

数字人技术建模方法与驱动策略综述

李振东^{1,2}, 朱振鑫¹, 赵世华¹, 武艺强^{1,2}, 刘昊^{1,2}

(1. 宁夏大学信息工程学院, 银川 750021; 2. 宁夏东数西算人工智能与信息安全重点实验室, 银川 750021)

摘要: 数字人作为智能交互的核心载体, 其技术的创新突破对人工智能与数字经济的深度融合具有关键意义。本文聚焦于计算机视觉领域, 围绕数字人技术的两大核心主题建模与驱动展开讨论, 并系统整理相关数据集的特性及评估方法。在建模方面, 研究从传统几何建模技术展开, 涵盖基于网格优化、点云处理的精细化三维重建方法, 进而探讨深度学习驱动的生成式建模新范式。在驱动方面, 重点分析基于视频序列的人体姿态估计与表情迁移技术, 以及结合音频特征的语音驱动口型同步生成算法。数据集的规模与多样性对数字人形象生成至关重要, 评估方法的完善程度则能更客观地衡量生成效果。本文对数字人建模与驱动领域的代表性工作进行系统归类和总结, 分析现有方法的优势与不足, 并结合当前技术发展趋势, 展望未来可能的研究方向。

关键词: 计算机软件与理论; 数字人; 建模方法; 驱动策略

中图分类号: TP391.41 **文献标志码:** A **文章编号:** 1671-5497(2026)02-0289-24

DOI: 10.13229/j.cnki.jdxbgxb.20250382

A review of digital human technology: modeling methods and driving strategies

LI Zhen-dong^{1,2}, ZHU Zhen-xin¹, ZHAO Shi-hua¹, WU Yi-qiang^{1,2}, LIU Hao^{1,2}

(1. School of Information Engineering, Ningxia University, Yinchuan 750021, China; 2. Ningxia Key Laboratory of Artificial Intelligence and Information Security for Channeling Computing Resources from the East to the West, Yinchuan 750021, China)

Abstract: As the core carrier of intelligent interaction, innovative breakthroughs in digital human technology are of crucial significance to the in-depth integration of artificial intelligence and the digital economy. Focusing on the field of computer vision, this paper discusses the two core themes of digital human technology—modeling and driving—and systematically sorts out the characteristics of relevant datasets and evaluation methods. In terms of modeling, the research starts with traditional geometric modeling techniques, covering refined three-dimensional reconstruction methods based on mesh optimization and point cloud processing, and further explores new paradigms of generative modeling driven by deep learning. In terms of driving, it focuses on analyzing human pose estimation and facial expression transfer techniques based on video sequences, as well as speech-driven lip-sync generation algorithms

收稿日期: 2025-04-29.

基金项目: 装备智能运用教育部重点实验室开放基金项目(AAIE-2023-0403); 宁夏回族自治区中央引导地方科技发展专项项目(2024FRD05050).

作者简介: 李振东(1990-), 男, 副教授, 博士. 研究方向: 计算机视觉, 数字人. E-mail: lizhendong@nxu.edu.cn

combined with audio features. The scale and diversity of datasets are vital to the generation of digital human appearances, while the sophistication of evaluation methods allows for a more objective measurement of generation performance. This paper systematically categorizes and summarizes representative works in the field of digital human modeling and driving, analyzes the advantages and limitations of existing methods, and prospects potential future research directions in combination with current technological development trends.

Key words: computer software and theory; digital human; modeling methods; driving strategies

0 引言

在数字化转型浪潮下,人工智能技术正重构人类社会运行范式。数字人技术通过整合计算机图形学、深度学习,实现了从静态建模向数据驱动的范式跃迁。本文聚焦数字人技术的建模范式演进与协同驱动机制,系统探讨其关键技术突破与未来发展方向。

数字人技术的演进历程可划分为4个关键发展阶段,其技术路径与应用形态的变迁反映了计算机图形学、人工智能及虚拟现实技术的交叉融合进程。数字人技术的突破标志着相关领域技术的协同发展和应用场景的拓展。

20世纪70至90年代为该技术的萌芽期,计算机图形学(Computer graphic image, CGI)技术的突破为数字人原型构建提供了基础框架。21世纪初期属于技术探索期,二维以及三维建模技术的迭代显著提升了数字人形态的写实性。随后进入技术融合初级阶段,由于深度学习算法的突破性进展重构了数字人的技术范式,正经历从传统三维建模到数据驱动范式的结构性转变。基于神经辐射场(Nerve radiation fields, NeRF)^[1]、生成对抗网络(Generative adversarial networks, GANs)^[2]与扩散模型(Diffusion model, DM)^[3]等深度学习架构,通过参数化特征提取与概率分布建模,显著提升了人体结构重建精度以及丰富的动态细节,为数字人的逼真呈现提供了坚实的技术支撑。2020年至今进入元宇宙驱动阶段,三维重建技术的革新推动数字人向超写实维度演进。数字人作为元宇宙的核心交互载体,在跨空间协作场景中展现出独特优势。当前研究焦点已转向多模态大语言模型与数字人的深度集成,旨在突破情感认知与情境推理的技术瓶颈。

数字人系统构建的核心在于建模与驱动两大关键技术。建模技术从传统人工建模向数据驱动

建模演进,包括手工建模、参数化建模、三维扫描技术,再到GANs、NeRF、扩散模型等深度学习方法的革新。驱动技术主要分为音频驱动与视频驱动:音频驱动通过解析语音信号生成同步唇动,虽具备实时交互优势,但受模型表达能力限制,易出现微表情缺失与动作僵硬;视频驱动基于NeRF或GANs实现动作迁移,能捕捉细腻表情纹理,却因依赖高质量数据而限制跨身份泛化性能。两者的融合推动了数字人系统的规模化应用与高质量发展。

本文首先聚焦数字人角色的建模与生成,梳理传统几何建模与深度学习建模的核心算法流程,剖析不同技术路径在建模精度、效率及真实感上的优势与瓶颈,继而深入动画驱动技术,按视频驱动与音频驱动两大类,阐释如何实现动作迁移、口型同步生成的技术原理;其次探讨基准测试体系,从视频驱动、音频驱动及多模态数据集的特性分析出发,构建包含图像质量、分布相似性、运动精度的量化评估指标体系,为技术优化提供客观标准;再次结合影视制作、广电媒体、教育领域等实践案例,展现数字人技术在多场景的落地成效;最后总结现有方法在数据依赖、跨模态一致性等方面的局限性,结合前沿技术探究数字人技术未来可能的研究方向。

1 数字人角色的建模与生成

数字人虚拟形象作为人机交互系统的核心载体,其构建技术已形成两大主流技术体系:基于计算机图形学的传统三维建模方法与数据驱动的深度生成式建模范式。这两种技术体系的核心差异在于建模范式的转变,从依赖人工操作的静态建模到基于数据学习的动态生成,具体对比如图1所示。

传统方法依托专业建模软件,通过人工建模流程实现参数化构建,该方法虽能实现高精度建



图 1 静态建模与数据驱动

Fig. 1 Static modeling and data-driven

模,但存在建模周期长、资源消耗大、表情动态生成受限等固有局限;而深度学习驱动的生成式建模体系通过引入 GANs、DM 及变分自编码器^[4](Variational autoencoder, VAE)等前沿算法,构建了端到端的面部特征学习框架。该技术路线通过大规模多模态数据集训练,可自动解耦身份特征、表情参数与光照条件等潜在变量,在面部微表情合成、实时动态光影渲染等维度展现出显著优势。这一演进过程与数字人技术整体发展阶段高度契合,其从萌芽到成熟的完整历程如图 2 所示。



图 2 数字人发展历程

Fig. 2 Development process of digital humans

1.1 传统几何建模方法

在数字人体建模领域,几何建模的核心挑战在于构建兼具高精度解剖学特征与高效可编辑性的数学表征体系,以支撑后续的物理仿真与动态交互需求。当前技术路线分为两类:直接建模法通过底层几何操作实现细节刻画,但其人工密集型流程导致建模效率与几何复杂度间的显著权衡;参数化建模法则依托统计形变模型,将人体建模为低维形状参数以实现快速建模,但面临局部特征保真度不足的瓶颈。

1.1.1 直接建模方法

直接建模方法通过几何重构实现人体三维建模,其核心优势在于模型的高度可编辑性。尽管 3ds Max、Maya 等主流软件可实现虚拟形象构建,但存在专业门槛高、制作周期长的局限。此类方法生成的模型在生物力学特征还原度方面存在表

现瓶颈。

三维人体扫描是使用扫描仪对保持特定姿态的人体进行扫描,并且使用三维建模软件进行处理转换为点云数据来显示人体模型,具有扫描迅速、重建模型准确和可编辑的优点^[5]。早期三维人体扫描设备多数采用结构光扫描或激光扫描等技术,通过对人体不同视角扫描获取数据,再使用扫描前标记在身体部位的标记点注册获得点云数据^[6],从而得到完整的人体点云数据,在此基础上进一步从点云数据中提取出人体网格模型。三维扫描人体重建的算法流程如图 3 所示。

扫描设备操作复杂以及数据采集效率低下,难以适应多样化的工作场景需求,这促使人体建模研究的重点转向三维人体模型的编辑与渲染优化。深度相机的问世推动了三维重建技术的进一步发展,不仅能实时捕捉动态数据,还可同步记录

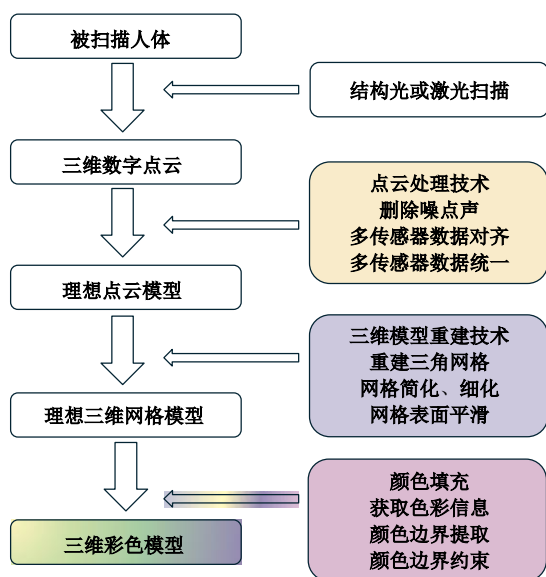


图3 三维扫描重建流程

Fig. 3 Process of 3D scanning and reconstruction

图像深度信息,为快速、低成本获取三维点云数据提供了可行路径。初始点云数据受到仪器、周围环境、被扫描目标本身的特性影响无法避免存在一些噪声。基于路径选择的点云去噪^[7]通过剔除不同水平的噪声从而提升数据质量,为后续配准提供可靠输入。

三维模型配准算法——迭代最近点 (Iterative closest point, ICP) 从出现以来,就成为了经典的配准算法。该算法利用最小二乘法对齐两组点云,找到它们之间的最优刚体变换,在不断的迭代中使它们之间的匹配误差最小化^[8]。林瑞等^[9]创新地引入了骨骼点数据作为对应点集进行点云配准,利用 Kinect 设备采集的骨骼数据,通过最小二乘法迭代计算得到变换矩阵,将不同角度采集的点云数据配准到同一坐标系下,实现了三维人体模型的重建,该方法在保证配准效果的同时,极大降低了配准时间,提高了三维重建的效率。庞浩等^[10]在前两种方法的基础上通过改进 ICP 算法和多视角数据融合实现全身三维重建,其创新在于结合摄像机标定与 ICP 算法,多视角点云融合中,体素化网格降采样通过保留关键特征点平衡精度与效率,实现全身模型的无缝拼接。网格优化是三维重建的关键环节,通过泊松表面重建^[11]将点云转化为连续网格,利用隐式函数拟合提升表面光滑度,它可修复扫描数据中的空洞与重叠区域,使重建模型满足动画驱动的精细形变需求。如图 4 所示,从左往右依次为人体模型在建模后的正视图、后视图以及侧视图。



图4 人体模型处理效果图

Fig. 4 Visualization of human model processing

1.1.2 参数化建模方法

参数化建模方法首先在系统中构建一个基础的通用三维模型,随后通过几何变换实现目标建模结果。该方法是基于统计学驱动的人体参数化模型,通过低维向量实现对人体形状的高效描述,并通过调整关键特征值实现对人体外部几何特征的精确表达。

