虚拟人形象合成技术综述

邓钇敏[1,2]，张旭龙[1]，司世景[1,3]，王健宗[1]，肖京[1]
1. 平安科技（深圳）有限公司，广东 深圳 518063；
2. 中国科学技术大学，安徽 合肥 230026；
3. 上海外国语大学国际金融贸易学院，上海 200083

摘要
随着元宇宙兴起，针对虚拟人形象化高效建模的需求日益迫切。从人类图像数据集中构建人类模型一直是计算机视觉的热门话题，其中3D虚拟人合成可以视作三维重建的子模块，重点在于对复杂的人体结构和表面细节的还原。对近年来虚拟人形象构建相关文献进行了全面调研，研究范围覆盖了全身形象、头部形象以及衣物建模等领域。分析归纳构建工作的基本原理，从各自技术路线层面出发将虚拟人合成方法分为基于网格、基于图像、基于体素、基于隐式表示、混合表示5类。首先介绍各类方法的基本原理，然后结合现有工作讨论具体技术，并指出各类方法的优缺点。此外还介绍了部分常见的模型质量评估的数据集和评价指标，简要介绍了虚拟人的常见应用。最后对虚拟人合成技术未来发展方向进行了展望，以合成高质量、高保真度、低延迟的虚拟人形象。

关键词
元宇宙；虚拟人；三维人体重建；计算机视觉；深度学习；人脸合成

中图分类号：TP391 文献标志码：A doi: 10.11959/j.issn.2096-0271.2022081

Human avatars synthesis technologies: a survey

DENG Yimin[1,2], ZHANG Xulong[1], SI Shijing[1,3], WANG Jianzong[1], XIAO Jing[1]
1. Ping An Technology (Shenzhen) Co., Ltd., Shenzhen 518063, China
2. University of Science and Technology of China, Hefei 230026, China
3. School of Economics and Finance, Shanghai International Studies University, Shanghai 200083, China

Abstract
Nowadays, the demand for efficient human avatars modeling is becoming increasingly urgent since metaverse has attracted more and more attention. Creating human avatars from human image datasets has always been a popular topic in the field of computer vision. 3D human avatars synthesis can be regarded as a sub-module of 3D reconstruction focusing on reproducing the complex articulated body and surface details of human. A comprehensive survey of the literature related to the human reconstruction in recent years was conducted, including the work of full-body avatars, talking-head and clothing modeling. By analyzing and summarizing existing work, human avatars synthesis technologies were divided into five categories: mesh-based methods, image-based methods, voxel-based methods, implicit methods and hybrid methods due to the features of their pipelines. Firstly, the basic principles of them were introduced respectively. Secondly, the realization based on related work was discussed and then the advantages and disadvantages of methods respectively were

pointed out. Thirdly, the datasets and metrics for model quality evaluation were introduced. Besides, an overview of various applications was given. Finally, the future directions of human avatars synthesis technology were prospected to synthesize high-quality, high-fidelity and low-latency human avatars.

Key words
metaverse, human avatars, three-dimensional human reconstruction, computer vision, deep learning, face synthesis

0 引言

人类虚拟形象常应用于远程视讯、VR/AR领域，近年来得益于计算机绘图和人工智能的发展，人类虚拟形象应用领域从文娱领域拓展至金融、医疗、教育、通信等领域。在未来，虚拟人作为现实人类的虚拟化身，是元宇宙的一个基础及核心载体，从事各项虚拟世界的活动。

通过图片数据集构建人类虚拟形象是一门结合计算机图形和几何学、图形渲染、深度学习的技术。近年来人工智能的发展使深度学习技术被应用于虚拟人形象的生成，如神经渲染、利用神经网络提升内存效率等应用，极大提高了虚拟人的合成效率和视觉质量。

虚拟人形象分为2D形象和3D形象，两者区别在于视角和制作技术[1]。2D虚拟人体通常从单一视角去观察，其在制作过程中已经固定了某一个视角，用户不可更换；3D虚拟人则支持以多个视角去观察，制作过程需要收集多视角数据，并计算推理出不同视角对应的图像。3D形象的制作技术与2D形象的区别在于其必须先生成对应的三维立体模型，3D虚拟人形象的合成技术可以被视作三维重建的一个重要的子模块[2]。本文主要对3D虚拟人形象的合成技术进行综述。

现有与虚拟人相关的综述中，参考文献[3]从应用层面出发将虚拟人构建技术分为人脸生成、姿势引导和面向衣物3类，分析3个方面的技术特点，并给出应用前景。参考文献[4]从技术层面出发，以深度学习为技术背景，将人体三维重建技术分为基于学习和基于优化两大类讨论，突出深度学习对比传统方法的高效，并分析了技术前景。本文将3D虚拟人合成技术视作三维重建的延伸，从技术路线出发，基于三维重建常用的数据表示方法，结合虚拟人重建对保真度的高要求和人体运动特点，将虚拟人合成技术分为基于网格、基于图像、基于体素、基于隐式表示、混合表示5类，讨论各类合成方法的技术特点和挑战，讨论现有的解决思路。

之所以如此分类，是出于以下考量。网格和体素是常见的三维重建过程中的数据表示方式，虽然点云也是一种数据表示，但在虚拟人领域中其一般是原始输入数据形式，需转换为网格以进行后续处理渲染；基于图像的方法处理主体是图片及其中包含的纹理、光照数据，具有多个图图转换网络；基于隐式表示的方法特点是设计一个连续的隐函数隐式描述数据属性，与离散的训练数据相比，找到一个连续的函数表示更能表现点之间的相互关系，更接近真实世界的复杂性。

本文的贡献总结如下：①介绍5类虚拟人合成技术的技术原理和现有挑战；②从5类方法中挑选出代表性算法，并剖析其优劣利弊；③给出虚拟人构建技术未来优化方向的建议。

第1节对虚拟形象合成技术涉及的主

要技术概念进行介绍，为技术原理解释做铺垫；第2节结合现有工作分类介绍各类技术的优缺点；第3节简要介绍模型性能评估所需数据集和常见的评价指标；第4节简要介绍了虚拟人技术的应用；第5节对文章进行总结，并提出高质量虚拟人合成技术的未来发展方向。

1 相关概念

虚拟人合成技术涉及计算机视觉、图形学和几何学等技术领域，本节主要介绍与之关系密切的图形学概念，包括用于刻画人体细节的纹理概念和常见的渲染方法，最后介绍近年提出的用于多视图合成的神经辐射场。

1.1 纹理

纹理的概念可以囊括为肉眼可见的物体表面细节，如颜色、粗糙程度、凹凸度等。在图像处理领域中，图像纹理用于量化图像的感知特征，提供有关图像或图像选定区域中颜色或强度的空间排列信息。纹理由表示纹理空间的纹素数组表示，纹素是纹理映射的基本单位。纹理映射是将纹理数据与模型关联的过程，一个经典的纹理映射方法是使用二维数组存储三维物体的纹理信息，三维空间的顶点除了空间坐标还引入u、v坐标以映射到纹理空间生成UV纹理贴图，将纹理空间信息与3D模型建立起联系[5-6]。凹凸贴图是在纹理贴图基础上，利用高度差信息展示凹凸纹理细节的方法。同理，UV凹凸纹理贴图需要映射到三维空间才能生成完整的网格模型。

神经纹理是纹理的一种表示，是指以基于卷积神经网络特征空间进行自然纹理的处理[7-8]。得益于神经网络，神经纹理能够存储可学习的高维特征图，可以作为场景捕获过程的一部分进行训练的学习特征图[9]。与传统纹理类似，神经纹理可存储在3D网格顶点内，与空间顶点建立相应的映射关系。对比传统纹理，神经纹理可以达到更高的维度，而高维特征图可以包含更多信息。

1.2 真实感渲染

渲染过程以三维场景和视角设置为输入，产生三维场景从特定视角看到的二维图像[10]。真实感渲染是为了客观呈现真实世界的场景，强调图像真实感。基本步骤是首先对场景几何建模，然后采集给定的环境光照条件，计算视点可见各物体表面颜色，以达到与真实世界相近、人眼可接受的视觉效果[11]。

物体在人眼所呈现的颜色与入射光线接触物体后反射到人眼有关，这是局部光照模型的基础。局部光照模型是计算机真实感图形学中的一种光照模型，能与光栅化渲染算法相适应，特点是易于计算，被广泛应用于游戏等领域。局部光照模型的计算体现于对物体的直接光照计算，包括环境光、漫反射光、镜面反射光的计算。

PhongBT[12]提出的模型是一种典型的局部光照模型，其真实感取决于环境光、漫反射、镜面反射数据。漫反射是光入射到物体表面后以同等光强反射到各个方向的过程，由物体粗糙表面产生，漫反射的存在可以刻画物体的体积感，用于表示物体的形状线条等粗糙特征。镜面反射是平行光入射到相对光滑的表面后平行地向一个方向反射的过程，镜面反射的存在可以用于刻画物体的高光和阴影。而不同环境光照射到物体，物体表面会呈现不同的视觉效果，譬如高光区的改变、色调的变

化，模拟这种变化也是真实感渲染的重要内容。在计算机图形学中，凹凸贴图用于突出物体的凹凸纹理，该技术除了使用更多的多边形表示，还可以从高精度图像用法线贴图[13]实现凹凸光照感，极大地增强低精度多边形的外观和细节。对虚拟形象进行重新照明时，意味着变换环境光后，人体需要呈现与之对应的反射效果，展现相应的颜色和明暗。

基于物理渲染（physically based rendering，PBR）是一种基于物理世界的渲染技术，旨在以物理上合理的方式模拟光线，比传统光照模型更具真实感[14]。与前述的光照模型原理相似，PBR使用从现实测量的表面参数表示现实世界的材质，突出效果的真实感。进行PBR的重要组件有漫反射、镜面反射及法线数据，渲染时需要根据给定的光照条件进行光照绘制。PBR所生成的虚拟形象支持重新照明和动态变化等操作，突出展现虚拟人的真实感。

1.3 体积渲染

体素是组成体数据的最小单元，类比像素之于二维平面，体素表示体空间的相对位置。体积渲染又称为体绘制，目的在于提供一种基于体素的绘制技术，生成二维结果图像。与光照模型重视展现光照效果不同，体绘制重视刻画物体内部细节。

