中图法分类号:TP391.414　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-22

论文引用格式:Yang Hang, Liu Na, Meng Lei, Mao Qi Rong, Li Man Yi, Li Xiang, Wang Cheng Jie, Zhu Jun Wei, Wang Peng Jie. A Review of Intelligent Digital Human Content Generation Technology[J/OL]. Journal of Image and Graphics, XXXX:1-22. DOI: 10.11834/jig.260074. (杨航, 柳娜, 孟雷, 毛启容, 李曼祎, 李祥, 汪铖杰, 朱俊伟, 王鹏杰. 智能数字人内容生成技术综述[J/OL]. 中国图象图形学报,XXXX:1-22. DOI: 10.11834/jig.260074.)[DOI:10.11834/jig.260074]

智能数字人内容生成技术综述

杨航[1],柳娜[2],孟雷[3],毛启容[2],李曼祎[3],李祥[3],汪铖杰[4],朱俊伟[4],王鹏杰[1]

1. 大连民族大学,大连　11660; 2. 江苏大学,镇江　212013; 3. 山东大学,济南　250100; 4. 腾讯优图实验室,盐城,224002

摘　要:数字人技术作为计算机图形学与人工智能交叉领域的核心方向,目前已经从单一的形象展示开始向智能化、拟人化与情感化的应用阶段发展。当前,视频到数字人生成则通过单目、多目及开放场景技术路线,推动低成本、高保真建模成为可能,不过复杂环境下的几何完整性与动态一致性仍需突破;扩散模型凭借优秀的生成质量,成为3D人体运动合成与编辑的主流框架,但长时序连贯、动作精确编辑、多人复杂交互等仍是关键挑战;情感数字人在单一情感生成与多情感交互建模上取得进展,却面临情感细腻表达与多模态协同的瓶颈。文章系统综述数字人技术三大核心领域的前沿进展,涵盖主流模型、技术分类、数据集与评估体系,最后总结待解决挑战,为未来研究提供前瞻性指引,数字人技术的实现需经历"形态重建-动作赋予-情感交互"的递进过程:首先通过视频到数字人生成技术完成3D形态的基础构建,再通过3D人体运动合成与编辑赋予动态能力,最终通过情感数字人技术实现自然交互,本文所提到的算法,数据集等整理在https://github.com/blue-cola-bc/Overview-of-Intelligent-Digital-Humans.git,最终对未来可能的研究方向进行了展望。

关键词:数字人技术;3D人体运动合成与编辑;数字人智能生成;扩散模型;视频到数字人生成;多模态情感交互

A Review of Intelligent Digital Human Content Generation Technology

Yang Hang[1], Liu Na[2], Meng Lei[3], Mao Qi Rong[2], Li Man Yi[3], Li Xiang[3], Wang Cheng Jie[4], Zhu Jun Wei[4], Wang Peng Jie[1]

1. Dalian Minzu University, Dalian 116600, China; 2. Jiangsu University, Zhenjiang 212013, China; 3. Shandong University, Jinan 250100, China; 4. Tencent YouTu Lab, Shenzhen Yancheng 224002, China

Abstract: Intelligent digital humans have rapidly evolved with the advancement of computer graphics, computer vision, speech synthesis, and multimodal generative modeling. From early virtual avatars focusing mainly on visual representation, digital humans are now developing toward dynamic motion modeling, emotion-aware interaction, and real-time deployment. This paper presents a systematic review of recent research progress in intelligent digital human content generation, organized around three core technical directions: video-to-digital human generation, 3D human motion synthesis and editing, and emotion-driven digital human generation. In addition, practical considerations for real-time on-device deployment are discussed. Video-to-digital human generation serves as the foundational stage of digital human construction. Its goal is to reconstruct animatable 3D human avatars from monocular, multi-view, or in-the-wild video inputs. Early approaches primarily relied on implicit neural representations such as neural radiance fields (NeRF), often combined with parametric body models such as the skinned multi-person linear model (SMPL). Although implicit methods provide continuous and high-fidelity geometry representation, their rendering efficiency limits real-time applicability. More recent studies have shifted toward explicit or hybrid representations, particularly 3D Gaussian splatting, which significantly

收稿日期:2026-02-03;**修回日期**:2026-03-06

© 中国图象图形学报版权所有

improves rendering speed while maintaining visual quality. Extensions incorporating SMPL or SMPL-X priors further enhance geometric stability and animatability. In multi-view settings, stronger geometric constraints improve reconstruction accuracy, while open-scene scenarios introduce additional challenges such as occlusion handling, multi-person interaction, and background interference. Despite notable progress, maintaining temporal consistency and geometric robustness in complex environments remains an open problem. On top of geometric reconstruction, 3D human motion synthesis and editing enable digital humans to exhibit realistic dynamic behaviors. Compared with static modeling, motion generation requires accurate modeling of high-dimensional temporal distributions under kinematic and physical constraints. Early approaches based on statistical models or variational autoencoders (VAE) improved representation capacity but often suffered from limited motion diversity. In recent years, diffusion models have become the dominant paradigm for motion generation due to their strong capability in modeling complex multi-modal distributions. Representative frameworks demonstrate improved motion realism, diversity, and semantic alignment with textual or conditional inputs. Latent diffusion strategies further enhance efficiency by performing denoising processes in compact latent spaces. Beyond unconditional generation, condition-driven and fine-grained motion editing have attracted increasing attention. Text-guided editing frameworks allow local modification of specific joints or temporal segments while preserving overall motion style. Skeleton-aware and physics-guided diffusion models introduce structural constraints to improve anatomical plausibility and reduce artifacts such as foot sliding. Moreover, research has gradually expanded toward multi-person interaction modeling and long-sequence coherence, addressing challenges in action composition, interaction synchronization, and environment-aware motion planning. Nevertheless, balancing physical consistency, computational efficiency, and controllability remains a critical challenge for practical applications. Emotion-driven digital human generation further enhances interactivity and human-likeness. This direction includes facial expression synthesis, emotional speech synthesis, and multi-turn empathetic interaction modeling. In facial animation, research has progressed from parameterized 3D morphable models to implicit neural rendering and, more recently, Gaussian-based explicit representations that achieve improved fidelity and real-time performance. In emotional speech synthesis, end-to-end neural architectures, non-autoregressive frameworks, and neural codec language models enable expressive, zero-shot, and fine-grained controllable speech generation. Meanwhile, emotion modeling in interactive dialogue systems has evolved from passive emotion recognition toward empathetic response generation, incorporating graph-based contextual modeling and large language model fine-tuning strategies. Although current systems can generate recognizable emotional expressions, challenges remain in maintaining emotional consistency over long interactions, decoupling emotion from underlying controllable factors, and ensuring cross-modal alignment between speech, facial motion, and semantic context. To support real-world deployment, real-time on-device digital human systems have also gained attention. Under constrained computational resources, lightweight models, reduced-resolution rendering, intermediate parameter representations, and efficient inference frameworks are commonly adopted. In practical applications, a collaborative architecture is often employed, where large language models handle semantic reasoning in the cloud while speech synthesis and avatar rendering are executed locally. This edge-cloud collaboration balances interaction latency and generation quality, facilitating scalable deployment in mobile and desktop environments. In addition to reviewing representative models and technical routes, this paper summarizes commonly used datasets and evaluation metrics across subfields, including text-to-motion benchmarks, emotional speech corpora, and multi-view reconstruction datasets. Performance is typically evaluated from multiple perspectives, such as perceptual realism, geometric accuracy, semantic alignment, motion stability, and subjective human assessment. To facilitate reproducible research and provide a centralized resource for the community, we have curated all surveyed datasets, benchmark links, and a structured list of representative models into a public GitHub repository, available at: https://github.com/blue-cola-bc/Overview-of-Intelligent-Digital-Humans.git. Despite rapid progress, unified benchmarks for long-term interactive digital humans are still lacking. Overall, intelligent digital human technology is advancing along a progressive pathway from geometric reconstruction to motion generation and finally to emotion-aware interaction. Future research is expected to focus on unified multimodal generative frameworks, improved long-term consistency modeling, physics-aware motion control, and efficient real-time deployment. By systematically organizing recent developments and open challenges, this review aims to provide a structured understanding of current progress and potential research directions in intelligent digital human content generation.

© 中国图象图形学报版权所有

Key words: digital human technology; 3D human motion synthesis and editing; digital human generation; diffusion models; video-to-digital human generation; multimodal emotional interaction

0 引 言

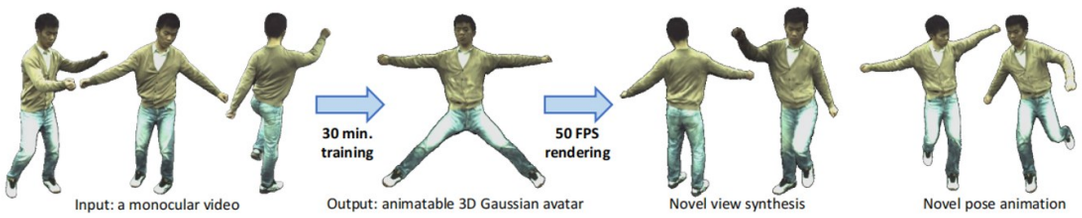
随着人工智能、虚拟现实与多模态交互技术快速发展，数字人已广泛应用于智能客服等领域，成为新一代信息技术与数字经济战略的重要组成部分。中国科协 2024 年十大前沿科学问题中，“情智兼备数字人与机器人的研究”位列首要。

在技术演进中，3D 人体运动合成与编辑依托扩散模型实现质量突破。Tevet 等人（2023）的 MDM、Chen 等人（2023）的 MLD 奠定生成质量基础；Han 等人（2024）的 AToM、DSDFM 在细粒度文本对齐与生成多样性上取得进展；MotionFix 与 SD（Athanasiou 等，2024；Curreli 等，2025）提供动作编辑与物理约束新思路。尽管 InterGen（Liang 等，2024）、PINO（Ota 等，2025）探索多人交互，但现有技术在复杂物理规律下的动作合理性与多人互动上仍需梳理。

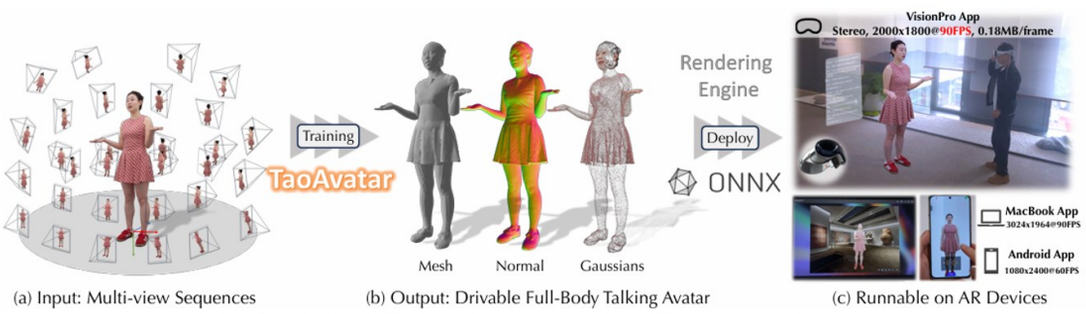
情感数字人推动从功能代理向情感伴侣转型，却在情感建模、细腻表达与多模态协同上存在不足，难以满足人机自然互动与长期陪伴需求，核心趋势是从“可见的情绪”迈向“可理解、可互动的情绪”。

视频到数字人生成彻底改变传统依赖激光扫描等设备的复杂流程。如图 1 所示，该方向按应用场景与采集条件划分为单目、多目与开放场景三类路线：三者均指向高效建模与高质量外观重建，单目需权衡速度与真实感，多目对细节精度要求更高，开放场景需处理遮挡等因素。

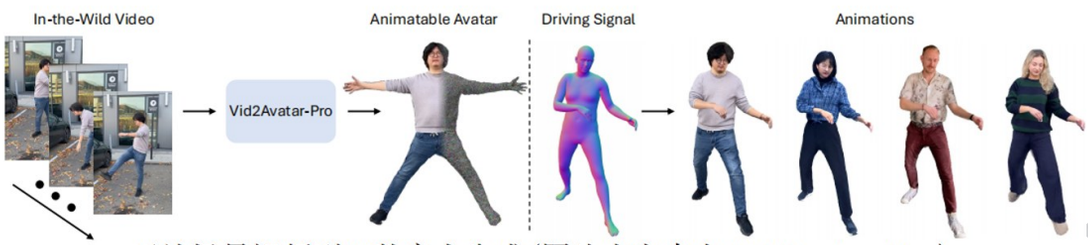
因此，文章整合三大技术领域，从基础模型、技术分类、数据集与未来展望等多个方面描写，在数字



