

基于注意力与专家混合网络的高保真数字人

谈华丰^{1,2,3}, 何儒汉^{1,2}

(1. 武汉纺织大学 计算机与人工智能学院; 2. 湖北省服装信息化工程技术研究中心;
3. 纺织服装智能化湖北省工程研究中心, 湖北 武汉 430200)

摘要: 针对动态人体渲染中非刚性运动难以准确建模的问题, 提出一种基于单目视频的高保真动态人体重建方法, 可以捕捉到丰富的人体姿态, 并且可以从新视角生成逼真的三维渲染结果, 该方法采用专家混合网络以分解复杂的运动模式, 利用多个专家子网络对非刚性变形进行细致的建模; 设计残差结构与通道注意力机制的阴影预测网络, 将残差结构和通道注意力机制融合起来, 从而增强对光照和阴影变化的建模能力; 采用纹理损失和梯度差异边缘感知损失相结合的方式, 促使模型恢复细粒度的褶皱和纹理细节。在 ZJU-MoCap 数据集上, 该方法在新视角合成任务中的 PSNR 为 30.44 dB、SSIM 为 0.969 0、LPIPS* 为 31.07, 在新姿态合成任务中的 PSNR 为 30.53 dB、SSIM 为 0.969 8、LPIPS* 为 29.95, 在 PeopleSnapshot 数据集上, 4 个测试序列的平均指标 PSNR 为 30.83 dB、SSIM 为 0.979 6、LPIPS 为 0.020 5, 都优于当前主流方法。该方法推理速度可以达到 40 FPS, 模型只有 3.92 MB, 实现了高精度和高效率的结合, 实验结果证明了该方法在动态细节建模、渲染质量和实时性方面具有优势, 在虚拟现实、数字人、影视制作、远程医疗等领域有巨大的应用潜力。

关键词: 高保真动态人体渲染; 专家混合网络; 单目视频; 注意力机制; 3D 数字人

DOI: 10.11907/rjdk.251564

扫描二维码阅读全文:



中图分类号: TP391.41

文献标识码: A

文章编号: 1672-7800(XXXX)0XX-0001-09

High-Fidelity Digital Humans Based on Attention and Expert Mixture Network

TAN Huafeng^{1,2,3}, HE Ruhan^{1,2}

(1. School of Computer Science and Artificial Intelligence, Wuhan Textile University; 2. Hubei Provincial Engineering Research Center for Intelligent Textile and Fashion; 3. Engineering Research Center of Hubei Province for Clothing Information, Wuhan 430200, China)

Abstract: In this study, a high-fidelity dynamic human body reconstruction method based on monocular video is proposed to solve the problem that non-rigid motion is difficult to accurately model in dynamic human body rendering. It can capture rich human posture and generate realistic 3D rendering results from a new perspective. This method uses expert mixture network to decompose complex motion patterns, and uses multiple expert sub-networks to model non-rigid deformation in detail. The shadow prediction network of residual structure and channel attention mechanism is designed, and the residual structure and channel attention mechanism are integrated to enhance the modeling ability of illumination and shadow changes. The combination of texture loss and gradient difference edge perception loss is used to promote the model to restore fine-grained wrinkles and texture details. On the ZJU-MoCap dataset, the performance of this method in the new perspective synthesis task reaches PSNR 30.44 dB, SSIM 0.969 0, LPIPS* 31.07, and the performance in the new pose synthesis task reaches PSNR 30.53 dB, SSIM 0.969 8, LPIPS* 29.95, on the PeopleSnapshot dataset. The average indexes of the four test sequences are PSNR 30.83 dB, SSIM 0.979 6 and LPIPS 0.020 5, which are better than the current mainstream methods. The inference speed of this method can reach 40 FPS, and the model is only 3.92 MB, which realizes the combination of high precision and high efficiency. The experimental results show that this method has advantages in dynamic detail modeling, rendering quality and real-time performance, and has the potential to be widely used in virtual reality, digital human, film and television production, telemedicine and other fields.

Key Words: high-fidelity dynamic human body rendering; expert mixture network; monocular video; attention mechanism; 3D digital human

收稿日期: 2025-09-16

作者简介: 谈华丰 (1999-), 男, CCF 学生会会员, 武汉纺织大学计算机与人工智能学院硕士研究生, 研究方向为 3D 数字人; 何儒汉 (1974-), 男, 博士, 武汉纺织大学计算机与人工智能学院教授、硕士生导师, 研究方向为计算机视觉和机器学习。本文通讯作者: 何儒汉。

0 引言

在影视制作、医疗应用、增强/虚拟现实及各种仿真场景中,高保真且可动画的数字化身正发挥着越来越重要的作用。传统构建方法往往需要昂贵的可穿戴设备或者复杂的多视角摄像系统,在专业动作捕捉(MoCap)工作室完成,既耗费资金又繁琐,很难做到个性化和大规模应用。于是,研发出一种依靠单目RGB视频就能完成数字化身构建的低成本、高质量方案成为亟待解决的问题。

近年来,隐式神经场(Implicit Neural Representations, INR)发展迅速,为低视角数量甚至单目视频的人体几何和外观重建带来了新的可能性^[1-2]。这些方法通常计算开销大,训练和推理过程慢,限制了它们在实际应用中的效率

和可扩展性^[3-7]。

相比之下,基于网格的方法具有很好的关节动画兼容性和渲染效率,但拓扑变化和细节还原仍有欠缺^[8]。基于点云的方法轻量灵活,但由于拓扑不完整、几何表达能力有限,难以满足高精度重建需求^[9]。

最近,高斯泼洒(Gaussian Splatting)在神经渲染领域取得了突破性进展^[10-15],这引起了人们对于它在非刚性动态场景中应用的广泛关注。基于网格的高斯渲染方法试图将高斯体素安排到网格表面上去,从而加强几何一致性并改进渲染速度,既保存了高斯渲染的高效特点,又改善了表面细节的还原水平^[14]。针对剧烈形变、拓扑改变或者明显运动模糊的动态情况,高斯泼洒办法在几何精确度和纹理保真程度上仍然存在明显的局限性,这制约着它在高质量动态重建与表现方面的普遍使用。

