

Putting People in their Place: Monocular Regression of 3D People in Depth: 复现

Yu Sun¹ Wu Liu² Qian Bao² Yili Fu¹ Tao Mei² Michael J. Black³ 1Harbin Institute of Technology, Harbin, China 2 Explore Academy of JD.com, Beijing, China 3Max Planck Institute for Intelligent Systems, Tübingen, Germany

摘要

给定一个有几个人的图像，本文的目标是直接回归到所有人的姿势和形状以及他们的相对深度。然而，在不知道一个人的身高的情况下，推断他的深度从根本上来说是模糊的。当场景中包含大小非常不同的人，例如从婴儿到成人时，这尤其成问题。要解决这个问题，首先，本文开发了一种新的方法来推断在单一图像中多个人的姿态和深度，这种方法称为 BEV，增加了一个额外的想象鸟瞰表示来明确地推理深度。BEV 同时解释图像和深度的身体中心，通过梳理这些，估计三维身体位置。与之前的工作不同，BEV 是一种端到端可微的单镜头方法。其次，身高随年龄而变化，如果不估计图像中的人的年龄，就不可能分辨出深度。为此，本文利用了一个 3D 身体模型空间，让 BEV 可以推断出从婴儿到成人的形状。第三，为了训练 BEV，本文创建了一个新的数据集，“相对人类”（RH）数据集，其中包括年龄标签和图像中的人之间的相对深度关系。对 RH 和 AGORA 的大量实验证明了该模型和训练方案的有效性。BEV 在深度推理、子形状估计和对遮挡的鲁棒性方面都优于现有的方法。

关键词：姿态估计；深度估计；单目回归；论文复现

1 引言

“Putting People in their Place: Monocular Regression of 3D People in Depth” 是一篇机器视觉领域的论文，主要研究的是单目图像中人体的三维深度信息的回归（即通过单眼图像预测人体的三维姿态和位置信息）^[1]。如图 1 所示，将一张彩色图片输入到网络中，最终会输出图像中所有人物的三维姿态和深度，其中无论是小孩还是成人，都能很好的检测出来。

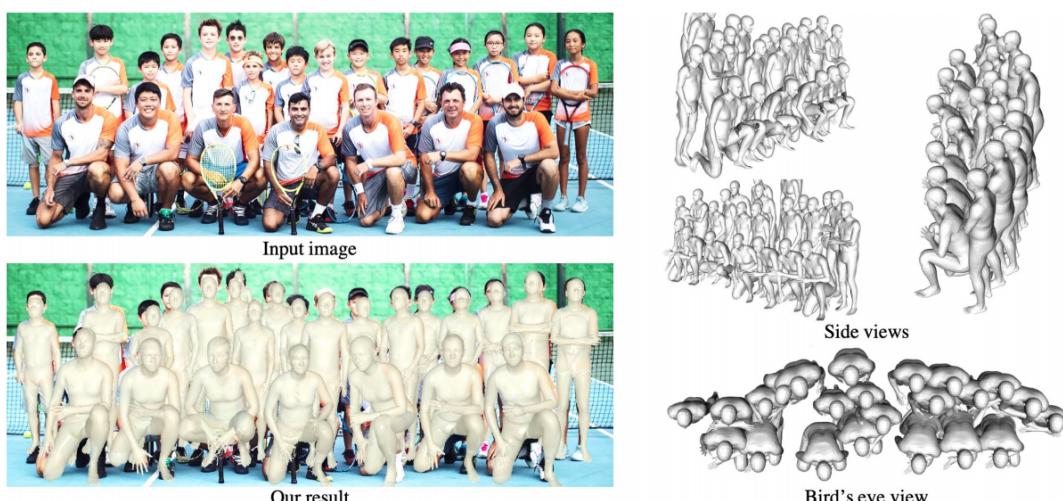


图 1：文章方法的功能展示

获取三维人体信息对于许多应用都是非常重要的，例如，在虚拟现实和增强现实中，需要准确地模拟人体的动作和位置；在机器人导航中，需要准确地识别和避开人体；在人体追踪和行为分析中，需要准确地跟踪人体的姿态和动作。然而，由于单眼图像的信息有限，在单眼图像中准确地预测人体的三维信息是一项挑战性的任务。

