中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)06-1717-27

论文引用格式:Liu Y B, Su H, Gao L, Yi L, Wang H, Liao Y Y, Shi B X, Cao Y P, Hong F Z, Dong H, Zhang J Y, Wang X T, Xu H Z, Yang J L, Kang B Y, Chu M Y, Sun H, Chen W Z, Ma Y X, Zhang H W, Guo Y L, Zhou X W, Zhang G F, Han X G, Dai Y C and Chen B Q. 2025. Research trends and major developments in 3D vision in 2024. Journal of Image and Graphics, 30(6):1717-1743(刘烨斌, 苏昊, 高林, 弋力, 王鹤, 廖依伊, 施柏鑫, 曹炎培, 洪方舟, 董豪, 张举勇, 王鑫涛, 许华哲, 杨蛟龙, 康炳易, 楚梦渝, 孙赫, 陈文拯, 马月昕, 张鸿文, 郭裕兰, 周晓巍, 章国锋, 韩晓光, 戴玉超, 陈宝权. 2025. 2024年度三维视觉前沿趋势与十大进展. 中国图象图形学报, 30(6):1717-1743)[DOI:10.11834/jig.250057]

2024年度三维视觉前沿趋势与十大进展

刘烨斌[1],苏昊[2],高林[3],弋力[1],王鹤[4],廖依伊[5],施柏鑫[4],曹炎培[6],洪方舟[7],董豪[4],张举勇[8],王鑫涛[9],许华哲[1],杨蛟龙[10],康炳易[11],楚梦渝[4],孙赫[4],陈文拯[4],马月昕[12],张鸿文[13],郭裕兰[14],周晓巍[5],章国锋[5],韩晓光[15],戴玉超[16],陈宝权[4*]

1. 清华大学自动化系,北京　100084; 2. 加州大学圣迭戈分校计算机科学与工程系,圣迭戈　92093,美国;
3. 中国科学院计算技术研究所,北京　100190; 4. 北京大学智能学院,北京　100871; 5. 浙江大学计算机科学与技术学院,杭州　310027;
6. VAST哇嘶嗒科技有限公司,北京　100083; 7. 南洋理工大学计算机与数据科学学院,新加坡　639798,新加坡;
8. 中国科学技术大学数学科学学院,合肥　230026; 9. 快手科技有限公司,北京　100085; 10. 微软亚洲研究院,北京　100080;
11. 字节跳动科技有限公司,北京　100098; 12. 上海科技大学信息科学与技术学院,上海　201210;
13. 北京师范大学人工智能学院,北京　100875; 14. 中山大学电子与通信工程学院,深圳　518107;
15. 香港中文大学(深圳)理工学院,深圳　518172; 16. 西北工业大学电子信息学院,西安　710129

摘　要: 三维视觉作为计算机视觉、图形学、人工智能与光学成像的交叉学科,是构建具身通用智能与元宇宙的核心基石。2024年,以神经辐射场和高斯泼溅为代表的可微表征技术持续发展和完善并逐渐突破传统三维重建边界,无论从微观细胞组织到宏观物理天体,还是从静态场景到动态人体,均取得显著的精度提升;在生成式人工智能技术和大模型规模定律(scaling law)的推动下,三维视觉迎来从优化到可泛化前馈生成的范式跃迁,并在可控数字内容生成方向取得重要进展和突破;具身智能持续备受关注,研究者逐渐意识到三维虚拟仿真数据和三维人体运动数据的捕捉和生成,是训练具身智能的核心关键;随着世界模型和空间智能的概念成为科技界热议的焦点,对物理世界进行建模、对空间关系进行理解、对未来状态进行预测成为重要研究方向,而这些都离不开三维视觉技术的支撑;此外,计算成像技术的革新则通过非传统视觉传感器与新型重建算法,突破了传统三维重建的物理限制与性能瓶颈。这些技术突破正推动着三维视觉进入"感知—建模—生成—交互"全链路智能化、规模化学习的新阶段。为促进学术交流,本文分析总结三维视觉领域前沿趋势,并遴选年度十大研究进展,为学术界与产业界提供参考观点。

关键词: 三维视觉;具身智能;三维表征;三维生成;三维重建

Research trends and major developments in 3D vision in 2024

Liu Yebin[1], Su Hao[2], Gao Lin[3], Yi Li[1], Wang He[4], Liao Yiyi[5], Shi Boxin[4], Cao Yanpei[6], Hong Fangzhou[7], Dong Hao[4], Zhang Juyong[8], Wang Xintao[9], Xu Huazhe[1], Yang Jiaolong[10], Kang Bingyi[11], Chu Mengyu[4], Sun He[4], Chen Wenzheng[4], Ma Yuexin[12], Zhang Hongwen[13], Guo Yulan[14], Zhou Xiaowei[5], Zhang Guofeng[5], Han Xiaoguang[15], Dai Yuchao[16], Chen Baoquan[4*]

1. Department of Automation, Tsinghua University, Beijing　100084, China;

收稿日期:2025-02-14;**修回日期**:2025-03-04;**预印本日期**:2025-03-11

*通信作者:陈宝权　baoquan@pku.edu.cn

基金项目:国家自然科学基金项目(62125107, 62306016,62376012,62322210,62271410,U24B20154,62441224,62441223,62136001,62371007)

Supported by: National Natural Science Foundation of China (62125107, 62306016, 62376012, 62322210, 62271410, U24B20154, 62441224, 62441223, 62136001, 62371007)

2. Department of Computer Science and Engineering, University of California, San Diego, La Jolla 92093, USA;
3. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
4. School of Intelligence Science and Technology, Peking University, Beijing 100871, China;
5. College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China; 6. VAST AI Research, Beijing 100083, China; 7. College of Computing and Data Science, Nanyang Technological University, Singapore 639798, Singapore; 8. Department of Mathematics, University of Science and Technology of China, Hefei 230026, China; 9. Kuaishou Technology, Beijing 100085, China; 10. Microsoft Research Asia, Beijing 100080, China; 11. ByteDance Ltd., Beijing 100098, China; 12. School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China; 13. School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China; 14. School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen 518107, China;
15. School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), Shenzhen 518172, China;
16. School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China

Abstract: As an interdisciplinary field that integrates computer vision, graphics, artificial intelligence (AI), and optical imaging, three-dimensional (3D) vision serves as the foundational cornerstone for building embodied general intelligence and the metaverse ecosystem. In 2024, differentiable representation technologies, exemplified by neural raddiance field (NeRF) and Gaussian splatting, continued to evolve and refine, gradually transcending the boundaries of traditional 3D reconstruction. From microscopic cellular structures to macroscopic physical celestial bodies, and from static scenes to dynamic human bodies, significant improvements in accuracy have been achieved across a spectrum. Propelled by advancements in generative AI technologies and the scaling laws of large models, the field of 3D vision has witnessed a paradigm shift from optimization to generalizable feedforward generation, marking important progress and breakthroughs in the direction of controllable digital content generation. Embodied intelligence remains a focal point of interest, with researchers increasingly recognizing the capture and generation of 3D virtual simulation data and 3D human motion data as central elements of training embodied intelligence. As the concepts of world models and spatial intelligence become hot topics among technology researchers, modeling the physical world, understanding spatial relationships, and predicting future states have emerged as crucial research directions, all of which rely on the support of 3D vision technologies. Furthermore, through nontraditional visual sensors and novel reconstruction algorithms, innovations in computational imaging technology have exceeded the physical limitations and performance bottlenecks of traditional 3D reconstruction. These technological breakthroughs are propelling 3D vision into a new era characterized by an intelligent, large-scale learning process that encompasses the processes of perception, modeling, generation, and interaction. Specifically, the development trends in 3D vision primarily manifested in the following aspects: 1) Controllable and physics-aware generation of the visual elements of AI-generated content (AIGC). With the rapid development of AIGC technology, visual content generation has evolved from simple two-dimensional (2D) image creation toward more controllable and physics-aware approaches. This trend requires the combination of physical prior knowledge and multidimensional control parameters, such as 3D viewpoints, lighting conditions, and 3D character motion, to achieve higher-quality content generation. 3D vision technology plays a pivotal role in this process, providing essential spatial-temporal and physical constraints for AIGC. 2) 4D spatial intelligence: bridging virtual and physical worlds. 4D (3D space + time dimension) spatial intelligence has emerged as a core technology connecting virtual worlds (e. g., the metaverse) and physical realities (e. g., embodied intelligent robots). This technology focuses on establishing a digital mapping of dynamic physical environments. By leveraging 3D vision and multimodal large model technologies, AI systems can construct 4D spatial models to understand spatial relationships, predict motion trajectories, and simulate future evolutions. In turn, intelligent agents can interactively learn within physical or virtual 4D environments to acquire intelligence. 3) Data-driven embodied intelligence: 3D virtual simulation and human motion capture. The advancement of embodied intelligence relies heavily on high-quality 3D virtual simulation data and the capture/generation of human 3D motion data. These datasets fuel the training of embodied intelligent robots to achieve sophisticated behavioral control. Through high-precision 3D vision technologies, robots can better understand and simulate human actions, resulting in their enhanced intelligence in complex tasks. 4) Differentiable 3D representation and integration with large model technologies. From microscopic cellular structures to indoor environments, human/animal modeling, autonomous driving/city modeling, and even astronomical black hole reconstruction, novel 3D representations, such as

NeRF and 3D Gaussian splatting(3DGS), drive performance improvements in scene generation and reconstruction across scales. Their efficiency and flexibility have opened up new possibilities for 3D vision applications. Furthermore, by integrating large-scale 3D data with transformer-based architectures and advanced generative methods, such as diffusion models, fundamental 3D vision tasks are being unified into efficient end-to-end frameworks, resulting in the scaled-up learning of core 3D vision paradigms. The breakthroughs in 3D vision in 2024 have injected new momentum into technological development, with future trends focused on several key directions. Spatiotemporal-consistent world models that integrate 4D spacetime and physical laws provide dynamic prediction and interaction support for complex scenarios, such as autonomous driving and embodied intelligence. It is expected that the deep integration of generative AI and 3D content generation technologies will overcome data bottlenecks, enabling the automated creation of high-fidelity and controllable 3D content. Enhanced cross-modal generalization capabilities will also strengthen the fusion of vision, language, and motion modalities, thus improving the adaptability and robustness of robotic strategies. Physics-driven dynamic reconstruction, combined with physical engines, will achieve high-precision modeling and interactive editing of dynamic scenes, thus advancing digital twins and virtual reality. Meanwhile, 3D imaging technologies will accelerate scientific exploration in astrophysics and cell biology. Efficient real-time processing and lightweight solutions will also boost the reconstruction and rendering efficiency of large-scale dynamic scenes, promoting edge device applications. Furthermore, ethical and privacy protection will emerge as critical concerns, balancing innovation and security through encryption technologies and regulatory frameworks that govern 3D data acquisition and generation. In summary, this article reveals that, in the future, 3D vision will evolve toward an era of more intelligent and universal "spatial intelligence" in which technological breakthroughs will reshape human-computer interaction, scientific exploration, and industrial ecosystems. Therefore, to foster academic exchange, this article analyzed cutting-edge trends in 3D vision and highlights the top ten research breakthroughs of the year, thus offering insights and references for both academia and industry.

Key words: 3D vision; embodied AI; 3D representation; 3D generation; 3D reconstruction

0 引　言

三维视觉作为计算机视觉、计算机图形学、人工智能以及光学成像等多学科交叉的前沿领域，近年来在相关技术进步和应用需求共同驱动下，迎来前所未有的发展机遇。

2024年，随着生成式人工智能和空间智能等前沿方向成为科技界关注的焦点，三维视觉的重要性愈发凸显，成为人工智能领域的核心研究方向之一。同时，三维视觉技术也在多个领域展现出广泛的应用潜力，其发展趋势主要体现在以下方面：

1）视觉内容AIGC（artificial intelligence-generated content）的可控生成与物理感知生成。随着生成式人工智能（AIGC）技术快速发展，视觉内容的生成正从简单的二维图像生成向可控性更强、物理感知更精准的方向演进。这一趋势要求引入三维视点、光照条件和人物三维运动等多维控制参数，并结合物理先验知识，以实现更高质量的内容生成。三维视觉技术在这一过程中扮演了关键角色，为AIGC提供了必要的时空和物理约束。

2）4D空间智能。虚拟世界与真实世界的桥梁。4D空间智能（三维空间 + 时间维度）正成为连接虚拟世界（如元宇宙）和真实世界（如具身智能机器人）的核心技术。4D空间智能技术在于建立动态物理世界的数字映射。借助三维视觉技术和多模态大模型技术，AI系统能够构建4D空间模型，理解空间关系，预测运动，演化生成未来。同时，智能体可在物理或虚拟的4D空间环境中交互学习，获得智能。

3）具身智能的数据驱动。3D虚拟仿真与人体运动捕捉。具身智能的发展高度依赖高质量的3D虚拟仿真数据和人体3D运动数据的捕捉与生成。这些数据是训练具身智能机器人实现智能行为控制的"燃料"。通过高精度的三维视觉技术，机器人能够更好地理解和模拟人类行为，从而在复杂任务中表现出更高的智能水平。

4）可微三维表征技术及其与大模型技术的融合。神经辐射场（neural radiance field，NeRF）（Mildenhall等，2020）和3D高斯泼溅（Kerbl等，2023）等三维表征推动从微观到宏观的各类场景生成与重建性能升级，无论是细胞组织，还是室内场景、人体/动物建模、智驾/城市建模，甚至天文黑洞的三维重建，这一技术都

展现出强大潜力，其高效性和灵活性为三维视觉的应用开辟了新的可能性。同时，借助大规模三维数据与Transformer（Vaswani等，2017）等大模型网络架构与Diffusion（Ho等，2020）等前沿生成方法，将三维视觉的基础任务连成一个高效端到端的框架，实现了三维视觉核心范式的规模化拓展（scale up）学习。

图1展示了2024年度三维视觉研究热点，本文从10个方面细化总结2024年三维视觉领域的10大科研进展。

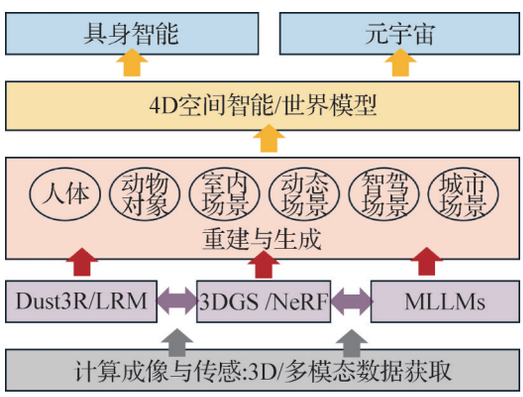


图1　2024年度三维视觉研究热点框架图

Fig. 1　The framework of 3D vision research hotspots in 2024

1　三维视觉领域科研进展

1.1　DUSt3R革新三维几何视觉研究范式

近两三年间，自然语言处理与二维计算机图像领域纷纷推出基础模型，以ChatGPT（Achiam等，2024；OpenAI，2022）与CLIP（Radford等，2021）等模型为代表，通过数据与模型的规模化拓展，在语言智能与图像智能上取得了巨大进展。另一方面，三维视觉领域在表征层面持续创新，在诸多应用上获得了大量突破，但是众多三维几何视觉任务仍然无法有效统一，导致无法发挥充分数据的规模效应，本领域亟须范式转变，寻求构建三维视觉基础模型。

在此背景下，芬兰阿尔托大学以及NAVER实验室的研究者提出DUSt3R（Wang等，2024e），如图2所示，改变了以往基于特征匹配和几何优化的思路，提出可解决多种三维几何视觉问题的通用框架：给定一组未标定图像对，通过前馈网络（feed-forward network）预测出一组在相同坐标系下的三维点云图，通过后处理优化实现相机内参标定、深度估计、像素匹配、相机位姿估计以及稠密点云三维重建等一系列三维几何视觉问题。DUSt3R首次证明了规模定律（scaling law）在解决三维视觉问题上的可行性：其使用基础的ViT架构，通过海量三维标注数据的预训练，将三维视觉的基础任务连成一个高效端到端的框架，为三维基础模型的规模化拓展（scale up）范式提供了一个有效的思路。

