计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2023/17(01)-0074-14
doi: 10.3778/j.issn.1673-9418.2205070

深度学习的三维人体姿态估计综述

王仕宸[1,2],黄　凯[2],陈志刚[2],张文东[1+]
1. 新疆大学 软件学院,乌鲁木齐 830046
2. 中南大学 计算机学院,长沙 410083
+ 通信作者 E-mail: wdzhang@xju.edu.cn

摘　要:三维人体姿态估计的目的是预测出人体关节点的三维坐标位置和角度等信息,构建人体表示(如人体骨骼),以便进一步分析人体姿态。随着深度学习方法的不断推进,越来越多的基于深度学习的高性能三维人体姿态估计方法被提出。然而由于图片的人体遮挡、训练规模需求较大等原因,三维人体姿态估计仍然存在挑战。该研究目的是通过对近年来的多篇研究论文进行回顾,分析和比较这些方法的推理过程和核心要素,从不同输入的角度入手,全面阐述近年来基于深度学习的三维人体姿态估计方法。此外,还介绍了相关数据集和评价指标,在Human3.6M、Campus和Shelf数据集上对部分模型进行实验数据比对,分析对比实验结果。最后,根据本次调查的结果,讨论目前三维人体姿态估计所面临的困难和挑战,对三维人体姿态估计的未来发展进行了探讨。

关键词:三维人体姿态估计;深度学习;神经网络;关键点检测
文献标志码:A　**中图分类号:**TP391

Survey on 3D Human Pose Estimation of Deep Learning

WANG Shichen[1,2], HUANG Kai[2], CHEN Zhigang[2], ZHANG Wendong[1+]
1. School of Software, Xinjiang University, Urumqi 830046, China
2. School of Computer Science and Engineering, Central South University, Changsha 410083, China

Abstract: The purpose of 3D human pose estimation is to predict information such as the 3D coordinate position and angle of human joint points, and construct human representations (such as human bones) for further analysis of human posture. With the continuous advancement of deep learning methods, more and more high-performance 3D human pose estimation methods based on deep learning have been proposed. However, due to the human occlusion of the picture and the large demand for training scale, there are still challenges in 3D human pose estimation. The research purpose of this paper is to review a number of research papers in recent years, analyze and compare the reasoning process and core elements of these methods, and comprehensively elaborate the 3D human pose estimation methods based on deep learning in recent years. In addition, this paper also introduces the relevant datasets and evaluation indicators, compares the experimental data of some models on the Human3.6M dataset, Campus dataset and Shelf dataset, and analyzes and compares the experimental results. Finally, according to the results of this survey, the difficulties and challenges faced by the current 3D human pose estimation are discussed, and the future development of 3D human pose estimation is discussed.

Key words: 3D human pose estimation; deep learning; neural networks; joints detection

基金项目:长沙市科技计划重大专项(kh2103016)。
This work was supported by the Major Special Projects of Changsha Science and Technology Plan (kh2103016).
收稿日期:2022-05-19　**修回日期:**2022-09-07

人体姿态估计在计算机视觉文献中得到了广泛的研究，它涉及到从传感器获取的输入数据中估计人体部位的信息，生成人体姿态，在运动分析[1]、虚拟现实[2]、医疗辅助[3]、电影制作[4]等领域有着广泛的应用前景。人体姿态估计这个任务，最终面向的使用场景是对视频流进行实时的姿态估计，而且至少要像人类一样能够适应各种复杂环境。然而实现起来需要循序渐进，因此最简单的样例场景就是：从单张图像中识别单个人体，且只需要二维的骨架。从图像和视频中提取二维姿态标注的二维人体姿态估计很容易实现，基于深度学习的单人人体姿态估计技术已经达到很高的性能。

近年来，随着深度学习的快速发展，在图像分类、语义分割和目标检测等任务中，基于深度学习解决方案明显优于传统方法。深度学习被引入姿态估计之后，基于深度学习的人体姿态估计方法可以通过建立网络模型，在图像数据上进行训练和学习，直接得到最有效的表征方法，其核心是深度神经网络，主要是利用神经网络从图像中提取出比人工特征语义信息更丰富、准确性更高和更具鲁棒性的图像特征，并且网络模型的表达能力会因网络堆叠数量的增加而呈指数增长，因此相较于传统方法可以进一步提升复杂环境下的人体姿态估计的精度和鲁棒性。

三维人体姿态估计的主要任务是在三维空间中预测出人体的三维结构信息，换种方式说就是在二维姿态估计结果的基础上加上深度信息。由于深度信息的引入，三维的人体姿态估计在描述人体姿态以及识别人体行为等方面，比二维姿态估计更加精准，拥有更高的研究价值。相比之下，对于三维人体姿态估计来说，获得准确的三维姿态标注要比二维人体姿态估计困难得多。深度学习在人体姿态估计任务中的应用已经取得了显著的进展，然而像遮挡、深度模糊和训练数据不足等挑战仍然是难以克服的。对于基于RGB图像的三维人体姿态估计，单目输入的挑战在于RGB图像固有的深度模糊，而多目输入的挑战在于如何在多个不同的输入视角中匹配正确的姿态。利用运动捕捉系统可以在受控的实验室环境中收集到准确的三维姿势注释，然而在野外环境中就会部分失效。其他的一些工作选择使用RGB-D摄像头和惯性测量单元（inertial measurement unit，IMU）等设备作为输入设备，然而这类设备通常成本较高，不具有商业化能力。基于三维人体姿态估计的重要性，本文主要总结三维人体姿态估计的研究进展。