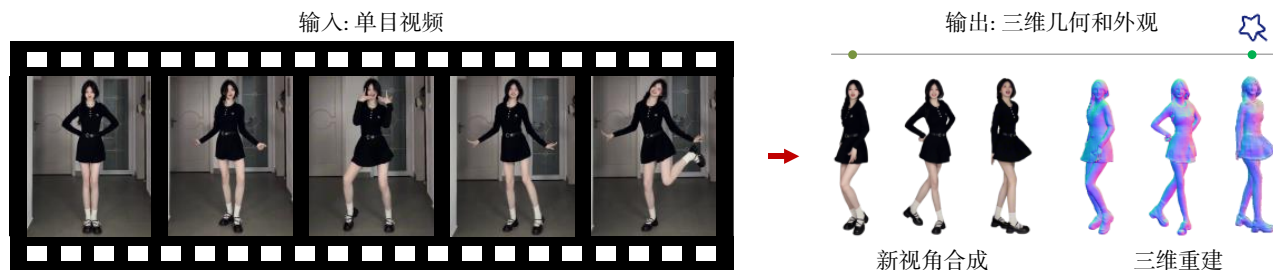


Fig. 1 High-fidelity dynamic human body rendering method

图1 高保真动态人体渲染方法

从动态人体的单目视频生成高保真的数字人仍具困难性,不仅要补全没有被观测到的角度,还要精确地建模衣物的非刚性变形和人体复杂的姿态变化。为此,本文提出了一种新的高保真动态人体渲染方法(见图1),可以详细捕捉人体动作和衣服细节,从而产生非常逼真且充满细节的人体渲染效果。本文主要贡献如下:①设计了融合残差结构与通道注意力机制的阴影预测网络(Residual Channel Attention Pseudo-Shadow Prediction Network, ResCAS-Net),准确捕捉视角依赖的阴影变化,使动态衣着人体在不同光照条件下的真实感和细节更加丰富;②提出了一种基于专家混合结构的非刚性运动建模范式,即专家混合网络(Expert Hybrid Network, EHN),将动态场景划分成多个具有某种特定运动模式的子空间,由3个专家子网络分别加以建模,从而可以更好地捕捉复杂动态细节,此设计既优化了运动建模的表达力,又大幅改善了模型在各种场景下的泛化性能;③在训练时引入了纹理损失和基于梯度差异的边缘感知损失,前者让图像在像素级、局部纹理上得到更好的还原,后者通过捕捉图像梯度域的差异,保留褶皱、轮廓这些高频细节,从而提升渲染质量与鲁棒性。

实验结果表明,本方法在高保真动态人体渲染任务中表现出优秀的视觉质量和细节再现能力,远超当前主流方法,此方法既改善了单目视频创建数字化身的实用程度,又为将来在虚拟人、交互娱乐、远程医疗等领域的应用给予了新的技术支持,具有重要理论意义和广泛应用前景。

1 相关工作

动态人体数字化身研究涉及多种表示与渲染技术,主要包含隐式神经表示、显式几何表示和高斯渲染等方法。

以神经辐射场(NeRF)为代表,利用体积渲染技术,在自由视角下实现了高质量的三维重建和图像合成^[5,16]。对于动态人体场景,有些方法将NeRF同人体参数或者特征编码融合起来,从而改善重建质量并加强动画效果^[17]。这类办法依靠较深的多层感知机(MLP)以预估体积密度和颜色,因而训练耗时长,推断速度慢,不能达到即时应用的要求。并且,隐式表达的计算复杂度较高,加上体积渲染开销,制约了它在动态、高复杂度场景中的广泛使用。

基于网格的显式几何方法结合运动学约束(线性混合蒙皮),实现几何变形和姿态驱动,具有较高的渲染效率^[8,18]。但是,其无法处理复杂的拓扑变化,难以准确捕捉到动态衣物褶皱等高频细节。点云表示灵活、轻量,却往往拓扑不完整、几何表征能力弱,导致其在动态建模中的表现力有限^[9]。

作为一种新兴的神经渲染技术,高斯泼洒通过将场景表示为一组三维高斯分布,并投影到二维图像平面以实现光栅化,在多视角静态场景中展现出极高的渲染效率与质量。为扩展至动态场景,研究者引入物理先验或基于神经网络的形变预测机制对高斯体素进行时序建模^[19-20]。但

在面对非刚性剧烈运动、拓扑变化和快速动态时,仍面临几何精度不足与细节还原受限等挑战。为提升动态建模能力,GART引入显式的人体骨架与模板网格,对高斯体素进行结构性约束,增强了复杂姿态与服饰变化下的稳定性与表达力,但对非结构化或极端形变的适应性有限^[12]。GoMAvatar将显式网格与高斯体素结合,提升了运动驱动效率与细节保真度,但在复杂衣物场景中仍存在性能瓶颈^[14]。Gaussianavatar从单目视频构建可动画3D高斯人偶,数据采集简单且可直接动画驱动,但在高频细节恢复方面受限^[11]。GGAvatar实现衣物与人体分离重建,细节表现力和动画灵活性较高,但细节融合度略显不足^[13]。TAGA采用无模板自监督范式,适应复杂形变与拓扑变化,但缺乏显式结构约束,可能导致局部几何扭曲与漂移^[15]。

综上所述,现有方法在渲染质量、建模效率与动态细节表达方面各具优劣:隐式神经场具备高质量渲染能力,但存在效率瓶颈;显式几何方法具有较高的渲染效率,但在细节建模方面表现不足;而高斯泼洒技术(Gaussian Splatting)在渲染速度与视觉效果之间取得了良好平衡,但在处理复杂动态变化时的适应性仍有限。为应对上述挑战,本文提出了一种高保真动态人体重建方法,核心包括3

项关键设计:①ResCASNet,融合残差结构与通道注意力机制,有效增强动态场景中光照与阴影细节的建模能力;②EHN,通过结构化的专家路由机制提升对非刚性运动(如衣物形变)的建模表现;③细节感知损失设计,纹理损失和基于梯度差异的边缘感知损失,提高几何重建精度与边缘清晰度。实验结果表明,该方法在动态人体渲染任务中展现出优异性能,尤其在细节保真度与感知真实感方面显著超越现有方法,在一定程度上缓解了隐式神经场效率低、显式几何方法细节不足以及高斯泼洒在复杂动态场景适应性有限的问题。但同时也注意到,对于极端快速运动与高度复杂的衣物拓扑变化,该方法仍存在改进空间。

2 高保真数字人渲染策略

在输入一段目标人物的单目视频后,本文目标是学习一个规范的 Gaussians-on-Mesh 表示(记作 GoM_θ^c),从而能够在任意相机内参 $K \in \mathbb{R}^{3 \times 3}$ 、外参 $E \in SE(3)$ 和人体姿态 P 下,实现人物高保真的实时渲染。本文参数 θ 表示可学习的变量,而上标 c 表示规范姿态空间。整体流程如图2所示。

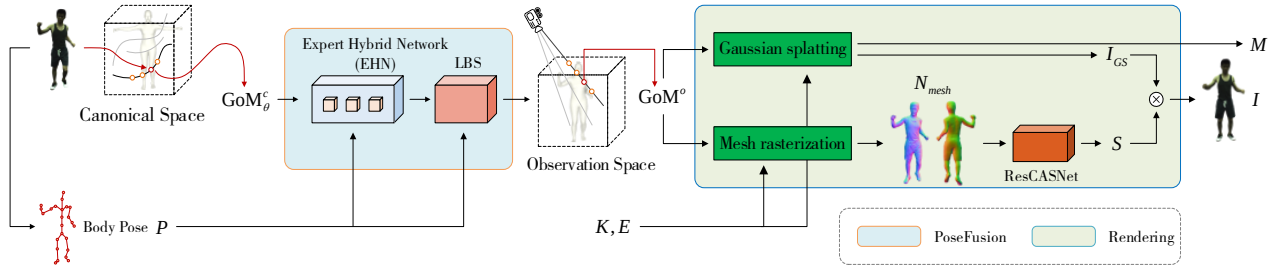


Fig. 2 Framework structure

图2 框架结构

为了在新的姿态下渲染人物,首先通过关节驱动机制(Pose Fusion, PoseFusion)将 GoM_θ^c 转换到观测空间,计算公式如式(1)所示。

$$\text{GoM}^o = \text{PoseFusion}_\theta(\text{GoM}_\theta^c, P) \quad (1)$$

本文关节驱动机制包含两个步骤。首先,利用基于混合专家(Mixture-of-Experts, MoE)架构的专家混合网络,对规范顶点进行与姿态相关的非刚性形变。该网络为每个规范顶点预测一个偏移量,用以捕捉由姿态引起的细微形变,生成非刚性变换后的规范位置 p_i^r 。接着,这些顶点位置通过线性混合蒙皮(Linear Blend Skinning, LBS)进一步变换到观测空间。LBS将关节的旋转和平移与学习得到的蒙皮权重结合,得到最终的关节驱动顶点位置 p_i^o 。

