vision transformer for 3D object recognition. Journal of Visual Communication and Image Representation, 95: #103906 [DOI: 10.1016/j.jvcir.2023.103906]

Li J H, Zhang J W, Bai X, Zheng J, Ning X, Zhou J and Gu L. 2024a. DNGaussian: optimizing sparse-view 3D Gaussian radiance fields with global-local depth normalization [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2403.06912.pdf

Li J N, Li D X, Savarese S and Hoi S. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: 19730-19742

Li J N, Li D X, Xiong C M and Hoi S. 2022b. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR: 12888-12900

Li K C, He Y, Wang Y, Li Y Z, Wang W H, Luo P, Wang Y L, Wang L M and Qiao Y. 2024b. VideoChat: chat-centric video understanding [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2305.06355.pdf

Li K L, Wang J B, Yang L X, Lu C W and Dai B. 2024c. Semgrasp:

semantic grasp generation via language aligned discretization [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2404.03590.pdf

Li M S, Chen X, Zhang C, Chen S J, Zhu H Y, Yin F K, Yu G and Chen T. 2023c. M3DBench: let's instruct large models with multi-modal 3D prompts [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2312.10763.pdf

Li P F, He F Z, Fan B and Song Y P. 2023d. TPNet: a novel mesh analysis method via topology preservation and perception enhancement. Computer Aided Geometric Design, 104: #102219 [DOI: 10.1016/j.cagd.2023.102219]

Li T Y, Slavcheva M, Zollhoefer M, Green S, Lassner C, Kim C, Schmidt T, Lovegrove S, Goesele M, Newcombe R and Lyu Z Y. 2022c. Neural 3D video synthesis from multi-view video//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5511-5521 [DOI: 10.1109/CVPR52688.2022.00544]

Li X, Ding J, Chen Z Y and Elhoseiny M. 2023e. Uni3DL: unified model for 3D and language understanding [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2312.03026.pdf

Li Y X, Yang W D and Fei B. 2024d. 3DMambaComplete: exploring structured state space model for point cloud completion [EB/OL]. [2024-09-20].https://arxiv.org/pdf/2404.07106.pdf

Li Z J, Zhang C, Wang X Y, Ren R L, Xu Y F, Ma R F and Liu X D. 2024e. 3DMIT: 3D multi-modal instruction tuning for scene understanding [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2401.03201.pdf

Liang D K, Zhou X, Xu W, Zhu X K, Zou Z K, Yu X Q, Tan X and Bai X. 2024. PointMamba: a simple state space model for point cloud analysis [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2402.10739.pdf

Liang F, Wu B C, Dai X L, Li K P, Zhao Y N, Zhang H, Zhang P Z, Vajda P and Marculescu D. 2023a. Open-vocabulary semantic segmentation with mask-adapted CLIP//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 7061-7070 [DOI: 10.1109/CVPR52729.2023.00682]

Liang J, Huang W L, Xia F, Xu P, Hausman K, Ichter B, Florence P and Zeng A. 2023b. Code as policies: language model programs for embodied control//Proceedings of 2023 IEEE International Conference on Robotics and Automation. London, UK: IEEE: 9493-9500 [DOI: 10.1109/ICRA48891.2023.10160591]

Lin C H, Gao J, Tang L M, Takikawa T, Zeng X H, Huang X, Kreis K, Fidler S, Liu M Y and Lin T Y. 2023a. Magic3D: high-resolution text-to-3D content creation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 300-309 [DOI: 10.1109/CVPR52729.2023.00037]

Lin K E, Lin Y C, Lai W S, Lin T Y, Shih Y C and Ramamoorthi R. 2023b. Vision transformer for NeRF-based view synthesis from a single input image//Proceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 806-815 [DOI: 10.1109/WACV56688.2023.00087]

Liu D N, Huang X S, Hou Y N, Wang Z H, Yin Z F, Gong Y S, Gao P and Ouyang W L. 2024a. Uni3D-LLM: unifying point cloud perception, generation and editing with large language models [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2402.03327.pdf

Liu H T, Li C Y, Wu Q Y and Lee Y J. 2023a. Visual instruction tuning//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 34892-34916

Liu M, Qin D X, Han Y B, Chen X and Wang Y N. 2024. A multi-scale dynamic visual network for surgical robots scene segmentation. Journal of Image and Graphics (刘敏, 秦敦璇, 韩雨斌, 陈祥, 王耀南. 2024. 多尺度动态视觉网络的手术机器人场景分割. 中国图象图形学报) [DOI: 10.11834/jig.240385]

Liu M H, Shi R X, Kuang K M, Zhu Y H, Li X L, Han S Z, Cai H, Porikli F and Su H. 2023b. OpenShape: scaling up 3D shape representation towards open-world understanding//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 44860-44879

Liu R, Wang X H, Wang W G and Yang Y. 2023c. Bird's-eye-view scene graph for vision-language navigation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 10934-10946 [DOI: 10.1109/ICCV51070.2023.01007]

Liu Y T, Wang L, Yang J, Chen W K, Meng X X, Yang B and Gao L. 2024b. NeUDF: learning neural unsigned distance fields with volume rendering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(4): 2364-2377 [DOI: 10.1109/TPAMI.2023.3335353]

Liu Z Z, Hu J Y, Hui K H, Qi X J, Cohen-Or D and Fu C W. 2023d. EXIM: a hybrid explicit-implicit representation for text-guided 3D shape generation. ACM Transactions on Graphics, 42(6): #228 [DOI: 10.1145/3618312]

Long X X, Lin C, Liu L J, Liu Y, Wang P, Theobalt C, Komura T and Wang W P. 2023. NeuralUDF: learning unsigned distance fields for multi-view reconstruction of surfaces with arbitrary topologies//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 20834-20843 [DOI: 10.1109/CVPR52729.2023.01996]

Loo J, Wu Z X and Hsu D. 2024. Open scene graphs for open world object-goal navigation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2407.02473.pdf

Loper M, Mahmood N, Romero J, Pons-Moll G and Black M J. 2015. SMPL: a skinned multi-person linear model. ACM Transactions on Graphics, 34(6): #248 [DOI: 10.1145/2816795.2818013]

Lu P, Peng B L, Cheng H, Galley M, Chang K W, Wu Y N, Zhu S

and Gao J F. 2023a. Chameleon: plug-and-play compositional reasoning with large language models//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 43447-43478

