细粒度语义解耦的三维人体表示及其应用

Fine-Grained Semantic Disentangled 3D Human Body Representation and Its Applications

专　业　类　别：电子信息

研究方向（领域）：计算机技术

作　者　姓　名：孙晓琨

指　导　教　师：李坤教授

企　业　导　师：谢英弟

答辩日期			
答辩委员会	姓名	职称	工作单位
主席			
委员			

天津大学智能与计算学部

二〇二四年一月

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 **天津大学** 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名:　　签字日期:　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解 **天津大学** 有关保留、使用学位论文的规定。特授权 **天津大学** 可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

(保密的学位论文在解密后适用本授权说明)

学位论文作者签名:　　导师签名:

签字日期:　　年　　月　　日　签字日期:　　年　　月　　日

摘　　要

三维人体表示问题，即如何利用表示空间内的一组参数把三维人体网格准确、灵活、可控地表示出来，作为数字人体领域的基础问题，受到越来越多的关注。尽管其意义重大且应用广泛，但是现有方法却受限于粗糙的语义和不令人满意的几何刻画能力，无法精确且灵活地表示网格，尤其是在缺乏配对监督数据的情况下。针对上述问题，本文提出基于部位解耦的三维人体表示方法，通过设计新颖的骨骼分离的局部解耦策略，从根本上解决了先前工作语义粗糙、重建精度低等问题，并且整个训练过程无需配对的监督网格数据，极大地降低了后续应用的开发门槛。此外在该表示的基础上，本文结合表示“骨骼分离”、“部位解耦”等特点提出局部与全局注意力引导的人体重建网络和基于几何意义明确隐空间的人体编辑与生成应用方案，体现了该表示方法在人体重建、编辑与生成等工程任务中的广阔前景。本文的工作成果和创新点如下：

1）提出了基于部位解耦的三维人体表示方法，该表示方法以骨骼分离的局部解耦策略为核心，能够在不依赖配对监督数据的前提下兼顾细粒度语义和几何刻画能力，并且得益于表示所学习的语义精细隐空间，用户可以通过调整局部隐编码来实现灵活且高质量的人体编辑。

2）提出了骨架引导的自动编码器架构，该架构以骨骼分离的局部解耦策略为灵感，分两路处理人体各部位的骨骼信息和形状特征，从而合理且高效地完成对人体几何信息的编码，并结合独特的三分支训练流程及其无监督损失，赋予隐编码精确几何重建、无监督语义解耦、灵活部位编辑的能力。

3）在上述表示的基础上，针对人体重建问题提出契合表示特点的局部与全局注意力引导的人体重建网络，该网络在部位语义分割结果的引导下，能够有侧重地关注各部位的局部图像特征，以实现更精确的人体重建；其次针对人体编辑与生成问题，得益于表示所学习的几何意义明确的隐空间，用户可以通过调节隐编码实现灵活可控的高质量人体编辑与生成，大幅降低了用户的操作难度。

在多个数据集上的对比实验印证，比起现有表示方法，本文提出的表示方法在重建和编辑两方面都至少取得了 10% 的平均精度提升，并且具有更合理且精细的人体几何细节。

关键词：三维人体表示，几何学习，人体编辑，无监督解耦

ABSTRACT

The 3D human body representation problem, i.e., how to represent the 3D human body mesh accurately, flexibly, and controllably using a set of parameters in the representation space, has received more and more attention as a fundamental problem in the field of digital human body. Despite its significance and wide range of applications, existing methods are limited by coarse semantics and unsatisfactory geometric inscribing ability, and are unable to accurately and flexibly represent meshes, especially in the absence of paired supervised data. To address the above problems, we propose a 3D human body representation based on part decoupling, which fundamentally solves the problems of rough semantics and low reconstruction accuracy of previous work by designing a novel skeleton-separated decoupling strategy, and the whole training process does not require paired supervised mesh data, which greatly reduces the development threshold of subsequent applications. In addition, based on this representation, we propose a local and global attention-guided human body reconstruction network that combines the representation features of ”skeleton separation” and ”part decoupling” and human body editing and generation application scheme based on geometrically explicit latent space, which reflects the broad potential of our method for engineering tasks such as human body reconstruction, editing and generation. The achievements and innovations of this thesis are as follows:

(1) A 3D human body representation based on part decoupling is proposed, which is centered on the local decoupling strategy of skeleton separation and can balance the fine-grained semantics and geometric inscribing capability without relying on pairwise supervised data, and thanks to the latent space with fine semantics learned by the representation, the user can adjust the local latent code to achieve flexible and high-quality human body editing.

(2) A skeleton-guided autoencoder architecture is proposed, which is inspired by the local decoupling strategy of skeleton separation, and processes the skeletal information and shape features of each part of the human body in two ways, to rationally and efficiently accomplish the encoding of geometric information of the human body, and combines with the unique three-branch training process and its unsupervised loss, which empowers the latent code with accurate geometric reconstruction, unsupervised

semantic decoupling, and flexible part editing.

(3) Based on the above representation, we propose a local and global attention-guided human reconstruction network for the human body reconstruction problem, which is suitable for the characteristics of the representation, which can focus on the local image features of each human body part under the guidance of the results of the semantic segmentation of the parts, to realize the accurate human body reconstruction; secondly, for the human body editing and generation problem, thanks to the geometrically clear latent space, the user can adjust the latent code to realize flexible and controllable high-quality human body editing and generation, which greatly reduces the user's operation difficulty.

Compared with the existing representation methods on multiple datasets, the representation method proposed in this thesis can achieve at least an average accuracy improvement of approximately 10% in both aspects of reconstruction and editing, and have more reasonable and fine human geometric details.

KEY WORDS： 3D Human Representation, Geometry Learning, Human Editing, Unsupervised Decoupling

目　录

第 1 章 绪论

1.1 研究背景与意义

近年来，伴随硬件和软件技术的迅猛发展，数字信息科技产业获得了显著的进步。而作为社会的核心参与者，人类无疑是数字化研究领域的重要焦点。因此三维人体表示问题，即如何利用表示空间内的一组参数把目标三维人体精确、灵活、可控地表示出来，作为数字人体领域的基础问题广受研究人员们的关注。如图1-1所示，三维人体表示技术与各种数字人体应用有着密切的联系。比如通过从图像中估计表示参数可以完成对三维人体的重建[1–4]，由于表示模型具有丰富的人体先验知识，所以这类基于人体表示的的人体重建方法比起其他方法更加高效鲁棒；比如通过调整表示参数可以实现对于人体属性的灵活可控编辑[5–7]，以满足用户的个性化需求；比如通过对于表示参数插值或采样可以快速生成具有多样性的高质量三维人体[8–10]，以方便后续研究工作的展开。

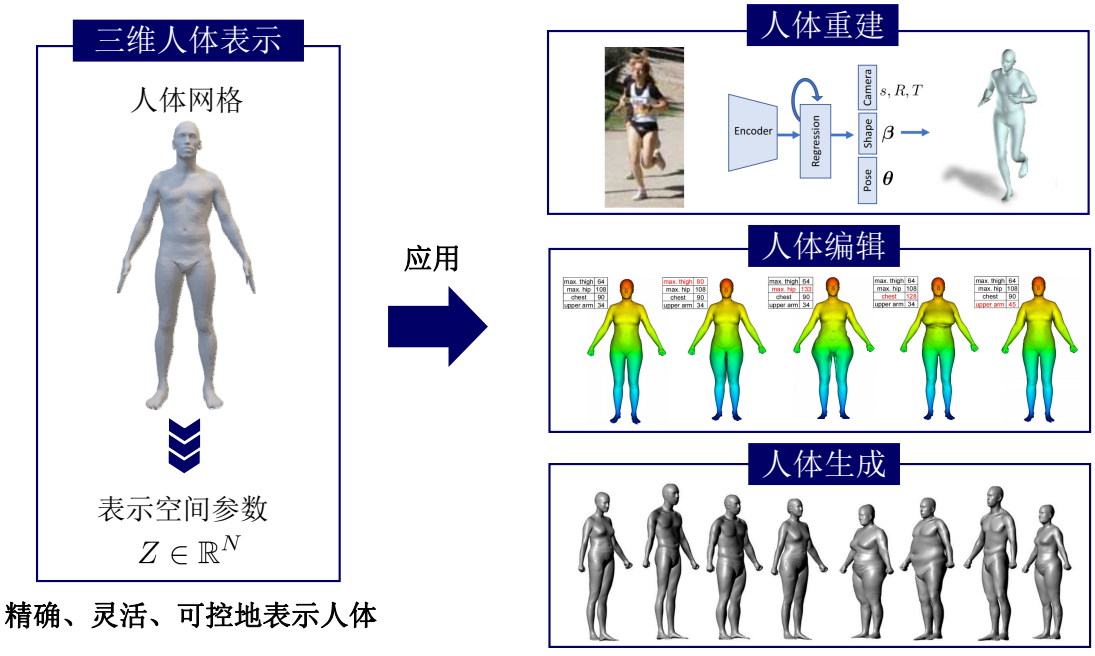


图 1-1 三维人体表示问题及应用

然而尽管前景广阔，由于人体几何结构复杂且具有丰富的形状、姿态变化，想要精确、灵活、可控地表示人体是一项充满挑战的任务。对于一种人体表示，业界通常从两个方面去评价它的好坏。首先是表示的几何刻画能力，即通过某种

表示重建出的人体在多大程度上保留了原始的输入人体的几何；其次是表示的语义颗粒度，即表示的参数是否有明确且精细的几何或物理意义。但是现有工作，无论是传统的基于统计模型的方法还是新兴的基于深度学习的方法在这两个方面都有所欠缺。传统方法[6,7,11–15]一般采用主成分分析（Principal Component Analysis, PCA）对人体形状进行建模，然而这种基于线性空间的方法无法处理人体网格中复杂的非线性几何结构。而基于学习的方法[10,16–24]大多以编码器-解码器框架为基础，通过在输入与输出间施加几何约束来确保隐编码尽可能多地保留输入网格的几何信息，得益于神经网络强大的非线性拟合能力，这类方法在几何刻画层面显著优于传统方法，但是语义的解耦依赖于制作耗时的配对监督数据。并且上述表示方法的语义都太过粗糙，只停留在对人体全局属性的描述，无法支持部位级别的灵活可控编辑。

因此本文以此为出发点，提出一种基于部位解耦策略的三维人体表示，得益于新颖的骨骼分离的局部解耦策略，该表示不仅具有细粒度的语义，以实现部件级别的灵活可控编辑，还可以准确且高效地刻画几何结构，并且整个训练过程不需要配对的监督数据。

1.2 问题与挑战

要想在无监督学习的前提下使表示兼顾精细语义与几何刻画绝非易事，所面对的问题与困难是多方面的。

首先人体表示语义的精细程度由该表示的人体解耦思路决定，所以第一个问题就是应采用怎样的思路对人体进行解耦。传统的思路是把人体解耦为形状和姿态[14,15]，然而这种针对人体全局属性的解耦思路既限制了表示语义的颗粒度又不便于设计有效且鲁棒的无监督损失。一种简单直接的想法是把人体拆分为一个个部位然后进行解耦，可是人体各部位形状各异，很难找到一种统一的思路对其进行解耦。其次人体网格数据几何结构复杂，尤其是人脸、人手等具有丰富几何细节的部位，并且不同个体间形状、姿态各异，传统的基于线性空间的建模方法[11–13]根本无法处理非线性几何结构，基于深度学习的方法[10,16,17]虽然通过构建精巧的类卷积算子提升了模型的重建精度，但是对于复杂几何细节的刻画仍然不尽如人意。最后现有基于深度学习的解耦工作[21]为了保障方法的几何表示能力严重依赖于制作复杂的配对监督数据，这极大地限制了这类方法的后续应用；而之后提出的无监督解耦工作又受其不够鲁棒的无监督损失的影响，重建出的人体往往存在明显的伪影。

综上，要设计出一种既不依赖配对监督数据又可以兼顾精细语义与几何刻画的三维人体表示主要面临以下三点挑战：1）传统解耦策略缺陷明显，新的解耦

思路设计难度大；2）人体结构复杂且细节丰富，现有表示方法难以精确表示；3）制作监督数据费时费力，现有无监督方法鲁棒性差。

1.3 本文内容及贡献

本文着眼于三维人体表示问题，以精确、灵活、可控地表示三维人体为研究目标，针对上述所提到的问题与挑战，提出一种不需要配对监督数据且兼具细粒度语义与高重建精度的三维人体表示方法。

首先为了赋予表示细粒度的语义，本文设计了一种具有人体解剖学先验的骨骼分离的部位解耦策略。具体来说，该策略首先将人体拆分为一个个部位，再在部位层级上将其形状变化解耦为与骨骼相关的变化（如骨骼长度和方向的变化）和与骨骼无关的变化（如部位围度和风格的变化）。与传统的基于人体全局属性的解耦人体姿态与形状的策略相比，这种聚焦于人体局部的解耦策略建立了隐编码与人体部位的几何属性之间的明确对应关系，这不仅有利于灵活可控的人体编辑，而且方便构建鲁棒且有效的无监督损失。

其次为了提升表示的几何刻画能力，本文提出一种骨架引导的自动编码器架构，该架构以经典的编码器-解码器框架为基础，融合解耦思路中的“部位感知”、“骨骼分离”的特点，把编码器拆分为编码人体骨骼信息和形状特征的骨骼支路和形状支路，并且将一个全局全连接层细化为多个局部全连接层，这样设计不仅有利于人体几何特征的提取与聚合，得以更加有效地建模几何细节，而且显著降低了模型参数量。同时为了摆脱对于配对监督数据的依赖，本文还设计了一系列新颖且有效的训练分支与无监督损失，以分别实现精确的几何重建、无监督的语义解耦、灵活的部件编辑的目的。

最后为了展现该表示在落地应用方面的巨大潜力，本文面向人体重建、人体编辑、人体生成等现实需求给出了新颖且有效的实践方案。具体而言，面向人体重建问题，受到表示“骨骼分离”、“部位解耦”等特点的启发，本文提出局部与全局注意力引导的人体重建网络，在语义分割图的显式引导下，该网络能够更精确地估计各部位的表示参数；另外针对人体编辑和生成问题，得益于表示隐空间几何意义明确的特点，该表示可以通过调整隐编码来编辑和生成高质量的三维人体，大幅降低了相关应用的技术门槛。本文主要贡献总结如下：

1）提出了一种不依赖于配对监督数据且兼具细粒度语义与几何刻画的三维人体表示。得益于细粒度的表示语义，用户可以通过调整局部隐编码来实现个性化的人体编辑。

2）提出了一种新颖的骨骼分离的部位解耦策略。不同于传统的基于人体全局属性的解耦思路，该策略建立了隐编码与人体部位的几何属性之间的对应关

系，这不仅赋予表示细粒度的语义，而且方便构建鲁棒且有效的无监督损失。

3）提出了一种骨架引导的自动编码器架构，并结合独特的三分支训练流程和新颖的无监督损失，赋予了表示精确几何重建、无监督语义解耦、灵活部件编辑的能力。

4）基于该三维人体表示，针对常见的人体重建、编辑、生成等应用给出了更契合表示特点的局部与全局注意力引导的人体重建网络和基于语义精细隐空间的人体编辑与生成实践方案，证明了该表示方法在实际工程应用中的广阔前景。

1.4 论文组织架构

本文聚焦于三维人体表示问题，第一章为绪论，第二章为相关理论与工作介绍，第三章和第四章是本文的主体，分别介绍了本文所提出的基于部位解耦的三维人体表示方法和该方法在不同场景下的应用实例，第五章是总结与展望。具体来说，本文的章节内容安排如下：

第一章为绪论，首先阐述了本文选题的背景及意义，然后分析了该课题面临的问题与挑战，最后引出本文的研究内容和核心创新点以及整体的内容安排。

第二章为相关理论与工作介绍，首先给出了三维人体表示问题的具体描述和评判标准，然后介绍了传统的基于统计模型的三维人体表示方法和新兴的基于深度学习的三维人体表示方法，并从原理和应用角度简要分析了这些方法的优缺点以及与本文所提出方法的区别与联系。

第三章介绍了本文提出的基于部位解耦的三维人体表示方法。首先深入分析了先前工作无法实现研究目标的根本原因，随后具体描述了本文所提出的骨骼分离的部位解耦策略，并基于此策略设计了一系列新颖且有效的网络架构、训练流程与损失，最后从重建精度和编辑精度两方面将该方法与现有工作进行了充分的实验对比，验证了方法的有效性，并补充了消融实验证明了所提出各模块的必要性。

第四章面向多个应用任务给出了行之有效的实践方案。首先针对基于人体掩膜的人体重建问题，通过在基础网络框架下与主流人体表示工作进行对比验证了该表示的优越性，并在此基础上结合该表示“骨骼分离”、“部位解耦”的特点，提出局部与全局注意力引导的人体重建网络，进一步展现其潜力和可拓展性。其次在人体编辑问题上，得益于表示局部隐编码与人体部位几何属性间的精确对应关系，该表示可以通过调整局部隐编码实现对于人体骨骼方向、骨骼长度、形状围度、形状风格的灵活可控编辑，证明了其灵活性和可编辑性。最后借助表示所学习到的具有明确几何意义的隐空间，该方法可以通过在隐空间上线性插值或随机采样生成大量高质量的人体网格数据，展现了该方法在人体生成问题上的应用

前景。

第五章为总结和展望，从全局角度对本文的研究内容做了详尽的总结和梳理，并指出了目前存在的不足以及未来的研究方向。

第 2 章　相关理论与工作介绍

本章主要介绍三维人体表示问题及其相关工作，主要包括基于统计模型的三维人体表示方法、基于深度学习的三维人体表示方法，最后从原理和应用角度对先前方法进行归纳与总结，并分析其优劣势，进而引出本文的研究内容，为之后的内容作铺垫。

2.1　三维人体表示问题描述

在计算机视觉和计算机图形学中对于三维数据的表示方法大致可以分为显式表示和隐式表示两类方法，常见的显式表示方法包括体素、网格、点云，而隐式表示包括占用场、符号距离场、神经辐射场等。其中网格表示因其内存占用少、对纹理友好等特点一直是学术界和工业界的主流三维数据表示方法，本文所研究的三维人体数据就是由网格表示的。

三维人体表示问题可以被建模为具有 n 个顶点的三维人体网格被压缩成 N 维的表示空间参数的降维问题，该问题一般从两个维度进行评判，即“几何保留是否完整”和“表示语义是否精细”，问题示意图如图2-1所示。

