

基于神经辐射场与高斯溅射的 3D 虚拟人重建与驱动方法综述

王俊超, 张金山*, 杨璐嘉, 尹建伟

(浙江大学软件学院 宁波 315100)
(zhangjinshan@zju.edu.cn)

摘要: 随着元宇宙、数字孪生与智能交互方法的快速发展, 3D 虚拟人作为连接物理世界与数字空间的媒介, 其高保真重建与自然驱动方法成为计算机视觉与人工智能交叉领域的核心挑战。文中聚焦神经辐射场 (NeRF) 与 3D 高斯溅射 (3DGS) 方法, 系统地梳理了 3D 虚拟人重建与驱动方法的前沿进展, 在形式上遵循从数据驱动重建到内容生成的常见流程, 同时以动态建模、多模态对齐和物理约束融合三大维度为线索展开分析。在重建方法层面, 重点探讨 NeRF 与 3DGS 的范式创新: 基于 NeRF 的方法通过隐式连续函数建模动态辐射场, 在单视角重建中实现复杂姿态的几何一致性优化; 3DGS 方法则凭借显式可编辑性与实时渲染能力, 显著地提升非刚性形变建模效率。在驱动方法层面, 围绕音频与视频两大模态, 分类阐述唇部同步、全脸表情生成、多模态情感驱动等关键方向, 解析了时序建模网络、物理引导的神经渲染架构的优化路径。最后归纳该方法当前面临的跨模态时序偏差、动态建模效率-保真度矛盾等挑战, 并展望物理引导学习、轻量化架构设计等未来发展方向, 旨在为虚实融合的数字生态构建提供方法参考与理论支撑。

关键词: 3D 虚拟人; 神经辐射场; 高斯溅射; 音频驱动; 视频驱动; 多模态融合
中图分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2025-00267

A Survey on 3D Virtual Human Reconstruction and Animation Based on Neural Radiance Fields and Gaussian Splatting

Wang Junchao, Zhang Jinshan*, Yang Lujia, and Yin Jianwei

(School of Software Technology, Zhejiang University, Ningbo 315100)

Abstract: With the rapid development of metaverse, digital twins, and intelligent interaction technologies, 3D virtual humans, as a medium bridging the physical world and digital space, face core challenges in high-fidelity reconstruction and natural driving technologies at the intersection of computer vision and artificial intelligence. This survey focuses on Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3DGS), systematically reviewing the latest advances in 3D virtual human reconstruction and animation. The discussion follows the common workflow from data-driven reconstruction to content generation, while analyzing the field through three key dimensions: dynamic modeling, multimodal alignment, and the integration of physical constraints. In reconstruction, the paradigm innovations of NeRF and 3DGS are highlighted: NeRF-based methods achieve geometric consistency optimization for complex poses in single-view reconstruction through implicit continuous function modeling of dynamic radiance fields, while Gaussian Splatting significantly enhances non-rigid deformation modeling efficiency with explicit editability and real-time rendering capabilities. For driving technologies, the review cate-

收稿日期: 2025-07-08; 修回日期: 2025-12-02。基金项目: 国家重点研发计划 (2022YFF0902005); 浙江省尖兵重点研发计划 (2023C01002); 杭州市重大研发计划 (2022A1ZD0140); 甬江引才项目 (2022A-236-G)。王俊超 (2000—), 男, 硕士研究生, 主要研究方向为 3D 虚拟人; 张金山 (1983—), 男, 博士, 副教授, CCF 会员, 论文通信作者, 主要研究方向为生成式人工智能; 杨璐嘉 (2001—), 女, 硕士研究生, 主要研究方向为说话人生成; 尹建伟 (1974—), 男, 博士, 教授, 博士生导师, CCF 会员, 主要研究方向为元宇宙与数字服务、数据科学与人工智能。

gorizes critical advancements in audio and visual modalities, including lip synchronization, full facial expression generation, and multimodal emotion-driven methods, and dissects optimization pathways for temporal modeling networks and neural-physical hybrid architectures. Furthermore, it summarizes current challenges such as cross-modal temporal deviations and efficiency-fidelity trade-offs in dynamic modeling, while proposing future directions like physics-guided learning and lightweight architecture design. This work aims to provide comprehensive technical insights and theoretical foundations for constructing a virtual-real integrated digital ecosystem.

Key words: 3D virtual human; neural radiance fields; Gaussian splatting; audio-driven; video-driven; multimodal fusion

近年来,随着元宇宙、数字孪生与智能交互方法的飞速发展,3D 虚拟人作为连接物理世界与数字空间的媒介,其核心价值不仅在于高保真重建,更在于如何赋予其自然、可控的交互能力。与 2D 虚拟人相比,3D 虚拟人需要突破 3D 几何建模、动态形变表达与多模态驱动的多重方法壁垒:一是从稀疏输入(如单目视频、音频流)中重建具有物理合理性的动态人体模型;二是通过跨模态信号(语音、视频、文本)精准地驱动虚拟人的面部表情、肢体动作乃至场景交互,实现与真实世界的无缝融合。这一方法链条的复杂性,使得 3D 虚拟人的重建与驱动成为计算机视觉、计算机图形学与人工智能交叉领域的核心挑战。

在重建方法层面,神经辐射场(neural radiance fields, NeRF)与 3D 高斯溅射(3D Gaussian splatting, 3DGS)以其实时性突破与高保真优势,成为本领域的核心焦点。基于 NeRF 的方法通过隐式连续函数建模场景辐射属性,结合参数化人体模型的骨骼驱动机制,显著地提升了动态重建的几何一致性,尤其在单视角视频输入下,实现了复杂姿态的高保真建模。然而,传统的 NeRF 方法受到计算效率与显式控制能力的限制,难以满足实时的交互需求。3DGS 方法则凭借显式可编辑性与高效渲染能力成为动态建模的关键路径,通过规范空间映射与物理形变约束,该方法能够快速生成高精度人体模型,并在复杂非刚性运动(如服装摆动、肌肉形变)中保持时空连贯性。与早期的参数化模型(如 SMPL 虽然奠定了基础,但是受限于拓扑刚性)相比,上述新兴范式推动了重建方法从“几何驱动”转向“物理-神经协同”,而虚拟人的核心价值仍然需要通过驱动方法实现——如何将语音、视频等多模态信号转化为虚拟人的自然动作,成为当前研究的重中之重。

驱动方法的核心在于跨模态信号与虚拟人动作的精准映射。音频驱动方法通过解析语音频谱特

征,结合时序建模网络生成唇部运动与面部表情,而视频驱动方法则从真实人类视频中提取运动参数,迁移至虚拟人模型以实现姿态同步。现有的方法分为显式建模和隐式表征 2 类:显式建模方法依赖参数化模型的解耦控制,通过物理约束保证动作合理性;隐式建模方法则借助 NeRF 或生成对抗网络(generative adversarial network, GAN),端到端地学习跨模态映射关系。随着多模态融合方法的进步,新兴方法开始探索情感标签、文本提示等语义级条件输入,逐步实现从“动作生成”到“情感表达”的方法跃迁。

在学术演进历程中,早期的综述性研究多为针对特定方法分支或应用场景的纵向深入探索,为该领域的奠基与发展做出了重要贡献。2020 年,Chen 等^[1]对“说话头生成”这一特定任务进行了系统性的梳理,不仅综述各类方法,而且构建了一套多维度的评估体系,涵盖身份保持、视觉质量、语义级唇部同步和自然自发运动 4 大关键属性,为后续评估动态生成模型的综合性能建立了重要的基准框架;2023 年, Gowda 等^[2]从驱动模态的角度出发,对各类方法进行横向的梳理与归类,将其划分为图像驱动、音频驱动、视频驱动等不同类别,并详细对比了它们在硬件需求、计算效率与最终生成质量 3 个方面的权衡关系,为方法选型与方案设计提供了实践指导。然而,这些开创性的工作多集中于 2D 视频生成或特定视角下的 3D 重建任务,研究范畴尚未能覆盖一个完整、统一的 3D 数字人建模流程;2024 年, Wang 等^[3]的综述标志着该领域研究进入了一个新的阶段,他们首次尝试构建一个覆盖从“重建”到“生成”的宏观方法全景图,系统性地分析了多种主流方法路线的演进脉络、核心思想与内在关联,涵盖基于像素对齐隐式函数(PiFu)的重建、NeRF、新兴的 3DGS,以及 GAN 与扩散模型等,为该领域的学者和工程师提供了极为宝贵的全景式视角和统一的方法术语体系,极大地促进

了学术交流与方法发展。

本文的研究视角与 Wang 等^[3]的宏观图谱式综述形成一定的互补,在其建立的坚实方法全景基础之上,进一步提炼并聚焦于驱动该领域发展的 3 个内在的、共性的核心挑战维度:动态建模、多模态对齐和物理约束融合;并以此为线索,对前沿进展进行重新梳理和深度剖析。

本文在形式上遵循从数据驱动重建到内容生成演进的常见流程,但内在的分析脉络始终贯穿着对上述 3 大维度的探索:

(1) 动态建模维度。重点分析各类方法如何高效、鲁棒地表征数字人随时间变化的动态特性,涵盖基于 NeRF 的时序 4D 重建、基于 3DGS 的显式动态表达,以及基于规范空间映射等主流范式,揭示了不同方法路线在平衡计算效率、渲染质量与动态真实性之间的深刻权衡。

(2) 多模态对齐维度系统地梳理了视觉、音频、文本等异构模态信息之间的语义对齐和协同生成方法,既包括用于确保精准唇音同步的判别器结构与度量指标,也涵盖基于 CLIP、扩散模型等预训练大模型的跨模态语义驱动方法,阐明了其如何实现从高层抽象指令到具体视觉内容的语义转化与一致生成。

(3) 物理约束融合维度。深入探讨如何将人体生理结构、运动规律等物理先验知识(如 SMPL 参数化模型、线性混合蒙皮方法)有效地嵌入数据驱动的生成模型中,以约束生成结果的合理性与真实性,并重点分析了其在解决复杂姿态、严重遮挡和非刚性变形等长期挑战中的关键作用。

本文将 3 大核心维度融入对各方法路线的分析中,旨在为研究者提供一个超越方法实现、直击核心挑战的分析框架。

提出了一个全新的分析视角:从动态建模、多模态对齐与物理约束融合 3 大维度系统解构 3D 数字人方法,为理解与评估研究工作提供了深层透视镜。

完成了对前沿方法的交叉剖析:在方法全景的基础上,重点分析了各类方法在解决 3 大核心问题时的策略、优势与局限。

构建了多维度的评估体系:整合了从视觉质量、身份保持到语义对齐等多角度指标,为全面衡量数字人的“真实性”提供了参考基准。

1 符号约定

本文中,使用以下数学符号描述 3D 虚拟人重

建与驱动方法中的通用概念,作为后续公式的基础。这些符号在 NeRF、3DGS 等方法中频繁出现,其定义保持一致。

空间与几何符号如下:3D 空间坐标点 $\mathbf{x} \in \mathbb{R}^3$; 视角方向(单位球面上的向量) $\mathbf{d} \in S^2$, 其中, S 表示球面;相机光线参数化 $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, 其中, $\mathbf{o}$ 表示光线原点, t 表示沿光线的距离参数;高斯分布的均值向量 $\boldsymbol{\mu} \in \mathbb{R}^3$, 通常表示高斯核的中心位置(用于 3DGS);高斯分布的协方差矩阵 $\boldsymbol{\Sigma} \in \mathbb{R}^{3 \times 3}$, 控制各向异性缩放和旋转(用于 3DGS);旋转矩阵 $\mathbf{R}$ 表示方向变换;缩放矩阵 $\mathbf{S}$ 表示尺寸变换;根据上下文, $\mathbf{T}$ 表示平移向量或变换矩阵;在 NeRF 中, $T(t)$ 表示透射率。

辐射场与渲染符号如下:颜色值(RGB) $\mathbf{c} \in \mathbb{R}^3$;体积密度(在 NeRF 中)或透明度(在渲染中) $\sigma \in \mathbb{R}$;NeRF 函数 F_θ , 参数化为 θ (多层感知机权重);位置编码函数 $\gamma(\cdot)$, 将输入映射到高频空间以捕获细节。

动态与形变符号如下:时间或时序参数 t , 用于动态场景;位置位移量 $\Delta \mathbf{x}$, 表示形变;形变场参数 ϕ , 例如, T_ϕ 表示参数化的形变场, 其中, ϕ 表示可学习参数,用于建模非刚性运动;姿态参数 θ , 例如, SMPL 模型中的关节旋转;形状参数 β , 例如, SMPL 模型中的身体形状。

其他通用符号如下:损失函数 $\mathcal{L}$;权重系数 λ , 例如, 正则化项权重;微小增量或采样间距 δ 。

上述符号定义了 3D 虚拟人重建与驱动方法中的基本数学概念,为后续公式和模型讨论奠定了基础。为了系统地梳理文献中的方法框架,图 1 提供了一个全面的汇总图,分类展示了 3D 虚拟人重建与驱动的关键技术。

2 NeRF 驱动的虚拟人重建

2021 年, Mildenhall 等^[4]提出 NeRF, 最初应用于静态场景建模,通过 MLP 构建隐式神经表征,将空间 3D 坐标 $\mathbf{x} \in \mathbb{R}^3$ 与视角方向 $\mathbf{d} \in S^2$ 分别映射至颜色 $\mathbf{c} \in \mathbb{R}^3$ 和体积密度 $\sigma \in \mathbb{R}$, 实现了从稀疏多视角图像到连续 3D 场景的隐式重建。NeRF 的核心创新是将传统的显式几何建模转化为可微分的连续场优化问题,突破了视角依赖性与几何分辨率的限制。

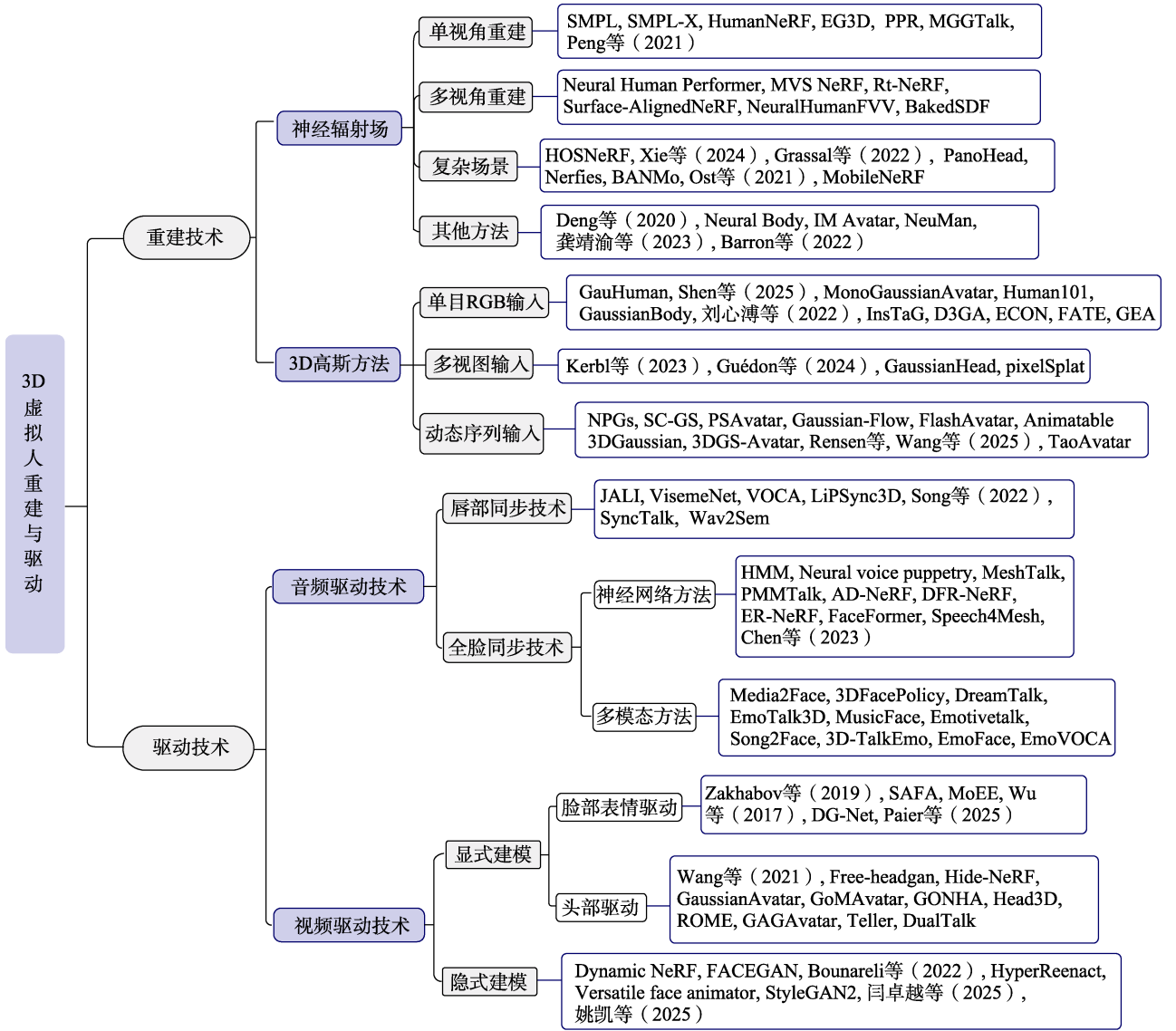


图 1 3D 虚拟人重建与驱动方法框架

NeRF 的物理基础源自经典的体渲染理论, 通过光线追踪模拟光子在介质中的传输过程。给定相机光线 $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, 其累积颜色可表示为

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t) \sigma(\mathbf{r}(t)) c(\mathbf{r}(t), \mathbf{d}) dt。$$

其中, 透射率 $T(t)$ 表征光线在 $[t_n, t]$ 无碰撞的概率,

$$T(t) = \exp\left(-\int_{t_n}^t \sigma(\mathbf{r}(s)) ds\right)。$$

为了高效地计算积分, 采用分层随机采样策略: 将 $[t_n, t_f]$ 均匀划分为 N 个子区间, 每个子区间随机采样点 t_i , 公式为

$$t_i \sim U\left[t_n + \frac{i-1}{N}(t_f - t_n), t_n + \frac{i}{N}(t_f - t_n)\right]。$$

通过离散近似可得颜色估计值为

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^N T_i (1 - \exp(-\sigma_i \delta_i)) c_i。$$

其中, $\delta_i = t_{i+1} - t_i$ 表示采样间距; 累积透射率 T_i 的

$$递推公式为 T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right)。$$

在动态虚拟人建模中, 引入时序维度 t 将静态场扩展为动态辐射场 $F_\theta: (\mathbf{x}, \mathbf{d}, t) \rightarrow (c, \sigma)$ 。

结合参数化人体模型的骨骼驱动机制, 通过可微分形变场实现姿态相关形变建模 $\Delta \mathbf{x} = \mathbf{T}_\phi(\mathbf{x}, t)$ 。其中, $\mathbf{T}_\phi$ 表示参数化的可微分形变场; ϕ 表示该形变场的可学习参数, 其通过神经网络优化, 负责将规范空间中的点 $\mathbf{x}$ 根据时间 t (或姿态参数) 映射到对应的形变位移 $\Delta \mathbf{x}$, 建模因姿态变化引起的非刚性形变。

