中图法分类号:TP391　　文献标识码:A　　文章编号:1006-8961(2024)09-2541-25

论文引用格式:Zhao B Q, Fu Y Y, Su Z, Wang R M, Lyu C L and Luo X N. 2024. A survey on multimodal information-guided 3D human motion generation. Journal of Image and Graphics, 29(09):2541-2565(赵宝全, 付一愉, 苏卓, 王若梅, 吕辰雷, 罗笑南. 2024. 多模态信息引导的三维数字人运动生成综述. 中国图象图形学报, 29(09):2541-2565)[DOI:10.11834/jig.230626]

多模态信息引导的三维数字人运动生成综述

赵宝全[1],付一愉[1],苏卓[2*],王若梅[2],吕辰雷[3],罗笑南[4]

1. 中山大学人工智能学院,珠海　519000; 2. 中山大学计算机学院,广州　510006;
3. 深圳大学计算机与软件学院,深圳　518060; 4. 桂林电子科技大学计算机与信息安全学院,桂林　541004

摘　要: 基于多模态信息的三维数字人运动生成技术旨在通过文本、音频、图像和视频等数据实现特定输入条件下的人体运动生成。这项技术在电影、动画、游戏制作和元宇宙等领域具有重要的应用价值和广泛的经济社会效益,是近年来计算机图形学和计算机视觉等领域研究的热点问题之一。然而,基于多模态信息的三维数字人运动生成面临着诸多挑战,包括跨模态信息的表征和融合困难、高质量数据集缺乏、生成的运动质量较差(如抖动、穿模和脚部滑动等)以及生成效率低等问题。虽然近年来研究者们提出了各式各样的解决方案来应对上述挑战,但如何根据不同模态数据的特点实现高效、高质量的三维数字人运动生成仍然是一个开放性问题。本文以数字人运动生成所采用的模型架构为分类标准,将现有的主流方法分为基于生成对抗网络(generative adversarial network,GAN)的方法、基于自编码器(autoencoder,AE)的方法、基于变分自编码器(variational autoencoder,VAE)的方法以及基于扩散模型的方法,总结并形成了一种数字人运动生成通用框架。本文还介绍了该领域常见的参数化人体模型、数据集以及评估指标。对于一些具有代表性的工作,本文在一些常用数据集上进行了对比实验,评估这些方法的性能表现。最后综合现有的数据集、算法和代表性研究,总结了该领域的问题和挑战,探讨了完善数据集、优化运动质量和多样性、融合跨模态信息和提高生成效率等潜在的研究方向。

关键词: 三维数字人;运动生成;多模态信息;参数化人体模型;生成对抗网络(GAN);自编码器(AE);变分自编码器(VAE);扩散模型

A survey on multimodal information-guided 3D human motion generation

Zhao Baoquan[1], Fu Yiyu[1], Su Zhuo[2*], Wang Ruomei[2], Lyu Chenlei[3], Luo Xiaonan[4]

1. School of Artificial Intelligence, Sun Yat-sen University, Zhuhai　519000, China;
2. School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou　510006, China;
3. College of Computer Science and Software Engineering, Shenzhen University, Shenzhen　518060, China;
4. School of Computer and Information Security, Guilin University of Electronic Science and Technology, Guilin　541004, China

Abstract: Three-dimensional (3D) digital human motion generation guided by multimodal information generates human motion under specific input conditions through data, such as text, audio, image, and video. This technology has a wide spectrum of applications and extensive economic and social benefits in the fields of film, animation, game production,

收稿日期:2023-09-05;修回日期:2024-01-26;预印本日期:2024-02-02

*通信作者:苏卓　suzhuo3@mail.sysu.edu.cn

基金项目:国家重点研发计划资助(2022YFF0903103);广东省自然科学基金项目(2023A1515011639);中央高校基本科研业务费专项资金资助(23xkjc019;24qnpy145)

Supported by: National Key R&D Program of China (2022YFF0903103); Natural Science Foundation of Guangdong Province, China (2023A1515011639); Fundamental Research Funds for the Central Universities (23xkjc019;24qnpy145)

metaverse, etc., and is one of the research hotspots in the fields of computer graphics and computer vision. However, such a task faces grand challenges, including the difficult representation and fusion of multimodal information, lack of high-quality datasets, poor quality of generated motion (such as jitter, penetration, and foot sliding), and low generation efficiency. Although various solutions have been proposed to address the aforementioned challenges, a mechanism for achieving efficient and high-quality 3D digital human motion generation based on the characteristics of distinct modal data remains an open problem to be solved. This paper comprehensively reviews 3D digital human motion generation and elaborates on related recent advances from the perspectives of parametrized 3D human models, human motion representation, motion generation techniques, motion analysis and editing, existing human motion datasets and evaluation metrics. Parametrized human models facilitate digital human modeling and motion generation through the provision of parameters associated with body shapes and postures and serve as key pillars of current digital human research and applications. This survey begins with an introduction to widely used parametrized 3D human body models, including shape completion and animation of people (SCAPE), skinned multi-person linear model (SMPL), SMPL-X, and SMPL-H, and their detailed comparison in terms of model representations and the parameters used to control body shapes, poses, and facial expressions. Human motion representation is a core issue in digital human motion generation. This work highlights the musculoskeletal model and classic skinning algorithms, including linear blending skinning and dual quaternion skinning, and their application in physics-based and data-driven methods to control human movements. We have also extensively studied approaches to existing multimodal information-guided human motion generation and categorized them into four major branches, i. e., generative adversarial network-, autoencoder-, variational autoencoder-, and diffusion model-based methods. Other works, such as generative motion matching, have also been mentioned and compared with data-driven methods. The survey summarizes existing schemes of human motion generation from the perspectives of methods and model architectures and presents a unified framework for the generation of digital human motion. A motion encoder extracts motion features from an original motion sequence and fuses them with the conditional characteristics extracted by the conditional encoder into latent variables or maps them to the latent space. This condition enables generative adversarial networks, autoencoders, variational autoencoders, or diffusion models to generate qualified human movements through a motion decoder. In addition, this paper surveys the current work on digital human motion analysis and editing, including motion clustering, motion prediction, motion in-betweening, and motion in-filling. Data-driven human motion generation and evaluation requires the use of a high-quality dataset. We collected publicly available human motion databases and classified them into various types based on two criteria. From the perspective of data type, existing databases can be classified into motion capture and video reconstruction datasets. Motion capture data sets rely on devices, such as motion capture systems, cameras, and inertial measurement units, to obtain real human movement data (i. e., ground truth). Meanwhile, the video reconstruction dataset was used to reconstruct a 3D human body model through estimation of body joints from motion videos and fitting them to a parametric human body model. From the perspective of task type, commonly used databases can be classified into text-, action-, and audio-motion datasets. The new datasets are usually obtained by processing motion capture and video reconstruction datasets based on specific tasks. A comprehensive briefing on the evaluation metrics of 3D human motion generation, including motion quality, motion diversity, and multimodality, consistency between inputs and outputs, and inference efficiency, is also provided. Apart from objective evaluation metrics, user study was employed to generate human motion quality and was discussed in this paper. To compare the performances of various generation methods used in digital human motion on public datasets, we selected a collection of the most representative work and carried out extensive experiments for comprehensive evaluation. Finally, the well-addressed and underexplored issues in this field were summarized, and several potential further research directions regarding datasets, the quality and diversity of generated motions, cross-modal information fusion, and generation efficiency were discussed. Specifically, existing datasets generally fail to meet the expectations concerning motion diversity and descriptions associated with motions, data distribution, and length of motion sequence. Future work should consider the development of a large-scale 3D human motion database to boost the efficacy and robustness of motion generation models. In addition, the quality of generated human motions, especially those with complex movement patterns, remains dissatisfactory. Physical constraints and postprocessing show promise in the integration into human motion generation frameworks to tackle issues. In addition, although human-motion generation methods

can generate various motion sequences from multimodal information, such as text, audio, music, actions and keyframes, work on cross-modal human motion generation (e. g. , generating a motion from a text description and a piece of background music) is scarcely reported. Investigation of such a task is worthy, especially in unlocking new opportunities in this area. In terms of the diversity of generated content, some researchers have explored harvesting rich, diverse, and stylized motions using variational autoencoders, diffusion models, and contrastive language-image pretraining neural networks. However, current studies mainly focus on the motion generation of a single human represented by an SMPL-like naked parameterized 3D model. Meanwhile, the generation and interaction of multiple dressed humans have huge untapped application potential but have not received sufficient attention. Finally, another nonnegligible issue is a mechanism for boosting motion generation efficiency and achieving a good balance between quality and inference overhead. Possible solutions to such a problem include lightweight parameterized human models, information-intensive training datasets, and improved or more advanced generative frameworks.

Key words: 3D avatar; motion generation; multimodal information; parametric human model; generative adversarial network (GAN); autoencoder (AE); variational autoencoder (VAE); diffusion model

0 引　言

数字人是通过计算机图形学、计算机视觉、动作建模与仿真、语音合成和人工智能等技术手段创建的具有外观、语言和行为等多重人类特征的虚拟形象。其中，高效、高质量的运动生成不仅是数字人制作的核心要素之一，也是满足众多应用场景中不同用户多样化需求的内在要求。数字人运动生成技术旨在通过文本、音频、图像、视频、动作类别和属性等多模态信息来控制数字人完成特定的运动。

基于物理模拟的数字人运动生成方法是一种利用物理学原理和方法，包括动力学、前向运动学、逆向运动学、碰撞检测与响应等，来对人体关节点施加力和约束，从而实现对人体运动的模拟方法。其主要优势在于能够模拟真实环境，使得数字人的运动更加真实，可以更自然地与周围环境和物体交互(Taheri等，2022)，从而避免由于数据不足而无法完成特定动作的问题。然而，基于物理的方法在仿真和控制方面也面临一些挑战。首先，这些方法通常需要对真实环境做出一定的假设和简化，因此可能会丢失一些细节。其次，在求解物理方程时，采用的是离散化的数值方法，因此只能对真实世界中的连续运动进行近似。最后，设计人物角色的控制器以确保其能够适应环境变化也是一项具有挑战性的任务。一种主流的解决方法是将这些物理方法与强化学习相结合(Peng 等，2018；Bergamin 等，2019；Lee 等，2021；Kwiatkowski 等，2022)，训练智能体来响应不同的物理条件，实现具有物理真实感的数字人运动生成和控制。

工业界常用的基于关键帧的方法需要依靠专业的建模和动画工作者，使用复杂的建模软件调整部分关键帧中数字人的姿态。然而，这种方法难以满足工业界对于快速、批量制作的需求。

随着人工智能生成内容(AI generated content，AIGC)技术的快速发展，以文本、语音、图像、视频、动作类别和属性等多模态信息为主要驱动数据的深度人体运动生成技术引发了广泛关注。这种技术的出现不仅降低了数字人制作的技术门槛，还大幅提升了内容创作效率和人机交互体验，推动了数字人在影视、动画、游戏、传媒和虚拟现实等多个领域中的应用。

根据生成方式分类，数字人运动生成(motion generation)可以分为条件生成(conditional generation)和无条件生成(unconditional generation)(Petrovich 等，2022)。其中，无条件生成是指模拟整个可能的运动空间分布，再从分布中采样产生不同的运动序列。这种生成方式具有随机性，因为事先并不知道动作空间的分布，也可以通过在潜在空间中使用聚类方法将不同类别的动作进行划分，进而通过采样获得特定的动作。而条件生成是指根据各种条件输入，如文本描述、音频、关键帧图像、动作类别和属性等，生成满足指定要求的人体运动序列。

根据数据模态分类，如图 1 所示，主流的数字人运动生成任务可以分为文本生成运动(text to motion)、音频生成运动(audio to motion)和动作生成运动(action to motion)。其中，文本生成运动是指根据指定的人体运动的自然语言描述(如："站立演奏

小提琴”)产生满足输入条件的运动。音频生成运动是指根据输入音频中包含的语音、音乐等信息生成数字人产生相应的运动。例如,根据演讲者的语音生成相应的手势动作(speech to gesture)或根据一段音乐或者韵律生成舞蹈序列(music to dance)等。Ao等人(2022)提出,手势在说话中起到辅助表达的作用,包含韵律信息(如节奏、速度)和语义信息(用手势辅助解释或表达情感),但这些信息都是难以被定义的。同时,手势的表达还存在多义性,同一句话可以使用不同的手势,同一个手势也可以对应不同的话语。从音乐生成舞蹈的挑战性在于生成的运动要同时满足舞蹈的艺术性(动作真实、流畅)和音乐的一致性(动作要与音乐的风格、节奏和旋律相一致)。动作生成运动是指输入一个动作的类别,产生对应该动作种类的运动。可以将它看做是动作识别(根据运动识别动作类别)的反任务。这里要注意“动作”和“运动”的区别,前者侧重于概括性描述动作的种类,后者侧重于刻画一段具体的动作序列。例如,“打篮球”是动作,它是对篮球运动的抽象描述,而生成的运动可能包含运球、上篮、扣篮等具体的动作。