经过上述步骤后,得到的观测空间 GoM^o 被输入到神经渲染器中,用以生成彩色图像 $I \in \mathbb{R}^{H \times W \times 3}$ 和 Alpha 掩码 $M \in \mathbb{R}^{H \times W \times 1}$,计算公式如式(2)所示。

$$(I, M) = \text{Render}_\theta(K, E, \text{GoM}^o) \quad (2)$$

最终,通过标准的 Alpha 混合操作,将生成的人物外观与背景图像进行融合。分别介绍 Gaussians-on-Mesh 表示

方法、渲染过程、伪阴影预测网络、姿态细化策略、关节驱动方式、细节感知损失、训练损失构成和正则化策略,形成从表示到优化再到驱动的技术链条。

2.1 基于网格的高斯表示

Gaussians-on-Mesh (GoM) 通过将高斯分布附着在网格表面上,使其形状能够灵活适应不同的人体姿态,从而实现形变的正则建模。同时,借助高斯点的高效渲染特性,GoM 避免了传统体渲染中对每条光线进行密集采样的计算开销,大幅提升了渲染效率。规范高斯网格表示由一组带有关联属性的点和面共同定义,计算公式如式(3)所示。

$$\text{GoM}_\theta^c \triangleq \left\{ \left\{ v_{\theta,i}^c \right\}_{i=1}^V, \left\{ f_{\theta,j}^c \right\}_{j=1}^F \right\} \quad (3)$$

其中, $\left\{ v_{\theta,i}^c \right\}_{i=1}^V$ 和 $\left\{ f_{\theta,j}^c \right\}_{j=1}^F$ 分别表示具有相关属性的 V 个顶点和 F 个三角面。一个顶点的定义如式(4)所示。

$$v_{\theta,i}^c = (p_{\theta,i}^c, w_i) \quad (4)$$

其中, $p_{\theta,i}^c \in \mathbb{R}^3$ 表示顶点的三维坐标, $w_i \in \mathbb{R}^J$ 表示相对

于 J 个关节的线性混合蒙皮权重。一个面的定义如式(5)所示。

$$f_{\theta,j} = \left(r_{\theta,j}, s_{\theta,j}, c_{\theta,j}, \left\{ \Delta_{j,k} \right\}_{k=1}^3 \right) \quad (5)$$

其中, $r_{\theta,j} \in so(3)$ 和 $s_{\theta,j} \in \mathbb{R}^3$ 分别表示与该面相关联的局部高斯分布的旋转和缩放参数。此外 $c_{\theta,j} \in \mathbb{R}^3$ 是颜色向量。 $\left\{ \Delta_{j,k} \right\}_{k=1}^3$ 是属于第 j 面 3 个顶点的索引, 其中 $\Delta_{j,k} \in \{1, \dots, V\}$ 。

2.2 渲染

与传统单目人体渲染方法直接预测图像颜色的方式不同^[6-7], 本文方法在 Gaussians-on-Mesh (GoM) 表示上, 将 RGB 图像 I 分解为伪反照率图 I_{CS} 与伪光照图 (阴影) 图 S 。具体而言, 最终图像满足式(6)所示关系。

$$I = I_{CS} \cdot S \quad (6)$$

其中, I_{CS} 由高斯溅射 (Gaussian splatting) 生成, 而 S 则通过基于网格光栅化 (Mesh rasterization) 的法线图加以预测。模型渲染过程在观测空间下的 GoM 表示中进行, 如(2)所示。

$$\text{GoM}^o \triangleq \left\{ \left\{ (p_i^o, w_i) \right\}_{i=1}^V, \left\{ (r_{\theta,j}, s_{\theta,j}, c_{\theta,j}) \right\}_{j=1}^F \right\} \quad (7)$$

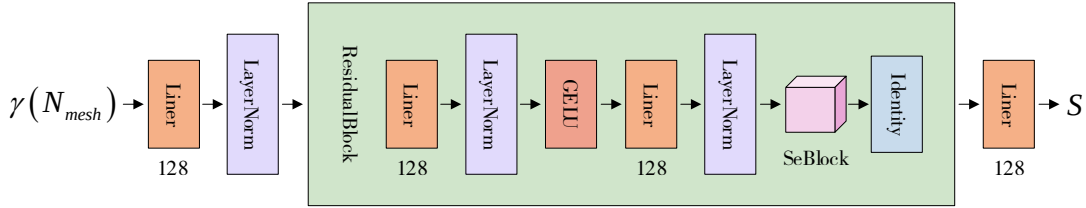


Fig. 3 Residual structure and channel attention pseudo-shadow prediction network

图3 残差结构与通道注意力伪阴影预测网络

2.4 姿势细化

由于图像估计的人体姿态往往存在误差, 本文使用姿态纠正模块对初始预测进行优化^[6]。具体而言, 给定预测的人体姿态 $\hat{P} = \left\{ \left(\hat{R}_j^p, t_j^p \right) \right\}_{j=1}^J$, 姿态纠正网络 Pose Refiner_o 会对各关节的旋转矩阵进行处理, 估计每个关节的修正值 $\xi_j \in SO(3)$, 计算公式如式(9)所示。

$$\left\{ \xi_j \right\}_{j=1}^J = \text{Pose Refiner}_o \left(\hat{R}_j^p \right)_{j=1}^J \quad (9)$$

随后, 更新每个关节的姿态, 计算公式如式(10)所示。

$$P = \left\{ \left(R_j^p, t_j^p \right) \right\}_{j=1}^J = \left\{ \left(\hat{R}_j^p \cdot \xi_j, t_j^p \right) \right\}_{j=1}^J \quad (10)$$

更新后的姿态将用于后续的图像合成模块。需要强调的是, 姿态纠正模块仅在训练阶段和新视角合成时启用, 用以减少初始估计误差的影响, 而在动画生成阶段不会启用。

2.5 关节驱动机制

2.5.1 专家混合网络

为更好地建模非刚性姿态变形, 本文引入专家混合网络 (Expert Hybrid Network, EHN), 如图4所示。EHN 采用

其中, GoM^o 的顶点坐标采用观测空间的 p_i^o 。下文详细介绍了伪光照图 S 的计算过程, 并在 2.5 节介绍了如何从规范空间的顶点坐标 $p_{\theta,i}^c$, 计算出 p_i^o 。

