arXiv:2301.01161v1 [cs.CV] 3 Jan 2023

用于计算机视觉的程序化人体

Charlie Hewitt、Tadas Baltrušaitis、ErrollWood、Lohit Petikam、Louis Florentin 和 Hanz Cuevas Velasquez

Mesh Labs Cambridge、微软

2023年1月

近期研究表明，合成数据用于计算机视觉具有诸多优势，应用范围从自动驾驶 [17, 18] 到面部标志点检测 [20] 和重建 [19]。使用合成数据具有诸多益处，包括隐私保护和偏差消除 [1, 12] ，以及标注的质量和可行性 [19]。就真实感和领域差距而言，生成以人为中心的合成数据尤其具有挑战性，不过近期研究表明，仅使用合成人脸数据也能训练出有效的机器学习模型 [20]。我们在 Wood 等人的流程基础上进行构建，证明这一方法可以扩展到包含全身 [20] ，从而生成完整人体的合成图像，并为计算机视觉应用提供真实标注。

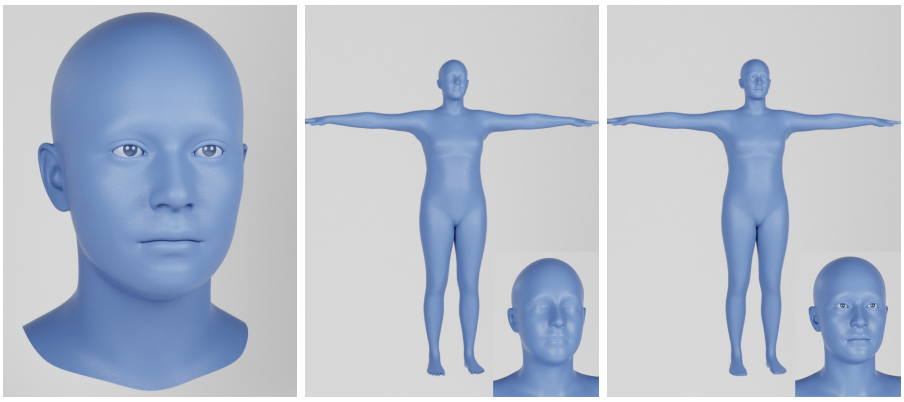
本报告介绍了我们如何构建包含关节式手部的人体面部参数化模型；如何基于该人体模型，通过渲染流程生成逼真的人体图像；如何训练深度神经网络回归覆盖全身的密集关键点集；以及如何将人体模型拟合到从多个视角预测的密集关键点。

1 形状模型

1.1 模型构建

我们的人体模型结合了 Wood 等人的高保真面部模型 [20] 与广受欢迎的 SMPL-H 人体和手部模型 [14], 而 SMPL-H 本身结合了 MANO 的关节式手部 [14] 与 SMPL 人体模型 [9]。因此，我们获得了一个完整的人体参数化模型，既能像 SMPL-H 一样控制身体形状和姿态 [14], ，也能像 Wood 等人一样控制面部、身份和表情。[20],参见图 1。

人体网格由 $N = 12943$ 个顶点和 12726 个多边形组成，骨架包含 $K = 54$ 个关节：人体有 22 个（SMPL 骨架 [9]），每只手有 15 个（与 MANO/SMPL-H 中相同 [14]），眼睛有 2 个。



(a) Wood 等人的头部模型等人 [20]　(b) SMPL-H 人体模型 [14]　(c) 我们组合的人体和头部模型

图 1：用于构建我们组合形状模型（右）的各组成模型模板网格（左侧两个），全身模型中的头部区域以局部放大图显示。与 SMPL-H 人体模型相比，我们的组合模型在面部具有显著更高的保真度 [14]。

人体网格顶点位置由网格生成函数$\mathcal{M}(\vec{\gamma},\ \vec{\beta},\ \vec{\psi},\ \vec{\theta})$定义：$\mathbb{R}^{|\vec{\gamma}|+|\vec{\beta}|+|\vec{\psi}|+|\vec{\theta}|} \rightarrow \mathbb{R}^{N\times 3}$该函数以参数 $\vec{\gamma} \in \mathbb{R}^{|\vec{\gamma}|}$表示面部身份，$\vec{\beta} \in \mathbb{R}^{|\vec{\beta}|}$表示身体身份，$\vec{\psi} \in \mathbb{R}^{|\vec{\psi}|}$表示表情，$\vec{\theta} \in \mathbb{R}^{K\times 3}$表示骨骼姿势。

$$\mathcal{M}(\vec{\gamma},\vec{\beta},\vec{\psi},\vec{\theta}) = \mathcal{L}(\mathcal{T}(\vec{\gamma},\vec{\beta},\vec{\psi},\vec{\theta}),\vec{\theta},\mathcal{J}(\vec{\gamma},\vec{\beta});\mathbf{W})$$

其中 $\mathcal{L}(\mathbf{X}\ \theta,\mathbf{J};\mathbf{W})$ 是一个标准的线性混合蒙皮（LBS）函数，它根据局部关节旋转 $\vec{\theta}$，围绕关节位置 $\mathbf{J} \in \mathbb{R}^{K\times 3}$ 旋转顶点位置 $\mathbf{X} \in \mathbb{R}^{N\times 3}$；每个顶点的手工设定蒙皮权重 $\mathbf{W} \in \mathbb{R}^{K\times N}$决定了旋转如何在网格上插值。

$\mathcal{T}(\vec{\gamma},\vec{\beta},\vec{\psi},\vec{\theta}) : \mathbb{R}^{|\vec{\beta}|+|\vec{\gamma}|+|\vec{\psi}|+|\vec{\theta}|} \rightarrow \mathbb{R}^{N\times 3}$通过向模板网格 $\mathbf{T} \in \mathbb{R}^{N\times 3}$添加位移，构建未摆姿态的人体网格。该模板网格表示处于$\overline{\mathrm{T}}$姿势且具有中性表情的平均人体：

$$\mathcal{T}(\vec{\gamma},\vec{\beta},\vec{\psi},\vec{\theta})^j_k = \overline{T}^j_k + \gamma_i S^{ij}_k + \beta_i U^{ij}_k + \psi_i E^{ij}_k + P(\vec{\theta})^j_k$$