因此，研究单眼图像中人体的三维信息回归问题具有重要的实际意义。选择这个课题的背景包括：

在许多应用中，使用单眼图像是很常见的，例如手机摄像头、普通摄像头等。获取单眼图像中人体的三维信息可能比使用多眼图像更加方便，因为单眼图像的设备更加便携、成本更低。

在单眼图像中准确地预测人体的三维信息是一个具有挑战性的任务，但同时也是一个有前景的任务。随着计算机视觉技术的不断发展，越来越多的算法和模型被提出用于解决这个问题。因此，研究单眼图像中人体的三维信息回归问题可能是一个有趣且有前途的选题。

在本文中，作者提出了一种新的方法来解决单眼图像中人体的三维信息回归问题。这种新的方法称为 BEV 方法（Bird's Eye View），BEV 方法使用了鸟瞰图（Bird's Eye View）来预测人体的三维信息。具体来说，BEV 方法首先将单眼图像转换为鸟瞰图，然后使用卷积神经网络对鸟瞰图进行训练，从而能更好地预测人体的深度和姿态等三维信息。鸟瞰图是一种特殊的投影图，它将摄像机的视角投影到一个平面上。在这种投影方式下，人体的深度和姿态信息可以更加清晰地呈现出来。因此，通过使用鸟瞰图，BEV 方法能够更加精确地预测人体的深度和姿态信息。在论文中，作者还对 BEV 方法进行了实验验证，证明了该方法在准确度和效率方面都优于其他现有方法。因此，BEV 方法是一种有效的解决单眼图像中人体三维信息回归问题的方法。

选择这个课题的意义在于，它为研究单眼图像中人体的三维信息回归问题提供了一种新的方法，这对于解决许多应用中的问题都是有益的。此外，该方法的实验结果表明，它在准确度和效率方面都优于其他现有方法，这表明它具有较高的实用价值。

2 相关工作

2.1 传统的人体三维信息回归方法

传统的人体三维信息回归方法可以分为基于模板和基于特征两类。基于模板的方法假设人体的形态有一定的固有规律，并建立一个模板库来描述人体的形态。这些模板包括人体的关键点坐标、边界框尺寸等信息。在回归过程中，首先对输入图像进行预处理，然后与模板库中的模板进行匹配，最后得到人体三维信息。基于特征的方法则是通过提取图像中的人体特征来回归人体三维信息。这些特征可以是人体的轮廓、肌肉线条、骨骼结构等。通过提取特征，可以使用各种回归模型，如线性回归、决策树、支持向量机等，来预测人体的三维信息。这些传统的方法已经被深度学习方法所取代，但是仍然有一定的参考价值。

2.2 基于单眼图像的人体三维信息回归方法

基于单眼图像的人体三维信息回归方法：这些方法使用单眼图像获取人体三维信息的方法。一些常见的方法包括使用深度学习模型、使用视差图像、使用多视角图像等。这些方法的优点在于可以使用较为便携的设备，但是存在一些不足，例如准确度较低、受到光线影响大等。

2.3 基于卷积神经网络的人体三维信息回归方法

基于卷积神经网络的人体三维信息回归方法：这些方法使用卷积神经网络（CNN）解决人体三维信息回归问题的方法。这些方法通常使用大量的数据进行训练，以便能够准确地预测人体的三维信息。这些方法的优点在于准确度较高、计算效率较高，但是存在一些不足，例如训练所需的数据量较大、对数据质量要求较高等。

2.4 来自自然场景的单目三维网格回归

这个方法关注回归一个参数模型如 SMPL 从单个 RGB 图像回归三维身体网格。大多数方法可分为多阶段方法和单阶段方法。对于一般的多人情况，大多数现有的方法^{[2][3][4][5][6]}都是基于一个典型的两阶段框架，它首先检测人，然后分别估计每个人的参数。最近的方法集中于探索各种监督^[7]信号，如时间相干^[8]、轮廓对齐^{[9][10][11]}、自接触^[12]、地面约束^{[13][14]}，或全局人体轨迹^[15]，以增强几何/动态一致性。然而，对于对场景中所有人的深度推理，这些多阶段的方法并不理想。个体裁剪个体的处理不能利用场景上下文或深度排序的原因。