Angelov 等^[12]提出了早期参数化人体模型 SCAPE,该方法定义两个低维参数空间人体体形空间与人体姿态空间。其中,人体体形空间是通过对相同姿态、不同体形的人体数据库进行主成分分析降维后得到的子空间,体型参数则表示为该子空间中各基向量的线性组合系数。随后,开源的基于顶点变形的人体参数化模型 SMPL^[13]通过人体体型参数与姿态参数共同控制人体变形。其中,体型参数与 SCAPE 模型一致,采用主成分分析提取的体型变形基的参数进行表示;姿态参数则是以人体的全局旋转以及 23 个关节的关节角旋转表示,并通过线性混合蒙皮进行人体姿态变形。SMPL 参数化人体模型如图 5 所示。

SMPL 凭借其简洁性和易用性,在学术界和工业界得到了广泛应用,并可无缝集成至主流三维建模软件及游戏引擎中,用于动画制作与生成。Bogo 等^[14]进一步提出基于卷积神经网络的人体二维姿态估计模型的方法,通过最小化合成人体三维姿态与检测得到的二维关节点的重投影误差优化 SMPL 参数,并引入人体穿透约束以降低从二维提升到三维提升过程中的歧义性。然而,该方法未对体型参数进行约束,且容易陷入局部最优解,从而导致重建失败。Lassner 等^[15]加入更多的人体标记点约束,提升了姿态重建的精度,并提出利用随机森林模型学习人体轮廓与 SMPL 体型参数的映射关系。尽管上述方法能够对人体整体进行建模并实现可控的躯体与四肢变形,但对

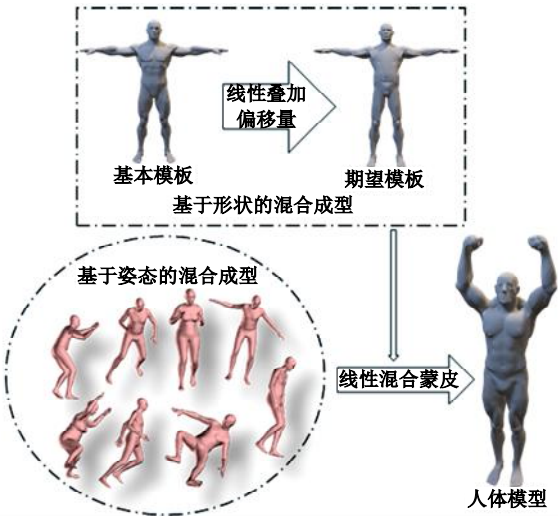


图 5 SMPL 参数化人体模型

Fig. 5 SMPL parametric human model

细节特征如微表情和皮肤褶皱的捕捉能力有限,难以达到高拟真效果。为此,研究者逐渐将重点转向人脸精细化重建领域。

Blanz 等^[16]提出的三维可变人脸模型(3D morphable model, 3DMM)为三维数字人脸生成提供了理论基础。3DMM 人脸模型如图 6 所示。

在 3DMM 出现之前,对于人脸的研究主要有两大类:一类是人脸控制和表征,另一类是人脸特征的低维压缩。3DMM 开创性地解决了计算机

视觉中单张图像估计头部三维形状以及仅通过图像恢复高质量三维表征的难题。其核心思想是人脸可以在三维空间中通过正交基的加权线性组合进行建模,即将人脸核心特征如前额长度、眼睛和鼻子大小等分解为一组基向量,通过组合这些基向量实现任意人脸的合成。Deep3Dface^[17]首次将可微分渲染技术与 3DMM 相结合,实现了基于单图的三维人脸重建。Feng 等^[18]开发的 DECA 方法通过低分辨率细节表示与分离,结合大规模户外数据训练,生成具有精细几何细节的三维人脸,并支持通过调整表情参数生成逼真的人脸几何特征。Daněček 等^[19]在 DECA 基础上引入可训练的表情分支,同时固定其他部分参数,仅利用富含情感的图像数据优化面部表情部分。这种结构使该方法在单一图像输入下生成的三维人脸模型在面部表情重建质量方面有很大进步,同时保持了身份与形状重建的高精度。

随着技术的演进,数字人建模的研究重点逐渐从整体建模转向面部精细建模,并从传统显式建模方法逐步过渡至基于深度学习的隐式建模方法。这一转变既提升了面部细节的捕捉能力,又通过数据驱动的方式实现更高效的建模流程。在此背景下,隐式建模方法通过学习数据分布的内在特征,避免了传统建模方法中对显式几何表示的依赖,为高拟真数字人生成提供了新的技术路径。

1.2 深度学习建模方法

基于深度学习的数字人建模技术可按数据维度与核心任务划分为两类:一类是二维图像生成模型,如 GAN、VAE,核心任务是从数据分布中学习并生成新的二维人脸图像,侧重提升图像的真实感与多样性;另一类是三维重建与渲染模型,如 NeRF,核心任务是通过多视角数据重建三维场景的隐式表示,并支持任意视角渲染,侧重还原三维结构的真实性与空间一致性。两者虽任务不同,但共同推动数字人建模从静态二维向动态三维演进:二维生成模型解决单视角逼真度问题,三维重建模型解决多视角一致性与空间交互性问题,为数字人在虚拟交互场景中的应用奠定基础。本节根据是否使用标签或文本引导数字人建模,分为无条件生成方法和条件生成方法。

1.2.1 无条件生成方法

无条件生成方法用于数字人脸生成时,不依

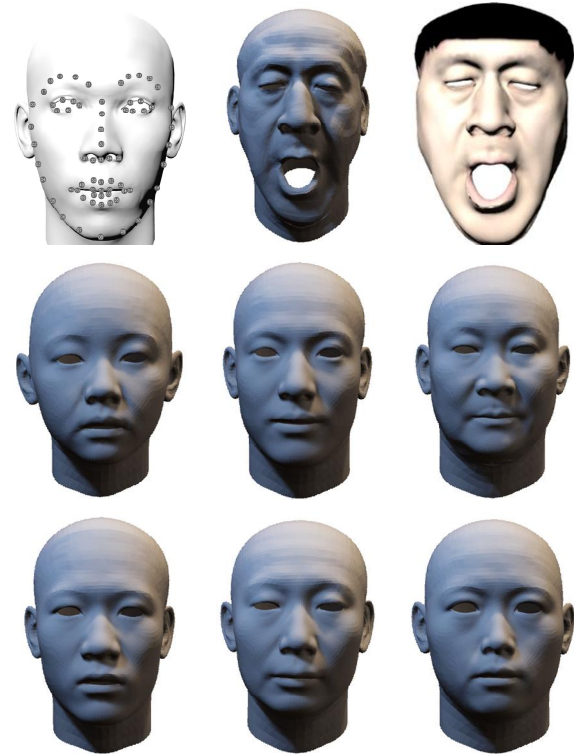


图 6 3DMM 人脸模型

Fig. 6 3DMM face model

赖于具体的输入信号,如人脸属性、标签或引导文本。此类方法通过学习人脸图像的特征分布构建生成模型,主要分为以下3类:基于噪声的生成模型、建模概率分布的自回归模型,以及基于体渲染表征场景。

基于噪声的生成模型以随机噪声为输入,将其映射到神经网络的数据空间以生成真实图像,如图7所示。近年来,此类模型在生成图像质量和效率方面取得了进展。Karras等^[20]提出一种改进的GAN训练方法,通过渐进式增加生成器和判别器的网络深度,从低分辨率开始逐步添加新层以模拟精细细节,从而提升图像生成的质量、稳定性和多样性。在此基础上,Karras等^[21]进一步改进了生成模型,包括分离调制与归一化避免水滴状伪影,采用替代架构避免细节位置偏好伪影以及增加网络容量和采用懒惰正则化等策略提升训练效率;而MoCoGAN^[22]创新性地视觉信号分解为内容和运动两个部分,通过对抗学习框架结合图像和视频判别器,实现内容与动作的分解。该框架通过随机向量序列映射到视频帧序列生成视频,其中内容部分在整个视频中保持固定,而运动部分作为随机过程实现。

IntroVAE^[23]结合了VAE和GANs的优点,

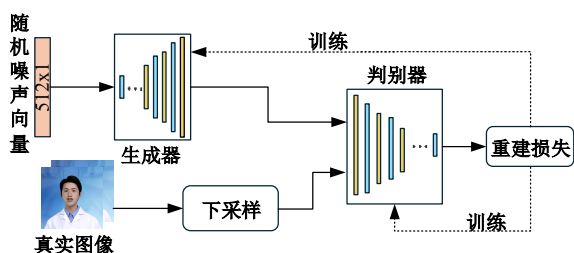


图7 基于GAN生成模型

Fig. 7 A generative model based on GAN

通过自省式的方法实现自我优化。该模型一方面要求生成器从推理模型的噪声输出中重建输入图像,另一方面鼓励推理模型区分生成样本与真实样本,而生成器则试图欺骗推理模型。这种结合方式在无须额外判别器的情况下,保持了VAE训练的稳定性良好的潜在流形特性。Jiang等^[24]提出了一种名为自适应伪增强的策略,用于在有限数据下训练GANs。该策略利用生成器自身生成的图像增强真实数据分布,从而缓解判别器的过拟合问题。基于噪声的生成方法,包括GANs和VAEs,能够灵活地建模复杂数据分布、学习潜在表示,生成高保真样本。然而,此类方法存在模式崩溃、训练不稳定以及评估难度等问题,限制了样本多样性,并需要精细调整超参数。为解决这些问题,自回归模型通过建模概率依赖关系生成连贯且多样化的输出,成为一种有效的替代方案。

自回归模型将生成任务定义为条件概率分布问题,通过逐步生成序列中的每个后续元素,最终形成完整数据样本,如图8所示。Esser等^[25]将自回归模型与多分类扩散过程相结合,通过引入双向上下文和层次化表示使模型能够从粗到细地逐步引入上下文信息。该方法克服了传统自回归模型在图像合成中的单向性偏差和全局上下文缺失问题。为解决传统自回归模型在场景信息覆盖的不足, Lee等^[26]提出了一种用于高分辨率图像生成的自回归模型框架。RQ-VAE通过残差量化方法,在不增加码本大小的情况下更精确近似图像特征图,将其表示为离散代码的堆叠并降低空间分辨率。RQ-Transformer则学习预测RQ-VAE提取的量化特征向量,通过空间变换器和深度变换器的架构高效学习代码序列,降低计算成

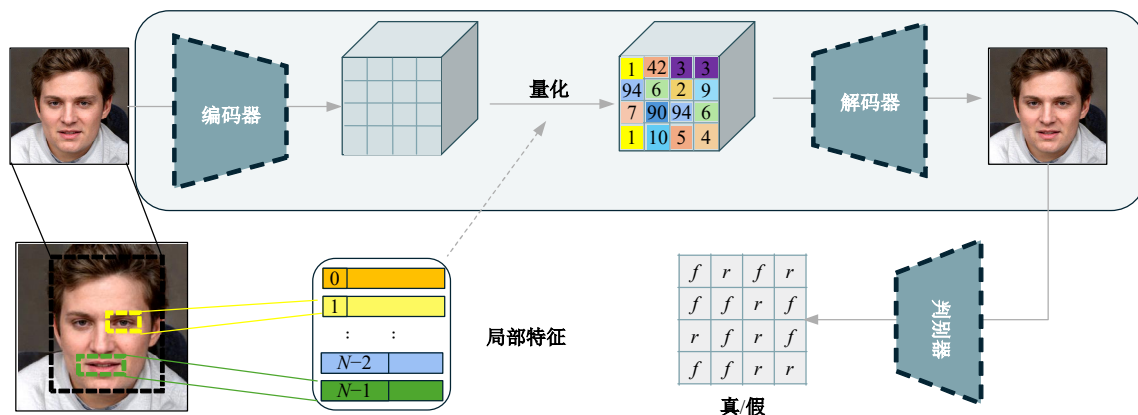


图8 自回归生成模型

Fig. 8 Autoregressive generative model

本,并通过软标记和随机采样等训练技巧解决暴露偏差问题,从而生成高质量图像。

自回归模型在二维肖像生成中展现出高逼真度,但受限于二维平面表示的固有局限,难以满足数字人对三维结构建模多视角渲染的需求,如虚拟交互中视角变化导致的面部结构失真。此时,以NeRF为代表的三维重建渲染技术成为重要补充,其通过体渲染技术对三维场景进行隐式建模,既能保留高保真细节,又能生成具有空间一致性的多视角图像,从而弥补了二维生成模型在三维结构还原上的不足。两者并非替代关系,而是数字人建模从单视角生成向三维全视角重建演进中的协同技术。

