通过 BridgeDiffusion 在统一潜在空间中进行联合几何–外观人体重建

Yingzhi Tang、Qijian Zhang 和 Junhui Hou，高级会员，*IEEE*

arXiv:2601.00328v1 [cs.CV] 1 Jan 2026

摘要**仅凭单张 RGB图像 对 三维数字人 进行一致且高保真的几何与外观重建，本质上是一项极具挑战性的任务。现有研究通常采用相互解耦的流程来进行几何估计和外观合成，这往往妨碍统一重建并导致不一致。本文提出 JGA-LBD，这是一种新颖的框架，可将几何与外观的建模统一到联合潜在表示中，并将生成过程构建为桥接扩散。我们发现，直接整合异构输入条件（例如深度图、SMPL模型）会导致显著的训练困难，因此将所有条件统一为三维高斯表示，并通过共享的稀疏变分自编码器（变分自编码器）进一步压缩到统一潜在空间。随后，桥接扩散的专门形式能够从目标潜代码的部分观测开始，并专注于推断缺失部分。最后，专用解码模块从推断出的潜在表示中提取完整的三维人体几何结构，并渲染新视角。实验表明，JGA-LBD 在几何保真度和外观质量方面均优于当前最先进的方法，包括具有挑战性的野外场景。我们的代码将公开发布于 https://github.com/haiantyz/JGA-LBD。**

索引词—**人体重建、桥接扩散、潜在扩散、三维高斯泼溅、变分自编码器。**

I 引言

RECONSTRUCTING 从单视图 RGB 图像重建高保真数字人是计算机视觉中的一项基础问题，在虚拟现实、游戏、自动驾驶等领域有着广泛应用。尽管近年来取得了进展，但由于人体形状复杂、服装多样且存在严重的自遮挡，从单张图像中准确重建人体几何和外观仍然极具挑战性。

一般而言，人体重建任务有两项需求，即几何重建和外观重建。现有解决方案大致可分为基于隐式函数的方法 [11], [38], [39], [59], [60],基于显式点的方法 [8], [44], [63], 以及基于图像生成的方法 [19], [55]。隐式函数方法 [11], [38], [39], [59], [60] 提取像素对齐特征、特征这些特征来自 SMPL 等参数化人体模型或其他线索，并使用多层感知机学习占据场或符号距离场，以进行表面重建。尽管这类方法在几何重建方面效果显著，但往往无法重建准确且细致的外观，因为三维空间中的查询点很少能获得精确的颜色监督；通常会使用最近的表面点作为代理，导致模型学习到的是近似值，而非真实颜色。基于显式点的方法 [8], [43], [44], [63] 使用从 RGB图像（通常借助估计的深度图）生成的点云来表示人体。这类方法能够重建细节丰富的几何形状，但通常忽略外观，或是在预先重建的几何形状或 SMPL模型等参数化模型上构建外观模型，因此需要多阶段流程，可能导致几何与外观之间不一致。基于图像生成的方法 [19], [55]利用大型生成模型，从单个输入视图合成多视图图像，再使用连续重网格化等技术重建几何形状 [31]。尽管前景可期，但这类方法同样需要多个步骤，且容易受到合成视图中伪影的影响。现有方法要么难以处理外观，要么只关注几何，或依赖复杂的多阶段流程。这些局限表明，仍有两项关键需求尚未得到满足：

1) 高质量的真实表示至关重要，它们需同时编码几何与外观，同时便于网络有效学习；2) 需要采用单阶段流程联合重建几何与外观，因为这种方式能够自然地确保几何结构与外观之间的一致性

对于第一项要求，我们采用三维高斯 [17]作为真实表示，因为它们是显式的，能够自然地捕捉粗略几何和精细外观。为进一步实现高质量的几何与外观编码，我们遵循 [43] 提出的两阶段真实三维高斯准备策略。通过将高斯限制在真实表面上，该策略使表示能够忠实地编码准确的几何信息以及细致的外观。

因此，我们着重关注第二项要求，并将任务重新表述为：从单张 RGB图像中通过单阶段预测三维高斯表示，从而统一实现几何重建与新视角渲染。一种直观的解决方案是学习三维高斯的紧凑潜在表示，并使用扩散模型在该潜在空间中进行生成建模。然而，现有

本研究部分由国家自然科学基金优秀青年科学基金 62422118 资助，部分由香港研究资助局通过项目 11219324、11202320 和 11219422 资助。（通讯作者：*Junhui Hou*）Yingzhi Tang 和 Junhui Hou 任职于香港特别行政区香港城市大学计算机科学系。电子邮箱：yztang4-c@my.cityu.edu.hk；jh.hou@cityu.edu.hk。Qijian Zhang 任职于中国广东深圳腾讯游戏。电子邮箱：keegan.zqj@gmail.com。

方法无法直接采用：基于 3DShape2VecSet 的方法 [54] 仅限于隐式几何，无法捕捉外观，而 Trellis [48] 依赖于中间稀疏结构生成，破坏了单阶段的任务表述。

在本文中，我们提出了 JGA-LBD，这是一种在统一潜在空间中进行学习的桥扩散模型，可通过单步重建数字人体的高分辨率三维高斯。具体而言，我们设计了一种稀疏变分自编码器，在几何与外观监督下进行联合训练，可将输入的三维高斯映射为紧凑的潜在表示。为充分利用图像中蕴含的丰富信息，我们从输入中提取了两种互补模态——深度估计和 SMPL预测。然而，二者固有的差异使得直接利用它们颇具挑战。为此，我们引入模态统一模块，将这两种模态转换为三维高斯表示，随后由稀疏变分自编码器将其压缩至同一潜在空间。这一设计确保后续所有扩散学习都在统一潜在空间中进行，从而大幅降低训练复杂度。基于这一统一潜在设计，我们发现桥接扩散意外地适用于人体重建，因为深度条件潜变量天然对应于目标潜代码的部分观测。桥扩散模型无需从噪声开始生成，只需补全缺失部分，从而显著降低生成难度，并提升所学习潜在表示的质量。最后，从潜代码解码得到的三维高斯既可用于提取几何表面，也可通过基于泼溅的光栅化实现高质量新视角渲染。在两个基准数据集上的大量实验，以及对野外图像的评估，始终表明 JGA-LBD 在定量精度和定性视觉真实感方面均优于最先进的方法。

总而言之，本工作的主要贡献如下：

- 我们设计了一种稀疏变分自编码器，可将高分辨率 3D 高斯表示的几何与外观联合压缩为紧凑的潜在编码，克服了以往方法只关注几何或依赖额外稀疏结构先验的局限；
- 我们引入了一种模态统一模块，通过稀疏U-Net和SMPL修补器，将深度估计和SMPL预测转换为潜在结构引导，从而确保不同模态之间的条件控制一致；并且
- 我们调整桥式潜扩散，使其在统一潜在空间中运行，从而能够以高效的单阶段方式生成完整的潜在编码，同时对几何与外观进行建模。

本文其余部分的组织如下。第II 节全面回顾现有文献，包括单目深度估计、点云生成、人体姿态与形状估计以及单视图人体重建。第III 节详细介绍我们提出的 JGA-LBD。第IV节开展 extensive experiments 和消融研究，以评估 JGA-LBD 的有效性。最后，第V 节对本文进行总结。

II. 相关工作

A. 扩散模型

扩散模型 [12] 在图像合成、视频生成和音频处理等多种领域的生成建模中取得了显著成功。其核心原理是通过逐步对高斯噪声进行去噪来学习数据分布。尽管这种方法功能强大，但直接在像素空间中执行扩散过程的计算成本高昂，而且往往存在冗余。为解决这一问题，潜在扩散 [37]会先将输入压缩到紧凑的潜在空间，再执行扩散过程，从而在保留高质量生成效果的同时实现高效训练。这种范式此后已成为大规模图像和视频扩散模型的标准 [2], [27], [34]。

另一条并行研究路线聚焦于条件扩散，旨在利用辅助输入引导生成。早期方法（如分类器引导和无分类器引导）[5], [13]会在采样过程中注入条件信号。后续方法（如 ControlNet） [57], 通过引入可训练网络来调节受外部条件影响的中间特征，进一步拓展了这一思路，实现了细粒度的可控性。尽管这些方法效果显著，但它们仍从高斯噪声初始化扩散过程，这可能限制其充分利用结构化先验的能力。桥扩散模型通过学习扩散桥来扩展扩散模型，直接映射两个端点分布，而非从纯噪声映射到目标数据分布。DDBM [62]统一了基于分数的扩散与最优传输流匹配，使扩散模型能够更自然地处理非噪声输入。BBDM [18] 将图像到图像的转换重新表述为随机布朗桥过程，直接对两个域之间的变换进行建模。与传统的条件扩散模型不同，BBDM [18]在源域和目标域之间执行双向扩散，从而缓解域差距问题。