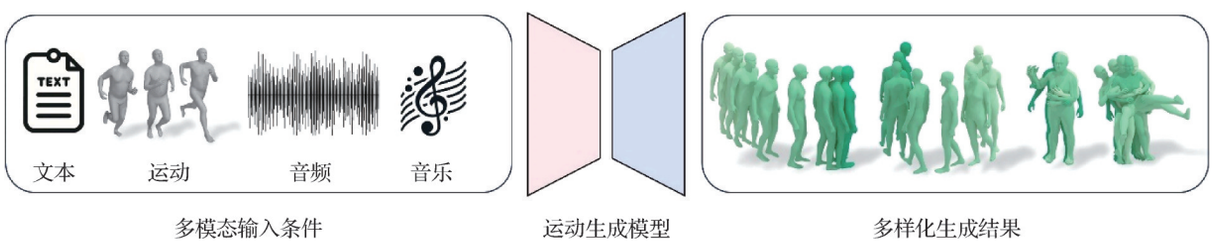


图1 多模态信息引导的三维人体运动生成

Fig. 1 Multimodal information guided 3D human motion generation

数字人动作生成领域面临着诸多挑战。以文本生成运动任务为例,主要存在以下问题:1)生成的结果缺乏多样性,即相同输入条件下可以产生多种不同的输出结果。例如,以“一个人在挥手”作为动作描述输入,生成的结果应在左右手的选择、挥手速度和幅度等方面呈现出多样化,从而获得更加丰富、多变的运动序列。2)复杂输入条件下的生成结果质量不佳。人体运动复杂多样,现有的大多数方法在面临较复杂的输入条件时,无法充分理解多模态数据中蕴含的动作语义信息,导致生成运动质量欠佳。3)长序列运动生成效果不理想。受限于训练数据中最长运动序列的时长,以及模型的学习和生成能力,当前的数字人运动生成方法无法生成超过一定时长的运动,或者随着序列的增长,生成运动的质量愈发不稳定。4)生成的结果缺乏真实性。生成运动存在一些肢体抖动、脚部滑动(foot slide)、漂浮、关节异位等不符合物理规律和人体运动学的结果,会产生强烈的不真实感。5)缺乏高质量数据集。首先,现有的数据集大部分来源于标注的动作捕捉数据,其中存在的噪声会严重干扰模型的推理过程,进而导致生成的动作会出现伪影(artifacts)和畸变;其次,数据集的分布不均衡会导致模型对于特定的生成动作很难有较好的表现(如大部分数据集中“走路”、“跑步”等常规动作的数量要远大于“瑜伽”、“游泳”等动作的数量),模型在训练阶段没有或者很少从这些动作中学习到特征,自然很难生成相应的高质量的人体运动。6)模型的生成效率亟待进一步优化和提高。现有的扩散模型和一些集成方法虽然具有优异的表达能力,但训练和推理都需要大量的计算资源和运算时间,无法较好地满足用户在硬件资源受限条件下对生成效率的要求。

本文以数字人运动生成的方法为主线,对近年来的主流算法、参数化人体模型、数据集和评估指标进行系统综述,总结归纳常用方法的设计思路优缺点,讨论该领域当下面临的挑战,并展望未来的研究方向。

1 参数化人体模型

参数化人体模型是基于参数化的方法构建的人体模型,通过改变模型的参数值可以实现对人体形状和动作姿态的灵活控制。参数化人体模型在简化人体建模流程的同时,还能提高模型的灵活性和逼真度,是数字人建模和运动生成的重要基础,已经成为目前数字人研究和应用的基石。

SCAPE (shape completion and animation of people)(Anguelov等,2005)是一种早期提出的、基于

三角面片变形的参数化人体模型。在该模型中，由姿态和身体形状变化引起的变形是分别建模的。SMPL（skinned multi-person linear model）（Loper 等，2015）是一种裸体的、基于顶点的三维人体模型，主要通过参数β描述人体的身材形状（如高矮胖瘦）以及通过参数θ描述人体的动作姿态。SMPL模型包含6 890个顶点和24个关节点，通过训练可以更好地拟合人体的形状和不同姿态下的形变，真实地模拟人的肌肉在肢体运动过程中的凸起和凹陷。由于SMPL模型与Blender、Maya、Unity3D等图形软件兼容，因此在数字人建模和运动生成领域得到广泛应用。

SMPLify（Bogo 等，2016）是第1个基于单目图像重建三维人体模型的方法，它使用DeepCut卷积神经网络从单目图像中估计出人体的二维关节参数，然后将SMPL模型拟合到这些关节点上。SMPLify训练了一个从SMPL的形状参数得到“胶囊”轴长和半径的回归器来近似人体部位，然后根据运动链引起的旋转对“胶囊”进行摆位，有效地解决了身体部位穿模问题。SMPLify在SMPL的男性模型和女性模型的基础上训练了中性性别的模型，扩展了SMPL模型的应用场景。Pavlakos 等人（2019）在SMPL模型和SMPLify方法的基础上提出了SMPL-X模型和一个基于单目RGB图像重建三维人体模型的方法SMPLify-X。相较于SMPL模型，SMPL-X增加了人物的手部和面部细节，由10 475个顶点和54个关节组成。SMPLify-X通过OpenPose（Cao 等，2021）检测骨骼、手部和面部的关键点，估计出人体二维特征，然后使用SMPL-X模型拟合这些关键点来重建人体。相较于SMPLify，它定义了一种新的既快速又准确的穿模惩罚，并训练了一个性别分类器，是首个可以在拟合三维人体姿势时自动考虑性别因素的方法。

人类的手部拥有大量关节，可以完成多种复杂的动作。一些场景中对于手部建模有很高的要求，例如，手部抓取识别（Zheng 等，2023）、手部目标跟踪（Chen 等，2023a）等。以往的人体手部建模存在两个问题：1）手和身体被当做独立的对象分别建模，但现实中人类的手势与身体其他部位的运动是高度协调的；2）三维人体模型的扫描往往存在视角受限且分辨率低引发的手部遮挡、存在噪声以及局部信息缺失等问题，导致手指难以分辨。为了解决这些问题，Romero 等人（2017）从高分辨率的手部数据集中学习到了手部参数化模型MANO（hand model with articulated and non-rigid deformations），然后将MANO与SMPL模型合并得到SMPL+H模型，该模型能够高精度、协调地表示身体运动和手部姿势。图2展示了常用的参数化人体模型SMPL-X、SMPL和SMPL-H，参数化模型可以由包含了姿态参数θ、体型参数β和表情参数ψ的函数M表示（见表1）。可以发现相较于SMPL和SMPL-H，SMPL-X新增了表情参数，这使得模型可以清晰表现出五官特征；相较于SMPL，SMPL-H有更加丰富的手部细节，手指可以完成更复杂的动作。

参数化人体模型的表达能力随着参数量的增加不断增强，但过多的参数意味着更高的训练难度和成本，一些工作开始研究更加轻量的参数化人体模型。STAR（sparse trained articulated human body regressor）（Osman 等，2020）为SMPL模型的运动树中每个关节都定义了一个姿势修正函数，这些函数从

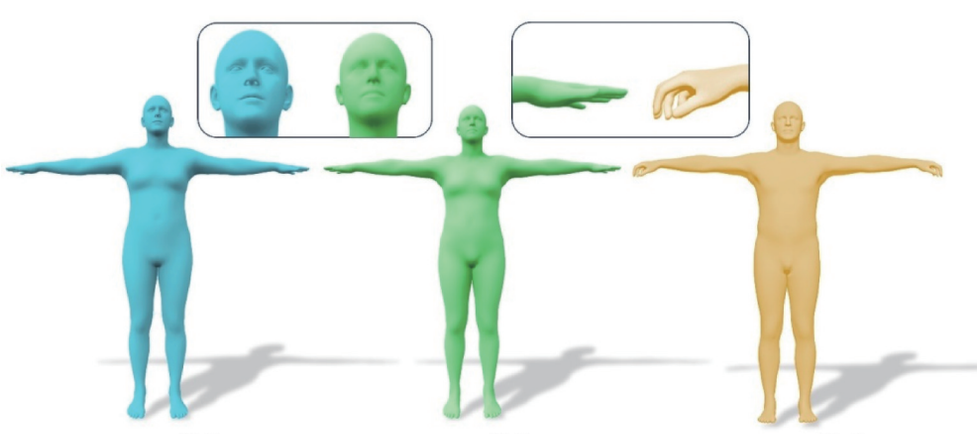


图2　典型的参数化三维人体模型

Fig. 2　Classic 3D human parametric models

表1　SMPL-X、SMPL和SMPL-H模型的表示与参数

Table 1　Representation and parameters of the SMPL-X, SMPL, and SMPL-H models

参数类型	SMPL-X	SMPL	SMPL-H
模型表示M	$M(\theta,\beta,\psi):\mathbf{R}^{\lvert\theta\rvert\times\lvert\beta\rvert\times\lvert\psi\rvert}\rightarrow\mathbf{R}^{3N}$	$M(\theta,\beta):\mathbf{R}^{\lvert\theta\rvert\times\lvert\beta\rvert}\rightarrow\mathbf{R}^{3N}$	$M(\theta,\beta):\mathbf{R}^{\lvert\theta\rvert\times\lvert\beta\rvert}\rightarrow\mathbf{R}^{3N}$
顶点数N	N = 10 475	N = 6 890	N = 6 890
姿态参数θ	$\lvert\theta\rvert$ = 75	$\lvert\theta\rvert$ = 72	$\lvert\theta\rvert$ = 78
体型参数β	$\lvert\beta\rvert$ = 24	$\lvert\beta\rvert$ = 10	$\lvert\beta\rvert$ = 20
表情参数ψ	$\lvert\psi\rvert$ = 10	–	–

注：“–”表示未提供该参数信息。

人体运动数据中学习，用于校正身体形状和姿势的相关变形，比SMPL模型的人体表示效果更加真实，而且模型参数量降低到SMPL的20%，在使用中具有更高的效率，且不太可能出现过拟合的问题。EllipBody（Wang等，2020）是一种可微渲染的轻量级参数化人体模型，以身体不同部位的长度、厚度和方向作为参数，使用多个椭球体表示身体的不同部位。这种局部分割的人体表示方法一方面能处理三维人体重建中不同身体部位的遮挡问题，另一方面可以减少参数量。该方法也可以通过拟合SMPL模型的形状参数，将椭圆体转换为SMPL人体模型。表2总结了常用的参数化人体模型及其特点。

表2　常用的参数化人体模型

Table 2　Commonly used parameterized human models

模型名称	特点	关节点数量	顶点数量
SCAPE（Anguelov等，2005）	早期提出的参数化人体模型	16	–
SMPL（Loper等，2015）	当前使用最广泛的参数化人体模型之一	24	6 890
SMPLify（Bogo等，2016）	提出了SMPL中性模型	24	6 890
SMPL+H（Romero等，2017）	将手部参数化模型MANO集成到SMPL模型上	73	6 890
SMPL-X（Pavlakos等，2019）	为SMPL模型增加了手部动作和面部表情，可以自动判别性别	54	10 475
STAR（Osman等，2020）	轻量化参数化人体模型，参数量降低到SMPL的20%	24	6 890
EllipBody（Wang等，2020）	可微渲染的轻量级参数化人体模型，可以转换为SMPL模型	–	–

注：“–”表示未提供该参数信息。

2　人体运动的表示

人体模型由大量的顶点组成，顶点越多，模型的表示越精细。但在运动中同时控制这些顶点会严重降低计算效率。在生物学上，骨骼是人体的支撑结构，关节（joint）是不同骨骼的连接，皮肤是外在表现，图3展示了Lee等人（2019）提出的肌肉骨骼模型，树状组织的关节包括8个转动关节（包括膝盖和肘部）和14个球窝关节（包括臀部、脚踝、肩膀和手腕）。借助人体骨骼、关节和皮肤的关系，通过控制关节点的移动和旋转可以实现对人体运动的控制。通过蒙皮（skinning）可以将骨骼和附近的顶点绑定，给不同的顶点分配合适的权重，确保皮肤在运动时的形状是自然的，主流的蒙皮算法有线性蒙皮（linear blending skinning，LBS）和双四元数蒙皮（dual quaternion skinning，DQS）。人体动作表示和数字人运动生成的核心问题是对关节点的控制。

在人体运动生成问题中，关节点通常以运动学树（kinematic trees）的层次结构组织的。通过定义父子节点的相对位置（三维坐标）和旋转（四元数）来控制运动。这种表示方法有利于模拟生物学上的约束，控制关节活动范围。例如通过限制肘关节的旋转角度，可以避免模型出现一些不符合生物学和人体运动学的动作；通过根关节控制人体的运动轨迹；通过关节点的速度、加速度等物理量更加精确地表

示人体运动的细节。

以深度学习为代表的数据驱动方法从大量人体运动数据中学习人体运动的特征。这些数据一方面源于动作捕捉，另一方面是通过从视频中重建人体序列或者姿态估计得到（Kocabas 等，2020；Cao 等，2021），二者的关键都是获取逐帧的关节点数据。深度学习模型从训练数据中学习人体动作的特征空间（动作隐空间），其中包含所有可能的人体动作集合，从中采样即可生成多样化的人体运动。如果与其他输入条件的特征空间进行跨模态融合，就能实现条件动作生成任务。

当前，基于物理的方法和基于深度学习的方法呈现逐渐融合的趋势。一些研究（Yuan 等，2023；Raab 等，2023；Tevet 等，2023）尝试将物理约束和物理模拟方法引入深度学习方法中，增强生成运动的真实性和环境交互表现。

图 3　肌肉骨骼模型（Lee 等，2019）

Fig. 3　Musculoskeletal model（Lee et al.，2019）

3　数字人运动生成技术

3.1　基于生成对抗网络（GAN）的方法

生成对抗网络（generative adversarial network，GAN）（Goodfellow 等，2020）是一种生成模型，它由生成器和判别器组成。生成器将输入的随机噪声映射到与真实数据分布类似的空间，以生成逼真的数据样本。判别器和生成器相互对抗，判别生成结果是真实数据还是生成的假数据。在训练的过程中，生成器和判别器的能力不断提升，直到生成器生成的数据足以欺骗判别器。

