DOI:10.19651/j.cnki.emt.2417736

基于 3D 高斯溅射的人物化身重建算法*

李 斌 张文慧 项 颖 郝禄国
(广东工业大学信息工程学院 广州 510006)

摘 要:针对基于神经辐射场的隐式建模技术在个性化三维人物化身的创建中存在训练效率低和人物姿势泛化能力不足的问题。本文提出了一种结合 3D 高斯溅射技术和人体参数化模型的显示表示方法,并引入基于注意力机制的 Point Transformer 架构,该架构能够深入学习并提取每一帧中的人物姿势信息,并将其有效融合到高斯属性参数中,从而增强模型的渲染能力。在 People-Snapshot 数据集上的实验中,本文方法与当前先进方法进行了对比。定量结果显示,本文方法在 PSNR 指标上平均达到 29.53,相较于基线方法提升了 13.7%,表现出显著优势。定性评估表明,即使在人物大幅度运动的情况下,本文算法仍能有效保证渲染结果的完整性和细节表现。

关键词:3D 高斯溅射;参数化人体模型;三维人物重建;Transformer
中图分类号:TP391;TN911.73 **文献标识码**:A **国家标准学科分类代码**:520.6030

Human avatar reconstruction algorithm based on 3D Gaussian splatting

Li Bin Zhang Wenhui Xiang Ying Hao Luguo
(School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: In this paper, we address the issues of low training efficiency and insufficient pose generalization ability in personalized 3D human avatar creation using neural radiation fields based implicit modeling techniques. We propose a novel method that combines 3D Gaussian splatting with parametric human models to provide an explicit representation. Additionally, we introduce a Point Transformer architecture based on attention mechanisms. This architecture can deeply learn and extract pose information from each frame and effectively integrate it into the Gaussian attribute parameters, thereby enhancing the rendering capabilities of the model. In experiments conducted on the People-Snapshot dataset, our method is compared with current state-of-the-art methods. Quantitative results show that our approach achieves an average PSNR of 29.53, which is a 13.7% improvement over the baseline method, demonstrating a significant advantage. Qualitative evaluations indicate that even with large avatar movements, our algorithm can effectively maintain the integrity and detail of the rendering results.

Keywords: 3D Gaussian splatting; parametric human models; 3D human avatar reconstruction; Transformer

0 引 言

创建高质量的 3D 人物化身在各个领域都具有广泛的实用性,包括 AR/VR、远程会议、视频游戏和电影制作。

随着技术的发展,基于神经辐射场[1](neural radiation field,NeRF)隐式表示的方法,可以从稀疏视角视频甚至单张图像中学习并构建三维人体化身[2-3],创建成本效益高且具有照片级真实感的三维化身成为可能,利用体积渲染技术并结合姿态条件的多层感知器变形场,允许化身根据特定姿态进行控制或驱动。尽管神经辐射场展现出良好的品质,但这种建模方法遇到了训练持续时间长、姿势泛化能力有限等挑战,尤其是在面对显著的姿势变形时。张超等[4]和朱代先等[5]基于 NeRF 技术在多视图甚至稀疏视角的三维重建取得了进展,尽管基于 NeRF 的方法在视图重建中显示出卓越的性能,但其隐式表征的特性使得实现人物的实时渲染面临挑战,最先进的 InstantAvatar[6]方法引入多分辨率哈希编码提高 NeRF 的训练速度,但是未能实现实时渲染。这主要归因于所采用的固有隐式表针的特性导致的。

在 3D 人物建模领域,不同方法各有优缺点,但都面临着一些挑战。例如,如 NeRF 的隐式方法通过逆蒙皮将图像空间中的图像变换到规范空间中,其 3D 体积表示在捕

收稿日期:2024-12-26
* 基金项目:广东省自然科学基金(2022A1515010777)、广东省科技计划(2022A0505050072)项目资助

捉人体表面细节时效率较低，逆蒙皮过程还可能引入模糊的对应关系。这些问题使得隐式方法在处理动态人物的细粒度细节时表现不佳。相比之下，如基于网格建模的显式表示方法虽然能够更高效地表示人体表面，但固定的网格拓扑结构限制了其对复杂细节（如皱纹）的捕捉能力[7]。基于点的表示方法则提供了一种更灵活的解决方案。例如，PointAvatar[8]通过点云表示人体几何结构，并学习连续的变形场来实现人物驱动。这种方法不仅能够高效渲染和变形，还能处理复杂的拓扑结构和细节，如头发和配饰。然而，基于点的方法也存在挑战，例如需要大量的点来捕捉详细外观，且在表面几何形状和拓扑完整性方面可能受到限制。

近期在基于点的渲染中的研究可以通过溅射一组有限的 3D 点来快速渲染静态和动态场景[9-10]。3DGS 的高斯函数具有更好的连续性和可微性，能够更好地表达人物的几何和外观特征，从而在可驱动 3D 人体建模中提供更平滑、更自然的视觉效果。作为显式表示方法，3D 高斯通过前向蒙皮直接从规范空间映射到运动空间，避免了传统基于 NeRF 方法中逆蒙皮过程导致的一对多歧义问题，这种表示方式不仅提高了渲染效率，还增强了模型的可编辑性和灵活性[11]。

然而，3D 高斯虚拟人物模型在实现的过程中遇到了两个主要的局限性。首先，目前流行的 3D 高斯虚拟人物模型[12-13]通常只支持对身体或者面部的控制，它们并没有提供对手部细节和面部表情进行精确控制的功能。这限制了虚拟人物在需要精细手势和表情传达的场景中的应用，例如在虚拟现实交流、在线教育或者动画制作中，手部动作和面部表情是传达情感和意图的关键因素。其次，现有技术在创建 3D 高斯虚拟人物时，普遍依赖于 SMPL（skinned multi-person linear）[14]模型的顶点来初始化 3D 高斯。SMPL 模型本身仅能表示人物最基本的衣着形态，而个体的具体特征，比如服装的褶皱和个性化的面部特征，需要依赖于 3D 高斯模型自带的分裂和克隆机制来捕捉。此外，这种基于 SMPL 模板的初始化方法在处理人物自遮挡部分时，尤其是在生成新视角视图时，容易产生明显的伪影，影响虚拟人物的视觉真实性。