一些单阶段方法同时估计多个三维人^{[16][17]}。给定单幅图像，ROMP^[17]方法输出二维身体中心热图、相机图和参数图，分别用于二维人体检测、定位和网格参数回归。在从二维身体中心热图中解析的位置上，ROMP 从相机和参数图中采样最终的网格参数。这些单阶段的方法具有图像的整体视角，更适合于深度推理。然而，它们是基于不表示深度的二维表示法的。像大多数方法一样，它们模拟成年人（使用 SMPL），对成年人的图像进行训练，因此只预测成年人。为了解决其二维表征和年龄偏差的局限性，本文提出了 BEV 及其约束身高的学习年龄前验训练方案。

2.5 单目深度推理

以前的大多数方法都是通过后处理来深入放置物体。由于他们的 2d 管道和不同年龄组的身高优势不足，他们的结果并不令人满意。一些基于学习的方法，如 3DMPPE^[18]和 CRMH^[19]，解决了多阶段的深度推理。3DMPPE 使用图像特征来细化基于边界盒的深度预测。CRMH 从实例分割中学习，以区分重叠的人之间的相对深度。然而，实例分割是昂贵的，在没有重叠的情况下，无法促进深度关系的学习。SMAP^[20]和 HMOR^[21]使用二维深度图来表示每个像素上 3 个三维姿态的根深度。然而，在拥挤的场景中，这些二维表示是模糊的。相比之下，BEV 采用了一种新的基于鸟瞰图的三维表示法来区分不同深度的人，因此，它对重叠的情况更为稳健。最近，乌格里诺维奇等人^[22]提出了一种基于优化的方法来细化估计体网格的三维平移。它们将 3D 身体网格与检测到的 2D 姿势相匹配，并迫使双脚接触地面。相比之下，我们基于学习的、单阶段的框架更高效和灵活，并且可以适应更多的场景，比如跳跃。Albiero 等人^[23]通过回归 6 自由度姿态来估计人群中所有面孔的深度；它们不处理形状变化或关节。

3 本文方法

3.1 本文方法概述

在本文中，作者提出了一种新的方法 BEV 来解决单眼图像中人体的三维信息回归问题。BEV 方法使用了鸟瞰图来预测人体的三维信息。具体来说，BEV 方法首先将单眼图像转换为鸟瞰图，然后使用卷积神经网络对鸟瞰图进行训练，从而能更好地预测人体的深度和姿态等三维信息。鸟瞰图是一种

特殊的投影图，它将摄像机的视角投影到一个平面上。在这种投影方式下，人体的深度和姿态信息可以更加清晰地呈现出来。因此，通过使用鸟瞰图，BEV 方法能够更加精确地预测人体的深度和姿态信息。在论文中，作者还对 BEV 方法进行了实验验证，证明了该方法在准确度和效率方面都优于其他现有方法。因此，BEV 方法是一种有效的解决单眼图像中人体三维信息回归问题的方法。该方法使用了一个单眼图像中人体的深度信息的卷积神经网络来对人体的三维信息进行回归。通过对大量的数据进行训练，该模型能够准确地预测人体的三维信息，包括人体的骨架信息和姿态信息。