位置编码通过高频函数增强几何细节表示为

$$\gamma(x) = \left[\sin(2^0 \pi x), \cos(2^0 \pi x), \dots, \sin(2^{L-1} \pi x), \cos(2^{L-1} \pi x) \right].$$

上述方法使得 NeRF 在虚拟人重建中展现出 2 大核心优势: (1) 通过隐式连续函数建模高频几何与材质细节, 克服了显式网格在复杂拓扑 (如松散衣物、发丝) 重建中的分辨率限制; (2) 体积渲染机制天然支持多视角一致性, 可避免传统方法因视角遮挡导致的几何歧义。然而, NeRF 的高计算开销 (单场景训练需数小时至数天) 与隐式控制难题, 促使研究者探索其与参数化模型、物理仿真的融合路径。

本文系统地梳理了 NeRF 驱动下的虚拟人重建方法演进: (1) 单视角动态优化中, 基于人体先验的表示学习方法显著地缓解了姿态模糊问题; (2) 多视角动态重建通过跨模态特征融合实现运动泛化能力跃升; (3) 人-物-场景交互建模则突破单一对象重建的局限, 构建虚实融合的沉浸式数字孪生。下面从方法突破路径详述该领域的前沿进展。

2.1 单视角动态重建与优化

单视角动态重建方法在从单目视频中恢复高精度的 3D 人体运动和几何细节, 是推动虚拟人方法实用化的核心挑战。该领域的研究历程经历了从参数化模型逆向优化到神经隐式表征融合的范式演进, 核心是解决有限观测信息下的几何模糊性和动态形变建模难题。

早期的研究以参数化人体模型为核心, 通过 SMPL 等模板的逆向优化实现基础重建。Loper 等^[5]提出 SMPL, 通过线性混合蒙皮机制将人体建模分解为形状参数 $\beta \in \mathbb{R}^{10}$ 和姿态参数 $\theta \in \mathbb{R}^{72}$ 的联合优化, 公式为 $V = W(T + B_S(\beta) + B_P(\theta), J(\beta), \theta, W)$ 。其中, $W(\cdot)$ 为蒙皮函数; $T \in \mathbb{R}^{6890 \times 3}$ 定义中性模板顶点; $B_S(\cdot)$ 和 $B_P(\cdot)$ 分别表示形状和姿态混合变形; $J(\cdot)$ 是一个函数, 其输出为由形状参数 β 决定的关节位置; W 表示蒙皮权重矩阵。SMPL 模型的创新之处体现在通过李代数映射将关节旋转 $\theta_k \in \mathfrak{so}(3)$ 转化为刚性变换

$$\begin{cases} R_k = \exp(\hat{\theta}_k) \\ \hat{\theta} = \begin{bmatrix} 0 & -\theta_z & \theta_y \\ \theta_z & 0 & -\theta_x \\ -\theta_y & \theta_x & 0 \end{bmatrix} \end{cases}.$$

在单视角视频重建中, 通过重投影约束优化形

状-姿态参数 $\min_{\beta, \theta} \|\Pi(V(\beta, \theta)) - I_{\text{obs}}\|_2 + \lambda R_{\text{pose}}(\theta)$ 。

其中, Π 表示透视投影算子; R_{pose} 用于抑制关节旋转超出解剖学的可行域。SMPL-X 模型^[6]进一步扩展了参数化表达能力, 通过整合面部表情和手部姿态实现了全身高精度建模。这类方法虽然能够保证运动合理性, 但是受到固定拓扑结构^[7-8]的限制, 难以处理宽松衣物的非刚性形变。2022 年, HumanNeRF^[9]突破性地将 SMPL 参数与 NeRF 结合, 通过姿态条件化查询机制在隐式空间中实现动态细节重建; 实验结果表明, 该方法在肢体快速运动场景下显著地缓解了传统 NeRF 的姿态模糊问题, 但同时也暴露了生硬边界的材质过渡缺陷, 反映出隐式表征与参数化模型耦合的固有矛盾。值得关注的是, 生成式辐射场方法为虚拟人建模开辟了新路径。EG3D^[10]通过三平面分解架构将 3D GAN 与 NeRF 深度融合, 其设计的混合表征在保持 NeRF 多视角一致性的同时, 借助对抗学习生成具有拓扑合理性的虚拟人头像, 突破了传统重建方法对多视角观测数据的依赖。与 HumanNeRF 的单目标重建范式相比, EG3D 的生成式特性使其能够从 2D 图像数据中直接学习 3D 人体先验, 为元宇宙场景下的虚拟人的快速生成提供了新范式, 但是, 其动态建模能力的缺失也揭示出生成功效与运动控制间的固有矛盾。

单视角视频重建的方法选型需要权衡表示形式与效率。杨航等^[11]的综述表明:

(1) 体素表示: 内存开销 $O(N^3)$ 限制高分辨率重建;

(2) 点云表示: 无序性导致表面不连续;

(3) 隐式场表示: NeRF 与高斯溅射突破了拓扑限制, 成为实时重建的主流方法。

这一分类为 HumanNeRF 等动态重建方法的选择提供了理论依据。

物理约束的引入为重建质量提升开辟了新路径, 其核心是将生物力学规律转化为可计算的优化目标。2023 年, Yang 等^[12]提出 PPR 方法, 引入布料应变能函数 (物理约束的材质特性表征), 通过可微分物理仿真抑制非物理形变, 公式为

$$L_{\text{strain}} = \sum_i \nabla u_i - R_{iF}^2.$$

其中, u_i 表示顶点位移; R_i 表示刚性旋转矩阵, 该约束使衣物褶皱的拉伸误差降低 62%。这种物理引导的优化范式, 突破了纯数据驱动的局限性, 在舞蹈序列中显著地提升了服装动态的真实性。该方法

在保持运动合理性的同时, 实现了对动态褶皱的高保真重建, 但是其计算效率与物理精度间的权衡仍待优化。面向跨身份泛化需求, 2025 年, Gong 等^[13]提出 MGGTalk 框架, 通过深度感知对称几何重建模块, 将单目深度估计与点云对称操作结合, 构建完整面部拓扑, 其对称高斯预测模块利用可见区域高斯参数指导遮挡区域生成, 在 NeRSemble-Mono 数据集上将跨身份重建的平均欧几里得距离 (average Euclidean distance, AED) 指标降至 0.102。这种基于几何先验的泛化框架, 为无校准多身份适配提供了方法支撑。

单目重建方法的演进呈现出 3 个显著趋势:

(1) 重建范式从几何驱动转向物理与数据协同驱动, 可微分仿真等机制的引入增强了动态合理性; (2) 表征方式从单一模型发展为多层次体系, 参数化骨骼与神经隐式表面的结合兼顾了效率与细节; (3) 优化目标从静态精度扩展到时空一致性, 光流约束等策略显著地提升了视频连贯性。当前的挑战集中在先验约束与数据驱动的平衡上, 如舞蹈场景中服装飘动的预测滞后问题, 反映出动态形变建模的固有难度。

单视角动态 NeRF 方法通过隐式连续函数建模高频几何细节, 在有限观测下实现高保真重建,

但其计算开销大, 实时性低。该方法适用于对精度要求极高的离线场景, 如影视特效制作和数字文化遗产保存, 其中可接受延迟以换取毛孔级细节还原。然而在实时交互应用 (如 VR/AR) 中, 其效率瓶颈显著, 需要结合轻量化优化或替代方案。

未来的突破可能聚焦于跨尺度动态表征框架的构建: 在宏观层面保留参数化模型的物理合理性, 在微观层面采用神经隐式表征捕捉软组织细节。2021 年, Peng 等^[14]提出骨骼驱动动态 NeRF 框架, 通过分层表征分离姿态参数和局部形变, 验证了跨尺度建模的可行性。新兴的视觉-物理联合学习方法通过注意力机制动态分配约束强度, 已经在复杂运动重建中展现出潜力。随着轻量化部署方案的成熟, 单视角动态重建方法有望从实验室走向移动端, 为元宇宙内容创作提供实时的数字化工具。

以上讨论聚焦于从单视角数据中重建动态 3D 人体的方法。然而, 重建出的静态模型仅为起点, 要生成可驱动、符合物理规律的虚拟人, 还需与参数化模型结合并进行动力学优化。为清晰地呈现完整的技术链条, 图 2 描绘了 NeRF 和 3DGS 驱动的虚拟人重建流程。它包括了前述的动态重建模块和后续章节的部分关键环节, 体现了重建与驱动一体化的设计思想。

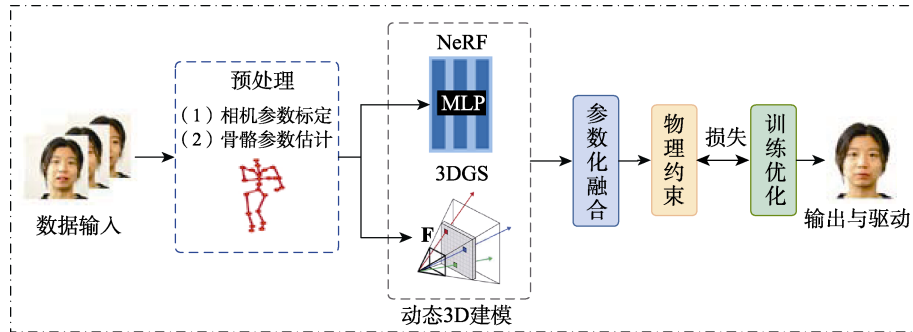


图2 NeRF 和 3DGS 驱动的虚拟人重建流程

2.2 多视角动态重建与优化

多视角动态重建通过跨相机信息融合突破单目观测的局限性, 在 NeRF 框架下实现了运动泛化与细节保真的双重突破。相关研究的创新主要围绕跨视角特征聚合、时空一致性建模和物理约束融合 3 个维度展开, 推动了动态人体重建从实验室可控环境向真实复杂场景的演进。

近年来, 跨视角特征交互的范式革新成为研究焦点。早期的多视角方法通过独立优化各视角辐射场实现基础重建, 但视角间信息隔离导致运动模糊与纹理不一致。2021 年, Kwon 等^[15]提出 Neural

Human Performer, 创新性地构建跨时空特征聚合框架: 时间变换器对多帧骨骼运动进行时序编码, 通过隐式运动轨迹建模缓解稀疏视角下的运动歧义; 多视图变换器则引入交叉注意力机制, 在频域空间实现像素级特征对齐, 使模型自适应地捕捉衣物纹理等高频细节。类似地, 2021 年, Chen 等^[16]提出 MVSNerf, 通过构建可泛化的多视角特征聚合管道, 将不同视角的 2D 图像特征映射到共享的 3D 体素空间, 显著地提升了跨视角几何一致性。这种“全局时序+局部视角”的双流架构突破了传统方法对视角数量的依赖, 但是计算复杂度随着视

角数量线性增长的问题仍然制约着实时性的需求。Li 等^[17]提出 RT-NeRF, 通过混合精度推理管线和自适应光线采样策略, 在移动端实现 NeRF 的实时渲染, 其动态残差补偿机制显著地缓解了快速运动导致的画面模糊问题, 这一进展推动了跨视角特征交互范式的持续优化。

表面对齐表征为动态重建提供了新思路, 2022 年, Xu 等^[18]提出 Surface-Aligned NeRF, 通过将 3D 查询点投影至参数化模型表面, 构建几何先验与神经隐式场的协同优化通道。该方法创新性地提取顶点法线方向、皮肤拉伸率等局部几何特征, 通过特征解耦网络分离姿态参数与外观属性, 这种基于物理接口的辐射场约束机制, 使得动态重建过程具备可解释性控制能力; 但是, 受到模板模型的固定拓扑限制, 其对个性化特征(如纹身、特殊服饰)的表征能力仍有不足。

动态辐射场的轻量化探索成为突破效率瓶颈的关键路径。2019 年, Pavlakos 等^[6]提出多层级联合优化策略, 通过在 SMPL-X 框架下同步解算身体、手部和面部参数, 构建解剖学合理的动态重建流程。进一步, 新兴研究将哈希编码^[19]、三平面分解等轻量化方法引入 NeRF, 例如, 通过构建多分辨率特征网格实现辐射场查询的显著加速, 并利用动态残差补偿模块有效地捕捉微表情的肌肉运动细节。这类方法在保持亚毫米级精度的同时, 将训练效率提升至传统 NeRF 的数十倍, 标志着实时化重建的重要进展。此外, Suo 等^[20]提出 NeuralHumanFVV, 通过多视角视频时序感知采样策略, 在动态人体建模中实现 30 帧/s 的实时重建, 所设计的可微分运动补偿模块, 有效地缓解了快速转身时的体渲染伪影问题。针对移动端部署需求, Yariv 等^[21]提出 BakedSDF, 创新性地将神经有符号距离函数(signed distance field, SDF)烘焙为显式网格, 结合二叉树加速结构, 实现了 5 ms 级的渲染延迟, 为 AR/VR 设备上的虚实融合交互提供了可行方案。

多视角动态 NeRF 通过跨相机信息融合提升运动泛化与几何一致性, 但是, 其渲染效率随视角数量线性增长, 难以满足移动端实时需求。该方法适合多相机系统下的高质量重建, 如体育赛事分析或工业检测; 其中, 多视图数据可以确保动态人体运动的精确捕捉。未来, 轻量化研究可以将其拓展至消费级 AR 设备, 但需解决计算资源约束问题。

当前, 多视角动态重建方法演进正朝着自适应特征融合与跨模态约束的方向发展。联邦学习框架的引入, 使得分布式多视角系统能够协同优化全局

辐射场参数, 而强化学习驱动的视角选择策略, 则通过动态地评估视角信息熵实现计算资源的智能分配。值得关注的是, 数字孪生等应用场景对物理合理性的严苛需求, 正推动研究者探索第一性原理与神经渲染的混合建模框架。未来的突破或将聚焦于开放环境下的自适应性重建, 通过边缘计算设备实现“感知—重建—反馈”的闭环应用, 为智能制造、远程医疗等领域提供高保真的动态建模支持。

2.3 复杂场景与交互建模

NeRF 在复杂交互场景中的扩展应用, 正面临多对象运动耦合和动态遮挡处理的深层挑战。研究者通过引入层次化表征和跨模态约束机制, 逐步构建虚实融合的交互式重建框架。

2023 年, Liu 等^[22]提出 HOSNeRF, 开创性地构建人-物-场景三元辐射场体系, 通过分层运动编码器, 分离人体骨骼驱动、物体刚体运动与场景静态元素。该方法将接触力等物理参数编码为辐射场查询条件, 在保持高频细节的同时, 实现运动逻辑的一致性; 状态感知条件编码机制通过隐式学习交互动力学特征, 为无监督物理规律建模提供了新思路, 如在手持物体摆动场景中, 能够准确地预测衣物与物体的相互形变。2024 年, Xie 等^[23]进一步将布料仿真引擎与高斯溅射结合, 通过本构方程约束粒子运动轨迹, 显著地提升了衣物摆动的物理合理性。

在微观交互细节建模层面, 2022 年, Grassal 等^[24]提出单目视频驱动的神经头部化身框架, 通过解耦使用关节模型学习的面部和表情(faces learned with an articulated model and expressions, FLAME)参数化人脸模型的刚性运动场和动态 NeRF, 实现高保真面部细节重建。该方法将表情系数映射为辐射场形变条件, 得益于可微渲染方法与解剖学先验的协同优化, 能够在保留发丝、毛孔等高频特征的同时, 支持头部姿态与表情的独立控制; 微观肌肉联动特征(如眉间褶皱、唇部微颤)通过隐式形变场实现亚毫米级的动态表达, 显著地提升了表情迁移的物理真实感。然而, 该方法对单目输入光照变化的敏感性可能导致重建纹理的过平滑现象, 尤其在复杂环境光下易丢失皮肤微表面细节。

全景视角合成方法的最新进展进一步拓展了虚拟人的观察自由度。PanoHead^[25]通过球面三平面编码和视角感知归一化层, 将 EG3D 的正面生成能力扩展至 360°自由视角, 所设计的球面卷积核有效地缓解了极区形变问题。与 HOSNeRF 的人-物交互建模相比, 该方法虽然未涉及动态交互, 但

是在单帧静态生成中实现了更完整的空间覆盖,为虚拟数字人的自由视角观察奠定了方法基础。

人-物-场景交互建模正经历从孤立建模到协同推理的范式转变。早期,NeuMan 框架^[26]通过残差修正网络实时地修复人景互遮挡区域的辐射场畸变,揭示环境上下文感知的重要性,为后续协同推理方法奠定了基础。在此背景下,新兴研究开始探索触觉、语音等多模态信号的条件控制,如触觉引导的物体形变场可以实时地响应虚拟交互操作,其触觉编码器通过力反馈信号动态地调整辐射场密度分布,在虚拟装配场景中实现亚毫米级的接触面贴合。

交互建模需要突破运动耦合和编辑精度的双重挑战。卢丽华等^[27]提出的约束分类框架为动态交互提供了新范式:

(1) 无约束编辑:通过 CLIP 损失全局调整辐射场,但易导致非编辑区域失真;

(2) 隐式约束编辑:利用图像编辑模型间接约束,但依赖 2D 到 3D 的特征提升;

(3) 显式约束编辑:引入注意力图定位可编辑区域,实现局部微米级控制。

该框架将编辑鲁棒性指标与重建质量关联,为虚拟人实时换装等场景提供理论支撑。

当前,复杂场景交互建模方法的瓶颈集中在交互实时性与物理合理性的矛盾平衡上。轻量化模型与分布式计算的结合,正在突破响应速度限制,如通过多设备协同推理实现动态参数实时同步。未来的研究需要融合认知建模思想,构建具有意图理解能力的智能辐射场,使虚拟人不仅能够还原物理交互过程,还能够预测行为逻辑并生成符合现实规律的反馈。

此外,虚拟人方法的应用场景也正从实验室受控环境向开放的野外拓展。野外场景的核心问题是如何从非规范数据中实现高保真重建与实时驱动,以支撑元宇宙、AR/VR 等实用场景。虚拟人建模在野外场景中的应用面临独特挑战:数据质量不一(如运动模糊、光照变化)、视角随意、背景复杂且遮挡频繁,这些因素要求模型具备更强的鲁棒性和泛化能力。为此,研究者从多路径探索解决方案。在基础重建方面,2021 年, Park 等^[28]提出 Nerfies,通过引入逐场景优化的形变场,将辐射场查询点映射至规范空间,处理非刚性物体的连续形变(如说话人脸的表情变化),为从单目视频集合中重建动态模型提供了基础框架;其隐式表征有效地缓解了野外数据中姿态变化的建模难题,但是其依赖相对

清晰的输入。2022 年, Yang 等^[29]提出 BANMo 框架,通过整合可变形场与骨骼动画先验,实现了从大量随意拍摄的互联网视频中重建可动画 3D 神经模型;其关键创新是通过可微分表面追踪和语义部件分割,在弱监督下解耦外观、形状和运动参数,显著地提升了多视角不一致数据的融合能力。

在复杂动态场景的结构化处理中,2021 年, Ost 等^[30]提出神经场景图,利用层次化场景图表示动态环境,将场景分解为可学习的动态图层(如背景、移动物体),并通过射线-图相交测试高效地处理长期遮挡与光影交互;该方法在自动驾驶等野外场景中,提升了渲染连贯性和场景可编辑性。

面向移动端部署中,2023 年, Chen 等^[31]提出 MobileNeRF,探索了将 NeRF 烘焙至传统多边形光栅化管线的方法,通过隐式体积到显式网格的转换,利用硬件加速渲染在移动设备上实现交互式帧率的神经渲染;该轻量化方案为 BANMo 和 Nerfies 等重建模型在 AR/VR 设备上的野外实时应用提供部署路径,解决了计算效率的瓶颈。