为了解决上述问题，本研究采用了 SMPL-X（skinned multi-person linear model eXpressive）[15]网格结合 3D 高斯的混合表示方法。在这种混合表示中，每个 3D 高斯被绑定到 SMPL-X 网格的三角面，成为网格上的一部分。使得 3D 高斯点遵循 SMPL-X 网格的拓扑结构，通过三角形面片相互连接，形成了预定义的连通性。这种连通性至关重要，因为它使得能够在保持人物整体网格结构的同时，精确捕捉和表达个体的具体特征。

此外，为了深入学习人物姿势对高斯点属性的影响，本研究引入了基于注意力机制的 Point Transformer 架构[16]。Point Transformer 中的注意力机制能够学习并捕捉每一帧中的人物姿势信息，特别关注局部邻域的特征。这一机制的输出以残差的形式被整合到高斯属性的参数中，从而进一步提升了渲染的精细度和质量。

1　提出的方法

1.1　模型整体框架

本文的目标是生成一个由身体形状、头发和服装几何形状以及人体底层骨架组成的个性化彩色网格 3D 模型。给定一个 n 帧视频序列 $(\boldsymbol{I}_t)_{t=1}^{n}$，其中包含一个固定摄像机前的单一人体，以及相应的身体姿势 $\{\boldsymbol{\theta}_t\}$，输出一个个性化的人体模型，该模型能够在新的姿势 $\{\boldsymbol{\theta}_j\}$ 下进行动画渲染。

使用 3D 高斯溅射构建可驱动的 3D 人体模型。在一个规范的 3D 人体网格上建模，将这个规范网格记为 M，由顶点 $V_C=\{v_0,v_1,\cdots,v_m\}$ 和三角形 $F=\{i_x\}$ 组成，因此 $M=(V_C,F)$。每个 3D 高斯绑定到一个三角形面的形心上。然后，通过 LBS 变形 3DGS，使其与图像空间一致，并从给定的相机进行渲染对应于真实图像。最终，通过最小化渲染的图像和输入图像之间的差异来得到每个高斯点性质。本文的网络模型如图 1 所示。

1.2　可微分高斯溅射

三维高斯溅射（3D Gaussian splatting，3DGS）是点云渲染中的一项开创性技术，它利用大量椭圆各向异性球显式地表示静态场景。该技术扩展了 EWA 体积溅射（EWA volume splatting）[17]，它便于将 3D 高斯核高效地投影到 2D 图像平面上。此外，可微分渲染优化了用于表征场景的高斯核的数量和属性。每个 3D 高斯由其在 3D 空间中的位置和协方差矩阵所表征，模型化为：

$$G(\boldsymbol{x})=\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{1}$$

其中，$\boldsymbol{x}$ 表示位置向量，$\boldsymbol{\mu}$ 代表位置，$\boldsymbol{\Sigma}$ 是高斯分布的 3D 协方差矩阵。为确保 3D 协方差矩阵 $\boldsymbol{\Sigma}$ 的半正定性，$\boldsymbol{\Sigma}$ 被分解为两个可学习组件，即表示旋转的四元数 $\boldsymbol{r}\in\mathbb{R}^4$ 和表示缩放的向量 $\boldsymbol{s}\in\mathbb{R}^3$。通过将四元数 $\boldsymbol{r}$ 转换为旋转矩阵 $\boldsymbol{R}$，将缩放向量 $\boldsymbol{s}$ 转换为缩放矩阵 $\boldsymbol{S}$，此外，每个高斯分布都具有其不透明度 $\boldsymbol{\alpha}\in[0,1]$ 和一组球谐系数（spherical harmonics，SHs），这对于重建视图依赖的颜色至关重要。相关的协方差矩阵 $\boldsymbol{\Sigma}'$ 定义为：

$$\boldsymbol{\Sigma}'=\boldsymbol{JW\Sigma W}^{\mathrm{T}}\boldsymbol{J}^{\mathrm{T}} \tag{2}$$

其中，$\boldsymbol{W}$ 是视图变换矩阵，$\boldsymbol{J}$ 是投影变换的仿射近似的雅可比矩阵。这种设置便于评估每个投影高斯对 2D 颜色和不透明度足迹的贡献。像素的颜色 C 随后通过混合所有贡献给该像素的 K 个 2D 高斯来确定：