总体框架如图 2 所示。BEV 采用了多头架构。给定一个 RGB 图像作为输入，BEV 输出 5 个图。对于从粗到细的定位，我们使用前 4 个图，它们是前视图和鸟瞰视图中的身体中心热图和定位偏移图。我们首先在深度/高度上展开前部/鸟的视线视图贴图，然后将它们组合起来生成 3D 中心/偏移量贴图。对于粗检测，我们从 3D 中心地图中提取人的粗略三维位置。为了进行精细定位，我们从相应的三维偏移图中对三维中心位置的偏移向量进行采样。添加这些可以得到 3 个三维翻译预测。对于三维网格参数回归，我们使用估计的三维平移 (x_i, y_i, d_i) 和网格特征图。三维翻译的深度值 d_i 被映射到一个深度编码。在 (x_i, y_i) 处，我们从网格特征图中抽取一个特征向量，并将其添加到深度编码中，进行动态参数回归。最后，我们利用 SMPL+A 模型将估计的参数转换为身体网格。

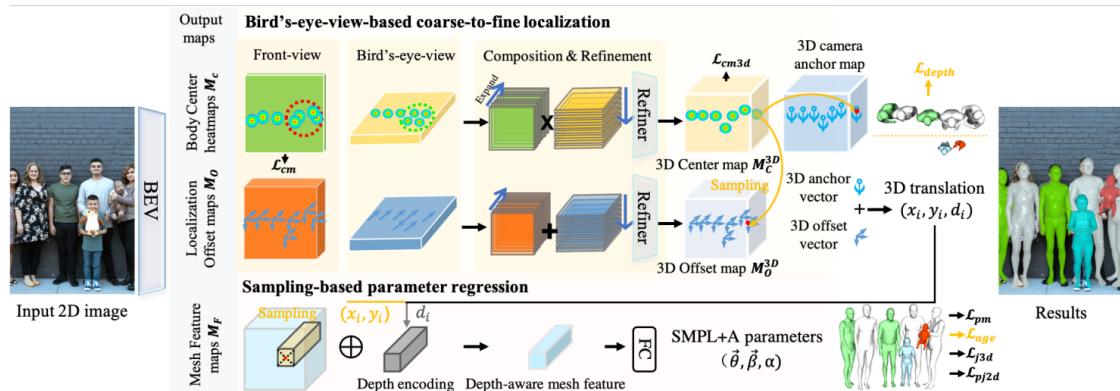


图 2: 检测流程图

3.2 训练数据

由于在野外收集地面真实的 3D 数据是困难的，本文使用成本效益高的野外图像的弱标签来训练 BEV。具体来说，本文收集了一个名为“相对人类”(RH) 的数据集，它包含了深度层和人类年龄的弱注释，分为成人、青少年、儿童和婴儿两组。此外，本文提出了一种弱监督训练方案 (WST) 来有效地从这些弱监督信号中学习。例如，本文使用一个分段损失函数，它利用深度层来惩罚不正确的相对深度顺序。利用年龄信息来限制身高是一件很棘手的事情。虽然年龄和身高是相关的，但身高在同一年龄组内可以有显著差异。因此，本文开发了一个模糊相容的混合损失函数，鼓励体型在每个年龄组的适当范围内的高度。

3.3 使用的损失函数

损失函数定义：在本文中，作者使用了多个损失函数来评估模型的性能。其中，包括基于欧几里得距离的损失函数和基于交叉熵的损失函数。这些损失函数能够帮助模型在训练过程中不断优化，从而提高模型的准确度。基于欧几里得距离的损失函数是用来评估预测人体骨架信息的准确度的。该损失函数计算预测骨架信息与真实骨架信息之间的距离，并以此来评估模型的性能。基于交叉熵的损失函数是用来评估预测人体姿态信息的准确度的。该损失函数计算预测姿态信息与真实姿态信息之间的

差异，并以此来评估模型的性能。在本文中，作者将这两个损失函数结合起来使用，从而更加准确地评估模型的性能。本文的损失函数分为两组：相对损失和标准网格损失。BEV 由所有损失项目的加权和来监督。首先，本文引入了弱监督训练（WST）的两个相对损失函数，分段深度层损失深度和深度设计监督预测深度。如果预测的深度差在一个可接受的范围内，则其深度为 0，即大于 DL 差和体宽 γ 的乘积。否则，Ldepth 将鼓励该模型来实现它。先前的顺序深度损失 [5,30] 鼓励模型尽可能扩大不同深度层的人之间的深度差异。相比之下，在深度中的惩罚被控制在一个范围内，这有助于避免训练发散。

