前沿论文复现

摘要

3D 人体姿态估计在计算机视觉领域得到了显著关注，尤其是在虚拟现实（VR）、增强现实（AR）、运动分析、健康监测、智能交互等多个领域中展现了巨大的应用潜力。目前最好的方法利用了大量的 3D 伪真实值（p-GT）和 2D 关键点数据集，从而实现了强大的性能。然而，使用这些方法时，我们观察到一个矛盾现象：随着 2D 精度的提高，3D 姿态精度反而下降。这是由于 p-GT 中的偏差以及使用了近似的相机投影模型所导致的。TokenHMR 由此提出了一种新的损失函数，“阈值自适应损失缩放”（Threshold-Adaptive Loss Scaling，TALS），惩罚较大的 2D 和 p-GT 误差，但不会对较小的误差进行惩罚，从而减小对 2D 关键点的拟合程度。之后利用了人体姿态的标记化表示，并将问题重新表述为标记预测。但在复现过程中，TokenHMR 的方法在随机图片中的表现仍存在一些问题。本文提出了一些改进方法，但由于时间问题并没有实现完成。

关键词：3D 人体姿态估计；阈值自适应损失缩放；人体姿态标记化表示

1 引言

人体姿态估计旨在从图像或视频中恢复人体的三维空间结构，包括关节位置、人体网格形状等关键信息。在实际应用中，人体姿态估计在诸如虚拟现实（VR）、增强现实（AR）、健康监测、智能交互、体育分析等多个领域都为关键技术。例如在医疗健康领域，姿态估计技术可以用于分析患者的运动模式，为个性化的康复方案提供依据。

目前在 3D 人体姿态估计领域，已有许多工作致力于提高从单张图像中精确回归 3D 姿态的能力。早期的工作如 SMPL 模型（SkinnedMulti-PersonLinear Model）通过建立人体的参数化模型，在形状和姿态建模方面取得了突破性进展 [11]。Humans in 4D 模型 [7] 进一步利用深度学习方法，将 2D 图像特征直接映射为人体网格参数，大大提高了 3D 姿态和形状回归的精度。然而，尽管这些方法在 2D 关键点和 3D 伪真实值（p-GT）监督下表现较好，但研究者们发现，随着 2D 精度的提高，3D 精度反而会出现下降。这一问题在 3DPW 和 Human3.6M 等数据集中尤为突出，一种原因在于 p-GT 数据由于未知的摄像机参数，导致模型在 2D 平面上的投影与实际动作有着一定偏差。

为了有效应对遮挡和人体姿态的多样性，使用视频推理人体动作，以利用时序信息成为了一个不错的选择。Shen 等提出了一种从单目视频恢复世界坐标系下人体运动的新方法，利用由世界重力和相机视角定义的重力视角（Gravity-View, GV）坐标系统，有效减少了图像与姿态映射时的歧义 [13]。在单目视频中，由于缺少时序信息等，解决姿态问题变得更为棘手。

发表在 CVPR2024 上的论文《TokenHMR:AdvancingHumanMeshRecovery with a Tokenized Pose Representation》通过专注于解决相机投影造成的姿态误差，实现了不错的效果。

本报告工作通过复现并测试 TokenHMR 项目效果，来深度理解最新三维人体姿态估计的研究形式，从而推动研究生期间工作进度。在复现过程中学习相关代码 API 和实现方式，并通过测试随机图片来观察项目的最终效果。若存在效果不好的地方，思考并尝试改进。

2 相关工作

2.1 人体姿态于形状回归

从单张图像中估计 3D 人体姿态和形状已经经过了广泛的研究，方法从基于优化的方案到最新的基于 Transformer 的回归器不等。基于优化的方法将参数化模型与 2D 图像线索（如关键点 [12]）进行匹配。一些基于学习的方法 [5] 直接从图像中估计参数化人体模型。最近的方法 [10] 使用 Transformer 估计 3D 人体，达到了当前最好的精度。为了应对泛化的挑战，最近的如 HMR2.0 和 CLIFF [9] 等方法在训练损失中使用了 2D 关键点和 p-GT，从而在投影的人体与图像之间实现良好的对齐。

