中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-16

论文引用格式:Han Xiaoguang, Xiu Yuliang, Xu Zhen, Lian Zhouhui, Peng Sida, Yao Yao, Chen Anpei, Huang Jingwei, Zhang Bang, Xu Lan, Xu Feng, Zhang Guofeng, Xu Weiwei, Yu Jingyi, Liu Ligang, Chen Baoquan, Liu Yebin, Zhou Xiaowei. Frontiers and prospects of 3D reconstruction and generation[J/OL]. Journal of Image and Graphics, XXXX:1-16. DOI: 10.11834/jig.260070.(韩晓光, 修宇亮, 徐震, 连宙辉, 彭思达, 姚遥, 陈安沛, 黄经纬, 张邦, 许岚, 徐枫, 章国锋, 许威威, 虞晶怡, 刘利刚, 陈宝权, 刘烨斌, 周晓巍. 三维重建与生成前沿进展与展望[J/

三维重建与生成前沿进展与展望

韩晓光[1],修宇亮[2],徐震[3],连宙辉[4],彭思达[3],姚遥[5],陈安沛[2],黄经纬[6],张邦[7],许岚[8],徐枫[9],章国锋[3],许威威[3],虞晶怡[8],刘利刚[10],陈宝权[4],刘烨斌[9],周晓巍[3]

1. 香港中文大学(深圳),广东省深圳市　518172; 2. 西湖大学,浙江省杭州市　310024; 3. 浙江大学,浙江省杭州市　310058; 4. 北京大学,北京市　100091; 5. 南京大学,江苏省南京市　210008; 6. 深圳市腾讯计算机系统有限公司,广东省深圳市　518057; 7. 杭州阿里巴巴网络技术有限公司,浙江省杭州市　311121; 8. 上海科技大学,上海市　201210; 9. 清华大学,北京市　100084; 10. 中国科学技术大学,安徽省合肥市　230026

摘　要: 近年来,三维视觉领域正经历一场深刻的范式转变,核心问题正从单一的"感知重建"逐步迈向"重建—生成—交互"的一体化新阶段。本文旨在系统梳理三维重建与生成技术的前沿进展,对三维重建、三维生成及三维数字人等方向进行综述,剖析优化式与前馈式重建方法的原理差异,评估物体级生成、CAD(computer-aided design)生成及具身智能场景生成的现状与挑战,并对比2D与3D数字人技术在实时渲染与复杂交互中的表现。分析显示,三维重建技术中,优化式方法虽精度占优但计算冗余,而前馈式方法虽推断迅速但细节不足,两者融合及多模态语义注入是当前主流;三维生成领域,技术焦点已从单纯的视觉质量转向部件级可控性,但CAD生成仍面临"脏几何"难以满足制造标准的难题;数字人技术方面,2D生成技术展现了非凡的生成能力与迭代速度,而3D技术在处理复杂空间交互时具有不可替代性。研究表明,三维领域正经历从"观测驱动重建"向"数据驱动生成"的范式转变,未来发展将集中在前馈式与优化式方法的深度融合,三维生成向工业可用性与可编辑性演进,三维技术与具身智能、数字人等场景深度耦合三方面。未来,三维重建与生成将不再是孤立的视觉问题,而是支撑虚实融合与智能决策的基础能力。

关键词: 三维重建;三维生成;数字人;空间智能;具身智能

Frontiers and prospects of 3D reconstruction and generation

Han Xiaoguang[1], Xiu Yuliang[2], Xu Zhen[3], Lian Zhouhui[4], Peng Sida[3], Yao Yao[5], Chen Anpei[2], Huang Jingwei[6], Zhang Bang[7], Xu Lan[8], Xu Feng[9], Zhang Guofeng[3], Xu Weiwei[3], Yu Jingyi[8], Liu Ligang[10], Chen Baoquan[4], Liu Yebin[9], Zhou Xiaowei[3]

1. The Chinese University of Hong Kong, Shenzhen, Shenzhen, Guangdong Province 518172; 2. Westlake University, Hangzhou, Zhejiang Province 310024; 3. Zhejiang University, Hangzhou, Zhejiang Province 310058; 4. Peking University, Beijing Municipality 100091; 5. Nanjing University, Nanjing, Jiangsu Province 210008; 6. Shenzhen Tencent Computer Systems Co., Ltd., Shenzhen, Guangdong Province 518057; 7. Hangzhou Alibaba Network Technology Co., Ltd., Hangzhou, Zhejiang Province 311121; 8. ShanghaiTech University, Shanghai Municipality 201210; 9. Tsinghua University, Beijing Municipality 100084; 10. University of Science and Technology of China, Hefei, Anhui Province 230026

收稿日期:2026-01-29;**修回日期**:2026-03-09

*通信作者:周晓巍,男,浙江大学计算机科学与技术学院教授。研究方向包括三维计算机视觉、混合现实与智能机器人等

基金项目:国家自然科学基金(批准号:U24B20154;62125107)

Supported by:Project supported by the National Natural Science Foundation of China (Grant No. U24B20154;62125107)

Abstract: The field of 3D vision is currently undergoing a profound and historical paradigm shift, transitioning from a traditional core focus on "perception and reconstruction"—which emphasizes the faithful recovery of geometry from observation—to a new, integrated stage characterized by "reconstruction-generation-interaction". This paper provides a systematic, comprehensive, and critical review of the frontier advances in 3D reconstruction and generation technologies, covering key directions including 3D reconstruction paradigms, 3D object and scene generation, and the evolution of 3D digital humans. Specifically, it analyzes the fundamental principle differences between optimization-based and feed-forward reconstruction methods, evaluates the current status and bottlenecks of object-level generation, CAD generation, and embodied AI scene generation, and compares the performance and future viability of 2D versus 3D digital human technologies in the context of real-time rendering and complex spatial interactions. The analysis indicates that in the domain of 3D reconstruction, two distinct technical paradigms have emerged, each with unique strengths and limitations. Optimization-based methods, exemplified by classical pipelines like COLMAP and modern neural representations such as Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3DGS), excel in achieving high-precision geometric recovery. By defining 3D representations as optimizable structures and iteratively minimizing the photometric error between rendered results and ground truth, these methods can achieve sub-millimeter level accuracy on standard benchmarks like the DTU dataset. However, they suffer from significant computational redundancy, often requiring hundreds or thousands of iterations, and exhibit poor robustness when dealing with ill-posed problems such as sparse-view inputs or textureless regions. In contrast, feed-forward reconstruction methods, represented by architectures like VGGT, DUSt3R, and Fast3R, represent a shift towards data-driven direct prediction. These models utilize neural networks to predict 3D geometry from input images in a single forward pass, offering rapid inference speeds and superior generalization capabilities in underdetermined conditions. However, they currently lack the fine-grained detail of optimization methods, often producing "blurry" geometry due to the domain gap between synthetic training data and real-world scenarios. The review identifies that the current mainstream trend is the deep fusion of these two paradigms—such as using feed-forward models to initialize optimization processes (e.g., InstantSplat)—and the injection of multimodal semantic information. Notably, optimization methods excel at fusing pixel-aligned modalities like depth and normals, whereas feed-forward methods are more adept at integrating abstract semantic signals such as text and audio. In the realm of 3D generation, the technical focus has shifted rapidly from pure visual quality to structural controllability and part-level manipulation. While "3D native" large models have achieved success in generating high-resolution meshes, significant challenges remain in bridging the gap between AI generation and industrial production standards. A critical bottleneck identified is the specific domain of CAD generation. Although recent approaches have attempted to model CAD designs as sequence learning problems or graph-based B-Rep generation, they frequently face the "dirty geometry" challenge. Generated models often exhibit non-watertight surfaces, irregular topologies, and defective assembly relations, failing to meet the strict constraints required for manufacturing, physics simulation, or downstream engineering tasks. Furthermore, the paper highlights the intersection of 3D generation and Embodied AI. For robotic simulation and training, the demand is not merely for static visual assets but for interactive scenes where objects possess articulated parts, physical properties, and realistic, messy layouts. Current generative methods struggle to produce such complex, unordered scenes that reflect the reality of the physical world, which is essential for training robust embodied agents. Regarding digital human technology, the field is witnessing a fierce competition between 2D and 3D approaches. 2D generation methods, driven by advanced video diffusion models (e.g., Sora, Animate Anyone), have shown tremendous progress, delivering hyper-realistic visuals and rapid iteration speeds that challenge the necessity of traditional 3D pipelines. These 2D models leverage massive video datasets to learn motion and appearance, often bypassing the need for explicit geometric modeling. However, the analysis argues that 3D technology remains irreplaceable in handling complex spatial interactions. 3D digital humans (e.g., Gaussian Avatars) provide the precise geometric consistency, explicit depth information, and collision data necessary for immersive VR/AR (Virtual Reality/Augmented Reality) experiences and autonomous system interactions—capabilities that 2D video generation currently lacks. The study suggests a future trend where 2D and 3D technologies converge, with massive 2D video data being used to supervise and refine 3D representations, moving toward "multimodal interactive avatars" capable of understanding and generating speech, gesture, and expression in real-time environments. The study reveals that the 3D field is experiencing a fundamental paradigm shift

from "observation-driven reconstruction" to "data-driven generation". Future developments will likely concentrate on three strategic directions: first, the deep fusion of feed-forward and optimization-based methods to solve the inherent trade-off between robustness and precision; second, the evolution of 3D generation towards industrial usability and editability, ensuring assets are physically valid and topologically sound rather than just visually plausible; and third, the deep coupling of 3D technology with scenarios such as embodied AI and digital humans, where generation serves functional interaction needs. Ultimately, 3D reconstruction and generation are no longer isolated visual problems but have evolved into foundational capabilities supporting virtual-real fusion and intelligent decision-making.

