ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 41,No. 2,Mar. 2026,pp. 543－565
DOI:10. 16337/j. 1004-9037. 2026. 02. 013

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

三维人脸生成技术综述

王　伟[1,2], 何一康[1,2], 魏云超[1,2], 赵　耀[1,2]

(1. 北京交通大学信息科学研究所,北京 100044;2. 视觉智能交叉创新教育部国际合作联合实验室,北京 100044)

摘　要: 近年来,计算机视觉与图形学的快速发展推动了三维人脸生成技术的突破,尤其在以数字化身构建领域,三维视觉技术在互联网快速普及,受到了学术界和工业界的广泛关注。该技术通过从显式或隐式的底层表征中重建几何结构与纹理细节来合成逼真的多视角人脸图像,并在娱乐与交互应用中取得显著成果,如通过文本描述修改面部特征的属性编辑,或生成说话视频的说话人脸技术。但早期基于线性参数化模型的技术存在生成的真实感和细节表现不佳的问题,随后兴起的隐式神经表示技术虽然大幅提升了视觉质量,却面临计算成本高昂、难以实时交互的难题,这给实际部署与应用均带来了极大限制。为了克服速度与质量之间的矛盾,众多学者对基于显式高斯基元的新型表征以及基于概率扩散的生成模型进行了深入研究,并从不同视角提出了一系列混合生成方法。此外,生成技术仍面临小样本泛化困难、头部物理建模不完整与动态一致性不足等挑战,使其在实现完全写实与实时交互的道路上仍有很长一段距离。目前,三维人脸生成与驱动技术的研究仍处在发展期。本综述对迄今为止的主要研究工作进行了科学系统的总结与归纳,并对现有技术的局限性做简要分析。最后,探讨了三维人脸生成与应用技术的潜在挑战与发展方向,旨在为领域内未来的研究工作提供借鉴。

关键词: 三维人脸生成;三维可变形模型;神经辐射场;三维高斯泼溅;扩散模型;生成对抗网络

中图分类号: TP391.41　　**文献标志码:** A

引用格式: 王伟,何一康,魏云超,等. 三维人脸生成技术综述[J]. 数据采集与处理,2026,41(2):543-565. WANG Wei,HE Yikang,WEI Yunchao, et al. A survey on 3D face generation technology[J]. Journal of Data Acquisition and Processing,2026,41(2):543-565.

引　　言

随着数字经济与人工智能技术的飞速发展,构建高保真、可编辑且具备实时交互能力的“数字人”已成为元宇宙、虚拟现实(Virtual reality, VR)、增强现实(Augmented reality, AR)以及下一代人机交互界面(Human-machine interface, HMI)的核心需求。三维人脸模型作为数字人最关键的身份标识与情感表达载体,其生成技术的研究具有重要的学术价值与广泛的应用前景。

过去二十年间,三维人脸生成领域经历了一场深刻的范式转换。早期的研究主要建立在Blanz和Vetter[1]提出的三维可变形模型(3D morphable model, 3DMM)之上,通过主成分分析(Principal component analysis, PCA)构建线性参数空间。然而,线性模型的表达能力有限,难以复现人脸的微表情皱纹、皮肤毛孔等高频细节。2020年,神经辐射场(Neural radiance field, NeRF)的提出[2]引发了隐式表示的革命,它通过神经网络编码场景并通过体渲染合成图像,实现了照片级的真实感。尽管NeRF效果已经很好,但其昂贵的计算开销限制了实时应用。2023年,3D高斯泼溅(3D Gaussian splatting, 3DGS)的出现[3]提供了新的解决思路,它以显式的高斯点基元表示结合可微光栅化,在保持高画质的同时实现了

基金项目: 国家自然科学基金(62372033);北京市自然科学基金“海淀联合-重点”项目(L252025)。
收稿日期: 2026-01-09;**修订日期:** 2026-02-26

超高帧率的渲染，为三维人脸生成的实时性应用带来了新的契机。

与此同时，深度生成模型，特别是生成对抗网络（Generative adversarial network，GAN）[4]与扩散模型（Diffusion models）[5]的引入，极大地拓展了三维人脸生成的边界。这些模型利用海量二维图像数据学习到的先验知识，使得从单张照片生成多视角一致的三维人脸成为可能，并赋予了模型强大的语义编辑能力。

为了系统地梳理这一快速发展的领域，本文对三维人脸生成技术进行了全面的综述。本文的结构如图1所示，主要包括三维人脸底层表征方式的演进、三维人脸生成的方法以及三维人脸生成的典型应用场景三大部分。

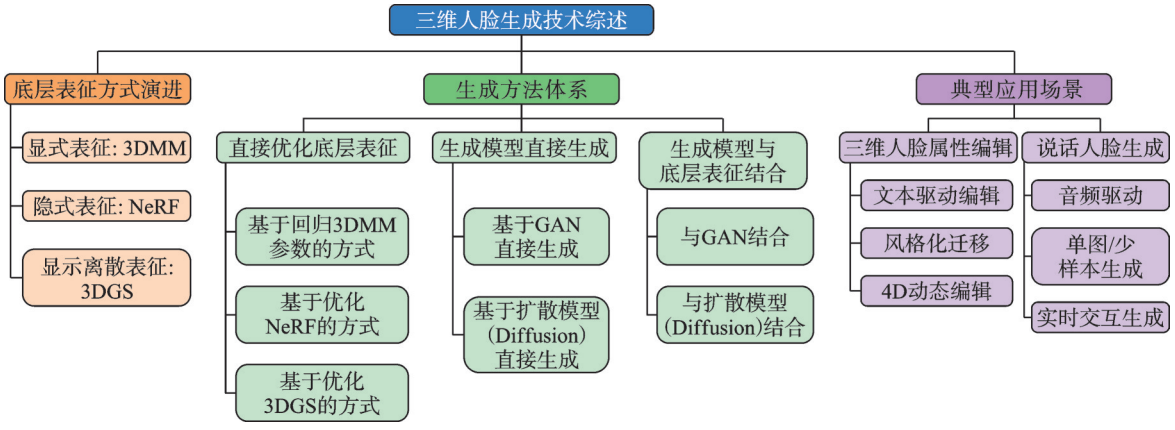

图1 三维人脸生成技术分类导图

Fig.1 Classification map of 3D face generation methods

本文的主要贡献总结如下：

（1）底层表征演进的系统回顾。本文详细梳理了从显式参数化模型（3DMM）到隐式神经场（NeRF），再到新一代显式表征（3DGS）的技术演进脉络，深入分析了每种表征方式的基本原理、核心优势及局限性，阐明了技术发展的内在逻辑。

（2）生成方法分类体系的构建。本文将现有的生成方法归纳为基于直接优化的方法、基于生成模型（GAN，Diffusion）的方法以及混合方法。特别重点分析了如何利用生成式先验解决数据稀缺（Few-shot/Zero-shot）条件下的三维一致性生成问题，对比了不同生成范式在训练效率、生成质量与可控性方面的差异。

（3）前沿应用与挑战的探讨。本文聚焦于属性编辑与说话人脸生成两大应用场景，展示了最新的研究成果如何利用前述底层技术实现文本驱动编辑与实时语音驱动。同时，深入剖析了当前技术在整个头部建模完整性、极端视角泛化能力以及移动端实时部署等方面面临的挑战，并对未来的研究方向进行了展望。

1 三维人脸的底层表征方式演进

三维人脸生成的核心挑战在于如何寻找一种既能高效存储与渲染，又能灵活支持几何与纹理编辑的底层表征。过去二十年间，这一领域经历了从显式几何参数化到隐式神经场，再到显隐式结合的螺旋式上升过程。

1.1 三维可变形模型

三维可变形模型（3DMM）[1]是计算机视觉和图形学中一个影响深远的范式，最早由Blanz和Vetter

提出,其核心思想是将高维度的三维人脸数据表示在一个低维的线性向量空间中。这一空间通常是通过对一个包含了大量配准后三维扫描人脸的数据集应用PCA来构建的。通过这种方式,任何新的人脸都可以被近似为“平均脸”与一系列“主成分”的线性组合。3DMM的关键功能在于它能够解耦人脸的关键属性,例如形状(Shape)和纹理(Texture/Albedo),并将其应用于诸如从单张二维图像重建三维人脸的“分析-合成”(Analysis-by-synthesis)任务中。这一框架为参数化地理解和操控三维人脸提供了强大的数学工具。3DMM的基本流程如图2所示。

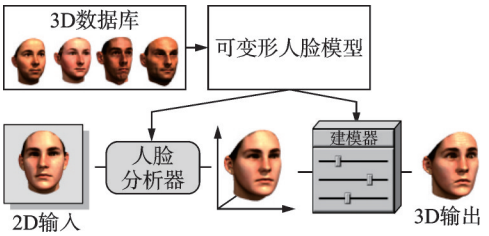

图2　3DMM基本流程[1]

Fig.2　Basic pipeline of 3DMM[1]

3DMM通过对人脸形状、表情和外观进行统计建模,构建出由多边形网格和纹理贴图组成的3D模型,进而可渲染出二维图像[6],其核心组成部分的基本原理如下:

(1) 形状建模(Shape modeling):旨在捕捉不同身份间的几何差异。通常采用统计方法(如PCA),基于平均人脸,使用低维参数来表示个体形状的线性变化。

(2) 表情建模(Expression modeling):核心在于解耦身份特征与动态表情。常见途径包括在静态身份形状上叠加表情偏移量(加法模型),利用多线性张量方法同时分解身份与表情,或采用非线性变换方法。

(3) 外观建模(Appearance modeling):用于定义人脸表面的纹理和色彩。主要方式是通过统计分析顶点的颜色变化,或者在UV纹理空间中定义能包含更丰富细节的纹理贴图。

根据数据来源、覆盖范围及建模方法的不同,学界发布了多种公开模型。表1展示了几种最具代表性的模型特点分析。

表1　主流三维可变形模型分析对比

Table 1　Comparative analysis of mainstream 3D morphable models

模型名称	核心特征与建模方法	数据规模与特点	适用场景
BFM(Basel face model)[7-8]	经典PCA模型。BFM 2009[7]主要包含形状和纹理(逐顶点);BFM 2017[8]增加了表情模型	约200名个体的扫描数据。数据质量高,但样本量相对较小,主要覆盖中性表情(早期版本)	静态人脸重建、作为基础模型进行学术研究
FaceWarehouse[9]	多线性张量模型。强调表情的多样性,采用Identity × Expression的张量结构	150名个体,每人20种表情。表情覆盖广泛,适合动态重建	实时面部表情捕捉和基于RGB-D的三维面部重建
LSFM (Large scale facial model)[10]	大规模统计模型。基于极大样本量构建的PCA模型,使用了自动化的非刚性ICP流程	9 663名个体。样本量极大,种族、年龄覆盖面广,主要针对形状(Identity)	需要高泛化能力的种群统计分析、大范围人脸形状重建
FLAME[11]	非线性/关节模型。结合了线性混合变形(Blendshapes)与关节变换(如头部和下颚旋转),相比纯线性模型更符合解剖学结构	3 800名个体的形状数据 + 21 000帧表情序列。包含头部姿态(Pose)建模	现代基于学习的重建流程、人体/人头联合建模及需要解剖学合理性的动画
CoMA(Convolutional mesh autoencoder)[12]	深度学习模型。使用卷积网格自编码器在网格空间进行卷积	12名个体的极端表情序列(共20k+网格)	非线性表示学习、极低维度的形状压缩与生成