2.3 残差结构与通道注意力伪阴影预测网络

为建模视角依赖的光照变化, 本文提出了一种融合残差结构与通道注意力机制的伪阴影预测网络 (ResCAS-Net), 用于从光栅化得到的法线图 N_{mesh} 生成伪阴影图 $S \in \mathbb{R}^{H \times W \times 1}$ 。预测过程定义为如式(8)所示。

$$S = \sigma \left(\text{ResCASNet}_\theta \left(\gamma(N_{mesh}) \right) \right) \quad (8)$$

其中, $\gamma(\cdot)$ 表示位置编码 (Positional Encoding) 函数, 用于将输入的法线向量映射到高维空间, 从而增强对高频几何特征的感知能力; $\sigma(\cdot)$ 表示 Sigmoid 函数, 用于将输出约束在区间 $[0, 1]$ 。函数 $\text{ResCASNet}_\theta(\cdot)$ 表示本文设计的伪阴影预测网络, 如图3所示。网络先将位置编码特征线性映射至 128 维隐藏空间并进行 LayerNorm 归一化; 接着, 特征输入由两层线性变换组成的残差块, 层间依次使用 LayerNorm 和 GELU 激活, 残差路径中嵌入 SE 通道注意力模块, 自适应调整通道响应, 增强关键区域建模能力。残差连接采用恒等映射保证信息流稳定, 最后通过线性映射输出每像素阴影强度。

带有动态路由子网络的混合专家 (MoE) 结构, 可以针对不同的输入动态地选择适合的专家子网络, 从而增强对多样化的非刚性运动模式的表达能力。在此基础上, 预测 1 个偏移并将其添加到第 i 个规范顶点 p_i^c , 计算公式如式(11)所示。

$$p_i^{nr} = p_{\theta,i}^c + \text{MoE}_\theta \left(\gamma(p_{\theta,i}^c), P \right) \quad (11)$$

$$\text{MoE}_\theta(x) = \sum_{k=1}^K G(x) \cdot E_k(x) \quad (12)$$

其中, $\gamma(\cdot)$ 表示正弦位置编码 (Sinusoidal Positional Encoding), MoE_θ 是本文提出的形变模块。每个 MoE 的专家 E_k 是独立的多层感知机 (MLP), 架构相同但参数各异。具体计算公式如式(13)所示。

$$E_k(x) = \text{MLP}_k(x) \quad (13)$$

其中, 输入 $x = \left[\gamma(p_{\theta,i}^c) \oplus P \right]$ 。所有专家的输入维度 $d_{in} = 3 + \dim(P)$, 输出为三维偏移向量。MoE 中的门控网络 G 用于计算专家输出的自适应权重值, 其定义如式(14)所示。

$$G(x) = \text{soft max} \left(W_2 \cdot \sigma \left(\text{Dropout}(W_1 x) \right) \right) \quad (14)$$

其中, $W_1 \in \mathbb{R}^{d_m \times 128}$, $W_2 \in \mathbb{R}^{128 \times K}$ ($K = 3$ 位专家), σ 为 LeakyReLU 激活函数, *Dropout* (rate=0.1) 用于防止过拟合。最终的形变偏移通过门控权重对各专家输出加权求和得到。

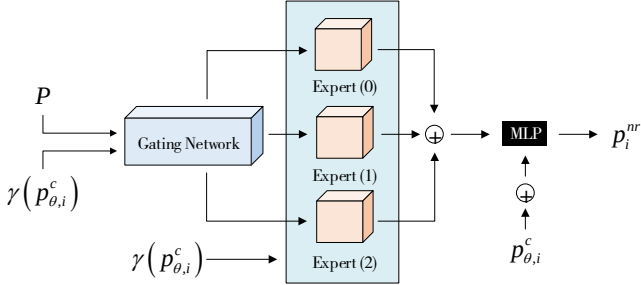


Fig. 4 Expert mixture network

图4 专家混合网络

2.5.2 线性混合蒙皮技术

本文采用标准的线性混合蒙皮(LBS)方法,将顶点从专家混合网络(非刚性变换)的规范空间映射到观测空间,计算公式如式(15)所示。

$$p_i^o = LBS(p_i^{nr}, w_i, P) = \frac{\sum_{j=1}^J w_j^i (R_j^p p_i^{nr} + t_j^p)}{\sum_{k=1}^J w_k^i} \quad (15)$$

其中,人体姿势 $P = \left\{ \left(R_j^p, t_j^p \right) \right\}_{j=1}^J$ 由 J 个关节的旋转和平移组成。每个顶点都关联有线性混合蒙皮权重 w_i 。

2.6 纹理损失函数

为了更好地渲染出衣服的动态褶皱和其他细节,本文使用了基于感知的纹理损失函数以生成图像的纹理细节,计算公式如式(16)所示。

$$L_{\text{texture}} = \frac{1}{N} \sum_{i=1}^N \left\| \phi(I_{\text{pred}}^{(i)}) - \phi(I_{\text{gt}}^{(i)}) \right\|_2^2 \quad (16)$$

其中, $\phi(I_{\text{pred}}^{(i)})$ 和 $\phi(I_{\text{gt}}^{(i)})$ 分别表示经过 VGG 网络提取后的预测图像和真实图像的特征向量。 N 是图像中的像素总数。函数 ϕ 表示从预训练的 VGG-19 网络中间层提取的激活输出,该函数常用于计算感知损失。 $I_{\text{pred}}^{(i)}$ 和 $I_{\text{gt}}^{(i)}$ 分别是第 i 个像素的预测 RGB 值和真实 RGB 值, $\|\cdot\|_2^2$ 是平方差。

2.7 梯度差异的边缘感知损失函数

为了保留边缘结构,本文使用了梯度差异的边缘感知损失以惩罚预测图像与真实图像之间梯度幅度的差异。在水平方向和垂直方向上采用 Sobel 算子计算梯度,并且定义了它们之间的平均绝对差异作为损失函数,计算公式如式(17)所示。

$$L_{\text{edge}} = \frac{1}{N} \sum_{i=1}^N \left\| \nabla I_{\text{pred}}^{(i)} - \nabla I_{\text{gt}}^{(i)} \right\| \quad (17)$$

其中, I_{pred} 和 I_{gt} 分别表示预测图像和真实图像的灰度图, ∇I 表示图像的梯度(使用 Sobel 滤波器提取, $\|\nabla I\| = \sqrt{(\partial_x I)^2 + (\partial_y I)^2} + \varepsilon$ 为梯度幅值(边缘强度))。 N 是图像中

像素的总数, ε 是一个小常数(如 1×10^{-6}),防止数值不稳定, $|\cdot|$ 是 L1 范数(绝对值差)。

2.8 训练

本文使用预测 RGB 图像 I 和主体掩码 M , 与其对应的真实值 I_{gt} 和 M_{gt} 进行监督,总体损失函数定义如式(18)所示。

$$L = \lambda_I L_I + \lambda_{\text{lpips}} L_{\text{lpips}} + \lambda_M L_M + \lambda_{\text{mask}} L_{\text{mask}} + \lambda_{\text{lap}} L_{\text{lap}} + \lambda_{\text{normal}} L_{\text{normal}} + \lambda_{\text{color}} L_{\text{color}} + L_{\text{reg}} \quad (18)$$