早期GAN、VAE等二维模型通过学习图像分布生成高逼真单视角人脸,解决了视觉真实感问题,但缺乏三维结构约束,在视角变化或动态交互中易出现面部结构失真。随着元宇宙等场景对虚拟交互的需求升级,数字人需具备三维空间一致性与动态可交互性。此时,NeRF等三维重建模型通过还原真实三维结构,支持任意视角渲染与动态形变,成为二维生成模型的重要延伸。

NeRF及其衍生模型生成新视角图像时不依赖文本或标签引导,仅通过多视角图像及相机位姿数据学习三维场景规律,如图9所示。

Mip-nerf^[27]提出的方法能够从任意视角渲染高质量的三维场景。在数字人形象生成中,该过程首先通过生成模型构建NeRF,学习数字人面部的辐射场表示,将三维空间位置与视角方向映射为颜色和体密度;其次通过多尺度锥形采样替代传统点采样实现视角变换,对于新视角下的每个像素,模型不再采样单一空间点,而是通过锥形区域覆盖该像素在三维空间中的可能贡献范围,结合多尺度特征融合计算辐射场值;最后通过体渲染方程积分生成新视角图像,不仅有效解决了原始NeRF在视角快速变化时的锯齿模糊问题,

还能更精准捕捉侧脸时的耳后轮廓、牙齿内侧等之前不可见的细节。这一改进使数字人多视角渲染的视觉一致性提升,尤其适用于虚拟交互中频繁视角切换的场景。GRAF^[28]率先实现从无姿态图像训练生成高分辨率图像的能力,通过引入多尺度判别器,采用连续辐射场表示,将三维空间位置与二维视角方向映射到颜色值和体密度,从而在不牺牲分辨率的情况下实现相机与场景属性的解耦。Bergman等^[29]提出名为GNARF的三维生成对抗网络框架,采用三平面特征表示和显式变形场,使生成器能够生成处于标准姿态的人体或人脸辐射场,并将其变形为目标姿态。GNARF首次实现了高质量的人体辐射场生成,并且生成的辐射场可直接编辑,解决了在人体生成和编辑方面的不足。Pix2NeRF^[30]作为首个基于NeRF的逆生成式对抗网络框架,将图像映射到潜在空间,解耦数字人面部的三维几何与纹理特征和相机视角参数,实现高质量新视角合成。其生成的高质量新视角具有两重核心作用:一是支撑数字人在虚拟交互场景中的多视角一致性展示,如元宇宙社交中,用户从正面、侧面等不同角度观察数字人时,面部结构、光影细节均保持连贯自然;二是通过新视角渲染验证三维结构的准确性,只有准确重建数字人的三维几何,才能在任意视角下生成符合物理规律的图像。

除此之外,基于NeRF重建的三维人体难以表示与姿势相关服装细节,Li等^[31]将3D高斯与可动画化的数字人联系起来,从输入中学习参数模板,然后将模板参数化为正面和反面的标准空间高斯图,学习到的模板对穿着的服装具有自适应性,可用于制作连衣裙等宽松服装的模型。SplattingAvatar^[32]在三角形网格上嵌入3D高斯的逼真数字人的混合3D表现形式,通过显式网格和隐式高斯进行拼接建模,将数字人的运动和外观分离开,其中网格代表低频运动和表面变形,而高斯则代表高频几何和细节外观,直接通过网格控制高斯的旋转和平移,这就增强了它与各种动画技术的兼容性。GoMAvatar^[33]核心方法同样使用了网格高斯表示法,但它能够无缝集成基于光栅化的图形管道,实现实时、自由视点渲染。

与传统生成方法相比,NeRF在构建精细且空间连贯的三维模型方面具有显著优势,使其在多视角图像生成任务中表现出色。但NeRF的局

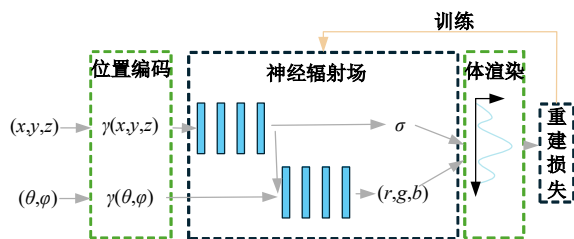


图9 基于NeRF生成模型

Fig. 9 Generative model based on NeRF

限性在于其对计算资源的高需求、对新场景的泛化性较差、可控性不足以及难以满足实时应用需求,且需要经过精细调优才能达到最佳性能。为克服上述问题,研究者提出条件生成方法,通过引入特定条件信息引导生成过程,将先验知识编码为条件约束,使模型能够根据这些条件动态调整生成策略,为在实际应用中的瓶颈提供了新路径。

1.2.2 条件生成方法

条件生成方法通过引入标签或文本指导图像合成,增强了无条件生成技术的可控性。标签是预定义的结构化属性标识,通常为离散值或有限区间,其特征维度固定且语义明确,依赖人工标注的标准化属性体系。文本是非结构化的自然语言描述,可包含离散属性、连续特征及场景上下文,其表达不受预设维度限制,能传递模糊语义及多特征关联信息,更贴近人类自然表达习惯。根据预定义属性生成图像,同时保持真实性和高分辨率,并允许对特征进行精确操作,如图10所示。

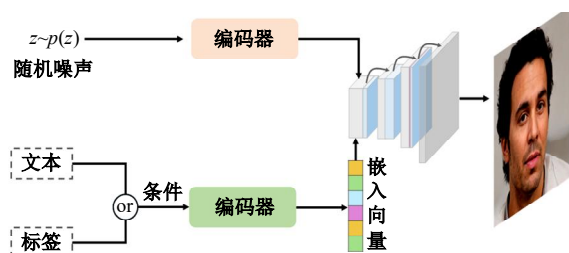


图10 文本或标签引导生成方法

Fig. 10 Generative methods guided by text or labels

标签引导方法通过特定标签的引入,能够精准调控生成图像的各类特征。这些标签涵盖从年龄、性别、面部表情等基础属性到化妆风格、配饰选择等细节特征,范围广泛。当这些标签被嵌入生成模型时,模型能够依据标签信息学习如何合成符合预设要求的图像内容。CycleGAN^[34]提出一种属性引导人脸生成方法,利用轻量级卷积网络提取身份特征向量,并通过修改对抗损失函数将条件特征向量作为生成器和判别器的输入部分,实现属性引导的人脸图像生成。该方法在低分辨率输入约束下生成的图像能够忠实反映给定属性或身份特征。Liu等^[35]则研究基于生成对抗网络的年龄引导生成模型,将面部属性嵌入生成器和判别器,同时引入小波包变换模块捕捉多尺度年龄相关纹理细节,从而提升视觉保真度。生成器采用残差连接结构,专注于建模输入输出人脸差异的老化特征,避免无关视觉内容干扰。

Men等^[36]将人物属性如姿势、头部、上衣、裤子等嵌入潜在空间中作为独立的编码,实现对属性的灵活和连续控制。标签引导的人像合成技术因其对生成图像属性的精准把控,为需要特定视觉特征的应用场景提供了极大便利,同时通过降低对大量变体数据的依赖,有效提升了数据利用效率。然而,标签引导方法也存在局限性,主要体现在对创作自由度的限制以及对准确标注训练数据的高度依赖,这种依赖可能导致模型对数据偏见的敏感性增加。为应对上述挑战,研究者进一步探索了文本引导合成技术,通过自然语言描述实现更加灵活且精准的生成控制,为图像合成领域开辟了新的研究方向。

文本引导方法通过自然语言描述扩展了标签引导技术的能力,为图像生成提供细腻且详细的控制。这些方法能够利用文本描述指定复杂的特征组合与场景,使表达更加丰富灵活。Mansimov等^[37]提出基于注意力机制的图像生成模型,用于从自然语言描述中生成图像。该方法采用双向循环神经网络将输入描述编码为语言特征表示,在每一步生成过程中结合描述信息,通过注意力机制跨模态对齐特征,逐步在画布上绘制图像。AttnGAN^[38]作为一种细粒度文本到图像生成的注意力生成对抗网络,包含注意力生成网络和深度注意力多模态相似性模型。前者通过多阶段过程生成图像,每个阶段利用注意力机制关注与当前绘制图像子区域最相关的文本词汇,从而逐步细化图像细节;后者则利用注意力机制计算生成图像与文本在全局和细粒度层面的相似性,为生成器训练提供额外的细粒度图像文本匹配损失。通过可视化的注意力层,首次展示了分层注意力能够自动关注相关单词以形成图像生成的条件,证明了其在细粒度文本到图像生成任务中的有效性。Canfes等^[39]提出一种结合文本和图像引导的3D头像生成与操控方法,利用CLIP模型实现复杂属性变化的3D操控,如年龄、性别、表情等,同时保持其他属性不变。DALL-E^[40]通过将文本和图像标记建模为单一数据流,实现了零样本文本到图像生成。利用大规模数据和模型规模提升零样本性能,并通过自回归变换器简化了传统复杂架构。Parti模型^[41]将文本到图像生成视为序列到序列建模问题,使用图像标记作为目标输出而非另一种语言的文本标记。该模型基于Trans-

former^[42]架构,包括图像标记器和编码器—解码器模型,通过大规模数据和模型尺寸的扩展提升性能。它在文本生成图像领域具有创新性,为生成高质量、内容丰富的图像提供了新的方法和思路。

Huang 等^[43]设计多模态人脸扩散模型,通过预训练的单模态模型协作,无须重新训练即可实现多模态控制,其核心在于使用一个元网络,能够自适应地预测每个预训练单模态模型的影响函数,从而在去噪过程中选择性地增强或抑制不同模态的贡献。最新的研究提出了一个基于预训练文本到图像扩散模型^[44]的一次性微调方法,用于文本到视频生成。该方法通过扩展模型到时空域并引入稀疏时空注意力机制,在推理阶段采用 DDIM^[45]反演提供结构引导,从而生成具有连续运动和时空一致性的视频。尽管文本引导图像生成技术在生成质量与可控性方面取得进展,但仍面临若干制约因素。首先,该技术对大规模预训练数据集存在高度依赖,可能导致训练数据中的潜在偏差被放大。其次,模型对未见过的文本描述的泛化能力不足,影响了其在开放场景中的适用性。

基于不同生成条件的生成图像样例如图 11 所示。当前研究已从静态肖像生成拓展至动态会说话的头像驱动机制,通过融合多种驱动方式生成更加逼真的头像视频,以满足包括虚拟助手、教育娱乐等领域的多样化应用需求。



图 11 条件生成方法生成图像样例

Fig. 11 Examples of images generated by conditional generative methods

2 数字人角色的驱动

驱动机制在会说话的人头合成领域发挥着关键作用,其核心功能在于通过精准调控合成角色

的面部动作与表情,实现逼真且富有表现力的动态效果。此类机制通过多种输入模式实现对合成过程的精确控制,从而确保最终输出的视频在视觉呈现上达到自然性与连贯性的统一。目前,该领域的主要技术路径可分为两类:一是基于视频驱动的方法;二是基于音频驱动的方法。前者通过提取源视频中的面部运动特征进行映射,后者则依赖音频信号解析实现口型与表情的同步生成,二者均以提升合成效果的逼真度与表现力为目标。

2.1 视频驱动方法

在深度学习技术兴起之前,基于视频输入生成面部动画的驱动任务面临诸多技术挑战。早期的数字人驱动技术主要依赖基于规则方法^[46]和传统图像处理技术^[47]实现面部特征的追踪与说话数字人制作。这些传统模型通常依赖于几何变换、基于模板的修改以及混合变换技术。