DUSt3R在2023年底一经推出便吸引大量社区的关注，并在2024年由众多后续工作进一步完善与扩展，其作为基础模型被成功应用在多个极具挑战性的任务上。在新视角合成（novel view synthesis）任务上，Splatt3R（Smart等，2024）、InstantSplat（Fan等，2025）和NoPoSplat（Ye等，2024a）等工作结合DUSt3R与高斯泼溅技术（3D Gaussian splatting），实现了基于未标定稀疏视图的前馈式高斯重建；在单

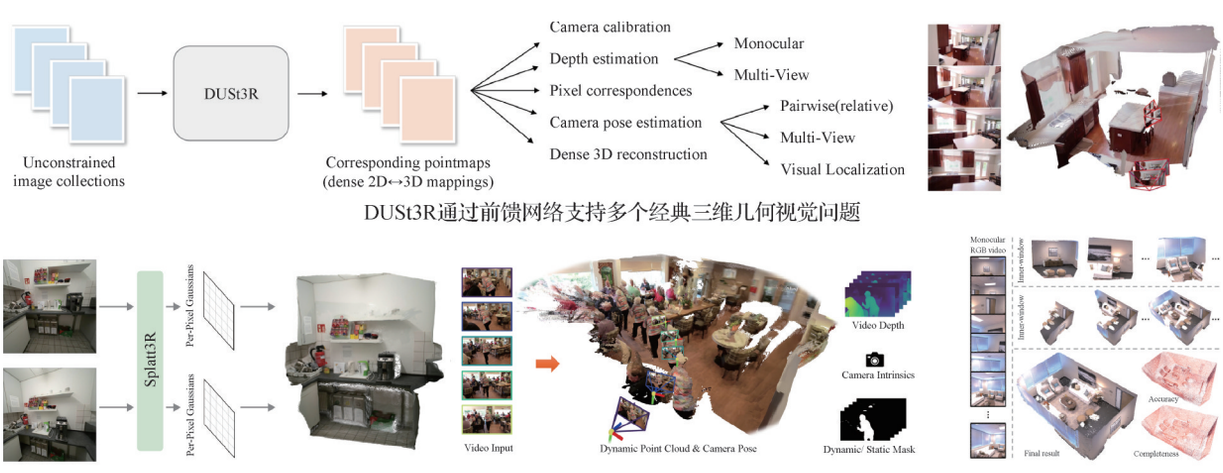


图2　DUSt3R与后续改进拓展工作

Fig. 2　DUSt3R and subsequent improvements

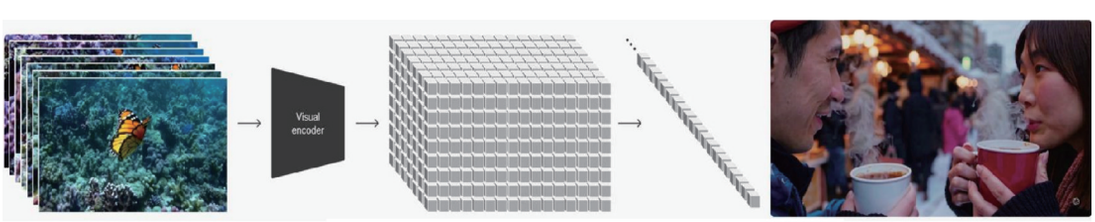
视角动态场景重建上，MonST3R（Zhang 等，2024b）将 DUSt3R 在少量动态三维场景数据上微调，实现了前馈式的动态场景点阵图预测，成功地将数据先验应用到这个高度不适定（Ill-posed）问题上。

当然，DUSt3R 也存在一些缺点，例如仅接受两幅图像作为输入的模式，在多视图或视频输入的场景下会带来高复杂度。为了解决这个问题，可以提升网络一次接受的输入图像数，例如 SLAM3R（Liu 等，2025c）中进行的尝试。另外，DUSt3R 预训练中需要使用带有点阵图标注的数据，这样的数据要求较高，给进一步规模化拓展带来了困难，如何放宽数据要求，利用互联网上的海量 RGB 视频数据，也是走向三维视觉基础模型的重要课题。

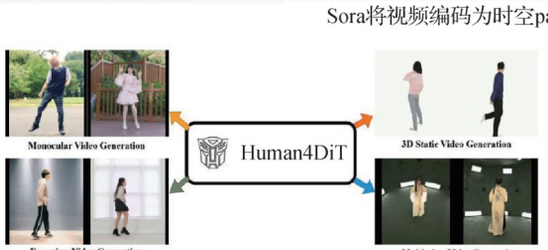
1.2　视频生成开启 4D 空间智能

2024 年，视频生成技术在大模型与海量互联网数据的驱动下实现跨越式发展。当前，视频生成模型正经历从“二维内容生成”向“物理规律感知的 3D 内容生成”的范式跃迁，这一转变在产业界得到强力印证——英伟达最新发布的世界基础模型 Cosmos（Agarwal 等，2025），基于 200 万小时视频训练形成兼具 3D 一致性与物理合理性的视频生成能力，不仅促进了视频生成时空一致性的显著提升，其直接生成合成数据的技术突破将缓解物理人工智能领域长期存在的数据饥渴问题，为 3D 游戏、具身智能和自动驾驶等领域的技术发展提供了全新的视角和支持。

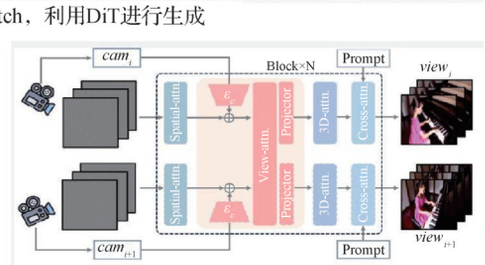
在这一领域，Sora 明确了视频生成技术路线，以 DiT（diffusion Transformer）（Peebles 和 Xie，2023）为基础结构将视频生成的效果推上新的台阶。其所表现出来的对物体规律的初步理解，一定程度上证明了视频模型具备理解和构造 3D 世界模型的潜力，然而当前视频生成模型仍然难以生成多视角严格一致、物理准确的结果。为改善视频生成模型在视角、3D 方面的生成能力，提升视角一致性并建立起视频生成与 3D 的联系，Human4DiT（Shao 等，2024）和 SynCamMaster（Bai 等，2024）等方法探索了如何将 3D 条件如 3D 人体、相机参数等引入视频模型。借助 DiT 结构，这类方法通过注意力机制将 3D 相关的条件引入到视频生成模型当中，实现了时间和视点更加一致的动态 4D 视频生成，如图 3 所示。



Sora将视频编码为时空patch，利用DiT进行生成



Human4DiT利用DiT实现4D人体视频生成



SynCamMaster将相机参数嵌入DiT实现视角控制

图 3　基于 DiT 的视频生成方法

Fig. 3　DiT-based video generation methods

此外，基于视频生成模型强大的可扩展能力，利用其中包含的隐式 3D 结构先验提升 3D 重建效果成为一个重点探索方向。传统的基于多视几何的 3D 重建技术存在若干局限性。首先，这些方法通常要求拍摄的多视图像之间有足够的重叠区域，其次，重建的几何结构和纹理质量通常依赖于图像的分辨率。此外，这些技术在重建弱纹理、高反光或半透明区域时效果较差，且未被捕获的区域无法进行重建。虽然基于 AIGC 的技术在理论上可以突破这些局限，但由于高质量 3D 数据的匮乏，不足以支撑直接的高质量 3D 生成。因此，CAT3D（Gao 等，2024c）和 ViewCrafter（Yu 等，2024b）利用多视角和视频 diffusion 模型中的先验来辅助生成高质量的 3D 多视角内容，使得用 2D 生成模型为 3D 重建提供先验成为

可能，为3D重建提供了新的技术路径。

World Labs（WorldLabs，2025）于2024年11月发布首个3D空间智能模型，只需一幅图像即可生成3D场景，并且具备可交互性和可编辑性，允许用户在3D场景中自由漫游，并实现景深控制、滑动变焦和重打光等多种3D效果。同时，Genie2（2025）将动态视频信息融入3D空间智能，基于单幅图像可生成长达1 min的视频场景画面，并且具备实时交互、物理模拟、空间记忆和多样化的环境生成等特点。这些技术的突破，标志着人工智能开始迈向基于4D的空间智能时代。空间生成作为空间智能中关键组成部分之一，可以通过3D重建与视频生成技术的结合实现。一方面可以通过结合3D重建技术（如large reconstruction model）生成的3D信息引导2D视频的生成，从而提高结果的时空一致性。另一方面，通过生成的时空一致2D视频，再结合3D重建技术可以生成完整的3D场景，而这些完整的高质量3D场景数据可以作为训练样本，通过强化学习机制持续优化3D场景生成模型的性能。

除了以扩散模型为主的视频生成模型外，自回归视频生成的空间智能模型如Genie2带来的交互式生成技术，为视频与3D技术在交互领域（如游戏、场景生成、具身、自动驾驶等）带来新的可能性。其中一类工作是结合多模态自回归大模型，如在自动驾驶方面，DrivingGPT（Chen等，2024b）将驾驶过程建模为交替的视频帧和指令序列，并将两个模态统一为驾驶语言，利用多模态自回归Transformer同时执行建模和端到端的规划，并基于给定的历史驾驶状态进行未来状态的预测。通过多模态统一的方式，加深了对于真实3D世界的理解。通过自回归预测的模式，有助于扩展智能的交互性和可控性。在场景生成方面，StarGen（Zhai等，2025）基于视频生成模型以自回归的方式生成三维空间一致的大范围场景。通过将空间和时序上相邻帧作为条件，StarGen能够用姿态或布局信息控制实现3D空间中的稀疏视角插值以及城市场景生成，展现了一定的空间智能。图4展示了空间智能模型和自回归生成模型的应用。

图4　空间智能模型和自回归生成模型的应用

Fig. 4　Applications of spatial intelligence modeling and autoregressive generative modeling

随着视频生成模型的快速发展，三维视觉研究开启了探索真实物理世界的历程。利用视频生成模型所展现的理解物理规律和可扩展能力，可以帮助对3D世界以及空间智能的建模，为真正的世界模型奠定基础。另一方面，通过新的自回归生成范式将多个模态统一到一起来预测未来状态，有助于模拟真实世界中的多信号动态行为，可以为具身智能体以及自动驾驶等应用带来新的可能。

1.3　3D AIGC——多方向突破下的持续演进

2024年，3D AIGC领域延续了近年来的发展态势，在学术研究与产业应用两方面均取得了显著进展。总体而言，该领域正处于稳步发展期，技术迭代迅速，应用场景不断拓展，但距离构建完善且成熟的技术体系仍存在一定距离。

学术界在2024年继续保持着高度活跃的研究态势，主要聚焦于几何生成、纹理生成、场景级生成以及动态4D生成等关键方向，并取得了系列创新成果。这些方向代表了当前3D AIGC领域的核心研究前沿，旨在解决三维内容创建中存在的精细度、逼真度、复杂度和动态性等关键问题。产业界则积极探索3D AIGC技术的落地应用，多家公司推出了相关的3D生成平台或工具，Tripo、Rodin和Meshy等公司的相关产品展示了文生3D、图生3D技术在各方面的应用潜力，预示着3D AIGC技术正逐步从实验室走向实际应用。图5展示了3D AIGC领域2024年部分代表性工作。

高质量几何生成(CLAY)　真实多样纹理生成(TEXGen)　大规模场景重建(Long-LRM)

功效性生成(MeshArt)　表达无关的结构化潜空间(TRELLIS)　动态4D生成(CAT4D)

图5　3D AIGC领域2024年部分代表性工作

Fig. 5　Selected representative works on 3D AIGC in 2024

1.3.1　几何生成

几何生成作为构建三维模型的基础，精细化与结构化并举，在2024年得到进一步发展。CLAY(Zhang等，2024d)在SIGGRAPH 2024上获得最佳论文提名，通过对3DShape2Vecset(Zhang等，2023)进行充分scale up，验证了扩散模型在原生3D数据领域的可泛化生成能力，实现了在三维物体级别上精细几何、准确结构的生成。2024年年底发布的TRELLIS(Xiang等，2024)提出一种表达无关的三维结构化潜空间，其稀疏性、局部化的表达特点使得重建和生成过程更容易被训练，效果也引发了开源社区讨论。

1.3.2　纹理生成

纹理生成旨在为三维模型赋予逼真的外观，实现逼真度与可控性的平衡。TEXGen(Huo等，2024)在SIGGRAPH Asia 2024上获得最佳论文提名，该方法提出一个大规模扩散模型，实现了在3D模型的UV空间进行直接、快速和高质量的纹理生成。通过创新的2D-3D混合架构，实现了高分辨率细节的保持和三维一致性。该方向的研究重点在于提升生成纹理的真实感和多样性，同时探索更有效的方法来控制纹理的生成过程，使用户能够根据需求生成特定风格或属性的纹理。

1.3.3　场景级生成

从单一物体到复杂场景，场景级生成的目标是构建包含多个物体、具有复杂布局和丰富细节的三维场景。2024年，大型三维重建模型(large reconstruction model，LRM)(Hong等，2024))如GS-LRM(Zhang等，2024c)、LongLRM(Chen等，2024c)以及结合图像、视频生成预训练模型能力的方法如CAT3D等，在该方向上取得了显著进展，展现出较强的泛化能力。通过大规模数据预训练，这些模型在重建规模、精度、速度和通用性上都取得了明显提升，为复杂场景的自动生成和重建奠定了基础。

1.3.4　动态4D生成

三维生成的另一个重要趋势是从静态三维物体的生成扩展到动态三维物体的生成，其主要挑战在于捕捉时空信息的新挑战。动态4D生成是近年来新兴的研究方向，旨在生成具有时间维度信息的三维动态场景。CAT4D(Wu等，2024b)、SynCamMaster

(Bai等,2024)等工作利用视频生成基础模型,实现了时间和视点一致的动态4D生成。这类研究的难点在于如何捕捉和表达三维场景随时间变化的复杂信息,并保证生成结果的时空一致性。尽管该方向仍处于起步阶段,但已展现出明显的潜力和广阔的潜在应用前景。

1.3.5 内在结构/功效性生成

随着三维生成技术的不断发展,生成精美外观的三维物体已变得可能。然而,三维物体不仅需要具备吸引人的"外表",还需要拥有实际的"功能"。因此,三维生成的下一步目标应聚焦于不仅呈现物体的外形,还能够生成物体的内在结构和功能性。为实现这一目标,加拿大西蒙菲莎大学提出Slice3D(Wang等,2023)方法,通过切片技术生成具有内在结构的三维物体,从而获取物体的分层结构和更为复杂的内部构造。此外,德国慕尼黑大学提出的MeshArt(Gao等,2024a)方法进一步推动了这一方向,支持直接生成带有铰链结构的物体,例如可以旋转的椅子等。这项技术突破了传统三维生成模型的局限,使得生成的物体不仅拥有逼真的外观,还具备实际的功效性。

1.3.6 小结

尽管3D AIGC领域在2024年取得可观进展,但现有技术仍存在明显的局限性。首先,生成模型的输出质量与真实世界相比仍有一定差距,尤其是在处理复杂场景和精细结构时,生成结果往往缺乏足够的细节,物理的准确性和语义的一致性也难以得到保证。其次,生成过程的可控性和可编辑性仍然不足,用户难以对生成结果进行精确的控制和调整,这限制了该技术的实用性和灵活性。此外,高质量训练数据的获取和标注仍然是一个挑战。

展望未来,3D AIGC领域将继续朝着更高质量、更可控、更具结构化和实时性的方向发展。未来的研究将致力于进一步提升生成模型的表达能力,生成更加精细、逼真、符合物理规律的三维内容。同时,在算法上,需要构建更加高效、自适应的生成框架与跨模态数据处理能力,例如通过深度学习与传统图形学的融合,打造多分辨率生成与训练机制;并在数据层面积极探索激光雷达、影像测量和传感器信息等多模态数据的采集及半自动监督标注方式,以降低人工成本,提升数据多样性和精度。此后,开发更有效、更直观的控制和编辑方法,使用户能够自由操控生成过程并对结果进行细致调整,也将是重要的研究方向。基于语义信息等结构化信息的生成方法有望得到更多关注,从而提高生成结果的可理解性和可控性。此外,随着硬件性能的提升和算法的优化,实时生成技术将持续进步,为虚拟现实、增强现实等应用带来更流畅、更具沉浸感的体验。最终,3D AIGC技术有望与智能体(AI agents)等技术深度融合,实现更智能、更自主的三维内容创作流程,推动相关产业的变革与发展。