$$C=\sum_{i\in N}c_i\alpha_i\prod_{j=1}^{i-1}(1-\alpha_j) \tag{3}$$

其中，c_i 和 α_i 分别代表高斯的视图依赖颜色和不透明度，根据从投影高斯的中心点的指数衰减进行调整。每个

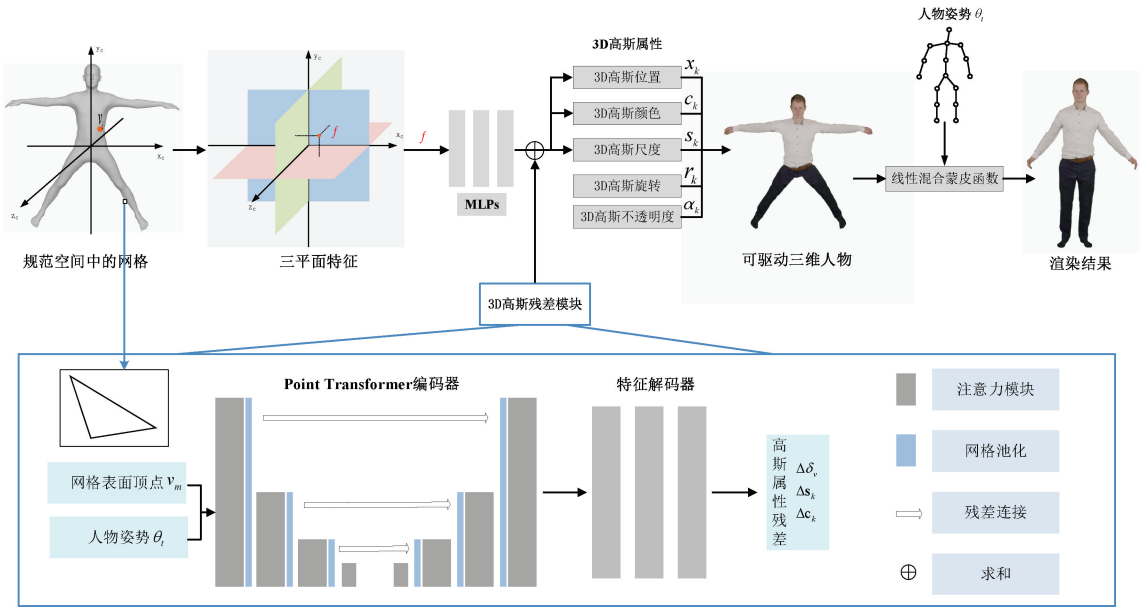

图 1　模型框架

Fig. 1　Model framework

3D 高斯的参数,如位置 $\boldsymbol{x}$、旋转 $\boldsymbol{r}$、缩放 $\boldsymbol{s}$、不透明度 $\boldsymbol{\alpha}$ 和球谐系数(SH)系数,都经过优化以确保渲染的 2D 高斯与训练图像对齐。在训练阶段,3D 高斯散射以可微分的方式高效渲染,产生 2D 图像。这一渲染过程涉及各向异性散射的混合、对它们进行排序,并使用基于瓦片的光栅化器。

在优化过程中,首先通过运动恢复结构(structure-from-motion,SfM)或随机采样来初始化 3D 高斯。接着,利用高斯的自适应密度控制(adaptive control of Gaussians)的方法来提高渲染质量,这主要涉及分裂、克隆和修剪。分裂和克隆是在位置梯度较大的 3D 高斯上进行的操作:1)当缩放矩阵的尺寸超出预设阈值时,会将高斯分裂成更细小的高斯;2)当缩放矩阵的尺寸低于阈值时,则会进行克隆。尽管分裂和克隆操作会增加高斯的数量,但修剪操作会针对性地移除那些不透明度过低或缩放尺寸过大的高斯。

1.3　人体参数化模型 SMPL-X

SMPL-X 模型是原始 SMPL 人体模型的扩展,包括面部和手部,旨在捕获更详细和富有表现力的人体变形。记 SMPL-X 人体模型为 $M(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\psi})$,其中 $\boldsymbol{\beta}$ 为人体体型参数、$\boldsymbol{\theta}$ 为人体姿态参数,$\boldsymbol{\psi}$ 为人体表情参数,具体公式可以表示为:

$$M(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\psi}) = F_{lbs}(T_p(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\psi}),J(\boldsymbol{\beta}),\boldsymbol{\theta},\boldsymbol{W}) \tag{4}$$

$$T_p(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\psi}) = T + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}) + B_e(\boldsymbol{\psi}) \tag{5}$$

其中,T 为平均形状模板,$B_s(\boldsymbol{\beta})$、$B_e(\boldsymbol{\psi})$ 和 $B_p(\boldsymbol{\theta})$ 分别对应体型、表情和姿势的混合形状。F_{lbs} 为线性混合蒙皮函数(linear blend skinning, LBS),它将考虑了体型、表情和姿势的偏移量的 $T_p(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\psi})$ 转换为目标姿势 $\boldsymbol{\theta}$,同时利用关节回归器 $J(\boldsymbol{\beta})$ 根据体型参数 $\boldsymbol{\beta}$ 来确定身体关节的位置,$\boldsymbol{W}$ 是 F_{lbs} 中使用的混合蒙皮权重矩阵。虽然 SMPL-X 模型提供了一个可驱动的人体网格,但它没有对头发和衣服进行建模。本文的方法只在初始化阶段利用 SMPL-X 网格和 LBS,并允许人体网格偏移,以模拟细节,如头发和服装。

本文提出的方法通过在规范空间中的 SMPL-X 网格中为每个顶点添加一个可学习的偏移量 $\boldsymbol{\delta}_v$,来增强头发和衣服的建模。具体而言,使用三平面来表示规范空间中的人物。为了充分利用三维人体结构的先验知识,高斯核的初始化位置被设置在 SMPL-X 网格三角面的形心,并与这些位置绑定。随后,利用多层感知器预测这些高斯点的属性,包括它们的位置、尺度和颜色等,并将蒙皮权重从 SMPL-X 模型顶点传播到最近的 3D 高斯。本文将保持 3D 高斯尺度的自由度限制为 1,并将旋转 $\boldsymbol{r}$ 设为[1,0,0,0]来将高斯限制为各向同性。为了避免网络在人物边界学习一个零的不透明度,将不透明度 $\boldsymbol{\alpha}$ 设置为 1,以保持所有 3D 高斯的可见性,这样做可以更好地推广到新的视点。

1.4　基于 Point Transformer 的人物姿势特征提取模块

在原始的 3D 高斯表示(3DGS)中,存在两个主要的不足之处。首先,直接优化 3D 高斯点属性以紧密匹配输入图像的方法虽然在细节上能够达到精确,但容易导致过拟合,因为这些高斯点会过度适应单个像素的特征。同时,3DGS 在实现过程中未能充分捕捉高斯点之间的邻近和远距离空间关系,这限制了模型在理解和模拟人物关节之间的结构相关性及其与输入数据的依赖关系方面的能力。其次,尽管每个 3D 高斯都存储着球谐系数 f 以编码视图依赖的颜色,由于训练期间只提供了一个相机视角 $\boldsymbol{d}$,导致世界空间中的视图方向是固定的,而 3D 高斯点的颜色 c