几何变换通过旋转、缩放和平移等操作对面部特征进行数学建模,该方法因计算效率高且实现简单而被广泛采用,能够依据预设规则或序列改变面部表情。基于模板的修改则通过调整预定义的面部模板以匹配源视频中的动作,通常采用主动外观模型^[48]和混合形状模型^[49],并逐帧修改模板适应面部动作^[50]。混合变换结合传统模型与深度学习技术,在提升数字人精度的同时降低了合成复杂度。Thies 等^[51]提出的单目视频实时面部表情捕捉与重建方法,通过密集光度一致性度量跟踪源和目标视频的面部表情,确保高精度的面部特征对齐,并通过检索和变形目标序列中的最佳匹配嘴部形状,生成逼真的嘴部内部区域,避免了直接复制原嘴部或使用通用嘴部导致的不一致结果。X2Face^[52]作为一种自监督网络架构,能够利用另一张人脸或模态控制给定人脸的姿态和表情,从而实现轻量级且复杂的视频和图像编辑。然而,传统方法由于缺乏适应性和可扩展性,存在显著局限性。每遇到新面孔或新表情时,这些方法需要大量人工设置和微调,且因依赖僵化的模板或简化的几何规则,难以捕捉人类面部表情的复杂性。相比之下,深度学习方法通过从大规模数据集中学习复杂的映射关系,实现了面部动画的自动化制作,并能够生成极为逼真且富有表现力的面部动画。这些技术主要分为两大范式:显式建模和隐式建模。显式建模通过直接参

数化面部特征实现精确控制,而隐式建模则通过学习数据驱动的映射关系实现高效合成。两类方法均显著提升了动画的质量和准确性,为视频驱动面部动画技术的发展奠定了基础。

在视频驱动的头部分成领域,显式建模范式通过对关键点、网格等明确定义的特征和参数进行直接操作,展现出高度的可解释性和易于实现的特点。基于关键点的方法通过学习源图像与驱动视频之间的关键点对应关系,生成运动序列实现特征变形^[53]。Siarohin等^[54]提出一种基于一阶运动模型的图像动画方法,通过无监督训练提取对象关键点及其局部仿射变换建模复杂运动,结合遮挡感知生成器,将运动信息与源图像外观信息融合,生成输出帧并处理遮挡问题。Zhao等^[55]提出的端到端无监督运动传递框架通过多分辨率遮挡掩码解决了源图像与驱动图像中对象姿态差异较大的问题,提升了生成图像的真实性。Da-GAN^[56]通过深度感知生成对抗网络结合深度引导的关键点检测和跨模态注意力机制,有效利用深度信息提升生成质量。Zhang等^[57]利用密集面部关键点准确预测几何感知的光流场,保持面部几何结构和人物的身份特征,提高生成视频的真实感。Doukas等^[58]结合稀疏3D关键点和注视方向估计网络实现了对生成图像中眼睛注视方向的精确控制,并通过注意力机制支持少样本学习。为了解决说话数字人身份缺失和合成大姿态问题,Zakharov等^[59]通过元学习和对抗训练,使模型能够基于少量图像快速生成高度个性化且逼真的模型。在元学习阶段,网络通过模拟少样本任务将面部标记位置转换为逼真图像,在微调阶段,模型通过少量新个体图像快速适应新任务。

基于网格的渲染方法依赖于3D头部重建模型和3D表示模型,为人脸面部动态提供更全面的理解。Yao等^[60]提出一种基于网格引导的面部重演框架,利用重建的3D网格作为指导光流学习,实现面部合成的精准重演。Wang等^[61]结合了3DMM和2D仿射变换模型分别建模面部区域、前景组件和背景,细致模拟面部图像的场景结构。ROME^[62]方法是一种基于网格的从单张照片生成逼真三维人头模型,使用卷积神经网络从单张照片中编码纹理,捕捉人物的外观和几何信息。Li等^[63]提出一种从单视图肖像进行3D头部头像重建和动画的方法,采用三平面分解方法分别生

成粗略几何、详细外观和表情信息。尽管该方法在效率上具有一定优势,但在处理复杂场景或高分辨率图像时,计算资源和时间成本仍较高。

在视频驱动的头部分成领域,基于关键点的变形方法能够有效保持主体身份,但在处理遮挡和大幅度动作时表现受限;而基于网格的渲染方法虽能提供更高的保真度,却面临模型复杂度高和计算需求大的挑战。这些局限性推动了隐式建模范式的兴起。隐式建模方法通过潜在嵌入或空间推断特征和关系,提供了更大的灵活性,并能够处理复杂的非线性交互。

潜在嵌入方法将图像表征为嵌入码,实现了对动态面部动画的高效建模。ReenactGAN^[64]方法是一种基于学习的面部重演框架,通过将源人脸映射到边界潜在空间,并使用变换器适应目标人脸的边界,最终通过目标特定的解码器生成重演的目标人脸。Wang等^[65]提出的模型不依赖于显式结构表示,如关键点或区域表示,而是在潜在空间中进行线性组合实现图像动画化。Meshry等^[66]开发一种少样本头部动画合成方法,通过将头部动画的表示分解为空间和风格两个部分,提高了合成图像的质量和身份保持性。基于神经渲染的系统^[67]能够从单张照片创建头部动画。该方法将头部图像分解为粗略图像层和高频纹理层,既提高合成图像的分辨率,又提升了神经渲染的速度。Ni等^[68]设计了一个双分支变换合成网络,用于视频运动重定向。该模型结合基于变形的变换分支和无变形的合成分支,能够在保持目标视频人物身份的同时,将驱动视频的运动模式转移到目标视频中。Bounareli等^[69]提出了一种新的神经面部重演方法,通过在预训练的生成对抗网络的潜在空间中发现控制面部姿态和表情变化的方向,以捕捉面部姿态、身份和表情的解耦方向。

NeRF在视频驱动任务中也发挥了关键作用。Gafni等^[70]提出了一种基于动态神经辐射场的方法,结合场景表示网络和低维可变形模型,通过体渲染生成图像,能够从单目输入数据中学习动态神经场景表示,无须特殊采集设备。Head-NeRF^[71]是基于NeRF的参数化人头模型,通过将神经辐射场集成到人头的参数化表示中,并结合神经渲染策略,显著加速了NeRF的渲染过程,同时保留了细粒度细节,如牙齿缝隙、皱纹和胡须

等。HiDe-NeRF方法^[72]通过多尺度通用外观模块保留人脸形状和细节,并通过轻量级表情感知变形模块解耦姿势和表情,显著提高了身份和表情的保真度。Mallya等^[73]提出了一种用于图像动画的隐式变形框架,通过跨模态注意力层在源图像和驱动图像之间找到对应关系,选择最合适的特征并进行变形。

视频驱动的方法通过保留静态图像的身份信息并复制驱动视频中的动作,推动了逼真头部动画的生成。然而,它们仍面临遮挡、身份完整性保持以及大幅度姿势变化处理等挑战。相比之下,音频驱动的方法利用音频信号使口型和面部表情与语音内容同步,进一步增强了头部动画的逼真度和自然度。

2.2 音频驱动方法

音频驱动的头部合成旨在将静态面部图像转化为与音频输入同步的动态视频。由于音频与视觉模态的内在异质性,这一多模态任务面临诸多技术挑战,包括口型同步精度、表情多样性及真实性的实现。为应对这些挑战,现有方法主要可分为两大类:统计参数模型与深度生成模型。统计参数模型通过提取音频特征与面部运动之间的统计关系,实现对口型和表情的参数化建模;而深度生成模型则利用深度学习技术从数据中学习复杂的模态映射关系,从而生成自然且逼真的动态面部动画。

2.2.1 统计参数模型

统计参数模型的演变可追溯至隐马尔可夫模型(HMM)^[74]、卷积神经网络(CNN)^[75]和循环神经网络(RNN)^[76]的发展与应用。这些方法通过捕捉音频与视觉模态之间的动态关系,逐步提升了面部动画的逼真度和自然度。

基于隐马尔可夫模型的方法能够有效捕捉视频和音序列中的动态变化。Yamamoto等^[77]提出了一种通过语音信号驱动唇部运动序列生成的方法,利用语音识别过程中的状态序列映射唇部参数,从而处理了语音上下文对唇部运动的影响。然而,由于音素与唇部运动之间缺乏一一对应关系,HMM在面部动画的精确控制方面存在局限性。为克服这一问题,Xie等^[78]提出了一种基于耦合隐马尔可夫模型的方法,通过两个隐藏链分别建模音频和视觉模态,能够更准确地捕捉两者之间的复杂时间关系,显著提升了动画的真实感

和自然度。

基于卷积神经网络的方法利用深度学习架构进一步提升面部动画质量和拟真度。Chung等^[79]开发了一种编码器—解码器卷积网络,能够根据目标人脸的静态图像和音频语音实时生成同步的说话人脸视频。该方法通过联合嵌入人脸与音频特征,精准生成与音频匹配的视频帧,并利用大量未标记视频数据进行训练,展现出良好的泛化能力。Cudeiro等^[80]提出的VOCA框架能够将任意语音信号与静态3D面部模板结合,生成逼真的3D角色动画。其核心在于分离身份与面部运动,并通过说话者标签约束实现对未见主体的泛化。DiNet^[81]通过空间变形参考图像的特征图保留高频纹理细节,进一步提升了视觉配音的真实感。

基于循环神经网络的方法通过长短期记忆网络(LSTM)等架构增强了对序列数据的建模能力。Fan等^[82]提出了一种基于深度双向长短期记忆网络的方法。该方法构建了一个包含两个LSTM层和一个前馈层的网络结构,通过最小化预测与实际视觉序列之间的误差,有效建模音频和视觉数据的长时间依赖关系,生成更自然的说话头动画。Zhou等^[83]使用语音转换网络将音频内容和说话者身份信息分离,音频内容用于控制嘴唇及周边区域的运动,而说话者信息则决定了面部表情和头部运动的细节。这种方法能够处理多种类型的肖像图像,包括艺术画作、素描和风格化漫画等,并且可以推广到训练过程中未见过的面孔和角色。与基于隐马尔可夫模型的方法相比,LSTM能够更有效地建模音频和视觉数据的长时间依赖关系,从而提供更准确的视觉特征预测。

尽管上述方法在提升逼真度和处理序列数据方面取得了显著进展,但在保持口型同步、实现多样且真实的表情以及个性化说话者风格方面仍存在不足。近年来,前沿的深度生成模型借助神经架构和大规模数据集,有效解决了传统方法的局限性。这些模型通过学习音频与视觉模态之间的复杂映射关系,进一步提升了生成视频的自然度和真实性,为该领域的研究提供了新的方向。

2.2.2 深度生成模型

近年来,前沿深度生成模型的快速发展显著推动了音频驱动头部合成技术的进步。通过结合

先进的神经网络架构与大规模数据集,这些模型在提升口型同步精度、增强表情多样性以及实现个性化风格表达方面取得了突破性进展。主要方法包括 GAN、VAE、Transformer、NeRF 以及扩散模型等,每种方法均通过独特创新提升了头部动画的逼真度与表现力。

基于生成对抗网络的音视频合成方法将听觉

特征与视觉特征无缝融合,实现了精确且拟真的面部动画生成。基于生成对抗模型的语音驱动方法的典型架构如图 12 所示。Chen 等^[84]开发了一种级联 GAN 方法,该方法将音频信号映射到面部标志,再进一步生成视频帧,这一过程避免了音频与视觉信号的不相关关联,从而提高了生成视频的自然性和一致性。

