Expressive Body Capture: 3D Hands, Face, and Body from a Single Image [19]

邝文婷

2024 年 1 月 10 日

摘要

将人体姿势、手部姿势和面部表情表达纳入 3D 人体模型的构建，对于分析人类行为、互动和情感具有重要的研究意义。尽管现有的建模工作已经有了丰富的人体姿势变化表现，但在人体模型中对于面部和手部姿势的建模方面仍然存在缺失，以及在身体各部位衔接的效果方面也有待进一步提升。为了克服这些问题，研究人员提出了一种根据单目图像数据自动捕捉富有表现力的人体姿势、手势和面部表情变化的方法以及 3D 模型。研究者使用数千个三维扫描数据来训练一个新的统一的人体三维模型 SMPL-X，该模型扩展了现有的 SMPL 模型，使之包含了可调节的手部和表情模型。为了提高模型拟合效果，还提出了一种基于 SMPLify 的改进方法 SMPLify-X，它在单目图像中检测与面部、手部和脚部对应的 2D 特征，并将完整的 SMPL-X 模型拟合到这些特征上。实验结果显示，该方法以及 3D 模型在受控环境和真实环境中的测试中表现出了有效性，这对于从单目图像中深入研究人类行为和情感具有显著的推动作用。

关键词： SMPLify-X；SMPL-X；单目图；自动捕捉

1 引言

在现实世界中，人们的行为表达、姿势互动以及情感交流是基于三维空间进行的，而不仅仅限于二维平面。在计算机建模领域中，当我们尝试从一张 2D 单目图像中深入理解或反映人类行为的复杂性，效果通常是不理想的。这是因为 2D 单目图像受平面的限制只能提供关于人体关节的投影信息，难以捕捉到物体的立体形态和细节改变，如身体姿态的扭曲、手势的旋转以及微小的表情变化。图 1就阐明了这个问题，其中从左到右的图像依次表示为 2D 单目图、主要关节点、骨骼分布架构、SMPL 模型以及 SMPL-X 模型。可以看出，具有身体、手部、面部综合表达的 3D 模型，即 SMPL-X 模型，能够更全面和准确地表示人类活动，并具有更丰富的表现力。

然而，目前要从 2D 单目图像中提取准确的完整 3D 人体模型信息仍然具有挑战性。主要原因是缺乏合适的 3D 模型和充足的 3D 模型训练数据。传统的 3D 模型专注于身体整体形状和姿势的捕捉，往往忽略了对人体的面部和手部等局部特征丰富性的表达 [1], [2], [3], [7]。尽管现有工作逐渐开始了对身体以及其他部位一起建模，但粗糙的缝合技术常常导致最终的模型不真实 [13], [20]。与此同时，大量的 3D 训练数据在基于深度神经网络的学习上容易出现过

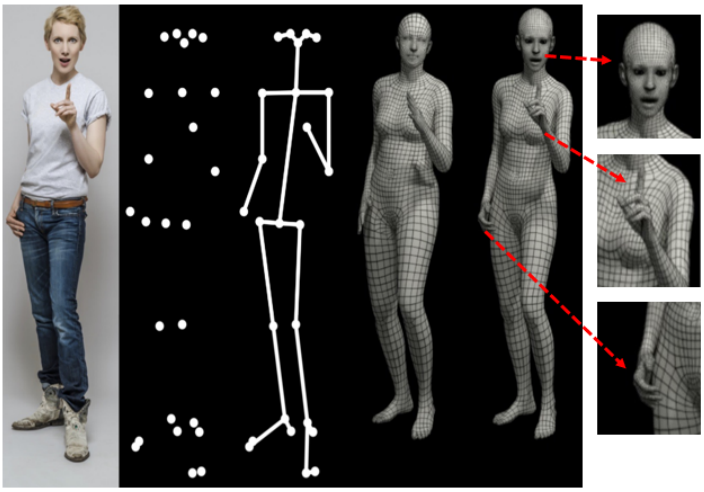

图 1. 特征比较示意图

度拟合，从而降低模型的泛化能力。为了解决这些问题，研究者们将身体模型、头部模型和手部模型相结合，捕获三种模型之间的自然相关性，提出了一种全新的全面人体模型 SMPL-X。此外，他们还提出了一种三维空间到二维平面的统计学拟合方法，使得人体模型的 3D 信息可以与单目图像的 2D 特征完整重合。这一进展有助于为克服 2D 图像对人体三维信息的构建和推断的局限性提供了新的可能，并且为深入理解和表达人类行为、姿势及情感交流的复杂性打开了新的突破口。