是通过与依赖于相机视角 $\boldsymbol{d}$ 的球谐基函数 γ 以及存储的球谐系数 f 进行点积来计算的的，即 $c=\langle\gamma(\boldsymbol{d}),f\rangle$ 这使得模型对未见过的、测试视图的泛化能力较差。

为了有效捕捉 3D 高斯点之间的空间关系并建模它们的交互作用，本文借鉴了 Transformer 架构的优势，特别是其在处理复杂数据关系方面的能力。采用了基于 Point Transformer V3(PTv3)架构的分层 Transformer 编码器-解码器层 f_θ。在人体姿势编码的过程中，f_θ 为每个 3D 高斯点分配了一个抽象的 Z 维特征向量 $\boldsymbol{z}_k\in\mathbb{R}^Z$，这些特征向量包含了与人体姿势相关的特征信息。

$$\{\boldsymbol{z}_k\}_{k=1}^K=f_\theta(V_C,\{\boldsymbol{\theta}_t\}) \tag{6}$$

随后，编码器将这些特征向量作为输入传递给特征解码器 g_θ。g_θ 利用这些特征向量来预测每个 3D 高斯点的属性残差，即 $\Delta\mathbf{G}_k=(\Delta\boldsymbol{x}_k,\Delta\boldsymbol{s}_k,\Delta\boldsymbol{c}_k)$，其中 k 取值为 $1\sim K$，表示所有高斯点的索引，$\Delta\boldsymbol{x}_k$ 和顶点偏移残差 $\Delta\boldsymbol{\delta}_v$ 相关。这一过程可以表示为：

$$\{\Delta\mathbf{G}_k\}_{k=1}^K=g_\theta(\{\boldsymbol{z}_k\}_{k=1}^K) \tag{7}$$

通过这种方式，能够细化每个 3D 高斯点的属性，包括位置、尺度、和颜色从而提高模型对新视角合成任务的性能和鲁棒性。

1)Point Transformer 编码器，规范空间中的高斯点和人物姿势 θ_t 作为特征通过 MLP 嵌入层输入，经过 MLP 嵌入层处理后，人物姿势特征通过 5 个 Flash 注意力[18]块和下采样网格池化层进一步处理，这些注意力块能够捕捉人物的局部特征和空间关系，接着，使用另外 4 个 Flash 注意力块和上采样网格池化层来恢复分辨率。这一过程有助于重建细节信息，提高模型对高频信息的捕捉能力。为了进一步改善梯度流动和捕捉高频细节，采用跳跃连接 MLP 模块，这些连接将中间的下采样输出映射为残差，然后将这些残差加入对应的分辨率的上采样层，增强特征的传递和融合。

2)特征解码器，提取的每个 3D 高斯点的特征 $\boldsymbol{z}$ 被送入共享的特征解码头，用于预测规范空间的相对于初始高斯点属性的残差。为了确保模型在训练初期的稳定性，对 MLP 的最后一层的权重和偏置进行了零初始化，使得初始残差特征为零，减少训练过程中的震荡。

1.5　损失函数

在训练过程中，不是直接对 3D 高斯属性进行更新，而是优化三平面的特征网格和属性预测网络。使用随机背景防止模型学习到背景颜色，并且将高斯点的不透明度 $\boldsymbol{\alpha}$ 保持为 1，保持所有的高斯点不透明，以有效的将图像空间的损失反向传播到顶点偏移参数，让模型优化网格的几何形状，而不产生大尺度的高斯。使用 SMPL-X 骨架变换将人物从规范空间驱动到图像空间，并通过可微渲染的得到渲染图像，渲染图像与真实图像进行比较，使用 L_2 损失、结构相似损失 SSIM 损失 L_{SSIM} 和感知损失 L_{LPIPS}[19] 来衡量两者之间的差异。为了获得边缘更为清晰的图像，引入 L_{Sobel} 损失。该损失的计算基于 Sobel 算子对渲染图像和实际图像处理后结果的 L_2 差异。具体而言，Sobel 算子是一种经典的边缘检测算子，通过计算图像的梯度幅度来检测边缘。它由水平和垂直的 Sobel 核组成。每个 Sobel 核可以分解为平滑核和差分核的乘积。水平和垂直的 Sobel 核 G_x，G_y 可以表示为：

$$\begin{cases}G_x=[1\quad 2\quad 1]^{\mathrm{T}}\times[+1\quad 0\quad -1]\\ G_y=[+1\quad 0\quad -1]^{\mathrm{T}}\times[1\quad 2\quad 1]\end{cases} \tag{8}$$

Sobel 算子通过差分核$[+1\quad 0\quad -1]$获取图像的梯度信息，同时利用平滑核$[1\quad 2\quad 1]$抑制噪声，从而在边缘检测中表现出色。图像 $\boldsymbol{I}$ 的高频分量可以通过 Sobel 算子处理后获得，具体如下：

$$f_{sobel}(\boldsymbol{I})=\sqrt{(G_x{}^*\boldsymbol{I})^2+(G_y{}^*\boldsymbol{I})^2} \tag{9}$$

本文的目标是渲染高保真的人物图像。为此，在 Sobel 边缘空间中直接最小化渲染图像与目标图像之间的距离，这样做可以确保在训练过程中对图像的高频区域进行额外的优化，从而提高渲染图像的清晰度。因此，本文提出的 Sobel 损失函数可以表示为：

$$L_{sobel}(\boldsymbol{R},\boldsymbol{T})=\frac{1}{N}\left|f_{sobel}(\boldsymbol{R})-f_{sobel}(\boldsymbol{T})\right| \tag{10}$$