Key words: 3D Reconstruction; 3D Generation; Digital Human; Spatial Intelligence; Embodied AI

0 引　言

近年来，三维视觉领域正从以“感知重建”为核心的问题设定，逐步迈向“重建—生成—交互”一体化的新阶段。随着前馈式三维重建与三维生成模型的快速发展，传统依赖迭代优化的重建范式在速度、鲁棒性与泛化能力上正面临新的挑战。本文系统梳理了三维重建与生成领域的关键技术，围绕优化式与前馈式重建范式、三维生成模型以及三维数字人应用等方向，分析当前能力边界与核心瓶颈，总结三维视觉从“几何恢复”走向“可用资产生成与可交互世界模型”的演进路径。

1 三维重建范式之争：优化式重建算法与前馈式重建算法

1.1 三维重建常用技术范式：优化式与前馈式

针对三维重建问题（给定图像，重建其三维结构），目前常用的技术范式有两种：优化式重建（如COLMAP（Schonberger等，2016））与前馈式重建（如VGGT（Wang等，2025））。除RGB外，输入还可能包括相机参数、基于深度相机、LiDAR等传感器获得的深度信息等；重建的三维结构包括稀疏或稠密点云、水密三角面片、体素，或近年来兴起的神经辐射场或三维高斯球等。优化式重建一般指代将三维表示设计为可优化结构，并设计损失或能量函数，以渲染结果和真值做比较或设计重投影流程，以迭代方法优化该三维表示的算法，例如NeRF（Mildenhall等，2021）、3DGS（3D gaussian splatting）（Kerbl等，2023）。前馈式重建指代近年来兴起的、基于神经网络的直接预测三维几何的方法。大部分方法都是基于像素和相机组成的射线，对预测深度值和其他属性做反投影的管线，如DUSt3R（Wang等，2024），Fast3R（Yang等，2025）等。两种范式在测量精度、算法鲁棒性、视觉质量、运行速度、多模态融合性等角度有较为显著的区别。

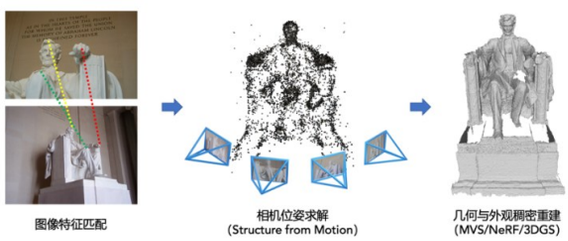


图1　优化式重建算法流程（Schonberger等，2016）

Fig. 1　Optimization-Based Reconstruction Algorithm

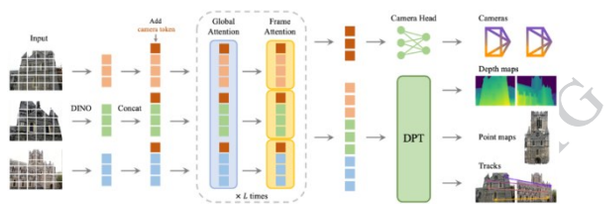
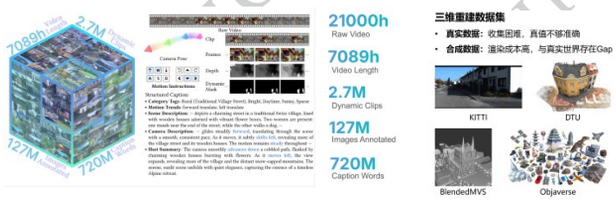


图2　前馈式重建网络架构（VGGT（Wang等，2025））

Fig. 2　Feed-forward Reconstruction Network Architecture

1.2 前馈式与优化式重建精度区别

从测量精度角度看，目前前馈式重建算法往往



（SpatialVid（Wang等，2025），KITTI（Geiger等，2013），DTU（Jensen等，2014），BlendedMVS（Yao等，2020），Objaverse（Deitke等，2023））

图3　相较于视频/图片数据，三维数据收集难/真值差

Fig. 3　Compared with video/image data, 3D data is harder to acquire and lacks high-quality ground truth.

无法达到优化式重建算法的精度级别。在DTU(Jensen等,2014)数据集上,COLMAP,Gipuma(Galliani等,2015),NeuS(Wang等,2021)等传统重建算法能够获得小于毫米级别的高精度重建结果,而目前最先进的前馈式算法Depth Anything 3(Lin等,2025)误差通常仍在毫米量级。限制前馈式三维重建算法精度进一步提高的最大因素是高质量三维数据的稀缺性,目前市面上的主流三维数据分为两类:真实数据与CG(computer generated)数据。真实数据的三维重建真值往往是通过优化式重建算法获得的,因此基于此类数据流程训练的前馈式重建算法精度只能无限逼近为其提供真值的优化式算法。除此之外,通过LiDAR或其他深度感知硬件获得的三维重建结果往往不够准确,其原因在于目前深度感知硬件发展不完全,扫描误差较大。基于游戏引擎或三维渲染引擎,由艺术家设计并构建的三维场景的确能生产出已知真实深度或完整三维几何结构的三维重建数据对,但目前此类数据与真实场景存在较大的数据分布差异(domain gap),对前馈式重建算法在一般场景下的表现效果提升不高。因此,如何更大规模采集多样性更充足的CG数据,以及如何更好地使用CG数据训练前馈模型以减弱数据分布差异的影响,是提高其重建精度表现的关键因素。

1.3 稀疏视角与欠定条件下的重建鲁棒性

从重建鲁棒性角度看,前馈式重建算法,如VGGT和Depth Anything 3对欠定问题(ill-posed problem)的处理能力强于优化式重建算法。目前最先进的前馈式重建可在稀疏视角或单目情况下获得视觉质量尚可的重建结果。前馈式模型的此种表现与神经网络的泛化能力有密切关系;对比之下,优化式算法仅能通过引入三维世界上的归纳偏置(inductive bias)信息进行重建,无法处理欠定问题。目前,能在欠定问题上获得一定成效的优化式算法,例如MonoSDF(Yu等,2022),Shape-of-Motion(Wang等,2024)和Mosca(Lei等,2025)都引入了其他前馈式模型作为补充,将优化式重建算法的输入做拓展,使其成为非欠定问题后,才可获得较为可观的重建效果。

1.4 重建结果视觉质量区别

从视觉质量角度看,受到重建精度限制,前馈式重建算法目前仍然无法获得与优化式重建相当的三维几何视觉效果,仍存在水密性不够、拓扑连续性差、几何破洞、内容重复、外观清晰度差、平滑度过高等视觉质量问题。而优化式重建算法对显存/内存开销的需求相对可控,能处理更为精细的几何与外观,目前优化式重建算法往往能获得更为精细完整的重建视觉质量。从三维生成领域的进展推断,前馈式重建算法有望在视觉质量角度获得与目前优化式算法相似的结果,在局部欠定问题上(例如欠观察的角度或区域),有望取得超越优化式算法的视觉精度。

(a) DUSt3R不重叠图像重建

(b) VGGT稀疏图像重建

(a) DUSt3R Non-Overlapping Image Reconstruction (b) VGGT Sparse Image Reconstruction

图4 前馈式重建稀疏视角/欠定条件重建

Fig. 4 Feed-forward Reconstruction for Sparse-View/Underdetermined-Condition Reconstruction

1.5 重建算法运行速度区别

从重建速度角度看,前馈式重建算法在实现与部署上存在性能优势。优化式重建算法无法避免迭代过程,成百上千步的优化迭代流程往往会拖慢重建流程。而前馈式重建算法从原理上将大批量计算开销移到了模型训练时,一次前向传播即可完成推理,例如Depth Anything 3在模型参数量规模较大的情况下也能做到较高、甚至实时的推理效率。目前业内也有将前馈式算法的运算速度和泛化性优势与优化式算法的精度优势做结合的工作,最简单有效的为InstantSplat(Fan等,2025)等将前馈式重建模型

