E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP18;TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2026)04-1256-16

论文引用格式:Zhang Y H, Liu L, Fu X D, Liu L J and Peng W. 2026. Clothed human generation via pose diffusion priors and multiview consistency. Journal of Image and Graphics, 31(4):1256-1271(张渊杭, 刘骊, 付晓东, 刘利军, 彭玮. 2026. 融合姿态扩散先验与多视图一致性的着装人体生成. 中国图象图形学报, 31(4):1256-1271)[DOI:10.11834/jig.250367]

融合姿态扩散先验与多视图一致性的着装人体生成

张渊杭[1],刘骊[1,2*],付晓东[1,2],刘利军[1,2],彭玮[1,2]

1. 昆明理工大学信息工程与自动化学院,昆明　650500; 2. 云南省计算机技术应用重点实验室,昆明　650500

摘　要:**目的**　针对单视图着装人体生成中不可见区域纹理缺失、局部细节模糊以及宽松服装几何生成困难等关键问题,提出一种融合姿态扩散先验与多视图一致性的生成方法。**方法**　首先,采用人体姿态估计算法提取25个关键点并将其编码为高斯热图,结合人体掩码与UV映射构建姿态特征向量,指导潜在扩散模型生成不可见视角的二维扩散图像;其次,将SMPLX(skinned multi-person linear model expressive)模板的法线信息与输入图像和生成的扩散图像进行对应视角的特征融合,并输入跨视角法线一致性网络,通过多视图一致性约束机制提取跨视角的三维空间特征;最后,融合SMPLX人体模板的体素化特征,输入分布预测网络进行空间占用概率估计,并在学习的概率分布中采样,将三维特征、体素化特征与采样结果输入占用预测网络,实现三维着装人体生成。**结果**　在THuman2.0(Tsinghua human 2.0 dataset)与CAPE(clothed auto-person encoding)公开基准数据集上的定量评估表明,所提方法的倒角距离(chamfer distance)和点到面距离(point-to-surface distance)在THuman2.0数据集上较最优对比方法分别降低6.27%和5.74%,在CAPE数据集上平均降低8.67%和2.38%。**结论**　本文提出的融合姿态扩散先验与多视图一致性的单视图三维着装人体生成方法,能够有效恢复局部纹理,并准确生成褶皱细节丰富和宽松服装等复杂拓扑结构的着装人体模型。

关键词:单视图着装人体生成;姿态扩散先验;多视图一致性约束;分布预测网络;概率分布

Clothed human generation via pose diffusion priors and multiview consistency

Zhang Yuanhang[1], Liu Li[1,2*], Fu Xiaodong[1,2], Liu Lijun[1,2], Peng Wei[1,2]

1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming　650500, China;

2. Computer Technology Application Key Laboratory of Yunnan Province, Kunming　650500, China

Abstract: **Objective** Clothed human generation, which aims to recover the 3D geometry and texture of the human body from input data to generate accurate 3D human models, is a challenging problem in the fields of computer vision and computer graphics. The need for high-quality generations has become increasingly critical with the growing demand for realistic 3D human models in applications such as virtual reality and augmented reality. Traditional multiview generation methods, which are often expensive and impractical for everyday use, typically require specialized equipment to capture images from

收稿日期:2025-08-18;**修回日期**:2025-10-08;**预印本日期**:2025-10-15

*通信作者:刘骊　ieall@kust.edu.cn

基金项目:国家自然科学基金项目(62262036, 62362043);兴滇英才支持计划项目(KKXY202203008);云南省科技计划项目(202503AA080013, 202502AD080003)

Supported by: National Natural Science Foundation of China(62262036, 62362043); Xingdian Talent Support Project (KKXY202203008); Yunnan Provincial Science and Technology Support Project(202503AA080013, 202502AD080003)

multiple viewpoints. By contrast, obtaining single-view images from the web is much easier than obtaining multiview images. Thus, single-view generation methods become more cost-effective than multiview generation methods, and the model creation process becomes simple. Given these advantages, we consider using a single view as input to recover the 3D model of a clothed human. However, single-view images lack comprehensive spatial information and structural details of occluded regions. Thus, recovering a complete 3D shape becomes difficult. As a result, existing methods based on implicit functions struggle to learn rear-view information effectively, thereby leading to overly smooth and unrealistic back regions in the generated 3D human model. Methods combining diffusion models show some potential in enhancing texture detail performance. However, most of these methods lack view consistency constraints, thereby making the full recovery of the local texture details of the human body difficult. Additionally, the absence of precise geometric constraints during the diffusion process causes discrepancies between the generated models and the true geometry, particularly when handling complex 3D structures. Existing methods typically assume a uniform point distribution across spatial regions by ignoring variations in the distribution of query points caused by differences in distance from the human body surface. This assumption makes adapting to the geometric complexity differences across various regions of the body difficult for these methods. As a result, these methods face limitations when generating the surfaces of loose clothing, which have complex and variable geometries. This study addresses these challenges by combining three mechanisms: pose diffusion priors generation, multiview consistency constraints, and adaptive geometry generation. This approach not only preserves the generative capabilities of the diffusion model but also introduces geometric constraints to ensure the accuracy of the generation. Furthermore, this method can generate high-quality 3D human models by incorporating the probability distribution of human body structure. This study proposes a generation method that integrates pose diffusion priors with multiview consistency. **Method** This study constructs a method for single-view clothed human generation. First, a human pose estimation algorithm is used to extract 25 key points, which are encoded into Gaussian heatmaps to achieve spatial continuity modeling. This approach enables the model to understand the spatial relationships around the key points. The Gaussian heatmaps, combined with the human mask and UV mapping, are used to construct a pose feature vector. This feature vector guides the denoising process of the latent diffusion model and generates 2D diffusion images for unseen viewpoints through an adaptive cross-attention mechanism. Second, after the normal information of the (skinned multi-person linear model expressive, SMPLX) human template estimated from the input image and the 2D diffusion image are fused, they are input into the cross-view normal consistency network, where the multiview consistency mechanism extracts the corresponding 3D spatial features for each viewpoint. Finally, the voxelized features of the SMPLX human template and the 3D spatial features are fused and input into the distribution prediction network for spatial occupancy probability estimation. The model can express geometric uncertainty at different spatial locations and sample from the learned probability distribution by learning the distribution parameters of each point. Then, the 3D features, voxelized features, and sampling results are input into the occupancy prediction network to achieve 3D clothed human generation. Our entire model is trained on the THuman2.0 (Tsinghua human 2.0 dataset) dataset, with 490 images being used for training and 21 images being used for testing. We tested the model on the CAPE (clothed auto-person encoding) dataset to evaluate the generalization ability of the model further. This dataset is divided into two subsets: CAPE fitted poses (CAPE-FP), which contains 75 images used to assess the geometric generation accuracy of the method under simple poses, and CAPE nonfitted poses (CAPE-NFP), which contains 75 images and focuses on evaluating the method's adaptability to complex poses. The experiments are conducted on an NVIDIA GeForce RTX 3090 GPU, with a learning rate being set to 1×10^{-4} and a batch size of 2. **Result** We conducted experiments on the THuman2.0 and CAPE datasets and compared the single-view clothed human generation results with the results of six other methods. Chamfer distance (CD) is used to evaluate the overall geometric similarity of the 3D human body, and point-to-surface distance (P2S) is used to assess the geometric accuracy of the reconstructed surface. Both metrics perform well when their values are small. On the THuman2.0 dataset, the CD and P2S metrics of the single-view clothed human generation method were reduced by 6.27% and 5.74%, respectively, compared with those of the best-performing method. On the CAPE-FP and CAPE-NFP subsets, the CD and P2S of the single-view clothed human generation method performed better than those of the other comparison methods. On the entire CAPE dataset, the CD metric of the single-view clothed human generation method decreased by an average of 8.67%, and the P2S metric decreased by an average of 2.38%. Quantitative

experiments show that our method has good generalization ability for unseen data and can effectively handle human generation tasks in complex poses. Inference efficiency comparison results show that the computational complexity of our method is lower than that of similar diffusion model methods. Experimental results indicate that combining pose diffusion priors and multiview consistency helps recover the texture details of the 3D human body, and adaptive geometry generation enables accurate recovery of complex clothing topologies. **Conclusion** The single-view 3D clothed human generation method proposed in this paper, which combines pose diffusion priors and multiview consistency, effectively recovers the local details of the clothed human and accurately generates 3D human models with complex topological structures, such as rich wrinkle details and loose clothing.