一些工作将生成对抗网络引入数字人运动生成任务。Text2Action（Ahn 等，2018）是一个基于序列到序列（sequence-to-sequence）（Sutskever 等，2014）的生成对抗网络模型，由一个基于 RNN（recurrent neural network）结构的文本编码器和运动解码器组成。文本编码器将输入文本编码为特征向量，基于注意力机制的运动解码器将特征向量解码为相应的人体运动，通过构建跨模态的联合嵌入（joint embedding），将文本和运动映射到相同的空间，再从中采样，通过生成对抗网络生成多样性的人体运动。DVGANs（dense validation GANS）（Lin 等，2018）是一种基于生成对抗网络的文本条件运动生成模型，采用密集验证（dense validation）方法，在每个时间尺度和每一帧上评估卷积层的输出，以减少伪影并提高所生成运动的质量。然而该方法生成的运动长度是固定的。GANimator（Li 等，2022）是一个渐进式运动生成框架，旨在从单个短运动序列中合成多样且高质量的动作。GANimator 由一系列生成对抗神经网络组成，通过在每一层对高斯噪声进行采样，并对上一层生成的运动执行上采样，以生成新的、包含更多细节的运动，实现对不同细粒度运动的分层控制。该框架使用一种骨骼感知算子（Aberman 等，2020）学习输入序列的骨骼拓扑结构，可以将运动推广到具有其他骨骼结构的两足或多足动物。Raab 等人（2023）受到 StyleGAN 的启发，提出了一个名为 MoDi 的架构。MoDi 首先通过编码器从运动数据集中学习到具有强语义信息和高度结构化的潜在空间（latent space），然后将高斯噪声通过一个映射网络投影到该潜在空间，最后使用一个专门针对人体关节设计的三维卷积神经网络生成人体运动。但三维卷积神经网络为每个关节都分配了专用的内核，这导致模型需要较大的内存占用和较长的运行时间。

然而，生成对抗网络难以训练而且不稳定，它们的损失值并不一定代表生成结果的质量，训练过程很容易崩溃。HP-GAN（human pose prediction GAN）（Barsoum 等，2018）通过在网络损失的基础上添加

一个自定义损失函数来提高训练的稳定性，但仍然无法判断模型何时收敛，继续训练有可能造成模型的发散。多项研究（Guo等人，2022a；Petrovich等人，2021；Zhang等人，2022）也指出生成对抗网络在人体运动生成任务上存在难以训练的问题。这主要归因于两方面：一方面，相较于图像生成问题，人体运动生成任务受到关节数量的限制，骨骼的自由度较少，人体关节之间缺乏明显的局部关联性。此外，人体运动具有不规则性，这使得传统卷积方法难以有效提取相关特征（Raab等，2023）。另一方面，运动的控制条件以及人体运动本身都是序列数据，这种时序性使得训练模型变得更具挑战性，特别是序列数据的长期依赖问题和可变长度特点。

为了有效处理这些问题，一些研究采用了编码器—解码器（encoder-decoder）架构，这种架构因其出色的特征提取和序列信息处理能力成为数字人运动生成任务的典型范式。无论是在常用的基于自编码器（autoencoder，AE）、变分自编码器（variational autoencoder，VAE）（Kingma和Welling，2014）的方法，还是当前最流行的基于扩散模型（denoising diffusion probabilistic model，DDPM）（Sohl-Dickstein等，2015；Ho等，2020）的方法，都能看到编码器—解码器的应用。

3.2 基于自编码器（AE）的方法

自编码器是一种无监督学习的神经网络模型，它通过编码获得输入数据的特征向量，然后用特征向量重构输入，可以用于特征提取和降维。早期的人体运动生成方法通常采用基于循环神经网络的自编码器架构。Lin等人（2018）提出了一种基于序列到序列（Sutskever等，2014）的自编码器框架，用于从文本描述中生成简单的人体骨骼动画。该框架由基于双层堆叠的长短期记忆网络（long short-term memory，LSTM）（Hochreiter和Schmidhuber，1997）文本编码器和基于门控循环单元（gated recurrent unit，GRU）（Cho等，2014）的自编码器组成。由于模型的表达能力和训练数据质量的限制，该方法生成的部分运动无法准确符合文本描述，并且可能会产生一些在物理上不合理的动作。

序列到序列的方法只是将一个序列转换为另一个序列，由于人体运动和输入条件（文本、音频、动作等）是不同模态的数据，直接将输入条件特征作为运动的控制信号会导致特征融合困难，降低建模能力。一些工作尝试通过自编码器分别提取输入条件和运动的特征，再将它们映射到一个联合嵌入空间（joint embedding space）。Language2Pose（Ahuja和Morency，2019）通过文本编码器和运动编码器，将输入文本和运动映射到一个联合嵌入空间，再通过一个运动解码器生成人体运动序列。在模型的训练阶段采用课程学习（curriculum learning）（Bengio等，2009）策略，首先生成容易获得的较短的人体运动序列，再逐渐增加生成序列的长度，让模型逐步适应更复杂的长序列运动，提高生成质量。DanceNet（Zhuang等，2022）设计了一种带有门控激活单元的基于扩展卷积的自编码器框架，采用高斯混合模型（Gaussian mixture model，GMM）损失训练网络，能够以音乐的风格、节奏和旋律等特征作为控制信号，生成具有高度真实感、多样性、与音乐节奏相融合的三维舞蹈动作。为了提高模型的性能，作者收集了几个由专业舞者表演的“音乐—舞蹈”数据集。但是生成的舞蹈动作质量与专业舞者相比还存在一定差距，脚部也存在轻微的滑步问题。Ghosh等人（2021）提出了一个分层、双流的自编码器架构，分别学习人体上下肢动作和自然语言之间的联合嵌入空间，再通过运动解码器生成叠加的全身运动，但模型的泛化性能差，在生成新的叠加运动时并不总是成功的。MotionCLIP（motion generation to clip space）（Tevet等，2022）提出了一个基于Transformer（Vaswani等，2017）的自编码器架构，通过将人体运动的特征空间与CLIP（contrastive language-image pretraining）（Radford等，2021）的特征空间对齐，建立了文本和人体运动之间的映射关系。利用CLIP在视觉领域的先验知识，MotionCLIP使文本生成运动任务可以使用抽象的、风格化的语言作为输入，例如输入“沙发”，可以生成坐下的动作，这进一步释放了自然语言在运动生成和控制方面的灵活性。

虽然可以通过函数映射或者CLIP等预训练模型将输入条件的特征和运动特征进行对齐，但自编码器从数据中学习到的是一个固定的特征向量，不适合多样性的运动生成任务。

3.3 基于变分自编码器（VAE）的方法

不同于自编码器，变分自编码器在编码过程中最小化潜在空间与预设先验分布（通常是高斯分布）的差异，然后从这个分布中采样得到隐变量（latent vector），从而引入了随机性，可以生成新的数据。

Action2Motion(Guo等,2020)是第1个实现了动作生成运动任务的方法。该方法使用李代数表示人体运动姿势,通过时序变分自编码器(temporal variational autoencoder)学习动作类别和人体运动的特征映射。然而,基于RNN的变分自编码器在生成长序列运动时存在一些挑战。长序列的复杂性使得模型难以有效捕捉长距离依赖关系,由于其递归式生成数据的特点,可能会导致误差的积累,进而造成生成运动崩溃。因此后续的工作中通常使用基于Transformer的编码器和解码器。

Actor(action-conditioned Transformer)(Petrovich等,2021)使用基于Transformer的变分自编码器从动作数据集中学习潜在空间,通过位置编码引导生成过程,生成具有可变长度的人体运动序列。Guo等人(2022a)提出了一种自回归的文本生成运动框架,根据文本编码器提取到的特征自回归地生成运动。TEMOS(text-to-motions)(Petrovich等,2022)是一个跨模态变分人体运动生成模型,通过Transformer编码器学习文本空间和运动空间的分布,用KL(Kullback-Leibler)散度将这两个空间对齐到一个高斯分布空间,再从中采样,通过解码器生成人体运动序列。但由于自然语言和运动序列的分布差异很大,仅通过KL散度强行将文本和运动空间对齐到简单的高斯分布空间会严重压缩文本中的信息,使得模型难以生成准确符合文字描述的人体运动,而且在长序列运动生成任务中表现较差,会出现抖动、滑步等问题。

传统的变分自编码器在潜在空间中使用连续的高斯分布表示数据,这种表示方法会压缩信息,影响不同模态数据的映射,进而降低生成运动的质量。最近一些工作开始使用向量量化变分自编码器(vector quantized variational autoencoder, VQ-VAE)(van den Oord等,2017)取代传统的变分自编码器,实现运动的离散表示。VQ-VAE是基于向量量化的变分自编码器,在潜在空间中使用离散的向量表示数据,可以解耦潜在空间,有利于模型学习数据的特征。

TM2T(Guo等,2022b)提出一种结合了文本到动作和动作到文本任务的联合训练框架,通过VQ-VAE能够更好地学习文本和人体运动之间的相互映射关系。TM2T首次提出了一种新的称为“运动token”(motion token)的运动表示方法,并将其与文本token结合。这种利用token的离散短序列表示有助于在文本和运动之间建立映射关系,可以将文本生成运动任务转化为自然语言与运动序列之间的翻译任务。T2M-GPT(Zhang等,2023a)是一个基于VQ-VAE和生成式预训练Transformer(generative pre-trained Transformer, GPT)的文本—运动生成框架。该框架通过VQ-VAE的编码器学习运动数据与离散码序列之间的映射,以文本描述为条件生成索引,再通过VQ-VAE的解码器从索引中恢复运动。Jiang等人(2023)认为人体运动表现出与人类语言相似的语义耦合,因此可以看做是一种特殊的“运动语言”,这种观点与TM2T提出的“运动token”不谋而合。他们提出了一个运动语言模型MotionGPT,通过VQ-VAE将运动数据编码为运动token,并与文本token结合起来学习文本和运动之间的语义耦合,训练“运动—语言”预训练模型,再根据指令对预训练模型进行微调,完成运动生成、运动描述、运动预测和运动中间帧生成等任务。除了运动token和文本token的结合,Yi等人(2023)将VQ-VAE引入了音频生成运动任务,提出了第1个从语音生成三维人体全身网格的方法TalkSHOW。作者认为人类语言与面部关节的相关性高,与身体姿势和手势的相关性较低,因此使用简单的编码器—解码器架构生成音唇匹配的面部,使用VQ-VAE来学习身体和手部动作的空间分布,将脸、身体和手分别建模,再通过一个交叉条件的自回归模型来生成连贯的身体运动和手势。

Cervantes等人(2022)提出了一种基于隐式神经表示(implicit neural representation, INR)的变分自编码器框架用于动作生成运动任务。INR通过神经网络来隐式地表示物体或场景,常用于三维重建和渲染。在传统变分自编码器中,编码器的权重是相对于整个数据集进行优化的,无法保证每个单独序列的最优表示,这项工作用隐式神经表示方法代替了传统变分自编码器中的显式编码器,可以直接针对每个单独的序列进行优化,生成的运动更加真实和多样。但隐式神经表示方法是通过简单的多层感知器拟合运动序列,当数据量增加时,生成运动的质量会下降。

变分自编码器对数据潜在空间的假设有利于通过操纵隐变量来控制生成的样本,在运动编辑方面具有优势。然而这种假设也限制了其捕捉复杂数据分布的能力,使模型难以充分学习输入条件和人体

运动数据的细节，因此在生成复杂的运动时存在困难。

3.4 基于扩散模型（Diffusion）的方法

扩散模型的基本原理是对原始数据不断添加噪声，使其逼近高斯分布，再通过逐步去除噪声恢复原始数据，在这个过程中模型学习到数据的特征，因此可以在高斯分布中随机采样噪声并去噪，生成全新的数据（Sohl-Dickstein等，2015；Ho等，2020）。扩散模型是目前最受关注的生成模型之一，在图像生成任务中展示出强大性能（Rombach等，2022；Saharia等，2022；Hertz等，2023）。一些研究将扩散模型生成结果丰富多样、不受目标分布假设的限制等优点引入数字人运动生成任务中。

MotionDiffuse（Zhang等，2022）是第1个基于扩散模型的文本生成运动方法。通过一种基于跨模态线性Transformer的扩散模型，可以根据运动的持续时间生成任意长度的运动。类似于隐变量插值，MotionDiffuse提出了专门针对扩散模型的“噪声插值”（noise interpolation），可以实现对不同身体部位之间的细粒度控制。GestureDiffuCLIP（gesture diffusion model with clip latents）（Ao等，2023）可以将文本、动作和视频作为条件输入模型，使用CLIP在潜在空间编码输入和手势动作，从多个输入模态中提取有效的风格表示，再通过扩散模型生成逼真多样的手势动画。

基于扩散模型的图像生成任务通常采用U-Net架构，MoFusion（Dabral等，2023）和隐空间运动扩散（motion latent diffusion，MLD）模型（Chen等，2023b）也采用了类似U-Net的架构。但在人体运动生成任务中，输入的条件和生成的运动都是可变长度的序列数据。鉴于Transformer在处理序列数据方面的优势，Dabral等人（2023）提出了一种基于去噪扩散模型的框架MoFusion。该框架采用基于Transformer的一维U-Net架构，可以用音乐和文本作为输入，生成长时间、高质量的人体运动。生成的舞蹈动作与音乐的节奏相匹配，而且能很好地推广到不在训练数据分布范围之内的新音乐。Kim等人（2023）设计了一个名为FLAME（free-form language-based motion synthesis and editing）的基于Transformer的扩散模型架构，可以实现文本生成运动与运动编辑任务。在文本编码部分，FLAME利用预训练语言模型RoBERTa来提取输入文本的特征，以增强模型对语言的理解能力。这种结合使得FLAME在运动生成和编辑的过程中能够更好地融合文本信息，提高生成动作的准确性和连贯性。

基于扩散模型的运动生成方法虽然能生成高质量和多样性的人体运动序列，但模型的训练和推理需要耗费大量时间和算力，而且生成可控性差，由此产生的随机性会导致一些方法生成不符合物理规律的人体运动序列，或者存在较大的误差，出现不合理的（夸张或严重畸变的）动作、身体漂浮、穿模和滑步等问题，后续一些工作致力于优化此类方法的效率和生成过程。