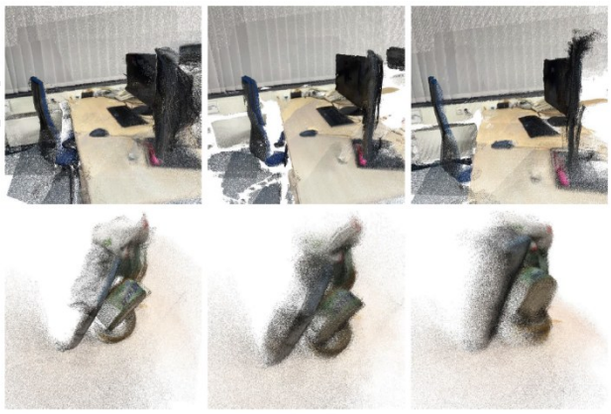


图5　前馈式重建精度、置信度低、视觉效果模糊(Depth Anything 3)，左至右分别为VGGT，Pi3，Fast3R结果

Fig. 5　Feed-forward reconstruction suffers from low accuracy, low confidence, and blurry visual effects (results from VGGT, Pi3 and Fast3R from left to right)

的重建结果直接输入优化算法进行微调的方式，以获得对场景更精确细致的三维表达。

1.6　多模态信息融合能力区别

此外，在多模态信息融合方面，优化式方法在像素对齐的多模态信息融合上有优势，而前馈式方法在抽象语义的多模态信息融合上有优势。针对像素级多模态信息，例如深度、法向、光流、语义分割、实例分割、渲染材质或高维像素级语义向量如DINOv2(Oquab等，2023)，优化式算法能方便地将其作为渲染结果中的额外通道并引入和RGB图像相同的像素优化损失；而引入额外通道容易导致前馈式算法模型中的神经网络出现容量不足现象。针对语义多模态信息，例如文本，语音，或非精确的运动控制信号，由于前馈式算法往往采用可拓展神经网络结构，这些方法能通过引入额外前馈网络层(multi-layer perceptron)或注意力头(attention head)较为方便地实现抽象语义信号的注入；而优化式方法很难接收非像素级的语义信息的引导或监督。目前，如何为前馈式重建算法引入多模态信息的研究仍处于起步阶段，大部分研究，包括VGGT的输入信号仅限于视觉信息，而三维生成领域的进展，例如Wonder3D(Long等，2024)，已经证明能很好地将其他模态，例如文本描述或三维脚手架，作为前馈式模型的输入。

2　从重建到生成的问题演变：三维生成领域的发展与挑战

在三维重建能力逐步成熟的背景下，研究重心正由“还原已有世界”转向“生成三维内容”。三维生成模型在物体级、部件级乃至场景级建模上的快速进展，显著降低了三维资产生产门槛，但也暴露出表征能力、工业可用性与可控性不足等新问题。本章聚焦三维生成领域的技术演进，系统分析当前主流方法的能力与核心挑战，探讨三维生成从研究走向实际生产工具所面临的瓶颈。

2.1　三维生成的发展和挑战

近年来，从DreamFusion(Poole等，2022)的“2D升维”路径到如今各种“3D原生”大模型的层出不穷(Lai等，2025)(Tochilkin等，2024)(Xiang等，2025)，展现出三维生成领域的快速发展。如今，给定一张照片，生成一个高质量的三维模型的任务在某些情况下已经能达到非常好的效果(例如图6展示的Hunyuan3D v3.0的效果)。2025年5月，Sparc3D(Li等，2025)的出现首次将三维生成带到了高分辨率阶段，7月，BANG(Zhang等，2025)及其一系列后续的工作将三维生成从物体级带到了部件级(如图7展示的是PartCrafter的效果)(Tang等，2025)。

图6　Hunyuan3D v3.0单张图片生成三维模型的效果

Fig. 6　Performance of Hunyuan3D v3.0 in 3D Model Generation from a Single Image

尽管发展迅速，但仍然有以下问题困扰着整个领域：

3D表征仍待完善：3D表征方式是决定生成模型上限的基石。当前业界主要有两类技术路径：全局表征方法(如3DShape2VecSet(Zhang等，2023))-该方法将物体编码为具有全局意义的离散token。其优势在于生成的几何整体结构稳定，但对局部细节的刻画能力相对有限；局部表征方法(如Sparc3D)-这类方法在细节还原方面表现更佳。但其中常引入渲染损失等具有较强归纳偏置的约束，可能导致潜在表征学习到非物理规律的信号，影响

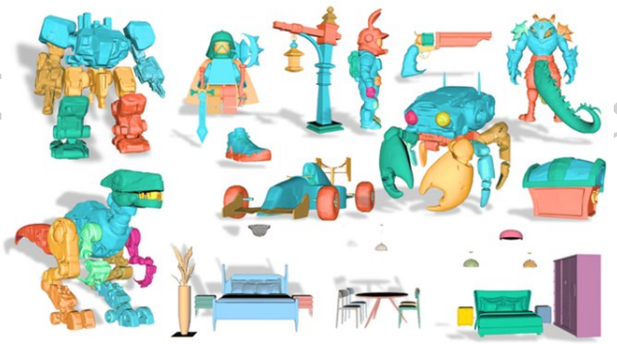


图7　现有三维生成方法已经达到了部件级生成效果(PartCrafter)

Fig. 7　Current 3D generative methods have achieved part-level generation performance

生成质量的稳定性和可控性。理想的表征应兼具全局一致性与局部细节保真度,且避免引入过多假设,目前尚无完美方案。

距离工业级“可用”仍存在巨大差距:其核心挑战在于3D资产的可编辑性和轻量性。一个仅仅是“形状相似”的模型,若不具备水密性及规整的中模布线拓扑,将无法满足后续的编辑诉求和曲面细分。而不合理的低模拓扑,会给动画绑定、实时渲染等流程带来灾难,艺术家仍需投入大量时间进行手动调整。这正是当前三维生成模型在融入现有生产管线时的主要瓶颈。

学术研究的空间被急剧压缩:整个领域正在从学术研究走向工业应用,革命性的创新已经很难,剩下了很多工程方面的任务有待解决。与此同时,我们也逐渐发现大家把目光转移到了一些交叉处,比如视频与3D、3D与物理、3D与具身等。这些交叉研究领域尚未进入Scaling Law能够充分发挥效益的阶段,因此学界可以发挥更大作用。

具体而言,接下来在三维生成方面值得探索的研究方向包括:

2.1.1 生成与重建的统一。

虽然目前单张或少量几张图已经可以通过生成模型很好的得到一个三维模型,但生成结果与图像的匹配却可能产生较大的差异。例如,图像的椅背有5根横杆,但得到的模型却可能变成6根。此外,目前基于3DGS的物体重建虽然可以得到很好的效果,但由于真实采集不可避免的遮挡,使得重建难以得到完整模型。因此,一个统一的模型使对于图像中可见区域进行重建而对于不可见区域基于生成模型进行补全将是一个必要的趋势。

2.1.2 结构化三维物体生成。

目前的部件级生成,仅考虑到几何层面上如何将现在的三维模型进行部件的划分。而真实场景中,物体不光是部件的简单拼接,部件之间一般都会通过一定的结构进行连接,如剪刀、微波炉门等的铰链结构、家具中的部件组装关系等。如今面向游戏场景的三维资产生成更多只是为了可观看,而面向未来可交互的游戏或者服务于具身训练的可交互仿真场景,物体则需要可操作或拆卸,因此对于部件的结构关系的推理和生产是必然的趋势。

2.1.3 三维场景生成。

虽然现在的三维物体生成获得了非常快速的发展,在场景生成方面的研究仍然比较匮乏。目前来看,有两种场景生成的范式:一是,将场景表达成一个完整的全景图或者多视角视频,并利用图像或者视频生成的方法进行;二是,首先生成场景中物体的布局,再对每个物体单独生成。不论是哪种范式,目前均面临着一些挑战,其中一个最大的难点在于数据的匮乏;另外,如何保证场景的全局闭环(loop closure)也是较大的挑战。当前,可以用到的场景数据大多都是居家生活类的,对于工厂、医院等复杂场景则极度匮乏,这也是未来亟待解决的问题。

2.2　三维CAD生成现状及挑战

3D模型的应用主要有两大方向:一个方向是通过3D建模输出一张图像,是在“虚拟世界”中的生成,让人眼能看到及理解这个虚拟世界;另一个方向是通过3D建模进行“制造”(“实体化”),生成我们生活中的实际物品,比如手机、遥控器等等,即,使用真实的材料去将3D数字文件加工成一个实体。因此,3D模型的两个应用方向,一个是虚,一个是实。虚的应用方向是面向娱乐、虚拟现实、元宇宙等领域;实的应用方向是面向CAD/CAE/CAM及智能制造领域。

CAD产品的外形表达并非采用离散点云或网格表示,而是基于数学表达形式(比如Bezier曲面、B-样条曲面)的分片表达,称为边界表达B-Rep(boundary representation)。

近年来,基于生成式AI的CAD设计及建模的研究已引起众多学者的关注。近期已有综述性工作(如中国科学技术大学的综述论文AI-driven Generation of 3D CAD Models: A Survey,CVM期刊)对该方

