arXiv:2601.05853v1 [cs.CV] 9 Jan 2026

LayerGS：通过 2D 高斯泼溅实现分层 3D 人体化身的分解与图像修复

Yinghan Xu、John Dingliana

都柏林圣三一学院，爱尔兰都柏林
{yixu,john.dingliana}@tcd.ie

摘要。 我们提出了一种新颖的框架，可将任意姿势的人体分解为可动画化的多层 3D 人体化身，并将身体与服装分离。传统的单层重建方法会将服装固定于单一身份，而以往的多层方法则难以处理遮挡区域。我们通过将每一层编码为一组二维高斯，实现精确的几何重建与逼真渲染，并利用预训练 2D 扩散模型通过分数蒸馏采样（SDS）修复隐藏区域，从而克服这两项局限。我们的三阶段训练策略首先通过单层重建恢复粗略的规范服装，随后进行多层训练，联合恢复内层身体与外层服装细节。在两个 3D 人体基准数据集（4D-Dress、Thuman2.0）上的实验表明，与以往最先进的方法相比，我们的方法在渲染质量以及层分解与重组方面表现更优，可在新颖视点和姿势下实现逼真的虚拟试穿，并推动面向沉浸式应用的高保真 3D 人体资产的实用化创建。我们的代码见 https://github.com/RockyXu66/LayerGS。

关键词： 3D 人体 · 虚拟试穿 · 高斯泼溅。

1 引言

创建逼真的数字化人体表示对于提升扩展现实应用的沉浸感至关重要。一项热门应用是逼真的三维虚拟试穿，但由于 3D 服装资产有限，且难以采集合适的 3D 人体模型，这项应用面临诸多挑战。目前已有多种方法利用单张图像 [24], 单目 [8], 多视图 [13], 或 RGBD 视频输入 [10]来重建高质量的人体。然而，大多数现有工作都聚焦于单层表示，不会将服装与内部身体分离，因此限制了其在虚拟试穿中的应用。由于采集输入中缺少信息，被遮挡的身体部位往往会被忽略；而服装的非线性材料属性和动态变形也使其难以准确建模。为了解决这一问题，同时避免深入研究服装建模，Kim 等人提出了 GALA [12], ，这是一种框架，利用二维扩散模型中的通用知识作为几何先验

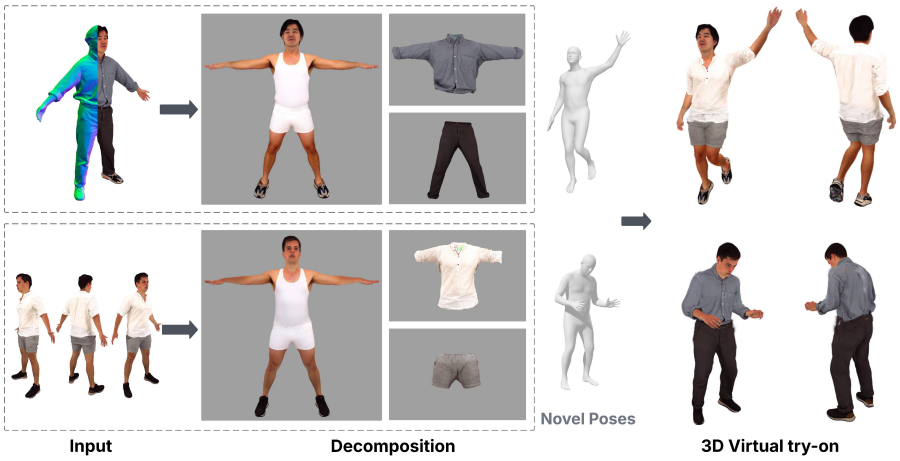

图 1：给定三维人体扫描或静态人物的多视角图像，我们的框架会将主体分解并进行图像修复，得到多个规范高斯层，用于动画和三维虚拟试穿。

以及外观先验，用于多层人体修复与生成。给定静态的单层三维人体扫描，GALA 可以训练并提取多层模型，从而分离身体与服装。然而，分解质量在很大程度上取决于原始三维扫描。虽然可以从单张图像 [24], 估计出三维人体模型，但要获得高质量扫描，可能需要一百多台相机 [29], ，这限制了该方法的潜在应用范围。

本文提出了一种框架，可从静态姿势下的多视角 RGB 图像创建多层人体表示。与以往方法不同，我们的方法不局限于仅从三维扫描重建多层人体，因此能够利用单目视频重建逼真的静态人体。与 GALA 类似，我们使用二维扩散模型作为几何和外观先验；但不同于 GALA 的顶点颜色和 DMTet [27] 表示，我们利用 2D 高斯泼溅（2DGS）的渲染和建模优势[9]。实验表明，我们的方法在分解和重新组合后能够提升图像质量，从而实现不同主体间的服装互换，并为人体被遮挡部位生成逼真的内部身体和外层服装外观。

总之，我们的主要贡献如下：

– 我们提出了一种框架，可将逼真的 3D 人体分解为多个 3D 层，效果优于现有的最先进方法。– 我们不依赖三维扫描，而是利用 2D 高斯泼溅，仅根据多视角 RGB 图像重建并修复三维人体模型。– 此外，我们使用 3D 网格显式建模这些分解后的层，使其适用于三维虚拟试穿应用。