其中, λ_* 是各损失项的权重, L_I 和 L_M 分别是 RGB 图像和主体掩码 L1 损失, L_{lpips} 是预测的 RGB 图像 I 与地面真实图像 I_{gt} 之间的 LPIPS 感知损失。 $L_{\text{mask}} = \|M_{\text{mesh}} - M_{\text{gt}}\|$ 是对网格轮廓(silhouette)的正则约束, $L_{\text{lap}} = \frac{1}{N} \sum_{i=1}^N \|\delta_i\|^2$ 是拉普拉斯平滑损失。 δ_i 表示第 i 个顶点的拉普拉斯坐标。 L_{normal} 是法线一致性损失,旨在最大化相邻法线的余弦相似度。 L_{color} 是颜色平滑损失,惩罚相邻面之间的颜色差异,原理类似于法线一致性。本文还引入了额外的正则化,它由纹理损失函数和梯度差异的边缘感知损失函数组成,计算公式如式(19)所示。

$$L_{\text{reg}} = \lambda_{\text{texture}} L_{\text{texture}} + \lambda_{\text{edge}} L_{\text{edge}} \quad (19)$$

2.9 ResCASNet 与 EHN 算法流程

算法1 ResCASNet 算法

输入: 法向量 normals
输出: 阴影预测(S)

1. function ResCASNet(normals)
2. pos_embed $\leftarrow$ PositionalEncoding(normals)
3. h $\leftarrow$ Linear(pos_embed $\rightarrow$ hidden_dim)
4. h $\leftarrow$ LayerNorm(h)
5. FOR each ResidualBlock in blocks DO
6. IF skip_connection THEN
7. h $\leftarrow$ Concatenate(h, pos_embed)
8. END IF
9. END FOR
10. out $\leftarrow$ Linear(h $\rightarrow$ 1)
11. RETURN Sigmoid(out)
12. END function

算法2 EHN 算法

输入: 骨架顶点 xyzs_skeleton, 目标姿态向量 dst_posevec, 当前迭代步数 i_iter, 可选旋转矩阵 R, 可选缩放向量 S

输出: 更新后的顶点 xyzs_new, 更新后的旋转 R_new, 更新后的缩放 S_new

1. function EHN(xyzs_skeleton, dst_posevec, i_iter, R, S)
2. xyzs $\leftarrow$ Permute(xyzs_skeleton)
3. pos_embed_fn $\leftarrow$ GetEmbedder(multires, i_iter)
4. pos_embed $\leftarrow$ pos_embed_fn(xyzs)
5. h $\leftarrow$ Concatenate(Expand(dst_posevec), pos_embed)
6. offset $\leftarrow$ MoE(h)
7. offset $\leftarrow$ Linear(offset)
8. xyzs_new $\leftarrow$ xyzs_skeleton + offset[...,:3]

```

9. IF update_rot THEN
10. R_off ← SO3ExpMap(offset[ ..., 3:6 ])
11. R_new ← R_off @ R
12. ELSE
13. R_new ← R
14. END IF
15. IF update_scale THEN
16. S_new ← S + offset[ ..., 6: ]
17. ELSE
18. S_new ← S
19. END IF
20. RETURN xyzs_new, R_new, S_new
21. END function

```

3 实验与结果分析

3.1 数据集

本文以 ZJU-MoCap 数据集^[3]和 PeopleSnapshot 数据集^[21]作为训练与评估的基础,以全面验证方法性能和泛化能力。为此,在 ZJU-MoCap 数据集中选取了 6 名表现出复杂运动模式的个体(对应序列编号为 377、386、387、392、393、394),并严格遵循 MonoHuman 中定义的训练与测试划分,采用其单目视频人体渲染设置。针对每个主体,使用 Camera 0 拍摄的前 4/5 帧数据进行训练,利用其他摄像头对应帧的数据评估新视角合成效果,进一步通过所有视角中的最后 1/5 帧数据对新姿态合成性能进行测试,确保模型在空间和动作上的泛化能力。

在 PeopleSnapshot 数据集中,选取了 4 名具有代表性的个体(编号分别为 m3c、m4c、f3c、f4c),涵盖不同性别和服饰类型,以增强实验的多样性和全面性。严格按照 InstantAvatar 中制定的评估协议对方法进行了系统验证,确保评测过程的科学性和可比性。所有数据均经过预处理与统一格式转换,以保证训练和评估过程的稳定性与一致性。

3.2 训练细节

本文使用 Adam 优化器^[22],其中 $\beta_1 = 0.9, \beta_2 = 0.999$ 。在 ZJU-MoCap 数据集上,使用单个 V100 GPU 对模型进行了 300 k 次迭代训练,单个数据集训练时长为 27 个小时。本文将 Pose Refiner _{θ} 的学习率设置为 $5e-5$,模型其余部分的学习率为 $5e-4$ 。本文分别在 100 k 和 150 k 迭代后开始训练 Pose Refiner _{θ} 和 MoE _{θ} 。对于 MoE _{θ} ,遵循 HumanNeRF^[6]在训练过程中采用 HanW 窗口。在式(18)中,设置 $\lambda_{\text{lips}} = 1.0, \lambda_M = 5.0, \lambda_{\text{lap}} = 10.0, \lambda_{\text{normal}} = 0.1, \lambda_{\text{color}} = 0.05$;在式(19)中,设置 $\lambda_{\text{texture}} = 0.5, \lambda_{\text{edge}} = 0.5$ 。本文在 50 k 迭代后对 GoM 进行细分。

在 PeopleSnapshot 数据集上,对模型进行了 200 k 次迭代训练,并在 100 k 次迭代后开始训练 MoE _{θ} ,单个数据集训练时长为 20 个小时。本文在 10 k 迭代后对 GoM 进行一次细分,不会在 InstantAvatar 之后使用 Pose Refiner _{θ} 以优化

训练姿势^[10]。

3.3 方法对比

本文对比评估了一系列当前主流的单视频人体建模方法,包括 NeuralBody^[3]、Anim-NeRF^[4]、NeuMan^[5]、HumanNeRF^[6]、MonoHuman^[7]、InstantAvatar^[10]、GaussianAvatar^[11]、GART^[12]、GGAvatar^[13]、GoMAvatar^[14]和 TAGA^[15]。这些方法与本文方法类似,均以一段视频及其对应的三维人体骨架作为输入,生成可驱动的人体神经表示,从而支持新视角渲染和新姿态合成。

在新视角合成与新姿态合成任务中,本文采用 PSNR、SSIM、LPIPS 和 LPIPS* 作为评估指标,以全面衡量方法在视觉保真度与感知一致性方面的表现。

峰值信噪比(Peak Signal-to-Noise Ratio, PSNR)计算公式如式(20)所示。

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}(I_{\text{pred}}, I_{\text{gt}})} \right) \quad (20)$$

