软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(5):1982−2005 [doi: 10.13328/j.cnki.jos.007539] [CSTR: 32375.14.jos.007539]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

姿态控制人物图像视频生成技术综述*

李玘芮[1], 励雪巍[2], 赵 奇[1], 李 杰[1], 李 玺[1]

[1](浙江大学 计算机科学与技术学院, 浙江 杭州 310013)

[2](上海电机学院 电子信息学院, 上海 201306)

通信作者: 李玺, E-mail: xilizju@zju.edu.cn

摘 要: 生成技术的飞速发展揭示了相关技术在实际应用中的潜力, 姿态控制人物图像视频生成技术 (pose-guided person image and video generation) 的核心目标是将输入信息的人物转换为指定姿态, 同时保持人物外观的高度一致性. 该技术可以广泛应用于虚拟试穿与时尚行业、广告内容生成领域的视频生成与编辑以及多模态结合生成等多个应用场景, 推动用户体验和技术创新的进步. 尽管该技术已经取得了显著进展, 仍面临着多个挑战, 包括姿态迁移过程中外观信息的有效提取和重排、不可见信息的生成、一致性保持、模型的高效训练与使用等. 基于现有技术的挑战, 详细分析了当前主流的姿态控制生成方法应对挑战的策略, 并探讨了它们在实际应用中的可行性和局限性. 同时, 还讨论了姿态控制生成技术的常用生成模型以及不同的姿态信息表示方法. 此外, 整理讨论了该技术常用的数据集大小、特点等信息、各项测试基准, 并从虚拟试穿、视频生成与编辑、多模态结合生成等应用场景展开了讨论. 此外, 还揭示了目前方法仍遇到的个性化信息的保留、复杂场景的生成以及模型效率与实时性能等挑战, 并讨论姿态控制生成技术可能的未来发展趋势, 旨在为相关领域的研究人员提供系统的总结与参考, 以期推动该技术在各行业中的应用与创新.

关键词: 姿态控制生成; 人物图像视频生成; 生成对抗网络; 扩散模型; 可控生成

中图法分类号: TP181

中文引用格式: 李玘芮, 励雪巍, 赵奇, 李杰, 李玺. 姿态控制人物图像视频生成技术综述. 软件学报, 2026, 37(5): 1982–2005. http://www.jos.org.cn/1000-9825/7539.htm

英文引用格式: Li QR, Li XW, Zhao Q, Li J, Li X. Review on Pose-guided Person Image and Video Generation Technologies. Ruan Jian Xue Bao/Journal of Software, 2026, 37(5): 1982–2005 (in Chinese). http://www.jos.org.cn/1000-9825/7539.htm

Review on Pose-guided Person Image and Video Generation Technologies

LI Qi-Rui[1], LI Xue-Wei[2], ZHAO Qi[1], LI Jie[1], LI Xi[1]

[1](College of Computer Science and Technology, Zhejiang University, Hangzhou 310013, China)

[2](School of Electronic Information Engineering, Shanghai Dianji University, Shanghai 201306, China)

Abstract: The rapid development of generative technologies has revealed their potential for real-world applications. The core objective of pose-guided person image and video generation is to transform a person from inputs into a specified pose while maintaining a high level of appearance consistency. This technology can be widely applied in various fields such as virtual try-on and fashion, advertising video generation and editing, and multimodal content creation, driving advancements in user experience and technological innovation. However, despite significant progress, the technology still faces multiple challenges, including effective extraction and rearrangement of appearance information during pose transfer, generation of unseen information, consistency preservation, and efficient model training and deployment. Based on the existing challenges, this study provides a detailed analysis of the strategies employed by current mainstream pose-guided generation methods to address these issues, discussing their feasibility and limitations in practical applications. Moreover, it explores the

* 本文由“多媒体智能理解与生成”专题特约编辑孙立峰教授、闵巍庆副研究员、马占宇教授、蒋树强研究员、彭宇新教授、田丰研究员、黄庆明教授推荐.
收稿时间: 2025-05-19; 修改时间: 2025-07-11; 采用时间: 2025-09-05; jos 在线出版时间: 2025-09-23
CNKI 网络首发时间: 2025-12-26

commonly used generative models and pose representation methods in pose-guided generation. It also reviews the datasets, their sizes, characteristics, and evaluation benchmarks used in this field. Furthermore, this study discusses the applications of this technology in virtual try-on, video generation and editing, and multimodal content generation. It highlights the remaining challenges, such as the retention of personalized information, generation in complex scenes, and model efficiency and real-time performance. Finally, this study discusses potential future development trends of pose-guided generation technology, aiming to provide researchers with a systematic summary and reference to promote its application and innovation across industries.

Key words: pose-guided generation; person image and video generation; generative adversarial network (GAN); diffusion model; controllable generation

随着计算机视觉、图形学和深度学习技术的飞速发展, 姿态控制人物图像视频生成技术 (pose-guided person image and video generation) 逐渐成为人工智能领域的重要研究方向. 如图 1 所示, 该任务通常在给定人物外观 (appearance) 和人物姿态 (pose) 的输入条件下进行, 人物的外观通常由一张二维图像表示, 而指定的人物姿态则由二维关键点姿态等方法表示, 其核心目标是将输入中的人物转换为指定姿态, 同时确保人物的外观特征得以保留. 通过精确控制人物的姿态并保持其外观特征, 我们能够生成逼真的人物图像或视频, 这对于提升用户体验和推动技术创新具有重要意义.

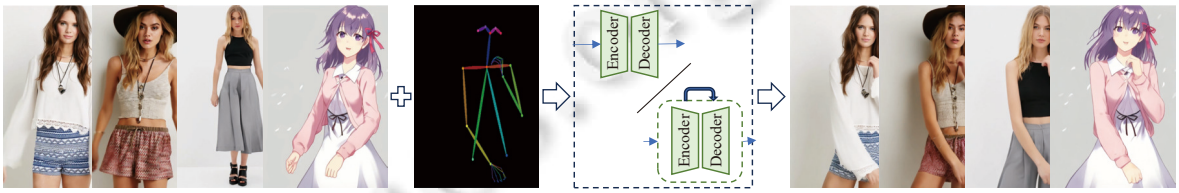


图 1 姿态控制人物图像视频生成任务

这一技术不仅在娱乐、游戏、虚拟现实等行业中具有重要应用[1–4], 还在时尚电商、广告内容生成和社交媒体内容创作等多个领域[5–11]展示了其潜力. 在娱乐和游戏行业, 姿态控制人物图像视频生成技术可以用于创建逼真的虚拟角色, 增强玩家的沉浸感; 在时尚电商领域, 姿态控制生成技术使得消费者能够在没有试穿的情况下, 通过虚拟试衣体验快速看到自己穿上不同服装后的效果[6,7,9]; 在广告内容生成领域, 品牌商可以利用这项技术通过自动生成广告短片或个性化内容来减少拍摄成本, 提高创作效率; 此外, 社交平台也逐步采用这一技术, 通过虚拟人物和增强现实 (AR) 互动, 增强用户体验和娱乐性[10,11].

姿态控制人物图像视频生成技术的广泛应用, 离不开近年来各类生成模型等前沿技术的进步. 特别是生成对抗网络 (generative adversarial network, GAN)[12]和扩散模型[13]等生成模型, 这些方法的成功使得姿态生成任务在视觉效果上逐渐突破传统方法的局限, 达到前所未有的精度和真实感. 生成对抗网络通过对抗训练的方式有效提升了生成图像的质量, 尤其在处理高复杂度姿态时, 能够生成高质量、自然流畅的人物图像. 扩散模型作为近年来崭露头角的生成模型, 凭借其逐步去噪的生成机制, 使得图像细节可以被更加精确地捕捉, 从而进一步提升了生成人物的真实感. 与此同时, 随着更高效的姿态估计和 3D 建模的发展, 生成的虚拟人物不仅能够在二维平面上表现出自然流畅的动作, 更能够模拟复杂的三维空间姿态和动态行为. 这一进步使得虚拟人物图像视频生成的应用场景不再局限于静态图像的生成, 而扩展到视频生成、实时互动和增强现实等领域, 极大地丰富了应用场景.

尽管这一领域取得了显著进展, 姿态控制人物图像视频生成仍然面临诸多技术挑战. 首先, 如何在完成多变的姿态迁移任务的情况下保留外观信息并保持外观的高度一致性是一个关键的问题. 尤其是在人物姿态发生较大变化时, 需要保持面部、衣物、身体细节等外观特征的一致性, 避免在姿态转换过程中出现失真或不自然的现象. 其次, 由于姿态生成任务通常需要处理复杂场景中的遮挡和不可见部分, 如何推测并生成符合风格一致性的外观来补充这些缺失的外观信息, 也是一项严峻的挑战. 此外, 姿态、背景和时间一致性问题, 即如何确保在生成过程中保持人物动作的流畅可控、背景的稳定和谐和时间一致性也是动态视频生成中的一大难题. 这些挑战的解决对于提升生成质量、增强应用场景中的用户体验至关重要.

现有的综述文献虽然从多个角度探讨了姿态控制人物图像视频生成技术的进展, 但在基于扩散方法学习和视频生成方法的论述上仍显不足, 特别是在从解决技术挑战的角度进行文献综述方面的讨论相对薄弱. 本文旨在围绕现有技术的挑战出发, 对当前主流的姿态控制生成方法进行详细分类和分析, 并探讨它们在实际应用中的可行性与局限性.

具体而言, 本文将从以下几个方面展开讨论: (1) 介绍姿态控制人物图像视频生成技术的基础理论和前置知识, 重点包括生成模型、姿态表示等关键技术; (2) 基于姿态控制人物图像视频生成任务遇到的姿态迁移过程中外观信息的有效提取和重排、不可见信息的生成、一致性保持、模型的高效训练与使用等几大难点, 详细分析和探讨现有研究方法的相关解决策略; (3) 讨论现有方法的常用的数据集大小、特点等信息, 并整理了相关测试基准, 并从虚拟试衣、视频生成和编辑和多模态结合生成等应用场景展开讨论, 分析其在实际应用中遇到的限制和现有技术仍遇到的技术难点; (4) 讨论姿态控制人物图像视频生成技术可能的未来发展趋势.

本文的目的在于为研究人员提供一个系统的总结和参考, 从而推动姿态控制人物图像视频生成技术在各行业中的应用与创新.

1 前置知识

姿态控制人物图像视频生成技术涉及多个前沿领域的交叉与融合. 为便于理解, 本文将首先介绍该技术的核心理论与基础技术, 包括生成模型、姿态数据表示等关键技术.

1.1 生成模型

生成模型是姿态控制人物图像视频生成技术的核心, 决定了生成效果的质量与多样性. 生成模型的目标是从输入数据中学习到潜在的分布, 并利用该分布生成与输入条件匹配的图像或视频. 在姿态控制人物图像视频生成任务中, 生成模型不仅需要根据输入的姿态数据生成人物图像, 还需要确保生成结果符合源图像中人物的外观, 也符合现实世界的物理规律和美学要求.

1.1.1 生成对抗网络 (GAN)

GAN 是一种由生成器和判别器组成的模型, 通过两者之间的对抗训练来学习数据分布. 在图像生成领域, GAN 主要是结合卷积神经网络实现生成能力, 生成器尝试生成逼真的图像以欺骗判别器, 而判别器则努力区分真实图像和生成图像. 通过这种“博弈”过程, 生成器逐渐学习到如何生成更为真实的图像. 在姿态控制人物图像视频生成任务中, 条件生成对抗网络 (cGAN)、pix2pix[14]和 pix2pixHD[15]等方法常用于生成符合特定姿态要求的虚拟人物图像. 例如, PG^2[16]使用条件生成对抗网络来学习将特定姿态信息映射到目标图像, 生成具有预定姿态的人物图像. 但由于其特殊优化目标, GAN 容易出现不稳定的训练结果, 导致生成的样本多样性有限.

1.1.2 变分自编码器 (VAE)

变分自编码器 (variational autoencoder, VAE)[17]是一种通过最大化数据的似然函数来学习潜在空间表示的生成模型. 与 GAN 相比, VAE 通常在图像生成的精度上有所欠缺, 但在多样性和模型训练的稳定性方面表现更好.

VAE 在姿态生成中常用于将图像转化为潜在空间表示, 通过操控潜在空间的特征来控制生成图像的姿态和外观. VAE+GAN 组合模型也常被应用, 比如 VU-Net[18]组合 VAE 与 GAN, 以平衡生成质量和多样性, 提高姿态迁移的生成水平.

1.1.3 扩散模型 (diffusion model)

扩散模型是近年来在图像合成领域崛起的一种生成模型. 与 GAN 不同, 扩散模型通过逐步去噪的方式生成图像或视频. 扩散模型的工作原理包括两个关键阶段: 正向扩散过程和逆向生成过程, 它通过定义一个扩散步骤的马尔可夫链, 在正向扩散过程中逐渐向数据添加随机噪声, 然后在逆向生成过程中学习逆扩散过程, 从噪声中构建所需的生成结果. 由于其生成过程中的精细控制, 扩散模型能够产生更加细致且真实感较强的图像或视频效果.

在人工智能生成领域解决 GAN 模型训练不稳定和生成数据多样性不足等问题上, 扩散模型逐渐成为一种重要的替代方法. 然而, 扩散模型在速度方面存在明显的缺点, 因为它需要大量的迭代来逐步去噪, 使得生成过程相对较慢. 此外, 扩散模型的训练过程也可能出现不稳定, 需要精心设计的网络架构和训练策略来确保模型的稳定性

和生成质量.