单目视频到3D数字人生成（图片内容来自 3DGS-Avatar）



多目视频到3D数字人生成（图片内容来自TaoAvatar）



开放场景视频到3D数字人生成(图片内容来自Vid2Avatar-Pro)

图 1 单目、多目和开放场景视频生成数字人的相关研究

Figure1 Related research on generating digital humans from monocular, multi-ocular, and open-scene videos

© 中国图象图形学报版权所有

人技术在可控性、鲁棒性与语义对齐方面的最新突破进行阐述。数字人技术的实现遵循“形态重建-动作赋予-情感交互”的递进逻辑：首先通过视频到数字人生成技术完成基础构建，再通过3D人体运动合成与编辑赋予动态能力，最终通过情感数字人技术实现自然交互，数字人技术的三大方向之间层次关系如图2所示本报告即按此逻辑展开综述。

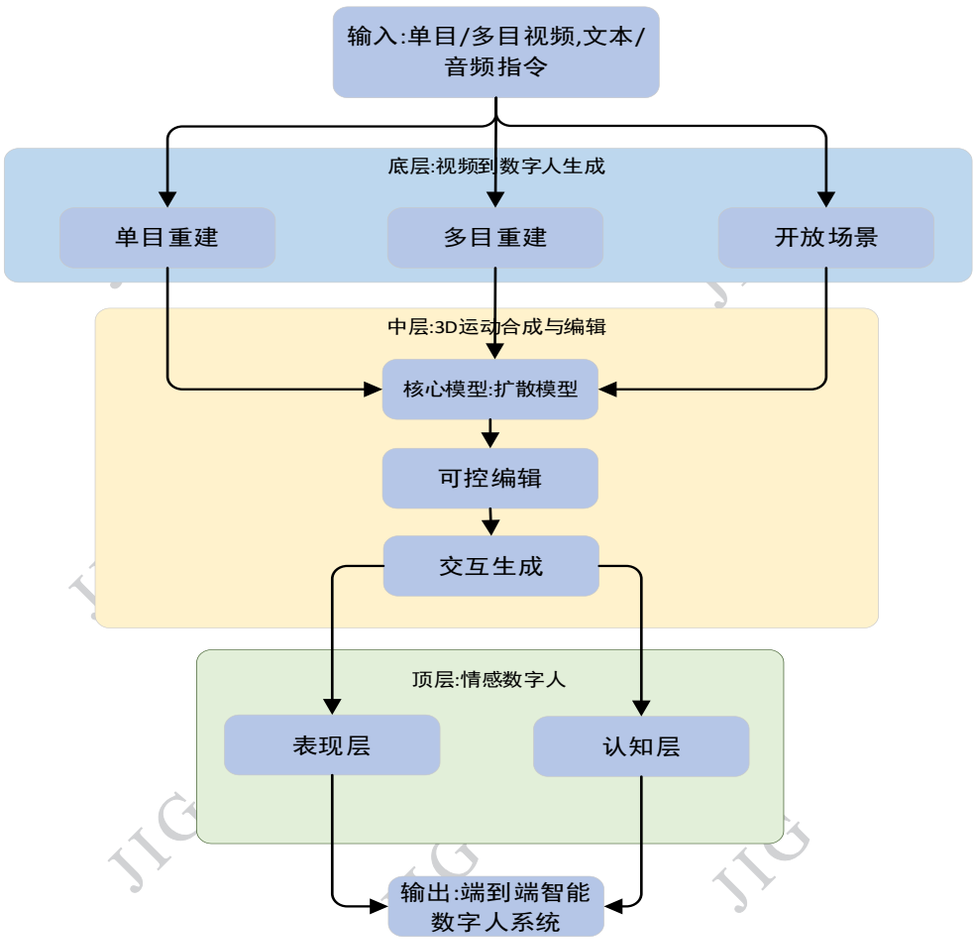


图2　全局框架图

Figure2　Overall framework diagram

框架图中“视频到数字人生成”为底层建形技术，“3D运动合成与编辑”为中层赋动技术，“情感数字人”为顶层注情技术，三者层层递进，结合端上实时交互技术，共同支撑智能数字人落地应用。

1　视频到数字人生成技术

视频到数字人生成技术是智能数字人系统中实现三维几何建模的重要基础，其核心目标是从单目或多目视频序列中恢复具有可驱动性的三维人体表示。相较于基于扫描或多传感器设备的建模方式，视频驱动方法在数据获取成本和应用灵活性方面具有明显优势，因此逐渐成为当前研究的重点方向之一。根据视频采集条件和场景复杂度的不同，相关研究通常可分为单目视频、多目视频以及开放场景视频到数字人的建模方法。

1.1　单目视频到3D数字人

单目视频到三维数字人的重建任务旨在仅依赖单一视角的RGB视频（如手机等），构建可用于动画和交互的虚拟化身。由于单目输入天然缺乏深度信息，该任务在几何恢复、动态外观建模以及非刚性形变处理等方面存在较大挑战，研究工作通常需要在重建精度、计算效率和系统稳定性之间进行权衡。

早期方法主要基于隐式神经表示框架，通过神经网络学习规范空间到观测空间的连续映射关系。代表性工作包括Jiang等人（2023）提出的InstantAva-

© 中国图象图形学报版权所有

tar以及Weng等人(2022)提出的HumanNeRF,这类方法通常基于神经辐射场(neural radiance field, NeRF)模型,并结合参数化人体先验(如SMPL)进行约束。隐式神经场能够提供连续且细致的几何表示,但其依赖体积渲染的计算方式,在推理效率和实时性方面存在明显限制,从而影响了在实际应用场景中的部署。

随Kerb等人(2023)提出3D Gaussian Splatting方法,研究开始转向基于显式表示的高效建模策略。3D高斯泼溅通过各向异性高斯原语进行场景表达,在渲染效率方面具有显著优势。然而,原始3D高斯泼溅(3D gaussian splatting,3DGS)方法并未针对人体的非刚性运动进行设计,因此后续研究通常结合人体结构先验对其进行扩展。Hu等人(2024)提出的GauHuman以及Li等人(2024)提出的GoMAvatar利用蒙皮多人线性模型(skinned multi-person linear model,SMPL)顶点作为高斯点云的初始化先验,从而加速模型收敛并提升几何稳定性. 在此基础上,ExAvatar将高斯点进一步绑定至蒙皮多人线性模型扩展版(skinned multi-person linear model extended, SMPL-X)网格表面,通过拓扑约束实现了对手部和面部细节的兼容建模(Li等, 2024)。

针对长时序动作中可能出现的几何失真问题,Leung等人(2024)在GART中建立了高斯点与骨骼关节之间的刚性绑定关系,以增强形变过程中的结构一致性。此外,3DGS-Avatar通过引入可学习的位置和旋转速度参数,并结合曲线插值策略,实现了对连续动作序列的平滑建模(Li等, 2024)。在系统层面,单目视频在线密集人体场景重建模型(online dense 3D reconstruction of humans and scenes from monocular videos,ODHSR)通过联合优化相机追踪与人体姿态估计,提高了在线处理场景下的稳定性,而Link to the Past则利用时序传播机制复用历史帧信息,以降低重复计算带来的开销(Liu等, 2024; Zhang等, 2023)。

在动态序列建模中,姿态估计误差往往会导致外观模糊或局部失真。为缓解这一问题,Zhang等人(2024)提出了GaussianAvatar方法通过引入动态外观网络,对姿态误差和外观参数进行联合优化;SplattingAvatar则采用基于网格的显式运动控制机制,实现了运动参数与外观建模的解耦(Shao等, 2024)。

在头部重建这一细分方向,Chen等人(2024)提出MonoGaussianAvatar结合参数化人脸先验,缓解了单视角条件下侧面几何缺失的问题。Xiang(2024)所提出的方法FlashAvatar通过高斯嵌入压缩策略进一步提升了渲染效率,而贴片级反射率扩散模型(patch-level reflectance diffusion model, DoRA)利用贴片级反射率扩散模型,引入纹理先验以补充单目观测中缺失的信息(Khakhulin等, 2023)。

1.2 多目视频到3D数字人

多目视频到三维数字人的重建依赖多摄像头同步采集的数据,通过多视角约束提升几何恢复的完整性和准确性。这类方法在处理遮挡、复杂姿态和细节建模方面具有天然优势,但同时也对系统设计和计算资源提出了更高要求。

早期代表性工作如peng等人(2021)设计了Neural Body通过结构化隐式编码实现了较为鲁棒的多目重建,但其推理速度受限于隐式渲染框架。Gaussian Avatars将3DGS引入多目场景,在提升渲染效率的同时,仍依赖由Zhang等人(2023)所提出来的多层感知机(multi-layer perceptron,MLP)网络预测高斯属性,在复杂形变下存在一定瓶颈。

为在效率与质量之间取得更好的平衡,Animatable Gaussians采用基于卷积结构的网络直接从姿态条件预测高斯属性,从而减少低频偏差并提升渲染稳定性(Li等, 2024)。Yang(2024)所使用的TaoAvatar进一步引入知识蒸馏策略,将复杂形变网络压缩为轻量模型,实现了在移动端条件下的高帧率渲染。

在几何精度方面,随着Zhang(2023)所提出的方法MultiGO通过分层先验融合策略,将重建过程解耦为骨骼、关节和细节层级,以提升整体几何一致性。DiffuMan4D则构建了时空扩散框架,通过迭代式生成与修正过程,改善了动态人体重建中的拓扑准确性(Zhao等, 2024)。此外,针对复杂服饰建模问题,ToMiE提出了自适应关节树增长策略(Zhang等, 2024),而基于位置动力学的动态高斯模型(position based dynamic Gaussians,PBDyG)将位置动力学方法引入3DGS表示,用于对衣物物理属性进行反向优化(Chen等, 2024)。

1.3 开放场景视频到3D数字人

开放场景视频到数字人生成旨在解决真实环境

© 中国图象图形学报版权所有

中的遮挡、多主体干扰及复杂背景等问题，是实现通用数字人建模的关键方向之一。由于观测条件受限，这类任务通常需要融合更强的先验知识与跨模态建模策略。

针对单目视角下因信息缺失导致的几何不完整问题，近期研究通过引入先验模型来提升重建质量。例如，Guo等人（2024）在WonderHuman工作当中利用大规模二维扩散模型作为先验，通过分数蒸馏采样融合未观测区域与可见区域的几何信息。类似的，Chen等人（2024）在工作Vid2Avatar-Pro当中通过构建通用人体先验模型，约束逆向渲染过程，从而增强了复杂场景下的重建稳定性。

在严重遮挡场景中，方法的鲁棒性面临更大考验。OccNeRF采用表面渲染策略，将局部可见区域与SMPL人体模型先验进行关联，以提升遮挡条件下的重建效果（Zhang等，2023）。OccGaussian则进一步引入特征补偿机制，通过聚合可见点的特征来补全被遮挡区域（Zhang等，2024）。

多人交互场景的建模还受限于高质量数据的稀缺。为此，Harmony4D与Hi4D等数据集的构建为复杂交互研究提供了重要支持（Jiang等，2024；Yin等，2023）。在方法学上，Rempe等人（2024）提出Guess The Unseen方法从局部二维观测推理完整四维场景，为处理严重遮挡提供了新思路。针对密集交互，Huang等人（2024）设计了双分支矢量量化变分自编码器（vector quantized variational autoencoder，VQ-VAE）架构，并利用交叉注意力机制实现多主体运动的协同建模。

2 3D人体运动合成与编辑技术

在数字人内容生成的整体技术体系中，3D人体运动合成与编辑是连接底层几何建模与高层语义交互的关键环节。相较于外观建模主要关注几何一致性与视觉真实感，人体运动直接决定了数字人在时间维度上的行为合理性，其生成质量往往更容易被第一时间观察到。在虚拟人交互、数字内容制作以及具体的应用场景中，即便外观模型保真度高，但是缺乏自然、连贯且可控的运动表现，仍会削弱整体在真实感上的表现。