1）针对生成过程的优化。ReMoDiffuse（retrieval-augmented motion diffusion model）（Zhang等，2023b）是一个检索增强的人体运动扩散模型。ReMoDiffuse使用了一种混合检索技术，利用文本语义和动作的相似性从训练数据中提取额外的知识来优化生成过程，从而增强生成结果的泛化性和多样性。在常见和不常见的文本提示下都能生成高质量的运动序列，但是信息检索会大幅度增加模型运算量，降低效率。不同于传统的扩散模型，MDM（human motion diffusion model）（Tevet等，2023）在去噪过程中预测的是运动序列信号本身，而不是噪声。相较于预测噪声，预测运动序列有助于设计人体关节的速度、位置等损失来约束运动，让生成过程更可控。具体的过程是通过基于Transform的编码器将文本、动作等输入条件（也可以是空条件）、噪声时间步长和高斯噪声编码，再送入扩散模型中迭代地产生运动序列。

2）针对生成效率的优化。Chen等人（2023b）认为造成计算开销大和推理速度慢的一个重要原因是这些方法直接对原始运动序列使用了扩散模型，因此提出了一种在潜在空间上执行扩散过程的方法MLD。该方法首先通过基于Transformer的变分自编码器得到一个具有代表性的低维人体运动序列潜在编码（latent code），该变分自编码器使用了与U-Net类似的长跳连接（long skip connection），它可以增强隐变量的表达能力，提升模型性能，并且加速收敛。然后在潜在空间上进行扩散过程，在训练和推理阶段大大减少计算开销，比其他在原始运动序列上的扩散模型快两个数量级。

3.5 其他方法

GenMM（generative motion matching）（Li等，

2023)是一个基于运动匹配方法的生成模型,可以从若干个示例序列中挖掘出尽可能多的不同动作,完成运动补全、关键帧引导生成、无限循环运动和运动重组等任务。运动匹配方法使得GenMM能在几分之一秒内生成出高质量的运动,效率远超其他需要长时间训练的基于数据驱动的方法。

3.6 总结

表3总结了近年来数字人运动生成领域的研究工作。

1)从使用方法的角度来看。文本、音频等输入条件和人体运动都是序列数据,采用传统卷积神经网络的生成对抗网络由于训练困难和生成结果多样性差,逐渐被采用编码器—解码器架构的自编码器和变分自编码器架构取代。自编码器学习到数据的特征是一个固定长度的向量,因此不能很好地实现跨模态特征映射,难以生成多样性结果。变分自编码器可以将数据的特征编码到一个潜在空间,但它对数据分布的假设会压缩信息。扩散模型不需要假设数据的分布,可以生成更高质量和多样性的人体运动序列,但它高昂的计算成本和不可控性也带来了巨大挑战。

2)从模型架构的角度来看。当前研究的一个趋势是多种方法的融合,自编码器和变分自编码器可以作为特殊的编码器—解码器,用于编码和解码不同模态数据的特征,再通过一些特殊的神经网络将这些跨模态特征和噪声融合,输入扩散模型来生成人体运动序列。早期的LSTM(Hochreiter和Schmidhuber,1997)、GRU(Cho等,2014)等循环神经网络结构经常作为编码器—解码器的组成部分,但在生成人体运动,特别是生成较长的运动序列时,这些方法通常会在一段时间后回归到平均的人体姿势,生成静止或漂移的结果(Petrovich等,2021;Zhuang等,2022),并且有可能累计误差,造成生成过程的崩溃。而引入了自注意力机制(self-attention)的Transformer可以很好地处理长距离的依赖关系,并且可以通过位置编码引导模型生成可变长度的运动序列。

本文通过总结当前常用的数字人运动生成方法,提出了一个如图4所示的通用框架。运动编码器从原始运动序列中提取运动特征,再与条件编码器提取到的条件特征融合为隐变量或者映射到潜在空间,然后由生成对抗网络、自编码器、变分自编码器和扩散模型等方法从隐变量或潜在空间中学习,最后通过运动解码器输出符合条件的人体运动。当然,在实际应用中模型会更加复杂,例如MLD(Chen等,2023b)是在潜在空间上执行扩散过程的。

4 运动分析与编辑

运动分析和编辑是指在运动生成的基础上完成运动聚类(motion clustering)(Raab等,2023)、运动预测(motion prediction)(Lyu等,2022)、运动插值(motion in-betweening)(Mourot等,2022)和运动填充(motion in-filling)(Duan等,2021)等任务。运动分析和编辑通常是在潜在空间和隐变量上实现的。潜在空间融合了多模态数据的特征,通过解耦这些特征,可以实现一些下游任务。MotionCLIP(Tevet等,2022)通过计算特征向量的距离实现运动聚类和动作识别,通过特征向量的加减实现动作组合和风格迁移(style transfer)。例如一个运动是“一个人在饮水”,另一个运动是“一个人在行走”,将这两个特征向量相加,再减去“一个站立的人”的特征向量,就能生成“一个人边走路边饮水”的运动效果(图5(a))。MoDi(Raab等,2023)通过K-means聚类和t-SNE算法将潜在空间可视化,实现了运动聚类(图5(b))。

运动插值和运动补全是运动编辑中重要的任务,有广泛的应用场景。在虚拟角色动画制作中,可以对选定的关键动作帧实现动作序列的自动化补全,从而提高设计者的工作效率;在游戏开发中,可以批量生成可编辑的人体运动序列作为游戏资产。运动预测和运动补全可以视为运动插值的子任务,运动插值的“外插”可以看做是运动预测,运动插值的“内插”可以看做是运动补全。FLAME(Kim等,2023)和MotionGPT(Jiang等,2023)可以实现给定运动序列开头的关键帧,让模型预测后续的运动序列(图5(c)),MoFusion(Dabral等,2023)和MDM(Tevet等,2023)可以实现给定运动序列开头和结尾的关键帧,根据输入条件引导模型补全中间动作(图5(d))。Harvey等人(2020)提出了一种基于对抗循环神经网络的运动生成模型,可以实现时间稀疏的运动插值任务。

表3　数字人运动生成方法总结

Table 3　Summary of digital human motion generation methods

名称	数据模态	核心方法	方法评价
HP-GAN(Barsoum等,2018)	动作	生成对抗网络	模型不容易收敛,质量欠佳。
Text2Action(Ahn等,2018)	文本	生成对抗网络	受限于训练数据,只能生成人体上肢动作。
DVGANs(Lin等,2018)	文本	生成对抗网络	质量和多样性欠佳,生成的运动长度是固定的。
Harvey等人(2020)	关键帧	生成对抗网络	质量较高,可以实现时间稀疏的运动插值任务。
GANimator(Li等,2022)	运动序列	生成对抗网络	质量高,可以实现对生成运动的细粒度控制。
MoDi(Raab等,2023)	动作	生成对抗网络	质量高,但泛化性能差,计算开销大。
Lin等人(2018)	文本	自编码器	质量和多样性欠佳。
Language2Pose(Ahuja和Morency,2019)	文本	自编码器	生成的运动长度较短,运动轨迹简单,多样性欠佳。
DanceNet(Zhuang等,2022)	音乐	自编码器	质量较高,生成的舞蹈动作和专业舞者的动作仍有差距,有轻微的滑步问题。
Ghosh等人(2021)	文本	自编码器	不同部分肢体动作的叠加可能失败,泛化性欠佳。
MotionCLIP(Tevet等,2022)	文本	自编码器	泛化性能强,但生成的部分动作是静止的。
Action2Motion(Guo等,2020)	动作	变分自编码器	首个动作生成运动工作,质量较高。
Guo等人(2022a)	文本	变分自编码器	质量较高,生成动作的质量随着长度增加而减小。
Actor(Petrovich等,2021)	动作	变分自编码器	通过数据增强方法提高了运动生成质量。
TEMOS(Petrovich等,2022)	文本	变分自编码器	质量欠佳,会有抖动、滑步和漂浮等问题。
INR(Cervantes等,2022)	动作	变分自编码器	相较于传统的VAE方法,可以针对每个单独的序列进行优化,生成更加真实多样和可变长度的运动。
TM2T(Guo等,2022b)	文本	变分自编码器	质量高,提出了新的运动表示形式。
Rhythmic Gesticulator(Ao等,2022)	文本、音频	变分自编码器	质量高,能够充分挖掘手势中蕴含的语义和韵律。
TalkSHOW(Yi等,2023)	音频	变分自编码器	质量高,在生成人体动作时还能准确生成面部表情。
T2M-GPT(Zhang等,2023a)	文本	变分自编码器	质量高,针对VQ-VAE难以训练的问题进行了优化,鲁棒性强。
MotionGPT(Jiang等,2023)	文本	变分自编码器	质量高,动作编辑能力强,但计算开销大。
MotionDiffuse(Zhang等,2022)	文本	扩散模型	首个基于扩散模型的文本生成运动方法,质量较高,计算开销大。
PhysDiff(Yuan等,2023)	文本、动作	扩散模型	物理真实性高,但计算开销大。
MLD(Chen等,2023b)	文本、动作	扩散模型	质量较高,计算开销低。
FLAME(Kim等,2023)	文本	扩散模型	质量高,动作编辑能力强,但计算开销大。
MDM(Tevet等,2023)	文本、动作	扩散模型	质量高,可以实现多种运动编辑任务和细粒度控制。
ReMoDiffuse(Zhang等,2023b)	文本	扩散模型	相较于MotionDiffuse有很大提升,泛化性和多样性更好,但计算开销大。
GestureDiffuCLIP(Ao等,2023)	文本、动作、视频	扩散模型	质量高,但泛化性能欠佳。
MoFusion(Dabral等,2023)	文本、音乐	扩散模型	质量较高,泛化性能欠佳,计算开销大。
GenMM(Li等,2023)	运动序列	运动匹配	质量高,计算开销低,但多样性欠佳,不能处理过长的序列。

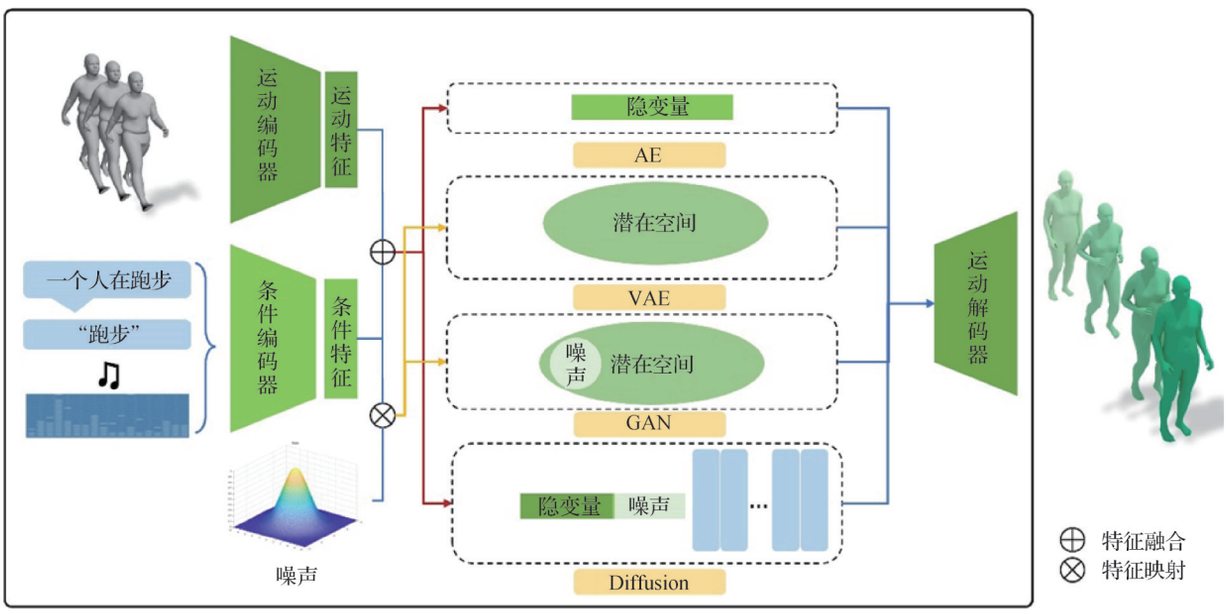
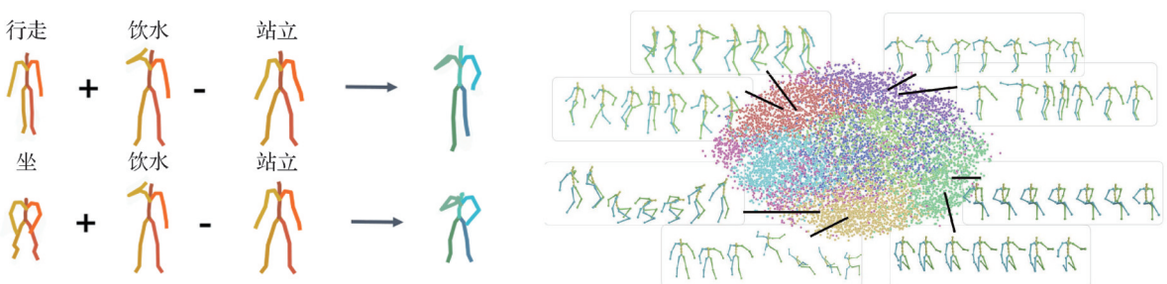
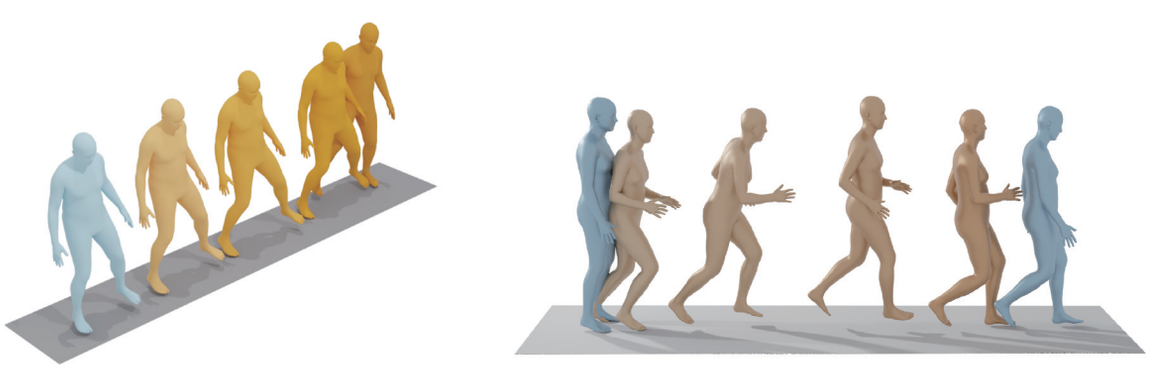