直接体绘制算法根据不同的绘制次序，分为基于图像空间序列和基于物体空间序列两类[15]。基于图像空间序列的体绘制是从反方向模拟光线穿过物体的过程，从屏幕上每一个像素点出发，沿视点方向发射一条射线，穿过三维数据场，沿着射线进行等距采样，使用一条光线上所有采样点的不透明度和颜色，通过运算推理出屏幕上该像素点的颜色值，也称光线追踪法。基于物体空间序列的体绘制事先根据每个数据点的函数值计算该点的颜色及不透明度，根据给定的视平面和观察方向将数据点投影到图像平面上，参考数据点在空间的先后遮挡顺序合成计算不透明度和颜色，也称光线投射法。体绘制形成的图像一般是半透明的，体绘制的光线投射法简单示例如**图1**所示，假设某一点发出多条光线r_0~r_4，从f_0~f_4点进入体数据，从l_0~l_4出去，按照一定的距离间隔等距采样体纹理数据，直至穿出体数据。

1.4 神经辐射场

神经辐射场NeRF最早被提出是用于静态场景表示的[16]，通过多视角发出的相机光线收集数据合成多视角图像，结合经典体绘制技术将输出的颜色和体密度投影到图像中。其中，用多层感知机（multilayer perception，MLP）网络隐式学习静态3D场景，如式（1）所示：

$$F_\theta:(x,d)\rightarrow(c,\sigma) \qquad (1)$$

采用一个5D向量值函数F_θ表示一个连续场景，输入负责存储三维坐标信息的x和球坐标表示的视角方向d，输出指定位置

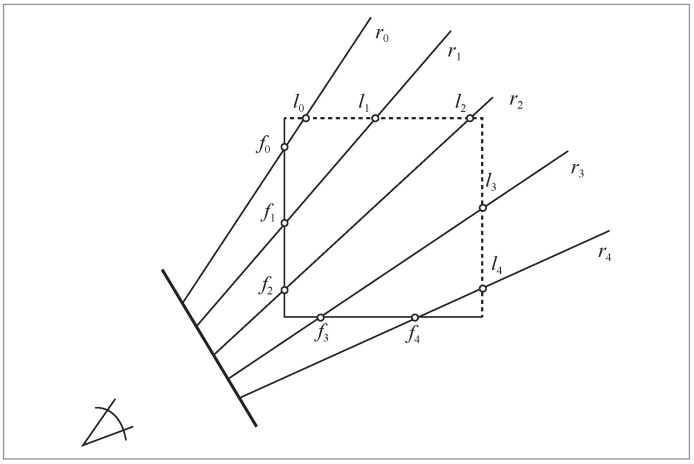

图1 体绘制的光线投射法简单示例

的颜色c和体素密度预测σ，体密度σ为光射线在指定位置的无穷小粒子处终止的微分概率。

颜色估计通过分层采样，采用数值近似的方法估计每条相机射线上连续的空间采样点数据，基于立体神经渲染绘制出2D图像中对应像素点的颜色$C(r)$，如式（2）、式（3）所示：

$$C(r)=\int_{t_n}^{t_f}T(t)\sigma\left(r(t)\right)c\left(r(t),d\right)\mathrm{d}t \quad (2)$$

$$T(t)=\exp\left(-\int_{t_n}^{t}\sigma\left(r(s)\right)\mathrm{d}s\right) \quad (3)$$

其中，$T(t)$表示射线从最近点t_n到最远点t_f的累积透过率，即从两点间传播而不撞击其他粒子的概率。$r(t)$表示相机光线的采样点，c表示采样点颜色，d表示采样间隔。

最初的NeRF算法需要结合经典的体绘制算法，因体素存储能力有限，限制了结果的分辨率，加上场景优化的需求，总体耗时较长。因此Point-NeRF[17]算法结合立体渲染方法和深度多视图立体能快速重建场景几何的优点，提出用神经点云计算体积属性，模拟辐射场的方法。点云可以被看作观察空间中海量点的集合，神经点云具有神经特征。研究应用一个基于点云的神经渲染过程，该研究提出基于光线行进法，聚合场景表面附近的神经点特征进行渲染，避免空区域的采样，通过直接网络推理加速辐射场的初始化，并通过点增长和剪枝优化场景表示。目前神经辐射场已被应用于虚拟人合成领域执行多角度合成视觉图像的任务，展现出良好的性能。

2 虚拟形象合成技术

本节将依次介绍基于网格、基于图像、基于体素、基于隐式、混合表示5类方法的基本原理和存在的技术挑战。

2.1 基于网格构建

基于网格构建虚拟人形象，核心在于人体的网格模型，本文着重介绍人体参数化模型。人体参数化模型是支持参数对模型的属性进行动态调整的人体网格模型。最经典的人体参数化模型生成算法是SCAPE[18]，利用主成分分析法（principal component analysis，PCA）提取体型和姿态两个独立的低维参数合成人体参数化模型，网格变形依赖于三角形的旋转变形。而如今动画制作常常应用网格顶点进行变形，对应经典的蒙皮技术，接下来对此进行介绍。

虚拟人形象可被视作骨架和表皮两部分，骨架由关节树构成，表皮是由多个三维空间点组成的面。要建立人体的网格模型，应先生成一副骨架，将网格顶点按照一定的权重绑定在关节上，这一个过程被称作“蒙皮”[19]。人体运动可以被视作人体内部发生了铰接运动，对应关节发生了旋转和位移。对人体运动的模拟体现在通过计算运动相关的关节受到的影响，获得活动后的关节位置。线性混合蒙皮（linear blending skinning，LBS）算法[20]根据人体特定运动对每个绑定的关节点的影响进行加权求和的线性运算，实现网格变形。传统的LBS算法单纯对旋转进行线性运算，造成“糖果包装”的肢体扭曲现象，而且关节连接处可能会出现断裂。在LBS算法的基础上，SMPL模型[21]实现了一个人体参数化三维模型，在混合变形过程中利用数据学习参数为关节连接处提供平滑的过渡。SMPL模型支持从外界输入姿势参数和体型参数，模拟人体肌肉在肢体运动过程中的形变，从而控制人体的形态变化。

近年来，SMPL模型的版本已经衍生

到SMPL-X[22]，支持从单帧RGB图像建立起包含身体、手部姿态和面部表情的三维立体模型，扩展了对手部和面部的建模。SMPL-X模型增加了表情和手势等细节，**图2**展示了模型建立的基本步骤及效果对比。此外规则的人体网格还可用于特征提取[23]，为动态虚拟人合成提供强大的人体先验知识[24]，包括基本的姿势、体型数据，作为构建4D虚拟人的初始化辅助手段。H4D[24]以每帧采集的点云组成序列为输入，分别进行形状、姿势、运动等编码构建初始的人体参数化网格模型，再设计一个辅助编码器处理细粒度的衣物和头发进行几何组合，得到完整的人体网格。

除此以外，Osman A A A等人[25]提出了将STAR作为SMPL的替代方法，将姿势相关的形变分解为一组空间局部姿势校正的混合形状函数，姿势形变会根据人的体型进行校正。SMPL生成的模型是一个基于顶点的线性模型，是目前最广为应用的人体参数化模型；而采用非线性策略的模型以GHUM/GHUML[26]为代表。基于VAE[27]的隐空间表示，GHUM/GHUML依赖于标准流[28]的分布近似和推导计算，生成一个非线性参数模型表示骨架运动。

虽然基于网格的方法可以生成逼真的人体模型，具有足够的铰接式运动的仿真模拟能力，但是前提是需要建模对象具有固定的拓扑，其对于衣物、头发等精细结构的建模能力很差。现有工作[29-34]给SMPL模型的网格顶点添加偏移以表示衣服的几何（SMPL+displacement，SMPL+D），结果仅适用于紧身衣物的表面模拟，且难以恢复衣物的边界细节。尤其对于宽松的衣物，由于与顶点绑定后具有相同的蒙皮权重，移动时会造成明显的伪影。为了改善SMPL模型的性能，BCNet[35]采用神经网络为特定类型的衣物建立独立的蒙皮权重，使衣物网格独立于SMPL模型并且能够叠加在SMPL模型上，通过位移网络表示衣物随运动的形变，规避了部分类型衣物绑定身体网格顶点带来的伪影问题，进一步优化基于网格的方法对服装迁移的解决方案。除此以外，近年来基于网格的方法所生成的人体模型被用作建模的基础[36]，寻找对衣物、头发建模能力更强的方案作为替代细化模型，或者协助完成动作推导等任务。

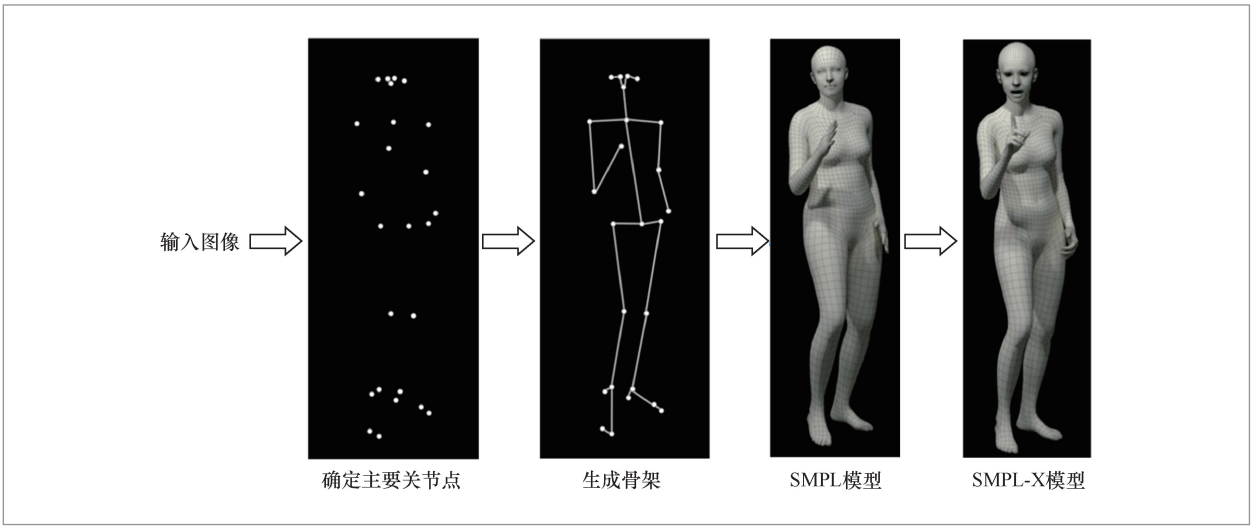

图 2　SMPL 模型及 SMPL-X 模型合成流程

2.2 基于图像构建

基于图像的方法的核心是通过若干幅二维图像恢复场景的三维结构，实现多个图像到图像的转换网络，常见于人脸重建。人脸重建关系着对应的3D可变形面部模型（3D morphable model，3DMM）[37]，其创建过程要求提供大量的光照变化、姿势和表情数据，分为线性和非线性两类。低维性是线性3DMM的特点之一，利用PCA采集低维空间的纹理和面部形状特征，或采用学习网络推理线性面部模型，生成具有真实感的、用于物理渲染的数据，例如反射率、法线等。采用PCA建立的线性3DMM难以复现人体纹理和几何的高频细节，对自然场景下的图像集的泛化能力很差[38]。基于无监督或弱监督学习生成的非线性可变形面部模型[39]可以处理大量自然场景下的图片，但不适合重新照明的人像化身和动画，因为环境光照条件和表情等数据已经被保存在输出的纹理图片中。主流方法是采用基于深度学习进行图像后处理以推理线性面部模型用于重照明的渲染组件[40]。