Key words: single-view clothed human generation; pose diffusion priors; multiview consistency constraints; distribution prediction network; probability distribution

0 引　言

着装人体生成(Saito等,2019)旨在从二维图像恢复完整的三维人体几何与纹理,在虚拟现实、数字人等领域有重要应用。当前方法主要分为三大类:1)基于隐式函数的方法通过学习三维坐标到占用值的映射关系,能够高效表达复杂的人体几何结构(Lim和Lee,2024);2)基于神经辐射场(neural radiance fields,NeRF)的方法通过体渲染技术生成基于多视角图像的三维人体,能够获得高质量的光照与纹理细节(Hu等,2023);3)基于高斯泼溅的方法采用高斯椭球集合实现三维人体的快速渲染(Abdal等,2024)。基于NeRF和高斯泼溅的方法依赖多个视角的图像来恢复完整的三维信息,当视角信息有限时,恢复精度会受到显著影响。相比之下,隐式函数通过单一视角推断三维人体模型,表现出较强的表达能力和较高的几何精度。

单视图着装人体生成因其输入便捷性而更具实用性,但缺乏完整的空间信息、深度数据以及遮挡区域的结构细节,使得准确生成三维着装人体成为挑战。黄千芃等人(2024)结合姿态特征学习与服装柔性变形,提升在肢体遮挡情况下的姿态估计准确性。Zhang等人(2023)采用混合先验融合策略增强三维特征与人体先验的融合。然而,这些方法未考虑人体背面信息生成,导致背面区域过于平滑。

视图一致性约束通过确保同一3D点在不同视角下的投影表示保持语义和几何一致性,Xiu等人(2022)提出具有代表性的单视图人体生成方法,基于SMPL(skinned multi-person linear model)模板的法线渲染预测着装法线映射,通过局部特征回归进行隐式表面表示。但这类方法采用的分离式法线贴图预测模式存在根本性缺陷:通过两个独立的网络分支分别预测可见与不可见视角的纹理信息,各分支间缺乏有效的跨视角协同学习机制和统一的多视图监督约束,仅在网络输出阶段进行特征拼接,而非在特征学习过程中建立视角间的一致性关联,导致不同视角的特征表示缺乏内在连贯性,局部细节模糊。近期,Ho等人(2024)通过扩散模型生成不可见的后视角图像,结合SMPLX(skinned multi-person linear model expressive)人体模板作为几何指导,从输入图像和生成的后视角图像中恢复完整纹理网格。然而,上述方法在训练数据有限的条件下普遍依赖SMPLX进行人体生成,且缺乏有效的视图一致性约束,特别是在处理宽松服装时出现边缘破碎等伪影,限制了其泛化能力。

综上所述,单视图三维人体生成面临以下难点:1)单视角图像输入无法提供着装人体的背面细节,降低整体生成的质量;2)现有方法缺乏多视角法线一致性约束,导致生成视图间的几何不一致性;3)基于SMPLX人体模板的方法难以学习远离人体表面的服装结构,在恢复宽松服装表面时适应性差。为此,本文提出融合姿态扩散先验与多视图一致性的着装人体生成方法,构建姿态扩散先验生成、多视图一致性约束与自适应几何生成3个协同机制。先验生成为一致性约束提供数据基础,一致性约束为几何生成提供特征指导,几何生成结果反馈优化整体框架,以确保生成结果的准确性,并提升对复杂服装拓扑的适应能力。

本文主要工作如下:1)构建姿态扩散先验生成模块,通过人体姿态引导高质量的后视角图像生成,为后续的多视图一致性约束模块提供必要的多视角纹理数据,解决单视图输入信息缺失的问题;2)建立多视图一致性约束机制,通过法线方向引导和几何

关系对应，确保扩散生成的多视角纹理在空间上连贯一致，提升生成视图与输入视图之间的空间一致性；3）设计自适应几何生成模块，通过学习空间变化的概率分布并进行自适应采样，适应远离人体表面的复杂服装结构，提高三维着装人体生成的质量。

1 相关工作

1.1 单视图人体生成

单视图人体生成技术旨在从单幅图像中恢复三维人体结构，隐式表示作为一种连续的函数表征能够灵活处理三维拓扑结构变化。Saito 等人（2019）首次提出像素对齐隐式函数，利用多层感知器（multi-layer perceptron，MLP）建模从像素特征到占用值的映射关系。Saito 等人（2020）提出的 PIFuHD（multi-level pixel-aligned implicit function for high-resolution 3D human digitization）引入从粗到精的网络，集成低分辨率和高分辨率图像特征以及预测的正面和背面法线图像。He 等人（2020）通过结合几何对齐的 3D 特征和像素对齐的 2D 特征，增强网格生成的细节和形状一致性。Chen 等人（2023）结合 Transformer 和像素对齐隐式函数，通过 Transformer 探索 3D 特征块序列间的几何关联。针对隐式函数产生的肢体断裂问题，Lim 和 Lee（2024）提出三向隐式函数（tri-directional implicit function for high-fidelity 3D character reconstruction，TIFu），结合像素级特征学习和体素表示的优点。Chan 等人（2022）将相对深度图和人体解析图信息集成到像素对齐隐式模型中，改善深度歧义问题并关注几何细节。

为进一步提升隐式表示的几何生成精度，研究者在采样策略与特征融合方面提出多项改进。Yang 等人（2024）提出高低频分解范式，通过渐进式高频符号距离场（signed distance field，SDF）学习详细几何形状。Chan 等人（2024a）提出精细结构感知采样训练方案，通过自适应调整采样点位置与标签提升模型对细节的感知能力。

针对复杂姿态变化与自遮挡问题，研究者引入 SMPL/SMPLX 人体模板作为三维先验。Zheng 等人（2022）从 SMPL 模型中提取三维体素对齐特征和语义信息。Cao 等人（2023）提出自进化符号距离模块，利用二维像素对齐和空间对齐特征细化 SMPLX 派生的符号距离场。Xiu 等人（2023）提出 ECON（explicit clothed humans optimized via normal integration），预测前后法线与深度图，采用深度感知双边法线积分优化器分别恢复前后表面。Yang 等人（2023）提出基于隐式分布场 D-IF（uncertainty-aware human digitization via implicit distribution field）的方法，利用分布预测网络估计查询点的占用概率分布。Liao 等人（2023）提出以粗到精的方式创建着装人体网格，通过正则法向量引导的细化过程优化服装细节表面。此外，结合注意力机制（Li 等，2023；Zhuang 等，2024）和混合表示（Albahar 等，2023；Jiang 等，2023）的方法也在不断发展。

1.2 扩散生成

扩散模型在三维人体生成中展现出强大的生成能力。Zhang 等人（2024b）提出的 SIFU（side-view conditioned implicit function for real-world usable clothed human reconstruction）通过侧视图条件隐式函数构建粗糙纹理网格，采用侧视图去耦变换器处理输入图像与 SMPLX 侧视图特征，最终进行基于扩散的三维一致性纹理细化。Zhang 等人（2024a）提出 HumanRef（single image to 3D human generation via reference-guided diffusion），基于 SMPLX 人体模板估计姿态形状，利用哈希编码符号距离场网络进行三维表示。

Huang 等人（2024）提出 TeCH（text-guided reconstruction of lifelike clothed humans），通过服装解析与视觉问答模型生成文本描述，利用文本到图像扩散模型隐式指定外观细节。Song 等人（2023）提出深度指导隐式函数（depth-guided implicit function for clothed human reconstruction，DIFu），通过生成器产生背面图像与深度图，将深度图投影到三维空间提取精确体素对齐特征。Chen 等人（2024）提出多视图图像生成模型，采用视图选择策略全面覆盖人体。Lee 等人（2024）提出 PIDiffu（pixel-aligned diffusion model for high-fidelity clothed human reconstruction），通过像素对齐光线采样解决深度模糊问题。Chan 等人（2024b）引入循环调节机制，检验生成视图序列的一致性，提高对人体结构的理解。

1.3 视图一致性约束

视图一致性约束确保从不同视图获取的三维数据在几何和纹理上保持一致，避免由于视图差异导致的几何生成错误。Hong 等人（2021）将立体视觉的几何约束与隐式函数相结合，通过相对 z 偏移概

念恢复几何细节，同时利用深度图为可见区域提供完整表面约束。Zhou等人（2024）引入像素对齐空间变换器，通过计算多视角图像相关性融合特征，并采用几何引导的可见性推理机制，仅集成可见源视图特征以解决稀疏视图遮挡问题。Li等人（2024）提出基于扩散的傅里叶占用场方法，在预测后视角图像与参考图像间引入样式一致性约束。大多现有方法主要集中在后处理阶段，噪声和视角偏差的影响导致几何生成结果的细节难以准确恢复。