1.2 神经辐射场

3DMM已经可以表示一个三维人脸，但是其直接渲染生成的人脸缺乏真实感，并且对于头发等变化形式多样、区域不固定的特征难以直接进行建模，因此，需要使用其他方式来表征整个三维头部。

传统上，三维场景主要通过显式表征(Explicit representations)来描述。这些方法直接定义场景的几何形状，主要包括：

(1) 点云(Point clouds)。由三维空间中的离散点集合构成。

(2) 体素网格(Voxel grids)。将三维空间划分为规整的立方体单元。

(3) 多边形网格(Meshes)。由顶点(Vertices)、边(Edges)和面(Faces)构成的物体表面拓扑结构。

相对地，隐式表征不直接存储几何实体，而是用一个连续函数来定义场景。如图3所示，神经辐射场(NeRF)的本质是将静态的三维场景隐式地存储在一个连续的5D函数(由MLP网络近似)中，并通过可微的体渲染技术从任意视角合成二维图像[2]。

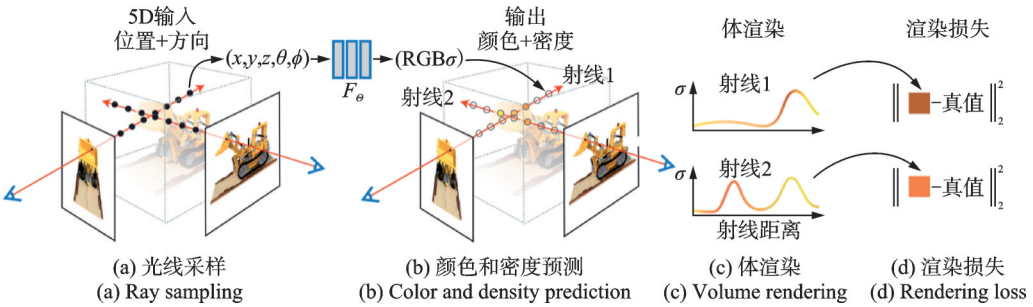

图3 NeRF基本原理[2]

Fig.3 NeRF's core principles[2]

具体来说，NeRF使用一个多层感知机F_Θ来编码整个场景。其接收空间中的三维坐标(x, y, z)和观察视角$\boldsymbol{d}=(\theta, \varphi)$作为输入，输出该位置的体积密度$\sigma$和颜色$c=(r, g, b)$。NeRF采用的可微体渲染技术模拟光线穿过场景的过程，通过沿着光线路径对采样点的颜色和密度进行累加(积分的离散近似)，从而计算出二维图像上每个像素的最终颜色。由于整个体渲染过程是可微分的，因此可以通过计算渲染出的图像与真实图像之间的误差，并利用梯度下降算法来反向优化神经网络的权重，从而实现对三维场景的学习和重建。

NeRF可以被视为一种高效的场景压缩算法。它将海量的三维视觉信息“压缩”到了只有几兆字节的MLP权重参数中，通过求解逆渲染问题来重建三维世界。最终通过将3DMM与NeRF相结合便可以建模一个隐式的三维人脸模型，通过体渲染就可以生成照片级真实感的不同视角下的人脸图像。

1.3 3D高斯泼溅

虽然NeRF已经可以渲染出高保真度的人脸图像，但是其面临一个核心的矛盾：高质量的渲染效果通常需要以高昂的计算成本为代价。例如，达到当时最先进视觉质量的Mip-NeRF 360[13]模型，其训练时间长达数天，渲染速度远低于实时应用的要求。后续工作如Instant-NGP[14]等虽然显著缩短了训练时间，却不得不在一定程度上牺牲渲染质量。正是在这样的背景下，3D高斯泼溅(3DGS)技术应运而生，它放弃了隐式的神经网络权重编码，转而采用显式的、基于高斯点的渲染技术，通过数百万个可学习的3D高斯基元集合来表示场景。这种显式表示避免了渲染过程中耗时的神经网络查询，结合高度优化的可微光栅化渲染模块，3DGS在保持媲美甚至超越顶尖NeRF方法视觉质量的同时，成功实现了高分辨率(1 080P)下的实时渲染(>100 fps)和数十分钟级的快速训练，迅速成为新视角合成领域的新标准。

3DGS的基础是其场景表示的基本单元:3D高斯基元。每个基元都是一个在三维空间中的高斯分布,它是一个显式的、可学习的实体,由一组精确的参数定义,这些参数共同决定了它在场景中的位置、形状、颜色和透明度。每个高斯基元由以下参数定义:

(1) 位置(Mean, μ)。一个三维向量,确定了高斯基元在世界坐标系中的中心点。

(2) 协方差(Covariance, Σ)。描述了高斯基元的形状(各向异性椭球体)、大小和方向。为了保证优化的稳定性和物理有效性,它一般通过被分解为可学习的缩放和旋转组件来表示。

(3) 颜色(Color, c)。通过球谐函数(SH)系数来表示,这使得基元能够模拟高光反射等依赖于观察方向(视角)的颜色变化,从而呈现出高度真实的、与视角相关的外观。

(4) 不透明度(Opacity, α)。一个直接可学习的标量值,表示基元的透明程度,用于渲染阶段的混合计算。

3DGS的训练是一个端到端的优化过程,目标是调整高斯基元的属性,使其渲染结果匹配真实图像。主要包括以下3个步骤:

(1) 初始化(Initialization)。利用运动恢复结构(Structure-from-motion, SfM)生成的稀疏点云作为起点,每个点初始化为一个高斯基元:位置取点坐标,颜色取点颜色,形状根据最近邻距离初始化为球体。

(2) 损失函数(Loss function)。在二维图像层面计算渲染图像与真实图像的差异(不同于NeRF的光线采样)。结合L_1损失(像素绝对差异)和D-SSIM损失(结构相似性),损失函数L可表示为

$$L=(1-\lambda)\mathcal{L}_{\mathrm{L1}}+\lambda\mathcal{L}_{\mathrm{D\text{-}SSIM}} \tag{1}$$

(3) 自适应密度控制(Adaptive density control)。为了精细化几何结构,系统依据梯度信息动态调整基元的数量和分布,包括① 致密化。针对重建不足的区域,通过“克隆”(复制并移动小基元以填充)或“分裂”(将大基元拆分为小基元以细化)来增加基元密度;② 剪枝。移除不透明度过低的基元,以去除冗余噪声并保持模型紧凑。

最终训练得到的3DGS基元便可以通过一个类似于光栅化的渲染流程快速渲染得到多视角的高质量人脸图像,完成了对于三维人脸的建模。图4所示为3DGS的训练流程图[3]。

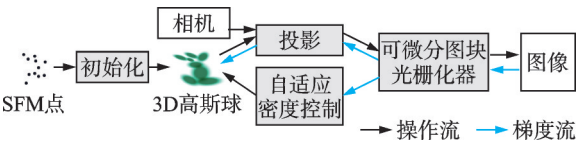

图4　3DGS训练流程图[3]

Fig.4　3DGS training pipeline[3]

综上所述,三维人脸表征在过去二十年间经历了螺旋式上升的演进:从早期侧重结构解耦但缺乏真实感的3DMM显式低维线性参数化,转向虽能实现照片级真实感却计算代价昂贵的NeRF隐式神经辐射场,最终发展为兼顾实时性与高保真的3DGS显式高斯基元表示。这一演变历程为三维人脸生成这一复杂任务奠定了坚实基础。

2　三维人脸生成方法

有了3DMM、NeRF和3DGS等基本的三维人脸底层表示方法,就可以实现从单张/多张图像中直接建模三维人脸模型,并进行新视角图像的生成这一最终目标。当拥有大量多视角单一目标人脸的图像时,直接优化这些底层表征模型便可以得到比较好的结果,但是在仅有少量甚至单张图像的情况下,直接优化的方式难以获得较好的结果。而生成式模型如生成对抗网络、扩散模型具有强大的泛化生成能力,于是便可以将其与底层三维表征相结合,借助其内部蕴含的强大先验知识在数据缺乏的状态下生成高质量的新视角人脸图像。此外,部分研究直接训练生成模型,不需对三维人脸进行显示建模,便可以生成对应的多视角图像。

2.1 直接优化的方式

通过直接优化三维人脸底层表示的参数来实现三维人脸生成的方法往往需要大量的多视角图像用于训练，同时对于计算资源的要求也较高，这些方法生成的人脸图像质量较高，同时也可以作为其他方法的基础，参与到生成模型的整个流程中。

2.1.1 基于回归三维可变形模型参数的方式

将3DMM应用于三维人脸生成，主要依赖于深度学习方法对参考图像进行参数估计。接着，使用得到的3DMM参数进行推理构建出由多边形网格(Mesh)表示的三维人脸几何结构，并同步优化获取纹理贴图。最终，结合几何与纹理渲染得到多视角人脸图像，这一过程也被称为3D人脸重建。3D人脸重建领域在不同维度上取得了显著进展，代表性方法如Deep 3D Face Reconstruction[15]、DECA[16]、EMOCA[17]和MICA[18]分别解决了弱监督训练、动态细节、情感表达与度量准确性方面的核心挑战。Deep 3D Face Reconstruction率先采用了基于可微渲染的弱监督学习框架，利用ResNet-50回归3DMM参数，并创新性地提出了基于置信度的多图聚合策略，通过加权组合而非简单平均来有效利用多张图像信息以提升身份重建的鲁棒性。在此基础上，DECA致力于解决几何细节难以随表情动画化的问题，它不仅回归粗糙形状，还通过引入细节一致性损失函数，成功将个人特有的静态细节(如毛孔)与随表情变化的动态皱纹解耦，实现了可动画化的高频细节重建。针对传统方法难以捕捉细腻情感的缺陷，EMOCA在DECA架构上增加了可训练的表情编码器，并引入深度感知情感一致性损失，利用预训练的情感识别网络监督几何生成，确保重建结果能较好地体现出输入数据的表情特征。不同于上述方法主要依赖自监督视觉对齐，MICA旨在解决透视投影导致的深度与尺度模糊问题，它利用在海量2D数据上预训练的人脸识别网络(ArcFace)提取鲁棒身份特征，并结合统一的3D扫描数据集进行监督学习，从而预测出具有正确度量尺度的中性人脸几何形状。通过这些方法可以有效地获得一个完整的三维人脸头部模型，但是如图5所示的结果[1]，其渲染得到的人脸图像真实感较差，并且缺乏头发等其他头部属性的建模，因此这些方法往往用于作为一种提取先验信息的方式，与其他生成式模型相结合进而实现可驱动的、真实感更强的三维人脸生成。

图5 3DMM渲染得到的图像[1]

Fig.5 Image obtained via 3DMM rendering[1]

2.1.2 基于优化神经辐射场的方式

NeRF通过隐式表示有效克服了传统显式网格在拓扑固定与精细结构(如毛发、口腔内部)表达上的固有缺陷，同时也解决了二维生成模型在三维一致性方面的不足。针对不同的应用场景，现有研究主要从通用参数化模型、个性化语义重建以及高保真动态合成3个维度进行了深入探索：