图 4　数字人运动生成的通用框架

Fig. 4　A general framework of digital human motion generation



(a) MotionCLIP(Tevet等，2022)　　(b) MoDi(Raab等，2023)



(c) MotionGPT(Jiang等，2023)　　(d) MDM(Tevet等，2023)

图 5　运动分析和编辑

Fig. 5　Motion analysis and editing ((a) MotionCLIP (Tevet et al. , 2022); (b) MoDi (Raab et al. , 2023); (c) MotionGPT (Jiang et al. , 2023); (d) MDM (Tevet et al. , 2023)

5　实验与评估

5.1　数据集

5.1.1　根据数据类型划分

从数据类型的角度，数据集可以分为动作捕捉(mocap)类数据集和视频重建类数据集。动作捕捉类数据集依靠动作捕捉系统、摄像机和惯性测量系统(inertial measurement unit，IMU)等设备获取真实的人体运动数据(ground truth)。Sigal 等人(2010)提出的小型动作捕捉数据集 HumanEva，采集了室内 4 名受试者 3 次重复地执行一组 6 个预定义动作的数

据。Human3. 6M(Ionescu等,2014)收集了360万个三维人体姿势数据,该数据集包含了17个动作场景下,5名女性和6名男性受试者在4个不同视角下的动作,可以提供日常生活中丰富的动作和姿势,例如拍照、打电话及问候等,常用于人体姿势估计的研究。3DPW(3D poses in the wild dataset)(von Marcard等,2018)从手机摄像头和惯性传感器IMU采集的户外视频中重建出拟合人体的SMPL模型,构造了一个包含7个受试者、60个运动序列和超过51 000帧的人体户外运动数据集,内容包含购物、运动、拥抱和讨论等日常动作。MoVi(Ghorbani等,2021)是一个人体运动视频数据集,采集了60名女性和30名男性受试者的日常动作和运动数据,包含9 h的运动捕捉数据,17 h的视频数据和6. 6 h的IMU数据。可以用于人体姿势估计、动作识别、运动建模、步态分析和三维人体重建。

然而,由于其复杂烦琐的流程和昂贵的成本,动作捕捉导致数据集存在一系列问题:数据量不足、动作种类受限、表示方法多样,并且动作捕捉数据集中还存在伪影。例如图6中展示的SMPL+H扫描数据,存在身体部位缺失、手指脚趾粘连的问题。这些问题在很大程度上制约了数字人运动生成的效果。

为了解决这些问题,Mahmood等人(2019)通过一种从运动捕捉数据中恢复人体运动形态的方法MoSh++和人工辅助矫正,可以将具有不同数据格式和不同人体运动学结构的动作捕捉数据统一转换成SMPL人体模型。然后将15个不同的人体动作捕捉数据集整合成目前最大的人体运动数据集AMASS(archive of motion capture as surface shapes)。AMASS有超过40 h的运动数据,涵盖了344个对象,11 265个运动。在AMASS数据集的基础上,Punnakkal等人(2021)构造了新的数据集BABEL(bodies, action and behavior with English labels)。作者认为一段运动序列可能包含了多种动作,不能简单地用一个标签来描述,因此为AMASS中的运动序列添加了动作标签,其中包括37. 5 h的逐帧动作标签和43. 5 h的整体运动描述标签。BABEL含有超过28 000个序列标签和260种动作类别的63 000个帧数标签,丰富了AMASS的应用场景,可以用于动作识别、时间动作定位和动作生成等任务。

视频重建类数据集通过从包含人体运动的视频中估计关节点,并将其拟合到参数化人体模型上。Lin等人(2023)提出了一个全身模型预测框架OSX,从大量新闻直播、在线教育和视频对话等视频中回归SMPL-X模型的参数,构建了一个涵盖日常生活场景的大规模上半身数据集UBody(upper-body dataset),极大地拓展了参数化人体模型在手语识别、手势生成和情感识别等下游任务中的应用场景。Actor(Petrovich等,2021)通过VIBE(video inference for human body pose and shape estimation)(Kocabas等,2020)从包含40个动作类别、25 600个序列的UESTC(University of Electronic Science and Technology of China)(Ji等,2018)数据集中重建出人体运动序列的SMPL模型作为训练数据,并且将生成的动作序列添加到训练数据中实现数据增强。视频重建类数据集的优点是制作容易、成本低,缺点是部分重建结果会出现穿模、抖动和肢体畸变等问题,此外高质量的人体运动视频并不容易获取,遮挡、镜头快速切换和复杂场景等都会影响数据质量。Text2Action(Ahn等,2018)的训练数据是从MSR-VTT数据集的视频中提取二维人体姿势,再转换成三维人体姿势。由于数据集中大部分样本存在下肢遮挡的问题,Text2Action只能完成人体上肢的运动生成。表4总结了以上数据集的详细情况。

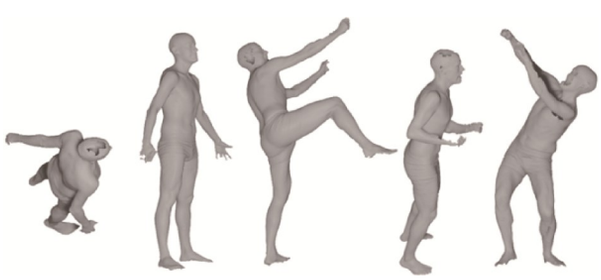


图6　SMPL+H扫描数据中存在的伪影(Romero等,2017)
Fig. 6　Artifacts in SMPL+H scan data (Romero et al. ,2017)

5. 1. 2　根据任务类型划分

从任务类型的角度,常用的数据集可以分为文本—运动数据集、动作—运动数据集和音频—运动数据集等。它们通常是根据具体任务对动作捕捉类数据集和视频重建类数据集加工处理得到的新数据集。

1)文本生成运动方面。Plappert等人(2016)使用动作捕捉数据处理框架MMM(the master motor map)从动作捕捉数据集中构造标准化人体模型,并为运动添加了文本描述,制作了KIT-ML(KIT motion-language)数据集。该数据集总共包含111个

表 4　根据数据类型划分的人体动作数据集

Table 4　Human action datasets classified by data types

数据集	数据类型	内容
HumanEva（Sigal等，2010）	动作捕捉	室内4名受试者3次重复地执行一组6个预定义动作
Human3.6M（Ionescu等，2014）	动作捕捉	17个动作场景下，5名女性和6名男性受试者在4个不同视角下的动作
3DPW（von Marcard等，2018）	动作捕捉	7个受试者，60个运动序列和超过51 000帧的人体户外运动数据
AMASS（Mahmood等，2019）	动作捕捉	包含15个不同的人体动作捕捉数据集，超过40 h的运动数据，涵盖了344个对象和11 265个运动
MoVi（Ghorbani等，2021）	动作捕捉	60名女性和30名男性受试者的日常动作和运动数据，包含9 h的运动捕捉数据、17 h的视频数据和6.6 h的IMU数据
BABEL（Punnakkal等，2021）	动作捕捉	在AMASS数据集基础上，为运动序列添加了逐帧动作标签和整体运动描述标签，包含超过28 000个序列标签和260种动作类别的63 000个帧数标签
Ubody（Lin等，2023）	视频重建	从大量视频中回归SMPL-X模型参数的涵盖日常生活场景的大规模上半身数据集

对象，3 911个动作，总时长11.23 h，文本描述包含52 903个单词的6 278个自然语言注释。Guo等人（2022a）为HumanAct12和AMASS数据集的每个运动序列添加了3～4个文本描述，并通过动作镜像变换和关键词替换实现数据增强，得到了HumanML3D数据集，它包括14 616个动作和44 970个文本描述，由5 371个不同的单词组成，运动序列总时长为28.59 h，平均运动长度为7.1 s，平均描述长度为12个单词。涵盖了广泛的人类行为，包括日常活动、运动、杂技和舞蹈等动作。

2）动作生成运动方面。Guo等人（2020）对NTU-RGB-D数据集、CMU MoCap数据集和PHSPD（polarization human shape and pose dataset）数据集的部分数据进行重构和动作类型标注，构建了新的数据集HumanAct12。HumanAct12将动作划分为12个类别，34个更加精细的子类，包含1 191个运动序列和90 099帧。

3）音频生成运动任务方面。Yi等人（2023）构建了一个来自真实场景（in-the-wild）视频的语音同步的三维全身网格数据集SHOW，用于从语音中生成人体运动序列。表5总结了以上数据集的详细情况。

5.2　评价指标

在数字人运动生成任务中，理想的生成结果应该具备以下特点：1）生成的运动接近真实运动数据集的分布；2）生成的结果具有多样性，同一个输入条件可以产生不同的输出结果；3）生成的运动要符合输入条件的描述；4）生成的结果要自然流畅，符合人类的运动模式和物理规律。针对这些预期特点，本文分别从生成运动质量、生成结果多样性、条件匹配程度、其他评估指标和用户调查几个方面介绍常用的评估指标。

5.2.1　生成运动质量

FID（Frechet inception distance）（Heusel等，2017）用于评估生成运动和真实运动之间分布的相似性。通过运动编码器从生成的结果和ground truth运动序列中提取到的特征（即均值与方差），计算这两个

表 5　根据任务类型划分的人体动作数据集

Table 5　Human action datasets categorized by task types

数据集	任务类型	内容
KIT-ML（Plappert等，2016）	文本生成运动	包含111个对象、3 911个动作和52 903个单词的6 278个自然语言注释，总时长11.23 h
HumanAct12（Guo等，2020）	动作生成运动	包含1 191个运动序列和90 099个姿势，划分为12个类别和34个子类
HumanML3D（Guo等，2022a）	文本生成运动	包含14 616个动作、44 970个文本描述和5 371个不同单词的数据集，总时长28.59 h，平均动作长度7.1 s，平均描述长度12个单词
SHOW（Yi等，2023）	音频生成运动	包含4个不同说话风格的对象在真实场景中的谈话视频，配有同步音频，总时长27 h

分布之间的距离，值越小表示生成结果和真实结果的分布差异越小，模型的性能越好。

值得注意的是，数据集在一定程度上限制了评估指标的使用。TEMOS（Petrovich等，2022）使用的KIT-ML数据集非常小，导致对于相同的文本没有足够多的运动样本可以用于计算基于分布的指标（例如FID）。因此TEMOS通过计算生成运动和ground truth运动的对应关节点的平均位置误差（average positional error，APE）和平均方差误差（average variance error，AVE）来评估生成运动的质量。在最新的工作中，MoDi（Raab等，2023）使用KID（kernel inception distance）（Bińkowski等，2018）评估生成运动的质量，KID相较于FID更适合中小型数据集。

5.2.2　生成结果多样性

多样性（diversity）是通过计算整个集合的运动方差来衡量生成运动的多样性。它反映了模型的生成结果在整个数据集上的变化程度。生成运动的多样性与真实数据的多样性越接近越好。另一个与之相似的评估指标是多模态性（multimodality）。该指标用来度量每个输入条件下生成运动的多样性，关注的是在特定输入条件下生成模型输出结果的多样性水平。多模态性越高，生成结果的多样性越好。同时考虑多样性和多模态性可以更加全面地对生成结果进行评估。

5.2.3　条件匹配程度

检索精度（retrieval precision，R-precision）用于衡量输入条件与生成运动之间的相关性。将每对输入条件和相应的生成运动与测试集中随机选择的31个其他输入条件生成的运动混合，然后计算输入条件与这些生成运动的top-k（k通常取1～3）匹配精度。多模态距离（multimodal distance）是指测试集中每个生成运动的特征与其对应输入条件的特征之间的平均欧几里德距离。检索精度越高，多模态距离越低，表示输入条件与生成运动越匹配。

5.2.4　其他评估指标

一些研究针对其目标任务和提出的方法使用了其他评估指标。MLD（Chen等，2023b）的核心工作是提高运动生成扩散模型的效率，因此提出了以秒为单位的每句平均推理时间（average inference time per sentence，AITS）来评估扩散模型的推理效率。PhysDiff（physics-guided human motion diffusion model）（Yuan等，2023）的核心工作是引入物理约束来优化生成运动的质量，因此提出了基于物理的度量标准Phys-Err来度量人体模型与地面的穿透、漂浮和滑步程度。

5.2.5　用户调查

即使大部分工作在实验中都会测试这些评估指标20次，并给出95%的置信区间，但前文提及的定量评估指标存在一定缺陷：FID需要假设数据的分布，检索精度、多样性和多模态性在计算过程中的随机采样会造成很大的不确定性。因此人类的主观感受对于生成运动的评估必不可少。

用户调查通常是这样设计的：从测试集中选择多个输入条件，对于每一个输入条件，将待测评方法和其他方法生成的运动序列混合后展示给用户，让受试者分别根据测试对象的真实感、流畅度、测试对象与输入条件的匹配程度和测试对象的多样性对这些结果进行排序。一些测试中还会使用ground truth数据作为参照。

5.3　不同算法在现有数据集上的性能评估

为了对比不同数字人运动生成方法在公开数据集上的性能表现，本文选用了上述常见的评价指标并开展了大量实验进行综合评估。当前基于多模态信息的数字人运动生成任务中，最常见的是文本生成运动和动作生成运动，这些任务也提供了相对齐全的数据集和评估指标，再结合方法类别，对比实验选择了基于GAN的方法MoDi（Raab等，2023）；基于VAE的方法Actor（Petrovich等，2021）；基于扩散模型的方法MotionDiffuse（Zhang等，2022）、MDM（Tevet等，2023）和MLD（Chen等，2023b）。此外，还有一些方法的任务类型、评估指标和测试数据集不同于对比实验中大多数方法。例如TEMOS的评估指标是APE和AVE、MoDi提供了KID评估指标，以及在Mixamo数据集上的表现、MDM（Tevet等，2023）还有无条件运动生成任务。为了统一格式和对比性能，本文不展示这些结果。