其中，$\boldsymbol{R}$ 和 $\boldsymbol{T}$ 分别表示渲染图像和目标图像。损失值通过总像素数量 N 进行归一化。

虽然 L2、SSIM、LPIPS 损失函数确保了图像的整体准确性，但 Sobel 损失函数为网络提供了高频重建的指导，从而生成清晰锐利的图像。最终，本文提出的损失函数为：

$$L=\lambda_{\text{L2}}L_2+\lambda_{\text{LPIPS}}L_{\text{LPIPS}}+\lambda_{\text{SSIM}}L_{\text{SSIM}}+\lambda_{\text{Sobel}}L_{\text{Sobel}} \tag{11}$$

其中，λ_{L2}、λ_{LPIPS}、λ_{SSIM} 和 λ_{Sobel} 分别代表 L2 损失函数、LPIPS 损失函数、SSIM 损失函数和 Sobel 损失函数的权重。这些权重用于平衡不同损失函数在整体损失计算中的影响，以优化模型的性能。

2　实验结果与分析

2.1　实验细节

实验环境为 Ubuntu20.04，Python3.9.18，Pytorch1.13.0，CUDA11.6，硬件环境为 CPU 配置为 Intel Xeon Silver 4120R，GPU 为 RTX A5000，128 GB RAM，所有实验均在相同配置进行。

在数据预处理阶段，首先用 SAM[20] 算法提取人体前景掩码，为了减少边界处的不自然，对人体蒙版图像放大到 0～255 的范围，应用一个 5×5 的矩形结构元素进行腐蚀操作，减少边界伪影，接着使用 3×3 的均值滤波器进一步平滑人体蒙版图像。在处理每一帧的人物姿势和体型参数时，单目人体姿态估计算法并不能直接提供精确的 SMPL-X 参数。为了解决这一问题，本文采用 HybrIK-X[21] 模型获取 SMPL-X 参数，并使用 MMPose 得到的人物 2D 关键点和人物掩膜板作为限制条件来优化每一帧的

SMPL-X 参数,最后通过 Savitzky-Golay 平滑来减少相邻帧之间的姿势差异,对于同一个人物的体型参数,由于它们是固定的,本文采用了估计出的体型参数的均值来进行处理。

模型训练阶段,初始化一个全零的可学习三平面 $\boldsymbol{T} \in \mathbb{R}^{3\times C\times H\times W}$,其中 $C=32$ 代表通道维度,$H=128$ 和 $W=128$ 分别代表三平面的高度和宽度。在规范空间中使用星形姿势的 SMPL-X 进行初始化,并用 Pytorch3D 提供的网格细分函数对该网格进行上采样得到 $\mathcal{M}$。通过将 $\mathcal{M}$ 正交投影到每个平面上并执行双线性插值,从三平面中提取每个顶点的特征。

提取后的特征随后被拼接成一个新的特征集合,记为 $\boldsymbol{F} \in \mathbb{R}^{N\times 96}$。将 $\boldsymbol{F}$ 传递给两个多层感知机,它们预测规范空间中的 SMPL-X 网格顶点的 3D 偏移量 $\boldsymbol{\delta}_v \in \mathbb{R}^{N\times 3}$ 和尺度 $\boldsymbol{S} \in \mathbb{R}^{N\times 1}$ 以及 RGB 值 $C \in \mathbb{R}^{N\times 3}$。高斯点的颜色由一个宽度为 64 的两层多层感知机(MLP)来确定,激活函数采用 GELU(高斯误差线性单元)。此外,网格顶点的偏移量和高斯点的尺度则通过另一个宽度为 128 的两层 MLP 来计算,同样使用 GELU 作为激活函数。在训练初期,为防止 MLP 不稳定生成过大的高斯点占用过多显存,使用固定尺度参数,以保持模型稳定并确保训练顺利进行。

Point Transformer 编码器以一个 MLP 嵌入层开始,输入规范空间中的 3D 高斯点位置,并嵌入人物姿势的特征,随后是 5 个下采样和 4 个上采样阶段,最终产生维数为 Z=96 的特征。下采样阶段包含(2,2,2,6,2)个注意力块,并且具有(64,96,128,256, 512)的隐藏维度。除了第 1 个阶段外,每个阶段后面都跟着一个下采样网格池化层。上采样阶段由(2,2,2,2)个注意力块组成,隐藏维度为(256, 128, 96,96)。除了第 1 个阶段外,每个阶段前面都先有一个上采样网格池化层。使用 50 的网格分辨率来体素化高斯点,网格池化层的步长设置为(1,2,2,2)。特征解码器由 5 个独立的 MLP 分支组成,它们负责预测网格偏移量、尺度和颜色的残差。每个 MLP 分支由两个宽度为 128 的线性层组成,除了最后一层外,所有层都使用 ReLU 激活函数。对残差均值应用 Tanh 激活函数,以将其规范化到[0,1]范围内。

使用 Adam 优化器进行训练,初始学习率设置为 1×10^{-3},训练过程学习率阶梯型衰减。损失函数的权重 λ_{L2}、λ_{LPIPS}、λ_{SSIM} 和 λ_{Sobel} 分别设置为 1.0、0.01、0.2 和 1.0。