1.4 高斯泼溅处理方法和工具链日趋完善

三维高斯泼溅(3D Gaussian splatting, 3DGS)(Kerbl等,2023)技术发表于SIGGRAPH 2023,并获得最佳论文奖,一经发布就受到学术界和工业界广泛关注,成为三维计算机视觉方向最热点的研究内容之一。其具有渲染速度与训练速度快、真实感强的优势,被业界认为是下一代的高真实感三维表示技术,因此相应的处理方法和工具链是高斯泼溅技术取得广泛应用的前提条件,并将有望形成全新的高真实感三维重建与编辑的平台。这一点在学术界和工业界得到普遍共识。在2024年学术界和工业界围绕高斯泼溅在基础表示、重建与编辑、重光照、物理仿真以及压缩存储等方面开展了大量基础性的研究工作,有多篇综述论文(Chen和Wang,2025;Fei等,2024;Wu等,2024d)对这些工作进行梳理。这些基础性研究工作逐渐完善了高斯泼溅的处理方法和工具链并为高斯泼溅的广泛应用奠定了基础。高斯泼溅正在成为三维数字人、智能驾驶和数字城市等应用领域的关键底座技术,成为构建高真实感数字世界模型的基础三维表征。同时人工智能的发展正在从数字世界走向实体世界,对三维世界的理解与重建并在三维世界中进行具身操作将成为新一代人工智能的基石,平台化的高斯泼溅技术将对人工智能的应用从虚入实起到关键推动作用。高斯泼溅的原理如图6所示。

本节在高斯泼溅的基础表示、变形编辑、重光照、物理仿真和存储压缩方面回顾2024年度的代表性的研究工作。

在基础表征方面,SuGaR(Guédon和Lepetit,2024)首先通过添加自监督损失提升了三维高斯泼溅表示重建几何表面的能力,2DGS(Huang等,2024)和Gaussian Surfels(Dai等,2024)都提出基于二维高斯泼溅的几何表示方法,该类表示方法能够

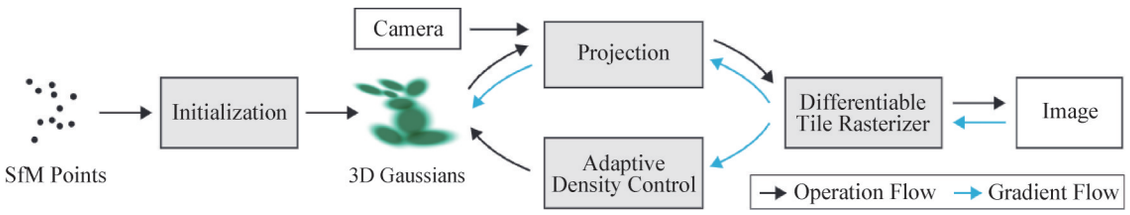


图 6　高斯泼溅的原理示意图

Fig. 6　Schematic diagram of the principle of Gaussian splatting

对具有高度复杂性的几何结构进行精确还原，进一步提升了几何细节的重建精度，为高精度三维场景的建模奠定了重要基础。在几何重建的基础上，可以将网格和高斯泼溅进行混合表示来提升对高斯泼溅的变形效果，如 GaussianMesh（Gao 等，2024b）和 GaussianAvatar（Hu 等，2024b）将三维高斯绑定在网格表面，通过网格表示的几何变形来驱动三维高斯泼溅表示的变形，以支持大尺度的变形驱动。为了进一步仿真满足物理规律的动态三维场景，Phys-Gaussian（Xie 等，2024）通过在重建的高斯场中加入物理属性，利用物质点法等软体流体数值仿真解算，从而实现物体在不同环境下的物理动态变化和不同的交互效果。PhysDreamer（Zhang 等，2025）在 PhysGaussian（Xie 等，2024）的基础上，利用视频模型生成的视频来反推 3D GS 场景中的物理属性，进而利用物理仿真，生成物理更准确且语义和视频一致的动态。类似地，Spring-Gaus（Zhong 等，2025）将弹簧质点仿真和高斯泼溅表示相结合，支持从视频估计弹性系数，并利用仿真预测弹性体动态。这些技术在三维场景的仿真与交互式应用中具有广泛的潜力。图 7 展示了高斯泼溅在基础表示、变形编辑和物理仿真方面的代表性研究工作。

在外观编辑方面，RelightableGS（Gao 等，2025）在三维高斯泼溅表示的渲染中引入基于物理的渲染方程实现该表示的重光照编辑，DeferredGS（Wu 等，2024c）和 GSDeferred（Ye 等，2024b）提出延迟渲染策略来建模更加复杂的镜面反射输入。为了进一步拓展三维高斯泼溅的渲染质量和应用场景，3DGRT（Moenne-Loccoz 等，2024）、RayGauss（Blanc 等，2025）和 EVER（Mai 等，2024）等工作探索了与传统渲染方式的融合，使用光线追踪方法（ray tracing）对高斯泼溅进行渲染。通过结合光线追踪技术，实现高精度的光线反射、折射模拟，三维高斯泼溅在真实感渲染中的表现力得到了显著提升，并且支持和传统的三维表示如网格等表示的联合渲染。为了对高斯泼溅进行压缩存储与传输，FCGS（Chen 等，2025）将不同的高斯属性分配到独立的熵约束路径上并设计了高斯间和高斯内的上下文模型，进一步提升了压缩效率。图 8 展示了高斯泼溅在重光照、基于光线跟踪的渲染方法和存储压缩方面的代表性研究工作。

通过这一系列的探索和优化，三维高斯泼溅表示的处理方法和工具链日益完备，为未来应用提供

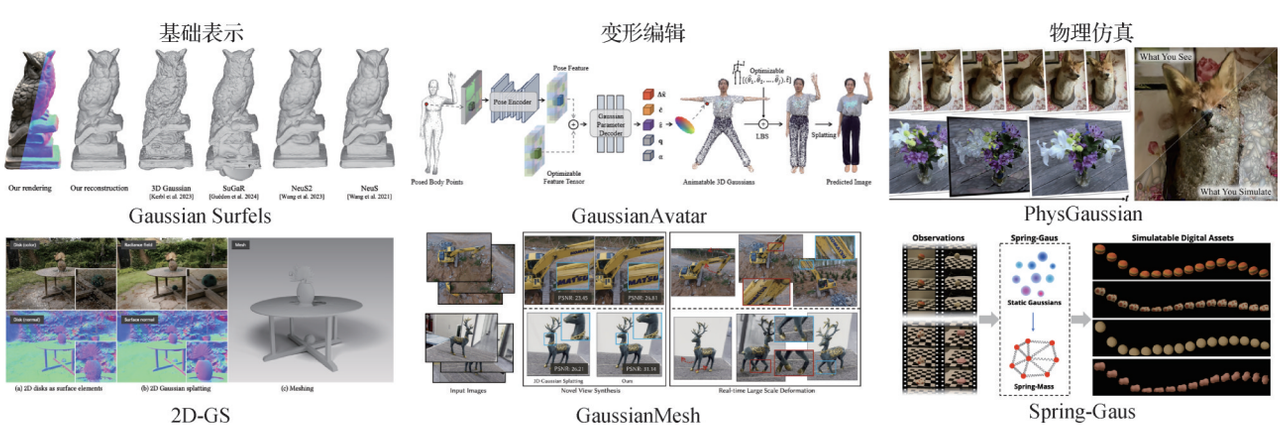


图 7　高斯泼溅在基础表示、变形编辑、物理仿真方面的代表性研究工作

Fig. 7　Representative research works on Gaussian splatting in fundamental representation, deformation editing, and physical simulation

了平台化的技术支持。当然现有的建模、编辑与渲染技术仍然存在需要密集的视角输入和准确的相机位姿问题，如何降低采集的图像要求，利用已有的图像或视频生成模型为三维高斯泼溅的建模、编辑与渲染提供先验，是值得继续探索的方向。如何将平台化的高斯泼溅技术应用在具身智能、空间智能、生成式人工智能，进一步推动人工智能从虚到实，赋能实体经济仍是需要发力的方向。

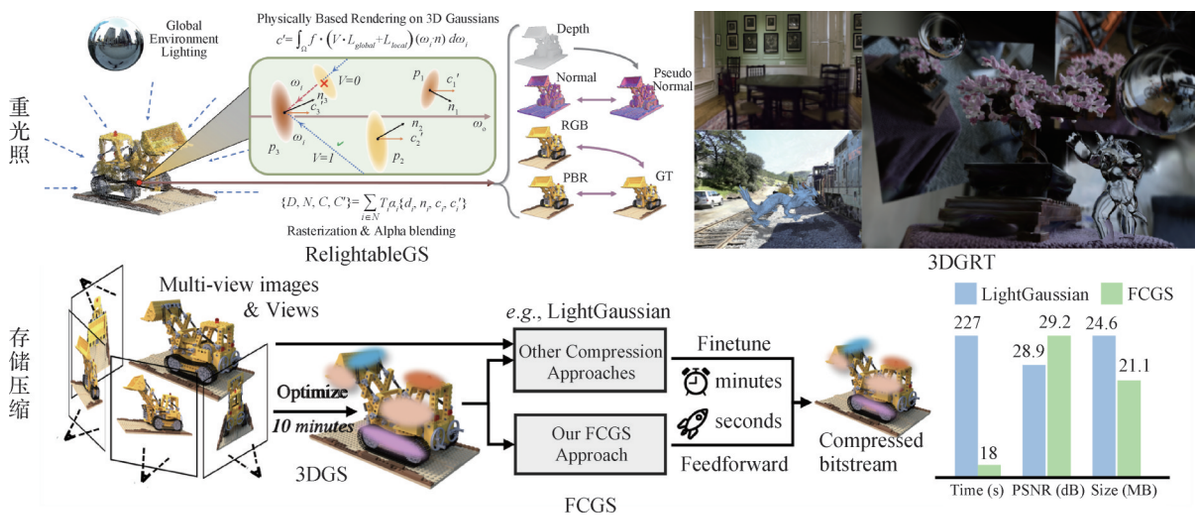


图8　高斯泼溅在重光照、基于光线跟踪的渲染方法和存储压缩方面的代表性研究工作

Fig. 8　Representative research works on Gaussian splatting in relighting, ray-tracing based rendering methods and storage compression

1.5　3DGS走向应用：静到动、小到大的场景重建进化

3DGS（Kerbl等，2023）针对静态、尺度较小的场景设计，无法直接适用于动态与大规模场景。2024年度，基于3DGS的场景重建在从静态到动态、从小场景到大场景方面都取得可观进展，这些时间维度和空间维度的扩展推动了3DGS在沉浸式媒体、自动驾驶和城市建模等方向的应用。

在时间维度上，针对通用动态场景的3DGS重建进展丰富，主要朝着表征轻量化、采集平民化的方向发展。首先，不同的4D表征呈现多家争鸣的状态，如4DGS的4维高斯基元建模（Wu等，2024a）、3DGStream的逐帧三维高斯建模（Sun等，2024）、4DGaussians（Wu等，2024a）和Deformable 3DGS（Yang等，2024）的规范空间与逐时刻形变联合建模等方式，这些表征方式的参数化形式虽然各不相同，其背后的核心逻辑具有一定相似性，即在每个时刻输出一组三维高斯，并建立多个时刻三维高斯之间的相关性，从而提升时域一致性，以及通过紧凑的参数化方式减小建模动态场景所需的逐帧参数量。其次，在视频采集方式方面，单目视频相比于多目视频更易获取、更加平民化，因此基于单目视频的动态场景重建也是领域的研究热点，为了提升单目视频动态重建的质量，主要思路是引入更多先验知识，从而克服单目视频在时空上的稀疏性，例如Shape of Motion（Wang等，2024d）引入单目深度估计和光流等先验，CAT4D（Wu等，2024b）利用扩散模型的先验生成更多训练图像。

动态场景中还有一类典型且具有实用价值的场景是驾驶场景，因为驾驶场景的自由视角合成为面向自动驾驶的写实仿真平台提供了可能性。2024年度3DGS在动态驾驶场景重建方面也取得了系列进展，主要关注问题在于如何建模场景中的动态车辆和行人，并实现动态场景的编辑。考虑到驾驶场景的动态车辆都属于刚性运动，DrivingGaussian（Zhou等，2024c）等多数动态驾驶场景重建方法利用动态车辆的三维标注框实现了动静解耦，从而进一步赋能动态车辆的添加、删除以及位姿控制等编辑操作，Street Gaussians（Yan等，2025）和HUGS（Zhou等，2024b）额外考虑了对车辆三维标注框的联合优化。针对非刚性运动的行人，OmniRe（Chen等，2024d）进一步基于行人的三维标注框和人体网格模板参数实现了街景动态行人的重建。

在空间维度上，2024年也有多个工作将3DGS

扩展到城市级别的大规模场景重建。大场景重建的一个主要挑战在于三维高斯数量过多时如何减小占用显存和提升渲染实时性。一类思路是通过分而治之的分块方式将场景划分为多个部分，每部分单独训练，例如 VastGaussian(Lin等，2024)；一类是通过细节层次(level of detail, LOD)的方式由粗到细地表达大规模场景，根据相机距离来选择渲染的三维高斯基元的层级颗粒度，从而保证大规模场景的渲染实时性，例如 Hierarchical 3DGS(Kerbl 等，2024)、CityGaussian(Liu等，2025b)和Octree-GS(Ren等，2024)等。

随着 3DGS 在时空扩展上的可观发展，相关原型应用也相继推出，例如 V3(Wang等，2024c) 实现了基于 3DGS 的动态三维视频在轻量化端侧设备的流式播放，Long Volumetric Video(Xu等，2024e)实现了 10 min的长视频重建，推进了沉浸式媒体的发展；HUGSIM(Zhou等，2024a) 构造了基于 3DGS 的端到端闭环仿真平台，实现了动态车辆的交互式插入，构建了智能驾驶算法和写实仿真算法的闭环交互；LetsGo(Cui等，2024) 实现了大规模场景的细节层次重建，并且实现了轻量化端侧设备的实时渲染，赋能了车库场景的实时定位导航。这些原型应用为3DGS的落地提供了正面范例。

图 9 展示了 3DGS 从静到动、从小到大的场景重建及其相关应用。

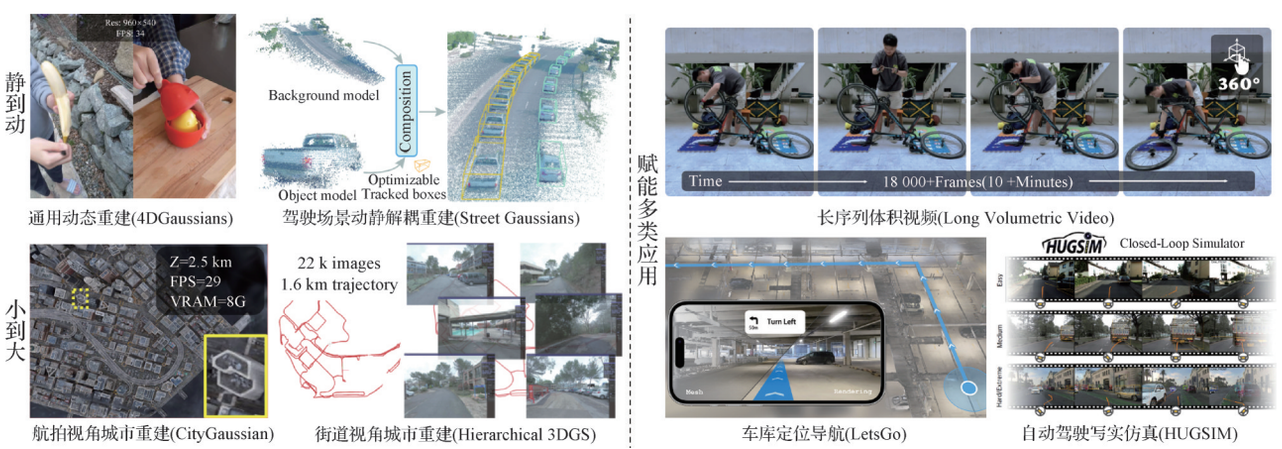


图 9　3DGS 从静到动、从小到大的场景重建及其相关应用

Fig. 9　3DGS scene reconstruction from static to dynamic, from small to large, and its related applications