2 相关工作

2.1 单层人体建模

PIFu [24] 通过训练三维隐式场，从单张二维图像中捕获几何信息，以此重建人体。随后使用移动立方体算法提取单层几何，并预测顶点颜色。SIFU [37] 引入了三维一致的纹理细化，以生成高质量的三维着装人体模型。HumanSplat [20] 结合二维多视图扩散模型和人体结构先验，从单张图像生成三维高斯人体。IP-Net [1] 以点云为输入，生成带衣物的单层配准可动画化虚拟化身。一些方法 [3] [16] [26] 以一系列人体三维扫描为输入，生成单层姿态相关的服装形变。 [16] 专注于单层几何重建与表示，而 X-Avatar [26]则将其扩展到带纹理的几何。然而，所有这些方法都将人体虚拟化身视为单层，忽略了衣物下方的几何和纹理信息。这些方法也限制了虚拟试穿应用以及不同主体之间的服装互换能力。

2.2 多层人体建模

采用解耦服装层对人体进行建模，对于可编辑虚拟化身至关重要。SMPLicit [5] 在给定二维图像的情况下，将人体服装拟合到 SMPL [15]参数化模型之上，但缺少纹理生成能力。CaPhy [28] 在学习可动画化服装的物理属性的同时，根据一组三维扫描生成身体和服装模板。LayGA [14] 以多视图视频为输入，分别对身体和服装建模，并根据时序观测学习服装属性。然而，获取如此密集的视频数据成本高昂，而且该方法无法准确重建被遮挡的内层身体，导致虚拟化身穿着比原始服装更短的衣物时出现伪影。DeClothH [18] 使用专门设计的扩散模型，从单张图像重建多层三维着装人体，但难以保持高保真度的三维一致性。LayerAvatar [36] 提出了一种前馈扩散模型，利用分层 UV 特征平面表示生成解耦的着装虚拟化身。尽管速度很快，LayerAvatar 的重点是从潜在噪声生成新的身份，而不是根据特定的用户引导编辑现有虚拟化身。

2.3 虚拟试穿

一些研究利用大规模预训练文生图模型的能力来进行二维图像修复，例如稳定扩散 [23]。ControlNet-图像修复 [34] 可根据描述提示或二维人体姿态生成逼真的二维人体图像。为专门提升服装相关任务的效果， [31][38] 专注于虚拟试穿和二维服装编辑。然而，这些方法仍局限于固定视点，缺乏沉浸式三维支持。

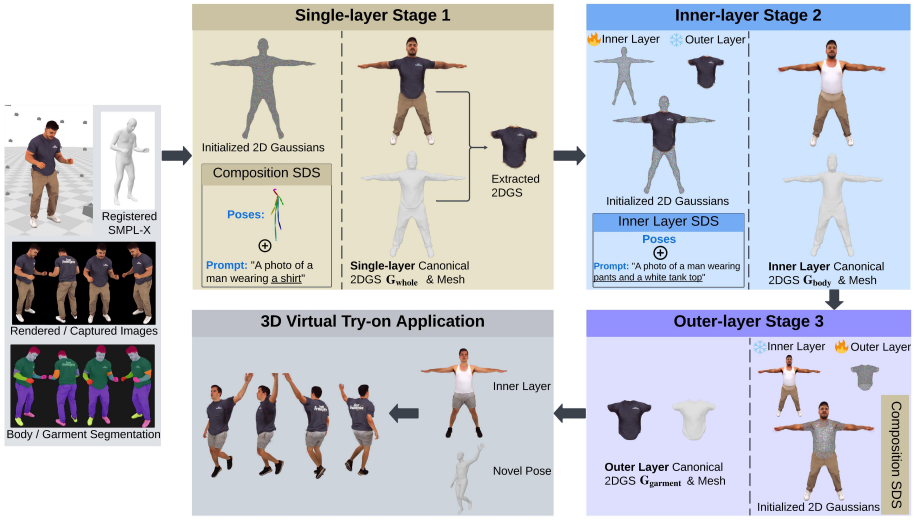

图 2：**概述**。（左）输入的多视图 RGB 图像和分割掩码。在**阶段 1**，我们通过结合组合式 SDS 损失，学习单层规范化 2D 高斯集合，并通过分割提取粗略的服装网格。在**阶段 2**，我们分解并优化内部高斯层（身体和内层服装），同时保持外层固定。在**阶段 3**，我们在冻结内层的情况下细化外层服装层。随后将得到的两层高斯附着到网格上，以支持在新姿势下实现高保真三维虚拟试穿。

研究人员将二维图像修复拓展至三维：以二维扩散模型作为先验，并将其与 3D 表示相结合。例如，Instruct-NRF2NeRF [6]支持基于指令编辑神经辐射场 [17]。GaussianEditor [4] 利用 2D 扩散模型，为 3DGS 资产提供 3D 编辑能力。VTON360 [7] 将 2D 虚拟试穿重新表述为多视角 2D 图像修复任务。该方法先将 3D 人体投影到二维视图中，再应用以服装图像为条件的基于扩散模型的图像修复器，最后通过高斯泼溅将编辑后的视图重建回三维。尽管该编辑流程对特定视图效果良好，但缺乏显式的层分解，往往会导致不一致。

3 方法

我们通过三阶段流程，从静态人体采集的 2D 图像中重建高斯人体数字化身层 G_{body} 和 $G_{garment}$，从而能够在新视角和新姿势下获得逼真的渲染结果（图 2）。在阶段 1 中，我们训练一个单层规范化人体数字化身，用于粗略重建外层服装并进行图像修复。在阶段 2 中，我们冻结阶段 1 得到的外层服装，并优化规范化内部层。在阶段 3 中，我们冻结规范化内部层，并训练经过细化的外层，以解决由层间重叠导致的渲染伪影