Lu S Y, Chang H N, Jing E P, Boularias A and Bekris K. 2023b. OVIR-3D: open-vocabulary 3D instance retrieval without training on 3D data//Proceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR: 1610-1620

Lu T, Yu M L, Xu L N, Xiangli Y B, Wang L M, Lin D H and Dai B. 2023c. Scaffold-GS: structured 3D Gaussians for view-adaptive rendering [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2312.00109.pdf

Lu Y H, Xu C F, Wei X B, Xie X D, Tomizuka M, Keutzer K and Zhang S H. 2023d. Open-vocabulary point-cloud object detection without 3D annotation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 1190-1199 [DOI: 10.1109/CVPR52729.2023.00121]

Lu Y X, Zhang J Y, Li S W, Fang T, McKinnon D, Tsin Y, Quan L, Cao X and Yao Y. 2024. Direct2.5: diverse text-to-3D generation via multi-view 2.5D diffusion [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2311.15980.pdf

Lundell J, Verdoja F and Kyrki V. 2020. Beyond top-grasps through scene completion//Proceedings of 2020 IEEE International Conference on Robotics and Automation. Paris, France: IEEE: 545-551 [DOI: 10.1109/ICRA40945.2020.9197320]

Luo H S, Bao J W, Wu Y Z, He X D and Li T R. 2023a. SegCLIP: patch aggregation with learnable centers for open-vocabulary semantic segmentation//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR: 23033-23044

Luo J Y, Fu J H, Kong X H, Gao C, Ren H B, Shen H, Xia H X and Liu S. 2022. 3D-SPS: single-stage 3D visual grounding via referred point progressive selection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 16433-16442 [DOI: 10.1109/CVPR52688.2022.01596]

Luo T G, Rockwell C, Lee H and Johnson J. 2023b. Scalable 3D captioning with pretrained models//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 75307-75337

Lyu R, Wang T, Lin J L, Yang S, Mao X H, Chen Y L, Xu R S, Huang H F, Zhu C M, Lin D H and Pang J M. 2024. MMScan: a multi-modal 3D scene dataset with hierarchical grounded language annotations [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2406.09401.pdf

Lyu Z, Wang J Y, An Y W, Zhang Y, Lin D H and Dai B. 2023. Controllable mesh generation through sparse latent point diffusion models//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 271-280 [DOI: 10.1109/CVPR52729.2023.00034]

Ma C F, Jiang Y, Wen X, Yuan Z H and Qi X J. 2023. CoDet: co-occurrence guided region-word alignment for open-vocabulary object detection//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 71078-71094

Ma X J, Yong S L, Zheng Z L, Li Q, Liang Y T, Zhu S C and Huang S Y. 2023. SQA3D: situated question answering in 3D scenes [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2210.07474.pdf

Man Y Z, Gui L Y and Wang Y X. 2024. Situational awareness matters in 3D vision language reasoning [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2406.07544.pdf

Mao J H, Xu W, Yang Y, Wang J, Huang Z H and Yuille A. 2015. Deep captioning with multimodal recurrent neural networks (M-RNN) [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/1412.6632.pdf

Maturana D and Scherer S. 2015. VoxNet: a 3D convolutional neural network for real-time object recognition//Proceedings of 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems. Hamburg, Germany: IEEE: 922-928 [DOI: 10.1109/IROS.2015.7353481]

Mescheder L, Oechsle M, Niemeyer M, Nowozin S and Geiger A. 2019. Occupancy networks: learning 3D reconstruction in function space//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4455-4465 [DOI: 10.1109/CVPR.2019.00459]

Metzer G, Richardson E, Patashnik O, Giryes R and Cohen-Or D. 2023. Latent-NeRF for shape-guided generation of 3D shapes and textures//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 12663-12673 [DOI: 10.1109/CVPR52729.2023.01218]

Michel O, Bar-On R, Liu R, Benaim S and Hanocka R. 2022. Text2Mesh: text-driven neural stylization for meshes//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13482-13492 [DOI: 10.1109/CVPR52688.2022.01313]

Mikami Y, Melnik A, Miura J and Hautamäki V. 2024. Natural language as policies: reasoning for coordinate-level embodied control with LLMs [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2403.13801.pdf

Mikolov T, Chen K, Corrado G and Dean J. 2013. Efficient estimation of word representations in vector space [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/1301.3781.pdf

Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R. 2020. NeRF: representing scenes as neural radiance fields for view synthesis//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 405-421 [DOI: 10.1007/978-3-030-58452-8_24]

Minderer M, Gritsenko A and Houlsby N. 2023. Scaling open-vocabulary object detection//Proceedings of the 37th International

Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 72983-73007

Mittal P, Cheng Y C, Singh M and Tulsiani S. 2022. AutoSDF: shape priors for 3D completion, reconstruction and generation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 306-315 [DOI: 10.1109/CVPR52688.2022.00040]

Mo K C, Zhu S L, Chang A X, Yi L, Tripathi S, Guibas L J and Su H. 2019. PartNet: a large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 909-918 [DOI: 10.1109/CVPR.2019.00100]

Morrison D, Corke P and Leitner J. 2018. Closing the loop for robotic grasping: a real-time, generative grasp synthesis approach [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/1804.05172.pdf

Mu Y, Zhang Q L, Hu M K, Wang W H, Ding M Y, Jin J, Wang B, Dai J F, Qiao Y and Luo P. 2023. EmbodiedGPT: vision-language pre-training via embodied chain of thought//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 25081-25094 [DOI: 10.5555/3666122.3667212]

Navaneet K L, Meibodi K P, Koohpayegani S A and Pirsiavash H. 2024. CompGS: smaller and faster Gaussian splatting with vector quantization [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2311.18159.pdf

OpenAI. 2024. GPT-4 technical report [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2303.08774.pdf

Osher S, Fedkiw R and Piechor K. 2004. Level set methods and dynamic implicit surfaces. Applied Mechanics Reviews, 57(3): #B15 [DOI: 10.1115/1.1760520]

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C L, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J and Lowe R. 2022. Training language models to follow instructions with human feedback//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 27730-27744

Pang Y T, Wang W X, Tay F E H, Liu W, Tian Y H and Yuan L. 2022. Masked autoencoders for point cloud self-supervised learning//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 604-621 [DOI: 10.1007/978-3-031-20086-1_35]