生成式对抗网络（generative adversarial networks，GAN）通过生成器和判别器网络的不断博弈提高建模能力，如今被广泛应用于脸部重建的纹理提取，结合3DMM完成人脸重建任务，优势在于能够处理高分辨率图片，实现图像到图像的转换。在图像处理领域，GAN的生成器主动学习人脸特征，输入高斯噪声控制面部细节的变化，生成虚假图像“欺骗”判别器，而判别器需要不断提高判别能力甄别生成结果的真伪。二者在对抗过程中改善模型生成效果。

GANFIT[41]实现从单张自然场景下的图片重建高质量的纹理和形状数据，采用GAN训练出大规模的高分辨率的纹理数据，且身份特征得到了保留。与前述一样，在GANFIT获得的纹理数据中，光照条件已经被保存进去，不能从中重建高频法线和镜面反射等数据，因此不能直接进入渲染阶段。根据GANFIT获得的纹理和形状数据，AvatarMe[38]对输入图像的非线性3DMM，通过消除纹理照明网络以提取其中的漫反射数据，并将已生成的、可靠的漫反射数据设计通过多个图转换网络推算出镜面反射、镜面法线和漫反射法线数据等真实感渲染的重要组件，所得的人脸模型支持重新照明操作，呈现不同光照条件下面部对应的变化，可以直接用于渲染。尽管AvatarMe的训练数据集相对庞大，但由于欠缺深色皮肤的人脸数据，不能很好地处理深色皮肤人种的面部重建工作，此外还较为依赖输入图片的分辨率、照明等条件。

styleGAN[42]提出一个基于样式的GAN算法，实现输入图片中的高级属性和生成的图像自学习、无监督地分离，并且可以直观控制合成。其中的style表示输入的数据风格主要体现为人脸的主要属性，如表情、面的朝向、发型之类。与传统生成器网络相似，styleGAN的生成器网络每一层的图片分辨率是递增的，呈现渐进式的生长。传统GAN和styleGAN的结构对比如**图3**所示。styleGAN相对于传统GAN的一个明显改进是对输入z的特征解耦，产生一个不受训练数据分布影响的中间向量$\boldsymbol{w}$，减少某个特征与向量中元素的关联数，并且投喂给生成器网络的每一层，使得每一层输入的噪声对其他特征的影响降到最低。StyleRig[43]实现一个通过3DMM对styleGAN进行面部绑定的网络，控制语义参数（如人脸表情）以实现面部变换，然而变换能力是非常依赖3DMM的，且不能显式控制不被3DMM解释的场景特征。

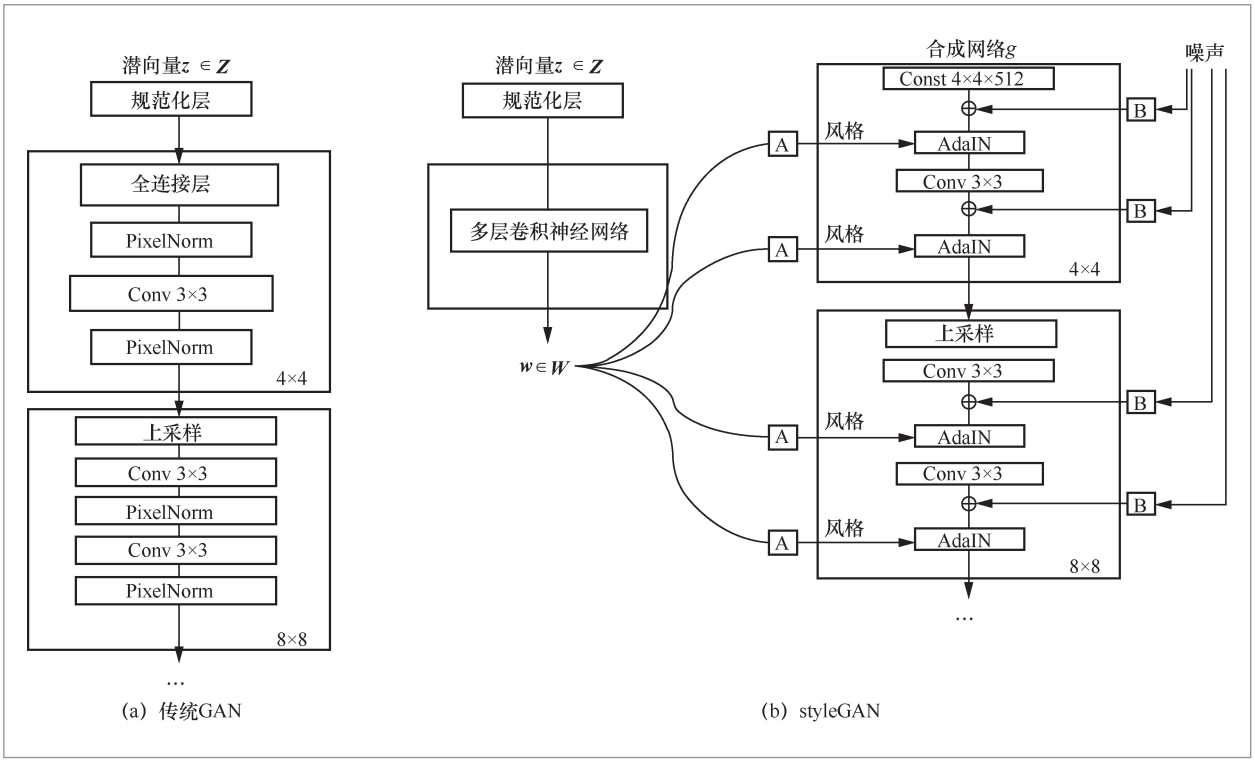

图 3 传统 GAN 和 styleGAN 结构对比

styleGAN2[44]是styleGAN的改进版本，消除了水滴伪影，并且结合残差网络直接将低分辨率特征映射到最终生成结果。参考文献[45]利用styleGAN2的架构训练具有3D几何和反射率纹理的3DMM，并对渲染后的人脸采用感知细化，能克服极端光照条件带来的困难，生成高分辨率的标准化、中性表情的人脸，然而不完善的训练数据会使该模型难以从肤色中完全分离光照信息，光照信息和表情信息不能与人脸完全解耦。除了人脸重建领域，近年来GAN也被应用于虚拟着装领域[46-47]，用于模拟具有不同拓扑的服装，通过设计映射网络将服装定位到不同的人体模型上。

基于图像生成的人类化身可以达到高分辨率，达到足以“欺骗”观众的视觉质量。然而，纯粹基于图像生成的人类化身依赖于训练的图像数据，通常只适合对正脸图像的训练，而且由于缺乏3D空间信息，难以保持多视角下的模型一致性。

2.3 基于体素构建

基于体素的虚拟人构建方法能够生成多视角一致的模型，要求具有三维空间体素化操作，以及将三维空间的对象进行二维投影操作，图像重构的结果具有与原图一致的纹理和分辨率。理想状态下的体素重构算法应具有范围不确定性表示、增量和顺序的独立更新、一定的时空效率、无限制的拓扑类型4种理想属性[48]。早期的基于体素的重建方法依赖3D扫描数据[48-50]，对实验设备有一定的要求；近年来基于体素的重建工作尝试将体素化概念嵌入重建过程。Deep Voxels[51]先提取源图片的2D特征，引入体素表示一种固定视点

的3D特征网格，将2D特征提升到3D空间观察并集成到特征网格中，依次进行3D空间推理和2D特征合成，特点在于无须3D监督学习。参考文献[52]收集所有体素的“上下文”关键特征信息来更新当前体素中的关节点特征，约束肢体长度，实现从单张图像估计3D姿势。Deep Human[53]将人体参数化模型体素化，提出一个图像引导的体积到体积转换网络，使用多尺度体积变换结合3D体积和2D图像的知识。

存储3D空间点信息意味着高成本的内存，提高精度就会大幅增加计算耗时。因此一个可行的基于体素的构建方法要解决如何将图片缩放到更高分辨率的问题，以处理更精细的细节，如皱纹之类。为了规避内存限制的低分辨率问题，Deep Voxels采用局部分辨率换空间策略，但这也意味着牺牲了数据利用率，存在细节丢失的问题。Neural Volumes[54]提出warping fields，将存储空间尽可能分配给对合成图片贡献更大的区域。但是要想达到传统纹理网格表面实现的保真度，还需要进一步改进。

2.4 基于隐式表示

隐式表示直观来说是定义一个连续的三维空间标量函数表示曲面，近年来可以结合神经网络进行场景的隐式表示。在虚拟形象合成领域，隐函数借助上下文内容提供的局部特征信息，推断出整体形状信息[55-57]，神经辐射场常用于多视图合成。与体素表示方法相比，基于隐式表示方法的内存效率更高；对比基于图像的表示方法，隐式表示方法还可以推理出不宜观察的区域颜色。

TextureFields[9]提出一个基于回归的神经网络参数化的连续3D函数表示纹理场，独立于3D对象的形状表示，学习将示例图的纹理转移到源网格以合成新视图。SRNs[58]提出一个连续的、3D结构感知的场景表示，能够通过学习的定向距离场定义表面，无须3D监督条件下对3D场景的几何和外观建模，并保持多视图一致。BANMo[59]利用隐函数隐式表示物体，结合NeRF的概念通过MLP网络给出3D空间点的颜色、体密度和训练所得的规范嵌入。规范嵌入用于对3D空间点的语义信息进行编码，注册不同时间示例中的像素观察值，其中应用MLP计算点到表面的定向距离函数（signed distance function，SDF）给出3D形状，用连续表面嵌入（continuous surface embeddings，CSE）[60]初始化像素嵌入，生成像素对应的特征。与SMPL建立的可参数化模型相比，BANMo所需的数据量更小；与NeRF相比，BANMo更适用于表示物体更大幅度的运动。