。最后，我们从每一层提取一个 3D 网格，并将相应的高斯附着到该网格上，从而能够通过在新姿势下重组 3D 资产，实现无缝的三维虚拟试穿。阶段 1–3 的详细内容见第 3.1 节、第 3.2 节和第 3.3 节。

3.1 单层标准化头像重建与图像修复

与需要静态人体及其 3D 网格的 GALA [12], 不同，我们的方法只需要 2D 多视角图像。这些图像可以通过单目视频采集：摄像机围绕随机姿态的人体主体做球形运动；也可以从围绕三维人体扫描布置的球形虚拟摄像机视点渲染得到。SMPL(-X) 参数已知或经过拟合。我们从标准姿态的 SMPL(-X) 网格表面采样 30k 个点，并初始化 2D 高斯。根据 SMPL(-X) 参数预先计算 3D 网格线性混合蒙皮（LBS）权重，并通过插值计算 2D 高斯的蒙皮权重。

在此阶段，人体的姿态化 2D 高斯表示会被渲染为单层。2D 高斯的属性 $\mathbf{G_{whole}}$ 可以使用 2D 真值图像进行优化。RGB 图像损失函数 $\mathcal{L}^{rgb}$ 如下：

$$\mathcal{L}^{rgb}_{whole} = (1-\lambda_c)\mathcal{L}^{L1}_{whole} + \lambda_c \mathcal{L}^{\text{D-SSIM}}_{whole} \tag{1}$$

其中，损失 $\mathcal{L}^{L1}_{whole}$ 和 $\mathcal{L}^{\text{D-SSIM}}_{whole}$ 分别表示由单层渲染的 RGB 图像 g（$\mathbf{G_{whole}}$, $\mathbf{P}$）的 $L1$ 损失和 D-SSIM 损失，权重系数为 λ_c。$\mathbf{P}$ 表示相机模型。

我们沿用原始 2DGS 工作中的两种正则化，以获得更准确的几何结构。深度失真损失 $\mathcal{L}_d$ 用于减轻高斯沿交集的扩散。法线一致性损失 $\mathcal{L}_n$ 用于确保相交的 2D 泼溅共享同一法线，该法线与实际表面对齐。深度失真损失如下：

$$\mathcal{L}^{d}_{whole} = \sum_{i,j} \omega_i \omega_j |z_i - z_j| \tag{2}$$

其中，ω_i 是第 i 个交集的混合权重，z_i 是交集点的深度。法线一致性损失如下：

$$\mathcal{L}^{n}_{whole} = \sum_{i} \omega_i (1 - \mathrm{n}_i^T \mathrm{N}) \tag{3}$$

其中，ω_i 是第 i 个交集的混合权重，n_i 表示 2D 泼溅的法线，N 是通过深度图的梯度估计得到的法线。

为了与 GALA 进行公平比较，我们使用相同的扩散模型——稳定扩散 V1.5 [23], 进行分数蒸馏采样（SDS）损失 [21]。扩散模型通过 ControlNet [34] 和 OpenPose [2]进行增强。SDS 损失作用于渲染图像，从而能够对单层 $\mathbf{G}_{\mathbf{whole}}$中的遮挡区域进行图像修复。每个主体的 SDS 文本提示均由人工指定（例如，“一张男子穿着 T 恤的照片”）。遮挡区域中的高斯

通过致密化和剪枝进行优化，尤其是在上臂与躯干之间的区域。SDS 引导损失的梯度如下：

$$\nabla_{\mathbf{G}_{\text{whole}}}\mathcal{L}^{\text{SDS}}_{\text{whole}} = \mathbb{E}_{\epsilon,t}\left[w_t\left(\epsilon_\phi(\mathbf{z}_t^{\text{whole}}; y_{\text{whole}}, t) - \epsilon\right)\frac{\partial \mathbf{x}_{\mathbf{whole}}}{\partial \mathbf{G}_{\mathbf{whole}}}\right] \quad (4)$$

其中，$\mathbf{x}_{\mathbf{whole}}$ 表示使用单层参数$\mathbf{G}_{\mathbf{whole}}$渲染得到的图像。变量$\mathbf{z}_t^{\text{whole}}$指相应的噪声图像。文本嵌入 y_{whole} 由描述主体穿着该服装的文本提示生成。阶段 1 的总损失是四项损失的加权和。

获得完整的单层规范化化身后，我们采用 2DGS 中的相同技术提取网格。具体而言，我们渲染训练视角的深度图，并利用截断符号距离融合（TSDF）将其融合为网格。接着，我们为该规范网格渲染多个视角的二维图像，并通过二维分割和三维投票得到分割后的规范化服装网格。假设 2D 高斯分布在网格表面上，我们通过设置从高斯中心位置到网格的最近距离阈值$\alpha = 0.015$，得到粗略的规范化外层服装。