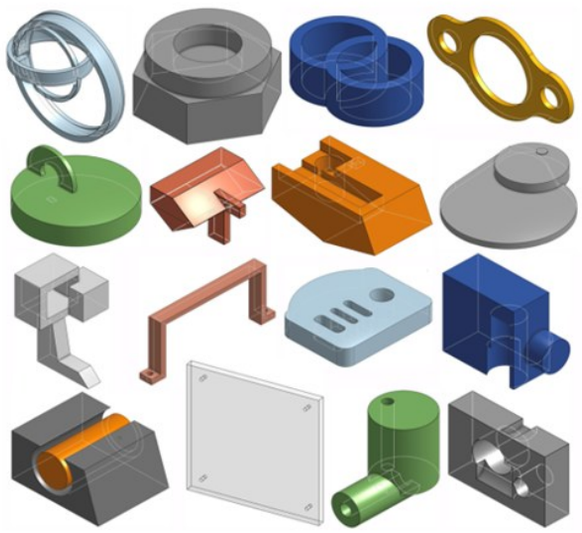


图8　使用AI生成的CAD模型(DeepCAD(Wu等,2021))

Fig. 8　AI-generated CAD models

向的研究进展进行了系统梳理。该方向的早期研究主要由AI领域研究者推动,其核心观察在于,CAD设计过程能够被表达为指令序列,其形式与自然语言序列在结构上具有一定相似性,因此,可以将CAD建模问题视为序列学习问题,通过直接学习指令序列实现模型生成。各类生成模型已被引入用于建模文本描述与CAD指令序列之间的映射关系,例如Transformer、Llama及Mamba等模型架构。第二种方法是基于B-Rep表达的方式,通过生成点、边、面构成的图结构的形式来直接生成CAD模型的B-Rep表达。第三种方法是基于大语言模型作为设计代理(agents)推断设计过程来生成CAD模型。

在具体实现上,相关研究首先依赖于大规模数据集的构建。近年来,CAD数据集不断丰富,从早期广泛使用的ABC数据集(Koch等,2019),到DeepCAD数据集(Wu等,2021),再到国内研究团队(如武汉大学的WHUCAD数据集等)发布的多个CAD数据集(Fan等,2021),其规模和多样性均持续提升。在此基础上,各类生成模型被引入用于建模文本描述与CAD指令序列之间的映射关系,例如Transformer、Llama及Mamba等模型架构(Xu等,2023)(Li等,2025)(Li等,2025)。该方向的早期研究主要由AI领域研究者推动,其核心思路在于将CAD建模问题视为序列学习问题,通过直接学习指令序列实现模型生成。在数据规模足够充足的前提下,设计师可通过文本描述生成对应的CAD指令序列,从而完成模型设计。

然而,需要指出的是,CAD模型并不仅仅是几何外形的三维表示,其在实际工业应用中还涉及装配关系、物理仿真以及制造约束等复杂环节。B-rep采用B样条函数对曲面进行表达,受其表达形式本身的限制,在实际建模过程中不可避免地会产生“脏几何”、非水密曲面等问题,这也是长期以来工业界在CAD建模与应用中面临的重要挑战之一。

2.3　三维生成与具身智能

随着具身智能概念的兴起,三维生成领域的研究者们也在开始考虑如何服务于具身智能场景,其中一个重要的应用场景便是利用三维生成技术构建虚拟的仿真环境用于机器人的训练。那么具身场景给三维生成提出了什么新的要求呢?

在服务具身智能应用时,首先需要明确当前亟待解决的关键问题。现阶段的三维生成技术在物体级资产生成方面已取得较为成熟的进展,诸如Objaverse(Deitke等,2023)、ShapeNet(Chang等,2013)等公开资产库已在一定程度上满足需求,基于这些现有资产开展几何分析任务(如抓取规划)在多数情况下亦具备可行性。因此,单纯引入更多物体资产本身的边际价值有限。事实上,在具身场景中,许多任务并不依赖于大规模资产数量,仍可在相对有限的资产条件下实现较好的通用性。相比之下,当前具身应用中更具挑战性的环节在于场景内的交互建模。一方面,物体资产需要具备可交互性,例如在虚拟环境中,剪刀不仅能够被拿起和操控,还应能够真实地完成剪切动作,并使被作用物体产生符合物理规律的变化。这意味着系统不仅需要生成物体的几何形态与纹理外观,还需进一步推理其部件结构关系,并同时建模物体的材质属性与物理特性。另一方面,场景布局的真实性同样至关重要。当前多数具身仿真环境中,物体往往被规整、有序地摆放,且内部空间(如冰箱内部、柜体内部等)资产严重不足,而真实世界中的物体分布通常更加密集且无序。如何生成这种贴近现实、结构复杂且可交互的场景布局,仍然是一个具有高度挑战性的研究问题。

3　重建与生成的典型应用:三维数字人前沿与展望

作为三维重建与生成技术最具代表性的应用场

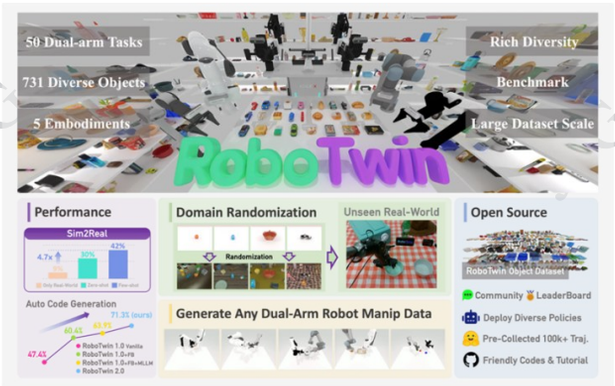


图9　现有的三维生成方法可用于生成具身智能场景所需的数字资产(RoboTwin(Mu等,2025))

Fig. 9　Existing 3D generative methods enable the generation of digital assets for embodied intelligence scenarios

景之一,数字人是通过计算机图形学、计算机视觉、机器学习等多学科融合技术构建的,能够模拟人类外观、动作、交互行为甚至情感表达的数字化实体。其核心特征在于对“人”的数字化还原与功能拓展,既涵盖视觉层面的写实呈现,也包含行为层面的交互及场景互动能力。数字人集中体现了几何精度、动态一致性与交互能力的综合需求,从技术形态上可分为2D数字人(如Animate Anyone(Li等,2025),OmniHuman-1.5(Jiang等,2025))与3D数字人(如AnimatableGaussians(Li等,2023),(Zhan等,2025),(Xu等,2024))。随着2D生成模型能力的持续提升,数字人领域正面临技术路线与研究定位的重构。本章围绕三维数字人的技术演进与现实约束,分析2D与3D数字人之间的互补与竞争关系,探讨数字人在具身交互、沉浸式体验及功能性应用中的长期价值与发展方向。

3.1　2D和3D数字人的替代性之争

近两年视频生成技术在人物视频生成所展现的非凡能力(如Sora 2, HeyGen, Kling-Avatar(Ding等,2025),EMO(Tian等,2024), Animate Anyone(Li等,2025), VASA-3D(Xu等,2025), Media2Face(Zhao等,2024), SCAIL(Yan等,2025)等),让学界和产业界对2D数字人能否完全替代3D数字人这一核心问题上存在显著分歧。部分学者对3D数字人的长期前景持偏悲观态度,认为2D视频生成技术已可在常规硬件上实现数字人实时渲染,3D技术的速度优势已不显著,仅在简单交互场景的逼真度上略具优势,面对复杂交互场景的适配性反而不足;3D数字人在

(a)　基于重建算法的3D数字人

(b)　基于生成算法的2D数字人

(a) Reconstruction-Based 3D Digital Humans (b) Generative-Based 2D Digital Humans

图10　基于重建的3D数字人与基于算法的2D数字人

Fig. 10　Reconstruction-Based 3D Digital Humans vs. Generative-Based 2D Digital Humans

沉浸式内容生产中的优势具有暂时性,未来将被实时渲染视频扩散模型等新一代范式取代,仅在AR/VR(augmented reality / virtual reality)眼镜等端侧强交互、多模态应用中仍能保留适配性价值。另有部分学者认为,3D数字人在功能交互、显示适配及特定实时场景中具备不可替代性。从渲染速度和三维一致性看,采用3DGS的3D多视点显示推理速度优于DiT(diffusion transformer)类2D模型,可满足VR及游戏等实时交互场景的核心需求;交互层面,具身智能、自动驾驶等场景需精准的数字人几何位置以实现避障、物体交互,3D技术可提供关键的位置信息支撑;显示适配方面,AR/VR头显等设备对效率与分辨率要求严苛,2D技术难以满足,3D仍是必要选择。3D数字人在功能性三维场景交互,如导购数字人TaoAvatar(Chen等,2025);手语数字人,如MIO(Cai等,2025);沉浸式体积视频,如FreeTimeGS(Wang等,2025)等领域具备2D难以覆盖的独特