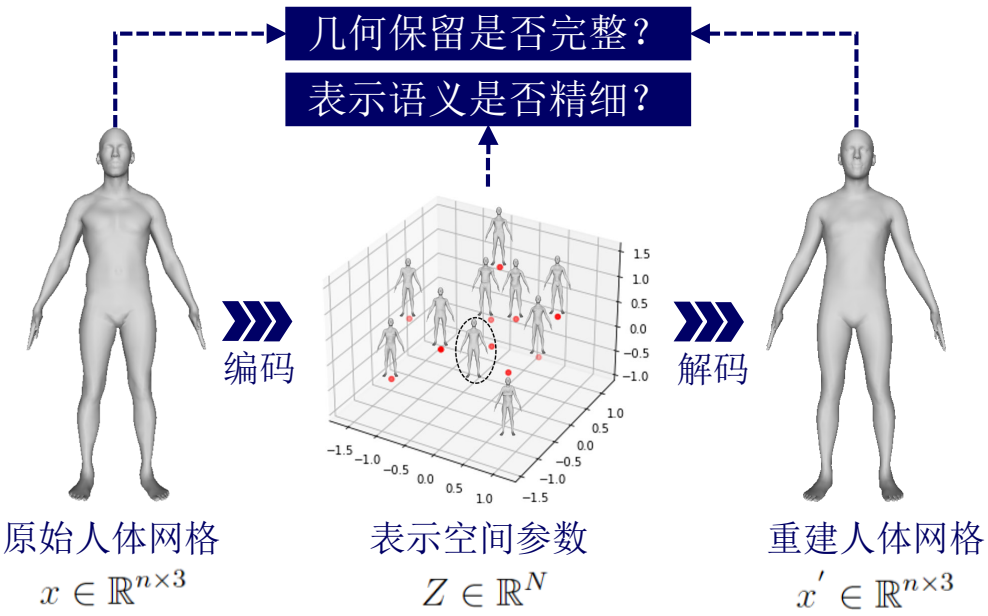


图 2-1　　三维人体表示问题示意图

2.2 基于统计模型的三维人体表示方法

由于人体是一种具有强先验的几何体，所以利用统计模型去对三维人体进行建模是一种非常简单且有效的思路。Allen 等[13] 先通过一种基于混合目标函数的优化方法得到一系列具有相似姿态的高分辨率人体网格，并在其基础上结合 PCA 技术完成对于三维人体网格的降维表示。Yang 等[6] 则在上述工作基础上利用局部映射的思路，在人体部位层级上对人体网格进行降维表示，不仅提高了重建精度并且建立了人体测量参数与人体部件间的对应关系。但是以上工作在处理网格时都没有考虑人体姿态的变化，而在实际生活中姿态又是人体不可或缺的一部分属性，因此 SCAPE[15]、SMPL[14] 等同时建模人体姿态与形状变化的工作应运而生。在人体形状层面，SCAPE 和 SMPL 仍然采用 PCA 技术对其建模，在人体姿态层面，它们通过给人体内嵌骨架然后利用骨骼蒙皮技术去处理姿态相关的人体形变。得益于其对人体姿态和形状的解耦，用户可以通过控制形状参数和姿态参数来控制人体的姿态和形状属性。其中 SMPL 因其更详细的人体关节定义和与图形学软件更好的兼容性得到更广泛的应用，现已成为数字人体领域的奠基性工作。SMPL[14] 的形状参数 β 为 PCA 线性空间下的基系数（取前 10 个维度），可以简单地理解为人体沿某方向的膨胀或收缩度，但是并无实际物理或几何意义。姿态参数 θ 则为预先定义的 24 个关节点相对于其父节点的旋转角的轴角表示，其维度大小为 $24 \times 3 = 72$ 维，输入姿态与形状参数即可输出具有 6890 个顶点的目标人体网格，图2-2展示了 SMPL 模型生成人体的全过程。

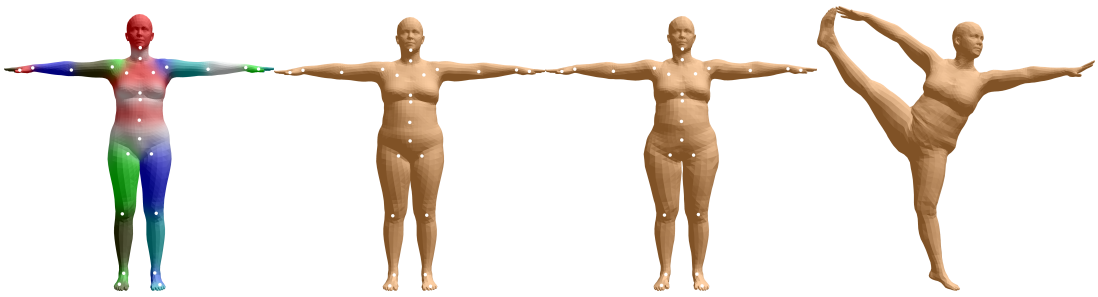


图 2-2　　SMPL 模型[14] 示意图

整个生成过程分为四个阶段，首先在 (a) 阶段定义一个对大量真实人体网格求均值得到的基模板 $T \in \mathbb{R}^{6890\times3}$，可简单理解为该模版定义了一个具有中等身材和标准姿态的基础人体。然后 (b) 阶段和 (c) 阶段分别在基模板的基础上施加基于形状的混合形状（Shape Blend Shapes）$B_S(\beta)$ 和基于姿态的混合形状（Pose Blend Shapes）$B_P(\theta)$，$B_S(\beta)$ 和 $B_P(\theta)$ 分别建模了姿态和形状对于人体几何的影响。最后根据姿态参数 θ 结合混合蒙皮（Blend Skinning）实现人体姿态的驱动，上述

过程的公式化表述可总结为下式：

$$\beta = \mathbb{R}^{10}, \tag{2-1}$$

$$\theta = \mathbb{R}^{72}, \tag{2-2}$$

$$T_P(\beta, \theta) = T + B_S(\beta) + B_P(\theta), \tag{2-3}$$

$$M(\beta, \theta) = W(T_P(\beta, \theta), J(\beta), \theta, \mathcal{W}), \tag{2-4}$$

其中 $W(\cdot)$ 是驱动人体的函数，$J(\beta)$ 代表人体关节点，$\mathcal{W}$ 代表蒙皮权重。然而上述基于统计模型的方法对于人体形状的表示仍然局限于线性空间，因此往往很难做到对原始网格的精确重建，并且其采用把人体解耦为姿态和形状的全局属性解耦思路，无法为表示提供细粒度的语义，也就不支持灵活可控的人体编辑。

2.3 基于深度学习的三维人体表示方法

随着硬件算力的不断提升，神经网络凭借其强大的非线性拟合能力在各个领域大放异彩，特别是基于编码器-解码器的网络架构在语音图像等低维数据的表示学习领域展现出了巨大的潜力。然而由于三维网格具有更高的数据维度和不规则的结构，循环神经网络（Recurrent Neural Network, RNN）、卷积神经网络（Convolutional Neural Networks, CNN）等传统网络结构无法被直接应用，所以许多工作[25–30] 致力于提出能够有效提取与聚合三维网格特征的类卷积算子。这类方法虽然在重建精度上较传统方法有显著的性能优势，但是其学习到的隐空间在语义层面仍然是耦合的，因此这类方法可以统称为语义耦合的三维人体表示方法；反之具有明确语义的就被称作语义解耦的三维人体表示方法，同时这类方法根据其是否需要配对的监督网格数据又可以被分为监督解耦工作与无监督解耦工作，下面就针对各类方法展开详细介绍。

2.3.1 语义耦合的三维人体表示方法

根据特征域的不同，拓展到网格上的卷积操作可以分为基于空间域特征的方法和基于谱域特征的方法。基于谱域特征的方法[25,26,31] 利用图拉普拉斯算法对网格特征进行卷积运算。Ranjan 等[16] 通过在网格上进行基于截断切比雪夫多项式的谱域特征提取在三维人脸表示任务上展现了出色的性能。另外基于空间域特征的方法以[28–30] 各个顶点的局部空间结构和特征为输入构建卷积算子。[32,33] 等先采用测地线极坐标对局部网格表面进行参数化，然后利用各向异性核函数对空域特征进行处理。Bouritsas[10] and Gong 等[17] 提出在空间域利用螺旋序列来定义局部网格表面，也获得了惊艳的效果。Hanocka 等[34] 则在顶点连线上定义卷积并把其拓展到网格细分[35] 上。Chen[19] 和 Gao[18] 等则想到利用注意力机

制去提升网格空间域特征处理的鲁棒性和有效性。除此之外，Tan 等[8,9] 则另辟蹊径，决定摒弃欧式坐标空间，转而采用对于网格大尺度变形更鲁棒的 ACAP（As-consistent-as-possible）[36] 特征去分析三维人体，同样也表现出了不错的性能。但是这些方法不仅没有明确的语义，并且对于那些细节丰富的人体几何刻画也不够到位。

2.3.2 语义解耦的三维人体表示方法

Jiang 等[21] 提出一种分层自动编码器架构，并基于此架构实现了一种具有高重建精度的姿态与形状解耦的人体表示方法。但是此方法严重依赖于一条数据限制，即每个用于训练的网格都必须有一个与之配对的具有相同形状和中性姿态的网格，通过这种配对数据带来的显式监督来实现语义的解耦。然而制作监督数据费时费力，这严重限制了这类工作的后续应用，因此一系列无监督解耦的表示工作[20,22–24] 涌出。这类工作大都通过构建捕捉人体姿态或形状特征的无监督损失来为表示提供隐式的监督，从而实现隐空间的解耦。比如 Cosmo 等[23] 提出采用几何度量保留作为学习可变形三维形状的潜在表示的强大先验，此创新点的关键是引入一个直接定义在解码的形状上的几何失真准则，从而将解码的几何度量保留转化为在潜在空间中形成的线性路径。Zhou 等[22] 则通过组合使用自一致性和交叉一致性约束来学习人体网格的姿态和形状空间。此外他们还将 ARAP（As-rigid-as-possible）变形引入训练循环，以避免退化解的出现。但是这两篇工作严格意义上不能称作无监督解耦，因为它们仍然对训练数据的姿态与形状有限制。随后完全不约束训练网格的姿态与形状的 IEP-GAN[24] 被 Chen 等提出，具体地说，该工作提出了一种内在-外在信息保留的生成对抗网络，用于同时捕捉人体的内在（即形状）和外在（即姿态）信息。理论上，他们提出了利用鉴别器来捕获不同拉普拉斯网格的外在/姿态不变性；同时引入一个内在保留损失，在保持测地线先验的同时避免了大量的计算。然而这些无监督解耦工作虽然摆脱了对配对监督数据的依赖，但是由于其无监督损失的鲁棒性较差，生成出的人体网格往往有明显的伪影，这使其实用性大打折扣。并且上述解耦工作仍遵循姿态和形状的全局性解耦策略，这严重限制了人体表示的语义颗粒度，灵活可控的人体编辑更是无从谈起。一言以蔽之，现有的三维人体表示工作无法在无监督学习的前提下兼顾精细语义与高重建精度。

2.4 三维人体表示的应用

三维人体表示作为数字人体领域的基础性问题，无论是在计算机视觉还是计算机图形学领域都有广阔的应用前景。在人体重建方面，SMPL 模型[14] 因其提

出较早且与现有图形学软件良好的兼容性，被广泛应用在基于模板的人体重建方法[1,3,4]。尽管这些方法在重建精度和运行速度方面都给人留下了深刻的印象，但是 SMPL 本身的局限性使得其存在无法改变的缺陷，比如 SMPL 的形状参数 β 没有实际的物理意义，基于 PCA 的形状建模使其注定无法精确灵活地表示人体，这都极大地限制了这些方法的性能；另外 SMPL 解耦人体姿态和形状的解耦思路使其模型参数 β 和 θ 都是以整个人体为单位进行属性描述，这无疑增大了网络训练时需要拟合的映射空间，也就无形中增大了人体重建的误差。在人体编辑方面，由于 SMPL 的形状参数 β 并没有明确的几何意义，所以其不支持灵活的人体形状编辑。HBR[6,7] 等工作通过建立人体测量参数与人体面片偏移间的映射实现了灵活可控的人体形状编辑，但是其泛化性太差，一旦编辑目标离训练数据分布的距离较远，其生成出的人体网格往往具有严重的伪影和不合理的变形，并且该类方法在建模人体时并没有考虑人体姿态，也就无法实现姿态编辑。而基于深度学习的方法，由于其隐编码必须由现有网格编码得到，所以严格意义上他们不兼容任何编辑性操作，他们所能实现的仅仅是不同网格间形状或姿态的迁移。在人体生成方面，由于 SMPL 形状参数 β 没有物理意义，通过随机采样参数生成出的人体容易具有极端的形状，不利于后续的研究与分析。而基于深度学习的工作可以通过在隐空间插值的方法实现大量人体的生成，但是因为其方法的局限性，表示学习到的隐空间往往连续性较差，生成出的人体网格容易出现伪影与不合理变形，并且上述工作参数空间的语义还是传统的姿态与形状，也就导致它们无法实现部位级的灵活可控人体生成。

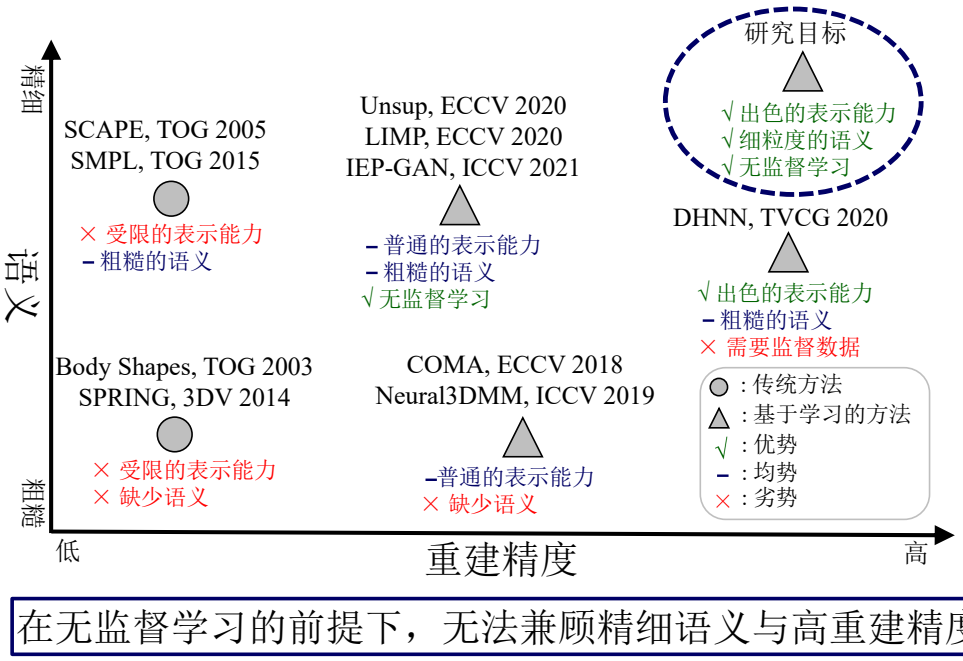


图 2-3 现有工作总览图

2.5 三维人体表示工作总结

综上所述，现有的三维人体表示工作在无监督学习的条件下无法兼顾细粒度语义与高重建精度，并且在主流应用场景下往往存在各式各样的缺陷与问题。图2-3总结了上述工作在语义和重建精度方面的性能表现，并指出各方法的优势与缺陷。Body Shapes[13] 和 SPRING[6] 缺少语义且表示能力有限，SCAPE[15] 和 SMPL[14] 虽然有初步的语义但是重建精度仍然不理想，COMA[16] 等工作通过结合神经网络的非线性拟合能力有效提升了方法的几何表示能力，但是其隐空间仍然是耦合的，DHNN[21] 虽然为表示提供了粗糙的语义但是依赖于监督数据的显式监督，IEP-GAN[24] 等工作通过构建无监督损失摆脱了数据限制，但是无法生成高质量的网格。而本文的研究目的就是针对上述方法的不足，实现一种不依赖配对监督数据的兼具精细语义与几何刻画的三维人体表示方法。

2.6 本章小结

本章围绕三维人体表示问题，首先给出了问题的具体表述，然后介绍了从基于统计模型到基于深度学习的一系列相关工作，并分析了各类表示方法的优势与缺陷，进而引出本文的研究内容。具体来说，第 2.1 节详细定义了三维人体表示问题，并且给出了三维人体表示的评判标准，即重建精度与语义颗粒度。第 2.2 节介绍了传统的基于统计模型的三维人体表示方法，并给出了该领域的代表性工作——SMPL 的具体定义和计算流程。第 2.3 节介绍了时下流行的基于深度学习的三维人体表示方法，从谱域到空间域，从有监督到无监督，对该类方法进行了详细且完备的梳理。第 2.4 节从人体重建、人体编辑、人体生成三个具有强实际需求的应用场景出发，分析了现有表示方法在工程应用上的优劣势。第 2.5 节对前面四小节的内容进行总结与归纳，用二维坐标图的方式直观分析了先前方法的优劣势与关键痛点，最后引出并强调了本文的研究目标。

第 3 章　基于部位解耦的三维人体表示

本章为本文的主体，主要介绍所提出的基于部位解耦的三维人体表示方法。具体而言，先分析先前工作存在明显缺陷背后的原因，随后介绍该表示方法的核心，即骨骼分离的部位解耦策略，然后引入基于此解耦策略设计的一系列新颖且有效的网络框架、训练流程与损失，最后通过对比实验验证了该表示在重建精度与编辑精度上的性能优势，并补充了消融实验证明了所提出各模块的必要性。

3.1　骨骼分离的部位解耦策略

研究目的已在上文阐明，即实现一种不依赖配对监督数据的且兼具精细语义与几何刻画的三维人体表示方法。那么先前工作为什么无法实现这一目标呢？本文认为这是因为先前工作都遵循把人体解耦为姿态和形状的解耦思路，这种基于人体全局属性的解耦思路存在两大缺陷，首先这种解耦策略所能提供的语义是粗糙的，这点很好理解，无论是姿态还是形状描述的都是整个人体的属性特征，自然限制了表示语义的颗粒度；其次，这种解耦策略增大了鲁棒且有效的无监督损失的设计难度。先前的无监督解耦工作[22–24] 都是通过构建捕捉人体姿态和形状特征的损失来实现无监督解耦。而刻画人体形状的难度并不大，先前工作就提出利用测地线矩阵、欧式距离矩阵等方法描述几何形状，尽管这些方法计算耗时较长但在效果层面表现良好。主要限制这类工作性能的是捕捉人体姿态特征损失的设计，由于人体关节众多且高度灵活，根据三维人体网格去精确估计出其关节姿态显然不是一件简单的事，事实也正是如此，先前的无监督解耦工作往往在形状保留方面表现良好，可是一旦大幅度改变姿态就会暴露其容易出现伪影和不合理变形的缺陷。那么如何解决上述问题呢？本文的解决方案是，摒弃原有的把人体解耦为姿态和形状的解耦思路，提出一种全新的骨骼分离的部位解耦策略，进而从根本上解决上述问题，该解耦策略示意图如图3-1所示。

该解耦思路来源于一个关键观察，即人体是由 $K = 17$ 个部位组成的（部位划分的详细信息见第3.3.4节），并且每个部位都有一根由三维关节点定义的骨骼，正如图3-1 (a) 和 (b) 所示。同时对于一些几何结构比较简单的人体部位，如腰部、手臂和腿部，其几何形状可近似成以其骨骼为主轴的圆柱体，这也就意味着这些部位的几何形变可以被建模为沿其骨骼方向 o_b（实心箭头）的变化以及沿其骨骼正交方向 $o_{b\perp}$（空心箭头）的变化，正如图3-1(b) 中小图所示。受这一观察的启

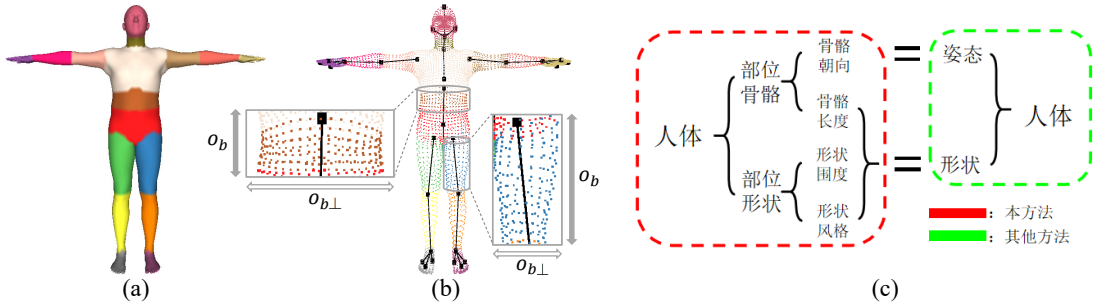


图 3-1　解耦策略示意图：(a) 人体部位，(b) 人体骨骼与关节点，(c) 解耦思路概述

发，本节提出一种骨骼分离的部位解耦策略，解耦思路概述如图3-1(c) 所示。具体来说，该思路首先将人体拆分为多个部位，再把各部位的几何形变解耦为骨骼相关的变化（比如骨骼长度和方向的变化）和与骨骼无关的形状变化（比如形状尺寸和风格的变化），它们分别由第 k 个部位的骨骼隐编码 z_b^k 和形状隐编码 z_s^k 表示。这种全新的基于人体局部的解耦策略不仅可以为表示提供细粒度的语义，以实现部位级别的灵活可控编辑，还可以准确且高效地刻画几何结构，并且有利于无监督损失的构建，从而摆脱对配对监督数据的依赖。