从研究任务本身来看，人体运动合成需要生成连续的动态姿势，同时在多个维度对实际应用需求的满足也要规划。一方面，生成模型需要覆盖足够丰富的动作分布，以避免结果模式单一；另一方面，在文本、音频或结构条件驱动下，系统还需支持对动作语义、时间结构以及局部关节的精细控制；与此同时，人体运动受到明确的运动学与物理约束，生成结果若出现关节异常、脚部滑移或接触不稳定等问题，将直接破坏整体的运动动作。这些特点使得人体运动生成在建模难度上显著高于静态几何或单帧图像方向的生成任务。

随着生成式模型的发展，人体运动生成的研究范式逐渐从早期依赖规则和数据拼接的方法，转向以数据驱动的端到端生成建模为主。尤其是近年来扩散模型在高维复杂分布建模方面展现出的优势，使得人体运动生成在生成质量、多样性以及可控性方面取得了显著进展（Tevet等，2023；Chen等，2023）。本章将围绕生成模型演进、条件驱动与编辑方法，以及复杂交互与物理约束等方面，对3D人体运动合成与编辑领域的研究进展进行系统梳理。

2.1 主流生成模型与技术演进

随着人体运动动画生成方向的持续发展，对于高维时序运动分布建模能力的逐步提升。在当前领域的早期工作当中主要依赖统计建模方法，通过对动作数据进行降维与参数化描述，实现对有限动作模式的重建与合成。这类方法通常假设运动数据分布具有较强的线性结构，能够较好地生成行走、跑步等周期性和简单性动作，但是该方法在复杂动作组合和风格变化方面存在明显短板。

随着深度学习技术的成熟，研究者开始引入神经网络对人体运动进行非线性建模。其中，基于变分自编码器（variational autoencoder，VAE）的生成方法成为较早被广泛采用的技术路线。Petrovich等人（2021）提出的ACTOR模型通过条件VAE框架，将动作生成任务与文本或动作类别条件相结合，实现了较为灵活的动作控制。Guo等人（2020）在Action2Motion中进一步引入图卷积网络，对人体骨架的拓扑结构进行显式建模，从而在一定程度上提升了生成动作的合理性。这类方法相较统计模型显著增强了表达能力，但由于潜在空间通常被约束为单峰高斯分布，在面对多模态动作分布时容易出现均值回归问题，生成结果在多样性上往往很欠缺。

近年来，扩散模型逐渐成为人体运动生成领域的主流方法。与一次性生成完整动作序列的生成模

© 中国图象图形学报版权所有

型不同，扩散模型通过逐步去噪的随机过程生成目标数据，使模型能够在生成过程中不断修正中间状态，从而在随机性与结构约束之间取得更稳定的平衡。Tevet等人（2023）提出的Motion Diffusion Model（MDM）首次系统性地验证了扩散模型在人体运动生成任务中的可行性，并通过引入几何约束损失，有效缓解了关节长度不一致和脚部滑移等问题。该工作已经是人体动画生成领域在扩散模型上的基础。

然而，直接在高维关节空间中执行扩散采样通常伴随着较高的算力要求，这在一定程度上限制了模型在实际系统中的应用。针对这一问题，Chen等人（2023）提出Motion Latent Diffusion（MLD），通过预训练变分自编码器（VAE）将运动序列映射至紧凑的潜在空间，并在该潜在空间中执行扩散过程，从而在显著提升推理效率的同时，保持较高的生成质量。这一思路不仅提升了模型的实用性，也为后续在实时生成和交互式应用中的探索提供了可能。

为了更直观地比较不同生成范式的特点，如表1对当前主流的人体运动生成技术路线在建模能力、生成质量和适用场景等方面进行了总结。可以看出，不同方法在生成质量与推理效率之间存在明显权衡，这也直接影响了其在编辑与交互任务中的应用方式。

表1　3D人体运动生成关键技术路线深度对比

Table1　In-depth comparison of key technology routes for 3D human motion generation

技术路线	建模空间	代表模型	优势	局限	适用场景
变分自编码器（VAE）	潜在空间正态分布	Action2Motion	推理速度快，结构简单	易趋同缺乏多样性	实时性要求极高的简单动作生成
自回归模型（Auto-Regressive）	姿态序列逐帧分析	T2M-GPTMotionGPT	适合长序列建模，语义推理能力强	易产生误差累积	结合大语言模型的动作推理任务
扩散模型（Diffusion Models）	原始姿态空间	Motion diffusion model	多模态建模能力强	推理速度慢，计算开销大	高质量动画制作、离线生成任务
潜在扩散模型	潜在编码空间	Motion latent diffusion	提升效率，兼顾质量	依赖VAE质量	交互式生成
条件编辑扩散模型	潜在+条件控制	MotionFix	支持精细局部编辑	结构设计复杂，训练成本高	动作编辑与修改
结构约束扩散模型	骨架感知空间	Skeleton Diffusion	物理合理性强	结构设计复杂	工业动画生成

2.2　条件驱动的运动合成与精细化编辑

在实际应用中，仅生成完整的动作序列往往难以满足具体交互需求。更复杂的交互生成方式在于如何在保持原有动作结构和风格的前提下，对动作的局部内容进行精确修改。围绕这一问题，近年来的研究开始从整体动作生成，转为基于条件引导下的局部细节编辑框架。

MotionFix是该方向的代表性工作之一。Athanasiou等人（2024）提出了一种双流扩散架构，使模型能够同时接收源动作和编辑指令，并通过跨模态注意力机制定位需要修改的时间片段或关节区域。该方法能够在不破坏整体动作连贯性的情况下，实现语义级别的局部编辑。为进一步保持原动作的节奏和风格特征，Li等人（2025）提出SimMotionEdit，在训练过程中引入运动相似性预测任务，从而约束编辑结果与源动作在高层运动特征上的一致性。

除文本条件外，人体自身的结构属性也是影响运动真实感的重要因素。Hong等人（2025）提出的SALAD在潜空间扩散过程中显式引入骨架约束，使模型在编辑过程中能够感知人体结构，从而支持基于关节拖拽的交互式编辑。在体型建模方面，Liao等（2025）人将SMPL-X的体型参数作为生成条件，

© 中国图象图形学报版权所有

使模型能够根据不同体型自动调整动作幅度，减少穿模等几何不合理现象。此外，PersonaBooth通过少样本参考序列提取个性化风格嵌入，实现了动作内容与个体风格的解耦迁移，为个性化数字人资产的快速构建提供了新的思路（Kim 等，2025）。

2.3 复杂运动的组合、交互与长时序连贯性

随着数字人应用从单一动作展示逐步走向复杂环境要求与多主体之间的交互场景，人体动画生成面临的生成难度显著提升。相比单人短时动作生成，复杂运动场景通常同时涉及多重动作语义、多主体之间的时序协同，以及长时间跨度内运动风格和物理状态的一致性约束。这些因素的叠加，使得传统以单一动作或短时序为建模目标的方法难以直接适用。

在复杂运动场景中，首先需要解决的是多动作语义的组合问题。在实际应用中，用户往往希望数字人能够同时执行多个动作指令，例如“边走边打电话”或“移动过程中完成取物操作”。这类任务并非简单地将两个动作序列进行叠加，而是需要在不同身体部位之间进行合理的功能分配，并在时间维度上保持整体协调。针对这一问题，Ruiz-Ponce 等人（2025）提出的 MixerMDM 采用混合专家架构，根据不同身体部位动态分配控制权重，使模型能够在生成过程中对多个动作语义进行解耦与融合。相比直接在姿态空间中进行动作叠加，该方法在动作自然性和稳定性方面表现出明显优势。另一方面，Zhang 等人（2024）提出的能量驱动运动生成模型（energy-based motion generation，ENERGYMOGEN）通过在潜在空间中构建可加的能量函数，实现了对不同动作语义与风格语义的组合建模，为复合动作生成提供了另一种建模思路。

除多动作组合外，多主体交互建模是复杂运动生成中的另一核心挑战。在多人交互场景中，个体动作不再是相互独立的，而是需要根据对方的动作状态进行动态调整。例如在拥抱、或协作等强交互场景中，动作的时序同步性和空间协调性直接决定了交互效果的真实感。Liang 等人（2024）提出的 InterGen 构建了大规模双人交互数据集，并设计了双流协作去噪架构，使交互双方在生成过程中能够显式感知对方的运动状态，从而显著提升紧密接触动作的时空一致性。该类方法表明，在多人交互任务中，将单人运动生成模型简单并行往往难以获得理想效果，而需要在模型结构层面引入显式的交互建模机制。

进一步地，随着数字人逐渐被置于复杂虚拟环境中，人与环境的交互建模也成为不可忽视的问题。在这类场景中，运动生成不仅需要考虑人体自身的运动合理性，还需满足场景几何约束和路径规划要求。Cong 等人（2025）提出的 SemGeoMo 将场景几何信息引入运动生成过程，使模型能够在生成动作时显式考虑人体与环境之间的空间关系。Huang 等人（2024）在 Move-in-2D 中进一步结合路径规划算法，使数字人能够在复杂环境中完成具有目标导向的移动与避障行为。这些研究表明，人体运动生成正逐步从“仅关注人体本体”向“感知环境约束”的方向演进。

在上述复杂场景中，长时序一致性问题对于生成质量很重要。随着动作序列长度的增加，模型在时间维度上的误差往往会逐渐累积，导致运动风格漂移或物理状态突变。为缓解这一问题，部分研究开始引入显式的记忆机制或结构约束，以保持生成序列在长时间尺度上的一致性。

如图 3 对上述典型复杂运动场景进行了直观示意，分别展示了单人动作在时间维度上的连贯性、多主体交互动作的语义组合，以及基于条件约束的精细化运动编辑方式。

2.4 工业落地挑战与实际应用瓶颈

近年来，基于扩散模型的人体运动生成方法在生成质量和动作多样性方面取得了明显进展。然而，从实际应用和复杂交互场景的角度来看，现有方法在物理一致性和工程可用性方面仍存在一些需要进一步研究的问题。

从建模角度来看，当前多数人体运动生成方法主要以关节姿态、骨骼结构和全局位移等运动学信息作为建模对象。这类方法在无接触或弱接触场景中通常能够生成较为自然的动作序列，但在涉及脚部着地、身体支撑或人与物体接触等情形时，仍可能出现滑步或接触不稳定等现象。这在一定程度上反映了现有模型对复杂物理交互过程的建模能力仍然有限。

针对上述问题，部分研究开始尝试在生成过程中引入结构性或物理相关的约束信息。Curreli 等人（2025）提出的 Skeleton Diffusion 通过在扩散过程中引入解剖学约束，对关节活动范围和骨骼结构进行

© 中国图象图形学报版权所有

图3　单人动作的连贯性(左)；多人交互的组合(中)；通过控制实现对运动精确、灵活编辑(右)

Figure3　Coherence of individual actions (left); Combination of multiple interactions (middle); Precise and flexible editing of movements through control (right)

限制，从而在一定程度上减少了明显不合理姿态的产生。该类方法通过在模型层面引入人体结构先验，为提升生成稳定性提供了一种可行思路。

进一步地，也有研究探索将物理模拟结果引入生成流程，以辅助生成过程的约束与修正。Yuan等人(2023)提出的PhysDiff在扩散推理过程中引入物理模拟反馈，使生成姿态在一定程度上满足物理可行性要求。尽管该类方法在物理一致性方面表现出不错的效果，但对于算力的要求较大，在实时或交互式应用中的适用性方面表现不好。

3　情感数字人的智能生成技术

随着人工智能、虚拟现实与多模态交互技术的快速发展，数字人正逐渐从单一的形象展示走向智能化、拟人化与情感化表达。国内外诸多企业纷纷布局虚拟人及情感交互系统，其应用场景已覆盖智能客服、虚拟主播、在线教育、心理健康辅助和沉浸式娱乐等领域，推动数字人从功能代理向情感伴侣转型。在中国科协发布的2024年十大前沿科学问题中，“情智兼备数字人与机器人的研究”被列为首要课题之一，这充分体现了情感数字人在未来智能社会中的战略地位。然而，现阶段的数字人在情感建模、细腻表达与多模态协同生成方面仍存在明显不足，尚难以满足人机自然互动与长期陪伴的需求，因此情感数字人的智能生成技术进展主要聚焦于情绪表达与交互建模，可分为单一情感生成与多情感交互生成。单一情感生成从基础的面部表情与语音合成拓展到细粒度的多模态调控，实现自然、可控且个性化的情绪呈现；多情感交互生成从单向情绪反应发展为复杂情感互动，不仅需要生成符合语境的表达，还需建模个体间的情感共鸣、情绪传递与交互意图，从而实现高质量、长期化和情境自适应的情感交流。总体而言，情感数字人的发展正经历由静态情绪呈现向动态、多模态与社会化交互的转变，其核心趋势是从“可见的情绪”迈向“可理解、可互动的情绪”。例如图4展示了从单一情感生成拓展至多情感交互生成的关键相关任务。