在姿态控制生成中, 扩散模型已被应用于生成符合特定姿态要求的人物图像和视频, 并展示了其在细节生成和质量上的优势. 比如 PIDM[19]作为完成姿态控制生成任务的第 1 个基于扩散的方法, 用渐进变换的方法实现了高质量的合成结果.

1.2 姿态数据的表示

人体的姿态是指一个人某一个时刻静态的身体形态, 包含对身体各部位如: 头、手臂、腿部等所在位置的描述. 在姿态控制的人物图像视频生成任务中, 人体姿态的表示对人体姿态描述的完整程度一定程度上影响了相关生成模型的生成能力[20]. 常用的姿态表示方法如图 2 所示.

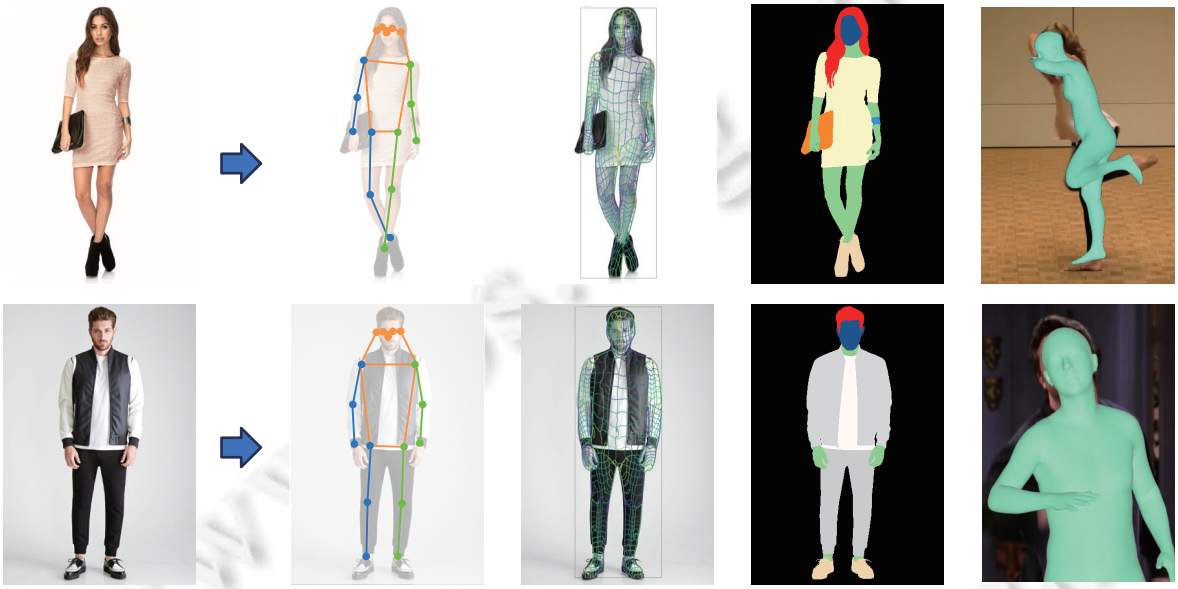


图 2 常用姿态表示方法示意图

1.2.1 二维关键点姿态

较为直接和便于操作的姿态表示是二维关键点姿态表示, 它通过人物不同关键点在图像中的位置来表示姿态, 这些关键点对应着人体上 25 个有一定自由度的关节, 比如: 颈部、肩部、肘关节、腕关节、膝关节等. 该姿态表示除了人工标注, 通常通过二维人体姿态估计方法来获得, 比如完成单人姿态估计任务的 DeepPose[21]以及完成多人姿态估计任务的 OpenPose[22]和 AlphaPose[23]. 二维关键点表示姿态的方法的优点是直接且便于操作, 也保留了一部分关键的姿态信息, 但是由于二维表示的局限性, 在人体姿态迁移的任务中不能有效地使生成模型获得空间信息[24], 从而影响了生成模型对一些有交叠肢体的姿态的理解. 同时由于这种方法只包含了对关键点位置信息的描述, 没有对关节扭转角度的关键信息, 增加了生成模型的学习负担.

1.2.2 深度姿态

深度姿态技术通过将二维图像中的每个像素映射到三维人体表面, 为姿态估计和生成任务提供了更加丰富的细节信息[25]. DensePose[26]是这一领域的代表性技术, 可以输出在包含人物的图像上预测的 IUV 图 (图像像素与三维人体模型之间的映射关系图), IUV 图代表了稠密网格顶点与图像像素之间的对应关系, 每一个有效的像素点都对应了一个三维向量, 包含了该点所在人体模型的位置, 例如头部、手臂等以及对应的贴图坐标. IUV 图可以通过预定义的双射映射将信息映射回三维的 SMPL (skinned multi-person linear) 模型, 在人体姿态迁移任务中, 可以帮助确保转移的姿态和人体的外观纹理对齐. 但同时, 使用 IUV 图也需承担一定的计算开销.

1.2.3 人体解析图像

人体解析图像指在图像解析过程中生成的分割图像, 也称为标注图或者是掩膜, 其根据语义为人体图像上的像素标注类别或标签, 这些类别或标签指示了背景或者人体不同的部位, 比如头发、手臂、躯干等, 也会指示不同的着装类别, 比如上衣、裤子、连衣裙等. 它相对于仅指示关键点位置的姿态表示来说, 包含了语义信息, 从而能够帮助保留不同区域上的外观纹理细节, 也能更有效地处理姿态遮挡的问题. 但同时, 获取准确的人体解析图像相较于基于关键点表示的姿态更具挑战性, 并且其准确程度也会影响模型的训练效果.

1.2.4 SMPL 模型

SMPL 模型是一种基于顶点的参数化人体模型, 可以通过参数来控制以及准确地表示人体姿态[27]. 具体来说, SMPL 模型包括 6 890 个顶点、23 个分别拥有 3 个自由度的关节以及 1 个代表全局方向的根节点, 通过 75 个 θ 参数和 10 个 β 参数来分别控制姿态和人体体型. 除了 SMPL 模型之外, 还有能模拟软组织的 DMPL[28]、加入了手部信息的 SMPL+H 以及加入了手部和面部信息的 SMPL-X[29]等扩展模型. 这些模型通过参数化的方式能够实现对人体体型和姿态的精确控制, 能够帮助模型定位人体外观纹理位置, 但在姿态控制人物图像视频生成任务中使用 SMPL 模型同时也意味着需要承担相当大的计算开销.

2 姿态控制人物图像视频生成挑战与解决策略

姿态控制人物图像视频生成任务通常给定两种信息: 人物外观和人物目标转移姿态. 当前, 外观与姿态信息的表达方式尚不能足够精细、完整地提供生成人物所需的信息, 这导致该任务在生成合理、准确且真实的人物图像方面面临巨大挑战. 所以, 要在有限的信息输入下精准地控制生成人物的姿态和合理推理其服饰在不同姿态下的表现是极具挑战的任务, 这些挑战主要集中在以下几个方面: 人物外观信息的保留与一致性、不可见外观信息的生成、一致性问题以及高效训练与使用模型. 如图 3 所示, 本节将围绕这些挑战展开讨论, 结合当前的研究成果与技术发展, 对每项挑战背后的核心问题进行分析, 并探讨现有方法的解决策略和局限性.

图 3 姿态控制人物图像视频生成挑战、数据集与应用示意图

2.1 人物外观信息的保留

人物外观信息的保留是姿态控制生成任务中的核心问题之一. 其目标是确保生成结果的外观细节 (如面部、衣物纹理等) 与输入人物图像高度一致, 尤其是在姿态发生显著变化的情况下, 这一问题尤为重要. 若外观信息未能有效保留, 生成结果可能会出现失真、模糊或与原始图像不一致等现象, 从而降低生成效果的可用性和可信度.

在姿态迁移任务中, 人物的姿态变化可能会导致: 在语义复杂区域 (面部或衣物的复杂图案、纹理部分) 人物面部 ID 和服装纹理等细节无法完整保留, 生成结果在特定区域出现模糊伪影, 或者是出现错误的纹理、不合理的衣服褶皱和人体轮廓, 导致生成结果不自然不可信. 这些问题与生成模型处理复杂图像特征, 特征提取能力密切相关.

2.1.1 提取有效的整体外观特征

在人物姿态生成领域, 一个有效的解决思路是尽可能完整地提取和保留外观特征, 尤其是面部和衣物纹理等细节信息. 这些外观特征在姿态变化过程中容易受到模糊或丢失, 因此需要借助合适的网络结构来准确提取和保持这些细节.

为此, 如图 4(a) 所示, 一些方法[1,30–32]提出了通过利用大规模数据集预训练的模型 (如 CLIP[33]和 VAE[17]) 来提取图像的外观特征, 再将这些特征作为条件输入到生成网络中, 以此增强生成图像在细节保留方面的能力. DreaMoving[31]和 MotionFollower[32]采用单一模型来提取全局特征, 但由于 CLIP 在处理细粒度纹理信息时存在一定限制, 一些研究 (如 DreamPose[30]和 DisCo[1]) 进一步结合了 VAE 模型和适配器模块, 对特征进行多维度提取, 从而获得更加完整的外观特征. 此外, 一些基于 U-Net 结构的方法[34–37]也被提出. 如图 4(b) 所示, 这些方法通过 U-Net 架构的编码器提取图像的全局特征, 再通过解码器将提取的特征应用到生成过程中. MagicPose[34]提出使用多源注意力模块来提供详细的外观指导, 以增强生成过程中的细节表现; MagicAnimate[35]主张通过提取密集的视觉特征来保留参考图像中的身份和背景信息. 如图 4(c) 所示, Animate Anyone[36]设计了一个对称的 U-Net 结构, 并在每一层利用空间注意力将特征注入扩散方法中的去噪过程, 从而有效保留外观条件.

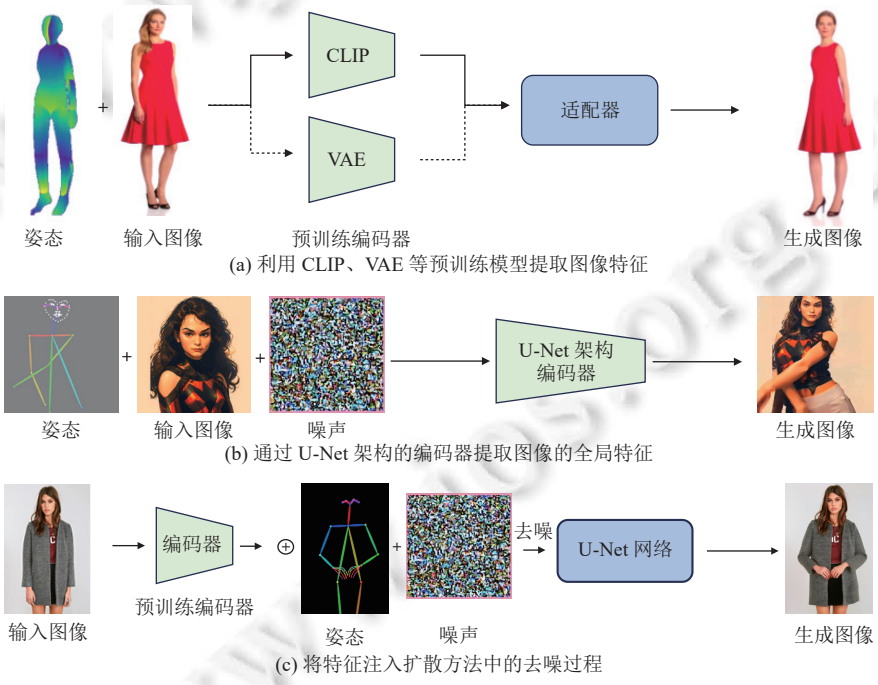


图 4 整体外观特征提取示意图

2.1.2 对不同属性的特征进行解耦

另一种有效的解决思路是通过对参考图像的不同属性特征进行解耦, 包括前景和背景的分离、外观和姿态的

分离、人体肢体属性的分离以及对人脸的关注. 这种方法通过局部和全局特征的分离, 使生成模型能够在关注细节的局部区域的同时, 保持全局结构的一致性, 从而提高生成结果的精度和稳定性.

例如, BodyROI7[38]、Yoon 等人[2]和 LWG[39]通过分离前景与背景, 使得背景成为可编辑对象, 同时确保人物外观特征在复杂场景中的有效保留. BodyROI7 通过将输入图像的前景、背景和姿态解耦, 并编码为不同的嵌入特征, 从而实现对这些属性在生成过程中的可控性. 然而, 这种方法在生成质量上出现了下降, 提示在特征解耦时需要平衡精度与灵活性. Animate Anyone 2[40]在人物掩码上做了改进, 通过动态边界避免传统精确掩码导致的形状泄露 (比如服装变形等). 为了避免姿态变化对外观的干扰, 一些方法[18,41,42]进一步探索将参考图像的外观和姿态进行解耦, 从而避免姿态变化对外观的干扰. 这些方法能够有效将姿态变化和外观特征进行独立处理, 使得生成模型更专注于各自的特征, 从而提高生成效果的精细度.

为了更加关注人体的细粒度信息, 许多研究[2,38,43–47]选择使用人体解析图像的方法, 分别提取躯干、四肢等不同人体属性的特征, 从而在生成过程中更好地保留服装等外观细节. 这些方法通过分离人体各部分的特征, 使得生成过程能更灵活地控制各部位的细节表现. 在这些方法中, 一些使用现有的自动解析器来进行人体解析图像的生成[44,45], 并将不同人体部位的嵌入拼接形成风格编码. Pu 等人[45]意识到当属性类别数量庞大时, 该方法需要承担巨大的内存开销, 所以提出 ADGAN++使用串行编码策略来缓解这一问题.