3.2 内层规范化化身重建与修补

我们通过从 SMPL(- X) 网格表面采样点来初始化规范化内部层。在训练过程中，我们冻结上一步得到的外层服装。除了针对内层 2D 高斯的 RGB 图像损失、法线一致性损失和分数蒸馏采样（SDS）损失外，我们还对身体、服装和背景应用二维分割损失。每一层都使用二维分割标签 $M \in \{0,1\}^{H\times W}$。我们根据二维分割标签，将内层的深度失真损失划分为可见区域和遮挡区域。可见区域按常规方式计算；对于遮挡区域，我们创建一个虚拟 SMPL(-X) 二维高斯集合，并将其与可优化的内层一同渲染，以引导深度信息。这些分割和深度正则化对于根据二维观测引导内层身体的形状、防止身体形状超出服装范围，以及修正服装与身体之间不准确的相交至关重要。损失函数如下：

$$\mathcal{L}^{rgb}_{\text{body}} = (1-\lambda_c)\mathcal{L}^{L1}_{\text{body}} + \lambda_c\mathcal{L}^{\text{D-SSIM}}_{\text{body}} \quad (5)$$

$$\mathcal{L}_{\text{seg}} = \textstyle\sum_k \left\| M_k \odot (\mathbf{S} - \mathbf{S}_k^{gt}) \right\|_1 \quad \text{where } k \in \{\text{body, garment, bg}\} \quad (6)$$

$$\mathcal{L}^{d}_{body} = \sum_{i_s,j}\omega_{i_s}\omega_j|z_{i_s} - z_j| + \sum_{i_o,j}\omega_{i_o}\omega_j|z_{i_o} - z_j| \quad (7)$$

$$\nabla_{\mathbf{G}_{\text{body}}}\mathcal{L}^{\text{SDS}}_{\text{body}} = \mathbb{E}_{\epsilon,t}\left[w_t\left(\epsilon_\phi(\mathbf{z}_t^{\text{body}}; y_{\text{body}}, t) - \epsilon\right)\frac{\partial \mathbf{x}_{\mathbf{body}}}{\partial \mathbf{G}_{\mathbf{body}}}\right] \quad (8)$$

其中，损失 $\mathcal{L}^{l1}_{body}$ 和 $\mathcal{L}^{\text{D-SSIM}}_{body}$分别表示掩码 RGB 图像 $g(\mathbf{(G^{body}, G^{garment})}, \mathbf{P})$ 的 $l1$ 损失和 D-SSIM 损失，这些图像通过对.....进行整合渲染得到

方法	00122-Inner-Take8			00127-Inner-Take5			00152-Inner-Take4		
	结构相似性指数↑	峰值信噪比↑	学习感知图像块相似度↓	结构相似性指数↑	峰值信噪比↑	学习感知图像块相似度↓	结构相似性指数↑	峰值信噪比↑	学习感知图像块相似度↓
GALA（使用三维扫描）	0.9801	34.60	0.0257	0.9766	32.53	0.0321	0.9805	33.38	0.0268
GALA（不使用三维扫描）	0.9697	31.70	0.0382	0.9684	31.44	0.0405	0.9715	31.74	0.0381
我们的方法（无三维扫描）	**0.9873**	**35.40**	**0.0195**	**0.9880**	**35.93**	**0.0191**	**0.9869**	**35.91**	**0.0214**

方法	00174-Inner-Take10			00175-Inner-Take4			00190-Inner-Take2		
	结构相似性指数↑	峰值信噪比↑	学习感知图像块相似度↓	结构相似性指数↑	峰值信噪比↑	学习感知图像块相似度↓	结构相似性指数↑	峰值信噪比↑	学习感知图像块相似度↓
GALA 使用三维扫描	0.9801	34.93	0.0340	0.9792	31.18	0.0292	0.9732	33.37	0.0250
GALA 未使用三维扫描	0.9718	30.91	0.0453	0.9674	28.88	0.0448	0.9628	31.25	0.0346
我们的方法（无三维扫描）	**0.9866**	**36.09**	**0.0306**	**0.9882**	**36.67**	**0.0207**	**0.9857**	**35.02**	**0.0153**

表 1： **4D-Dress 数据集上的定量重组评估**。

身体层和服装层。目标分割图像 $\mathbf{S}^{gt}$ 被赋予统一颜色，表示为 $\mathbf{S}^{gt} = v \cdot \mathbf{1}^{H\times W}$，其中 v 是用于表示不同层的任意 RGB 值。除 RGB 图像外，我们还渲染分割图像 $\mathbf{S} = g((\mathbf{G}'_{\mathbf{body}}, \mathbf{G}'_{\mathbf{garment}}), \mathbf{P})$，其中 $\mathbf{G}' = (\mu, c', s, \alpha, q)$。这里，每个二维高斯的颜色都设为与相应层的颜色一致，即 $c'_{body} = v_{body}$，$c'_{garment} = v_{garment}$。对于深度失真损失 $\mathcal{L}^d_{body}$，我们定义了第 i 个交集的掩码：i_s 表示身体的可见区域，而 i_o 表示遮挡区域；后者通过服装进行掩码处理，并使用虚拟身体形状计算得出。内层 SDS 损失的文本提示根据用户需求设定（例如，“一张男子穿着裤子和白色背心的照片”）。与阶段 1 类似，阶段 2 的总损失由 RGB 重建、法线一致性、深度失真和 SDS 引导损失加权求和得到，并为内层、外层和背景额外添加二维分割损失。

3.3 外层细化

由于阶段 1 中将外层作为单层进行重建和图像修复，在将内层与外层组合时可能会产生一些伪影，尤其是在使用 SDS 优化的遮挡区域。为缓解这一问题，我们冻结内层的 2D 高斯，仅细化外层。初始化时，从第 3.1 节生成的粗略规范外层中采样点。为进一步加快训练过程，我们采用与第 3.2 节类似的虚拟引导策略，将粗略外层网格用作虚拟层。为减少虚拟试穿过程中严重缩放形变导致的渲染伪影，我们在致密化过程中施加缩放约束：当 2D 高斯达到手动设定的 0.01 缩放阈值时，将其拆分，并在拆分后将缩放值裁剪为 0.01。