2 相关工作

2.1 3D 模型构建

现有的人体模型大多偏重于对身体姿势的构建，而忽略了手部和面部的动作和表情。通常在构建中他们默认手部动作是张开或者握拳的姿势，而面部表情是保持中立的 [6], [9], [8], [17]，这对学习人体情感以及互动信息是不利的。对于现有的统一人体模型构建，其具有一定表达意义的模型是 Frank 模型 [20] 和 SMPL+H 模型 [13]。Frank 模型分别将身体模型 SMPL 模型 [17]、艺术家创建的手部模型 [21] 以及面部模型 FaceWarehouse [5] 缝合在一起，但是产生的模型因重组时缺乏相关性影响的考虑并不完全真实。SMPL+H 模型是将身体模型 SMPL 和通过 3D 扫描学习的 3D 手部模型相结合，其中与 Frank 模型不同的是，SMPL+H 模型的手部模型中的形状变化来自于全身扫描数据，这是因为身体和手部的形状是高度相关的，而手部的姿态变化，则是从一个手部扫描数据集中学习得到，但是 SMPL+H 模型缺少可变形的面部模型表达。

2.2 3D 模型拟合

2D 图像只能提供物体在投影平面上的信息，无法提供物体在深度方向上的信息。而要准确地表示 3D 结构，需要考虑视角、距离、旋转和形状等多个维度的信息。因此，利用配对 3D 参数来反映 2D 和 3D 信息之间的映射关系变得非常重要。其中，SMPLify 方法 [4] 采用了一种“自下而上”的方式来检测 2D 图像特征，然后再在一个优化框架中将 SMPL 模型从

图 2. 综合表达效果图

“自上而下” 地拟合到这些特征上。HMR [14] 通过利用 2D 关键点和 3D 结构的对抗模型来训练无配对的数据模型。NBF [18] 使用分割身体部位，并选择 2D 特征的之间值表示来推断 3D 姿态。这些方法可以实现良好的 3D 模型到 2D 图像的拟合。此外，利用多摄像头设置 [12] 获取多眼图像，并将模型拟合到 3D 关键点和 3D 点云，来捕捉丰富的且具有表现力的人类互动，但这也面临着更多的摄像头采集和更复杂的数据处理的挑战。

3 本文方法

3.1 本文方法概述

本文提出了一种全新的身体模型 SMPL-X，该模型在身体模型 SMPL 的基础上结合了头部模型 FLAME 和手部模型 MANO，从而学习到一种综合的且具有丰富表现力的 3D 人体模型。与传统的简单组合模型不同，SMPL-X 的结合过程并未采用像 Frank 模型那样的简单缝合技术，而是通过使用 3D 扫描仪学习来捕捉身体、面部和手部形状之间的自然相关性。

在将 SMPL-X 模型拟合到单目图像上时，本文还提出了一种改进的方法，即基于 SMPLify 的改良版本 SMPLify-X。它通过 OpenPose 获取人体节点和骨骼结构表示，并对此采用了自下而上的 2D 图像特征估计，随后以自上而下的方式将 SMPL-X 模型拟合到这些 2D 特征上，从而实现了从单目图像中捕捉身体、手部和面部的表达。此外，在捕捉过程中，本文还定义了新的姿势先验项、自身渗透惩罚项以及性别检测器，以提高拟合的准确性。

通过开发这种新的身体模型和拟合方法，本文展示了一种更准确和更丰富表达的途径，使得我们可以从单目 RGB 图像中捕捉人类的身体、手部和面部的信息。此外，为了评估模型的准确性，还构建了一个包含全身 RGB 图像和相应的 3D 真实身体数据的评估数据集，以进行定量评估。结果表明，该模型和方法明显优于其他类似模型，能够产生更加自然的表达效果，其 3D 模型和拟合方法的综合表达效果如图 2所示。