2.2 姿态先验

人体姿态先验在诸如将 2D 姿态提升为 3D 和从图像/视频中估计人体姿态 [8] 等各种应用中起着至关重要的作用。早期的姿态先验侧重于学习关节限制 [1]，以避免出现不可能的姿态。高斯混合模型（GMMs）[4] 和生成对抗网络（GANs）[6] 也被用于在训练期间引入先验知识。这些方法中的许多偏向于常见的姿态，与现有工作相比，TokenHMR 通过离散的基于标记的先验来学习有效的 SMPL 姿态，减少了姿态偏差，并提高了对遮挡的鲁棒性，同时易于集成到 HPS 训练中。其文章中说明使用了 VQ-VAE [14]，一种 VAE 的变体，通过量化 3D 训练姿态的过程称为“标记化”来学习离散先验。这种方法能够将回归问题重新表述为姿态标记分类问题，从而利用标记化来表示有效姿态，有效地提供了姿态先验。

3 本文方法

3.1 本文方法概述

这篇文章针对从单张图像回归 3D 人体姿态和形状（HPS）的问题进行了研究。对于一个从图片中回归的姿态模型，其应该满足两个目标：一是应该正确回归 3D 姿态，二是要与图像证据相一致。然而，现有的模型无法同时做到这两点。TokenHMR 中讨论了一个看似矛盾的现象，即方法在拟合 2D 关键点越精准，其预测的 3D 姿态反而越不准确。通过研究这个问题，文章发现常见的弱透视相机假设是主要原因。该相机模型与获取图像时的真实相机不匹配，导致投影的 3D 关节点与检测到的 2D 关节点之间存在差异。目前，尚无可靠的方法能够从单张图像中估计相机参数。TokenHMR 通过使用合成数据集 BEDLAM [2]，利用 Human4D 的相机模型将 3D 数据投影到 2D，以量化 3D 动作投影到 2D 平面上的的偏差。并通过引入新的损失函数，实现了利用大量野外数据的丰富信息的同时，避免 3D 精度的下降。

为了控制对于单个 2D 姿态对应的 3D 姿态多解问题，TokenHMR 使用向量量化变分自编码器实现了不错的效果。现有的基于高斯混合模型或变分自编码器（VAE）的姿态先验偏向于训练数据中出现频繁的姿态。为了找到一个无偏的先验，使其仅限制网络输出有效的姿态，而不偏向任何特定的姿态，TokenHMR 通过将连续姿态回归问题转化为姿态标记预测问题，即通过标记化人体姿态来处理。

具体来说，TokenHMR 首先使用向量量化变分自编码器（VQ-VAE），在大规模运动捕捉数据集（AMASS 和 MOYO）上进行预训练，通过一个自编码器，将 SMPL 人体参数（包括每个关节的姿态）编码为离散的姿态标记。这个标记化表示为回归器提供了一个有效姿态的“词汇表”，将姿态先验有效地表示为一个码本 1。在自编码器的优化时，作者采用重建损失、嵌入损失和承诺损失等损失函数。此外，为了避免 VQ-VAE 可能出现的代码本崩塌问题，作者使用了指数移动平均（EMA）和代码本重置策略以提高代码本的利用率。与回归连续姿态不同，VQVAE 通过分类生成离散的标记。文章中作者将一个 HPS 方法中的连续姿态替换为 VQVAE 标记化姿态方法，结果表现出 3D 准确度得到了可观的提升。