Parelli M, Delitzas A, Hars N, Vlassis G, Anagnostidis S, Bachmann G and Hofmann T. 2023. CLIP-guided vision-language pre-training for question answering in 3D scenes//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver, Canada: IEEE: 5607-5612 [DOI: 10.1109/CVPRW59228.2023.00593]

Park J J, Florence P, Straub J, Newcombe R and Lovegrove S. 2019. DeepSDF: learning continuous signed distance functions for shape representation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 165-174 [DOI: 10.1109/CVPR.2019.00025]

Patel B, Dorbala V S, Manocha D and Bedi A S. 2024. Embodied question answering via Multi-LLM systems [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2406.10918v3.pdf

Peng H Y, Li B P, Zhang B, Chen X, Chen T and Zhu H Y. 2024. Multi-view vision fusion network: can 2D pre-trained model boost 3D point cloud data-scarce learning? IEEE Transactions on Circuits and Systems for Video Technology, 34(7): 5951-5962 [DOI: 10.1109/TCSVT.2023.3343495]

Peng S Y, Genova K, Jiang C Y, Tagliasacchi A, Pollefeys M and Funkhouser T. 2023. OpenScene: 3D scene understanding with open vocabularies//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 815-824 [DOI: 10.1109/CVPR52729.2023.00085]

Peng S Y, Niemeyer M, Mescheder L, Pollefeys M and Geiger A. 2020. Convolutional occupancy networks//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 523-540 [DOI: 10.1007/978-3-030-58580-8_31]

Piekenbrinck J, Hermans A, Vaskevicius N, Linder T and Leibe B. 2024. RGB-D Cube R-CNN: 3D object detection with selective modality dropout//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1997-2006 [DOI: 10.1109/CVPRW63382.2024.00205]

Po R and Wetzstein G. 2024. Compositional 3D scene generation using locally conditioned diffusion//Proceedings of 2024 International Conference on 3D Vision. Davos, Switzerland: IEEE: 651-663 [DOI: 10.1109/3DV62453.2024.00026]

Qi C, Yin J Q, Zhang Z C and Tang J. 2024a. Dynamic scene graph generation of point clouds with structural representation learning. Tsinghua Science and Technology, 29(1): 232-243 [DOI: 10.26599/TST.2023.9010002]

Qi C R, Su H, Mo K C and Guibas L J. 2017a. PointNet: deep learning on point sets for 3D classification and segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 77-85 [DOI: 10.1109/CVPR.2017.16]

Qi C R, Su H, Nießner M, Dai A, Yan M Y and Guibas L J. 2016. Volumetric and multi-view CNNs for object classification on 3D data//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 5648-5656 [DOI: 10.1109/CVPR.2016.609]

Qi C R, Yi L, Su H and Guibas L J. 2017b. PointNet++: deep hierarchical feature learning on point sets in a metric space//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 5105-5114

Qi Z K, Dong R P, Zhang S C, Geng H R, Han C R, Ge Z, Yi L and Ma K S. 2024b. ShapeLLM: universal 3D object understanding for embodied interaction [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2402.17766.pdf

Qi Z Y, Fang Y, Sun Z Y, Wu X Y, Wu T, Wang J Q, Lin D H and Zhao H S. 2023. GPT4POINT: a unified framework for point-language understanding and generation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2312.02980.pdf

Qian G, Li Y, Peng H, Mai J, Hammoud H, Elhoseiny M and Ghanem B. 2022. PointNeXt: revisiting pointnet++ with improved training and scaling strategies//Proceedings of the 35th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 23192-23204

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Vienna, Austria: PMLR: 8748-8763

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y Q, Li W and Liu P J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(1): 5485-5551

Rajpal A, Cheema N, Illgner-Fehns K, Slusallek P and Jaiswal S. 2023. High-resolution synthetic RGB-D datasets for monocular depth estimation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver, Canada: IEEE: 1188-1198 [DOI: 10.1109/CVPRW59228.2023.00126]

Ramakrishnan S K, Gokaslan A, Wijmans E, Maksymets O, Clegg A, Turner J, Undersander E, Galuba W, Westbury A, Chang A X, Savva M, Zhao Y L and Batra D. 2021. Habitat-Matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2109.08238.pdf

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

Rana K, Haviland J, Garg S, Abou-Chakra J, Reid I and Suenderhauf N. 2023. SayPlan: grounding large language models using 3D scene graphs for scalable robot task planning//Proceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR: 23-72

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

Rubenstein P K, Asawaroengchai C, Nguyen D D, Bapna A, Borsos Z, de Chaumont Quitry F, Chen P, El Badawy D, Han W, Kharitonov E, Muckenhirn H, Padfield D, Qin J, Rozenberg D, Sainath T, Schalkwyk J, Sharifi M, Ramanovich M T, Tagliasacchi M, Tudor A, Velimirović M, Vincent D, Yu J H, Wang Y Q, Zayats V, Zeghidour N, Zhang Y, Zhang Z S, Zilka L and Frank C. 2023. AudioPaLM: a large language model that can speak and listen [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2306.12925.pdf

Sanghi A, Chu H, Lambourne J G, Wang Y, Cheng C Y, Fumero M and Malekshan K R. 2022. CLIP-forge: towards zero-shot text-to-shape generation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18582-18592 [DOI: 10.1109/CVPR52688.2022.01805]

Schuhmann C, Beaumont R, Vencu R, Gordon C, Wightman R, Cherti M, Coombes T, Katta A, Mullis C, Wortsman M, Schramowski P, Kundurthy S, Crowson K, Schmidt L, Kaczmarczyk R and Jitsev J. 2022. LAION-5B: an open large-scale dataset for training next generation image-text models//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 25278-25294

Schumann R, Zhu W R, Feng W X, Fu T J, Riezler S and Wang W Y. 2024. VELMA: verbalization embodiment of LLM agents for vision and language navigation in street view//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 18924-18933 [DOI: 10.1609/aaai.v38i17.29858]

Schuster M and Paliwal K K. 1997. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11): 2673-2681 [DOI: 10.1109/78.650093]

Sella E, Fiebelman G, Hedman P and Averbuch-Elor H. 2023. Vox-E: text-guided voxel editing of 3D objects//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 430-440 [DOI: 10.1109/ICCV51070.2023.00046]