尽管取得了以上的可观发展，3DGS场景重建在进一步落地实用方面仍有许多值得进一步探讨的问题。首先，目前领域缺少一个公认的具有突出优势的4D表征，而未来的沉浸式媒体应用中是否需要一个统一的4D表征、如果需要统一的话该采用什么表征都仍然是未知数；其次，动态场景和大规模场景的泛化高效重建目前尚处于起步阶段，2024年虽然有这方面的相关探讨，然而其质量相比于静态小规模的泛化3DGS重建仍值得提升；最后，目前的工作相对独立地在时间维度或空间尺度上进行扩展，未来构建长动态视频、大规模场景的高效重建与实时渲染也是值得探索的方向。

1.6　3DGS助力3D数字人突破恐怖谷效益瓶颈

2023年底以来，3DGS(Kerbl等，2023)技术被引入到3D数字人重建、动画与生成中。通过将三维数字人表示为一组高斯点及其属性(如位置、半径、颜色等)，实现光栅化高效渲染，得以保证实时渲染。同时高斯点的离散分布方式使得数字人表面能够表现出更自然和真实的细节，尤其是能对一些褶皱或毛发等细微结构建模。3DGS技术的引入带来了在渲染速度慢、建模时间长、缺失细节等各项瓶颈中的突破性进展。在3D数字人技术的落地进程中，具有重要的里程碑意义。

在 3D 数字人重建方面，GPS-Gaussian(Zheng等，2024)和GHG(Kwon等，2025)等工作提出基于多视点图像或在人体参数模型上学习三维高斯点云表征，通过在较大规模的人体扫描数据上训练，可以通过前馈式网络从多视点图像输入直接预测三维人体高斯模型，取得了高质量的三维人体重建。

在 3D 数字人动画方面，Animatable Gaussians(Li等，2024b)提出将人体高斯点定义在标准姿态下的正反投影图上，并基于卷积神经网络回归姿势相

关的高斯图实现从多视角视频输入建模3DGS人体数字人，相较于基于NERF的表征方法，其细节建模能力更强、渲染速度更快。之后，ExAvatar（Moon等，2025）提出解耦式外观学习，在单目视频输入下取得了精细的数字人驱动效果。除了肢体动画，许多工作，包括GaussianAvatars（Qian等，2024）以及Gaussian Head Avatar（Xu等，2024c）更关注人头驱动，它们通过将3DGS点云绑定至Flame（Li等，2017）人头模型或自适应重建模型上，实现了可表情驱动的人头数字人建模。这两项工作都证明其方法在重建精度和渲染速度上大幅度超越先前的人头数字人。在此基础上，NPGA（Giebenhain等，2024）引入预训练的隐式表情作为驱动信号，进一步提升复杂表情下的表情准确性和渲染质量。此外，GPHM（Xu等，2024d）提出基于3DGS的人头模板并实现外貌和表情的解耦。随后的URAvatar（Li等，2024a）等系列工作，首先在3DGS模型中加入对法向的光传输方程的预测以实现重光照，其次利用多人数据集预训练先验模型以实现从手机端快速重建个人定制3DGS数字人。图10展示了3DGS数字化身生成。

图10　3DGS数字化身生成

Fig. 10　3DGS digital avatar generation

借鉴图像生成大模型，在3D数字人生成方面，Human-Gaussian（Liu等，2024a）、Human3Diffusion（Xue等，2024）以及FaceLift（Lyu等，2024）等工作提出基于Stable Diffusion（Rombach等，2022）实现基于单图或文本的三维高斯数字人生成，得益于基于Diffusion生成式模型的强大能力以及3DGS的高效表征，这类方法可以实现快速、高质量的三维数字人生成，超越以往NeRF类方法。2024年底以来，随着视频生成技术迅猛发展，围绕视频生成模型的可控性研究涌现出众多成果，为基于单幅图像的高真实感2D数字人驱动带来突破。一方面，借助基础视频模型强大的生成能力，AnimateAnyone（Hu等，2024a）和MagicAnimate（Xu等，2024f）等工作通过结合2D动捕信息控制视频生成，成功实现了可由视频驱动的2D数字人；EMO（Tian等，2025）和Hallo（Xu等，2024a）等工作则引入1D音频信息控制视频生成，构建可由音频驱动的2D口播数字人；Human4DiT更进一步，集成相机视角控制信息，实现动态4D数字人生成。另一方面，VASA（Xu等，2024b）等工作通过预先解耦动作与人物外观的隐空间，仅在一维动作空间中使用扩散模型进行建模，大幅降低了全流程计算开销，在生成多样化动作的同时支持实时画面渲染。得益于扩散模型强大的可扩展性，这类方法能够在海量互联网人物视频数据上进行训练，使得现阶段2D数字人在渲染视频真实感方面显著超越3D数字人，特别是在风格泛化性和动态驱动效果上展现出明显优势。图11展示了2D数字人生成与3D高斯数字人蒸馏。

基于此，一方面，研究者开始探索如何利用2D数字人模型的高质量生成数据进一步推动3D数字人的发展。CAP4D（Taubner等，2024）和PERSE（Cha等，2024a）等工作通过单幅正面图像输入，利用视频生成模型合成多视角或多表情数据，并蒸馏到3DGS数字人模型中，实现了基于2D视频生成的高效3D重建。这类方法不仅验证了视频生成技术在3D数字人建模中的潜力，还显著提升了重建的精度与表现力。另一方面，通过联合建模2D/3D统一表征，构建3D数字人大模型仍是一个有待探索的方向。在

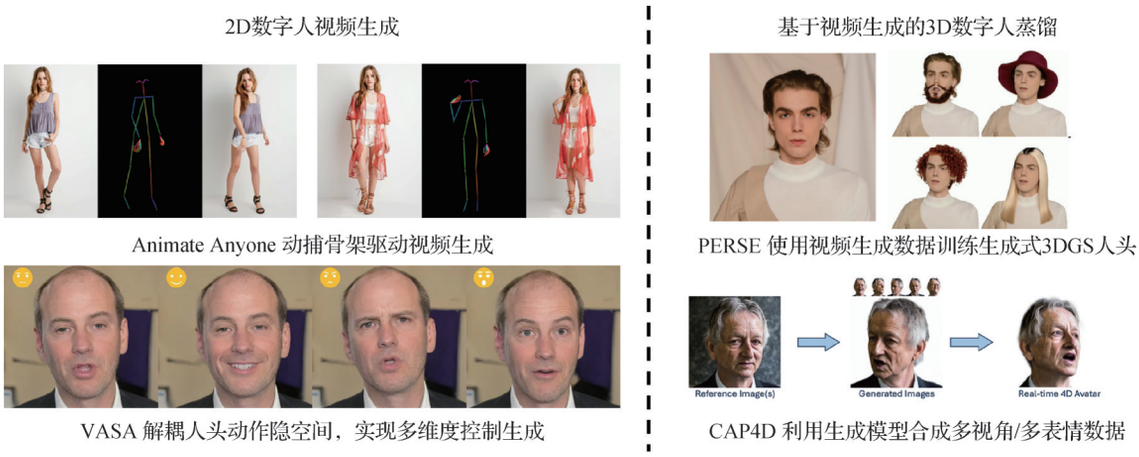


图 11　2D 数字人生成与 3D 高斯数字人蒸馏

Fig. 11　2D digital human generation and 3D Gaussian digital human distillation

现有的 2D 视频生成框架中嵌入 3DGS 特征空间，可以借助视频模型的扩展能力直接基于海量互联网视频数据建模通用 3D 数字人表征。这种方法有望显著降低 3D 数字人的建模成本，同时进一步提升其在动态表现和真实感方面的能力。

以上系列研究初步验证了 3DGS 技术在数字人建模领域带来的重大变革，并展示了其在精度和渲染速度上的优势。这些进展也为数字人的生成式重建提供了新的研究范式。然而，尽管在形象建模方面取得了显著进展，但依然存在如下问题：1）相比 2D 数字人，3D 数字人在语音驱动下表情和口型自然性问题依然存在；2）对于动作驱动，依赖从动作参数生成到外观形象驱动生成的两步骤方案，仍面临穿模、外形动作不逼真等难题；3）由于缺少相应的大规模 3D 数据，3D/4D 数字人仍未形成基础大模型，数字人的条件生成仍依赖复杂长时间的优化步骤，难以高质量前馈泛化生成。上述都是未来高斯数字人需要解决的核心难题。

1.7　三维视觉助力具身大数据构建

2024 年，具身智能成为科技领域最受瞩目的焦点。相比于大语言模型（large language model，LLM）和多模态大模型所依赖的互联网数据，训练具身智能所依赖的海量三维动作及交互数据等无法轻易获得，数据层面的大规模、高质量和高效获取成为具身智能致胜的核心关键。具身数据收集与高效利用方面取得诸多进展，数据获取来源主要包括 3 类：海量人类动作视频数据、人工在环遥操作交互动作数据和三维虚拟仿真数据。三维视觉技术在以上数据获取技术上都发挥了重要作用。

1.7.1　从人类动作视频学习机器人策略

互联网视频作为一种丰富的数据源，蕴含大量物理信息和运动行为，但其由于缺乏动作标签难以被提取和利用。对此，谷歌 Deep-Mind Vid2Robot（Jain 等，2024）收集了人类视频与机器视频动作数据对，借助直接视觉模仿学习，训练机器人完成与人类视频相同的机器动作。然而由于成对的人类视频与机器视频动作数据匮乏，Video-Diff（He 等，2024a）通过视频预测指导策略学习，将人类视频与机器视频压缩到统一的嵌入空间，并利用大规模人类视频进行预训练后在少量机器动作数据进行微调，从而将人类视频中蕴含的物理世界的动态知识迁移到机器人策略的学习过程。进一步地，由于视频预测模型物理合理性欠缺、计算开销大，Track2act（Bharadhwaj 等，2025）、Dreamitate（Liang 等，2024b）和 ATM（any-point trajectory modeling）（Wen 等，2024）等方法选择忽略像素级别的细节，转而从大规模人类视频数据集中预训练模型以预测物体关键点的移动方向，进而将其映射为机械操作指令，从而在效率上超越了直接基于视频预测的方法。值得注意的是，目前上述研究大多集中在二指抓取器的操作任务上，而如何从人类视频中学习五指灵巧手的操作策略，仍是一个亟待探索的研究领域。图 12 展示了从人类动作视频学习机器人策略方面年度代表性工作。

1.7.2　人工在环遥操作数据采集与模仿学习

尽管人类视频数据资源丰富，但由于人体和人手与机器人物理形态存在差异，从人类视频训练的控制策略难以准确映射到现实机器人中。遥操作数

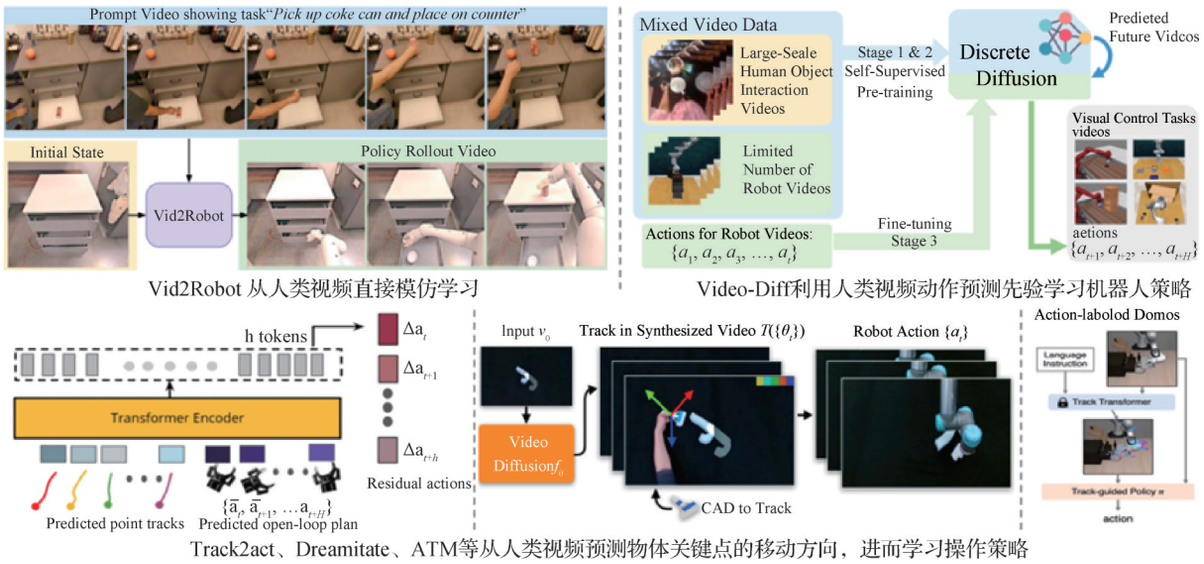


图 12　从人类动作视频学习机器人策略方面年度代表性工作

Fig. 12　Annual representative works on learning robot strategies from human action videos

据采集通过人类直接控制机器人在真实世界中执行任务，不存在跨域鸿沟，数据质量更高。工业界主流遥操方案包括光学动捕、惯性动捕等。光学动捕通过高速红外相机捕获标记点的三维位姿，能够实现亚毫米级精度，但其成本高便携性差。惯性动捕通过可穿戴设备内置的惯性测量单元（inertial measurement unit，IMU）推算运动物体位姿，成本较低但存在长时间漂移的问题。另外在遥操数据采集过程往往结合多相机系统，通过多视角视觉信息融合，实现3D空间感知理解。

美国斯坦福大学 Mobile Aloha（Fu 等，2024b）设计的一个低成本可移动的全身遥操数据采集系统利用所收集数据基于行为克隆，完成炒菜上菜、打电话并进入电梯等复杂移动操作任务。另一方面，常见遥操数据采集过程需要人类操作真实机器人执行任务，斯坦福 UMI（universal manipulation interface）（Chi 等，2024）设计一种低成本的手持平行夹爪，简单通过人类握持夹爪执行任务并录制数据，无需机器人实体，成本更低，便携性更高，数据采集更高效。同时，UMI 记录夹爪的六维空间运动轨迹，而非具体的关节角度，可以映射到任何具有 6 个自由度的机器人。另外，由于光学动捕与惯性动捕的不足，DexCap（Wang 等，2024a）使用 EMF（electromagnetic fields）电磁式动捕手套联合多视角相机来完成简单高效的数据采集，并通过逆运动学动作映射与基于点云的生成式行为克隆策略学习灵巧操作。近期发布的人工在环遥操作数据集 AgiBot World（Contributors，2024）是全球首个基于全域真实场景、全能硬件平台和全程质量把控的百万量级真机数据集，相较于 Google 的 Open X-Embodiment（Open X-Embodiment Collaboration，2024）数据集，AgiBot World 长程数据规模高出 10 倍，场景范围覆盖面扩大 100 倍，数据质量从实验室级上升到工业级标准，但其采集成本高，无法跨越不同本体进行泛化。图 13 展示了人工在环遥操作数据采集年度代表性工作。

1.7.3　三维虚拟仿真数据提高收集高效性与精确可控性

对于抓取放置等灵巧操作任务，要求实现精确的位姿控制，而人工遥操数据难以达到足够控制精度；同时，真机数据采集风险高，容易对机器人本体或物体造成损坏；此外，要保证训练后机器人策略的泛化性，需要数据集包含物体几何结构、外观材质、空间位置、背景及光照等各个方面指数级的多样化数据样本，真实世界难以通过控制变量完成数据采集；另外，基于人工在环遥操作采集大规模数据成本高，效率低。对此，三维虚拟仿真系统基于物理引擎模拟真实世界的物体材质、光照和力学等信号，通过代码编程精确控制数据的多样性变量，通过图形处理器（graphics processing unit，GPU）并行计算进行高效数据合成，可以生成大规模具身数据。由美国加州大学圣地亚哥分校（University of California San Diago，UCSD）和 Hillbot 公司共同开发的 ManiSkill3