其中，SDF和定向距离场核心都是通过体积场表示物体表面的，计算场中的点到物体表面的最短距离，距离在物体表面上为0，在物体内部为负，外部为正。为了提高表示效率，DeepSDF[61]结合MLP实现一个连续的SDF表示形状，是如今常用的隐式表示之一。SDF被广泛用于非刚性重建领域[48]，经过MLP的优化更有效执行非刚性的重建和变形跟踪任务[62]，其变体也能为多模态的三维重建任务提供强大的先验知识[63]。而CSE用于为2D图像中每个像素预测物体网格中相应顶点的嵌入向量，将其与3D对象几何建立密集对应关系。

PIFu[64]/PIFuHD[65]提出一个局部对齐2D图像像素对齐隐式函数，与其他隐式表示方法相比，PIFu更关注像素级的特征，保持输出图像的对齐。简单来说，该隐函数对于任意一个3D顶点，根据相机参数进行投影，获得对应的2D位置信息以及深度信息，同时学习该点的特征向量以保留局部

细节，同时进行不可视区域的信息推理。其将表面表示为一个水平集，如式（4）所示：

$$f\left(F(x),z(X)\right)=s,s\in R \quad (4)$$

其中，对于一个3D点X，$x=\pi(X)$是对应的2D投影，$z(X)$表示相机坐标空间的深度值，$F(x)=g(I(x))$表示2D投影点x的特征。因此表面信息的保存不需显式的存储空间，提高了内存利用率。PIFu可以被拓展到多图像、多视角的输入，给出完整的、高分辨率的3D模型的表面和纹理，能有效执行复杂的着装人体建模任务[66]。

以PIFu为代表的隐式表示方法的连续性支持以较高的内存利用率生成具有任意拓扑的几何图形，还可以拓展到颜色的合成。隐式表示的内存高效性能被用于改善内存限制的模型构建方法，出色的推理能力能用于弥补人体参数化模型在衣物建模方面的不足。

基于隐式表示的方法对现实生活中的遮挡问题有具体的解决方案，这对隐函数的推理能力提出更高的要求。而且对于形状的隐式表示受限于缺乏网格拓扑、骨架和蒙皮权重等结构信息而无法展现新的姿势，只能从固定视点去控制着装化身的身体形状。对此MVP-Human[67]借助3D扫描等技术获取三维信息；BANMo结合神经蒙皮模型，利用显式3D高斯椭球体随骨骼移动以调整权重，从而展现大范围的铰接变换；并且结合NeRF进行多视图合成，展现新的视角。而NASA[68]则提出一种以姿势为条件的隐式占用函数替代多边形人体网格，用于表示铰接可变形的人体对象。针对单视图人脸重建，JIFF[69]采用了3DMM提供的形状先验，结合空间对齐的三维特征和像素对齐的二维特征，共同预测隐式人脸函数，以改善隐函数在人脸重建应用中的质量。

2.5 混合方法构建

近年来对于上述4类合成技术的改进均取得了很大的进步，深度学习的普遍应用提高了训练效率。除此以外研究人员尝试结合多种方法的优点进行互补，致力于提高虚拟人的质量。而且各类构建方法的改善思路类似，总结如下。

基于网格的训练中，以SMPL模型为基础建立的赤裸人体参数化模型仍是主流方法之一，在此基础上对人体纹理的生成方法加以研究，以弥补其对衣物和头发建模能力的不足；基于图像的方法，考虑为模型添加3D空间信息，并提高数据的利用率；基于体素的方法，降低内存成本仍是主要改善方向；隐式表示类的方法的性能可以通过不受约束的观察环境以及更高分辨率的训练数据得到提高。混合应用这几类方法，能够互补各自技术路线存在的不足，有效提高模型质量。基于以上思路，本节提出混合合成方法作为独立的一类，并介绍部分现有工作的混合构建技术，以提供优化思路参考。

网格-隐函数混合。隐式3D表示具有很强的表现力，结合可学习的参数化模型如SMPL能更好地捕捉和还原着装人体的形状和外观。参考文献[70]联合学习两个隐函数用于预估着装人体及身体部位标签，将隐函数和参数化模型建立联系。其中被衣物覆盖的身体内表面由SMPL模拟，受预测的身体部位信息约束；应用SMPL+D将内表面注册到外表面，优化每个顶点的偏移D以拟合外部的隐式表示。所设计的隐函数特点在于将点的位置扩展到身体内部、身体与衣物间、身体外部3类。

SCANimate[71]能够对着装的人进行3D扫描并转换成参数驱动的虚拟人。其利用SMPL获得参数化人体三维模型，结合弱监督模型进行姿态修正，设计一个局部

位姿感知隐函数表示该人体模型，并模拟衣物在运动中的形变，据此生成新姿态。试验证明SMPL提供的人体结构信息有利于改善隐式表示的性能，提高人体姿态的泛化能力。美中不足的是SCANimate的模型表示适用于与身体拓扑相似的贴身衣物，不适用那些较为宽松的衣物，而且构建的模型是确定性的，也就是同一个姿势对应同样的衣物褶皱程度，因此对衣物的形变预测不能涵盖所有的随机性。

ICON[72]则将SMPL-X模型与自设计的法线预测网络进行循环优化，推断出的着装人体法线贴图用于回归着装人体的隐式3D表面，其中使用了不受全局姿势变换影响的局部特征来执行隐式3D重建任务。ICON可以从单张图片中恢复3D着装人体形象，并且能够应用于自然场景下的虚拟形象构建，还能结合SCANimate生成动态的人类化身。由于ICON基于正交视图训练，也就是由2D投影图描述3D属性，透视效果不够理想，容易产生不协调的肢体。

结合多项工作，可以看出SMPL参数化模型支持多种人体运动的可控形变，而NeRF作为一种先进的场景表示方法，可有效预测空间点的颜色和体积密度，提供多视图合成。因此SMPL结合NeRF能够提供人体模型的形状和衣物的控制[72-73]。其中参考文献[74]提出表面定向神经辐射场用于合成可控的人类形象，可以基于少量的多视角视频和SMPL模型的先验知识重建一个3D人体模型。

网格-图像混合。人体参数化模型能充分表示人体关节运动，而结合其他更先进的纹理合成方法能弥补其对衣物、毛发的建模。GAN的应用可以生成高分辨率的结果，其中包括了高质量的纹理[75-76]。StylePeople[76]提出一个神经穿着模型，核心是利用styleGAN2学习输入图像的神经纹理，叠加在SMPL-X生成的赤裸人体模型上，得到高质量的着装虚拟形象。StylePeople利用全卷积网络生成身体部位坐标和身体部分分配的堆栈，对身体纹理进行采样映射，结合堆栈指定的权重生成RGB图像。StylePeople将styleGAN2从人脸重建推广至人体的全身形象构建，所习得的神经纹理可有效模拟头发和衣服，弥补了SMPL模型的不足。与基于图像的方法一样，建模的效果依赖大量训练数据，对模型的数据利用率要求高。

体素-隐函数混合。基于体素的方法能够展现多个角度的视觉效果，而隐式表示方法内存效率高，能有效提高基于体素方法生成结果的分辨率。像素对齐立体虚拟人（pixel-aligned volumetric avatars，PVA）[77]结合体绘制和神经辐射场进行图片渲染，其中采用了PIFu提出的像素对齐特征保留高频细节，并用于调节多身份神经辐射场的参数，采用MLP将空间位置和像素对齐的特征转换为颜色和占有率来规避体素化带来的内存限制问题。PVA能根据少量的样本数据生成高保真的虚拟形象，然而不具有捕捉光照条件和背景变化的能力，意味着其不能应用于自然场景下的图片。

参考文献[78]基于神经辐射场采用隐函数表示面部和头发的几何外观，结合体积渲染恢复头发的体积感，实现头部的动态变化。对比传统的基于体素的体积渲染方法，结合神经场景表示网络的体绘制表示更紧凑，所渲染的图片分辨率得到进一步提高。而该实验只是止步于动态的头部形象生成，没有拓展到对全身的动态表示，该项拓展会带来更复杂的体积模型和光照计算。

网格-体素-隐函数。同时采集输入数据的二维特征和三维特征，有效提高数据的利用率。该类混合解决方案充分利用了人体网格的规范及隐函数强大的表现力和高效内存，并且能够实现多视角效果。PaMIR[79]

采用非参数化的深度隐式场表示表面，用SMPL提供的参数化模型规范人体，收集像素级和体素级特征，绑定每个3D点对应的隐函数的值。DeepMultiCap[80]将隐函数与姿势和体素化网格结合，从图像像素级恢复局部细节，提高姿势变化的健壮性。S3[81]将输入的点云数据体素化为一个体素网格，用于表示体积特征；结合2D图像特征提取将行人的形状、姿势、蒙皮权重表示为直接从数据中学习的神经隐函数，构建动态的人体模型。