除了基于扩散的生成模型之外，基于流的方法 [2], [23], [25] 还提供了另一种范式：学习数据分布之间确定且可逆的映射，并进行精确的似然估计。流匹配[23] 提出了一种无需模拟的连续归一化流训练方法，通过学习固定概率路径上的向量场，将扩散模型与基于流的模型统一起来。它支持最优传输插值等非扩散路径，从而实现更高效的训练和更快的采样，并提升泛化能力。整流流 [25] 学习确定性常微分方程，使其沿近乎直线的路径在两个分布之间进行传输，为基于扩散的生成提供了一种高效替代方案。其整流过程逐步提升传输效率，从而仅需少量离散化步数即可实现快速、准确的推理。

本文结合潜在扩散与桥接扩散来解决人体重建问题，主要有两个原因。首先，采用潜在扩散来处理大规模三维高斯表示，这种表示通常包含

超过 $100k$ 个基元。其次，桥接扩散使模型能够专注于生成被遮挡区域。

*B. 3D*生成模型

由于三维表示种类繁多，且3D监督的复杂性更高，生成3D模型本质上比合成2D图像或视频更具挑战性，因此形成了两个主要研究方向：基于多视图的生成和直接3D表示生成。

基于多视图的方法通常首先使用扩散模型，根据输入合成多个2D视图，然后通过连续重网格化 [31], 3、3D高斯泼溅 [17], 或 NeRF 等技术，从生成的多视图图像中重建3D内容 [28]。例如，Zero123[24]采用 Stable Diffusion 生成多视图图像，随后按照 SJC 公式 [45]优化 NeRF 以拟合这些合成视图，并从学习到的密度场中通过行进立方体提取网格。鉴于3D高斯泼溅（3DGS）相比 NeRF 效率更高，DreamGaussian[41] 和 LGM [40] 等方法使用图像扩散生成多视图图像，随后拟合3DGS；然而，这些方法通常只生成少量三维高斯（约 $20k$ 个点），因此需要采用连续重网格化等额外技术来构建网格。由于点密度有限，合成的新视角图像通常分辨率较低，导致重建网格的表面相对粗糙。

直接进行三维生成缺乏统一的流程，因为三维信息可以有多种表示形式，例如点云、隐式表示、体数据、NeRF 或三维高斯泼溅。Luo 等人 [26] 提出了一种点云生成概率模型，将点视为热力学扩散过程中的粒子。DiffGS [61]将三维高斯泼溅场景编码为三平面潜变量，并利用 DiT [34]在潜在空间中进行学习；CraftsMan3D [21] 将形状压缩为向量集合 [54] ，并训练 DiT 学习隐式场，再在推理阶段提取网格；Trellis [48] 则利用稀疏卷积神经网络压缩三维高斯泼溅，以支持多种下游表示，但需要额外阶段提供几何线索（稀疏结构），且无法在单一潜变量中统一几何与外观。

相比之下，我们的框架将几何与外观联合压缩为统一潜在表示，并采用 BridgeDiffusion 在单一阶段对其进行学习。

*C. 3D*人体重建

基于隐式表示的3D人体重建。 PIFu [38]是利用像素对齐特征重建带颜色3D人体的先驱工作。后续方法通过引入额外线索增强隐式表示：SiTH [11] 通过ControlNet生成背视图图像，并使用蒙皮网格解决3D歧义；GTA [59] 引入基于ViT的编码器—解码器，利用三平面特征重建着装化身；SIFU [60] 则利用SMPL-X引导的交叉注意力和基于扩散的纹理细化流程，提高其在真实场景中的稳健性。尽管取得了这些进展，基于隐式表示的方法仍缺乏真实颜色监督——外观是根据最近表面点近似得到的——限制了其建模高保真纹理的能力。

基于显式表示的3D人体重建。 Natsume [29]提出了 SiCloPe，该方法结合二维轮廓和三维关节，为深度视觉外壳形状重建生成一致的新视角轮廓，随后使用条件生成对抗网络预测背面纹理。ECON [49] 将隐式表面表示与显式人体正则化相结合，以重建具有挑战性的姿势和宽松服装下的高保真3D人体。Han 等人 [8]使用低分辨率深度网络预测全局形状，并通过分部图像到法线网络生成局部细节，再将二者合并，以生成用于完整3D重建的高分辨率深度图。Tang 等人 [42] 采用基于三维卷积的体积由粗到细策略，从稀疏 RGB 图像重建着装人体。HaP [44]提出了一种以深度图和 SMPL模型为条件的新型流程，利用扩散模型生成人体点云，随后通过屏蔽泊松重建将其重建为表面[16]。

基于 3DGS 的三维人体重建。 近年来，三维高斯泼溅已成为人体重建的一种强大显式表示。IDOL [63] 利用大规模数据集和基于 Transformer 的预测器，高效重建可动画化的高斯化身。MultiGo [55] 引入多层级几何学习，并结合骨架、关节和皱纹细化；LHM [35] 则采用多模态 Transformer，以保留精细的服装和面部细节。 [3] 提出了先生成再细化的流程和 HGM 模块，用于生成高质量的人体三维高斯属性。HumanSplat [32] 在通用 Transformer 框架中使用视频扩散模型生成人体三维高斯属性。HuGDiffusion [43]引入了一种三维高斯属性扩散模型，该模型以 SMPL 语义特征和像素对齐特征为条件。这些方法展现了三维高斯泼溅在捕捉几何和外观方面的优势，但紧凑表示学习和高效生成建模仍面临挑战。

基于优化的三维人体重建。 TeCH [15]、Human-SGD [1]、、WonderHuman [47]和 GeneMAN [46] 也能处理这一任务，并具有良好的泛化能力。PSHuman[19]和 MagicMan [9] 使用扩散模型生成多视图图像，并利用生成的多视图图像重建三维人体。Human-GIF [14]则将人体重建任务表述为单视图条件人体扩散生成任务，利用大型扩散模型生成缺失信息。

III. 提出的方法

JGA-LBD 是一种基于三维高斯溅射的方法，可从单视图 RGB 图像重建三维高斯表示 $\mathcal{G} = \{G_1, ..., G_n\}$，其中每个元素 $G_i = \{p_i, c_i, s_i, r_i, o_i\}$编码其对应的三维高斯属性，例如位置 p_i、颜色 c_i、尺度 s_i、旋转 r_i 和不透明度 o_i。

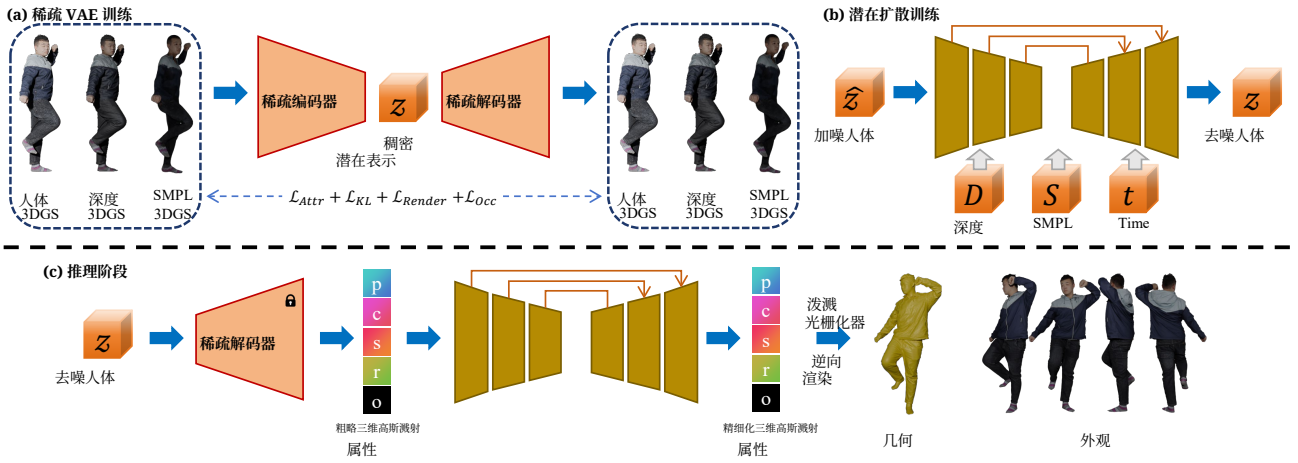

图 1. JGA-LBD 的流程。给定单视图 RGB 图像，将深度和 SMPL 先验转换为三维高斯，再由稀疏变分自编码器压缩为潜在编码。桥扩散模型在深度和 SMPL 先验的条件下生成潜在编码，解码器则将其细化为高保真三维高斯表示，用于表面重建和新视角渲染。**人体 3DGS**：真实人体三维高斯属性 $\mathcal{G}$，**深度 3DGS**：转换后的深度三维高斯属性 $\mathcal{D}$，**SMPL 3DGS**：SMPL 三维高斯属性 $\mathcal{S}$。

所得表示支持高保真的三维表面重建和新视角外观合成。如图 1所示，JGA-LBD 包含四个关键模块：

1）模态统一模块，用于将深度图和 SMPL 顶点统一为相同的稀疏三维高斯格式；2）稀疏变分自编码器，用于将人体三维高斯和转换后的条件压缩为统一潜在空间中的潜在编码；3）桥扩散模型，用于学习结构先验条件下潜在人体三维高斯的分布；4）解码器，用于将去噪后的潜在编码转换回三维高斯表示，随后进行细化以提高其保真度，再用于表面重建和新视角渲染。