3.2 SMPL-X：综合性人体三维模型

为了构建一个新的综合性 SMPL-X 模型，我们需要联合训练面部、手部和身体的形状参数。在此过程中，我们采用了基于标准顶点的线性混合蒙皮函数来定义 SMPL-X 模型。混合蒙皮技术常用于人体行为的估计中，它将人体的骨骼结构和表面顶点相连，使得骨骼变化能够影响到表面形状变化。当人体骨骼结构发生变化时，这些权重会相应地进行插值和调整，使得蒙皮表面呈现出流畅的形变和自然的外观。公式(1)通过蒙皮函数获得人的运动动作矩阵。

$$M\left(\beta,\theta,\psi\right)=W\left(T_{p}\left(\beta,\theta,\psi\right),J\left(\beta\right),\theta,\mathcal{W}\right) \tag{1}$$

其中，β 表示身体、面部以及手部的形状参数；θ 表示为身体关节、下颌关节以及手部关节参数；ψ 则表示面部表情参数，这些参数与人体模型的顶点位置相关，分别用函数 $T_{p}\left(\beta,\theta,\psi\right)$ 和 $J\left(\beta\right)$ 表示。$\mathcal{W}$ 表示为混合权重，用于平滑人体模型的运动轨迹。基于标准顶点的线性混合蒙皮函数的计算，为模型的视觉表现提供丰富的可能性，使得模型在形状、姿势和表情变化时具有高度的逼真度和细节还原性。

此外，基于标准顶点的线性混合蒙皮函数中还引入了混合形状的校正学习来进一步提升模型的表现力。其表达式为(2)所示：

$$T_{P}\left(\beta,\theta,\psi\right)=\bar{T}+B_{S}\left(\beta;\mathcal{S}\right)+B_{E}\left(\psi;\mathcal{E}\right)+B_{P}\left(\theta;\mathcal{P}\right) \tag{2}$$

校正混合形状函数 $B_{S}\left(\beta;\mathcal{S}\right)$，$B_{E}\left(\psi;\mathcal{E}\right)$，$B_{E}\left(\psi;\mathcal{E}\right)$ 可以根据具体的身体形状、姿势以及表情参数进行进一步的微调，以更好地适应真实的人体变化。$\bar{T}$ 是 SMPL-X 模型的主成分，用于建模人体形状的变化。这种综合的方法充分考虑了人体各部分之间的相关性，使得生成的人体模型更加真实和细致。

3.3 SMPLify-X：三维模型到单目图的拟合

为了使 SMPL-X 模型更好地拟合到单目图像，我们遵循了 SMPLify 方法，并在其基础上进行了改进，称为 SMPLify-X。其主要思路是将三维模型与单目图像的拟合转化为优化问题，通过优化目标函数来寻找最佳参数表达模型，其表达式为(3)所示：

$$\begin{aligned}E(\beta,\theta,\psi)=\ & E_{J}+\lambda_{\theta_{b}}E_{\theta_{b}}+\lambda_{\theta_{f}}E_{\theta_{f}}+\lambda_{\theta_{h}}E_{\theta_{h}}+\\ & \lambda_{\theta_{\alpha}}E_{\theta_{\alpha}}+\lambda_{\beta}E_{\beta}+\lambda_{\mathcal{E}}E_{\mathcal{E}}+\lambda_{\mathcal{C}}E_{\mathcal{C}}\end{aligned} \tag{3}$$

在这里，E_{J} 是测量预测的 2D 关节点和真是关节点之间的重投影误差,$\lambda_{\theta}E_{\theta}$ 表示不同身体部位的姿势先验项,$\lambda_{\beta}E_{\beta}$ 表示形状先验项，$\lambda_{\mathcal{E}}E_{\mathcal{E}}$ 为表情先验项，$\lambda_{\mathcal{C}}E_{\mathcal{C}}$ 为互穿渗透惩罚项，用于身体部分极端弯曲部位如肘部、膝部的条件先验。

由上述可知，我们可以目标函数的优化实则是求损失函数框架最小化。其中，目标函数是一个包含多个项的组合，每个项都有一个对应的权重 λ，在最小化目标函数的基础上，通过各个权重的合理调整，我们可以找到一组最佳的姿势、形状和表情参数，使得 SMPL-X 模型能够更好地拟合单目图像。