本文在两个不同的实验环境下进行实验。实验环境1采用Ubuntu 18.04操作系统，CUDA 11.1，NVIDIA GeForce RTX 3090 GPU和Intel（R）Xeon（R）Platinum 8358P 2.60 GHz 15 v CPU；实验环境2采用Ubuntu 20.04操作系统，CUDA 12.2，NVIDIA GeForce RTX 4090 GPU和12th Gen Intel（R）Core（TM）9-12900K CPU；其中，MDM（Tevet等，2023）和Actor（Petrovich等，2021）在实验环境1中运行，而

MoDi(Raab 等,2023)、MLD(Chen 等,2023b)、Guo 等人(2022a)和 MotionDiffuse(Zhang 等,2022)在实验环境 2 中运行。本文评估流程遵循这些方法的指示,测试 20 次结果,计算各个评估指标的 95% 置信区间。具体结果见表 6 和表 7。由于不同项目对真实数据的计算结果不一致,因此本文对多样性(diversity)不进行比较。

根据表 6 的结果,在 HumanML3D 数据集上,MotionDiffuse(Zhang 等,2022)在检索精度和多模态距离这两个衡量条件匹配程度的指标上表现最好,因此,在这些方法中它生成的运动最符合输入条件。MLD(Chen 等,2023b)在 FID 和多模态性上表现最好,因此,在这些方法中它生成的运动兼具多样性和质量。在 KIT-ML 数据集上,Guo 等人(2022a)生成的运动既符合输入条件,又能保持结果的多样性。

根据表 7 的结果,在 HumanAct12 数据集上,MLD(Chen 等,2023b)的生成运动质量更好,而 Actor(Petrovich 等,2021)的生成多样性更好;在 UESTC 数据集上,MDM(Tevet 等,2023)的生成运动质量更好,而 Actor(Petrovich 等,2021)的生成多样性更好。

6 结 语

如上所述,数字人运动生成任务已经成为计算机图形学和三维视觉领域的一个研究热点,引起了广泛的关注。本文在基础研究方面,介绍了常用的参数化人体模型,根据数据类型和任务类型分别介绍了常用的数据集和评估指标;在生成算法方面,根据模型框架将已有的研究分为基于生成对抗网络、自编码器、变分自编码器和扩散模型的方法,分析了

表 6 文本生成运动任务中不同方法的性能对比

Table 6 Performance comparison of different methods in text generation motion task

数据集	方法	R-分数(R-precision)↑			FID↓	多样性→	多模态性↑	多模态距离↓
		Top 1	Top 2	Top 3				
Human-ML3D	MDM	0.317±0.005	0.493±0.006	0.604±0.007	0.528±0.053	9.407±0.100	/	5.579±0.028
	MD	**0.490±0.003**	**0.683±0.003**	**0.781±0.002**	0.678±0.014	9.444±0.076	1.556±0.064	**3.088±0.008**
	Guo	0.456±0.003	0.638±0.003	0.738±0.002	1.067±0.025	9.230±0.100	2.262±0.115	3.349±0.011
	MLD	0.465±0.003	0.655±0.003	0.757±0.002	**0.417±0.010**	9.828±0.074	**2.603±0.085**	3.282±0.010
KIT-ML	MDM	0.163±0.003	0.291 ±0.005	0.398±0.005	**0.509±0.032**	10.686±0.118	/	9.166±0.028
	MD	0.340±0.006	0.542±0.008	0.670±0.006	2.169±0.095	9.940±0.108	1.735±0.070	3.654±0.027
	Guo	**0.359±0.006**	**0.555±0.006**	**0.674±0.006**	3.115±0.133	10.758±0.119	**2.113±0.114**	**3.514±0.025**

注:加粗字体表示各数据集中各列的最优结果。"/"表示不适用。"↑"表示数值越大越好,"↓"表示数值越小越好,"→"表示数值与真实数据越接近越好。MDM、MD、Guo、MLD 分别代表 MDM(Tevet 等,2023)、MotionDiffuse(Zhang 等,2022)、Guo 等人(2022a)以及 MLD(Chen 等,2023b)。

表 7 动作生成运动任务中不同方法的性能对比

Table 7 Performance comparison of different methods in action generation motion task

数据集	方法	FID↓	多样性→	多模态性↑	准确性↑
HumanAct12	MoDi(Raab 等,2023)	14.425±0.298	17.651±0.230	-	-
	MDM(Tevet 等,2023)	0.103±0.006	6.867±0.075	2.555±0.045	0.984±0.002
	MLD(Chen 等,2023b)	**0.079±0.004**	6.850±0.060	2.570±0.050	**0.985±0.001**
	Actor(Petrovich 等,2021)	0.116±0.004	6.841±0.048	**2.619±0.039**	0.953±0.003
UESTC	MDM(Tevet 等,2023)	**12.920±0.419**	33.029±0.235	14.225±0.116	**0.948±0.004**
	Actor(Petrovich 等,2021)	23.136±0.365	32.463±0.262	**14.382±0.064**	0.926±0.002

注:加粗字体表示每列最优值,"-"表示未提供该评估指标。"↑"表示数值越大越好,"↓"表示数值越小越好,"→"表示数值与真实数据越接近越好。

每种方法的特点和优缺点，并为这些方法总结了一种通用技术框架；在人体运动生成技术方面，介绍了运动分析和编辑及其应用场景。

得益于深度学习技术和三维视觉领域的快速发展，数字人运动生成的研究取得了显著的进展。在参数化人体模型方面，SMPL-X和SMPL-H等具有更多参数的模型取代了基本的SMPL模型，目前正向着模型轻量化的方向发展；在模型框架方面，如图7所示，具有强大生成能力的扩散模型取代了传统图像生成领域常用的生成对抗网络，具有更强特征提取能力和序列数据处理能力的Transformer取代了LSTM和RNN等传统循环神经网络，并且一些研究综合了这些框架的优势，将它们集成使用；在数据集方面，具有更多标注信息的大规模人体运动数据集取代了原始的动作捕捉数据集。但仍有以下几个方面需要完善。

6.1 更为完善的三维人体动作数据集构建

数据集样本量对模型的性能和泛化能力有重要影响，更大的数据集通常有助于提高模型的性能和鲁棒性。然而，相对于ImageNet（Deng等，2009）、MS COCO（Microsoft common objects in context）（Chen等，2015）等图像领域样本量庞大的数据集，人体运动生成领域的大规模、高质量数据集非常稀缺。

1）缺少运动种类和描述条件。图8展示了VIBE（Kocabas等，2020）从视频中重建运动序列的效果，这类从视频中构建数据的方法存在抖动、穿模和拟合异常等问题。因此，当前主流的人体运动数据还是以动作捕捉设备采集为主，但这些动作捕捉数据被“反复使用”。HumanML3D（Guo等，2022a）和HumanAct12（Guo等，2020）等新提出的数据集是在原有的一个或多个数据集上重新编辑得到的，这就导致了数据集中全新种类的运动和相应的描述条件很少被扩充，模型无法学习到新的数据特征，模型的泛化能力也受到限制（Raab等，2023），生成的运动不够丰富。对于数据集中未出现过的词汇，模型无法完成正确的推理，可能生成静止的运动序列（Dabral等，2023）。

2）数据分布不平衡。常规动作的数量远大于其他类型动作的数量，模型无法充分学习不常见运动的特征，导致生成质量较差。例如TEMOS（Petrovich等，2022）采用KIT-ML作为训练数据，其中大部分数据是走路、跑步等日常动作，模型能够稳定地生成此类运动，但对于一些复杂运动的生成结果较差，例如“一个人正在喝水”，生成的结果中人体上下肢动作不协调，下肢动作有明显的位移。

3）动作序列长度较短。由于大多数训练数据都是较短的运动序列，模型无法从中学习长序列运动的特征，因此生成的长序列运动质量不佳，可能产生静止、不符合输入条件描述的结果。例如MLD（Chen等，2023b）生成的运动序列长度受限于训练数据中的最大长度。

综上所述，为了解决数字人运动生成领域数据集欠缺的问题，未来的研究可以从以下几个方面进行探索：1）采集更多的人体运动数据，额外增加稀缺

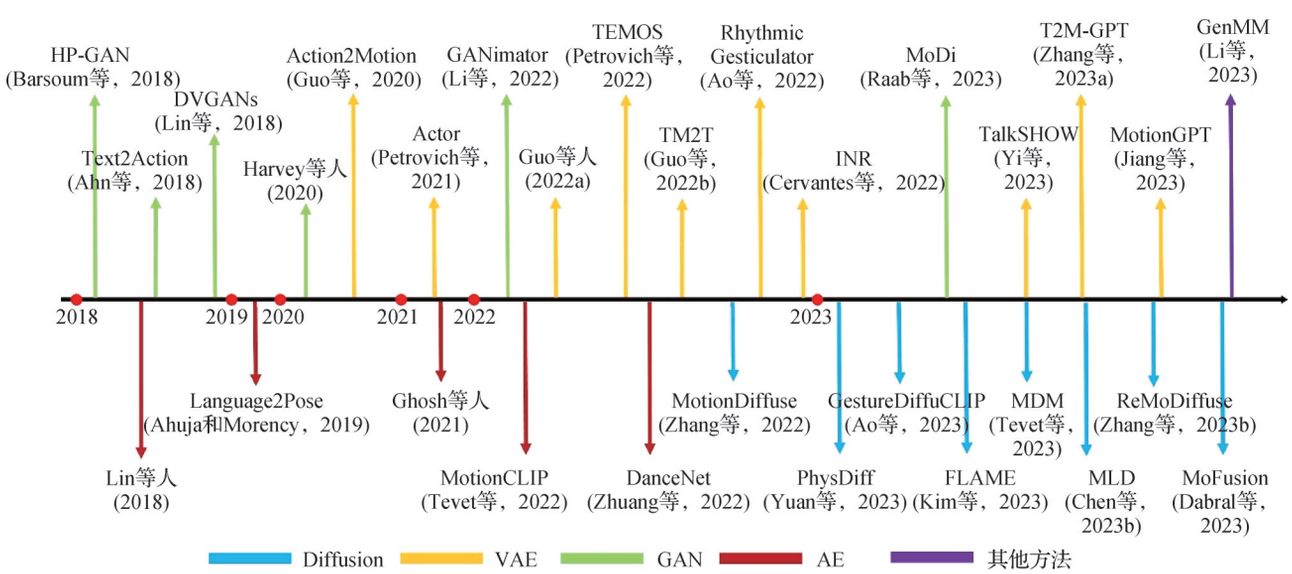


图7 基于多模态信息的三维数字人运动生成技术发展历程概览

Fig. 7 Overview of the development history of 3D digital human motion generation technology based on multimodal information

图 8　从视频中重建的人体模型（Kocabas 等，2020）

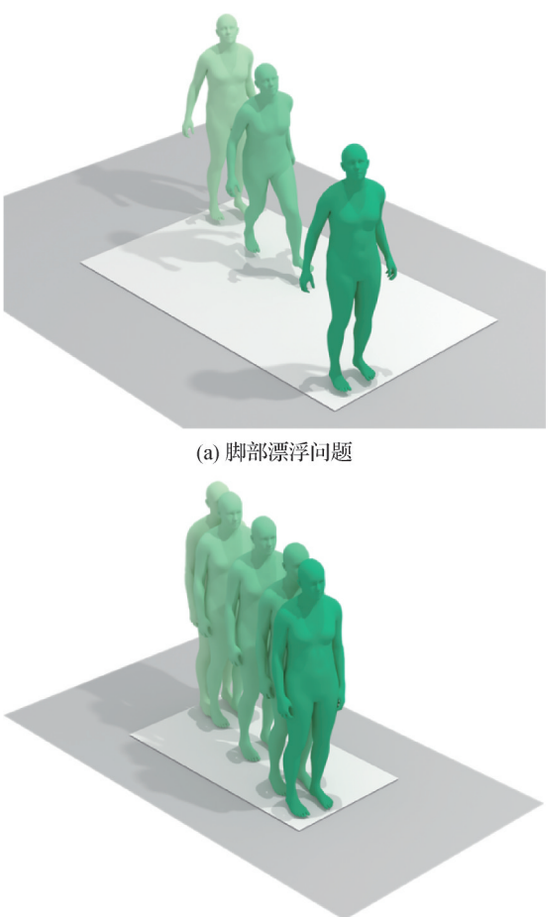
Fig. 8　Reconstructed human models from videos (Kocabas et al. ,2020)

运动和长序列运动的数量，确保数据分布均衡。此外，还可以对已有动作样本增加更多的描述数据，实现对数据集的快速扩充。例如文本生成运动的数据集可以基于已有的文本描述利用大语言模型给对应的运动添加多样化的文本描述。2）开发更稳定的从视频中重建人体序列的算法，解决抖动、穿模和复杂场景下人体模型拟合异常等问题，以便于大批量地制作人体运动数据集。