3.2　骨架引导的自动编码器网络

先前基于深度学习的三维人体表示工作大多以经典的编码器-解码器结构为基础，对三维人体进行低维编码，但是由于其网络设计缺少人体先验知识，对人体精细几何细节的刻画不够到位。基于此，本节提出骨架引导的自编码网络，该网络在保留自动编码器框架的基础上融合上述解耦思路中“骨骼分离”、“部位解耦”的特点，把编码器拆分为骨骼支路和形状支路，以用于分别编码人体各部位的骨骼信息和形状特征，并且通过把单个全局全连接层拆成多个局部全连接层显著减小了网络训练时需要拟合的映射空间，这样设计不仅有利于人体特征的提取与聚合，得以更加有效地建模几何细节，而且显著降低了模型参数量，网络设计细节如图3-2所示。

对于输入网格 x，编码器的骨骼支路 $E_b(\cdot)$ 和形状支路 $E_s(\cdot)$ 分别编码其每个部位的骨骼信息与形状特征，得到骨骼隐编码 $Z_b = \{z_b^1, ..., z_b^K\}$ 和形状隐编码 $Z_s = \{z_s^1, ..., z_s^K\}$，其中 z_b^k 和 z_s^k 表示第个 k 部位的局部隐编码，然后局部隐编码被送入对应的局部全连接层中得到局部特征，最终通过图卷积网络聚合不同部位的

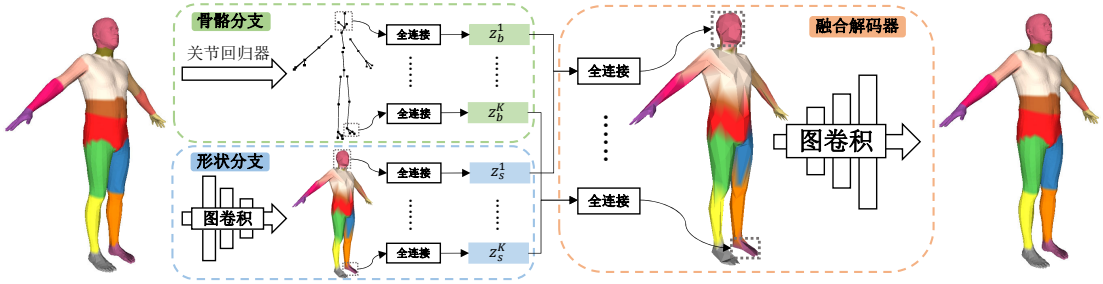


图 3-2　网络架构示意图

特征重建出输入的网格。以上过程的公式化表述如下式所示：

$$Kp = J(x), \tag{3-1}$$

$$E_b(Kp), E_s(x) = E(x), \tag{3-2}$$

$$Z_b = E_b(Kp), \tag{3-3}$$

$$Z_s = E_s(x), \tag{3-4}$$

$$x' = D(Z_b, Z_s), \tag{3-5}$$

其中 $E(\cdot)$ 和 $D(\cdot)$ 分别为编码器与解码器网络，$E_b(\cdot)$ 和 $E_s(\cdot)$ 分布代表编码器的骨骼支路与形状支路，Kp 为由关节点回归器 $J(\cdot)$ 计算得到的网格关节点位置，x' 则为根据表示参数重建出的人体。

具体来说，当网格 x 进入网络时，编码器中的骨骼分支首先根据由关节点线性回归器 $J(\cdot)$ 得到的关节点三维位置推断出网格的骨骼编码 $Z_b = \{z_b^1, ..., z_b^K\}$，由于关节点位置大致定义了人体部位的方向、长度等全局信息，因此这一步可以看作是在提取各部位的全局信息。另外形状分支采用螺旋卷积[17] 作为主要的特征提取算子，构建了层次化图卷积编码器以便学习人体网格的多尺度几何特征，然后根据顶点的部位语义标签将各部位顶点的多尺度特征输入相应的局部全连接层，得到包含局部几何细节的形状编码 $Z_s = \{z_s^1, ..., z_s^K\}$，因为网格几何结构由顶点决定，所以这一步可以看作是在提取各部位的局部几何特征。最后利用与形状分支网络结构完全镜像的解码器，通过聚合各个人体部位的局部和全局信息，准确且有效地重建原始网格。

3.3　网络训练

上述网络架构虽然体现了骨骼分离的部位解耦思路的特点，但是如果想让表示在无监督学习设置下兼顾细粒度语义和重建精度，以便实现灵活可控的高质量编辑，还必须设计对应的无监督训练损失。因此在上述网络框架的基础上，本节提出一系列新颖且有效的训练分支和损失函数以实现上述目标，网络训练总览图

如图3-3所示。

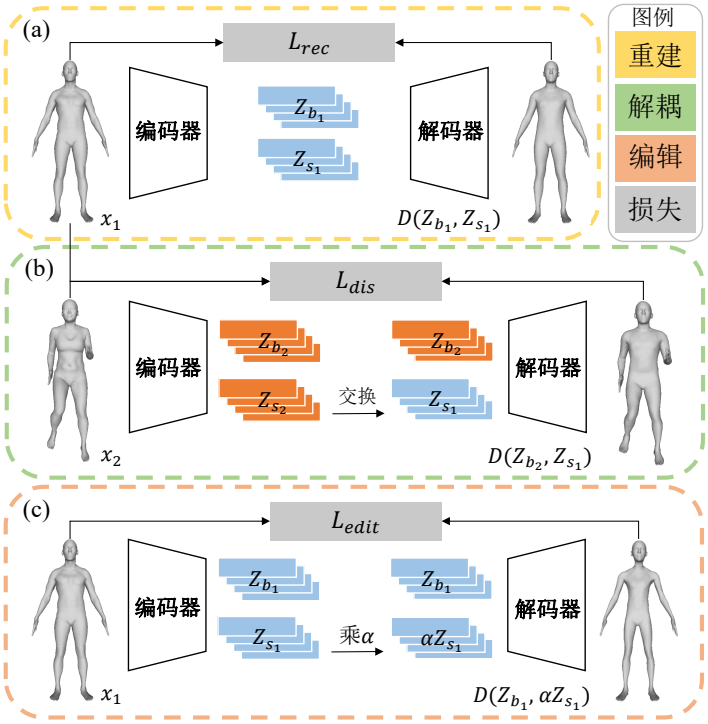


图 3-3　　训练流程示意图：(a) 重建分支、(b) 解耦分支、(c) 编辑分支

整个训练流程由 (a) 重建分支、(b) 解耦分支、(c) 编辑分支组成。在重建分支 (a) 中，编码器 $E(\cdot)$ 将输入网格 x_1 映射成骨骼隐编码 Z_{b_1} 和形状隐编码 Z_{s_1}，接着解码器 $D(\cdot)$ 通过聚合骨骼与形状隐编码中携带的人体全局与局部特征，在几何重建损失 L_{rec} 的监督下精确地恢复出输入的原始网格 $D(Z_{b_1}, Z_{s_1})$。然后在解耦分支 (b) 中，通过引入解耦损失 L_{dis}，由交换隐编码 (Z_{b_2}, Z_{s_1}) 生成得到的人体网格将会被要求同时具有 x_2 的骨架和 x_1 的局部形状特征。最后在编辑分支中，在编辑损失 L_{edit} 的帮助下，由缩放隐编码 $(Z_{b_1}, \alpha Z_{s_1})$ 解码得到的网格会被强迫按照期望变形，其中 α 是缩放因子，为一个标量。综上，基于骨架引导的自动编码器架构，本节通过利用三种损失来达成对应任务，整个训练流程的损失函数如下：

$$L_{full} = L_{rec} + L_{dis} + L_{edit}, \tag{3-6}$$

其中 L_{rec} 是用于实现精确人体重建的几何重建损失，L_{dis} 是用于实现人体部位骨骼与形状解耦的解耦损失，L_{rec} 是用于实现部位级灵活编辑的编辑损失，在下面章节中将会详细介绍这三个训练分支及其损失函数。

3.3.1 重建分支

如图3-3 (a) 所示，为了确保重建出的网格与原始网格尽可能相似，采用如下的几何监督损失：

$$L_{rec} = L_{vert} + \lambda_{edge} \cdot L_{edge}, \tag{3-7}$$

其中 λ_{edge} 是边长正则项的权重。顶点损失 L_{vert} 通过施加顶点间的 $L1$ 损失监督重建网格 $D(E(x))$ 的几何结构尽可能接近原始网格 x，顶点损失定义为：

$$L_{vert} = \|x - D(E(x))\|_1. \tag{3-8}$$

然而仅仅在顶点层面施加几何监督无法给网络训练提供足够的约束，不能避免生成的网格出现过长边，进而使网格的平滑性和合理性大打折扣。受到 NPT 等工作[37,38] 的启发，本节通过引入边长正则项 λ_{edge} 来解决该问题，其计算公式为：

$$L_{edge} = \sum_{p} \sum_{v \in N(p)} \|p - v\|_2^2, \tag{3-9}$$

其中 $N(p)$ 表示顶点 p 的一阶邻域。从计算公式中可知该损失通过保持具有共同顶点的一组边的长度一致性来使得网格表面更加紧致，从而提升了输出网格的平滑性和合理性。

3.3.2 解耦分支

在上述重建分支的约束下，本表示已经具备刻画复杂几何结构的能力，可是其隐空间仍然是耦合的，隐编码并没有明确的语义，因此遵循骨骼分离的部位解耦策略，本节采用如下的解耦损失来实现无监督的骨骼与形状解耦：

$$L_{dis} = L_{dis\_b} + \lambda_{dis\_s} \cdot L_{dis\_s}. \tag{3-10}$$

给定两个输入网格 x_1 和 x_2，x_{swp} 表示生成的网格 $D(Z_{b_2}, Z_{s_1})$，它由来自 x_2 的骨骼隐编码和来自 x_1 的形状隐编码组成的交换隐编码解码得到，如图3-3(b) 所示。L_{dis_b} 和 L_{dis_s} 分别用于监督 x_{swp} 保留属于 x_2 的骨骼信息和属于 x_1 的几何特征。通过这两种损失的协作，该表示就可以在无监督损失的隐式约束下实现语义的解耦。那么骨骼信息如何保留呢？由于人体骨骼完全由关节点位置决定，所以想要使得 x_2 的骨骼信息被 x_{swp} 完全保留，x_{swp} 应该具有与 x_2 一致的关节点位置。因此得益于骨骼分离的部位解耦策略，仅需利用一个简单且有效的线性关节回归器就可以完成对于骨骼信息的高效提取，从根本上避免了姿态特征保留损失设计难度大的问题，L_{dis_b} 的定义如下，其中 $J(\cdot)$ 为关节回归器：

$$L_{dis\_b} = \|J(x_2) - J(x_{swp})\|_1. \tag{3-11}$$

然而如何保留人体的局部几何特征是一个很有挑战性的问题，先前工作设计

的几何保留损失虽然效果尚可但是由于表示解耦思路的转变无法直接沿用。一个简单的可行方案是利用欧式距离矩阵来描述人体部位的几何结构，如下式所示：

$$L_{dis\_s} = \sum_{k=1}^{K} \| D_e(x_1^k) - D_e(x_{swp}^k) \|_1, \tag{3-12}$$

其中 x^k 代表 x 的第 k 个部位，$D_e(x^k)$ 表示 x^k 的顶点间的欧式距离矩阵，如果 x^k 有 n^k 个顶点那么该矩阵大小为 $[n^k, n^k]$。然而部位的欧式距离矩阵描述的是该部分的全部几何结构，也就是说骨骼的长度信息是耦合在该矩阵中的，因此这种方案会导致不彻底的骨骼与形状解耦从而影响后续人体编辑等应用的精度。

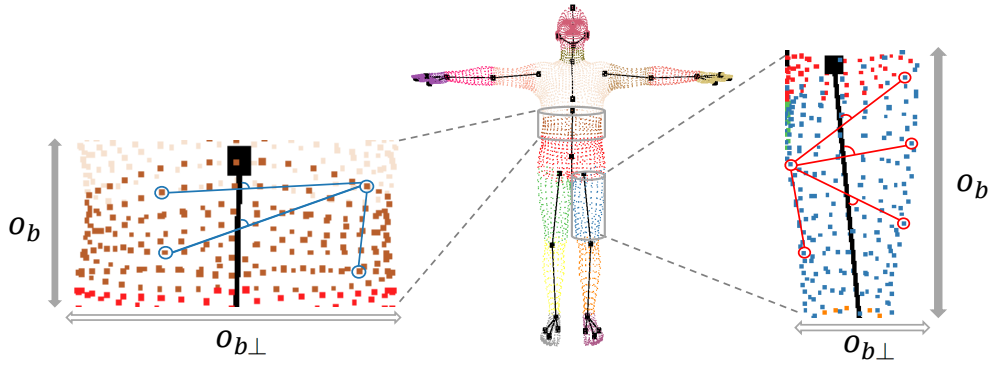


图 3-4　　方向自适应权重机制示意图

为了缓解这个问题，本节提出方向自适应权重机制（Orientation-adaptive Weighting, OAW）。具体来说，对于人体部位 x^k，其每对顶点间存在一条直线，通过计算该直线与骨骼方向 o_b 所形成的夹角可以得到一个大小为 $[n^k, n^k]$ 的角度矩阵，夹角如图3-4所示。显然，该夹角越大代表该连线所描述的几何信息对沿骨骼正交方向 $o_{b\perp}$ 的形状变化贡献越大，也就是说越与骨骼无关。举两个比较极端的例子便于读者理解，假如连线方向与骨骼方向平行（参考图3-4小图中的竖直连线），此时该连线所描述的几何信息就是骨骼的长度，即连线信息等于骨骼信息，与形状信息无关；而假如连线方向与骨骼垂直（参考图3-4小图中的水平连线），此时该连线所描述的几何信息就是部位宽度，即连线信息等于形状信息，与骨骼信息无关。直观来说，夹角角度一定程度上反映了该连线所携带的几何信息对刻画沿骨骼正交方向 $o_{b\perp}$ 几何形变的重要程度，角度越大重要程度越大，夹角为 90 度时重要程度达到最大，权重则为 1，角度越小重要程度越小，夹角为 0 度时重要程度达到最小，权重则为 0。接着只需要在上述角度矩阵的基础上建立值域为 $[0, 90]$ 的角度到值域为 $[0, 1]$ 的权重间的映射就可以利用权重矩阵来对前面的局部几何特征描述方案改进，改善骨骼长度信息耦合在欧式距离矩阵中的问题。具体而言本节通过采用如下的阈值处理和归一化函数 $f(\cdot)$ 得到大小同为 $[n^k, n^k]$ 的

权重矩阵 $W(x^k)$，其中 $A(x^k)$ 为部位 x^k 的夹角矩阵:

$$W(x^k) = f(A(x^k)), \tag{3-13}$$

$$f(a) = \begin{cases} a/90, & a > \sigma \\ 0, & others \end{cases} \tag{3-14}$$

加权后的欧式距离矩阵定义为：

$$D_e^w(x^k) = W(x^k) \otimes D_e(x^k), \tag{3-15}$$

其中 $\otimes$ 表示矩阵间的点乘运算。可以看到 OAW 机制通过利用由连线与骨骼间夹角转化得来的权重矩阵，为保留人体部位沿骨骼正交方向 $o_{b\perp}$ 的几何特征提供了显式指导，从而尽可能地将骨长信息从欧式距离矩阵中分离出来，实现了更彻底的语义解耦和更高精度的人体编辑（见后续消融实验3.5节）。

然而该机制也会导致在训练时忽略部分有用的沿骨骼方向 o_b 的几何信息，这会导致在大幅度编辑人体骨骼方向时出现不合理的网格变形和伪影（见后续消融实验3.5节）。为了解决这个问题，本节进一步提出部位级别的体积正则项来惩罚那些不合理的形状变化。理论上，当人体部位沿骨骼正交方向 $o_{b\perp}$ 的几何特征被完全保留时，该部位体积的变化是与该部位长度变化成正比的，比如某网格胳膊长度变为原来的 1.2 倍长，那么在其沿骨骼正交方向 $o_{b\perp}$ 几何特征不变的前提下，该部位的体积也应变为原有的 1.2 倍。所以本节以此为原理，引入可以提供强大几何监督以实现自然且合理人体编辑的体积正则项，该项可被定义为：

$$L_{vol} = \sum_{k=1}^{K} \| v(x_1^k)/l(x_1^k) - v(x_{swp}^k)/l(x_{swp}^k) \|_1, \tag{3-16}$$

其中 $v(\cdot)$ 是根据四面体体积公式来计算人体部位体积的函数，$l(\cdot)$ 是根据关节点位置来测量人体部位长度的函数。最终 L_{dis_S} 被改写为：

$$L_{dis\_s} = \sum_{k=1}^{K} \| D_e^w(x_1^k) - D_e^w(x_{swp}^k) \|_1 + L_{vol}. \tag{3-17}$$

3.3.3 编辑分支

在解耦分支的约束下，表示学习到的隐空间已经实现了部位级别骨骼与形状的无监督解耦，并且得益于局部解耦思路，在无监督学习前提下的语义解耦并没有影响表示在重建分支中掌握的精细几何刻画能力，使得表示真正做到兼顾细粒度语义与复杂几何刻画。然而尽管语义空间已经解耦，用户可以通过改变输入到骨骼分支的关节点位置来控制生成人体各部位的骨骼长度和方向，但是对于各部位的形状表示仍然无法实现灵活可控的编辑，为了解决这个问题，本节提出了编辑分支，如图3-3(c) 所示，该分支通过编辑损失 L_{edit} 监督由缩放隐编码解码得到的网格按照期望变形，从而实现部位级别的灵活形状编辑，具体计算公式如下式

所示，其中 α 是从均匀分布 $(\alpha_{min}, \alpha_{max})$ 中随机采样的标量，x_{sca} 表示生成的网格 $D(Z_{b_1}, \alpha Z_{s_1})$，它由来自 x_1 的骨骼隐编码和被缩放的来自 x_1 的形状隐编码组成的缩放隐编码解码得到：

$$L_{edit} = L_{edit\_b} + \lambda_{edit\_s} \cdot L_{edit\_s} + \lambda_{norm} \cdot L_{norm}, \tag{3-18}$$

$$L_{edit\_s} = \sum_{k=1}^{K} \|\alpha D_e^w(x_1^k) - D_e^w(x_{sca}^k)\|_1, \tag{3-19}$$

$$L_{edit\_b} = \|J(x_1) - J(x_{sca})\|_1, \tag{3-20}$$

$$L_{norm} = \frac{1}{K} \sum_{k=1}^{K} \|\|z_{s_1}^k\|_2 - circ(x_1^k)\|. \tag{3-21}$$

具体而言，编辑损失 L_{edit_s} 会监督由缩放隐编码生成的 x_{sca} 的第 k 个部件具有 $\alpha D_e^W(x_1^k)$ 所描述的几何结构。而对于欧式距离矩阵 $D_e^W(x_1^k)$ 来说，其矩阵内元素的绝对大小代表着第 k 个部件的形状尺寸，元素间的相对大小代表着第个 k 部件的形状风格，又因为标量 α 只会改变矩阵元素的绝对大小不会影响其相对大小，因此第 k 个部件会被要求随着矩阵内元素大小的缩放而沿着骨骼正交方向 $o_{b\perp}^k$ 缩放，同时不会影响该部位的形状风格。简单来说，L_{edit_s} 的引入成功使得形状隐编码向量的模长和方向分别代表部件的围度和风格，因此用户可以通过给形状隐编码乘标量来精确改变部件的围度，也可以通过在不同网格间交换形状隐编码的方向来实现形状风格的迁移，这不仅实现了部件级别的灵活人体编辑还在一定程度上保证了表示学习到的隐空间的连续性。同时表示期望在编辑部位形状的同时人体骨架不会被影响，本节通过引入 L_{edit_s} 约束人体网格编辑前后的关节点位置保持一致来达到这一目的。