3.4 三维模型性别分类器

我们提出一种基于深度学习的模型性别分类器，该分类器可以用于根据单目图拟合相应性别的人物形态三维模型。首先，在一个包含多个数据集的图像数据集中标注了人物的性别，其中包括 LSP 数据集 [10]、LSP-extended 数据集 [11]，MS-COCO 数据集 [16] 以及 LIP 数据集 [15]，并使用这些标注数据对一个预训练的 ResNet18 模型进行了微调，以训练一个性别分类器。当给定一个带有 OpenPose 关节点的全身图像时，性别分类器可以自动识别人的性别，并对应使用相应的人物形态模型进行拟合。为了进一步提高模型的性能，研究者使用了一个类平衡的验证集来确定阈值，使得分类器的预测结果既能够准确地被接受，又不会过于保守。具体而言，在测试时，当分类器预测的概率小于所设定的阈值时，系统会使用一个性别中立的人体形态模型进行拟合；而当分类器预测的概率高于阈值时，则会使用与该性别相应的人体形态模型进行拟合。通过使用性别分类器，可以根据人的性别来选择适合的人体形态模型，可以进一步提高 3D 人体姿势估计的准确性和可靠性。

4 复现细节

4.1 与已有开源代码对比

此工作已对外公开源代码。(开源代码链接：https://github.com/vchoutas/smplify-x)

本次工作不仅完成了论文的复现，还采用了先前的 SMPL 模型、SMPL+H 模型工作与本文的 SMPL-X 模型和 SMPL-op 模型工作进行了比较。对于改进部分，根据现有的 SMPL-X 模型观察，在损失函数优化中对身体姿势以及部分关节点的先验项权重作出调整。由于库的更新，原有 vposer 版本出现了不兼容的问题。为了解决这个问题，我参考了 CSDN 网站上提供的解决方案(参考链接：https://blog.csdn.net/weixin_44034102/article/details/123523443)，对 vposer 中的变分自动编码器的统计方式进行了修改。

4.2 实验环境搭建

本次复现工作基于 Linux 操作系统，使用 Python 版本和 PyTorch 深度学习框架。在 CPU 硬件平台上，安装了图像处理库、数组处理库、计算几何库以及人体建模库。其实验环境型号以及版本如下列所示：

- 操作系统：Linux Ubuntu 20.04
- 硬件平台：CPU x86_64
- Python 版本：Python 3.6
- 深度学习框架：PyTorch 1.0.1.post2
- 图像处理库：OpenCV-Python 4.3.0.38，Pillow
- 数组处理库：NumPy >= 1.16.2
- 计算几何库：TorchGeometry >= 0.1.2
- 人体建模库：SMPL-X，SMPL，SMPL+H，SMPL-op

4.3 使用说明与界面分析

4.3.1 使用说明

完成实验环境配合后，本项工作需要实现三维人体模型到二维的单目图像的拟合，其中二维单目图像数据选择 EHF 数据集，放置于 Data 文件中。当我们使用不同的三维人体模型进行拟合时，需要替换 yaml 配置文件以及模型文件。具体而言，对于 SMPL、SMPL+H、SMPL-X 以及 SMPL-op 模型分别更改为 fit_smpl.yaml 和 smpl、fit_smplh.yaml 和 smplh、fit_smplx.yaml 和 smplx、op_smplx.yaml 和 smplx 模型文件表示。此外，我们还需要其输出的 3D 模型表达结果输出到我们自定义的文件，以便后续的实验评估工作。以下是对于不同的三维模型拟合图像的执行代码：

- SMPL 模型：python smplifyx/main.py –config cfg_files/fit_smpl.yaml –data_folder data –output_folder output_smpl –visualize=False –model_folder models/smpl –vposer_ckpt vposer
- SMPL+H 模型：python smplifyx/main.py –config cfg_files/fit_smplh.yaml –data_folder data –output_folder output_smplh –visualize=False –model_folder models/smplh –vposer_ckpt vposer
- SMPL-X 模型： python smplifyx/main.py –config cfg_files/fit_smplx.yaml –data_folder data –output_folder output_smplx –visualize=False –model_folder models/smplx –vposer_ckpt vposer –part_segm_fn smplx_parts_segm.pkl
- SMPL-op 模型：python smplifyx/main.py –config cfg_files/op_smplx.yaml –data_folder data –output_folder output_op –visualize=False –model_folder models/smplx –vposer_ckpt vposer –part_segm_fn smplx_parts_segm.pkl