Shao S W, Pei Z C, Wu X M, Liu Z, Chen W H and Li Z G. 2023. IEBins: iterative elastic bins for monocular depth estimation//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 53025-53037

Shen Y L, Song K T, Tan X, Li D S, Lu W M and Zhuang Y T. 2023. Hugginggpt: solving ai tasks with chatgpt and its friends in hugging face [EB/OL]. [2023-03-30]. https://arxiv.org/pdf/2303.17580.pdf

Shi S S, Guo C X, Jiang L, Wang Z, Shi J P, Wang X G and Li H S. 2020. PV-RCNN: point-voxel feature set abstraction for 3D object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10526-10535 [DOI: 10.1109/CVPR42600.2020.01054]

Shi S S, Jiang L, Deng J J, Wang Z, Guo C X, Shi J P, Wang X G and Li H S. 2023. PV-RCNN++: point-voxel feature set abstraction with local vector representation for 3D object detection. International Journal of Computer Vision, 131(2): 531-551 [DOI: 10.1007/s11263-022-01710-9]

Shim J, Kang C and Joo K. 2023. Diffusion-based signed distance fields for 3D shape generation//Proceedings of 2023 IEEE/CVF Confer-

ence on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 20887-20897 [DOI: 10.1109/CVPR52729.2023.02001]

Singer U, Sheynin S, Polyak A, Ashual O, Makarov I, Kokkinos F, Goyal N, Vedaldi A, Parikh D, Johnson J and Taigman Y. 2023. Text-to-4D dynamic scene generation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2301.11280.pdf

Singh K P, Salvador J, Weihs L and Kembhavi A. 2023. Scene graph contrastive learning for embodied navigation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 10850-10860 [DOI: 10.1109/ICCV51070.2023.00999]

Singh S, Pavlakos G and Stamoulis D. 2024. Evaluating zero-shot GPT-4v performance on 3D visual question answering benchmarks [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2405.18831.pdf

Singh V V, Sheshappanavar S V and Kambhamettu C. 2021. MeshNet++: a network with a face//Proceedings of the 29th ACM International Conference on Multimedia. [s. l.]: ACM: 4883-4891 [DOI: 10.1145/3474085.3475468]

Slaney J and Thiébaux S. 2001. Blocks world revisited. Artificial Intelligence, 125 (1/2): 119-153 [DOI: 10.1016/S0004-3702 (00) 00079-5]

Song C H, Sadler B M, Wu J M, Chao W L, Washington C and Su Y. 2023. LLM-planner: few-shot grounded planning for embodied agents with large language models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 2986-2997 [DOI: 10.1109/ICCV51070.2023.00280]

Stechly K, Valmeekam K and Kambhampati S. 2024. Chain of thoughtlessness? An analysis of cot in planning [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2405.04776.pdf

Su H, Maji S, Kalogerakis E and Learned-Miller E. 2015. Multi-view convolutional neural networks for 3D shape recognition//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 945-953 [DOI: 10.1109/ICCV.2015.114]

Sun C, Wu X, Sun J, Sun C Y, Xu M Z and Ge Q B. 2023a. Saliency-induced moving object detection for robust RGB-D vision navigation under complex dynamic environments. IEEE Transactions on Intelligent Transportation Systems, 24(10): 10716-10734 [DOI: 10.1109/TITS.2023.3275279]

Sun Q H, Li Y G, Liu Z X, Huang X S, Liu F G, Liu X H, Ouyang W L and Shao J. 2023b. UniG3D: a unified 3D object generation dataset [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2306.10730.pdf

Tai H C, He Q D, Zhang J N, Qian Y J, Zhang Z Y, Hu X B, Wang Y B and Liu Y. 2024. Open-vocabulary SAM3D: anyunderstand 3D scene [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2405.15580v1.pdf

Takmaz A, Fedele E, Sumner R W, Pollefeys M, Tombari F and Engelmann F. 2023. OpenMask3D: open-vocabulary 3D instance segmentation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2306.13631.pdf

Tang J X, Zhou H, Chen X K, Hu T S, Ding E R, Wang J D and Zeng G. 2023a. Delicate textured mesh recovery from NeRF via adaptive surface refinement//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 17693-17703 [DOI: 10.1109/ICCV51070.2023.01626]

Tang Y, Han X, Li X Z, Yu Q, Hao Y X, Hu L and Chen M. 2024. MiniGPT-3D: efficiently aligning 3D point clouds with large language models using 2D priors [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2405.01413.pdf

Tang Z N, Yang Z Y, Khademi M, Liu Y, Zhu C G and Bansal M. 2023b. CoDi-2: in-context, interleaved, and interactive any-to-any generation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2311.18775.pdf

Thirunavukarasu A J, Ting D S J, Elangovan K, Gutierrez L, Tan T F and Ting D S W. 2023. Large language models in medicine. Nature Medicine, 29 (8): 1930-1940 [DOI: 10.1038/s41591-023-02448-8]

Ton T, Hong J W, Eom S, Shim J Y, Kim J and Yoo C D. 2024. Zero-shot dual-path integration framework for open-vocabulary 3D instance segmentation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7598-7607 [DOI: 10.1109/CVPRW63382.2024.00755]

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E and Lample G. 2023. LLaMA: open and efficient foundation language models [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2302.13971.pdf

Tsalicoglou C, Manhardt F, Tonioni A, Niemeyer M and Tombari F. 2024. TextMesh: generation of realistic 3D meshes from text prompts//Proceedings of 2024 International Conference on 3D Vision. Davos, Switzerland: IEEE: 1554-1563 [DOI: 10.1109/3DV62453.2024.00154]

Turki H, Zhang J Y, Ferroni F and Ramanan D. 2023. SUDS: scalable urban dynamic scenes//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 12375-12385 [DOI: 10.1109/CVPR52729.2023.01191]

Tychola K A, Tsimperidis I and Papakostas G A. 2022. On 3D reconstruction using RGB-D cameras. Digital, 2(3): 401-421 [DOI: 10.3390/digital2030022]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Vemprala S H, Bonatti R, Bucker A and Kapoor A. 2024. ChatGPT for robotics: design principles and model abilities. IEEE Access, 12: 55682-55696 [DOI: 10.1109/ACCESS.2024.3387941]

Vinyals O, Toshev A, Bengio S and Erhan D. 2015. Show and tell: a

neural image caption generator//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3156-3164 [DOI: 10.1109/CVPR.2015.7298935]