在一些应用场景下, 使用者除了关注服装的纹理, 通常也高度关注人脸的真实性与一致性. 为了提高面部生成的自然性和一致性, 一些方法[48,49]专门设计了保留人脸细节的策略. 例如, Chan 等人[49]使用 pix2pixHD 框架结合专门的人脸生成对抗网络, 通过学习二维关键点姿态与图像之间的映射关系, 来实现更精确的面部生成. APS[50]则提出了一种更为灵活的策略, 旨在解耦和重新耦合多个不同的属性, 例如前景与背景、外观与姿态、局部细节与全局结构等. 通过这一方法, 生成模型能够在高灵活性和可控性的基础上, 生成高质量的行人图像, 从而满足不同应用需求的多样化要求.

2.2 对外观信息进行重排

在姿态控制生成任务中, 为了有效地保留外观信息并提升生成质量, 以上这些方法尝试提取具有强大表达能力的特征向量. 然而, 这些方法通常将特征向量均匀、平等地应用在对生成过程的控制中, 一些关键细节可能在最终输出中丢失或模糊. 为了实现空间自适应调整, 更多的技术通过估计密集变形对参考图像的外观进行扭曲以对齐特征. 如图 5 所示, 这些方法从寻求局部纹理细节指导的角度出发, 进行了从像素级或者特征级对信息进行扭曲的尝试.

图 5 重排外观信息的不同方法

2.2.1 纹理扭曲

纹理扭曲方法[43,47,51–57]旨在通过几何变换对输入图像的纹理信息进行调整, 以适配目标姿态, 从而更好地保留外观特征的细节和一致性. 这类方法通常借助仿射变换或结合 3D 建模技术, 完成纹理和外观特征的空间对齐,

进而优化生成图像的质量.

早期的工作[43]受空间变换网络 (spatial Transformer network, STN)[58]的启发, 提出了通过拟合源二维关键点姿态与目标二维关键点姿态之间的仿射变换矩阵, 利用 STN 将源图像调整为参考姿态. 这一方法的核心思想是通过粗略的几何变换来对齐输入图像的外观特征, 从而使其适应目标姿态. 然而, 由于人体姿态的变形是高度非刚性的, 这一假设并不总是适用于复杂的姿态变化, 因此, 假设原始姿态和目标姿态之间存在全局仿射变换的做法存在局限性. 为了解决这一问题, 后续的研究提出了更为精细的特征对齐方法. 例如, CoCosNet[55]和 CoCosNet v2[56]通过将输入信息映射到共享的中间域, 计算出相应的相关性矩阵, 进而建立密集的对应关系. 这一方法有效克服了仿射变换对细节捕捉的不足, 更好地处理了姿态间的复杂变换. 此外, NTED[47]则通过引入双重注意力机制来提取语义信息, 使得纹理信息在目标姿态下的对齐更加精确. 这一机制使得纹理扭曲过程不仅仅依赖几何对齐, 还融入了语义层面的匹配, 从而进一步提升了生成结果的精细度与真实感. 针对更复杂的姿态变换, 一些方法则结合了显式或隐式的三维人体信息, 以在三维空间内进行精细的纹理扭曲. 例如, Neverova 等人[54]和 M2E-try on Net[57]采用了预训练的姿态估计器 DensePose 对三维人体模型进行非刚性对齐, 借此实现了更为细致的纹理调整. 通过这种方式, 模型能够在三维空间中更准确地捕捉到人体姿态的变化, 从而提升生成结果的自然性和真实感.

另一类方法则基于显式的参数化人体模型, 进行姿态迁移任务. 这类方法能够比传统的深度姿态估计方法更好地关注身形的个性化特征, 尤其是四肢的旋转和细节. Zanfir 等人[53]通过直接使用基于重心坐标的变形对网格三角形的色彩进行转移, 从而实现了精确的纹理扭曲. Grigorev 等人[51]则采用深度卷积网络, 将外观信息映射为完整的身体纹理, 进一步提升了纹理调整的精准度和可控性. 此外, Li 等人[52]提出了一种新的方法, 通过根据可见性图将特征图划分为互斥的区域, 并通过门控层生成完整的被扭曲的特征图. 这一方法通过划分特征区域, 能够更加精确地控制哪些部分的纹理需要调整, 从而优化了生成结果的质量, 避免了传统方法中可能出现的纹理扭曲不自然或不协调的问题. 由于这种借助三维人体信息的对齐无法处理由剧烈姿态变化引起的遮挡区域, 因此使用修复来填充遮挡区域. 然而, 在被遮挡的区域结果通常是模糊的, 同时值得注意的是无论显式还是隐式, 也无论是深度姿态还是三维参数模型这样的密集信息, 除了数据集标注稀缺的问题外, 还会增加模型的信息处理负担, 也会引入相关估计方法不准确所带来的误差. 此外, 三维参数模型更侧重于身体重建而非衣服表面的重建, 所以在特定姿态的生成任务中, 生成结果的外观上所产生的褶皱并不总是精确的.

2.2.2 深层特征扭曲

深层特征扭曲方法旨在克服基于像素的传统纹理扭曲在姿态变化过程中无法有效保留源图像色彩风格和细节的局限性. 不同于直接在图像空间进行纹理扭曲, 这些方法通过在深层特征空间内进行扭曲, 试图更深入地理解和适应显著姿态变化下的纹理变化. 这种方法尤其适用于需要对复杂姿态变化下的细节进行精细处理的任务.

早期的研究 (如 Def-GAN[59]) 通过引入可变形的跳跃链接和局部仿射变换来处理由姿态差异引起的像素错位. 尽管该方法能够对源图像进行一定程度的姿态对齐, 但由于仿射变换是线性的, 难以应对复杂的非刚性姿态变形, 生成的边界通常较为模糊, 且容易受到姿态估计不准确的影响, 导致伪影或外观丢失等问题. 为了解决这一问题, 许多后续研究引入了更为灵活的非线性变换. 例如, Warping GAN[5]结合了仿射变换和薄板样条变换 (thin plate spline, TPS), 采用软门控机制来对空间进行更精细的对齐. 尽管 TPS 能够在一定程度上处理非刚性变形, 但其在高维空间的适用性仍然受到限制, 尤其是在复杂姿态迁移中, TPS 可能无法完美捕捉高频细节变化.

考虑到非刚性姿态形变对纹理扭曲的影响, 一部分方法选择基于全局流从源图像迁移外观细节[6,39,52,60–63]. Li 等人[52]除了上一节提到的纹理扭曲之外, 也考虑了深层特征扭曲, 考虑到引入三维模型信息的推理负担问题, 提出了仅从二维表示中集成关于三维几何信息的隐式推理方法, 通过外观流生成模块编码从源到目标密集的对应关系, 并据此扭曲纹理和特征. Zheng 等人[60]也通过学习到的姿态流来学习复杂的姿态变形, 但提出一种无监督的流学习方案, 避免了效率较低的流真值的生成步骤. 与基于三维信息学习流的 LWG[39]和 Li 等人[52]提出的方法不同的是基于解析图像的外观流生成方案 ClothFlow[63], 它首先合成一个目标人体解析图, 然后估计一个密集的流来扭曲源图像特征, DwNet[62]则通过估计变形网格来扭曲特征. 而考虑图像特征和像素位置之间的相互依赖问题, Tabejamaat 等人[61]则鼓励流只关注外观未正确生成的区域, 生成独立图像块. 基于流的方法虽然可以生成逼真的

细节, 但是也容易受到复杂形变和严重遮挡的影响.

一些方法考虑加入注意力机制, 在特征级别对原特征进行扭曲[47,61,64–68]. GFLA[67]提出的网络可分为全局流场估计器和局部神经纹理渲染器两部分, 全局流场估计器负责提取全局相关性并生成流场, 局部神经纹理渲染器根据得到的流场利用局部注意力块为目标采样逼真的源纹理. NTED 则引入双重注意力机制, 首先从参考特征图中提取语义神经纹理特征, 然后根据从目标姿态中学习到的空间分布, 对提取的神经纹理进行分布操作. attLWG[66]在 LWG 的基础上加入注意力机制, 以解决网络假定每个元素的原特征对目标生成成果的影响程度相同的问题.

在基于流的方法中引入注意力机制使合成结果更加真实, 但有显著姿态变化的情况仍对生成模型影响较大, 在一次生成中对模型学习能力要求极高, 很难产生鲁棒的生成效果, 因此一些方法考虑使用渐进式的生成方案[65,68,69]. PATN[68]使用渐进式生成的方案, 依靠注意力机制避免捕获全局流形的复杂结构的挑战, APATN[69]在 PATN 的基础上进行改进, 显式计算条件姿态和目标姿态之间的对齐关系, 为特征转移提供了更直接和平滑的指导, 而考虑到 PATN 只参考姿态来生成注意力, XingGAN[65]让人体和外观特征相互参考进行改进. BiGraphGAN[64]则提出的二部图推理 (BGR) 模块推理二部图中源姿态和目标姿态之间的交叉长程关系, 从而减轻推理负担.

2.3 不可见外观信息的生成

在姿态控制人物图像视频生成任务中, 当目标姿态发生显著变化或出现肢体遮挡时, 部分区域的外观信息可能无法从输入图像中直接获取. 针对这一挑战, 研究人员提出了多种解决方案, 主要通过引入人体先验信息以及遮挡补全技术来生成高质量且风格一致的外观特征. 本文将从利用人体先验信息和遮挡补全技术两个方面展开讨论.

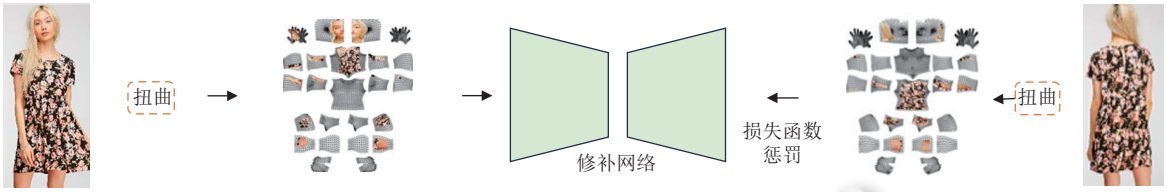
2.3.1 利用人体先验信息

人体先验信息 (如对称性、几何形状等) 为不可见外观信息的生成提供了关键的约束条件. 通过利用这些先验知识, 生成模型能够更合理地推测缺失部分的外观特征, 使生成结果更符合物理和视觉规律. 基于三维几何和 UV 补全的方法[51,70–74]主要通过将人体表面纹理映射到 UV 空间, 并对遮挡区域进行补全来实现. UV-GAN[71]通过构建 UV 补全网络来生成任意姿态的人脸, 将生成的 UV 图与三维可变形模型 (3DMM)[75]相结合, 从而合成二维图像. 在处理自遮挡问题时, UV-GAN 利用 UV 映射避免了传统基于颜色映射的模糊问题. 考虑到人体的自然对称性为不可见区域提供了可靠的推测依据, 一些方法[51,70,72]则利用了人体的对称先验. Pose with Style 方法[70]利用了人体对称性先验学习修复人体表面纹理和源图像之间的对应场, 并据此将从源姿态提取的局部特征迁移到目标姿态. 而 StylePoseGAN[73]通过将姿态条件化的 StyleGAN[76]与代理几何相结合, 能够生成具有全局一致性的外观特征, 但 StylePoseGAN 也指出它的全局解耦的方案限制了它保留复杂局部外观细节的能力. Grigorev 等人[51]则在合成了人脸的基础上, 提出了区别于基于颜色的基于坐标的纹理补全方法, 使得生成模型即使在可见信息有限的情况下也可以保留高层次的纹理细节. 3D-GCL[74]方法则联合编码了全局语义相关性、局部变形和 3D 人体几何先验, 使流具备三维感知能力, 并更好地处理姿态和视角的变化.