给定线性面部身份基 $\mathbf{S} \in \mathbb{R}^{|\vec{\gamma}|\times N\times 3}$、身体身份基 $\mathbf{U} \in \mathbb{R}^{|\vec{\beta}|\times N\times 3}$、表情基 $\mathbf{E} \in \mathbb{R}^{|\vec{\psi}|\times N\times 3}$以及 $P(\vec{\theta})$，后者表示依赖姿态-$\vec{\theta}$的混合形状偏移，对应姿势参数 θ（更多详情请参见 SMPL [9]）。请注意，此定义及下文中使用了爱因斯坦求和约定。最后，$\mathcal{J}(\vec{\gamma},\vec{\beta}) : \mathbb{R}^{|\vec{\gamma}|+|\vec{\beta}|} \rightarrow \mathbb{R}^{K\times 3}$会移动关节位置，以补偿身份变化：

$$\mathcal{J}(\vec{\gamma},\vec{\beta})^j_k = J(\overline{T}^j_k + \gamma_i S^{ij}_k + \beta_i U^{ij}_k)$$

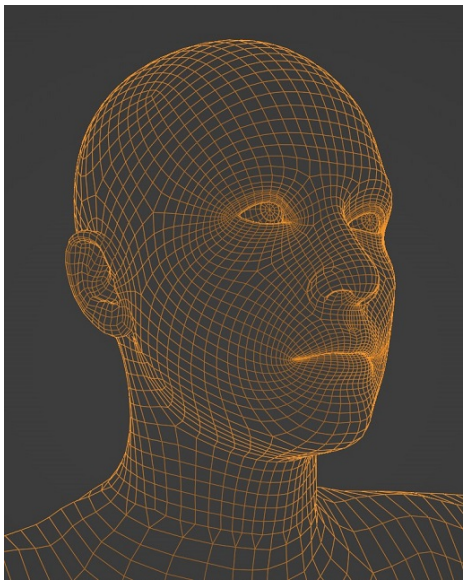


图 2：我们的人体模型的头部和颈部拓扑。

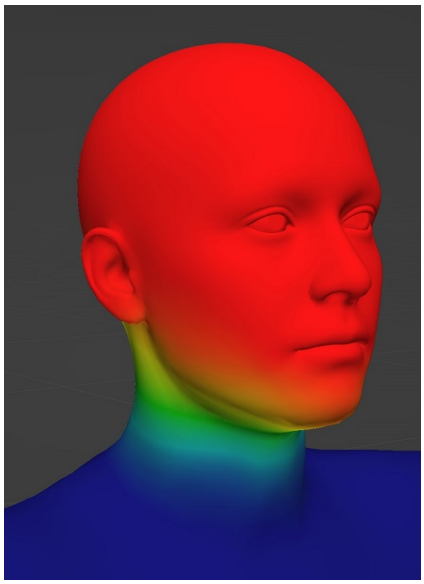


图 3：用于混合基的头部蒙版。

其中， J 是 SMPL 关节回归器的一个修改版本，在 SMPL模型的组成过程中经过了学习。

面部身份、$\mathbf{S}$和表情$\mathbf{E}$ 基来自 Wood 等人。 [20],人体身份基$\mathbf{U}$ 来自中性SMPL-H模型，该模型是根据人体扫描数据学习得到的主成分分析基。姿态相关混合形状同样取自中性SMPL-H模型。

面部身份基的大小为 $|\vec{\gamma}| = 260$ ，人体身份基的大小为 $|\vec{\beta}| = 300$。表情基包含 $|\vec{\psi}| = 224$ 个分量。

1.1.1 模板网格

为了构建模板网格，T我们手动将 Wood 等人的模板 [20]与 SMPL 模板的头部对齐。对齐后，移除 SMPL 的头部和新头部的下颈部，并将两个部分网格合并。考虑到两个网格密度不同，我们手工构建了连接处周围的拓扑，以形成平滑过渡，见图 2。

1.1.2 UV空间

为了创建 UV空间，我们以 SMPL 的 UV空间为基础，手动对齐新头部顶点的 UV坐标，使其与 SMPL UV空间中头部原有的边界和特征相匹配。这意味着 SMPL 与我们的模型在功能上具有相同的 UV空间，因此可以直接重用纹理

。SMPL 中不存在的元素，例如眼睛和口腔部件，则被添加到 UV空间中此前未使用的区域。

1.1.3 基底迁移

鉴于我们对网格拓扑所做的改动相较于 Wood 等人 [20] 和 SMPL 的模型都非常显著，因此需要调整这些源模型关联的所有基底，使其适用于我们的模型。这些基底包括面部和人体身份、表情、姿态相关混合形状、蒙皮权重以及关节回归器。

我们计算一个映射函数 $\mathcal{F}_Q : \mathbb{R}^{|Q|\times m} \to \mathbb{R}^{N\times m}$，它将给定拓扑 Q 的顶点数据转换为我们模型的拓扑，其中 $|Q|$ 是模型 Q 中的顶点数。具体而言，我们计算 $\mathcal{F}_{head}$ 和 $\mathcal{F}_{smpl}$。

映射函数 $\mathcal{F}_Q$ 的确定方法是：针对模板网格 $\mathbf{T}$ 中的每个顶点，找到模型 Q 的模板网格表面上距离最近的点。然后，将该点在所属三角形上的重心坐标存储下来。这样，我们便可以通过对该三角形各顶点在 Q 中的数据按重心坐标加权求和，来转换模型中给定顶点的输入数据。这种方法之所以有效，是因为所有模型的所有数据（如身份基底、姿态基底等）都是按顶点存储的。