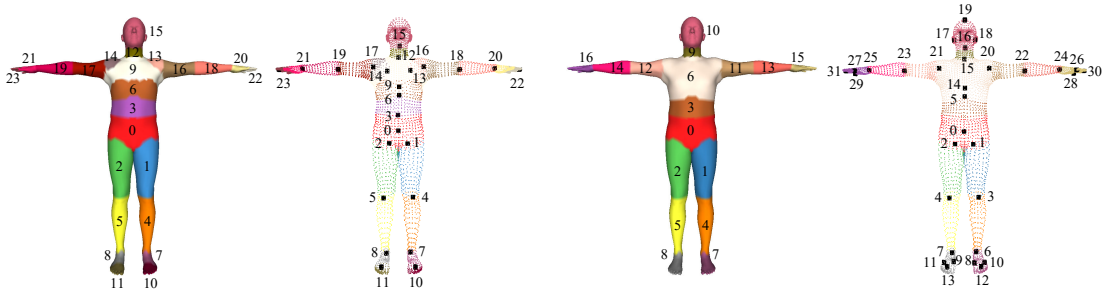
然而编辑分支容易导致训练崩溃。本节通过施加向量模长正则 L_{norm} 来解决这个问题，简单来说该正则项通过在部位围度和形状隐编码模长间建立明确的数值对应关系从而赋予形状隐编码的模长明确的几何意义，这种结合人体测量先验的正则项不仅有利于网络训练的收敛，而且也方便用户在统一尺度下编辑部位围度，该正则项计算公式如式3-21所示，其中 $circ(\cdot)$ 是利用标注点来测量人体部位围度的函数。

3.3.4 实现细节

对于螺旋卷积编码器，本节采用与 Neural3DMM[10] 相似的网络架构。具体而言，该编码器由四个螺旋卷积层和下采样层组成；解码器结构是编码器结构的镜像版本，另外把下采样层替换成上采样层，其中螺旋卷积的滤波器尺寸、扩张率等超参数均沿用 Neural3DMM[10] 中的设置。在代码实现方面，本章方法均在 Pytorch 框架[39] 下实现，所有的训练和测试实验均在装载着 RTX 3090 GPU 的 Ubuntu 服务器上实现。

在训练方面，网络学习率设置为 1×10^{-3}，并且在每个训练轮次结束后以 0.99 的衰减率衰减一次，网络在 Adam 优化器[40] 下训练 300 轮次，整个训练时间不超过 24 小时。在超参数设置方面，本文所有实验均用 16 维隐编码（骨骼隐编码和形状隐编码各 8 维）来表示人体各个部位。对于 DFAUST 数据集和 SPRING 数据集，训练时的下采样因子队列分别为 $[2,2,2,2]$ 和 $[4,2,2,2]$。超参数 $\lambda_{edge}, \lambda_{dis\_s}, \lambda_{edit\_s}, \lambda_{norm}$ 被设置为 1×10^{-2}，$(\alpha_{min}, \alpha_{max})$ 被设置为 $(0.8, 1.2)$，阈值 σ 被设置为 72 度。在训练细节方面，由于 SPRING 数据集[6] 中人体网格姿态基本维持在 A-pose，姿态变化幅度不大，因此本章方法在该数据集上训练时并没有用体积正则项。另外为了更好地重建人体局部的小尺度细节，本章借鉴 Limp[23] 的处理，在相对尺度下在计算 $L_{dis\_s}, L_{edit\_s}, L_{norm}$。具体而言，对于真值 V_T 和预测值 V_P，本章计算其相对误差 $\|(V_T - V_P)/V_T\|_1$ 而不是绝对误差 $\|(V_T - V_P)\|_1$，这有助于生成网格细节质量的提升。

在损失计算方面，对于某一批次输入进来的网格数据，首先将其送入到重建分支得到 x'，并根据损失函数计算得到重建损失 L_{rec}；接着进入解耦分支，训练数据内部交换骨骼与形状隐编码，并对其解码得到 x_{swp}，再利用损失函数计算解耦损失 L_{dis}；然后来到编辑分支，利用提前随机采样好的标量 α 对形状隐编码进行缩放并对其解码得到 x_{sca}，随后按照公式计算编辑损失 L_{edit}；最后将上述三种损失加合得到最终的训练损失 L_{full}，然后根据误差进行梯度反传，在优化器的帮助下对网络参数进行迭代优化。



(a) (b)

图 3-5 人体部位与关节点定义示意图：(a) SMPL[14]，(b) 本章方法

在人体部位与关节点定义方面，本章方法基本沿用了 SMPL 模型[14] 的设置，并在其基础上结合具体实践情况进行了改动与调整。图3-5展示了 SMPL 模型[14] 与本章方法的不同，表3-1给出了两种定义间具体的对应关系。具体来说，本章方法合并了部分人体部位的标签并且去除了冗余的关节点以便简化人体的结构，在不影响方法性能的前提下降低工程实现的难度，另外在此基础上为人头、人手和人脚定义了额外的关节点以便更精确地表示这些人体部位的姿态。

表 3-1 SMPL 和本章方法对于人体部位和关节点定义间的对应关系，i_{part} 和 i_{joint} 分别代表人体部件和关节点的索引，-：没有对应的关节点

部位	i_{part}^{ours}	i_{part}^{smpl}	i_{joint}^{ours}	i_{joint}^{smpl}
头	10	15	16,17,18,19	15,-,-,-
脖子	9	12	15,16	12,15
胸	6	6,9,13,14	14,15	9,12
腰	3	3	0,5	0,6
臀	0	0	0,1,2	0,1,2
左大腿	1	1	1,3	1,4
左小腿	4	4	3,6	4,7
左脚	7	7,10	6,8,10,12	7,-,-,10
右大腿	2	2	2,4	2,5
右小腿	5	5	4,7	5,8
右脚	8	8,11	7,9,11,13	8,-,-,11
左大臂	11	16	20,22	16,18
左小臂	13	18	22,24	18,20
左手	15	20,22	24,26,28,30	20,-,-,22
右大臂	12	17	21,23	17,19
右小臂	14	19	23,25	19,21
右手	16	21,23	25,27,29,31	21,-,-,23

3.4 对比实验

本节将在拥有不同拓扑且形状与姿态多样性互补的 DFAUST 数据集[41] 和 SPRING[6] 数据集上与现有 SOTA 方法进行重建与编辑方面的实验对比，并且通过收集调查问卷对模型编辑能力进行了更全面的对比。

3.4.1 数据集

DFAUST 数据集捕获了 10 个人体的 14 段不同的动作序列，比如跑步、跳跃、扭髋等等。该数据集中的人体姿态多样性丰富，但是形状多样性较差，所有网格的连通性均一致，采用 SMPL 模型[14] 的拓扑结构，即每个网格拥有 6890 个顶点和 12500 个三角面片。每段动作序列大概 300-500 帧，由于数据量过大且存在信息冗余，本节事先对该数据集进行抽帧处理，均匀抽取二十分之一作为最后的训练集与测试集。最终将抽帧后的数据集随机划分，得到有 182 个网格的测试集与有 1936 个网格的训练集。SPRING 数据集来源于 CAESAR 数据集[42]，采用非刚性变形算法将 CAESAR 数据集注册成网格连通性与 SCAPE 模型一致（即 12500 个顶点，25000 个三角面片）的人体数据。该数据集由三千多个形状各异的人体

网格构成，所以身材多样性丰富，但是人体姿态大致维持在 A-pose，故姿态多样性有限，同样将其随机划分，得到有 305 个网格的测试集与有 2743 个网格的训练集。对于一组训练数据，本章方法需要该组数据的关节回归器、各个顶点的部位语义以及用于测量人体围度的标注点才能开始训练。由于训练数据的网格连通性是一致的（所有的三维人体表示学习工作都需要这一点），所以对于一组训练数据只需要为其注册一次 SMPL 模型[14] 就可以借助 SMPL 模型中的设置来满足上述训练条件。具体来说，对于 DFAUST 数据集[41]，由于其拓扑结构与 SMPL 模型一致，所以可以直接沿用 SMPL 中的设置；对于 SPRING 数据集[6]，本节通过基于关节点损失、倒角损失、掩膜损失等损失的优化算法获得 SCAPE 模型与 SMPL 模型的拓扑对应关系，从而借助 SMPL 模型满足上述训练要求。

3.4.2 重建实验

首先与四类方法比较人体重建精度验证本章方法在几何刻画方面的优势：基于谱域分析的方法（COMA[16]），基于螺旋卷积的方法（Neural3DMM[10]，Spiralplus[17]），基于注意力机制的方法（Pai3DMM[18]，Deep3DMM[19]），形状与姿态解耦的表示方法（DHNN[21]，Unsup[22]）。为了保证实验的公平性，本节均采用上述方法的官方实现，且所有对比实验都采用相同的实验设置，包括隐编码维数、损失函数、学习率、训练轮次等等。由于 DHNN 并未公开训练代码，所以为了与其比较本章方法在其训练集上单独训练了一个特殊版本。最后在评价指标方面，本节利用顶点间欧式距离误差的平均值 E_{avd}（毫米）来评估表示方法的重建精度。

表 3-2 在 DFAUST[41] 和 SPRING[6] 数据集上的重建实验定量结果，Param(M)：模型可学习参数量（百万），- ：不支持该数据集

方法	DFAUST		SPRING	
	E_{avd}	Param(M)	E_{avd}	Param(M)
COMA[16]	6.06	7.54	6.04	6.84
Neural3DMM[10]	5.49	30.35	6.11	27.56
Spiralplus[17]	5.35	15.15	4.99	13.75
Pai3DMM[18]	5.76	15.18	4.45	13.78
Deep3DMM[19]	9.91	8.35	10.88	7.84
Unsup[22]	10.18	12.89	-	-
本方法	**4.70**	**1.59**	**4.33**	**1.47**

定量实验结果如表3-2所示，可以看到本章方法显示出了最出色的性能，并

表 3-3　在 DHNN[21] 数据集上的重建实验定量结果，Param(M)：模型可学习参数量（百万）

方法	DHNN	
	E_{avd}	Param(M)
DHNN[21]	**3.16**	91.63
本章方法	3.96	**1.47**

且得益于骨架引导的自动编码器架构，模型的参数量也是远远小于其他方法。图3-6对部分重建结果和误差图进行了可视化，该可视化结果直观地体现了本章方法在重建精度上的优势，尤其是在几何细节丰富的人体部位（如人手人脸），这也印证了其网络架构设计的合理性和有效性。

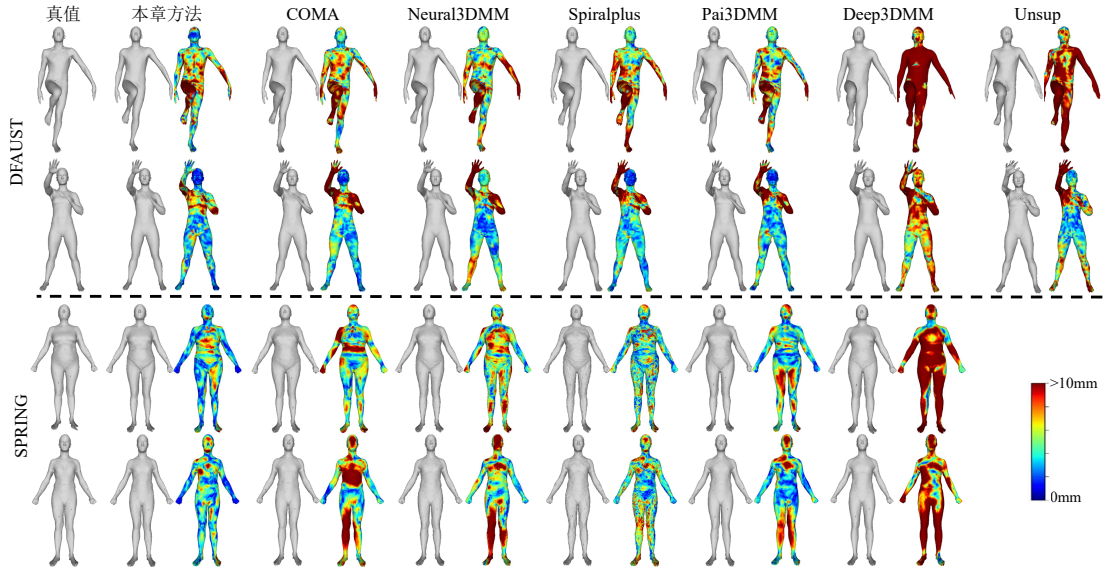


图 3-6　在 DFAUST[41] 和 SPRING[6] 数据集上的重建实验定性结果，其中每个例子左侧为重建网格，右侧为可视化误差。由于 Unsup[22] 对训练数据有限制，即必须给出具有同一形状和不同姿态的配对网格，所以该方法无法在 SPRING 数据集[6] 上训练，而本章方法不需要对人体网格的姿态和形状做出任何限制。

另外表3-3给出了在 DHNN 数据集[21] 上的定量对比结果。值得注意的是 DHNN[21] 依赖于一个严苛的数据假设，即每个网格都必须有一个配对的具有相同形状和中性姿态的网格，这样才能通过监督数据以显式约束的形式实现语义的解耦，而本章方法没有利用任何人体属性上的先验约束。即便在这种不公平的对比条件下，本章方法的重建精度仍然只是略低于 DHNN[21]，并且在模型参数量上也是远远小于 DHNN，另外表示的语义也是更加精细，图3-7展示了一些可视化结果。

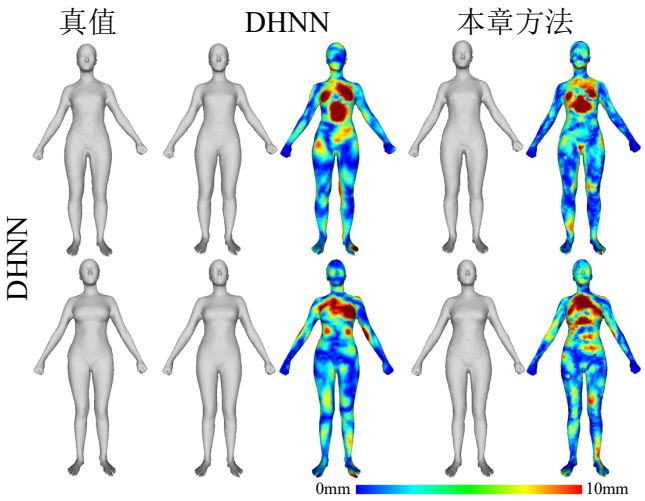


图 3-7 在 DHNN[21] 数据集上的重建实验定性结果，其中每个例子左侧为重建网格，右侧为可视化误差。

3.4.3 编辑实验

另外为了验证本章方法的灵活编辑能力，本节在三个编辑任务上与其他方法进行了对比实验：编辑骨骼方向、编辑骨骼长度、编辑形状尺寸（即围度）。因为骨骼方向等同于人体姿态，所以在该任务上本节与效果最稳定的无监督形状与姿态解耦工作 Unsup[22] 在 DFAUST 数据集[41] 上对比。由于 Unsup[22] 严格意义上不支持严格意义上的人体姿态编辑（本章方法支持这一点），只支持人体姿态迁移，即需要有一个现成网格当做编辑目标，所以为了与其对比本节对于每个测试网格，为其随机挑选另外一个测试网格作为目标姿态网格，然后对于 Unsup[22] 只需要交换两个网格的姿态隐编码再解码即可得到生成网格，对于本章方法只需要根据当前关节点位置和目标骨骼方向计算出姿态改变后的关节点位置，然后把目标关节点位置送入到骨骼分支得到新的骨骼隐编码再解码即可得到生成网格。而骨骼长度和部位围度属于人体形状信息，所以在这两个任务上本章方法与人体重塑工作 HBR[7] 在 SPRING 数据集[6] 上对比。在形状编辑任务中，本节为每个人体部位 x^k 随机设置 $\alpha \cdot l(x^k)$ 和 $\alpha \cdot circ(x^k)$ 为目标编辑长度和围度，其中 α 是从 $(\alpha_{min}, \alpha_{max})$ 随机均匀采样的标量，$l(\cdot)$ 和 $circ(\cdot)$ 是测量人体骨骼长度和部位围度的函数。对于 HBR[7] 只需要输入目标骨骼长度和部位围度即可得到生成网格。对于本章方法，只需要计算出骨骼长度变化后的关节点位置再解码即可得到骨长编辑后的网格，而对于部位围度编辑只需要对相应形状隐编码做缩放操作再解码即可得到围度编辑后的网格。

然而由于人体编辑任务没有真值，所以如何衡量方法的编辑性能是一个值得讨论的问题，本节利用关节点误差 E_{joint} 和围度误差 E_{circ} 来评价编辑骨骼和形状的准确性。具体而言，对于每个测试用例，本节利用 $J(\cdot)$ 和 $circ(\cdot)$ 来计算人体

属性目标值与实际值间的误差 $E_{joint}E_{circ}$（毫米），公式可定义为：

$$E_{joint} = \|T_{joint} - J(x_{edited})\|_2, \tag{3-22}$$

$$E_{circ} = \frac{1}{K}\sum_{k=1}^{K}\left|T_{circ}^{k} - circ(x_{edited}^{k})\right|, \tag{3-23}$$

其中 T_{joint} 和 T_{circ} 是目标关节点位置和目标围度。表3-4给出了编辑实验的定量对比结果。显而易见本章方法在所有的编辑任务上都展现出了最出色的性能，并且在编辑的同时也很好地保留了其他未被编辑的属性。图3.4.3展示了部分可视化结果，本章方法不仅在重建精度上远远领先其他方法（见图3.4.3左侧），并且对于人体属性的编辑更加精确合理自然（见图3.4.3右侧），即便是在非常规的编辑目标下也能得到高质量的人体网格。

表 3-4　编辑实验定量结果，-：不支持该任务

方法	编辑骨骼方向		编辑骨骼长度		编辑形状尺寸	
	E_{joint}	E_{cric}	E_{joint}	E_{cric}	E_{joint}	E_{cric}
Unsup[22]	36.79	14.16	-	-	-	-
HBR[7]	-	-	38.37	15.27	12.87	18.78
本章方法	**3.26**	**9.75**	**1.57**	**5.92**	**0.45**	**17.64**

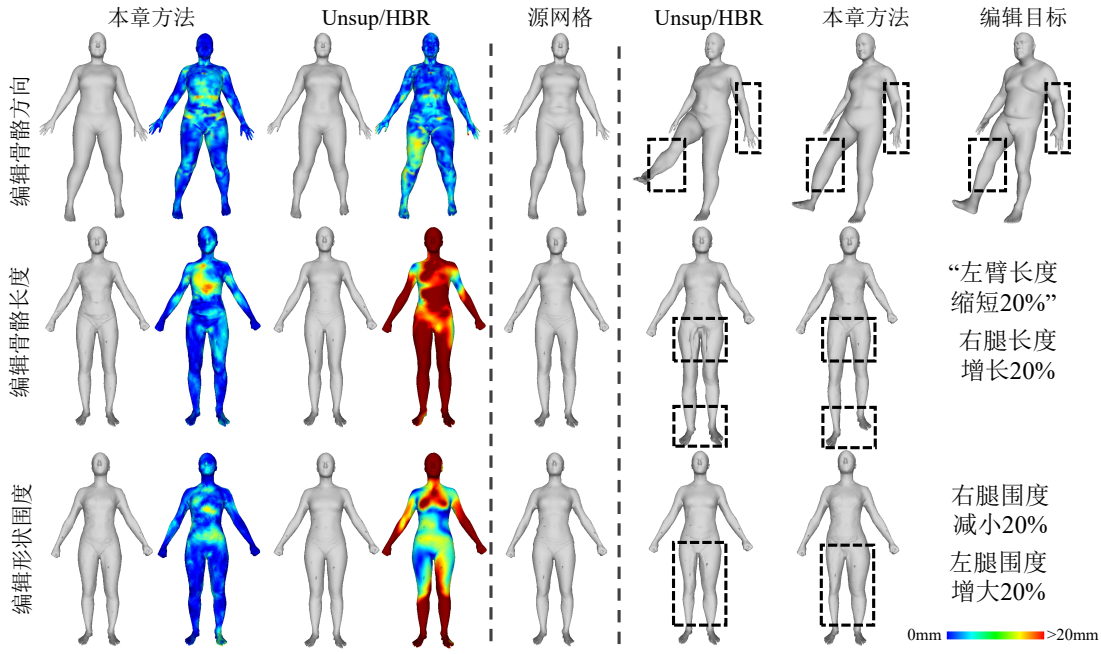


图 3-8　与 Unsup[22]（第一行）和 HBR[7]（第二行和第三行）对比的编辑实验定性结果，在原始网格的左侧和右侧分别为重建网格和编辑后网格

3.4.4 调查问卷

另外由于人体属性误差并不能完全反映出方法的编辑性能，所以为了更公正且完善地评估模型的编辑能力本节还设计了用户调查问卷。调查问卷由三组测试组成，第一组测试展示 Unsup[22] 和本章方法在骨骼方向编辑任务上的四个编辑结果，第二组测试展示 HBR[7] 和本章方法在骨骼长度编辑任务上的三个编辑结果，第三组测试展示 HBR[7] 和本章方法在部位围度编辑任务上的三个编辑结果。用户需要从两个方面来评价方法的编辑能力：首先编辑属性是否被自然合理且精确地被改变为目标值，另外其他未编辑属性是否被完整地保留下来。本节一共收集了 102 份问卷，包括 28 位女士和 74 位男士，其中年龄处于 18 岁以下的有 2 个人，年龄处于 18 岁到 40 岁的有 96 个人，年龄处于 40 岁到 60 岁的有 4 个人。统计了被认为在改变属性和保留属性方面具有更好性能的每种方法的百分比 P_{cha} 和 P_{pre}。表3-5给出了最终的统计结果，该结果有力地印证了本章方法拥有更灵活且精确的编辑能力。