Wald J, Avetisyan A, Navab N, Tombari F and Nießner M. 2019. RIO: 3D object instance re-localization in changing indoor environments// Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 7657-7666 [DOI: 10.1109/ICCV.2019.00775]

Wald J, Dhamo H, Navab N and Tombari F. 2020. Learning 3D semantic scene graphs from 3D indoor reconstructions//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3960-3969 [DOI: 10.1109/CVPR42600.2020.00402]

Wang J Q, Wu Z H, Li Y W, Jiang H Q, Shu P, Shi E Z, Hu H W, Ma C, Liu Y H, Wang X H, Yao Y C, Liu X, Zhao H Q, Liu Z L, Dai H X, Zhao L, Ge B, Li X, Liu T M and Zhang S. 2024b. Large language models for robotics: opportunities, challenges, and perspectives [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2401.04334.pdf

Wang J Y, Chakraborty R and Yu S X. 2022a. Transformer for 3D point clouds. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(8): 4419-4431 [DOI: 10.1109/TPAMI.2021.3070341]

Wang L W, Li Y, Huang J and Lazebnik S. 2019. Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2): 394-407 [DOI: 10.1109/TPAMI.2018.2797921]

Wang M Y, Hu L H, Bai Y T, Yao X L, Hu J H and Zhang S L. 2024c. AMNet: a new RGB-D instance segmentation network based on attention and multi-modality. The Visual Computer, 40(2): 1311-1325 [DOI: 10.1007/s00371-023-02850-w]

Wang P S. 2023. OctFormer: octree-based transformers for 3D point clouds. ACM Transactions on Graphics, 42(4): #155 [DOI: 10.1145/3592131]

Wang T, Mao X H, Zhu C M, Xu R S, Lyu R Y, Li P S, Chen X, Zhang W W, Chen K, Xue T F, Liu X H, Lu C W, Lin D H and Pang J M. 2023a. EmbodiedScan: a holistic multi-modal 3D perception suite towards embodied AI [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2312.16170v1.pdf

Wang X Y, Zhuang B H and Wu Q. 2024d. ModaVerse: efficiently transforming modalities with LLMs [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2401.06395.pdf

Wang Z Q, Cheng B W, Zhao L C, Xu D, Tang Y and Sheng L. 2023b. VL-SAT: visual-linguistic semantics assisted training for 3D semantic scene graph prediction in point cloud//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 21560-21569 [DOI: 10.1109/CVPR52729.2023.02065]

Wang Z R, Yu J H, Yu A W, Dai Z H, Tsvetkov Y and Cao Y. 2022b. SimVLM: simple visual language model pretraining with weak supervision [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2108.10904.pdf

Wang Z Y, Lu C, Wang Y K, Bao F, Li C X, Su H and Zhu J. 2023c. ProlificDreamer: high-fidelity and diverse text-to-3D generation with variational score distillation//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 8406-8441

Wei J, Wang X Z, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E H, Le Q V and Zhou D. 2022. Chain-of-thought prompting elicits reasoning in large language models//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 24824-24837

Wei J C, Wang H, Feng J S, Lin G S and Yap K H. 2023a. TAPS3D: text-guided 3D textured shape generation from pseudo supervision// Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 16805-16815 [DOI: 10.1109/CVPR52729.2023.01612]

Wei X, Yu R X and Sun J. 2020. View-GCN: view-based graph convolutional network for 3D shape analysis//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1847-1856 [DOI: 10.1109/CVPR42600.2020.00192]

Wei X, Yu R X and Sun J. 2023b. Learning view-based graph convolutional network for multi-view 3D shape analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(6): 7525-7541 [DOI: 10.1109/TPAMI.2022.3221785]

Wen B W, Yang W, Kautz J and Birchfield S. 2024. FoundationPose: unified 6D pose estimation and tracking of novel objects [EB/OL]. [2024-09-20]. https://doi.org/10.48550/arXiv.2312.08344

Weng C Y, Curless B, Srinivasan P P, Barron J T and Kemelmacher-Shlizerman I. 2022. HumanNeRF: free-viewpoint rendering of moving people from monocular video//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 16189-16199 [DOI: 10.1109/CVPR52688.2022.01573]

Werby A, Huang C G, Büchner M, Valada A and Burgard W. 2024. Hierarchical open-vocabulary 3D scene graphs for language-grounded robot navigation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2403.17846.pdf

Wu H, Wen C L, Shi S S, Li X and Wang C. 2023a. Virtual sparse convolution for multimodal 3D object detection//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 21653-21662 [DOI: 10.1109/CVPR52729.2023.02074]

Wu Q Y, Liu X, Chen Y D, Li K J, Zheng C X, Cai J F and Zheng J M. 2022. Object-compositional neural implicit surfaces//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 197-213 [DOI: 10.1007/978-3-031-19812-0_12]

Wu S C, Wald J, Tateno K, Navab N and Tombari F. 2021. SceneGraphFusion: incremental 3D scene graph prediction from RGB-D sequences//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 7511-7521 [DOI: 10.1109/CVPR46437.2021.00743]

Wu T, Zhang J R, Fu X, Wang Y X, Ren J W, Pan L, Wu W, Yang L, Wang J Q, Qian C, Lin D H and Liu Z W. 2023b. OmniObject3D: large-vocabulary 3D object dataset for realistic perception, reconstruction and generation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 803-814 [DOI: 10.1109/CVPR52729.2023.00084]

Wu T Y, Huang S Y and Wang Y C F. 2024. DOrA: 3D visual grounding with order-aware referring [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2403.16539v1.pdf

Wu Y M, Cheng X H, Zhang R R, Cheng Z S and Zhang J. 2023c. EDA: explicit text-decoupling and dense alignment for 3D visual grounding//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 19231-19242 [DOI: 10.1109/CVPR52729.2023.01843]

Wu Z R, Song S R, Khosla A, Yu F, Zhang L G, Tang X O and Xiao J X. 2015. 3D ShapeNets: a deep representation for volumetric shapes//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1912-1920 [DOI: 10.1109/CVPR.2015.7298801]

Xian W Q, Huang J B, Kopf J and Kim C. 2021. Space-time neural irradiance fields for free-viewpoint video//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9416-9426 [DOI: 10.1109/CVPR46437.2021.00930]