然后，我们可以将此映射函数直接应用于基底，将它们从原始头部或人体模型映射到我们的模型中，例如面部身份基：$\mathbf{S} = \mathcal{F}_{head}(\mathbf{S}_{head})$。通过创建一个掩码来映射顶点组：组内顶点对应的位置设为 1，不在组内的位置设为 0，然后将 $\mathcal{F}$ 应用于此掩码。如果某个顶点映射后的值高于某个阈值，则该顶点会被判定为属于我们模型中的顶点组。

为了避免头部身份基底影响颈部下方区域，我们使用掩码将其限制为仅作用于头部。同样，为了避免 SMPL 身份基底和依赖姿态的混合形状影响头部，我们使用掩码将它们限制为仅作用于人体。为了保留身体身份导致的头部位置变化，我们对掩码区域内的 SMPL 身份基底取平均值，并将其均匀应用于所有被掩码的顶点。该掩码通过以下方式构建：将 SMPL 头部关节的蒙皮权重映射到我们的模型，并加入 Wood 等人的眼睛和口腔部件。 [20],参见 图 3。

仅使用 $\mathcal{F}$ 将 SMPL 关节回归器映射到我们的模型中也不够，因为逐关节回归器的权重之和必须恰好为一。我们改为在两个模板网格上寻找彼此最近的顶点对，据此计算从 SMPL 到我们模型的一一顶点映射。由于我们的模型基于 SMPL，除头部和颈部关节外，其余关节都能精确匹配；对于头部和颈部关节，我们也能得到非常接近的近似结果。由于我们为眼睛添加了额外的关节，因此还需要构建眼部关节回归器。具体做法是，分别沿 x 和 y 方向取每个眼球的四个极端点。

同样，需要覆盖眼球的蒙皮权重，移除其他关节的所有影响，并将相应眼部关节对所有眼球顶点的影响设为 1。口腔部件的蒙皮权重也会被覆盖，使其完全跟随头部关节，因为上述映射函数可能会导致颈部关节产生一定影响。

1.2 身份采样

为采样面部身份，我们根据 Wood 等人的训练集中男性、女性及所有身份拟合多元高斯分布。 [20] （所有性别标签均为自我报告）。这样我们就能随机采样男性、女性和中性（非二元性别）的面部身份。

对于身体身份，SMPL有三个版本的模型：男性、女性和中性；但我们的模型只使用一个中性模型。因此，我们需要能够将性别化模型的形状参数转换为中性模型的形状参数：对于性别化参数 $\vec{\beta}_g$，找到中性参数 $\vec{\beta}_n$，使得 SMPL$_g$($\vec{\beta}_g$) $=$ SMPL$_n$(β_n)，其中 SMPL$_g$ 是性别为 g 的 SMPL 模型的网格生成函数。

因此，对于模板顶点 $\mathbf{T}$、形状基 $\mathbf{S}$ 和形状参数 β，我们希望找到模板偏移量 $\vec{o}$ 和身份映射 $\mathbf{M}$，使得

$$\mathbf{T}_g + \vec{\beta}_g \cdot \mathbf{S}_g = \mathbf{T}_n + (\vec{o}_g + \vec{\beta}_g \cdot \mathbf{M}_g) \cdot \mathbf{S}_n$$

为此，我们将问题分解为两个子问题，分别求解以下各项的最小二乘解:

$$\begin{aligned} \mathbf{T}_g &= \mathbf{T}_n + \vec{o}_g \cdot \mathbf{S}_n \\ \mathbf{T}_g - \mathbf{T}_n &= \vec{o}_g \cdot \mathbf{S}_n \\ \vec{o} &= \mathbf{S}_n / (\mathbf{T}_g - \mathbf{T}_n) \end{aligned}$$

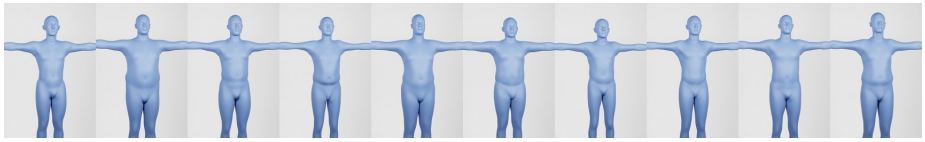
and

$$\begin{aligned} \vec{\beta}_g \cdot \mathbf{S}_g &= \vec{\beta}_g \cdot \mathbf{M}_g \cdot \mathbf{S}_n \\ \mathbf{S}_g &= \mathbf{M}_g \cdot \mathbf{S}_n \\ M_g^i &= \mathbf{S}_n / S_g^i \end{aligned}$$

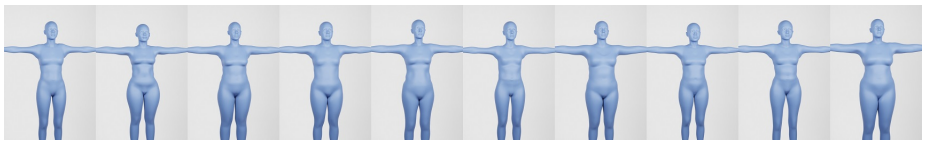
对于后者，我们分别求解形状基的每个元素， S_g^i 。

因此，有了这些映射信息 $\vec{o}_g$ 和 $\mathbf{M}_g$，对于男性和女性的 SMPL 模型（ $g \in [m, f]$） ，我们只需对单位正态分布采样即可获得身份$\vec{\beta}_g$；对于给定的性别 g，再将其转换为中性身份（可供我们的模型使用）。 $\vec{\beta}_n = \vec{o}_g + \beta_g \cdot \mathbf{M}_g$。