(1)在通用参数化与实时驱动方面，HeadNeRF[19]提出了一种创新的实时参数化头部模型。该方法将NeRF作为通用的3D几何与纹理表示，通过输入解耦的潜在代码(Latent codes)分别控制身份、表情、反照率和光照，从而实现了对头部属性的灵活解耦与编辑。为了突破传统体积渲染的计算瓶颈，HeadNeRF采用了一种高效的混合渲染策略，即首先通过体积渲染生成低分辨率特征图，随后利用2D神经渲染模块将其上采样为高分辨率图像。这种设计在保持严格多视角一致性的同时，大幅降低了计

算成本,成功在现代GPU上实现了超过40 fps的实时渲染性能。

(2)在个性化模型的快速构建与语义控制方面,Gao等[20]提出了一种将传统混合蒙皮(Blend shape)语义融入隐式表示的单目视频重建方法。其核心在于构建一组基于多分辨率哈希编码的体素场(Voxel fields)作为语义基底,并通过在潜在特征空间中利用表情系数对这些基底进行线性加权组合来驱动体素变形。这种"隐式线性混合"(Implicit linear blending)架构不仅继承了传统混合蒙皮的语义可解释性,支持对面部属性的直观编辑与重演,还通过调节局部特征分布显著降低了多层感知机(MLP)的学习负担,使得模型仅需10至20 min即可从短视频中完成训练,并能精准捕捉皱纹、发丝等高频细节。

(3)在复杂动态场景的高保真重建方面,为了解决极端非刚性形变(如快速头发运动、剧烈表情变化)的捕捉难题,NeRSemble[21]针对多视角视频输入,设计了变形场(Deformation field)与哈希编码集合(Hash ensembles)相结合的联合表征机制。其中,变形场用于处理粗略的场景运动与空间对齐,而哈希集合则通过基于时间权重的特征混合来精细刻画随时间演变的高频纹理细节。特别地,该研究引入了"预热"(Warm-up)训练策略,强制模型先学习几何对应关系再引入高频特征,有效避免了局部极小值并消除了空洞伪影。依托于其构建的7.1MP高分辨率、73 fps高帧率的大规模多视角数据集,NeRSemble在动态人脸渲染的时间一致性与视觉保真度上均取得了显著突破。图6展示了其生成效果,与图5中3DMM渲染得到的图像相比,NeRF生成的图像真实感大幅提升,并且能够展现出更多的纹理细节。

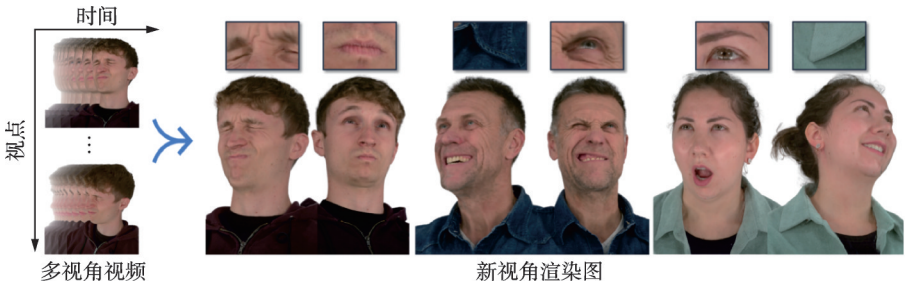

图6　NeRSemble[21]生成的图像

Fig.6　Image generated by NeRSemble[21]

2.1.3　基于直接优化3D高斯泼溅参数的方式

基于NeRF的方法已经可以较为高效地生成高质量的多视角人脸图像,但是其较长的训练时间,无法实现高分辨率下的实时渲染等缺点限制了其实际应用。最近,基于3DGS的方法提供了解决这些问题的新思路,其技术路线大致呈现出从"显式网格驱动"向"几何代理扩展",再到"隐式神经变形"的演进逻辑。

在显式驱动方面,GaussianAvatars[22]与SplattingAvatar[23]均利用参数化3DMM(如FLAME)作为强先验来驱动3D高斯场,以实现对于头部姿态、表情等的控制。GaussianAvatars的基本原理是"刚性绑定",即将3D高斯基元视为附着在三角面片局部的静态粒子,其位移与旋转完全由父三角形在全局空间中的刚性变换所决定,这种设计确保了高斯基元能够精确继承底层的网格运动逻辑,并通过绑定继承机制维持训练中自适应调整过程的拓扑一致性。SplattingAvatar则在此基础上深化了"几何嵌入"的数学表达,提出了一种基于Phong[24]表面的网格嵌入机制。不同于简单的刚性附着,该方法通过重心坐标和法向位移将高斯基元参数化,并引入了"提升优化"与"三角形游走"算法,允许高斯基元在优化过程中跨越三角形边界在网格表面"滑动"以寻找最佳几何锚点,从而实现了运动控制与外观表征的彻底解耦,在保证高帧率渲染的同时提升了纹理精度。

为了突破标准参数化网格在拓扑结构上的局限性,PSAvatar[25]提出了一种"几何代理扩展"的思

路。该方法认为仅依赖FLAME[11]网格表面无法有效表示头发、眼镜等“网格外部”的结构,因此构建了“基于点云的可变形形状模型”(Point-based morphable shape model, PMSM)。其核心原理是通过在网格表面及其法向延伸空间进行双重采样,构建一个既继承网格运动先验、又具备空间体积表达能力的泛化点云,并利用分析-合成的方式将这些点与3D高斯基元进行对齐。这种半显式的策略在保留网格驱动优势的同时,显著拓展了生成的3D人脸模型的几何表达范围,解决了传统3DMM难以处理非面部区域的痛点。

在追求极致的非线性变形与细节表现时,Gaussian Head Avatar[26]与MonoGaussianAvatar[27]转向了“神经变形场”的探索。Gaussian Head Avatar指出线性混合蒙皮(Linear blend skinning, LBS)难以拟合极端表情(如大张嘴)和纹理细节如皱纹,因此该方法摒弃了显式网格绑定,转而采用全学习的MLP动态生成器。该方法直接将标准空间的高斯属性与表情、姿态系数映射为观察空间的几何偏移和属性变化,利用神经网络的非线性拟合能力来捕捉高频动态细节,并辅以符号距离场(Signed distance field, SDF)几何引导初始化来解决离散高斯点难以收敛的问题。针对单目重建这一更具挑战性的场景,MonoGaussianAvatar则结合了显式LBS与隐式变形场的优势,设计了包含残差变形网络的混合驱动机制,并引入了动态的点云插入与删除策略。这一策略旨在通过动态调整高斯点的密度与分布,解决单目视角下的几何歧义以及夸张动作(如大张嘴显露出口腔内部)带来的拓扑空洞问题,从而在缺乏多视角约束的情况下依然保持结构的稳定性与真实感。

这些方法从不同的角度寻找如何更好地利用3DMM的先验知识,进而实现了通过使用3DMM参数驱动3D高斯场的变形,生成对应表情、姿态的高质量三维人脸。其渲染速度与训练速度均大幅领先于基于NeRF的方法,是目前具有大量多视角图像数据时最优的三维人脸生成方法。

2.2 生成模型直接生成三维人脸的方式

三维人脸生成任务本身的目的是在有部分视角图像数据的输入下,能获得新的视角下的图像,不同于NeRF或3DGS等依赖显式或隐式几何建模的传统路线,生成式模型的突破带来了一种全新的技术路径。其核心策略在于挖掘预训练生成模型中蕴含的丰富先验,并将其与相机参数解耦;通过在特定相机位姿下的条件化生成,即可在不显式重建三维模型的情况下,完成高质量的新视角人脸图像生成。

2.2.1 基于生成对抗网络直接生成的方式

生成对抗网络(GAN)[4]的核心原理基于博弈论中的极小极大博弈框架。该框架包含两个相互竞争的神经网络,即生成器(Generator)和判别器(Discriminator)。其中,生成器致力于学习真实数据的概率分布,试图将随机噪声映射为逼真的合成样本以“欺骗”对手;判别器则作为一个二分类器,旨在最大化区分真实样本与生成样本的准确率。

在三维可控人脸生成与编辑领域,如何在真实感保持的同时实现属性解耦是一个核心难题。相关研究从基于先验的解耦生成向基于物理条件的精确控制持续演进。DiscoFaceGAN[28]开创性地在生成对抗网络中引入3DMM先验,构建了“3D模仿-对比学习”框架,其核心原理在于利用变分自编码器(Variational autoencoder, VAE)将潜在变量映射为3DMM系数,通过强制生成器模仿物理渲染过程,确立潜在空间与身份、姿态、光照等物理属性的对应关系。为进一步解决渲染域与真实图像域差异导致的属性纠缠,该方法引入对比学习机制,在潜在空间微调单一属性时使用一个惩罚非目标属性变化的损失函数来实现一定程度上的解耦控制,从而在无监督或弱监督条件下实现了各属性在生成过程中的独立控制。

然而,上述方法在对真实图像进行操作时依赖于有损且低效的GAN反演操作,这一过程往往导致编辑后的图像质量下降。针对这一局限,3D-FM GAN[29]确立了专门面向人脸属性编辑的条件生成范式。该方法摒弃了传统的噪声驱动生成,转而采用物理驱动的编辑方式,即利用3D重建与渲染网络直

接为生成器提供显式的几何与光照引导。其架构创新的核心在于“乘性共同调制”(Multiplicative co-modulation)机制,该机制将包含身份信息的源图像特征与包含编辑信号的渲染特征分别编码至不同的调制空间,并通过乘性融合策略有效解决了身份保持与几何形变之间的权衡难题。配合基于合成数据与真实数据的双重训练策略,该框架成功实现了无需反演的高保真度三维人脸生成与属性编辑。

2.2.2　基于扩散模型直接生成的方式

关于如何突破传统几何建模与神经辐射场在纹理高频细节与泛化能力上的瓶颈,同时克服生成对抗网络生成的图像质量较低的困境,将二维扩散模型强大的图像生成先验与三维参数化模型相结合,成为解决单目重建不适定问题、实现零样本新视角合成及高保真个性化编辑的关键路径。

在单张图像的零样本新视角合成方面,DiffPortrait3D[30]展示了如何在无需微调的情况下实现身份保持与视角解耦。该方法解决的核心挑战在于如何在改变视角的同时,维持原图的身份、表情及背景内容。DiffPortrait3D利用在大规模图像数据集上预训练的Stable Diffusion[31]作为骨干,设计了一个外观参考模块(Appearance reference module),将参考图像的特征注入到冻结的UNet自注意力层中,使模型在生成过程中能够持续“查询”参考图像的纹理语义,从而在不同视角下保持外观一致性。为了实现视角的精确控制并避免参考图像中的姿态信息泄露,DiffPortrait3D引入了一个新颖的视角控制模块(View ControlNet)。该模块不直接使用目标视角的姿态参数,而是通过观察一个代理三维渲染图来解析相机姿态,这种设计有效地将几何控制与外观生成解耦。为了解决生成过程中的视角不一致问题,DiffPortrait3D在推理阶段引入了跨视角注意力机制(Cross-view attention),并采用了一种3D感知噪声生成策略,即利用轻量级3D网络预测的粗糙新视角图像作为扩散过程的初始噪声,从而为生成过程提供了强有力的结构引导,显著减少了多视角生成中的闪烁与伪影。这种设计使得DiffPortrait3D能够处理各种风格的人像,包括写实照片与艺术画作,展现了强大的泛化能力。