表 3-5 认为该方法编辑效果最好的人占所有人的比例，- ：不支持该任务

方法	编辑骨骼方向		编辑骨骼长度		编辑形状尺寸	
	P_{cha}	P_{pre}	P_{cha}	P_{pre}	P_{cha}	P_{pre}
Unsup[22]	27.95%	36.27%	-	-	-	-
HBR[7]	-	-	27.43%	34.63%	40.20%	40.16%
本章方法	**72.05%**	**63.73%**	**72.57%**	**65.37%**	**59.80%**	**59.84%**

3.5 消融实验

本节将通过完备的消融实验证明所提出的各个模块的必要性和有效性。由于重建损失 L_{rec} 是所有表示学习类工作的基础损失，所有的其他损失必须建立在该损失的基础上，所以本节只对除了重建损失之外的其他损失进行消融。

3.5.1 边长正则项、解耦损失、编辑损失

为了更好地约束重建网格的平滑性与合理性，本章在第3.3.1节引入了边长正则项 L_{edge}，为了摆脱对配对监督网格的依赖实现无监督的语义解耦，本章在第3.3.2节引入了解耦损失 L_{dis}，为了赋予表示部位级别灵活形状编辑的能力，本章在第3.3.3节引入了编辑损失 L_{edit}。本节在训练过程中分别去除这些监督以评估这些损失的影响。表3-6给出了定量消融结果，对表分析可得 L_{edge} 有效地降低了

重建误差，L_{dis} 和 L_{edit} 的使用分别赋予了表示灵活控制骨骼和形状的能力，该实验成功证明了上述所有损失的必要性和有效性。

表 3-6　在重建与编辑任务上的定量消融结果，-：不支持该任务

方法	重建实验			编辑实验						
	DFA.	SPR.	均值	骨骼方向		骨骼长度		形状尺寸		均值
				E_{joint}	E_{circ}	E_{joint}	E_{circ}	E_{joint}	E_{circ}	
完整方法	4.70	**4.33**	**4.52**	3.26	**9.75**	1.57	**5.92**	**0.45**	17.64	**6.43**
无 OAW	**4.66**	4.39	4.53	3.02	13.66	1.58	10.95	0.61	20.40	8.37
无 L_{edge}	5.23	5.01	5.12	3.20	9.76	**1.52**	8.37	0.53	29.48	8.81
无 L_{dis}	4.81	4.88	4.85	-	-	-	-	1.56	**17.50**	-
无 L_{edit}	4.87	4.43	4.65	**3.00**	10.19	1.88	8.01	-	-	-

3.5.2　方向自适应权重机制

由于骨骼的长度信息耦合在人体部位的欧式距离矩阵中，从而导致不彻底的骨骼与形状解耦，影响后续人体编辑等应用的精度，为了缓解这个问题本章在第3.3.2节出了方向自适应权重机制。本节通过在训练过程消除该机制来验证其有效性。正如表3-6所示，该机制的引入能够帮助本表示更好地关注沿着骨骼正交方向 $o_{b\perp}$ 的几何特征，进而实现更彻底的解耦和更精确的人体编辑。

3.5.3　体积正则项

OAW 机制的引入在缓解不彻底解耦的同时也会导致网络在训练时忽略部分有用的沿骨骼方向 o_b 的几何信息，从而导致在大幅度编辑人体骨骼方向时出现不合理的网格变形和伪影，为了解决这个问题本章在第3.3.2节提出体积正则项。为了分析体积正则项的影响，本节在训练过程中将其去掉。图3-9展示了是否使用该正则项的情况下编辑人体的对比结果，显然该体积正则项可以为保留人体几何提供强有力的约束，进而使得编辑结果更加自然且合理。

3.6　本章小结

本章详细介绍了基于部位解耦的三维人体表示方法，首先基于第二章对于现有方法的介绍，深入分析了这些工作存在的问题及其背后的原因，接着引出本表示方法的核心——骨骼分离的部位解耦策略。随后基于此局部解耦策略，介绍了本章设计的一系列新颖且有效的网络框架、训练流程、损失函数、训练细节。最

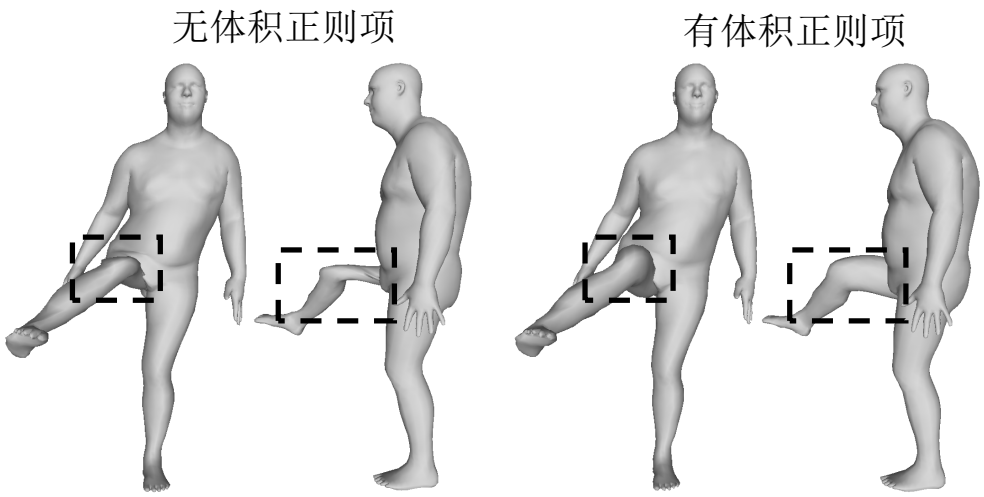


图 3-9　体积正则项的定性消融结果

后在实验分析部分介绍了本章使用的人体网格数据集，并在该数据集上与其他方法在人体重建和编辑方面进行了完备的定量和定性对比，充分证明了本章方法在几何刻画和人体编辑方面的出色性能，并补充了完善的消融实验，进一步验证了所提出的损失和正则项的有效性和必要性。具体来说，第 3.1 节介绍了骨骼分离的部位解耦策略，并结合相关示意图形象且直观地说明了该策略的具体细节和显著优势。第 3.2 节介绍了契合局部解耦思路的骨架引导的自动编码器网络架构，这种能够聚合人体部位的局部与全局特征的架构设计不仅有利于人体多尺度特征的分析与提取，而且极大地降低了模型的可学习参数，使表示更加的轻量化。第 3.3 节介绍了具体的训练流程和损失函数，训练流程主要分为三个分支：重建分支、解耦分支、编辑分支，在对应损失函数的帮助下三个分支分别赋予表示精确几何重建、无监督语义解耦、部位级别灵活编辑的能力。第 3.4 节则在具有不同拓扑结构且形状与姿态多样性互补的 DFAUST 数据集[41] 和 SPRING[6] 数据集上与现有 SOTA 方法进行重建与编辑方面的实验对比，并且通过收集调查问卷对模型编辑能力进行多方面的评估。在第 3.5 节中通过完备的消融实验证明所提出的各个损失和正则项的必要性和有效性。

第 4 章　面向人体重建、编辑与生成任务的解耦表示应用

本章承接上一章提出的基于部位解耦的三维人体表示方法，以该表示方法为基础面向人体重建、编辑与生成应用任务给出了新颖且具体的实践方案。首先针对基于单目人体掩膜的人体重建问题，通过在基础方案下与主流人体表示工作进行对比验证了该表示方法在重建精度上的优势，并在此基础上结合该表示“骨骼分离”、“部位解耦”的特点，提出能够有侧重地关注各部位图像特征的局部与全局注意力引导的人体重建网络，充分证明了该表示方法的潜力与可拓展性。其次面向人体编辑问题，得益于该表示细粒度的语义，用户可以通过调整对应隐变量来实现人体部位级别的骨骼方向、骨骼长度、形状围度、形状风格的灵活可控编辑，进一步展现了其灵活性与可编辑性。最后借助表示所学习到的具有明确几何意义的隐空间，该方法可以通过在隐空间上线性插值或随机采样生成大量高质量的人体网格数据，为后续的人体网格研究提供了有力的数据支持。

4.1　基于单目人体掩膜的人体重建

人体重建即对人体的三维几何结构进行重建，是对人体进行数字化的重要手段，作为一个十分具有挑战性的问题受到研究者们的广泛关注。在过去很长一段时间中，该问题的解决依赖于专业且笨重的采集系统，该系统一般由深度相机、相机阵列、扫描仪等设备组成，且有的方法[43–45] 需要长时间的离线处理才能得到最终的三维人体。另外有的方法[46–50] 仅需要一台深度相机即可快速重建人体，但是即便是消费级深度相机其价格仍比较昂贵，不适用于大多数用户。因此基于 2D 图像的人体重建问题因其简单便捷、对用户友好等特点引起人们的广泛研究，且随着深度学习的快速发展，一系列令人印象深刻的工作[1,3,4,51–58] 随之涌出，这些方法可以根据不同的方案设置被分类为基于单目/多目输入的人体重建方法、基于 RGB/掩膜人体重建方法、基于图像/视频的人体重建方法、基于显式模板/隐式场函数的人体重建方法。其中基于单目人体掩膜的人体重建得益于其独特的问题设置，既不需要担心数据集间的域差异（Domain Gap）问题，在虚拟数据集上训练的网络无需额外处理即可应用在真实场景中，又便于获取，对用户友好，具有重大的实际应用意义，也是本节研究的核心应用场景。其问题具体定义为输入人体掩膜 $x \in \mathbb{R}^{h \times w}$，即人体二值图，属于人体区域部分的像素值为 1，其余区域为 0，输出三维人体网格 $x' \in \mathbb{R}^{n \times 3}$，重建人体与真值 $x^{gt} \in \mathbb{R}^{n \times 3}$ 越接近代

表重建越精确，问题示意图如图4-1所示。

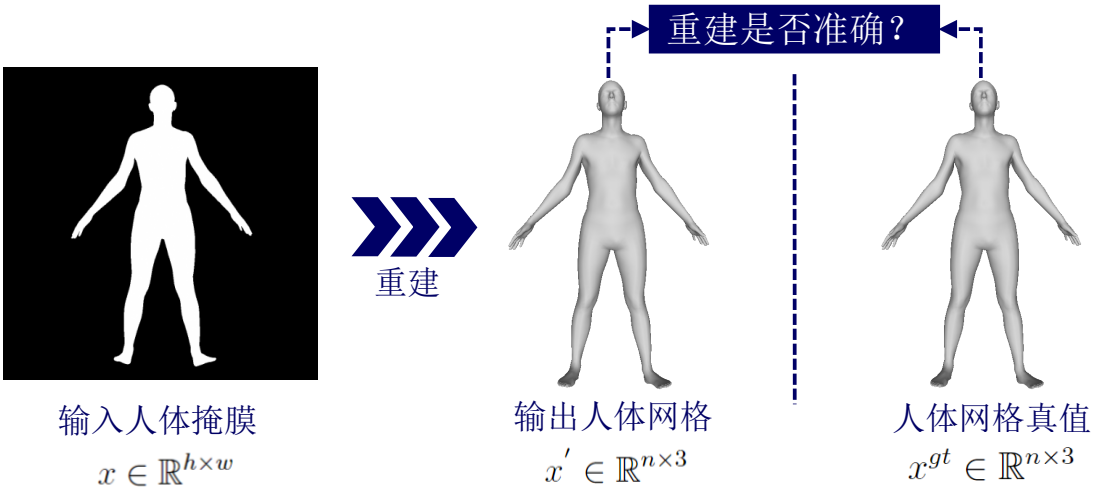


图 4-1　　人体重建问题示意图

4.1.1　基于迭代求解的人体重建网络

正如上文中介绍，人体重建工作可以分为基于显式模板（即人体表示）和隐式场函数的方法，其中基于显式模板的方法采用预先定义的人体表示模型，将人体重建任务从一个 2D 到 3D 的高度不适定问题简化为人体表示的表示参数回归问题，因为这类方法利用了人体几何结构的强先验知识，所以重建结果比较稳定，因此这类方法也成为了人体重建工作的主流。基于人体表示的方法大都遵循相似的网络框架，即输入图像先通过预训练模型提取图像特征，特征再进入不同的参数回归头来估计表示参数，最终参数通过人体表示模型转化成为三维人体网格进而完成对人体的重建。其中基于迭代求解的人体重建网络 HMR[3] 是最为经典且有效的网络框架之一，其网络框架如图4-2所示。

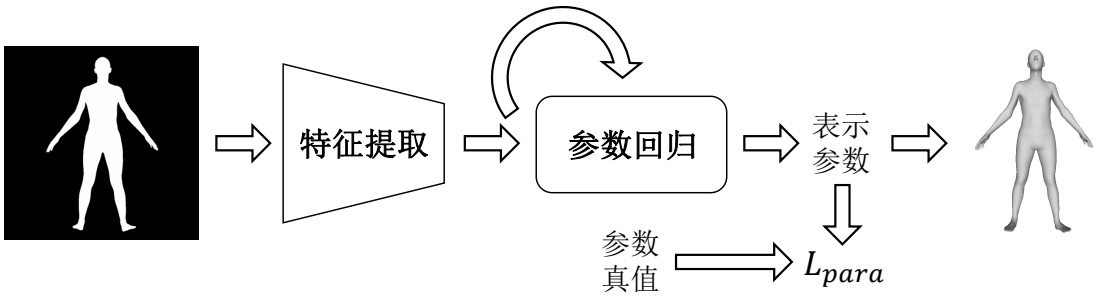


图 4-2　　基于迭代求解的人体重建网络示意图

具体而言，输入图像 I 先进入特征提取器（如 Resnet、Hrnet）获得其图像特征 F，常规的做法是后面接回归头直接根据图像特征完成对表示参数的回归，然

而直接从 2D 特征回归 3D 人体表示参数难度太大，所以 HMR[3] 提出了迭代求解的机制，即回归模块不再只以图像特征 F 为输入，而是以图像特征 F 和当前估计参数 θ_t 为输入，输出残差 $\Delta\theta_t$，并将此残差与当前估计值相加得到新的估计值，即 $\theta_{t+1} = \theta_t + \Delta\theta_t$，其中初始估计值 θ_0 一般设为定值。这种迭代求解的思路通过逐步预测表示参数的残差来逐渐缩小预测值与真值间的误差，从而比起普通的直接回归参数的重建网络更能准确且有效地重建人体。同时该网络架构简单直接，对回归的人体表示参数没有特定要求，因此该架构可被拓展到多种人体表示上，如 SMPL[14] 和本文提出的表示。

4.1.2 局部注意力引导的人体重建网络

上述基于迭代求解的人体重建网络虽然结构简单直接，可拓展性强，可由于图像特征域与表示参数域之间的差异过大，根据 2D 图像全局特征去直接估计全局人体参数的难度过大，这导致该网络在精度和鲁棒性方面表现不佳。针对该问题，本节基于本表示方法的“部位解耦”、“骨骼分离”的特点，提出了局部注意力引导的人体重建网络，不同于上述基于迭代求解的人体重建网络根据全局图像特征回归全局人体参数的思路，该网络在提取全局图像特征的同时对输入人体掩膜进行人体部位语义分割，并根据分割结果计算局部注意力以帮助网络有侧重地关注各部位的局部图像信息，接着将各部位图像特征送入对应的局部人体参数回归器得到各人体部位的表示参数，从而最终在由上一章提出的融合解码器 $D(\cdot)$ 的帮助下恢复出三维人体网格。这种局部图像特征回归局部人体参数的思路不仅完美契合了本文提出的基于部位解耦的三维人体表示，并且极大地简化了神经网络需要学习的特征映射空间，在增强网络鲁棒性的同时也大幅提升了人体重建精度。

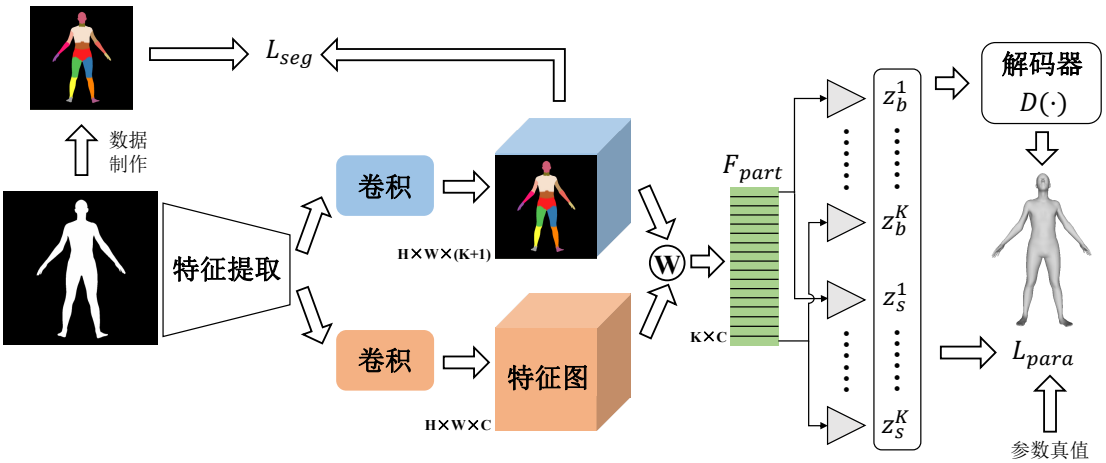


图 4-3 局部注意力引导的人体重建网络

其网络框架如图4-3所示，具体来说，对于输入图像 I，先将其送入特征提取器获得初步特征图 F_l，接着进入特征提取分支（图4-3红色部分）得到进一步提取后的特征 $F_h \in \mathbb{R}^{H\times W\times C}$，其中 H 和 W 分别是特征图的高与宽，C 为通道数。与此同时语义分割支路（图4-3蓝色部分）根据 F_l 得到 K 个人体部位加背景的分割结果 $P \in \mathbb{R}^{H\times W\times (K+1)}$，再对分割结果做空间维度的归一化处理就可以得到用于提取局部人体部位特征的局部注意力，接着局部注意力与特征图点乘得到代表着人体各部位局部信息的特征 $F_{part} \in \mathbb{R}^{K\times C}$（忽略背景维度），计算过程如下式所示：

$$F_{part}^{k,c} = \sum_{h,w} \sigma(P^k)^T \otimes F_h^c, \tag{4-1}$$

其中 P^k 和 F_h^c 分别代表 P 和 F_h 的第 k 和 c 个通道，$\sigma(\cdot)$ 代表在空间维度的 softmax 归一化操作，$\otimes$ 为矩阵间的点乘运算。直观理解，局部注意力 $\sigma(P^k)$ 可以在空间域更好地聚合特征，显式地帮助网络关注到各个部位对应的局部图像，从而更精确地估计对应部位的表示参数。