到目前为止的虚拟人合成技术分类如**图4**所示。基于上述对5类虚拟人构建方法原理的介绍，**图5**根据时间线展示了5类合

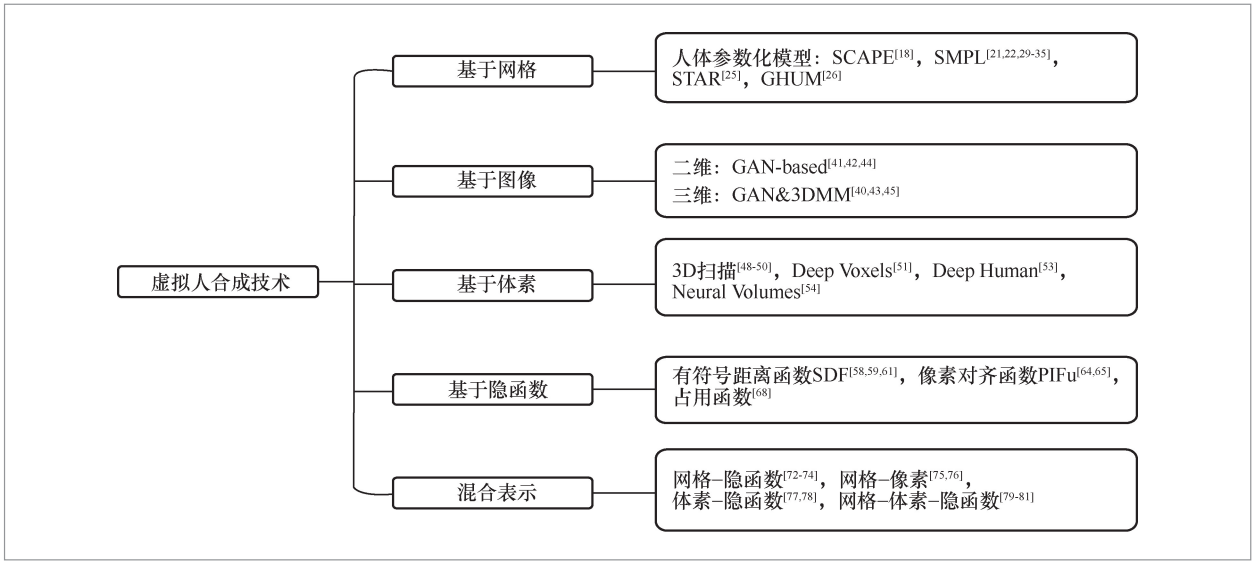

图 4　虚拟人合成技术分类

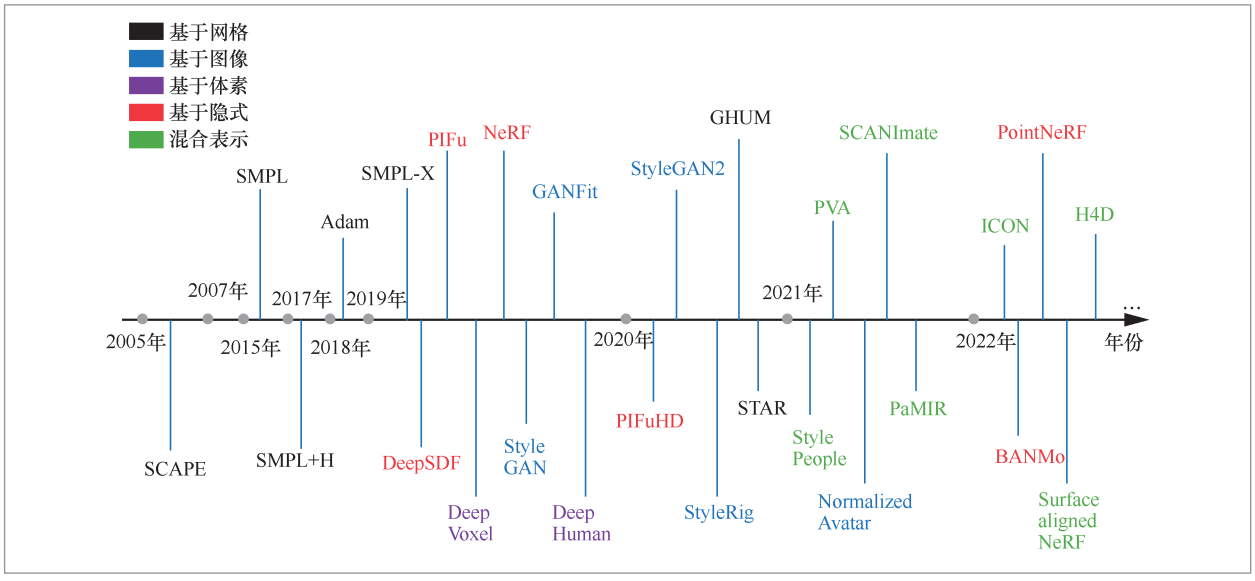

图 5　虚拟人合成技术发展历程

成技术中具有代表性的技术发展历程及最新应用，近年来比较多的工作采用混合方法合成虚拟人，也有工作在不断创新或改进其他4类技术。

3 性能评价

本节首先介绍合成虚拟人的实验所需数据集和性能评价中常见的定量指标，再给出各类合成技术的性能评价结果。介绍常见数据集形式、主要内容及应用方向，同时具体介绍部分数据集为性能的定量比较做铺垫，介绍感知类指标和非感知类指标、比较典型合成算法的性能。

3.1 数据集

在现有工作中定性比较的方法通常是比较直观、简单的，通常采取用户调查等方法收集信息。而在定量比较方面，则涉及复杂的计算和场景。

虚拟人合成技术的应用涉及了人脸建模、行为预测、虚拟着装等多个领域，着重点不同的合成实验所用的数据集各有特点。常见的训练数据集包含了图片、视频两种形式，通过单一视角或多视角的相机系统采集。为了提高模型性能，近年来有团队通过对原始数据进行预处理，获取2D和3D的人体数据作为模型输入，系统化地生成专门面向虚拟人的脸部建模、行为预测等应用方向的数据集。以下将结合具体文献工作列举总结现有数据集的形式和特点。

二维图像集通常用于提供二维信息，也可以结合深度图通过多角度展示三维信息。除深度图外，三维数据类型还包括多边形网格、点云数据，其中点云适用于表示稀疏结构，可以转换为标准的三维多边形网格。Neural body[82]创建了一个多视角视频数据集ZJU-Mocap用于评价模型基于稀疏数据集合成新视角结果的表现，总共包含了9个动态人类视频，由21个同步相机多视角摄制，其中的人类展示了太极、热身、拳击等复杂动作，该数据集广泛应用于多视角合成的质量评估。S3[81]利用了2D图像集和雷达扫描的3D点云数据，数据输入形式包含单张图片和单个体素化雷达扫描数据。

面向关注姿势的虚拟人合成技术，Market-1501[83]是行人重识别常用的数据集，包含了1 501个行人的32 643个注释数据图像，每个身份最多由6个相机捕捉。Human3.6M[84]是数据量级更大的3D人类姿势数据集，有由4个相机摄制的多视角视频，使用基于标记的动作捕捉系统，其中包含由5名女性和6名男性展示的复杂动作。

除了常规形式的数据输入外，3D人体模型也可以作为虚拟人合成模型的输入。STATE[85]根据实验需要合成Human图像数据集，数据来源于twindom提供的真实扫描3D人类模型，每个模型由496张多视角的图片渲染而成。近几年的虚拟人数据集还有CAPE[33]和AGORA[86]。CAPE是第一个将直接打扮的3D人体网格推广到多姿势，从3D扫描中生成以姿势和衣物为条件的人体模型。AGORA通过扩展SMPL-X身体模型安装到3D扫描中，创建3D姿势和身体形状，具有多种姿势和服装的人体扫描数据。

现有公开的标注数据集的体量是很庞大的，在虚拟人合成技术中的性能评估往往只取其中的子集进行分析，输入形式仍以图片和视频为主，**表1**给出了部分文献采用的数据集信息，涵盖了2D和3D数据。

表1 数据集信息

数据集	时间	实验数据规模	类型	信息维度
People Snapshot[87]	2018年	24个视频序列，11个人	视频	2D
短多目RGB视频序列[78]	2021年	2分钟6 000帧，分辨率为512×512	视频	2D
ZJU-MoCap[82]	2021年	9个动态人类视频	视频	3D
Market-1501[83]	2015年	32 668张，1 501人	图片	2D
Human[85]	2021年	立体模型	模型	3D

3.2 评价指标

本节介绍常见的用于量化虚拟人模型性能的精度指标，指标的选取因模型训练方法而异。

● L1-Loss：即平均绝对误差（mean absolute error，MAE），模型预测值和真实值之间绝对差值的平均值。

● LPIPS：学习感知图像块相似度指标，也被称为“感知损失”，能系统地评估不同结构和任务的深层特征，适用于多种不同的结构和监督级别。给定真实图像x和重建图像x_0，感知相似度[88]如式（5）所示：

$$d(x,x_0)=\sum_l \frac{1}{H_l W_l}\sum_{h,w}\left\| w_l \odot \left(\hat{y}^l_{hw}-\hat{y}^l_{0hw}\right)\right\|_2^2 \quad (5)$$

式（5）中，从L个层提取特征堆栈，并在通道维度上进行单元规范化，用向量W_l缩放激活通道，并计算$L2$距离，LPIPS的值越低，建模效果越好，两张图像越相似。

● IS：即inception score，基于inception model计算的、用于衡量GAN模型的性能，将生成图像结果输入用真实图像训练的分类模型中，衡量与真实图像整体的分布距离，距离越小该得分越高，性能越好。

● FID：计算真实样本、生成样本在特征空间之间的距离，常见于GAN模型的测量，比较不同生成器结构产生的图像的质量。FID分数计算的是真实图像分布和生成器结果分布之间的差异[89]，如式（6）所示：

$$\mathrm{FID}(x,g)=\left\|\mu_x-\mu_g\right\|+Tr\left(X+G-2\sqrt{XG}\right) \quad (6)$$

真实图像和生成图像在Inception Net-V3输出2 048维特征向量，X,G表示各自特征向量集合的协方差矩阵，μ_x和μ_g表示集合的均值。FID的值越小，表示距离越小，模型性能越好。

● SSIM：用于衡量两个图片间的结构相似性，从图像组成角度定义图像信息[90]。该指标分析量化图像的亮度、对比度以及结构。用均值估计亮度，用标准差估计对比度，用协方差估计结构相似程度。结构相似性的范围为-1~1，因此SSIM的值越高越好，当两个图片完全一样时，SSIM的值为1。

假设两张图片x和y的亮度为l，对比度为c，结构s衡量表示如式（7）~式（10）所示：

$$l(x,y)=\frac{2\mu_x\mu_y+c_1}{\mu_x^2+\mu_y^2+c_1} \quad (7)$$

$$c(x,y)=\frac{2\sigma_{xy}+c_2}{\sigma_x^2+\sigma_y^2+c_2} \quad (8)$$

$$s(x,y)=\frac{\sigma_{xy}+c_3}{\sigma_x\sigma_y+c_3} \quad (9)$$

$$\mathrm{SSIM}(x,y)=l(x,y)\cdot c(x,y)\cdot s(x,y) \quad (10)$$

其中，μ表示图像均值，σ_x和σ_y表示图像标

准差，σ_{xy}表示图像x和y的协方差，c_1~c_3表示用于防止除0的常数。设定$c_3=c_2/2$时，SSIM表示为：

$$\mathrm{SSIM}(x,y)=\frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{\left(\mu_x^2+\mu_y^2+c_1\right)\left(\sigma_x^2+\sigma_y^2+c_2\right)} \tag{11}$$

● MSE：均方误差（mean square error，MSE），也称为L2-Loss，给定参考图像f和测试图像g，给定灰度图像，假设参考图像f和测试图像g大小为$M\times N$，MSE定义如式（12）所示[90]：

$$\mathrm{MSE}(f,g)=\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(f_{ij}-g_{ij})^2 \tag{12}$$

● PSNR：峰值信噪比（peak signal-to-noise ratio，PSNR），常见于图像和信号处理的测量实验中，通过计算图像均方误差的对数得出[90]。给定灰度图像，假设参考图像f和测试图像g，大小为$M\times N$，f的最大像素值为MAX_f，f和g之间的PSNR定义如式（13）所示：

$$\mathrm{PSNR}(f,g)=\frac{10\log_{10}\mathrm{MAX}_f^{\,2}}{\mathrm{MSE}(f,g)} \tag{13}$$