请注意，我们遵循 HuGDiffusion 提出的两阶段三维高斯属性准备策略 [43]来构建用于训练的真实人体三维高斯。本文将三维高斯的数量从 $20k$增加到 $200k$。随后，对所有三维高斯进行体素化，以训练稀疏三维卷积神经网络；体素化后，每个场景将包含约$130k \sim 150k$ 个高斯。

A. 模态统一模块

给定一张 RGB 图像，我们推断两种类型的三维信息，即深度图和 SMPL模型，它们将作为后续扩散模型训练的条件输入。深度图可以利用相机参数投影为局部点云，而 SMPL 模型则提供完整的几何先验。尽管两者都存在于三维空间中，但它们属于不同模态：点云编码了 (x, y, z, r, g, b) 数值及外观信息，而 SMPL 网格仅包含几何顶点和面。在 HaP 中，使用 FPS 从深度图生成的局部点云和 SMPL 网格中采样稀疏点，且不使用任何颜色信息。因此，这些稀疏几何表示无法直接用作统一的条件输入，以进行几何与外观重建。为解决这一问题，我们首先将两种模态都转换为一致的三维高斯表示，从而实现有效的模态统一。

对于部分点云，我们首先执行最近邻搜索，将每个点与真实人体三维高斯中距离最近的高斯相关联（准备真实 SMPL 三维高斯时也采用相同流程），并将这些属性作为监督的真实值。随后，基于 Minkowski Engine 的稀疏 U-Net [4] 经过训练，将部分点云的 $(r,\ g,\ b)$ 值映射为三维高斯属性。对于 SMPL 网格，我们首先将其投影到图像平面上，仅为可见顶点赋予颜色信息，遮挡顶点则不赋予颜色。然后，将这个部分观测到的网格输入另一个稀疏 Minkowski U-Net，为所有顶点预测完整的三维高斯属性，从而生成三维高斯表示。请注意，得到的 SMPL 三维高斯较为粗略，并非用于提供精确的外观监督；这一过程主要确保 SMPL 以统一的三维高斯泼溅格式提供全局结构指导。通过将异构模态转换为三维高斯，我们为扩散过程获得了一致且互补的条件输入。

B. 几何-外观联合压缩变分自编码器

体素是一种常见的三维表示，与标准卷积神经网络兼容。然而，要准确表示三维物体，通常需要极高分辨率的体素网格（至少为 512^3），由于 GPU 显存需求过高，这对于训练而言是不可行的。受潜在扩散 [37],启发，我们采用稀疏变分自编码器，将三维高斯压缩为紧凑的潜在表示，从而能够在统一潜在空间中使用标准卷积神经网络进行高效建模。

给定真实人体三维高斯 $\mathcal{G}$、转换后的深度三维高斯 $\mathcal{D}$以及 SMPL 三维高斯 $\mathcal{S}$，我们的目标

是将它们编码为统一潜在表示，以共同捕获几何和外观（所有高斯的位置在输入稀疏三维卷积神经网络之前都会被体素化）。具体而言，我们使用 Minkowski Engine [4]构建稀疏变分自编码器。稀疏变分自编码器的编码器 $E(\cdot)$ 由若干 ResNet块组成，其输出 z 作为编码器 $E(\cdot)$ 的输出，充当扩散模型的真实值。为避免学习到高方差的潜在空间，我们施加轻微的 KL惩罚项$\mathcal{L}_{\text{KL}}$ ，以 z 使其学习服从标准正态分布的潜在表示。解码器 $D(\cdot)$ 是稀疏变分自编码器中的关键模块，因为它应能独立解码扩散模型去噪后的 z ，而无需 Trellis [48]等稀疏结构提示。因此，我们采用生成式稀疏转置卷积层构建解码器 $D(\cdot)$，从而能够生成新坐标；与标准稀疏转置卷积不同，这些坐标无需依赖编码器缓存的坐标。它从 z 开始，并通过占用损失 $\mathcal{L}_{\text{Occ}}$逐步剪除多余的体素：

$$\begin{aligned}\mathcal{L}_{\text{Occ}} = -\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\Big[o_g(v)\,\log o_p(v)+\\ \big(1-o_g(v)\big)\,\log\big(1-o_p(v)\big)\Big],\end{aligned} \tag{1}$$

其中， $\mathcal{V}$ 表示活跃体素集合， $o_p(v) \in (0,1)$表示预测的占用概率，而 $o_g(v) \in\{0,1\}$ 表示真实占用标签；经过若干层后，最终达到 $\mathcal{G}$ 的分辨率。我们使用均方误差损失监督三维高斯属性的重建。然而，预测体素网格与真实体素网格并未严格对齐，因此无法直接对稀疏张量应用均方误差损失。将稀疏张量转换为稠密形式会引入大量非活跃体素（例如，仅有约 $130k \sim 150k$ 个活跃体素，而总数为 512^3），这会严重稀释梯度并妨碍有效学习。因此，我们仅在预测结果与真实值中活跃体素的交集上计算均方误差损失：

$$\mathcal{L}_{\text{Attr}} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\|\mathbf{a}_p(i)-\mathbf{a}_g(i)\|_2^2, \tag{2}$$

其中， $\mathcal{I}$ 表示活跃体素索引的交集，且$\mathbf{a}_p(i)$、$\mathbf{a}_g(i)$分别表示位置 i处预测的和真实值的三维高斯属性（即 p,c,s,r,o）。然而，仅在交集上进行监督不可避免地会导致某些区域未得到监督。为解决这一局限，我们引入一种损失，通过将预测的三维高斯泼溅渲染为2D图像，并使其与真实图像保持一致。渲染损失由L1损失、SSIM损失和LPIPS损失组成：

$$\begin{aligned}\mathcal{L}_{\text{Render}} = \lambda_1\| I_p - I_g\|_1 + \lambda_2\big(1-\text{SSIM}(I_p,I_g)\big)\\ +\lambda_3\,\text{LPIPS}(I_p,I_g),\end{aligned} \tag{3}$$

其中， I_p 和 I_g 分别表示渲染图像和真实值图像， λ_1, λ_2,λ_3 为平衡权重。稀疏变分自编码器的总体训练目标为：

$$\mathcal{L}_{\text{VAE}} = \lambda_4\mathcal{L}_{\text{KL}} + \lambda_5\mathcal{L}_{\text{Occ}} + \lambda_6\mathcal{L}_{\text{Attr}} + \lambda_7\mathcal{L}_{\text{Render}}, \tag{4}$$

其中， λ_4， λ_5， λ_6， λ_7 是平衡权重。 $\mathcal{G}$， $\mathcal{D}$ 和 $\mathcal{S}$ 的编码结果被转换为稠密潜在表示，用于扩散训练，分别记为 $\mathcal{G}_L$， $\mathcal{D}_L$ 和 $\mathcal{S}_L$。

C. 统一潜在空间中的桥接扩散

扩散模型通常旨在将数据分布传输到标准高斯先验。在我们的设定中，深度导出的潜在编码 $\mathcal{D}_L$ 可视为完整人体高斯表示的一个结构子集。因此，我们不采用朴素的扩散模型，而是采用更强大的桥扩散模型 [62]，它能够学习两个任意分布之间的传输路径。具体而言，目标是在以 SMPL先验 $\mathcal{S}_L$为条件的情况下，将深度导出的潜在编码 $\mathcal{D}_L$ 的部分先验分布 $p_{\mathcal{D}_L}$ 转换为目标分布 $p_{\mathcal{G}_L}$；桥接扩散过程主要对缺失信息进行建模，从而降低生成任务的不确定性和难度。

形式上，桥接扩散过程可表示为一系列带有时间索引的变量 $\{x_t\}_{t=0}^{T}$。利用 Doob 的h-变换 [6]，，条件随机桥可表示为：

$$dx_t = f(x_t,t\mid\mathcal{S}_L)\,dt + g(t)^2\,h(x_t,t,y,T\mid\mathcal{S}_L)\,dt + g(t)\,dw_t, \tag{5}$$

其中， $f(x_t,t\mid\mathcal{S}_L)$为漂移项， $g(t)$为扩散系数， $x_0 \sim p_{\mathcal{G}_L}(x\mid\mathcal{S}_L)$、 $x_T = y$和$y \sim p_{\mathcal{D}_L}$。项$h(x,t,y,T\mid\mathcal{S}_L) = \nabla_x \log\, p(x_T = y\mid x_t = x,\mathcal{S}_L)$表示由 h-变换引入的漂移调整，以确保该过程在两个端点之间插值。反转此桥接过程可得到条件反向随机微分方程：

$$\begin{aligned}dx_t = \Big[f(x_t,t\mid\mathcal{S}_L) - g(t)^2\big(U_\theta(x_t,t,y,T\mid\mathcal{S}_L)-\\ h(x_t,t,y,T\mid\mathcal{S}_L)\big)\Big]dt + g(t)\,d\bar{w}_t,\end{aligned} \tag{6}$$