4 复现过程与实验结果

4.1 与已有开源代码对比

我没有在源代码上做出有效改进，因为项目文件太多，结构复杂，而且是作者上一个方法的延伸，调用了许多未曾接触过的包，光是看懂代码流程我已经是尽力了。而且我认为这篇文章考虑到的因素相当完善，我没想到改进的点。但是为了体现工作量，我将这种方法应用到了实际中。

4.2 复现实验结果

将代码在项目中打开，安装运行时所需要的包后，我拍摄了一些多人场景中的图片来运行本文代码，输出结果如图 3、图 4、图 5 所示。



图 3: 运行结果图展示 1



图 4: 运行结果图展示 2



图 5: 运行结果图展示 3

这里就是一些图片的预测结果，从左到右分别是原图，预测后的展示图，俯视和正视图，以及姿态。它能在多人环境下预测出人之间的相对位置，特别的，在第二张图中成功预测出小孩，并正确还原出它的位置。

4.3 额外工作

复现实验完成后，讲一讲我所做的额外工作，通过模型去预测图像后，不仅会生成一张样例图片或视频，还会生成一个 npz 的数据文件，我使用 python 解读这个文件后发现这个文件里存放了很多人物的位置和姿态信息，如图 6 所示，去查阅官方文档发现，其中最值得关注的信息就是 joints，它包含了 71 个关节点信息，包含 SMPL joints 和 h36m joints 等，有了这些信息我们就可以尝试着在其他 3D

软件中还原这个人物姿态了。

```
results = np.load(path, allow_pickle=True)['results'][()]
join
pass
  results = {dict: 10} {'params_pred': array([[ 0.57758474,  0.53969306,  0.79097843, ...,  0.31619072,
   'params_pred' = {ndarray: (7, 146)} [[ 0.57758474  0.53969306  0.79097843 ... 0.31619072,
   'pred_batch_ids' = {ndarray: (7,)} [0 0 0 0 0 0 0] ...作为Array查看
   'center_confs' = {ndarray: (7,)} [0.3677744  0.31749383 0.24820681 0.20107174 0.168459
   'cam' = {ndarray: (7, 3)} [[ 0.57758474  0.53969306  0.79097843], [ 0.17164034  0.331749
   'smpl_thetas' = {ndarray: (7, 72)} [[ 3.04599023e+00 -3.03497940e-01  3.46466124e-01 ...
   'smpl_betas' = {ndarray: (7, 11)} [[-3.47464114e-01  8.53261709e-01  1.20492458e+00 4
   'cam_trans' = {ndarray: (7, 3)} [[ 1.3653643  0.93160266  2.9898164], [-3.7171257  1.913
   'verts' = {ndarray: (7, 6890, 3)} [[-9.46128070e-02 -6.78667188e-01 -8.98713320e-02], [
   'joints' = {ndarray: (7, 71, 3)} [[[ 2.08286382e-03  1.73510015e-02  8.64059012e-03], [ 8.0
   'pj2d_org' = {ndarray: (7, 71, 2)} [[[1526.7788  1107.5754 ], [1560.1831  1141.159 ], [150
   01 __len__ = {int} 10
```

图 6: 运行输出结果解读

于是，我，使用 blender 这个软件，并搭建了如 svn, vs, cmake 等环境。然后下载基础的 3d 模型文件，在检测完一个视频并输出成 npz 文件后，通过转 fbx 的转换代码，将关节点信息转换成能被 blender 软件所识别 fbx 模型文件，最后成功在 blender 中运行模型的结果。如图 7 所示为视频中的一帧图片。