优化完成后,可直接显式构建 3D 高斯并用 LBS 驱动人物,无需对三平面和 MLPs 进行推理来渲染新姿势,相比隐式神经场表示方法有明显优势。

2.2　数据集和评价指标

为了验证本章提出的方法可以从单目视频中重建动态人体,选择文献[22]提出的 People-Snapshot 数据集,该数据集包含了在自然场景中捕获的单目 RGB 视频序列,视频中的参与者在固定相机前大致处于的 A 形姿势并进行转动,使用数据集中的人物转动 2 圈作为训练帧,剩余的帧用于测试模型在新颖视图合成上的性能。对数据集中的所有实验均采用了统一的 SMPL 模型参数。为了适应 SMPL-X 模型,利用 SMPL 官方网站提供的转换工具,将原有的 SMPL 参数转换成 SMPL-X 所需的参数形式。此外,采用了 HybrIK-X 模型来精细调整手部和头部的姿态,目的是使模型的姿态更加贴近于实际观测到的姿势。以确保比较的公正性。

指标包括峰值信噪比(PSNR)、结构相似性指数(SSIM)以及感知损失指标(LPIPS)。PSNR 和 SSIM 是两个经典的图像质量评价指标,它们分别衡量图像的信号保真度和结构相似性。PSNR 值越高,表明重建图像与原始图像在像素级别上越接近;SSIM 值越高,意味着两者在亮度、对比度和结构上越相似。而 LPIPS 则是一种基于深度学习特征的感知质量评价指标,它通过比较重建图像与真实图像在神经网络特征层面的差异来评估视觉质量,LPIPS 值越低,表示重建图像在人类视觉感知上越接近真实图像。

2.3　对比实验

为了评估本文重建出的三维人物模型的外观质量,本文采用了 People-Snapshot 数据集中的测试帧姿势来驱动模型并进行渲染。通过比较渲染所得图像与实际图像,量化两者之间的差异性。

对所提出的人物化身重建方法进行了全面的定量评估对比。选择了几个具有代表性的算法作为对比基准,包括基于 NeRF 的隐式表示人物化身重建算法 HumanNeRF[23] 和 InstantAvatar,以及最近先进的基于 3DGS 的显示表示算法的 GART 和 GaussianAvatar 进行定量评估对比。表 1 中的数据展示了本文提出的方法在四个人物序列的数据集中,针对视图合成任务的表现与其他基线方法的对比。结果显示,本文的方法在峰值信噪比上超越了所有其他方法。尽管 InstantAvatar 方法在感知损失指标和结构相似性指数上获得了较高的评分,但它并不具备与基于 3DGS 的方法相媲美的实时渲染能力。

在定量分析的基础上,进一步通过可视化方法展示了本文方法与 GART 和 GaussianAvatar 这两种先进算法的性能对比,如图 2 所示。具体来说,针对 People-Snapshot 数据集中的 female4-casual 和 male-4-casual 两个视频序列,在星形姿势,以及测试视图下的渲染效果进行了详细的对比分析。在正面测试视图中,GART 和 GaussianAvatar 都能取得较好的渲染质量,但在人物侧面的渲染质量上,GART 缺乏衣服上的褶皱细节,而 GaussianAvatar 方法产生了一条绿色的线,这显然是不合理的。在星形姿势下的高斯人物进行了可视化渲染,GART 的方法在人物自遮挡区域产生大量颜色、尺度及位置不合理的高斯点,严重影响重建出的人物在大幅度运动时的渲染结果。

表 1　People-Snapshot 数据集的定性比较

Table 1　Qualitative comparison on the People-Snapshot dataset

人物序列	female-3-casual			female-4-casual			male-3-casual			male-4-casual		
方法	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
HumanNeRF	24.46	0.951 6	0.026 9	27.07	0.961 5	**0.017 8**	26.90	0.960 5	0.018 1	25.50	0.939 7	0.035 7
InstanAvatar	27.66	**0.970 9**	**0.021 0**	29.11	0.968 3	0.021 2	29.53	0.971 6	**0.011 5**	27.67	0.962 6	0.030 7
GART	26.29	0.966 8	0.049 8	29.21	**0.973 0**	0.038 0	30.39	**0.977 7**	0.038 3	27.63	**0.967 3**	0.060 1
GaussianAvatar	24.80	0.960 9	0.031 0	26.16	0.961 0	0.022 7	27.71	0.971 3	0.021 8	25.09	0.961 4	**0.030 6**
本文方法	**29.84**	0.961 2	0.047 9	**29.39**	0.963 6	0.040 9	**30.95**	0.966 3	0.031 3	**27.92**	0.952 0	0.049 8