以及相应的概率流常微分方程：

$$\begin{aligned}dx_t = \Big[f(x_t,t\mid\mathcal{S}_L) - g(t)^2\big(\tfrac{1}{2}U_\theta(x_t,t,y,T\mid\mathcal{S}_L)-\\ h(x_t,t,y,T\mid\mathcal{S}_L)\big)\Big]dt,\end{aligned} \tag{7}$$

其中， $U_\theta(\cdot)$表示一个参数为 θ的神经网络，用于近似桥评分函数。为学习该评分函数，我们采用去噪桥评分匹配，最小化预测评分与高斯桥闭式条件评分之间的差异：

$$\begin{aligned}\mathcal{L}_\theta = \mathbb{E}_{x_t,x_0,x_T,t}\Big[w(t)\,\big\|U_\theta(x_t,x_T,t\mid\mathcal{S}_L)-\\ \nabla_{x_t}\log q(x_t\mid x_0,x_T,\mathcal{S}_L)\big\|^2\Big],\end{aligned} \tag{8}$$

其中， $w(t)$表示一个时变加权函数，用于调整训练过程中不同扩散步的相对重要性。

备注。我们为稠密潜在表示中的每个网格补充一个占据值 $\{0,1\}$，使桥扩散模型能够联合学习潜在特征及其占据情况。在推理过程中，通过仅保留预测占据值大于 0.5 的网格，将稠密潜在表示转换为稀疏潜在表示。

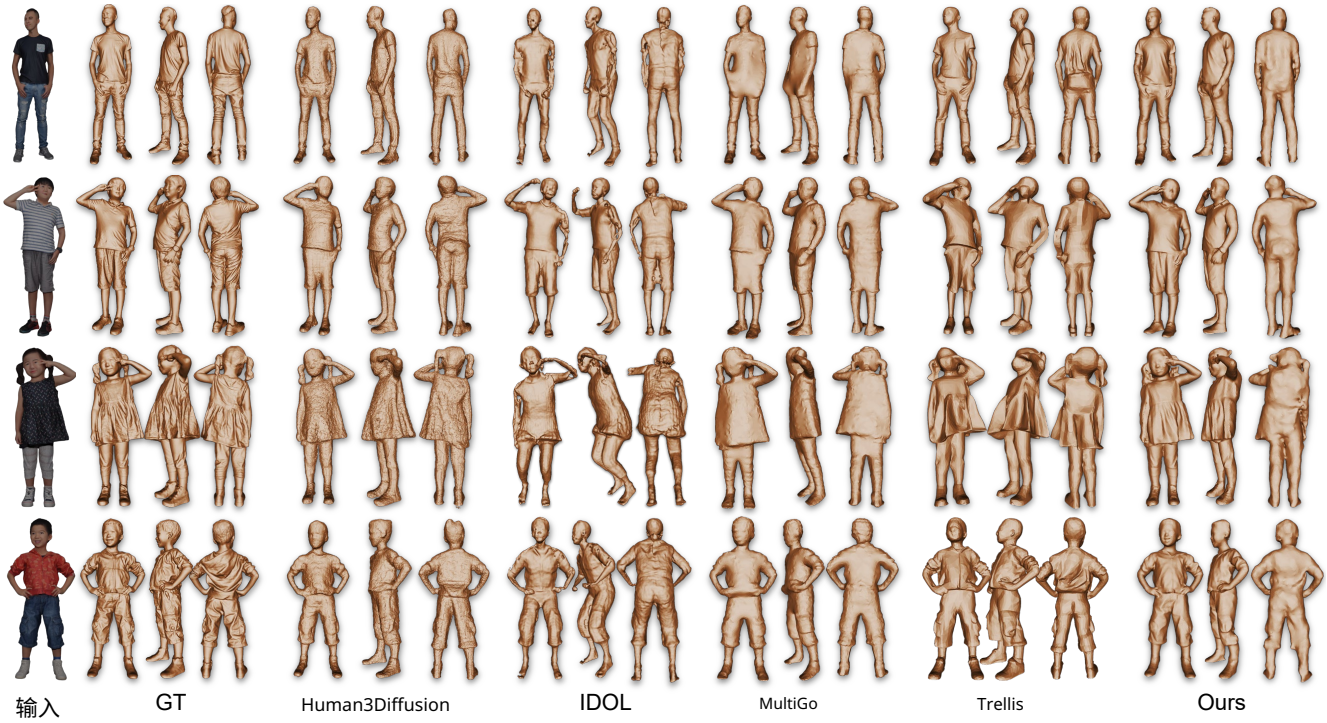

图 2. 我们的方法与基于三维高斯溅射的方法（即 Human3Diffusion [50], IDOL [63], MultiGo [55]和 Trellis [48]进行几何比较。 放大查看细节。

D. 解码模块

我们使用稀疏变分自编码器将大小为 $(512^3, 20)$ 的原始稀疏三维高斯属性压缩为大小为 $(64^3, 4)$的潜在表示，从而显著降低 GPU 内存消耗，并实现可行的训练。然而，这种激进的压缩不可避免地会导致信息丢失，而经过扩散生成过程后，信息损失会进一步加剧。为缓解这一问题，借鉴近期大规模 3D生成模型 [21], [36], [48], 的进展，我们在变分自编码器解码器之后添加了一个基于 Minkowski Engine 的 U-Net，以细化扩散模型的输出。

网格提取。 对于重建得到的每组三维高斯，我们利用其位置属性恢复人体表面。使用 WNNC [22], 估算法线，并通过屏蔽泊松重建 [16]重建表面。为进一步提升几何保真度，我们使用深度监督对重建表面进行优化：通过 PyTorch3D 将表面渲染为深度图，并针对预测的深度图计算 L1损失。

新视角合成。 对于新视角渲染，我们采用标准的 3D高斯泼溅流程，使用相应的相机参数从任意视点渲染优化后的三维高斯。

IV. 实验

A. 数据集和评估指标

数据集。 我们在 Thuman2.1 [52],2K2K [8]和 CustomHuman [10]上进行了实验。具体而言，我们使用 Thuman2.1 中的 1600 个扫描作为真实值，以准备真实值 3D 高斯属性，用于训练稀疏变分自编码器和桥接潜在扩散模型。在评估中，我们使用了来自 2K2K 的 25 个扫描和来自 CustomHuman 的 40 个扫描。此外，我们还在从互联网收集的野外图像上评估了 JGA-LBD 的泛化能力。

评估指标。 所有三维高斯泼溅结果和网格输出均归一化至立方体 $(-1,\ 1)$。对于外观重建，我们报告峰值信噪比（PSNR）、结构相似性指数（SSIM）和学习感知图像块相似度（LPIPS）。对于几何重建，我们评估倒角距离（CD）、点到表面距离（P2S）和法线误差。

B. 实现细节

稀疏变分自编码器使用 Minkowski Engine 0.5.4 进行训练，批量大小为 8，共训练 200,000 次迭代。采用 Adam优化器，学习率为 0.00035。损失权重 $\lambda_1 \sim \lambda_7$分别设置为 0.8、0.2、0.1、 5×10^{-7}、1、1、1。（为提高训练稳定性， $\mathcal{L}_{\text{Attr}}$ 和 $\mathcal{L}_{\text{Render}}$仅在 10,000 次迭代后引入。）对于桥接扩散，训练和推理均遵循原始 DDBM 设置 [62], , 其中参数 CHURN_STEP_RATIO 和 GUIDANCE 分别设置为 0.1 和 1。批量大小设置为 16，训练迭代次数为 100,000。采用 Adam优化器，学习率为 0.00035。采用 Depth Anything V2 [51] 作为深度估计的骨干网络，并选用 PIXIE [7] 来预测 SMPL模型。稀疏3D UNet 的训练

表 I 不同方法在 2K2K 和 CustomHuman 上的定量比较。最佳结果以**粗体**突出显示。 ↑：越高越好。 ↓：越低越好。

方法 \ 指标	2K2K						CustomHuman					
	峰值信噪比↑	结构相似性指数↑	学习感知图像块相似度↓	CD↓	P2S↓	法线↓	峰值信噪比↑	结构相似性指数↑	学习感知图像块相似度↓	CD↓	P2S↓	法线↓
GTA (NeurIPS 23)	24.15	0.921	0.080	1.156	1.114	2.127	28.86	0.920	0.088	1.249	1.123	2.552
SIFU (CVPR 24)	23.47	0.910	0.088	1.154	1.135	2.180	29.62	0.928	0.092	1.365	1.205	2.696
SiTH (CVPR 24)	24.30	0.920	0.076	0.891	0.944	2.019	26.47	0.911	0.095	2.244	2.367	3.365
PSHuman(计算机视觉与模式识别会议 25)	24.72	0.917	0.067	0.575	0.608	1.440	30.26	0.931	0.082	1.055	1.146	1.899
IDOL (计算机视觉与模式识别会议 25)	27.18	0.929	0.076	1.095	1.138	2.454	31.02	0.934	0.076	1.119	1.188	2.416
Human3Diffusion(神经信息处理系统大会 24)	29.05	0.942	0.062	0.503	**0.415**	1.429	**33.75**	0.952	0.067	0.809	0.768	1.755
MultiGO (计算机视觉与模式识别会议 25)	28.80	0.939	0.059	0.636	0.655	1.474	31.72	0.934	0.075	1.750	1.809	2.440
Trellis(计算机视觉与模式识别会议 25)	25.47	0.927	0.069	0.771	0.743	1.929	31.33	0.934	0.069	1.202	1.219	2.370
JGA-LBD	**30.16**	**0.946**	**0.055**	**0.489**	0.507	**1.202**	33.44	**0.957**	**0.061**	**0.674**	**0.670**	**1.469**