同时，表2进一步从核心技术维度出发，对情感数字人智能生成的关键路线与方法进行了系统性梳理，涵盖面部表情合成、情感语音合成、情感共鸣建模、交互式情感生成四大核心领域，详细对比了不同方法分类的代表性技术/模型、优势、局限及对应数据集，为理解各技术路径的特点与适用场景提供了全面参考。

3.1　单一情感生成

在情感数字人发展早期，研究重点主要集中在单一情感生成任务，其目标是通过对面部、语音等单模态或跨模态信号的建模，实现可控、自然的情绪呈现。这一方向不仅奠定了情感表达合成的技术基础，也为后续多情感交互生成提供了方法与数据支持。

3.1.1　面部表情合成

情感数字人的构建离不开面部表情动画的合成与渲染。面部表情合成是指从单个图像、视频、音频或文本生成或编辑人脸的表情变化，面部表情合成主要经历了从“参数化模型”的线性控制，到“隐式神经场”的高保真渲染，再到“显式高斯表示”的实时交互演进，逐步突破了精度与效率的平衡瓶颈。早期的方法主要依赖三维可变形模型(3D Morphable

© 中国图象图形学报版权所有

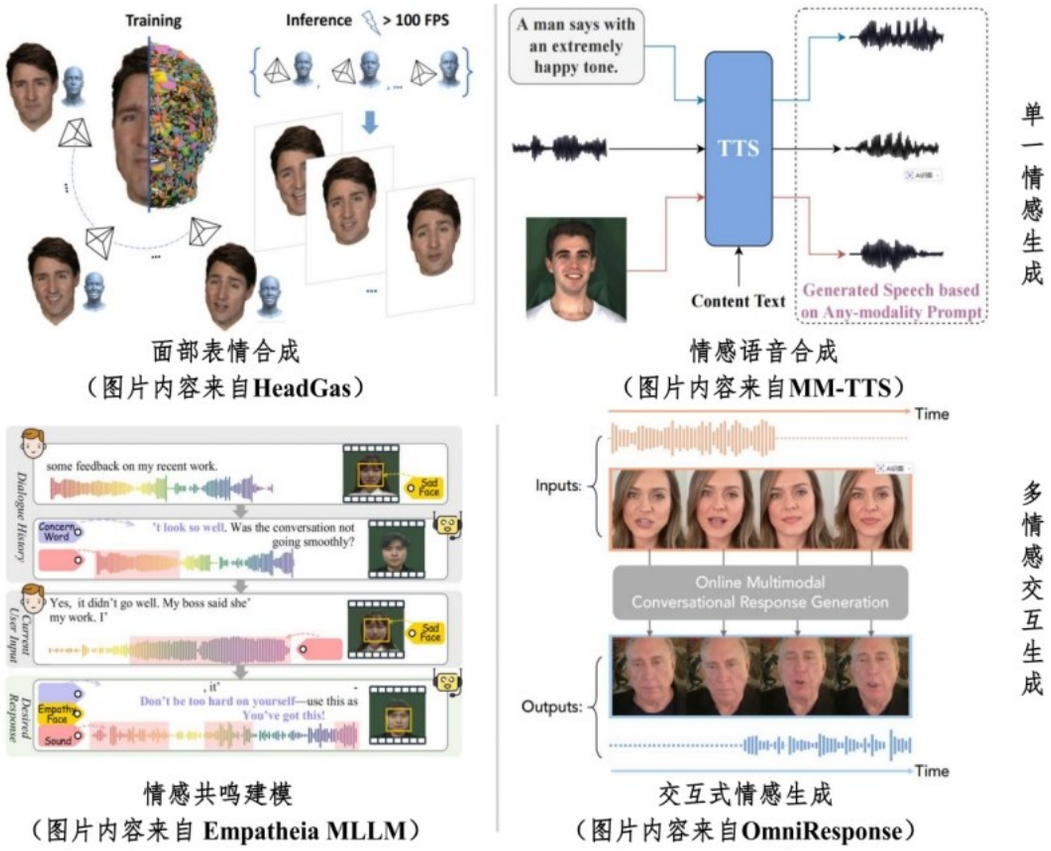


图4　单一和多情感交互生成相关任务

Figure4　Single and multi-emotion interactive generation-related tasks

Model，3DMM）作为表情参数化的基本框架。研究者们先后探索了通过卷积神经网络从单张图像直接回归3DMM参数的思路（Tuan 等，2017），并逐步拓展至在自编码器、对抗学习以及可微渲染框架下实现端到端的无监督训练（Tewari 等，2017）。随后，一些工作进一步将动作单元AU回归为3DMM的表情系数（Chang 等，2018），或将3DMM参数与生成模型结合以提升合成质量（Tewari 等，2020）。这些研究显著提高了表情建模的精度，但基于统计学的线性插值方法在处理个体差异和复杂表情组合时仍存在表达能力有限的问题，难以生成细腻且个性化的表情变化。随着神经辐射场（Neural Radiance Field，NeRF）（Gafni 等，2021）的提出，表情合成逐渐从传统的线性子空间方法转向基于体积渲染的隐式表示框架。典型工作包括：利用动态NeRF实现随表情变化的4D人脸重建Hong等人（2022）提出基于HeadNeRF与FDNeRF的参数化控制方法，实现少样本条件下的实时合成与表情编辑；将统计形变模型融入NeRF的FLAME-in-NeRF，实现基于参数驱动的自由视角表情动画（Athar 等，2023）；以及Face-CLIPNeRF，通过语义嵌入探索自然语言对表情与外观的可控操控（Hwang 等，2023）。这些方法共同突破了3DMM等线性模型在细粒度表情建模上的局限，使得生成结果质量有了显著提升。然而，NeRF的计算复杂度成为其实用化的主要瓶颈。新兴的高斯泼溅技术在保证接近NeRF渲染质量的同时显著降低了计算开销，为提升表情合成的实时性与可控性提供了新的解决方案。相关研究提出以高斯点云替代隐式体积作为表示基础，将blendshape、3DMM或FLAME等统计模型与高斯泼溅相结合，在保证表情准确性的同时捕捉精细的动态细节（Xu 等，2024）。此外，也有工作面向表情合成任务，探索由表情感知的高斯点云驱动头部与面部动画（Dhamo 等，2024），并结合音频条件实现嘴部动作与表情的实时同步，从而在细节保真度与交互性上取得显著提升。这一系列方法在表情合成方面实现了低计算复杂度，高保真面部表情细节，对实时性与交互性要求较高的情感数字人具有重要的意义。

© 中国图象图形学报版权所有

表2　情感数字人智能生成的关键技术路线与方法对比

Table2　Comparison of key technical routes and methods for intelligent generation of emotional digital humans

核心领域	方法分类	代表性技术/模型	优势	局限	数据集
面部表情合成	参数化模型驱动	单张图像回归为三维可变形模型(3D morphable model，3DMM)，AU回归为3DMM，生成对抗网络(generative adversarial network，GAN)与3DMM结合	计算量小，易于控制	表达能力受限，难以生成个性化细节	MICC，CK+，EmotiW-17，Flickr-HQ
	隐式神经场渲染	Dynamic NeRF, HeadNeRF, FDNeRF, FLAME-in-NeRF, FaceCLIPNeRF	照片级高保真渲染	高计算复杂度，难以实时渲染	FaceSEIP，FaceScape，FFHQ，VoxCeleb
	显式高斯表示	Gaussian Head Avatar, 3D Gaussian Blendshapes, HeadGaS, GaussianTalker	高保真与实时性的平衡	动态交互场景下存储开销大	NeRSemble，HAvatar，NeRFBlendShape，INSTA，AD-NeRF，GeneFace
情感语音合成	统计与端到端建模	Tacotron，WaveNet]，GST	语音自然连贯，情感表达清晰	难以解耦，精度控制不足	LJSpeech，VCTK
	显式特征解耦	FastSpeech，FastSpeech 2(Variance Adaptor)	生成速度快，韵律可控性强	耦合复杂，难以细调情感	LJSpeech
	零样本提示生成	VALL-E, NaturalSpeech, Voicebox, EmoCtrl-TTS, EmoSphere++, ECE-TTS	泛化能力强，情感复刻真实	跨文本迁移时的情感一致性挑战	LibriLight，MLS，LibriSpeech，VCTK，ESD
情感共鸣建模	图网络结构感知	DialogueGCN，DAG Network，Dynamic Graph Neural ODE，Long-Short Distance GNN	上下文依赖建模良好	被动识别为主，缺乏主动共情生成能力	IEMOCAP，AVEC，MELD，DailyDialog，EmoryNLP
	知识与大模型增强	Knowledge-Enhanced Generation，LLM Fine-tuning	深层语义理解	长期记忆维持难，灾难性遗忘	Empathetic-Dialogues，SMILECHAT
交互式情感生成	情感空间参数化	离散情感类别，连续维度情感空间	情感表达可量化，直观	难以涵盖所有复杂混合情绪	N/A
	条件生成模型	GAN条件生成，扩散模型，多模态融合，多模态大模型	动态多样的生成，跨模态一致性强	高实时渲染成本	IEMOCAP，EmoSet，CMU-MOSI，ICT-MMMO，MOUD，POM，YouTube2M

3.1.2　情感语音合成

情感语音合成旨在通过文本、音频或参数控制信号(如情感标签、韵律特征、风格向量)生成具有情感特征的语音，使合成结果在保持自然度和清晰度的同时，有效传递说话者的情绪与风格。其发展可视为传统语音合成向具备深度可控表达能力的演进

© 中国图象图形学报版权所有

过程。纵观这一历程,技术范式从早期的“统计参数化建模”,逐步转向基于深度学习的“显式特征解耦”、“非自回归细粒度控制”以及“零样本提示生成”,以不断逼近人类的情感表达效果。早期方法主要包括拼接合成和统计参数化方法,其中拼接合成通过人工录制的语音单元进行拼接,而统计参数化方法则依托声学模型预测参数,并通过声码器生成语音。这一阶段奠定了语音生成的基本框架及评价体系(Barakat 等,2024)。随着深度神经网络的发展,传统声学模型逐渐被端到端架构取代,其中Tacotron的兴起进一步提升了语音的自然度和可控性,并通过与WaveNet等高质量神经声码器结合后,实现了与真实人类语音几乎难以区分的合成效果。然而,人类语音富含情感、语调、重音及个体说话风格,因此研究者开始关注构建富有情感的类人语音合成。早期工作主要通过直接参考编码器、无监督全局风格令牌以及基于变分自编码器的潜变量方法,在有标注或无标注条件下提取并操纵韵律与情感特征(Wang 等,2018)。随后,为提高生成效率并实现细粒度韵律控制,一些研究者提出非自回归框架如FastSpeech系列与显式方差适配器,可精确调控语音的时长、基频和能量并加速推理(Ren 等,2019)。近期,受大规模自监督学习与神经音频编解码器影响,出现了基于上下文学习策略训练的零样本模型,如VALL-E、NaturalSpeech2和Voicebox,这些方法能够从少量音频提示中重建说话人的特性与情感风格,将情感语音合成推进到零样本泛化与提示式控制的新高度(Le 等,2023)。近年来,情感语音合成进一步向零样本泛化、细粒度情感控制与动态表达方向发展。一方面,以NaturalSpeech3、VALL-E 2为代表的模型在因式分解建模与提示式控制上显著提升了零样本语音的自然度与鲁棒性(Ju 等,2024);另一方面,EmoCtrl-TTS、Emo-Sphere++、ECE-TTS等方法强调对情感强度、非语言成分及时序变化的建模,从而增强语音的真实感与互动性(Wu 等,2024);此外,Prompt-Unseen-Emotion等方法探索通过提示生成未见或混合情绪,进一步拓展了情感语音合成的表达边界(Gao 等,2025)。