4.3.2 界面分析

为了评估不同模型在人体特征表达上的丰富性，我们对同一张二维单目图像分别使用不同的三维模型进行拟合，并通过分析每个模型输出文件中的网格文件来观察其对人体身体、面部和手部特征的表达情况。

图 3展示了论文中对 SMPL 模型、SMPL+H 模型和 SMPL-X 模型的有效性拟合示意结果。从图中可以观察到，SMPL 模型缺乏手部和面部表情特征，而 SMPL+H 模型则缺乏面部表情特征。然而，SMPL-X 模型同时具备了身体、面部和手部特征的表达能力，是最富有表达效果的三维人体模型。这种差异是由于模型的构建和拟合选择所导致的。在复现结果分析中，我们也会对不同的三维模型在人体特征表达上的丰富性进行比较，同时在三维模型表达界面上应该观察到相同的实验效果。

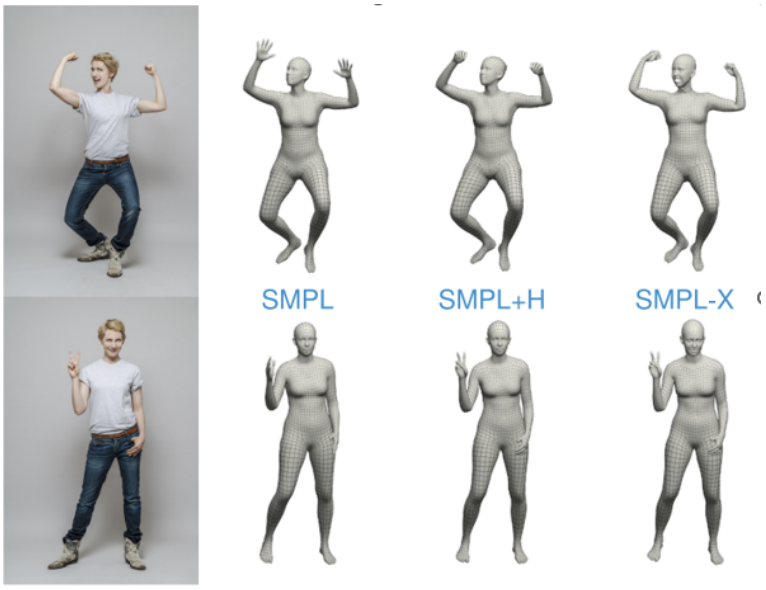

图 3. 模型比较示意图

4.4 创新点

本次复现工作的创新点主要是对 VPoser 中的统计方式进行了修改，解决了库更新导致的不兼容性问题，确保代码的正确运行；其次对损失函数进行优化，调整权重大小以及修改先验项计算，减少三维模型在二维模型拟合过程的所需时间以及误差。最后是构建先前相关工作的 SMPL+H 三维人体模型（原论文没有提供构建方法），并且对不同的三维模型在人体特征表达上进行效果评估。

5 实验结果分析

5.1 可视化实验结果分析

我将四种三维人体模型在 100 张二维单目图像 EHF 数据进行依次拟合，得到了 100 个拟合后的网格文件。接下来，在其中选择两张有效拟合的表达结果，对各个拟合后的三维人体模型进行了比较，如下图 4所示。从图中可以看出，SMPL 模型仅表现出身体变化特征，而 SMPL+H 模型除了身体变化特征之外，还表现了手部变化特征。相比之下，SMPL-X 和 SMPL-op 模型在身体、手部和面部表情特征方面均表现出卓越的效果，这与原论文的工作结果相符。其中，改进后的 SMPL-op 模型在视觉上表达人体特征更为纤细，这与改进过程中对损失函数进行调整有关。