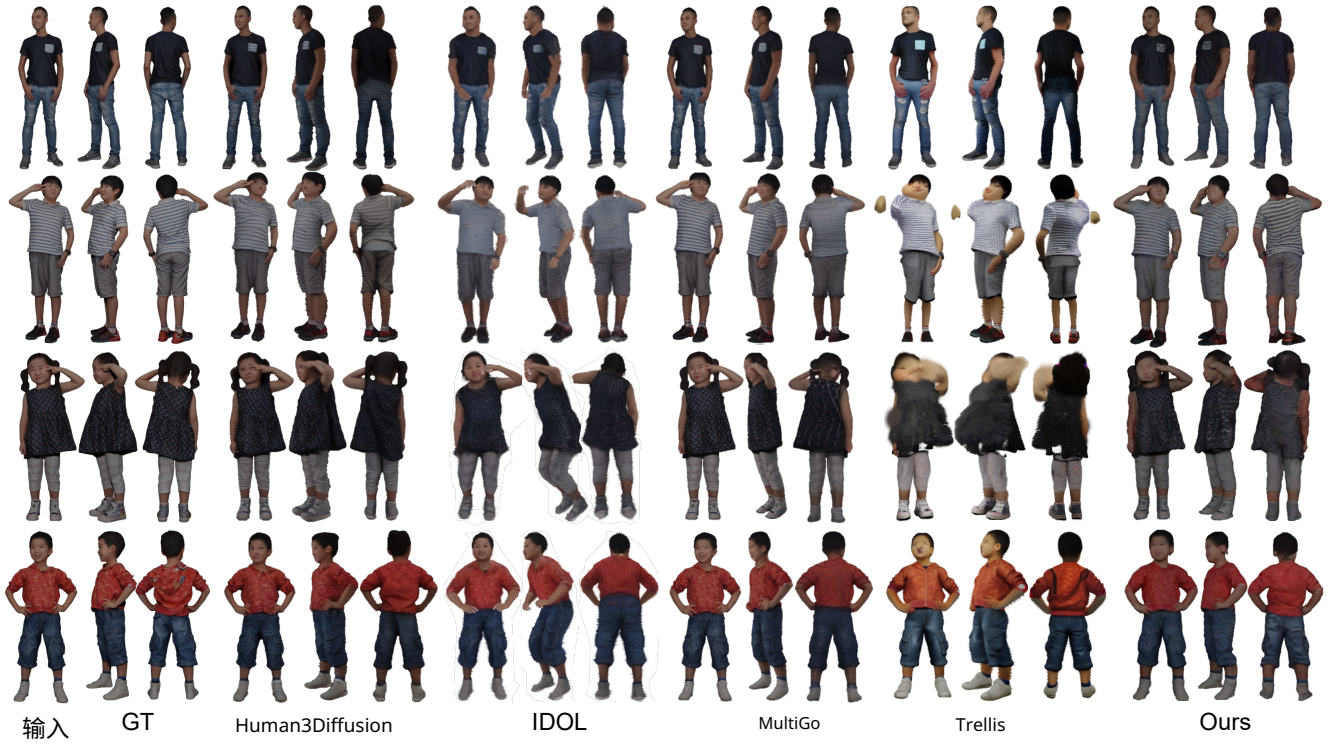

图 3. 我们的方法与基于3D高斯泼溅的方法的外观比较，即Human3Diffusion [50], IDOL [63], MultiGo [55] 和 Trellis [48]。 放大查看详细信息。

批量大小为 8，共训练 40,000 次迭代。采用 Adam优化器，学习率为 0.00035。所有训练和测试均在一台配备四块英伟达 A6000图形处理器的服务器上进行。

稀疏变分自编码器、桥接扩散三维U-Net 和稀疏 U-Net 的参数量分别为 9.04M、229.76M 和 36.99M。这些模型的训练时间分别为 2 天、4 天和 1 天。这些模型的 GPU 显存占用分别为 40 GB 、 4×45 GB 和 30 GB。单个样本的推理时间约为 2 分钟。

C. 与最先进方法的比较

我们主要将 JGA-LBD 与三种具有代表性的基于三维高斯溅射的方法进行比较。IDOL [63] 利用显式的 SMPL模型作为几何先验，以引导三维高斯的生成。MultiGo [55] 利用大型模型生成完整的三维高斯场景，同时将 SMPL 编码为傅里叶特征以提供结构指导。Trellis [48] 学习紧凑的潜在空间，将几何与外观联合编码，用于通过多个阶段进行结构化生成建模。我们还将 JGA-LBD 与近期的几种基于网格的方法进行比较 [11], [19], [59], [60]。如表I所示，JGA-LBD 在两个基准数据集的所有指标上均取得了最佳表现。

几何比较。Human3Diffusion [50] 生成的几何明显带有噪声，且无法保留精细的正面皱纹细节。IDOL [63] 严重依赖未经任何细化的SMPL模型，因此，如图 2所示，它经常生成错误的姿势。此外，由于

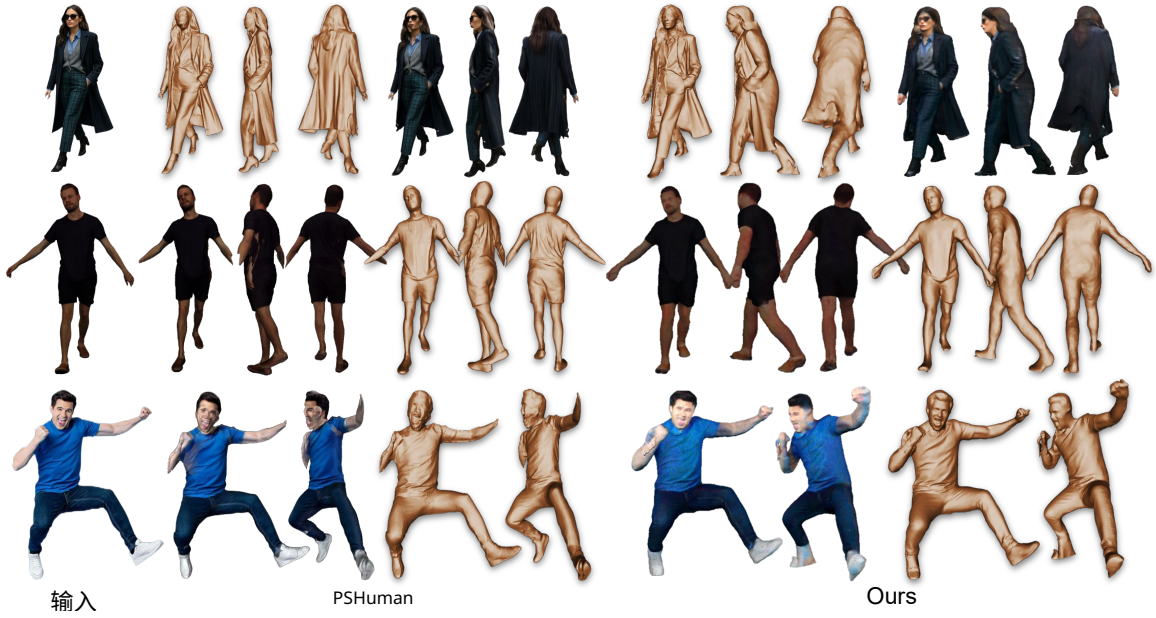

图4. 我们的方法与 PSHuman 在几何和外观方面的比较 [19]。 放大查看细节。



图5. 我们的 JGA-LBD 在野外图像上的重建结果。 放大查看细节。

由于 SMPL 施加了强正则化，因此无法处理连衣裙等宽松服装（见第三个案例）。尽管 MultiGo [55]采用了皱纹细化网络，但其几何重建仍缺乏精细细节。此外，如图2中的第三和第四个案例所示，重建的人体呈现出前倾趋势。这表明，尽管 MultiGo 避免了直接使用 SMPL 所引入的姿势误差，但其二维扩散模型不足以纠正三维空间中的姿势误差。Trellis [48]受限于较低的网格分辨率，精细细节的重建能力受到严重影响。此外，重建出的姿势通常不够准确，在所有案例中，头部旋转都与输入不一致。相比之下，我们的 JGA-LBD 能够重建精细的几何细节，并且处理宽松服装，同时保持准确的整体人体姿势。

外观比较。Human3Diffusion [50] 生成的渲染图像中存在许多锯齿状伪影，且无法准确恢复纹理细节。IDOL [63] 因 SMPL 姿势错误而出现严重错位。如图 3所示，它只能捕捉相对简单的颜色图案，无法表现第二个案例中条纹等精细纹理。此外，边缘处还可观察到明显的锯齿状伪影。MultiGo [55] 在正面表现良好，但背面重建效果仍然较差。例如，在第二个案例中，它未能恢复条纹图案；在第三个案例中，头部后侧错误地