价值。

图11　3D信息控制2D数字人生成

Fig. 11　3D Information-Driven 2D Digital Human Generation

3D信息在2D数字人技术发展中的角色正从显式基础支撑逐步向隐式特征演变，其核心价值随模型Scaling能力与数据可得性发生动态变化。

在技术发展历程中，3D信息的作用经历了多阶段更迭：数字人起源阶段，3DMM（3D morphable model）（Blanz等，1999）、SMPL（Loper等，2015）等3D信息奠定了数字人驱动的控制表征基础；GAN（generative adversarial network）时代，人脸Blendshape、关键点等3D信息成为2D数字人视频生成框架的核心支撑；第一代Diffusion时代（如Animate Anyone（Li等，2025），Animate Anyone 2（Li等，2025）），骨骼关键点、SMPL等3D信息仍作为控制信号保障生成精准度；进入第二代Diffusion（DiT架构）时代（如X-UniMotion（Song等，2025）），显式3D信息逐渐被一维/二维隐式特征替代，但3D承载的几何结构、运动规律等核心信息仍被隐式编码在模型参数中。专家预判，未来或出现“大模型时代的3DMM/SMPL”，以隐式特征形式持续支撑2D数字人的生成与交互。

从价值衰减逻辑来看，早期2D生成模型能力有限，必须依赖3D信息保障生成合理性；随着DiT架构推动Scaling Law落地，2D音视频海量数据的优势凸显，模型可直接从数据中学习动作与姿态的内在关联，显式3D信息的必要性随之降低，这本质是2D海量数据与3D数据稀缺性矛盾导致的技术路线倾斜。

3.3　数字人领域的独立研究定位及核心发展方向

当前，数字人领域正面临“存在主义危机”，其独立研究价值受到冲击，核心争议集中于领域边界的定义与技术路线的选择。

从历史贡献来看，数字人领域对“保真度”的极致追求，曾推动建模、渲染、仿真等多项技术突破；同时，人体作为融合刚体、非刚体、铰链物体的“无穷几何”，其高保真解决方案可为通用物体领域研究提供重要借鉴，衍生技术已广泛赋能General Objects相关方向。但近年来，AIGC（artificial intelligence generated content）技术的发展呈现“本末倒置”态势，数字人领域的重要成果多源于通用2D视频生成和通用3D对象生成技术在人体数据集上的特化，导致其独立研究地位受到挑战。

尽管如此，数字人在未来若干年仍将作为独特技术方向来归类和研究，其重点关注问题将从外观建模过渡到多模态交互，基于多模态生成与理解大模型的可实时交互数字人是热点前沿问题。交互的概念涵盖数字人与真人之间以及数字人与数字人之间进行实时双工交流，包括数字人的表情、手势、动作、声音（语音、语调、语气等），与物体和环境的交互，乃至行为和情感的拟真人化生成，从音容笑貌，到举手投足。数字人交互将向多维度一体化演进：实现理解-生成、多模态、生成-编辑的一体化能力；融入LLM/MLLM（large language model / multi-modal large language model）构建的“导演系统”，提升生成自然度；实现视音（包括人物发出的说话、唱歌、呻吟及环境对象声音）频同步生成（如OmniHuman-1.5（Jiang等，2025），OVI（Chetwin等，2025））；从单一数字人驱动升级为人-物-景协同交互生成（如ZeroHSI（Li等，2024），AHA（Mir等，2025），AvatarGO（Cao等，2024），CHOIS（Li等，2024））；在视频生成技术正

过渡到世界模型研究的关键阶段，视频生成技术同时也需特别关注多模态可交互数字人的生成问题（如 MAViD（Pang 等，2025），M3-Agent（Long 等，2025），M. I. O（Cai 等，2025），X-Streamer（Xie 等，2025），ORCA（He 等，2025），InteractiveOmni（Tong 等，2025））。诚然，最终的数字人能力将融入多模态大模型，成为视频生成、人机交互的核心模块。但其“人体 ID（identity）高保真、强交互、情感寄托、终生进化”需求是通用 2D 视频模型难以完全覆盖和需要特殊对待的，是视频生成技术的独特挑战。

另一方面，不管是基于视频生成，还是基于 3D 动作生成，数字人与场景和对象之间的交互生成，如手物交互和操作生成，正成为具身智能大脑的预训练基座大模型，推动具身智能世界模型的研究。互联网视频中人与环境交互（如 Human3R（Chen 等，2025））、第一视角人类交互数据（如 VITRA（Li 等，2025））、高质量人体运动捕捉数据（如 AMASS（Mahmood 等，2019）），这些数字人领域独特数据，皆是具身智能重要的数据基础；基于这些数据实现的人体和人手交互运动生成可通过隐空间映射直接驱动机器人身体关节、夹爪和灵巧手，相关探讨可参考 Mini3DV 观点报告《具身智能前沿与展望：数据、模型与系统演进》。

在数字人向外拓展，与环境或其他数字人及真人交互的同时，数字人技术的向内拓展也将是重要的未来研究方向。传统图形学中一直存在通过人体肌肉力学建模仿真人体外观运动的研究方向，近年来随着物理仿真技术、人体内部信息感知技术及深度强化学习等技术的不断发展（如 MyoSuite（Caggiano 等，2022），SKEL（Keller 等，2023）），这一方向有望在未来几年取得突破性进展，从而实现对人体内部运动生理相关信息，如肌肉激活、关节力矩等的感知与建模。相关技术在更广阔的领域，如运动疾病诊断与预测，面部及肢体手术模拟及人体外骨骼设计等方向，具有重要的应用价值。

3.4 3D 数字人技术的核心数据瓶颈与真实感提升的破解路径

诚然，当前数字人研究更多关注交互式生成，但从长远角度看，数字人的 3D 外观表征生成研究仍具有重要市场和价值。从现状来看，2D 数字人在静态写实渲染上已接近以假乱真，而 3D 数字人在泛化性和真实感上存在明显差距。然而，未来的 VR 内容

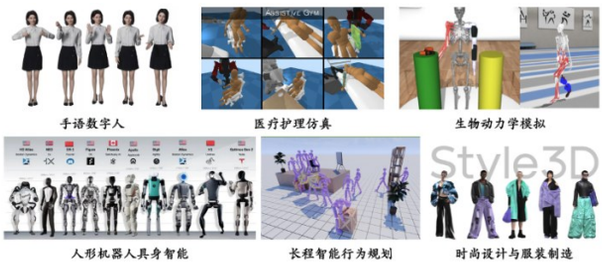


图 12　数字人交互与沉浸感的核心发展方向

Fig. 12　Core Development Directions for Digital Human Interaction and Immersion

生成和视点可交互视频生成方面，纯 2D 生成方式实现“隐式”的视点可交互性获得的沉浸体验仍难与更高质量体积视频等技术采用的显式 3D 表征所呈现的“可交互沉浸感”匹敌。如何获得 Mesh、3DGS 这类显式的 3D 数字人表达，兼容现有的各类 3D 设计软件，仍具有极其重要的意义和价值。

数据瓶颈是制约 3D 数字人技术迭代的核心因素，主要原因在于多视点同步采集困难，包括人物 ID 数量难以突破百万级、采集环境单一数据分布偏差显著，导致 3D 数字人技术短时间内难以显著突破。未来需针对性采取分层破解策略，如何有效使用日常拍摄的不完美数据（如 PuzzleAvatar（Xiu 等，2024），UP2You（Cai 等，2025）），如何充分利用 2D 视频数据（如 Vid2Avatar（Guo 等，2023），PGHM（Peng 等，2025））及巧用视频生成算法来辅助三维建模（如 IDOL（Zhuang 等，2025））是突破口。

具体而言，当前 2D 视频生成效果能够达到动作流畅自然，唇形语音同步，表情丰富细腻。然而，目前视频生成仍无法保证绝对的人物 ID 一致以及动态三维一致性，在实现复杂人物动作的自由视点视频时无法确保能够生成出高质量的 3DGS（人头效果较好，但人体效果较差）。另外，2D 视频生成在动态细节（如微表情、肢体运动自然惯性）、复杂场景交互（如衣物与物体的真实碰撞、光影精准映射）等方面仍存瑕疵，长时生成还易出现特征漂移、人脸 ID 改变等问题。若无法更好地挖掘 2D 视频数据中的三维信息，如视频帧之间精准的时空对齐，未来仍无法直接通过 2D 视频生成来实现 3D 数字人技术。

未来，通过少量图像前馈生成特定对象 3D 数字化身（如 IDOL（Zhuang 等，2025），LHM（Qiu 等，2025），PSHuman（Li 等，2024）等）仍有巨大的应用空间。如何融合大规模 2D 视频数据与有限 3D 人体

图13　2D数字人的视觉质量问题

Fig. 13　Visual Quality Issues of 2D Digital Humans

数据，充分探索2D视频生成技术与3D数字人生成技术的融合之道是一个值得研究的方向。譬如，通过2D视频生成技术实现质量尚可的海量3D数据以补充3D信息缺失，并通过低质3D数据预训练结合高质3D数据后训练方式实现3D数字人前馈生成（如HuaPi(Zhang等，2025)）。亦或，将现有2D视频生成技术升维到3D/4D，通过少量的目标人物图片输入条件控制，实现前馈式多模态表情和动作参数控制下的目标人物4D高斯生成。总的来说，前馈式、多模态可交互的带3D表征的数字人重建与生成，是数字人外观建模技术的终极问题。