针对特定人物的高保真外观编辑,DiffusionRig[32]深入探讨了如何通过学习个性化先验来实现对光照、表情和头部姿态的物理级“绑定”(Rigging)。如图7所示,DiffusionRig提出了一种两阶段训练策略:首先在FFHQ[33]等大规模人脸数据集上训练模型,使其具备将由DECA估算的3DMM参数推理得到的粗糙特征(如表面法线图、纹理图、朗伯(Lambertian)渲染图)映射为真实照片的通用人脸先验的能力;随后,在包含约20张特定人物照片的小型数据集上进行微调,使模型学习该人物特有的高频细节(如皱纹、痣)。为了处理3DMM无法建模的非刚性特征(如头发、眼镜、背景),DiffusionRig引入了一个全局编码器来提取表征身份信息的全局潜在代码(Global latent code),与3DMM推理得到的特征图共同作为扩散模型的条件来控制其生成方向。这种设计使得用户可以通过修改3DMM参数来精确控制生成的几何与光照属性,同时利用全局潜在代码保持背景与纹理特征的稳定性,实现了可驱动性与图像真实感的有机结合。

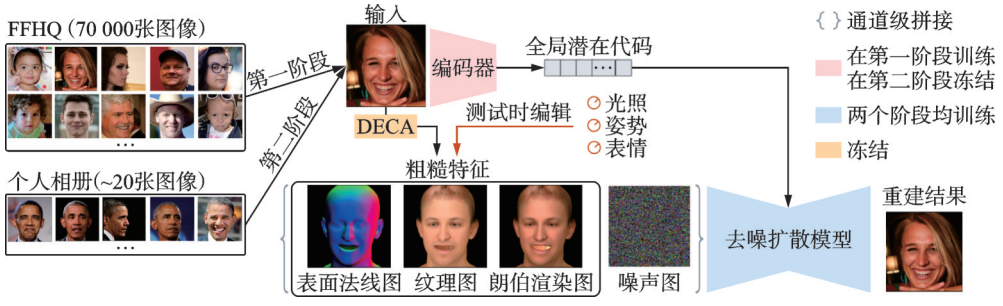

图7　DiffusionRig[32]基本流程

Fig.7　Pipeline of DiffusionRig[32]

Morphable Diffusion[34]进一步将三维一致性与精细动画控制推向新高度。该工作旨在从单张图像生成可驱动的三维头像,解决了现有方法在生成新表情时难以保持视角一致性的问题。Morphable Diffusion的核心创新在于将三维可变形模型(如NPHM[35]或FLAME[11])深度集成到扩散模型中。它提出了一种基于3DMM的特征体素化方法,通过将2D图像的噪声特征反投影到3DMM的网格顶点上,并利用稀疏3D卷积网络(Sparse conv net)处理这些顶点特征,构建出一个包含几何信息的3D特征体(Feature volume)。这种特征提升机制确保了生成的图像在几何结构上与输入网格结构严格对齐。为了捕捉微小的面部表情变化,Morphable Diffusion在UNet的交叉注意力层中直接注入了表情代码(Expression codes),使模型能够感知并生成细微的肌肉运动,弥补了仅靠粗糙网格难以表达细腻情感的缺陷。此外,该方法采用了一种解耦的训练策略,即在训练时输入图像的表情与目标生成的表情不一致,强制模型学习几何变形与纹理生成的解耦,从而赋予了模型对未见主体进行表情驱动的能力。实验表明,这种结合了显式几何引导与像素对齐特征的方法,在生成新视角和新表情时均优于现有的基线模型。

综上所述,这些方法共同确立了一个核心技术范式:利用三维参数化模型(如FLAME、NPHM)提供精确的几何与物理控制,同时利用二维扩散模型提供丰富的纹理细节与生成先验。这一系列方法不仅克服了传统显式建模在真实感上的不足,也解决了纯二维生成在三维一致性上的短板,在输入数据视角和数量不够丰富时也可以实现新视角图像的生成,同时可以快速泛化到不同的人脸身份上。

2.3 生成模型与显示三维表示相结合的方式

2.1节中介绍的方法大多都适用于拥有大量多视角单一目标人脸的图像的情况,并且这些方法很难直接泛化到不同身份(Identity)的人脸生成上,而2.2节中介绍的方法往往由于缺乏直接的3D建模而缺失多视角一致性,因此,使用生成模型与NeRF和3DGS相结合的方式应运而生,借助生成模型中蕴含的丰富先验知识和强大泛化能力,可以在数据缺乏的情况下也能实现高质量,多视角一致的三维人脸生成。

2.3.1 与生成对抗网络相结合的方式

近年来,生成对抗网络(GAN)与神经辐射场(NeRF)的深度融合推动了3D感知图像合成领域的快速演进,其核心发展逻辑遵循着从高效静态表征的构建,到语义一致性的增强,再到高保真动态驱动与显式解耦控制的技术路径。最近的研究工作通过对三维场景表示、渲染机制及驱动方式的不断革新,逐步克服了计算效率低下、多视角不一致以及动态控制僵硬等关键瓶颈。

在高效3D生成表征的探索阶段,研究重点在于解决传统基于NeRF的体渲染计算昂贵且难以捕捉高频细节的问题。GRAM[36]创新性地提出了一种基于流形约束的辐射场生成策略,其核心思想是将点采样和辐射场学习限制在一组在3D空间中联合学习的隐式2D流形上,而非整个体积空间。这种方法将随机的蒙特卡洛采样转变为确定性的表面采样,在显著降低计算开销的同时,有效消除了因采样不足导致的图像噪声,并提升了对发丝等高频几何细节的捕捉能力。与之相呼应,EG3D[37]则通过提出“三平面”(Tri-plane)混合表示进一步推动了效率的革命。如图8所示,该方法利用StyleGAN2[38]生成3个正交特征平面,通过投影与聚合高效查询空间特征,并结合神经渲染与2D超分辨率模块实现高分辨率生成。针对EG3D中2D超分辨率模块可能破坏物理3D一致性的隐患,Mimic3D[39]提出了一种“3D模仿2D”的训练策略,利用2D超分辨率分支作为“教师”来监督纯3D渲染分支的训练,从而在保留高频细节的同时维持严格的多视角一致性。

在确立了高效且一致的静态生成范式后,研究焦点转向了对生成内容的语义理解与局部编辑能力的探索。FENeRF[40]通过在神经辐射场中引入与几何、纹理严格空间对齐的语义场,实现了对3D生成内容的语义解耦。该方法利用解耦的潜在形状编码与潜在纹理编码控制不同的语义场,并通过2D语义标签的弱监督联合训练,使得用户能够通过修改2D语义分割图反向优化潜在代码,从而实现对发

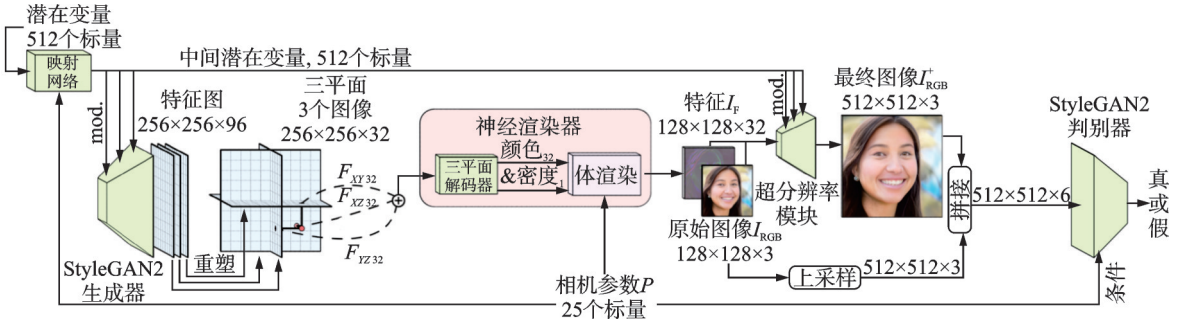

图8　EG3D[37]的三平面表示训练流程

Fig.8　Training pipeline of the tri-plane representation in EG3D[37]

型、表情等属性的精确3D编辑且保持多视角一致。

然而,仅实现静态语义编辑尚不足以满足复杂的交互需求,如何实现从静态到动态的跨越,即赋予模型高保真的动态驱动能力,成为了后续研究的核心。AniFaceGAN[41]在此方向上做出了重要尝试,旨在解决无3D数据监督下的表情驱动难题。其核心原理是将3D表示分解为表征规范身份的"模板辐射场"和表征表情运动的"3D变形场"。通过学习从观测空间到模板空间的逆向位移映射,并引入创新的3D模仿学习机制,强制生成器在几何和变形行为上拟合参数化人脸模型(3DMM)的先验知识,从而实现了纹理与几何高度一致的动态表情驱动,有效避免了纯2D方法中常见的纹理粘连问题。在EG3D基础上,Next3D[42]进一步结合了显式网格的驱动先验与隐式表示的高质量渲染效果,提出"生成式纹理光栅化三平面"表示。该方法将FLAME网格光栅化到神经纹理特征平面上作为变形驱动信息,在保留体渲染处理复杂拓扑能力的同时,实现了对视线、眨眼及全头部姿态的精细控制。

为了提升动态控制的解耦性并解决非刚性区域的稳定性问题,3DFaceShop[43]提出了一种基于三平面混合表示与体积融合(Volume blending)的显式控制框架。面对动态编辑时头发与背景容易出现纹理闪烁的问题,3DFaceShop在学习辐射场的同时联合学习了一个3D语义场,将空间精确划分为面部与非面部区域。在推理阶段,系统通过语义场生成的3D掩码,显式地将包含动态表情的辐射场与包含静态背景的辐射场进行体积融合。这种基于物理空间的融合策略,结合3DMM的显式参数引导,不仅实现了对身份、表情、光照和姿态的全维度解耦控制,还从根本上保证了复杂动态场景下的视觉稳定性与几何一致性。

综上所述,从GRAM的流形采样到EG3D的三平面表征,再到AniFaceGAN的变形场驱动以及3DFaceShop的语义体积融合,这一系列工作勾勒出了一条清晰的技术演进脉络:即从隐式到混合表示的效率提升,从无条件生成到显式语义控制的维度扩展,以及从静态几何到动态物理一致性的深度优化。

2.3.2　与扩散模型相结合的方式

随着扩散模型(Diffusion models)的不断发展,相较于生成对抗网络,使用扩散模型与基本的三维表示(NeRF,3DGS)相结合的方法在生成的图像真实感,训练的稳定性等方面均有一定的提升。

针对三维人脸生成中计算成本与渲染分辨率大小的矛盾,Rodin[44]提出了基于三平面(Tri-plane)表示的"展开扩散网络"原理。该方法并未直接在计算密集的体素上进行扩散,而是将神经辐射场压缩分解为3个正交的二维特征平面,利用高效的二维卷积架构处理三维数据。为了解决降维过程中空间关联性丢失的问题,Rodin创新性地设计了3D感知卷积机制,通过显式的跨平面通信,强制模型根据三维空间投影关系同步更新特征,从而在二维处理中保留了三维结构的一致性归纳偏置。结合隐空间条件化策略,Rodin有效协调了全局几何生成质量,证明了扩散模型在构建高保真、语义可控的三维人脸方