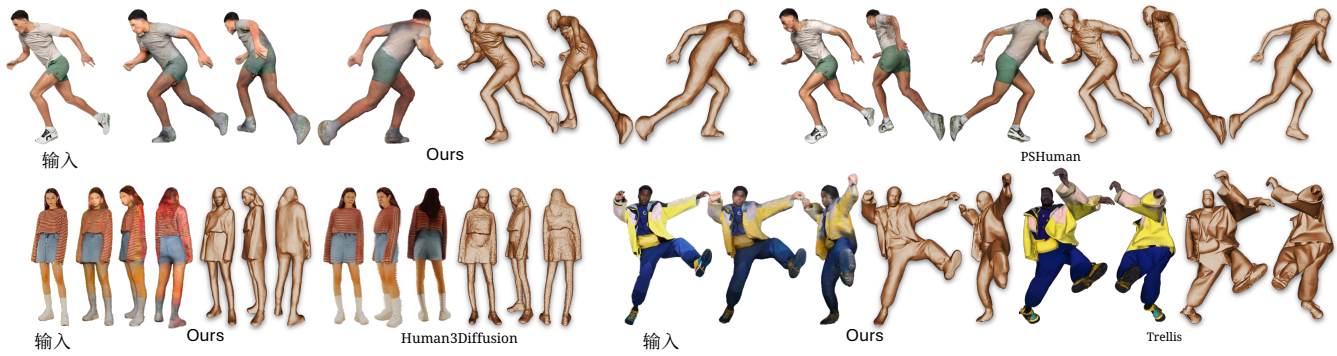

图 6. 我们的方法与 PSHuman [19], Human3Diffusion [50] 和 Trellis [48] 在野外图像上的几何与外观比较。 放大查看细节。

出现了面部细节，而不是黑色头发。除了无法重建条纹等精细细节外，Trellis [48] 还存在严重的几何与外观不一致。例如，在第二和第三个案例中，重建出的手臂与图 2 中所示的手臂明显不一致。相比之下，我们的 JGA-LBD 不仅能够重建条纹和背面皱纹等精细细节，还能保持与图 2 中参考图像的几何一致性。

与 PSHuman 的比较 [19]。 我们还将 JGA-LBD 与广受欢迎的 PSHuman 进行了比较。如图 4所示，PSHuman 经常出现身体部位缺失的问题。在第一个示例中，PSHuman 也难以处理腿部的相对位置。在第二个示例中，它同样无法保留面部细节，导致结果明显退化。相比之下，我们的方法有效避免了这些问题。

总体而言，定量指标和定性比较都充分表明，我们的 JGA-LBD 始终优于当前最先进的方法。我们还在具有挑战性姿势的野外图像上进行了实验，如图 5所示，JGA-LBD 成功重建出可信的外观和细致的三维表面。我们还在野外图像上将本方法与具有代表性的方法进行了比较，包括基于网格的 PSHuman [20], 基于三维高斯溅射的 Human3Diffusion [50], 以及通用方法 Trellis[48],, 如图 6所示。PSHuman[20] 无法保留精细的面部细节，且手部重建不准确。Human3Diffusion [50] 难以生成正确的人体姿势，且表面质量和外观细节均有所退化。作为通用重建框架，Trellis [48]，通常无法生成结构上合理的人体形状。相比之下，在复杂的真实世界条件下，我们的方法能够实现更准确的几何重建和更逼真的外观重建。请参阅随附的 **视频演示**。

D. 消融研究

稀疏变分自编码器的可视化结果。 为评估稀疏变分自编码器的有效性，我们直接从压缩后的潜在空间重建人体三维高斯表示。具体而言，输入的三维高斯从 $(512^3, 20)$ 压缩为尺寸为

表 II 2K2K 上的消融研究。

方法 \ 指标	峰值信噪比↑	结构相似性指数↑	学习感知图像块相似度↓	CD↓	P2S↓	法线↓
整流流	28.32	0.931	0.074	0.568	0.535	1.436
特征条件	Fail	Fail	Fail	Fail	Fail	Fail
无解码模块	27.68	0.931	0.073	0.498	**0.497**	1.287
EcoDepth + PIXIE	30.01	0.944	0.059	0.536	0.544	1.288
DepthAnything + PyMAF	29.51	0.943	0.059	0.540	0.543	1.367
完整模型	**30.16**	**0.946**	**0.055**	**0.489**	0.507	**1.202**



图 7. 变分自编码器结果与真实值的可视化对比。 放大查看细节。

$(64^3, 4)$，随后解码以恢复三维高斯属性。如图7所示，稀疏变分自编码器能够保留人体整体几何形状和粗略外观，这表明潜在空间能够有效编码结构和视觉信息。然而，由于压缩率较高，锐利的几何边界和高频纹理等细节明显有所损失。这一观察促使我们在变分自编码器解码器之后引入解码模块，以提高重建保真度。

桥接扩散的有效性。 我们进一步将本文采用的桥接扩散与流行的整流流方法进行比较 [25],定量结果如表 II所示。图 8 中的视觉结果表明，整流流



图8. 整流流与桥接扩散的比较。 放大查看细节。

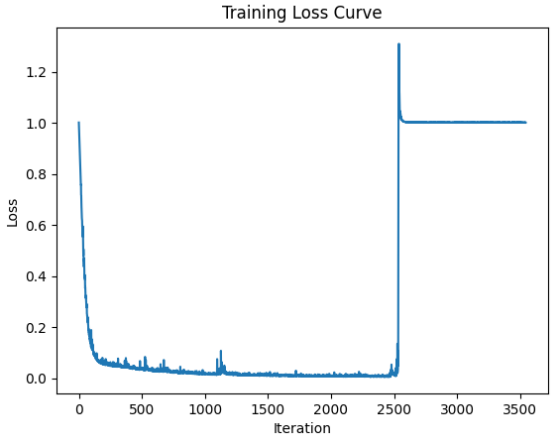

图9. 图像和 SMPL 特征级监督的训练损失曲线。 放大查看细节。

往往会生成带有许多孔洞的三维高斯，重建出的背面也受到严重噪声干扰。这一观察凸显了桥接扩散的优势：由于起始深度已包含完整三维高斯的一部分，桥接扩散无需将过多容量分配给可见的正面，而是专注于学习缺失区域。因此，我们采用桥接扩散的策略实现了更优的重建质量。

结构先验的必要性。我们进一步验证了引入结构先验的必要性。遵循 GaussianCube[53]，我们使用 DINOv 提取图像特征2 [30]，并使用 Point-M2AE 提取 SMPL 特征 [58]，，并利用这两种特征监督桥扩散模型的训练过程。然而，训练往往无法收敛：如图9所示，损失在初始阶段有所下降，但经过若干个周期后突然上升，最终导致发散。我们将此归因于图像与 SMPL 之间的模态鸿沟，这使模型难以学习有意义的表征——

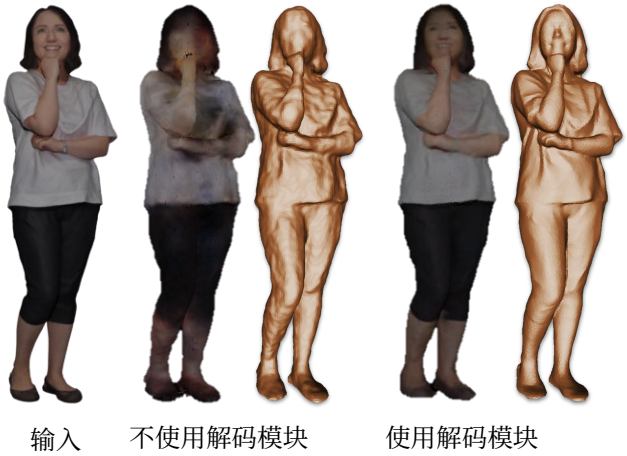

图10. 不使用和使用解码模块时视觉结果的比较。 放大查看细节。

当直接使用这类异构特征时，便难以学习有意义的表征。这一观察凸显了统一潜在空间的重要性：来自不同模态的特征被映射到共享表征中，从而显著降低训练难度并提升稳定性。

解码模块的必要性。由于变分自编码器的压缩比过高，不可避免地会引入信息损失。此外，扩散模型本身也无法实现完全无误差的重建，这些残余误差会进一步放大压缩造成的损失，使高频细节难以捕获和恢复。为解决这一问题，我们在扩散模型之后引入解码模块，以细化和增强生成结果。如表II所示，解码模块在外观和几何指标方面均带来了显著提升。从图10中还可以观察到，外观和几何细节均得到增强。此外，值得注意的是，即使不使用额外的解码模块，结果在几何方面已经达到最佳性能，并且外观重建效果仍具备与最先进的方法相竞争的水平。为量化不同深度估计方法的影响，我们比较了两种深度估计器——Depth Anything V2 [51]（在 2K2K 上的均方根误差为 0.014）和 EcoDepth [33]（在 2K2K 上的均方根误差为 0.016）—— 并观察到，更好的深度能够持续带来更优的重建性能（表II）。我们还评估了不同的 SMPL 回归器（即 PIXIE [7] 与 PyMAF [56]），发现它们产生的结果相当，仅存在细微差异。