4　总结与展望

综上所述，三维领域正经历从“观测驱动重建”向“数据驱动生成”的范式转变。未来的发展将集中体现在三方面：(1)前馈式与优化式方法的深度融合，以兼顾精度、效率与鲁棒性；(2)三维生成向工业可用性与可编辑性演进，突破资产生产与制造约束；(3)三维技术与具身智能、数字人等场景深度耦合，服务真实交互与功能需求。最终，三维重建与生成将不再是孤立的视觉问题，而是支撑虚实融合与智能决策的基础能力。

参考文献(References)

Blanz V and Vetter T. 1999. A morphable model for the synthesis of 3D faces//Proceedings of the SIGGRAPH. Los Angeles: ACM Press. [DOI: 10.1145/311535.311556]

Caggiano V, Wang H W, Durandau G, Sartori M and Kumar V. 2022. MyoSuite: A Contact-rich Simulation Suite for Musculoskeletal Motor Control//Proceedings of The 4th Annual Learning for Dynamics and Control Conference. Stanford: Proceedings of Machine Learning Research. [DOI: 10.48550/arXiv.2205.13600]

Cai Y Y, Chu X G, Gao X W, Gong S T, Huang Y F, Kang C X, et al. 2025. Towards Interactive Intelligence for Digital Humans. arXiv preprint arXiv:2512.13674. [DOI: 10.48550/arXiv.2512.13674]

Cai Z Y, Li Z Y, Li X B, Li B Q, Wang Z Y, Zhang Z Y, et al. 2025. UP2You: Fast Reconstruction of Yourself from Unconstrained Photo Collections. arXiv preprint arXiv: 2509.24817. [DOI: 10.48550/arXiv.2509.24817]

Cao Y K, Pan L, Han K, Wong K-Y K and Liu Z W. 2024. Avatargo: Zero-shot 4d human-object interaction generation and animation. arXiv preprint arXiv: 2410.07164. [DOI: 10.48550/arXiv. 2410.07164]

Chang A X, Funkhouser T, Guibas L, Hanrahan P, Huang Q X, Li Z M, et al.2015. Shapenet: An information-rich 3d model repository. arXiv preprint arXiv: 1512.03012. [DOI: 10.48550/arXiv. 1512.03012]

Chen J C, Hu J C, Wang G G, Jiang Z H, Zhou T S and Chen Z W, et al. 2025. TaoAvatar: Real-Time Lifelike Full-Body Talking Avatars for Augmented Reality via 3D Gaussian Splatting//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 10723-10734. [DOI: 10.1109/CVPR52734.2025.01002]

Chen T-S, Siarohin A, Menapace W, Deyneka E, Chao H W, Jeon B E, et al. 2024. Panda-70m: Captioning 70m videos with multiple cross-modality teachers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 13320-13331. [DOI: 10.1109/CVPR52733.2024.01265]

Chen Y, Chen X Y, Xue Y X, Chen A P, Xiu Y L and Pons-Moll G. 2025. Human3R: Everyone Everywhere All at Once. arXiv preprint arXiv:2510.06219. [DOI: 10.48550/arXiv.2510.06219]

Cheng G, Gao X, Hu L, Hu S Q, Huang M Y, Ji C N, et al. 2025. Wan-Animate: Unified Character Animation and Replacement with Holistic Replication. arXiv preprint arXiv:2509.14055. [DOI: 10.48550/arXiv.2509.14055]

Deitke M, Schwenk D, Salvador J, Weihs L, Michel O, Vanderbilt E, et al. 2023. Objaverse: A Universe of Annotated 3D Objects//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE: 13142-13153. [DOI: 10.1109/CVPR52729.2023.01263]

Ding Y K, Liu J W, Zhang W Y, Wang Z K, Hu W T, Cui L Y, et al. 2025. Kling-avatar: Grounding multimodal instructions for cascaded long-duration avatar animation synthesis. arXiv preprint arXiv:2509.09595. [DOI: 10.48550/arXiv.2509.09595]

Fan R B, He F Z, Liu Y X, Song Y P, Fan L K, Yan X H, et al. 2024. A parametric and feature-based CAD dataset to support human-computer interaction for advanced 3D shape learning. Integrated Computer-Aided Engineering, 32 (1). [DOI: 10.3233/ICA-240744]

Fan Z W, Cong W Y, Wen K R, Wang K, Zhang J, Ding X H, et al.

2024. InstantSplat: Unbounded Sparse-view Pose-free Gaussian Splatting in 40 Seconds. arXiv preprint arXiv: 2403.20309. [DOI: 10.48550/arXiv.2403.20309]

Forte M-P, Kulits P, Huang C-H, Choutas V, Tzionas D, Kuchenbecker K J, et al. 2023. Reconstructing signing avatars from video using linguistic priors//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE: 12791-12801. [DOI: 10.1109/CVPR52729.2023.01230]

Galliani S, Lasinger K and Schindler K. 2015. Gipuma: massively parallel multiview stereopsis by surface normal diffusion//Proceedings of the IEEE International Conference on Computer Vision. Santiago: IEEE: 873-881. [DOI: 10.1109/ICCV.2015.106]

Geiger A, Lenz P, Stiller C and Urtasun R. 2013. Vision meets robotics: The kitti dataset. The international journal of robotics research. 32(11). [DOI: 10.1177/0278364913491297]

Guo C, Jiang T J, Chen X, Song J and Hilliges O. 2023. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE: 12858-12868. [DOI: 10.1109/CVPR52729.2023.01236]

He X H, Yang T Y, Cao K, Wu R Q, Meng C, Zhang Y, et al. 2025. Active Intelligence in Video Avatars via Closed-loop World Modeling. arXiv preprint arXiv: 2512.20615. [DOI: 10.48550/arXiv.2512.20615]

Hu L. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 8153-8163. [DOI: 10.1109/CVPR52733.2024.00779]

Hu L, Wang G Y, Shen Z, Gao X, Meng D C, Zhuo L, et al. 2025. Animate Anyone 2: High-Fidelity Character Image Animation with Environment Affordance. arXiv preprint arXiv: 2502.06145. [DOI: 10.48550/arXiv.2502.06145]

Jensen R, Dahl A, Vogiatzis G, Tola E and Aanæs H. 2014. Large scale multi-view stereopsis evaluation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE: 406-413. [DOI: 10.1109/CVPR.2014.59]

Jiang J W, Zeng W H, Zheng Z R, Yang J Q, Liang C, Liao W, et al. 2025. OmniHuman-1.5: Instilling an Active Mind in Avatars via Cognitive Simulation. arXiv preprint arXiv: 2508.19209. [DOI: 10.48550/arXiv.2508.19209]

Jiang Y H, Shen Z H, Hong Y, Guo C C, Wu Y Z, Zhang Y L, et al. 2024. Robust dual gaussian splatting for immersive human-centric volumetric videos. ACM Transactions on Graphics (TOG), 43(6). [DOI: 10.1145/3687926]

Jiang Y H, Yao K X, Su Z, Shen Z H, Luo H M and Xu L. 2023. Instant-nvr: Instant neural volumetric rendering for human-object interactions from monocular rgbd stream//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE: 595-605. [DOI: 10.1109/CVPR52729.2023.00065]

Keller M, Werling K, Shin S Y, Delp S, Pujades S, Liu C K, et al. 2023. From skin to skeleton: Towards biomechanically accurate 3D digital humans. ACM Transactions on Graphics (TOG), 42 (6). [DOI: 10.1145/3618381]

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

Koch S, Matveev A, Jiang Z S, Williams F, Artemov A, Burnaev E, et al. 2019. Abc: A big cad model dataset for geometric deep learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, California: IEEE: 9593-9603. [DOI: 10.1109/CVPR.2019.00983]

Lai Z Q, Zhao Y F, Liu H L, Zhao Z B, Lin Q X, Shi H W, et al. 2025. Hunyuan3D 2.5: Towards High-Fidelity 3D Assets Generation with Ultimate Details. arXiv preprint arXiv: 2506.16504. [DOI: 10.48550/arXiv.2506.16504]

Lei J H, Weng Y J, Harley A W, Guibas L and Daniilidis K. 2025. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 6165-6177. [DOI: 10.1109/CVPR52734.2025.00578]

Li H J, Yu H X, Li J M and Wu J J. 2024. ZeroHSI: Zero-Shot 4D Human-Scene Interaction by Video Generation. arXiv preprint arXiv: 2412.18600. [DOI: 10.48550/arXiv.2412.18600]

Li J H, Ma W J, Li X Y, Lou Y Z, Zhou G Y, Zhou X D, et al. 2025. CAD-Llama: leveraging large language models for computer-aided design parametric 3D model generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 18563-18573. [DOI: 10.1109/CVPR52734.2025.01730]

Li J M, Clegg A, Mottaghi R, Wu J J, Puig X and Liu C K. 2024. Controllable human-object interaction synthesis//Proceedings of European Conference on Computer Vision. Milan: Springer Nature Switzerland: 54-72. [DOI: 10.1007/978-3-031-72940-9_4]