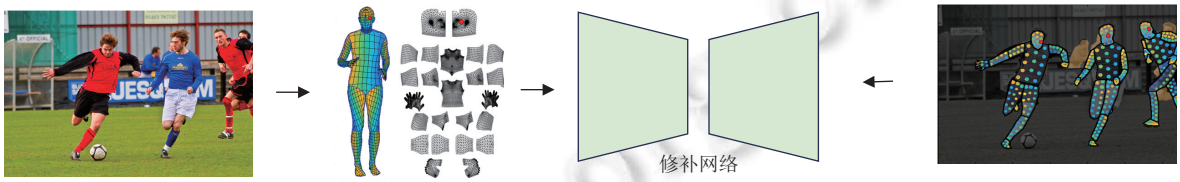
2.3.2 遮挡补全技术

遮挡补全技术的核心在于对目标图像的遮挡部分进行合理推测和补全, 使生成结果在视觉上更自然且符合逻辑. 这些方法通常结合三维建模技术与深度学习网络进行实现. 如第 2.2.1 节中所提到的利用三维人体信息进行辅助纹理扭曲的方法[51–54,57], 由于不能从输入图像中完整获取人体外观, 因此模型中从扭曲后的特征生成最终结果的过程, 也均需要在无法观察完整表面纹理的情况下有对不可见部分进行修复生成的能力. 这部分能力通常是使用基于 U-Net 的网络实现的. Neverova 等人[54]提出的方法允许从 STN 填充的表面节点扩展推理整个身体表面的外观, 由于在训练期间不能观察到完整的表面纹理, 因此, 如图 6(a) 示意图所示, 该方法通过仅计算 UV 图中观察到的部分的重建损失来进行训练. 此外, NHRR[72]也借助深度姿态来表示人的姿态, 如图 6(b) 所示, 有别于传统基于颜色的 UV 纹理映射, 该方法使用学习到的高维 UV 特征映射来编码外观, 再将中间特征图渲染成最终结果. 考虑结合多源的条件信息, 可以使得方法生成更精确. AFMHIG[77]利用多源条件输入对生成过程进行指导, 通过引入局部注意力机制来从不同的源图像区域选择相关信息, 避免为每个特定的源图像数量构建特定的生成器, 从而能

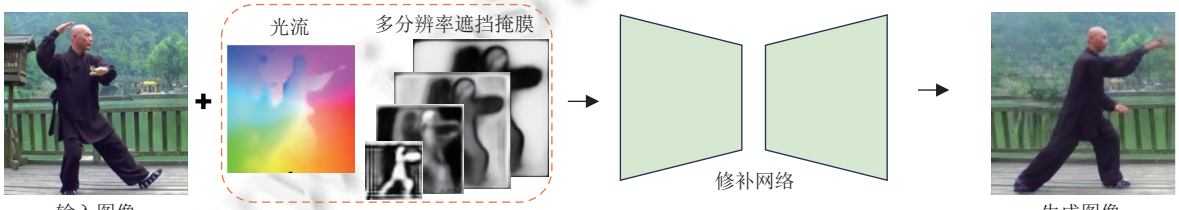
更有效地捕捉原人物的外观信息. 而 TPSMM[78], 如图 6(c) 所示, 为了更真实地恢复纹理缺失的区域, 为每一层扭曲特征图预测遮挡掩膜, 并通过多分辨率处理实现特征融合优化. 该方法能够针对不同层次的特征重点进行调整, 在恢复纹理缺失区域时表现优异.



(a) 仅计算 UV 图中观察到的部分的重建损失, 推理整个身体表面的外观



(b) 使用学习到的高维 UV 特征映射来编码外观



(c) 为每一层扭曲特征图预测遮挡掩膜, 通过多分辨率处理实现特征融合优化

图 6 遮挡补全技术代表方法示意图

2.4 一致性问题

在姿态控制人物图像视频生成任务中, 一致性问题是实现高质量生成效果的关键挑战之一. 它包括背景一致性、姿态一致性以及时间一致性. 这些问题不仅涉及对图像局部细节的生成质量, 还关乎整体生成结果在空间和时间维度上的自然性和稳定性.

2.4.1 背景一致性

背景的一致性是指在生成不同姿态人物的同时, 保持场景背景的视觉稳定性. 由于背景与目标人物之间一般相关性较弱, 背景部分在受到人物动作或姿态变化的干扰之后, 人体的遮挡会导致生成结果中的背景出现错位或不连贯. 在图像生成领域, 常见的解决背景一致性的方法是在前后景分离后, 单独对背景进行修复[39,66]. 而在视频生成领域则通常通过分离前后景运动, 追踪背景的运动来保证一致性[37,79,80]. Liu 等人[37]提出了一种稀疏跟踪点方法, 将前景和背景的运动表示分开, 通过跟踪背景的运动轨迹来确保背景稳定性. 而 Follow-your-pose v2[80]观察到, 直接在有噪声的数据集上优化模型会出现背景不稳定的问题, 因此提出利用光流的引导, 并将其与其他条件导引器集成, 进一步增强背景的整体稳定性. 此外, MRAA[79]通过引入额外的仿射变换, 对非目标相关的全局运动进行建模, 显式地对训练帧之间的背景或摄像机运动进行建模, 这样也能帮助模型更集中关注前景目标, 使识别的点更加稳定.

2.4.2 姿态一致性

姿态一致性指姿态控制生成任务中, 生成结果严格遵守输入姿态的要求. 由于图像生成人物不涉及时间维度的建模, 姿态迁移通常表现较好. 相较之下, 由于姿态序列的动态变化和数据质量等原因, 姿态一致性主要在姿态

迁移的视频生成任务中被关注. 视频生成的姿态迁移方法可以依据姿态控制方式分为用参考视频中人物动作进行控制[3,4,15,32,36,41,49,78,81–85]和使用抽象姿态序列进行控制[1,30,31,34,35]. 使用参考视频中的姿态进行控制的方法面临着准确提取运动信息的压力. 许多方法试图对运动表示进行建模, 解耦运动和外观信息以提高姿态生成的准确性. 一部分方法使用现成的姿态估计器对原视频的动作进行提取: Chan 等人[49]、Yang 等人[41]、TPSMM[78]通过使用预训练的姿态估计器 OpenPose[22]从原视频中提取二维关键点姿态, 考虑到不同视频中的人物可能有不同的肢体比例、视角以及与相机的距离, Chan 等人[49]通过全局姿态归一化对提取的姿态序列进行变换, 以适应目标人物的身形和位置. MotionFollower、MimicMotion[4]、Animate Anyone、UniAnimate[3]则使用 DWPose 工具[86]来从视频中提取二维关键点姿态序列. 而另一部分方法使用估计流的方法对动作进行建模[78,81,82]. Vid2vid[87]使用光流预测网络来估计视频帧之间的运动, 并使用 FlowNet2 提取的光流作为真值进行监督, Monkey-Net[81]通过预测无监督关键点估计运动的光流, 并用该光流驱动生成器生成目标姿态. FOMM[82]则引入了局部仿射变换, 在 Monkey-Net 的基础上在每个关键点附近进行一阶泰勒展开, 并使用局部仿射变换在每个关键点的邻域内估计运动, 然后用密集运动网络基于运动表示生成密集光流指导姿态控制. 但由于通过关键点来估计运动存在不明确的问题, 生成结果会混入背景的视觉信息, 帧间的对应关系也难以建立. MRAA 则改进了 FOMM 的缺点, 直接使用主成分分析计算运动热图中的平移、旋转和缩放等变换, 使用与有意义的语义部分相对应的区域而非关键点来表示运动, 能提供更丰富的运动信息. 此外, TPSMM 通过引入薄板样条变换来模拟复杂的人体姿态变化, 通过结合多个薄板样条变换和背景的仿射变换生成最终的光流, 并据此在生成过程中对源图像进行变形, 以匹配目标图像的运动.

在基于姿态序列的控制中, 一些方法通过利用人体先验知识, 让姿态的输入包含密集的姿态信息, 以提高姿态控制的准确性. 人体先验知识通常通过深度信息, 密集姿态、三维参数模型和网格等形式来传递. 比如 DreaMoving 加入 DWPose 工具提取的深度信息帮助模型理解身体不同部位与环境的空间关系, DreamPose[30]和 MagicAnimate 则利用 DensePose 提取密集姿态, 将这些细节融入生成器中以增强生成质量. 但此类姿态表示通常不能 表示出人体的关节约束等先验, Human4DiT[88]、Champ[89]和 VividPose[84]则引入了三维参数模型, VividPose 使用三维信息与深度和网格信息同时对姿态控制进行更高层次的条件约束, 从而更好地理解复杂运动. 为了更好地利用姿态序列约束生成结果, 不同方法探索了不同的姿态指导模块. DreamPose 和 Champ 直接连接姿态信息和噪声输入到去噪 U-Net 中, 不同之处在于, Champ 结合了渲染的深度图像、法线图和从 SMPL 序列获得的语义图, 以及基于骨架的运动指导, 以全面的 3D 形状和详细的姿态属性组成多层次的运动信息进行指导, MagicPose[34]、MagicAnimate、Animate Anyone、DisCo[1]基于 ControlNet[90]利用 U-Net 作为参考控制模块将姿态注入生成模型. 一些方法比如 MotionFollower、MimicMotion 和 UniAnimate 则设计了仅由卷积层组成的轻量神经网络作为姿态指导模块, 不涉及大量的注意力计算, 降低了计算代价.

2.4.3 时间一致性

姿态可控的角色视频生成技术在广告制作和内容创作等领域具有重要应用价值. 但是现有的视频生成方法在面对复杂场景, 例如存在身体遮挡、多角色互动或者复杂变化的背景的时候, 现有方法常常面临画面不连贯或细节跳动的问题. 因此, 现有方法通常需要依赖于具有稳定背景和良好时间一致性的大规模高质量视频数据集进行训练, 也就提高了在实际应用中的门槛. 研究者就如何解决确保连续帧之间的动作流畅和视觉连贯性的问题, 提出了多种方法, 例如三元组损失[91]、时间平滑策略[49]等. 其中, Yan 等人[91]引入三元组损失来约束时间一致性, 追求连续帧之间的外观平滑, 提高生成质量. Chan 等人[49]则通过时间平滑策略, 预测两个连续的帧, 减少了帧间闪动的现象. 此外, 受 AnimateDiff[92]的启发, 一些方法还通过合并时间运动模块来进一步增强时间一致性[35,36,80], 利用 U-Net 网络来捕捉具有复杂细节的视频特征. Animate Anyone 通过引入时间层建模多帧之间的关系, 从而显著提高了视频生成的稳定性, 也在模拟连续和平滑的时间运动过程的同时, 保留了视觉质量中的高分辨率细节. Liu 等人[37]提出了一种逐片段生成策略, 将长序列视频划分为多个片段, 其中每个前一个片段的终帧作为后一个片段的条件输入. 为了防止预测视频帧潜在的错误在整个生成过程中被累积, 他们设计了将从初始参考图像中提取全局特征持续注入生成过程的策略, 同时支持了视觉一致性和时间连贯性. HumanDiT[93]则在姿态迁移任务中, 插入 τ 帧过渡姿态, 平滑连接初始帧与目标姿态, 减少动作突变.

2.5 高效训练和使用模型

姿态控制人物图像视频生成任务中, 由于训练需要高标准的数据集以及存在一些如虚拟试衣等应用场景的需求, 如何提高模型的训练效率与实际使用中的推理性能是研究的关键问题之一.

2.5.1 数据集局限

在姿态控制任务中, 数据集的限制主要表现在两方面: 一是对数据集的内容要求较高, 一些全监督学习方法需要同一人在相同衣装下的不同姿态图像, 而一些模型对数据集复杂的背景比较敏感, 此类数据获取成本较高; 二是数据的多样性和规模不足, 导致模型在真实场景中的泛化能力受到一定的制约. 为了解决这一问题, 研究者提出了无监督学习、少样本泛化能力增强和数据集扩展等方法.

一部分方法提出通过无监督学习, 使用非配对数据引入循环一致性进行映射学习[94–96], 缓解了对成对数据集的依赖. UPIS[94]提出了完全无监督的策略, 通过以姿态为条件的双向生成过程, 将姿态迁移后的图像再映射回原姿态, 从而可以在使用非配对数据的情况下使用重建损失进行监督. 但同时由于其对数据映射的控制能力不足, 其生成效果较为有限. SPT[96]进一步引入语义信息, 在没有配对监督的情况下, 为语义解析转换创建伪标签, 利用循环一致性进行训练, 以完全无监督的方式生成人物图像. C2GAN[95]则引入三重循环一致性, 即一个图像生成循环和两个关键点生成循环之间相互隐式地约束, 提供了互补的信息. Wang 等人[85]则关注少样本泛化能力, 通过引入注意力机制的网络权重生成模块, 提升了少样本场景下的泛化能力, 使模型能够在测试时利用少量示例图像来学习合成以前未见过的主题或场景的视频. 一些方法尝试为数据稀缺提供更直接的手段, 通过利用更多的训练数据[97]或引入新的训练任务[98]来提升模型的表现. Follow your pose[97]利用图像-姿态对、无姿态视频以及预训练的文本到图像模型来生成姿态可控的视频, 扩大了可用数据的范围. 此外, DPTN[98]通过引入源姿态指导合成源图像的辅助任务, 探索双任务的相关性, 进一步增强了模型的训练效果.

2.5.2 模型性能

在实际应用中, 姿态控制生成模型的性能不仅取决于训练数据的质量, 还与模型架构设计和在推理阶段的高效性有关.

由粗到细的生成策略常用于应对模型参数的初始化问题, 也能保留不同尺度层面上的信息作为指导[16,45,48,56,60,99–102]. 采用由粗到细的生成策略的方法通常具有两个部分, 一个部分生成粗糙结果, 在第 2 个部分再进行细化补充. PG^2[16]、DRN[48]、ADGAN++[45]、PCDM[99]、Zheng 等人[60]的方法和 SCA-GAN[101]在第 2 阶段使用同一尺度的粗粒度结果进行指导. PG^2[16]开创性地提出了自上而下的基于生成式对抗网络的解决方案, 它首先提出了两阶段模型, 直接将带有外观信息的源图像和目标姿态连接起来作为模型输入, 在第 1 阶段主要关注全局的外观和结构, 生成相对粗略的结果, 第 2 阶段再在第 1 阶段初步结果的基础上进行外观细节的补充. SCA-GAN[101]利用像素级边缘映射的高频信号来指导生成过程, 从而增强服饰纹理和边缘细节的保留. 考虑到在应用中使用者更有可能将注意力集中在语义丰富和细节丰富的位置, DRN 方法[48]则除了在整张图上进行细节补充生成外, 还在人脸区域进行精细的转移补充. 但是只使用单尺度粗糙结果进行指导在一些显著姿态变化的情况下, 细节保留能力较差. 除了单尺度的指导外, 一些研究团队则考虑到将由粗到细的策略应用到多尺度的学习中[56,100,102], 能够在复杂姿态变化中更好地保留外观特征. CoCosNet v2[56]考虑到直接在全分辨率上建立由姿态到图的对应关系会增加模型拟合难度, 还容易受到细节信息的干扰, 于是采用分层策略, 将低分辨率的预测结果作为高分辨率的指导, 再通过 PatchMatch[103]高效地计算对应关系, 此举不仅节省了时间开销, 而且考虑了更大尺度上的上下文匹配和历史估计结果, 为最终生成提供多尺度的指导. Roy 等人[100]的方法利用不同分辨率级别的注意力链接来保留生成图像中的粗、细外观属性. CFLD[102]通过其所提出的感知细化解码器来逐步细化可学习查询来获得粗粒度的提示, 由此获得多尺度的纹理细节信息, 以保留源图像更准确的纹理细节.