E. 局限性

我们在图11中展示了两类具有代表性的失败案例。第一项局限源于高质量三维人体扫描数据的有限可用性。由于多样化训练样本的缺乏，所学习的潜在空间无法充分覆盖复杂的服装分布，从而限制了模型的泛化能力。如图11所示，尽管我们的方法能够成功重建出合理的人体，并从



图 11. JGA-LBD 局限性的示意图。

正面视角生成视觉上合理的结果，但无法恢复宽松服装（如长裙）的正确体积结构。具体而言，裙子的背面本应呈现出蓬松的形状，而重建结果却错误地向腿部塌陷。这一局限源于训练数据中此类服装风格的表示不足，使模型无法学习其特有的几何分布。第二类失败案例涉及外观重建。在某些情况下，模型无法忠实地恢复输入图像中明确可见的正面颜色信息。我们认为，这一问题主要归因于训练数据在规模和多样性方面的不足。具体而言，高质量三维人体扫描的稀缺限制了所学习VAE潜在空间的表示能力，进而限制了其对复杂场景或不常见外观配置的泛化能力，最终影响后续扩散模型的性能。

V. 结论

在本研究中，我们提出了 JGA-LBD，这是一种能够在单次生成过程中同时重建人体几何与外观的框架。实验结果表明，与最先进的方法相比，我们的方法在几何和外观重建方面均表现出更优的性能，并且能够很好地处理具有高难度姿势和宽松服装的野外图像。不同于将几何与外观分开处理的现有方法，JGA-LBD 采用联合建模，从而确保几何与外观之间具有更好的一致性。

未来，我们将利用功能更强大的稀疏变分自编码器来捕捉高频细节，并探索无需额外细化解码器的扩散架构，以进一步提升效率和重建质量。此外，我们还将研究更有效的策略，以从有限的三维人体扫描中学习具有判别性且鲁棒的潜在空间，旨在更好地支持真实世界挑战性条件下的基于扩散的生成。

参考文献

[1] Badour AlBahar、Shunsuke Saito、Hung-Yu Tseng、Changil Kim、Jo- hannes Kopf 和 Jia-Bin Huang。利用形状引导扩散进行单图像三维人体数字化。载于*SIGGRAPH Asia 2023 Conference Papers*，第 1–11 页，2023年。

[2] Stephen Batifol、Andreas Blattmann、Frederic Boesel、Saksham Consul、Cyril Diagne、Tim Dockhorn、Jack English、Zion English、Patrick Esser、Sumith Kulal 等。Flux. 1 Kontext：潜在空间中的上下文图像生成与编辑的流匹配。arXiv电子印本，页码 arXiv–2506，2025年。[3] Jinnan Chen、Chen Li、Jianfeng Zhang、Lingting Zhu、Buzhen Huang、Hanlin Chen 和 Gim Hee Lee。基于单视图图像的可泛化人体高斯表示。国际学习表征会议，2025年。[4] Christopher Choy、JunYoung Gwak 和 Silvio Savarese。4D时空卷积网络：闵可夫斯基卷积神经网络。载于 IEEE/CVF计算机视觉与模式识别会议，页码 3075–3084，2019年。[5] Prafulla Dhariwal 和 Alexander Nichol。扩散模型在图像合成方面胜过生成对抗网络。神经信息处理系统进展，34：8780–8794，2021年。[6] Joseph L Doob 和 JI Doob。经典势理论及其概率对应理论，第262卷。1984年。[7] Yao Feng、Vasileios Choutas、Timo Bolkart、Dimitrios Tzionas 和 Michael J Black。使用调节机制对富有表现力的人体进行协同回归。载于国际三维视觉会议，页码 804–792，2021年。[8] Sang-Hun Han、Min-Gyu Park、Ju Hong Yoon、Ju-Mi Kang、Young-Jae Park 和 Hae-Gon Jeon。基于单张2K分辨率图像的高保真三维人体数字化。载于 IEEE/CVF计算机视觉与模式识别会议，页码 12869–12879，2023年。[9] Xu He、Zhiyong Wu、Xiaoyu Li、Di Kang、Chaopeng Zhang、Jiangnan Ye、Liyang Chen、Xiangjun Gao、Han Zhang 和 Haolin Zhuang。MagicMan：利用三维感知扩散和迭代细化进行人体新视角生成。载于 AAAI，第39卷，页码 3437–3445，2025年。[10] Hsuan-I Ho、Lixin Xue、Jie Song 和 Otmar Hilliges。学习局部可编辑的虚拟人。载于 IEEE/CVF计算机视觉与模式识别会议，页码 21024–21035，2023年。[11] I Ho、Jie Song、Otmar Hilliges 等。SiTH：基于图像条件扩散的单视图带纹理人体重建。载于 IEEE/CVF计算机视觉与模式识别会议，页码 538–549，2024年。[12] Jonathan Ho、Ajay Jain 和 Pieter Abbeel。去噪扩散概率模型。神经信息处理系统进展，33：6840–6851，2020年。[13] Jonathan Ho 和 Tim Salimans。无分类器扩散引导。神经信息处理系统进展，2021年。[14] Shoukang Hu、Takuya Narihira、Kazumi Fukuda、Ryosuke Sawata、Takashi Shibuya 和 Yuki Mitsufuji。Humangif：带有生成先验的单视图人体扩散。arXiv预印本 arXiv:2502.12080，2025年。[15] Yangyi Huang、Hongwei Yi、Yuliang Xiu、Tingting Liao、Jiaxiang Tang、Deng Cai 和 Justus Thies。TeCH：文本引导的逼真着装人体重建。载于国际三维视觉会议，页码 1531–1542，2024年。[16] Michael Kazhdan 和 Hugues Hoppe。屏蔽泊松表面重建。ACM图形学汇刊，32：1–13，2013年。[17] Bernhard Kerbl、Georgios Kopanas、Thomas Leimk¨uhler 和 George Drettakis。用于实时辐射场渲染的3D高斯泼溅。ACM图形学汇刊，42(4)：139–1，2023年。[18] Bo Li、Kaitao Xue、Bin Liu 和 Yu-Kun Lai。BBDM：利用布朗桥扩散模型进行图像到图像的转换。载于 IEEE/CVF计算机视觉与模式识别会议，页码 1952–1961，2023年。[19] Peng Li、Wangguandong Zheng、Yuan Liu、Tao Yu、Yangguang Li、Xingqun Qi、Xiaowei Chi、Siyu Xia、Yan-Pei Cao、Wei Xue 等。PSHuman：利用跨尺度多视图扩散和显式重新网格化进行照片级真实感单图三维人体重建。载于 IEEE/CVF计算机视觉与模式识别会议，页码 16008–16018，2025年。[20] Peng Li、Wangguandong Zheng、Yuan Liu、Tao Yu、Yangguang Li、Xingqun Qi、Mengfei Li、Xiaowei Chi、Siyu Xia、Wei Xue 等。PSHuman：利用跨尺度扩散进行照片级真实感单视图人体重建。载于 IEEE/CVF计算机视觉与模式识别会议，2024年。[21] Weiyu Li、Jiarui Liu、Hongyu Yan、Rui Chen、Yixun Liang、Xuelin Chen、Ping Tan 和 Xiaoxiao Long。CraftsMan3D：利用三维原生生成和交互式几何细化器进行高保真网格生成。IEEE/CVF计算机视觉与模式识别会议，2025年。[22] Siyou Lin、Zuoqiang Shi 和 Yebin Liu。基于绕数法和法线一致性的快速全局一致法线定向。ACM图形学汇刊，43(6)：1–19，2024年。[23] Yaron Lipman、Ricky TQ Chen、Heli Ben-Hamu、Maximilian Nickel、