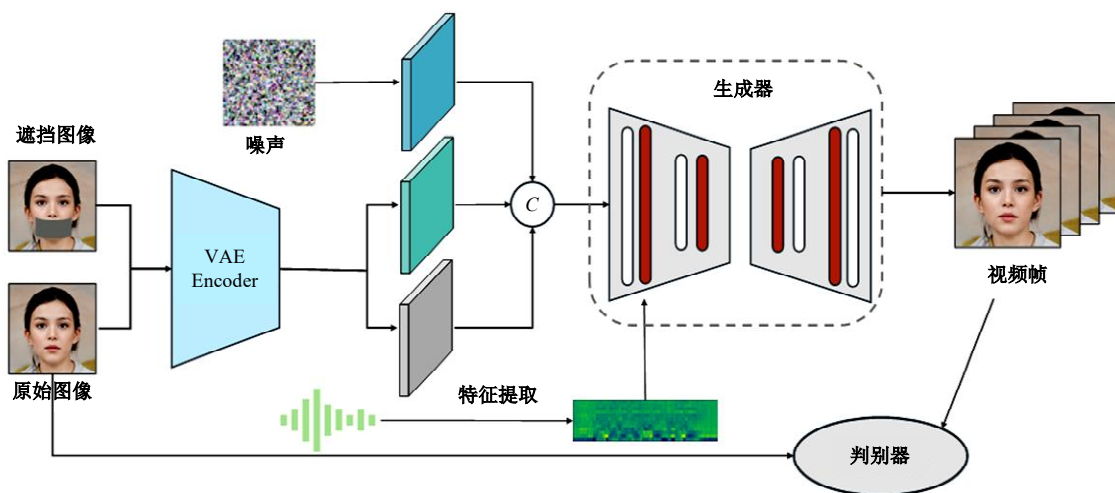


图 12 基于生成对抗模型的语音驱动方法

Fig. 12 Speech-driven method based on generative adversarial model

Wav2Lip^[85]通过引入预训练的唇形同步判别器,首次实现了在野外环境中对任意身份和语音的说话人脸视频进行高度准确的唇同步。该方法的核心在于唇同步判别器的设计,该判别器能够有效识别生成视频中唇部运动与音频信号之间的同步性,并通过对抗训练优化生成器的输出。这种对抗训练机制不仅提高了生成视频的唇同步精度,还增强了生成视频在复杂环境下的鲁棒性,使其能够在不同光照、角度和背景条件下保持高质量的输出。Tan 等^[86]的情感运动记忆网络通过解耦面部表情与嘴部运动,并利用记忆网络存储相关情感特征,确保了生成表情与嘴部运动的一致性。这种解耦与记忆机制使生成的视频不仅在口型同步方面表现出色,还能更自然地表达情感,但如何避免虚假关联并确保表情一致性仍是挑战。虚假关联可能导致生成视频中出现不自然的面部运动或情感表达,而表情一致性问题则可能影响生成视频的可信度和用户体验。

基于变分自编码器的音视频映射方法在捕捉音频和视觉模态之间的相互依赖关系方面表现出色,能够生成栩栩如生且富有表现力的面部动画。基于 VAE 的音频映射模型如图 13 所示。

Mittal 等^[87]使用变分自编码器将音序列分

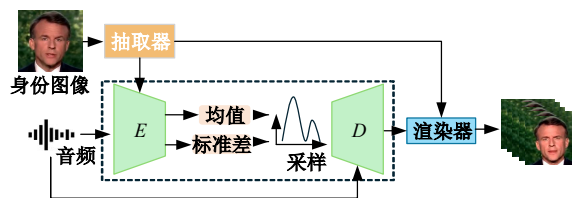


图 13 基于 VAE 的音频映射模型

Fig. 13 VAE-based audio mapping model

解为语音内容、情感语调和和其他因素的表示,从而提高头部生成对音频变化如背景噪音和情感的鲁棒性。Liu 等^[88]提出了一种名为 FONT 的流引导单次说话头生成模型,该方法采用概率建模头部姿态预测中的不确定性,利用无监督关键点捕捉面部结构信息,通过流引导生成器实现自然的头部运动和逼真的图像生成。SadTalker^[89]提出的是一种基于音频驱动的单图像说话人脸动画生成系统,该系统的核心创新在于将 3DMM 的运动系数作为中间表示,通过学习音频与不同运动系数之间的联系,生成头部姿态和表情系数,从而驱动静态图像生成自然的说话人脸视频。然而,这些方法如果数据不足或存在偏差会造成对高质量训练数据的依赖,可能导致过拟合或泛化能力差。

基于 Transformer 的方法通过解决长距离依

赖问题,改进了序列到序列任务的性能。Fan 等^[90]提出的自回归模型设计了偏置注意力机制。为了更好地对齐音频和面部运动模态,设计了两种偏置注意力机制:一种是有偏的跨模态多头注意力机制,用于有效对齐音频和运动模态;另一种是有偏的因果多头自注意力机制,结合周期性位置编码策略,用于处理更长的音序列,如图 14

所示。Bernardo 等^[91]利用基于 Transformer 模型从语音输入中提取丰富的上下文信息并生成面部关键点,然后在第二阶段采用基于 GAN 的模型将这些面部关键点转化为逼真的视频帧。这种两阶段的安排将由语音驱动的动态形状变化建模与由形状驱动的外观变化建模分离开来,部分解决了泛化问题。

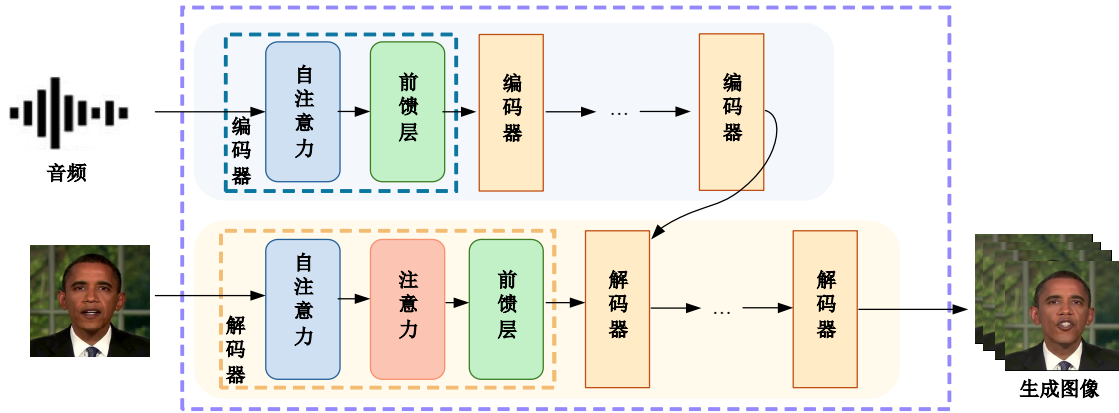


图 14 基于 Transformer 的生成模型

Fig. 14 Transformer-based generative model

隐式辐射场的建模方法 AD-NeRF^[92]直接将音频信号映射到动态神经辐射场,无须依赖二维标签等中间表示,从而避免了信息损失。Yao 等^[93]将嘴巴运动和眼睛眨动等面部属性解耦为独立的表示,使生成的说话头图像既能够保持唇形与音频的同步,又具有自然的面部运动。基于神经辐射场的生成模型如图 15 所示。

ER-NeRF^[94]方法在于其高效的三平面哈希表示减少了哈希冲突,提高了收敛速度和动态渲染质量,显式的区域注意模块能够捕捉音频条件与空间区域的相关性,提升唇音同步准确性和图像质量以及自适应姿态编码通过将头部姿态变换映射到空间坐标,解决了头部与躯干分离的问题,提高了躯干部分的渲染质量和姿态准确性。Ye 等^[95]开发了能够生成自然的、对应各种音频的高保真视频肖像的方法,通过可学习音频到 3D 面部标志的映射的变分自编码器和让躯干 NeRF 能够感知头部的位置和渲染结果,从而在渲染躯干时

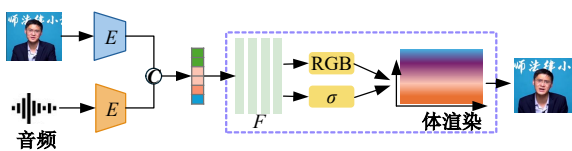


图 15 基于神经辐射场的生成模型

Fig. 15 Generative model based on neural radiance fields

能够与头部保持一致性和自然过渡。这些方法在调整音频信号、视角和背景方面具有灵活性,从而提升了生成动画的逼真度和适应性。不过,它们通常需要大量的计算资源和复杂的模型才能取得最佳效果。

基于扩散的合成方法利用扩散模型逐步向数据中添加和移除噪声,将其转换为高斯分布再还原,从而生成高质量图像。基于扩散模型的生成方法如图 16 所示。

Yu 等^[96]将扩散模型应用于音频到视觉的映射学习,利用预训练的身份和外观无关的面部运动编码器,学习唇部特征和非唇部特征,有效解决了音频与非唇部运动之间的多对一映射问题,生成多样且自然的面部运动。Shen 等^[97]通过潜在空间学习与多模态信息结合,提升了生成过程的可控性,并且不用进一步微调即可推广到不同身份。Bigioi 等^[98]开发了一种基于去噪扩散模型的

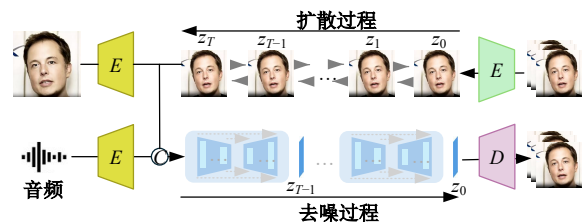


图 16 基于扩散模型的生成方法

Fig. 16 Generation method based on diffusion models

端到端语音驱动视频编辑方法,通过将音频的梅尔频谱特征作为条件,实现对视频中人物唇部和下颌运动的重新同步,无须依赖面部特征点等中间结构表示。扩散模型通过渐进式噪声生成与重建机制,逐步捕捉复杂数据分布的内在规律和人脸动态的细微变化,这种能力使扩散模型在语音驱动的视觉生成技术中成为关键驱动力。这一突破不仅推动了语音驱动视觉生成技术的发展,还为虚拟数字人、智能交互系统等应用提供了解决方案。

表1 常用数据集详细信息

Table 1 Detailed information translation of commonly used datasets

数据集	年份	时长/h	发言人	句子数	头部运动	图像尺寸	模态	视角
GRID ^[99]	2006	27.5	33	33 000	×	360×288, 720×576	视频、音频	正面视角
CREMA-D ^[100]	2014	11.1	91	12	√	960×720	视频、音频	—
TCD-TIMIT ^[101]	2015	11.1	62	6.9 000	×	1 920×1 080	视频、音频	多视角
LRW ^[102]	2016	173	1 000+	539 000	×	256×256	视频、文本	正面视角
MSP-IMMPROV ^[103]	2016	18	12	652	√	1 440×1 080	音频、视频、图像	正面视角
Voxceleb ^[104]	2017	352	1 200	153 500	√	224×224	视频	—
ObamaSet ^[105]	2017	14	1	—	√	—	视频	—
MODALITY ^[106]	2017	31	35	5 800	×	1 920×1 080	音频、视频	—
Voxceleb2 ^[107]	2018	2 400	6 100	1.1 m	√	224×224	视频	—
LRS2-BBC ^[108]	2018	224.5	500+	140 000+	√	224×224	视频、音频、文本	多视角
LRS3-TED ^[109]	2018	438	5 000+	152 000+	√	224×224	视频、音频、文本	多视角
Lombard ^[110]	2018	3.6	54	5 400	√	720×480, 864×480	视频、音频	多视角
MELD ^[111]	2018	13.7	407	13 700	√	—	视频、音频、文本	—
RAVDESS ^[112]	2018	7	24	2	√	1 920×1 080, 1 280×720	视频、音频	正面视角
LRW-1000 ^[113]	2019	57	2 000+	718 000	√	1 024×576, 1 920×1 080	视频、音频、文本	正面视角
Faceforensics++ ^[114]	2019	5.7	1 000	1 000+	√	512×512	视频、图像	正面视角
FFHQ ^[115]	2019	—	—	—	×	1 024×1 024	图像	—
MEAD ^[116]	2020	39	60	20	√	1 920×1 080	视频、音频	多视角
HDTF ^[117]	2021	15.8	362	10 000	√	1 280×720, 1 920×1 080	视频	—
TalkingHead-1KH ^[118]	2021	1 000	—	—	√	—	视频	—
CelebV-HQ ^[119]	2022	65	15 653	—	√	512×512	视频	—
Multiface ^[120]	2022	—	13	—	√	2 048×1 334, 1 024×1 024	图像、音频	多视角

视频驱动类数据集作为数字人建模的基础资源,其核心价值在于通过高质量时序视频数据实现头部运动与面部表情的精准建模。例如,Voxcele^[104]和 Voxceleb2^[107]数据集通过多视角采集策略捕获了大规模人体动作数据,为数字人建模提供了丰富的运动轨迹信息。然而,其分辨率限制导致微表情特征易被忽略,这在一定程度上限制了其在精细表情建模中的应用潜力。与之形成对比的是 TalkingHead-1KH^[118],该数据集虽以高分辨率记录了复杂头部运动轨迹,但其样本规模不足万例的特点可能引发过拟合风险。此外,