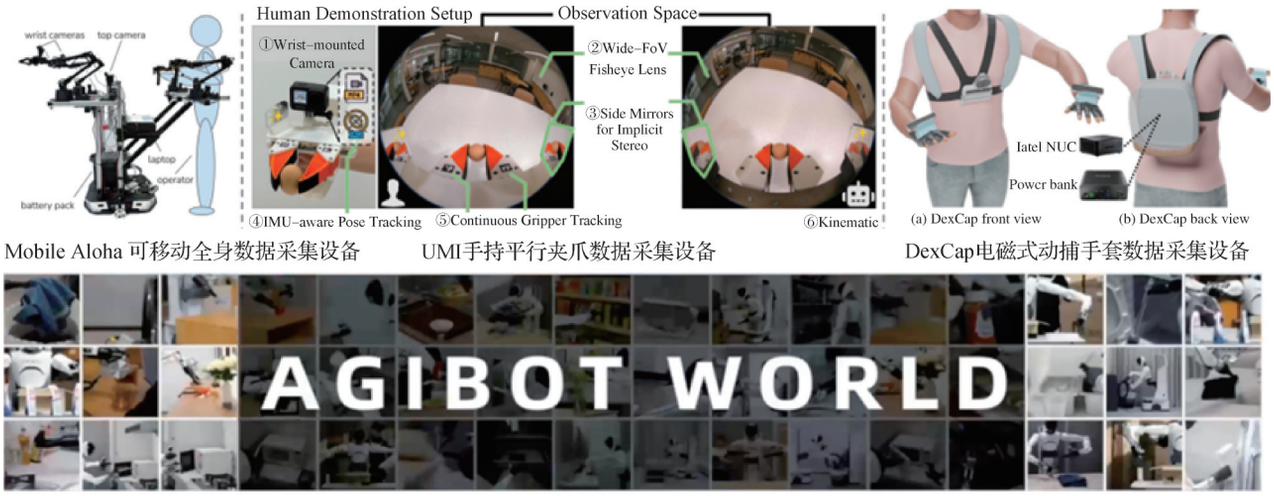


图 13　人工在环遥操作数据采集年度代表性工作

Fig. 13　Annual representative works of human-in-the-loop teleoperation data collection

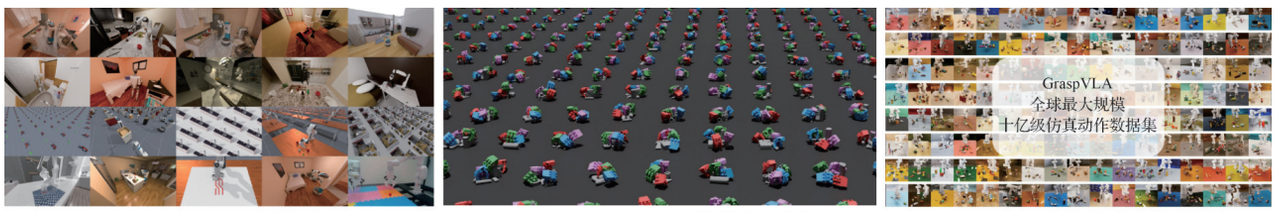
(Tao 等，2024) 提出一种先进的机器人仿真与渲染框架。该框架在仿真与渲染、异构仿真以及点云/体素视觉输入等多个方面实现了 GPU 并行化，能够达到 30 000 帧/s 以上的超高速度，同时使用比传统方法少 2～3 倍的 GPU 内存。ManiSkill3 不仅支持对多种机器人控制器的仿真，还集成了多种具身智能任务，包括四足机器人、双足机器人、固定臂物体操作、全身控制移动物体操作、基于触觉反馈的操作以及复杂环境中的长程操作等，并支持基于虚拟现实的遥操作系统。此外，ManiSkill3 通过 Sim2Real 技术缩小了仿真环境与真实环境之间的差距，使得在仿真器中训练的机器人操作能力能够直接迁移到真实机器人上。同时，结合 Real2Sim 方法，ManiSkill3 能够构建高精度数字孪生环境，帮助对机器人操作策略的能力进行全面评测(参考 UCSD 与谷歌公司合作项目 SIMPLER)，为机器人技术的开发与评估提供了高效且可靠的平台。DexGraspNet 2.0(Zhang 等，2024a)针对嘈杂场景灵巧抓取任务，合成了包含 1 319 个物体、8 270 个场景和 4.27 亿个抓取样本的数据集，并提出一种两阶段抓取模型，实现了真实环境 90.7% 的成功率。GraspVLA 合成了全球最大规模 10 亿级数据集，并使用统一表征实现与互联网数据的高效融合，训练了全球首个全面泛化的端到端具身抓取基础大模型。尽管仿真合成数据大大加速了数据收集，但仍然存在现实差距，难以完全模拟真实的传感信号、复杂物体材质以及真实行为的变异性与不可预测性等，存在一定的虚拟—现实鸿沟。

2024 年，具身智能在数据收集与高效利用方面取得显著进展，三维视觉技术成为关键推动力。从人类视频中提取运动规律、通过低成本设备实现高效数据采集，到利用仿真技术合成海量高质量数据，这些方法在一定程度上缓解了数据匮乏的困境。展望未来，数据高效利用仍是推动通用具身智能发展的核心动力。当前，人工在环遥操作数据虽然为核心，但面临高成本、低效率、动作灵活性损失及跨本体应用受限等挑战。未来研究将聚焦于提升数据真实感与多样性，采集力触视觉多模态数据并挖掘深层关联，利用非在环控制的人手动作交互数据，以及通过仿真与现实的闭环优化缩小“模拟—现实”差距。图 14 展示了仿真数据合成方面代表性工作。

1.8　人形机器人从人类运动中学习通用交互技能

随着人形机器人硬件技术快速发展，具身智能领域的研究者越来越关注人形机器人交互技能的学习。由于人形机器人的本体结构与人体高度相似，从人类动作中汲取灵感以学习交互技能已成为一个富有潜力的研究方向。2024 年，三维视觉领域的研究者一方面专注于逼真的人类交互动作捕捉与生成，提出众多具有物理真实性和类人可信度的数字人交互动作生成模型；另一方面，通过从大量人类交互运动中学习，多项研究工作实现了人形机器人模仿人类运动的能力。

在动作捕捉与生成领域，许多人体运动交互数据集相继提出，如 Motion-X(Lin 等，2023)、InterHuman(Liang 等，2024a) 和 TACO(Proença 和 Simões，

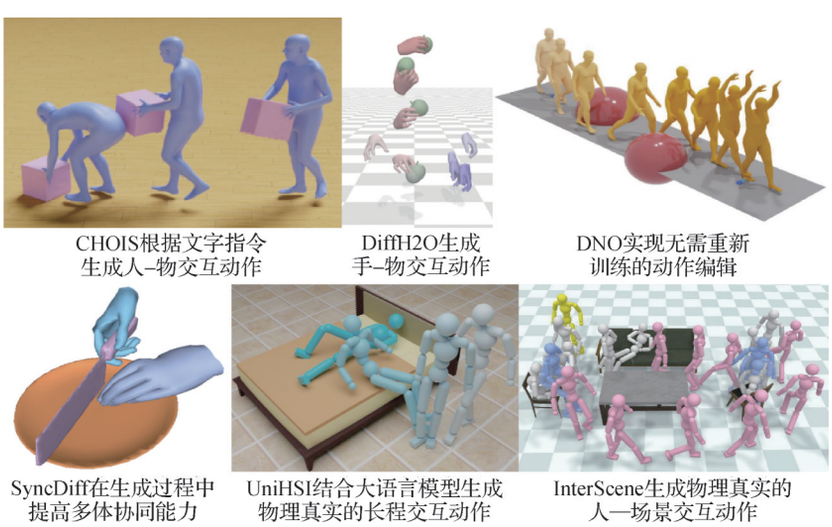


ManiSkill3 仿真渲染框架　DexGraspNet 2.0虚拟仿真合成数据　GraspVLA10亿级仿真动作数据集

图 14　仿真数据合成方面年度代表性工作

Fig. 14　Annual representative works on simulation data synthesis

2020)等。为丰富交互运动的多样性,大量研究致力于从这些数据集中学习,以生成包含人类与周围场景、操作对象或其他参与者互动的动作。这些研究主要覆盖全身交互(如 ROAM(Yang 等,2020)、CHOIS(Li 等,2025)和 InterGen(Liang 等,2024a)等)与手物交互(如 Text2HOI(Cha 等,2024b)、DiffH2O(Christen 等,2024)和 MACS(Zhang 等,2008)等)两个层面。通过对交互行为进行合理表征,并结合扩散模型等强大的生成技术,动作生成在 2024 年实现了显著的真实性和多样性提升。其中,SyncDiff(He 等,2025)提出一种在扩散模型推理过程中提升人类与物体运动同步性的方法,同时支持了全身交互、手物交互等多种类型复杂交互的生成。而 DNO(Karunratanakul 等,2024)则聚焦于提升生成模型的灵活性,提出一种无需重新训练的模型应用技术,使用户能够根据需求自由编辑动作。此外,UniHSI(Xiao 等,2024)和 InterScene(Pan 等,2025)等研究将动作生成与物理仿真相结合,实现了人类在场景中符合物理规律的动作生成。图 15 展示了人体动作生成相关代表性工作。



CHOIS根据文字指令生成人–物交互动作　DiffH2O生成手–物交互动作　DNO实现无需重新训练的动作编辑

SyncDiff在生成过程中提高多体协同能力　UniHSI结合大语言模型生成物理真实的长程交互动作　InterScene生成物理真实的人—场景交互动作

图 15　人体动作生成相关代表性工作

Fig. 15　Representative works on human motion generation

随着人类运动交互捕捉与生成技术的进步,人形机器人通过模仿人类获取多样化运动技能,成为全球多个研究团队在 2024 年共同关注的重点方向。HumanPlus(Fu 等,2024a)、OmniH2O(He 等,2024b)和 ExpressiveHumanoid(Cheng 等,2024)等研究通过建模人类行走、奔跑等运动数据,为人形机器人的规划与控制提供了关键的先验知识,从而推动该领域的研究范式从机器人自主探索转向模仿人类动作。为了研究人类交互动作对机器人模仿完成交互任务的影响,BiGym(Chernyadev 等,2024)和 Mimicking-Bench(Liu 等,2024b)在仿真环境中设计了多种交互任务,包括开柜门、收拾餐具、坐下椅子、搬运箱子等。另一方面,DexCap(Wang 等,2024a)、DexTrack(Liu 等,2025a)和 CyberDemo(Wang 等,2024b)等工作则聚焦于从人类手物交互运动中学习灵巧手操作技能,为通用灵巧手操作技能学习拓展了思路。如图 16 所示,总的来说,通过利用采集和生成的人类动作,这些研究显著提升了人形机器人在完成复杂

交互任务时的表现。

虽然2024年从人类运动数据中学习人形机器人操作交互技能方向涌现出了大量探索，但是人形机器人类人交互技能发展仍充满挑战。在交互数据捕捉方面，需要解决如何大量从互联网视频数据中提取高质量的交互运动数据这一关键问题。在交互运动生成方面，泛化性、物理真实性和交互复杂性仍是重要挑战。在人形机器人的运动智能方面，复杂接触场景中更为稳定精细的控制仍需要更多的研究关注。展望2025 年，期待人形机器人能够充分利用多样化的人类动作生成技术，模仿人类完成更具挑战性的任务，并进一步拓展其在真实场景中的应用能力。

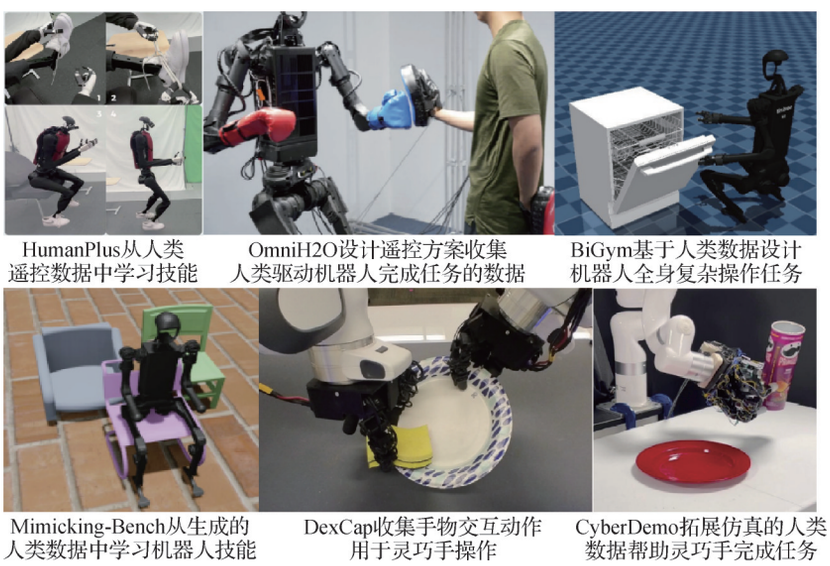


图 16　从人体运动数据学习人形机器人动作生成

Fig. 16　Learning humanoid robot motion generation from human motion data

1.9　具身 VLA 大模型吞吐虚实大数据利用三维模态增进泛化性

寻找机器人的通用操作策略是具身智能领域一直以来的关键问题。2023 年 7 月谷歌在 RT-2（Brohan 等，2023）工作中提出的 VLA（vision-language-action）模型给出一种端到端具身大模型的研究范式，输入连续视觉观测（V）和语言指令（L），模型直接输出机器人的末端执行器或全身关节的瞬时运动（A）。然而 RT-2 的全部动作数据只有 Everyday Robots 在有限的几个房间里采集的 13 万条数据，其动作训练数据不足使得 RT-2 在关于环境、物体等的泛化性及任务的多样性上仍有较大的限制。

2024 年，VLA 在世界范围内成为具身智能和大模型领域关注的焦点，成为机器人通用控制架构的有力角逐者，各团队提出不同手段应对数据和模型层面的挑战。

为了解决数据的不足，研究学者尝试利用各种各样的机器人采集数据。2024 年 6 月，谷歌团队发布开源且支持多种本体结构的大模型 OpenVLA（Kim 等，2024）。该大模型基于百万轨迹量级的跨本体真实机器人数据集 Open X-Embodiment 进行跨本体的预训练。跨本体训练去掉了机器人数据必须来源于同款机器人的限制，降低了数据采集门槛，此类数据被用于大规模预训练。OpenVLA 同时提供了在测试的机器人上进行多种后训练的方法，能够快速适应新的本体和任务。

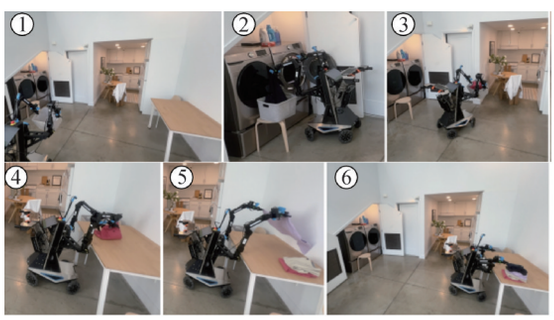
Physical Intelligence 团队进一步提出 π_0 模型（Black 等，2024），使用 Flow Matching 的方法提升了 VLA 模型的性能，展示出在真实世界处理复杂长程任务的能力。OpenVLA 和 π_0 模型的跨本体预训练虽然对模型带来了一定帮助，但不同本体的相机位置和动作空间都有所不同，因此预训练后的模型在测试机器人零样本（zero-shot）情况下直接使用并不能达到理想的工作水平，较依赖在测试机器人上采集数百到上千条数据进行后训练。

除了跨本体的思路以外，还可以采集单本体的大量数据，或使用单本体的大量合成数据进行训练。字节研究团队提出 GR-2（Cheang 等，2024），对特定场景中 55 个物体使用单一机械臂采集了 94 000 条抓放轨迹，实现了在此场景中对物体抓放的泛化性。北京大

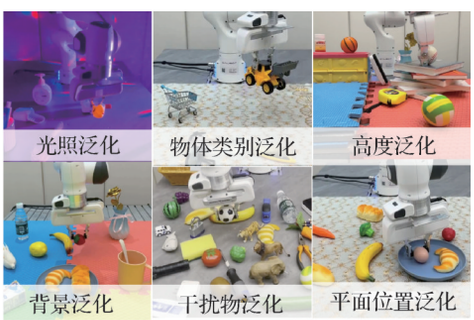
学和银河通用等团队提出GraspVLA，利用图形学手段合成了千万条、10亿帧场景随机、物体随机、物理真实以及高逼真渲染的Franka arm单一本体抓取动作数据。通过完全使用分式动作数据进行预训练，GraspVLA展示了很强的零样本(zero-shot)学习能力，对于闭环抓取中的物体种类、背景、前景、光照和干扰物都体现了很强的泛化性，并拥有更好的后训练效率。