3.1.3 单一情感生成的关键挑战

总体而言,单一情感生成虽然在局部技术指标上不断取得进展,但在高保真与实时性的平衡、情感控制因子的解耦建模,以及跨场景动态一致性等方面仍存在系统性不足,这也从根本上揭示了仅依赖单一情感建模难以支撑复杂自然交互需求的内在局限。

1) 高保真与实时性之间的结构性矛盾仍未根本解决

在面部表情生成中,高质量几何与纹理细节往往依赖计算复杂度较高的建模范式,而轻量化与实时友好的方法则在动态精细表达能力上存在明显不足。如何在保证微表情、关键点非线性运动和局部细节稳定性的同时,兼顾低延迟推理与长序列生成的计算与存储效率,仍是单一情感表情生成难以回避的核心瓶颈。这一矛盾在交互式场景中被进一步放大,直接制约了方法的实际可部署性(Tang 等,2024)。

2) 情感与底层可控因子的高度耦合限制了细粒度建模能力

无论是面部表情还是情感语音,情感并非独立变量,而是与韵律、语义、时序节奏等多因素强耦合的高层语义属性。现有方法虽然在时长、基频、能量或表情幅度等单一因子上具备一定可控性,但在情感驱动下,这些因子的协同变化往往呈现非线性和个体差异特征,导致情感强度、风格与表达细节难以进行稳定、可解释的联合调节。这使得生成结果在合理与情感准确之间仍存在明显落差(Li 等,2025)。

3) **跨条件泛化与动态一致性不足制约真实应用:**

单一情感生成在受控场景下可以取得较好效果,但在跨文本、跨说话人或跨交互阶段时,情感表达的连续性和一致性难以保证。一方面,提示式或零样本方法对外部条件高度敏感,容易在复杂语义或多模态输入下产生情绪漂移;另一方面,情感与语义之间的强相关性使得模型在迁移过程中容易出现情感失真或语义—情绪冲突。这一问题在长时序生成与实时交互中尤为突出,成为限制系统稳定性的关键因素。

3.2 多情感交互生成

与静态表情或单帧动作生成不同,情感数字人强调在多轮交互与长时序条件下,生成与语义、情境及个体特征相一致的连续面部反应序列。在多情感交互生成中,系统不仅需要识别用户的情绪状态,还需建模个体间的情感共鸣、情绪传递与交互意图,以

© 中国图象图形学报版权所有

实现长期化、情境自适应的情感交流。

3.2.1 情感共鸣建模

情绪传递与共鸣主要通过构建情感传递机制、多模态情感融合和因果关系图谱来实现个体间的情感理解与响应。为此，现有研究主要演化出了“基于图网络的结构感知”、“基于知识增强的共情推理”以及“基于大模型的生成式交互”三大核心建模范式。一些研究者提出了多种基于图神经网络情感感知模型来捕捉话语内部及多模态特征之间的依赖关系与长程影响(Shou 等，2025)。这类方法通常通过节点表示个体情绪状态、边表示情感影响关系，并利用图卷积操作模拟情绪在社会互动网络中的传播过程。为进一步实现自然的情感共鸣，近年来学界借鉴共情理论提出了计算模型，通过构建大规模共情对话数据集、引入混合专家机制以及设计情感模仿策略，实现了从被动情感识别到主动共情理解与生成的技术跃迁(Sharma 等，2025)。近年来，知识增强方法通过引入敏感情感识别与知识选择机制，有效提升了共情响应的质量与相关性(Fu 等，2023)，与此同时，部分研究通过情感增强与特征转换建模，将对话中的情绪变化模式进行了更为细致的刻画。随着大语言模型的快速发展，基大语言模型(large language model，LLM)的情感共鸣建模逐渐成为新趋势。相关工作依托多轮共情对话数据集和专门化的微调策略，显著增强了模型在心理支持等场景中的共情表达、倾听能力与安慰效果(Qian 等，2025)。

3.2.2 交互式情感生成

交互式情感生成与响应旨在构建能够实时感知用户情绪、理解语境并生成动态、个性化回应的智能数字人系统，其核心在于赋予数字人多模态情绪感知与调控能力。具体而言，该领域主要涵盖了“情感空间参数化”、“条件生成式调控”以及“多模态推理融合”三大关键技术路径，分别侧重于解决情感的量化表达、生成的动态可控性以及跨模态语义对齐问题。研究者首先通过构建情感空间实现参数化表达，既包括基于离散类别的直观建模，也涵盖能够连续刻画强度与复杂性的维度模型，从而支持更自然的情感生成与调节(Ji 等，2021)。在此基础上，交互式调控技术不断发展，通过在生成模型中引入条件变量，使用户或系统能够动态影响生成结果。典型代表包括生成对抗网络在语音、图像与音乐等情感生成任务中的应用，以及扩散模型在跨模态高保真情感生成与可控性提升方面的探索(Yang 等，2024)。近年来，多模态推理逐渐成为该领域的重要研究方向。大模型在整合文本、语音、视觉等多源信息方面展现出强大的推理与生成潜力，不仅提升了跨模态情感对齐的准确性，也推动了动态情感生成能力的发展(Jin 等，2024)。与此同时，随着情感计算应用规模的扩大，研究者开始日益关注低计算开销的生成策略。通过在模型架构、训练范式与推理机制上的优化，这类方法能够在有效降低资源消耗的同时保持较高的生成质量，从而显著提升系统的可用性与普适性(Chen 等，2024)。总体而言，交互式情感生成与调控正引领情感计算向高精度、强灵活性与高效能的方向演进，并为构建具备深层情绪理解与共情能力的智能数字人奠定了坚实的技术基础。

3.2.3 多情感生成的关键挑战

综上所述，多情感生成的核心挑战已从单一情绪的生成转向是否能够在复杂语义条件下理解情感、在长期交互中保持一致性，并在资源受限环境中实现高质量动态表达。这些问题共同揭示了现有方法在情感结构建模、时间一致性与系统可部署性方面的深层不足，也为后续研究指明了需突破的关键方向。

1) 跨模态高阶语义消歧能力不足

多情感生成不再局限于显性情绪的识别与合成，而需要在复杂语境中综合理解语言、语音、表情及上下文所隐含的情感意图。然而，现有模型在处理反讽、隐喻、含蓄表达等高阶语言现象时，仍主要依赖表层语义或单模态线索，难以有效解析文本语义与非言语信号之间的潜在冲突与补充关系。这种跨模态语义消歧能力的不足，直接导致情感理解在复杂场景下出现偏差，使生成的多情感反应缺乏真实共鸣基础(Maharana 等，2024)。

2) 长期交互场景下情感记忆保持与人格一致性约束不足

面向陪伴式与持续交互场景，多情感生成不仅要求当前情绪合理，还要求在长时间尺度上维持稳定的人格设定与情感反应逻辑。然而，大语言模型与现有情感生成框架普遍缺乏显式的长期记忆与人格约束机制，随着交互轮次增加，模型容易遗忘关键历史信息，甚至出现情感态度与行为风格前后矛盾的问题。这种共享记忆缺失与人格漂移现象显著削

© 中国图象图形学报版权所有

弱了情感连续性,使多情感生成难以支撑深层、可信的人机情感关系(Zhu 等,2025)。

3) 高保真动态情感生成与端侧实时可部署性协同不足

在交互式多情感生成中,情感表达往往涉及多区域、多尺度的动态协同变化。尽管扩散模型在生成质量上具备明显优势,但在音频驱动条件下,唇形运动与面部微表情之间的解耦仍然困难,容易造成非语音相关区域的僵化或过度平滑,从而破坏整体情感表现的自然度。与此同时,个性化风格迁移与持续学习需求进一步放大了这一问题:模型在适应新用户情感偏好的过程中,往往伴随原有表达能力的退化,而高昂的推理与更新成本又使其难以在端侧实时运行。这种质量、可塑性与效率之间的冲突,成为多情感生成系统落地应用的关键障碍(Park 等,2025)。

3.3 情感数字人的展望

总体来看,情感数字人的发展将持续在跨模态融合、个性化共情交互以及长期自主学习等方面取得突破。依托大模型与生成式人工智能的快速演进,数字人有望实现更加自然的情绪表达与多模态协同,逐步具备对个体差异的敏感建模与自适应调控能力。与此同时,如何在大规模交互中保持情感一致性与可信度,以及在医疗康养、教育辅导、虚拟社交等应用场景中实现安全可控的部署,将成为未来待解决的核心问题与研究重点。

4 端上实时交互数字人

端上数字人具有低延迟交互,离线场景可用等优势,因此同样具有较高的应用与研究价值。目前行业中也已出现可以运行在Windows、IOS、Android、小程序等多个终端操作系统的端上实时数字人,并适配了PC、手机、大屏等多种硬件场景。

4.1 端上实时数字人的实现方式

在端上设备上实现实时交互的数字人系统,通常需要在计算资源受限、响应延迟严格以及渲染帧率要求较高的条件下进行系统设计。因此,相关方法在模型结构、生成框架以及部署方式等方面,往往需要针对端上场景进行专门优化。

在2D渲染场景下,端上数字人多以视频或图像序列的形式呈现,其核心目标是在有限算力条件下实现可接受的视觉质量和稳定的实时响应。生成算法的计算开销主要与模型规模和输出分辨率相关,因此现有方法普遍采用参数量较小的网络结构作为骨干,并将生成分辨率控制在相对较低的水平。

除模型规模和生成区域的控制外,一些方法还通过引入中间表示来进一步降低端上计算负担。以RealTalk为例,其采用轻量化模型先对语音特征与面部三维参数之间的映射关系进行建模,再通过渲染得到中间结果,从而降低后续图像生成阶段的难度。在实际部署过程中,端上系统通常结合面向移动设备优化的推理框架(如NCNN),以提升模型推理效率并改善整体交互体验。

相比2D渲染方案,基于3D渲染的端上数字人实现路径在系统结构上相对直接。该类方法通常依赖预先构建的三维数字人模型,并通过一组多维的blend shape来描述面部表情和唇部运动变化。blend shape的维度一般在数十维范围内,维度数量的增加有助于提升说话过程中唇部细节和表情变化的表现能力。在驱动方式上,可以复用语音到三维面部参数映射的模型结构,将语音特征转换为对应的blend shape系数,从而实现实时的语音驱动效果。这类方案在渲染阶段主要依赖图形管线完成,因此在端上设备上通常具有较为稳定的帧率表现。

4.2 端上数字人行业成熟方案举例

随着相关技术的逐步成熟,端上数字人已在实际应用中形成了一定数量的工程化解决方案,并实现了对多种操作系统和硬件设备的适配。从行业实践来看,成熟方案通常以SDK的形式提供,支持2D和3D数字人两种表现形式,并覆盖主流移动端和桌面端平台。

以腾讯云智能数智人为例,其端上数字人方案提供了面向iOS和Android的移动端SDK,同时支持手机应用内嵌、小程序、H5形态以及Windows平台的部署,并可运行于智能手机、平板电脑、个人计算机及智慧屏等多类终端设备。这类方案在系统设计上通常将计算密集型任务与端上实时渲染进行合理拆分,从而在保证交互流畅性的同时,降低端侧的算力压力。

4.3 端上数字人与大模型对话模型的结合

随着大语言模型在自然语言理解和生成能力上的提升,端上数字人与大模型对话系统的结合逐渐成为一种常见的系统架构。在该模式下,文本内容

© 中国图象图形学报版权所有

生成和复杂语义推理通常在云端完成，而语音合成、数字人驱动与实时渲染则在端上执行，从而形成端云协同的交互流程。

在实际应用中，这种架构已被用于新闻播报、智能客服和内容交互等场景。例如《焦点访谈》节目中主持人与其数字分身进行对话的案例，即采用了大语言模型与端上数字人相结合的方式。在大模型的支持下，系统能够整合语音交互、自然语言理解等多种能力，并通过引入领域知识，使数字人在对话过程中具备一定的知识覆盖和推理能力。通过端云协同的方式，既可以利用大模型的表达和理解优势，又能够在端上保证数字人交互的实时性和稳定性。