注：加粗数值为该列的最优值。

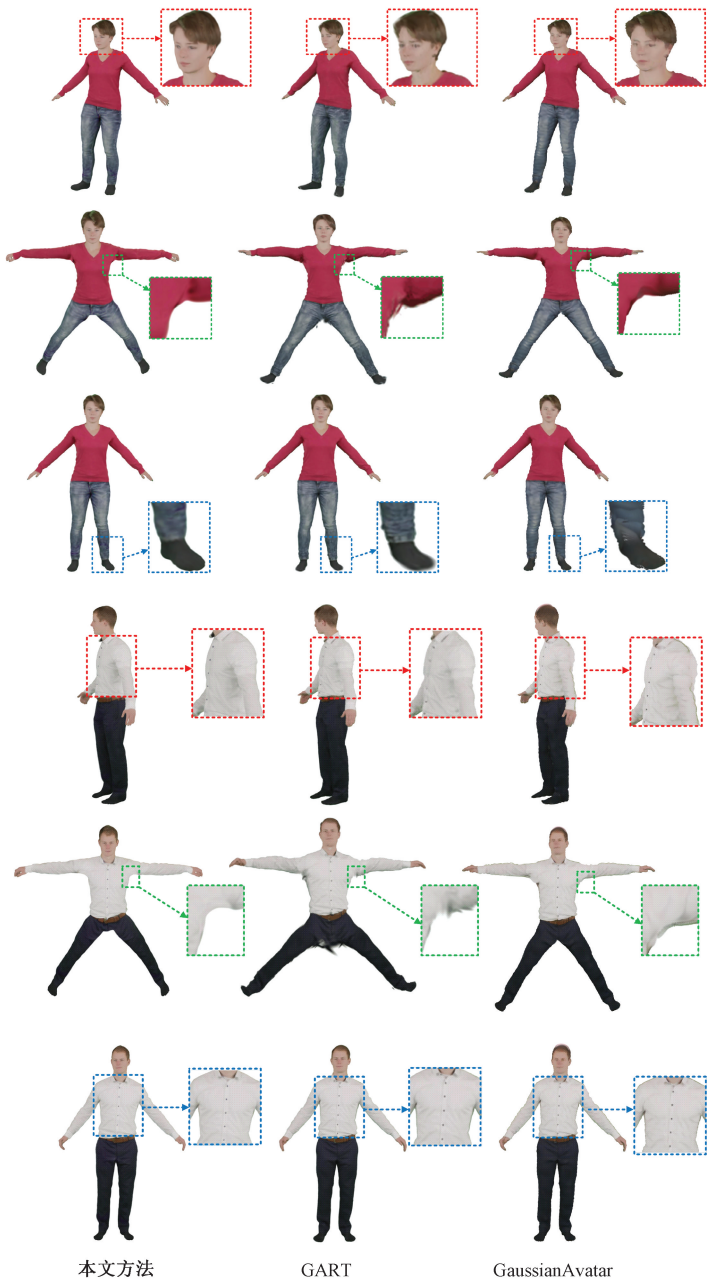

图 2　图像渲染结果

Fig. 2　Rendering results

相比之下,本文通过引入姿势特征提取模块,增强了对于人物动态姿势的捕捉能力以及克服单一视角下的高斯点颜色建模不足的问题。在颜色尺度的改善方面,采用了残差的形式来调整每个高斯点的颜色。这种方法允许在保持高斯点原有颜色信息的同时,对其进行微调,并且可以合理推理相邻位置的高斯点的颜色,从而提高了渲染的真实性和准确性。

2.4　消融实验

1)损失函数权重消融实验

为了验证不同损失函数权重的必要性,进行了消融实验。表 2 展示了不同权重在 PSNR、SSIM 和 LPIPS 3 个指标上的表现。结果表明,每种损失函数各有优势,但单独使用存在局限性。当所有损失函数权重设置为 1 时模型性能表现不佳。LPIPS 损失关注高级语义特征,实验发现,当其权重权重较小时,可以在保留 PSNR 和 SSIM 指标的情况下,在 LPIPS 指标上性能有所提升。这表明 LPIPS 损失能够引导模型生成更符合人类视觉感知的结果,但权重需适度调小,以避免其过度主导训练过程,同时保持像素级别的局部最优。同样,SSIM 损失注重图像的结构和纹理细节,权重需要调小以平衡优化过程的影响。Sobel 损失通过引入边缘信息,增强了模型对细节的捕捉能力,帮助 L2 损失避免陷入局部最优解。

表 2　不同损失函数权重的消融实验定性结果

Table 2　Qualitative results of ablation study on different loss function weights

λ_{L2}	λ_{LPIPS}	λ_{SSIM}	λ_{Sobel}	PSNR ↑	SSIM ↑	LPIPS ↓
1	0	0	0	30.48	0.947 4	0.053 8
0	1	0	0	23.72	0.932 7	0.036 4
0	0	1	0	30.08	0.956 4	0.044 0
1	1	1	1	29.99	0.953 4	0.032 0
1	0	0	1	30.52	0.953 0	0.041 0
1	0.01	0	0	30.33	0.950 3	0.038 9
1	0.2	0	0	29.38	0.946 9	0.033 1
1	0.01	0.2	0	30.74	0.956 4	0.034 1
1	0.2	0.2	0	29.97	0.952 7	0.041 0
1	0.01	0.2	0.01	30.71	0.957 9	0.034 0
1	0.01	0.2	0.1	30.82	0.961 2	0.031 7
1	0.01	0.2	1	**30.95**	**0.966 3**	**0.031 3**

注:加粗数值为该列的最优值。

从表 2 的 Sobel 损失函数权重变化结果看分析中,Sobel 损失函数权重为 0.01 时,PSNR 指标相较于不使用 Sobel 损失后出现下降,SSIM 和 LPIPS 指标略微提升,当权重增加到 1 时,模型学习到了如何同时优化这两种损失,Sobel 损失可以帮助模型学习图像的边缘和结构信息,这能间接帮助模型在像素级上做出更好的预测,结果是 PSNR 指标的提升。

通过实验,合理设置各损失函数的权重能够显著提升模型的最终性能。LPIPS 损失有助于模型在感知层面的收敛,SSIM 损失有助于结构和纹理细节的收敛,而 Sobel 损失则通过增强细节捕捉能力,帮助模型避免陷入局部最优解。这种多损失函数的优化策略能够有效平衡不同损失函数的优势,避免单一损失函数的局限性,从而实现更优的训练效果。

2)模型其他组件消融实验

为了进一步验证本文方法的有效性,对整体网络进行了消融实验和定性定量结果实验分析。消融实验定量结果如表 3 所示,在 People-Snapshot 数据集上的 male-3-casual 视频序列进行消融实验,以评估各个组件对整体性能的贡献。PTv3 代表 Point Transformer 编码器的人物姿势特征提取模块。

表 3　消融实验的定量评估

Table 3　Quantitative evaluation of ablation study

方法	PSNR ↑	SSIM ↑	LPIPS ↓
w/o PTv3	28.56	0.935 0	0.068 1
w/o Sobel loss	30.74	0.956 4	0.034 1
w/o triplane	28.01	0.930 2	0.084 3
完整模型	**30.95**	**0.966 3**	**0.031 3**