上述方法存在的局限性如表1所示，对此，本文提出融合姿态扩散先验与多视图一致性的生成方法。受到SiTH（single-view textured human reconstruction with image-conditioned diffusion）（Ho等，2024）的思路启发，本文采用先估计后生成的策略。然而，不同于SiTH，引入姿态引导机制，通过关键点的高斯热图提供精确的几何约束，确保生成视图与输入图像的姿态一致性。与法线预测依赖于单一输入图像的方法（Xiu等，2023）不同，通过双视图特征融合确保多视图几何信息的一致性。相较于D-IF仅在单一分布中采样（Yang等，2023），本文的分布预测网络基于SMPLX学习空间变化的概率分布，能够更好地适应人体不同部位的几何复杂度。

表1　单视图着装人体生成方法对比

Table 1　Comparison of single-view clothed human generation methods

方法类型	代表方法	优势	局限性
隐式函数	PIFu（Saito等，2019）/ICON（Xiu等，2022）	拓扑灵活	几何背面过度平滑
NeRF方法	SHERF（Hu等，2023）	渲染质量高	计算开销大
高斯泼溅	Gaussian shell maps for efficient 3D human generation（Abdal等，2024）	能实现实时渲染	难以恢复细节
扩散生成	SIFU（Zhang等，2024b）/SiTH（Ho等，2024）	纹理丰富	几何约束不足
混合方法	本文	兼具优势	计算复杂度适中

2　着装人体生成

本文方法流程如图1所示，包含3个核心模块：姿态扩散先验生成、多视图一致性约束与自适应几何生成。姿态扩散先验生成模块利用关键点编码的姿态信息引导扩散模型生成高质量后视角图像；多视图一致性约束模块基于SMPLX人体模板的法线信息整合前后视图特征，确保不同视角下的几何一致性；自适应几何生成模块通过学习空间变化的概率分布和位置相关方差场，自适应不同部位的几何复杂性，使复杂拓扑结构的生成更加稳定。

2.1　姿态扩散先验生成

传统扩散模型缺乏明确的几何约束，容易产生姿态不一致的后视角图像，本文设计高斯热图编码机制，将25个人体关键点集合$\boldsymbol{k}=\{(x_i,y_i)|i\in\{0,1,2,3,\cdots,24\}\}$（结构如图2所示）转化为可学习的几何先验，其中$i$是关键点索引，$(x_i,y_i)$是第$i$个关键点的坐标。自适应高斯姿态热图计算为

$$g_i(x_i,y_i,\sigma)=\exp\left(-\frac{(x-x_i)^2+(y-y_i)^2}{2\sigma^2}\right) \tag{1}$$

式中，(x,y)为像素位置，$g_i(x_i,y_i,\sigma)$是以k中第i个关键点为中心、标准差为σ的高斯函数。与直接使用关键点坐标不同，高斯热图编码可实现空间连续性建模，以理解关键点周围的空间关系。

现有方法将所有姿态信息混合编码，无法有效区分不同身体部位的语义特征。如表2所示，为了保持关键点信息的语义独立性，提出语义独立的多通道编码策略，每个关键点对应独立热图通道，避免不同身体部位间的特征耦合。具体为

$$G=\sum_{i=0}^{24}g_i(x_i,y_i,\sigma) \tag{2}$$

通过解耦设计使模型能够独立学习每个关键点的纹理变化模式，特别是在处理服装褶皱等局部细节时表现出更强的表达能力。

传统条件扩散模型主要基于单一模态，缺乏几何与纹理的协同建模。本文设计几何—纹理协同的姿态条件融合机制，对高斯热图$\boldsymbol{G}$进行多模态融合。具体为

$$c=\boldsymbol{\Psi}(\boldsymbol{G},U,S) \tag{3}$$

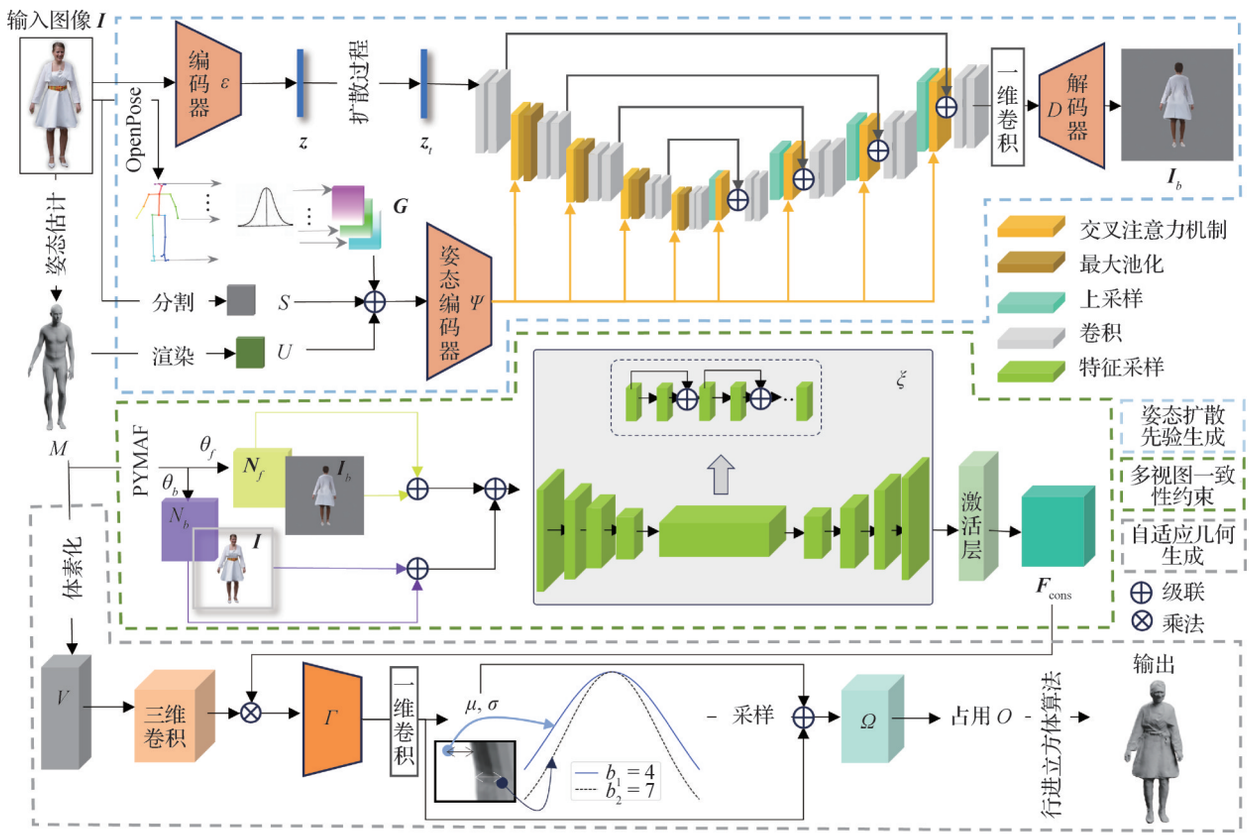


图 1　融合姿态扩散先验与多视图一致性的着装人体生成流程图

Fig. 1　Flowchart of clothed human generation via pose diffusion priors and multiview consistency

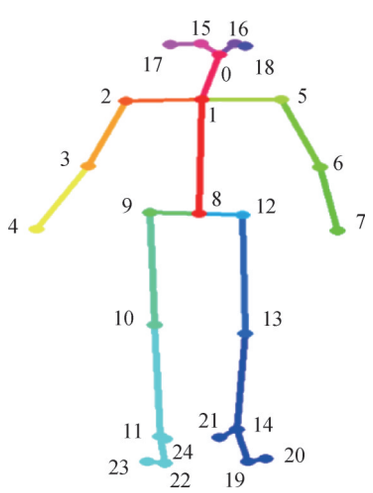


图 2　人体关键点结构图

Fig. 2　Human body keypoint structure diagram

式中，Ψ 为姿态编码器。UV 映射 U 提供纹理空间的几何对应关系，人体掩码 S 确保生成区域的精确控制，高斯热图 G 提供关键点约束。三者融合形成了几何—纹理—空间的三重约束机制。传统扩散模型的条件注入缺乏针对性，无法确保生成结果在关键区域的准确性。本文将姿态特征 c 通过自适应交叉注意力机制注入 U-Net 各层，实现精细化的去噪引导。具体定义为

$$f_{\text{Attention}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = f_{\text{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{4}$$

式中，查询 $\boldsymbol{Q}$ 来自图像特征，键值对 $(\boldsymbol{K},\boldsymbol{V})$ 由 c 得到。与全局条件注入不同，本文的交叉注意力机制使模型能够在每个去噪步骤中动态关注不同的关键点区域，实现空间自适应的生成控制。