2024年也有一些团队试图将3D视觉模态加入VLA，利用单视角或多视角RGB-D输入中的几何信息增强VLA。3D-VLA（Zhen等，2024）使用Diffusion model生成任务目标图像或点云，并得到对应的开始状态及结束状态的3D feature field，将这两者嵌入LLM中预测动作。3D Diffusion Actor（Ke等，2024）也利用了3D feature field，并成功通过denoised Transformer将其与Diffusion policy融合输出action，实现了比基于2D的VLA方法更好的视角泛化性。这些方法目前受制于RGB-D动作数据的体量，其对任务和环境的泛化性仍有待提升。

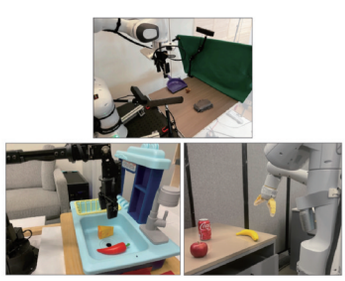
图17展示了具身VLA大模型相关代表性工作。展望2025年，期待三维视觉模态和多模态合成大数据大力推进VLA的发展，在通用性和泛化性取得长足的进展。



π_0处理真实世界中的复杂长程任务

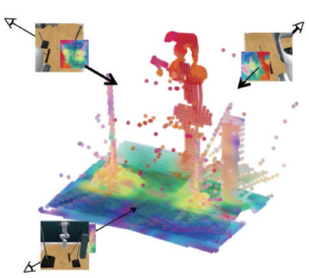


GraspVLA模型展现零样本泛化“金标准”



OpenVLA开源支持多种本体结构的大模型

GR-2针对单场景采集抓放轨迹



3D Diffusor Actor通过三维模态增进视角泛化性

图17　具身VLA大模型相关代表性工作

Fig. 17　Representative works on embodied VLA large models

1.10　三维计算成像与微观宏观领域科学研究突破

传统视觉传感器采集的图像/视频在动态范围、时间分辨率和波长谱段等方面只能记录完整光场中非常有限的一部分信息，对于三维重建的性能势必会带来一定的制约。2024年度涌现了一些利用非传统视觉传感器的研究工作，通过采用神经形态相机、单光子相机和热成像相机等传感器，在动态范围、时间和波长等维度上展示了相对传统成像的优异性能，突破了传统三维重建的局限。图18展示了三维计算成像领域应对有挑战的三维重建场景的代表性进展，其中，左上是利用事件相机，左下是单光子相机，右图是热成像相机。

北京大学团队的研究工作EventPS（Yu等，2024a）提出一种高速连续光源下的差分光度立体成像模型，将法线估计转化为零空间向量求解的线性问题，配合高速转台和事件相机，实现了30帧/s的实时表面法线重建。为了保证采集图像的动态范围足够，传统方法需要数秒融合多次不同曝光采集某一光照下的高动态范围图像，采集大量不同方向变化的光照图像则需要耗时数十秒（例如融合多次曝光获得一幅高动态范围图像需要5 s，10幅则耗时50 s），利用事件相机的方法数据采集（0.03 s）比传统相机提高上千倍；该方法为光度立体视觉方法提供了新范式，是一种高精度、高实时三维重建的全新解决方

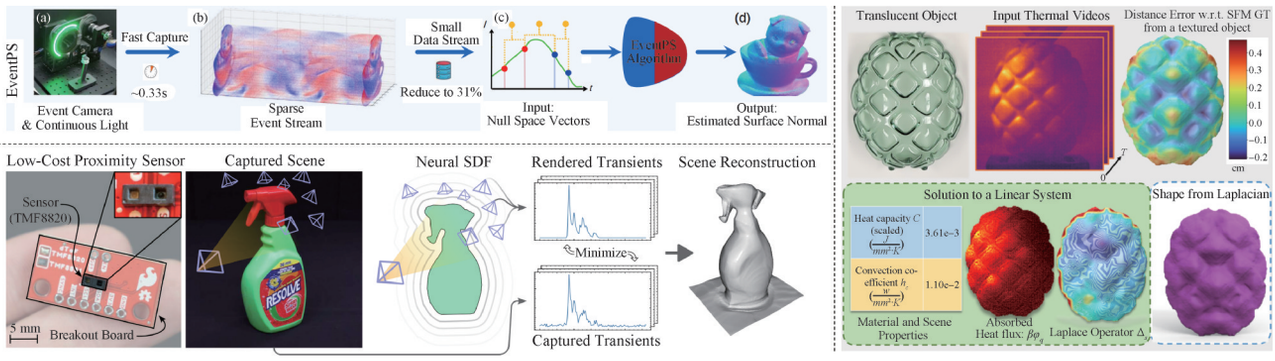


图 18　三维计算成像领域应对有挑战的三维重建场景的代表性进展

Fig. 18　Representative advances in 3D computational imaging: tackling challenging 3D reconstructed scenes

案。美国威斯康星大学麦迪逊分校的研究工作 Towards 3D Vision with Low-Cost Single-Photon Cameras(Mu 等，2024)提出一种使用单光子雪崩二极管(single photon avalanche diode，SPAD)传感器进行低成本、高精度三维重建的方法；该方法从多视角 SPAD 观测优化场景几何，SPAD 通过瞬态直方图记录光子的飞行时间，利用神经场可微渲染技术建模瞬态直方图成像过程，相较于传统的主动深度感知算法，实现了更高的重建精度，能够在低光和强光干扰环境中稳定工作，弥补了基于图像的三维重建方法的应用限制。美国卡内基梅隆大学团队的研究工作 Shape from Heat Conduction(Narayanan 等，2025)提出一种通过热成像相机捕捉物体在长波红外光谱下发出的热辐射，进而通过热传导计算实现三维重建的方法，该方法能够有效重建透明和半透明物体的形状，精度达到 1～2 mm，弥补了传统可见光方法在此类场景的局限。

上述几项研究展示了针对一些有特殊挑战的场景利用传感器的特性优势和计算摄像方法突破三维重建性能方面的进展。结合可微渲染、3D AIGC 等三维视觉算法方面的进展，如何深层次挖掘非传统传感器的性能优势值得进一步思考。然而，考虑到非传统传感器的普及程度，其在三维重建问题上的“杀手级应用”怎样构建和何时涌现还需要持续关注。

三维计算成像与三维表征的迅猛发展不仅在计算机领域取得了突破，还为多个科学领域带来了显著进步，成为解决一系列复杂科学问题的重要工具。例如，美国加州理工大学利用神经辐射场(NeRF)技术，成功重建了银河系中心黑洞的三维动态变化(Levis 等，2024)，极大推动了天体物理学的研究，为人类理解宇宙提供了前所未有的视角；美国伯克利加州大学利用将隐式时空间表征应用于活细胞超分辨三维动态成像(Cao 等，2024)，为揭示细胞生物学现象提供了强大的观测工具；上海科技大学利用 neural SDF(Park 等，2019)首次建立了能够从多视图超声扫描影像中精准恢复三维解剖结构的形状重建框架(Chen 等，2024a)，极大减少了超声视角依赖和探头层厚等限制带来的椎体结构畸变问题，有助于医生在术中更加简单、快速和精准地找到目标，降低医生和病人的辐射暴露。这些应用展示了三维表征与三维计算成像在科学探索中的巨大潜力，特别在涉及复杂生命动态现象、极端物理环境和天体结构等研究领域时，其独特的优势愈发凸显。

2　结语和展望

本文梳理了 2024 年三维视觉领域的发展趋势与十大前沿进展，但仍有一些重要方向未能详尽覆盖，例如点云的三维处理、三维感知与语义分割、三维动态物理仿真等。这些方向作为三维视觉的基础与关键环节，具有不可替代的价值。

2024 年，三维视觉的突破为技术发展注入新动力，未来趋势集中于多个关键方向。时空一致性世界模型将通过融合 4D 时空与物理规律，为自动驾驶和具身智能等复杂场景提供动态预测与交互支持；生成式 AI 与 3D 内容生成技术的深度结合将突破数据瓶颈，实现高保真、可控的 3D 内容自动生成；跨模态泛化能力的提升将强化视觉、语言和动作的多模态融合，提升机器人策略的泛化性与鲁棒性；物理仿真驱动的动态重建将结合物理引擎，实现动态场景的高精度建模与交互编辑，为数字孪生和虚拟现实

提供技术支持。同时,三维成像技术将在天体物理与细胞生物学等领域加速科学探索,高效实时化与轻量化技术将提升大规模动态场景的实时重建与渲染效率,推动端侧设备应用。此外,伦理与隐私保护将成为重要议题,通过加密技术与法规规范三维数据的采集与生成,平衡技术创新与隐私安全。三维视觉正迈向更智能、普适的"空间智能"时代,技术突破将重塑人机交互、科学探索与产业生态。

致谢:本文源于中国图象图形学学会三维视觉专委会2024年10月在北京举办的首届Mini3DV会议。会议受到字节跳动公司独家赞助,在此表示感谢。全体参会成员总结三维视觉领域过去一年的前沿研究,形成对三维视觉发展趋势和重要进展的关键共识。在此对未列入作者名单的参会人员表示感谢,包括:查红彬、连宙辉、仉尚航(北京大学),许威威(浙江大学),杨睿刚(上海交通大学),刘利刚(中国科学技术大学),吴毅红、张兆翔、申抒含(中国科学院自动化研究所),徐凯(国防科技大学),谭平(香港科技大学),王程(厦门大学),张力(复旦大学),许岚(上海科技大学),武玉伟(北京理工大学),朱昊(南京大学),庞江淼(上海人工智能实验室),李修、靳潇杰(字节跳动)。此外,本文得到学生的协助整理,在此表示感谢。包括:庞有鑫、邵睿智、徐乐朗、陈雨硕、刘昀(清华大学)、党灵伟(华南理工大学)、汤佳骏、于博涵、费凡、严汨(北京大学)、温怡健、李玉慧(北京师范大学)、吴桐(中国科学院计算技术研究所)。

参考文献

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman F L, Almeida D, Altenschmidt J, Altman S, Anadkat S, Avila R, Babuschkin I, Balaji S, Balcom V, Baltescu P, Bao H M, Bavarian M, Belgum J, Bello I, Berdine J, Bernadett-Shapiro G, Berner C, Bogdonoff L, Boiko O, Boyd M, Brakman A L, Brockman G, Brooks T, Brundage M, Button K, Cai T, Campbell R, Cann A, Carey B, Carlson C, Carmichael R, Chan B, Chang C, Chantzis F, Chen D, Chen S, Chen R, Chen J, Chen M, Chess B, Cho C, Chu C, Chung H W, Cummings D, Currier J, Dai Y X, Decareaux C, Degry T, Deutsch N, Deville D, Dhar A, Dohan D, Dowling S, Dunning S, Ecoffet A, Eleti A, Eloundou T, Farhi D, Fedus L, Felix N, Fishman S P, Forte J, Fulford I, Gao L, Georges E, Gibson C, Goel V, Gogineni T, Goh G, Gontijo-Lopes R, Gordon J, Grafstein M, Gray S, Greene R, Gross J, Gu S S, Guo Y F, Hallacy C, Han J, Harris J, He Y C, Heaton M, Heidecke J, Hesse C, Hickey A, Hickey W, Hoeschele P, Houghton B, Hsu K, Hu S L, Hu X, Huizinga J, Jain S, Jain S, Jang J, Jiang A, Jiang R, Jin H Z, Jin D, Jomoto S, Jonn B, Jun H, Kaftan T, Kaiser Ł, Kamali A, Kanitscheider I, Keskar N S, Khan T, Kilpatrick L, Kim J W, Kim C, Kim Y, Kirchner J H, Kiros J, Knight M, Kokotajlo D, Kondraciuk Ł, Kondrich A, Konstantinidis A, Kosic K, Krueger G, Kuo V, Lampe M, Lan I, Lee T, Leike J, Leung J, Levy D, Li C M, Lim R, Lin M, Lin S, Litwin M, Lopez T, Lowe R, Lue P, Makanju A, Malfacini K, Manning S, Markov T, Markovski Y, Martin B, Mayer K, Mayne A, McGrew B, McKinney S M, McLeavey C, McMillan P, McNeil J, Medina D, Mehta A, Menick J, Metz L, Mishchenko A, Mishkin P, Monaco V, Morikawa E, Mossing D, Mu T, Murati M, Murk O, Mély D, Nair A, Nakano R, Nayak R, Neelakantan A, Ngo R, Noh H, Ouyang L, O'Keefe C, Pachocki J, Paino A, Palermo J, Pantuliano A, Parascandolo G, Parish J, Parparita E, Passos A, Pavlov M, Peng A, Perelman A, de Avila Belbute Peres F, Petrov M, de Oliveira Pinto H P, (Rai) Pokorny M, Pokrass M, Pong V H, Powell T, Power A, Power B, Proehl E, Puri R, Radford A, Rae J, Ramesh A, Raymond C, Real F, Rimbach K, Ross C, Rotsted B, Roussez H, Ryder N, Saltarelli M, Sanders T, Santurkar S, Sastry G, Schmidt H, Schnurr D, Schulman J, Selsam D, Sheppard K, Sherbakov T, Shieh J, Shoker S, Shyam P, Sidor S, Sigler E, Simens M, Sitkin J, Slama K, Sohl I, Sokolowsky B, Song Y, Staudacher N, Such F P, Summers N, Sutskever I, Tang J, Tezak N, Thompson M B, Tillet P, Tootoonchian A, Tseng E, Tuggle P, Turley N, Tworek J, Uribe J F C, Vallone A, Vijayvergiya A, Voss C, Wainwright C, Wang J J, Wang A, Wang B, Ward J, Wei J, Weinmann C J, Welihinda A, Welinder P, Weng J Y, Weng L L, Wiethoff M, Willner D, Winter C, Wolrich S, Wong H, Workman L, Wu S, Wu J, Wu M, Xiao K, Xu T, Yoo S, Yu K, Yuan Q M, Zaremba W, Zellers R, Zhang C, Zhang M, Zhao S J, Zheng T H, Zhuang J T, Zhuk W and Zoph B. 2024. GPT-4 technical report [EB/OL]. [2025-01-02]. https://arxiv.org/pdf/2303.08774.pdf

Agarwal N, Ali A, Bala M, Balaji Y, Barker E, Cai T, Chattopadhyay P, Chen Y C, Cui Y, Ding Y F, Dworakowski D, Fan J J, Fenzi M, Ferroni F, Fidler S, Fox D, Ge S W, Ge Y H, Gu J W, Gururani S, He E, Huang J H, Huffman J, Jannaty P, Jin J Y, Kim S W, Klár G, Lam G, Lan S Y, Leal-Taixe L, Li A Q, Li Z S, Lin C H, Lin T Y, Ling H, Liu M Y, Liu X, Luo A, Ma Q L, Mao H Z, Mo K C, Mousavian A, Nah S, Niverty S, Page D, Paschalidou D, Patel Z, Pavao L, Ramezanali M, Reda F, Ren X W, Sabavat V R N, Schmerling E, Shi S, Stefaniak B, Tang S T, Tchapmi L, Tredak P, Tseng W C, Varghese J, Wang H, Wang H X, Wang H, Wang T C, Wei F Y, Wei X Y, Wu J Z, Xu J S, Yang W, Yen-Chen L, Zeng X H, Zeng Y, Zhang J, Zhang Q S, Zhang Y X, Zhao Q Q and Zolkowski A. 2025. Cosmos world foundation model platform for physical AI [EB/OL]. [2025-01-02].

https://arxiv.org/pdf/2501.03575.pdf

Bai J H, Xia M H, Wang X T, Yuan Z Y, Fu X, Liu Z Z, Hu H J, Wan P F and Zhang D. 2024. SynCamMaster: synchronizing multi-camera video generation from diverse viewpoints [EB/OL]. [2025-01-02]. https://arxiv.org/pdf/2412.07760.pdf

Bharadhwaj H, Mottaghi R, Gupta A and Tulsiani S. 2025. Track2Act: predicting point tracks from internet videos enables generalizable robot manipulation//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 306-324 [DOI: 10.1007/978-3-031-73116-7_18]

Black K, Brown N, Driess D, Esmail A, Equi M, Finn C, Fusai N, Groom L, Hausman K, Ichter B, Jakubczak S, Jones T, Ke L, Levine S, Li-Bell A, Mothukuri M, Nair S, Pertsch K, Shi L X, Tanner J, Vuong Q, Walling A, Wang H H and Zhilinsky U. 2024. π0: a vision-language-action flow model for general robot control [Z/OL]. https://www.physicalintelligence.company/blog/pi0