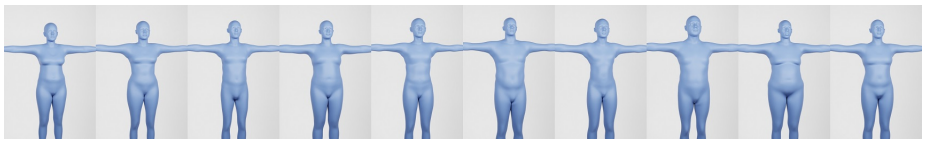
目前，我们尚未建立人体与面部身份之间依赖关系的概念，这意味着形状可能存在显著的不匹配。在实践中，我们发现，采用性别一致的方式进行采样，在大多数情况下都能得到合理的结果。对面部和身体身份进行联合采样，可能是未来工作中一个有趣的方向。采样得到的身份示例见图 4。通常，我们按相同比例采样男性、女性和中性身份，以确保充分覆盖性别谱。



(a) 男性。



(b) 女性。



(c) 中性。

图 4：随机采样的身份。

2 渲染流程

我们基于Wood等人的流程 [20]，使用Cycles渲染引擎 [2]。许多元素保持不变，例如头发和环境库。由于现在包含了腿和脚，我们还添加了一个接影平面，使主体更好地融入场景。图5 展示了该流程生成的一些最终渲染结果。

主要区别在于人体其余部分所用的纹理库、人体服装（现在需要适应姿态变化，而非像 Wood 等人 [20]那样保持静止），以及姿势库。这些新增内容的细节将在下一节介绍。图6 展示了我们如何利用各个组成部分构建人体合成图像。

我们能够通过我们的流程生成各种各样的真实标注数据和 RGB 图像，方式与 Wood 等人相同。 [20]。图7展示了使用我们的流程生成的图像的一些真实标注示例。

2.1 纹理

对于面部，我们使用 Wood 等人的高质量皮肤纹理库。 [20], 如图8a所示。对于人体，我们使用一组由三维人体扫描生成的 25 种高质量纹理 [16], 如图8b所示。对于每次扫描，我们都会提取反照率、位移，以及用于表现高频细节的近似凹凸贴图，均位于上文所述的 SOMA UV空间中。

独立采样面部和人体纹理可能会导致明显的不匹配



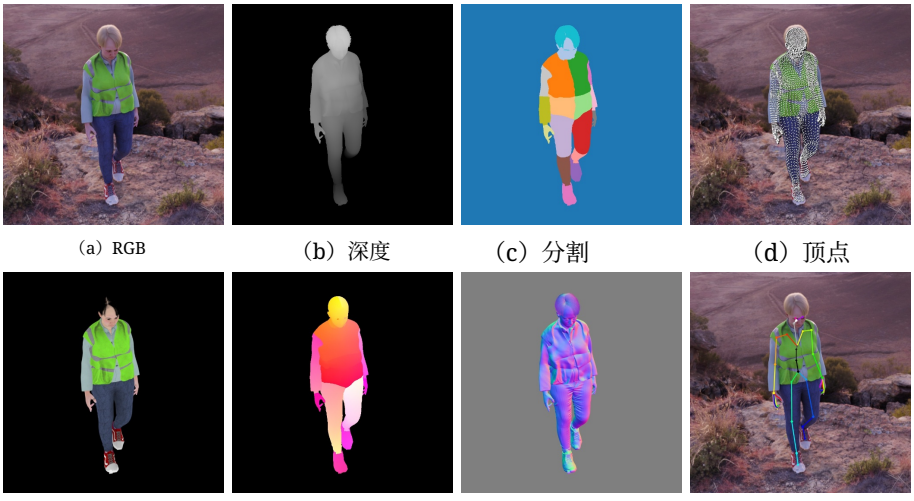
图5：使用我们的人类合成流程生成的示例图像。



(a) 采样身份。 (b) 服装。 (c) 头发。

(d) 纹理/着色器。 (e) 姿态。 (f) 环境。

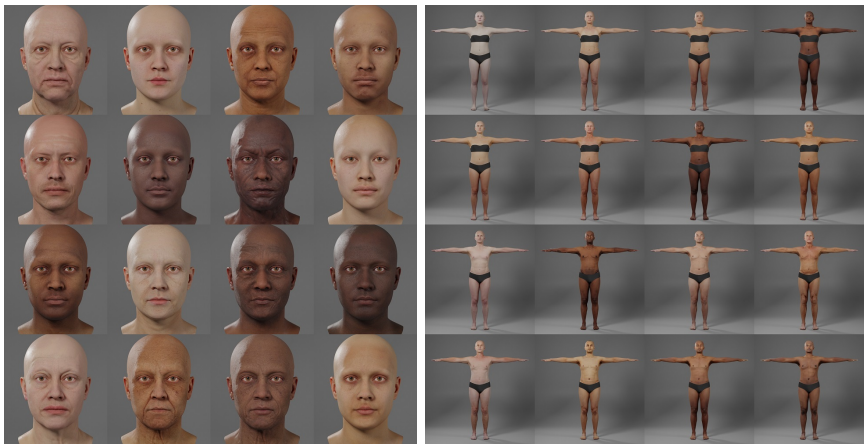
图6：构建合成人体的流程阶段。



（a）RGB （b）深度 （c）分割 （d）顶点

（e）反照率 (f) UV坐标 (g) 法线 (h) 骨架

图7：使用我们的流程可以生成多种真实标签类型。



(a) 面部 [20]。 (b) 人体。

图 8：我们纹理库中的皮肤纹理示例。

肤色匹配（参见图 9a）。

为解决这一问题，我们首先从纹理库中采样头部纹理，因为我们的头部纹理库具有更高的多样性；然后使用公式 1 根据输入的 RGB 值确定感知色差，从而在与面部具有一定感知相似度范围内，选择具有平均肤色的随机人体纹理。

$$\Delta C = \sqrt{\left(2 + \frac{\bar{r}}{256}\right) + \Delta R^2 + 4 \times \Delta G^2 + \left(2 + \frac{255 + \bar{r}}{256}\right) \times \Delta B^2} \qquad (1)$$