面的潜力。

为了解决隐式表示生成的资产难以直接嵌入工业化创作的痛点，HeadEvolver[45]确立了基于显式网格变形(Mesh deformation)的生成原理。其核心在于利用雅可比场(Jacobian fields)作为中间表示来参数化网格变形，从而避免了直接预测顶点位移可能引发的网格自交和法线噪声问题，保证了生成网格的平滑性。为了增强模型在局部区域的表现力，HeadEvolver进一步引入了可学习的向量场(Vector fields)原理，允许对网格进行非各向同性的局部缩放，在保持旋转一致性的同时实现了夸张的几何特征表达。通过利用预训练的2D扩散模型作为先验引导变形，并结合人脸关键点与轮廓正则化约束，该方法成功生成了既保留原始拓扑结构与绑定属性，又具备丰富几何细节的风格化头像。

在追求静态高保真外观与动态灵活驱动的平衡中，HeadStudio[46]提出了一种将3DGS与3DMM(FLAME)相结合的混合表示原理。该方法将3D高斯点绑定于FLAME网格表面，利用FLAME模型提供基础的运动结构，赋予模型语义驱动能力；同时，利用高斯点作为“残差项”，补偿FLAME在精细纹理和几何细节(如头发、配饰)表达上的不足。为解决稀疏监督信号下的收敛问题，HeadStudio采用超密集高斯初始化与自适应几何正则化原理，根据网格面积动态调整约束力度，允许在特定区域生成超出基础网格表面的复杂几何结构。配合基于扩散模型的去噪分数蒸馏策略，该方法有效克服了纹理过平滑问题，实现了实时的高保真动态渲染。

为了降低对大规模3D训练数据的依赖，Zero-1-to-A[47]探索了利用视频扩散模型作为先验，实现单图零样本4D生成的原理。针对视频扩散模型生成的空间视角不一致和时间动作不连贯问题，该研究提出了“共生生成”(Symbiotic generation)原理，构建了可更新的伪真值数据集与3D头像重建之间的互惠循环。具体而言，利用3D头像渲染出几何一致的视频并提取控制信号，结合视频扩散模型的逆过程修正数据集帧，进而反哺头像优化。此外，该方法引入渐进式学习原理，将学习过程解耦为空间一致性学习和时间一致性学习两个阶段，从简单视角过渡到复杂表情，有效规避了噪声干扰，实现了从单张图像到高质量可驱动4D头像的稳健重建。

针对从单目视频重建高保真三维头像这一极具挑战性的任务，GAF[48]框架解决的是训练数据稀疏导致的视角缺失问题。通常情况下，智能手机拍摄的短视频缺乏侧面或极端视角的覆盖，导致纯基于3D重建的方法的结果往往在新视角下出现伪影。如图9所示，GAF的核心策略是利用3DGS作为显式场景表示，并引入一个多视角头部扩散模型作为生成先验来补全未观测区域。在具体实现上，GAF并未简单地使用相机参数作为条件，而是利用从FLAME模型重建得到的法线图作为扩散模型的引导，这种像素对齐的归纳偏置使得生成的图像能够与几何结构严格对齐。同时，为了保留面部的身份与外观细节，GAF在扩散过程中不仅使用了法线图，还引入了从输入图像中提取的变分自编码器(VAE)特征

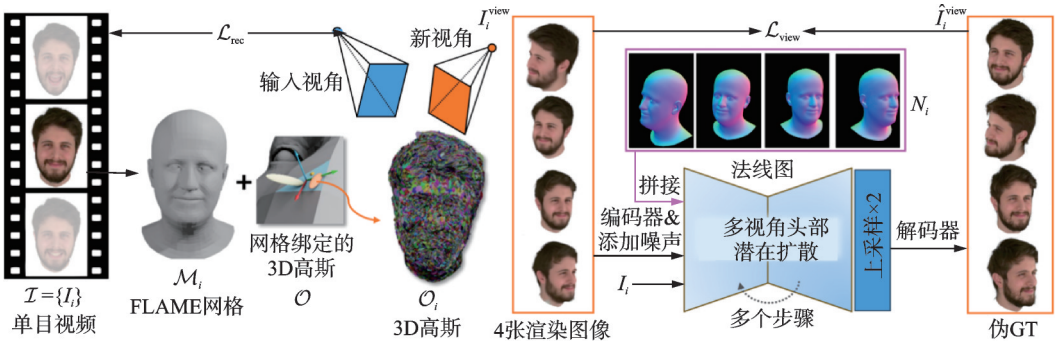

图9 GAF[48]的训练流程

Fig.9 Training pipeline of GAF[48]

作为条件。在训练策略上,GAF摒弃了常导致纹理过饱和或平滑的单步分数蒸馏采样(SDS)损失,转而采用一种迭代去噪策略:将多视角扩散模型生成的去噪图像作为“伪真值”(Pseudo ground truths)来监督3D高斯基元的优化。此外,为了进一步提升渲染的逼真度,GAF引入了一个潜在上采样器(Latent up sampler),在解码前对低分辨率的潜在特征进行超分辨率处理,从而生成更加锐利的面部细节。实验证明,这种结合了显式几何表示与多视角生成先验的方法,能够有效处理手持设备拍摄的非受控视频,重建出视角一致且纹理细腻的动态头像。表2展示了主要方法的对比分析。

表2　生成式模型与3D基本表征相结合的主要方法对比

Table 2　Comparison of primary methods combining generative models and 3D base representations

方法	训练/输入数据	生成质量(FFHQ)(FID↓)	主要特点
GRAM[36]	单类图像(256^2)	17.9	流形约束采样解决噪声,发丝/牙齿细节逼真
Next3D[42]	单类图像(512^2)	3.9	3DMM先验结合神经纹理三平面,支持高保真表情驱动
EG3D[37]	单类图像(512^2)	4.7	三平面表示大幅提升生成分辨率与渲染效率
AniFaceGAN[41]	单类图像(256^2)	19.9	仅靠2D图像实现身份/表情解耦与动态驱动
Mimic3D[39]	单类图像	5.37	通过模仿策略去除2D伪影,实现严格的3D一致性
HeadStudio[46]	零样本(文本)		无需微调,利用分数蒸馏先验实现任意人物新视角合成
Zero-1-to-A[47]	单张图像		利用视频扩散模型实现从单图到动态4D头像的转换
Rodin[44]	3D扫描/多视角		3D卷积保留空间结构,适合高质量数字资产创建
GAF[48]	单目视频		利用扩散模型先验解决手持拍摄视频的视角稀疏问题

尽管上述基于扩散模型的方法在面部区域取得了显著进展,但它们多依赖于传统的参数化模型(如FLAME),这导致其在全头部结构(如颅骨、颈部)、口腔内部以及复杂发型的物理建模上仍显不足,特别是头发的动态仿真与高光渲染,成为制约真实感进一步提升的瓶颈。为突破这一限制,最新的研究开始探索支持局部精细编辑的大规模隐式模型及具备强生成先验的复杂毛发建模技术。在全头部结构建模方面,针对传统3DMM难以捕捉高频几何细节且拓扑固定的缺陷,ImHead[49]提出了一种大规模隐式可变形模型。该方法利用身份分解网络(DecNet)将整个头部对应的全局潜在空间划分为局部区域嵌入编码,实现了对头发、鼻子、嘴部等特定区域的独立平滑编辑,有效避免了复杂头部形状在参数化建模中的拓扑对齐难题。针对稀疏视角下的重建难题,SIRA++[50]框架展示了如何利用概率性的形状与外观先验解决全头部重建的歧义性,通过将有符号距离函数场(Signed distance function, SDF)解耦为参考场与变形场,在仅有单张视角输入时即可精准重建包含面部、发型及肩部在内的完整三维结构。针对建模中最具挑战性的非刚性毛发结构,DiffLocks[51]利用扩散Transformer(Diffusion Transformer, DIT)模型直接预测头皮纹理映射中的发丝潜码,首次实现了对非洲式卷发(Afro-like)等极复杂发型的细粒度重建,并能直接导入实时引擎进行高保真物理仿真。这种从面部生成向全头部物理仿真的转变,显著提升了数字人的身份特征保持与视觉真实感,这些方法可以为生成模型提供一种更完整的,可编辑性更强的三维头部几何先验,扩散模型与这些模型相结合的方法将会是一个新的主流研究方向。

综上所述,这一系列工作清晰地展示了三维人脸生成技术通过优化几何表示与利用多模态先验,逐步实现更高保真度、更强兼容性及更低数据门槛的发展逻辑。

3　三维人脸生成应用

在第2节中已经详细介绍了三维人脸生成技术的发展脉络,可以了解到目前的三维人脸生成技术

已经可以生成具有“照片级”真实感的多视角人脸图像，则其已经具备了发展一些下游应用如三维人脸属性编辑（3D face attribute editing）和说话人脸生成（Talking head generation）的基本能力，本节将详细讨论这两大应用场景相关技术的最新进展。

3.1 三维人脸属性编辑

3D人脸编辑技术已从线性统计模型演进至结合深度学习、NeRF及3DGS的新阶段。尽管3DMM已经可以实现初步的属性编辑，但其线性本质导致高频细节（如皱纹）缺失，渲染的图像难以达到照片级真实感。此外，它在极端表情、非线性形变及头发口腔等区域建模上也表现不佳。为此，研究者转向结合2D生成模型，旨在利用几何先验的同时增强视觉真实感。

针对3DMM生成图像纹理细节缺失的问题，ClipFace[52]结合StyleGAN2[38]与CLIP[53]，实现了高保真纹理生成及文本驱动编辑。该方法创新性地将StyleGAN2特征映射至UV空间，利用可微渲染在无3D真值条件下进行对抗训练。在编辑方面，通过CLIP引导预测潜在空间与表情参数的偏移，并利用方向性CLIP损失确保了如“僵尸妆”等风格化编辑的一致性。尽管ClipFace显著提升了纹理质量，但受限于显式网格的表征方式，仍难以处理头发和背景，这促使后续研究转向结合3DMM与隐式NeRF的混合架构。

为突破传统限制，Exp-GAN[54]提出了一种具备显式表情控制的3D感知混合架构。该模型解耦了人脸与非人脸区域：人脸部分利用DECA[16]回归FLAME[11]参数生成显式几何代理，结合神经纹理实现细粒度表情控制；头发与背景等非刚性区域则通过基于EG3D[37]的神经体素生成器维持多视角一致性。核心在于引入基于深度的特征融合机制，在体渲染积分过程中利用3DMM深度图显式整合面部特征，从而兼顾了3D一致性与参数化的精确控制。

随着3D GAN技术的成熟，跨域风格化成为新的研究热点。如图10所示，3DAvatarGAN[55]聚焦于跨域风格化（真实至艺术）迁移。针对艺术域数据相机参数缺失与几何拓扑结构夸张的挑战，该方法提出一个域适应框架：通过对齐源域与目标域的平均人脸投影自动估计相机分布，并利用深度与几何正则化防止结构退化，其核心创新在于EG3D[37]的三平面空间引入薄板样条（Thin plate spline, TPS）变形模块，通过非线性扭曲实现了夸张的几何结构生成并可保持多视角一致性。此后，研究焦点转向利用CLIP挖掘潜在空间，以探索自然语言驱动的细粒度属性控制。