其中, MAX_I 表示图像像素可能的最大值(如 255 或 1), MSE 为预测图像 I_{pred} 与真实图像 I_{gt} 之间的均方误差。

结构相似度(Structural Similarity Index, SSIM)计算公式如式(21)所示。

$$\text{SSIM}(I_{\text{pred}}, I_{\text{gt}}) = \frac{(2\mu_{I_{\text{pred}}}\mu_{I_{\text{gt}}} + C_1)(2\sigma_{I_{\text{pred}} \cdot I_{\text{gt}}} + C_2)}{(\mu_{I_{\text{pred}}}^2 + \mu_{I_{\text{gt}}}^2 + C_1)(\sigma_{I_{\text{pred}}}^2 + \sigma_{I_{\text{gt}}}^2 + C_2)} \quad (21)$$

其中, $\mu_{I_{\text{pred}}}, \mu_{I_{\text{gt}}}$ 分别为图像 $I_{\text{pred}}, I_{\text{gt}}$ 的均值, $\sigma_{I_{\text{pred}}}^2, \sigma_{I_{\text{gt}}}^2$ 为方差, $\sigma_{I_{\text{pred}} \cdot I_{\text{gt}}}$ 为协方差, C_1, C_2 为稳定常数。

感知相似度(Learned Perceptual Image Patch Similarity, LPIPS)计算公式如式(22)所示。

$$\text{LPIPS}(I_{\text{pred}}, I_{\text{gt}}) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(f_l(I_{\text{pred}})_{hw} - f_l(I_{\text{gt}})_{hw} \right) \right\|_2^2 \quad (22)$$

其中, $f_l(\cdot)$ 表示在预训练深度网络的第 l 层提取的特征, w_l 为学习到的权重, H_l, W_l 为特征图的大小。LPIPS* 则为 $\text{LPIPS*} = \text{LPIPS}(I_{\text{pred}}, I_{\text{gt}}) \times 1000$ 。

3.4 实验结果

为验证本文方法的有效性,在 ZJU-MoCap 与 PeopleSnapshot 数据集上进行了对比实验,结果如表 1—表 4 所示。对比方法包括 Neural Body、HumanNeRF、NeuMan、MonoHuman、GoMAvatar 等具有代表性的动态人体重建方法。评价指标采用 PSNR、SSIM 和 LPIPS*,并额外报告了推理耗时(Inference, MS)与显存占用(Memory, MB)。

从表 1 可以看出,本文方法在新视角合成和新姿态合成任务上都取得了更佳结果,在 ZJU-MoCap 6 个序列的平均表现上,本文方法的 PSNR 分别为 30.44 dB 和 30.53 dB,均高于目前最好的 GoMAvatar 方法的 30.37 dB 和 30.34 dB,并且在很多指标上都有稳定的领先优势。更值得一提的是,本文方法在 SSIM、LPIPS* 指标上也实现了全面优化,说明本文方法在保持结构的同时,还能够达到感知质量上

的双重优势。推理效率方面,本文方法的平均推理时间仅为 25.0 ms,在保证高保真的同时,几乎接近 GoMAvatar (23.2 ms) 的速度,远远超过了 HumanNeRF (1 776.7 ms) 和 MonoHuman (5 970.0 ms),兼顾了精度和速度。本文方法的显存消耗仅为 3.92 MB,在性能优异的情况下,极大降低了资源成本,进一步提高了实用性和扩展性。

逐场景对比(见表 2、表 3,)可知,本文方法在大部分

场景下均为最优或者次优,在 Subject386 的 PSNR 上达到了 33.39 dB(新视角)、33.66 dB(新姿态),明显高于其他方法;在 Subject392 的 SSIM 上达到了 0.971 1,视觉结构保留更好。对于较为复杂的人体动作(如 Subject393),本文方法略占优势,说明本文方法具有较好的泛化性和鲁棒性,图 5 的可视化结果也验证了上述结论。

Table 1 Quantitative results on the ZJU-MoCap dataset

表 1 ZJU-MoCap 数据集上的定量结果

Method	Novel view synthesis			Novel pose synthesis			Inference	Memory
	PSNR ↑/dB	SSIM ↑	LPIPS* ↓	PSNR ↑/dB	SSIM ↑	LPIPS* ↓	MS ↓	MB ↓
Neual Body	28.72	0.961 1	52.25	28.54	0.960 4	53.91	212.3	16.76
HumanNeRF	29.61	0.962 5	38.45	29.74	0.965 5	34.79	1 776.7	245.73
NeuMan	28.96	0.947 9	60.74	28.75	0.940 6	62.35	3 412.5	2.27
MonoHuman	28.81	0.949 8	54.29	29.83	0.967 2	34.23	5 970.0	280.67
GoMAvatar	30.37	0.968 9	32.53	30.34	0.968 8	32.39	23.2	3.63
Ours	30.44	0.969 0	31.07	30.53	0.969 8	29.95	25.0	3.92

Table 2 Scene-by-scene decomposition of new perspective synthesis based on the ZJU-MoCap dataset

表 2 基于 ZJU-MoCap 数据集的新视角合成逐场景分解

Method	Subject377			Subject386			Subject387			Subject392			Subject393			Subject394		
	PSNR	SSIM	LPIPS*	PSNR	SSIM	LPIPS*	PSNR	SSIM	LPIPS*	PSNR	SSIM	LPIPS*	PSNR	SSIM	LPIPS*	PSNR	SSIM	LPIPS*
	↑/dB	↑	↓	↑/dB	↑	↓	↑/dB	↑	↓	↑/dB	↑	↓	↑/dB	↑	↓	↑/dB	↑	↓
Neual Body	29.08	0.967 9	41.17	29.76	0.964 7	46.96	26.84	0.953 5	60.82	29.49	0.964 0	51.06	28.50	0.959 1	57.07	28.65	0.957 2	55.78
HumanNeRF	29.79	0.971 4	28.49	32.10	0.964 2	41.84	28.11	0.962 5	37.46	30.20	0.963 3	40.06	28.16	0.957 7	40.85	29.28	0.955 7	41.97
MonoHuman	28.36	0.954 9	44.83	31.90	0.958 2	52.60	27.46	0.947 5	58.47	29.23	0.946 8	58.44	27.30	0.943 0	57.80	28.60	0.948 5	53.57
GoMAvatar	30.60	0.976 8	23.91	32.97	0.975 2	30.36	28.34	0.963 5	36.30	31.04	0.970 8	33.25	28.80	0.962 2	37.77	30.44	0.964 6	33.56
Ours	30.62	0.976 8	22.96	33.48	0.976 8	26.05	28.34	0.963 4	35.48	31.14	0.971 1	32.07	28.71	0.961 7	37.08	30.45	0.964 6	32.65

Table 3 Scene-by-scene decomposition of new pose synthesis based on the ZJU-MoCap dataset