一些方法则考虑引入渐进式生成模型, 以减轻每一步模型建模的复杂度, 以达成更好的生成效果. 但考虑到提高生成效果同时也承担了更大的计算开销, 一些方法则提出设计轻量级模型和优化模型能力的方法, 以降低计算复杂度、提高推理效率. 比如 MotionFollower、MimicMotion、Animate Anyone、UniAnimate 和 UniAnimate-Dit[104]

尝试设计轻量级姿态引导器, 在保持生成质量的同时, 减少了模型参数量和计算成本. 此外, 还能通过削减冗余计算进一步加速生成过程, 适合实时应用场景. 而 CoP[105]为减轻推理负担, 引入姿态链结构, 通过对生成器的条件约束进行分层建模, 避免了从源姿态一次性转换到目标姿态的挑战, 降低了高维输入对生成过程的复杂性, 同时确保生成结果的姿态和外观一致性.

3 数据集与评价指标

3.1 数据集

数据集是训练和评估姿态控制生成模型的重要基础, 根据模型设计, 不同方法对数据集的需求不同. 根据任务的具体需求, 本文对相关数据集进行收集整理, 按照其模态、内容类别将其分类, 其详细信息显示如表 1 和表 2, 分别按照图像类和视频类分类. 其中图像类较为常见的数据集是 Market-1501 和 DeepFashionHD, 姿态控制的图像生成方法通常同时在这两个数据集上进行训练和测试. 常用的视频类数据集包含 Human3.6M、Tai-Chi、TikTok 等, 其中, Tai-Chi 不是现成的数据集, 使用 Tai-Chi 需要自行通过提供的视频 id 和裁剪框从 YouTube 上下载和裁剪.

表 1 图像类姿态控制人物图像视频生成数据集

分类	名称	数据大小	分辨率	特点
行人图像	Market-1501[106]	1 501个人的32 668张图像	128×64	行人数据集, 姿态变化不显著
时尚服饰	DeepFashion[107]	52 712张图像, 约200 000对不同姿态相同外观图像	256×256	模特服装展示图, 人物体型、背景单一, 姿态偏向直立
	MVC[108]	161 260张图像, 每件相同衣物有4个视图	1960×2240	展示服装的图像, 部分展示下装的图像不包含人物上半身
	LookBook[109]	75 016张模特图像, 与对应的9 732件服装图像	256×256	背景较DeepFashion更复杂
	DeepFashionHD[107]	52 712张图像, 约200 000对不同姿态相同外观图像	1101×750	比DeepFashion有更高的分辨率, 达到1101×750
	MPV[110]	62 780个包含同一人物两张不同姿态的图像以及对应衣服图像的三元组	256×192	比DeepFashion多加了衣服平铺的图像, 但是在姿态控制任务中作用一致
	FashionOn[111]	11 283个包含同一人物两张不同姿态的图像以及对应衣服图像的三元组	288×192	暂未开源
	FashionTryOn[112]	28 714个包含同一人物两张不同姿态的图像以及对应衣服图像的三元组	256×192	与DeepFashion相同
	DeepFashion-MM[113]	44 096张图像, 对应的解析图像、二维关键点以及深度姿态	1101×750	完整人体图像, 含有对服装样式和材质以及对整体穿着的文字描述, 但缺少同一外观的不同姿态的成对图像
	SHHQ[114]	230 000张图像, 对应的解析图以及二维关键点	1024×512–2240×1920	手动注释的解析图像与二维关键点, 缺少同一外观的不同姿态的成对图像

表 2 视频类姿态控制人物图像视频生成数据集

分类	名称	数据大小	分辨率	特点
动作	UCF101[115]	13 320段视频	320×240	包含人与人、人与物交互的内容
	Penn action[116]	2 326段15个动作的视频片段, 对应的二维关键点姿态, 相机视角	640×480	室外场景
	Human3.6M[117]	11人的总共超过3 600 000帧的视频片段, 对应的二维关节点姿态, 像素级的身体部位的标注	1000×1000	室内场景, 17个动作, 每个动作有4个视角, 提供三维关节位置
	Tai-Chi[82]	280段太极视频	256×256	提供下载和预处理脚本, 需要自行从YouTube下载
	iPER[39]	30个人的206段视频, 总共包含241 654帧	256×256	包含不同体型的人物、不同风格的衣服
	iPER-HD[39]	30个人的206段视频, 总共包含241 654帧	1024×1024	比iPER的分辨率大, 达到1024×1024帧
	MotionVid[118]	1 200 000个文本-姿态配对	多分辨率	目前该领域最大规模的数据集

表 2 视频类姿态控制人物图像视频生成数据集 (续)

分类	名称	数据大小	分辨率	特点
时尚服饰	Fashion dataset[62]	600段视频, 每段视频350帧	940×720	模特服装展示视频, 动作包含转圈, 姿态偏向直立, 背景空白单一
	VVT[119]	791段视频, 总共190 101帧	256×192	时装模特走秀, 背景空白单一
舞蹈	Everybody dance[49]	105段视频, 对应的二维姿态	1024×512	包含人脸边框
	TikTok[120]	340段视频, 对应的UV坐标	1080×604	包含人物遮罩以及UV坐标

3.2 评价指标

评价指标是衡量姿态生成模型性能的重要工具, 主要分为标准化计算的客观指标与主观指标.

3.2.1 客观指标

在姿态控制人物图像视频生成任务中, 客观指标通常包含有衡量质量、多样性、美观度以及服装、姿态、时间上的一致性程度, 具体信息如表 3 所示. 表 4 引用了部分姿态控制生成方法在 TikTok 数据集上的评测, 结果引用自对应方法后标注的论文. 可以看出, 常用的客观指标仍然以衡量生成图像或视频的质量为主.

表 3 姿态控制人物图像视频生成任务指标

分类	名称	描述	英文全称
图像质量	L1误差[121]	测量预测帧和真实帧之间的绝对像素级差异	L1 error
	峰值信噪比 (PSNR)[122]	量化生成图像与真实图像之间的相似度, 单位为dB	Peak signal-to-noise ratio
	结构相似性指数 (SSIM)[123]	从亮度、对比度和结构这3个方面评价结构相似度	Structural similarity index
	学习感知图像块相似性 (LPIPS)[124]	基于深度学习的视觉相似性度量. 较低的LPIPS表明较高的相似性	Learned perceptual image patch similarity
	切片Wasserstein距离 (SWD)[125]	衡量从真实图像和生成图像中提取的投影特征图的分布之间的Wasserstein距离	Sliced Wasserstein distance
	Fréchet 初始距离 (FID)[126]	衡量生成图像和真实图像之间的分布距离	Fréchet inception distance
视频质量	核心视频距离 (KVD)[126]	测量生成视频序列与真实视频序列之间的分布距离	Kernel video distance
	Fréchet视频距离 (FVD)[126]	测量生成视频和真实视频的分布之间的距离	Fréchet video distance
	平均内容距离 (ACD)[127]	评估生成视频中动作序列的一致性, 尤其是对于姿态生成任务	Average content distance
	变形误差 (WE)[128]	获得每两帧图像的光流, 然后计算变形图像与预测图像的像素级差异	Warping error
	Fréchet姿态距离 (FGD)[129]	衡量真实姿态与生成姿态在特征空间中的分布差距	Fréchet gesture distance
	Fréchet模板距离 (FTD)[129]	与FGD类似, 度量生成特征与真实特征在特征空间中的分布相似性	Fréchet template distance
	视频的Fréchet inception距离 (FID-VID)[34]	衡量生成视频帧与真实视频帧之间的分布距离, 同时包含空间特征和时间特征. FID-VID越低, 质量越好	Fréchet inception distance for videos
多样性	Inception得分[130]	衡量生成图像的多样性和清晰度, 有时也用于视频质量	Inception score
	多样性 (Div)[131]	平均计算生成姿态之间的特征距离	Diversity
外观一致性	前k个服装属性保留率 (AttrRec-k)[63]	使用预训练的服装属性识别模型预测服装属性, 前k召回率作为最终分数	Top-k clothing attribute retention rate
姿态一致性	正确关键点的百分比 (PCK)[132]	测量与真实关键点在指定阈值距离内被正确定位的关键点的比例	Percentage of correct keypoints
	平均关键点距离 (AKD)[133]	通过比较与真实关键点的距离来评估生成视频中人体关键点的准确性	Average keypoint distance
	缺失关键点率 (MKR)[82]	测量在检测或生成期间遗漏关键点的比例	Missing keypoint rate
时间一致性	帧一致性 (FC)[134]	通过计算连续帧的特征向量之间的余弦相似度来评估视频中的时间一致性	Frame consistency

表 4 姿态控制人物图像视频生成方法在 TikTok 数据集上的评测

方法	数据来源	Image					Video	
		FID↓	SSIM↑	PSNR↑ (dB)	LPIPS↓	L1↓	FID-VID↓	FVD↓
FOMM	[1]	85.03	0.648	29.01	0.335	3.61E–04	90.09	366.39
MRAA	[79]	54.47	0.672	18.14	0.296	3.21E–04	66.36	284.82
TPSMM	[78]	53.78	0.673	18.32	0.299	3.23E–04	72.55	306.17
DreamPose	[1]	79.46	0.509	13.19	0.450	6.91E–04	80.51	551.56
MagicPose	[34]	25.50	0.752	29.53	0.292	0.81E–04	46.30	—
DisCo	[1]	30.75	0.668	29.03	0.292	3.78E–04	59.90	292.80
MagicAnimate	[35]	32.09	0.714	29.16	0.239	3.13E–04	21.75	179.07
Animate Anyone	[36]	—	0.718	29.56	0.285	—	—	171.90
Animate Anyone 2	[40]	—	0.812	30.82	0.223	—	—	144.65

3.2.2 主观指标

在姿态控制人物图像视频生成任务的评估中, 除了客观的定量指标外, 主观指标也发挥着至关重要的作用. 主观指标主要通过用户实验来直接评判生成图像的视觉效果与自然性, 从而弥补客观评价指标的局限性. 具体而言, 研究团队通常采用用户评分方法衡量图像的整体质量、姿态准确性以及细节保留效果等关键维度, 并以平均分作为最终的质量评判标准. 然而, 这种调研方式也存在一些明显的不足, 例如较强的主观性以及较高的人力成本.

在实际操作中, 研究团队所采用的主观评测方法大致可以分为评级打分和偏好性比较两大类. 所选择的评测方法的依据多种多样, 涵盖了真实性、姿态转移的准确性、视频生成中的时间一致性、运动的真实性、平滑度以及重演的准确性等诸多方面.

4 应 用

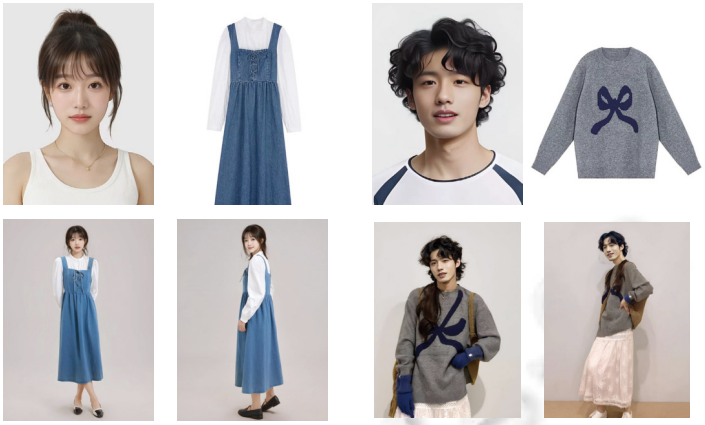
姿态控制人物图像视频生成技术的多样化应用场景展现了其广泛的实用性和商业价值. 在虚拟试穿、广告内容创作、视频生成、虚拟现实增强现实、与多模态技术结合等场景下, 相关技术不断提升生成结果的可控度、可行性和质量[135], 提升用户体验的同时也不断开辟着新的应用场景.

4.1 虚拟试穿与时尚行业

虚拟试穿是姿态控制人物图像视频生成技术的核心应用场景之一, 通过将目标服装映射到指定姿态下的虚拟人物图像中, 可以极大地提升用户的试衣体验, 同时相关电商也可以以更低的成本完成广告信息的制作.