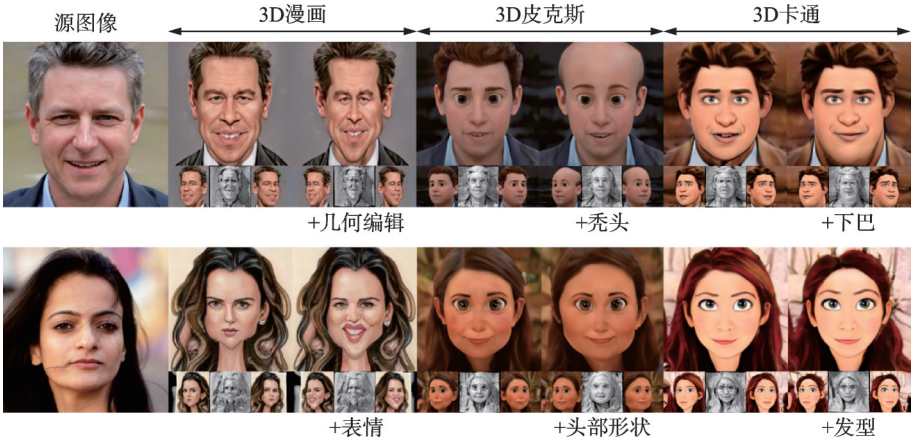

图10 3DAvatarGAN[55]的编辑效果

Fig.10 Editing results of 3DAvatarGAN[55]

TG-3DFace[56]展示了仅利用2D图像-文本对进行文本驱动3D生成的能力。该方法基于EG3D[37]架构,将CLIP文本嵌入注入StyleGAN2映射网络以控制三平面特征。为实现局部(如“蓝眼睛”)的精确控制,设计了结合人脸解析图与注意力机制的细粒度对齐模块。此外,在推理阶段引入CLIP方向引导策略微调生成器,显著增强了生成结果与复杂文本描述的语义一致性。

针对连续编辑(如先变老再戴眼镜)中“灾难性遗忘”的问题,FaceG2E[57]提出了“生成+编辑”的一体化框架。在生成阶段,采用几何-纹理解耦策略,先利用分数蒸馏损失优化几何细节,然后使用一个ControlNet引导纹理生成。其能处理灾难性遗忘的核心在于编辑阶段的自导向一致性保留机制:该机制利用InstructPix2Pix[58]的注意力图构建UV掩码,对非编辑区域施加严格正则化,从而在允许目标区域自由变化的同时,有效维持了多轮操作中非编辑区域的稳定性。

针对动态4D场景编辑中单一潜在代码导致的“关联局部属性”问题,FaceCLIPNeRF[59]提出了一种基于可变形NeRF的方案。该方法引入位置条件锚点合成器(Position-conditional anchor compositor,PAC),摒弃了全局控制策略,转而通过预测空间中每一点如何组合预学习的“锚点代码”,生成空间解耦的变形潜在代码。这种机制实现了局部形变的解耦控制,结合CLIP引导优化,成功完成了对动态视频中复杂表情的文本引导编辑。

针对动态NeRF的效率瓶颈,Control4D[60]引入3DGS并提出“Gaussian Planes”表示。该方法通过将高斯属性分解为三平面特征,将运动流分解为4D平面特征,以结构化方式解决了离散点云的噪声问题。针对2D扩散模型编辑带来的时间不一致性(抖动),Control4D并未直接优化高斯点,而是训练GAN生成器,利用对抗训练学习平滑流形,从而有效过滤高频噪声,实现了高保真且时空一致的4D编辑。

综上所述,3D人脸编辑技术正经历从传统的显式网格模型向隐式神经场及显隐式结合的高效表示演进,驱动方式也从参数化控制转向更自然的文本语义控制。未来的研究将继续致力于解决整个头部结构的完整性建模、实时交互编辑的效率问题以及对极度夸张形变的拓扑适应性,推动三维人脸创作与虚拟交互技术的进一步发展。表3展示了三维人脸属性编辑主要方法的对比分析。

表3 三维人脸属性编辑主要方法对比

Table 3 Comparison of key techniques in 3D face attribute editing

方法	核心架构	编辑驱动方式	核心特点
3DFaceShop[43]	体积融合(Volume blending)	3DMM参数	物理空间融合解决动态编辑时的纹理闪烁,解耦能力强
DiffusionRig[32]	Diffusion+3DMM特征图	3DMM参数	通过法线/反照率特征图,精确控制光照与姿态
ClipFace[52]	UV映射+CLIP	文本	在UV空间生成纹理,支持“僵尸妆”等风格化生成
Control4D[60]	Gaussian Planes (4D)	文本	高斯属性分解为4D平面,实现视频时空一致编辑
FENeRF[40]	语义神经辐射场	语义掩码	通过修改2D语义图反向控制3D几何与纹理
TG-3DFace[56]	细粒度文本对齐	文本	精确修改“蓝眼睛”等局部属性而不破坏整体结构
FaceG2E[57]	自导向一致性	文本	支持多轮编辑(如先变老再戴眼镜)而不遗忘特征
Faceclipnerf[59]	可变形NeRF+PAC	文本	引入条件锚点合成器,实现动态视频中的局部变形控制

3.2 说话人脸生成

随着计算机视觉与图形学技术的飞速发展,说话人脸生成(Talking head generation)的研究重心已逐步从早期的2D图像序列生成向具备高时空一致性、高保真度及实时交互能力的3D生成范式转变。

这一领域的演进脉络清晰地展现了从神经辐射场到三平面混合表示，再到结合扩散模型以及最新的3DGS的技术迭代路径。

早期的研究主要致力于解决音频驱动下的口型同步与个性化属性解耦问题，代表性工作如DFA-NeRF[61]提出了一种基于属性解耦的神经渲染框架。鉴于语音与唇部运动强相关而与头部姿态、眨眼弱相关，该方法利用Transformer-VAE结合高斯过程显式解耦这些属性，以生成自然连贯的动作序列。同时，引入对比学习策略对齐音视特征，并将解耦的特征输入动态NeRF进行渲染。尽管该方法在口型准确性与个性化风格上表现优异，但纯NeRF架构导致其训练推理效率低下，且依赖特定身份训练限制了其泛化能力。

为了克服NeRF在单样本(One-shot)场景下的泛化与效率瓶颈，研究者们开始探索基于三平面混合表示的生成方法，OTAvatar[62]利用EG3D[37]先验，通过“解耦反演”策略交替优化身份与运动代码，可实现从单图重建一个可驱动的3D头像，渲染速度达到35 fps。针对自监督训练的视角不一致问题，Portrait4D[63]提出合成数据强监督范式。该方法先生成大规模4D合成数据，然后训练一个Transformer重建器直接回归三平面特征。此外，通过解耦学习策略(随机屏蔽运动模块)弥补域差距，大幅提升了对真实照片的泛化能力与生成的3D头像的几何稳定性。

针对3D GAN/NeRF在高频细节(如头发、牙齿)上的不足，DiffusionAvatars[64]提出“延迟扩散”机制。该方法以NPHM[35]为几何表征方法，将几何特征光栅化作为ControlNet的条件，引导Stable Diffusion[31]生成高保真纹理。为捕捉皱纹等微细表情，进一步通过交叉注意力将表情编码注入U-Net，实现了对细微表情的精确控制。尽管该方法在视觉真实感上表现卓越，但依赖扩散模型的去噪过程导致推理缓慢，无法满足实时交互需求。

在高效三维人脸生成领域，3DGS的出现为解决渲染质量与计算效率之间的长期矛盾提供了新的技术路径。研究者们首先在特定身份的超高性能渲染上取得了突破：FlashAvatar[65]通过将3D高斯基元嵌入到FLAME网格表面，利用网格的刚性运动驱动整体姿态，并结合轻量级偏移网络建模发丝与皱纹等高频细节，成功实现了300 fps的高速渲染；3D Gaussian blendshapes[66]则进一步将传统的混合蒙皮(Blendshapes)概念引入线性混合高斯表示，通过叠加中性与差分基底生成表情，在达成370 fps峰值帧率的同时，实现了与工业界动画管线的无缝兼容。

随后，研究的重心开始从特定身份的建模向单图泛化的通用重建演进，旨在摆脱对长视频训练数据的依赖。GAGAvatar[67]针对单张图片难以估计完整3D结构的挑战，创新地提出了“双向提升(Dual-lifting)”策略以构建闭合的3D高斯壳体。在驱动层，该方法利用FLAME点云构建表情场，并通过多三平面注意力(Multi tri-plane attention, MTA)模块融合多视角特征以补全遮挡细节，成功在保持67 fps实时性的前提下，实现了对未见身份的高质量重建。其生成效果如图11所示。

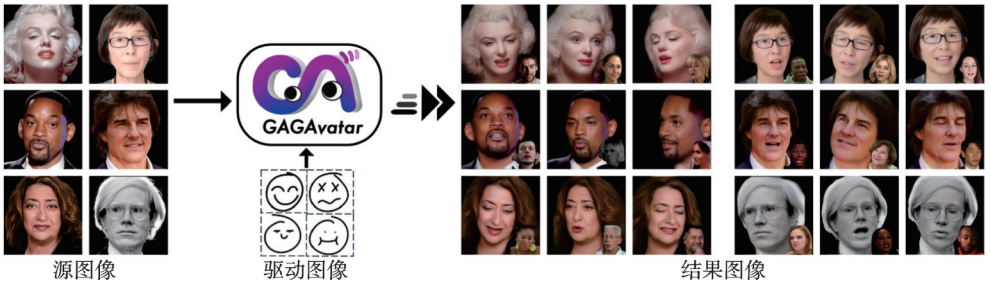

图11 GAGAvatar[67]的生成效果

Fig.11 Generation results of GAGAvatar[67]

尽管算法层面的效率与泛化性已取得显著进展,但如何在算力受限的移动端及增强现实(AR)设备上实现高保真、低延迟的工程化落地,已成为衡量该技术成熟度的核心指标,同时也代表了该领域一个极具挑战性的未来重要研究方向。为了克服复杂非刚性变形在移动设备上的巨大计算开销,TaoAvatar[68]提出了一种“教师-学生”训练框架,利用知识蒸馏技术将基于复杂的 StyleUnet结构的教师网络捕捉到的高频变形细节“烘焙”(Baking)进紧凑的MLP学生网络中。通过结合 FP16量化、异步推理及UInt16高斯排序优化,该方案在 Apple Vision Pro设备上实现了 2K分辨率下 90 fps的立体渲染性能。针对扩散模型推理速度慢且难以实现流式生成的痛点,Teller[69]提出了首个实时流式自回归生成框架。该模型通过残差矢量量化(Residual vector quantization, RVQ)将面部运动隐变量编码为离散Token,并在第一阶段利用自回归Transformer以 200 ms的音频块为粒度进行流式处理,实现了低延迟的音频到动作映射。此外,为了解决单步生成中的物理一致性问题(如耳环摆动、颈部肌肉拉伸),Teller引入了基于3D U-Net的高效时序模块(Efficient temporal module, ETM),在无需昂贵迭代扩散过程的情况下捕捉精细的时序依赖,最终在保持高保真细节的同时实现了高达 25 fps的实时生成速度。针对移动端神经头像生成中现有模型计算量过大(通常超过 100 GFLOPs)且结构复杂的难题,MobilePortrait[70]则从“外部知识注入”的角度突破,设计了基于标准轻量级U-Net的高效架构。在运动建模层面,该方法引入了混合关键点(Mixed keypoints)表征,通过融合显式面部关键点与隐式神经关键点,在大幅削减计算量的同时有效抑制了背景“液化”伪影;在图像合成层面,模型采用“开卷考试”策略,利用预计算的伪多视图特征和伪背景作为外观先验辅助生成,从而摒弃了复杂的动态卷积与注意力模块,使得模型在iPhone 14 Pro上以仅 16 GFLOPs的极低算力消耗实现了约 60 fps的实时渲染。