3 数据集与评估指标

3.1 数据集

在数字人技术的研究中,数据集的选择对模型的训练效果和泛化能力具有重要影响。根据数据模态、采集视角、分辨率特性及适用场景的差异性,现有数据集可系统划分为三大类别:视频驱动类、音频驱动类以及多模态数据集。表1详细列出了各类数据集的特性及其适用范围,其特性与适用场景的匹配程度成为决定模型效能的关键因素。

HDTF数据集^[117]通过工业级动捕设备实现了亚毫米级的面部细节捕捉,这种高精度的采集设备能够捕捉到面部肌肉的细微运动和皮肤的纹理变化,为数字人面部的精细建模提供了细节信息。但该数据集的单样本采集成本高昂,这在很大程度上制约了其在开源研究领域的应用广度。

在语音驱动建模领域,音频—视觉同步数据集发挥着不可替代的作用。CREMA-D数据集^[100]通过专业演员的情感演绎构建了多维情感特征空间,但其固定机位采集方式难以满足多视角合成需求。LRW数据集^[102]凭借百万级唇部运动样本

成为语音口型映射研究的基准数据集,然而其低分辨率特性与情感标注缺失限制了高级语义建模能力。MEAD数据集^[116]在技术指标上取得突破,不仅实现4K分辨率采集,还构建了32视角同步拍摄矩阵,能够更全面地记录数字人的面部和身体表现,但其有限的规模难以支撑深度神经网络的全参数优化需求。

多模态数据集的创新发展为复杂数字人系统的构建提供了全方位数据支撑。这类数据集整合了多种模态的信息,如语音、文本、视觉等,能够更全面地描述数字人的特征和行为。LRS3-TED数据集^[109]整合了语音文本对齐、面部表情编码等多维度信息,但其视觉模态分辨率不足导致微表情建模存在瓶颈。FFHQ数据集^[115]虽然以高分辨率构建了面部细节数据库,但静态图像的本质属性使其难以应用于时序动作合成场景。CelebV-HQ数据集^[119]通过融合高帧率视频与多模态语义标注,在表情肌群运动建模方面展现出独特优势,然而其样本多样性不足可能影响跨种族模型的泛化性能。

综上所述,视频驱动类数据集在运动轨迹建模方面具有先天优势,但受限于硬件采集成本,音频驱动类数据集为语音视觉映射研究奠定基础,却面临情感维度单一的发展瓶颈。多模态数据集虽提供全景数据支持,不同模态的数据在时间、空间和语义层面上通常存在不一致,如何实现这些异构数据的高效融合是一个有待解决的问题。未来研究应着力构建标准化多模态基准数据集,以解决当前数据集在规模、质量和一致性上的不足。同时,通过跨模态注意力机制等新型算法框架提升数据利用率,充分挖掘多模态数据之间的关联关系,从而推动数字人技术向高保真、强泛化的方向持续演进。

3.2 评估指标

全面评估说话头合成模型对了解其有效性至关重要。Chen等^[121]建议根据4个关键标准对这些模型进行评估:身份保持、视觉质量、唇部同步和自然动作。虽然这些方面都很重要,但模型很少能同时所有方面都表现出色,通常都是集中在某一个方面的进步。在说话头合成领域,定性和定量评估都得到了广泛应用。定性评估依赖于直接观察带有主观性,而定量指标则能提供更客观的评估。本节将介绍几种广受认可的量化指标。

结构相似性指数 (Structural similarity index measure, SSIM)^[122]是一种图像质量评估方法,其理论基础源于人类视觉系统对结构信息的高度敏感性,传统图像质量评估方法主要关注像素级误差的可见性,而忽略了图像结构信息在人类视觉感知中的核心作用。SSIM通过量化图像结构信息的退化程度,提供了一种更符合人类视觉感知的图像质量评估框架。给定原图像 x 与生成图像 y ,SSIM计算公式如下:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (1)$$

式中: μ_x 和 μ_y 分别为图像 x 和 y 的均值; σ_x 和 σ_y 分别为图像 x 和 y 的标准差; σ_{xy} 为图像 x 和 y 的协方差; C_1 和 C_2 为常数,避免分母为零的情况。

峰值信噪比 (Peak signal-to-noise ratio, PSNR)是一种广泛使用的图像质量评估指标,通过计算生成图像与原始图像之间的均方误差衡量生成图像的保真度。较高的PSNR值表示生成图像与原始图像之间的差异较小,图像质量较好。计算公式如下:

$$\begin{cases} \text{PSNR} = 10 \cdot \lg_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right) \\ \text{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (I(i, j) - K(i, j))^2 \end{cases} \quad (2)$$

式中:MAX为图像的峰值信号值;MSE为均方误差; I 为原始图像; K 为生成图像; m 和 n 分别为图像的行数和列数。

FID (Frechet inception distance)^[123]是一种用于评估生成对抗网络性能的指标,它通过比较生成图像和真实图像的特征分布衡量两者的相似性。首先,使用一个预训练的模型提取图像的特征向量。其次,将这些特征向量建模为多元高斯分布,分别计算生成数据和真实数据的均值和协方差矩阵。最后,利用弗雷歇距离公式计算这两个高斯分布之间的距离:

$$\text{FID} = \|\mu - \mu_w\|^2 + \text{Tr}(\Sigma + \Sigma_w - 2(\Sigma\Sigma_w)^{1/2}) \quad (3)$$

式中: μ 和 Σ 分别为生成数据的均值和协方差矩阵; μ_w 和 Σ_w 分别为真实数据的均值和协方差矩阵。

LPIPS (Learned perceptual image patch similarity)^[124]是一种学习得到的感知图像相似性度量方法,旨在更好地模拟人类对图像相似性的主观判断。LPIPS需要几个关键步骤,首先给定两个图像 x 和 x_0 ,使用预训练的深度卷积网络提取它们第 l 层的特征表示 y_l 和 y_l^0 ;其次进行归一化,得到特征 $\hat{y}_l$ 和 $\hat{y}_l^0$ 消除不同特征通道之间的尺度差异,计算两个归一化特征之间的加权欧几里得距离;最后将上述距离在所有空间位置上平均,并对所有卷积层的距离进行求和,得到最终的LPIPS距离,计算公式如下:

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h, w} \|w_l \odot (\hat{y}_{l, hw} - \hat{y}_{l, hw}^0)\|^2 \quad (4)$$

式中: H_l 和 W_l 分别为第 l 层特征图的高度和宽度; $\odot$ 为逐元素乘法。

LMD (Landmark distance)^[125]是一个用于评

估生成的唇部运动视频准确性的新指标。通过计算合成视频和真实视频中对应唇部特征点之间的欧几里得距离,并按时间长度和标志点数量进行归一化,衡量合成视频中嘴唇运动的精度。较低的LMD值表示生成的唇部运动与真实唇部运动更加接近,从而反映了模型在唇部运动生成任务上的性能表现。计算公式如下:

$$\text{LMD} = \frac{1}{T} \times \frac{1}{P} \sum_{t=1}^T \sum_{p=1}^P \| \text{LR}_{t,p} - \text{LF}_{t,p} \|_2 \quad (5)$$

式中: T 为视频的时间长度; P 为每幅图像上唇部特征点总数; $\text{LR}_{t,p}$ 和 $\text{LF}_{t,p}$ 分别为真实视频和合成视频在时间 t 和 p 处的唇部特征点坐标。

AKD (Average keypoint distance)用于衡量生成图像中的关键点与真实图像中对应关键点之间的平均距离。它被用来评估输入驱动图像中的运动在生成结果中保留的效果。

表2 视频驱动任务与音频驱动任务在不同数据集的定量评估比较

Table 2 Comparison of quantitative evaluations between bideo-driven tasks and audio-driven tasks on different datasets

基于视频和图像作为输入的合成方法定量对比						
数据集	方 法	评估指标				
		SSIM ↑	PSNR ↑	LPIPS ↓	FID ↓	AKD ↓
Voxceleb ^[104]	FOMM ^[54]	0.723	30.390	0.199	—	1.294
	Face-vid2vid ^[118]	0.761	30.690	0.212	—	1.620
	DAGAN ^[56]	0.804	31.220	0.185	—	1.279
	MCNet ^[126]	0.825	31.940	0.174	—	1.203
	HeadGAN ^[127]	—	21.460	0.112	36.100	—
	X2Face ^[52]	0.719	22.540	—	—	7.687
	FOMM ^[54]	0.790	23.280	0.160	33.220	2.950
TalkHead-1KH ^[118]	Face-vid2vid ^[118]	0.810	23.590	0.160	35.120	3.100
	TPSM ^[55]	—	25.530	—	32.770	0.983
	LIA ^[65]	—	24.430	—	38.890	0.932
	DAGAN ^[56]	—	23.930	—	34.350	2.405
	IWA ^[72]	—	23.320	0.150	—	3.480
基于语音和图像作为输入的合成方法定量对比						
数据集	方 法	评估指标				
		SSIM ↑	PSNR ↑	LMD ↓		
HDTF ^[117]	DiNET ^[81]	0.943	30.008	—		
	Wav2Lip ^[85]	0.908	29.288	2.890		
	MakeItTalk ^[83]	0.751	19.860	5.460		
	SadTalker ^[89]	0.770	—	4.070		
	DiNET ^[81]	0.920	29.118	—		
MEAD ^[116]	Wav2Lip ^[85]	0.899	28.539	4.050		
	EMMN ^[86]	0.660	29.380	2.780		
	SadTalker ^[89]	0.690	—	4.370		

表2中汇总了当前经典方法和SOTA方法在视频驱动任务和音频驱动任务常用数据集上的量化结果,将不同方法在同一数据集上的性能进行了比较,以提供直观的参考。

4 应用

在影视行业,数字人合成技术正引领一场视觉叙事的革新。以《阿凡达》电影制作为例^[128],FACS(Facial action coding system)重塑了传统面部动画制作模式。它采集演员各种不同表情,形成表情库,并在片场与演员的面部表情数据匹配后,通过肌肉曲线,动画师可以直接控制面部肌肉,形成逼真的数字演员。该技术还能降低后期特效成本,加速创作流程,融入会说话的头部合成技术,不仅提升了制作的视觉吸引力与叙事深度,而且为创新叙事手法打开了新的大门。

作为新闻和娱乐内容的传播者,央视网虚拟数字人小C^[129]从2021年至2024年,已经在多次重大活动和节目中成功应用。央视网的小C首次亮相于2021年,其在两会真探系列直播节目中担任起了记者的角色。小C在央视活动中的多次成功应用,展示了虚拟数字人在增强用户互动体验方面的巨大潜力。

随着技术的不断进步和应用的深化,虚拟数字人将在广电新媒体领域发挥越来越重要的作用,进而推动媒体互动形式的不断创新和用户体验的持续提升。数字人合成技术在教育领域的应用催生了虚拟教师这一新兴角色,为教学模式的创新提供了新的可能性。虚拟教师能够通过实时交互与学生展开动态沟通,并通过富有表现力的面部反应增强教学效果。基于其标准化的知识传授能力,虚拟教师能够在统一的教学框架下提供高质量的课程内容,同时根据学生的课堂反应和学习进度,动态调整教学策略,提供个性化辅导。

数字人技术的快速发展正在重塑多个行业的未来格局。从虚拟主播到智能客服,从医疗辅助到教育创新,数字人已经从概念走向实际应用,成为推动智能化转型的重要工具。未来,随着技术的进一步成熟和应用场景的拓展,数字人将在更多领域释放潜力,推动人类社会迈向智能化新时代。

5 结论

本文对数字人技术的建模方法与驱动策略常用的方法和技术进行了总结。经过数十年发展,已形成以建模与驱动的技术体系,在方法创新与场景落地中取得进展,但仍存在技术瓶颈与应用限制。具体可归纳为以下3个维度:

(1)在建模技术层面,形成了传统几何建模与深度学习建模并行的路径。传统方法中,基于网格优化、点云处理的三维重建与参数化模型通过显式几何控制实现高精度结构建模,适用于对精度要求高的影视制作等场景,但存在人工成本高、细节捕捉有限的问题。深度学习方法则通过GAN、VAE、NeRF等架构实现数据驱动的生成式建模,其中GAN能生成高逼真度二维图像,NeRF通过隐式辐射场实现多视角一致的三维渲染,扩散模型则在文本引导生成中展现灵活可控性。但深度学习方法依赖大规模高质量数据集,且复杂模型的计算成本高,实时性较低。