彩色RGB图像的PSNR计算，可以分别计算3个通道的PSNR再取均值；分别计算各通道的均方差的均值，统一求PSNR；或者将RGB转换为图像彩色编码YUV，只计算亮度分量的PSNR。显而易见，该值越高，表现越好。

PSNR对于保真度信号的测量与不可感知性有关，而SSIM是基于人体视觉系统设计的，更能体现重建结果的保真性，而通过深度学习网络抽取的LPIPS是近年来被认为最接近人体感知的指标，能更有效地展现两张图片的感知相似度。

3.3 算法评价

本节将依次从定量和定性两方面，依次进行各类技术具体算法的性能比较。首先基于3.2节的定量评价指标和**表1**的数据集，给出了具体算法的定量比较，见**表2**。其中**表2**数据来源于现有文献研究，针对不同数据集、同一实验环境下从罗列的定量指标中选取部分指标进行测量比较。由于数据都来源于现有工作的测量，部分研究的指标选取没有涉及上述所有指标，故存在缺失情况。

接下来进行的是各类技术的直观定性比较，**表3**给出了各类技术中部分具体算法的定性比较，合成范围由头部及至全身，从保真度、泛化性、能否支持动态虚拟人的合成这几方面进行比较。其中保真度和泛化性分为低、中、高3个等级，根据生成效果对3个衡量标准的满足程度进行定性评价。

● 时间信息的利用。体现在模型能否训练视频数据，生成动态的虚拟人形象。

● 保真度。模型的保真度注重视觉效果，能否尽可能还原图片的细节，从生成结果的分辨率、人像的完整性、与原图的视觉偏差等主观保真度准则出发进行比较。

● 泛化性。模型的泛化性通过能否展现新姿势、采用新视角、适用于自然场景图片这3个方面衡量。自然场景下的图片数据集具有光照条件多变、背景复杂等特点，往往体现为无标记的数据集。因此是否能应用于自然场景的虚拟人构建方法可以作为模型泛化性的一种度量方法。

4 虚拟人应用

目前市场上常见的虚拟人应用根据其定位可分为服务型和身份型两类。其中服务型虚拟人是具有一定功能性的服务提供者；身份型虚拟人则需具有身份特征，能够替代真人进行一定的娱乐、社交活动。

表 2 虚拟人合成技术定量比较

方法	种类	IS	FID	LPIPS	SSIM	PSNR
People Snapshot (one shot)[87]						
StylePeople[76]	网格–图像	1.7469[76]	272.1[76]	0.0836[76]	0.9012[76]	–
LWGAN[75]	图像	1.7159[76]	1771.9[76]	0.2727[76]	0.2876[76]	–
360Degree[91]	网格	1.8643[76]	1383.1[76]	0.2123[76]	0.8079[76]	–
短多目RGB视频序列[78]						
DA[78]	体素–隐式	–	–	0.06[78]	0.95[78]	26.85[78]
FOMM[92]	图像	–	–	0.16[78]	0.91[78]	23.77[78]
DVP[93]	图像	–	–	0.10[78]	0.93[78]	25.67[78]
ZJU–MoCap[82]						
SANeRF[74]	网格–隐式	–	–	–	0.902[74]	24.42[74]
NB[82]	网格–隐式	–	–	0.0762[94]	0.885[74]	23.49[74]
NV[54]	体素	–	–	0.0999[94]	0.821[74]	21.39[74]
NeRF[16]	隐式	–	–	–	0.885[74]	23.41[74]
Market–1501[83]						
VU–Net[95]	图像	3.214[3]	20.144[96]	–	0.353[3]	–
DGANs[97]	图像	3.185[3]	25.364[3]	–	0.290[3]	–
PSG[98]	图像	3.750[3]	16.742[3]	–	0.732[3]	–
PPAT[99]	图像	3.323[3]	22.657[96]	–	0.311[3]	–
Human[85]						
TBN[100]	图像	–	52.262[85]	0.080[85]	–	–
pixelNeRF[101]	图像–隐式	–	61.453[85]	0.068[85]	–	–
STATE[85]	图像	–	57.055[85]	0.068[85]	–	–

表 3 虚拟人合成具体算法性能定性比较

方法	种类	目标部位	是否使用时间信息	保真度	泛化性
SMPL[21]	网格	身体	否	低	低
H4D[24]	网格	身体	是	中	中
GANFit[41]	图像	身体	否	中	低
NA[45]	图像	头部	否	中	低
Deep Human[53]	体素	身体	否	低	中
NV[54]	体素	头部	是	中	中
PIFuHD[65]	隐函数	身体	否	高	低
BANMo[59]	隐函数	身体	是	高	高
SCANImate[71]	网格–隐函数	身体	否	中	中
ICON[72]	网格–隐函数	身体	否	中	中
SANeRF[74]	网格–隐函数	身体	否	低	高
StylePeople[76]	网格–图像	身体	是	高	中
LWGAN[75]	网格–图像	身体	是	中	中
DA[78]	体素–隐函数	头部	是	高	中
PVA[77]	体素–隐函数	头部	否	高	低
S3[81]	网格–体素–隐函数	身体	是	高	高
PaMIR[79]	网格–体素–隐函数	身体	否	高	低

服务型虚拟人的一类常见应用是虚拟主播，常见于新闻播报、直播带货等场景。虚拟主播系统结合了语音合成、人像建模、形象驱动等多种技术，能够实现从文本到语音和视频的转化。国内代表性的工作有PaddleBoBo、讯飞的AI虚拟主播系统，支持用户通过图像和文本快速定制形象属性，生成虚拟主播视频。其中PaddleBoBo是基于深度学习框架PaddlePaddle开发的一个开源项目，在人像建模方面应用了PaddleGAN技术完成表情迁移、唇形合成等任务，以驱动虚拟人的面部活动。

身份型虚拟人常见的应用有游戏和社交工具，其作为现实人类在虚拟世界的化身。在游戏领域中，用户可以创建专属于自己的虚拟形象，并在游戏世界中活动。这类虚拟人的重点通常在于人体建模，强调高保真性，代表性的工具有Epic Games推出的MetaHuman Creator（MHC）[102]。作为虚幻引擎的线上应用，MHC支持用户在线制作虚拟人，通过Quixel Bridge导出并在游戏当中使用。MHC包含骨骼网格体、用于定义头部的Groom和细节等级LOD等重要组件，其面部数据来源于大量的现实图像扫描数据，身体建模则应用基于网格模型进行操作，在渲染方面主要采用了基于物理方法渲染以增强人物真实感。

而在社交领域，远程视讯是一个重要的应用场景，特别是出于某些隐私保护的目的，需要在视讯中用到虚拟化身。麻省理工学院媒体实验室开源了一个虚拟角色生成工具AIC[103]，其技术核心基于Machinoia项目[104]，通过分析输入的视频和音频信息，实时生成一个2D虚拟人。该工具可应用于医疗通话等隐私的社交场景，其中合成人脸模块应用了基于图像的合成技术，主要采用GAN完成。

在传媒和营销领域，为了极大地发掘虚拟形象的营销价值，在社会上发挥其独特的影响力，其推出的虚拟人不仅具有服务性功能[105]，还具备一定的社交属性，与人类进行互动。近年来，国内外纷纷在社交平台[106]和营销广告[107]中推出虚拟形象，进行内容产出，如虚拟超模Shudu Gram[108]、日本的虚拟模特imma、上海燃麦打造的超写实数字人AYAYI[109]以及北京次世和上海魔珐共同研发的虚拟形象“翎”。其中AYAYI是基于虚幻引擎开发的一个3D高保真虚拟人，其参与各大美妆品牌在天猫的推广活动，同时也进驻了社交媒体，由幕后人员进行账号的日常运营。

在关注多种虚拟人的应用方式时，也要关注虚拟形象带来的社会影响。未来随着元宇宙和现实生活的关系更加密切，现实人类还会继续通过各类接口工具创建具有自然表情神态[105]、模仿人体运动的虚拟化身，参与各类商业和娱乐活动。多项调查显示，以虚拟形象方式为主的交流方式[110]能更有效地表达自我，但同时也存在网络欺凌、身份欺诈等隐患[111-112]。社区管理者应合理利用社交数据加强对虚拟形象所有者的管理，平衡隐私保护和社区安全的治理。未来，随着合成技术进一步提升，更逼真的虚拟人还会继续以各种身份、各种形式活跃在人们的视线里。如何更合理地应用虚拟形象，是一个值得探讨的方向。

5 总结与展望

本文主要介绍了5类虚拟人合成技术，首先将虚拟形象合成技术围绕网格、体素、图像、隐式表示4类经典技术展开，再提出第5类混合表示方法展现传统技术间的优势互补思路。其中网格模型具有可视化训练和贴近人体骨骼构造的特点，在进行姿势模拟上具有优势。体素表示规

范，易于存储，可以处理任意拓扑的物体。基于图像的方法中GAN已经在人脸重建领域得到广泛应用，可以生成高质量的模型。隐式表示方法高效利用内存，其连续性支持解决复杂的服装建模任务，还可以拓展到多图像、多视角的输入数据。第5类的混合方法展示了近年来对经典合成技术的融会贯通，多方向逐步提高虚拟形象的质量。随后简要介绍了性能评价的数据集和评价指标，更具体地展现了方法的定量和定性比较。最后介绍了虚拟人的综合应用及其场景。目前来说，提高模型的泛化能力、提高时空效率仍是未来虚拟人构建技术的重要优化方向。

综合近年来相关文献的工作，提高虚拟形象质量的方向可以从模型的泛化能力和生成结果的质量等方向出发，可分为以下5类。

● 提高模型的泛化能力，体现在对自然场景下图片的处理和对未观察区域的外推能力的提升。应用于自然场景下的图片的重建模型要求具有足够的能力去捕捉照明光谱和背景变化，以处理复杂多变的照明环境和背景数据；而对未观察区域的外推能力的提升，可以结合全局场景先验知识和扩展多视图的合成进行增强。当已标记的3D几何数据充足时，主流的SMPL人体参数化模型能为SDF等隐式表示方法提供强大的先验知识；当可用数据限制时，已有工作[113]证明能够引入先进的NeRF技术通过单张多平面图片进行重建，实现多视图合成和深度估算。

● 提高数据质量和模型的数据利用率，大部分基于学习的构建工作仍受图像数据的质量和数量的限制。基于视频序列训练能得到更多样的不同姿势和服装的生成模型，但要考虑视频序列存在的偏差问题。此外，一些微表情和姿势变化依赖于参数化人体模型，在保证前置步骤精度的前提下充分利用现有数据推测更细微的人体变化。未来可以考虑生成更丰富的用于训练的数据集，能够提供良好的光照条件和中性表情，并具有高分辨率以体现丰富的细节。

● 动态的虚拟人头部的重建工作已经初有成效，而对于自然场景下虚拟人的全身形象的动态表示工作还需要更复杂的可变性模型以及光照计算工作，面临的挑战包括模拟多样化的人体姿势和运动时着装的形变，以及保证重新照明时的产生的视觉质量。

● 为了提高虚拟人全身模型的质量，需要考虑进一步丰富身体表面细节，包括面部微表情和附着物（如头发、衣物）的动态变化，可以尝试将人脸重建、运动预测等细分的成熟领域的技术进行迁移，进一步提高虚拟人的视觉质量。

● 提高精度的同时应节约计算耗时和内存成本，对基于体素类方法，改善内存效率一直是热门话题，应设计更合理的内存分配算法或寻找内存效率更高的表示方法，以进一步提高虚拟形象的质量和细节。

参考文献:

[1] 中国人工智能产业发展联盟总体组，中关村数智人工智能产业联盟数字人工作委员会. 2020年虚拟数字人发展白皮书[R]. 2020.
Artificial Intelligence Industry Alliance, Digital Human Work Committee of Zhongguancun Shuzhi Artificial Intelligence Industry Alliance. 2020 virtual digital human development white paper[R]. 2020.