表 2　关键点名称对应通道编号

Table 2　Key point name corresponding channel number

序号	关节名	序号	关节名	序号	关节名
0	鼻子	9	右臀	17	右耳
1	脖子	10	右膝盖	18	左耳
2	右肩	11	右脚踝	19	左大拇指
3	右手肘	12	左臀	20	左小拇指
4	右手腕	13	左膝盖	21	左脚跟
5	左肩	14	左脚踝	22	右大拇指
6	左手肘	15	右眼	23	右小拇指
7	左手腕	16	左眼	24	右脚跟
8	中臀				

在姿态条件c的引导下，采用概率采样机制生成后视角图像，具体定义为

$$z_{t-1}=\frac{1}{\sqrt{\alpha_t}}\left(z_t-\frac{\beta_t}{\sqrt{1-\alpha_t^2}}\varepsilon_\theta(z_t,t,c)\right)+\sigma_t\varepsilon \tag{5}$$

$$\boldsymbol{I}_b=D(z_0) \tag{6}$$

式中，t表示时间步，α_t和β_t为扩散过程参数，ε_θ为去噪网络，最终解码器D将目标潜在向量z_0转换为后视角图像$\boldsymbol{I}_b$。通过姿态条件c的引导，采样过程能够在保持随机性的同时确保几何一致性。

2.2 多视图一致性约束

基于2.1节生成的后视角图像，本文提出多视图一致性约束方法。通过SMPLX的几何先验建立前后视图的显式关联，设计跨视角法线一致性网络学习几何一致的特征表示。

给定SMPLX人体模板$M\in\mathbf{R}^{N\times 3}$，首先通过可微分渲染器从前后视角分别渲染法线图，具体为

$$\boldsymbol{N}_f,\boldsymbol{N}_b=Render(M,\theta_f,\theta_b) \tag{7}$$

式中，θ_f和θ_b分别表示前视角和后视角的相机参数，$\boldsymbol{N}_f,\boldsymbol{N}_b\in\mathbf{R}^{H\times W\times 3}$为渲染的法线图。

为有效处理视角间的几何差异和纹理变化，基于输入图像$\boldsymbol{I}$与生成的后视角图像$\boldsymbol{I}_b$进行几何对齐，即

$$\boldsymbol{F}_f=\boldsymbol{\Phi}_f(\boldsymbol{I},N_f),\boldsymbol{F}_b=\boldsymbol{\Phi}_b(\boldsymbol{I}_b,N_b) \tag{8}$$

式中，$\boldsymbol{\Phi}_f$和$\boldsymbol{\Phi}_b$为前后视角的特征提取器，$\boldsymbol{F}_f,\boldsymbol{F}_b\in\mathbf{R}^{H\times W\times C}$分别为前后视角的几何引导特征。通过将图像特征与对应的法线信息融合，以实现纹理与几何协同建模，使特征具备明确的几何语义，有效解决视角转换时的特征漂移问题。

然后，考虑到可微分渲染中每个像素的RGB值编码法线方向：R通道表示法线的X分量、G通道表示法线的Y分量、B通道表示法线的Z分量，前视角人体法线N_f表示指向相机的表面方向，与摄像头朝向一致。后视角人体法线N_b指向相机相反方向，形成天然的方向性差异。本文利用法线图的物理特性构建跨视角法线一致性网络ξ，具体为

$$\boldsymbol{F}_{\text{cons}}=\xi(\boldsymbol{F}_f,\boldsymbol{F}_b,\boldsymbol{N}_f,\boldsymbol{N}_b) \tag{9}$$

网络ξ通过学习前后视角法线的符号差异模式，使模型能够理解并利用视图间的几何关系进行特征融合。由于单一的几何生成损失无法充分约束视图间的几何一致性，为了度量真实法线和预测法线的差异，定义双重约束机制为

$$L_{\text{geo}}=\left|\boldsymbol{N}_{\text{pred}}-\boldsymbol{N}_{\text{gt}}\right| \tag{10}$$

$$L_{\text{perc}}=\sum_l\lambda_l\left\|\phi_l(\boldsymbol{N}_{\text{pred}})-\phi_l(\boldsymbol{N}_{\text{gt}})\right\|_2^2 \tag{11}$$

式中，$\boldsymbol{N}_{\text{pred}}$为预测法线，$\boldsymbol{N}_{\text{gt}}$为真实法线，$\phi_l$为预训练VGG（Visual Geometry Group）网络的第l层特征，λ_l为加权系数。几何损失L_{geo}确保像素级精度，感知损失L_{perc}保证高层语义一致性，该机制能提升视图间的几何连续性。

2.3 自适应几何生成

基于2.1节生成的高质量后视角图像和2.2节融合的多视图特征，本文进一步构建自适应几何生成模块实现最终的三维几何生成。

传统隐式函数方法假设全局统一的分布模式，无法适应人体不同区域的几何复杂度差异。为此，设计空间感知的概率分布预测网络，根据查询点的空间位置和几何上下文预测个性化的高斯分布参数。首先以SMPLX人体模板M作为输入，进行体素化操作获得体素网格V，通过三维卷积提取几何特征，具体为

$$\boldsymbol{F}_{\text{3d}}=Conv3D(V) \tag{12}$$

再将来自多视图一致性约束模块的跨视角特征$\boldsymbol{F}_{\text{cons}}$与几何特征融合：$\boldsymbol{F}_{\text{hybrid}}=\boldsymbol{F}_{\text{cons}}\odot\boldsymbol{F}_{\text{3d}}$。其中，$\odot$表示逐元素乘积操作。由于固定的占用预测无法表达查询点的不确定性，导致边界区域生成模糊，本文将混合特征$\boldsymbol{F}_{\text{hybrid}}$输入分布预测网络$\Gamma$，提出参数化高斯分布建模，具体形式定义为

$$N(\mu_p,\sigma_p^2)=\Gamma(F_{\text{hybrid}},p) \tag{13}$$

式中，p为查询点坐标，μ_p和σ_p^2分别为预测的均值和方差。通过学习每个点的分布参数，模型能够表达不同空间位置的几何不确定性，特别适用于处理宽松服装等复杂拓扑结构。

然后，针对确定性的占用预测容易在边界处产生锯齿状结果，设计概率密度采样机制，从预测分布中进行连续采样，具体为

$$p_{\text{sample}}=\int N(o\left|\mu_p,\sigma_p^2\right)\mathrm{d}o \tag{14}$$

式中，o表示占用值，进一步引入占用精化网络Ω，结合采样结果和分布参数预测最终占用，具体为

$$o_p=\Omega(p_{\text{sample}},\mu_p,\sigma_p^2,\boldsymbol{F}_{\text{hybrid}}) \tag{15}$$

针对Yang等人（2023）提出的KL（Kullback Leibler）散度损失忽略了人体结构的内在特性，为了更精确地进行概率建模，本文设计人体结构感知的

分布学习损失，以区分人体内外部区域的不同特性。采用KL散度衡量预测分布$N(\mu_{\text{pred}}, \sigma^2_{\text{pred}})$与目标分布$N(\mu_{gt}, \sigma^2_{gt})$的差异，定义分布约束损失为

$$L_{\text{dist}} = \text{KL}(N(\mu_{\text{pred}}, \sigma^2_{\text{pred}}) \| N(\mu_{\text{gt}}, \sigma^2_{\text{gt}})) \quad (16)$$

式中，均值μ_{gt}设定为真实占用值，方差σ^2_{gt}根据人体结构特性自适应调整。利用SDF标签区分点p是否位于网格外部，为内外部点设计目标方差衰减策略，具体为

$$\sigma_{\text{gt}}(p) = k \cdot \left[I_0 \cdot \mathrm{e}^{-b_1(\mu - 0.5)^2} + I_1 \cdot \mathrm{e}^{-b_2(\mu - 0.5)^2} \right] \quad (17)$$

式中，k用于控制方差的缩放比例，b_1、b_2是缩放超参数（实验设置：$b_1 = 4, b_2 = 7$），I_0、I_1是指示函数，当$SDF(p)<0.5$时I_0为0，其他为1，且$I_0 + I_1 = 1$，通过此策略，使得离表面较远的外部点保持高不确定性，靠近表面或位于内部的点具有更低不确定性，以提升对复杂服装结构的适应能力。

为确保最终占用场的几何精确性，采用L2损失约束预测值与真实值的差异，具体为

$$\mathcal{L}_{\text{recon}} = \frac{1}{N} \sum_{i=1}^{N} \left\| \boldsymbol{o}_{\text{pred}}(p_i) - \boldsymbol{o}_{\text{gt}}(p_i) \right\|_2^2 \quad (18)$$