训练损失包括外层区域的掩码 RGB 图像损失、法线一致性损失、深度失真损失、二维分割损失，以及合成

SDS 损失，使用与阶段 1 相同的文本提示。法线一致性损失和深度失真损失应用于（a）内层与外层的合成，以及（b）虚拟层与外层的合成。阶段 3 的总损失是所有这些损失的加权和。

3.4 网格提取与虚拟试穿

我们针对三维虚拟试穿应用采用一种由网格驱动的方法。首先，我们分别从相应的二维高斯资产中提取规范人体层和规范服装层的 3D 网格。接着，将 2D 高斯附着到规范网格上距离其最近的三角形。为适应新姿势，我们对规范网格变换应用线性混合蒙皮（LBS），并对外层应用拉普拉斯变形，以解决内层与外层之间可能出现的穿插问题。最后，将 2D 高斯从规范网格变换到目标网格。

4 实验

4.1 预处理

据我们所知，GALA 是目前唯一一种通过图像修复将真实感三维人体分解为多个层的方法。为进行公平比较，我们采用了与 GALA 类似的预处理流程。为增强对任意多视角图像集的适应性，我们进一步从头开始联合估计相机内参和外参，以及人体形状和姿态参数。给定静态人体的多视角 RGB 图像，我们使用 COLMAP 计算相机外参和内参 [25]。接着，我们利用现成的二维分割模型 [11] 生成身体和服装组件的标注掩码。主体的人体形状（表示为 $\beta \in \mathbb{R}^{10}$）和人体姿态（记为 $\theta \in \mathbb{R}^{144}$）通过基于多视角图像及相应相机姿态优化 SMPL(-X) 参数获得。

4.2 数据集与评估指标

4D-Dress [29] 是一个真实世界的着装人体数据集，包含 64 套人体服装及其运动序列。每一帧都包含高质量的三维纹理扫描及顶点级语义标签，以及对应的服装网格和拟合后的 SMPL(-X) 身体网格。Thuman2.0 数据集 [33] 包含 526 个重建的着装人体扫描。在评估中，训练数据分为两种类型。第一种由 72 张图像组成，这些图像由分布在主体周围的虚拟相机位置均匀采集。第二种包括 GALA 原始流程中使用的真实三维扫描，以及训练数据：从随机旋转角度和距离渲染得到的法线图像与分割掩码，还有头部和手部放大视图。我们的方法和 VTON360 仅使用第一种数据，而 GALA 的框架则同时使用两种数据。在评估中，我们

方法	00122-Inner-Take8		00127-Inner-Take5		00152-Inner-Take4	
	CLIP↑	IR↑	CLIP↑	IR↑	CLIP↑	IR↑
GALA（使用三维扫描）	**30.50**	0.313	30.10	0.151	30.11	-0.513
GALA（不使用三维扫描）	30.08	0.175	29.34	0.138	29.90	-0.638
我们的方法（不使用三维扫描）	30.48	**0.534**	**30.97**	**0.217**	**31.08**	**-0.362**
方法	00174-Inner-Take10		00175-Inner-Take4		00190-Inner-Take2	
	CLIP↑	IR↑	CLIP↑	IR↑	CLIP↑	IR↑
GALA（使用三维扫描）	31.27	1.715	32.21	1.783	31.30	0.689
GALA（不使用三维扫描）	30.30	1.641	31.18	1.676	29.64	0.491
我们的方法（不使用三维扫描）	**31.82**	**1.767**	**33.26**	**1.843**	**33.34**	**0.993**

表2： **在 4D-Dress 数据集上的定量图像修复评估**。

为每个主体渲染 72 个测试视角：对于 GALA，相机位置均匀分布在观察球面上；对于 VTON360，这些位置则沿主体周围的水平圆周等间距排列。

评估指标：为了与 GALA 进行比较，我们使用 SSIM [30], 、峰值信噪比和学习感知图像块相似度 [35] 来评估重建质量。我们使用 CLIP [22]和 ImageReward (IR) [32] 分数来评估图像修复质量。与 VTON360 类似，我们计算参考图像与三维虚拟试穿的渲染测试视角图像之间的平均 DINO相似度 [19]分数。

实验设置：为了与 GALA 进行比较，我们从 4D-Dress 数据集中选取了六个身体形状各异的主体，并从各自的运动序列中随机采样一个姿势。我们的研究涵盖的服装包括 T 恤、长袖衬衫和毛衣。为了与 VTON360 进行比较，我们从 Thuman2.0 中选取六个主体作为内部身体重建的基础。在此设置下，我们的方法将从 4D-Dress 数据集中重建的六件服装与这些内部身体相结合。相比之下，VTON360 使用相同的 Thuman2.0 主体进行虚拟试穿，但仅将相应 4D-Dress 套装的正面和背面服装图像作为输入。所有定量评估均仅关注上身服装，以便进行比较。