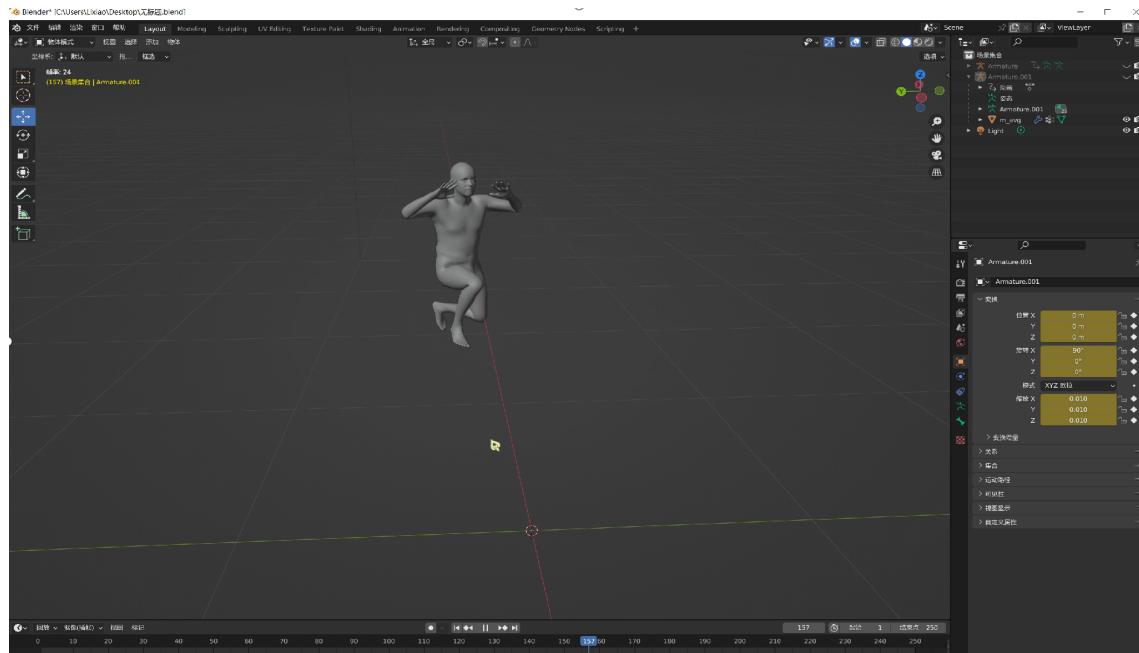


图 7: 运行结果在 3D 软件中展示

5 总结与展望

综上所述，本文讲的是三维人的深度的单目回归，目的是给定有多个人的 RGB 图像，检测每个人在 3D 世界中的相对位置和姿态，本文的方法称为 BEV，它是作者前一代方法 ROMP 的改进，ROMP 是一个单阶段、多人的回归方法，可以直接估计存在多个人的二维前视图，用于二维人体检测、定位和网格参数回归，无需深度推理。在此基础上，作者引入了一个新的想象的 2D “鸟瞰图”，它代表了身体可能的深度中心，将这个纳入模型的训练，使模型能执行基于鸟眼视图的粗检测和精细定位，有了鸟瞰图后便能更好的进行深度估计，而且推断人的深度的前提是知道人的身高，人的高度未知导致深度估计模糊，很难获得具有地面真实高度和深度的图像的训练数据。又因为身高随着年龄的变化而

变化，通过估计图像中的人的年龄，推断人的高度，进而分辨出深度，作者创建了数据集 RH：包括年龄标签和图像中人之间的相对深度关系，并且能体现身高和年龄上的多样性。

网络的效果非常不错，它是单阶段，速度快，而且能在一张图片上同时准确测出婴儿、青年、成人等不同类型的人的 3d 信息，并且基于鸟瞰图的三维表示法的 BEV 对重叠的情况更为稳健，高效且灵活，可以处理跳跃等复杂场景。

展望：历来大部分科学的进步是基于仿生这一原理，在真实世界中大多数生物都有成双的一对眼睛，一只眼睛便可以看清事物，一双眼睛就为看东西赋予了立体效果。本文的方法相当于单个眼睛“估计”人体的立体信息，单个眼睛观察事物的深度很难把控，本文也是基于大量的训练数据才估计的效果，因此我有一个未经过验证的猜想：现在或未来，已经出现或将会出现，将这种方法推广到双目视觉，将人的 3 维估计推广到任何事物的 3 维估计，到时候也许会真正意义上为计算机赋予了观看世界的能力。