(2)在驱动技术层面,视频驱动与音频驱动各有侧重。视频驱动通过提取源视频运动特征实现动作迁移,显式方法如基于关键点的FOMM、基于网格的HeadNeRF能保留身份信息,但处理遮挡和大幅度姿态变化时易失真;隐式方法通过辐射场变形提升细节保真度,却受限于数据泛化性。音频驱动通过解析语音信号生成同步口型,统计参数模型实现初步时序建模,深度生成模型Wav2Lip的唇同步判别器显著提升口型一致性,但仍存在微表情缺失、情感与语音协同不足的问题。

(3)在数据集与评估层面,现有资源呈现模态分化特征。视频驱动数据集Voxceleb侧重运动轨迹但微表情信息不足,音频驱动数据集LRW、HDTF聚焦唇动同步但情感标注有限,多模态数据集LRS3-TED、CelebV-HQ尝试整合多维度信息,却面临模态对齐如音频与视觉时空同步的挑战。评估指标虽通过SSIM、FID、LMD等量化图像质量与运动精度,但缺乏对情感一致性、跨场景泛化性的统一标准,难以全面衡量数字人的自然度。

6 展 望

数字人技术的突破需围绕高精度建模、自然化驱动、高效化落地的核心目标,结合多学科交叉与技术融合,重点在以下方向推进:

(1)细粒度动态建模与情感认知增强针对动态细节捕捉不足,如皮肤褶皱、微表情缺失的问题,需从几何、纹理、运动多维度突破。在建模层面,结合3D高斯溅射与神经辐射场,实现服装细节的高精度重建;在驱动层面,融合心理学情感模型与深度生成模型,通过扩散模型的渐进式去噪实现语音内容、情感强度、面部动作的协同生成。此外,需开发细粒度评估指标,如微表情变化速率,补充现有LMD、FID等指标的局限。

(2)轻量化建模与实时交互优化为突破计算效率瓶颈,如NeRF渲染慢、大模型部署难,需推动模型架构与渲染技术创新。一方面,通过知识蒸馏将模型压缩为轻量化版本;另一方面,优化渲染策略,如GoMAvatar的光栅化集成,提升NeRF等隐式模型的实时性。同时,探索边缘计算与云边协同模式,将轻量级驱动模型部署于移动端,如手机虚拟助手,复杂建模任务交由云端处理,平衡精度与实时性需求。

(3)跨场景泛化与行业适配深化针对当前技术在特定场景泛化不足的问题,需结合行业需求定制解决方案。在广电领域,优化虚拟主播的实时唇动同步与多风格切换。在教育领域,增强虚拟教师的情感反馈能力,如结合学生表情调整教学节奏。在医疗领域,开发基于高精度动捕数据的康复训练数字人,实现动作规范性实时矫正。通过场景化数据增强与模型微调,推动数字人从形似向场景适配的落地。

参考文献:

- [1] Mildenhall B, Srinivasan P P. Nerf: representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99-106.
- [2] Goodfellow I, Pouget A J. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.
- [3] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.
- [4] Kingma D P, Welling M. Auto-encoding variational bayes[C] // Proceedings of the International Conference on Learning Representations, Banff, Canada, 2014: No. 13126114.
- [5] 徐继红, 张文斌. 非接触式三维人体扫描技术的综述[J]. 扬州职业大学学报, 2006, 2006(3): 49-53.
Xu Ji-hong, Zhang Wen-bin. A review of non-contact 3D human body scanning technology[J]. Journal of Yangzhou Polytechnic College, 2006, 2006(3): 49-53.
- [6] Stefanie W, Peng C X, Chang S. Human shape correspondence with automatically predicted landmarks[J]. Machine Vision and Applications, 2012, 23(4): 821-830.
- [7] Wei Z, Chen H, Nan L, et al. PathNet: path-selective point cloud denoising[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(6): 4426-4442.
- [8] Besl P J, McKay H D. A method for registration of 3-D shapes[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1992, 14(2): 239-256.
- [9] 林瑞, 王俊英, 孙水发, 等. 基于Kinect的骨骼配准的人体三维重建[J]. 信息通信, 2016, 2016(12): 206-209.
Lin Rui, Wang Jun-ying, Sun Shui-fa, et al. 3D human body reconstruction based on Kinect skeletal registration[J]. Changjiang Information & Communications, 2016, 2016(12): 206-209.
- [10] 庞浩, 李吉平. ICP算法的改进及两台Kinect对人体的重建[J]. 大连工业大学学报, 2017, 36(6): 459-463.
Pang Hao, Li Ji-ping. Improvement of ICP algorithm and human body reconstruction using two Kinects[J]. Journal of Dalian Polytechnic University, 2017, 36(6): 459-463.
- [11] Kazhdan M, Hoppe H. Screened poisson surface reconstruction[J]. ACM Transactions on Graphics, 2013, 32(3): 1-13.
- [12] Anguelov D, Srinivasan P, Koller D, et al. SCAPE: shape completion and animation of people[J]. ACM Transactions on Graphics, 2005, 24(3): 408-416.
- [13] Loper M, Mahmood N, Romero J, et al. SMPL: a skinned multi-person linear model[J]. ACM Transactions on Graphics, 2015, 34(6): 1-16.
- [14] Bogo F, Kanazawa A, Lassner C, et al. Keep it SMPL: automatic estimation of 3D human pose and

- shape from a single image[C]//Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, Netherlands, 2016: 561–578.
- [15] Lassner C, Romero J, Kiefel M, et al. Unite the people: closing the loop between 3D and 2D human representations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 6050–6059.
- [16] Blanz V, Vetter T. A morphable model for the synthesis of 3D faces[C]//Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, New York, USA, 1999: 187–194.
- [17] Deng Y, Yang J, Xu S, et al. Accurate 3D face reconstruction with weakly-supervised learning: from single image to image set[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, USA, 2019: No. 201900038.
- [18] Feng Y, Feng H, Black M J, et al. Learning an animatable detailed 3D face model from in-the-wild images[J]. *ACM Transactions on Graphics*, 2021, 40(4): 1–13.
- [19] Daněček R, Black M J, Bolkart T. Emoca: emotion driven monocular face capture and animation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 20311–20322.
- [20] Karras T, Aila T, Laine S, et al. Progressive growing of GANs for improved quality, stability, and variation[C]//International Conference on Learning Representations, Vancouver, Canada, 2018: No. 171010196.
- [21] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of stylegan[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 2020: 8110–8119.
- [22] Tulyakov S, Liu M Y, Yang X D, et al. MoCo-GAN: Decomposing motion and content for video generation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 1526–1535.
- [23] Huang H, He R, Sun Z, et al. Introvae: introspective variational autoencoders for photographic image synthesis[J]. *Advances in Neural Information Processing Systems*, 2018, 31: 52–63.
- [24] Jiang L M, Dai B, Wu W E, et al. Deceive d: adaptive pseudo augmentation for gan training with limited data[J]. *Advances in Neural Information Processing Systems*, 2021, 34: 21655–21667.
- [25] Esser P, Rombach R, Blattmann A, et al. Imagebart: bidirectional context with multinomial diffusion for autoregressive image synthesis[J]. *Advances in Neural Information Processing Systems*, 2021, 34: 3518–3532.
- [26] Lee D, Kim C. Autoregressive image generation using residual quantization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 11523–11532.
- [27] Barron J T, Mildenhall B, Verbin D, et al. Mip-nerf 360: unbounded anti-aliased neural radiance fields[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 5470–5479.
- [28] Schwarz K, Liao Y, Niemeyer M, et al. Graf: generative radiance fields for 3D-aware image synthesis[J]. *Advances in Neural Information Processing Systems*, 2022, 33: 20154–20166.
- [29] Bergman A W, Kellnhofer P, Wang Y F, et al. Generative neural articulated radiance fields[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 19900–19916.
- [30] Cai S, Obukhov A, Dai D, et al. Pix2nerf: unsupervised conditional p-gan for single image to neural radiance fields translation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 3981–3990.
- [31] Li Z, Zheng Z, Wang L, et al. Animatable gaussians: learning pose-dependent gaussian maps for high-fidelity human avatar modeling[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024: 19711–19722.
- [32] Shao Z, Wang Z, Li Z, et al. Splattingavatar: realistic real-time human avatars with mesh-embedded gaussian splatting[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024: 1606–1616.
- [33] Wen J, Zhao X, Ren Z, et al. Gomavatar: efficient animatable human modeling from monocular video using gaussians-on-mesh[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024: 2059–2069.

- [34] Lu Y, Tai Y W, Tang C K. Attribute-guided face generation using conditional cyclegan[C] //Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018: 282-297.
- [35] Liu Y, Li Q, Sun Z. Attribute-aware face aging with wavelet-based generative adversarial networks[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 11877-11886.
- [36] Men Y, Mao Y, Jiang Y, et al. Controllable person image synthesis with attribute-decomposed gan[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 5084-5093.
- [37] Mansimov E, Parisotto E, Ba J L, et al. Generating images from captions with attention[C] //International Conference on Learning Representations, San Diego, USA, 2015: 1-12.
- [38] Xu T, Zhang P, Huang Q, et al. Attngan: fine-grained text to image generation with attentional generative adversarial networks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 1316-1324.
- [39] Canfes Z, Atasoy M F, Dirik A, et al. Text and image guided 3D avatar generation and manipulation [C] //Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Waikoloa, USA, 2023: 4421-4431.
- [40] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation[C] //International Conference on Machine Learning, Virtual, 2021: 8821-8831.
- [41] Yu J, Xu Y, Koh J Y, et al. Scaling autoregressive models for content-rich text-to-image generation[J]. Transactions on Machine Learning Research, 2022, 6: No. 220610789.
- [42] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 6: 1-11.
- [43] Huang Z, Chan K C, Jiang Y, et al. Collaborative diffusion for multi-modal face generation and editing [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2023: 6080 - 6090.
- [44] Wu J Z, Ge Y X, Wang X T, et al. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation[C] //Proceedings of the IEEE International Conference on Computer Vision, Paris, France, 2023: 7623-7633.
- [45] Song J, Meng C, Ermon S. Denoising diffusion implicit models[C] //International Conference on Learning Representations, Virtual, 2021: No. 201002502.
- [46] Lee Y, Terzopoulos D, Waters K. Realistic modeling for facial animation[C] //Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, USA, 1995: 55-62.
- [47] Chuang E, Bregler C. Performance driven facial animation using blendshape interpolation[J]. Computer Science Technical Report, 2002, 2(2): 1-3.
- [48] Cootes T F, Edwards G J, Taylor C J. Active appearance models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(6): 681-685.
- [49] Pighin F, Hecker J, Lischinski D, et al. Synthesizing realistic facial expressions from photographs[C] // Computer Graphics Proceedings, Annual Conference Series. Association for Computing Machinery Siggraph, Orlando, USA, 1998: 75-84.
- [50] Bouaziz S, Wang Y, Pauly M. Online modeling for realtime facial animation[J]. ACM Transactions on Graphics, 2013, 32(4): 1-10.
- [51] Thies J, Zollhofer M, Stamminger M, et al. Face2face: real-time face capture and reenactment of rgb videos[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 2387-2395.
- [52] Wiles O, Koepke A S, Zisserman A. X2face: a network for controlling face generation using images, audio, and pose codes[C] //Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018: 670-686.
- [53] Siarohin A, Lathuilière S, Tulyakov S, et al. Animating arbitrary objects via deep motion transfer[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 2377 - 2386.
- [54] Siarohin A, Lathuilière S, Tulyakov S, et al. First order motion model for image animation[C] //Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, Canada, 2019: 7137-7147.