6.2　运动质量的优化与提升

符合物理规律的人体运动可以模拟现实世界中的人体运动规律和生理限制，增强真实感。然而目前数字人的运动生成并不完美，一些生成的人体运动存在异常抖动，完成复杂动作时部分身体部位可能出现穿模，在行走和跑步等动作中可能出现滑步和漂浮问题，以及运动转向过于迅速。此处以滑步问题为例，探讨数字人运动生成中引入物理约束的必要性和一些解决思路。在基于数据驱动的人体运动生成工作中，参数化人体模型对于准确表现运动细节至关重要。然而，在主流的人体运动模型中，脚部通常仅由脚踝（ankle）和脚趾（toe）这两个关节点控制。脚部骨骼模型的细节缺失一方面限制了算法从数据中学习脚部运动的细节，即使有大量的训练数据，也难以完整表现脚部动作的细节；另一方面也限制了利用人体解剖学的知识建立脚部骨骼结构和关节运动约束的方法。这导致在舞蹈生成和其他涉及脚部运动的任务场景中表现不佳。例如，图 9 展示的 TEMOS（Petrovich 等，2022）中存在的滑步和漂浮问题。具体表现在数字人的脚部在接触地面时不能准确定位，导致在行走、跑步和跳跃等涉及脚部的动作在视觉效果上呈现出滑动、漂浮等现象。图 9（a）中人体双脚悬空，不符合人类的行走习惯和物理规律；图 9（b）中人体运动没有明显的脚步跨越动作，双脚在同步平移而不是走动，这些都会影响数字人运动的真实感。

一些工作通过约束脚部与地面的接触位置和接触速度缓解了脚部滑动问题。DanceNet（Zhuang 等，2022）用一个二维矢量来描述脚是否接触并固定到地面上，通过检测每个帧的左右脚趾端点的位置和速度，将足部约束添加到运动特征中，但生成的舞蹈序列中仍然会有轻微的滑步现象。MoDi（Raab 等，2023）和 GANimator（Li 等，2022）通过设计脚部与地



（a）脚部漂浮问题

（b）滑步问题

图 9　脚部漂浮和滑步问题（Petrovich 等，2022）

Fig. 9　Foot floatation and sliding problems (Petrovich et al. , 2022) ((a) foot floatation; (b) foot sliding)

面接触时的激励函数，引入约束接触速度的一致性损失，缓解了脚部漂浮和滑步现象，使生成的动作更加真实和自然。Guo等人（2022a）为了缓解滑步，用解码器额外地预测脚在每个帧数的接触。MDM（Tevet等，2023）通过约束运动序列的位置和速度来设计几何损失，其中减轻滑步是通过抵消脚部接触地面的速度来实现的，几何损失使得MDM生成的运动更加真实自然。以上研究提出的改进方法只是缓解了滑步问题，当输入条件复杂或者生成序列变长时，异常结果依然存在。

PhysDiff（Yuan等，2023）的提出很好地解决了滑步问题，它是一个物理引导的运动扩散模型。PhysDiff引入了一个基于物理的运动投影模块，通过物理模拟器将每一个扩散步骤生成的去噪运动投影成符合物理规律的运动，这些经过投影矫正的运动随后在下一个扩散步骤中用于指导去噪扩散过程，迭代地将运动约束在合理的物理范围内。但物理模拟的使用严重降低了模型的推理速度，PhysDiff推理时间开销高出其他基于扩散模型的人体运动生成方法2～3倍。

在基于数据驱动的数字人运动生成方法中，仅通过约束关节点位置是不够的，而引入物理模拟可以有效缓解甚至消除肢体抖动、穿模和滑步等问题，优化生成结果，这也是数字人运动质量优化的一个趋势。一方面，可以通过设置针对速度、加速度和转动角度等物理量的约束条件；另一方面，可以参考PhysDiff，使用强化学习方法学习人体运动策略，引导人体运动的生成过程。

后处理工作也能有效地提升生成动作的质量。GenMM（Li等，2023）通过将运动序列的初始帧和结束帧设置为相同的动作，从而生成无限循环的动画，实现运动的无缝连接。运动平滑（motion smooth）处理旨在消除或减少动画中的不连续性和突然的跳跃，以提供更自然和连续的视觉体验。Dagioglou等人（2021）使用贝塞尔曲线来平滑人体运动，并将其迁移到机器人上实现平滑的运动控制。Memar和Yan（2022）使用基于B样条的滤波方法为人体动作数据信号降噪，实现运动平滑。

此外，一些与人体运动生成领域相近的研究方向也采用了一些后处理工作来优化人体运动序列。Physical Inertial Poser（Yi等，2022）通过预测脚与地面的接触概率施加物理约束，通过可以控制关节点角加速度和线加速度的双PD控制器优化对人体运动的控制，有效地抑制了滑步问题。HiMoReNet（Wang等，2023）通过分层架构的后处理运动细化神经网络可以更加精确地估计姿态参数，缓解抖动问题。在运动填充课题上，一些研究工作（Kaufmann等，2020；Tang等，2022）设计特殊的神经网络架构来平滑人体运动序列。在视频重建人体运动序列课题上，Zhang等人（2021）通过约束相邻帧的差异来平滑重建后的人体运动序列。Zou等人（2020）设计了一个用于检测足部和地面接触的神经网络并添加物理约束，缓解了视频重建人体序列的滑步问题。在数字人运动生成领域，可以借鉴这些工作的后处理经验。

6.3 从多模态到跨模态

当前的数字人运动生成方法可以利用文本、音频、音乐、动作和关键帧等多模态信息生成丰富逼真的局部动作或者全身运动，但是这些研究只是将多个单模态运动生成方法集成到了一个框架，并没有真正实现跨模态生成。

跨模态生成是指基于模态之间的语义一致性，实现不同模态数据形式上的相互转换，有助于提高不同模态间的迁移能力（刘华峰 等，2023）。例如输入的信息是“一个舞蹈演员在跳芭蕾舞”和舞曲《天鹅湖》的一个片段，文本提供的信息让模型生成一个人的舞蹈动作，而音乐提供的信息可以使生成的舞蹈符合音乐的旋律。相较于单模态方法，跨模态生成可以提供更全面和丰富的信息优化生成结果。

6.4 丰富生成内容

在生成内容方面，一些基于变分自编码器和扩散模型的方法已经可以生成丰富多样的运动，引入了CLIP的方法可以生成风格化的运动，这些研究的关键点都在运动本身。除此之外，还有很多问题尚未得到关注和系统性研究：1）当前研究主要针对单个数字人的运动生成，而多个数字人的生成和交互（如图10所示）能更好地满足更多场景下的建模需求。例如，MoDi（Raab等，2023）提出的人群模拟方法，在潜在空间中采样多个运动特征接近的数字人；2）当前数字人运动序列多为骨骼模型或者SMPL系列裸体模型，未来工作可以研究如何生成穿着服装的数字人。该研究方向充满挑战，因为服装的褶皱形变一方面与服装类型（或面料）有关，另一方面要与人体运动匹配，要生成高质量的结果需要考虑丰

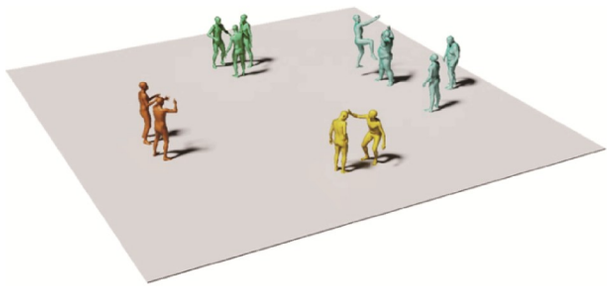


图 10　多数字人交互场景

Fig. 10　Multiple digital human interaction scenarios

富的物理细节。

6.5　提高生成效率

如何在生成质量和算法效率之前取得平衡，是数字人运动生成研究走向实际应用的关键。使用含有更多参数量的参数化人体模型，设计更复杂的模型架构和引入更细节的物理过程可以显著提高生成人体运动的质量，但这会大幅度增加计算量，降低生成效率。可以考虑从以下几个方面提高生成效率：1）使用轻量化的参数化人体模型，例如 STAR（Osman 等，2020），可以在保证生成质量的同时减少冗余的参数。2）使用信息密度更大的训练数据。在训练模型之前，先对训练数据降维，提取主要的特征，以此提升训练数据的信息密度。例如基于扩散模型的方法可以在潜在空间中执行扩散过程，通过学习编码器压缩过的数据提高生成效率（Chen 等，2023b）。3）改进生成模型。可以尝试模型加速策略，例如在图像生成领域，Song 等人（2021）提出了一种使用非马尔可夫过程的扩散模型 DDIM（denoising diffusion implicit models）。通过采样加速，DDIM 的生成速度是原有扩散模型的 10～50 倍。

参考文献（References）

Aberman K, Li P Z, Lischinski D, Sorkine-Hornung O, Cohen-Or D and Chen B Q. 2020. Skeleton-aware networks for deep motion retargeting. ACM Transactions on Graphics, 39(4): #62 [DOI: 10.1145/3386569.3392462]

Ahn H, Ha T, Choi Y, Yoo H and Oh S. 2018. Text2Action: generative adversarial synthesis from language to action//Proceedings of 2018 IEEE International Conference on Robotics and Automation. Brisbane, Australia: IEEE: 5915-5920 [DOI: 10.1109/icra.2018.8460608]

Ahuja C and Morency L P. 2019. Language2Pose: natural language grounded pose forecasting//Proceedings of 2019 International Conference on 3D Vision. Quebec City, Canada: IEEE: 719-728 [DOI: 10.1109/3dv.2019.00084]

Anguelov D, Srinivasan P, Koller D, Thrun S, Rodgers J and Davis J. 2005. SCAPE: shape completion and animation of people. ACM Transactions on Graphics, 24 (3): 408-416 [DOI: 10.1145/1073204.1073207]

Ao T L, Gao Q Z, Lou Y K, Chen B Q and Liu L B. 2022. Rhythmic gesticulator: rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. ACM Transactions on Graphics, 41(6): #209 [DOI: 10.1145/3550454.3555435]

Ao T L, Zhang Z Y and Liu L B. 2023. GestureDiffuCLIP: gesture diffusion model with CLIP latents. ACM Transactions on Graphics, 42(4): #42 [DOI: 10.1145/3592097]

Barsoum E, Kender J and Liu Z C. 2018. HP-GAN: probabilistic 3d human motion prediction via GAN//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 149901-149909 [DOI: 10.1109/CVPRW.2018.00191]

Bengio Y, Louradour J, Collobert R and Weston J. 2009. Curriculum learning//Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada: ACM: 41-48 [DOI: 10.1145/1553374.1553380]

Bergamin K, Clavet S, Holden D and Forbes J R. 2019. DReCon: data-driven responsive control of physics-based characters. ACM Transactions on Graphics, 38 (6): #206 [DOI: 10.1145/3355089.3356536]

Bińkowski M, Sutherland D J, Arbel M and Gretton A. 2018. Demystifying MMD GANs//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Bogo F, Kanazawa A, Lassner C, Gehler P, Romero J and Black M J. 2016. Keep it SMPL: automatic estimation of 3D human pose and shape from a single image//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 561-578 [DOI: 10.1007/978-3-319-46454-1_34]

Cao Z, Hidalgo G, Simon T, Wei S E and Sheikh Y. 2021. OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1): 172-186 [DOI: 10.1109/TPAMI.2019.2929257]

Cervantes P, Sekikawa Y, Sato I and Shinoda K. 2022. Implicit neural representations for variable length human motion generation//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 356-372 [DOI: 10.1007/978-3-031-19790-1_22]

Chen J Y, Yan M, Zhang J Z, Xu Y Z, Li X L, Weng Y J, Yi L, Song S R and Wang H. 2023a. Tracking and reconstructing hand object interactions from point cloud sequences in the wild//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 304-312 [DOI: 10.1609/aaai.v37i1.25103]

Chen X, Jiang B, Liu W, Huang Z L, Fu B, Chen T and Yu G. 2023b.

Executing your commands via motion diffusion in latent space//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 18000-18010 [DOI: 10.1109/cvpr52729.2023.01726]

Chen X L, Fang H, Lin T Y, Vedantam R, Gupta S, Dollár P and Zitnick C L. 2015. Microsoft COCO captions: data collection and evaluation server [EB/OL]. [2023-04-16]. https://arxiv.org/pdf/1504.00325.pdf

Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H and Bengio Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: Association for Computational Linguistics: 1724-1734 [DOI: 10.3115/v1/d14-1179]

Dabral R, Mughal M H, Golyanik V and Theobalt C. 2023. MoFusion: a framework for denoising-diffusion-based motion synthesis//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 9760-9770 [DOI: 10.1109/cvpr52729.2023.00941]

Dagioglou M, Tsitos A C, Smarnakis A and Karkaletsis V. 2021. Smoothing of human movements recorded by a single RGB-D camera for robot demonstrations//Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference. Corfu, Greece: ACM: 496-501 [DOI: 10.1145/3453892.3461627]

Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/cvpr.2009.5206848]

Duan Y L, Shi T Y, Zou Z X, Lin Y N, Qian Z H, Zhang B H and Yuan Y. 2021. Single-shot motion completion with Transformer [EB/OL]. [2023-07-21]. https://arxiv.org/pdf/2103.00776.pdf

Ghorbani S, Mahdaviani K, Thaler A, Kording K, Cook D J, Blohm G and Troje N F. 2021. MoVi: a large multi-purpose human motion and video dataset. PLoS One, 16(6): #e0253157 [DOI: 10.1371/journal.pone.0253157]

Ghosh A, Cheema N, Oguz C, Theobalt C and Slusallek P. 2021. Synthesis of compositional animations from textual descriptions//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 1376-1386 [DOI: 10.1109/ICCV48922.2021.00143]

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2020. Generative adversarial networks. Communications of the ACM, 63(11): 139-144 [DOI: 10.1145/3422622]

Guo C, Zou S H, Zuo X X, Wang S, Ji W, Li X Y and Cheng L. 2022a. Generating diverse and natural 3D human motions from text//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5142-5151 [DOI: 10.1109/cvpr52688.2022.00509]

Guo C, Zuo X X, Wang S and Cheng L. 2022b. TM2T: stochastic and tokenized modeling for the reciprocal generation of 3D human motions and texts//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 580-597 [DOI: 10.1007/978-3-031-19833-5_34]

Guo C, Zuo X X, Wang S, Zou S H, Sun Q Y, Deng A N, Gong M L and Cheng L. 2020. Action2Motion: conditioned generation of 3D human motions//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 2021-2029 [DOI: 10.1145/3394171.3413635]