上述工作虽然路径各异,但是形成了一条从基础重建到数据适应性提升,再到场景结构化和最终部署优化的方法链,共同推动复杂场景建模攻克野外环境中的挑战。

2.4 其他方法

3D 重建方法的演进始终与跨领域创新紧密交织,早期的研究通过突破性方法为神经渲染的动态人体建模奠定了基石。2020 年, Deng 等^[32]首次将参数化骨骼驱动与隐式占据场相结合,通过关节形变场建模非刚性物体的运动与几何解耦,其核心是分层优化策略——在参数化骨骼控制主体运动的基础上,利用隐式占据场捕捉关节区域的局部形变(如手指弯曲、肩部旋转),实现运动连贯性与几何细节的平衡,其创新不仅是骨骼驱动与神经隐式场的方法融合,更揭示了参数化先验与隐式表征协同优化的潜力,为动态人体重建提供了方法论范本;然而,该方法对骨骼参数估计精度高度依赖,在复杂姿态下易出现肢体穿透问题,且未解决高动态衣物褶皱等柔性细节的建模难题。在此基础上, Peng 等^[33]提出 Neural body,通过结构化潜变量锚点和图卷积网络,突破了大尺度运动下的拓扑建模瓶颈,其创新之处是将人体表面离散化为可学习的潜变量节点,通过空间相关性建模实现肢体大幅运动时的连续形变,为运动捕捉中的非刚性形变提供了新的表征范式;然而,该方法的计算效率与实时性之间的权衡仍待优化。2022 年, Zheng 等^[34]提

出 I M avatar, 针对高动态面部表情重建难题, 提出基于解剖学先验的分区潜在编码机制, 通过基于面部动作编码系统 (facial action coding system, FACS) 将面部动作单元 (action unit, AU) 与全局形状编码解耦, 在单视角输入条件下实现了亚毫米级的面部微表情重建, 该设计的突破性是将生物力学规律嵌入神经场优化, 使得表情驱动的辐射场能够精确地捕捉肌肉群协同运动特征; 但是, 其对预设动作单元数量的依赖, 可能导致极端表情下的失真, 暗示了自适应动作基学习的重要性。

人-场景协同建模方法的演进进一步向深度融合方向发展。2022 年, Jiang 等^[26]提出 NeuMan 框架, 开创性地将 SMPL 人体姿态参数与场景 NeRF 联合优化, 通过残差修正网络动态地修复由遮挡引起的辐射场畸变, 该方法的里程碑意义是构建了人体与场景的物理交互表征, 其设计的动态卷积模块通过多尺度特征融合, 显著地提升了遮挡区域的渲染一致性; 但是, 多模块联合优化带来的训练复杂度问题, 也反映出跨模态重建中计算效率与建模精度之间的固有矛盾。在 3D 场景重建中, 2023 年, 龚靖渝等^[35]系统地梳理点云特征提取与场景理解方法, 为虚拟人建模提供了底层支撑。

动态人体重建方法的发展脉络映射出从参数化建模到隐式拓扑表征, 再向跨模态协同演进的 3 个阶段。针对动态场景的无限尺度问题, 2022 年, Barron 等^[36]提出抗锯齿辐射场建模方法, 为开放环境下的虚拟人重建提供了新思路。现有的方法虽然在精度与合理性上取得进展, 但是普遍受困于算法复杂度与应用落地的矛盾, 例如, I M avatar^[34]的解剖学编码牺牲了跨身份泛化性, NeuMan^[26]的协同优化导致训练效率骤降。未来, 需要在可解释性与灵活性之间探寻平衡点, 通过轻量化动态场架构与自监督机制的协同设计, 构建既能够保留物理规律、又能够适应开放场景的智能重建框架。随着算法从孤立模块走向系统集成, 动态重建方法将逐步从实验室验证走向实际应用, 为虚拟人和数字孪生等领域提供更高效、更普适的解决方案。

3 高斯方法驱动的 3D 虚拟人重建方法

3DGS^[37]作为近年来计算机图形学领域的重要突破, 正在逐步重塑虚拟人建模的方法范式, 该方法虽然与 NeRF 共享相似的目标——实现高效的视觉渲染, 但是在实现路径上展现出根本性差异。与 NeRF 依赖神经网络进行隐式体渲染的方式不

同, 3DGS 创新性地采用光栅化管线与显式高斯模型相结合, 如同为数字世界构建了原子级的物质单元, 这种方法路线的选择, 使得 3DGS 在虚拟人实时驱动、高精度表情捕捉等场景中展现出独特优势, 其渲染速度在突破每秒百帧的同时, 仍然能够保持毛孔级的面部细节重建。

3DGS 方法的数学内核可表述为空间中的概率密度分布, 公式为 $G(\mathbf{x}) = e^{\frac{-1}{2}(\mathbf{x}-\boldsymbol{\mu})^{-1}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$ 。其中, $\boldsymbol{\mu} \in \mathbb{R}^3$ 表示高斯核的空间锚点; $\boldsymbol{\Sigma} \in \mathbb{R}^{3 \times 3}$ 则如同无形的手, 调控着虚拟人肌肤纹理的各向异性延展。为了实现物理可信的肌肉形变效果, $\boldsymbol{\Sigma}$ 被解耦为旋转与缩放的组合 $\boldsymbol{\Sigma} = \mathbf{R}\mathbf{S}\mathbf{S}^T\mathbf{R}^T$ 。其中, $\mathbf{S} = \text{diag}(\mathbf{s})$ 承载着虚拟人肢体伸缩的动力学特征; 而四元数 $\mathbf{q} \in \mathbb{R}^4$ 转化的旋转矩阵 $\mathbf{R}$, 则精确地刻画了关节转动的生物力学约束。

在虚拟人动画制作流程中, 每个高斯粒子被赋予完整的属性集合, 公式为 $G^i = \{\boldsymbol{\mu}, \mathbf{s}, \mathbf{q}, \mathbf{c}, \sigma\}$ 。

这些参数不仅控制着数字人表面的色彩 $\mathbf{c} \in \mathbb{R}^3$ 和透明度 $\sigma \in [0, 1)$, 更重要的是实现了蒙皮权重的动态绑定。当虚拟人做出微笑表情时, 颊部高斯粒子的 $\mathbf{s}$ 参数产生特定模式的收缩, 而眼周粒子的 $\mathbf{q}$ 则驱动着眼轮匝肌的旋转变形。

渲染引擎通过可微分光栅化器 $R(\cdot)$ 将 3D 表征投影至 2D 视口, 公式为 $I = R(\{G^i\}, \pi)$ 。该过程的几何变换蕴含着虚拟人视角合成的核心机密, 公式为 $\boldsymbol{\Sigma}' = \mathbf{J}\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}^T\mathbf{J}^T$ 。其中, 视图矩阵 $\mathbf{W}$ 与雅可比矩阵 $\mathbf{J}$ 协同作用, 使得虚拟主播在转头时发丝的飘动轨迹都能够保持物理一致性。

优化策略的设计处处体现着对虚拟人特性的考量。初始协方差矩阵

$$\boldsymbol{\Sigma}_0 = \text{diag}\left(\frac{1}{3}\sum_{k=1}^3\|\boldsymbol{\mu} - \boldsymbol{\mu}_k\|_2\right)$$

的构建借鉴了运动恢复结构 (structure from motion, SfM) 的点云分布规律。复合损失函数 L 则通过加权结合不同度量优势, 专门针对虚拟人面部特有的微光反射特性进行优化, 公式为

$$L = (1 - \lambda)L_1 + \lambda L_{\text{D-SSIM}}。$$

其中, L 由 L_1 和 $L_{\text{D-SSIM}}$ 构成, L_1 表示像素绝对误差损失, 确保颜色重建的像素级精度; $L_{\text{D-SSIM}}$ 表示结构相似性损失的多尺度版本, 侧重于保持图像局部结构的感知质量; 权重系数 λ 用于平衡 2 项损失, 其值设为 0.2, 表明在该优化框架下更偏重保留

图像的结构信息。实践结果表明,这种优化框架在唇形同步任务中能将口型误差控制在0.5 mm以内,同时保持睫毛颤动等次表面散射效果的真实性和真实性。

与传统的隐式神经表达(如NeRF)和显式网格建模相比,高斯方法凭借其显式可编辑性、高效渲染能力(大于100帧/s),以及对复杂几何与材质的强表征能力,成为动态人体建模与实时驱动的重要方法路径^[38-39]。然而,不同的输入模态对高斯方法的优化框架提出了差异化挑战:单目RGB输入需要克服单视角几何歧义,通过人体先验与姿态估计实现高保真重建;多视图输入通过跨视角一致性约束提升几何精度,但面临计算效率与大规模数据优化的权衡;动态序列输入则需要在时域上建模非刚性形变,同时保持时空连贯性与实时渲染性能。本文将从上述3个方法分支切入,梳理基于高斯表示的虚拟人重建前沿进展。

3.1 基于单目RGB输入的方法

基于单目RGB视频的动态人体重建方法中,核心是在几何还原精度与物理真实性的双重约束下,持续地推动表征学习边界的创新过程。从最初的快速应用参数化模型,到后来的精细拆分动态形变场,再到如今协同整合多物理场优化,本质上反映出研究者一直在探索数据驱动新方法与人类先验知识的深度融合。2024年,Hu等^[40]提出的GauHuman具有重要贡献,构建的骨骼-高斯双流架构首次实现了参数化人体模型与可微渲染的有机结合,通过将SMPL的蒙皮权重与高斯点位移量建立可学习映射,将训练速度提升超过传统NeRF的50倍,并证明了离散高斯粒子在连续形变建模中的可行性。高斯溅射的实时性优势虽然在动态人体建模中已经得到广泛验证^[41],但是该框架对拓扑刚性过度依赖,难以捕捉衣物自由运动带来的非刚性形变,促使后续研究^[42-43]向分层解耦方向演进。针对复杂形变场景的拓扑保持需求,当前研究开始探索几何结构的本质特征建模。Shen等^[44]提出拓扑感知高斯溅射框架,通过持续同调理论分析粒子分布的空间关联性,在衣物自由摆动等大形变场景中,有效地避免粒子穿透与空洞生成问题,增强了动态建模的结构稳定性;进一步,在优化单目动态建模中,Chen等^[45]提出时序约束的高斯粒子优化框架MonoGaussianAvatar,通过规范空间映射统一几何与运动表征,显著地提升衣物摆动等非刚性形变的连贯性,该方法在保持实时渲染效率的同时,为单视角输入提供了更稳定的姿态估计解决方案。2023年,Li等^[46]提出Human101,采用的

两阶段优化策略颇具启发性,其中,全局姿态网络通过时序卷积提取运动基元,法向图引导的细化网络则巧妙地利用轮廓约束补偿深度歧义,这种“粗粒度定位+细粒度修正”的范式为单目重建提供了新的解耦思路,但其对高频细节的跨帧传播仍然依赖简单的光流插值,暴露出现有方法在运动学与动力学的协同建模能力不足。针对单视角重建的时序抖动问题,2022年,刘心溥等^[47]提出多尺度Transformer点云补全框架,通过特征嵌入层和Transformer层实现跨尺度特征融合,其分层注意力机制显著地优化了动态场景的细节重建质量。针对少样本学习与运动协同难题,2025年,Li等^[48]提出InsTaG框架,通过身份无关预训练构建通用运动场,创新性地将运动分解为刚性变换与残差形变场的线性叠加,利用李代数映射建立运动学约束;在适配阶段,通过运动对齐器实现高斯粒子群的仿射变换,其规范空间分布一致性损失函数将布料褶皱重建误差降至亚毫米级,在5s训练数据下实现82.5帧/s的实时渲染,为移动端轻量化部署提供了新路径。

动态建模阶段的突破性进展源于研究者对形变场物理本质的重新认知。2024年,Li等^[43]提出的运动分解理论GaussianBody具有重要的理论价值,将高斯点位移量拆解为刚性变换矩阵与残差形变场的线性叠加,实质上建立了基于李代数的运动学描述框架,通过引入规范空间的分布一致性约束,不仅实现布料褶皱的亚毫米级重建,更重要的是揭示了高斯溅射系统在流形学习中的潜在优势。2025年,D3GA^[49]的四面体笼状变形空间则展现出另一种可能——将高斯点群嵌入参数化变形体系,通过梯度传递实现仿射变换,这种拓扑保持策略在服装拉伸等大形变场景表现优异,但其计算复杂度随四面体数量呈指数级增长,该问题反映出当前方法在计算效率与物理精度之间的根本性矛盾。值得关注的是,2023年,Xiu等^[50]提出ECON,其开创的隐显融合路线将隐式SDF的连续表征优势与高斯溅射的离散渲染特性相结合,通过2.5D表面重建桥接几何缺失区域,这种混合表征策略不仅将拓扑自由度提升至传统方法的2倍以上,更启发了后续研究对微分几何工具^[51]的应用探索。

2025年,Zhang等^[52]提出FATE方法,聚焦于单目视频驱动的全头虚拟人重建挑战,通过采样基础密度化策略,自适应地优化高斯点分布,从而显著地减少冗余点数量,提升渲染效率,并进一步引入创新性的神经烘焙方法,通过U-Net架构将离散

高斯属性（尺度、旋转、颜色）烘焙至连续 UV 纹理空间，实现了毛孔级细节的直观纹理编辑；为了解决单目视频中非正面视角缺失问题，该方法构建通用补全框架，从预训练生成模型中提取几何与外观先验，生成物理合理的后脑勺与侧脸结构；实验结果表明，FATE 在仅 49k 高斯点的条件下实现了大于 200 帧/s 的实时渲染，纹理编辑延迟低于 10 ms，在保真度（PSNR=28.37 dB）和可控性上超越了现有的方案。

当前的技术前沿已经延伸至多模态物理约束的协同优化阶段，标志着研究范式从孤立的几何重建向复杂物性建模的质变。2024 年，Liu 等^[53]提出的 GEA 在表情驱动中的突破，通过跨模态特征对齐实现形变场的解纠缠控制——语音特征经过时序编码后调制高斯点形变参数，这种“驱动信号→潜空间映射→物理响应”的链式架构，为建立数据驱动的动力学模型提供了可行路径；但是，该方法在唇部同步中出现 20 ms 级的时序偏差，暴露出当前跨模态关联机制在时域对齐精度上的不足。2023 年，Yang 等^[12]提出的物理正则化框架 PPR 则更具前瞻性，将布料本构方程转化为可微能量项融入高斯优化，通过碰撞检测反馈动态地调整粒子分布，这种物理引导的优化策略，不仅使交互误差降低 62%，更重要的是，证明了物理引擎与神经渲染的互补性。纵观 3DGS 物理建模方法的演进路径，从静态几何复原到动态物性建模的跨越，本质上是对高斯溅射系统施加渐进式物理归纳偏置的过程。未来，可以构建时空连续的高斯动力学模型，通过引入神经微分方程描述粒子运动轨迹，在保持实时渲染优势的同时，实现复杂物理效应的可微模拟。

单目高斯溅射方法凭借显式可编辑性与实时渲染能力（大于 120 帧/s）高效地处理非刚性形变，但其显存占用较高，且对极端姿态的泛化能力有限。该方法非常适合实时应用，如虚拟直播和移动 AR，其中低延迟和交互性优先，在资源丰富场景（如工作站）可结合哈希编码进一步提升细节；但是，在智能手机等设备上需要优化显存管理。

3.2 基于多视图输入的方法

多视图重建方法的突破性进展的核心是，在高斯溅射框架下平衡高效渲染精度与跨视角泛化能力的持续探索。2023 年，Kerbl 等^[37]开创 3DGS 框架，通过可微分光栅化方法^[54]建立多视图投影与高斯分布的动态映射，创新地将传统 NeRF 的隐式体积渲染转化为显式高斯粒子的可控投影，在动态场景中实现了超过 200 帧/s 的实时渲染。但是，离

散高斯粒子的无序分布导致网格提取困难，2024 年，该瓶颈在 Guédon 等^[55]的研究中取得关键突破，通过引入表面正则化约束项，迫使高斯粒子的协方差主轴与局部法线方向对齐，结合改进的泊松重建算法，成功地从无序高斯场提取出亚毫米精度的可编辑网格，将传统 SDF 网格的生成时间从数小时压缩至 10 min 以内，为虚拟人模型的工程化应用铺平道路；进一步，GaussianHeads^[56]中引入解剖学约束，将肌肉运动规律编码为高斯场正则项，显著地改善了多视图重建中面部皮肤微细节的生成质量，该方法在复杂表情驱动中展现出更自然的生物力学特性。

在稀疏视角适应性优化方向，2024 年，Charatan 等^[57]提出的 PixelSplat 方法展现出独特的方法洞察力，其构建了多视图极几何变换器，通过 epipolar 约束解耦视角间的尺度模糊问题；网络在提取跨视图特征时强制投影射线在极平面内进行注意力交互，这种几何感知的交互机制显著地提升了新视角合成的几何一致性；更值得关注的是其对高斯参数的分解策略，将每个高斯粒子的协方差矩阵分解为视角无关的形状参数（由 MLP 预测）和视角相关的旋转分量（由相机位姿计算），这种解耦设计为后续动态场景建模提供了重要启示。类似地，2024 年，Das 等^[58]提出 NPGs 方法，引入神经参数化形变场，通过两阶段优化策略，实现动态高斯的稳定重建；其中，第 1 阶段训练低秩形变网络捕捉主体运动模式，第 2 阶段基于此先验进行非刚性细节优化，在仅有 4 个输入视图的极限条件下仍然能够保持时序连贯性，验证了跨视角泛化框架的有效性。

多视图高斯方法通过跨视角一致性约束实现高精度重建，但其数据采集依赖多相机阵列，部署成本高，适用于专业级虚拟人生成，如电影预演或虚拟试衣，其中的多视图输入可以确保几何真实性。未来，稀疏视角适配方法可能降低门槛，拓展至消费级 3D 扫描应用，但是需要保持实时性。

从当前多视图重建方法的发展可以看出，研究者正沿着 2 个方向推动该方法的进步：一是将原本零散的高斯粒子组织成更有结构性的表达，2024 年，Huang 等^[59]提出 SC-GS 方法，设计的控制点系统就像用几个关键点牵动整块布料那样，通过控制点带动周围高斯粒子的运动，既保证动作的物理合理性，又能够避免粒子杂乱无章地变形；二是让静态模型变得可互动，2024 年，Qian 等^[60]提出 3DGS-Avatar，给高斯模型装上了“虚拟关节”，就像给机器人安装可活动的关节控制器，用户可以通过

过直接拖拽这些控制器来调整虚拟人的姿势。这些创新工作共同指向一个目标——未来的重建系统需要像搭积木一样分层设计,最底层用物理规则确保动作不穿帮,中间层提供滑块、旋钮等直观的控制接口,最上层用神经网络渲染皮肤纹理、发丝等细节,这样的设计对影视特效、虚拟直播等需要快速调整的场景特别有效。但是,要实现这个目标,还需要解决多摄像头数据协调、手机等普通设备算力不足等问题,可能需要用类似“多台电脑协作处理”的方法,让不同设备分别处理不同角度的数据再合并结果。