Li P, Zheng W G D, Liu Y, Yu T, Li Y G, Qi X Q, et al. 2025. PSHuman: Photorealistic Single-image 3D Human Reconstruction using Cross-Scale Multiview Diffusion and Explicit Remeshing// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE. [DOI: 10.1109/CVPR52734.2025.01492]

Li Q X, Deng Y, Liang Y B, Luo L, Zhou L, Yao C T, et al. 2025. Scalable Vision-Language-Action Model Pretraining for Robotic Manipulation with Real-Life Human Activity Videos. arXiv preprint arXiv: 2510.21571. [DOI: 10.48550/arXiv.2510.21571]

Li X Y, Lou Y Z, Song Y and Zhou X D. 2025. Mamba-cad: State space model for 3D computer-aided design generative modeling//Proceedings of the AAAI Conference on Artificial Intelligence. Philadel-

phia: AAAI Press.[DOI: 10.1609/aaai.v39i5.32531]

Li Z, Zheng Z R, Wang L Z and Liu Y B. 2024. Animatable Gaussians: Learning Pose-Dependent Gaussian Maps for High-Fidelity Human Avatar Modeling//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 19711-19722.[DOI: 10.1109/CVPR52733.2024.01864]

Li Z H, Wang Y F, Zheng H L, Luo Y H and Wen B H. 2025. Sparc3D: Sparse Representation and Construction for High-Resolution 3D Shapes Modeling. arXiv preprint arXiv:2505.14521.[DOI: 10.48550/arXiv.2505.14521]

Lin H T, Chen S L, Liew J H, Chen D Y, Li Z Y, Shi G, et al. 2025. Depth Anything 3: Recovering the Visual Space from Any Views. arXiv preprint arXiv: 2511.10647.[DOI: 10.48550/arXiv. 2511.10647]

Liu Y B, Su H, Gao L, Yi L, Wang H, Liao Y Y, et al. 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,等. 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]

Long L, He Y C, Ye W T, Pan Y Y, Lin Y, Li H, et al. 2025. Seeing, Listening, Remembering, and Reasoning: A Multimodal Agent with Long-Term Memory. arXiv preprint arXiv:2508.09736.[DOI: 10.48550/arXiv.2508.09736]

Long X X, Cheng X J, Zhu H, Zhang P J, Liu H M, Li J, et al. 2021. Recent progress in 3D vision. Journal of Image and Graphics, 26(6): 1389-1428 (龙霄潇, 程新景, 朱昊, 张朋举, 刘浩敏, 李俊,等. 2021. 三维视觉前沿进展. 中国图象图形学报, 26(6): 1389-1428)[DOI: 10.11834/jig.210043]

Long X X, Guo Y C, Lin C, Liu Y, Dou Z Y, Liu L J, et al. 2024. Wonder3d: Single image to 3D using cross-domain diffusion//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 9970-9980.[DOI: 10.1109/CVPR52733.2024.00951]

Loper M, Mahmood N, Romero J, Pons-Moll G and Black M J. 2015. SMPL: A skinned multi-person linear model//Proceedings of the ACM SIGGRAPH Asia. Kobe: ACM.[DOI: 10.1145/2816795.2818013]

Low C W, Wang W M and Katyal C. 2025. Ovi: Twin Backbone Cross-Modal Fusion for Audio-Video Generation. arXiv preprint arXiv: 2510.01284.[DOI: 10.48550/arXiv.2510.01284]

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. Communications of the ACM, 65(1): 99-106.[DOI: 10.1145/3503250]

Mir A, Wang J, Guler R A, Guo C, Pons-Moll G and Zhou B. 2025. AHA! Animating Human Avatars in Diverse Scenes with Gaussian Splatting. arXiv preprint arXiv: 2511.09827.[DOI: 10.48550/arXiv.2511.09827]

Mu Y, Chen T X, Chen Z X, Peng S J, Lan Z Q, Gao Z Yet al. 2025. RoboTwin: Dual-Arm Robot Benchmark with Generative Digital Twins//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 27649-27660.[DOI: 10.1109/CVPR52734.2025.02575]

Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. 2023. DINOv2: Learning Robust Visual Features without Supervision. arXiv preprint arXiv: 2304.07193.[DOI: 10.48550/arXiv.2304.07193]

Pang Y X, Liu J J, Tan L F, Zhang Y, Gao F, Deng X, et al. 2025. MAViD: A Multimodal Framework for Audio-Visual Dialogue Understanding and Generation. arXiv preprint arXiv: 2512.03034.[DOI: 10.48550/arXiv.2512.03034]

Peng C, Sun J X, Chen Y S, Su Z Q and Liu Y B. 2025. Parametric Gaussian Human Model: Generalizable Prior for Efficient and Realistic Human Avatar Modeling. arXiv preprint arXiv: 2506.06645.[DOI: 10.48550/arXiv.2506.06645]

Poole B, Jain A, Barron J T and Mildenhall B. 2022. DreamFusion: Text-to-3D using 2D Diffusion. arXiv preprint arXiv: 2209.14988.[DOI: 10.48550/arXiv.2209.14988]

Qiu L T, Gu X D, Li P H, Zuo Q, Shen W C, Zhang J F, et al. 2025. LHM: Large Animatable Human Reconstruction Model from a Single Image in Seconds. arXiv preprint arXiv: 2503.10625.[DOI: 10.48550/arXiv.2503.10625]

Schönberger J L and Frahm J M. 2016. Structure-from-Motion Revisited//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 4104-4113.[DOI: 10.1109/CVPR.2016.445]

Song G X, Xu H Y, Zhao X C, Xie Y, Gu T P, Li Z N, et al. 2025. X-UniMotion: Animating Human Images with Expressive, Unified and Identity-Agnostic Motion Latents//Proceedings of the SIGGRAPH Asia, Tokyo: ACM.[DOI: 10.1145/3757377.3763952]

Tang J X, Lu R J, Li Z H, Hao Z K, Li X, Wei F Y, et al. 2025. Efficient Part-level 3D Object Generation via Dual Volume Packing. arXiv preprint arXiv: 2506.09980.[DOI: 10.48550/arXiv. 2506.09980]

Tian L R, Wang Q, Zhang B and Bo L F. 2024. EMO: Emote Portrait Alive - Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions//Proceedings of the European Conference on Computer Vision. Milan: Springer.[DOI: 10.1007/978-3-031-73010-8_15]

Tochilkin D, Pankratz D, Liu Z X, Huang Z X, Letts A, Li Y G, et al. 2024. TripoSR: Fast 3D Object Reconstruction from a Single Image. arXiv preprint arXiv: 2403.02151.[DOI: 10.48550/arXiv.2403.02151]

Tong W W, Guo H W, Ran D C, Chen J N, Lu J F, Wang K B, et al. 2025. InteractiveOmni: A Unified Omni-modal Model for Audio-Visual Multi-turn Dialogue. arXiv preprint arXiv: 2510.13747.

[DOI: 10.48550/arXiv.2510.13747]

Wang J H, Yuan Y F, Zheng R J, Lin Y T, Gao J, Chen L Z, et al. 2025. SpatialVID: A Large-Scale Video Dataset with Spatial Annotations. arXiv preprint arXiv: 2509.09676. [DOI: 10.48550/arXiv.2509.09676]

Wang J Y, Chen M H, Karaev N, Vedaldi A, Rupprecht C and Novotny D. 2025. VGGT: Visual Geometry Grounded Transformer// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 5294-5306. [DOI: 10.1109/CVPR52734.2025.00499]

Wang P, Liu L J, Liu Y, Theobalt C, Komura T, Wang W P, et al. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. Advances in Neural Information Processing Systems, 34: 27171-27183. [DOI: 10.48550/arXiv.2106.10689]

Wang P H, Zhang Z R, Wang L, Yao K X, Xie S Y, Yu J Y, et al. 2024. V^3: Viewing Volumetric Videos on Mobiles via Streamable 2D Dynamic Gaussians. ACM Transactions on Graphics (TOG), 43(6). [DOI: 10.1145/3687935]

Wang Q Q, Ye V, Gao H, Zeng W J, Austin J, Li Z Q, et al. 2025. Shape of Motion: 4D Reconstruction from a Single Video//Proceedings of the IEEE/CVF International Conference on Computer Vision. Hawaii: IEEE. [DOI: 10.48550/arXiv.2407.13764]

Wang S Z, Leroy V, Cabon Y, Chidlovskii B and Revaud J. 2024. DUSt3R: Geometric 3D Vision Made Easy//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 20697-20709. [DOI: 10.1109/CVPR52733.2024.01956]

Wang X Y, Zhang X Y, Zhu Y H, Guo Y C, Yuan X Y, Xiang L Y, et al. 2020. Panda: A Gigapixel-Level Human-Centric Video Dataset// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 3265-3275. [DOI: 10.1109/CVPR42600.2020.00333]

Wang Y F, Yang P S, Xu Z, Sun J M, Zhang Z H, Chen Y, et al. 2025. FreeTimeGS: Free Gaussian Primitives at Anytime Anywhere for Dynamic Scene Reconstruction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 21750-21760. [DOI: 10.1109/CVPR52734.2025.02026]