5 数据集、评估指标与未来展望

5.1 关键数据集与评估指标

数字人相关技术的发展高度依赖高质量数据集与合理评估体系的支撑。由于不同研究方向在建模目标、任务设定和应用场景上存在明显差异，各子领域逐步形成了具有针对性的数据基础和评估方式。从整体来看，当前研究主要围绕3D人体运动合成与编辑、情感数字人建模以及视频到数字人生成三个方向展开。

在3D人体运动合成与编辑领域，通用数据集为模型训练和方法对比提供了基础支撑。其中，HumanML3D数据集（Guo 等，2022）和KIT-ML数据集（Plappert 等，2018）是目前使用较为广泛的文本到运动数据集，覆盖了多种常见人体动作类型，被用于基础模型预训练和通用能力评估。在此基础上，研究逐渐向更具针对性的任务拓展，同时出现了一批面向细分问题的数据集。例如，MotionFix数据集主要服务于运动编辑任务，强调源动作与编辑指令之间的对应关系（Athanasiou 等，2024）；PerMo关注运动风格和个性化建模问题（Kim 等，2025）；而InterHuman、Harmony4D与Hi4D等数据集则面向双人或多人交互场景，为研究复杂时序协同和空间关系提供了数据支持（Liang 等，2024；Jiang 等，2024；Yin 等，2023）。

在情感数字人研究中，数据集通常围绕表情、语音和对话三个层面构建。面部表情合成相关数据多来源于3D人脸扫描或多视角采集，涵盖多种基础表情和表情强度变化，被广泛用于表情参数建模和动画生成任务（Blanz and Vetter，1999；Tewari 等，2017）。情感语音数据集则包含不同情感类别和强度的语音样本，为情感语音合成和情感识别提供训练数据（Shen 等，2018；Ren 等，2020）。此外，共情对话数据集逐渐受到关注，例如Rashkin等人（2018）构建的大规模共情对话语料，为情感理解和共情响应生成提供了重要数据基础。

在视频到数字人生成领域，数据集通常根据采集方式和场景复杂度进行区分。单目视频数据集为单视角人体重建和时序建模提供了基础样本（Kocabas 等，2020）；多目视频数据集通过多摄像头同步采集，为高精度几何重建和动态建模提供支持（Pons-Moll 等，2017）；而面向真实应用场景的开放数据集则包含遮挡、多主体交互和复杂环境因素，更有助于评估模型在复杂条件下的鲁棒性（Zhang 等，2024）。

在评估方面，不同方向的模型性能通常从多个维度进行衡量。在人体运动合成与编辑任务中，运动质量和分布相似性是常用评估指标，例如通过FID或MM-Dist等度量生成运动与真实数据之间的差异（Tevet 等，2023）。此外，研究中也常通过脚部滑动、姿态抖动等统计指标，对生成运动的稳定性和物理合理性进行分析。在文本或条件驱动任务中，生成结果与语义指令之间的匹配程度同样是重要评估维度。

对于情感数字人，评估通常结合客观指标与主观评价。表情自然度可通过面部关键点误差等方式进行量化，也常结合人工评分进行综合判断（Chang 等，2018）；语音情感匹配度通常通过情感分类准确率或相似度指标进行评估（Akuzawa 等，2018）；而共情响应质量更多依赖人工评测，用于判断生成回复在语境理解和情感支持方面的合理性（Rashkin 等，2018）。

在视频到数字人生成任务中，几何精度、纹理保真度和动态一致性是常用评估维度。几何误差通常通过顶点距离或重建误差进行衡量，纹理质量关注外观细节与真实人体的一致程度，而动态一致性则用于评估数字人在运动过程中的稳定性和连贯性（Guo 等，2024）。

© 中国图象图形学报版权所有

6 结 论

本文系统综述了智能数字人内容生成技术的前沿进展，深度梳理了“视频到数字人生成”、“3D人体运动合成与编辑”以及“情感数字人智能生成”三大核心领域的主流模型与技术演进。总体而言，数字人技术正遵循“形态重建—动作赋予—情感交互”的递进逻辑，从静态、单一形态向动态、多模态与情感化交互深刻变革。尽管相关技术取得了显著突破，但面向真实的复杂应用场景，各子领域仍面临关键瓶颈：在视频到数字人生成领域，复杂环境下的鲁棒性与实时性难以兼顾，严重遮挡与多人交互场景的重建精度易下降；在3D人体运动合成与编辑方向，高质量生成与计算效率的平衡仍是核心挑战，多人交互场景中的动作协同与物理合理性有待提升；在情感数字人研究中，细腻情感的表达及多模态间的一致性存在不足，长期交互场景下的情感记忆与人格连贯性亟待解决。展望未来，智能数字人技术将在跨领域深度融合中持续演进，构建出更加自然拟真的端到端系统。多模态大模型与生成式前沿方法的引入，为复杂开放场景中的意图理解和行为生成提供了全新范式；同时，依托模型压缩与端云协同加速的轻量化实时化研究，将全面推动情智兼备的数字人系统在实际应用中的规模化部署。

7 致 谢

致谢：本文由中国图象图形学学会数字娱乐与智能生成专业委员会组织撰写，该专业委员会链接为 https://www.csig.org.cn/16/201612/49316.html，在此表示感谢。

参考文献(References)

Akuzawa K, Iwasawa Y and Matsuo Y. 2018. Expressive speech synthesis via modeling expressions with variational autoencoder[EB/OL]. [2024-02-20]. https://arxiv.org/pdf/1804.02135.pdf

Athanasiou N, Mathis P, Black M J, Tang S, Bolkart T and Romero J. 2024. MotionFix: text-driven 3D human motion editing[EB/OL]. [2024-08-05]. https://arxiv.org/pdf/2408.00712.pdf

Athar S R, Shu Z and Samaras D. 2023. FLAME-in-NeRF: neural control of radiance fields for free view face animation//Proceedings of the 17th IEEE International Conference on Automatic Face and Gesture Recognition. Waikoloa, USA: IEEE: 1-8[DOI: 10.1109/FG57933.2023.10042583]

Barakat H, Turk O and Demiroglu C. 2024. Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources. EURASIP Journal on Audio, Speech, and Music Processing, 2024 (1): 11 [DOI: 10.1186/s13636-024-00325-w]

Blanz V and Vetter T. 1999. A morphable model for the synthesis of 3D faces//Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. Los Angeles, USA: ACM: 187-194 [DOI: 10.1145/311535.311556]

Cao Y, Wen H, Rong K, Liu Y, Li Z and Liu Z. 2020. Nonparallel emotional speech conversion using VAE-GAN//Proceedings of the Interspeech 2020. Shanghai, China: ISCA: 3885-3889 [DOI: 10.21437/Interspeech.2020-1178]

Chang F J, Tran A T, Hassner T, Masi I, Nevatia R and Medioni G. 2018. ExpNet: landmark-free, deep, 3D facial expressions//Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition. Xi'an, China: IEEE: 122-129[DOI: 10.1109/FG.2018.00027]

Chen H, Jiang B, Liu W, Huang Z, Fu B, Chen T, et al. 2024. PBDyG: position based dynamic Gaussians for motion-aware clothed human avatars[EB/OL]. [2024-05-10]. https://arxiv.org/pdf/2405.17890.pdf

Chen J, Mao X, Chen X and Cai L. 2021. An improved StarGAN for emotional voice conversion//Proceedings of the Interspeech 2021. Brno, Czech Republic: ISCA: 3800-3804 [DOI: 10.21437/Interspeech.2021-1234]

Chen M, Wei Y, Zhang C, Liu L, Tang Y and Lu J. 2024. Vid2Avatar-Pro: authentic avatar from videos in the wild via universal prior [EB/OL]. [2024-03-15]. https://arxiv.org/pdf/2403.05678.pdf

Chen S, Liu S, Zhou L, Li J, Wu X, Liu S, et al. 2024. VALL-E 2: neural codec language models are human parity zero-shot text to speech synthesizers[EB/OL]. [2024-06-15]. https://arxiv.org/pdf/2406.05370.pdf

Chen X, Jiang B, Liu W, Huang Z, Fu B, Chen T, et al. 2023. Executing your commands via motion diffusion in latent space//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE/CVF: 1984-1993[DOI: 10.1109/CVPR52729.2023.00196]

Chen X, Wang H and Liu Y. 2024. FATE: full head Gaussian avatar with textural editing from monocular video [EB/OL]. [2024-04-20]. https://arxiv.org/pdf/2411.15604.pdf

Chen Y, Wang L, Li Q, et al. 2024. MonoGaussianAvatar: monocular

© 中国图象图形学报版权所有

Gaussian point-based head avatar//ACM SIGGRAPH 2024 Conference Papers. Denver, USA: ACM.

Chen Y, Xing X, Lin J, et al. 2023. SoulChat: improving LLMs′ empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations[EB/OL]. [2023-11-01]. https://arxiv.org/pdf/2311.00273.pdf

Chen Z, Wang Y and Liu H. 2024. RGBAvatar: reduced Gaussian blend shapes for online modeling of head avatars[EB/OL]. [2024-02-10]. https://arxiv.org/pdf/2402.05678.pdf

Cho D H, Oh H S, Kim S B and Choi S. 2025. EmoSphere++: emotion-controllable zero-shot text-to-speech via emotion-adaptive spherical vector. IEEE Transactions on Affective Computing, 2025, early access [DOI:10.1109/TAFFC.2024.3406692]

Cho K, Lee J, Yoon H, et al. 2024. GaussianTalker: real-time high-fidelity talking head synthesis with audio-driven 3D Gaussian splatting[EB/OL]. [2024-04-25]. https://arxiv.org/pdf/2404.16012.pdf

Cong Y, Zhang H, Liu J, Wang Q, Li S and Chen R. 2025. SemGeoMo: dynamic contextual human motion generation with semantic and geometric guidance[EB/OL]. [2025-03-05]. https://arxiv.org/pdf/2503.01291.pdf

Curreli G, De Martini D, Melzi S, Dotti N, Rodola E and Ovsjanikov M. 2025. Skeleton diffusion: nonisotropic Gaussian diffusion for realistic 3D human motion prediction[EB/OL]. [2025-01-12]. https://arxiv.org/pdf/2501.06035.pdf

Dhamo H, Nie Y, Moreau A, et al. 2024. HeadGaS: real-time animatable head avatars via 3D Gaussian splatting//Proceedings of the European Conference on Computer Vision. Cham, Switzerland: Springer: 459-476[DOI: 10.1007/978-3-031-72983-6_26]

Ekman P. 1992. An argument for basic emotions. Cognition & Emotion, 6(3－4): 169－200 [DOI: 10.1080/02699939208411068]

Fu F, Zhang L, Wang Q, et al. 2023. E-CORE: emotion correlation enhanced empathetic dialogue generation//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore, Singapore: ACL: 10568-10586 [DOI: 10.18653/v1/2023.emnlp-main.653]

Gafni G, Thies J, Zollhofer M, et al. 2021. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 8649-8658 [DOI: 10.1109/CVPR46437.2021.00854]

Gao X, Zhang H and Chen N F. 2025. Prompt-unseen-emotion: zero-shot expressive speech synthesis with prompt-LLM contextual knowledge for mixed emotions[EB/OL]. [2025-06-05]. https://arxiv.org/pdf/2506.02742.pdf

Ghosal D, Majumder N, Poria S, et al. 2019. DialogueGCN: a graph convolutional neural network for emotion recognition in conversation//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Hong Kong, China: ACL: 154-164 [DOI: 10.18653/v1/D19-1015]

Gong J, Foo L G, Fan Z, Ke Q, Rahmani H and Liu J. 2023. TM2D: bimodality driven dance generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE/CVF: 1792-1801 [DOI: 10.1109/CVPR52729.2023.00178]

Guo C, Zou S, Zuo X, Wang S, Ji W and Li X. 2020. Action2motion: conditioned generation of 3D human motions with graph convolutional networks//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 2021-2029 [DOI: 10.1145/3394171.3413941]

Guo Y, Liu H and Wang Z. 2024. WonderHuman: hallucinating unseen parts in dynamic 3D human reconstruction//Proceedings of the European Conference on Computer Vision. Cham, Switzerland: Springer: 1-16.