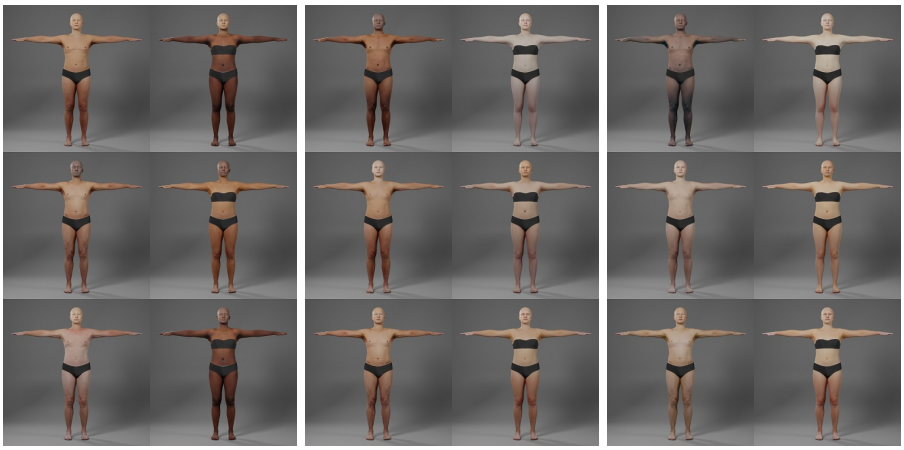
其中，$\bar{r}$ 是两种颜色的平均红色分量。

总体而言，这种方法能使面部与人体的肤色相当接近，不过仍存在细微的不匹配（见图 9b）。为解决这一问题，我们调整人体纹理中像素值的均值和方差，使其与面部纹理相匹配，从而确保人体与面部的肤色高度一致（见图 9c）。

2.2 服装

Wood 等人 [20]使用基于网格的素材，为面部添加服装和配饰。对于头饰、面部配饰（面具、眼罩）和眼镜，也可以采用相同的技术，只需将这些素材关联到全身模型的头部骨骼上即可。但其他服装则必须适应人体的动态姿势。因此，我们使用位移贴图来模拟服装 [10] 我们将这些素材分为上装和下装（包括鞋子），并使用这一技术来模拟其他一些配饰，例如手套、手表、手镯和戒指。

动态细分让我们能够运用这项技术生成极高保真的结果。对于每个资产，我们都会制作法线、粗糙度和金属度贴图，



(a) 随机采样。 (b) 筛选采样。 (c) 颜色调整。

图 9：肤色匹配过程。我们在采样时进行筛选以获得近似匹配，随后进一步进行颜色校正。

此外还包括反照率和位移贴图，从而在着色方面实现高度逼真的效果。

服装单品使用 Marvelous Designer 制作 [6]，并使用 Marmoset Toolbag 烘焙位移贴图 [7]。手动清理和材质细节则在 Substance Painter 中制作 [5]。我们的置换贴图服装库中的部分资产示例见 图 10。

这种方法存在一些明显的缺陷，其中最突出的是无法使用位移贴图表现宽松服装。此外，也无法进行模拟；动画过程中，服装必须直接贴合其下方的人体网格。我们发现，位移贴图的效果出人意料地好，适用范围不止于人们通常预期的非常紧身的服装，但它无法用于连衣裙和裙子等单品；而领带或夹克等单品在某些姿势下也无法呈现真实效果。为了解决这些问题，我们计划在未来采用基于网格的服装，并加入布料模拟。

2.3 姿势库

对于面部，我们使用 Wood 等人的表情库。 [20]对于人体，我们使用 AMASS 数据集 [11] 作为初始姿势库。在此基础上，我们加入了通过动作捕捉舞台采集并使用 MoSh 处理的数据 [8]。其中一部分动作捕捉数据专门用于填补现有库中的空缺，例如交叉腿姿势。我们的身体姿态库总共包含超过 200 万帧，帧率为 30 fps，即约 19 小时的动作数据。

在某些情况下，采集的姿态数据包含关节化手部姿势，但很多情况下并没有。对于没有手部姿势的帧，我们会随机抽取

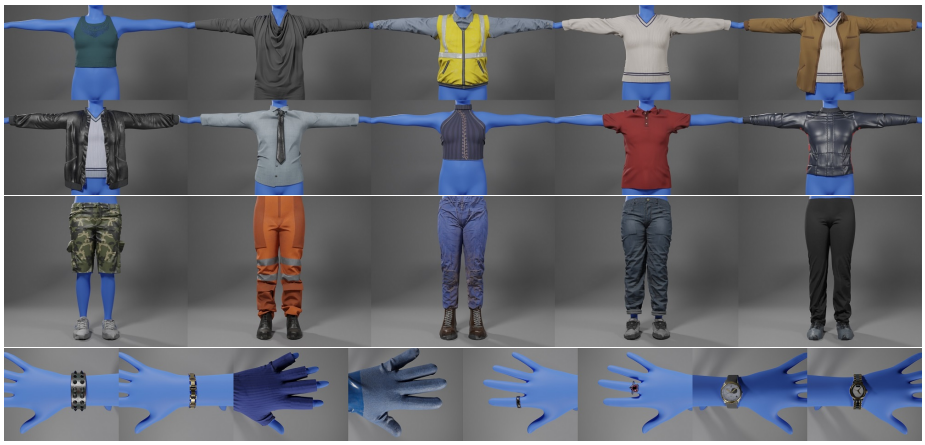


图 10：库中的位移服装示例，包括上装、下装、手镯、手套、戒指和手表。

MANO 数据集中的每只手的姿势 [14] 并将其拼接进去。面部表情（以及眼部姿势）和身体姿势是独立采样后再拼接起来的。总体而言，我们发现这样能产生非常逼真的结果，尤其适用于单帧（即非序列）数据。

采集动作捕捉数据时，通常会以标准姿势（通常是 T姿势）开始（并结束）动作序列。因此，我们发现，均匀采样姿势时，T姿势出现的频率非常高。为此，我们使用高斯混合模型（GMM）将姿势划分为一组粗略类别，其中一类为 T姿势。这样一来，我们便能大幅降低 T姿势在最终采样结果中的权重。