Wang Z J, Zhang P, Qi J W, Wang G Y, Ji C N, Xu S, et al. 2025. OmniTalker: One-shot Real-time Text-Driven Talking Audio-Video Generation With Multimodal Style Mimicking. arXiv preprint arXiv: 2504.02433. [DOI: 10.48550/arXiv.2504.02433]

Wu R D, Xiao C and Zheng C X. 2021. DeepCAD: A Deep Generative Network for Computer-Aided Design Models//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 6752-6762. [DOI: 10.1109/ICCV48922.2021.00670]

Xiang J F, Lv Z L, Xu S C, Deng Y, Wang R C, Zhang B W, et al. 2025. Structured 3D Latents for Scalable and Versatile 3D Generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 21469-21480. [DOI: 10.1109/CVPR52734.2025.02000]

Xie Y, Gu T P, Li Z N, Zhang C X, Song G X, Zhao X C, et al. 2025. X-Streamer: Unified Human World Modeling with Audiovisual Interaction. arXiv preprint arXiv: 2509.21574. [DOI: 10.48550/arXiv.2509.21574]

Xiu Y L, Ye Y F, Liu Z, Tzionas D and Black M J. 2024. PuzzleAvatar: Assembling 3D avatars from personal albums. ACM Transactions on Graphics (TOG), 43(6). [DOI: 10.1145/3687771]

Xu S C, Chen G J, Yang J L, Zhang Y Z, Deng Y, Lin S, et al. 2025. VASA-3D: Lifelike Audio-Driven Gaussian Head Avatars from a Single Image//Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS). [DOI: 10.48550/arXiv.2512.14677]

Xu X, Jayaraman P K, Lambourne J G, Willis K D D and Furukawa Y. 2023. Hierarchical Neural Coding for Controllable CAD Model Generation//Proceedings of the 40th International Conference on Machine Learning. Hawaii: Proceedings of Machine Learning Research. [DOI: 10.48550/arXiv.2307.00149]

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

Xu Z, Peng S D, Lin H T, He G Z, Sun J M, Shen Y J, et al. 2024. 4K4D: Real-Time 4D View Synthesis at 4K Resolution//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 20029-20040. [DOI: 10.1109/CVPR52733.2024.01893]

Xu Z, Xu Y H, Yu Z Y, Peng S D, Sun J M, Bao H J, et al. 2024. Representing long volumetric video with temporal Gaussian hierarchy. ACM Transactions on Graphics (TOG), 43(6). [DOI: 10.1145/3687919]

Yan W H, Ye S, Yang Z Y, Teng J Y, Dong Z H, Wen K R, et al. 2025. SCAIL: Towards Studio-Grade Character Animation via In-Context Learning of 3D-Consistent Pose Representations. arXiv preprint arXiv: 2512.05905. [DOI: 10.48550/arXiv.2512.05905]

Yang J N, Sax A, Liang K J, Henaff M, Tang H, Cao A, et al. 2025. Fast3R: Towards 3D Reconstruction of 1000+ Images in One Forward Pass//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE. [DOI: 10.1109/CVPR52734.2025.02042]

Yang Y H, Zhou Y F, Guo Y C, Zou Z X, Huang Y K, Liu Y T, et al. 2025. OmniPart: Part-Aware 3D Generation with Semantic Decoupling and Structural Cohesion//Proceedings of the SIGGRAPH Asia. Tokyo: ACM. [DOI: 10.1145/3757377.3763872]

Yao K X, Zhang L W, Yan X H, Zeng Y, Zhang Q X, Yang W, et al. 2025. CAST: Component-aligned 3D scene reconstruction from an

RGB image. ACM Transactions on Graphics，44(4).［DOI：10.1145/3730841］

Yao Y，Luo Z X，Li S W，Zhang J Y，Ren Y F，Zhou L，et al. 2020. BlendedMVS：A Large-Scale Dataset for Generalized Multi-View Stereo Networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle：IEEE：1787-1796.［DOI：10.1109/CVPR42600.2020.00186］

Yu Z H，Peng S Y，Niemeyer M，Sattler T and Geiger A. 2022. Monosdf：Exploring monocular geometric cues for neural implicit surface reconstruction. Advances in Neural Information Processing Systems，35.［DOI：10.48550/arXiv.2206.00665］

Zhan Y Y，Shao T J，Yang Y and Zhou K. 2025. Real-time High-fidelity Gaussian Human Avatars with Position-based Interpolation of Spatially Distributed MLPs//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville：IEEE.［DOI：10.1109/CVPR52734.2025.02449］

Zhang J W，Chu L，Li J H，Zang Z Y，Li C，Li X，et al. 2025. Bringing Your Portrait to 3D Presence. arXiv preprint arXiv：2511.22553.［DOI：10.48550/arXiv.2511.22553］

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).［DOI：10.1145/3592442］

Zhang L W，Zhang Q X，Jiang H R，Bai Y N，Yang W，Xu L，et al. 2025. BANG：Dividing 3D Assets via Generative Exploded Dynamics. ACM Transactions on Graphics，44(4).［DOI：10.1145/3730840］

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

Zhang W X，Cun X D，Wang X，Zhang Y，Shen X，Guo Y，et al. 2023. Sadtalker：Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver：IEEE：8652-8661.［DOI：10.1109/CVPR52729.2023.00836］

Zhao Q C，Long P Y，Zhang Q X，Qin D F，Liang H，Zhang L W，et al. 2024. Media2Face：Co-Speech Facial Animation Generation with Multi-Modality Guidance//Proceedings of the ACM SIGGRAPH. Denver：ACM.［DOI：10.1145/3641519.3657413］

Zhu S H，Chen J L，Dai Z Z，Su Q K，Xu Y H，Cao X，et al. 2024. Champ：Controllable and Consistent Human Image Animation with 3D Parametric Guidance//Proceedings of the European Conference on Computer Vision. Milan：Springer Nature Switzerland.［DOI：10.1007/978-3-031-73001-6_9］

Zhuang Y Y，Lv J X，Wen H，Shuai Q，Zeng A L，Zhu H，et al. 2025. Idol：Instant photorealistic 3D human creation from a single image//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville：IEEE：26308-26319.［DOI：10.1109/CVPR52734.2025.02450］

作者简介

修宇亮，男，西湖大学工学院助理教授。研究方向包括虚拟数字人、运动捕捉、角色动画、计算机视觉及图形学、可控多模态内容生成等 E-mail：xiuyuliang@westlake.edu.cn

徐震，男，浙江大学博士研究生。研究方向包括四维重建，视频世界模型等 E-mail：zhenx@zju.edu.cn

连宙辉，男，北京大学王选计算机研究所副教授。研究方向包括计算机图形学、计算机视觉、人工智能等 E-mail：lianzhouhui@pku.edu.cn

彭思达，男，浙江大学软件学院研究员。研究方向包括三维计算机视觉等 E-mail：pengsida@zju.edu.cn

姚遥，男，南京大学智能科学与技术学院准聘副教授。研究方向包括三维重建、可微渲染及三维内容生成等 E-mail：yaoynju@gmail.com

陈安沛，男，西湖大学工学院助理教授。研究方向包括计算机图形学、计算机视觉等 E-mail：chenanpei@westlake.edu.cn

黄经纬，男，腾讯混元3D技术专家。研究方向包括计算机图形学、计算机视觉等 E-mail：jingweih@stanford.edu

张邦，男，阿里巴巴通义实验室算法科学家。研究方向包括数字人 AIGC 等 E-mail：zhangbang.zb@alibaba-inc.com

许岚，男，上海科技大学信息科学与技术学院助理教授、研究员。研究方向包括计算机视觉、计算机图像学、机器学习等 E-mail：xulan1@shanghaitech.edu.cn

徐枫，男，清华大学软件学院副教授。研究方向包括三维重建、运动捕捉、人脸建模与动画生成等 E-mail：feng-xu@tsinghua.edu.cn

章国锋，男，浙江大学计算机科学与技术学院教授。研究方向包括计算机视觉、增强现实等 E-mail：zhangguofeng@cad.zju.edu.cn

许威威，男，浙江大学计算机科学与技术学院长聘教授。研究方向包括三维视觉、物理仿真、数字孪生和虚拟现实等 E-mail：xww@ cad.zju.edu.cn

虞晶怡，男，上海科技大学副校长、信息科学与技术学院讲席教授。研究方向包括计算机视觉、计算机图形学、计算成像和摄影、医学图像处理、生物信息学等 E-mail：yujingyi@shanghaitech.edu.cn

刘利刚，男，中国科学技术大学数学科学学院教授。研究方向包括几何建模与处理、图像与视频处理、计算几何等 E-mail：lgliu@ustc.edu.cn

陈宝权，男，北京大学智能学院教授。研究方向包括计算机图形与可视化等 E-mail：baoquan@pku.edu.cn

刘烨斌，男，清华大学自动化系教授。研究方向包括三维视觉、数字人体技术、计算摄像等 E-mail：liuyebin@tsinghua.edu.

cn