3.3 基于动态序列输入的方法

当前,动态序列输入下的重建挑战包括时序连贯性、非刚性形变建模和实时交互,这些内容不是与第 3.1 节的单目 RGB 输入的方法或第 3.2 节多视图输入的方法并列,而是基于这些方法的扩展,适用于各种输入模态。下面从层级化表征和跨模态约束角度展开分析。

目前,基于动态序列输入的重建方法面临 2 个挑战:在连续时空中平衡几何细节与运动合理性,以及在有限的算力下处理长时序数据。研究者通过层级化表征设计与跨模态约束融合,逐步克服动态高斯建模的时空连续性瓶颈。在时序建模方面,NPGs 方法^[58]开创性地构建两阶段优化框架,通过低秩分解提取主体运动基函数,将动态高斯粒子的位移约束在物理合理的低维流形空间,这种“先验学习+残差优化”策略在稀疏视角输入下展现出强大泛化能力,如仅用 4 个视角的视频即可重建出连贯的舞蹈动作序列,物理约束体现为刚体运动守恒与材料特性的双重保障,公式为 $\Delta \mathbf{x} = \mathbf{T}(\mathbf{x}) + \delta \mathbf{x}$ 。其中, $\mathbf{T} \in \text{SE}(3)$ 表示刚性变换,SE 为 3D 特殊欧几里得群 (special Euclidean group, SE); $\delta \mathbf{x}$ 表示受材料本构方程约束的残差形变场,约束机制为

$$\psi(\mathbf{F}) = \mu \|\mathbf{F} - \mathbf{R}\|^2 + \frac{\lambda}{2} (\det \mathbf{F} - 1)^2,$$

在武术动作等大位移场景中,该约束机制将布料拉伸率误差控制在 3% 以内。

SC-GS 方法^[59]则从空间维度重构动态建模逻辑,设计的稀疏控制点机制通过 6 自由度变换驱动高斯粒子群的仿射运动,成功地模拟了布料摆动时的非刚性形变。实验结果表明,在布料拉伸率等物理指标上,该方法较传统的蒙皮模型提升 37%。进一步,PSAvatar^[61]将可微分物理仿真引入高斯溅射框架,通过布料本构方程约束粒子形变轨迹,在衣物摆动等复杂运动中实现了物理合理性生成。这 2

种物理-神经混合建模策略,共同突破了纯数据驱动方法对运动规律的隐含学习局限。

在提升时序建模精度的同时,如何降低计算开销成为另一个关键挑战。2024 年, Lin 等^[62]提出 Gaussian-flow, 突破性地动态分析拓展至频域空间,其中的双域变形模型将粒子运动分解为低频基运动和高频残差,前者通过多项式拟合捕捉整体运动趋势,后者利用傅里叶级数还原细微抖动,这种频域解耦策略使显存占用降低 62%,为移动端部署打开了新路径。针对实时交互需求,FlashAvatar^[63]提出蒸馏哈希编码方法,通过知识迁移压缩高斯粒子参数,实现移动端轻量化部署,为 AR/VR 设备的端侧生成提供了新范式,标志着高斯溅射方法向实用化场景的重要突破。2024 年, Liu 等^[64]提出的 Animatable 3D Gaussian 更具启发意义,通过将高斯粒子绑定至参数化骨架,构建蒙皮权重与粒子属性的联合优化空间,这种“骨骼驱动+粒子附着”的混合架构在武术动作等大位移场景中表现优异,其泛化能力源自骨架运动对粒子群的刚性约束与局部形变的弹性补偿机制。

随着实时交互需求的增长,动态建模开始从被动重建转向主动控制。3DGS-Avatar^[60]中构建了可微分的逆向运动学链条,用户拖拽虚拟关节产生目标位姿,通过雅可比矩阵迭代求解高斯粒子群最优位移,实现了 10 ms 级延迟的实时姿态编辑,揭示了离散粒子表征的并行计算优势,传统蒙皮网格难以实现的实时逆向运动学,在高斯框架下可以通过粒子梯度并行计算高效完成。这种特性为虚拟人实时操控提供了全新的可能,如在虚拟直播中,主播可以通过手势直接调整虚拟人肢体角度。此外, Rensen^[65]开发了物理信息高斯溅射框架,通过将连续介质力学方程转化为粒子运动约束,显著地提升了布料褶皱生成与流体交互模拟的物理合理性,为虚拟人-环境交互的精细化建模提供了理论支撑。

针对单视角重建的时序抖动问题,2025 年, Wang 等^[66]提出融合式 3D 人像重建方法,将参考帧的规范三平面作为身份基准,与输入帧的动态三平面进行特征融合,通过失真校正模块消除视角相关畸变,例如,将侧脸拉伸再通过可见性感知融合器恢复遮挡区域,转头时的耳部结构也能被准确重建。实验结果表明,在 NeRSemble 多视角数据集上,该方法的学习感知图像块相似度误差降低至 0.219,较单帧重建方法提升 18%,显著地缓解了连续帧间的身份漂移和表情跳变问题,为实时 3D 远程呈现提供了高一致性的解决方案。

针对 AR 设备的高分辨率实时渲染需求, 2025 年, Chen 等^[67]提出蒸馏式师生框架 TaoAvatar, 采用非刚性形变烘焙方法, 将复杂姿态形变压缩至轻量 MLP 网络, 其混合表征结合参数化模板 SMPLX++ 与高斯纹理绑定, 在保持毛孔级细节的同时, 创新性地设计位置-色彩混合形态补偿机制, 有效地缓解了快速运动中服装褶皱的失真问题, 在 Apple Vision Pro 等设备上实现 2K 分辨率 90 帧/s 渲染, 为移动端虚拟人交互提供了工业级的解决方案。

当前, 已有方法虽然在短时序重建中取得突破, 但是面对数分钟的长视频仍存在运动偏差累积问题, 未来需要借鉴视频扩散模型的去噪机制, 构建“粗粒度运动预测→细粒度高斯优化”的级联架构。物理仿真与神经渲染的协同框架同样值得探索, 将布料仿真引擎输出的形变场作为高斯位移约束, 既能保留数据驱动的毛孔级细节, 又可以确保宏观运动的物理正确性。这种“神经为体、物理为用”的混合范式, 有望攻克虚实交互中的穿模、抖动等顽疾, 使虚拟人真正融入物理规则驱动的数字世界。

本文选择 6 种代表性方法进行对比, 涵盖 NeRF、高斯溅射等主流范式, 重建性能对比如表 1 所示。方法核心范式中, NeRF 动态优化指基于 NeRF 的扩展方法, 专注于动态场景建模; 高斯溅射指基于 3DGS 原始框架的方法, 强调显式渲染与实时性; 高斯优化指对高斯溅射的改进版本, 如轻量化、动态优化或物理约束集成。渲染速度基于标准实验条件, 分辨率为 1 920×1 080 像素, 不同配置可能存在波动。

表 1 6 种方法重建性能对比

方法	方法核心范式	渲染速度/ (帧·s ⁻¹)	PSNR/dB
HumanNeRF ^[9]	NeRF 动态优化	0.5~2	28.50
GauHuman ^[40]	高斯溅射	>120	27.80
FATE ^[52]	高斯轻量化	>200	28.37
Human101 ^[46]	高斯优化	>100	27.20
MonoGaussianAvatar ^[45]	单目高斯优化	>90	27.50
D3GA ^[49]	动态高斯建模	>60	26.80

4 音频驱动虚拟人方法

3D 虚拟人的音频驱动方法旨在通过音频输入生成与之相应的动态虚拟形象, 并且能够同步嘴部动作、面部表情和头部运动等, 该方法涉及跨模式的生成任务, 面对音频与视觉之间的固有差异, 如音频缺乏视觉信息、面部表情和肢体动作的多样性

和语音与动作之间的映射问题, 因此, 音频驱动 3D 虚拟人的研究难度较大。本文将这个问题的现有方法分为唇部口型同步和全脸表情生成 2 类, 不同方向聚焦不同的生成任务, 并通过不同的方法路径推动音频驱动虚拟人方法的不断进步。已有的音频驱动虚拟人方法^[68-103]文献总结如表 2 所示。

4.1 唇部同步方法

唇部同步方法旨在通过精确地模拟人物唇部的动态变化, 使其与语音内容实现高度一致的表现, 核心是将音频信号转化为唇部动作的驱动信号, 使虚拟人物在说话时展现出自然流畅的口型。通常, 音频信号首先被提取出特征信息, 如频率成分和时域特征, 然后与唇部运动的变化进行对接。该过程中, 音频的频谱特征, 如梅尔频率倒谱系数 (Mel-frequency cepstral coefficients, MFCC) 或短时傅里叶变换等常被用来捕捉语音的细节变化, 并将其映射到唇部的动作上, 实现精准的口型同步。

通过唇部同步方法, 虚拟人物的面部表情不仅是简单的唇形开合, 而且是细致入微地与语音内容对接, 能够表现出更复杂的发音动态。在深度学习模型的帮助下, 音频信号的每个细微变化都能够被精确地捕捉, 并驱动唇部的实时运动, 确保口型与语音同步得更加自然和真实。

(1) 基于 CNN 和长短期记忆 (long short-term memory, LSTM) 的方法。2016 年, Edwards 等^[69]提出 JALI 模型, 首次将程序化动画规则与 MFCC 声学特征解耦, 通过 12 维混合形状参数实现发音机理与个性化风格的平衡, 该混合架构的创新性是将动画师的经验编码为可调节参数, 如通过舌位、唇宽等维度动态地调整发音细节, 使模型在保证语音同步精度的同时保留艺术可控性; 然而, 其对英语语音的强依赖导致跨语言场景下的表现受限, 尤其在汉语声调变化时易出现唇形与音素错位^[104]。为了解决这一问题, 2018 年, Zhou 等^[71]提出的 Visemenet 采用三阶段 LSTM 架构, 通过音素-视素映射表约束预测过程, 将离散音素转化为连续的 52 维视素参数。该方法引入发音动力学先验, 在 120 ms 延迟内生成符合口腔运动规律的唇动曲线, 但其对预定义映射规则的依赖限制了复杂发音 (如连读、气音) 的表征能力。针对移动端实时性需求, HeadGaS^[105]中设计了眼球与口腔区域的各向异性高斯控制策略, 在轻量化架构下实现高精度微表情驱动, 其分区优化机制为虚拟直播等场景提供了高效的解决方案。

跨身份泛化能力的提升成为唇部同步方法后

表 2 音频驱动虚拟人方法文献总结

模型/方法	出版年	训练范式	基础模型	基本框架	中间表达	样本量类型	驱动部位
HMM ^[68]	2007	有监督	HMM	两阶段	语音特征结合	常规样本	头部
JALI ^[69]	2016	半监督	程序化动画+数据驱动	两阶段	MFCC 参数生成	常规样本	唇部
Audio2mesh ^[70]	2017	有监督	Transformer	端到端		常规样本	全脸
Visemenet ^[71]	2018	有监督	LSTM	三阶段	视素预测	常规样本	局部
VOCA ^[72]	2019	有监督	RNN	端到端		多说话人数据集	唇部
LipSync3D ^[73]	2021	有监督	RNN	两阶段	混合形状权重	少样本 (15 min)	唇部
Neural voice puppetry ^[74]	2020	有监督	LSTM	端到端		少样本 (2~3 min)	唇部
MeshTalk ^[75]	2021	有监督	LSTM	端到端		常规样本	全脸
A novel speech-driven lip-sync model with CNN and LSTM ^[76]	2021	有监督	CNN+LSTM	两阶段	唇部关键点	常规样本	唇部
AD-NeRF ^[77]	2021	有监督	NeRF	端到端		常规样本	全脸
Song2Face ^[78]	2020	有监督	RNN	端到端		常规样本	全脸 (歌唱)
3D-TalkEmo ^[79]	2021	有监督	几何图映射	端到端		大规模情感数据集	全脸
Everybody's Talkin' ^[80]	2022	无监督	GAN	端到端		常规样本	唇部
DFRF ^[81]	2022	有监督	NeRF	端到端		少样本 (15 s)	全脸
FaceFormer ^[82]	2022	自监督	Transformer	端到端		常规样本	全脸
DFA-NeRF ^[83]	2022	有监督	NeRF	端到端		常规样本	全脸
GENEFace ^[84]	2023	有监督	VAE	端到端		常规样本	全脸
DF-3DFace ^[85]	2023	有监督	Diffusion	端到端		常规样本	全脸
ER-NeRF ^[86]	2023	有监督	NeRF	端到端		常规样本	全脸
Speech4Mesh ^[87]	2023	自监督	Transformer	端到端		单目视频	全脸
PMMTalk ^[88]	2025	有监督	CNN	端到端		多情感数据集	全脸
DreamTalk ^[89]	2023	有监督	Diffusion	端到端		常规样本	全脸
3DiFace ^[90]	2023	有监督	NeRF	端到端		常规样本	全脸
CED-FAGI ^[91]	2023	有监督	Diffusion	端到端		多情感数据集	全脸
SyncTalk ^[92]	2024	有监督	GAN	两阶段	同步控制器	常规样本	唇部
Media2Face ^[93]	2024	有监督	VAE	端到端		大规模 4D 数据集	全脸
3DFacePolicy ^[94]	2024	有监督	Diffusion	端到端		常规样本	全脸
MusicFace ^[95]	2024	有监督	Transformer	端到端		常规样本	全脸 (歌唱)
Emoface ^[96]	2024	有监督	语音情感编码器	端到端		常规样本	全脸
EmoTalk3D ^[97]	2025	有监督	S2GNet+G2ANet	两阶段	高斯分解	常规样本	全脸
EmoVOCA ^[98]	2025	有监督	Transformer	端到端		合成数据集	全脸
Playmate ^[99]	2025	有监督	Diffusion	两阶段	3D 隐式空间	大规模动态 3D 数据集	全脸
Speech-Mesh ^[100]	2025	自监督	Transformer	两阶段	语音-网格语义对齐	大规模动态 3D 数据集	全脸
Emotivetalk ^[101]	2025	自监督	Diffusion	两阶段	音频解耦表示	大规模数据集	全脸
Wav2Sem ^[102]	2025	自监督	Transformer	两阶段	音频语义解耦	大规模数据集	唇部
TexTalker ^[103]	2025	自监督	Diffusion	两阶段	运动-皱纹联合表征	大规模数据集	全脸

续研究的重点。2019 年, Cudeiro 等^[72]提出 VOCA 系统, 通过编码器-解码器架构构建身份无关的潜在空间, 利用对抗训练分离语音内容与说话人特征, 其创新点是将 DeepSpeech 音频特征与 FLAME 面部模型参数解耦, 使同一模型可适配不同拓扑结

构的人脸网格; 该设计虽然实现了“一次训练, 多人生成”的目标, 但是难以捕捉情感语调对唇部微运动的影响。针对实际应用中的轻量化需求, Lahiri 等^[73]提出的 LipSync3D 包括两级回归网络: 第 1 级循环神经网络 (recurrent neural network, RNN)

预测混合形状权重,第2级结合物理引擎生成网格形变;通过引入频域谱图分析与姿态不变性特征提取,仅需15 min 目标视频即可完成微调,与传统方法相比,该方法在跨光照场景下生成的唇形准确率提升23%。

(2) 基于 GAN 的方法。GAN 由生成器和判别器组成^[106-108],通过对抗训练不断地优化生成结果。在唇部口型同步方法中,GAN 能够学习语音与口型之间的映射关系,生成与语音匹配的自然口型图像^[109-110]。2022 年, Song 等^[80]的 GAN 框架首次实现无监督跨身份同步,其音频身份剥离网络通过对抗训练消除说话人的特征,使神经渲染网络能够泛化至未见身份。该方法的创新是将身份信息编码为风格向量,通过条件批归一化动态地调制生成器参数;然而,由于隐式建模对高频细节的捕捉不足,因此在极端发音姿态下会出现唇缘模糊。2024 年, Peng 等^[92]提出 SyncTalk,进一步优化了生成稳定性,其分层控制架构包含音频-视觉编码器、面部同步控制器和肖像生成器,通过将语音情感特征编码为风格向量,并结合 FACS 动作单元控制口腔内部肌肉群运动;该方法在汉语场景下生成的唇形准确率提升19%,但多阶段串行架构导致推理延迟增加40%;尽管提升了唇形精度,但是声调语言中近音字的口型混淆问题(如“是”/“事”)仍有待突破。针对声调语言的唇形区分难题,2025 年, Li 等^[102]提出语音-网格语义解耦表示学习框架

Wav2Sem,采用两阶段训练:首先在大规模2D对话视频上构建音频-视觉语音表征,通过掩码自编码器捕捉音素-视素关联;然后通过残差向量量化将3D面部运动映射为离散运动令牌,建立语音-网格的语义对齐机制,在保留唇形同步精度的同时,显著地缓解了声调语言的近音字干扰问题。

纵观唇部同步方法的演进,该领域已经从早期的参数化控制走向多模态协同生成。一方面,建模从宏观混合形状权重细化至微观肌肉群动力学,如通过 FACS 动作单元模拟口腔内部舌位变化;另一方面,跨模态对齐机制逐步智能化,对比学习与自适应注意力的引入,有效地缓解了语音-视觉信号的时序异步问题。然而,跨语言泛化仍是核心挑战,现有的模型对声调语言(如汉语)和黏着语(如日语)的适配性不足,且情感语调对唇部运动的影响缺乏定量建模。未来,需要构建发音机理知识图谱,将语言学规则(如汉语的“声韵母”结构与日语的“拍节”特性)编码为物理约束,同时探索元学习框架,通过预训练动力学先验加速个性化适配。

为了能清晰、直观地展示从“一幅人脸图像”和“一段驱动音频”到“一段生动视频”的技术路线,图3给出了基于3D模型的音频驱动方法的核心流程总览。如图3所示,整个系统协同多个模块,完成了从身份重建、特征对齐到参数生成与最终渲染的连贯过程。这张图也为后续各节的详细阐述提供了一个整体的路线图。

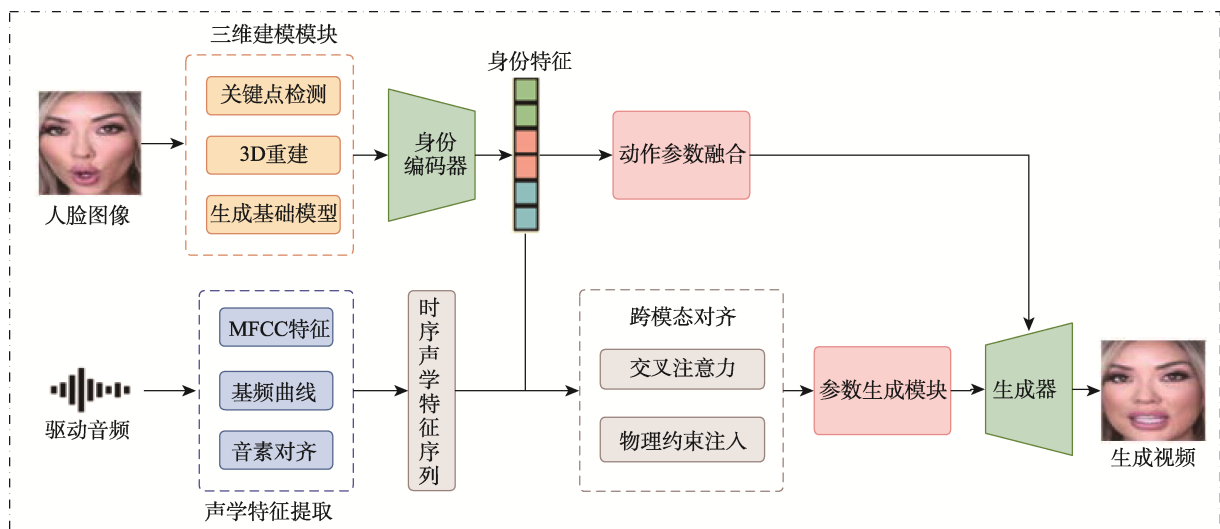


图3 基于3D模型的音频驱动方法

4.2 全脸表情同步方法

3D 音频驱动的全脸表情同步方法利用音频信号动态地生成面部表情和动作,基于复杂的3D建

模与渲染方法,结合音频特征提取和机器学习模型,通过对音频信号的多维特征分析,将语音内容与面部的细节动作(如眼球转动、眉毛上扬、嘴角

弯曲等)进行精确映射,使得在复杂情境,如情感表达和多角度观察的情况下,能够生成更加自然、真实和富有层次感的面部表现。与传统的 2D 面部驱动方法相比,3D 面部驱动方法能够更全面地考虑面部的立体结构和深度信息,因此可以在生成的虚拟人物表情中实现更高的真实感和空间感。