另外，我还对不同性别的二维单目图像使用三维人体模型进行了拟合，并对其性别分类的有效性进行了检验，如图 5所示。结果显示，仅有 SMPL-X 模型和 SMPL-op 模型具备性别分类的能力。这主要是因为这两种模型中都引入了经过多个数据集学习训练好的性别分类器。其中，SMPL-op 模型可以更好地体现到性别特征，如在女性的二维图像上更能表达到其纤细的身材比例，在男性的二维图像上更能表达到其健壮的身材比例，这种优化仍然与改进过程中对损失函数进行修改相关。在复现的过程中，我还发现到，当人体的双手被其他部位遮盖，四种三维模型均难以对手部进行特征表达，如图 6所示。从 OpenPose 结果中观察到，

当双手被遮盖时，它没有对人体手部的关节点以及骨骼结构进行推断，并且选择不表达此特征，因此在三维模型的拟合过程中，就无法对此手部的姿势以及形状进行丰富性表达。

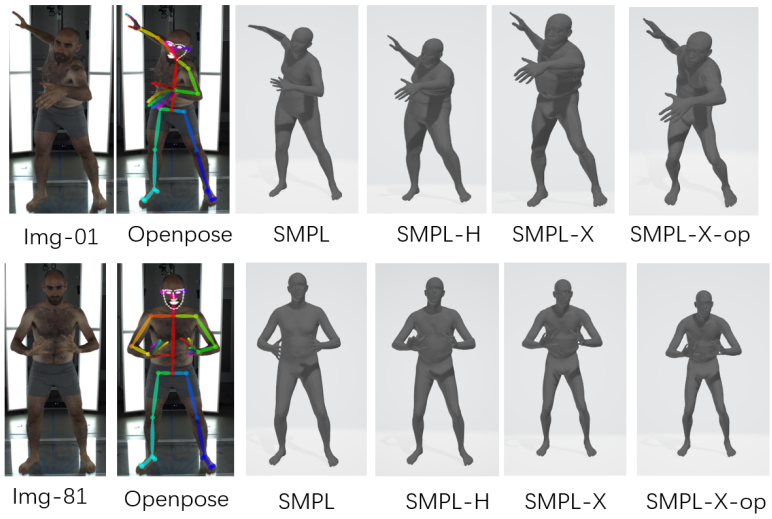

图 4. 基于同一性别复现实验效果显示图

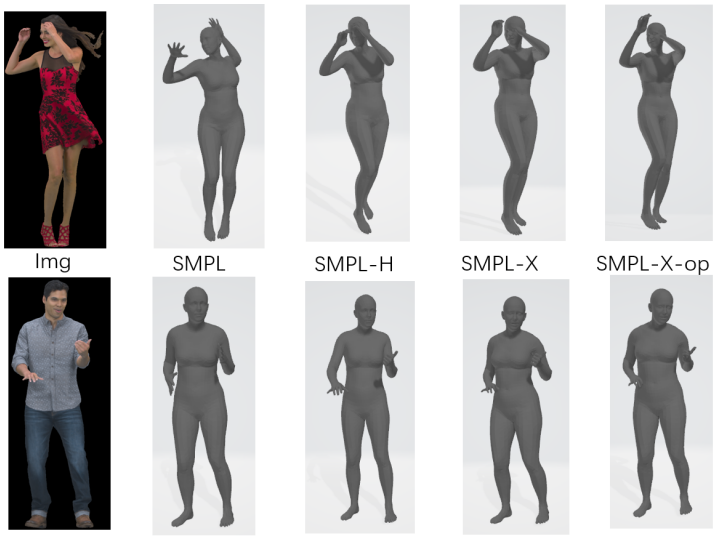

图 5. 基于不同性别复现实验效果显示图

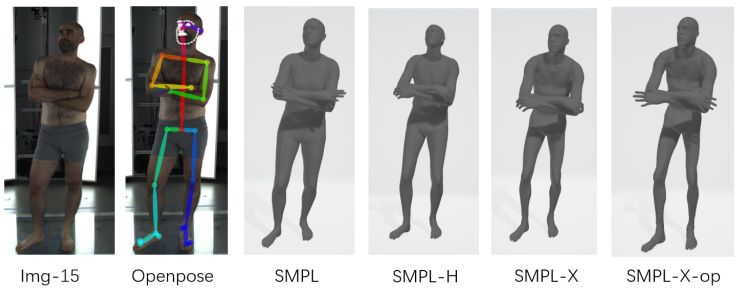

图 6. 基于双手遮盖复现实验效果显示图

5.2 数据化实验结果表达

根据原论文的实验结果，三种模型的拟合效果标准采用了模型顶点与真实网格顶点之间的误差大小进行比较。其中，SMPL-X 模型展现了丰富的人体特征表达，并且实现了最小的误差，表明了 SMPL-X 模型在拟合人体姿态方面具有优势性。