注:加粗数值为该列的最优值。

如图 3 所示,在不同消融实验条件下,本研究方法的人物在测试视图的渲染效果的细节。去除 PTv3 模块后,高斯点的颜色没有得到人物运动姿势的补偿,总体结果偏向于模糊,去除 Sobel 损失后,模型的细节捕捉能力明显减少,在框选的区域中,衣服上的图像出现失真,因为 Sobel 算子能够有效监督渲染结果的边缘信息,对于增强人物轮廓和细节特征具有重要作用。此外,在模型训练时引入可学习的三平面,高斯点的颜色信息可以帮助高斯点分布在规范空间中正确的位置,去除三平面的编码方式后会影响最终的渲染质量。

图 3　消融实验定性评估结果

Fig. 3　Qualitative assessment results of ablation study

3　结　　论

本文提出了一种从单目视频序列中重建可驱动的三

维人物模型的方法，该方法无需依赖于 3D 监督数据。采用了 SMPL-X 网格绑定 3D 高斯，并通过 2 D 监督进行训练。此外，本文将人物姿势特征整合到 Transformer 模型中，赋予模型更灵活的数据理解能力，以残差形式补充原始 3DGS 在视角变化上的不足，有效解决了人物自遮挡问题，减少了不合理的高斯分布，显著降低了渲染伪影，并提升了人物表面的外观质量。与现有模型相比，本文方法在定量和定性评估中均展现出优越的性能。

本研究提出的方法确实存在一些限制。目前，该方法依赖于参数化变形模型来实现人物的关节活动，但在模拟如头发和宽松衣物这类不受身体关节直接驱动的部分的物理运动方面存在不足。在未来的研究中，计划探索一种组合式的表征方法，以便更准确地模拟人物各个部分的动态变化。

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] JIANG B, HONG Y, BAO H, et al. Selfrecon: Self reconstruction your digital avatar from monocular video[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 5605-5615.

[3] WENG ZH ZH, WANG Z, YEUNG S. Zeroavatar: Zero-shot 3D avatar generation from a single image[J]. ArXiv preprint arXiv:2305.16411, 2023.

[4] 张超，袁亮，肖文东，等. 基于神经辐射场的稀疏视角三维重建方法[J]. 电子测量技术, 2024, 47(20): 159-166.
ZHANG CH, YUAN L, XIAO W D, et al. Sparse perspective 3D reconstruction method based on neural radiance fields [J]. Electronic Measurement Technology, 2024,47(20):159-166.

[5] 朱代先，孔浩然，秋强，等. 注意力机制与神经渲染的多视图三维重建算法[J]. 电子测量技术, 2024,47(5): 158-166.
ZHU D X, KONG H R, QIU Q, et al. Attention mechanismand neural rendering algorithm for multi-view 3D reconstruction[J]. Electronic Measurement Technology, 2024,47(5):158-166.

[6] JIANG T J, CHEN X, SONG J, et al. Instantavatar: Learning avatars from monocular video in 60 seconds[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 16922-16932.

[7] ZHAO H, ZHANG J S, LAI Y K, et al. High-fidelity human avatars from a single rgb camera[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 15904-15913.

[8] ZHENG Y F, YIFAN W, WETZSTEIN G, et al. Pointavatar: Deformable point-based head avatars from videos[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 21057-21067.

[9] KERBL B, KOPANAS G, LEIMKÜHLER T, et al. 3D Gaussian splatting for real-time radiance field rendering[J]. ACM Transactions Graphics, 2023, 42(4): 139:1-14.

[10] WU G, YI T, FANG J, et al. 4D Gaussian splatting for real-time dynamic scene rendering[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 20310-20320.

[11] CHEN X, ZHENG Y, BLACK M J, et al. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes[C]. IEEE/CVF International Conference on Computer Vision, 2021: 11594-11604.

[12] LEI J H, WANG Y F, PAVLAKOS G, et al. Gart: Gaussian articulated template models[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 19876-19887.

[13] HU L X, ZHANG H W, ZHANG Y X, et al. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d Gaussians[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 634-644.

[14] LOPER M, MAHMOOD N, ROMERO J, et al. SMPL: A skinned multi-person linear model[M]. Seminal Graphics Papers: Pushing the Boundaries, 2023: 851-866.

[15] PAVLAKOS G, CHOUTAS V, GHORBANI N, et al. Expressive body capture: 3D hands, face, and body from a single image[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 10975-10985.

[16] WU X Y, JIANG L, WANG P S, et al. Point Transformer V3: Simpler faster stronger[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 4840-4851.

[17] ZWICKER M, PFISTER H, VAN BAAR J, et al. EWA volume splatting [C]. Visualization, IEEE, 2001: 29-538.

[18] DAO T, FU D, ERMON S, et al. Flashattention: Fast and memory-efficient exact attention with io-awareness [J]. Advances in Neural Information Processing Systems, 2022, 35: 16344-16359.

[19] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a

perceptual metric[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 586-595.

[20] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything [C]. IEEE/CVF International Conference on Computer Vision, 2023: 4015-4026.

[21] LI J F, BIAN S Y, XU CH, et al. HybrIK-X: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(4): 2754-2769.

[22] ALLDIECK T, MAGNOR M, XU W P, et al. Detailed human avatars from monocular video[C]. 2018 International Conference on 3D Vision(3DV). IEEE, 2018: 98-109.

[23] WENG C Y, CURLESS B, SRINIVASAN P P, et al. Humannerf: Free-viewpoint rendering of moving people from monocular video [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 16210-16220.

作者简介

李斌,硕士研究生,主要研究方向为三维人体重建、单目人体姿态估计。

E-mail:243202019@qq.com

张文慧,硕士,讲师,主要研究方向为三维重建。

E-mail: zwh83@gdut.edu.cn

项颖(通信作者),博士,教授,博士生导师,主要研究方向为 3D 视觉、AR\VR 视觉技术。

E-mail: xiangy@gdut.edu.cn

郝禄国,博士,讲师,主要研究方向为三维重建、视频编解码。

E-mail: haolg@gdut.edu.cn