表 3 基于 ZJU-MoCap 数据集的新姿态合成逐场景分解

Method	Subject377			Subject386			Subject387			Subject392			Subject393			Subject394		
	PSNR	SSIM	LPIPS*	PSNR	SSIM	LPIPS*	PSNR	SSIM	LPIPS*	PSNR	SSIM	LPIPS*	PSNR	SSIM	LPIPS*	PSNR	SSIM	LPIPS*
	↑/dB	↑	↓	↑/dB	↑	↓	↑/dB	↑	↓	↑/dB	↑	↓	↑/dB	↑	↓	↑/dB	↑	↓
Neual Body	29.29	0.969 3	39.40	30.71	0.966 1	45.89	26.36	0.952 0	62.21	28.97	0.961 5	57.03	27.82	0.957 7	59.24	28.09	0.955 7	59.66
HumanNeRF	29.91	0.975 5	23.87	32.62	0.967 2	39.36	28.01	0.963 4	35.27	30.95	0.968 7	34.23	28.43	0.960 9	36.26	28.52	0.957 3	39.75
MonoHuman	30.23	0.977 1	23.87	32.41	0.968 3	39.45	27.61	0.961 3	38.34	30.88	0.971 4	33.52	28.75	0.963 4	35.43	29.10	0.961 9	34.75
GoMAvatar	30.68	0.977 6	23.41	32.86	0.973 7	32.25	28.18	0.962 6	36.43	31.44	0.971 6	33.20	29.09	0.963 5	36.02	29.79	0.963 8	33.00
Ours	30.44	0.977 3	22.17	33.66	0.976 5	26.43	28.37	0.963 7	34.70	31.55	0.972 2	31.16	28.98	0.963 5	34.59	30.19	0.965 5	30.67

在 PeopleSnapshot 数据集(见表 4)上,本文方法优势也很明显,m3c、m4c、f3c、f4c 这 4 个场景的 PSNR 值分别为 32.54 dB、30.28 dB、30.61 dB、31.87 dB,均高于已有方法。在 SSIM 和 LPIPS 指标上,本文方法也取得了最优或者次优结果,在单目视频人体重建任务中表现出较高的保真度和鲁棒性。图 6 所示的视觉结果进一步佐证了这一点,左侧图是本文方法与 GoMAvatar 的定性对比,右侧图展示了本文几何信息,渲染表面法线可视化。

综合来看,本文方法在重建精度上远超已有方法,在推理速度和显存消耗方面也达到了较好平衡,表明本文基于残差结构注意机制与专家混合网络的人体建模方法可以较好地解决非刚性运动带来的建模问题,在动态人体渲染任务中具有较好的应用前景。

值得注意的是,虽然本文方法在大多数场景和指标上

都领先,但是在一些指标或者数据集上还没有达到绝对最优,这是因为 ResCASNet 在姿势高度变化的情况下,对于不同的光照条件存在一定偏差,后续可以引入光照自适应机制、多光照数据增强策略以及结合高频纹理约束的优化方法以进一步提高模型在复杂姿态和光照条件下的表现,并且能够更好地恢复细节纹理。

3.5 消融实验

本文在 ZJU-MoCap 数据集上,分别通过新的视角合成和新的姿态开展消融实验,对 6 个序列(377、386、387、392、393、394)进行平均评估,结果如表 5 所示。通过移除 EHN、ResCASNet 以及纹理损失函数、基于梯度差异的边缘感知损失函数以确定每个模块对整体性能的影响。

实验结果表明,完整模型在 PSNR、SSIM 和 LPIPS*等指标上均达到当前最优或接近最优。去掉 EHN 会使 LPIPS*下降很多,说明 EHN 对边缘细节的恢复起着重要作

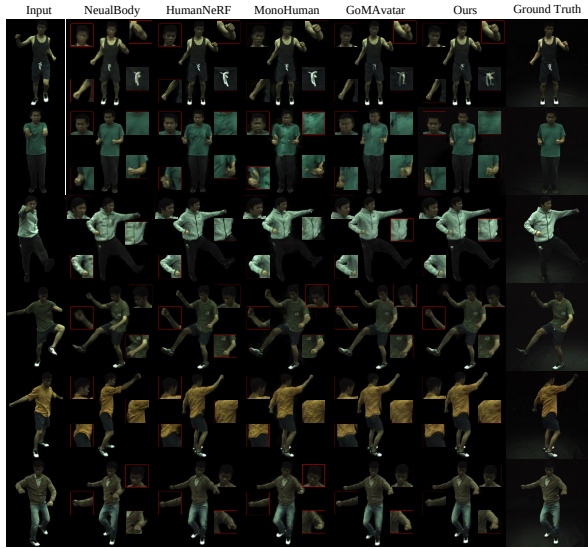


Fig. 5 Qualitative results of the new perspective synthesis task on the ZJU-MoCap dataset

图5 ZJU-MoCap 数据集上的新视角合成任务定性结果

用, 去掉 ResCASNet 会使结构保真度略微降低, 说明 ResCASNet 有助于提高几何一致性; 去掉纹理与边缘感知损失项会影响外观细节的还原能力, 如图 7 所示, 各组件都

Table 4 Decomposition of each scene of novel view synthesis on the PeopleSnapshot dataset

表4 PeopleSnapshot 数据集上的新视图合成每场景分解

Method	m3c			m4c			f3c			f4c		
	PSNR ↑ /dB	SSIM ↑	LPIPS ↓	PSNR ↑ /dB	SSIM ↑	LPIPS ↓	PSNR ↑ /dB	SSIM ↑	LPIPS ↓	PSNR ↑ /dB	SSIM ↑	LPIPS ↓
Neural Body	24.94	0.942 8	0.032 6	24.71	0.946 9	0.042 3	23.87	0.950 4	0.034 6	24.37	0.945 1	0.038 2
Anim-NeRF	29.37	0.970 3	0.016 8	28.37	0.960 5	0.026 8	28.91	0.974 3	0.021 5	28.90	0.967 8	0.017 4
HumanNeRF	23.82	0.948 7	0.037 6	26.76	0.960 5	0.023 7	26.13	0.955 8	0.027 7	24.46	0.936 5	0.050 2
InstantAvatar	29.65	0.973 0	0.019 2	27.97	0.964 9	0.034 6	27.90	0.972 2	0.024 9	28.92	0.969 2	0.018 0
GaussianAvatar	30.98	0.979 0	0.014 5	28.78	0.975 5	0.022 8	29.55	0.976 2	0.022 5	30.84	0.977 1	0.014 0
GART	30.40	0.976 9	0.037 7	27.57	0.965 7	0.060 7	26.26	0.965 6	0.049 8	29.23	0.972 0	0.037 8
GGAvatar	31.01	0.981 2	0.034 9	29.40	0.972 8	0.050 9	27.96	0.975 3	0.045 2	30.29	0.975 0	0.034 3
GoMAvatar	31.74	0.979 3	0.018 7	29.78	0.973 8	0.028 2	29.83	0.975 8	0.020 9	31.38	0.978 0	0.017 4
TAGA	29.69	0.974 9	0.018 2	27.17	0.965 6	0.036 8	25.81	0.962 9	0.026 3	28.63	0.971 8	0.019 3
Ours	32.54	0.982 9	0.017 7	30.28	0.976 0	0.027 3	30.61	0.978 3	0.019 8	31.87	0.981 2	0.017 2