然后 F_{part} 中各部位通道的特征被输入到对应的回归器中用于估计该部位的骨骼隐编码与形状隐编码，进而最终通过表示的融合解码器 $D(\cdot)$ 恢复得到三维人体网格 x'，计算过程如下式所示：

$$z_s^k = R_s^k(F_{part}^k), \tag{4-2}$$

$$z_b^k = R_b^k(F_{part}^k), \tag{4-3}$$

$$x' = D(Z_s, Z_b), \tag{4-4}$$

其中 R_s^k 和 R_b^k 分别是估计各部位骨骼隐编码和形状隐编码的回归器。直观理解，在局部注意力的引导下，网络在估计各部位表示参数时能够更聚焦于对应的局部图像特征，比如估计腿部的表示参数时会更关注腿部的图像特征，估计胸部的表示参数时会更关注胸部的图像特征，这无形中简化了网络需要学习的特征映射空间，降低了训练难度，从而实现精确性和鲁棒性更好的参数估计和网格重建。

在损失方面，该网络的损失函数由表示参数损失 L_{para} 和语义分割损失 L_{seg} 构成，具体细节如下式所示：

$$L = L_{para} + \lambda_{seg} L_{seg}, \tag{4-5}$$

$$L_{para} = \|Z_s - Z_s^{gt}\|_1 + \|Z_b - Z_b^{gt}\|_1, \tag{4-6}$$

$$L_{seg} = \frac{1}{HW} \sum_{h,w} cross(sig(P_{h,w}), P_{h,w}^{gt}), \tag{4-7}$$

其中 λ_{seg} 是语义分割损失的权重因子，Z_s^{gt} 和 Z_b^{gt} 是真值网格 x^{gt} 的真值表示参数，参数损失 L_{para} 通过施加参数间的 $L1$ 约束来监督网络准确地估计表示参数。$P_{h,w} \in \mathbb{R}^{(K+1)}$ 代表在 (h,w) 位置上的语义分割结果，$P_{h,w}^{gt} \in \mathbb{R}^{(K+1)}$ 代表该位置的经过 one-hot 编码的语义分割真值，该真值可以由 x^{gt} 直接渲染得来；$sig(\cdot)$ 为

sigmoid 归一化函数，$cross(\cdot)$ 为用于多分类任务的交叉熵函数，语义分割损失 L_{seg} 通过实现像素级别的损失约束来显式地指导网络对输入图像中的人体部位进行语义分割，进而帮助网络更好地聚焦局部图像特征。

4.1.3 局部与全局注意力引导的人体重建网络

上述局部注意力引导的人体重建网络得益于部位语义分割分支的引入，能够在由分割结果计算得来的局部注意力的引导下实现更精确的几何重建，但是局部注意力的引入是一把双刃剑，它在帮助网络聚焦各部位局部图像特征同时也会使其忽略存在于其他部位图像中的有用信息（比如对于人体胸部参数的估计，腰部、臀部、脖颈部的图像特征也能提供一定的帮助），因此我们提出局部与全局注意力引导的人体重建网络，该网络在原有的语义分割分支的基础上引入自适应学习的全局注意力分支，意在辅助网络捕获在语义分割分支中丢失掉的其余人体部位信息，从而进一步提升网络的重建精度。

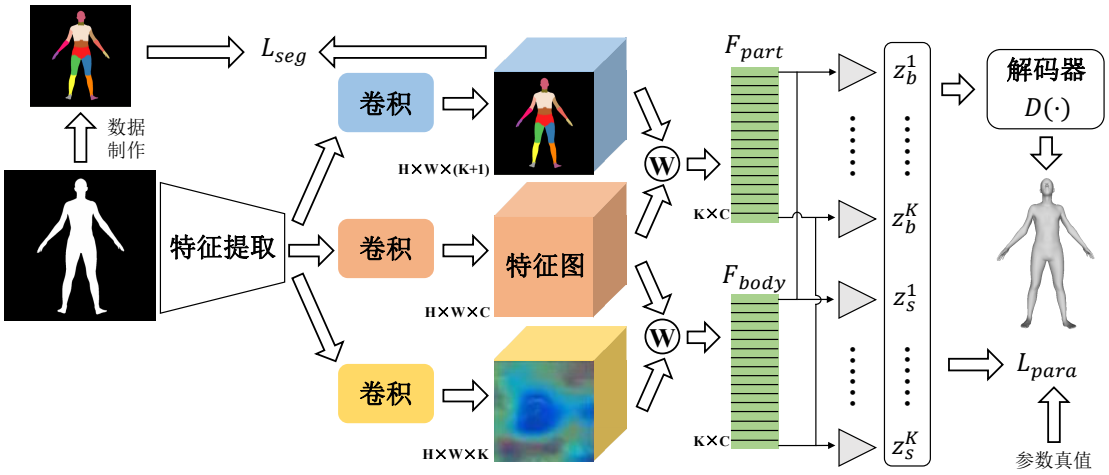


图 4-4　　局部与全局注意力引导的人体重建网络

其网络框架如图4-4所示，具体来说，全局注意力分支（图4-4黄色部分）与语义分割分支类似，都是根据 F_l 计算 K 个人体部位的全局注意力特征 $B \in \mathbb{R}^{H\times W\times K}$，再对其做空间维度的归一化处理得到全局注意力，最后全局注意力与特征图点乘得到具有人体所有部位形状信息的全局特征 $F_{body} \in \mathbb{R}^{K\times C}$，计算过程如下式所示：

$$F_{body}^{k,c} = \sum_{h,w} \sigma(B^k)^T \otimes F_h^c. \tag{4-8}$$

上式定义与式4-1类似，但是与语义分割分支不同的是在训练时并不对该分支的输出施加语义分割损失约束，而是让网络自适应地去学习全局注意力分布，从而辅助网络捕获丢失的人体信息。然后 F_{part} 和 F_{body} 中属于相同部位通道的特征合

并在一起输入到对应的回归器中去估计该部位的骨骼隐编码与形状隐编码，进而最终通过表示的解码器恢复得到三维人体网格 x'，计算过程如下式所示：

$$F_{final}^{k} = F_{part}^{k} \oplus F_{body}^{k}, \tag{4-9}$$

$$z_{s}^{k} = R_{s}^{k}(F_{final}^{k}), \tag{4-10}$$

$$z_{b}^{k} = R_{b}^{k}(F_{final}^{k}), \tag{4-11}$$

$$x' = D(Z_{s}, Z_{b}), \tag{4-12}$$

其中 $\oplus$ 代表特征向量间的合并操作，R_{s}^{k} 和 R_{b}^{k} 分别是估计各部位骨骼隐编码和形状隐编码的回归器。直观理解，在局部注意力和全局注意力的通力合作下，网络在估计各部位表示参数时既可以聚焦于对应的局部图像特征，又可以自适应地捕获其他身体部分中有用的信息，从而进一步实现精确且鲁棒的参数估计和网格重建。

在损失方面，由于新引入的全局注意力分支并不需要损失约束，所以该网络的损失函数仍然沿用第4.1.2节的定义。

4.1.4 实现细节

在代码实现方面，本节的算法均在 Pytorch 框架[39] 下实现，特征提取器采用 ResNet-50 网络，所有的训练和测试实验均在装载着 RTX 3090 GPU 的 Ubuntu 服务器上实现。在训练方面，网络学习率设置为 1×10^{-3}，并且在每个训练轮次结束后以 0.99 的衰减率衰减一次，网络在 Adam 优化器[40] 下训练 200 轮次，整个训练时间不超过 12 个小时。在超参数设置方面，损失权重因子 λ_{seg} 被设置为 0.1。

4.1.5 数据集

为了评估表示在基于单目人体掩膜的人体重建任务上的性能，本节利用 SMPL[14] 模型随机生成的虚拟三维人体网格进行训练与测试。具体而言通过对 SMPL 的 β 和 θ 随机采样生成了 1000 个身材丰富多样且大致保持 A-pose（为了降低网络学习的难度）的人体网格，其中 900 个用于训练，100 个用于测试。除此之外本节还从 FAUST[59] 和 DFAUST[41] 挑选了部分真实扫描的三维人体模型作为额外的真实测试集。然后每个人体模型渲染一张分辨率为 224×224 的正面二值掩膜图作为网络的输入，其中训练集根据第三章中对人体部位的划分额外渲染一张同等尺寸的语义分割图作为语义分割分支的监督，最后为了计算 L_{para} 还需准备训练集的真值参数，对于基于 SMPL 的重建实验本节直接采用生成数据时使用的参数作为真值；对于基于本表示方法的重建实验，本节在训练数据集上进行训练然后对人体网格进行编码以获取真值参数。

4.1.6 对比实验

为了证明本表示在人体重建任务上的优势以及提出的注意力引导的网络结构的有效性，本节将在虚拟人体数据集（简写为 Vir_Hm）和真实人体数据集（简写为 Real_Hm）上对 SMPL 表示下的基于迭代求解的人体重建网络（简写为 HMR_SMPL）、本表示下的基于迭代求解的人体重建网络（简写为 HMR_Ours）、本表示下的局部注意力引导的人体重建网络（简写为 Att_Sin）、本表示下的局部与全局注意力引导的人体重建网络（简写为 Att_Dou）这四种网络框架进行重建精度上的对比。在评价指标方面，本节沿用第三章中的设置，采用顶点间欧式距离误差的平均值 E_{avd}（毫米）来评估网络的重建精度。

表 4-1 四种网络框架在虚拟人体数据集和真实人体数据集数据集上的重建实验定量结果

方法	Vir_Hm	Real_Hm
HMR_SMPL	41.84	74.41
HMR_Ours	35.79	61.74
Att_Sin	34.05	61.78
Att_Dou	**32.89**	**58.55**

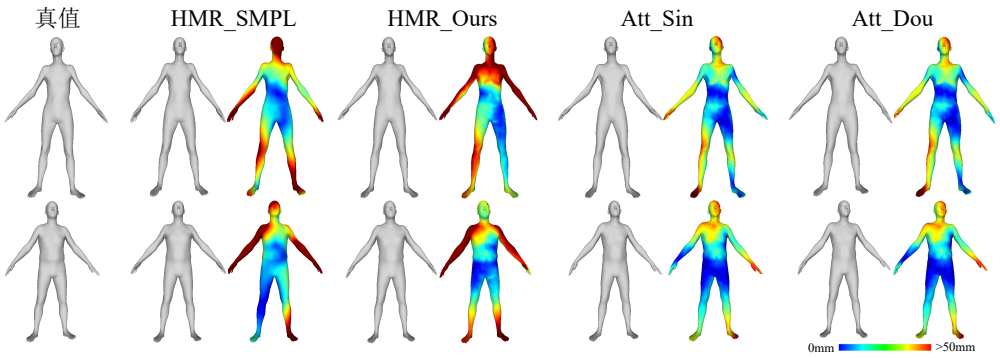


图 4-5 四种方案的重建实验定性结果，其中每个例子左侧为重建网格，右侧为可视化误差。

定量实验结果如表4-1所示，比较 HMR_SMPL 和 HMR_Ours 的结果可以得出在相同的网络架构下，由于本文提出的表示具有部位解耦的特点，所以比起 SMPL[14] 更容易被网络学习，从而达到了更高的人体重建精度；比较 HMR_Ours、Att_Sin 和 Att_Dou 的结果可得本文提出的局部与全局注意力引导的网络框架比起基本的迭代求解和仅有局部注意力的框架更能合理有效地提取与聚合图像特征，提升回归表示参数的准确性，以实现准确的人体重建。图4-5对部分重建结果和误差图进行了可视化，该结果直观地体现了本文提出的表示与局部与全局注

意力引导的网络结构在人体重建任务上的性能优势，此外图4-6还对该网络的局部与全局注意力进行了可视化，显而易见局部注意力更集中，通常聚焦于人体各部位，而全局注意力则更加分散，对输入图像各部分都有所关注，正是这种局部注意力为主，引导网络关注对应部位特征，全局注意力为辅，帮助网络补充遗漏信息的组合显著地提升了该模型的重建性能。

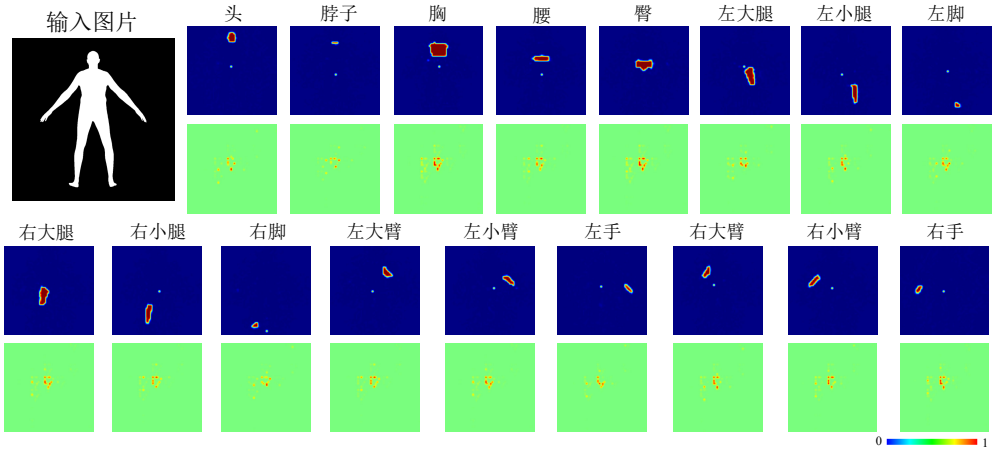


图 4-6　Att_Dou 的局部与全局注意力可视化结果，其中第一行为局部注意力可视化，第二行为全局注意力可视化

4.1.7　应用实例

通过借助 Inner-Body[60] 中提出的 SUNet 网络（输入为人体着衣 RGB 图像，输出为人体裸身掩膜的神经网络模型，具体细节见论文[60]），即可实现根据单张着衣人体照片重建裸身人体三维模型，应用实例如图4-7所示，输入一张用户穿着日常衣服的自拍照片，SUNet 在语义分割图和关节点图的引导下去除衣服影响，得到裸身人体的二值化掩膜图，接着人体掩膜图被送入到上文提出的局部与全局注意力引导的人体重建网络中得到最终的三维人体模型。该方案可以在不侵犯用户隐私权的前提下完成对裸身三维人体模型的重建，对于虚拟试衣、人体测量、AR/VR 等领域有重大意义[61]。

4.2　部位级别的灵活可控人体编辑

人体编辑即根据编辑目标对三维人体属性（如姿态、胸围、腰围、腿长等）进行编辑。解决该问题一般采用两类思路，第一类思路采用机器学习的方法[6,7,13]，根据大规模人体扫描数据集建立人体测量参数与网格顶点位移间的映射，映射建立完成后就可以通过调整目标人体测量参数来编辑符合预期的人体网格，比如第三章中对比的 HBR[7]，但是这类方法的泛化性有限，一旦编辑目标的人体属性分布与训练集的属性分布差别较大，编辑结果往往会出现严重的伪影。第二

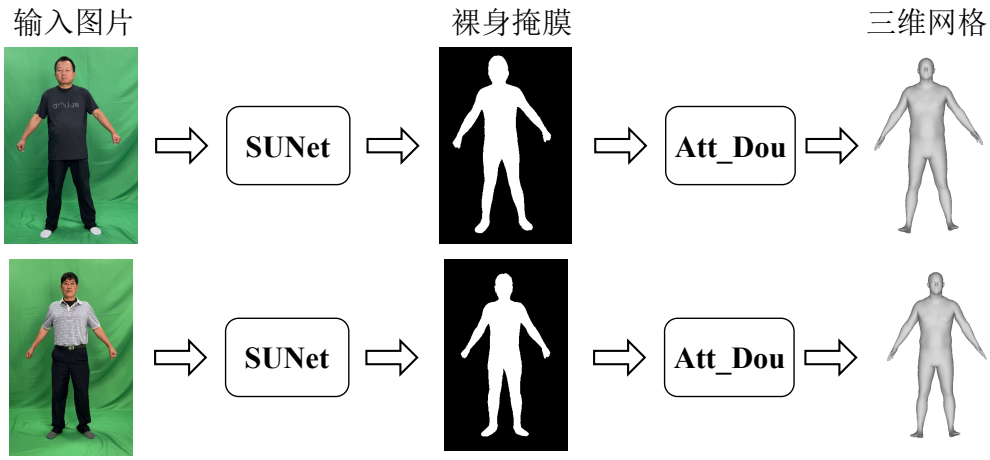


图 4-7　　人体重建应用实例

类思路采用计算机图形学中的网格变形方法，比如拉普拉斯变形[62]，但是这类方法需要人工选择锚点，对用户不友好，且该算法并没有考虑人体网格的结构先验，所以变形后的网格合理性也无法得到保障。而本表示由于具有几何意义明确的隐空间，可以直接被应用于人体编辑问题，且得益于表示语义的细粒度，用户可以在部位级别对人体进行灵活编辑，下面将详细介绍在多个人体编辑任务上的具体实践方案。

4.2.1　骨骼方向编辑

对于输入网格 x，首先利用关节点回归器 $J(\cdot)$ 计算其关节点位置 Kp，接着计算出 x 骨骼的当前方向，再根据目标骨骼方向得出编辑后的关节点位置 Kp^{new}，再把目标关节点位置 Kp^{new} 输入到编码器的骨骼分支 $E_b(\cdot)$ 得到新的骨骼隐编码 Z_b^{new}，接着把网格 x 送入到编码器的形状分支 $E_s(\cdot)$ 得到原始的形状隐编码 Z_s，最后把骨骼隐编码 Z_b^{new} 和形状隐编码 Z_s 送入到融合解码器 $D(\cdot)$ 得到全新的骨骼方向编辑后的人体网格 x^{new}。以上过程的公式化表述如下式所示：

$$Z_b^{new} = E_b(Kp^{new}), \tag{4-13}$$

$$Z_s = E_s(x), \tag{4-14}$$

$$x^{new} = D(Z_b^{new}, Z_s). \tag{4-15}$$

图4-8展示了骨骼方向编辑的网格变化过程。具体而言，本节把一段跑步动作序列作为编辑目标，然后遵循上述方法对全身或局部骨骼方向进行编辑，最终从结果中挑选若干帧作为可视化结果，其中图4-8第一行为编辑全身骨骼方向的结果，第二行和第三行分别为仅编辑腿和胳膊骨骼方向的结果。显而易见本表示方法支持部件级别的高质量人体编辑，而这种灵活的可编辑性也是该表示最大的亮点与特点，并且更注意的是整个训练过程不依赖配对的监督数据，这极大地降

低了表示应用于实际场景的难度。

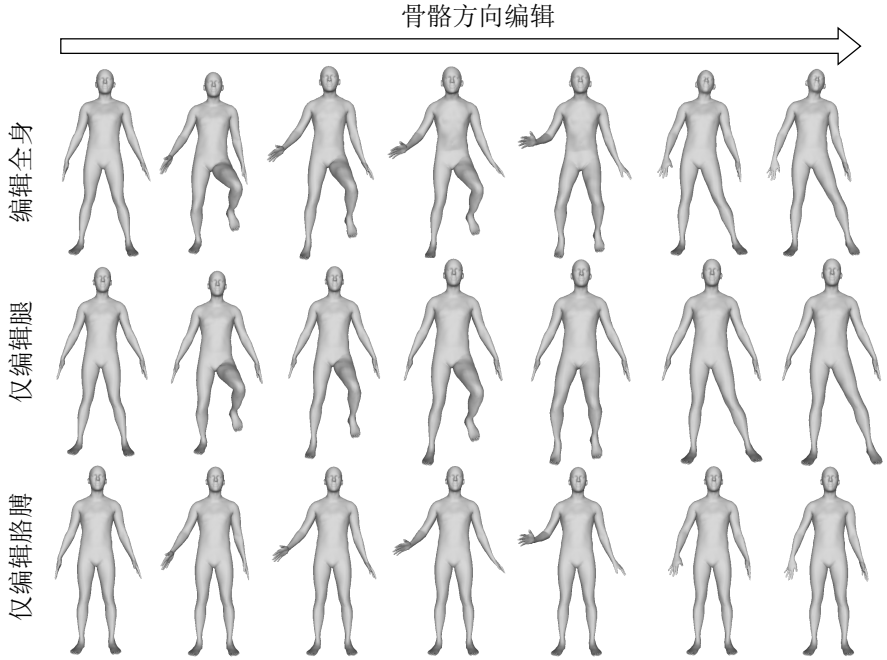


图 4-8　　骨骼方向编辑示意图

4.2.2　骨骼长度编辑

骨骼长度编辑与上述的骨骼方向编辑类似，都是通过改变目标关节点位置来实现人体属性的编辑，只不过在改变关节点位置时不再是修改骨骼的方向而是修改骨骼的长度，目标关节点位置确定后采用和上述一样的方法来得到骨骼长度编辑后的人体网格。