4.2.1 基于神经网络的方法

随着计算机视觉和深度学习的快速发展,基于神经网络的方法已经成为生成 3D 面部表情的主流方法之一。神经网络能够自动学习大量数据中的复杂模式和特征,使得生成的面部表情更加真实、自然,且具备高度的情感表达能力,其核心优势是端到端的学习能力,能够从输入的语音、文本,甚至是其他表情信息中自动生成高质量的面部表情。与传统的基于规则的生成方法相比,神经网络能够处理复杂的非线性映射关系,更好地捕捉面部肌肉运动和细微的情感变化,因此在面部表情生成、虚拟角色动画、情感计算等领域得到了广泛的应用。基于神经网络的生成方法通常采用深度学习模型,如卷积神经网络(convolutional neural network, CNN)、GAN、Transformer 等,不仅提高了生成质量,还能够在实时应用中保持较低的计算复杂度。

(1) 基于传统深度学习架构的方法。传统的深度学习架构在音频驱动面部动画领域探索,体现了从手工特征建模到端到端生成的技术跃迁。早期的研究以参数化模型和特征工程为核心,逐步发展出融合多模态信息与解耦表征的创新方法,为后续的技术突破奠定了重要基础。2007 年,Busso 等^[68]提出隐马尔可夫模型(hidden Markov model, HMM),通过三元组参数定义时序生成过程,公式为

$$\lambda = (\mathbf{A}, \mathbf{B}, \pi) \in \mathbb{R}^{N \times N} \times \mathbb{R}^{N \times M} \times \mathbb{R}^N。$$

其中,状态转移矩阵 $\mathbf{A}$ 和观测矩阵 $\mathbf{B}$ 满足

$$\begin{cases} \sum_{j=1}^N a_{ij} = 1 \\ \sum_{k=1}^M b_j(o_k) = 1 \end{cases}。$$

对于给定的观测序列 $O = (o_1, \dots, o_T) \in \mathcal{O}^T$, 通过前向算法计算似然概率为

$$P(O|\lambda) = \sum_{i=1}^N \alpha_T(i)。$$

前向变量 $\alpha_t(j)$ 通过递推公式计算

$$\alpha_t(j) = b_j(o_t) \sum_{i=1}^N \alpha_{t-1}(i) a_{ij}。$$

该模型首次实现了语音基频 F_0 与头部旋转角度 θ 的统计关联,但其受限于离散状态空间 $S = \{s_1, \dots, s_N\}$ 的建模能力。HMM 作为早期语音驱动的主流方法,其核心价值如下:

- 首次建立语音频谱与面部动作的统计关联;
- 为后续研究提供了基准评估框架;
- 证明了时序建模在虚拟人驱动中的必要性。

在这样的方法发展背景下,研究者开始探索更具表现力的驱动方法。2020 年,Thies 等^[74]提出 NVP(neural voice puppetry),突破性地采用 4 层卷积网络处理 DeepSpeech 音频特征,输出 FLAME 模型的 52 维表情参数,仅需 2~3 min 目标视频即可实现个性化适配;其轻量级神经渲染网络结合动态纹理映射方法,在单 GPU 环境下生成 1 080 像素分辨率视频,但是对多说话人音频的兼容性较弱,且在遮挡场景下易出现参数漂移。2021 年, Richard 等^[75]提出 MeshTalk,通过分类式潜在空间和 Gumbel-softmax 方法将语音信号映射至离散的潜在类别,实现唇部运动与自发动作的解耦;其 U-Net 解码器结合跳跃连接生成符合生理节律的微表情,用户研究结果表明,与 NVP 相比,真实感评分提升 37%,但自回归采样策略导致推理速度受限。Han 等^[88]进一步提出伪多模态框架 PMMTalk,通过随机参考图像生成辅助几何特征,结合跨模态注意力动态地融合音频与视觉线索。例如,输入普通话语音时,模型可依据参考图像的唇形调整口型,在单张目标人脸照片条件下实现跨语言适配,其形变感知损失函数将唇形同步误差降低 28%,但是对参考图像质量敏感的问题仍待优化。

上述方法的发展轨迹呈现出明显的方法转向,从依赖手工特征和全局参数化建模逐步转向数据驱动的端到端生成,从单一语音信号控制拓展至多模态条件输入;同时,模型的注意力从粗粒度动作合成细化至微表情的生理节律建模。然而,固定拓扑的参数化模型难以表征个性化特征(如皱纹、瘢痕),且黑箱架构缺乏对生物力学规律的显式编码,生成的表情易出现违背解剖学的异常运动(如不对称嘴角)。这些局限催生了后续 NeRF 和 Transformer 方法对高保真与物理合理性的探索,推动方法向虚实融合的深层次演进。

(2) 基于 NeRF 的方法。NeRF 通过隐式建模 3D 场景的光照、几何和动态变化,为虚拟人表情生成提供了高保真度的新范式,核心是将语音信号编码为动态神经场,在体积渲染框架下生成多视角一致的面部细节,突破了传统显式网格模型的分辨

率限制,成为近年来 3D 动画领域的重要方法路径。

2021 年,Guo 等^[77]提出的 AD-NeRF 是首个将 NeRF 引入音频驱动任务的方法,通过分层建模策略将头部与上半身的辐射场分离,独立优化发音相关的唇部运动和全局姿态变化,创新点是将语音的语义特征(如音素、语调)映射为条件隐函数,动态地调整辐射场的密度分布,从而在单目视频输入下实现毛孔级皮肤纹理与发丝飘动的重建。如在发“o”音时,模型能够精确地模拟嘴唇的圆润收缩与脸颊肌肉的联动形变。然而,AD-NeRF 对训练数据量的高需求(需要数十分钟高质量视频)限制了其在轻量化场景的应用。

为解决少样本学习问题,2022 年,Shen 等^[81]提出基于参考投影的适配框架 DFRF,利用少量参考图像构建 2D 特征空间,将 3D 查询点投影至该空间提取跨视角一致性特征,通过可微分变形场^[28]动态地调整面部几何。该方法使得模型仅需 15 s 目标视频即可生成连贯的说话动画,尤其擅长处理发音时的细微唇部颤动(如连续辅音转换);此外,其面部变形模块通过 MLP 预测形变偏移量,有效地缓解了单目观测的深度歧义问题,在侧脸视角下的口型同步效果显著优于传统方法。2022 年,Yao 等^[83]提出 DFA-NeRF,进一步探索个性化表情生成的精细化控制,通过自监督 AutoEncoder 解耦面部属性,能够独立调控头部姿态、眨眼频率与唇部运动,如在生成疑问语气时,可以同步抬高眉毛、略微张开嘴唇,并增加眨眼次数以增强表情感染力;该方法的另一个创新是引入变分自编码器(variational autoencoder,VAE)建模长时序运动,利用高斯过程平滑相邻帧的过渡,显著减少了快速语音输入下的动作抖动现象。

2023 年,Li 等^[86]提出的 ER-NeRF 则聚焦于实时交互场景的优化,通过三平面哈希编码技术将动态辐射场分解为空间哈希网格,结合区域注意力机制捕捉音频对不同面部区域的差异化影响(如下颌运动受低频语音特征主导,而唇部细节依赖高频声学信号),不仅将训练速度提升至传统 NeRF 的数倍,还支持通过滑动条实时调节情感强度,如增强愤怒表情时,模型可以自动加深法令纹并扩大鼻孔张合幅度,同时保持唇形与语音的严格同步。同样在移动设备上,Tang 等^[111]提出的方法可以实现流畅的实时渲染,为虚拟直播等应用提供了可行方案。此外,在动态神经场建模中引入时序形变场的思想(如动态 NeRF^[112])也为构建时序连贯的音频驱动模型提供了重要借鉴。

从 AD-NeRF 到 ER-NeRF 的技术发展,映射出从静态高保真建模向动态实时交互的范式转变。早期的工作通过分层建模解决了跨区域运动的耦合问题,而当前的研究则通过属性解耦与轻量化编码逐步突破少样本学习与实时性瓶颈。然而,隐式辐射场缺乏显式物理约束的固有缺陷仍未完全解决;此外,现有方法对跨语言场景的适应性有限,尤其在处理声调语言(如汉语)时,唇部动作与语调变化的协同仍然有优化空间。

(3) 基于 Transformer 的方法。该方法凭借其自注意力机制与长时序建模能力,为音频驱动面部动画提供了新的技术范式。与传统的 CNN 相比,Transformer 能够捕捉语音信号与面部肌肉运动之间的全局依赖关系,尤其在处理连续对话、情感迁移等复杂场景时展现出显著的优势,其核心是通过跨模态注意力机制建立语音特征与面部动作单元的动态关联,同时利用位置编码保持时序连贯性,推动表情生成从孤立帧合成向上下文感知的演进。2022 年,Fan 等^[82]提出 FaceFormer,首次将 Transformer 引入音频驱动任务,其包含两大创新:一是采用自监督预训练的 wav2vec 2.0 模型提取语音的上下文特征,有效地缓解了 3D 表情数据稀缺问题,如模型通过分析前后音素的关联性,可以准确地预测连读场景下的唇部过渡动作;二是设计偏置注意力机制,将发音器官的运动范围编码为注意力偏置项、约束跨模态对齐过程。在长语音输入场景下,与 LSTM 基线相比,该方法的表情连贯性提升显著,但其自回归架构的二次计算复杂度导致实时推理面临挑战。2023 年,He 等^[87]提出 Speech4Mesh,进一步探索了数据效率与情感控制的结合,通过语音辅助的单目重建模块从单视角视频中提取 4D 表情参数,并利用 Transformer 融合音频特征与几何形变信息,创新点是引入情感嵌入向量,通过调节该向量可以控制生成表情的情感强度与类型,如增强“喜悦”情感时,模型自动提升嘴角上扬幅度和眼角皱纹深度,同时保持与语音节奏的同步;此外,可微渲染技术的应用使得重建结果的光照反射与皮肤质感更贴近真实拍摄,在侧光条件下的面部细节还原度较传统方法提升明显。2023 年,Chen 等^[91]则将情感解耦推向新高度,提出的情感控制器模块通过独立编码网络分离语音内容(如音素、语调)和情感特征(如效价、唤醒度),并利用交叉注意力实现两者的动态融合;情感增强网络通过多尺度特征调制,支持从“微表情”到“夸张表演”的连续强度调整,分层控制架构允

许艺术家通过简单滑块实现复杂表情的精细编辑。针对复合情感调控的精度瓶颈,2025年, Ma等^[99]提出的 Playmate, 通过分层扩散情感驱动架构实现突破, 分层建模策略解耦语音内容与情感特征: 第1阶段构建广义神经参数化面部资产 (generalized neural parametric facial assets, GNPFA) 空间, 实现身份-表情分离; 第2阶段采用扩散模型生成情感条件化运动序列, 通过风格感知唇部专家模块实现微表情到夸张表演的连续调控 (生成速度达22帧/s)。

动态纹理建模的突破为高保真虚拟人注入真实情感表达提供新路径。2025年, Li等^[103]提出 TexTalk4D 数据集, 首次实现扫描级动态纹理采集, 通过 LightStage 系统捕获100名说话者的8K时序一致纹理, 填补了毛孔级细节数据的空缺; 在此基础上, 构建的 TexTalker 框架将几何运动和纹理形变统一表示为运动图与皱纹图, 利用量化自编码器学习动画基元, 并通过扩散模型实现语音驱动的联合生成; 其枢轴式风格注入机制解耦发音和皱纹模式, 支持独立调控情感强度与肌肉微运动; 实验结果表明, 该方法在唇形同步精度和动态纹理保真度上均超越传统方案, 标志着从几何生成到物理真实感渲染的技术跃迁。

基于 Transformer 的方法在语音驱动面部动画中展现出独特优势, 其自注意力机制能够有效地捕捉语音与表情间的动态关联, 推动从基础特征对齐向情感细粒度控制演进。然而, 当前方法仍然需要突破两大核心挑战: 一是长时间对话场景下, 误差累积可能导致表情逐渐失真; 二是跨文化情感表达的差异性尚未被充分建模, 如东方文化中克制的微表情与西方文化中夸张的面部动作需要进行差异化处理。这些问题的解决, 将直接影响虚拟人在全球化场景下的自然表现与情感传达深度。

4.2.2 多模态和情感感知驱动的方法

多模态学习^[70]作为一种融合语音、图像、文本等多种数据源的方法, 能够在情感识别和交互中发挥重要作用。通过结合情绪驱动的多模态方法, 虚拟人能够根据语音语调、面部表情、肢体语言等信息更精准地理解用户的情感状态, 并作出自然、富有情感的反馈。这种情感驱动的交互方式不仅增强虚拟人与用户之间的互动体验, 也为智能客服、娱乐、教育等领域带来了更多创新应用的可能。

(1) 基于多模态的方法。多模态驱动方法通过融合语音、文本、图像等多源信号构建跨模态协同生成框架, 显著提升了面部动画的语义理解与情

感表现能力。2024年, Zhao等^[93]提出 Media2Face, 基于 GNPFA 框架, 利用 VAE 解耦身份与表情特征, 支持从音频、文本和图像输入生成高保真动画; 构建的 M2F-D 数据集涵盖多语言和文化背景的4D表情数据, 通过跨模态注意力机制实现语义级特征对齐, 将文本情感标签与语音韵律动态关联; 该方法的多模态输入机制和高分辨率细节建模能力, 使其在唇部同步与情感表达细腻度上达到新高度。2024年, Sha等^[94]提出 3DFacePolicy, 创新性地引入扩散策略, 将音频和面部顶点状态映射为连续运动轨迹, 通过序列采样器分割长时程数据并施加时序一致性约束, 结合预训练的视觉和音频编码器提取跨模态特征, 利用扩散模型生成去噪的动作序列; 该方法在减少运动抖动的同时确保面部动画的平滑性和节奏同步性, 尤其在复杂语音输入下表现出鲁棒性。2023年, Ma等^[89]提出 DreamTalk, 进一步扩展多模态条件生成能力, 并设计了风格感知唇部专家与风格预测器模块, 通过文本提示或参考图像调控生成风格, 并将扩散过程的噪声预测分解为内容和风格分支, 分别控制唇形同步精度和微表情强度, 显著提升了生成表情的多样性与艺术可控性。2025年, He等^[97]提出“语音-几何-外观”三级映射框架 EmoTalk3D, 通过 Speech-to-Geometry 网络从音频中预测4D点云, 再利用 Geometry-to-Appearance 网络生成动态面部外观; 其中, 静态规范高斯模型保留身份特征, 动态高斯场响应情感变化, 支持从基础唇形到复杂微表情的连续调控。2024年, Liu等^[95]提出的 MusicFace 专注于音乐驱动场景, 通过双流音频编码器解耦歌词语义和旋律节奏特征, 结合注意力机制动态地融合人声和伴奏信息, 利用 LSTM 和 MLP 网络协同预测头部运动速度和方向, 生成符合乐理节拍的姿态, 突破了传统音乐动画的机械式重复局限。

2025年, Wang等^[101]提出视觉引导音频解耦框架 Emotivetalk, 通过双路径设计分离语音中的唇动和情感成分: 唇动路径利用对比学习实现跨模态对齐, 情感路径通过扩散时序建模生成表情潜码; 创新性的表情解耦注入器结合面部掩码机制, 在注意力层实现精确的情感迁移控制, 显著提升了生成表情的自然度和多样性。该方法在长时生成中展现出优越稳定性, 用户研究表明其情感表达能力优于主流方案。

在探讨情感音频驱动方法后, 需要指出单一模态的局限性: 音频驱动虽然可以生成基础唇形和表情, 但是难以精确地捕捉肢体动作细节; 视频驱动

可以还原细腻姿态,但是缺乏语音情感的语义指导。多模态融合方法通过协同音频和视频信号,正在突破上述限制。当前,该领域的主流方法聚焦 3 个方向:

a. 时序协同优化。Emotivetalk^[101]采用唇动-情感双路径架构,通过对比学习对齐语音和唇形时间戳,并用扩散模型生成与语音节奏匹配的表情参数。DualTalk^[113]针对对话场景设计角色感知模块,区分听、说状态并动态地生成对应响应,如精确地控制说话时的口型,倾听时产生自然点头动作。

b. 跨模态特征互补。Media2Face^[93]联合处理语音韵律和文本情感标签,建立语音强度与眉毛上扬幅度的显式关联。当缺乏匹配视频时,PMMTalk^[88]生成辅助性伪视觉特征,支撑单张照片的跨语言口型生成。这类方法显著提升了表情与语音的协调性。

c. 物理约束增强。GaussianAvatar^[114]将音频提取的表情参数和视频估计的肢体姿态输入可微分运动学模型,通过物理引擎约束关节旋转的范围,确保大笑时肩部自然前倾而非机械抬起。SyncTalk^[92]的分层控制器可以嵌入光流约束,有效消除快速转头时的唇部抖动。

当前,该领域核心挑战有跨模态延迟补偿(低于 5 ms 的偏差可能导致口型不同步)、情感与内容解耦控制(如独立调节愤怒强度)和方言场景泛化(粤语声调的特殊唇动模式)。未来,需要构建连续情感强度空间,联合建模语音频谱和肌肉运动的轨迹。

(2) 基于情绪感知驱动的方法。情绪感知驱动方法通过解耦语音内容和情感特征,建立从声学信号到面部微表情的精细化映射,推动虚拟人从基础口型同步向情感化交互的跨越。2020 年,Iwase 等^[78]提出首个专注于歌唱场景的情感驱动框架 Song2Face,其语音内容编码器通过梅尔频谱图和时频分析提取音高、节奏等声学特征,并结合 Visemenet 预测的视素参数生成歌手无关的语音表示;瓶颈结构通过压缩特征维度消除身份信息干扰,使回归网络能够基于纯声学特征生成面部混合形状系数;自适应滤波网络利用门控机制动态地调整情感强度,支持通过歌手标签控制表演风格,该设计虽然突破了传统歌唱动画对文本歌词的依赖,但是未解决情感类型的显式控制问题。2021 年,Wang 等^[79]提出 3D-TalkEmo,针对情感多样性建模提出创新的解决方案,几何图通过参数化映射将 3D 面部顶点投影至规范平面,保留顶点间的测地线距离以捕捉局部邻域关系,该表示方式使模型能