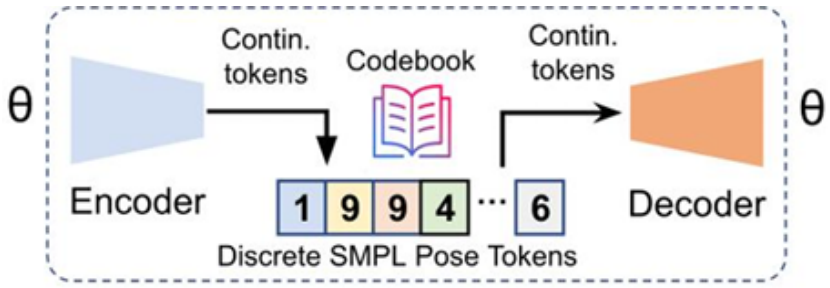

图 1. Tokenization

3.2 相机误差分析

通常，估计 3D 人体姿态与形状的方法旨在实现两个主要目标：一是准确估计 3D 姿态，二是确保其与 2D 图像特征的精确对齐。将一个 3D 物体渲染到相机平面上时，需要计算相机的内参（如焦距）或外参（如旋转和平移）。但将参数化人体模型拟合到图片上时，我们很难获得图片原本的相机参数。如果使用固定且大概率不正确的相机参数，依靠缩放的正射投影或透视投影来估计图像中的人体位置，就会导致真实的 3D 关节点与其 2D 投影之间的不匹配。比如，由于照片通常是从大约眼睛高度拍摄的，人的腿部由于位于身体的下方且远离相机，会呈现出缩短的现象，这使得通过训练模型最小化 2D 误差的方式，迫使模型生成不准确的 3D 姿态，如图 2 所示。唯一能够使人体适应这种现象的方式就是弯曲膝盖或在 3D 空间中倾斜身体，使腿部距离相机更远，这种方式将导致不自然甚至不稳定的姿态，严重影响了结果的可靠性。

目前的 3D 姿态数据集无论是场景还是动作的丰富度都有较大发展空间。而 2D 姿态数据集的场景和动作更加多样。使用 2D 姿态数据集作为伪真实值，会由于使用了错误相机参数引入误差。

这种问题已经成为当前所有方法的根本障碍，如果不考虑相机参数，就无法同时实现低的 3D 和 2D 误差。为了更为清晰地定量评估这种不匹配的影响，文章使用合成数据集 BEDLAM，其包含精确的 3D 数据和 2D 数据，并包含相机的真实信息，从而消除噪声的干扰。如图 3 所

示，作者将 BEDLAM 数据集中的 3D 真值通过 HMR2.0 相机进行投影。结果表明，当使用错误的相机时，所得到的 2D 投影误差会显著增加。

文章采用标准的正确关键点百分比（PCK）作为度量指标，评估了错误相机对 2D 投影误差的影响。通过 HMR2.0b 模型计算出的 3D 身体在 PCK0.5 和 PCK1.0 上的误差分别为 0.78 和 0.88。然而，当将 HMR2.0b 相机应用于真实的 3D 数据时，PCK 得分下降至 0.66（PCK0.5）和 0.86（PCK1.0）。理想情况下，使用正确的相机模型时，PCK0.5 和 PCK1.0 应当都能达到 1.0，这表明 HMR2.0b 模型尽管在 2D 误差上表现较好，但在实际应用中，由于相机参数的偏差，它的 3D 姿态和形状估计明显偏离了真实值。这一结果也进一步证明了，像 HMR2.0b 这样的模型，虽然能够实现较高的 PCK 值，但却是以牺牲 3D 准确度为代价的。因此，文章得出结论：在没有准确相机模型的情况下，单纯追求高 PCK 值对于 3D 准确度的提升是有害的。

3.3 阈值自适应损失缩放：TALS

在上一节中，文章的分析揭示了使用伪真实值和 2D 关键点进行学习的一个显著障碍——相机/姿态偏差。尽管存在这一问题，但 2D 关键点作为能从图片当中高质量推理出的数据，其对实现高的泛化能力和鲁棒性仍然至关重要。文章为了有效利用这些数据，提出了一个想法：通过建立一个有效的阈值，来识别那些不会提供额外收益的训练信号。当损失超过这一阈值时，传统的学习机制会指导姿态估计；相反，当损失低于这一有效阈值时，减少其影响，以防止过拟合相机/姿态偏差。

为了确定这一有效的阈值，文章分析了使用 GroundTruth3D 姿态和标准（不正确）相机模型所得到的误差。对于 2D 关键点，使用统一的相机模型（非正确参数）将预测的 SMPL 参数替换为来自 BEDLAM 的 GT，从而获得真实 3D 人体的 2D 关键点投影。然后计算这些投影与 GT2D 关键点之间的平均 L1 范数，并将其作为 2D 关键点监督的阈值。在建立了 2D 关键点和 SMPL 伪真实值的有效阈值之后，使用阈值自适应损失缩放（TALS）作为损失函数。函数定义为仅在损失低于阈值时缩小损失：