Blanc H, Deschaud J E and Paljic A. 2025. RayGauss: volumetric Gaussian-based ray casting for photorealistic novel view synthesis [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2408.03356.pdf

Brohan A, Brown N, Carbajal J, Chebotar Y, Chen X, Choromanski K, Ding T L, Driess D, Dubey A, Finn C, Florence P, Fu C Y, Arenas M G, Gopalakrishnan K, Han K H, Hausman K, Herzog A, Hsu J, Ichter B, Irpan A, Joshi N, Julian R, Kalashnikov D, Kuang Y H, Leal I, Lee L, Lee T W E, Levine S, Lu Y, Michalewski H, Mordatch I, Pertsch K, Rao K, Reymann K, Ryoo M, Salazar G, Sanketi P, Sermanet P, Singh J, Singh A, Soricut R, Tran H, Vanhoucke V, Vuong Q, Wahid A, Welker S, Wohlhart P, Wu J L, Xia F, Xiao T, Xu P, Xu S C, Yu T H and Zitkovich B. 2023. RT-2: vision-language-action models transfer web knowledge to robotic control [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2307.15818.pdf

Cao R M, Divekar N S, Nuñez J K, Upadhyayula S and Waller L. 2024. Neural space-time model for dynamic multi-shot imaging. Nature Methods, 21(12): 2336-2341 [DOI: 10.1038/s41592-024-02417-0]

Cha H, Lee I and Joo H. 2024a. PERSE: personalized 3D generative avatars from a single portrait [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2412.21206.pdf

Cha J, Kim J, Yoon J S and Baek S. 2024b. Text2HOI: text-guided 3D motion generation for hand-object interaction//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1577-1585 [DOI: 10.1109/CVPR52733.2024.00156]

Cheang C L, Chen G Z, Jing Y, Kong T, Li H, Li Y F, Liu Y X, Wu H T, Xu J F, Yang Y H, Zhang H B and Zhu M Z. 2024. GR-2: a generative video-language-action model with web-scale knowledge for robot manipulation [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2410.06158.pdf

Chen G K and Wang W G. 2025. A survey on 3D Gaussian splatting [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2401.03890.pdf

Chen H B, Gao Y C, Zhang S H, Wu J J, Ma Y X and Zheng R. 2024a. RoCoSDF: row-column scanned neural signed distance fields for freehand 3D ultrasound imaging shape reconstruction//Proceedings of the 27th International Conference on Medical Image Computing and Computer-Assisted Intervention. Marrakesh, Morocco: Springer: 721-731 [DOI: 10.1007/978-3-031-72083-3_67]

Chen Y H, Wu Q Y, Li M Y, Lin W Y, Harandi M and Cai J F. 2025. Fast feedforward 3D Gaussian splatting compression [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2410.08017.pdf

Chen Y T, Wang Y Q and Zhang Z X. 2024b. DrivingGPT: unifying driving world modeling and planning with multi-modal autoregressive transformers [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2412.18607.pdf

Chen Z W, Tan H, Zhang K, Bi S, Luan F J, Hong Y C, Li F X and Xu Z X. 2024c. Long-LRM: long-sequence large reconstruction model for wide-coverage Gaussian splats [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2410.12781.pdf

Chen Z Y, Yang J W, Huang J H, de Lutio R, Esturo J M, Ivanovic B, Litany O, Gojcic Z, Fidler S, Pavone M, Song L and Wang Y. 2024d. OmniRe: Omni urban scene reconstruction [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2408.16760.pdf

Cheng X X, Ji Y D, Chen J M, Yang R H, Yang G and Wang X L. 2024. Expressive whole-body control for humanoid robots. [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2402.16796.pdf

Chernyadev N, Backshall N, Ma X, Lu Y F, Seo Y and James S. 2024. BiGym: a demo-driven mobile bi-manual manipulation benchmark [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2407.07788.pdf

Chi C, Xu Z J, Pan C E, Cousineau E, Burchfiel B, Feng S Y, Tedrake R and Song S R. 2024. Universal manipulation interface: in-the-wild robot teaching without in-the-wild robots [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2402.10329.pdf

Christen S, Hampali S, Sener F, Remelli E, Hodan T, Sauser E, Ma S G and Tekin B. 2024. DiffH2O: diffusion-based synthesis of hand-object interactions from textual descriptions//Proceedings of 2024 SIGGRAPH Asia Conference Papers. Tokyo, Japan: ACM: #145 [DOI: 10.1145/3680528.3687563]

Contributors A W C. 2024. AgiBot-World Colosseum [EB/OL]. [2025-02-14]. https://github.com/OpenDriveLab/AgiBot-World

Cui J D, Cao J M, Zhao F Q, He Z P, Chen Y F, Zhong Y H, Xu L, Shi Y J, Zhang Y L and Yu J Y. 2024. LetsGo: large-scale garage modeling and rendering via LiDAR-assisted Gaussian primitives. ACM Transactions on Graphics, 43(6): #172 [DOI: 10.1145/3687762]

Dai P X, Xu J M, Xie W X, Liu X G, Wang H M and Xu W W. 2024. High-quality surface reconstruction using Gaussian surfels//Proceedings of 2024 ACM SIGGRAPH Conference Papers. Denver, USA: ACM: #22 [DOI: 10.1145/3641519.3657441]

Fan Z W, Wen K R, Cong W Y, Wang K, Zhang J, Ding X H, Xu D F, Ivanovic B, Pavone M, Pavlakos G, Wang Z Y and Wang Y.

2025. InstantSplat: sparse-view Gaussian splatting in seconds [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2403.20309.pdf

Fei B, Xu J Y, Zhang R, Zhou Q Y, Yang W D and He Y. 2024. 3D Gaussian splatting as new era: a survey. IEEE Transactions on Visualization and Computer Graphics: #3397828 [DOI: 10.1109/TVCG.2024.3397828]

Fu Z P, Zhao Q Q, Wu Q, Wetzstein G and Finn C. 2024a. HumanPlus: humanoid shadowing and imitation from humans [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2406.10454.pdf

Fu Z P, Zhao T Z and Finn C. 2024b. Mobile ALOHA: learning bimanual mobile manipulation with low-cost whole-body teleoperation [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2401.02117.pdf

Gao D Y, Siddiqui Y, Li L and Dai A. 2024a. MeshArt: generating articulated meshes with structure-guided transformers [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2412.11596.pdf

Gao J, Gu C, Lin Y T, Li Z H, Zhu H, Cao X, Zhang L and Yao Y. 2025. Relightable 3D Gaussians: realistic point cloud relighting with BRDF decomposition and ray tracing//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 73-89 [DOI: 10.1007/978-3-031-72995-9_5]

Gao L, Yang J, Zhang B T, Sun J M, Yuan Y J, Fu H B and Lai Y K. 2024b. Real-time large-scale deformation of Gaussian Splatting. ACM Transactions on Graphics, 43(6): #200 [DOI: 10.1145/3687756]

Gao R Q, Holynski A, Henzler P, Brussee A, Martin-Brualla R, Srinivasan P, Barron J T and Poole B. 2024c. CAT3D: create anything in 3D with multi-view diffusion models [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2405.10314.pdf

Genie2. 2025. Genie2 —— AI-powered solutions [Z/OL]. [2025-02-14]. https://genie2.co

Giebenhain S, Kirschstein T, Rünz M, Agapito L and Nießner M. 2024. NPGA: neural parametric Gaussian avatars//Proceedings of 2024 SIGGRAPH Asia Conference Papers. Tokyo, Japan: ACM: #127 [DOI: 10.1145/3680528.3687689]

Guédon A and Lepetit V. 2024. SuGaR: surface-aligned Gaussian splatting for efficient 3D mesh reconstruction and high-quality mesh rendering//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5354-5363 [DOI: 10.1109/CVPR52733.2024.00512]

He H R, Bai C J, Pan L, Zhang W N, Zhao B and Li X L. 2024a. Learning an actionable discrete diffusion policy via large-scale actionless video pre-training [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2402.14407.pdf

He T R, Luo Z Y, He X L, Xiao W L, Zhang C, Zhang W W, Kitani K, Liu C L and Shi G Y. 2024b. OmniH2O: universal and dexterous human-to-humanoid whole-body teleoperation and learning. arXiv preprint [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2406.08858.pdf

He W K, Liu Y, Liu R T and Yi L. 2025. SyncDiff: synchronized motion diffusion for multi-body human-object interaction synthesis. arXiv preprint [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2412.20104.pdf

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 6840-6851

Hong Y C, Zhang K, Gu J X, Bi S, Zhou Y, Liu D F, Liu F, Sunkavalli K, Bui T and Tan H. 2024. LRM: large reconstruction model for single image to 3D [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2311.04400.pdf

Hu L, Gao X, Zhang P, Sun K, Zhang B and Bo L F. 2024a. Animate anyone: consistent and controllable image-to-video synthesis for character animation [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2311.17117.pdf

Hu L X, Zhang H W, Zhang Y X, Zhou B Y, Liu B N, Zhang S P and Nie L Q. 2024b. GaussianAvatar: towards realistic human avatar modeling from a single video via Animatable 3D Gaussians//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 634-644 [DOI: 10.1109/CVPR52733.2024.00067]

Huang B B, Yu Z H, Chen A P, Geiger A and Gao S H. 2024. 2D Gaussian splatting for geometrically accurate radiance fields//Proceedings of 2024 ACM SIGGRAPH Conference Papers. Denver, USA: ACM: #32 [DOI: 10.1145/3641519.3657428]

Huo D, Guo Z X, Zuo X X, Shi Z H, Lu J W, Dai P, Xu S C, Cheng L and Yang Y H. 2024. TexGen: text-guided 3D texture generation with multi-view sampling and resampling [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2408.01291.pdf

Jain V, Attarian M, Joshi N J, Wahid A, Driess D, Vuong Q, Sanketi P R, Sermanet P, Welker S, Chan C, Gilitschenski I, Bisk Y and Dwibedi D. 2024. Vid2Robot: end-to-end video-conditioned policy learning with cross-attention transformers [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2403.12943.pdf

Karunratanakul K, Preechakul K, Aksan E, Beeler T, Suwajanakorn S and Tang S Y. 2024. Optimizing diffusion noise can serve as universal motion priors//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1334-1345 [DOI: 10.1109/CVPR52733.2024.00133]

Ke T W, Gkanatsios N and Fragkiadaki K. 2024. 3D diffuser actor: policy diffusion with 3D scene representations [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2402.10885.pdf

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]

Kerbl B, Meuleman A, Kopanas G, Wimmer M, Lanvin A and Drettakis G. 2024. A hierarchical 3D Gaussian representation for real-time rendering of very large datasets. ACM Transactions on Graphics, 43(4): #62 [DOI: 10.1145/3658160]

Kim M J, Pertsch K, Karamcheti S, Xiao T, Balakrishna A, Nair S, Rafailov R, Foster E, Lam G, Sanketi P, Vuong Q, Kollar T, Burchfiel B, Tedrake R, Sadigh D, Levine S, Liang P and Finn C. 2024. OpenVLA: an open-source vision-language-action model [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2406.09246.pdf

Kwon Y, Fang B L, Lu Y X, Dong H Y, Zhang C, Carrasco F V, Mosella-Montoro A, Xu J J, Takagi S, Kim D, Prakash A and De La Torre F. 2025. Generalizable human Gaussians for sparse view synthesis//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 451-468 [DOI: 10.1007/978-3-031-73229-4_26]

Levis A, Chael A A, Bouman K L, Wielgus M and Srinivasan P P. 2024. Orbital polarimetric tomography of a flare near the Sagittarius A* supermassive black hole. Nature Astronomy, 8(6): 765-773 [DOI: 10.1038/s41550-024-02238-3]

Li J M, Clegg A, Mottaghi R, Wu J J, Puig X and Liu C K. 2025. Controllable human-object interaction synthesis//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 54-72 [DOI: 10.1007/978-3-031-72940-9_4]

Li J X, Cao C, Schwartz G, Khirodkar R, Richardt C, Simon T, Sheikh Y and Saito S. 2024a. URAvatar: universal relightable Gaussian codec avatars//Proceedings of 2024 SIGGRAPH Asia Conference Papers. Tokyo, Japan: ACM: 128 [DOI: 10.1145/3680528.3687653]

Li T Y, Bolkart T, Black M J, Li H and Romero J. 2017. Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics, 36(6): #194 [DOI: 10.1145/3130800.3130813]

Li Z, Zheng Z R, Wang L Z and Liu Y B. 2024b. Animatable Gaussians: learning pose-dependent Gaussian maps for high-fidelity human avatar modeling//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 19711-19722 [DOI: 10.1109/CVPR52733.2024.01864]

Liang H, Zhang W Q, Li W X, Yu J Y and Xu L. 2024a. InterGen: diffusion-based multi-human motion generation under complex interactions. International Journal of Computer Vision, 132(9): 3463-3483 [DOI: 10.1007/s11263-024-02042-6]

Liang J B, Liu R S, Ozguroglu E, Sudhakar S, Dave A, Tokmakov P, Song S R and Vondrick C. 2024b. Dreamitate: real-world visuomotor policy learning via video generation [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2406.16862.pdf

Lin J, Zeng A L, Lu S L, Cai Y H, Zhang R M, Wang H Q and Zhang L. 2023. Motion-X: a large-scale 3D expressive whole-body human motion dataset//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 25268-25280

Lin J Q, Li Z H, Tang X, Liu J Z, Liu S Y, Liu J Y, Lu Y D, Wu X F, Xu S C, Yan Y L and Yang W M. 2024. VastGaussian: vast 3D Gaussians for large scene reconstruction//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5166-5175 [DOI: 10.1109/CVPR52733.2024.00494]

Liu X, Zhan X H, Tang J X, Shan Y, Zeng G, Lin D H, Liu X H and Liu Z W. 2024a. HumanGaussian: text-driven 3D human generation with Gaussian splatting//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6646-6657 [DOI: 10.1109/CVPR52733.2024.00635]

Liu X Y, Adalibieke J, Han Q W, Qin Y Z and Yi L. 2025a. DexTrack: towards generalizable neural tracking control for dexterous manipulation from human references [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2502.09614.pdf

Liu Y, Luo C C, Fan L, Wang N Y, Peng J R and Zhang Z X. 2025b. CityGaussian: real-time high-quality large-scale scene rendering with Gaussians//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 265-282 [DOI: 10.1007/978-3-031-72640-8_15]

Liu Y, Yang B W, Zhong L C, Wang H and Yi L. 2024b. Mimicking-bench: a benchmark for generalizable humanoid-scene interaction learning via human mimicking [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2412.17730.pdf

Liu Y Z, Dong S Y, Wang S Z, Yin Y D, Yang Y C, Fan Q N and Chen B Q. 2025c. SLAM3R: real-time dense scene reconstruction from monocular RGB videos [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2412.09401.pdf

Lyu W, Zhou Y, Yang M H and Shu Z X. 2024. FaceLift: single image to 3D head with view generation and GS-LRM [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2412.17812.pdf

Mai A, Hedman P, Kopanas G, Verbin D, Futschik D, Xu Q G, Kuester F, Barron J T and Zhang Y D. 2024. EVER: exact volumetric ellipsoid rendering for real-time view synthesis [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2410.01804.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]

Moenne-Loccoz N, Mirzaei A, Perel O, de Lutio R, Esturo J M, State G, Fidler S, Sharp N and Gojcic Z. 2024. 3D Gaussian ray tracing: fast tracing of particle scenes. ACM Transactions on Graphics, 43(6): #232 [DOI: 10.1145/3687934]

Moon G, Shiratori T and Saito S. 2025. Expressive whole-body 3D Gaussian avatar//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 19-35 [DOI: 10.1007/978-3-031-72940-9_2]

Mu F Z, Sifferman C, Jungerman S, Li Y Q, Han M, Gleicher M, Gupta M and Li Y. 2024. Towards 3D vision with low-cost single-photon cameras//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE:

5302-5311 [DOI: 10.1109/CVPR52733.2024.00507]

Narayanan S, Ramanagopal M, Sheinin M, Sankaranarayanan A C and Narasimhan S G. 2025. Shape from heat conduction//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 426-444 [DOI: 10.1007/978-3-031-72920-1_24]