4.3 与 GALA 的比较

重组结果： 在 4D-Dress 数据集上的定量评估（表1）表明，我们的方法在结构相似性指数、峰值信噪比和学习感知图像块相似度等多项指标上始终优于 GALA。具体而言，当主体穿着紧身服装或带有复杂纹理和图案的衣物时，我们的方法能够有效地重建并修复身体和服装。相比之下，GALA 高度依赖三维扫描的几何信息以及拟合得到的 SMPL(-X) 模型。如果 SMPL(-X) 模板不准确，就会导致 GALA 在第一阶段的几何重建出现缺陷，进而在合成时产生重叠渲染。GALA 通过额外的细化步骤来解决这一问题，但

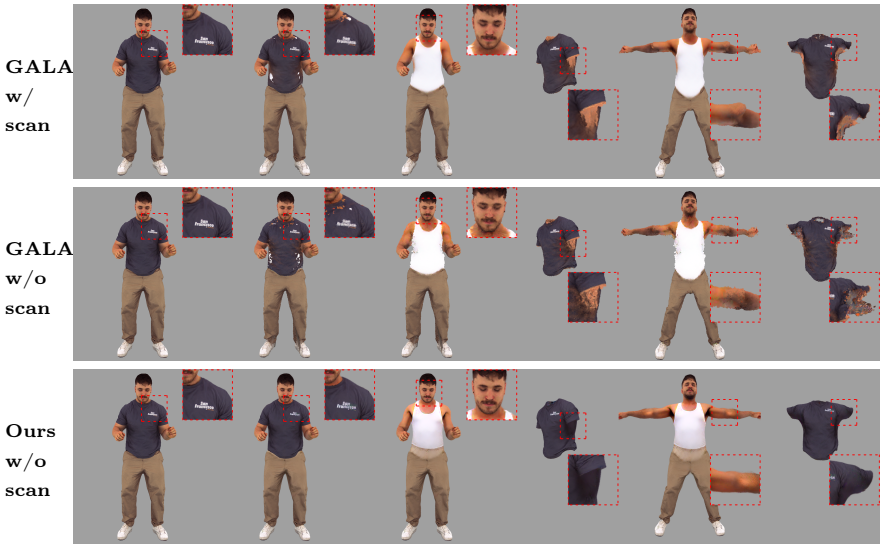

图3：**GALA（使用和不使用三维扫描）与我们的方法（不使用三维扫描）的定性比较。**顶行展示使用三维扫描训练的 GALA 结果，中间行展示未使用三维扫描训练的 GALA 结果，底行展示我们的方法的结果，该方法同样不需要三维扫描。从左到右，每列依次展示：（1）真实图像，（2）重新组合的身体和服装，（3）姿态调整后的身体，（4）姿态调整后的服装，（5）规范身体，以及（6）规范服装。

仍然存在穿透伪影。由于细化代码尚未公开，我们报告的是其细化步骤之前的结果。图3展示了说明这些差异的定性示例。

修复结果：我们在测试视角下渲染经过图像修复的规范内部身体，并使用文本提示“一个穿着裤子和白色背心、背景为黑色的男人/女人”对分数取平均值。图2中的定量结果表明，与 GALA 相比，我们的方法获得了更高的 CLIP 和 ImageReward 分数。如图3所示，由于结合了 2DGS 和扩散模型，我们的方法生成的图像修复结果更加逼真且连贯。

4.4 与 VTON360 的比较

在评估中，VTON360 [7] 需要分别训练 6 个主体和 6 件上身服装，而我们的方法只需重建全部 12 个主体一次，之后即可任意重新组合内层和外层。平均 DINO 相似度分数的定量结果如表 3 所示。图 4 中的定性对比表明，我们的方法能够保留高频细节，例如 T 恤上的文字和毛衣上的图案，而 VTON360 生成的结果则较为模糊。与我们的方法只需分解一次图层即可通用不同，VTON360 需要针对每一组特定的服装与人物重新训练或优化重建结果。



图 4：**VTON360 与我们的方法的定性比较**。上：原始主体、正面/背面服装图像，以及 VTON360 编辑后的三维高斯泼溅渲染结果。下：规范身体、规范服装，以及我们方法重组后的结果。

方法	DINO_{sim} ↑
VTON360	0.455
Ours	**0.506**

表 3：**在 Thuman2.0/4D-Dress 设置下，使用 VTON360 进行的定量虚拟试穿评估**。

4.5 自定义单目视频结果

与 GALA 相比，我们方法的一项优势是，无需显式 3D 信息，仅凭多视角图像即可将人体分解为多个图层。为展示这一能力，我们将系统应用于自定义数据：使用手持单目摄像机，以球形轨迹围绕场景中的静态人体主体移动采集数据。我们从视频中提取了约 100 张图像，并对这些图像应用了预处理流程（第 4.1 节）。如图 5 所示，我们的方法成功重建并修复了分层的 3D 人体化身。

4.6 消融实验

我们开展了三项消融研究，以评估我们的方法中不同组件的贡献。首先，我们对比了阶段 2 和阶段 3。如图 6c 所示，细化过程会移除附着不准确的区域，并加密稀疏的外层二维高斯。表 4a 中的定量结果表明，基于描述两层标准构成的提示词计算得到的 CLIP 和 ImageReward 得分均有所提升。其次，我们评估了虚拟 SMPL(-X) 身体引导的效果。图 6a 显示，尽管使用或不使用虚拟引导时，二维高斯泼溅（2DGS）的渲染输出在视觉上相似，但在未应用虚拟引导时，图像修复后身体的网格中会出现凹陷伪影