$$\mathcal{L}_{\boldsymbol{\theta}_{pGT}} = \begin{cases} \|\boldsymbol{\theta} - \boldsymbol{\theta}_g\|^2 & \text{if } \mathcal{L}_{\boldsymbol{\theta}_{pGT}} > \varepsilon_{\boldsymbol{\theta}} \\ \alpha_\theta \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}_g\|^2 & \text{otherwise} \end{cases}$$
$$\mathcal{L}_{J_{2D_{pGT}}} = \begin{cases} |\boldsymbol{J_{2D}} - \boldsymbol{J_{2D_g}}| & \text{if } \mathcal{L}_{\boldsymbol{J_{2D_{pGT}}}} > \varepsilon_{J_{2D}} \\ \alpha_{J_{2D}} \cdot |\boldsymbol{J_{2D}} - \boldsymbol{J_{2D_g}}| & \text{otherwise} \end{cases} \tag{1}$$

3.4 模型架构

TokenHMR 首先用一个 VisionTransformer 将输入图像 I 转换为输入潜特征，然后由另一个 Tranformer 处理以生成另一组潜特征，这个 Transformer 具有多头自注意力，并通过交叉注意力将一个零输入标记与图像潜特征融合，以从 ViT 的输出中提取特征。HMR2.0 在此处使用三个线性层将 Transformer 输出的潜特征映射到 SMPL 姿态、形状和相机。但 TokenHMR 使用了新的方法。其将 SMPL 姿态参数划分为身体姿态和全局方向，并使用不同的线性层从标记器中预测全局方向和身体姿态。另一个问题是，因为从代码本中选择嵌入的过程是不可

微分的。TokenHMR 在这里使用了基于 logit 的方式，不直接估计代码索引，而是输出每个标记的 logit 值。这些 logit 值与代码本相乘，得到加权的嵌入。

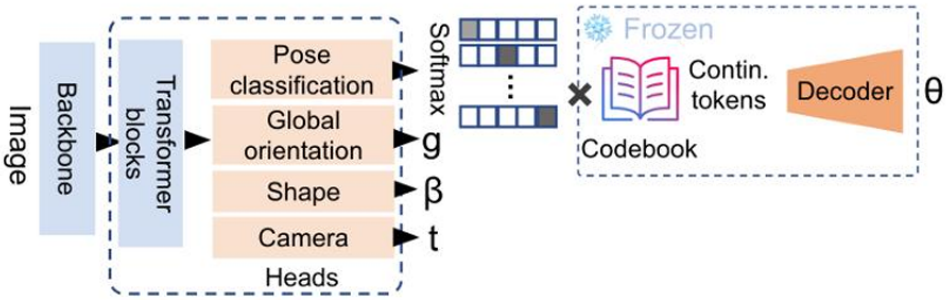

图 2. TokenHMR 架构

3.5 损失函数

TokenHMR 的损失函数使用之前一些工作思路，定义了 2D 和 3D 关节、SMPL 姿态和形状参数的损失。其中对于来自 2D 和 3D 数据集的数据处理方式有所不同。对于 3D 真实数据集，定义了标准损失为：

$$\mathcal{L}_{GT} = \lambda_\theta \mathcal{L}_\theta \left(\boldsymbol{\theta}, \boldsymbol{\theta_g}\right) + \lambda_\beta \mathcal{L}_\beta \left(\boldsymbol{\beta}, \boldsymbol{\beta_g}\right) + \lambda_{3D} \mathcal{L}_{3D} \left(\boldsymbol{J_{3D}}, \boldsymbol{J_{3D_g}}\right) + \lambda_{2D} \mathcal{L}_{2D} \left(\boldsymbol{J_{2D}}, \boldsymbol{J_{2D_g}}\right) \tag{2}$$

其中，L_β 是 SMPL 形状损失，L_{J3D} 是 3D 关节损失，L_{J2D} 是关节重投影损失。λ_β、λ_{3D} 和 λ_{2D} 是每个项的权重系数。

对于学习 SMPL 伪真实数据，使用之前的阈值自适应损失缩放。由此总损失定义为：

$$\mathcal{L}_{Total} = \mathcal{L}_{GT} + \mathcal{L}_{\theta_p GT} + \mathcal{L}_{Total} = \mathcal{L}_{GT} + \mathcal{L}_{\theta_p GT} + \mathcal{L}_{J_2 D_P GT} \tag{3}$$