图4-9展示了骨骼长度编辑的网格变化过程。具体而言，本节把骨骼长度延长为原来的 1.2 倍作为编辑目标，然后遵循上述方法对全身或局部骨骼长度进行编辑，然后截取变化过程中的某几帧作为可视化结果，其中图4-9第一行为编辑全身骨骼长度的结果，第二行、第三行和第四行分别为编辑腿、胳膊和躯干骨骼长度的结果，显然无论是编辑全身还是局部的骨骼长度，本表示方法都能得到自然且合理的人体网格。

4.2.3　部位围度编辑

部位围度编辑则与上述两种编辑不同，是通过改变形状隐编码来实现编辑目标。具体而言，对于输入网格 x 的形状隐编码 Z_s，其由多个局部形状隐编码 $\{z_s^1, ..., z_s^K\}$ 组成，每个局部形状隐编码是一个向量，而在第三章中的编辑分支引入的编辑损失已经成功地让形状隐编码的模长和方向分别代表部件的围度和风格，因此可以在通过让形状隐编码与缩放因子相乘来改变其模长，最终达到精确

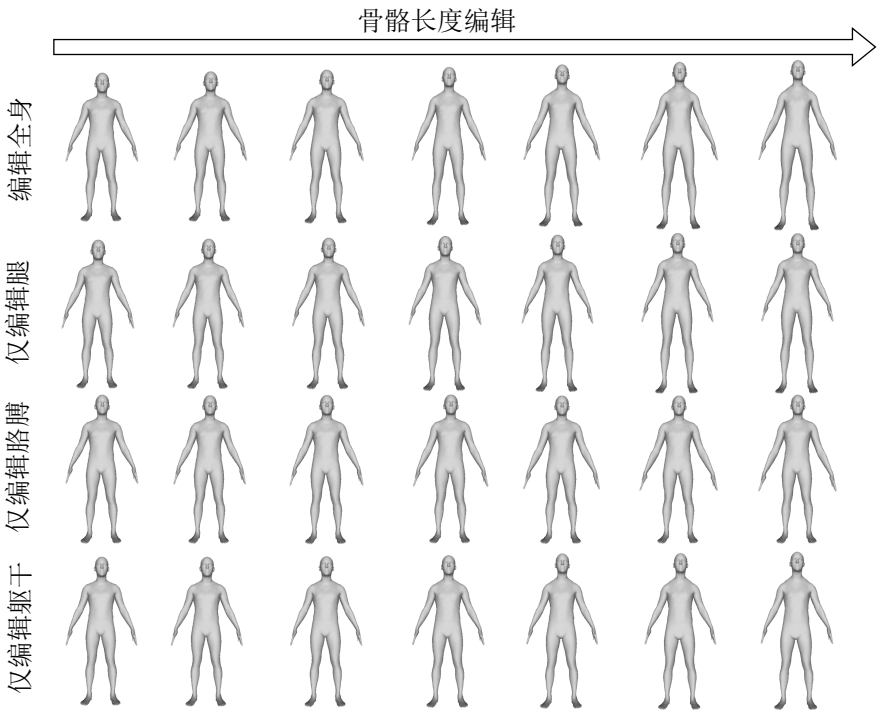


图 4-9　　骨骼长度编辑示意图

编辑部位围度的目的。比如用户想让部件围度扩大为原来的 1.2 倍，只需要令标量 $\alpha = 1.2$，然后将其与对应的形状隐编码 Z_s 相乘得到新的形状隐编码 Z_s^{new}，接着把网格 x 送入到编码器的骨骼分支 $E_b(\cdot)$ 得到原始的骨骼隐编码 Z_b，最后把骨骼隐编码 Z_b 和形状隐编码 Z_s^{new} 送入到融合解码器 $D(\cdot)$ 得到全新的部位围度编辑后的人体网格 x^{new}。以上过程的公式化表述如下式所示：

$$Z_b = E_b(P), \tag{4-16}$$

$$Z_s = E_s(x), \tag{4-17}$$

$$Z_s^{new} = \alpha \cdot Z_s, \tag{4-18}$$

$$x^{new} = D(Z_b, Z_s^{new}). \tag{4-19}$$

图4-10展示了形状围度编辑的网格变化过程。具体而言，将形状围度扩大为原来的 1.2 倍作为编辑目标，然后遵循上述方法对全身或局部围度进行编辑，然后截取变化过程中的某几帧作为可视化结果，其中图4-10第一行为编辑全身围度的结果，第二行、第三行和第四行分别为编辑腿、胳膊和躯干围度的结果，显然本表示方法可以通过编辑形状隐变量的模长来自然且合理地改变人体围度。

4.2.4　形状风格迁移

形状风格迁移与部件围度编辑一样，都是通过改变形状隐编码来实现编辑目标。正如上节中介绍的，编辑分支引入的编辑损失已经成功地让形状隐编码

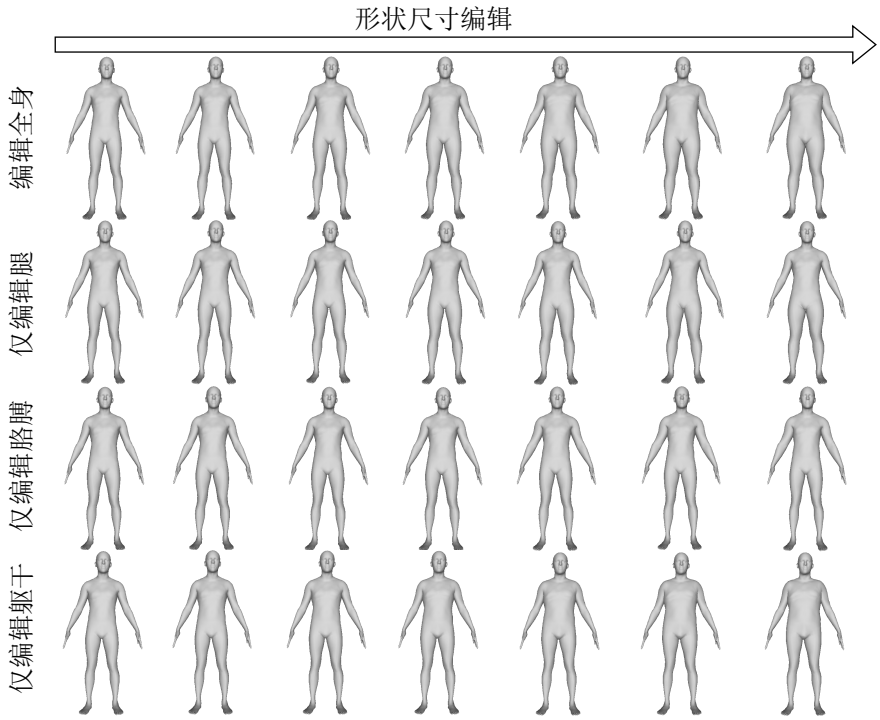


图 4-10　　部位围度编辑示意图

的模长和方向分别代表部件的围度和风格，所以用户可以通过改变形状隐编码的方向来编辑人体形状的风格，对于拥有目标风格（比如性别特征、体脂率）的网格 x^{style}，随后将其送入到形状编码分支 $E_s(\cdot)$，得到拥有目标风格的形状隐编码 Z_s^{style}，然后将其除以向量模长得到其单位向量，再让该单位向量与原形状编码的模长相乘就可以得到最终的拥有目标风格和原有部位围度的目标形状隐编码 Z_s^{new}，接着采用与上节类似的步骤得到形状风格迁移后的人体网格 x^{new}。以上过程的公式化表述如下式所示，其中 $norm(\cdot)$ 为计算向量模长的函数：

$$Z_b = E_b(P), \tag{4-20}$$

$$Z_s = E_s(x), \tag{4-21}$$

$$Z_s^{new} = norm(Z_s) \cdot (Z_s^{style} / norm(Z_s^{style})), \tag{4-22}$$

$$x^{new} = D(Z_b, Z_s^{new}). \tag{4-23}$$

图4-11展示了形状风格迁移的网格变化过程。具体而言，本节展示了三个形状风格迁移的例子，分别为把女性形状风格迁移到男性人体网格上（第一行）、把男性形状风格迁移到女性人体网格上（第二行）、把肥胖形状风格迁移到精瘦人体网格上（第三行），正如图4-11所示，每个例子都逐渐具有了目标形状风格，并且值得注意的是在风格迁移过程中初始网格的其他属性（如骨骼方向和长度、部件围度）并没有发生变化，这也印证了本表示实现了较为彻底的语义解耦。

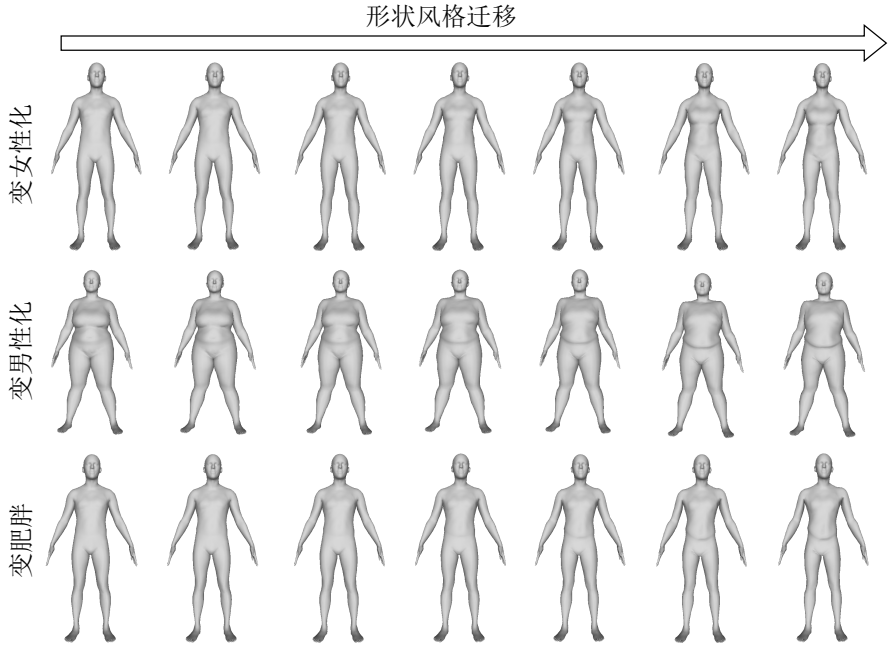


图 4-11　　形状风格迁移示意图

4.3　基于语义精细隐空间的高质量人体生成

人体生成，即生成大量高质量的三维人体网格，以便于为后续人体几何学习提供强大的数据支持。针对该问题一般有两种处理思路，第一种是利用图形学算法在网格几何层面进行处理，比如在两个网格 A 和 B 间进行插值得到一系列几何逐渐由 A 变化为 B 的人体网格，但是这类方法由于缺少人体先验，往往无法保证生成人体的合理性；第二种先将网格压缩降维为一组隐空间内的参数，再在隐空间中对参数进行插值、采样等处理得到人体网格。其中第二类方法可以根据降维方法的不同分为基于机器学习的方法（如 PCA 降维[11,12]）和基于深度学习的方法（如自动编码器、变分自动编码器[8,9,20]、生成对抗网络[24]），可也以根据隐空间语义是否解耦分为语义耦合的方法[10,16,17] 和语义解耦的方法[21–24]。基于深度学习方法凭借神经网络强大的非线性拟合能力往往能保障隐空间的连续性和平滑性，因此生成的人体网格质量远高于基于机器学习的方法，而语义解耦的方法凭借隐空间明确的语义可以显式地控制生成人体网格的属性，极大地方便了后续应用。而本表示方法不仅具有出色的几何刻画能力，以便保证生成三维人体网格的质量，更值得注意的是由于本表示的语义是细粒度的，因此不同以往只能在整个人体层面控制生成人体属性的解耦工作，本表示可以做到在部位级别灵活控制生成人体的属性。

4.3.1 基于隐空间线性插值的人体生成

由于本表示方法将人体网格映射到了具有一定连续性的隐空间，因此可以采用在已有隐编码间插值的方法生成新的人体网格。具体而言，对于输入网格 x_1 与 x_2，先将其送入编码器 $E(\cdot)$ 得到其骨骼隐编码 (Z_b^1, Z_b^2) 和形状隐编码 (Z_s^1, Z_s^2)，然后在两个网格的隐编码间线性插值得到新的骨骼隐编码 Z_b^{new} 和形状隐编码 Z_s^{new}，接着对其做解码得到新生成的高质量人体网格。以上过程的公式化表述如下式所示，其中 α 为值域为 $[0,1]$ 的标量：

$$Z_b^1, Z_s^1 = E(x_1), \tag{4-24}$$

$$Z_b^2, Z_s^2 = E(x_2), \tag{4-25}$$

$$Z_b^{new} = \alpha \cdot Z_b^1 + (1-\alpha) \cdot Z_b^2, \tag{4-26}$$

$$Z_s^{new} = \alpha \cdot Z_s^1 + (1-\alpha) \cdot Z_s^2, \tag{4-27}$$

$$x^{new} = D(Z_b^{new}, Z_s^{new}). \tag{4-28}$$

图4-12展示了基于隐空间线性插值的人体生成结果。具体而言，图4-12(a) 的横纵坐标分别代表形状隐编码与骨骼隐编码，其中被虚线标注的是输入网格，其余皆是通过对输入网格全身隐编码进行线性插值生成的人体网格，图4-12(b) 和 (c) 则分别为对上半身和下半身隐编码进行线性插值生成的结果，这充分地证明了本表示方法的灵活性和细粒度语义。显而易见无论是在全身隐空间还是在局部隐空间本表示都能够通过隐编码线性插值生成高质量的人体网格，这对后续的人体几何研究有重大意义。

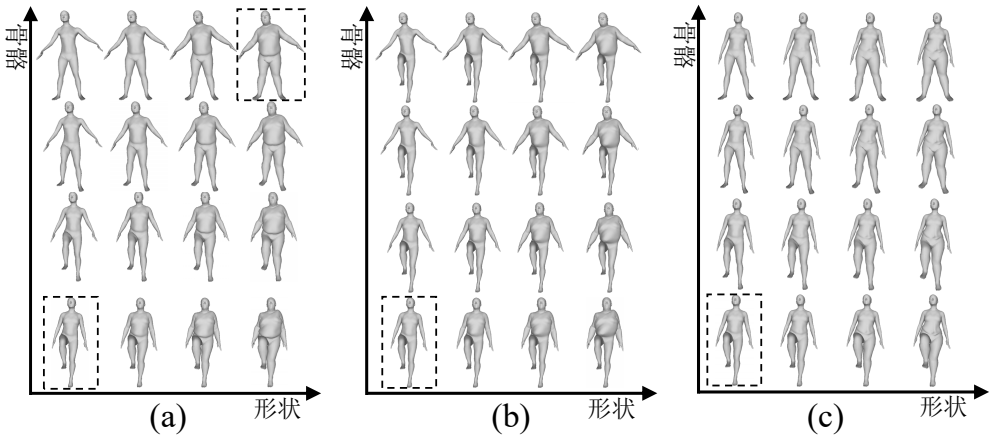


图 4-12　基于隐空间线性插值的人体生成结果：(a) 全身插值，(b) 上半身插值，(c) 下半身插值

4.3.2 基于隐空间随机采样的人体生成

得益于本表示方法学习到的连续性良好的隐空间，可以通过在已有隐编码附近采样生成新的人体网格。具体而言，对于输入网格 x，本节通过利用编码器 $E(\cdot)$ 得到其骨骼隐编码 Z_b 与形状隐编码 Z_s，接着在其形状隐编码周围做符合高斯分布的随机采样得到新的形状隐编码 Z_s^{new}，最后对其做解码得到新生成的人体网格。以上过程的公式化表示如下式所示，其中 α 为从高维高斯分布中随机采样的噪声向量：

$$Z_b, Z_s = E(x), \tag{4-29}$$

$$Z_s^{new} = Z_s + \alpha, \tag{4-30}$$

$$x^{new} = D(Z_b, Z_s^{new}). \tag{4-31}$$

图4-13展示了基于隐空间随机采样的人体生成结果，其中被虚线标注的是输入网格，其余皆是通过在隐空间随机采样生成的人体网格。观察该图可得，得益于表示所学习到的连续性良好的隐空间，生成的人体网格不仅兼具合理性与高质量，而且有些网格甚至表现出了训练集中不存在的形状风格，展现了该方法在人体生成任务上的广阔前景。

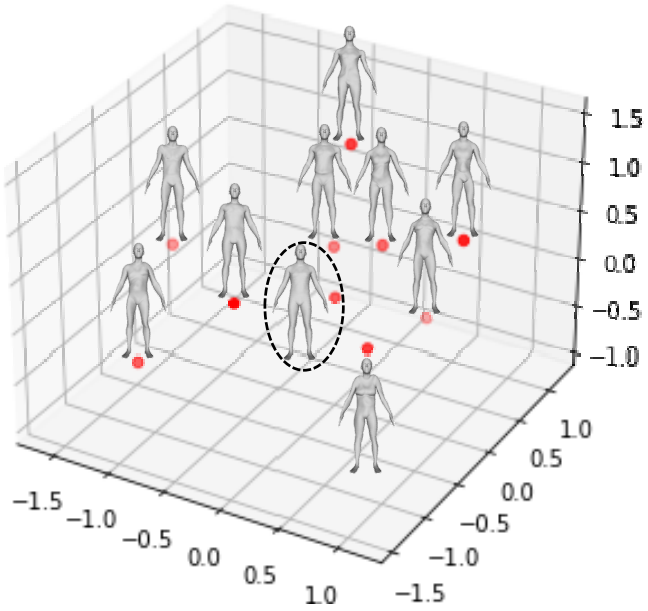


图 4-13 基于隐空间随机采样的人体生成结果

4.4 本章小结

本章以第三章介绍的基于部位解耦的三维人体表示方法为基础，首先简要介绍了该表示潜在的三个应用场景及其意义，并针对不同场景给出了新颖且详尽的实践方案，最终结合大量且完备的实验结果印证了该表示方法的广阔应用前景。具体而言，第 4.1 节针对基于单目人体掩膜的人体重建问题，先在基础框架下证明了本文提出该表示较先前主流方法的优势，并且结合该表示“骨骼分离”、“部位解耦”的特点进一步提出局部与全局注意力引导的人体重建网络，充分验证了方法在该任务上的潜力与可拓展性。第 4.2 节研究人体编辑任务，得益于表示的细粒度语义与几何意义明确的隐空间，本表示方法可以通过调整局部隐编码实现对人体骨骼方向、骨骼长度、形状围度、形状风格的灵活可控编辑，有力地证明该表示方法的灵活性与可编辑性。第 4.3 节聚焦于人体生成问题，在表示所学习到的语言精细隐空间的帮助下，本表示方法可以通过在全局或局部隐空间线性插值或随机采样生成大量高质量的人体网格数据，展现了该表示在人体生成问题上的无限潜力。