从用户端出发, 用户对线下购物的需求有一部分体现在对试衣的需要, 但是通过姿态控制生成技术, 如 M2E-try on Net[57]等技术可以将服装从提供的模特图像中提取, 产生目标人物的试衣图像, 用户便可以在线上购物中获得试衣的体验. 为了满足用户对虚拟试衣的需要, 现有技术也拓展出更多的可控性、灵活性. 为了满足用户对服装个性化的需求, 生成模型也开始支持对服装组件的编辑. 一些方法[44–46]通过编辑特定的服装部件 (如改变衣领、袖口的样式), 进一步增强了虚拟试穿系统的灵活性. DiOr[6]提出循环的生成流水线, 可以实现穿衣顺序的可控性以及服装之间的不同交互, 比如先穿 T 恤再穿外套, 或者是将上衣的底部塞到裤子里之类的要求, 从而实现更真实的试穿体验. 而目前新兴的整合完整的虚拟穿衣技术也层出不穷, 如代表了最新技术前沿的 OOTDiffusion[136]模型和 OutfitAnyone[7]借助扩散模型, 在试衣中可以支持不同的身材, 支持不同衣物的组合搭配, 也实现了对高分辨率试穿效果的支持. 其中 OutfitAnyone 还可以结合 Animate Anyone[36]算法进行视频生成的下游任务. 从设计者出发, 一些方法[18,38]提供了从学习到的特征空间采样生成新样本的技术, Yang 等人[41]则提供了改变部件色彩的技术, 可以使效果呈现的成本降低, 从而辅助相关设计师进行新服装的设计工作. 从商户端出发, 虚拟试穿技术则可以帮助商家进行低成本的产品宣传图的制作, Fang 等人[23]提出在静态图像之间合成连续的图像序列来动态展示服装的方法, 而如绘蛙、linkfox 等平台可以只通过产品图像完成多姿态多视角模特试穿、展示视频生成等任务. 但同时由于目前技术对细节保留, 布料仿真不够真实、精确, 虚拟试衣难以准确描述真实产品的版型在不同姿态及不同

体型上出现的效果, 从而容易影响消费者的判断. 图 7 展示了虚拟试衣效果和一些失败案例. 同时, 目前在虚拟试衣上的场景也逐渐开始向三维人体扩展.



(a) 虚拟试衣应用 (b) 失败案例

图 7 虚拟试衣应用案例

4.2 视频生成与编辑

姿态控制生成技术在视频生成与编辑中的应用, 为影视、艺术、广告创作和娱乐等行业提供了前所未有的创新工具. 通过将姿态信息注入生成器中, 模型可以生成动态视频内容或对已有视频进行精细化编辑, 这不仅降低了内容制作成本, 还显著提升了创作的灵活性和效率.

在广告内容创作中, 由于常需展示服装、配饰等产品的动态效果, 一些技术提出可以从静态图像生成连续图像序列来展示服装[8], 这类技术能够显著降低传统拍摄和后期制作的复杂性, 使得创作者能够快速生成高质量的动态广告内容. 此外在娱乐、艺术等领域, MimicMotion 和 UniAnimate 等技术可以生成目标人物模仿参考动作的视频, 而 DisCo 和 BodyROI7 方法可以实现视频中背景和人物的可控, 使用户能够实现更高的视频生成灵活性. 同时, 与视频生成领域的问题一致, 目前应用的挑战集中在生成的实时性和时间一致性, 实现包含多角色交互、角色与环境互动的视频生成技术仍有待进一步发展.

4.3 多模态结合生成

姿态控制任务结合多模态进行生成, 通过将多种输入模态 (比如: 文本、姿态、视觉提示等) 整合为条件信息, 进一步提高了生成结果的多样性和实用性. UPGPT[137]提出一个接受文本、姿态和视觉提示的多模态扩散模型进行生成任务, 同时可以继续接受编辑操作. 这些方法通过增强条件的表达能力, 使生成器能够更好地捕捉目标姿态与外观的细节, 也提升了技术的易用性. HumanDreamer[118]则将视频生成分为文本到姿态 (text-to-pose) 和姿态到视频 (pose-to-video) 两阶段, 通过文本控制生成多样化姿态, 再基于姿态生成高质量视频, 解决了传统方法依赖预定义姿态的局限性.

5 挑战与展望

姿态控制人物图像视频生成技术的不断发展使其在虚拟试穿、影视、广告、艺术等领域应用中取得了显著进展. 然而, 在实际应用和研究过程中, 仍然存在许多尚未完全解决的技术挑战. 本节将从个性化信息的保留、复杂场景生成、生成即时性这 3 方面总结讨论关键的代表性挑战.

5.1 研究挑战

5.1.1 外观信息的保留

姿态控制生成任务要求生成给定参考图像人物外观下, 具有目标姿态的人物, 而现有技术在已经应用至商用

平台的情况下, 如图 7 所示, 虽然可以生成具有真实感的成果, 但也难以在一些规律的图案、细节上完全生成准确的纹理. 此外, 大多数技术方案难以保留人脸身份和身材维度信息等个性化信息, 如体型、四肢比例, 容易在面向用户的虚拟试衣应用中造成割裂感, 降低试穿效果的可信度. 在一些方法借助三维参数人体模型信息的情况下, 由于模型主要对人体表面进行建模而非服装, 生成结果倾向于产生类似贴身衣物的效果. 同时, 在人物穿着宽松衣物的情况下, 现有方法也难以产生自然、物理上合理的褶皱和光影.

5.1.2 复杂场景生成

一些方法通常通过 DeepFashion[107]等背景单一的数据集进行训练, 而专注于解决复杂背景不稳定问题的方法, 通过前后景的解耦以及背景的修复单独对背景进行了处理, 在静态的生成结果上产生了稳定的效果. 在视频生成下, 一些方法也通过跟踪来解决相机视角的移动以及背景的移动. 但它们是基于全局的移动或是关键点的粗略的运动, 而在存在多人运动的、需要人物与背景交互的或是背景存在复杂个体运动的场景却研究较少, 在一定程度上限制了相关技术在复杂场景中的应用.

5.1.3 模型的实时性能

随着虚拟试穿、游戏和视频生成等实际应用对实时性的要求提高, 现有模型的推理速度仍不足以满足商业需求. 如何在保证生成质量的前提下, 优化模型架构和计算流程 (如轻量化设计、知识蒸馏) 也是未来的关键挑战之一.

5.2 技术发展趋势

面对姿态控制生成技术的现有挑战, 未来的发展趋势可以从模型能力的提升、生成模型的扩展以及实时性和效率这 3 个方向展开. 技术的不断演进, 不仅能够解决当前的技术瓶颈, 还将进一步拓展其在虚拟试穿、影视娱乐、广告创作以及社交平台中的实际应用潜力.

5.2.1 模型能力的提升

逼真的姿态控制生成技术是未来可探索的关键方向之一. 目前, 技术在保留外观信息和布料质感方面存在局限, 特别是在视频生成领域, 对于人物服饰动态运动的真实感表现 (例如宽松衣物的褶皱和拖曳效果) 仍然不足. 此外, 在生成细节方面, 如服装配饰的独立编辑或肢体的精确控制 (例如手部的具体动作或肩部姿态的调整), 现有方法的可控性还有待提高. 通过结合深度信息或三维参数模型, 生成器可以更准确地捕捉复杂场景中的布料动态和精细部件控制. 例如, 利用 DensePose[26]等技术提取的三维人体特征, 有助于在生成过程中解决遮挡问题, 并生成更自然的姿态和动作. 同时, 三维参数模型的优化可能进一步增强人物与环境之间的交互建模能力, 这在虚拟试穿和影视制作中具有重要的应用前景.

在高分辨率生成方面, 尤其是在广告和影视制作领域, 需求日益增长. 一些方法通过分层生成器从低分辨率逐步优化细节, 以平衡生成质量和计算效率. 然而, 如何在更高分辨率 (如 8K) 下同时实现性能和细节表现, 仍是一个关键挑战. 此外, 减少数据需求对于实际应用至关重要. 无监督学习和数据增强技术可能通过减少对标注的依赖和扩展样本多样性来提升模型的泛化能力, 但这些方法在复杂场景中的实际效果仍需进一步研究和探索.

5.2.2 模型生成的实时性

实时性的优化是姿态生成技术向实际应用迈进的关键挑战之一. 在商业应用场景, 如虚拟试穿和游戏互动等领域, 对生成速度有着极高的要求. 目前, 许多生成模型在推理速度和显存需求方面存在瓶颈, 难以达到实时交互的标准. 为了解决这一问题, 可以采用知识蒸馏和模型剪枝等轻量化设计方法, 这些方法能够有效减少模型的计算负担, 从而提升推理速度和效率. 此外, 通过并行计算技术和分布式训练框架的应用, 可以显著提高多模态生成任务的处理效率, 这对于支持更复杂的实时交互场景至关重要.

5.2.3 多模态生成模型

多模态生成模型为姿态控制生成带来了更多的可能性. 通过引入语音、文本等多模态数据, 生成任务可以变得更加灵活. 例如, 用户通过语音或自然语言描述, 可能实现对人物动作和情感表达的细致控制. 这种多模态输入方式不仅能够提升生成内容的多样性, 还为虚拟社交平台和创意内容生成提供了新的交互方式. 在应用层面, 姿态

生成技术或可扩展至虚拟社交和个性化数字人生成领域. 例如, 结合表情控制和动态姿态, 虚拟社交平台中高度个性化的虚拟人物形象能够为用户提供更加沉浸式的体验. 此外, 在社交媒体内容创作中, 生成技术有望进一步降低内容创作者的技术门槛, 例如用于自动化生成个性化短视频或动态表情包.

6 总结和结论

姿态控制人物图像视频生成技术作为生成建模与计算机视觉领域的重要方向, 在虚拟试穿、广告创作、影视制作等领域展现了巨大的应用潜力. 本文介绍了完成相关任务的主要生成器与常见姿态数据的表示方法, 并以解决任务核心挑战为导向综述了相关技术的发展现状, 包括信息保留与重拍、推理不可见外观信息、一致性问题与模型的高效训练和使用, 整理了相关数据集与任务评价指标, 同时分析了个性化信息保留、复杂背景生成以及提升模型的实时性与效率等技术挑战并分析了其发展趋势.

References

[1] Wang T, Li LJ, Lin K, Zhai YH, Lin CC, Yang ZY, Zhang HW, Liu ZC, Wang LJ. DisCo: Disentangled control for realistic human dance generation. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2024. 9326–9336. [doi: 10.1109/CVPR52733.2024.00891]

[2] Yoon JS, Liu LJ, Golyanik V, Sarkar K, Park HS, Theobalt C. Pose-guided human animation from a single image in the wild. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 15034–15043. [doi: 10.1109/cvpr46437.2021.01479]

[3] Wang X, Zhang SW, Gao CX, Wang JY, Zhou XQ, Zhang YY, Yan LX, Sang N. UniAnimate: Taming unified video diffusion models for consistent human image animation. arXiv:2406.01188, 2024.

[4] Zhang Y, Gu JX, Wang LW, Wang H, Cheng JQ, Zhu YF, Zou FY. MimicMotion: High-quality human motion video generation with confidence-aware pose guidance. arXiv:2406.19680, 2025.

[5] Dong HY, Liang XD, Gong K, Lai HJ, Zhu J, Yin J. Soft-gated warping-GAN for pose-guided person image synthesis. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 472–482.

[6] Cui AY, McKee D, Lazebnik S. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 14618–14627. [doi: 10.1109/iccv48922.2021.01437]

[7] Sun K, Cao J, Wang Q, Tian LR, Zhang XD, Zhuo L, Zhang B, Bo LF, Zhou WB, Zhang WM, Gao DH. OutfitAnyone: Ultra-high quality virtual try-on for any clothing and any person. arXiv:2407.16224, 2024.

[8] Fang NY, Qiu LM, Zhang SY, Wang ZL, Hu KR, Dong LY. A novel human image sequence synthesis method by pose-shape-content inference. IEEE Trans. on Multimedia, 2023, 25: 6512–6524. [doi: 10.1109/TMM.2022.3209924]

[9] Shi CC, Chen YX, Lei BR, Chen JC. FashionPose: Text to pose to relight image generation for personalized fashion visualization. arXiv:2507.13311, 2025.

[10] Randhavane T, Bera A, Kapsaskis K, Gray K, Manocha D. FVA: Modeling perceived friendliness of virtual agents using movement characteristics. IEEE Trans. on Visualization and Computer Graphics, 2019, 25(11): 3135–3145. [doi: 10.1109/TVCG.2019.2932235]

[11] Xiang J, Guo YD, Hu LP, Guo BY, Yuan YC, Zhang JY. One shot, one talk: Whole-body talking avatar from a single image. arXiv:2412.01106, 2024.

[12] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. arXiv:1406.2661, 2014.

[13] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. arXiv:2006.11239, 2020.

[14] Isola P, Zhu JY, Zhou TH, Efros AA. Image-to-image translation with conditional adversarial networks. In: Proc. of the 2017 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 5967–5976. [doi: 10.1109/CVPR.2017.632]

[15] Wang TC, Liu MY, Zhu JY, Tao A, Kautz J, Catanzaro B. High-resolution image synthesis and semantic manipulation with conditional GANs. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018. 8798–8807. [doi: 10.1109/CVPR.2018.00917]

[16] Ma LQ, Jia X, Sun QR, Schiele B, Tuytelaars T, van Gool V. Pose guided person image generation. arXiv:1705.09368, 2018.