4 复现细节

4.1 与已有开源代码对比

与原代码相比，本文虽然进行了一定的修改，但由于设备限制（显存不足），目前还没有测试出结果。在代码中，本文工作修改了其网络结构。原论文中的模型在测试一张图片时，出现了明显的肢体错误（图3）。为了解决这个问题，引入具有深度感知能力的模型，并通过特征融合的方式，能够提高原模型处理具有一定透视的动作时的能力（图4为使用 Depth-pro 深度估计的结果 [3]，对人体深度有着足够的理解能力）。

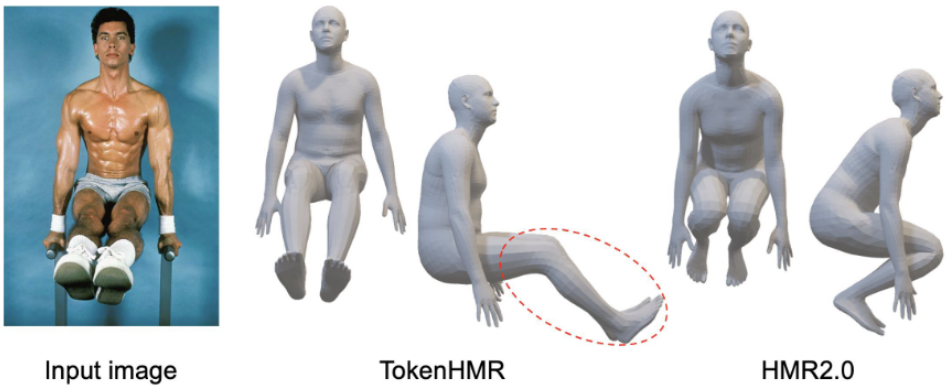

图 3. TokenHMR 结果图

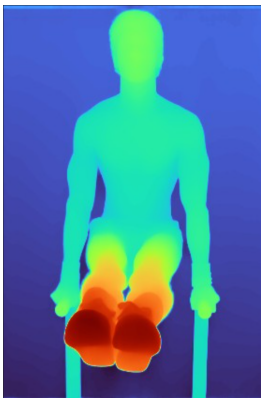

图 4. Depth-pro 深度估计

4.2 实验环境搭建

作者于 https://github.com/saidwivedi/TokenHMR 提供了完整的项目源代码。对于配置环境，将项目克隆到本地，并创建新的 conda 环境（python 版本 3.10）。在安装 pytorch 时，可运行项目主页上的相关指令。然后安装项目根文件夹下的 requirements.txt 即可完成。

在无图形界面的系统中,需要将项目文件夹下的 train.py,utils 文件夹下的 mesh_render.py 以及 renderer.py 中的 os.environ['PYOPENGL_PLATFORM'] 由'egl' 改为'osmesa'。之后还需要额外安装 PyOpenGL 以及 osmesa 库才能运行成功。

4.3 改进思路

目前关于单图的 3D 人体姿态估计存在很多问题。TokenHMR 通过改进 4DHuman，使用大量的 3D 和 2D 姿态数据集，在之前取得了不错的效果。由于在相机外参和内参未知的情况下，3D 姿态在相机平面上的投影会由于透视的作用带来误差。TokenHMR 设置了阈值 2D 关键点损失 Loss，通过降低 2D 关键点在高拟合情况下的权重实现了一定的提升。但其实际结果仍然存在很多问题，因为 2D 姿态关键点是图片中的高可信度证据，降低了 2D 关键点的拟合程度也会损失掉部分精度。

深度估计模型能够在图片上推理出人体区域的深度。融合深度信息能够弥补先前姿态估计工作在深度和姿态复杂情况下的表现。3D 姿态数据集也能够进一步联合优化深度估计模型和姿态估计模型的性能。

2D 姿态数据集拥有更复杂的场景和更多样的动作。为了解决训练 2D 伪真值时所带来的透视误差，模型将首先估计出三个主要数据：人体姿态、相机内参，以及相机与人物的距离。之后对这三个目标进行优化，使其在相机上的投影与 2D 关键点拟合，并满足深度估计结果。先前工作在解决 2D 伪真值训练时的问题时，并没有给出合理的方法。我们的方法在流程中能够针对每一张图片进行调整。