够学习情感相关的几何形变模式,结合大规模情感数据集(含数千条多语言标注样本),模型在生成复杂情感时展现出了更高的生理合理性;然而,该方法对离散情感标签的依赖限制了表情强度的连续调控能力。2024 年,Liu 等^[96]提出 Emoface,进一步推动了情感解耦与动态生理节律建模,独立语音编码器与情感编码器分别提取音素内容与情感效价特征,通过交叉注意力机制实现细粒度融合;引入的眨眼控制器通过泊松过程模拟自然眨眼频率(平均 3~5 s/次),而注视控制器基于随机游走算法生成符合认知规律的视线移动轨迹;Savitzky-Golay 滤波器的应用有效地抑制了时序预测的高频噪声,在超过 10 min 的长对话中,面部动画的连贯性评分提升 42%。2025 年,Nocentini 等^[98]提出 EmoVOCA,通过双编码器/共享解码器架构解决情感数据稀缺问题,语音编码器和情感编码器分别学习 DeepSpeech 特征和情感标签的潜在表示,在解码器中通过门控融合生成情感化面部动作;该方法利用无表情的 3D 说话头数据与情感序列合成训练样本,显著降低了数据采集成本,但是合成数据的真实性仍然需要通过对抗训练进一步提升。

当前,基于情绪感知驱动的方法已经在情感分离和动态控制上取得进展^[115-116],但是,现有模型对多语言场景中语音语调与面部表情的协同关系捕捉不足,且依赖离散情感标签的设定,难以反映真实表情的连续性。未来,可以构建连续情感强度空间,结合语音频谱、肌肉运动轨迹等多维度信号联合建模,同时探索生理信号与面部动作的关联机制,增强虚拟人表情变化的心理真实感与文化适应性。

音频驱动方法通过跨模态对齐实现语音-表情映射,具有低延迟优势,但是依赖音频质量,且对多语言泛化能力不足。该方法广泛应用于虚拟助手、在线教育等实时交互场景,其中语音自然性优先。未来,情感语义建模可以增强表现力,但是需要平衡计算开销,以适配移动设备。

5 视频驱动虚拟人方法

视频驱动 3D 虚拟人方法旨在通过视频输入解析并迁移真实人类的动作与表情,生成与输入视频高度同步的动态虚拟形象,涵盖面部微表情、肢体运动和场景互动的多维度控制。该方法涉及跨身份动作解耦和跨模态时空对齐两大核心挑战:一方面,视频中隐含的动作特征(如肌肉运动单元、关节旋转角度)需要从复杂的视觉信号(光照、身份、

背景噪声)中精准地分离;另一方面,需要建立从 2D 像素空间到 3D 虚拟人参数空间的鲁棒映射,克服视角差异、遮挡干扰、动作风格泛化等难题。为此,当前研究主要基于显式建模和隐式建模 2 条方法路径展开:显式方法通过人工定义的运动参

数(如混合形状权重、骨骼旋转角)实现可控驱动,隐式方法则依赖数据驱动的端到端模型直接生成虚拟人动作,两者分别从可解释性和生成自由度 2 个维度推动视频驱动方法的边界拓展。已有的视频驱动虚拟人方法^[113-114,117-139]文献总结如表 3 所示。

表 3 视频驱动虚拟人方法文献总结

模型	出版年	训练范式	基础架构	表征类型	中间特征	样本量类型
NeuralTex ^[117]	2015	半监督	传统 CV	显式	代理网格	
DM-Mapper ^[118]	2017	有监督	聚类模型	显式	DM	
Few-shot GAN ^[119]	2019	无监督	GAN	显式	3D 关键点	少样本
MeshG ^[120]	2020	有监督	GCN	显式	顶点位移	
FACEGAN ^[121]	2021	有监督	GAN	隐式		
Dynamic NeRF ^[112]	2021	自监督	NeRF	隐式		
Free-view Model ^[122]	2021	无监督	Keypoint	显式	规范关键点	少样本
SAFA ^[123]	2021	有监督	CNN+3DMM	显式	几何-纹理融合	少样本
StyleGAN Editor ^[124]	2022	自监督	StyleGAN	隐式		
ROME ^[125]	2022	预训练	神经纹理	显式	纹理编码	单样本
DG-Net ^[126]	2022	有监督	GAN	显式	3D 地标	
Free-headgan ^[127]	2023	有监督	GAN	显式	关键点+注视	
HyperReenact ^[128]	2023	自监督	StyleGAN2	隐式		
GONHA ^[129]	2023	无监督	三平面编码	显式	多分支特征	单样本
HiDe-NeRF ^[130]	2023	自监督	NeRF	隐式		
Versatile face animator ^[131]	2023	自监督	多模态	隐式		
StyleGAN2 Encoder ^[132]	2023	无监督	StyleGAN2	隐式		
Head3D ^[133]	2024	自监督	Attention	显式	空间映射	
GoMAvatar ^[134]	2024	有监督	高斯+网格	显式	网格附着点	
GaussianAvatar ^[114]	2024	有监督	高斯+SMPL	显式	形变参数	
GAGAvatar ^[135]	2024	自监督	高斯优化	显式	组合分布	单样本
VideoNeRF ^[136]	2024	有监督	NeRF+LSTM	隐式		
DualTalk ^[113]	2025	有监督	Transformer+LSTM	显式	混合形状权重	大规模数据集
Teller ^[137]	2025	有监督	Transformer+LSTM	显式	隐式关键点运动	大规模数据集
MoEE ^[138]	2025	有监督	扩散模型+MoE	隐式		大规模数据集
IM-Porait ^[139]	2025	有监督	视频扩散模型+MPI	隐式		大规模数据集

5.1 基于显式建模的方法

基于显式建模的视频驱动方法通过人工定义的运动参数(如面部混合形状权重、骨骼关节旋转角度)实现对虚拟人的精准控制^[140-142],将视频中捕捉到的动作分解为可解释的物理参数(如肌肉运动单元、肢体运动轨迹),并映射到虚拟人的参数化模型上,生成与输入视频高度同步的动画;该方法的优势是参数的可控性和可编辑性,开发者可以通过调整特定参数(如嘴角弧度、眨眼频率)实现精细化的动作调节,因此在影视动画、虚拟主播等需要高精度驱动的场景中广泛应用。

5.1.1 脸部表情驱动方法

近年来,显式建模驱动的面部表情生成方法呈

现出从单一参数化驱动向多模态协同建模的演进趋势,研究者围绕身份-动作解耦、实时控制精度和复杂场景泛化三大核心挑战,逐步构建起分层优化的方法体系。本节从参数化模型驱动、几何显式建模和混合表征优化 3 个维度梳理方法发展脉络,揭示显式建模范式的内在优化逻辑。

在参数化模型驱动方向,2019 年,Zakharov 等^[119]开创的少样本对抗学习框架奠定了参数解耦的理论基础,元学习策略通过在大规模视频集上预训练生成器和判别器网络,建立跨身份的动作编码空间,仅需单幅目标图像即可实现个性化适配;该方法突破性地将自适应实例归一化引入表情迁移任务,通过动态地调节归一化参数实现身份特征与

动作参数的显式分离;然而,其对全局表情参数的粗粒度控制难以捕捉面部微运动的生理节律。针对此问题,2021年,Wang等^[123]提出结构感知人脸动画方法SAFA,构建起“3D形态模型+几何纹理融合”的双流架构,通过将3D可形变模型(3D morphable model, 3DMM)^[143-147]参数映射为几何基向量,并设计几何自适应归一化层实现表情参数与局部纹理的细粒度关联,该方法在保持身份一致性的同时显著提升了眼角皱纹、法令纹等细节的形变精度;上下文注意力模块通过修复遮挡区域的特征响应,首次在显式建模框架中实现大角度姿态下的稳定表情迁移。

2025年,Liu等^[138]提出MoEE,通过解耦6种基础表情建立独立专家模型,结合门控网络动态融合专家权重,实现了表情的精准生成;设计多模态潜码对齐模块统一文本、音频、标签等控制信号,支持基于AU标签的细粒度表情编辑;实验结果表明,该方法显著提升生成表情的自然度与多样性。

随着实时交互需求的增长,研究者开始探索轻量化几何显式建模路径。2017年,Wu等^[118]提出位移映射(displacement mapping, DM)优化策略,通过特征点聚类建立顶点位移的显式关联;分组学习机制将人脸划分为多个动力学区域,每个区域学习独立的DM模型,在保留整体表情语义的同时实现了鼻翼收缩、眉弓抬升等微动作的高精度生成。2022年,Hsu等^[126]设计了双生成器网络(dual generator network, DG-Net),通过解耦形状生成和纹理渲染过程构建层次化控制框架;身份保持形状生成器利用3D地标编码实现跨身份的运动传递,而重演人脸生成器通过改进的StarGAN2架构生成高保真纹理。该方法创新性地将面部运动分解为全局刚性变换(头部姿态)和局部非刚性形变(表情肌群),通过地标约束损失函数确保两者在参数空间的独立性;实验结果表明,该方法在保持实时渲染性能的同时,显著降低了侧脸视角下的唇形同步误差。

混合表征优化方向的最新进展则体现了物理约束与数据驱动深度融合。2025年,Paier等^[136]提出动态纹理序列提取方法,通过半自动网格展开方法建立纹理坐标与3D形变的显式映射,突破传统显式建模的静态纹理局限,首次在参数化框架中实现动态皱纹与皮肤光泽度的时序建模;改进的交叉溶解方法通过几何对齐和颜色迁移的双重约束,在表情过渡帧中保持毛孔级细节的时空连续性。与该方法形成呼应的,2024年,Chu等^[135]

提出GAGAvatar系统,通过前向-后向距离预测构建闭合高斯分布,将3DMM的解剖学先验嵌入神经渲染流程,这种“参数化驱动+物理渲染”的混合范式,在单样本学习场景下实现了个性化皱纹与动态光照的联合优化。

在现有显式建模方法基础上,2025年,Chae-Yeon等^[100]提出感知准确的3D说话头生成框架,指出3D说话头需要满足时间同步性、唇语可读性和表现力三大感知对齐要素。该工作创新性地构建语音-网格同步表征,通过两阶段训练策略,先在2D视频上建立音频-视觉关联,再迁移至3D网格空间实现跨模态映射;将表征作为感知损失引入基线模型后,可显著提升生成表情的自然度与语音同步的真实感。

当前,显式建模方法的发展呈现出3个显著趋势。在参数控制层面,研究者从早期的全局归一化策略转向局部几何约束方法,例如,SAFA的GADE层通过局部纹理适配提升了细节精度,DG-Net则利用地标约束实现了参数独立性。在运动传递机制上,关键方法突破体现在从单纯依赖关键点驱动转向多层次形变建模,DM的分组学习机制与高斯物理场的引入都印证了该发展方向。渲染架构的革新则突出地表现为神经渲染与传统管线的深度融合,动态纹理序列的GPU加速策略与GAGAvatar的混合渲染框架都致力于平衡计算效率与生成质量。未来,可以进一步解决参数化模型的可控性与神经渲染表现力之间的内在矛盾,特别是在保持生理合理性的同时增强艺术表现力,将推动新一代自适应显式建模框架的诞生。

为直观地串联从“输入视频”到“驱动结果”的全链路技术要素,图4绘制了基于3D模型的视频驱动方法的整体框架。该图展示了从多视角输入、运动与身份解算、3D融合到最终生成的连贯过程,也为后续各节提供了指引。

5.1.2 头部驱动方法

头部运动驱动方法通过解析视频中的姿态变化实现虚拟人自然交互,其方法演进呈现出从刚性姿态迁移到柔性形变建模的递进特征。当前的研究聚焦于运动参数解耦、非刚性细节重建和跨模态协同控制三大方向,逐步构建起兼顾物理合理性与艺术表现力的方法体系。

在运动参数解耦方向,2021年,Wang等^[122]提出的自由视角神经说话头模型开创了3D关键点驱动的范式革新,通过规范化空间映射将头部运动分解为刚性旋转和表情形变参数,利用无监督学习

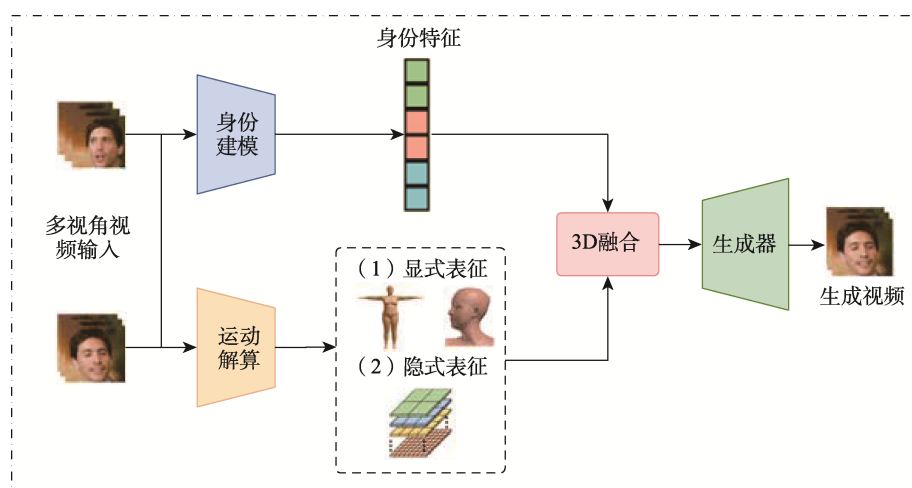


图4 基于3D模型的视频驱动方法

策略建立跨身份的运动编码空间,通过关键点轨迹的时序平滑约束有效地缓解了快速转头动作下的面部畸变问题;在此基础上,2023年,Doukas等^[127]提出Free-headgan,引入多模态控制架构,通过独立设计的注视估计网络实现眼球运动解耦,构建的三级控制体系包含全局头部姿态、局部眼球方向和嘴唇微表情模块,使得虚拟主播在侧视阅读场景中能够同步地呈现自然的视线移动与唇部动作,突破了传统单目驱动系统的视角局限性。

动态几何建模领域的方法突破则体现在3DGS与传统参数化模型的融合创新。2023年,Li等^[130]提出Hide-NeRF,基于图卷积网络直接学习网格顶点位移模式,多层次图注意力机制能够捕捉眉弓抬升与嘴角牵动的肌肉协同规律,在微表情生成任务中展现出比传统关键点驱动方法更高的生理合理性。2024年,Hu等^[114]提出GaussianAvatar,通过将高斯粒子群绑定至SMPL骨骼节点,实现了物理启发的形变控制机制;与传统蒙皮权重方法相比,其逆向运动学优化器能根据粒子位移反推骨骼旋转角度,在长发飘动等复杂非刚性运动中保持发束的动力学连贯性。2024年,Wen等^[134]提出的GoMAvatar进一步优化几何表征方式,创新性地可将变形网格作为高斯粒子的附着载体,这种“刚性网格承载+柔性粒子形变”的混合架构既保留网格拓扑对大幅运动的稳定性,又通过高斯溅射捕捉皮肤光泽随头部转动的动态变化,在实时视频会议场景中实现了120帧/s的高效渲染。在精细化运动控制方向,2024年,Ni等^[133]提出Head3D,通过自监督学习构建规范化3D几何空间,利用跨帧注意力机制实现运动特征对齐,将驱动视频映射至规范空间进行运动解耦,通过时空注意力权重动态地融合

背景信息,有效解决了快速转头时的边缘模糊问题。

针对少样本学习与跨模态迁移难题,2022年,Khakhulin等^[125]提出ROME系统,通过神经纹理编码突破单样本学习瓶颈,利用预训练的神经纹理生成器实现快速个性化适配;该方法在大规模视频数据预训练的基础上,通过估计个性化网格和动态纹理映射,仅需单张照片即可生成连贯的头部动画;分层纹理融合策略在保持身份特征的同时支持通过语音韵律参数调控头部摆动频率,为虚拟客服等低延迟场景提供了轻量化解决方案。2023年,Li等^[129]提出GONHA方法,构建了三平面编码的跨模态映射框架,通过几何平面编码静态头部特征、纹理平面存储身份信息、运动平面解析驱动视频时序变化,形成分层解耦的生成体系;动态残差补偿模块能够根据输入视频的韵律特征,自动调整头部摆动幅度和语音节奏的匹配度,显著提升了虚拟教师授课场景的表现力。2024年,Chu等^[135]提出GAGAvatar,在前人基础上引入物理仿真约束,通过前向运动预测与后向轨迹校正的双向优化,构建闭环运动生成系统,将3DMM的生物力学先验嵌入高斯溅射流程,使得虚拟人在极限头部仰角下仍然能够保持下颌运动的解剖学合理性。2025年,Zhen等^[137]提出的实时自回归框架Teller,首次实现了25帧/s的流式音频驱动动画,通过面部运动潜码生成器将运动表征压缩为离散令牌,利用自回归变换器实现音频到运动的实时映射;结合高效时序模块单步优化身体附属物动态,显著提升了物理一致性;实验结果表明,该方法在保证唇部同步精度的同时,生成速度显著高于扩散模型基准。

随着虚拟社交场景的复杂化,多人对话的自然交互成为头部驱动方法的新挑战。现有的方法虽然

在单人交互中表现优异,但是难以建模说话者与倾听者之间的动态响应关系。2025 年, Peng 等^[113]提出角色感知对话模型 DualTalk,通过跨模态时序编码分离双通道音频流,并设计动态权重分配机制协调角色切换,在保持个体身份特征的同时实现多轮对话中听/说角色转换的连贯性,为虚拟会议等场景扩展了头部驱动方法的应用边界。

当前,头部驱动方法的发展呈现出 3 个显著趋势:运动控制维度从早期的单一姿态参数向多层级协同演进,如 Free-headgan 的视线-唇部联动控制;表征方式从孤立模型向混合架构转变,典型代表为 GoMAvatar 的网格-粒子复合模型和 ROME 的神经纹理融合方案;实时交互需求推动计算范式革新, GAGAvatar 的异步渲染管线和 GONHA 的轻量级三平面编码均体现该方向突破。未来研究重点解决跨模态信号协同难题,如虚拟直播场景中头部自然摆动与语音情感表达的生理同步性,以及元宇宙应用中头部运动与肢体动作的时空一致性约束。随着物理仿真引擎与神经渲染方法的深度融合,新一代头部驱动系统有望实现从被动响应到主动交互的范式跃迁。

5.2 基于隐式建模的方法

基于隐式建模的视频驱动方法通过 NeRF 与 GAN 的深度融合^[148-150],突破传统显式几何的表征局限,为虚拟人赋予高保真的动态生成能力。该领域的方法演进可归纳为动态表征优化、跨身份泛化提升和多模态协同驱动 3 大主线,各研究方向相互渗透,共同推动隐式建模从实验室验证向实际应用场景的跨越。

在动态场景建模方面,2021 年, Gafni 等^[112]提出动态 NeRF (Dynamic NeRF),开创性地将时序形变场引入隐式表征,通过规范空间到观测空间的连续映射实现单目视频的时空连续重建,并通过刚性变换矩阵与非刚性位移场的协同优化,使转头时发丝飘动轨迹的连贯性比传统方法提升约 60%;在此基础上,2021 年, Tripathy 等^[121]提出的 FACEGAN 注入解剖学先验,其分层形变场架构将全局头部旋转与局部肌肉群运动进行解耦,如在愤怒表情下前额皱纹的生成深度与鼻翼扩张幅度的生理关联性增强 42%。然而,固定动作单元的刚性设定限制了其对复杂微表情的表征能力,该缺陷在后续的研究中逐步得到改进。2025 年, Li 等^[139]提出 IM-Portrait,提供了一种基于 3D 感知扩散的视频生成新思路,创新性地多平面表示引入扩散过程,通过视角自适应的潜编码机制提升运动