式中，N为采样点数量，$\boldsymbol{o}_{\text{pred}}(p_i)$和$\boldsymbol{o}_{\text{gt}}(p_i)$分别为预测和真实的占用值。将分布学习与几何生成精度两部分加权结合，得到最终训练目标。具体为

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{dist}} + \lambda \mathcal{L}_{\text{recon}} \quad (19)$$

式中，$\lambda = 0.5$在实验中实现了分布表达能力与几何生成精度的最佳平衡。双重损失机制确保了从粗糙几何到精细细节的渐进式学习，分布损失提供全局约束，几何生成损失确保局部精度。最后结合分布损失，对优化后的$\boldsymbol{F}_{\text{hybrid}}$进行人体隐式生成，输出最终的着装人体模型。

3 实验结果与分析

3.1 实验设置及数据集

实验使用的硬件平台如下：AMD 48核处理器@2.3 GHz，NVIDIA GeForce RTX 3090 GPU（24 GB显存），128 GB内存，基于PyTorch 1.12.0深度学习框架和CUDA 11.6实现。训练采用RMSprop优化器，学习率设置为1×10^{-4}，批次大小为2，总训练轮数为15个epoch，并引入权重衰减（1×10^{-5}）和梯度裁剪（1.0）确保训练稳定性。为提升模型泛化能力，采用分阶段训练策略：前5个epoch单独训练姿态扩散先验生成模块，第6～15个epoch进行全模块端到端训练。

实验在两个公开数据集上进行验证。Thuman2.0数据集包含525幅高质量的三维人体扫描数据，涵盖150种不同服装款式和多样化人体姿态，与SIFU（Zhang等，2024a）方法类似，本文选择490次扫描用于训练，21次扫描用于测试，确保训练数据的充分性和测试结果的可靠性。CAPE（clothed auto-person encoding）数据集遵循ICON（implicit clothed humans obtained from normals）（Xiu等，2022）的标准测试协议，从中精选150次扫描，划分为两个子集：CAPE-FP（fitted poses）包含75次相对简单的姿态扫描，用于测试方法在标准姿态下的几何生成精度；CAPE-NFP（non-fitted poses）包含75次复杂姿态扫描，重点评估方法对极端姿态和复杂动作的适应能力。

3.2 评估指标

与对比方法（Xiu等，2022，2023；Yang等，2023；Zhang等，2024a）保持一致，本文采用3个互补的几何评估指标，从不同维度量化几何生成质量。倒角距离（chamfer distance，CD）衡量预测点云与真实点云之间的双向最近邻距离，评估整体几何形状的相似度。点到面距离（point-to-surface distance，P2S）计算预测点集中每个点到真实表面的最小距离，评估生成表面的几何精度。法线差异（normals difference，Normal）计算预测表面法线与真实法线的夹角差异，评估表面方向的准确性和局部几何细节。CD评估整体形状一致性，P2S关注表面几何精度，Normal检测局部细节质量，三者结合提供了从宏观到微观的全面生成质量评估。所有指标均采用越小越好的评价标准。

3.3 对比和分析

3.3.1 定量对比

为确保对比的全面性和公平性，选择6个代表性的单视图着装人体生成方法进行对比，包括基于隐式函数的PIFu（Saito等，2019）、PaMIR（parametric model-conditioned implicit representation for image-based human reconstruction）（Zheng等，2022），基于参数模型的ICON（Xiu等，2022）、ECON（Xiu等，2023）、D-IF（Yang等，2023），以及基于扩散模型的SIFU（Zhang等，2024b）。对比方法的相关数据均来源于原文。表3展示了在THuman2.0数据集上的定量对比结果。本文方法在CD和P2S指标上均取得

最优性能，CD值为0.558 7，相比次优方法SIFU降低6.3%；P2S值为0.571 0，相比次优方法降低5.7%。这表明本文方法能够生成更精确的几何结构，生成的点云分布更接近真实模型。在Normal指标上，本文方法为0.048 2，略高于SIFU的0.040 7，主要原因是SIFU采用四视角法线特征而本文使用双视角特征。与使用双视角特征的方法（Xiu等，2022，2023；Yang等，2023）相比，本文方法仍保持较低的法线差异。

为验证本文方法的泛化能力，在THuman2.0数据集上训练，在CAPE数据集上测试。表4显示，本文方法在CAPE-FP和CAPE-NFP子集上均取得最优的CD和P2S性能。在CAPE-FP上，CD值为0.573 8，P2S值为0.588 6；在CAPE-NFP上，CD值为0.707 0，P2S值为0.711 9，均显著优于现有方法，验证了本文方法对未见数据具有良好的泛化能力，能够有效处理复杂姿态下的人体生成任务。

表5展示了推理效率的对比。本文方法推理时间为23.25 s，虽然比传统方法D-IF（Yang等，2023）和HiLo（detailed and robust 3D clothed human reconstruction with high-and low-frequency information of parametric models）（Yang等，2024）略高，但相比同类扩散模型SIFU（Zhang等，2024b）减少21.9%的推理时间，说明本文方法的潜在扩散模型将特征映射到低维空间，有效降低了计算复杂度。

表3　本文方法在THuman2.0数据集上与其他方法的定量对比结果

Table 3　Quantitative comparison results of the proposed method with other methods on the THuman2.0 dataset

方法	CD	P2S	Normal
PIFu（Saito等，2019）	1.599 1	1.433 3	0.084 3
PaMIR（Zheng等，2022）	1.215 2	1.058 2	0.073 0
ICON（Xiu等，2022）	0.949 1	0.984 6	0.062 1
ECON（Xiu等，2023）	1.258 5	1.418 4	0.061 2
D-IF（Yang等，2023）	1.169 6	1.290 0	0.093 6
SIFU（Zhang等，2024b）	0.596 1	0.605 8	**0.040 7**
本文	**0.558 7**	**0.571 0**	0.048 2

注：加粗、下划线字体表示各列最优、次优结果。

表4　本文方法在CAPE数据集上与其他方法的定量对比结果

Table 4　Quantitative comparison results of the proposed method with other methods on the CAPE dataset

方法	CAPE-FP			CAPE-NFP		
	CD	P2S	Normal	CD	P2S	Normal
PIFu（Saito等，2019）	1.813 9	1.510 8	0.079 8	2.560 9	1.997 1	0.102 3
PaMIR（Zheng等，2022）	1.481 0	1.163 1	0.072 7	1.631 3	1.266 6	0.073 0
ICON（Xiu等，2022）	0.724 7	0.697 9	0.037 1	0.884 6	0.856 9	0.043 4
ECON（Xiu等，2023）	0.903 9	0.893 8	0.037 3	0.946 2	0.933 4	0.038 2
D-IF（Yang等，2023）	0.762 5	0.769 0	0.050 3	0.823 7	0.835 7	0.057 5
SIFU（Zhang等，2024b）	0.629 7	0.598 0	**0.032 7**	0.772 5	0.735 4	**0.037 8**
本文	**0.573 8**	**0.588 6**	0.035 7	**0.707 0**	**0.711 9**	0.040 4

注：加粗、下划线字体表示各列最优、次优结果。

表5　本文方法与其他方法的推理效率比较结果

Table 5　Inference efficiency comparison results of the proposed method with other methods

方法	推理时间/s
D-IF（Yang等，2023）	**18.51**
HiLo（Yang等，2024）	19.17
SIFU（Zhang等，2024b）	29.77
本文	23.25

注：加粗字体表示最优结果。

3.3.2　定性对比

本文从复杂姿态和宽松服装2个方面与基线方法HiLo（Yang等，2024）、SIFU（Zhang等，2024b）和D-IF（Yang等，2023）进行定性对比。图3给出了在野外图像上的定性对比结果。从图3第1行可以看出，本文方法能有效恢复面部细节，并在不可见区域生成自然的褶皱纹理。相比之下，同样基于扩散模型的方法SIFU由于采用SMPLX人体模板拼接策略，当姿态估计出现误差时会导致人体网格出现姿

态偏差。第2行结果显示，在复杂姿态下，本文方法在手部和脚部的生成上更准确。与D-IF和SIFU方法相比差异显著。第3行进一步验证了本文方法在局部细节恢复的优势。第4行展示了在复杂纹理情况下，本文方法能更准确地捕获服装褶皱变形，并在背部区域恢复合理的几何细节。实验结果表明，本文方法在处理复杂姿态和丰富服装细节时能有效实现更准确的着装人体几何生成。