Model	Keypoints	Mean Vertex-to-Vertex error（mm）
SMPL	Body	57.6
SMPL+H	Body+Hands	54.2
SMPL-X	Body+Hands+Face	52.9

表 1. 原文实验效果定量分析表

在复现工作中，我对四种三维人体模型同样采用了模型顶点与真实网格顶点之间的误差大小进行比较，并且计算模型在 100 张二维单目图像中的拟合时间，由表所示。复现实验结果与原论文的实验结果在误差值上存在略微的偏差，这可能与操作系统或者硬件平台有关，但从整体上的结果上看，SMPL-X 模型和 SMPL-op 模型在其余相关模型中仍具备优势性，且改进的 SMPL-op 模型在拟合所需的时间上更少，但在误差效果改进上并不明显。

Model	Keypoints	Mean VtoV error（mm）	Time
SMPL	Body	60.3	00:38:41
SMPL+H	Body+Hands	57.4	02:01:58
SMPL-X	Body+Hands+Face	55.8	03:10:01
SMPL-op	Body+Hands+Face	55.1	02:21:29

表 2. 复现实验效果定量分析表

6 总结与展望

在本次复现工作中，我对将三维人体模型拟合到二维单目图像的过程中的目标损失函数进行了修改。并将改进后的模型与其他相关的三维模型在拟合效果上进行了比较。实验评估表明，改进的 SMPL-op 模型在拟合过程中所需的时间和误差都有所减少。接下来的研究方向是从处理被身体部位遮挡的推理开始。这是因为现有的二维特征检测 Openpose 对被遮挡的身体部位不敏感，无法表达它们的关节位置和骨架结构，导致人体三维模型无法准确拟合。因此，我们从推理被遮挡部位的相关特征进行改进是在三维模型实现丰富性表达是有必要的。

参考文献

[1] Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes: reconstruction and parameterization from range scans. *ACM transactions on graphics (TOG)*, 22(3):587–594, 2003.

[2] Brett Allen, Brian Curless, Zoran Popovic, and Aaron Hertzmann. Learning a Correlated Model of Identity and Pose-Dependent Body Shape Variation for Real-Time Synthesis. In Marie-Paule Cani and James O'Brien, editors, *ACM SIGGRAPH / Eurographics Symposium on Computer Animation*. The Eurographics Association, 2006.

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005.

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016.

[5] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013.

[6] Oren Freifeld and Michael J Black. Lie bodies: A manifold representation of 3d human shape. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 1–14. Springer, 2012.

[7] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and H-P Seidel. A statistical model of human pose and body shape. In *Computer graphics forum*, volume 28, pages 337–346. Wiley Online Library, 2009.

[8] Nils Hasler, Thorsten Thormählen, Bodo Rosenhahn, and Hans-Peter Seidel. Learning skeletons for shape and pose. In *Proceedings of the 2010 ACM SIGGRAPH symposium on Interactive 3D Graphics and Games*, pages 23–30, 2010.

[9] David A Hirshberg, Matthew Loper, Eric Rachlin, and Michael J Black. Coregistration: Simultaneous alignment and modeling of articulated 3d shape. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 242–255. Springer, 2012.

[10] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, volume 2, page 5. Aberystwyth, UK, 2010.

[11] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR 2011*, pages 1465–1472. IEEE, 2011.

[12] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015.

[13] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018.

[14] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018.

[15] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized convolutional neural network. In *Proceedings of the IEEE international conference on computer vision*, pages 1386–1394, 2015.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

[18] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018.

[19] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[20] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

[21] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *Proceedings of the IEEE international conference on computer vision*, pages 2456–2463, 2013.