Han L, Huang J, Wang H, Zhang X, Li Y, Ma Z, et al. 2024. AToM: aligning text-to-motion model at event-level with GPT-4Vision reward[EB/OL]. [2024-06-01]. https://arxiv.org/pdf/2405.15570.pdf

Hong S, Lee J, Kim D, Park Y, Choi M and Kwon T. 2025. SALAD: skeleton-aware latent diffusion for text-driven motion generation and editing[EB/OL]. [2025-03-22]. https://arxiv.org/pdf/2503.13836.pdf

Hong Y, Peng B, Xiao H, et al. 2022. HeadNeRF: a real-time NeRF-based parametric head model//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 20374-20384 [DOI: 10.1109/CVPR52688.2022.01974]

Hu S, Zhang Y, Liu H and Wang Y. 2024. GauHuman: articulated Gaussian splatting from monocular human videos//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE

Hua W, Liu Y, Zhang J, Wang Z, Chen S and Li Q. 2025. DSDFM: deterministic-to-stochastic diverse latent feature mapping for humanmotionsynthesis[EB/OL]. [2025-05-10]. https://arxiv.org/pdf/2505.00998.pdf

Huang S, Wang J, Li Y, Chen H, Zhou T and Liu X. 2024. Move-in-2D: 2D-conditioned human motion generation[EB/OL]. [2024-12-18]. https://arxiv.org/pdf/2412.13185.pdf

Huang Y, Liu Z, Wang H and Chen X. 2024. Closely interactive human reconstruction with proxemics and physics-guided//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1-10.

Hwang S, Hyung J, Kim D, et al. 2023. FaceCLIPNeRF: text-driven 3D face manipulation using deformable neural radiance fields//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3469-3479 [DOI: 10.1109/

© 中国图象图形学报版权所有

ICCV51070.2023.00320]

Ji X, Zhou H, Wang K, et al. 2021. Audio-driven emotional video portraits//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14080-14089 [DOI: 10.1109/CVPR46437.2021.01386]

Jiang B, Chen X, Liu W, Yu J, Yu G and Chen T. 2024. MotionGPT: human motion as a foreign language//Advances in Neural Information Processing Systems. New Orleans, USA: NeurIPS [DOI: 10.48550/arXiv.2306.14795]

Jiang B, Zhang Y, Wei Y, Li X and Liu Z. 2023. InstantAvatar: learning avatars from monocular video in 60 seconds//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE/CVF: 16922-16932 [DOI: 10.1109/CVPR52729.2023.01620]

Jiang B, Zhang Y, Wei Y, et al. 2024. Harmony4D: a video dataset for in-the-wild close human interactions//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1-10.

Jin Y, Choi M, Verma G, et al. 2024. MM-SoC: benchmarking multimodal large language models in social media platforms [EB/OL]. [2024-02-21]. https://arxiv.org/pdf/2402.14154.pdf

Ju Z, Wang Y, Shen K, et al. 2024. NaturalSpeech 3: zero-shot speech synthesis with factorized codec and diffusion models [EB/OL]. [2024-03-05]. https://arxiv.org/pdf/2403.03100.pdf

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

Khakhulin T, Sklyarova V, Lempitsky V and Zakharov E. 2023. Facial appearance capture at home with patch-level reflectance prior. ACM Transactions on Graphics, 42(4): Article 145 [DOI: 10.1145/3592418]

Kim J, Park S, Lee H, Choi Y, Jung H and Kwon S. 2025. PersonaBooth: personalized text-to-motion generation [EB/OL]. [2025-03-15]. https://arxiv.org/pdf/2503.07390.pdf

Kim W, Ahn Y, Kim D, et al. 2022. Emp-RFT: empathetic response generation via recognizing feature transitions between utterances//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics. Seattle, USA: ACL: 4118-4128 [DOI: 10.18653/v1/2022.naacl-main.306]

Le M, Vyas A, Shi B, et al. 2023. Voicebox: text-guided multilingual universal speech generation at scale//Advances in Neural Information Processing Systems. New Orleans, USA: NeurIPS: 14005-14034 [DOI: 10.48550/arXiv.2306.15687]

Leung J, Zhang Y and Liu H. 2024. GART: Gaussian articulated template models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE

Li T, Bolkart T, Black M J, Li H and Romero J. 2017. Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics, 36(6): 194: 1-194: 17 [DOI: 10.1145/3130800.3130813]

Li X, Wang G, Wang Y, et al. 2023. Mixed knowledge-enhance empathetic dialogue generation//Proceedings of the 2023 International Conference on Electronics, Computers and Communication Technology. Dalian, China: IEEE: 77-81 [DOI: 10.1109/ICECCT57849.2023.10183764]

Li X, Zhang Y and Liu H. 2025. Long-short distance graph neural networks and improved curriculum learning for emotion recognition in conversation//Proceedings of the European Conference on Artificial Intelligence. Santiago, Spain: ECAI: 1-8.

Li Y, Liu H and Wang Z. 2024. GoMAvatar: efficient animatable human modeling from monocular video using Gaussians-on-mesh//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1-10.

Li Z, Wang Y, Chen X, Liu J, Zhang H and Zhou K. 2025. SimMotionEdit: text-based human motion editing with motion similarity prediction [EB/OL]. [2025-03-28]. https://arxiv.org/pdf/2503.18211.pdf

Li Z, Zheng Z, Wang L, et al. 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, USA: IEEE: 19711-19722 [DOI: 10.1109/CVPR52733.2024.01867]

Li Z, Zheng Z, Wang L, et al. 2024. 3DGS-Avatar: animatable avatars via deformable 3D Gaussian splatting//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE

Li Z, Zheng Z, Wang L, et al. 2024. Expressive whole-body 3D Gaussian avatar//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1-10.

Liang H, Zhang W, Li W, Wang Y, Liu Z and Chen X. 2024. InterGen: diffusion-based multi-human motion generation under complex interactions//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1789-1799 [DOI: 10.1109/CVPR52733.2024.00175]

Liang S, Zhou R and Yuan Q. 2025. ECE-TTS: a zero-shot emotion text-to-speech model with simplified and precise control. Applied Sciences, 15(9): 5108 [DOI: 10.3390/app15095108]

Liao Y, Sun Z, Wang H, Chen Y, Li J and Zhang X. 2025. Shape my moves: text-driven shape-aware synthesis of human motions [EB/OL]. [2024-04-09]. https://arxiv.org/pdf/2504.03639.pdf

Lin Z, Madotto A, Shin J, et al. 2019. MoEL: mixture of empathetic listeners [EB/OL]. [2019-08-21]. https://arxiv.org/pdf/1908.07687.pdf

Liu H, Chen X, Zhang Y, Wang Q, Li J and Zhou K. 2024. ProgMo-

© 中国图象图形学报版权所有

Gen：programmable motion generation for open-set motion control tasks//Proceedings of the Advances in Neural Information Processing Systems. New Orleans，USA：NeurIPS［DOI：10.48550/arXiv.2312.16484］

Liu Y，Huang J，Wang H，et al. 2025. ODHSR：online dense 3D reconstruction of humans and scenes from monocular videos［EB/OL］.［2024-04-20］. https://arxiv.org/pdf/2504.13167.pdf

Liu Y，Wang H and Chen X. 2024. GAF：Gaussian avatar reconstruction from monocular videos via multi-view diffusion//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：1-10.

Loper M，Mahmood N，Romero J，Pons-Moll G and Black M J. 2015. SMPL：a skinned multi-person linear model. ACM Transactions on Graphics，34（6）：248：1-248：16［DOI：10.1145/2816795.2818013］

Ma S，Weng Y，Shao T，et al. 2024. 3D Gaussian blendshapes for head avatar animation//ACM SIGGRAPH 2024 Conference Papers. Denver，USA：ACM：1-10［DOI：10.1145/3641519.3657424］

Majumder N，Hong P，Peng S，et al. 2020. MIME：mimicking emotions for empathetic response generation//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Online：ACL：8968-8979［DOI：10.18653/v1/2020.emnlp-main.721］

Wu Z，Zhu H，Lin G F，et al. 2026. Lightweight stereo matching network for edge computing devices. Journal of Image and Graphics，31(2)：589-608（武忠，朱虹，蔺广逢，等. 2026. 面向边缘计算设备的轻量级双目立体匹配网络. 中国图象图形学报，31(2)：589-608）［DOI：10.11834/jig.250081］

Mehrabian A. 1996. Pleasure-arousal-dominance：a general framework for describing and measuring individual differences in temperament. Current Psychology，14（4）：261-292［DOI：10.1007/BF02686918］

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

Ota M，Sato Y，Tanaka K，Suzuki T，Nakamura R and Ito H. 2025. PINO：person-interaction noise optimization for long-duration and customizable motion generation of arbitrary-sized groups［EB/OL］.［2025-07-18］. https://arxiv.org/pdf/2507.19292.pdf

Pavlakos G，Choutas V，Ghorbani N，Bolkart T，Osman A A and Tzionas D. 2019. Expressive body capture：3D hands，face，and body from a single image//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach，USA：IEEE：10975-10985［DOI：10.1109/CVPR.2019.01123］

Peng S，Zhang Y，Xu Y，Wang Q，Shuai Q and Bao H. 2021. Neural Body：implicit neural representations with structured latent codes for novel view synthesis of dynamic humans//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville，USA：IEEE：9054-9063［DOI：10.1109/CVPR46437.2021.00894］

Petrovich M，Black M J and Varol G. 2021. Action-conditioned 3D human motion synthesis with transformer VAE//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal，Canada：IEEE：10985-10995［DOI：10.1109/ICCV48922.2021.01083］

Qian Y，Zhang W N and Liu T. 2023. Harnessing the power of large language models for empathetic response generation：empirical investigations and improvements［EB/OL］.［2023-10-10］. https://arxiv.org/pdf/2310.05140.pdf

Qiu H，He H，Zhang S，et al. 2023. Smile：single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support［EB/OL］.［2023-05-01］. https://arxiv.org/pdf/2305.00450.pdf

Rashkin H，Smith E M，Li M and Boureau Y L. 2018. Towards empathetic open-domain conversation models：a new benchmark and dataset//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence，Italy：ACL：5370-5381［DOI：10.18653/v1/P19-1534］

Rempe D，Birdal T，Hertzmann A，et al. 2022. Guess the unseen：dynamic 3D scene reconstruction from partial 2D glimpses//Proceedings of the IEEE/CVF International Conference on Computer Vision. Tel Aviv，Israel：IEEE：1-10.

Ren Y，Hu C，Tan X，et al. 2020. FastSpeech 2：fast and high-quality end-to-end text to speech//Proceedings of the International Conference on Learning Representations. Addis Ababa，Ethiopia：ICLR［DOI：10.48550/arXiv.2006.04558］

Ren Y，Ruan Y，Tan X，et al. 2019. FastSpeech：fast，robust and controllable text to speech//Advances in Neural Information Processing Systems. Vancouver，Canada：NeurIPS：3171-3180.