为验证本文方法在宽松服装人体生成中的有效性，在野外图像上与现有方法进行对比，结果如图4所示。从图4第1行可以看出，对于身穿飘逸长裙的人体，本文方法能够准确生成服装褶皱的自然形态，而对比方法在背部区域出现了几何破碎现象。同时，本文方法生成的面部细节更加丰富。图4第2行展示了百褶裙的生成结果，本文方法在服装边缘表现得更加平滑，优于其他方法。图4第3行进一步验证了本文方法在褶皱细节还原的准确性优势。

3.3.3 鲁棒性分析

为了验证本文方法在不同视角输入以及复杂服装形变下的稳定性，进行定性对比实验，结果如图5所示。在侧视图输入测试中，对比方法在手臂、鞋子等部位出现了不同程度的伪影和破损，表明其在处理视角变化时存在一定的不稳定性，相比之下，本文

(a) 输入图像　(b) D-IF　(c) SIFU　(d) HiLo　(e) 本文

图3　本文方法在复杂姿态和丰富褶皱细节情况下与其他方法的定性对比

Fig. 3　Qualitative comparison of the proposed method with other methods under complex poses and rich wrinkle details ((a) input images; (b) D-IF; (c) SIFU; (d) HiLo; (e) ours)

(a) 输入图像　(b) D-IF　(c) SIFU　(d) HiLo　(e) 本文

图4　本文方法在不同款式的宽松服装下与其他方法的定性对比

Fig. 4　Qualitative comparison of the proposed method with other methods under different styles of loose clothing

((a) input images; (b) D-IF; (c) SIFU; (d) HiLo; (e) ours)

方法能够在该视角下提供更为准确和可信的结果。在复杂服装形变的低分辨率图像测试中，对比方法在服装边缘表现出明显的锯齿状伪影，并且D-IF和SIFU方法估计了错误的姿态，而本文方法能够成功恢复远离人体的服装和头发，实现边缘更加平滑的几何生成。在长发角色测试中，对比方法难以生成合理的后视角细节，头发的纹理缺乏流动感和层次结构。本文方法不仅生成效果更佳，还能够准确处理叉腰等复杂手部姿态。实验结果表明，本文方法在具有挑战性的输入条件下具有更好的鲁棒性和稳定性。

3.4　消融实验

3.4.1　姿态扩散先验生成的有效性分析

为验证本文方法中各模块的有效性，选用基线模型（base）并依次叠加不同组件进行消融实验，训练数据集为THuman2.0。针对高斯热图生成中的标准差σ参数，选取4个代表性数值（σ = 2、5、10、15）进行消融实验。

图6展示了不同σ值对同一关键点高斯热图形态分布的影响。当σ = 2时，热图分布过于集中，细节信息丰富但鲁棒性较差；当σ = 15时，热图分布过于分散，鲁棒性提高但细节捕捉能力下降。

为平衡细节捕捉与鲁棒性，设置σ = 5，对应的高斯热图可视化结果示例如图7所示。

为了评估生成模块的有效性，将本文提出的姿态扩散先验生成模块（PD）与SiTH（Ho等，2024）的后视角图像幻觉生成的图像质量进行了定性对比，SiTH的后视角图像幻觉模块与本文的姿态扩散先验生成模块都是基于生成的方法，但采用了不同的策略。本文通过对SiTH生成的后视角图像可视化，对比这两种方法的效果，如图8所示。SiTH的结果容易出现纹理的扭曲，本文提出的PD更加注重学习人体结构以及关键点和纹理的联系，因此结果在视觉感知上更加真实。

为了深入验证姿态扩散先验生成在野外图像上

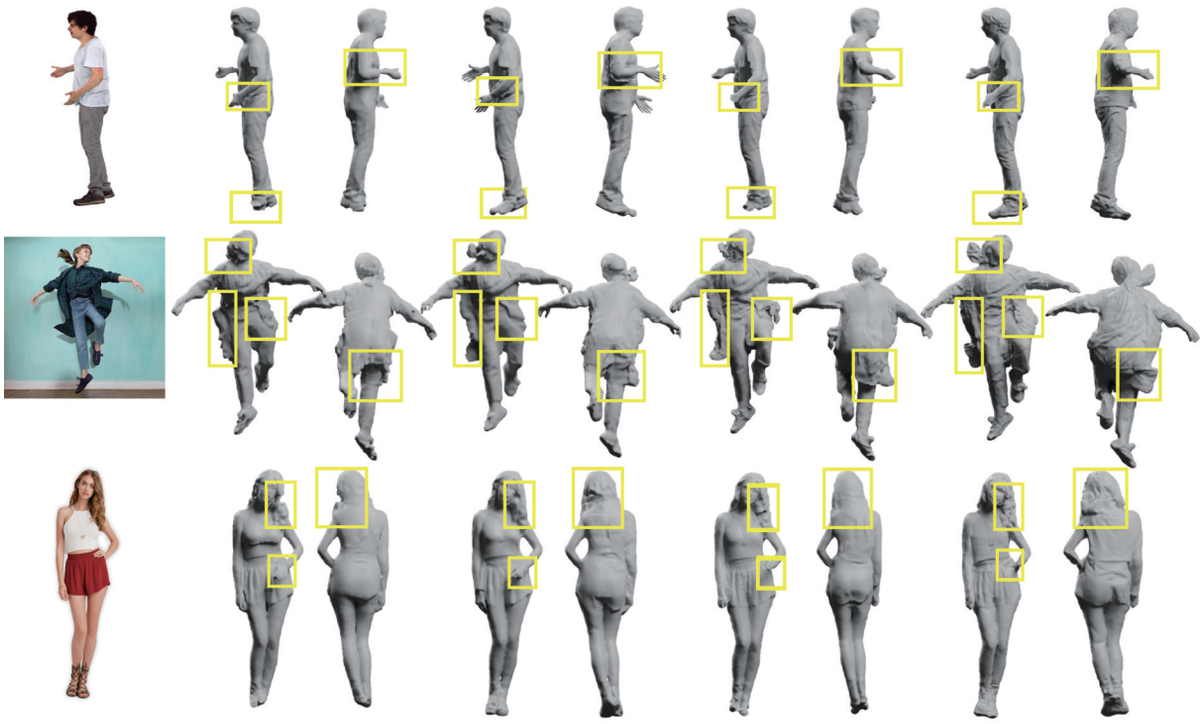


图 5　本文方法在侧视视图输入以及复杂服装形变下的定性对比结果

Fig. 5　Qualitative comparison results of the proposed method under side-view input and complex clothing deformation ((a) input images; (b) D-IF; (c) SIFU; (d) HiLo; (e) ours)

的生成能力，在 CAPE 数据集的野外图像上实验，可视化对比结果如图 9 所示。可以看出本文姿态扩散先验生成模块能更自然地生成与姿态变化一致的褶皱。当输入图像中的服装图案复杂时，SiTH 模型会出现局部区域的图案失真和错误学习，导致生成图像中的纹理与实际输入图像不一致。

3.4.2　多视图一致性约束的有效性分析

如表 6 所示，通过消融实验验证多视图一致性

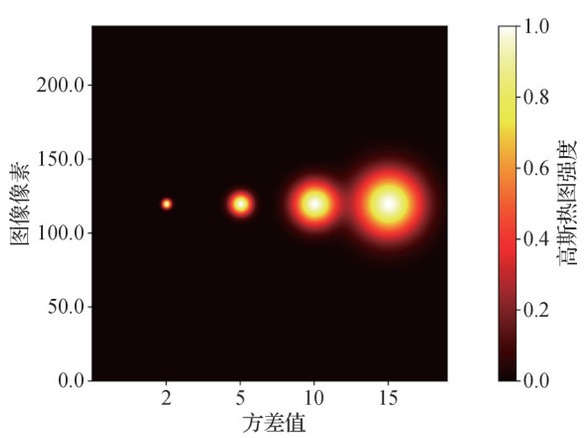


图 6　不同方差值的高斯热图对比

Fig. 6　Comparison of Gaussian heatmaps with different variance values

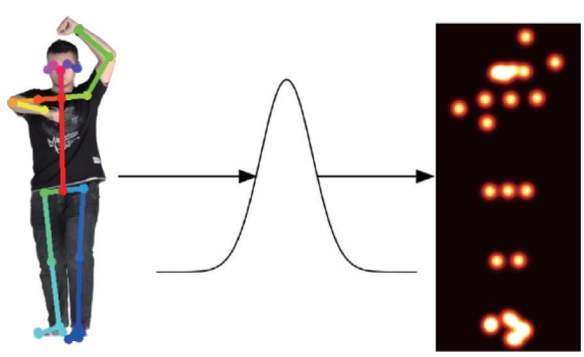


图 7　通道叠加后关键点编码的高斯热图可视化结果

Fig. 7　Visualization results of Gaussian heatmaps for keypoint encoding after channel addition