5 实验结果分析

由于暂未解决的设备问题（显存不足），我们还未对改进思路训练出结果，但已完成了代码的实现。这里展示的为原论文模型复现结果。

对于人体估计效果，我们从网络随机取了几张动作图片，使用 TokenHMR 进行估计，结果如图5。

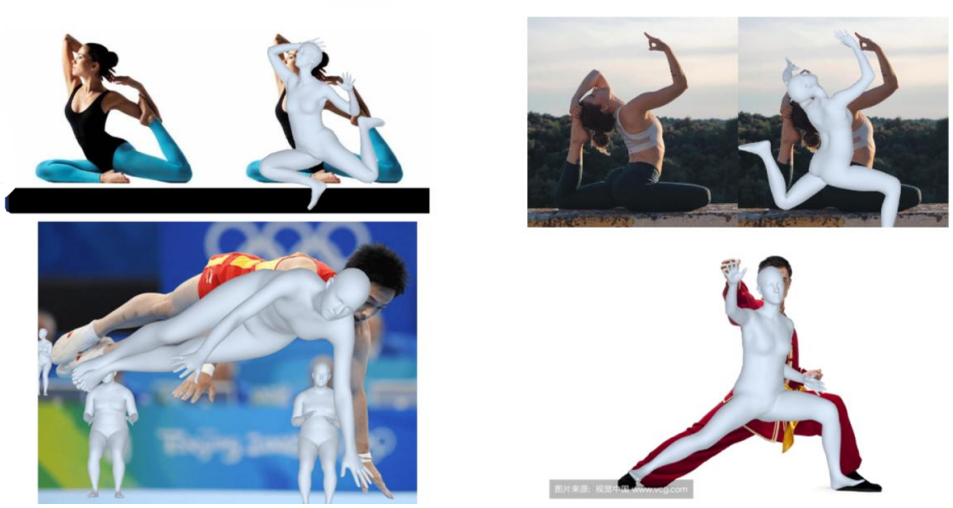

图 5. 模型估计结果

TokenHMR 在推理动作的 3D 准确度上展现出了一定的实力。但在使用论文模型推理一些具有挑战性的动作时，TokenHMR 表现出了不适应。原因在于 TokenHMR 的损失函数降低了 2D 关键点的拟合程度，导致其在动作细节上的结果与原图偏差较大。

对于量化指标，三位姿态估计常使用平均顶点误差（MVE）、每个关节位置的平均误差（MPJPE）和 Procrustes 对齐的每个关节位置平均误差（PA-MPJPE）。原论文模型在 3DPW 测试集上的结果为：

表 1. 3DPW 测试集上误差结果

Method	MVE(mm)	MPJPE(mm)	PA-MPJPE(mm)
HMR2.0	88.4	77.4	47.4
TokenHMR	84.6	71.0	44.3

6 总结与展望

目前还未成功实现创新点，这也是之后工作的内容。对于显存不足的问题，虽然将模型拆分到多卡上能够解决，但由此带来的训练速度损失难以接受（慢了 12 倍）。也许可以先将流程中的动作优化内容进行实现并评估效果，之后再尝试融合深度感知模型。除此之外，目前三维人体姿态的合成数据集图片都与现实图片有较大差别。使用 Diffusion 模型处理合成图片，以对齐现实特征能一定程度上提高合成数据集的性能。这个想法可以作为补充工作，在项目后期进行探索。

参考文献

[1] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015*, June 2015.

[2] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, June 2023.

[3] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun.

[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 Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 561–578, Cham, 2016. Springer International Publishing.

[5] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[6] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *Computer Vision –ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII*, page 768–784, Berlin, Heidelberg, 2020. Springer-Verlag.

[7] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023.

[8] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021.

[9] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information innbsp;full frames intonbsp;human pose andnbsp;shape estimation. In *Computer Vision –ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, page 590–606, Berlin, Heidelberg, 2022. Springer-Verlag.

[10] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. *CVPR*, 2023.

[11] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[12] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[13] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia Conference Proceedings*, 2024.

[14] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6309–6318, Red Hook, NY, USA, 2017. Curran Associates Inc.