此外，我们发现，相对中性、站立的姿势很常见，而在训练深度神经网络以执行地标检测等下游任务时，这类姿势通常用处不大。因此，我们还会提高平均关节角度较大的帧的权重，也就是说，提高我们认为姿势更“有趣”的帧的权重。对于从 Wood 等人的表情库中采样面部表情，我们也采用类似方法，并根据平均混合形状激活加权。最后，我们会随机镜像身体姿势，从而使独特姿势的数量有效翻倍。

使用上述技术从我们的库中采样得到的姿势示例如图 11所示。

3 地标回归

这类以人为中心的视觉数据最常见的用途之一，或许是检测和跟踪图像中的人物。因此，我们定义了与第 1 节中定义的人体模型顶点相对应的关键点定义。一个仅包含 36 个关键点的稀疏定义（图12a）用于

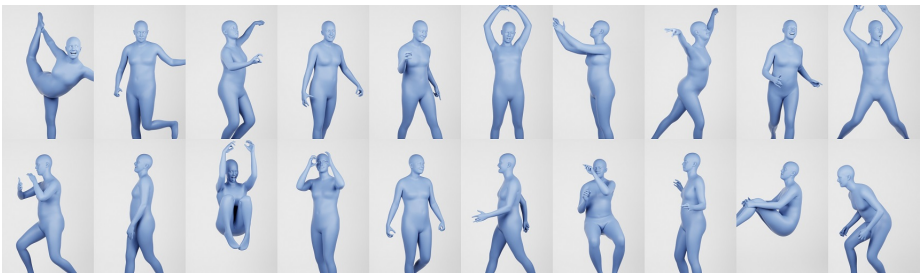
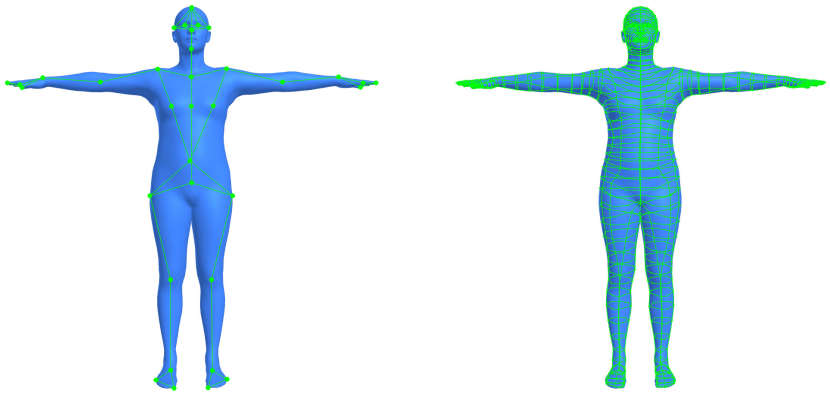


图11：从我们的人体姿态库中采样的示例姿态，其中拼接了 Wood 等人的面部表情。 [20] 以及来自 MANO 的手部姿态 [14] （见部分帧）。



(a) 稀疏（36）。 (b) 密集（1428）。

图12：用于 (a) 检测和 (b) 密集跟踪的全身关键点集。

按照 Wood 等人提出的滑动窗口方法进行检测和跟踪。 [19]。此外，我们还定义了一个包含 1428 个关键点的密集定义（图12b），用于模型拟合，见第 4 节。利用这些定义，我们可以使用顶点位置输出（图7d）轻松生成合成数据的二维关键点标注。

我们使用第 2 节中概述的流程，渲染了一个包含 100,000 张图像的数据集，其中每张图像中只有一个人，共有 20,000 个身份，每个身份包含 5 帧。每帧都具有不同的姿态和环境光照，以提高数据的多样性。用于训练的此类输入图像示例如上方图7a所示。

为了回归稀疏关键点，我们训练了一个 MobileNetV2 [15] 模型；为了回归密集关键点，我们训练了一个 ResNet101 [4] 模型，两者的输入图像尺寸均为 256×256 像素。两种情况下，我们都采用 Wood 等人提出的训练流程。 [19]



图13：手部数据集中的样例图像。

采用高斯负对数似然（GNLL）损失，并大量使用数据增强技术。因此，模型不仅预测二维关键点位置，还预测每个关键点的不确定性值。

3.1 手部和面部子网络

如上所述，在预测全身关键点时，我们发现手部的预测效果较差。这并不意外，因为在 256像素感兴趣区域中，手部所占区域很小。手部的形状及其运动方式也会造成严重的自遮挡，使这项任务尤具挑战性。因此，我们训练专门的深度神经网络来预测手部关键点，并以仅包含手部的感兴趣区域作为输入。

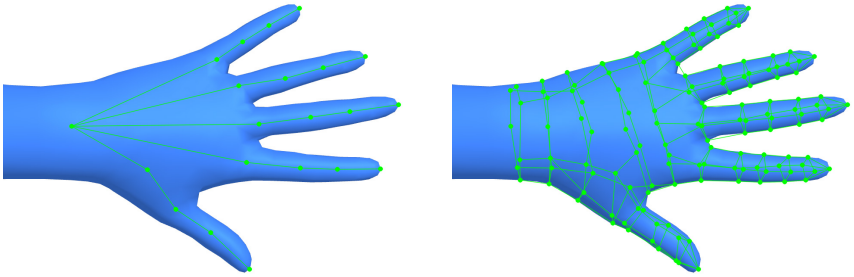
我们生成了一个包含 100,000 张合成图像的数据集，这些图像经过裁剪，仅包含左手，示例如图13所示。我们还为手部定义了稀疏和密集关键点，如图14所示。在密集情形下，手部关键点是图12b中全身定义的子集，因此两者可以直接对应。

同样，我们采用 Wood 等人的方法进行训练，[19], 在稀疏情形下使用 MobileNetV2，[15] 在密集情形下使用 ResNet，18 [4] 输入图像大小为 128×128 像素。我们增加了旋转增强的幅度，以进一步提高数据多样性。我们还提高了运动模糊增强的频率，以匹配真实数据中的观察结果，因为手部通常比身体其他部位运动得更快。