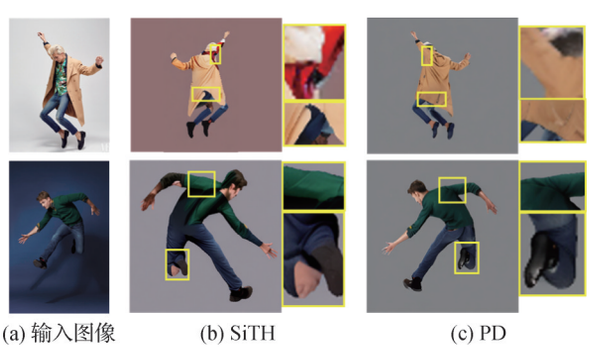
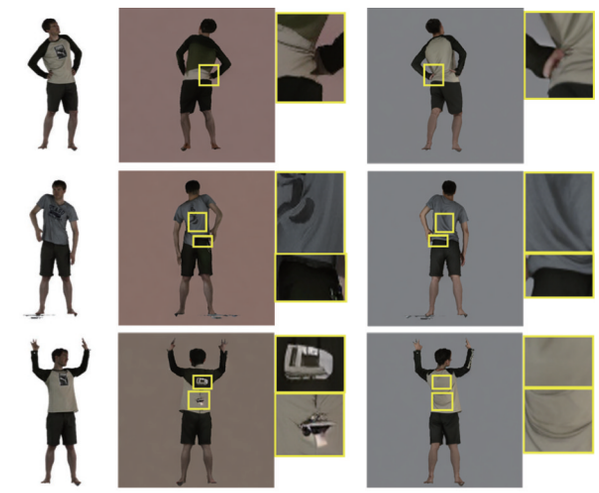


图 8　姿态扩散先验生成有效性的定性比较

Fig. 8　Qualitative comparison of the effectiveness of pose diffusion priors generation ((a) input images; (b) SiTH; (c) PD)



(a) 输入图像 (b) SiTH (c) PD

图9 姿态扩散先验生成在CAPE数据集上的定性比较

Fig. 9 Qualitative comparison of pose diffusion priors generation on the CAPE dataset ((a) input images; (b) SiTH; (c) PD)

约束模块(multiview consistency constraints, NC)对整体模型的贡献,实验设置、训练策略与基线模型(base)一致。结果显示,CD降低至0.875 2,P2S降低至0.884 1,两个几何精度指标均明显提升,表明模型生成的几何结构更加准确。Normal上升至0.075 9,这是由于姿态扩散先验生成模块在生成过程中融入了部分伪特征,导致法线估计存在偏差。总体而言,多视图一致性约束模块有效提升了生成结果的几何精度和视觉质量。

表6 多视图一致性约束的有效性分析定量结果

Table 6 Quantitative results of the effectiveness analysis of multiview consistency constraints

模块	CD	P2S	Normal
base	0.923 0	0.985 5	**0.073 2**
base + PD + NC	**0.875 2**	**0.884 1**	0.075 9

注:加粗字体表示各列最优结果。

3.4.3 自适应几何生成的有效性分析

为验证自适应几何生成的有效性,在多视图一致性约束模型上进行消融实验,并与基线模型在THuman2.0上比较,分别对参数b_1和b_2选择不同的值,其中,b_1控制人体外部点的方差,b_2控制人体内部点的方差。参数值越大对应方差越小。表7展示了不同参数组合的定量结果。

从表7可以看出,当$b1 = 10$,$b_2 = 6$时,3个指标(CD、P2S和Normal)均劣于基线设置,表明外部点过小的方差设置不适合人体结构建模。当$b_1 = 4$,$b_2 = 7$时,3项指标均取得最优性能,相比base,本文方法在CD降低0.101 6,P2S降低0.150 8,Normal降低0.001 9,验证了基于人体结构的分布方法的有效性,说明能有效提升生成几何与真实几何之间的相似度。

表7 自适应几何生成的有效性分析和参数消融定量结果

Table 7 Quantitative results of the effectiveness analysis of adaptive geometry generation and parameter ablation

模块	b_1	b_2	CD	P2S	Normal
Base	–	–	0.923 0	0.985 5	0.073 2
base + PD + NC + DI	10	6	0.894 1	0.910 3	0.079 9
	7	7	0.864 2	0.879 2	0.075 5
	4	6	0.879 8	0.904 3	0.074 7
	4	8	0.858 2	0.870 9	0.075 7
	4	7	**0.821 4**	**0.834 7**	**0.071 3**

注:加粗字体表示各列最优结果。“–”表示无相关信息。

图10展示了不同参数组合在野外图像上的定性结果。当$b_1 = 10$,$b_2 = 6$时,生成结果出现大量几何破洞;随着b_1减小,破洞逐渐减少,模型逐步感知服装分布;当$b_1 = 4$,$b_2 = 7$时生成效果最佳,能够恢复连续的几何形状;当$b_1 = 4$,$b_2 = 8$时,出现多层结构,表明模型过拟合。实验结果表明,最优的分布参数设置为$b_1 = 4$,$b_2 = 7$,对宽松服装生成更加适用。

4 结 论

针对单视图输入的着装人体生成结果存在不可见区域纹理缺失、局部细节不够清晰以及难以生成宽松服装几何等问题,本文通过构建姿态扩散先验生成、多视图一致性约束和自适应几何生成模块,提出了一种融合姿态扩散先验与多视图一致性的着装人体生成方法。姿态扩散先验生成模块利用高斯热图引导潜在扩散模型,补全不可见区域的合理纹理;多视图一致性约束模块基于姿态扩散先验生成的后视角图像,增强局部细节表达能力;自适应几何生成模块通过均值和方差学习查询点的概率分布,提升宽松服装生成的几何平滑性。

虽然本文方法对着装人体的纹理和宽松服装能够实现较为准确和鲁棒的生成,但是模型中包含的

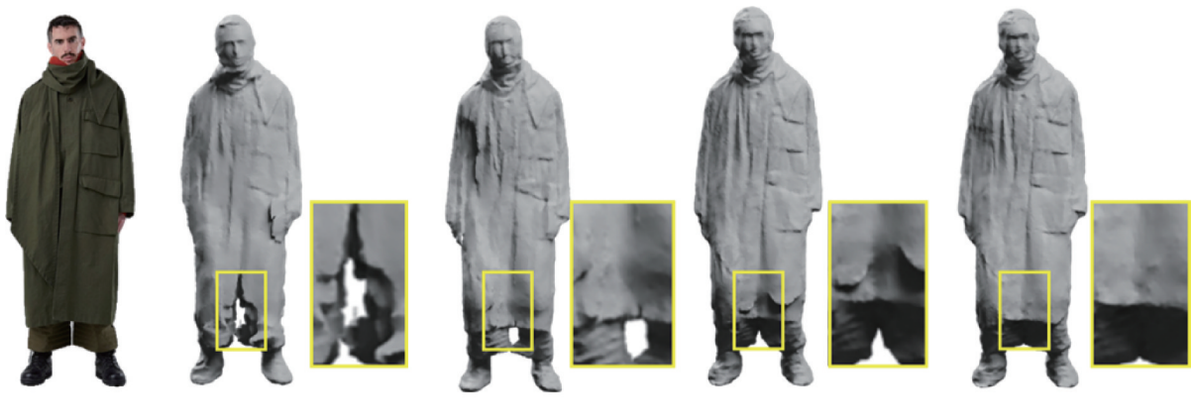


图 10　自适应几何生成的参数消融定性结果

Fig. 10　Qualitative results of parameter ablation for adaptive geometry generation

((a) input image; (b) $b_1 = 10$, $b_2 = 6$; (c) $b_1 = 7$, $b_2 = 7$; (d) $b_1 = 4$, $b_2 = 8$; (e) $b_1 = 4$, $b_2 = 7$)

扩散模型增加了额外的时空开销。此外，本文方法依赖姿态估计的质量，当关键点估计存在较大误差时，会影响生成结果。虽然本文在 CD 和 P2S 指标上取得最优性能，但 Normal 指标略低于 SIFU，主要原因是本文采用双视角特征而 SIFU 使用四视角特征。未来工作将围绕提高计算效率和增强姿态估计鲁棒性等方面进一步改进：通过网络轻量化等方法，减少模型的计算负担，提升处理速度；在保持计算效率的前提下提升法线估计精度并增强姿态估计的鲁棒性。

参考文献（References）

Abdal R, Yifan W, Shi Z F, Xu Y H, Po R, Kuang Z F, et al. 2024. Gaussian shell maps for efficient 3D human generation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9441-9451 [DOI: 10.1109/CVPR52733.2024.00902]