[2] FU K, PENG J S, HE Q W, et al. Single image 3D object reconstruction based on deep learning: a review[J]. Multimedia Tools and Applications, 2021, 80(1): 463-498.

[3] SHA T, ZHANG W, SHEN T, et al. Deep person generation: a survey from the perspective of face, pose and cloth synthesis[J]. arXiv preprint, 2021, arXiv: 2109.02081.

[4] CHEN L, PENG S D, ZHOU X W. Towards efficient and photorealistic 3D human reconstruction: a brief survey[J]. Visual Informatics, 2021, 5(4): 11–19.
[5] JOEYDEVRIES. Textures[Z]. 2022.
[6] ZENG W, OUYANG W L, LUO P, et al. 3D human mesh regression with dense correspondence[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 7052–7061.
[7] GATYS L A, ECKER A S, BETHGE M. Texture synthesis using convolutional neural networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems New York: ACM Press, 2015: 262–270.
[8] RISSER E, WILMOT P, BARNES C. Stable and controllable neural texture synthesis and style transfer using histogram losses[J]. arXiv preprint, 2017, arXiv: 1701.08893.
[9] OECHSLE M, MESCHEDER L, NIEMEYER M, et al. Texture fields: learning texture representations in function space[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 4530–4539.
[10] 阿娣提·玛珠德, M. 戈皮. 视觉计算基础:计算机视觉、图形学和图像处理的核心概念[M]. 赵启军, 涂欢, 梁洁, 译. 北京: 机械工业出版社, 2019.
MAJUMDER A, GOPI M. Introduction to visual computing: core concepts in computer vision, graphics, and image processing[M]. Translated by ZHAO Q J, XU H, LIANG J. Beijing: China Machine Press, 2019.
[11] JONES A, GARDNER A, BOLAS M, et al. Simulating spatially varying lighting on a live performance[C]//Proceedings of 3rd European Conference on Visual Media Production and the 2nd Multimedia Conference 2006. [S.l.:s.n.], 2006: 127–133.
[12] PHONG B T. Illumination for computer generated pictures[J]. Communications of the ACM, 1975, 18(6): 311–317.
[13] JOEYDEVRIES. Normal mapping[Z]. 2022.
[14] JOEYDEVRIES. PBR:theory[Z]. 2022.
[15] 洪锋, 梅炯, 李明禄. 医学图象三维重建技术综述[J]. 中国图象图形学报(A辑), 2003, 8(z1): 784–791.
HONG F, MEI J, LI M L. Study on the techniques for 3D reconstruction of medical images[J]. Journal of Image and Graphics, 2003, 8(z1): 784–791.
[16] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. NeRF: representing scenes as neural radiance fields for view synthesis[C]//Proceedings of 2020 European Conference on Computer Vision. Cham: Springer, 2020: 405–421.
[17] XU Q G, XU Z X, PHILIP J, et al. Point-NeRF: point-based neural radiance fields[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 5428–5438.
[18] 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.
[19] KAVAN L, COLLINS S, ŽÁRA J, et al. Geometric skinning with approximate dual quaternion blending[J]. ACM Transactions on Graphics, 2008, 27(4): 1–23.
[20] JACOBSON A, BARAN I, POPOVIĆ J, et al. Bounded biharmonic weights for real-time deformation[J]. ACM Transactions on Graphics, 2011, 30(4): 1–8.
[21] 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.
[22] PAVLAKOS G, CHOUTAS V, GHORBANI N, et al. Expressive body capture: 3D hands, face, and body from a single image[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 10967–10977.
[23] WU S Z, JIN S, LIU W T, et al. Graph-based 3D multi-person pose estimation using multi-view images[C]//Proceedings of 2021 IEEE/CVF International Conference

on Computer Vision. Piscataway: IEEE Press, 2022: 11128–11137.

[24] JIANG B Y, ZHANG Y D, WEI X K, et al. H4D: human 4D modeling by learning neural compositional representation[C]// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 19333–19343.

[25] OSMAN A A A, BOLKART T, BLACK M J. STAR: sparse trained articulated human body regressor[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 598–613.

[26] XU H Y, BAZAVAN E G, ZANFIR A, et al. GHUM & GHUML: generative 3D human shape and articulated pose models[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 6183–6192.

[27] KINGMA D P, WELLING M. Auto-encoding variational bayes[J]. arXiv preprint, 2013, arXiv: 1312.6114.

[28] REZENDE D J, MOHAMED S. Variational inference with normalizing flows[J]. arXiv preprint, 2015, arXiv: 1505. 05770.

[29] BHATNAGAR B, TIWARI G, THEOBALT C, et al. Multi-garment Net: learning to dress 3D people from images[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 5419–5429.

[30] ALLDIECK T, PONS-MOLL G, THEOBALT C, et al. Tex2Shape: detailed full human body geometry from a single image[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 2293–2303.

[31] WENG C Y, CURLESS B, KEMELMACHER-SHLIZERMAN I. Photo wake-up: 3D character animation from a single photo[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 5901–5910.

[32] ALLDIECK T, MAGNOR M, XU W P, et al. Detailed human avatars from monocular video[C]//Proceedings of 2018 International Conference on 3D Vision. Piscataway: IEEE Press, 2018: 98–109.

[33] MA Q L, YANG J L, RANJAN A, et al. Learning to dress 3D people in generative clothing[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 6468–6477.

[34] ALLDIECK T, MAGNOR M, BHATNAGAR B L, et al. Learning to reconstruct people in clothing from a single RGB camera[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 1175–1186.

[35] JIANG B Y, ZHANG J Y, HONG Y, et al. BCNet: learning body and cloth shape from a single image[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 18–35.

[36] WEI W L, LIN J C, LIU T L, et al. Capturing humans in motion: temporal-attentive 3D human pose and shape estimation from monocular video[C]// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 13201–13210.

[37] 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: ACM Press, 1999: 187–194.

[38] LATTAS A, MOSCHOGLOU S, GECER B, et al. AvatarMe: realistically renderable 3D facial reconstruction "In-the-wild" [C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 757–766.

[39] ZHENG M W, YANG H Y, HUANG D, et al. ImFace: a nonlinear 3D morphable face model with implicit neural representations[C]//Proceedings of 2022

IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 20311–20320.

[40] ZHENG Y F, ABREVAYA V F, BÜHLER M C, et al. I M avatar: implicit morphable head avatars from videos[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 13535–13545.

[41] GECER B, PLOUMPIS S, KOTSIA I, et al. GANFIT: generative adversarial network fitting for high fidelity 3D face reconstruction[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 1155–1164.

[42] KARRAS T, LAINE S, AILA T M. A style-based generator architecture for generative adversarial networks[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 4396–4405.

[43] TEWARI A, ELGHARIB M, BHARAJ G, et al. StyleRig: rigging StyleGAN for 3D control over portrait images[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 6141–6150.

[44] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of StyleGAN[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 8107–8116.

[45] LUO H W, NAGANO K, KUNG H W, et al. Normalized avatar synthesis using StyleGAN and perceptual refinement[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 11657–11667.

[46] SHEN Y, LIANG J B, LIN M C. GAN-based garment generation using sewing pattern images[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 225–247.

[47] RAFFIEE A H, SOLLAMI M. GarmentGAN: photo-realistic adversarial fashion transfer[C]//Proceedings of 2020 25th International Conference on Pattern Recognition. Piscataway: IEEE Press, 2021: 3923–3930.

[48] CURLESS B, LEVOY M. A volumetric method for building complex models from ange images[C]//Proceedings of the 23rd Annual Conference on Computer graphics and Interactive Techniques. New York: ACM Press, 1996: 303–312.

[49] IZADI S, KIM D, HILLIGES O, et al. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera[C]//Proceedings of the 24th annual ACM symposium on User Interface Software and Technology. New York: ACM Press, 2011: 559–568.

[50] DAI A, NIEßNER M, ZOLLHÖFER M, et al. BundleFusion: real-time globally consistent 3D reconstruction using on-the-fly surface reintegration[J]. ACM Transactions on Graphics, 2017, 36(4): 76a.

[51] SITZMANN V, THIES J, HEIDE F, et al. DeepVoxels: learning persistent 3D feature embeddings[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 2432–2441.

[52] MA X X, SU J J, WANG C Y, et al. Context modeling in 3D human pose estimation: a unified perspective[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 6234–6243.

[53] ZHENG Z R, YU T, WEI Y X, et al. DeepHuman: 3D human reconstruction from a single image[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 7738–7748.

[54] LOMBARDI S, SIMON T, SARAGIH J,

et al. Neural volumes: learning dynamic renderable volumes from images[J]. ACM Transactions on Graphics, 2019, 38(4): 1–14.

[55] MESCHEDER L, OECHSLE M, NIEMEYER M, et al. Occupancy networks: learning 3D reconstruction in function space[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 4455–4465.

[56] PARK J J, FLORENCE P, STRAUB J, et al. DeepSDF: learning continuous signed distance functions for shape representation[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 165–174.

[57] CHEN Z Q, ZHANG H. Learning implicit fields for generative shape modeling[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 5932–5941.

[58] SITZMANN V, ZOLLHÖFER M, WETZSTEIN G. Scene representation networks: continuous 3D-structure-aware neural scene representations[J]. arXiv preprint, 2019, arXiv: 1906.01618.