Xiao Z H, Jing L L, Wu S X, Zhu A Z, Ji J W, Jiang C M, Hung W C, Funkhouser T, Kuo W C, Angelova A, Zhou Y and Sheng S W. 2024. 3D open-vocabulary panoptic segmentation with 2D-3D vision-language distillation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2401.02402.pdf

Xie L, Xu G D, Cai D and He X F. 2023. X-View: non-egocentric multi-view 3D object detector. IEEE Transactions on Image Processing, 32: 1488-1497 [DOI: 10.1109/TIP.2023.3245337]

Xie T Y, Zong Z S, Qiu Y X, Li X, Feng Y T, Yang Y and Jiang C F F. 2024. PhysGaussian: physics-integrated 3D Gaussians for generative dynamics [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2311.12198.pdf

Xing Y J, Wang J B, Chen X K and Zeng G. 2019. Coupling two-stream RGB-D semantic segmentation network by idempotent mappings//Proceedings of 2019 IEEE International Conference on Image Processing. Taipei, China: IEEE: 1850-1854 [DOI: 10.1109/ICIP.2019.8803146]

Xiong H L, Muttukuru S, Upadhyay R, Chari P and Kadambi A. 2024. SparseGS: real-time 360° sparse view synthesis using Gaussian splatting [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2312.00206.pdf

Xu C F, Wu B C, Hou J, Tsai S, Li R L, Wang J L, Zhan W, He Z J, Vajda P, Keutzer K and Tomizuka M. 2023a. NeRF-Det: learning geometry-aware volumetric representation for multi-view 3D object detection//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 23263-23273 [DOI: 10.1109/ICCV51070.2023.02131]

Xu D J, Liang H W, Bhatt N P, Hu H Z, Liang H X, Plataniotis K N and Wang Z Y. 2024a. Comp4D: LLM-guided compositional 4D scene generation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2403.16993.pdf

Xu H J, He K, Plummer B A, Sigal L, Sclaroff S and Saenko K. 2019. Multilevel language and vision integration for text-to-clip retrieval//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI: 9062-9069 [DOI: 10.1609/aaai.v33i01.33019062]

Xu J L, Wang X T, Cheng W H, Cao Y P, Shan Y, Qie X and Gao S H. 2023b. Dream3D: zero-shot text-to-3D synthesis using 3D shape prior and text-to-image diffusion models//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 20908-20918 [DOI: 10.1109/CVPR52729.2023.02003]

Xu R S, Wang X L, Wang T, Chen Y L, Pang J M and Lin D H. 2024b. PointLLM: empowering large language models to understand point clouds [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2308.16911.pdf

Xu X Y, Chen L C, Cai C J, Zhan H Y, Yan Q, Ji P, Yuan J S, Huang H and Xu Y. 2023c. Dynamic voxel grid optimization for high-fidelity RGB-D supervised surface reconstruction [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2304.06178.pdf

Xu Y H, Peng S D, Yang C Y, Shen Y J and Zhou B L. 2022. 3D-aware image synthesis via learning structural and textural representations//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18409-18418 [DOI: 10.1109/CVPR52688.2022.01788]

Xue L, Gao M F, Xing C, Martín-Martín R, Wu J J, Xiong C M, Xu R, Niebles J C and Savarese S. 2023. ULIP: learning a unified representation of language, images, and point clouds for 3D understanding//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 1179-1189 [DOI: 10.1109/CVPR52729.2023.00120]

Xue L, Yu N, Zhang S, Panagopoulou A, Li J N, Martín-Martín R, Wu J J, Xiong C M, Xu R, Niebles J C and Savarese S. 2024. ULIP-2: towards scalable multimodal pre-training for 3D understanding [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2305.08275.pdf

Yan X, Yuan Z H, Du Y H, Liao Y H, Guo Y, Li Z and Cui S G. 2023. Comprehensive visual question answering on point clouds through compositional scene manipulation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2112.11691.pdf

Yang H Z, Sun Y X, Sundaramoorthi G and Yezzi A. 2023a. StEik: stabilizing the optimization of neural signed distance functions and finer shape representation//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 13993-14004

Yang J H, Ding R Y, Deng W P, Wang Z and Qi X J. 2024b. RegionPLC: regional point-language contrastive learning for open-world 3D scene understanding [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2304.00962.pdf

Yang J N, Chen X W Y, Madaan N, Iyengar M, Qian S Y, Fouhey D F and Chai J. 2024a. 3D-GRAND: a million-scale dataset for 3D-LLMs with better grounding and less hallucination [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2406.05132.pdf

Yang J N, Chen X W Y, Qian S Y, Madaan N, Iyengar M, Fouhey D F and Chai J. 2023b. LLM-grounder: open-vocabulary 3D visual grounding with large language model as an agent [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2309.12311.pdf

Yang S Q, Liu J M, Zhang R, Pan M J, Guo Z, Li X Q, Chen Z H, Gao P, Guo Y D and Zhang S H. 2023c. Lidar-LLM: exploring the potential of large language models for 3D lidar understanding [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2312.14074.pdf

Yang Y X, Lu J R, Zhao Z X, Luo Z, Yu J J Q, Sanchez V and Zheng F. 2024c. LLplace: the 3D indoor scene layout generation and editing via large language model [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2406.03866.pdf

Yang Z Y, Li L J, Wang J F, Lin K, Azarnasab E, Ahmed F, Liu Z C, Liu C, Zeng M and Wang L J. 2023d. MM-REACT: prompting ChatGPT for multimodal reasoning and action [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2303.11381.pdf

Yang Z Y, Yang H Y, Pan Z J and Zhang L. 2024d. Real-time photorealistic dynamic scene representation and rendering with 4D Gaussian splatting [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2310.10642.pdf

Yao L W, Han J H, Liang X D, Xu D, Zhang W, Li Z G and Xu H. 2023. DetCLIPv2: scalable open-vocabulary object detection pre-training via word-region alignment//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 23497-23506 [DOI: 10.1109/CVPR52729.2023.02250]

Yao L W, Pi R J, Han J H, Liang X D, Xu H, Zhang W, Li Z G and Xu D. 2024. DetCLIPv3: towards versatile generative open-vocabulary object detection [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2404.09216.pdf

Ye J L, Wang N Y and Wang X L. 2023. FeatureNeRF: learning generalizable NeRFs by distilling foundation models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 8928-8939 [DOI: 10.1109/ICCV51070.2023.00823]