综上所述,三维说话人脸生成技术正经历着从神经辐射场(NeRF)到高斯泼溅(3DGS)、从特定人物建模到通用泛化、从云端离线渲染到移动端实时交互的深刻演进。这一系列研究表明,未来的发展将更加注重生成模型先验知识与显式几何控制力的深度融合,通过算法与系统的协同设计,在单图输入的极简条件下,为用户提供兼具物理真实感与超高画质的移动端实时交互体验。表4展示了主要的说话人脸生成方法对比分析。

表4　说话人脸生成主要方法对比

Table 4　Comparative analysis of primary techniques for talking head generation

方法	核心架构	评价指标(PSNR↑/LPIPS↓)	主要特点
GAGAvatar[67]	3DGS (Dual-lifting)	(VFHQ数据集[71]) 21.83/0.122	单图泛化任务中质量最佳,且支持实时渲染,解决了几何闭合问题
Portrait4D[63]	Tri-plane＋Synthetic	(VFHQ数据集[71]) 20.35/0.191	利用合成数据强监督,对夸张表情和头部姿态的适应性强
OTAvatar[62]	Tri-plane (Optimization)	(VFHQ数据集[71]) 17.65/0.294	验证了Tri-plane实现单图驱动的可行性,但生成质量不高
Gaussian blendshapes [66]	3DGS Blendshapes	(INSTA数据集[72]) 33.34/0.052	线性混合模型,极易集成到现有引擎,目前速度最快
FlashAvatar [65]	3DGS (Mesh Embed)	(INSTA数据集[72]) 32.33/0.032	网格嵌入表面,兼顾超高帧率与发丝、皱纹等高频细节捕捉
Diffusion Avatars[64]	Diffusion＋NPHM	(NeRSemble[21]数据集) 24.9/0.081	利用扩散模型生成极致的皮肤纹理,解决了几何模型平滑问题
DFA-NeRF[61]	Dynamic NeRF	—	专注于音频驱动下的属性解耦(如眨眼、头部姿态与口型的分离)

4 未来发展与挑战

尽管三维人脸生成技术在表征效率与生成质量上取得了显著突破,但要实现真正普及化、低门槛且物理真实的全头部生成,仍面临诸多挑战:

(1)小样本与单视角的泛化。当前高质量的三维人脸生成往往依赖于多视角图像或特定身份的视频数据。虽然基于扩散模型的方法(如 Zero-1-to-A[47], GAGAvatar[68])在单图零样本生成上取得了进展,但在处理极端视角、遮挡以及未见过的复杂表情时,几何结构与纹理的一致性仍难以保证。未来需要进一步探索如何更有效地利用大规模预训练模型的2D先验知识,提升模型在极度稀疏数据下的泛化能力与鲁棒性。

(2)高保真与实时交互的统一。基于NeRF的方法难以满足实时性要求,而3DGS虽然实现了高帧率渲染,但在结合扩散模型进行生成时,推理速度仍受限于扩散模型的去噪过程。此外,在移动端等算力受限设备上部署高保真三维模型仍存在困难。未来的研究将致力于轻量化网络设计、高效的神经渲染流程优化以及端云协同计算,以实现消费级设备上的实时交互。

(3)全头部与精细结构的完整建模。现有的参数化模型(如FLAME)和生成方法多集中于面部区域,对于口腔内部(牙齿、舌头)、眼球动态以及复杂发型的物理建模仍显不足。特别是头发的动态仿真与高光渲染,仍是制约真实感的瓶颈。未来趋势将是发展解耦的混合表征,针对皮肤、毛发、眼球等不同材质采用专用的建模与渲染策略,实现全头部的物理一致性仿真。

(4)物理感知与语义解耦的精细控制。虽然目前的文本驱动编辑已实现了一定的语义控制,但缺乏对人脸解剖学结构和物理规律的深层理解。生成的动态往往缺乏重力、肌肉弹性等物理属性的约束。未来的生成模型需要更深入地结合生物力学模型与物理引擎,实现不仅视觉逼真,且符合生理规律的精细化表情与动作控制。

5 结束语

本文系统梳理了三维人脸生成技术的发展脉络,并对其核心算法架构与前沿应用场景进行了深入剖析。纵观现有研究成果,三维人脸生成技术的发展并非单一维度的线性迭代,而是在“高保真度(Fidelity)、可控性(Controllability)与实时性(Efficiency)”的多目标约束下,寻求最优解的螺旋式上升过程。

底层表征呈现出从“显式参数化”向“隐式神经化”演进,进而向“结构化高效表征”理性回归的辩证规律。早期的三维可变形模型(3DMM)奠定了基于拓扑一致性的语义解耦基础,但受限于线性子空间的低频表达能力,生成的图像质量较低。神经辐射场(NeRF)的引入通过连续的隐式场建模丰富的高频细节,实现了照片级的视觉合成,却引入了训练与渲染的高昂计算成本。当前最新的3D高斯泼溅(3DGS)技术,通过离散几何基元与可微光栅化的结合,在显式表征的可编辑性与隐式渲染的高质量之间建立了新的平衡,标志着领域研究从纯粹的视觉效果追求,转向了兼顾计算效率与工业化部署的务实探索。

生成范式已完成从“单一几何重建”向“多模态先验驱动的混合生成”的根本性范式转移。针对单目视角下三维重建的不适定性(Ill-posed)问题,现阶段的主流方法确立了“参数化模型引导几何、生成式模型补全纹理、神经渲染提升质感”的混合技术路线。这种范式有效利用了大规模预训练生成模型(如Diffusion models)中蕴含的海量二维视觉先验,与3DMM的物理结构先验相结合,突破了传统方法在小样本泛化与高频细节生成上的瓶颈,实现了在数据稀缺条件下的高保真三维一致性生成。

尽管现有技术已取得显著进展,但实现全物理感知的数字化身(Digital avatar)仍面临本质挑战,未来的研究亟须在以下3个维度寻求突破:

(1)从“视觉外观仿真”向“生物物理仿真”深化:突破仅基于表面的几何形变,引入解剖学与完整头

部建模约束。重点解决口腔内部结构、眼球微动及毛发动力学的物理一致性三维建模,以实现符合生理规律的动态交互。

(2)从“特定场景优化”向“通用基础模型”跃迁:克服现有方法对特定身份或多视角数据的依赖,探索构建通用的三维人脸基础模型(Foundation model)。利用大规模视频数据学习跨身份、跨表情的通用表征,提升模型在零样本(Zero-shot)与极端视角下的鲁棒性与泛化能力。

(3)从“离线高算力生成”向“端侧实时交互”下沉:针对移动端与XR设备的算力约束,研究轻量化网络架构与高效神经渲染算法。通过模型压缩、知识蒸馏及端云协同计算策略等解决高保真模型在消费级设备上的实时推理与低延迟驱动难题。

综上所述,三维人脸生成技术正处于从视觉层面的“复刻”向物理层面的“智能体构建”跨越的关键阶段。随着混合表征学习与生成式人工智能技术的深度融合,未来的研究将致力于构建具备物理属性完备性、语义理解深度化及交互响应实时化的三维数字人,为元宇宙、人机交互及虚拟现实应用提供坚实的技术支撑。

参考文献:

[1] 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. [S.l.]: ACM, 1999: 187-194.

[2] MILDENHALL B, PRATUL P P, TANCIK M, et al. NeRF: Representing scenes as neural radiance fields for view synthesis [J]. Communications of the ACM, 2021, 65(1): 99-106.

[3] KERBL B, KOPANAS G, LEIMKÜHLER T, et al. 3D Gaussian splatting for real-time radiance field rendering[J]. ACM Transactions on Graphics, 2023, 42(4): 139.

[4] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[5] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.

[6] EGGER B, SMITH WA, TEWARI A, et al. 3D morphable face models—Past, present, and future[J]. ACM Transactions on Graphics (TOG), 2020, 39(5): 157.

[7] PAYSAN P, KNOTHE R, AMBERG B, et al. A 3D face model for pose and illumination invariant face recognition[C]//Proceedings of 2009 sixth IEEE International Conference on Advanced Video and Signal Based Surveillance. [S.l.]: IEEE, 2009: 296-301.

[8] GERIG T, MOREL-FORSTER A, BLUMER C, et al. Morphable face models——An open framework[C]//Proceedings of 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). [S.l.]: IEEE, 2018: 75-82.

[9] CAO C, WENG Y, ZHOU S, et al. FaceWarehouse: A 3D facial expression database for visual computing[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(3): 413-425.

[10] BOOTH J, ROUSSOS A, PONNIAH A, et al. Large scale 3D morphable models[J]. International Journal of Computer Vision, 2018, 126(2): 233-254.

[11] LI T, BOLKART T, BLACK M J, et al. Learning a model of facial shape and expression from 4D scans[J]. ACM Transactions on Graphics (TOG), 2017, 36(6): 194.

[12] RANJAN A, BOLKART T, SANYAL S, et al. Generating 3D faces using convolutional mesh autoencoders[C]//Proceedings of the European Conference on Computer Vision (ECCV). Berlin, Heidelberg: Springer, 2018: 704-720.

[13] BARRON JT, MILDENHALL B, VERBIN D, et al. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2022: 5470-5479.

[14] MÜLLER T, EVANS A, SCHIED C, et al. Instant neural graphics primitives with a multiresolution hash encoding[J]. ACM Transactions on Graphics (TOG), 2022, 41(4): 102.

[15] 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/CVF Conference on Computer Vision and Pattern Recognition workshops. [S.l.]: IEEE, 2019.

[16] 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 (TOG), 2021, 40(4): 88.

[17] DANĚČEK R, BLACK M J, BOLKART T. EMOCA: Emotion driven monocular face capture and animation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2022: 20311-20322.

[18] ZIELONKA W, BOLKART T, THIES J. Towards metrical reconstruction of human faces[C]//Proceedings of the European Conference on Computer Vision (ECCV). Berlin, Heidelberg: Springer, 2022: 250-269.

[19] HONG Y, PENG B, XIAO H, et al. HeadNeRF: A real-time NeRF-based parametric head model[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2022: 20374-20384.

[20] GAO X, ZHONG C, XIANG J, et al. Reconstructing personalized semantic facial NeRF models from monocular video[J]. ACM Transactions on Graphics (TOG), 2022, 41(6): 1-12.

[21] KIRSCHSTEIN T, QIAN S, GIEBENHAIN S, et al. NeRSemble: Multi-view radiance field reconstruction of human heads [J]. ACM Transactions on Graphics (TOG), 2023, 42(4): 1-14.

[22] QIAN S, KIRSCHSTEIN T, SCHONEVELD L, et al. GaussianAvatars: Photorealistic head avatars with rigged 3D Gaussians[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2024: 20299-20309.

[23] SHAO Z, WANG Z, LI Z, et al. SplattingAvatar: Realistic real-time human avatars with mesh-embedded Gaussian splatting [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2024: 1606-1616.

[24] SHEN J, CASHMAN T J, YE Q, et al. The phong surface: Efficient 3D model fitting using lifted optimization[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: [s.n.], 2020: 687-703.