和 Matt Le。用于生成建模的流匹配。国际学习表征会议，2023年。[24] Ruoshi Liu、Rundi Wu、Basile Van Hoorick、Pavel Tokmakov、Sergey Zakharov 和 Carl Vondrick。Zero-1-to-3：单张图像零样本生成三维物体。载于 IEEE/CVF国际计算机视觉会议，页码 9298–9309，2023年。[25] Xingchao Liu、Chengyue Gong 和 Qiang Liu。流直且快：利用整流流学习生成和迁移数据。国际学习表征会议，2022年。[26] Shitong Luo 和 Wei Hu。用于三维点云生成的扩散概率模型。载于 IEEE/CVF计算机视觉与模式识别会议，页码 2837–2845，2021年。[27] Andrew Melnik、Michal Ljubljanac、Cong Lu、Qi Yan、Weiming Ren 和 Helge Ritter。视频扩散模型：综述。arXiv预印本 arXiv:2405.03150，2024年。[28] Ben Mildenhall、Pratul P Srinivasan、Matthew Tancik、Jonathan T Barron、Ravi Ramamoorthi 和 Ren Ng。NeRF：将场景表示为用于视图合成的神经辐射场。载于欧洲计算机视觉会议，页码 99–106，2020年。[29] Ryota Natsume、Shunsuke Saito、Zeng Huang、Weikai Chen、Chongyang Ma、Hao Li 和 Shigeo Morishima。SiCloPe：基于轮廓的着装人物。载于 IEEE/CVF计算机视觉与模式识别会议，页码 4480–4490，2019年。[30] Maxime Oquab、Timoth´ee Darcet、Th´eo Moutakanni、Huy Vo、Marc Szafraniec、Vasil Khalidov、Pierre Fernandez、Daniel Haziza、Francisco Massa、Alaaeldin El-Nouby 等。DINOv2：无需监督学习鲁棒视觉特征。arXiv预印本 arXiv:2304.07193，2023年。[31] Werner Palfinger。用于逆向渲染的连续重网格化。《计算机动画与虚拟世界》，33(5)：e2101，2022年。[32] Panwang Pan、Zhuo Su、Chenguo Lin、Zhen Fan、Yongjie Zhang、Zeming Li、Tingting Shen、Yadong Mu 和 Yebin Liu。HumanSplat：具有结构先验的可泛化单图像人体高斯泼溅。神经信息处理系统进展，37：74383–74410，2024年。[33] Suraj Patni、Aradhye Agarwal 和 Chetan Arora。EcoDepth：用于单目深度估计的扩散模型有效条件控制。载于 IEEE/CVF计算机视觉与模式识别会议，页码 28285–28295，2024年。[34] William Peebles 和 Saining Xie。基于变换器的可扩展扩散模型。载于 IEEE/CVF国际计算机视觉会议，页码 4195–4205，2023年。[35] Lingteng Qiu、Xiaodong Gu、Peihao Li、Qi Zuo、Weichao Shen、Junfei Zhang、Kejie Qiu、Weihao Yuan、Guanying Chen、Zilong Dong 等。LHM：数秒内从单张图像重建可动画人体的大型模型。IEEE/CVF国际计算机视觉会议，2025年。[36] Xuanchi Ren、Jiahui Huang、Xiaohui Zeng、Ken Museth、Sanja Fidler 和 Francis Williams。Xcube：利用稀疏体素层级结构进行大规模三维生成建模。载于 IEEE/CVF计算机视觉与模式识别会议，页码 4209–4219，2024年。[37] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer。利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，页码 10684–10695，2022年。[38] Shunsuke Saito、Zeng Huang、Ryota Natsume、Shigeo Morishima、Angjoo Kanazawa 和 Hao Li。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，页码 2304–2314，2019年。[39] Shunsuke Saito、Tomas Simon、Jason Saragih 和 Hanbyul Joo。PI-FuHD：用于高分辨率三维人体数字化的多层级像素对齐隐式函数。载于 IEEE/CVF计算机视觉与模式识别会议，页码 84–93，2020年。[40] Jiaxiang Tang、Zhaoxi Chen、Xiaokang Chen、Tengfei Wang、Gang Zeng 和 Ziwei Liu。LGM：用于高分辨率三维内容创作的大型多视图高斯模型。载于欧洲计算机视觉会议，页码 1–18，2024年。[41] Jiaxiang Tang、Jiawei Ren、Hang Zhou、Ziwei Liu 和 Gang Zeng。DreamGaussian：用于高效三维内容创作的生成式高斯泼溅。国际学习表征会议，2024年。[42] Sicong Tang、Guangyuan Wang、Qing Ran、Lingzhi Li、Li Shen 和 Ping Tan。着装人体的高分辨率体积重建。ACM图形学汇刊，2023年。[43] Yingzhi Tang、Qijian Zhang 和 Junhui Hou。HuGDiffusion：通过三维高斯扩散实现可泛化的单图像人体渲染。IEEE可视化与计算机图形学汇刊，2025年。[44] Yingzhi Tang、Qijian Zhang、Yebin Liu 和 Junhui Hou。人体即点：基于显式点的单视图RGB图像三维人体重建。《IEEE模式分析与机器智能汇刊》，2025年。[45] Haochen Wang、Xiaodan Du、Jiahao Li、Raymond A Yeh 和 Greg Shakhnarovich。Score Jacobian Chaining：将预训练二维扩散模型提升至三维生成。在*IEEE/CVF*计算机视觉与模式识别会议上发表，第12619–12629页，2023年。[46] Wentao Wang、Hang Ye、Fangzhou Hong、Xue Yang、Jianfu Zhang、Yizhou Wang、Ziwei Liu 和 Liang Pan。GeneMAN：基于多源人体数据的可泛化单图像三维人体重建。2025年。[47] Zilong Wang、Zhiyang Dou、Yuan Liu、Cheng Lin、Xiao Dong、Yunhui Guo、Chenxu Zhang、Xin Li、Wenping Wang 和 Xiaohu Guo。Wonder-human：在动态三维人体重建中幻觉生成不可见部位。*IEEE*可视化与计算机图形学汇刊，2025年。[48] Jianfeng Xiang、Zelong Lv、Sicheng Xu、Yu Deng、Ruicheng Wang、Bowen Zhang、Dong Chen、Xin Tong 和 Jiaolong Yang。用于可扩展多功能三维生成的结构化三维潜变量。在*IEEE/CVF*计算机视觉与模式识别会议上发表，第21469–21480页，2025年。[49] Yuliang Xiu、Jinlong Yang、Xu Cao、Dimitrios Tzionas 和 Michael J Black。ECON：从法线获得的显式着衣人体。*IEEE/CVF*计算机视觉与模式识别会议，2023年。[50] Yuxuan Xue、Xianghui Xie、Riccardo Marin 和 Gerard Pons-Moll。Human 3Diffusion：通过显式三维一致性扩散模型创建逼真化身。在神经信息处理系统进展上发表，2024年。[51] Lihe Yang、Bingyi Kang、Zilong Huang、Zhen Zhao、Xiaogang Xu、Jiashi Feng 和 Hengshuang Zhao。Depth Anything V2。神经信息处理系统进展，37:21875–21911，2024年。[52] Tao Yu、Zerong Zheng、Kaiwen Guo、Pengpeng Liu、Qionghai Dai 和 Yebin Liu。Function4D：利用极稀疏消费级RGBD传感器进行实时人体体积捕获。在*IEEE/CVF*计算机视觉与模式识别会议上发表，第5746–5756页，2021年。[53] Bowen Zhang、Yiji Cheng、Jiaolong Yang、Chunyu Wang、Feng Zhao、Yansong Tang、Dong Chen 和 Baining Guo。GaussianCube：用于三维生成建模的结构化显式辐射表示。在神经信息处理系统进展上发表，2024年。[54] Biao Zhang、Jiapeng Tang、Matthias Niessner 和 Peter Wonka。3DShape2VecSet：用于神经场和生成式扩散模型的三维形状表示。*ACM*图形学汇刊，42 16，2023 4 1( ): – 。[55] Gangjian Zhang、Nanjie Yao、Shunsi Zhang、Hanfeng Zhao、Guoliang Pang、Jian Shu 和 Hao Wang。MultiGo：面向单目三维纹理人体重建的多层级几何学习。在*IEEE/CVF*计算机视觉与模式识别会议上发表，第338–347页，2025年。[56] Hongwen Zhang、Yating Tian、Xinchi Zhou、Wanli Ouyang、Yebin Liu、Limin Wang 和 Zhenan Sun。PyMAF：采用金字塔网格对齐反馈回路的三维人体姿态与形状回归。在*IEEE/CVF*国际计算机视觉会议上发表，第11446–11456页，2021年。[57] Lvmin Zhang、Anyi Rao 和 Maneesh Agrawala。为文生图扩散模型添加条件控制。在*IEEE/CVF*国际计算机视觉会议上发表，第3836–3847页，2023年。[58] Renrui Zhang、Ziyu Guo、Peng Gao、Rongyao Fang、Bin Zhao、Dong Wang、Yu Qiao 和 Hongsheng Li。Point-M2AE：用于层次化点云预训练的多尺度掩码自编码器。神经信息处理系统进展，35:27061–27074，2022年。[59] Zechuan Zhang、Li Sun、Zongxin Yang、Ling Chen 和 Yi Yang。用于着装化身重建的全局相关三维解耦变换器。在神经信息处理系统进展上发表，2023年。[60] Zechuan Zhang、Zongxin Yang 和 Yi Yang。SIFU：侧视图条件隐式函数，用于重建可在现实世界中使用的着装人体。在*IEEE/CVF*计算机视觉与模式识别会议上发表，第9936–9947页，2024年。[61] Junsheng Zhou、Weiqi Zhang 和 Yu-Shen Liu。DiffGS：功能性高斯泼溅扩散。在神经信息处理系统进展上发表，2024年。[62] Linqi Zhou、Aaron Lou、Samar Khanna 和 Stefano Ermon。去噪扩散桥模型。国际学习表征会议，2024年。[63] Yiyu Zhuang、Jiaxi Lv、Hao Wen、Qing Shuai、Ailing Zeng、Hao Zhu、Shifeng Chen、Yujiu Yang、Xun Cao 和 Wei Liu。IDOL：从单张图像即时创建逼真人体。在*IEEE/CVF*计算机视觉与模式识别会议上发表，第26308–26319页，2025年。