在运行时，我们首先预测全身关键点，并利用这些关键点提取手部周围的感兴趣区域并对其进行近似。随后，我们使用稀疏深度神经网络迭代优化该感兴趣区域，最后运行密集深度神经网络以获得输出关键点。由于密集定义中的直接对应关系，我们可以覆盖初始预测中的手部关键点，并在手腕处进行插值。

由于我们的网络只见过左手，而初始的全身预测能够区分左手和右手，因此我们只需将右手的感兴趣区域镜像后输入左手关键点深度神经网络。返回的关键点随后会在使用前镜像回去。

与手部类似，我们发现，使用单个深度神经网络回归全身关键点时，人脸关键点也无法得到准确预测。这同样



(a) 稀疏手部（21）。　　(b) 密集手部（141）。

图14：用于 (a) 检测和 (b) 密集跟踪的手部关键点集。

可能是因为这些模型所用的 256像素感兴趣区域中，面部尺寸较小。因此，我们也会截取专用的人脸感兴趣区域，并使用 Wood 等人的深度神经网络 [19]来准确回归人脸关键点。由于使用的是同一个人脸模型，人脸关键点也保持了直接对应关系，因此我们可以再次覆盖初始预测中的相应关键点。

3.2 结果

我们稠密全身关键点预测方法（包括手部子网络）的一些结果如图15所示。我们能够应对各种姿态、形状、外观和环境。在某些情况下，即使宽松服装、儿童和假肢并未在合成数据中建模，我们也能处理这些情况。尤其有用的是，当出现部分遮挡时（例如一条手臂被人体完全遮住），网络能够预测出合理的关键点，并对其给出较高的不确定性。

我们方法失效的示例如图16所示。对于极端姿态、严重（自）遮挡、宽松服装、缺失肢体和假肢，我们尤其难以应对。这很可能是因为目前的合成数据生成流程尚未对其中许多要素进行显式建模。事实上，我们在第1节中介绍的参数化人体模型甚至无法表示缺失的肢体。不过，由于我们使用了高斯负对数似然损失，并据此预测不确定性，因此通常能够准确指出这些错误发生的位置，许多示例都证明了这一点。未来工作将改进我们的形状模型和渲染流程，以弥补真实世界数据表示方面的这些不足。



图 15：密集地标跟踪结果。置信度以颜色编码，绿色表示置信度高，红色表示置信度低。图像采集自 https://pexels.com。



图 16：我们的地标预测方法失效的示例。置信度通过颜色编码，绿色表示置信度高，红色表示置信度低。图片收集自https://pexels.com。

4 模型拟合

除了关键点之外，能够获得对人物形状和运动进行参数化表示，通常也很有用。我们将 Wood 等人 [19] 用于人脸重建的方法扩展到拟合我们在第 1 节中描述的完整人体模型，并将其应用于第 3 节流程输出的密集地标。

因此，给定概率性二维密集地标 L，我们的目标是找到能使以下能量函数最小化的最优模型参数 $\mathbf{\Phi}^*$：

$$E(\mathbf{\Phi}; L) = \underbrace{E_{\text{landmarks}}}_{\text{Data term}} + \underbrace{E_{\text{face_identity}} + E_{\text{body_identity}} + E_{\text{expression}} + E_{\text{pose}} + E_{\text{temporal}} + E_{\text{intersect}}}_{\text{Regularizers}}$$

这里，我们要优化的是 $\mathbf{\Phi}$，即面部身份 $\vec{\gamma}$、身体身份 β、每一帧的表情 $\mathbf{\Psi}$、每一帧的姿态 $\mathbf{\Theta}$，以及相机旋转 $\mathbf{R}$ 和位置 $\mathbf{T}$。

$$\mathbf{\Phi} = \{\underbrace{\vec{\gamma}, \vec{\beta}, \mathbf{\Psi}_{F\times|\vec{\psi}|}, \mathbf{\Theta}_{F\times|\vec{\theta}|}}_{\text{Human}}; \underbrace{\mathbf{R}_{C\times 3}, \mathbf{T}_{C\times 3}}_{\text{Cameras}}\}$$

其中，F 是给定序列中的帧数，C 是相机数量。$E_{\text{landmarks}}$ 采用与 Wood 等人相同的形式 [19]。$E_{\text{face identity}}$，$E_{\text{expression}}$，E_{temporal} 以及 $E_{\text{intersect}}$ 也遵循 Wood 等人的实现。[19]。

对于 $E_{\text{body identity}}$，我们采用 L2 先验，因为 SMPL-H 使用经过方差缩放的主成分分析基来表示身份。采用高斯混合模型身体身份先验（正如我们对面部所做的那样）或许有助于生成更合理的人体形状；我们将其留作未来工作的一个潜在方向。

对于姿态，我们没有像 Wood 等人 [19]，那样使用 L2 先验，而是使用三个高斯混合模型先验：一个用于人体姿势（不包括手部和眼部姿势），其高斯混合模型拟合自姿势库的一个子集；另外两个分别用于左右手，其高斯混合模型分别拟合自 MANO 数据集 [14] 。这种数据驱动的方法有助于生成合理的姿势。更先进的姿态先验（例如深度神经网络）或许能带来更好的结果 [13], ，但我们仍将其留待未来工作。

在这类二维到三维提升任务中，例如本例，人体比人脸更具挑战性。投影之后，肢体位置存在更多歧义，例如，与面部相比，某些身体部位的运动范围更大。此外，人体更容易发生自遮挡；而对于人脸，即使出现有限的自遮挡，对称性也能提供非常强的先验。因此，我们发现，与人脸相比，有效的模型拟合困难得多，单目情形通常也是病态的。我们发现，需要多个相机视角（$C \geq 3$），并且相机数量越多，结果就会显著改善。



图 17：使用三台 Azure Kinect RGB 摄像机采集的数据进行模型拟合的示例结果。图中展示了不同主体和姿势下的三个视角，以及右下角四种不同姿势的单一视角。