- [55] Zhao J, Zhang H. Thin-plate spline motion model for image animation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 3657-3666.
- [56] Hong F T, Zhang L, Shen L, et al. Depth-aware generative adversarial network for talking head video generation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 3397 - 3406.
- [57] Zhang B, Qi C, Zhang P, et al. Metaportrait: identity-preserving talking head generation with fast personalized adaptation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 22096-22105.
- [58] 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.
- [59] Zakharov E, Shysheya A, Burkov E, et al. Fewshot adversarial learning of realistic neural talking head models[C]//Proceedings of the IEEE International Conference on Computer Vision, Seoul, South Korea, 2019: 9459-9468.
- [60] 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, Seattle, USA, 2020: 1773-1781.
- [61] Wang Q, Zhang L, Li B. Safa: structure aware face animation[C]//2021 International Conference on 3D Vision, London, UK, 2021: 679-688.
- [62] Khakhulin T, Sklyarova V, Lempitsky V, et al. Realistic one-shot mesh-based head avatars[C]//European Conference on Computer Vision, Tel Aviv, Israel, 2022: 345-362.
- [63] Li X, De M S, Liu S, et al. Generalizable one-shot 3D neural head avatar[J]. *Advances in Neural Information Processing Systems*, 2023, 36: 47239-47250.
- [64] Wu W, Zhang Y, Li C, et al. Reenactgan: learning to reenact faces via boundary transfer[C]//Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018: 603-619.
- [65] Wang Y, Yang D, Bremond F, et al. Latent image animator: learning to animate images via latent space navigation[C]//ICLR 2022-The International Conference on Learning Representations, Virtual, 2022: No. 220309043.
- [66] Meshry M, Suri S, Davis L S, et al. Learned spatial representations for few-shot talking-head synthesis[C]//Proceedings of the IEEE International Conference on Computer Vision, Virtual, 2021: 13829-13838.
- [67] Zakharov E, Ivakhnenko A, Shysheya A, et al. Fast bi-layer neural synthesis of one-shot realistic head avatars[C]//Computer Vision - ECCV 2020: 16th European Conference, Virtual, 2020: 524-540.
- [68] Ni H, Liu Y, Huang S X, et al. Cross-identity video motion retargeting with joint transformation and synthesis[C]//Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Waikoloa, USA, 2023: 412-422.
- [69] Bounareli S, Argyriou V, Tzimiropoulos G. Finding directions in gan's latent space for neural face reenactment[J]. *Arxiv Preprint*, 2022, 2: 202200046.
- [70] Gafni G, Thies J, Zollhöfer M, et al. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 2021: 8649-8658.
- [71] Hong Y, Peng B, Xiao H Y, et al. Headnerf: a real-time nerf-based parametric head model[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 20374-20384.
- [72] Li W, Zhang L, Wang D, et al. One-shot high-fidelity talking-head synthesis with deformable neural radiance field[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 17969-17978.
- [73] Mallya A, Wang T C, Liu M Y. Implicit warping for animation with image sets[J]. *Advances in Neural Information Processing Systems*, New Orleans, USA, 2022, 35: 22438-22450.
- [74] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. *Proceedings of the IEEE*, 1989, 77(2): 257-286.
- [75] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. *Proceedings of the IEEE*, 1998, 86(11): 2278-2324.
- [76] Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization[J]. *Arxiv Preprint*, 2014, 9:

- 20142329.
- [77] Yamamoto E, Nakamura S, Shikano K. Lip movement synthesis from speech based on hidden Markov models[J]. *Speech Communication*, 1998, 26(2): 105-115.
- [78] Xie L, Liu Z Q. A coupled HMM approach to video-realistic speech animation[J]. *Pattern Recognition*, 2007, 40(8): 2325-2340.
- [79] Chung J S, Jamaludin A, Zisserman A. You said that?[C]//British Machine Vision Association and Society for Pattern Recognition, London, UK, 2017: 170502966.
- [80] Cudeiro D, Bolkart T, Laidlaw C, et al. Capture, learning, and synthesis of 3D speaking styles[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 10101-10111.
- [81] Zhang Z, Hu Z, Deng W, et al. Dinet: deformation inpainting network for realistic face visually dubbing on high resolution video[C] //Proceedings of the AAAI Conference on Artificial Intelligence, Washington, USA, 2023, 37(3): 3543-3551.
- [82] Fan B, Wang L, Soong F K, et al. Photo-real talking head with deep bidirectional LSTM[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, Brisbane, Australia, 2015: 4884-4888.
- [83] Zhou Y, Han X, Shechtman E, et al. Makeltalk: speaker-aware talking-head animation[J]. *ACM Transactions On Graphics*, 2020, 39(6): 1-15.
- [84] Chen L, Maddox R K, Duan Z, et al. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 7832-7841.
- [85] Prajwal K R, Mukhopadhyay R, Namboodiri 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, Rio de Janeiro, Brazil, 2020: 484-492.
- [86] Tan S, Ji B, Pan Y. Emmn: emotional motion memory network for audio-driven emotional talking face generation[C]//Proceedings of the IEEE International Conference on Computer Vision, Paris, France, 2023: 22146-22156.
- [87] Mittal G, Wang B. Animating face using disentangled audio representations[C] //Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, Aspen, USA, 2020: 3290-3298.
- [88] Liu J, Wang X, Fu X, et al. FONT: flow-guided one-shot talking head generation with natural head motions[C]//IEEE International Conference on Multimedia and Expo, Brisbane, Australia, 2023: 2099-2104.
- [89] Zhang W, Cun X, Wang X, et al. Sadtalker: learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 8652-8661.
- [90] Fan Y, Lin Z, Saito J, et al. Faceformer: speech-driven 3D facial animation with transformers[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 18770-18780.
- [91] Bernardo B, Costa P. A speech-driven talking head based on a two-stage generative framework[C]//Proceedings of the 16th International Conference on Computational Processing of Portuguese, Santiago, Spain, 2024: 580-586.
- [92] Guo Y, Chen K, Liang S, et al. Ad-nerf: audio driven neural radiance fields for talking head synthesis [C]//Proceedings of the IEEE International Conference on Computer Vision, Virtual, 2021: 5784-5794.
- [93] Yao S, Zhong R Z, Yan Y, et al. Dfa-nerf: personalized talking head generation via disentangled face attributes neural rendering[J]. *Arxiv Preprint*, 2022, 1: No. 220100791.
- [94] Li J, Zhang J, Bai X, et al. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis[C]//Proceedings of the IEEE International Conference on Computer Vision, Paris, France, 2023: 7568-7578.
- [95] Ye Z, Jiang Z, Ren Y, et al. Geneface: generalized and high-fidelity audio-driven 3D talking face synthesis[C]//International Conference on Learning Representations, Kigali, Rwanda, 2023: No. 230113430.
- [96] Yu Z, Yin Z, Zhou D, et al. Talking head generation with probabilistic audio-to-visual diffusion priors[C]//Proceedings of the IEEE International Conference on Computer Vision, Paris, France, 2023: 7645-7655.

- [97] Shen S, Zhao W, Meng Z, et al. DiffTalk: crafting diffusion models for generalized audio-driven portraits animation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 1982-1991.
- [98] Bigioi D, Basak S, Stypułkowski M, et al. Speech driven video editing via an audio-conditioned diffusion model[J]. *Image and Vision Computing*, 2024, 142: No. 104911.
- [99] Cooke M, Barker J, Cunningham S, et al. An audio-visual corpus for speech perception and automatic speech recognition[J]. *The Journal of the Acoustical Society of America*, 2006, 120(5): 2421-2424.
- [100] Cao H, Cooper D G, Keutmann M K, et al. Crema-d: crowd-sourced emotional multimodal actors dataset [J]. *IEEE Transactions on Affective Computing*, 2014, 5(4): 377-390.
- [101] Harte N, Gillen E. TCD-TIMIT: an audio-visual corpus of continuous speech[J]. *IEEE Transactions on Multimedia*, 2015, 17(5): 603-615.
- [102] Chung J S, Zisserman A. Lip reading in the wild[C]//Computer Vision-ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, China, 2017, 2(13): 87-103.
- [103] Busso C, Parthasarathy S, Burman A, et al. MSP-IMPROV: an acted corpus of dyadic interactions to study emotion perception[J]. *IEEE Transactions on Affective Computing*, 2016, 8(1): 67-80.
- [104] Badr A, Hassan A A. VoxCeleb1: speaker age-group classification using probabilistic neural network[J]. *International Arab Journal of Information Technology*, 2022, 19(6): 854-860.
- [105] Suwajanakorn S, Seitz S M, Kemelmacher S I. Synthesizing obama: learning lip sync from audio[J]. *ACM Transactions on Graphics*, 2017, 36(4): 1-13.
- [106] Czyzewski A, Kostek B, Bratoszewski P, et al. An audio-visual corpus for multimodal automatic speech recognition[J]. *Journal of Intelligent Information Systems*, 2017, 49: 167-192.
- [107] Chung J S, Nagrani A, Zisserman A. VoxCeleb2: deep speaker recognition[J]. *Arxiv Preprint*, 2018, 6: 180605622.
- [108] Afouras T, Chung J S, Senior A, et al. Deep audio-visual speech recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 44(12): 8717-8727.
- [109] Afouras T, Chung J S, Zisserman A. LRS3-TED: a large-scale dataset for visual speech recognition[J]. *Arxiv Preprint*, 2018, 9: No. 180900496.
- [110] Alghamdi N, Maddock S, Marxer R, et al. A corpus of audio-visual lombard speech with frontal and profile views[J]. *The Journal of the Acoustical Society of America*, 2018, 143(6): 523-529.
- [111] Poria S, Hazarika D, Majumder N, et al. Meld: a multimodal multi-party dataset for emotion recognition in conversations[J]. *Association for Computational Linguistics*, 2019, 7: 527-536.
- [112] Livingstone S R, Russo F A. The ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English[J]. *Plos One*, 2018, 13(5): No. e0196391.
- [113] Yang S, Zhang Y, Feng D, et al. LRW-1000: a naturally-distributed large-scale benchmark for lip reading in the wild[C]//2019 14th IEEE International Conference on Automatic Face & Gesture Recognition, Lille, France, 2019: 1-8.
- [114] Rossler A, Cozzolino D, Verdoliva L, et al. FaceForensics++: learning to detect manipulated facial images[C]//Proceedings of the IEEE International Conference on Computer Vision, Seoul, South Korea, 2019: 1-11.
- [115] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 4401-4410.
- [116] Wang K, Wu Q, Song L, et al. Mead: a large-scale audio-visual dataset for emotional talking-face generation[C]//European Conference on Computer Vision, Virtual, 2020: 700-717.
- [117] Zhang Z, Li L, Ding Y, et al. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 2021: 3661-3670.
- [118] Wang T C, Mallya A, Liu M Y. One-shot free-view neural talking-head synthesis for video conferencing[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 2021: 10039-10049.
- [119] Zhu H, Wu W, Zhu W, et al. CelebV-HQ: a large-

- scale video facial attributes dataset[C]//European Conference on Computer Vision, Tel Aviv, Israel, 2022: 650-667.
- [120]Wu C, Zheng N, Ardisson S, et al. Multiface: a dataset for neural face rendering[J]. Arxiv Preprint, 2022, 7: No. 220711243.
- [121]Chen L, Cui G, Kou Z, et al. What comprises a good talking-head video generation?[C]//IEEE Conference on Computer Vision and Pattern Recognition Workshops, Virtual, 2020: No. 200503201.
- [122]Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.
- [123]Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in Neural Information Processing Systems, 2017, 30: No. 170608500.
- [124]Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 586-595.
- [125]Chen L, Li Z H, Maddox R K, et al. Lip movements generation at a glance[C]//Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018: 520-535.
- [126]Hong F T, Xu D. Implicit identity representation conditioned memory compensation network for talking head video generation[C]//Proceedings of the IEEE International Conference on Computer Vision, Paris, France, 2023: 23062-23072.
- [127]Doukas M C, Zafeiriou S, Sharmanska V. Headgan: one-shot neural head synthesis and editing[C]//Proceedings of the IEEE International Conference on Computer Vision, Virtual, 2021: 14398-14407.
- [128]张雪. 维塔数字智能面部动画系统在电影《阿凡达: 水之道》得到成功应用[J]. 现代电影技术, 2023, 2023(5): 63-64.
- Zhang Xue. The successful application of weta digital's intelligent facial animation system in the film avatar: the way of water[J]. Advanced Motion Picture Technology, 2023, 2023(5): 63-64.
- [129]韦剑峰. 虚拟数字人技术在广电新媒体中的应用[J]. 卫星电视与宽带多媒体, 2024, 21(19): 22-24.
- Wei Jian-feng. Application of virtual digital human technology in radio, television and new media[J]. Satellite TV & IP Multimedia, 2024, 21(19): 22-24.