Yenamandra T, Tewari A, Yang N, Bernard F, Theobalt C and Cremers D. 2024. FIRe: fast inverse rendering using directional and signed distance functions//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 3065-3075 [DOI: 10.1109/WACV57701.2024.00305]

Yin F K, Chen X, Zhang C, Jiang B, Zhao Z B, Fan J Y, Yu G, Li T H and Chen T. 2023. ShapeGPT: 3D shape generation with a unified multi-modal language model [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2311.17618.pdf

Yin S K, Fu C Y, Zhao S R, Li K, Sun X, Xu T and Chen E H. 2024a. A survey on multimodal large language models [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2306.13549.pdf

Yin Z F, Wang J, Cao J J, Shi Z L, Liu D N, Li M K, Huang X S, Wang Z Y, Sheng L, Bai L, Shao J and Ouyang W. 2024b. LAMM: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark//Proceedings of the 37th Conference on Advances in Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 26650-26685

Yu A, Ye V, Tancik M and Kanazawa A. 2021. PixelNeRF: neural radiance fields from one or few images//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 4576-4585 [DOI: 10.1109/CVPR46437.2021.00455]

Yu H, Qin Z, Hou J, Saleh M, Li D S, Busam B and Ilic S. 2023a. Rotation-invariant transformer for point cloud matching//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 5384-5393 [DOI: 10.1109/CVPR52729.2023.00521]

Yu J H, Wang Z R, Vasudevan V, Yeung L, Seyedhosseini M and Wu Y H. 2022a. CoCa: contrastive captioners are image-text foundation models [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2205.01917.pdf

Yu Q H, He J, Deng X Q, Shen X H and Chen L C. 2023b. Convolutions die hard: open-vocabulary segmentation with single frozen convolutional CLIP//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 32215-32234

Yu T, Meng J J and Yuan J S. 2018. Multi-view harmonized bilinear network for 3D object recognition//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 186-194 [DOI: 10.1109/CVPR.2018.00027]

Yu X M, Tang L K, Rao Y M, Huang T J, Zhou J and Lu J W. 2022b. Point-BERT: pre-training 3D point cloud transformers with masked point modeling//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 19291-19300 [DOI: 10.1109/CVPR52688.2022.01871]

Yu Y, Ko H, Choi J and Kim G. 2017. End-to-end concept word detection for video captioning, retrieval, and question answering//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 3261-3269 [DOI: 10.1109/CVPR.2017.347]

Yu Z H, Chen A P, Huang B B, Sattler T and Geiger A. 2023c. Mip-

splatting: alias-free 3D Gaussian splatting [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2311.16493.pdf

Yuan Z H, Ren J K, Feng C M, Zhao H S, Cui S G and Li Z. 2024a. Visual programming for zero-shot open-vocabulary 3D visual grounding [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2311.15383.pdf

Yuan Z H, Yan X, Liao Y H, Guo Y, Li G B, Cui S S and Li Z. 2022. X-Trans2Cap: cross-modal knowledge transfer using transformer for 3D dense captioning//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 8553-8563 [DOI: 10.1109/CVPR52688.2022.00837]

Yuan Z Q, Lan H X, Zou Q and Zhao J B. 2024b. 3D-PreMise: can large language models generate 3D shapes with sharp features and parametric control? [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2401.06437.pdf

Yuksekgonul M, Bianchi F, Kalluri P, Jurafsky D and Zou J. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2210.01936.pdf

Zareian A, Rosa K D, Hu D H and Chang S F. 2021. Open-vocabulary object detection using captions//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14388-14397 [DOI: 10.1109/CVPR46437.2021.01416]

Zeng Y F, Jiang Y Q, Zhu S Y, Lu Y X, Lin Y T, Zhu H, Hu W M, Cao X and Yao Y. 2024. STAG4D: spatial-temporal anchored generative 4D Gaussians [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2403.14939.pdf

Zhan J, Dai J Q, Ye J S, Zhou Y H, Zhang D, Liu Z G, Zhang X, Yuan R B, Zhang G, Li L Y, Yan H, Fu J, Gui T, Sun T X, Jiang Y G and Qiu X P. 2024. AnyGPT: unified multimodal LLM with discrete sequence modeling [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2402.12226.pdf

Zhang C H, Zhou Y Y and Zhang L. 2024a. Voxel-mesh hybrid representation for real-time view synthesis [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2403.06505.pdf

Zhang D M, Li C, Zhang R R, Xie S H, Xue W, Xie X D and Zhang S H. 2024b. FM-OV3D: foundation model-based cross-modal knowledge blending for open-vocabulary 3D detection//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 16723-16731 [DOI: 10.1609/aaai.v38i15.29612]

Zhang J, Huang J C, Cai B W, Fu H, Gong M M, Wang C H, Wang J M, Luo H C, Jia R F, Zhao B Q and Tang X. 2022a. Digging into radiance grid for real-time view synthesis with detail preservation//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 724-740 [DOI: 10.1007/978-3-031-19784-0_42]

Zhang J B, Dong R P and Ma K S. 2023a. CLIP-FO3D: learning free open-world 3D scene representations from 2D dense CLIP//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision Workshops. Paris, France: IEEE: 2040-2051 [DOI: 10.1109/ICCVW60793.2023.00219]

Zhang J B, Li X Y, Wan Z Y, Wang C and Liao J. 2024c. Text2NeRF: text-driven 3D scene generation with neural radiance fields [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2305.11588.pdf

Zhang J H, Zhan F N, Xu M Y, Lu S J and Xing E. 2024d. FreGS: 3D Gaussian splatting with progressive frequency regularization [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2403.06908.pdf

Zhang K, Riegler G, Snavely N and Koltun V. 2020. NeRF++: analyzing and improving neural radiance fields [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2010.07492.pdf

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

Zhang L W, Wang Z Y, Zhang Q X, Qiu Q W, Pang A Q, Jiang H R, Yang W, Xu L and Yu J Y. 2024e. CLAY: a controllable large-scale generative model for creating high-quality 3D assets. ACM Transactions on Graphics, 43(4): #120 [DOI: 10.1145/3658146]