OpenAI. 2022. Introducing ChatGPT [Z/OL]. [2025-02-14] https://openai.com/blog/chatgpt

Open X-Embodiment Collaboration. 2024. Open X-embodiment: robotic learning datasets and RT-X models [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2310.08864.pdf

Pan L, Wang J B, Huang B Z, Zhang J Y, Wang H F, Tang X and Wang Y. 2025. Synthesizing physically plausible human motions in 3D scenes [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2308.09036.pdf

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]

Peebles W and Xie S N. 2023. Scalable diffusion models with transformers//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 4172-4182 [DOI: 10.1109/ICCV51070.2023.00387]

Proença P F and Simões P. 2020. TACO: trash annotations in context for litter detection [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2003.06975.pdf

Qian S H, Kirschstein T, Schoneveld L, Davoli D, Giebenhain S and Nießner M. 2024. GaussianAvatars: photorealistic head avatars with rigged 3D Gaussians//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 20299-20309 [DOI: 10.1109/CVPR52733.2024.01919]

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 [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2103.00020.pdf

Ren K R, Jiang L H, Lu T, Yu M L, Xu L N, Ni Z K and Dai B. 2024. Octree-GS: towards consistent real-time rendering with LOD-structured 3D Gaussians [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2403.17898.pdf

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]

Shao R Z, Pang Y X, Zheng Z R, Sun J X and Liu Y B. 2024. Human4DiT: 360-degree human video generation with 4D diffusion transformer [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2405.17405.pdf

Smart B, Zheng C X, Laina I and Prisacariu V A. 2024. Splatt3R: zero-shot Gaussian splatting from uncalibrated image pairs [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2408.13912.pdf

Sun J K, Jiao H, Li G Y, Zhang Z J, Zhao L and Xing W. 2024. 3DGStream: on-the-fly training of 3D Gaussians for efficient streaming of photo-realistic free-viewpoint videos//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 20675-20685 [DOI: 10.1109/CVPR52733.2024.01954]

Tao S, Xiang F B, Shukla A, Qin Y Z, Hinrichsen X, Yuan X D, Bao C, Lin X S, Liu Y L, Chan T K, Gao Y, Li X L, Mu T Z, Xiao N, Gurha A, Huang Z, Calandra R, Chen R, Luo S and Su H. 2024. ManiSkill3: GPU parallelized robotics simulation and rendering for generalizable embodied AI [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2410.00425.pdf

Taubner F, Zhang R H, Tuli M and Lindell D B. 2024. CAP4D: creating animatable 4D portrait avatars with morphable multi-view diffusion models [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2412.12093.pdf

Tian L R, Wang Q, Zhang B and Bo L F. 2025. EMO: emote portrait alive generating expressive portrait videos with Audio2Video diffusion model under weak conditions//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 244-260 [DOI: 10.1007/978-3-031-73010-8_15]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł 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

Wang C, Shi H C, Wang W Z, Zhang R H, Fei-Fei L and Liu C K. 2024a. DexCap: scalable and portable Mocap data collection system for dexterous manipulation [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2403.07788.pdf

Wang J, Qin Y Z, Kuang K M, Korkmaz Y, Gurumoorthy A, Su H and Wang X L. 2024b. CyberDemo: augmenting simulated human demonstration for real-world dexterous manipulation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 17952-17963 [DOI: 10.1109/CVPR52733.2024.01700]

Wang P H, Zhang Z R, Wang L, Yao K X, Xie S Y, Yu J Y, Wu M Y and Xu L. 2024c. V3: viewing volumetric videos on mobiles via streamable 2D dynamic Gaussians. ACM Transactions on Graphics, 43(6): #187 [DOI: 10.1145/3687935]

Wang Q Q, Ye V, Gao H, Austin J, Li Z Q and Kanazawa A. 2024d. Shape of motion: 4D reconstruction from a single video [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2407.13764.pdf

Wang S Z, Leroy V, Cabon Y, Chidlovskii B and Revaud J. 2024e. DUSt3R: geometric 3D vision made easy//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recogni-

tion. Seattle, USA: IEEE: 20697-20709 [DOI: 10.1109/CVPR52733.2024.01956]

Wang Y Z, Lira W, Wang W Q, Mahdavi-Amiri A and Zhang H. 2023. Slice3D: multi-slice, occlusion-revealing, single view 3D reconstruction [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2312.02221.pdf

Wen C, Lin X Y, So J, Chen K, Dou Q, Gao Y and Abbeel P. 2024. Any-point trajectory modeling for policy learning [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2401.00025.pdf

WorldLabs. 2025. WorldLabs-Innovation platform [EB/OL]. [2025-02-14]. https://www.worldlabs.ai/

Wu G J, Yi T R, Fang J M, Xie L X, Zhang X P, Wei W, Liu W Y, Tian Q and Wang X G. 2024a. 4D Gaussian splatting for real-time dynamic scene rendering//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 20310-20320 [DOI: 10.1109/CVPR52733.2024.01920]

Wu R D, Gao R Q, Poole B, Trevithick A, Zheng C X, Barron J T and Holynski A. 2024b. CAT4D: create anything in 4D with multi-view video diffusion models [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2411.18613.pdf

Wu R D, Gao R Q, Poole B, Trevithick A, Zheng C X, Barron J T and Holynski A. 2024b. CAT4D: create anything in 4D with multi-view video diffusion models [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2411.18613.pdf

Wu T, Sun J M, Lai Y K, Ma Y W, Kobbelt L and Gao L. 2024c. DeferredGS: decoupled and editable Gaussian splatting with deferred shading [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2404.09412.pdf

Wu T, Yuan Y J, Zhang L X, Yang J, Cao Y P, Yan L Q and Gao L. 2024d. Recent advances in 3D Gaussian splatting. Computational Visual Media, 10(4): 613-642 [DOI: 10.1007/s41095-024-0436-y]

Xiang J F, Lv Z L, Xu S C, Deng Y, Wang R C, Zhang B W, Chen D, Tong X and Yang J L. 2024. Structured 3D latents for scalable and versatile 3D generation [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2412.01506.pdf

Xiao Z Q, Wang T, Wang J B, Cao J K, Zhang W W, Dai B, Lin D H and Pang J M. 2024. Unified human-scene interaction via prompted chain-of-contacts [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2309.07918.pdf

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//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4389-4398 [DOI: 10.1109/CVPR52733.2024.00420]

Xu M W, Li H, Su Q K, Shang H L, Zhang L W, Liu C, Wang J D, Yao Y and Zhu S Y. 2024a. Hallo: hierarchical audio-driven visual synthesis for portrait image animation [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2406.08801.pdf

Xu S C, Chen G J, Guo Y X, Yang J L, Li C, Zang Z Y, Zhang Y Z, Tong X and Guo B N. 2024b. VASA-1: lifelike audio-driven talking faces generated in real time [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2404.10667.pdf

Xu Y L, Chen B W, Li Z, Zhang H W, Wang L Z, Zheng Z R and Liu Y B. 2024c. Gaussian head avatar: ultra high-fidelity head avatar via dynamic Gaussians//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1931-1941 [DOI: 10.1109/CVPR52733.2024.00189]

Xu Y L, Su Z Q, Wu Q Y and Liu Y B. 2024d. GPHM: Gaussian parametric head model for monocular head avatar reconstruction [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2407.15070.pdf

Xu Z, Xu Y H, Yu Z Y, Peng S D, Sun J M, Bao H J and Zhou X W. 2024e. Representing long volumetric video with temporal Gaussian hierarchy. ACM Transactions on Graphics, 43(6): #171 [DOI: 10.1145/3687919]

Xu Z C, Zhang J F, Liew J H, Yan H S, Liu J W, Zhang C X, Feng J S and Shou M Z. 2024f. MagicAnimate: temporally consistent human image animation using diffusion model//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1481-1490 [DOI: 10.1109/CVPR52733.2024.00147]

Xue Y X, Xie X H, Marin R and Pons-Moll G. 2024. Human-3Diffusion: realistic avatar creation via explicit 3D consistent diffusion models [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2406.08475.pdf

Yan Y Z, Lin H T, Zhou C X, Wang W J, Sun H Y, Zhan K, Lang X P, Zhou X W and Peng S D. 2025. Street Gaussians: modeling dynamic urban scenes with Gaussian splatting//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 156-173 [DOI: 10.1007/978-3-031-73464-9_10]

Yang T Y, Xu P F, Hu R B, Chai H and Chan A B. 2020. ROAM: recurrently optimizing tracking model//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6717-6726 [DOI: 10.1109/CVPR42600.2020.00675]

Yang Z Y, Gao X Y, Zhou W, Jiao S H, Zhang Y Q and Jin X G. 2024. Deformable 3D Gaussians for high-fidelity monocular dynamic scene reconstruction//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 20331-20341 [DOI: 10.1109/CVPR52733.2024.01922]

Ye B T, Liu S F, Xu H F, Li X T, Pollefeys M, Yang M H and Peng S Y. 2024a. No pose, no problem: surprisingly simple 3D Gaussian splats from sparse unposed images [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2410.24207.pdf

Ye K Y, Hou Q M and Zhou K. 2024b. 3D Gaussian splatting with deferred reflection//Proceedings of 2024 ACM SIGGRAPH Conference Papers. Denver, USA: ACM: 40 [DOI: 10.1145/3641519.3657456]

Yu B H, Ren J J, Han J, Wang F S, Liang J X and Shi B X. 2024a. EventPS: real-time photometric stereo using an event camera//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9602-9611 [DOI: 10.1109/CVPR52733.2024.00917]

Yu W B, Xing J B, Yuan L, Hu W B, Li X Y, Huang Z P, Gao X J, Wong T T, Shan Y and Tian Y H. 2024b. ViewCrafter: taming video diffusion models for high-fidelity novel view synthesis [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2409.02048.pdf

Zhai S J, Ye Z C, Liu J L, Xie W J, Hu J Q, Peng Z, Xue H, Chen D P, Wang X M, Yang L, Wang N, Liu H M and Zhang G F. 2025. StarGen: a spatiotemporal autoregression framework with video diffusion model for scalable and controllable scene generation [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2501.05763.pdf

Zhang B, Tang J P, Nießner M and Wonka P. 2023. 3DShape2VecSet: a 3D shape representation for neural fields and generative diffusion models. ACM Transactions on Graphics, 42(4): #92 [DOI: 10.1145/3592442]

Zhang J L, Liu H R, Li D S, Yu X Q, Geng H R, Ding Y F, Chen J Y and Wang H. 2024a. DexGraspNet 2.0: learning generative dexterous grasping in large-scale synthetic cluttered scenes [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2410.23004.pdf

Zhang J Y, Herrmann C, Hur J, Jampani V, Darrell T, Cole F, Sun D Q and Yang M H. 2024b. MonST3R: a simple approach for estimating geometry in the presence of motion [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2410.03825.pdf

Zhang K, Bi S, Tan H, Xiangli Y B, Zhao N X, Sunkavalli K and Xu Z X. 2024c. GS-LRM: large reconstruction model for 3D Gaussian splatting [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2404.19702.pdf

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. 2024d. 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 T Y, Yu H X, Wu R D, Feng B Y, Zheng C X, Snavely N, Wu J J and Freeman W T. 2025. PhysDreamer: physics-based interaction with 3D objects via video generation//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 388-406 [DOI: 10.1007/978-3-031-72627-9_22]

Zhang Y, Liu T, Meyer C A, Eeckhoute J, Johnson D S, Bernstein B E, Nusbaum C, Myers R M, Brown M, Li W and Liu X S. 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biology, 9(9): #R137 [DOI: 10.1186/gb-2008-9-9-r137]

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]. [2025-02-14]. https://arxiv.org/pdf/2403.09631.pdf

Zheng S Y, Zhou B Y, Shao R Z, Liu B N, Zhang S P, Nie L Q and Liu Y B. 2024. GPS-Gaussian: generalizable pixel-wise 3D Gaussian splatting for real-time human novel view synthesis//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 19680-19690 [DOI: 10.1109/CVPR52733.2024.01861]

Zhong L C, Yu H X, Wu J J and Li Y Z. 2025. Reconstruction and simulation of elastic objects with spring-mass 3D Gaussians//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 407-423 [DOI: 10.1007/978-3-031-72627-9_23]

Zhou H Y, Lin L Z, Wang J B, Lu Y C, Bai D F, Liu B B, Wang Y, Geiger A and Liao Y Y. 2024a. HUGSIM: a real-time, photo-realistic and closed-loop simulator for autonomous driving [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2412.01718.pdf

Zhou H Y, Shao J H, Xu L, Bai D F, Qiu W C, Liu B B, Wang Y, Geiger A and Liao Y Y. 2024b. HUGS: holistic urban 3D scene understanding via Gaussian splatting//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 21336-21345 [DOI: 10.1109/CVPR52733.2024.02016]

Zhou X Y, Lin Z W, Shan X J, Wang Y T, Sun D Q and Yang M H. 2024c. DrivingGaussian: composite Gaussian splatting for surrounding dynamic autonomous driving scenes//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 21634-21643 [DOI: 10.1109/CVPR52733.2024.02044]

作者简介

刘烨斌，男，教授，主要研究方向为三维视觉和计算成像。
E-mail: liuyebin@tsinghua.edu.cn
陈宝权，通信作者，男，教授，主要研究方向为计算机图形学和三维视觉。E-mail: baoquan@pku.edu.cn
苏昊，男，副教授，主要研究方向为计算机图形学和三维视觉。E-mail: haosu@ucsd.edu
高林，男，研究员，主要研究方向为计算机图形学和三维视觉。E-mail: gaolin@ict.ac.cn
弋力，男，助理教授，主要研究方向为具身智能和三维视觉。
E-mail: ericyi0124@gmail.com
王鹤，男，助理教授，主要研究方向为具身智能和三维视觉。
E-mail: hewang@pku.edu.cn
廖依伊，女，研究员，主要研究方向为计算机图形学和三维视觉。E-mail: yiyi.liao@zju.edu.cn
施柏鑫，男，研究员，主要研究方向为三维视觉和计算成像。
E-mail: shiboxin@pku.edu.cn
曹炎培，男，研究员，主要研究方向为计算机图形学和三维视觉。E-mail: caoyanpei@gmail.com
洪方舟，男，博士研究生，主要研究方向为具身智能和三维视觉。E-mail: fangzhouhong820@gmail.com
董豪，男，助理教授，主要研究方向为具身智能和三维视觉。

E-mail：hao.dong@pku.edu.cn
张举勇，男，教授，主要研究方向为三维视觉和计算成像。
E-mail：juyong@ustc.edu.cn
王鑫涛，男，研究员，主要研究方向为三维视觉和计算成像。
E-mail：xintao.wang@outlook.com
许华哲，男，助理教授，主要研究方向为具身智能和三维视觉。E-mail：huazhe_xu@mail.tsinghua.edu.cn
杨蛟龙，男，研究员，主要研究方向为计算机图形学和三维视觉。E-mail：jiaoyan@microsoft.com
康炳易，男，研究员，主要研究方向为计算机图形学和三维视觉。E-mail：bingykang@gmail.com
楚梦渝，女，助理教授，主要研究方向为三维视觉和计算成像。E-mail：mchu@pku.edu.cn
孙赫，男，助理教授，主要研究方向为三维视觉和计算成像。
E-mail：hesun@pku.edu.cn
陈文拯，男，助理教授，主要研究方向为三维视觉和计算成像。E-mail：wenzhengchen@pku.edu.cn
马月昕，女，助理教授，主要研究方向为具身智能和三维视觉。E-mail：mayuexin@shanghaitech.edu.cn
张鸿文，男，副教授，主要研究方向为计算机图形学和三维视觉。E-mail：zhanghongwen@bnu.edu.cn
郭裕兰，男，教授，主要研究方向为三维视觉和计算成像。
E-mail：guoyulan@sysu.edu.cn
周晓巍，男，教授，主要研究方向为计算机图形学和三维视觉。E-mail：xwzhou@zju.edu.cn
章国锋，男，教授，主要研究方向为计算机图形学和三维视觉。E-mail：zhangguofeng@cad.zju.edu.cn
韩晓光，男，助理教授，主要研究方向为计算机图形学和三维视觉。E-mail：hanxiaoguang@cuhk.edu.cn
戴玉超，男，教授，主要研究方向为三维视觉和计算成像。
E-mail：daiyuchao@nwpu.edu.cn