Harvey F G, Yurick M, Nowrouzezahrai D and Pal C. 2020. Robust motion in-betweening. ACM Transactions on Graphics, 39(4): #60 [DOI: 10.1145/3386569.3392480]

Hertz A, Mokady R, Tenenbaum J, Aberman K, Pritch Y and Cohen-Or D. 2023. Prompt-to-prompt image editing with cross-attention control//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net: 1-36

Heusel M, Ramsauer H, Unterthiner T, Nessler B and Hochreiter S. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc.: 6629-6640 [DOI: 10.5555/3295222.3295408]

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #574 [DOI: 10.5555/3495724.3496298]

Hochreiter S and Schmidhuber J. 1997. Long short-term memory. Neural Computation, 9(8): 1735-1780 [DOI: 10.1162/neco.1997.9.8.1735]

Ionescu C, Papava D, Olaru V and Sminchisescu C. 2014. Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(7): 1325-1339 [DOI: 10.1109/tpami.2013.248]

Ji Y L, Xu F X, Yang Y, Shen F M, Shen H T and Zheng W S. 2018. A large-scale RGB-D database for arbitrary-view human action recognition//Proceedings of the 26th ACM international conference on Multimedia. Seoul, Korea(South): ACM: 1510-1518 [DOI: 10.1145/3240508.3240675]

Jiang B, Chen X, Liu W, Yu J Y, Yu G and Chen T. 2023. MotionGPT: human motion as a foreign language//Proceedings of the 37th Conference on Neural Information Processing Systems. New Orleans, USA: NeurIPS

Kaufmann M, Aksan E, Song J, Pece F, Ziegler R and Hilliges O. 2020. Convolutional autoencoders for human motion infilling//Proceedings of 2020 International Conference on 3D Vision. Fukuoka, Japan: IEEE: 918-927 [DOI: 10.1109/3DV50981.2020.00102]

Kim J, Kim J and Choi S. 2023. FLAME: free-form language-based motion synthesis and editing//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 8255-8263 [DOI: 10.1609/aaai.v37i7.25996]

Kingma D P and Welling M. 2014. Auto-encoding variational Bayes//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: ICLR

Kocabas M, Athanasiou N and Black M J. 2020. VIBE: video inference for human body pose and shape estimation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5252-5262 [DOI: 10.1109/cvpr42600.2020.00530]

Kwiatkowski A, Alvarado E, Kalogeiton V, Liu C K, Pettré J, van de Panne M and Cani M P. 2022. A survey on reinforcement learning methods in character animation. Computer Graphics Forum, 41(2): 613-639 [DOI: 10.1111/cgf.14504]

Lee S, Lee S, Lee Y and Lee J. 2021. Learning a family of motor skills from a single motion clip. ACM Transactions on Graphics, 40(4): #93 [DOI: 10.1145/3450626.3459774]

Lee S, Park M, Lee K and Lee J. 2019. Scalable muscle-actuated human simulation and control. ACM Transactions on Graphics, 38(4): #73 [DOI: 10.1145/3306346.3322972]

Li P Z, Aberman K, Zhang Z H, Hanocka R and Sorkine-Hornung O. 2022. GANimator: neural motion synthesis from a single sequence. ACM Transactions on Graphics, 41(4): #138 [DOI: 10.1145/3528223.3530157]

Li W Y, Chen X L, Li P Z, Sorkine-Hornung O and Chen B Q. 2023. Example-based motion synthesis via generative motion matching. ACM Transactions on Graphics, 42(4): #94 [DOI: 10.1145/3592395]

Lin A S, Wu L M, Corona R, Tai K, Huang Q X and Mooney R J. 2018. Generating animated videos of human activities from natural language descriptions//Proceedings of the 32nd Conference on Neural Information Processing Systems. Montréal, Canada: NeurIPS

Lin J, Zeng A L, Wang H Q, Zhang L and Li Y. 2023. One-stage 3D whole-body mesh recovery with component aware Transformer//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 21159-21168 [DOI: 10.1109/cvpr52729.2023.02027]

Lin X and Amer M R. 2018. Human motion modeling using DVGANs [EB/OL]. [2023-07-12]. https://arxiv.org/pdf/1804.10652.pdf

Liu H F, Chen J J, Li L, Bao B K, Li Z C, Liu J Y and Nie L Q. 2023. Cross-modal representation learning and generation. Journal of Image and Graphics, 28(6): 1608-1629 (刘华峰, 陈静静, 李亮, 鲍秉坤, 李泽超, 刘家瑛, 聂礼强. 2023. 跨模态表征与生成技术. 中国图象图形学报, 28(6): 1608-1629) [DOI: 10.11834/jig.230035]

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

Lyu K D, Chen H P, Liu Z G, Zhang B Q and Wang R L. 2022. 3D human motion prediction: a survey. Neurocomputing, 489: 345-365 [DOI: 10.1016/j.neucom.2022.02.045]

Mahmood N, Ghorbani N, Troje N F, Pons-Moll G and Black M J. 2019. AMASS: archive of motion capture as surface shapes//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 5441-5450 [DOI: 10.1109/ICCV.2019.00554]

Memar A M and Yan H. 2022. Noise reduction in human motion-captured signals for computer animation based on b-spline filtering. Sensors, 22(12): #4629 [DOI: 10.3390/s22124629]

Mourot L, Hoyet L, Le Clerc F, Schnitzler F and Hellier P. 2022. A survey on deep learning for skeleton-based human animation. Computer Graphics Forum, 41(1): 122-157 [DOI: 10.1111/cgf.14426]

Osman A A A, Bolkart T and Black M J. 2020. STAR: sparse trained articulated human body regressor//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 598-613 [DOI: 10.1007/978-3-030-58539-6_36]

Pavlakos G, Choutas V, Ghorbani N, Bolkart T, Osman A A, Tzionas D and Black M J. 2019. Expressive body capture: 3D hands, face, and body from a single image//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 10967-10977 [DOI: 10.1109/CVPR.2019.01123]

Peng X B, Abbeel P, Levine S and van de Panne M. 2018. DeepMimic: example-guided deep reinforcement learning of physics-based character skills. ACM Transactions on Graphics, 37(4): #143 [DOI: 10.1145/3197517.3201311]

Petrovich M, Black M J and Varol G. 2021. Action-conditioned 3d human motion synthesis with Transformer VAE//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 10965-10975 [DOI: 10.1109/iccv48922.2021.01080]

Petrovich M, Black M J and Varol G. 2022. TEMOS: generating diverse human motions from textual descriptions//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 480-497 [DOI: 10.1007/978-3-031-20047-2_28]

Plappert M, Mandery C and Asfour T. 2016. The KIT motion-language dataset. Big Data, 4(4): 236-252 [DOI: 10.1089/big.2016.0028]

Punnakkal A R, Chandrasekaran A, Athanasiou N, Quirós-Ramirez A and Black M J. 2021. BABEL: bodies, action and behavior with English labels//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 722-731 [DOI: 10.1109/cvpr46437.2021.00078]

Raab S, Leibovitch I, Li P Z, Aberman K, Sorkine-Hornung O and Cohen-Or D. 2023. MoDi: unconditional motion synthesis from diverse data//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE:

13873-13883 [DOI: 10.1109/CVPR52729.2023.01333]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 8748-8763

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10674-10685 [DOI: 10.1109/CVPR52688.2022.01042]

Romero J, Tzionas D and Black M J. 2017. Embodied hands: modeling and capturing hands and bodies together. ACM Transactions on Graphics, 36(6): #245 [DOI: 10.1145/3130800.3130883]

Saharia C, Chan W, Saxena S, Li L L, Whang J, Denton E, Ghasemipour S K S, Ayan B K, Mahdavi S S, Lopes R G, Salimans T, Ho J, Fleet D J and Norouzi M. 2022. Photorealistic text-to-image diffusion models with deep language understanding//Proceedings of the 36th Conference on Neural Information Processing Systems. New Orleans, USA: NeurIPS

Sigal L, Balan A O and Black M J. 2010. HUMANEVA: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal of Computer Vision, 87(1/2): 4-27 [DOI: 10.1007/s11263-009-0273-6]

Sohl-Dickstein J, Weiss E A, Maheswaranathan N and Ganguli S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics//Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR.org: 2256-2265 [DOI: 10.5555/3045118.3045358]

Song J M, Meng C L and Ermon S. 2021. Denoising diffusion implicit models//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: OpenReview.net

Sutskever I, Vinyals O and Le Q V. 2014. Sequence to sequence learning with neural networks//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 3104-3112 [DOI: 10.5555/2969033.2969173]

Taheri O, Choutas V, Black M J and Tzionas D. 2022. GOAL: generating 4D whole-body motion for hand-object grasping//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13253-13263 [DOI: 10.1109/CVPR52688.2022.01291]

Tang X J, Wang H, Hu B, Gong X, Yi R F, Kou Q L and Jin X G. 2022. Real-time controllable motion transition for characters. ACM Transactions on Graphics, 41(4): #137 [DOI: 10.1145/3528223.3530090]

Tevet G, Gordon B, Hertz A, Bermano A H and Cohen-Or D. 2022. MotionCLIP: exposing human motion generation to CLIP space//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 358-374 [DOI: 10.1007/978-3-031-20047-2_21]

Tevet G, Raab S, Gordon B, Shafir Y, Cohen-Or D and Bermano A H. 2023. Human motion diffusion model//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net

van den Oord A, Vinyals O and Kavukcuoglu K. 2017. Neural discrete representation learning//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc.: 6309-6318 [DOI: 10.5555/3295222.3295378]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010 [DOI: 10.5555/3295222.3295349]

von Marcard T, Henschel R, Black M J, Rosenhahn B and Pons-Moll G. 2018. Recovering accurate 3D human pose in the wild using IMUs and a moving camera//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 614-631 [DOI: 10.1007/978-3-030-01249-6_37]

Wang M, Qiu F, Liu W T, Qian C, Zhou X W and Ma L Z. 2020. Monocular human pose and shape reconstruction using part differentiable rendering. Computer Graphics Forum, 39(7): 351-362 [DOI: 10.1111/cgf.14150]

Wang Z M, Wang J, Ge N and Lu J H. 2023. HiMoReNet: a hierarchical model for human motion refinement. IEEE Signal Processing Letters, 30: 868-872 [DOI: 10.1109/LSP.2023.3295756]

Yi H W, Liang H L, Liu Y F, Cao Q, Wen Y D, Bolkart T, Tao D C and Black M J. 2023. Generating holistic 3D human motion from speech//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 469-480 [DOI: 10.1109/cvpr52729.2023.00053]

Yi X Y, Zhou Y X, Habermann M, Shimada S, Golyanik V, Theobalt C and Xu F. 2022. Physical inertial poser (PIP): physics-aware real-time human motion tracking from sparse inertial sensors//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13157-13168 [DOI: 10.1109/CVPR52688.2022.01282]

Yuan Y, Song J M, Iqbal U, Vahdat A and Kautz J. 2023. PhysDiff: physics-guided human motion diffusion model//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 15964-15975

Zhang H, Wang J M and Liu H. 2021. Video-based reconstruction of smooth 3D human body motion//Proceedings of the 4th Chinese Conference on Pattern Recognition and Computer Vision. Beijing, China: Springer: 42-53 [DOI: 10.1007/978-3-030-88007-1_4]

Zhang J R, Zhang Y S, Cun X D, Zhang Y, Zhao H W, Lu H T, Shen X and Ying S. 2023a. Generating human motion from textual

descriptions with discrete representations//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 14730-14740 [DOI: 10.1109/CVPR52729.2023.01415]

Zhang M Y, Cai Z A, Pan L, Hong F Z, Guo X Y, Yang L and Liu Z W. 2022. MotionDiffuse: text-driven human motion generation with diffusion model [EB/OL]. [2023-03-21]. https://arxiv.org/pdf/2208.15001.pdf

Zhang M Y, Guo X Y, Pan L, Cai Z A, Hong F Z, Li H R, Yang L and Liu Z W. 2023b. ReMoDiffuse: retrieval-augmented motion diffusion model//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 364-373 [DOI: 10.1109/ICCV51070.2023.00040]

Zheng J T, Zheng Q Y, Fang L X, Liu Y and Yi L. 2023. CAMS: CAnonicalized manipulation spaces for category-level functional hand-object manipulation synthesis//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 585-594 [DOI: 10.1109/CVPR52729.2023.00064]

Zhuang W L, Wang C Y, Chai J X, Wang Y G, Shao M and Xia S Y. 2022. Music2Dance: DanceNet for music-driven dance generation. ACM Transactions on Multimedia Computing, Communications, and Applications, 18(2): #65 [DOI: 10.1145/3485664]

Zou Y L, Yang J M, Ceylan D, Zhang J M, Perazzi F and Huang J B. 2020. Reducing footskate in human motion reconstruction with ground contact constraints//Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass, USA: IEEE: 448-457 [DOI: 10.1109/WACV45572.2020.9093329]

作者简介

赵宝全，男，副教授，主要研究方向为计算机图形学、数字人、跨媒体智能和多模态信息智能处理与分析。
E-mail: zhaobaoquan@mail.sysu.edu.cn

苏卓，通信作者，男，副教授，主要研究方向为智能可视媒体计算、基于深度学习的新兴人工智能应用、计算机图形学与可视化、计算机视觉与图像处理。
E-mail: suzhuo3@mail.sysu.edu.cn

付一愉，男，硕士研究生，主要研究方向为数字人。
E-mail: fuyy28@mail2.sysu.edu.cn

王若梅，女，教授，主要研究方向为计算机图形学理论和应用、三维计算机仿真、视频数据分析与处理。
E-mail: isswrm@mail.sysu.edu.cn

吕辰雷，男，助理教授，主要研究方向为计算机图形学、三维视觉、几何分析和机器学习。E-mail: aliexken@163.com

罗笑南，男，教授，主要研究方向为数字家庭技术、移动计算、图形图象处理及三维仿真CAD技术。
E-mail: luoxn@guet.edu.cn