Ruiz-Ponce M，Perez J，Martinez A，Gomez R，Lopez D and Salas J. 2025. MixerMDM：learnable composition of human motion diffusion models［EB/OL］.［2025-04-08］. https://arxiv.org/pdf/2504.01019.pdf

Shafir Y，Tevet G，Kapon R，Bar A，Gordon B and Bermano A H. 2024. Human motion diffusion as a generative prior//Proceedings of the International Conference on Learning Representations. Kigali，Rwanda：ICLR［DOI：10.48550/arXiv.2303.01418］

Shao M，Weng Y and Zhou K. 2024. SplattingAvatar：realistic real-time human avatars with mesh-embedded Gaussian splatting//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE

Sharma A，Lin I W，Miner A S，Atkins D C and Althoff T. 2023. Human－AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nature Machine Intelligence，5(1)：46-57［DOI：10.1038/s42256-022-00593-2］

© 中国图象图形学报版权所有

Shen J, Pang R, Weiss R J, et al. 2018. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions//Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE: 4779-4783 [DOI:10.1109/ICASSP.2018.8461360]

Shen K, Ju Z, Tan X, et al. 2023. NaturalSpeech 2: latent diffusion models are natural and zero-shot speech and singing synthesizers [EB/OL]. [2023-04-18]. https://arxiv.org/pdf/2304.09116.pdf

Shen W, Wu S, Yang Y, et al. 2021. Directed acyclic graph network for conversational emotion recognition//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. Online: ACL: 1551-1560 [DOI: 10.18653/v1/2021.acl-long.123]

Shou Y, Meng T, Ai W, et al. 2024. Dynamic graph neural ordinary differential equation network for multi-modal emotion recognition in. conversation[EB/OL]. [2024-12-05]. https://arxiv.org/pdf/2412.02935.pdf

Tang H, Liu H and Wang Z. 2023. EmoMix: emotion mixing via diffusion models for emotional speech synthesis//Proceedings of the Interspeech 2023. Dublin, Ireland: ISCA: 4009-4013 [DOI: 10.21437/Interspeech.2023-228]

Tang Z, Li Y, Wang J, Chen M, Zhou T and Liu H. 2025. Stochastic human motion prediction with memory of action transition and action characteristic[EB/OL]. [2025-07-02]. https://arxiv.org/pdf/2507.04062.pdf

Tencent Cloud. 2024. Intelligent digital human technical documentation [EB/OL]. [2024-02-10]. https://cloud.tencent.com/document/product/1240/118294

Tencent.2024.Focus Interview: smart digital human use cases[EB/OL]. [2024-02-10]. https://cloud.tencent.com/developer/article/2384707

Tencent. 2024. NCNN: a high-performance neural network inference framework optimized for the mobile platform[EB/OL]. [2024-01-20]. https://github.com/Tencent/ncnn

Tevet G, Raab S, Gordon B, Bermano A H and Cohen-Or D. 2023. Human motion diffusion model//Proceedings of the International Conferenceon Learning Representations. Kigali, Rwanda: ICLR [DOI: 10.48550/arXiv.2209.14916]

Tevet G, Gordon B, Hertz A, Bermano A H and Cohen-Or D. 2022. MotionCLIP: exposing human motion generation to CLIP space//Proceedings of the European Conference on Computer Vision. TelAviv, Israel: Springer: 358-374 [DOI: 10.1007/978-3-031-19803-8_21]

Deng T S, Cai G Y, Dong K, et al. 2026. Cross-modal perception dialogue emotion recognition by decoupling emotion dependency relationships. Journal of Image and Graphics, 31(2): 525-540 (邓天生, 蔡国永, 董凯, 等. 2026. 解耦情绪依赖关系的跨模态感知对话情绪识别. 中国图象图形学报, 31(2): 525-540) [DOI:10.11834/jig.250309]

Tewari A, Elgharib M, Bharaj G, et al. 2020. StyleRig: rigging StyleGAN for 3D control over portrait images//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6142-6151[DOI:10.1109/CVPR42600.2020.00618]

Tewari A, Zollhofer M, Kim H, et al. 2017. MoFA: model-based deep convolutional face autoencoder for unsupervised monocular reconstruction//Proceedings of the IEEE International Conference on Computer Vision Workshops. Venice, Italy: IEEE: 1274-1283 [DOI: 10.1109/ICCVW.2017.153]

Tuan Tran A, Hassner T, Masi I, Paz E and Medioni G. 2017. Regressing robust and discriminative 3D morphable models with a very deep neural network//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5163-5172 [DOI: 10.1109/CVPR.2017.548]

Wang C, Chen S, Wu Y, et al. 2023. Neural codec language models are zero-shot text to speech synthesizers[EB/OL]. [2023-01-05]. https://arxiv.org/pdf/2301.02111.pdf

Wang L, Li J, Lin Z, et al. 2022. Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection[EB/OL]. [2022-10-25]. https://arxiv.org/pdf/2210.11715.pdf

Wang Y, Liu Z, Zhang J, Chen H, Li X and Zhou K. 2024. TIMotion: temporal and interactive framework for efficient human-human motion.generation[EB/OL]. [2024-08-30]. https://arxiv.org/pdf/2408.17135.pdf

Wang Y, Skerry-Ryan R J, Stanton D, et al. 2017. Tacotron: towards end-to-end Speechsynthesis[EB/OL]. [2017-03-29]. https://arxiv.org/pdf/1703.10135.pdf

Wang Y, Stanton D, Zhang Y, et al. 2018. Style tokens: unsupervised style modeling, control and transfer in end-to-end speech synthesis//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR: 5180-5189.

Wang Z, Liu H and Chen X. 2024. SeqAvatar: sequential Gaussian avatar with hierarchical motion context[EB/OL]. [2024-03-20]. https://arxiv.org/pdf/2411.16768.pdf

Weng C Y, Curless B, Srinivasan P P, et al. 2022. HumanNeRF: free-viewpoint rendering of moving people from monocular video//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 16210-16220 [DOI: 10.1109/CVPR52688.2022.01573]

Wu H, Wang X, Eskimez S E, et al. 2024. Laugh now cry later: controlling time-varying emotional states of flow-matching-based zero-shot text-to-speech//Proceedings of the 2024 IEEE Spoken Language Technology Workshop. Macau, China: IEEE: 690-697 [DOI:10.1109/SLT54892.2024.1002345]

Wu M, Wang Y, Zhang Q and Liu H. 2022. Structured local radiance fields for human avatar modeling//Proceedings of the IEEE/CVF

© 中国图象图形学报版权所有

Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 1-10 [DOI: 10.1109/CVPR52688.2022.01574]

Xiang J, Liu H and Wang Z. 2024. FlashAvatar: high-fidelity head avatar with efficient Gaussian embedding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1-10.

Xu Y, Chen B, Li Z, 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, USA: IEEE: 1931-1941 [DOI: 10.1109/CVPR52733.2024.00190]

Xu Z, Zhang J, Liew J H, et al. 2024. MagicAnimate: temporally consistent human image animation using diffusion model//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1-10.

Yang J, Feng J and Huang H. 2024. EmoGen: emotional image content generation with text-to-image diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6358-6368 [DOI: 10.1109/CVPR52733.2024.00609]

Yang L, Zhang H and Chen X. 2024. TaoAvatar: real-time lifelike full-body talking avatars for augmented reality via 3D Gaussian splatting [EB/OL]. [2024-04-15]. https://arxiv.org/pdf/2503.17032.pdf

Jia D, Zhao C, Zhang H X, et al. 2026. Frequency-guided lightweight RGB-D semantic segmentation network. Journal of Image and Graphics, 31(2): 479-498 (贾迪, 赵辰, 张华修, 等. 2026. 融合频域引导的RGB-D轻量级语义分割网络. 中国图象图形学报, 31(2): 479-498) [DOI:10.11834/jig.250212]

Yang Z, Liu H and Chen X. 2023. Creating your editable 3D photorealistic avatar with tetrahedron-constrained Gaussian splatting//ACM SIGGRAPH Asia 2023 Conference Papers. Sydney, Australia: ACM: 1-9.

Yin Y, Birdal T, Zhang Y, et al. 2023. Hi4D: 4D instance segmentation of close human interaction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE/CVF: 1-10.

Yuan Y, Song J, Iqbal U, Vahdat A and Kautz J. 2023. PhysDiff: physics-guided human motion diffusion model//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 16010-16021 [DOI: 10.1109/ICCV51070.2023.01472]

Li Y, Huang J, Zou Q, et al. 2026. Multi-frequency phase unwrapping and three-dimensional imaging using conditional diffusion model. Journal of Image and Graphics, 31(2): 609-627 (李妍, 黄霁, 邹勤, 等. 2026. 利用条件扩散模型的多频相位解包裹与三维成像. 中国图象图形学报, 31(2): 609-627) [DOI:10.11834/jig.250133]

Zadeh A, Liang P P, Poria S, et al. 2018. Multi-attention recurrent network for human communication comprehension//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI: 5642-5649.

Zhan F, Liu H and Wang Z. 2023. GaussianAvatars: photorealistic head avatars with rigged 3D Gaussians [EB/OL]. [2023-11-20]. https://arxiv.org/pdf/2311.12345.pdf

Zhang C, Yan J, Wei Y, Li J, Liu L, Tang Y, et al. 2023. OccNeRF: rendering humans from object-occluded monocular videos//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 1-10 [DOI: 10.1109/ICCV51070.2023.01480]

Zhang H, Song R, Wang L, et al. 2022. Classification of brain disorders in rs-fMRI via local-to-global graph neural networks. IEEE Transactions on Medical Imaging, 42(2): 444-455 [DOI: 10.1109/TMI.2022.3218765]

Gong J X, Xu J D and Sun H Q. 2026. Wavelet-suppressed interactive diffusion model for dual-channel blind image separation. Journal of Image and Graphics, 31(2): 465-478 (龚嘉鑫, 徐金东, 孙浩钦. 2026. 面向双通道盲图像分离的小波抑制交互扩散模型. 中国图象图形学报, 31(2): 465-478) [DOI:10.11834/jig.250230]

Zhang H, Wang Z and Liu H. 2024. ToMiE: towards explicit exoskeleton for the reconstruction of complicated 3D human avatars//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1-10.

Zhang J, Chen X, Liu H and Wang Z. 2024. OccGaussian: 3D Gaussian splatting for occluded human rendering [EB/OL]. [2024-03-10]. https://arxiv.org/pdf/2404.08449.pdf

Zhang J, Li X, Wan Z, et al. 2022. FDNeRF: few-shot dynamic neural radiance fields for face reconstruction and expression editing//ACM SIGGRAPH Asia 2022 Conference Papers. Daegu, Korea: ACM: 1-9 [DOI: 10.1145/3550469.3555386]

Zhang J, Liu H, Wang Y, Chen X, Li Z and Zhou K. 2024. EnergyMoGen: compositional human motion generation with energy-based diffusion model in latent space [EB/OL]. [2024-12-20]. https://arxiv.org/pdf/2412.14706.pdf

Zhang J, Liu Z, Wang H and Chen X. 2024. GaussianAvatar: towards realistic human avatar modeling from a single video via animatable 3D Gaussians//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1-10.

Zhang J, Zhang Y, Cun X, et al. 2023. T2M-GPT: generating human motion from textual descriptions with discrete representations//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE/CVF: 1472-1482 [DOI: 10.1109/CVPR52729.2023.00147]

Zhang Y, Black M J and Tang S. 2023. Link to the past: temporal propagation for fast 3D human reconstruction from monocular video//Proceedings of the IEEE/CVF International Conference on Computer

© 中国图象图形学报版权所有

Vision. Paris, France: IEEE: 1-10.

Wang Y W and Zhao X. 2026. A review of augmented reality human-computer interaction technology from input-output dual perspectives. Journal of Image and Graphics, 31(2): 349-373 (王怡雯, 赵玺. 2026. 输入输出双视角下的增强现实人机交互技术综述. 中国图象图形学报, 31(2): 349-373) [DOI: 10.11834/jig.250197]

Zhao W, Zhang Y, Liu H and Wang Z. 2024. Uni-ControlNet: all-in-one control for text-to-image diffusion models//Advances in Neural Information Processing Systems. Vancouver, Canada: NeurIPS.

Zhao Y, Liu H and Wang Z. 2024. Diffuman4D: 4D consistent human view synthesis from sparse-view videos with spatio-temporal diffusion models [EB/OL]. [2024-05-15]. https://arxiv.org/pdf/2507.13344.pdf

Zhu J Y, Park T, Isola P, et al. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2223-2232 [DOI: 10.1109/ICCV.2017.244]

Yang Y X, Deng Y Q, Gu H J, et al. 2026. Motion-centric siamese network for three-dimensional object tracking. Journal of Image and Graphics, 31(2): 512-524 (杨宇翔, 邓颖琦, 顾鸿杰, 等. 2026. 以运动为中心的孪生网络三维目标跟踪. 中国图象图形学报, 31(2): 512-524) [DOI: 10.11834/jig.250112]

作者简介

杨航,男,硕士研究生,主要研究方向为计算机视觉,动画生成。 E-mail:3046803985@qq.com

柳娜,女,博士研究生,主要研究方向为人脸生物信息识别、多模态情感计算。 E-mail:na.liu@ujs.edu.cn

孟雷,男,博士研究生,主要研究方向为复杂多媒体大数据分析的机器学习。 E-mail:lmeng@sdu.edu.cn

毛启容,女,博士研究生,主要研究方向为多媒体智能信息处理、计算机视觉。 E-mail:mao_qr@ujs.edu.cn

李曼祎,女,博士研究生,主要研究方向为计算机图形学、三维视觉、人工智能。 E-mail:manyili@sdu.edu.cn

汪铖杰,男,博士研究生,主要研究方向为计算机视觉,机器学习。 E-mail:jasoncjwang@tencent.com

王鹏杰,男,博士研究生,主要研究方向为计算机视觉,动画生成。E-mail:pengjiewang@qq.com

© 中国图象图形学报版权所有