第 5 章　总结与展望

本章将对全文内容进行总结与展望，首先阐述了本文对于三维人体表示问题及其应用的技术贡献，随后针对该表示方法的缺陷结合多种相关技术提出可行的改进思路。

5.1　工作总结

本文聚集于三维人体表示问题，即如何利用表示空间内的一组参数把目标三维人体精确、灵活、可控地表示出来，该问题作为数字人体领域的基础问题一直备受研究人员们的关注。然而现有方法不仅语义粗糙，不支持灵活的人体编辑，而且几何刻画能力有限，重建的人体往往会丢失几何细节，并且重建人体的质量严重依赖于配对的监督数据，这极大地限制了其后续应用。针对上述问题，本文提出一种不需要配对监督数据且兼具细粒度语义与高重建精度的三维人体表示方法。

首先，本文先具体描述了三维人体表示问题的定义，随后介绍了传统的基于统计模型的三维人体表示方法和新兴的基于深度学习的三维人体表示方法，并简要分析了这些方法在原理和应用层面的优缺点以及与本文所提出方法的区别与联系。

其次，本文详细介绍了提出的基于部位解耦的三维人体表示方法。首先深入分析了先前工作无法实现研究目标的根本原因，随后详细阐述了本文所提出的骨骼分离的部位解耦策略，并基于此策略提出了骨架引导的自动编码器架构、三分支的训练流程和一系列新颖且有效的训练损失，最后通过详尽的定量与定性实验将该方法与现有工作进行了充分对比，证明了所提出方法的有效性，最后补充了消融实验证明了所提出各模块的必要性。

最后，本文面向多个应用场景，结合所提出方法的特点给出了新颖且有效的实践方案。首先针对基于人体掩膜的人体重建问题，受该方法“骨骼分离”、“部位解耦”特点的启发，提出能更好聚焦各部位图像特征的局部与全局注意力引导的人体重建网络，体现了该方法在人体重建上的应用潜力。其次在人体编辑问题上，得益于表示局部隐编码具有明确的几何意义，该方法可以通过调整局部隐编码实现对于人体骨骼方向等属性的灵活可控编辑，证明了该方法的灵活性和可编辑性。最后借助该表示所学习到的具有良好连续性的语言精细隐空间，该方法可

以通过在全局或局部隐空间上线性插值或随机采样生成大量高质量的人体网格数据，展现了该方法在人体生成问题上的应用前景。

5.2 未来工作展望

虽然本文在三维人体表示问题上取得了一定进展，但是在以下几个方面仍有不足，需要在未来开展进一步工作对其进行完善。

1）首先本文提出的骨骼分离部位解耦策略依赖于对人体部位的几何限制，即部位几何形状可解耦为沿骨骼方向的形变和正交于骨骼方向的形变，对于那些几何过于复杂而不满足该条件的人体部位（如人头、人手、人脚），该方法并不能灵活可控地编辑其形状。

2）另外在编辑骨骼方向时，如果目标骨骼方向与训练数据差别过大，本表示方法会在人体局部区域出现伪影，正如图5-1所示，如何进一步挖掘人体结构先验以增强方法的泛化性仍需要后续工作继续探索。

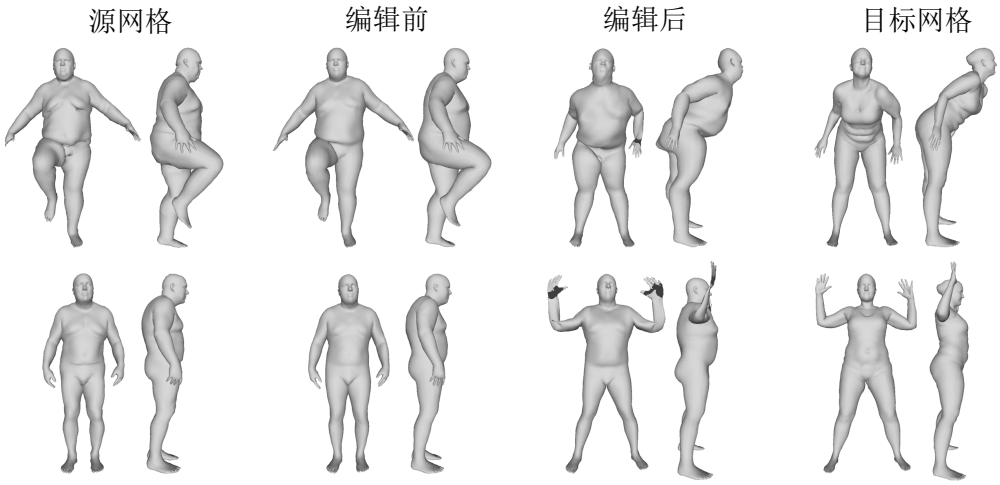


图 5-1　　失败案例示意图

3）最后本表方法属于三维人体的显式表示方法，对几何的刻画受限于网格分辨率，近些年来以符号距离场[63]、占用场[64] 为代表的隐式表示方法因其无限分辨率的特点备受瞩目，并且一些利用隐式场来建模人体的方法[65–70] 也取得了不错的效果，如何以一种合理的方式来结合显式表示与隐式表示也是未来需要探索的一个方向。

参考文献

[1] Tian Y, Zhang H, Liu Y, et al. Recovering 3d human mesh from monocular images: A survey [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[2] Bogo F, Kanazawa A, Lassner C, et al. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image [C]. In European Conference on Computer Vision, 2016: 561–578.

[3] Kanazawa A, Black M J, Jacobs D W, et al. End-to-end Recovery of Human Shape and Pose [C]. In Computer Vision and Pattern Recognition, 2018: 7122–7131.

[4] Zhang H, Tian Y, Zhou X, et al. PyMAF: 3D Human Pose and Shape Regression with Pyramidal Mesh Alignment Feedback Loop [C]. In International Conference on Computer Vision, 2021: 11446–11456.

[5] Zhou S, Fu H, Liu L, et al. Parametric reshaping of human bodies in images [J]. ACM Transactions on Graphics, 2010, 29 (4): 1–10.

[6] Yang Y, Yu Y, Zhou Y, et al. Semantic parametric reshaping of human body models [C]. In International Conference on 3D Vision, 2014: 41–48.

[7] Zeng Y, Fu J, Chao H. 3D human body reshaping with anthropometric modeling [C]. In International Conference on Internet Multimedia Computing and Service, 2017: 96–107.

[8] Tan Q, Gao L, Lai Y-K, et al. Mesh-based autoencoders for localized deformation component analysis [C]. In AAAI, 2018.

[9] Tan Q, Gao L, Lai Y-K, et al. Variational autoencoders for deforming 3D mesh models [C]. In Computer Vision and Pattern Recognition, 2018: 5841–5850.

[10] Bouritsas G, Bokhnyak S, Ploumpis S, et al. Neural 3D morphable models: Spiral convolutional networks for 3d shape representation learning and generation [C]. In International Conference on Computer Vision, 2019: 7213–7222.

[11] Seo H, Magnenat-Thalmann N. An automatic modeling of human bodies from sizing parameters [C]. In Proceedings of the Symposium on Interactive 3D Graphics, 2003: 19–26.

[12] Pishchulin L, Wuhrer S, Helten T, et al. Building statistical shape spaces for 3D human modeling [J]. Pattern Recognition, 2017, 67: 276–286.

[13] Allen B, Curless B, Popović Z. The space of human body shapes: reconstruction and parameterization from range scans [J]. ACM Transactions on Graphics, 2003, 22 (3): 587–594.

[14] Loper M, Mahmood N, Romero J, et al. SMPL: A skinned multi-person linear model [J]. ACM Transactions on Graphics, 2015, 34 (6).

[15] Anguelov D, Srinivasan P, Koller D, et al. SCAPE: shape completion and animation of people [J]. ACM Transactions on Graphics, 2005: 408–416.

[16] Ranjan A, Bolkart T, Sanyal S, et al. Generating 3D faces using convolutional mesh autoencoders [C]. In European Conference on Computer Vision, 2018: 704–720.

[17] Gong S, Chen L, Bronstein M, et al. SpiralNet++: A fast and highly efficient mesh convolution operator [C]. In International Conference on Computer Vision Workshops, 2019.

[18] Gao Z, Yan J, Zhai G, et al. Learning local neighboring structure for robust 3D shape representation [C]. In AAAI, 2021: 1397–1405.

[19] Chen Z, Kim T-K. Learning feature aggregation for deep 3D morphable models [C]. In Computer Vision and Pattern Recognition, 2021: 13164–13173.

[20] Aumentado-Armstrong T, Tsogkas S, Jepson A, et al. Geometric disentanglement for generative latent shape models [C]. In International Conference on Computer Vision, 2019: 8181–8190.

[21] Jiang B, Zhang J, Cai J, et al. Disentangled human body embedding based on deep hierarchical neural network [J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26 (8): 2560–2575.

[22] Zhou K, Bhatnagar B L, Pons-Moll G. Unsupervised shape and pose disentanglement for 3D meshes [C]. In European Conference on Computer Vision, 2020: 341–357.

[23] Cosmo L, Norelli A, Halimi O, et al. LIMP: Learning latent shape representations with metric preservation priors [C]. In European Conference on Computer Vision, 2020: 19–35.

[24] Chen H, Tang H, Shi H, et al. Intrinsic-extrinsic preserved GANs for unsupervised 3D pose transfer [C]. In International Conference on Computer Vision, 2021: 8630–8639.

[25] Boscaini D, Masci J, Melzi S, et al. Learning class-specific descriptors for deformable shapes using localized spectral convolutional networks [C]. In Computer Graphic Forum, 2015: 13–23.

[26] Bruna J, Zaremba W, Szlam A, et al. Spectral networks and locally connected networks on graphs [J]. arXiv preprint arXiv:1312.6203, 2013.

[27] Henaff M, Bruna J, LeCun Y. Deep convolutional networks on graph-structured data [J]. arXiv preprint arXiv:1506.05163, 2015.

[28] Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs [J]. Advances in Neural Information Processing Systems, 2017, 30.

[29] Li L, Gan Z, Cheng Y, et al. Relation-aware graph attention network for visual question answering [C]. In International Conference on Computer Vision, 2019: 10313–10322.

[30] Verma N, Boyer E, Verbeek J. Dynamic filters in graph convolutional networks [J]. arXiv preprint arXiv:1706.05206, 2017.

[31] Li R, Wang S, Zhu F, et al. Adaptive graph convolutional neural networks [C]. In AAAI, 2018.

[32] Boscaini D, Masci J, Rodolà E, et al. Learning shape correspondence with anisotropic convolutional neural networks [J]. Advances in Neural Information Processing Systems, 2016, 29.

[33] Masci J, Boscaini D, Bronstein M, et al. Geodesic convolutional neural networks on riemannian manifolds [C]. In International Conference on Computer Vision, 2015: 37–45.

[34] Hanocka R, Hertz A, Fish N, et al. Meshcnn: a network with an edge [J]. ACM Transactions on Graphics, 2019, 38 (4): 1–12.

[35] Liu H-T D, Kim V G, Chaudhuri S, et al. Neural subdivision [J]. arXiv preprint arXiv:2005.01819, 2020.

[36] Gao L, Lai Y-K, Yang J, et al. Sparse data driven mesh deformation [J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 27 (3): 2085–2100.

[37] Wang J, Wen C, Fu Y, et al. Neural pose transfer by spatially adaptive instance normalization [C]. In Computer Vision and Pattern Recognition, 2020: 5831–5839.

[38] Groueix T, Fisher M, Kim V G, et al. 3D-CODED: 3D correspondences by deep deformation [C]. In European Conference on Computer Vision, 2018: 230–246.

[39] Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library [J]. Advances in Neural Information Processing Systems, 2019, 32.

[40] Kingma D P, Ba J. Adam: A method for stochastic optimization [J]. arXiv preprint arXiv:1412.6980, 2014.

[41] Bogo F, Romero J, Pons-Moll G, et al. Dynamic FAUST: Registering human bodies in motion [C]. In Computer Vision and Pattern Recognition, 2017: 6233–6242.

[42] Robinette K M, Daanen H, Paquet E. The CAESAR project: a 3-D surface anthropometry survey [C]. In International Conference on 3-D Digital Imaging and Modeling, 1999: 380–386.

[43] Schonberger J L, Frahm J-M. Structure-from-motion revisited [C]. In Computer Vision and Pattern Recognition, 2016: 4104–4113.

[44] Schönberger J L, Zheng E, Frahm J-M, et al. Pixelwise view selection for unstructured multi-view stereo [C]. In European Conference on Computer Vision, 2016: 501–518.

[45] Collet A, Chuang M, Sweeney P, et al. High-quality streamable free-viewpoint video [J]. ACM Transactions on Graphics, 2015, 34 (4): 1–13.

[46] Newcombe R A, Fox D, Seitz S M. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time [C]. In Computer Vision and Pattern Recognition, 2015: 343–352.

[47] Yu T, Guo K, Xu F, et al. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera [C]. In International Conference on Computer Vision, 2017: 910–919.

[48] Yu T, Zheng Z, Guo K, et al. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor [C]. In Computer Vision and Pattern Recognition, 2018: 7287–7296.

[49] 宋诗超，禹素萍，许武军. 基于 Kinect 的三维人体扫描、重建及测量技术的研究 [J]. 天津工业大学学报, 2012, 31 (34-37+41).

[50] 周瑾，潘建江，童晶，刘利刚，潘志庚. 使用 Kinect 快速重建三维人体 [J]. 计算机辅助设计与图形学学报, 2013, 25 (873-879).

[51] Kocabas M, Athanasiou N, Black M J. Vibe: Video inference for human body pose and shape estimation [C]. In Computer Vision and Pattern Recognition, 2020: 5253–5263.

[52] Kolotouros N, Pavlakos G, Black M J, et al. Learning to reconstruct 3D human pose and shape via model-fitting in the loop [C]. In International Conference on Computer Vision, 2019: 2252–2261.

[53] Kolotouros N, Pavlakos G, Jayaraman D, et al. Probabilistic modeling for human mesh recovery [C]. In International Conference on Computer Vision, 2021: 11605–11614.

[54] Saito S, Huang Z, Natsume R, et al. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization [C]. In International Conference on Computer Vision, 2019: 2304–2314.

[55] Saito S, Simon T, Saragih J, et al. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization [C]. In Computer Vision and Pattern Recognition, 2020: 84–93.

[56] Xiu Y, Yang J, Tzionas D, et al. ICON: implicit clothed humans obtained from normals [C]. In Computer Vision and Pattern Recognition, 2022: 13286–13296.

[57] Xiu Y, Yang J, Cao X, et al. ECON: Explicit Clothed humans Obtained from Normals [C]. In Computer Vision and Pattern Recognition, 2023.

[58] 刘峰，周弈帆. 基于参数模型和法线推理的单视图三维人体隐式重建 [J]. 南京邮电大学学报 (自然科学版), 2023, 43 (1-10).

[59] Bogo F, Romero J, Loper M, et al. FAUST: Dataset and evaluation for 3D mesh registration [C]. In Computer Vision and Pattern Recognition, 2014: 3794–3801.

[60] Li X, Huang J, Zhang J, et al. Learning to Infer Inner-Body under Clothing from Monocular Video [J]. IEEE Transactions on Visualization and Computer Graphics, 2022.

[61] 周忠，周颐，肖江剑. 虚拟现实增强技术综述 [J]. 中国科学: 信息科学, 2015, 45 (157-180).

[62] Sorkine O. Laplacian mesh processing [J]. Eurographics (State of the Art Reports), 2005, 4 (4).

[63] Park J J, Florence P, Straub J, et al. Deepsdf: Learning continuous signed distance functions for shape representation [C]. In Computer Vision and Pattern Recognition, 2019: 165–174.

[64] Mescheder L, Oechsle M, Niemeyer M, et al. Occupancy networks: Learning 3d reconstruction in function space [C]. In Computer Vision and Pattern Recognition, 2019: 4460–4470.

[65] Deng B, Lewis J P, Jeruzalski T, et al. Nasa neural articulated shape approximation [C]. In European Conference on Computer Vision, 2020: 612–628.

[66] Mihajlovic M, Zhang Y, Black M J, et al. LEAP: Learning articulated occupancy of people [C]. In Computer Vision and Pattern Recognition, 2021: 10461–10471.

[67] Lombardi S, Yang B, Fan T, et al. Latenthuman: Shape-and-pose disentangled latent representation for human bodies [C]. In International Conference on 3D Vision, 2021: 278–288.

[68] Mihajlovic M, Saito S, Bansal A, et al. COAP: Compositional articulated occupancy of people [C]. In Computer Vision and Pattern Recognition, 2022: 13201–13210.

[69] Chen X, Zheng Y, Black M J, et al. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes [C]. In International Conference on Computer Vision, 2021: 11594–11604.

[70] Saito S, Yang J, Ma Q, et al. SCANimate: Weakly supervised learning of skinned clothed avatar networks [C]. In Computer Vision and Pattern Recognition, 2021: 2886–2897.

发表论文和参加科研情况说明

（一）发表的学术论文

[1] **Xiaokun Sun**, Qiao Feng, Xiongzheng Li, Jinsong Zhang, Yu-Kun Lai, Jingyu Yang, Kun Li. Learning semantic-aware disentangled representation for flexible 3D human body editing[C]. In Computer Vision and Pattern Recognition. 2023: 16985-16994.

[2] Xiongzheng Li, Jing Huang, Jinsong Zhang, **Xiaokun Sun**, Haibiao Xuan, Yu-Kun Lai, Yingdi Xie, Jingyu Yang, Kun Li. Learning to infer inner-body under clothing from monocular video[J]. IEEE Transactions on Visualization and Computer Graphics, 2022.

[3] 李坤，李万鹏，**孙晓琨**，方璐．大场景多对象的深度社交分组网络 [J]．中国科学：信息科学，2021，051(8)：1287-1301.

（二）参与的科研项目

[1] 国家自然科学基金面上项目 (62171317)，偏振光场采集及其动态三维重建应用，2022.01-2025.12。

[2] 日本 VRC 公司横向项目，2021GKF-0336，人体 3D Inner Body 推测和重建技术，2021.06-2022.05。

致　　谢

行文至此，代表我的硕士生涯即将结束，按理来说我应该说些光阴似箭岁月如梭之类的话，可是每当我闭上眼北洋园的日与夜又时常闪烁起来，我就明白这两年半的真切感受过的日子既不是“箭”也不是“梭”，而是成为了我鲜活的一部分，会在未来陪着我继续走下去。

人要常怀感恩之心是我在硕士期间最为重要的“科研成果”。首先我要感谢我的爸爸妈妈、姥姥姥爷，年纪越大就越是感激他们四人对我的养育之恩，感谢他们用自己的言行教会我如何做一个正直、真诚、温柔、善良的人，被他们四人共同用心呵护的童年是我一辈子取之不尽用之不竭的精神财富。其次是我的导师李坤老师，感谢李老师在生活和科研上对我的指导和帮助，带我走入三维视觉的世界，并提供雄厚的资源与宽广的平台让我得以尽情挖掘我的兴趣与能力；另外还要感谢来煜坤老师在科研上对我的指导，感谢老校区的杨敬钰老师和岳焕景老师当初给我保研面试机会，这才让我有机会进入智能成像与重建课题组。还有感谢我可爱的朋友们，读研的过程并非是一帆风顺的，幸好有你们才让那一个个平常的日子鲜活起来，你们都是我毕业论文的“第二作者”！最后我想感谢我自己，我平时话很多，脑子里想的更多，活跃的思维导致我经常陷入对未来的焦虑中，好处是在焦虑的压迫下我常常比别人更努力，也就有更大的机会获得世俗意义上的成果；坏处是我时常迷茫，目标的达成只能给我带来片刻的满足，等到新的目标出现我又跳入下一个痛苦的循环……我想感谢自己并非因为获得的那些世俗成就，而是因为自己在迷茫痛苦的日子从来没有放弃过思考，对于科研的思考也好，对于人生的思考也罢，虽然这些问题大多我都没有找到答案，甚至有些问题本就不存在答案，可我仍然乐此不疲，正是这些思考让我变得更加完整。

最后的最后挥手道个别吧，和你也和我。人海茫茫，能被同一阵风拥抱几百个日夜已是莫大的缘分，希望未来再见到大家时仍然能如今天这般推心置腹地交流，也希望再见到自己时能问心无愧地说：“那些问题的答案我还在找，虽然不一定能找到，但是我还在找。”