图 5：**用于分解的自定义单目视频演示。**

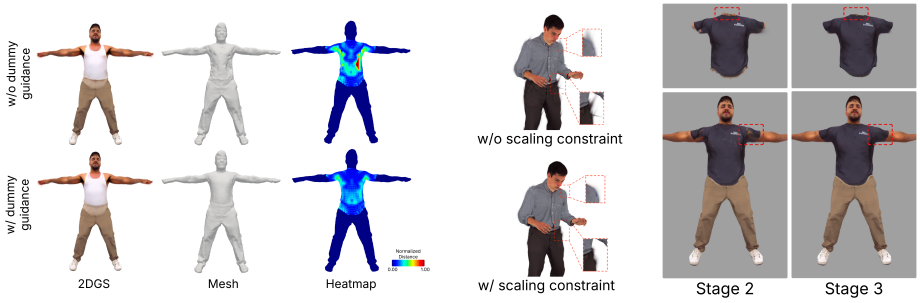

图 6：**消融实验。** (a) 虚拟形状引导的可视化，展示了渲染的内层、网格和误差热图。(b) 缩放约束对虚拟试穿结果的影响。(c) 阶段 2 与阶段 3 的比较。

区域。网格与真实 SMPL(-X) 之间差异的热力图显示，使用虚拟引导后误差有所降低（表 4b），从而提高了几何精度，并防止虚拟试穿中的渲染重叠。最后，我们评估了缩放约束策略。最初为静态场景设计的 2DGS 训练策略倾向于为低频区域生成较大的高斯。由于二维高斯附着在网格上，在严重形变（例如极端拉伸、急剧折叠）下进行虚拟试穿模拟时，较大的高斯会导致视觉伪影。在训练期间约束高斯尺寸可以有效缓解这些伪影（图 6a）。

5 讨论

我们提出了一种新方法，可根据静态姿势主体的多视角图像生成分层 3D 人体化身。通过利用扩散模型引导的 2D 高斯泼溅，我们的方法能够创建照片级真实感三维服装和人体资产，非常适用于虚拟试穿等应用。实验表明，我们的方法在重建和分解方面均优于最先进的分解

方法	CLIP↑	红外↑
阶段 2	30.83	1.378
阶段 3	**30.96**	**1.388**

(a)

方法	均值↓	标准差↓	最大值↓
无引导	0.021	0.012	0.091
有引导	**0.015**	**0.006**	**0.048**

(b)

表 4：**消融实验**。（a）服装细化的定量结果。（b）有无引导时的热力图差异，展示遮挡区域中的误差指标。

方法，并且同样适用于图像修复；此外，该方法还兼容易于采集的自定义数据，提升了虚拟试穿的实用性。

局限性：我们的方法需要静态人体的多视角图像。由于依赖 SMPL(-X) 线性混合蒙皮，在处理宽松服装或存在自交的服装时，分解可能会失败。当服装明显小于身体时，虚拟试穿也可能产生穿透伪影。将该方法扩展到从单张图像对人体进行建模和分解，是未来工作中一个有前景的方向。基于物理的仿真可以进一步提高整体鲁棒性，而在物理仿真和静态姿势下实现高斯服装的高保真渲染，仍是未来研究的重要领域。

致谢 本研究在爱尔兰科学基金会数字增强现实研究培训中心（d-real，位于都柏林圣三一学院）的资助下开展，资助编号为 18/CRT/6224。我们还感谢欧洲地平线框架计划项目的支持，资助协议编号为 101070109（TRANSMIXR）。

参考文献

1. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.：结合隐式函数学习与参数化模型进行三维人体重建。载于：欧洲计算机视觉会议（ECCV），第 511–529 页（2020） 2. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.：OpenPose：利用部件亲和场进行实时多人二维姿态估计。IEEE模式分析与机器智能汇刊（TPAMI）**43**(1)，172–186（2021） 3. Chen, X., Zheng, Y., Black, M.J., Hilliges, O., Geiger, A.：SNARF：用于动画化非刚性神经隐式形状的可微前向蒙皮。载于：IEEE/CVF国际计算机视觉会议（ICCV），第 11574–11584 页（2021） 4. Chen, Y., Chen, Z., Zhang, C., Wang, F., Yang, X., Wang, Y., Cai, Z., Yang, L., Liu, H., Lin, G.：GaussianEditor：借助高斯泼溅实现快速、可控的三维编辑。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR），第 21476–21485 页（2024） 5. Corona, E., Pumarola, A., Alenyà, G., Pons-Moll, G., Moreno-Noguer, F.：SM-PLicit：面向穿着衣物的人体的拓扑感知生成模型。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR），第 11870–11880 页（2021）