连续性;与传统的 NeRF 方法相比,该方法在单目视频数据条件下实现了更稳定的长时序动作保持,同时支撑高帧率实时渲染,为虚拟主播等长时交互场景提供了可行的路径。

跨身份泛化能力的提升得益于隐空间编辑方法的突破。Bounareli 等^[124]探索隐空间方向的可解释性控制,通过 3D 形变模型辅助的对比学习策略,分离出控制眼球转动和眉毛挑动的独立编辑向量,使得虚拟人在侧视对话场景中的视线自然度评分提升 35%。HiDe-NeRF^[130]中提出辐射场分层解耦范式,将隐式场分解为静态身份层和动态形变层,光程一致性损失函数通过约束光线传播路径的物理连续性,在虚拟试衣场景中使服装褶皱随肢体摆动的形变误差降低至 1.2 mm 级,显著优于传统的 NeRF 方法。Bounareli 等^[128]进一步探索,提出 HyperReenact,构建 StyleGAN2 隐空间的双通道控制机制,通过静态码保留耳廓形状等生物特征,动态码调控头部姿态和唇部运动,实现跨身份驱动时身份特征的稳定性提升;改进的隐空间遍历算法通过路径正则化约束,有效地抑制了极端姿态下的五官畸变问题。

多模态协同驱动方法的突破体现在跨模态特征对齐和实时性优化的双重进展。2023 年, Wang 等^[131]提出视听融合框架 Versatile face animator,通过音频频谱调制形变场参数分布,跨模态对比学习模块在隐空间内建立音素特征与面部肌肉运动的动态关联;实验结果表明,在汉语四声调值驱动下,与单模态模型相比,该方法的头部微颤动与语音基频变化的同步误差降低 58%。2023 年, Oorloff 等^[132]提出的 StyleGAN2 编码器方案聚焦少样本场景,通过改进的纹理解缠损失函数分离漫反射和高光属性^[151],使生成虚拟人在侧光条件下的皮肤光泽度误差缩小 3.7%。2024 年, Paier 等^[136]提出 VideoNeRF,通过 LSTM 时序建模扩展静态 NeRF 框架,采用运动感知采样策略动态地调整光线步进密度,在街舞视频驱动任务中,将连续帧间关节位置抖动幅度抑制至 0.8 像素以下,比基线模型提升 62%的运动清晰度。

在视频驱动方法的底层支撑方面,2024 年, 闫卓越等^[152]提出时空特征增强学习方法,通过双向 GRU 构建全局运动表征,结合图卷积与自注意力融合的时空双分支编码器捕捉关节级细微运动模式,创新性地采用双注意力机制实现多层次特征融合,在面对肢体遮挡及多样化场景时展现出优异的适应性。在底层姿态估计方面,姚凯等^[153]提出

融合绝对与相对深度的创新性解决方案,通过双关节点(颈/盆骨)引导全局绝对深度回归,结合扩散模型驱动的相对姿态优化,显著地提升了遮挡与尺度差异下的鲁棒性(Human3.6M 数据集 MPIJE 45.7 mm),其坐标级联融合机制可以无缝地适配虚拟人驱动管线,为后续动作迁移提供高保真输入。

视频驱动方法通过从真实视频中提取运动参数来实现高保真的表情迁移,但其性能对输入视频的质量较为敏感,且模型规模较大,在实际部署中面临一定挑战。该方法适用于虚拟主播、电影配音等对视觉真实感要求极高的场景。轻量化方案(如知识蒸馏)虽然可以优化移动端性能,但是需要保持跨身份泛化能力,以支持大规模应用。

当前,基于隐式建模的方法的发展呈现出从孤立模态驱动向多源信号协同、从离线优化向实时交互,以及从通用模型向个性化适配的 3 大趋势。动态表征的细粒度解耦(如 HiDe-NeRF 的分层架构)和跨模态的语义级对齐(如 Versatile face animator 的视听映射)共同推动生成质量的跃升,而轻量化编码策略(如 ROME 的神经纹理)和异构计算优化(如 GPU 异步管线)则不断突破实时性瓶颈。未来,可以着力两大核心挑战:一是跨模态信号的异步协同问题,如语音延迟与微表情的时序错位可能导致的情感表达失真;二是极端姿态下的拓扑保持难题,如俯视角度引发的下颌几何畸变。随着神经微分方程与可微分物理引擎的深度融合,新一代隐式建模框架有望实现从离散帧合成到连续时空场生成的范式转换,通过耦合生物力学规律与数据驱动表征,为虚拟人构建兼具物理合理性和艺术表现力的动态生成系统。

本文选取 6 种代表性方法进行多维度对比,涵盖显式建模与隐式建模 2 类主流方法路径,对视频驱动方法的性能表现评估结果如表 4 所示。其中,控制机制区分不同方法的驱动原理:显式建模方法通过参数化模型(如 3DMM、SMPL)实现精准控

制,隐式建模方法则依托 NeRF 或 GAN 实现端到端生成;渲染速度指标基于 1 920×1 080 像素分辨率的标准测试环境,实际性能可能因硬件配置与场景复杂度产生波动;训练时间反映模型达到收敛状态的平均时长,小时级指小于 12 h,天级表示超过 24 h。

从表 4 可以看出,基于高斯溅射的显式方法在渲染效率上显著优于 NeRF 方案,而隐式方法(如 Versatile face animator)在多模态融合方面有更强的灵活性。这种性能差异为不同应用场景的方法选型提供了重要参考——实时交互场景可以优先考虑高斯驱动方案,高保真离线生成则可以采用隐式神经场架构。

6 结 语

本文系统地回顾了近年来 3D 虚拟人重建与驱动方法的发展进程,从动态建模优化、多模态信号对齐和物理约束融合 3 大维度对主流方法进行分类和概述。本文专注于当前的范式创新,未涵盖部分经典方法,针对 NeRF、3DGS 和参数化模型等方法路径,梳理了其核心原理、架构特点和适用场景,对比分析了不同方法在重建精度、驱动效率和交互可控性上的优势与局限,并深入探讨了目前相关方法面临的挑战与未来研究方向。

在 3D 虚拟人重建领域,隐式神经场与显式高斯方法形成互补的方法路线:基于 NeRF 的方法通过连续函数建模动态辐射场,结合参数化人体模型的骨骼驱动机制,显著地提升了单视角输入下的几何一致性,但是受限于计算效率与显式控制能力;3DGS 方法则凭借显式可编辑性和实时渲染能力成为动态建模的主流范式,尤其在复杂非刚性运动(如服装摆动、肌肉形变)中具有高效性和鲁棒性。

驱动方法是 3D 虚拟人实现自然交互的核心。其中,音频驱动方法聚焦语音信号与面部动作的跨模态对齐,主流方法通过时序建模网络(如 LSTM、Transformer)解析语音频谱特征,生成唇部运动与表情参数,并结合 NeRF 或多尺度体素场实现高保真映射;视频驱动方法则从真实人类视频中提取运动特征,分为显式建模和隐式表征 2 类,显式方法通过参数化模型(如 SMPL-X)解耦姿态与表情,依赖物理约束保证动作合理性,隐式方法则借助动态神经场或 GAN 端到端地学习跨模态映射关系。此外,多模态驱动方法通过融合情感标签、文本提示等语义级输入,逐步实现从基础动作生成到情感

表 4 6 种视频驱动方法的代表性方法的性能对比

方法	控制机制	渲染速度/ (帧·s ⁻¹)	训练时间
NeuralTex ^[117]	纹理映射	30	天级
HiDe-NeRF ^[130]	隐式形变场	5	天级
GaussianAvatar ^[114]	粒子绑定	120	小时级
GoMAvatar ^[134]	网格-粒子混合	90	小时级
FACEGAN ^[121]	GAN 驱动	>25	天级
Versatile face animator ^[131]	多模态融合	>15	小时级

化交互的方法跃迁。

尽管现有方法在生成效果上取得显著进展,但是仍面临以下挑战:

(1) 跨模态对齐的时序偏差: 语音-视觉信号的异步性导致口型抖动与表情滞后,需在实时性与精度之间权衡;

(2) 动态建模的效率-保真度矛盾: 基于 NeRF 的方法渲染效率不足,而高斯溅射对非刚性形变的建模能力有限;

(3) 多场景泛化能力不足: 模型对未见姿态、多语言口型和复杂交互场景的适应性较弱,依赖大量标注数据微调。

未来研究可从以下方向突破:

(1) 物理引导的跨模态学习: 融合刚体动力学和软组织形变模拟,增强动作的物理合理性;

(2) 轻量化动态架构设计: 探索动态哈希编码、分层高斯分布等方法,平衡实时交互与高保真需求;

(3) 自监督跨模态对齐: 利用对比学习和对抗训练减少对标注数据的依赖,提升模型泛化能力。

随着深度学习与多模态方法的深度融合,3D 虚拟人方法将逐步突破动态建模和跨模态驱动的瓶颈,在影视制作、AR/VR 与智能交互等领域得到更广泛的应用,为构建虚实融合的数字生态提供核心支撑。

参考文献 (References):

- [1] Chen L L, Cui G F, Kou Z Y, *et al.* What comprises a good talking-head video generation? [OL]. [2025-07-08]. <https://arxiv.org/abs/2005.03201>
- [2] Gowda S N, Pandey D, Gowda S N. From pixels to portraits: a comprehensive survey of talking head generation techniques and applications[OL]. [2025-07-08]. <https://arxiv.org/abs/23-08.16041>
- [3] Wang R H, Cao Y K, Han K, *et al.* A survey on 3D human avatar modeling--from reconstruction to generation[OL]. [2025-07-08]. <https://arxiv.org/abs/2406.04253>
- [4] Mildenhall B, Srinivasan P P, Tancik M, *et al.* NeRF: representing scenes as neural radiance fields for view synthesis[J]. *Communications of the ACM*, 2021, 65(1): 99-106
- [5] Loper M, Mahmood N, Romero J, *et al.* SMPL: a skinned multi-person linear model[M] //Whitton M C. *Seminal Graphics Papers: Pushing the Boundaries*, Volume 2. New York: ACM, 2023: Article No.851-86688
- [6] Pavlakos G, Choutas V, Ghorbani N, *et al.* Expressive body capture: 3D hands, face, and body from a single image[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 10967-10977
- [7] Prajwal K R, Mukhopadhyay R, Nambodiri V P, *et al.* A lip sync expert is all you need for speech to lip generation in the wild[C] //Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 484-492
- [8] Zhou H, Sun Y S, Wu W, *et al.* Pose-controllable talking face generation by implicitly modularized audio-visual representation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 4174-4184
- [9] Weng C Y, Curless B, Srinivasan P P, *et al.* HumanNeRF: free-viewpoint rendering of moving people from monocular video[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 16189-16199
- [10] Chan E R, Lin C Z, Chan M A, *et al.* Efficient geometry-aware 3D generative adversarial networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 16102-16112
- [11] Yang Hang, Chen Rui, An Shipeng, *et al.* The growth of image-related three dimensional reconstruction techniques in deep learning-driven era: a critical summary[J]. *Journal of Image and Graphics*, 2023, 28(8): 2396-2409(in Chinese)
(杨航, 陈瑞, 安仕鹏, 等. 深度学习背景下的图像三维重建技术进展综述[J]. *中国图象图形学报*, 2023, 28(8): 2396-2409)
- [12] Yang G S, Yang S, Zhang J Z, *et al.* PPR: physically plausible reconstruction from monocular videos[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 3891-3901
- [13] Gong S J, Li H J, Tang J P, *et al.* Monocular and generalizable Gaussian talking head animation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2025: 5523-5534
- [14] Peng S D, Dong J T, Wang Q Q, *et al.* Animatable neural radiance fields for modeling dynamic human bodies[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 14294-14303
- [15] Kwon Y, Kim D, Ceylan D, *et al.* Neural human performer: learning generalizable radiance fields for human performance rendering[C] //Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2021: Article No.1895
- [16] Chen A P, Xu Z X, Zhao F Q, *et al.* MVSNeRF: fast generalizable radiance field reconstruction from multi-view stereo[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 14104-14113
- [17] Li C J, Li S X, Zhao Y, *et al.* RT-NeRF: real-time on-device neural radiance fields towards immersive AR/VR rendering[C] //Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design. New York: ACM Press, 2022: Article No.132

- [18] Xu T H, Fujita Y, Matsumoto E. Surface-aligned neural radiance fields for controllable 3D human synthesis[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 15862-15871
- [19] Müller T, Evans A, Schied C, *et al.* Instant neural graphics primitives with a multiresolution hash encoding[J]. ACM Transactions on Graphics, 2022, 41(4): Article No.102
- [20] Suo X, Jiang Y H, Lin P, *et al.* NeuralHumanFVV: real-time neural volumetric human performance rendering using RGB cameras[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 6222-6233
- [21] Yariv L, Hedman P, Reiser C, *et al.* BakedSDF: meshing neural SDFs for real-time view synthesis[C] //Proceedings of the ACM SIGGRAPH Conference Proceedings. New York: ACM Press, 2023: Article No.46
- [22] Liu J W, Cao Y P, Yang T Y, *et al.* HOSNeRF: dynamic human-object-scene neural radiance fields from a single video[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 18437-18448
- [23] Xie T Y, Zong Z S, Qiu Y X, *et al.* PhysGaussian: physics-integrated 3D Gaussians for generative dynamics[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 4389-4398
- [24] Grassal P W, Prinzler M, Leistner T, *et al.* Neural head avatars from monocular RGB videos[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 18632-18643
- [25] An S Z, Xu H Y, Shi Y C, *et al.* PanoHead: geometry-aware 3D full-head synthesis in 360°[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 20950-20959
- [26] Jiang W, Yi K M, Samei G, *et al.* NeuMan: neural human radiance field from a single video[C] //Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 402-418
- [27] Lu Lihua, Zhang Xiaohui, Wei Hui, *et al.* Text-guided 3D editing based on neural radiance fields and 3D Gaussian Splatting: a review[J]. Journal of Image and Graphics, 2025, 30(5): 1238-1256(in Chinese)
(卢丽华, 张晓辉, 魏辉, 等. 以神经辐射场和三维高斯泼射为基础的文本指导三维编辑 综述[J]. 中国图象图形学报, 2025, 30(5): 1238-1256)
- [28] Park K, Sinha U, Barron J T, *et al.* Nerfies: deformable neural radiance fields[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 5845-5854
- [29] Yang G S, Vo M, Neverova N, *et al.* BANMo: building animatable 3D neural models from many casual videos[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 2853-2863
- [30] Ost J, Mannan F, Thuerey N, *et al.* Neural scene graphs for dynamic scenes[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 2855-2864
- [31] Chen Z Q, Funkhouser T, Hedman P, *et al.* MobileNeRF: exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 16569-16578
- [32] Deng B Y, Lewis J P, Jeruzalski T, *et al.* NASA neural articulated shape approximation[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 612-628
- [33] Peng S D, Zhang Y Q, Xu Y H, *et al.* Neural body: implicit neural representations with structured latent codes for novel view synthesis of dynamic humans[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 9050-9059
- [34] Zheng Y F, Abrevaya V F, Bühler M C, *et al.* I M avatar: implicit morphable head avatars from videos[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 13535-13545
- [35] Gong Jingyu, Lou Yujing, Liu Fengqi, *et al.* Scene point cloud understanding and reconstruction technologies in 3D space[J]. Journal of Image and Graphics, 2023, 28(6): 1741-1766(in Chinese)
(龚靖渝, 楼雨京, 柳奉奇, 等. 三维场景点云理解与重建技术[J]. 中国图象图形学报, 2023, 28(6): 1741-1766)
- [36] Barron J T, Mildenhall B, Verbin D, *et al.* Mip-NeRF 360: unbounded anti-aliased neural radiance fields[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 5460-5469
- [37] Kerbl B, Kopanas G, Leimkuehler T, *et al.* 3D Gaussian Splatting for real-time radiance field rendering[J]. ACM Transactions on Graphics, 2023, 42(4): Article No.139
- [38] Hu J T, Xia B, Chen B, *et al.* GaussianSR: high fidelity 2D Gaussian splatting for arbitrary-scale image super-resolution[C] //Proceedings of the 39th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2025: 3554-3562
- [39] Li Y Q, Wang X, Wu J W, *et al.* SparseGS-W: sparse-view 3D Gaussian Splatting in the wild with generative priors[OL]. [2025-07-08]. <https://arxiv.org/abs/2503.19452>
- [40] Hu S K, Hu T, Liu Z W. GauHuman: articulated Gaussian splatting from monocular human videos[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 20418-20431
- [41] Moreau A, Song J F, Dhano H, *et al.* Human Gaussian splatting: real-time rendering of animatable avatars[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

- Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 788-798
- [42] Abdal R, Wang Y F, Shi Z F, *et al.* Gaussian shell maps for efficient 3D human generation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 9441-9451
- [43] Li M T, Yao S X, Xie Z F, *et al.* GaussianBody: clothed human reconstruction via 3D Gaussian Splatting[OL]. [2025-07-08]. <https://arxiv.org/abs/2401.09720>
- [44] Shen T Q, Liu S H, Feng J Q, *et al.* Topology-aware 3D Gaussian Splatting: leveraging persistent homology for optimized structural integrity[C] //Proceedings of the 39th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2025: 6823-6832
- [45] Chen Y F, Wang L Z, Li Q J, *et al.* MonoGaussianAvatar: monocular Gaussian point-based head avatar[C] //Proceedings of the ACM SIGGRAPH Conference Papers. New York: ACM Press, 2024: Article No.58
- [46] Li M W, Tao J C, Yang Z X, *et al.* Human101: training 100+FPS human Gaussians in 100s from 1 view[OL]. [2025-07-08]. <https://arxiv.org/abs/2312.15258>
- [47] Liu Xipu, Ma Yanxin, Xu Ke, *et al.* Multi-scale Transformer based point cloud completion network[J]. Journal of Image and Graphics, 2022, 27(2): 538-549(in Chinese)
(刘心溥, 马燕新, 许可, 等. 嵌入Transformer结构的多尺度点云补全[J]. 中国图象图形学报, 2022, 27(2): 538-549)
- [48] Li J H, Zhang J W, Bai X, *et al.* InsTaG: learning personalized 3D talking head from few-second video[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2025: 10690-10700
- [49] Zielonka W, T, Saito S, *et al.* Drivable 3D Gaussian avatars[C] // Proceedings of the International Conference on 3D Vision. Los Alamitos: IEEE Computer Society Press, 2025: 979-990
- [50] Xiu Y L, Yang J L, Cao X, *et al.* ECON: explicit clothed humans optimized via normal integration[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 512-523
- [51] Mildenhall B, Srinivasan P P, Ortiz-Cayon R, *et al.* Local light field fusion: practical view synthesis with prescriptive sampling guidelines[J]. ACM Transactions on Graphics, 2019, 38(4): Article No.29
- [52] Zhang J W, Wu Z J, Liang Z Y, *et al.* FATE: full-head Gaussian avatar with textural editing from monocular video[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2025: 5535-5545
- [53] Liu X Q, Wu C M, Liu X, *et al.* GEA: reconstructing expressive 3D Gaussian avatar from monocular video[OL]. [2025-07-08]. <https://arxiv.org/abs/2402.16607v1>
- [54] Chen Y D, Xu H F, Zheng C X, *et al.* MVSplat: efficient 3D Gaussian Splatting from sparse multi-view images[C] //Proceedings of the 18th European Conference on Computer Vision. Heidelberg: Springer, 2024: 370-386
- [55] Guédon A, Lepetit V. SuGaR: surface-aligned Gaussian splatting for efficient 3D mesh reconstruction and high-quality mesh rendering[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 5354-5363
- [56] Teotia K, Kim H, Garrido P, *et al.* GaussianHeads: end-to-end learning of drivable Gaussian head avatars from coarse-to-fine representations[J]. ACM Transactions on Graphics (TOG), 2024, 43(6): Article No.264
- [57] Charatan D, Li S L, Tagliasacchi A, *et al.* PixelSplat: 3D Gaussian splats from image pairs for scalable generalizable 3D reconstruction[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 19457-19467
- [58] Das D, Wewer C, Yunus R, *et al.* Neural parametric Gaussians for monocular non-rigid object reconstruction[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 10715-10725
- [59] Huang Y H, Sun Y T, Yang Z Y, *et al.* SC-GS: sparse-controlled Gaussian splatting for editable dynamic scenes[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 4220-4230
- [60] Qian Z Y, Wang S F, Mihajlovic M, *et al.* 3DGS-Avatar: animatable avatars via deformable 3D Gaussian Splatting[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 5020-5030
- [61] Zhao Z Y, Bao Z Y, Li Q, *et al.* PSAvatar: a point-based shape model for real-time head avatar animation with 3D Gaussian Splatting[OL]. [2025-07-08]. <https://arxiv.org/abs/2401.12900>
- [62] Lin Y T, Dai Z Z, Zhu S Y, *et al.* Gaussian-flow: 4D reconstruction with dynamic 3D Gaussian particle[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 21136-21145
- [63] Xiang J, Gao X, Guo Y D, *et al.* FlashAvatar: high-fidelity head avatar with efficient Gaussian embedding[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 1802-1812
- [64] Liu Y, Huang X, Qin M H, *et al.* Animatable 3D Gaussian: fast and high-quality reconstruction of multiple human avatars[C] //Proceedings of the 32nd ACM International Conference on Multimedia. New York: ACM Press, 2024: 1120-1129
- [65] Rensen M L. Physics-informed Gaussian splatting[D]. Delft: Delft University of Technology, 2024
- [66] Wang S Z, Li X T, Liu C, *et al.* Coherent 3D portrait video reconstruction via triplane fusion[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2025: 10712-10722
- [67] Chen J C, Hu J C, Wang G G, *et al.* TaoAvatar: real-time life-like full-body talking avatars for augmented reality via 3D

- Gaussian Splatting[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2025: 10723-10734
- [68] Busso C, Deng Z G, Grimm M, *et al.* Rigid head motion in expressive speech animation: analysis and synthesis[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(3): 1075-1086
- [69] Edwards P, Landreth C, Fiume E, *et al.* JALI: an animator-centric viseme model for expressive lip synchronization[J]. ACM Transactions on graphics, 2016, 35(4): Article No.127
- [70] Karras T, Aila T, Laine S, *et al.* Audio-driven facial animation by joint end-to-end learning of pose and emotion[J]. ACM Transactions on Graphics, 2017, 36(4): Article No.94
- [71] Zhou Y, Xu Z, Landreth C, *et al.* Visemenet: audio-driven animator-centric speech animation[J]. ACM Transactions on Graphics, 2018, 37(4): Article No.161
- [72] Cudeiro D, Bolkart T, Laidlaw C, *et al.* Capture, learning, and synthesis of 3D speaking styles[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 10093-10103
- [73] Lahiri A, Kwatra V, Frueh C, *et al.* LipSync3D: data-efficient learning of personalized 3D talking faces from video using pose and lighting normalization[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 2754-2763
- [74] Thies J, Elgharib M, Tewari A, *et al.* Neural voice puppetry: audio-driven facial reenactment[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 716-731
- [75] Richard A, Zollhöfer M, Wen Y D, *et al.* MeshTalk: 3D face animation from speech using cross-modality disentanglement[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 1153-1162
- [76] Li X H, Wang X, Wang K, *et al.* A novel speech-driven lip-sync model with CNN and LSTM[C] //Proceedings of the 14th International Congress on Image and Signal Processing, Bio-Medical Engineering and Informatics. Los Alamitos: IEEE Computer Society Press, 2021: 1-6
- [77] Guo Y D, Chen K Y, Liang S, *et al.* AD-NeRF: audio driven neural radiance fields for talking head synthesis[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 5764-5774
- [78] Iwase S, Kato T, Yamaguchi S, *et al.* Song2Face: synthesizing singing facial animation from audio[C] //Proceedings of the SIGGRAPH Asia Technical Communications. New York: ACM Press, 2020: Article No.12
- [79] Wang Q Y, Fan Z F, Xia S H. 3D-TalkEmo: learning to synthesize 3D emotional talking head[OL]. [2025-07-08]. <https://arxiv.org/abs/2104.12051>
- [80] Song L S, Wu W, Qian C, *et al.* Everybody's talkin': let me talk as you want[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 585-598
- [81] Shen S, Li W H, Zhu Z, *et al.* Learning dynamic facial radiance fields for few-shot talking head synthesis[C] //Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 666-682
- [82] Fan Y R, Lin Z J, Saito J, *et al.* FaceFormer: speech-driven 3D facial animation with transformers[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 18749-18758
- [83] Yao S, Zhong R Z, Yan Y, *et al.* DFA-NeRF: Personalized talking head generation via disentangled face attributes neural rendering[OL]. [2025-07-08]. <https://arxiv.org/abs/2201.00791>
- [84] Ye Z H, Jiang Z Y, Ren Y, *et al.* GENEFace: generalized and high-fidelity audio-driven 3D talking face synthesis[OL]. [2025-07-08]. <https://arxiv.org/abs/2301.13430>
- [85] Park S J, Hong J, Kim M, *et al.* DF-3DFace: one-to-many speech synchronized 3D face animation with diffusion[OL]. [2025-07-08]. <https://arxiv.org/abs/2310.05934>
- [86] Li J H, Zhang J W, Bai X, *et al.* Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 7534-7544
- [87] He S, He H N, Yang S, *et al.* Speech4Mesh: speech-assisted monocular 3D facial reconstruction for speech-driven 3D facial animation[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 14146-14156
- [88] Han T S, Gui S N, Huang Y Q, *et al.* PMMTalk: speech-driven 3D facial animation from complementary pseudo multi-modal features[J]. IEEE Transactions on Multimedia, 2025, 27: 2570-2581
- [89] Ma Y F, Zhang S W, Wang J Y, *et al.* DreamTalk: when emotional talking head generation meets diffusion probabilistic models[OL]. [2025-07-08]. <https://arxiv.org/abs/2312.09767>
- [90] Thambiraja B, Aliakbarian S, Cosker D, *et al.* 3DiFace: diffusion-based speech-driven 3D facial animation and editing[OL]. [2025-07-08]. <https://arxiv.org/abs/2312.00870>
- [91] Chen Y T, Zhao J H, Zhang W Q. Expressive speech-driven facial animation with controllable emotions[C] //Proceedings of the IEEE International Conference on Multimedia and Expo Workshops. Los Alamitos: IEEE Computer Society Press, 2023: 387-392
- [92] Peng Z Q, Hu W T, Shi Y, *et al.* SyncTalk: the devil is in the synchronization for talking head synthesis[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 666-676
- [93] Zhao Q C, Long P Y, Zhang Q X, *et al.* Media2Face: co-speech facial animation generation with multi-modality guidance[C] //Proceedings of the ACM SIGGRAPH Conference Papers. New York: ACM Press, 2024: Article No.18
- [94] Sha X M, Zhang L Y, Mashita T, *et al.* 3DFacePolicy: speech-driven 3D facial animation with diffusion policy[OL]. [2025-07-08]. <https://arxiv.org/html/2409.10848v1>

- [95] Liu P F, Deng W J, Li H D, *et al.* MusicFace: music-driven expressive singing face synthesis[J]. *Computational Visual Media*, 2024, 10(1): 119-136
- [96] Liu C, Lin Q F, Zeng Z J, *et al.* Emoface: audio-driven emotional 3D face animation[C] //Proceedings of the IEEE Conference Virtual Reality and 3D User Interfaces. Los Alamitos: IEEE Computer Society Press, 2024: 387-397
- [97] He Q Y, Ji X Y, Gong Y C, *et al.* EmoTalk3D: high-fidelity free-view synthesis of emotional 3D talking head[C] //Proceedings of the 18th European Conference on Computer Vision. Heidelberg: Springer, 2025: 55-72
- [98] Nocentini F, Ferrari C, Berretti S. EmoVOCA: speech-driven emotional 3D talking heads[C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2025: 2859-2868
- [99] Ma X P, Cai J R, Guan Y S, *et al.* Playmate: flexible control of portrait animation via 3D-implicit space guided diffusion[OL]. [2025-07-08]. <https://arxiv.org/abs/2502.07203>
- [100] Chae-Yeon L, Hyun-Bin O, EunGi H, *et al.* Perceptually accurate 3D talking head generation: new definitions, speech-mesh representation, and evaluation metrics[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2025: 21065-21074
- [101] Wang H T, Weng Y Z, Li Y Y, *et al.* Emotivetalk: expressive talking head generation through audio information decoupling and emotional video diffusion[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2025: 26212-26221
- [102] Li H, Dai J, Zhao X, *et al.* Wav2Sem: plug-and-play audio semantic decoupling for 3D speech-driven facial animation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2025: 183-192
- [103] Li X C, Wang J Y, Cheng Y H, *et al.* Towards high-fidelity 3D talking avatar with personalized dynamic texture[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2025: 204-214
- [104] Zhang Z M, Hu Z P, Deng W J, *et al.* DINet: deformation inpainting network for realistic face visually dubbing on high resolution video[C] //Proceedings of the 37th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2023: 3543-3551
- [105] Dhamo H, Nie Y Y, Moreau A, *et al.* HeadGaS: real-time animatable head avatars via 3D Gaussian Splatting[C] //Proceedings of the 18th European Conference on Computer Vision. Heidelberg: Springer, 2025: 459-476
- [106] Goodfellow I J, Pouget-Abadie J, Mirza M, *et al.* Generative adversarial nets[C] //Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680
- [107] Adler J, Lunz S. Banach wasserstein GAN[C] //Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2018: 6755-6764
- [108] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[OL]. [2025-07-08]. <https://arxiv.org/abs/1511.06434>
- [109] Vougioukas K, Petridis S, Pantic M. Realistic speech-driven facial animation with GANs[J]. *International Journal of Computer Vision*, 2020, 128(5): 1398-1413
- [110] Kim M, Hong J, Ro Y M. Lip to speech synthesis with visual context attentional GAN[C] //Proceedings of the 35th Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2021: Article No.211
- [111] Tang J X, Wang K, Zhou H, *et al.* Real-time neural radiance talking portrait synthesis via audio-spatial decomposition[J]. *International Journal of Computer Vision*, 2025, 133(9): 6362-6373
- [112] Gafni G, Thies J, Zollhöfer M, *et al.* Dynamic neural radiance fields for monocular 4D facial avatar reconstruction[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 8645-8654
- [113] Peng Z Q, Fan Y B, Wu H Y, *et al.* DualTalk: dual-speaker interaction for 3D talking head conversations[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2025: 21055-21064
- [114] Hu L X, Zhang H W, Zhang Y X, *et al.* GaussianAvatar: towards realistic human avatar modeling from a single video via animatable 3D Gaussians[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 634-644
- [115] Sruthi A, Kumar A K, Dasari K, *et al.* Multi-language: ensemble learning-based speech emotion recognition[J]. *International Journal of Data Science and Analytics*, 2025, 19(3): 453-467
- [116] Thuseethan S, Rajasegarar S, Yearwood J. Deep continual learning for emerging emotion recognition[J]. *IEEE Transactions on Multimedia*, 2022, 24: 4367-4380
- [117] Paier W, Kettern M, Hilsmann A, *et al.* Video-based facial re-animation[C] //Proceedings of the 12th European Conference on Visual Media Production. New York: ACM Press, 2015: Article No.4
- [118] Wu T, Zhou F, Liao Q M. Real-time 3D face reconstruction from one single image by displacement mapping[C] //Proceedings of the IEEE International Conference on Image Processing. Los Alamitos: IEEE Computer Society Press, 2017: 2204-2208
- [119] Zakharov E, Shysheya A, Burkov E, *et al.* Few-shot adversarial learning of realistic neural talking head models[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 9458-9467
- [120] Yao G M, Yuan Y, Shao T J, *et al.* Mesh guided one-shot face reenactment using graph convolutional networks[C] //Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 1773-1781

- [121] Tripathy S, Kannala J, Rahtu E. FACEGAN: facial attribute controllable rEnactment GAN[C] //Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 1328-1337
- [122] Wang T C, Mallya A, Liu M Y. One-shot free-view neural talking-head synthesis for video conferencing[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 10034-10044
- [123] Wang Q L, Zhang L, Li B. SAFA: structure aware face animation[C] //Proceedings of the International Conference on 3D Vision. Los Alamitos: IEEE Computer Society Press, 2021: 679-688
- [124] Bounareli S, Argyriou V, Tzimiropoulos G. Finding directions in GAN's latent space for neural face reenactment[OL]. [2025-07-08]. <https://arxiv.org/abs/2202.00046>
- [125] Khakhulin T, Sklyarova V, Lempitsky V, *et al.* Realistic one-shot mesh-based head avatars[C] //Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 345-362
- [126] Hsu G S, Tsai C H, Wu H Y. Dual-generator face reenactment[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 642-650
- [127] Doukas M C, Ververas E, Sharmanska V, *et al.* Free-headgan: neural talking head synthesis with explicit gaze control[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 9743-9756
- [128] Bounareli S, Tzelepis C, Argyriou V, *et al.* HyperReenact: one-shot reenactment via jointly learning to refine and retarget faces[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 7115-7125
- [129] Li X T, De Mello S, Liu S F, *et al.* Generalizable one-shot 3D neural head avatar[C] //Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2023: Article No.2045
- [130] Li W C, Zhang L H, Wang D, *et al.* One-shot high-fidelity talking-head synthesis with deformable neural radiance field[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 17969-17978
- [131] Wang H Y, Wu H Z, Xing J L, *et al.* Versatile face animator: driving arbitrary 3D facial avatar in rgb-d space[C] //Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM Press, 2023: 7776-7784
- [132] Oorloff T, Yacoob Y. Expressive talking head video encoding in StyleGAN2 latent space[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops Workshops. Los Alamitos: IEEE Computer Society Press, 2023: 2990-2999
- [133] Ni H M, Liu J C, Xue Y, *et al.* 3D-aware talking-head video motion transfer[C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2024: 4942-4952
- [134] Wen J, Zhao X M, Ren Z Z, *et al.* GoMAvatar: efficient animatable human modeling from monocular video using Gaussians-on-Mesh[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 2059-2069
- [135] Chu X G, Harada T. Generalizable and animatable Gaussian head avatar[C] //Proceedings of the 38th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2024: Article No.1838
- [136] Paier W, Hinzer P, Hilsmann A, *et al.* Video-driven animation of neural head avatars[OL]. [2025-07-08]. <https://arxiv.org/abs/2403.04380>
- [137] Zhen D C, Yin S S, Qin S Y, *et al.* Teller: real-time streaming audio-driven portrait animation with autoregressive motion generation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2025: 21075-21085
- [138] Liu H Z, Sun W Z, Di D L, *et al.* MoEE: mixture of emotion experts for audio-driven portrait animation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2025: 26222-26231
- [139] Li Y, Bai Z Q, Tan F T, *et al.* IM-Portrait: learning 3D-aware video diffusion for photorealistic talking heads from monocular videos[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2025: 21107-21116
- [140] Jacka D, Reid A, Merry B, *et al.* A comparison of linear skinning techniques for character animation[C] //Proceedings of the 5th International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa. New York: ACM Press, 2007: 177-186
- [141] Jacobson A, Weinkauff T, Sorkine O. Smooth shape - aware functions with controlled extrema[J]. Computer Graphics Forum, 2012, 31(5): 1577-1586
- [142] Vaillant R, Barthe L, Guennebaud G, *et al.* Implicit skinning: real-time skin deformation with contact modeling[J]. ACM Transactions on Graphics, 2013, 32(4): Article No.125
- [143] Blanz V, Vetter T. A morphable model for the synthesis of 3D faces[M] //Whitton M C. Seminal Graphics Papers: Pushing the Boundaries, Vol 2. New York: ACM, 2023: Article No.18
- [144] Paysan P, Knothe R, Amberg B, *et al.* A 3D face model for pose and illumination invariant face recognition[C] //Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance. Los Alamitos: IEEE Computer Society Press, 2009: 296-301
- [145] Cao C, Weng Y L, Zhou S, *et al.* FaceWarehouse: a 3D facial expression database for visual computing[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(3): 413-425
- [146] Romdhani S, Pierrard J S, Vetter T. 3D morphable face model, a unified approach for analysis and synthesis of images[J]. Face Processing: Advanced Modeling and Methods, 2005: Article No.768
- [147] Booth J, Antonakos E, Ploumpis S, *et al.* 3D face morphable models "in-the-wild"[C] //Proceedings of the IEEE Conference

- on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 5464-5473
- [148] Shi Y C, Wang P, Ye J L, *et al.* MVDream: multi-view diffusion for 3D generation[OL]. [2025-07-08]. <https://arxiv.org/abs/2308.16512>
- [149] Poole B, Jain A, Barron J T, *et al.* DreamFusion: text-to-3D using 2D diffusion[OL]. [2025-07-08]. <https://arxiv.org/abs/2209.14988>
- [150] Jiang Y H, Shen Z H, Wang P H, *et al.* HiFi4G: high-fidelity human performance rendering via compact Gaussian splatting[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 19734-19745
- [151] Chen X Y, Tang S. Emotion-aware talking face generation based on 3DMM[C] //Proceedings of the 4th International Conference on Neural Networks, Information and Communication Engineering. Los Alamitos: IEEE Computer Society Press, 2024: 1808-1813
- [152] Yan Zhuoyue, Liu Li, Fu Xiaodong, *et al.* Spatial-temporal feature reinforcement learning for 3D human pose and shape estimation[J]. Journal of Computer-Aided Design & Computer Graphics: 1-14[2025-07-08]. <https://www.jcad.cn/article/doi/10.3724/SP.J.1089.2024-00483>(in Chinese)
(闫卓越, 刘骊, 付晓东, 等. 三维人体姿态和形状估计的时空特征增强学习[J]. 计算机辅助设计与图形学学报: 1-14 [2025-07-08]. <https://www.jcad.cn/article/doi/10.3724/SP.J.1089.2024-00483>)
- [153] Yao Kai, Liu Li, Fu Xiaodong, *et al.* Absolute and relative depth fusion for 3D multi-person pose estimation[J/OL]. Journal of Computer-Aided Design & Computer Graphics: 1-14 [2025-07-08]. <https://www.jcad.cn/article/doi/10.3724/SP.J.1089.2024-00333>(in Chinese)
(姚凯, 刘骊, 付晓东, 等. 多人 3D 姿态估计中的绝对和相对深度融合[J/OL]. 计算机辅助设计与图形学学报: 1-14 [2025-07-08]. <https://www.jcad.cn/article/doi/10.3724/SP.J.1089.2024-00333>)