[17] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proc. of the 18th Int'l Conf. on Medical Image Computing and Computer-assisted Intervention (MICCAI). Munich: Springer, 2015. 234–241. [doi: 10.1007/978-3-

319-24574-4_28]

[18] Esser P, Sutter E. A variational U-Net for conditional appearance and shape generation. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8857–8866. [doi: 10.1109/CVPR.2018.00923]

[19] Bhunia AK, Khan S, Cholakkal H, Anwer RM, Laaksonen J, Shah M, Khan FS. Person image synthesis via denoising diffusion model. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023. 5968–5976. [doi: 10.1109/cvpr52729.2023.00578]

[20] Xu LH, Zhao L, Sun XX, Yan KD, Li GY. A comprehensive review of progress in deep-learning-based occluded human pose estimation. Journal of Image and Graphics, 2024, 29(12): 3529–3542 (in Chinese with English abstract). [doi: 10.11834/jig.230730]

[21] Toshev A, Szegedy C. DeepPose: Human pose estimation via deep neural networks. In: Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Columbus: IEEE, 2014. 1653–1660. [doi: 10.1109/cvpr.2014.214]

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

[23] Fang HS, Li JF, Tang HY, Xu C, Zhu HY, Xiu YL, Li YL, Lu CW. AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2023, 45(6): 7157–7173. [doi: 10.1109/tpami.2022.3222784]

[24] Yang HH, Liu HX, Zhang YM, Wu XJ. Parallel multi-scale spatio-temporal graph convolutional network for 3D human pose estimation. Ruan Jian Xue Bao/Journal of Software, 2025, 36(5): 2151–2166 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/7200.htm [doi: 10.13328/j.cnki.jos.007200]

[25] He JH, Sun JY, Liu Q. Multi-person 3D pose estimation using human-and-scene contexts. Ruan Jian Xue Bao/Journal of Software, 2024, 35(4): 2039–2054 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6837.htm [doi: 10.13328/j.cnki.jos.006837]

[26] Güler RA, Neverova N, Kokkinos I. DensePose: Dense human pose estimation in the wild. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7297–7306. [doi: 10.1109/CVPR.2018.00762]

[27] Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ. SMPL: A skinned multi-person linear model. Seminal Graphics Papers: Pushing the Boundaries, 2023, 2: 851–866. [doi: 10.1145/3596711.3596800]

[28] Romero J, Tzionas D, Black MJ. Embodied hands: Modeling and capturing hands and bodies together. ACM Trans. on Graphics, 2017, 36(6): 245. [doi: 10.1145/3130800.3130883]

[29] Pavlakos G, Choutas V, Ghorbani N, Bolkart T, Osman AA, Tzionas D, Black MJ. Expressive body capture: 3D hands, face, and body from a single image. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 10967–10977. [doi: 10.1109/cvpr.2019.01123]

[30] Karras J, Holynski A, Wang TC, Kemelmacher-Shlizerman I. DreamPose: Fashion image-to-video synthesis via stable diffusion. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Paris: IEEE, 2023. 22623–22633. [doi: 10.1109/iccv51070.2023.02073]

[31] Feng MY, Liu JL, Yu K, Yao Y, Hui Z, Guo XF, Lin XH, Xue HL, Shi C, Li XW, Li AJ, Kang XY, Lei BW, Cui MM, Ren PR, Xie XS. DreaMoving: A human video generation framework based on diffusion models. arXiv:2312.05107, 2023.

[32] Tu SY, Dai Q, Zhang ZH, Xie SC, Cheng ZQ, Luo C, Han XT, Wu ZX, Jiang YG. MotionFollower: Editing video motion via lightweight score-guided diffusion. arXiv:2405.20325, 2024.

[33] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. arXiv:2103.00020, 2021.

[34] Chang D, Shi YC, Gao QK, Fu J, Xu HY, Song GX, Yan Q, Zhu YZ, Yang X, Soleymani M. MagicPose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. arXiv:2311.12052, 2024.

[35] Xu ZC, Zhang JF, Liew JH, Yan HS, Liu JW, Zhang CX, Feng JS, Shou MZ. MagicAnimate: Temporally consistent human image animation using diffusion model. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2024. 1481–1490. [doi: 10.1109/CVPR52733.2024.00147]

[36] Li H. Animate Anyone: Consistent and controllable image-to-video synthesis for character animation. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2024. 8153–8163. [doi: 10.1109/cvpr52733.2024.00779]

[37] Liu JL, Yu K, Feng MY, Guo XF, Cui MM. Disentangling foreground and background motion for enhanced realism in human video generation. arXiv:2405.16393, 2024.

[38] Ma LQ, Sun QR, Georgoulis S, van Gool V, Schiele B, Fritz M. Disentangled person image generation. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 99–108. [doi: 10.1109/cvpr.2018.00018]

[39] Liu W, Piao ZX, Min J, Luo WH, Ma L, Gao SH. Liquid warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 5903–5912. [doi: 10.1109/ICCV.2019.00600]

[40] Li H, Wang GY, Shen Z, Gao X, Meng DC, Zhuo L, Zhang P, Zhang B, Bo LF. Animate Anyone 2: High-fidelity character image animation with environment affordance. arXiv:2502.06145, 2025.

[41] Yang LB, Zhao ZH, Wang SQ, Wang SS, Ma SW, Gao W. Disentangled human action video generation via decoupled learning. In: Proc. of the 2019 IEEE Int'l Conf. on Multimedia & Expo Workshops. Shanghai: IEEE, 2019. 495–500. [doi: 10.1109/icmew.2019.00091]

[42] Li NN, Shih KJ, Plummer BA. Collecting the puzzle pieces: Disentangled self-driven human pose transfer by permuting textures. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Paris: IEEE, 2023. 7092–7103. [doi: 10.1109/ICCV51070.2023.00655]

[43] Balakrishnan G, Zhao A, Dalca AV, Durand F, Guttag J. Synthesizing images of humans in unseen poses. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018. 8340–8348. [doi: 10.1109/cvpr.2018.00870]

[44] Men YF, Mao YM, Jiang YN, Ma WY, Lian ZH. Controllable person image synthesis with attribute-decomposed GAN. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 5083–5092. [doi: 10.1109/CVPR42600.2020.00513]

[45] Pu G, Men Y, Mao YM, Jiang YN, Ma WY, Lian ZH. Controllable image synthesis with attribute-decomposed GAN. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2023, 45(2): 1514–1532. [doi: 10.1109/tpami.2022.3161985]

[46] Zhang JS, Li K, Lai YK, Yang JY. PISE: Person image synthesis and editing with decoupled GAN. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 7978–7986. [doi: 10.1109/cvpr46437.2021.00789]

[47] Ren YR, Fan XQ, Li G, Liu S, Li TH. Neural texture extraction and distribution for controllable person image synthesis. In: Proc. of the 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 13525–13534. [doi: 10.1109/CVPR52688.2022.01317]

[48] Yang LB, Wang P, Liu C, Gao ZN, Ren PR, Zhang XF, Wang SS, Ma SW, Hua XS, Gao W. Towards fine-grained human pose transfer with detail replenishing network. IEEE Trans. on Image Processing, 2021, 30: 2422–2435. [doi: 10.1109/tip.2021.3052364]

[49] Chan C, Ginosar S, Zhou TH, Efros A. Everybody dance now. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 5932–5941. [doi: 10.1109/iccv.2019.00603]

[50] Huang SY, Xiong HY, Cheng ZQ, Wang QZ, Zhou XR, Wen BH, Huan J, Dou DJ. Generating person images with appearance-aware pose stylizer. In: Proc. of the 29th Int'l Joint Conf. on Artificial Intelligence (IJCAI). Yokohama: Morgan Kaufmann, 2021. 623–629.

[51] Grigorev A, Sevastopolsky A, Vakhitov A, Lempitsky V. Coordinate-based texture inpainting for pose-guided human image generation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 12127–12136. [doi: 10.1109/cvpr.2019.01241]

[52] Li YN, Huang C, Loy CC. Dense intrinsic appearance flow for human pose transfer. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 3688–3697. [doi: 10.1109/cvpr.2019.00381]

[53] Zanfir M, Popa AI, Zanfir A, Sminchisescu C. Human appearance transfer. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5391–5399. [doi: 10.1109/CVPR.2018.00565]

[54] Neverova N, Güler RA, Kokkinos I. Dense pose transfer. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 128–143. [doi: 10.1007/978-3-030-01219-9_8]

[55] Zhang P, Zhang B, Chen D, Yuan L, Wen F. Cross-domain correspondence learning for exemplar-based image translation. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 5142–5152. [doi: 10.1109/CVPR42600.2020.00519]

[56] Zhou XR, Zhang B, Zhang T, Zhang P, Bao JM, Chen D, Zhang ZF, Wen F. CoCosNet v2: Full-resolution correspondence learning for image translation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 11460–11470. [doi: 10.1109/CVPR46437.2021.01130]

[57] Wu ZH, Lin GS, Tao QY, Cai JF. M2E-try on Net: Fashion from model to everyone. In: Proc. of the 27th ACM Int'l Conf. on Multimedia. Nice: ACM, 2019. 293–301. [doi: 10.1145/3343031.3351083]

[58] Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial Transformer networks. arXiv:1506.02025, 2016.

[59] Siarohin A, Sangineto E, Lathuilière S, Sebe N. Deformable GANs for pose-based human image generation. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 3408–3416. [doi: 10.1109/cvpr.2018.

00359]

[60] Zheng HT, Chen LL, Xu CL, Luo JB. Unsupervised pose flow learning for pose guided synthesis. arXiv:1909.13819, 2019.

[61] Tabejamaat M, Negin F, Bremond FF. Guided flow field estimation by generating independent patches. 2021. https://www.bmvc2021-virtualconference.com/assets/papers/1038.pdf

[62] Zablotskaia P, Siarohin A, Zhao B, Sigal L. DwNet: Dense warp-based network for pose-guided human video generation. arXiv: 1910.09139, 2019.

[63] Han XT, Huang WL, Hu XJ, Scott M. ClothFlow: A flow-based model for clothed person generation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 10470–10479. [doi: 10.1109/ICCV.2019.01057]

[64] Tang H, Bai S, Torr PHS, Sebe N. Bipartite graph reasoning GANs for person image generation. arXiv:2008.04381, 2020.

[65] Tang H, Bai S, Zhang L, Torr PHS, Sebe N. XingGAN for person image generation. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 717–734. [doi: 10.1007/978-3-030-58595-2_43]

[66] Liu W, Piao Z, Tu Z, Luo WH, Ma L, Gao SH. Liquid warping GAN with attention: A unified framework for human image synthesis. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(9): 5115–5133. [doi: 10.1109/TPAMI.2021.3078270]

[67] Ren YR, Yu XM, Chen JM, Li TH, Li G. Deep image spatial transformation for person image generation. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 7687–7696. [doi: 10.1109/CVPR42600.2020.00771]

[68] Zhu Z, Huang TT, Shi BG, Yu M, Wang BF, Bai X. Progressive pose attention transfer for person image generation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 2342–2351. [doi: 10.1109/CVPR.2019.00245]

[69] Zhu Z, Huang TT, Xu MD, Shi BG, Cheng WQ, Bai X. Progressive and aligned pose attention transfer for person image generation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(8): 4306–4320. [doi: 10.1109/TPAMI.2021.3068236]

[70] Albahar B, Lu JW, Yang JM, Shu ZX, Shechtman E, Huang JB. Pose with style: Detail-preserving pose-guided image synthesis with conditional StyleGAN. ACM Trans. on Graphics, 2021, 40(6): 218. [doi: 10.1145/3478513.3480559]

[71] Deng JK, Cheng SY, Xue NN, Zhou YX, Zafeiriou S. UV-GAN: Adversarial facial UV map completion for pose-invariant face recognition. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7093–7102. [doi: 10.1109/cvpr.2018.00741]

[72] Sarkar K, Mehta D, Xu WP, Golyanik V, Theobalt C. Neural re-rendering of humans from a single image. In: Proc. of the 16th European Conf. on Computer Vision (ECCV). Glasgow: Springer, 2020. 596–613. [doi: 10.1007/978-3-030-58621-8_35]

[73] Sarkar K, Golyanik V, Liu LJ, Theobalt C. Style and pose control for image synthesis of humans from a single monocular view. arXiv:2102.11263, 2021.

[74] Huang ZY, Li HH, Xie ZY, Kampffmeyer M, Cai QL, Liang XD. Towards hard-pose virtual try-on via 3D-aware global correspondence learning. arXiv:2211.14052, 2022.

[75] Booth J, Antonakos E, Ploumpis S, Trigeorgis G, Panagakis Y, Zafeiriou S. 3D face morphable models "in-the-wild". In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 5464–5473. [doi: 10.1109/cvpr.2017.580]

[76] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2021, 43(12): 4217–4228. [doi: 10.1109/TPAMI.2020.2970919]

[77] Lathuilière S, Sangineto E, Siarohin A, Sebe N. Attention-based fusion for multi-source human image generation. In: Proc. of the 2020 IEEE Winter Conf. on Applications of Computer Vision (WACV). Snowmass: IEEE, 2020. 428–437. [doi: 10.1109/WACV45572.2020.9093602]

[78] Zhao J, Zhang H. Thin-plate spline motion model for image animation. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 3647–3656. [doi: 10.1109/cvpr52688.2022.00364]

[79] Siarohin A, Woodford OJ, Ren J, Chai ML, Tulyakov S. Motion representations for articulated animation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 13648–13657. [doi: 10.1109/cvpr46437.2021.01344]

[80] Xue JY, Wang HF, Tian Q, Ma Y, Wang AD, Zhao ZY, Min SB, Zhao WZ, Zhang KH, Shum HY, Liu W, Liu MY, Luo WH. Follow-your-pose v2: Multiple-condition guided character image animation for stable pose control. arXiv:2406.03035, 2024.

[81] Siarohin A, Lathuilière S, Tulyakov S, Ricci E, Sebe N. Animating arbitrary objects via deep motion transfer. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 2372–2381. [doi: 10.1109/CVPR.2019.00248]

[82] Siarohin A, Lathuilière S, Tulyakov S, Ricci E, Sebe N. First order motion model for image animation. arXiv:2003.00196, 2020.