参考文献

- [1] SUN Y, LIU W, BAO Q, et al. Putting People in their Place: Monocular Regression of 3D People in Depth[J]. In CVPR, 2022.
- [2] CHOI H, MOON G, PARK J, et al. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes[J]. In CVPR, 2022.
- [3] KANAZAWA A, BLACK M J, JACOBS D W, et al. End-to-end recovery of human shape and pose[J]. In CVPR, 2018: 7122-7131.
- [4] KOLOTOUROS N, PAVLAKOS G, BLACK M J, et al. Learning to reconstruct 3D human pose and shape via model-fitting in the loop[J]. In ICCV, 2019: 2252-2261.
- [5] MOON G, LEE K M. Pose2Pose: 3d positional pose-guided 3d rotational pose prediction for expressive 3d human pose and mesh estimation[J]. ArXiv, 2020.
- [6] PAVLAKOS G, KOLOTOUROS N, DANIILIDIS K. Putting People in their Place: Monocular Regression of 3D People in Depth[J]. In ICCV, 2019: 803-812.
- [7] RONG Y, ZIWEI LIU C L, CAO K, et al. Putting People in their Place: Monocular Regression of 3D People in Depth[J]. In ICCV, 2019: 5340-5348.
- [8] KOCABAS M, ATHANASIOU N, BLACK M J. VIBE: Video inference for human body pose and shape estimation[J]. In CVPR, 2020: 5253-5263.
- [9] DWIVEDI S K, ATHANASIOU N, KOCABAS M, et al. Learning to regress bodies from images using differentiable semantic rendering[J]. In ICCV, 2021: 11250-11259.
- [10] PAVLAKOS G, ZHU L, ZHOU X, et al. Learning to estimate 3d human pose and shape from a single color image[J]. In CVPR, 2018: 459-468.
- [11] XIU Y, YANG J, TZIONAS D, et al. ICON: Implicit Clothed humans Obtained from Normals[J]. In CVPR, 2022.

- [12] MULLER L, OSMAN A A, TANG S, et al. On self-contact and human pose[J]. In CVPR, 2021: 9990-9999.
- [13] REMPE D, BIRDAL T, HERTZMANN A, et al. HuMoR: 3d human motion model for robust pose estimation[J]. In ICCV, 2021: 11488-11499.
- [14] YI H, HUANG C H P, TZIONAS D, et al. Human-aware object placement for visual environment reconstruction[J]. In CVPR, 2022.
- [15] YUAN Y, IQBAL U, MOLCHANOV P, et al. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras[J]. In CVPR, 2022.
- [16] MEHTA D, SOTNYCHENKO O, MUELLER F, et al. Single-shot multi-person 3d pose estimation from monocular rgb[J]. In 3DV, 2018: 120-130.
- [17] SUN Y, BAO Q, LIU W, et al. Monocular, one-stage, regression of multiple 3d people[J]. In ICCV, 2021: 11179-11188.
- [18] MOON G, CHANG J Y, LEE K M. Camera distance-aware top-down approach for 3D multiperson pose estimation from a single RGB image[J]. In CVPR, 2019: 10133-10142.
- [19] JIANG W, KOLOTOUROS N, PAVLAKOS G, et al. Coherent reconstruction of multiple humans from a single image[J]. In CVPR, 2020: 5579-5558.
- [20] ZHEN J, FANG Q, SUN J, et al. SMAP: Single-shot multiperson absolute 3D pose estimation[J]. In ECCV, 2020: 550-566.
- [21] WANG C, LI J, LIU W, et al. HMOR: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation[J]. In ECCV, 2020: 242-259.
- [22] UGRINOVIC N, RUIZ A, AGUDO A, et al. Body size and depth disambiguation in multi-person reconstruction from single images[J]. In 3DV, 2021: 53-63.
- [23] ALBIERO V, CHEN X, YIN X, et al. Img2pose: Face alignment and detection via 6dof, face pose estimation[J]. In CVPR, 2021: 7617-7627.