[25] ZHAO Z, BAO Z, LI Q, et al. PSAvatar: A point-based shape model for real-time head avatar animation with 3D Gaussian splatting[EB/OL]. (2024-01-23). https://doi.org/10.48550/arXiv.2401.12900.

[26] XU Y, CHEN B, LI Z, et al. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic Gaussians[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2024: 1931-1941.

[27] CHEN Y, WANG L, LI Q, et al. MonoGaussianAvatar: Monocular Gaussian point-based head avatar[C]//Proceedings of ACM SIGGRAPH 2024 Conference. New York, NY, USA: ACM, 2024: 58.

[28] DENG Y, YANG J, CHEN D, et al. Disentangled and controllable face image generation via 3D imitative-contrastive learning [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 5154-5163.

[29] LIU Y, SHU Z, LI Y, et al. 3D-FM GAN: Towards 3D-controllable face manipulation[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: [s.n.], 2022: 107-125.

[30] GU Y, XU H, XIE Y, et al. DiffPortrait3D: Controllable diffusion for zero-shot portrait view synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2024: 10456-10465.

[31] STABILITY AI. Stable diffusion v1.5 model card. https://huggingface.co/runwayml/stable-diffusion-v1-5, 2022. 2, 4

[32] DING Z, ZHANG X, XIA Z, et al. DiffusionRig: Learning personalized priors for facial appearance editing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2023: 12736-12746.

[33] KARRAS T, LAINE S, AND AILA T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 4401-4410.

[34] CHEN X, MIHAJLOVIC M, WANG S, et al. Morphable diffusion: 3D-consistent diffusion for single-image avatar creation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2024: 10359-10370.

[35] GIEBENHAIN S, KIRSCHSTEIN T, GEORGOPOULOS M, et al. Learning neural parametric head models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2023: 21003-21012.

[36] DENG Y, YANG J, XIANG J, et al. GRAM: Generative radiance manifolds for 3D-aware image generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2022: 10673-10683.

[37] CHAN E R, LIN C Z, CHAN M A. et al. Efficient geometry-aware 3D generative adversarial networks[C]//Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2022: 16123-16133.

[38] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of stylegan[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 8110-8119.

[39] CHEN X, DENG Y, WANG B . Mimic3D: Thriving 3D-aware GANs via 3D-to-2D imitation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2023: 2338-2348.

[40] SUN J, WANG X, ZHANG Y, et al. FENeRF: Face editing in neural radiance fields[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2022: 7672-7682.

[41] WU Y, DENG Y, YANG J, et al. AniFaceGAN: Animatable 3D-aware face image generation for video avatars[J]. Advances in Neural Information Processing Systems, 2022, 35: 36188-36201.

[42] SUN J, WANG X, WANG L, et al. Next3D: Generative neural texture rasterization for 3D-aware head avatars[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2023: 20991-21002.

[43] TANG J, ZHANG B, YANG B, et al. 3DFaceShop: Explicitly controllable 3D-aware portrait generation[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 30(9): 6020-6037.

[44] WANG T, ZHANG B, ZHANG T, et al. Rodin: A generative model for sculpting 3D digital avatars using diffusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2023: 4563-4573.

[45] WANG D, MENG H, CAI Z, et al. HeadEvolver: Text to head avatars via expressive and attribute-preserving mesh deformation[C]//Proceedings of 2025 International Conference on 3D Vision (3DV). Singapore: IEEE, 2025: 211-221.

[46] ZHOU Z, MA F, FAN H, et al. HeadStudio: Text to animatable head avatars with 3d Gaussian splatting[C]//Proceedings of the European Conference on Computer Vision (ECCV). Berlin, Heidelberg: Springer, 2024: 145-163.

[47] ZHOU Z, MA F, FAN H, et al. Zero-1-to-A: Zero-shot one image to animatable head avatars using video diffusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2025: 15941-15952.

[48] TANG J, DAVOLI D, KIRSCHSTEIN T, et al. GAF: Gaussian avatar reconstruction from monocular videos via multi-view diffusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2025: 5546-5558.

[49] POTAMIAS RA, GALANAKIS S, DENG J, et al. ImHead: A large-scale implicit morphable model for localized head modeling [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2025: 10196-10206.

[50] CASELLES P, RAMON E, GARCIA J, et al. Implicit shape and appearance priors for few-shot full head reconstruction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(5): 3691-3705.

[51] ROSU R A, WU K, FENG Y, et al. DiffLocks: Generating 3D hair from a single image using diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2025: 10847-10857.

[52] ANEJA S, THIES J, DAI A, et al. ClipFace: Text-guided editing of textured 3D morphable models[C]//Proceedings of ACM SIGGRAPH 2023 Conference. [S.l.]: ACM, 2023: 1-11.

[53] RADFORD A, KIM JW, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of International Conference on Machine Learning. Vienna, Austria: PMLR, 2021: 8748-8763.

[54] LEE Y, CHOI T, GO H, et al. Exp-GAN: 3D-aware facial image generation with expression control[C]//Proceedings of the Asian Conference on Computer Vision. [S.l.]: IEEE, 2022: 3812-3827.

[55] ABDAL R, LEE HY, ZHU P, et al. 3DAvatarGAN: Bridging domains for personalized editable avatars[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2023: 4552-4562.

[56] YU C, LU G, ZENG Y, et al. Towards high-fidelity text-guided 3D face generation and manipulation using only images[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2023: 15326-15337.

[57] WU Y, MENG Y, HU Z, et al. Text-guided 3D face synthesis-from generation to editing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2024: 1260-1269.

[58] BROOKS T, HOLYNSKI A, EFROS A A. InstructPix2Pix: Learning to follow image editing instructions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2023: 18392-18402.

[59] HWANG S, HYUNG J, KIM D, et al. FaceCLIPNeRF: Text-driven 3D face manipulation using deformable neural radiance fields[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2023: 3469-3479.

[60] SHAO R, SUN J, PENG C, et al. Control4D: Efficient 4D portrait editing with text[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2024: 4556-4567.

[61] YAO S, ZHONG R, YAN Y, et al. DFA-NeRF: Personalized talking head generation via disentangled face attributes neural rendering[EB/OL]. (2022-01-03). https://doi.org/10.48550/arXiv.2201.00791.

[62] MA Z, ZHU X, QI G J, et al. OTAvatar: One-shot talking face avatar with controllable tri-plane rendering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2023: 16901-16910.

[63] DENG Y, WANG D, REN X, et al. Portrait4D: Learning one-shot 4D head avatar synthesis using synthetic data[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2024: 7119-7130.

[64] KIRSCHSTEIN T, GIEBENHAIN S, NIEßNER M. DiffusionAvatars: Deferred diffusion for high-fidelity 3D head avatars [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2024: 5481-5492.

[65] XIANG J, GAO X, GUO Y, et al. FlashAvatar: High-fidelity head avatar with efficient Gaussian embedding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2024: 1802-1812.

[66] MA S, WENG Y, SHAO T, et al. 3D Gaussian blendshapes for head avatar animation[C]//Proceedings of ACM SIGGRAPH 2024 Conference. [S.l.]: ACM, 2024: 1-10.

[67] CHU X, AND HARADA T. Generalizable and animatable Gaussian head avatar[J]. Advances in Neural Information Processing Systems, 2024, 37: 57642-57670.

[68] CHEN J, HU J, WANG G, et al. TaoAvatar: Real-time lifelike full-body talking avatars for augmented reality via 3D Gaussian splatting[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2025: 10723-10734.

[69] ZHEN D, YIN S, QIN S, et al. Teller: Real-time streaming audio-driven portrait animation with autoregressive motion generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2025: 21075-21085.

[70] JIANG J, LIN G, RONG Z, et al. MobilePortrait: Real-time one-shot neural head avatars on mobile devices[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2025: 15920-15929.

[71] XIE L, WANG X, ZHANG H, et al. VFHQ: A high-quality dataset and benchmark for video face super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2022: 657-666.

[72] XU Y, ZHANG H, WANG L, et al. LatentAvatar: Learning latent expression code for expressive neural head avatar[C]//Proceedings of ACM SIGGRAPH 2023 Conference. [S.l.]: ACM, 2023: 1-10.

作者简介:

王伟(1990-),男,教授,博士生导师,研究方向:计算机视觉、机器学习,E-mail:wei.wang@bjtu.edu.cn。

何一康(2001-),男,硕士研究生,研究方向:计算机视觉。

魏云超(1987-),男,教授,博士生导师,研究方向:计算机视觉、机器学习。

赵耀(1967-),**通信作者**,男,教授,博士生导师,研究方向:数字媒体信息处理与智能分析、人工智能、计算机视觉、AIGC、AI视频编码,E-mail:yzhao@bjtu.edu.cn。

(编辑:王静)

A Survey on 3D Face Generation Technology

WANG Wei[1,2], HE Yikang[1,2], WEI Yunchao[1,2], ZHAO Yao[1,2*]

(1. Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China; 2. Visual Intelligence + X International Cooperation Joint Laboratory of the Ministry of Education, Beijing 100044, China)

Abstract: In recent years, benefiting from the rapid development of computer vision and graphics, 3D face generation technology has achieved significant breakthroughs; 3D vision technologies, such as digital avatar creation, have become increasingly popular on the internet, attracting extensive attention from both academia and industry. This generation technology synthesizes realistic multi-view face images by reconstructing geometric structures and texture details from explicit or implicit underlying representations. 3D face generation technology has sparked many related entertainment and interactive applications, such as using attribute editing technology to modify facial features via text descriptions, or using talking head generation technology to drive a static portrait to generate a talking video. However, early technologies based on linear parametric models suffered from poor realism and detail performance. And the subsequently emerging implicit neural representation technologies, while significantly improving visual quality, face the challenges of high computational costs and difficulty in achieving real-time interaction, which have brought great limitations to practical deployment and application. In order to overcome the contradiction between speed and quality, numerous scholars have conducted in-depth research on novel representations based on explicit Gaussian primitives and generative models based on probabilistic diffusion, and have proposed a series of hybrid generation methods from different perspectives. However, due to problems such as difficulty in generalizing from small sample data, incomplete modeling of full-head physical structures, and insufficient consistency in dynamic driving, there is still a long way to go for generation technology on the path to becoming fully photorealistic and capable of real-time interaction. In fact, research on 3D face generation and driving technology is still in a developmental stage, and the connotations and extensions of its technology are rapidly updating and iterating. This review provides a systematic summary of the main research works to date, along with a brief analysis of the limitations of current technologies. It also explores potential challenges and future directions for 3D face generation and application technologies, offering a guidance for future research.

Highlights:

1. Evolution of 3D representations. Systematically reviews the technological transition from explicit parametric models (3DMM) to implicit neural fields (NeRF), and to the latest explicit 3D Gaussian Splatting (3DGS).

2. Taxonomy of generation methods. Categorizes current techniques into direct optimization, generative models (GAN/Diffusion), and hybrid approaches, emphasizing their performance in data-scarce scenarios.

3. Applications & challenges. Highlights advancements in text-driven attribute editing and real-time talking head generation, while identifying future bottlenecks like complete full-head modeling and mobile deployment.

Key words: 3D face generation; 3D morphable models; neural radiance field; 3D Gaussian splatting; diffusion models; generative adversarial network

Foundation items: National Natural Science Foundation of China (No.62372033); Natural Science Foundation of Beijing (No.L252025).
Received: 2026-01-09; **Revised:** 2026-02-26
corresponding author, E-mail: yzhao@bjtu.edu.cn.