[83] Wang BC, Zheng HB, Liang XD, Chen YM, Lin L, Yang M. Toward characteristic-preserving image-based virtual try-on network. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 607–623. [doi: 10.1007/978-3-030-01261-8_36]

[84] Wang QL, Jiang ZK, Xu CM, Zhang JN, Wang YB, Zhang XY, Cao Y, Cao WJ, Wang CJ, Fu YW. VividPose: Advancing stable video diffusion for realistic human image animation. arXiv:2405.18156, 2024.

[85] Wang TC, Liu MY, Tao A, Liu GL, Kautz J, Catanzaro B. Few-shot video-to-video synthesis. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 5013–5024.

[86] Yang ZD, Zeng AL, Yuan C, Li Y. Effective whole-body pose estimation with two-stages distillation. arXiv:2307.15880, 2023.

[87] Wang TC, Liu MY, Zhu JY, Liu GL, Tao A, Kautz J, Catanzaro B. Video-to-video synthesis. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 1152–1164.

[88] Shao RZ, Pang YX, Zheng ZR, Sun JX, Liu YB. Human4DiT: 360-degree human video generation with 4D diffusion Transformer. arXiv:2405.17405, 2024.

[89] Zhu SH, Chen JL, Dai ZZ, Dong ZL, Xu YH, Cao X, Yao Y, Zhu H, Zhu SY. CHAMP: Controllable and consistent human image animation with 3D parametric guidance. In: Proc. of the 18th European Conf. on Computer Vision. Milan: Springer, 2025. 145–162. [doi: 10.1007/978-3-031-73001-6_9]

[90] Zhang LM, Rao AY, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Paris: IEEE, 2023. 3813–3824. [doi: 10.1109/iccv51070.2023.00355]

[91] Yan YC, Xu JW, Ni BB, Zhang WD, Yang XK. Skeleton-aided articulated motion generation. In: Proc. of the 25th ACM Int'l Conf. on Multimedia. Mountain View: ACM, 2017. 199–207. [doi: 10.1145/3123266.3123277]

[92] Guo YW, Yang CY, Rao AY, Liang ZY, Wang YH, Qiao Y, Agrawala M, Lin DH, Dai B. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv:2307.04725, 2024.

[93] Gan QJ, Ren Y, Zhang C, Ye ZH, Xie P, Yin X, Yuan ZH, Peng BY, Zhu JK. HumanDiT: Pose-guided diffusion Transformer for long-form human motion video generation. arXiv:2502.04847, 2025.

[94] Pumarola A, Agudo A, Sanfeliu A, Moreno-Noguer F. Unsupervised person image synthesis in arbitrary poses. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8620–8628. [doi: 10.1109/cvpr.2018.00899]

[95] Tang H, Xu D, Liu GW, Wang W, Sebe N, Yan Y. Cycle in cycle generative adversarial networks for keypoint-guided image generation. In: Proc. of the 27th ACM Int'l Conf. on Multimedia. Nice: ACM, 2019. 2052–2060. [doi: 10.1145/3343031.3350980]

[96] Song SJ, Zhang W, Liu JY, Mei T. Unsupervised person image generation with semantic parsing transformation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 2352–2361. [doi: 10.1109/CVPR.2019.00246]

[97] Ma Y, He YQ, Cun XD, Wang XT, Chen SR, Li X, Chen QF. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: Proc. of the 38th AAAI Conf. on Artificial Intelligence. Vancouver: AAAI, 2024. 4117–4125. [doi: 10.1609/aaai.v38i5.28206]

[98] Zhang PZ, Yang LX, Lai JH, Xie XH. Exploring dual-task correlation for pose guided person image generation. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 7703–7712. [doi: 10.1109/CVPR52688.2022.00756]

[99] Shen F, Ye H, Zhang J, Wang C, Han X, Yang W. Advancing pose-guided image synthesis with progressive conditional diffusion models. arXiv:2310.06313, 2024.

[100] Roy P, Bhattacharya S, Ghosh S, Pal U. Multi-scale attention guided pose transfer. Pattern Recognition, 2023, 137: 109315. [doi: 10.1016/j.patcog.2023.109315]

[101] Yu WY, Po LM, Zhao YZ, Xiong JJ, Lau KW. Spatial content alignment for pose transfer. In: Proc. of the 2021 IEEE Int'l Conf. on Multimedia and Expo (ICME). Shenzhen: IEEE, 2021. 1–6. [doi: 10.1109/ICME51207.2021.9428146]

[102] Lu YZ, Zhang ML, Ma AJ, Xie XH, Lai JH. Coarse-to-fine latent diffusion for pose-guided person image synthesis. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2024. 6420–6429. [doi: 10.1109/CVPR52733.2024.00614]

[103] Barnes C, Shechtman E, Finkelstein A, Goldman DB. PatchMatch: A randomized correspondence algorithm for structural image editing. ACM Trans. on Graphics, 2009, 28(3): 24. [doi: 10.1145/1531326.1531330]

[104] Wang X, Zhang SW, Tang LX, Zhang YY, Gao CX, Wang YH, Sang N. UniAnimate-DiT: Human image animation with large-scale video diffusion Transformer. arXiv:2504.11289, 2025.

[105] Fu XM, Wang X, Liu J, Wang SH, Dai J, Han JZ. CoP: Chain-of-pose for image animation in large pose changes. In: Proc. of the 31st ACM Int'l Conf. on Multimedia. Ottawa: ACM, 2023. 6969–6977. [doi: 10.1145/3581783.3611772]

[106] Zheng L, Shen LY, Tian L, Wang SJ, Wang JD, Tian Q. Scalable person re-identification: A benchmark. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision (ICCV). Santiago: IEEE, 2015. 1116–1124. [doi: 10.1109/iccv.2015.133]

[107] Liu ZW, Luo P, Qiu S, Wang XG, Tang XO. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 1096–1104. [doi: 10.1109/CVPR.2016.124]

[108] Liu KH, Chen TY, Chen CS. MVC: A dataset for view-invariant clothing retrieval and attribute prediction. In: Proc. of the 2016 ACM on Int' l Conf. on Multimedia Retrieval. New York: ACM, 2016. 313–316. [doi: 10.1145/2911996.2912058]

[109] Yoo D, Kim N, Park S, Paek AS, Kweon IS. Pixel-level domain transfer. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 517–532. [doi: 10.1007/978-3-319-46484-8_31]

[110] Dong HY, Liang XD, Shen XH, Wang BC, Lai HJ, Zhu J, Hu ZT, Yin J. Towards multi-pose guided virtual try-on network. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 9025–9034. [doi: 10.1109/ICCV.2019.00912]

[111] Hsieh CW, Chen CY, Chou CL, Shuai HH, Liu JY, Cheng WH. FashionOn: Semantic-guided image-based virtual try-on with detailed human and clothing information. In: Proc. of the 27th ACM Int'l Conf. on Multimedia. Nice: ACM, 2019. 275–283. [doi: 10.1145/3343031.3351075]

[112] Zheng N, Song XM, Chen ZZ, Hu LM, Cao D, Nie LQ. Virtually trying on new clothing with arbitrary poses. In: Proc. of the 27th ACM Int'l Conf. on Multimedia. Nice: ACM, 2019. 266–274. [doi: 10.1145/3343031.3350946]

[113] Jiang YM, Yang S, Qiu HN, Wu W, Loy CC, Liu ZW. Text2Human: Text-driven controllable human image generation. ACM Trans. on Graphics, 2022, 41(4): 162. [doi: 10.1145/3528223.3530104]

[114] Fu JL, Li SK, Jiang YM, Lin KY, Qian C, Loy CC, Wu W, Liu ZW. StyleGAN-human: A data-centric odyssey of human generation. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 1–19. [doi: 10.1007/978-3-031-19787-1_1]

[115] Soomro K, Zamir AR, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402, 2012.

[116] Zhang WY, Zhu ML, Derpanis KG. From actemes to action: A strongly-supervised representation for detailed action understanding. In: Proc. of the 2013 IEEE Int'l Conf. on Computer Vision. Sydney: IEEE, 2013. 2248–2255. [doi: 10.1109/ICCV.2013.280]

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

[118] Wang BY, Wang XF, Ni CJ, Zhao GS, Yang ZQ, Zhu Z, Zhang MY, Zhou YK, Chen XZ, Huang G, Liu LH, Wang XG. HumanDreamer: Generating controllable human-motion videos via decoupled generation. arXiv:2503.24026, 2025.

[119] Dong HY, Liang XD, Shen XH, Wu BW, Chen BC, Yin J. FW-GAN: Flow-navigated warping GAN for video virtual try-on. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 1161–1170. [doi: 10.1109/iccv.2019.00125]

[120] Jafarian Y, Park HS. Self-supervised 3D representation learning of dressed humans from social media videos. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2023, 45(7): 8969–8983. [doi: 10.1109/TPAMI.2022.3231558]

[121] Bishop CM. Pattern Recognition and Machine Learning. New York: Springer, 2006. 47.

[122] Horé A, Ziou D. Image quality metrics: PSNR vs. SSIM. In: Proc. of the 20th Int'l Conf. on Pattern Recognition. Istanbul: IEEE, 2010. 2366–2369. [doi: 10.1109/ICPR.2010.579]

[123] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Trans. on Image Processing, 2004, 13(4): 600–612. [doi: 10.1109/TIP.2003.819861]

[124] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 586–595. [doi: 10.1109/CVPR.2018.00068]

[125] Rabin J, Peyré G, Delon J, Bernot M. Wasserstein barycenter and its application to texture mixing. In: Proc. of the 3rd Int'l Conf. on Scale Space and Variational Methods in Computer Vision. Ein-Gedi: Springer, 2011. 435–446. [doi: 10.1007/978-3-642-24785-9_37]

[126] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6629–6640.

[127] Tulyakov S, Liu MY, Yang XD, Kautz J. MoCoGAN: Decomposing motion and content for video generation. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1526–1535. [doi: 10.1109/CVPR.2018.00165]

[128] Liu YF, Cun XD, Liu XB, Wang XT, Zhang Y, Chen HX, Liu Y, Zeng TY, Chan R, Shan Y. EvalCrafter: Benchmarking and evaluating large video generation models. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2024. 22139–22149. [doi: 10.1109/CVPR52733.2024.02090]

[129] Yoon Y, Cha B, Lee JH, Jang M, Lee J, Kim J, Lee G. Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Trans. on Graphics, 2020, 39(6): 222. [doi: 10.1145/3414685.3417838]

[130] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 2234–2242.

[131] Liu X, Wu QY, Zhou H, Xu YH, Qian R, Lin XY, Zhou XW, Wu W, Dai B, Zhou BL. Learning hierarchical cross-modal association for co-speech gesture generation. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 10452–10462. [doi: 10.1109/CVPR52688.2022.01021]

[132] Wu HN, Zhang EL, Liao L, Chen CF, Hou JW, Wang AN, Sun WX, Yan Q, Lin WS. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Paris: IEEE, 2023. 20087–20097. [doi: 10.1109/ICCV51070.2023.01843]

[133] Yang Y, Ramanan D. Articulated human detection with flexible mixtures of parts. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2878–2890. [doi: 10.1109/TPAMI.2012.261]

[134] Esser P, Chiu J, Atighehchian P, Granskog J, Germanidis A. Structure and content-guided video synthesis with diffusion models. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Paris: IEEE, 2023. 7312–7322. [doi: 10.1109/ICCV51070.2023.00675]

[135] Chen HH, Tao YF, Zhang JY. A survey on 3D clothed human body reconstruction: From traditional methods to high-fidelity models. Journal of Image and Graphics, 2024, 29(9): 2566–2595 (in Chinese with English abstract). [doi: 10.11834/jig.230646]

[136] Xu YH, Gu T, Chen WF, Chen CC. OOTDiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. arXiv:2403.01779, 2024.

[137] Cheong SY, Mustafa A, Gilbert A. UPGPT: Universal diffusion model for person image generation, editing and pose transfer. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision Workshops (ICCVW). Paris: IEEE, 2023. 4175–4184. [doi: 10.1109/ICCVW60793.2023.00451]

附中文参考文献

[20] 徐琳皓, 赵林, 孙辛欣, 颜克冬, 李广宇. 基于深度学习的遮挡人体姿态估计进展综述. 中国图象图形学报, 2024, 29(12): 3529–3542. [doi: 10.11834/jig.230730]

[24] 杨红红, 刘泓希, 张玉梅, 吴晓军. 基于平行多尺度时空图卷积网络的三维人体姿态估计算法. 软件学报, 2025, 36(5): 2151–2166. http://www.jos.org.cn/1000-9825/7200.htm [doi: 10.13328/j.cnki.jos.007200]

[25] 何建航, 孙郡瑶, 刘琼. 基于人体和场景上下文的多人 3D 姿态估计. 软件学报, 2024, 35(4): 2039–2054. http://www.jos.org.cn/1000-9825/6837.htm [doi: 10.13328/j.cnki.jos.006837]

[135] 陈鸿鹄, 陶云帆, 张举勇. 三维穿衣人体重建综述——从传统方法到高保真模型. 中国图象图形学报, 2024, 29(9): 2566–2595. [doi: 10.11834/jig.230646]

作者简介

李玘芮, 硕士生, CCF 学生会员, 主要研究领域为人工智能生成, 推理加速.

励雪巍, 博士, 特聘副研究员, CCF 专业会员, 主要研究领域为计算机视觉, 人工智能, 如知识蒸馏, 场景图生成, 扩散模型.

赵奇, 硕士生, 主要研究领域为人工智能生成, 推理加速.

李杰, 硕士生, 主要研究领域为人工智能生成, 推理加速.

李玺, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为人工智能, 计算机视觉, 机器学习, 模式识别, 数据挖掘.