我们还发现，相比人脸，合理的初始化更加重要，也难得多。对于人脸，只需使用 PnP 进行简单的六自由度对齐，就足以为优化器提供一个很好的起点；但对于全身，这种方法的结果相当差。对于手部姿势等精细细节尤其如此：即使关键点非常准确，若没有良好的初始化，也很难收敛。

为了解决这个问题，我们采用机器学习方法初始化姿态，根据一个可用视角直接预测姿态 [3], 从而为优化器提供一个非常好的初始值。随后，优化器利用多视角关键点，实现各视角间高度精确的三维一致性，并在世界坐标空间中进行非常精确的配准。许多机器学习方法无法将多视角数据作为输入；即使能够，也难以实现此处所需的各视角间高度精确的对齐与一致性。未来工作可以尝试改进这一初始化步骤，使其能够利用多视角数据，从而减少对第二阶段优化步骤的需求。

图中展示了我们模型拟合方法的一些结果图 17。

致谢

感谢 Rodney Brunet 及其团队在整个过程中为服装素材和艺术创作提供的宝贵帮助，感谢 DarrenCosker 协助收集姿势库的动作捕捉数据，并感谢 Kendall Robertson 协助处理姿势库的部分内容。

参考文献

[1] Gwangbin Bae、Martin de La Gorce、Tadas Baltrušaitis、Charlie Hewitt、Dong Chen、Julien Valentin、Roberto Cipolla 和 Jingjing Shen。《DigiFace- 1M：用于人脸识别的 100 万张数字人脸图像》。载于：~~WACV~~。IEEE。2023年。[2] Blender Foundation。Cycles 渲染器。https://www.cycles-renderer.org/。2021年。[3] Vasileios Choutas、Georgios Pavlakos、Timo Bolkart、Dimitrios Tzionas 和 Michael J. Black。《通过人体驱动注意力实现单目表现力人体回归》。载于：ECCV。2020年。URL：https://expose.is.tue.mpg.de。[4] K. He、X. Zhang、S. Ren 和 J. Sun。《用于图像识别的深度残差学习》。载于：CVPR。2016年。[5] Adobe Inc. Substance Painter。https://www.adobe.com/products/substance3d-painter~~.html~~。2022年。[6] CLO Virtual Fashion Inc. Marvelous Designer~~。https://~~marvelousdesigner.com/。2022年。[7] Marmoset LLC。Marmoset Toolbag。https://marmoset.co/Toolbag/。2022年。[8] Matthew Loper、Naureen Mahmood 和 Michael J Black。“MoSh：基于稀疏标记点的运动与形状捕捉”。载于：ACM ToG 33.6（2014年），第 1–13 页。[9] Matthew Loper、Naureen Mahmood、Javier Romero、Gerard ~~Pons-Moll~~ 和 Michael J. Black。“SMPL：蒙皮多人线性模型”。载于：ACM ToG 34.6（2015年10月），248:1–248:16。[10] Qianli Ma、Jinlong Yang、Anurag Ranjan、Sergi Pujades、Gerard Pons-Moll、Siyu Tang 和 Michael J Black。“学习为三维人物穿上生成式服装”。载于：CVPR。2020年，第 6469–6478 页。[11] Naureen Mahmood、Nima Ghorbani、Nikolaus F. Troje、Gerard Pons-Moll 和 Michael J. Black~~。~~“AMASS：表面形状动作捕捉档案”。载于：ICCV。2019年10月，第 5442–5451 页。[12] Daniel McDuff、Xin Liu、Javier Hernandez、Erroll Wood 和 Tadas Baltrusaitis。“用于多参数摄像头生理传感的合成数据”。载于：EMBC。IEEE。2021年，第 3742–3748 页。

[13] Georgios Pavlakos、Vasileios Choutas、Nima Ghorbani、Timo Bolkart、Ahmed AA Osman、Dimitrios Tzionas 和 Michael J Black。“表现力人体捕捉：通过单幅图像获取三维手部、面部和人体”。载于：CVPR。2019年，第 10975–10985 页。[14] Javier Romero、Dimitrios Tzionas 和 Michael J. Black。“具身手部：手部与人体的协同建模及捕捉”。载于：ACM ToG。245:1–245:17 36.6（2017年11月）。[15] Mark Sandler、Andrew Howard、Menglong Zhu、Andrey Zhmoginov 和 Liang-Chieh Chen。“MobileNetV2：倒置残差与线性瓶颈”。载于：CVPR。2018年，第 4510–4520 页。[16] Ten24 动画就绪人体扫描。Ten24 Media LTD。https://www.3dscanstore.com/retopologised-body-models/animation-ready-body-scans。2022年。[17] Tao Sun、Mattia Segu、Janis Postels、Yuxuan Wang、Luc Van Gool、Bernt Schiele、Federico Tombari 和 Fisher Yu。“SHIFT：用于连续多任务域适应的合成驾驶数据集”。载于：CVPR。2022年，第 21371–21382 页。[18] Phillip Thomas、Lars Pandikow、Alex Kim、Michael Stanley 和 James Grieve。用于提升骑行者检测效果的开放合成数据集。2021年11月。URL：https://paralleldomain.com/。[19] Erroll Wood、Tadas Baltrusaitis、Charlie Hewitt、Matthew Johnson、Jingjing Shen、Nikola Milosavljevic、Daniel Wilde、Stephan Garbin、Chirag Raman、Jamie Shotton、Toby Sharp、Ivan Stojiljkovic、Tom Cashman 和 Julien Valentin。基于密集地标的三维人脸重建。2022年。 DOI：10.48550/ARXIV.2204.02776。 URL：https://arxiv.org/abs/2204.02776。[20] Erroll Wood、Tadas Baltrušaitis、Charlie Hewitt、Sebastian Dziadzio、Matthew Johnson、Virginia Estellers、Thomas J. Cashman 和 Jamie Shotton。先假装，直到做到：使用合成数据进行真实环境中的人脸分析

e.

2021年。arXiv：2109。15102 [cs.CV]。