6. Haque, A., Tancik, M., Efros, A.A., Holynski, A., Kanazawa, A.：Instruct-NeRF2NeRF：使用指令编辑三维场景。载于：IEEE/CVF国际计算机视觉会议（ICCV），第 19683–19693 页（2023） 7. He, Z., Ning, Y., Qin, Y., Wang, W., Yang, S., Lin, L., Li, G.：VTON 360：从任意观察方向实现高保真虚拟试穿。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR），第 26388–26398 页（2025） 8. Hu, L., Zhang, H., Zhang, Y., Zhou, B., Liu, B., Zhang, S., Nie, L.：GaussianAvatar：通过可动画化三维高斯，从单段视频实现逼真的人体虚拟形象建模。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR），第 634–644 页（2024） 9. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.：用于几何精确辐射场的二维高斯泼溅。载于：SIGGRAPH会议论文集（2024） 10. Jiang, B., Ren, X., Dou, M., Xue, X., Fu, Y., Zhang, Y.：LoRD：用于高保真动态人体建模的局部四维隐式表示。载于：欧洲计算机视觉会议（ECCV），第 307–326 页（2022） 11. Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.：Sapiens：人体视觉模型的基础。载于：欧洲计算机视觉会议（ECCV），第 206–228 页（2024） 12. Kim, T., Kim, B., Saito, S., Joo, H.：GALA：从单次扫描生成可动画化分层资产。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR），第 1535–1545 页（2024） 13. Li, Z., Zheng, Z., Wang, L., Liu, Y.：可动画化高斯：学习依赖姿态的高斯映射，实现高保真人体虚拟形象建模。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR），第 19711–19722 页（2024） 14. Lin, S., Li, Z., Su, Z., Zheng, Z., Zhang, H., Liu, Y.：LayGA：用于可动画化服装迁移的分层高斯虚拟形象。载于：SIGGRAPH会议论文集（2024） 15. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.：SMPL：蒙皮多人线性模型。ACM图形学汇刊（TOG）**34**(6)，248:1–248:16（2015） 16. Ma, Q., Yang, J., Tang, S., Black, M.J.：点在着装人体建模中的作用。载于：IEEE/CVF国际计算机视觉会议（ICCV），第 10954–10964 页（2021） 17. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.：NeRF：将场景表示为神经辐射场以进行视图合成。载于：欧洲计算机视觉会议（ECCV），第 405–421 页（2020） 18. Nam, H., Kim, D., Oh, J., Lee, K.M.：DeClotH：从单张图像重建可分解的三维衣物与身体。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR），第 5636–5645 页（2025） 19. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., 等：DINOv2：无监督学习稳健的视觉特征。arXiv预印本 arXiv:2304.07193（2023） 20. Pan, P., Su, Z., Lin, C., Fan, Z., Zhang, Y., Li, Z., Shen, T., Mu, Y., Liu, Y.：HumanSplat：利用结构先验实现可泛化的单图人体高斯泼溅。载于：神经信息处理系统进展（NeurIPS）（2024） 21. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.：DreamFusion：利用二维扩散实现文本生成三维。载于：国际学习表征会议（ICLR）（2023） 22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., 等：通过自然语言监督学习可迁移视觉模型。载于：国际机器学习会议（ICML），第 8748–8763 页（2021）

23. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 使用潜在扩散模型进行高分辨率图像合成。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）。第10674–10685页（2022） 24. Saito, S., Huang, Z., Natsume, R., Morishima, S., Li, H., Kanazawa, A.: PIFu：用于高分辨率着装人体数字化的像素对齐隐式函数。载于：IEEE/CVF国际计算机视觉会议（ICCV）。第2304–2314页（2019） 25. Schönberger, J.L., Frahm, J.M.: 重新审视运动恢复结构。载于：IEEE计算机视觉与模式识别会议（CVPR）。第4104–4113页（2016） 26. Shen, K., Guo, C., Kaufmann, M., Zarate, J.J., Valentin, J., Song, J., Hilliges, O.: X-Avatar：表现力人体化身。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）。第16911–16921页（2023） 27. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: 深度行进四面体：一种用于高分辨率三维形状合成的混合表示。载于：神经信息处理系统进展（NeurIPS）。第34卷，第6087–6101页（2021） 28. Su, Z., Hu, L., Lin, S., Zhang, H., Zhang, S., Thies, J., Liu, Y.: CaPhy：为可动画人体化身捕捉物理属性。载于：IEEE/CVF国际计算机视觉会议（ICCV）。第14104–14114页（2023） 29. Wang, W., Ho, H.I., Guo, C., Rong, B., Grigorev, A., Song, J., Zárate, J.J., Hilliges, O.: 4D-DRESS：带语义标注的真实世界人体服装四维数据集。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）。第550–560页（2024） 30. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: 图像质量评估：从误差可见性到结构相似性。《IEEE图像处理汇刊》**13**(4)，600–612（2004） 31. Xie, Z., Huang, Z., Dong, X., Zhao, F., Dong, H., Zhang, X., Zhu, F., Liang, X.: GP-VTON：通过协同局部流与全局解析学习，实现通用虚拟试穿。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）。第23550–23559页（2023） 32. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: ImageReward：学习并评估人类对文本到图像生成的偏好。载于：神经信息处理系统进展（NeurIPS）。第15903–15935页（2023） 33. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4D：利用极稀疏的消费级RGBD传感器实时捕捉人体体积。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）。第5742–5752页（2023） 34. Zhang, L., Rao, A., Agrawala, M.: 为文本到图像扩散模型添加条件控制。载于：IEEE/CVF国际计算机视觉会议（ICCV）。第3813–3824页（2023） 35. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: 深度特征作为感知度量的非凡效用。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）。第586–595页（2018） 36. Zhang, W., Yan, Y., Wu, S., Liao, M., Yang, X.: 基于分层表示的解耦着装化身生成。载于：IEEE/CVF国际计算机视觉会议（ICCV）。第11327–11338页（2025） 37. Zhang, Z., Yang, Z., Yang, Y.: SIFU：用于真实世界可用着装人体重建的侧视图条件隐式函数。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）。第9936–9947页（2024） 38. Zhu, L., Li, Y., Liu, N., Peng, H., Yang, D., Kemelmacher-Shlizerman, I.: M&M VTO：多服装虚拟试穿与编辑。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）。第1346–1356页（2024）