Albahar B, Saito S, Tseng H Y, Kim C, Kopf J and Huang J B. 2023. Single-image 3D human digitization with shape-guided diffusion//Proceedings of 2023 SIGGRAPH Asia Conference Papers. Sydney, Australia: ACM: #62 [DOI: 10.1145/3610548.3618153]

Cao Y K, Han K and Wong K Y K. 2023. SeSDF: self-evolved signed distance field for implicit 3D clothed human reconstruction//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 4647-4657 [DOI: 10.1109/CVPR52729.2023.00451]

Chan K Y, Lin G S, Zhao H Y and Lin W S. 2022. IntegratedPIFu: integrated pixel aligned implicit function for single-view human reconstruction//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 328-344 [DOI: 10.1007/978-3-031-20086-1_19]

Chan K Y, Liu F Y, Lin G S, Foo C S and Lin W S. 2024a. Fine structure-aware sampling: a new sampling training scheme for pixel-aligned implicit models in single-view human reconstruction//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 964-971 [DOI: 10.1609/aaai.v38i2.27856]

Chan K Y, Liu F Y, Lin G S, Foo C S and Lin W S. 2024b. R-cyclic diffuser: reductive and cyclic latent diffusion for 3D clothed human digitalization//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10304-10313 [DOI: 10.1109/CVPR52733.2024.00981]

Chen L F, Su J H and Luo S Y. 2023. TransPIFu: combining transformer and pixel-aligned implicit function for single-view clothed human reconstruction. Computers and Graphics, 111: 1-13 [DOI: 10.1016/j.cag.2022.12.009]

Chen M J, Chen J H, Ye X J, Gao H A, Chen X X, Fan Z X, et al. 2024. Ultraman: single image 3D human reconstruction with ultra speed and detail [EB/OL]. [2025-08-18]. https://arxiv.org/pdf/2403.12028.pdf

He T, Collomosse J, Jin H L and Soatto S. 2020. Geo-PIFu: geometry and pixel aligned implicit functions for single-view human reconstruction//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #778

Ho H I, Song J and Hilliges O. 2024. SiTH: single-view textured human reconstruction with image-conditioned diffusion//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 538-549 [DOI: 10.1109/CVPR52733.2024.00058]

Hong Y, Zhang J Y, Jiang B Y, Guo Y D, Liu L G and Bao H J. 2021. StereoPIFu: depth aware clothed human digitization via stereo vision//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 535-545

[DOI: 10.1109/CVPR46437.2021.00060]

Hu S K, Hong F Z, Pan L, Mei H Y, Yang L and Liu Z W. 2023. SHERF: generalizable human NeRF from a single image//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 9318-9330 [DOI: 10.1109/ICCV51070.2023.00858]

Huang Q P, Liu L, Fu X D, Liu L J and Peng W. 2024. Clothed feature learning for single-view 3D human reconstruction. Journal of Image and Graphics, 29(9): 2610-2624 (黄千芃, 刘骊, 付晓东, 刘利军, 彭玮. 2024. 单视角三维人体重建的着装特征学习. 中国图象图形学报, 29(9): 2610-2624) [DOI: 10.11834/jig.230623]

Huang Y Y, Yi H W, Xiu Y, Liao T T, Tang J X, Cai D, et al. 2024. TeCH: text-guided reconstruction of lifelike clothed humans//Proceedings of 2024 International Conference on 3D Vision (3DV). Davos, Switzerland: IEEE: 1531-1542 [DOI: 10.1109/3DV62453.2024.00152]

Jiang S Y, Jiang H R, Wang Z Y, Luo H M, Chen W Z and Xu L. 2023. HumanGen: generating human radiance fields with explicit priors//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 12543-12554 [DOI: 10.1109/CVPR52729.2023.01207]

Lee J, Kim S, Lee H, Adiya T and Lim H. 2024. PIDiffu: pixel-aligned diffusion model for high-fidelity clothed human reconstruction//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 5160-5169 [DOI: 10.1109/WACV57701.2024.00509]

Li J H, Yang Z X, Wang X H, Ma J X, Zhou C and Yang Y. 2023. JOTR: 3D joint contrastive learning with transformers for occluded human mesh recovery//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 9076-9087 [DOI: 10.1109/ICCV51070.2023.00836]

Li Y Z, Luo F and Xiao C X. 2024. Diffusion-FOF: single-view clothed human reconstruction via diffusion-based Fourier occupancy field//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9525-9534 [DOI: 10.1109/CVPR52733.2024.00910]

Liao T T, Zhang X M, Xiu Y, Yi H W, Liu X D, Qi G J, et al. 2023. High-fidelity clothed avatar reconstruction from a single image//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 8662-8672 [DOI: 10.1109/CVPR52729.2023.00837]

Lim B and Lee S W. 2024. TIFu: tri-directional implicit function for high-fidelity 3D character reconstruction//Proceedings of the 4th International Conference on Pattern Recognition and Artificial Intelligence. Jeju Island, Korea (South): Springer, 2024: 151-165 [DOI: 10.1007/978-981-97-8705-0_10]

Saito S, Huang Z, Natsume R, Morishima S, Li H and Kanazawa A. 2019. PIFu: pixel-aligned implicit function for high-resolution clothed human digitization//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 2304-2314 [DOI: 10.1109/ICCV.2019.00239]

Saito S, Simon T, Saragih J and Joo H. 2020. PIFuHD: multi-level pixel-aligned implicit function for high-resolution 3D human digitization//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 81-90 [DOI: 10.1109/CVPR42600.2020.00016]

Song D Y, Lee H, Seo J and Cho D. 2023. DIFu: depth-guided implicit function for clothed human reconstruction//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 8738-8747 [DOI: 10.1109/CVPR52729.2023.00844]

Xiu Y, Yang J L, Cao X, Tzionas D and Black M J. 2023. ECON: explicit clothed humans optimized via normal integration//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 512-523 [DOI: 10.1109/CVPR52729.2023.00057]

Xiu Y, Yang J L, Tzionas D and Black M J. 2022. ICON: implicit clothed humans obtained from normals//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 13286-13296 [DOI: 10.1109/CVPR52688.2022.01294]

Yang X T, Luo Y H, Xiu Y, Wang W, Xu H and Fan Z X. 2023. D-IF: uncertainty-aware human digitization via implicit distribution field//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 9088-9098 [DOI: 10.1109/ICCV51070.2023.00837]

Yang Y F, Liu D, Zhang S H, Deng Z S, Huang Z X and Tan M K. 2024. HiLo: detailed and robust 3D clothed human reconstruction with high-and low-frequency information of parametric models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10671-10681 [DOI: 10.1109/CVPR52733.2024.01015]

Zhang J B, Li X Y, Zhang Q, Cao Y P, Shan Y and Liao J. 2024a. HumanRef: single image to 3D human generation via reference-guided diffusion//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1844-1854 [DOI: 10.1109/CVPR52733.2024.00181]

Zhang Z C, Sun L, Yang Z X, Chen L and Yang Y. 2023. Global-correlated 3D-decoupling transformer for clothed avatar reconstruction//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #343

Zhang Z C, Yang Z X and Yang Y. 2024b. SIFU: side-view conditioned implicit function for real-world usable clothed human reconstruction//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9936-9947 [DOI: 10.1109/CVPR52733.2024.00948]

Zheng Z R, Yu T, Liu Y B and Dai Q H. 2022. PaMIR: parametric

model-conditioned implicit representation for image-based human reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6): 3170-3184 [DOI: 10.1109/TPAMI.2021.3050505]

Zhou T S, Huang J, Yu T, Shao R Z and Li K. 2024. HDhuman: high-quality human novel-view rendering from sparse views. IEEE Transactions on Visualization and Computer Graphics, 30(8): 5328-5338 [DOI: 10.1109/TVCG.2023.3290543]

Zhuang Y Y, Lyu J X, Wen H, Shuai Q, Zeng A L, Zhu H, et al. 2024. IDOL: instant photorealistic 3D human creation from a single image//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 26308-26319 [DOI: 10.1109/CVPR52734.2025.02450]

作者简介

张渊杭，女，硕士研究生，主要研究方向为图形图像处理。E-mail: 20232204305@stu.kust.edu.cn

刘骊，通信作者，女，教授，博士生导师，主要研究方向为计算机图形学与计算机视觉、图像处理。E-mail: ieall@kust.edu.cn

付晓东，男，教授，博士生导师，主要研究方向为服务计算、决策理论与方法。E-mail: xdfu@kust.edu.cn

刘利军，男，副教授，主要研究方向为图像处理云计算和信息检索。E-mail: cloneiq@kust.edu.cn

彭玮，女，教授，博士生导师，主要研究方向为机器学习和数据挖掘。E-mail: weipeng@kust.edu.cn