[59] YANG G S, VO M, NEVEROVA N, et al. BANMo: building animatable 3D neural models from many casual videos[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 2853–2863.

[60] NEVEROVA N, NOVOTNY D, KHALIDOV V, et al. Continuous surface embeddings[J]. arXiv preprint, 2020, arXiv: 2011.12438.

[61] PARK J J, FLORENCE P, STRAUB J, et al. DeepSDF: learning continuous signed distance functions for shape representation[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 165–174.

[62] BOžIČ A, PALAFOX P, ZOLLHÖFER M, et al. Neural deformation graphs for globally-consistent non-rigid reconstruction[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 1450–1459.

[63] MITTAL P, CHENG Y C, SINGH M, et al. AutoSDF: shape priors for 3D completion, reconstruction and generation[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 306–315.

[64] SAITO S, HUANG Z, NATSUME R, et al. PIFu: pixel-aligned implicit function for high-resolution clothed human digitization[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 2304–2314.

[65] SAITO S, SIMON T, SARAGIH J, et al. PIFuHD: multi-level pixel-aligned implicit function for high-resolution 3D human digitization[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 81–90.

[66] JADHAV O, PATIL A, SAM J, et al. Virtual dressing using augmented reality[J]. ITM Web of Conferences, 2021, 40.

[67] ZHU X, LIAO T, LYU J, et al. MVP-human dataset for 3D human avatar reconstruction from unconstrained frames[J]. arXiv preprint, 2022, arXiv: 2204.11184.

[68] DENG B Y, LEWIS J P, JERUZALSKI T, et al. NASA Neural articulated shape approximation[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 612–628.

[69] CAO Y K, CHEN G Y, HAN K, et al. JIFF: jointly-aligned implicit face function for high quality single view clothed human reconstruction[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 2719–2729.

[70] BHATNAGAR B L, SMINCHISESCU C, THEOBALT C, et al. Combining implicit function learning and parametric models for 3D human reconstruction[C]// Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 311–329.

[71] SAITO S, YANG J L, MA Q L, et al. SCANimate: weakly supervised learning of skinned clothed avatar networks[C]// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 2885–2896.

[72] XIU Y L, YANG J L, TZIONAS D, et al. ICON: implicit clothed humans obtained from normals[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 13286–13296.

[73] ZHENG Z R, HUANG H, YU T, et al. Structured local radiance fields for human avatar modeling[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 15872–15882.

[74] XU T H, FUJITA Y, MATSUMOTO E. Surface-aligned neural radiance fields for controllable 3D human synthesis[C]// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 15862–15871.

[75] LIU W, PIAO Z X, MIN J, et al. Liquid warping GAN: a unified framework for human motion imitation, appearance transfer and novel view synthesis[C]// Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 5903–5912.

[76] GRIGOREV A, ISKAKOV K, IANINA A, et al. StylePeople: a generative model of fullbody human avatars[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 5147–5156.

[77] RAJ A, ZOLLHÖFER M, SIMON T, et al. Pixel-aligned volumetric avatars[C]// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 11728–11737.

[78] GAFNI G, THIES J, ZOLLHÖFER M, et al. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 8645–8654.

[79] ZHENG Z R, YU T, LIU Y B, et al. PaMIR: parametric model-conditioned implicit representation for image-based human reconstruction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3170–3184.

[80] ZHENG Y, SHAO R Z, ZHANG Y X, et al. DeepMultiCap: performance capture of multiple characters using sparse multiview cameras[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 6219–6229.

[81] YANG Z, WANG S L, MANIVASAGAM S, et al. S3: neural shape, skeleton, and skinning fields for 3D human modeling[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 13279–13288.

[82] PENG S D, ZHANG Y Q, XU Y H, et al. Neural body: implicit neural representations with structured latent codes for novel view synthesis of dynamic humans[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 9050–9059.

[83] ZHENG L, SHEN L Y, TIAN L, et al. Scalable person re-identification: a benchmark[C]//Proceedings of 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2016: 1116–1124.

[84] IONESCU C, PAPAVA D, OLARU V, et al. Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(7): 1325–1339.

[85] JING X Y, FENG Q, LAI Y K, et al. STATE: learning structure and texture representations for novel view synthesis[C]//Proceedings of IEEE International Conference on Computer Vision. [S.l.:s.n.], 2022.

[86] PATEL P, HUANG C H P, TESCH J, et al. AGORA: avatars in geography optimized for regression analysis[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 13463–13473.

[87] ALLDIECK T, MAGNOR M, XU W P, et al. Video based reconstruction of 3D people models[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8387–8397.

[88] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 586–595.

[89] BROWNLEE J. How to implement the frechet inception distance (FID) for evaluating GANs[Z]. 2019.

[90] SETIADI D R I M. PSNR vs SSIM: imperceptibility quality assessment for image steganography[J]. Multimedia Tools and Applications, 2021, 80(6): 8423–8444.

[91] LAZOVA V, INSAFUTDINOV E, PONS-MOLL G. 360-degree textures of people in clothing from a single image[C]// Proceedings of 2019 International Conference on 3D Vision. Piscataway: IEEE Press, 2019: 643–653.

[92] SIAROHIN A, LATHUILIÈRE S, TULYAKOV S, et al. First order motion model for image animation[J]. Advances in Neural Information Processing Systems, 2019, 32.

[92] SIAROHIN A, LATHUILIÈRE S, TULYAKOV S, et al. First order motion model for image animation[J]. arXiv preprint, 2020, arXiv: 2003.00196.

[93] KIM H, GARRIDO P, TEWARI A, et al. Deep video portraits[J]. ACM Transactions on Graphics, 2018, 37(4): 1–14.

[94] ZHAO F Q, YANG W, ZHANG J K, et al. HumanNeRF: efficiently generated human radiance field from sparse inputs[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 7733–7743.

[95] ESSER P, SUTTER E. A variational U-net for conditional appearance and shape generation[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8857–8866.

[96] REN Y R, YU X M, CHEN J M, et al. Deep image spatial transformation for person image generation[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 7687–7696.

[97] SIAROHIN A, SANGINETO E, LATHUILIÈRE S, et al. Deformable GANs for pose-based human image generation[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 3408–3416.

[98] LIU M C, WANG K J, JI R H, et al. Pose transfer generation with semantic parsing attention network for person re-identification[J]. Knowledge-Based Systems, 2021, 223.

[99] ZHU Z, HUANG T T, SHI B G, et al.

Progressive pose attention transfer for person image generation[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 2342-2351.

[100]OLSZEWSKI K, TULYAKOV S, WOODFORD O, et al. Transformable bottleneck networks[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 7647-7656.

[101]YU A, YE V, TANCIK M, et al. pixelNeRF: neural radiance fields from one or few images[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 4576-4585.

[102]FANG Z X, CAI L B, WANG G. MetaHuman creator the starting point of the metaverse[C]//Proceedings of 2021 International Symposium on Computer Technology and Information Science. Piscataway: IEEE Press, 2021: 154-157.

[103]PATARANUTAPORN P, DANRY V, LEONG J, et al. AI-generated characters for supporting personalized learning and well-being[J]. Nature Machine Intelligence, 2021, 3(12): 1013-1022.

[104]PATARANUTAPORN P, DANRY V, MAES P. Machinoia, machine of multiple me: integrating with past, future and alternative selves[C]//Proceedings of Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2021: 1-7.

[105]KATO R, KIKUCHI Y, YEM V, et al. Reality Avatar for Customer Conversation in the Metaverse[C]//Proceedings of International Conference on Human-Computer Interaction. Cham: Springer, 2022: 131-145.

[106]CONTI M, GATHANI J, TRICOMI P P. Virtual influencers in online social media[J]. IEEE Communications Magazine, 2022, 60(8): 86-91.

[107]SILVA E S, BONETTI F. Digital humans in fashion: will consumers interact?[J]. Journal of Retailing and Consumer Services, 2021, 60.

[108]KÁDEKOVÁ I Z, HOLIENČINOVÁ I M. Influencer marketing as a modern phenomenon creating a new frontier of virtual opportunities[J]. Communication Today, 2018, 9(2): 90-105.

[109]沈浩, 刘亭利. 虚实共融，若即若离：全面进击的虚拟数字人[J]. 视听界, 2022(3): 5-10.
SHEN H, LIU T L. Integration of reality and reality, at arm's length: an all-round attack on virtual digital people[J]. Broadcasting Realm, 2022(3): 5-10.

[110]PARK I, SAH Y J, LEE S, et al. Avatar-mediated communication in video conferencing: effect of self-affirmation on debating participation focusing on moderation effect of avatar[J]. International Journal of Human-Computer Interaction, 2023, 39(3): 464-475.

[111]TAKANO M, YOKOTANI K. Online social support via avatar communication buffers harmful effects of offline bullying victimization[J]. Proceedings of the International AAAI Conference on Web and Social Media, 2022, 16: 980-992.

[112]CHEONG B C. Avatars in the metaverse: potential legal issues and remedies[J]. International Cybersecurity Law Review, 2022, 3(2): 467-494.

[113]LI J X, FENG Z J, SHE Q, et al. MINE: towards continuous depth MPI with NeRF for novel view synthesis[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 12558-12568.

作者简介

邓钇敏（1999– ），女，中国科学技术大学硕士生，中国计算机学会会员，主要研究方向为深度学习、计算机视觉、元宇宙等。

张旭龙（1988– ），男，博士，平安科技（深圳）有限公司高级算法研究员，主要研究方向为语音合成、语音转换、音乐信息检索、机器学习和深度学习方法在人工智能领域应用。

司世景（1988– ），男，博士，平安科技（深圳）有限公司资深算法研究员，深圳市海外高层次人才。美国杜克大学人工智能博士后，中国计算机学会会员，主要研究方向为机器学习和及其在人工智能领域应用。

王健宗（1983– ），男，博士，平安科技（深圳）有限公司副总工程师，资深人工智能总监，联邦学习技术部总经理。美国佛罗里达大学人工智能博士后，中国计算机学会高级会员，中国计算机学会大数据专家委员会委员，曾任美国莱斯大学电子与计算机工程系研究员，主要研究方向为联邦学习和人工智能等。

肖京（1972– ），男，博士，中国平安集团首席科学家，2019年吴文俊人工智能杰出贡献奖获得者，中国计算机学会深圳分部副主席，主要研究方向为计算机图形学学科、自动驾驶、3D显示、医疗诊断、联邦学习等。

收稿日期： 2022-05-25
通信作者： 王健宗，jzwang@188.com
基金项目： 广东省重点领域研发计划“新一代人工智能”重大专项”（No.2021B0101400003）
Foundation Item: The Key Research and Development Program of Guangdong Province (No.2021B0101400003)