Zhang M X, Feng Q, Su Z, Wen C, Xue Z and Li K. 2024f. Joint2Human: high-quality 3D human generation via compact spherical embedding of 3D joints//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1429-1438 [DOI: 10.1109/CVPR52733.2024.00142]

Zhang Q H, Wang C Y, Siarohin A, Zhuang P Y, Xu Y H, Yang C Y, Lin D H, Zhou B L, Tulyakov S and Lee H Y. 2023c. SceneWiz3D: towards text-guided 3D scene composition [EB/OL]. [2024-09-20]. https://doi.org/10.48550/arXiv.2312.08885

Zhang R R, Guo Z Y, Zhang W, Li K C, Miao X P, Cui B, Qiao Y, Gao P and Li H S. 2022b. PointCLIP: point cloud understanding by CLIP//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 8542-8552 [DOI: 10.1109/CVPR52688.2022.00836]

Zhang R R, Wang L H, Guo Z Y, Wang Y L, Gao P, Li H S and Shi J B. 2023d. Parameter is not all you need: starting from non-parametric networks for 3D point cloud analysis [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2303.08134.pdf

Zhang Y M, Gong Z M and Chang A X. 2023e. Multi3DRefer: grounding text description to multiple 3D objects//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 15225-15236 [DOI: 10.1109/ICCV51070.2023.01397]

Zhang Y P, Zhu Z and Du D L. 2023f. OccFormer: dual-path transformer for vision-based 3D semantic occupancy prediction//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 9399-9409 [DOI: 10.1109/ICCV51070.2023.00865]

Zhang Y Q, Luo H and Lei Y J. 2024g. Towards CLIP-driven language-

free 3D visual grounding via 2D-3D relational enhancement and consistency//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13063-13072 [DOI: 10.1109/CVPR52733.2024.01241]

Zhang Z H, Cao S C and Wang Y X. 2024h. TAMM: triadapter multi-modal learning for 3D shape understanding [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2402.18490.pdf

Zhao Z B, Liu W, Chen X, Zeng X F, Wang R, Cheng P, Fu B, Chen T, Yu G and Gao S H. 2024. Michelangelo: conditional 3D shape generation based on shape-image-text aligned latent representation//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 73969-73982

Zhen H Y, Qiu X W, Chen P H, Yang J C, Yan X, Du Y L, Hong Y N and Gan C. 2024. 3D-VLA: a 3D vision-language-action generative world model [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2403.09631.pdf

Zheng D, Huang S J, Zhao L, Zhong Y W and Wang L W. 2024. Towards learning a generalist model for embodied navigation [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2312.02010.pdf

Zheng H M and Gao W. 2024. End-to-end RGB-D image compression via exploiting channel-modality redundancy//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 7562-7570 [DOI: 10.1609/aaai.v38i7.28588]

Zheng J, Zhang J F, Li J, Tang R, Gao S H and Zhou Z H. 2020. Structured3D: a large photo-realistic dataset for structured 3D modeling//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 519-535 [DOI: 10.1007/978-3-030-58545-7_30]

Zhou C, Zhang Y N, Chen J X and Huang D. 2023. OcTr: octree-based transformer for 3D object detection//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 5166-5175 [DOI: 10.1109/CVPR52729.2023.00500]

Zhou H, Qi L, Wan Z L, Huang H and Yang X. 2020. RGB-D co-attention network for semantic segmentation//Proceedings of the 15th Asian Conference on Computer Vision. Kyoto, Japan: Springer: 519-536 [DOI: 10.1007/978-3-030-69525-5_31]

Zhou X Y, Ran X J, Xiong Y J, He J L, Lin Z W, Wang Y T, Sun D Q and Yang M H. 2024. GALA3D: towards text-to-3D complex scene generation via layout-guided generative Gaussian splatting [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2402.07207.pdf

Zhu C M, Wang T, Zhang W W, Chen K and Liu X H. 2024a. ScanReason: empowering 3D visual grounding with reasoning capabilities [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2407.01525.pdf

Zhu C M, Zhang W W, Wang T, Liu X H and Chen K. 2023a. Object2Scene: putting objects in context for open-vocabulary 3D detection [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2309.09456.pdf

Zhu D Y, Chen J, Shen X Q, Li X and Elhoseiny M. 2023b. Minigpt-4: enhancing vision-language understanding with advanced large language models [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2304.10592.pdf

Zhu X Y, Zhang R R, He B W, Guo Z Y, Zeng Z Y, Qin Z P, Zhang S H and Gao P. 2023c. PointCLIP V2: prompting CLIP and GPT for powerful 3D open-world learning//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 2639-2650 [DOI: 10.1109/ICCV51070.2023.00249]

Zhu Z H, Fan Z W, Jiang Y F and Wang Z Y. 2024b. FSGS: real-time few-shot view synthesis using Gaussian splatting [EB/OL]. [2024-09-20]. https://arxiv.org/pdf/2312.00451.pdf

Zhu Z Y, Ma X J, Chen Y X, Deng Z D, Huang S Y and Li Q. 2023d. 3D-VisTA: pre-trained transformer for 3D vision and text alignment//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 2899-2909 [DOI: 10.1109/ICCV51070.2023.00272]

作者简介

冯明涛，男，副教授，主要研究方向为三维视觉感知与理解。
E-mail: mintfeng@hnu.edu.cn
王耀南，通信作者，男，教授，中国工程院院士，主要研究方向为智能机器人。E-mail: yaonan@hnu.edu.cn
沈军豪，男，硕士研究生，主要研究方向为三维视觉。
E-mail: 15372435952@163.com
武子杰，男，博士研究生，主要研究方向为机器人三维视觉。
E-mail: wuzijieeee@hnu.edu.cn
彭伟星，男，副研究员，主要研究方向为机器人三维视觉。
E-mail: wexing_peng@hnu.edu.cn
钟杭，男，副教授，主要研究方向为智能机器人控制。
E-mail: zhonghang@hnu.edu.cn
郭裕兰，男，教授，主要研究方向为计算机三维视觉。
E-mail: gfkdguoyulan@163.com
舒祥波，男，教授，主要研究方向为多模态视觉与多媒体处理。E-mail: shuxb@njust.edu.cn
张辉，男，教授，主要研究方向为智能机器人感知与控制。
E-mail: zhanghui1983@hnu.edu.cn
董伟生，男，教授，主要研究方向为图像处理与计算机视觉。
E-mail: wsdong@mail.xidian.edu.cn