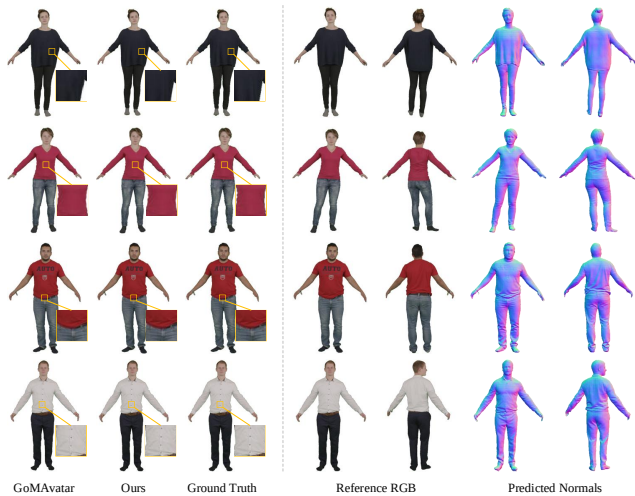


Fig. 6 The qualitative results presented on the PeopleSnapshot dataset

图6 PeopleSnapshot 数据集上展示定性结果

在提升模型整体重建质量上发挥着不可替代的作用, 进一步验证了所提出的方法在结构和外观建模上的有效性。

3.5 失败案例

该方法在无遮挡场景中效果较好, 但是对遮挡区域重建效果不太理想, 细节信息容易丢失, 如图 8 所示。未来研究将针对这些不足加以改进, 提高模型在复杂场景中的鲁棒性和细节恢复能力。

4 结语

针对动态人体中复杂的非刚性运动、衣物褶皱变化等建模问题, 本文提出了一种新的基于单目视频的高保真动态人体重建方法。引入 EHN 对非刚性运动进行精细建模, 提高了其对各种运动模式的适应性; 在关键子模块中设计了 ResCASNet, 进一步加强了模型对动态细节的建模能力和特征表达能力。并且, 还提出了结合纹理损失和基于梯度差异的边缘感知损失函数, 更好地保留衣物褶皱和局部纹理等细粒度外观信息, 使重建结果更加真实, 细节更丰富。在复杂动态场景下, 本方法比已有方法具有更好的适应性, 但是当姿势高度变化较大或存在严重遮挡时,

其仍然存在一定局限。

Table 5 Ablation experiments on the ZJU-MoCap dataset

表5 ZJU-MoCap 数据集上的消融实验

	Novel view synthesis			Novel pose synthesis		
	PSNR ↑ /dB	SSIM ↑	LPIPS* ↓	PSNR ↑ /dB	SSIM ↑	LPIPS* ↓
w/o EHN	30.26	0.968 3	31.60	30.45	0.969 2	30.33
w/o ResCASNet	30.17	0.968 2	30.78	30.24	0.968 8	29.78
w/o $L_{texture} + L_{edge}$	30.42	0.968 8	31.42	30.51	0.969 7	30.50
Full model	30.44	0.969 1	31.07	30.53	0.969 8	29.95

大量实验结果表明, 该方法在重建精度和渲染质量上相比已有方法具有明显优势, 可以生成具有较高感知质量和结构一致性的动态人体图像, 既能在影视制作、虚拟现实、数字人建模等实际场景中得到广泛应用, 又为动态场景下的高质量重建与神经渲染提供了新的技术途径和理论依据, 具备较好的创新性和前瞻性。

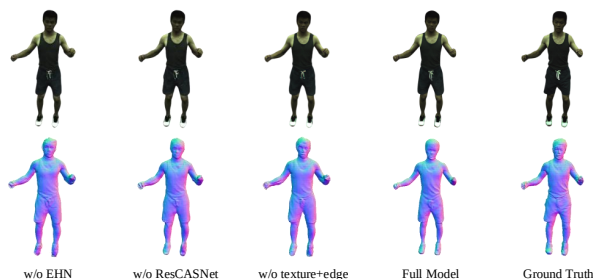


Fig. 7 Ablation experiment

图 7 消融实验



Fig. 8 Failure cases

图 8 失败案例

参考文献:

- [1] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. Nerf: representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99–106.
- [2] SUHAIL M, ESTEVES C, SIGAL L, et al. Light field neural rendering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 8269–8279.
- [3] PENG S, ZHANG Y, XU Y, et al. Neural body: implicit neural representations with structured latent codes for novel view synthesis of dynamic humans[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 9054–9063.
- [4] CHEN J, ZHANG Y, KANG D, et al. Animatable neural radiance fields from monocular RGB videos[DB/OL]. <https://arxiv.org/abs/2106.13629>.
- [5] JIANG W, YI K M, SAMEI G, et al. Neuman: neural human radiance field from a single video[C]//European Conference on Computer Vision, 2022: 402–418.
- [6] WENG C Y, CURLESS B, SRINIVASAN P P, et al. Humannerf: free-viewpoint rendering of moving people from monocular video[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 16210–16220.
- [7] YU Z, CHENG W, LIU X, et al. Monohuman: animatable human neural field from monocular video[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 16943–16953.
- [8] ROCCO I, MAKAROV I, KOKKINOS F, et al. Real-time volumetric rendering of dynamic humans[DB/OL]. <https://arxiv.org/abs/2303.11898>.
- [9] ZHENG Y, YIFAN W, WETZSTEIN G, et al. Pointavatar: deformable point-based head avatars from videos[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 21057–21067.
- [10] JIANG T, CHEN X, SONG J, et al. Instantavatar: learning avatars from monocular video in 60 seconds[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 16922–16932.
- [11] HU L, ZHANG H, ZHANG Y, et al. Gaussianavatar: towards realistic human avatar modeling from a single video via animatable 3d gaussians[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 634–644.
- [12] LEI J, WANG Y, PAVLAKOS G, et al. Gart: Gaussian articulated template models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 19876–19887.
- [13] CHEN J. GGAAvatar: reconstructing garment-separated 3D Gaussian splatting avatars from monocular video[C]//Proceedings of the 6th ACM International Conference on Multimedia in Asia, 2024: 1–7.
- [14] WEN J, ZHAO X, REN Z, et al. Gomavatar: efficient animatable human modeling from monocular video using Gaussians-on-mesh[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 2059–2069.
- [15] ZHAI Z, CHEN G, WANG W, et al. TAGA: self-supervised learning for template-free animatable gaussian articulated model[C]//Proceedings of the Computer Vision and Pattern Recognition Conference, 2025: 21159–21169.
- [16] SUHAIL M, ESTEVES C, SIGAL L, et al. Light field neural rendering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 8269–8279.
- [17] SU S Y, YU F, ZOLLHÖFER M, et al. A-nerf: articulated neural radiance fields for learning human shape, appearance, and pose[J]. Advances in Neural Information Processing Systems, 2021, 34: 12278–12291.
- [18] ROCCO I, MAKAROV I, KOKKINOS F, et al. Real-time volumetric rendering of dynamic humans[DB/OL]. <https://arxiv.org/abs/2303.11898>.
- [19] LUITEN J, KOPANAS G, LEIBE B, et al. Dynamic 3d Gaussians: tracking by persistent dynamic view synthesis[C]//2024 International Conference on 3D Vision, 2024: 800–809.
- [20] WU G, YI T, FANG J, et al. 4D Gaussian splatting for real-time dynamic scene rendering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 20310–20320.
- [21] ALLDIECK T, MAGNOR M, XU W, et al. Video based reconstruction of 3d people models[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8387–8397.
- [22] KINGMA D P, BA J. Adam: a method for stochastic optimization[DB/OL]. <https://arxiv.org/abs/1412.6980>.

(责任编辑:孙娟)