本文将对三维人体姿态估计按照如图1进行综述。从基于RGB输入的三维姿态和基于其他输入的三维姿态两个角度进行介绍，基于RGB的三维姿态中，从单目和多目两类进行论述。其次，在Human 3.6M数据集[5]中对部分方法进行模型对比，分析不同模型间的差异对模型性能带来的影响。最后，根据研究需要对三维人体姿态估计的数据集及评价指标进行系统性介绍，并且本文将对当前研究面临的问题以及未来的发展趋势进行概述，为这个领域的研究者提供参考。

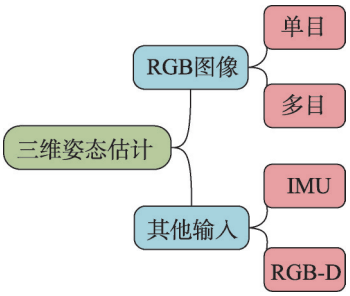

图1 三维人体姿态估计分类

Fig.1 Classification of 3D human pose estimation

1 基于RGB输入的三维人体姿态估计

在人体姿态估计领域，单目RGB摄像头是最常用的输入工具。在野外场景中大多使用单个单目RGB摄像头作为采集设备，然而从单一视图中估计三维人体姿态是一项艰巨的任务。单张RGB图像存在关键点遮挡、深度模糊等问题，并且由于不同的三维人体姿态可以投影成相似的二维姿态，这是一个严重的不适定问题。遮挡问题最直接的解决方法就是从不同角度采集目标图像，在三维姿态估计中使用多个RGB摄像机作为输入可以相对缓解遮挡问题。然而多个摄像机的使用又引入了另一个问题——如何匹配不同视角中的姿态。因此，基于RGB的三维人体姿态估计可以分为单目姿态估计和多目姿态估计两类。

1.1 单目三维人体姿态估计

与二维姿态估计的发展相似，单目三维姿态估计同样从单人姿态估计入手，最后发展到多人姿态估计。因此，单目三维人体姿态估计方法可分为单目单人三维姿态估计和单目多人姿态估计两类。

根据是否使用二维姿态结果作为中间表示，单目单人三维姿态估计进一步可以分为直接估计法和二维提升到三维两种方法。在二维提升到三维的过

程中，由于模型的分段执行，可以灵活地添加人体先验知识、时间序列和参数化人体模型（skinned multi-person linear model，SMPL）[6]等模块提升模型性能。单目多人三维姿态估计分为两类，自顶向下的方法和自底向上的方法。自顶向下的方法首先检测每个人的边界框，再在每个边界框中进行三维姿态估计。自底向上的方法首先检测图中所有关键点，生成关键点坐标和深度图，再对所有关键点进行聚类组合构造人体。

1.1.1 基于直接估计法的单人三维姿态估计

直接估计法没有使用二维姿态结果作为中间表示，而是利用一个完整的大型神经网络端到端从RGB图像中直接推理出三维姿态。通常来讲，很多二维数据对于三维姿态是有帮助的，同时三维姿态也能对二维位置点估计提供额外的信息辅助。文献[7]就把二维骨架以及三维骨架的估计问题关联到一起来做优化。文献[8]使用关节点之间的相对深度进行训练，不需要知道每一个关节点的绝对物理深度，只需要知道关节点之间的深度顺序。文献[9]沿用二维姿态估计的方法，回归出一个三维热图估计各个关键点。文献[10]提出了一种单阶段分布感知式模型（distribution-aware single-stage model，DAS），该模型将三维人体姿态表示为2.5维人体中心点和三维人体关键点偏移，这一表示有效地适配了基于RGB图片域的深度信息预测。文献[11]将输入空间从二维像素空间转换为规范化坐标系中的三维光线，这种简单的设计有效地规范化了摄像机固有参数变化以及摄像机俯仰角变化带来的变化。然而热图的下采样会产生量化误差，文献[12]利用积分回归方法，将热图和回归结合，避免了量化误差的产生并且可以端到端训练。文献[13]将训练中的误差作为样本，利用极大似然估计和基于流的生成模型学习潜在的误差分布。

1.1.2 基于二维到三维的单人三维姿态估计

由于没有二维姿态结果作为中间表示，直接估计法的性能一般低于二维提升到三维的方法，这是因为二维到三维方法利用先进的二维姿态估计器获取人体关键点二维信息，然后由二维人体姿态预测三维人体姿态坐标。文献[14]首先对图像做二维姿态估计，然后利用最近邻匹配寻找最佳三维姿态。文献[15]将二维和三维姿态公式化为距离矩阵回归问题。文献[16]直接使用二维姿态通过神经网络回归出三维姿态。然而以上方法过于依赖二维姿态估计器检测的二维姿态结果，可能会导致次优性能。文献[17]提出了一个双分支框架预测二维热图，利用关键点热图作为中间表示，以获得最终的三维关键点坐标。在此基础上文献[18]利用积分实现端到端训练。

1.1.3 基于先验知识的单人三维姿态估计

在三维人体姿态估计中人体结构的先验知识受到了越来越多的关注，利用先验知识对生成姿态进行约束能有效提高模型性能。文献[19]利用长短期记忆网络（long short-term memory，LSTM）在整个骨骼中传递各个关节点信息；文献[20]引入了人体不同关节的自由度；文献[21]使用顺序双向递归网络（sequential bidirectional recursive network, SeBiReNet）来模拟人类骨骼数据；文献[22]将图神经网络与人体结构模型结合传递上下文信息，生成和修正人体骨骼。然而以上方法没有考虑到二维输入数据的精度，文献[23]发现二维骨架精确度越高，对应获得的三维骨架精度也会提高，通过对二维噪声进行优化再结合人体结构先验知识对结果进行修正，获得了不错的结果。

1.1.4 基于时间序列的单人三维姿态估计

对于从单个RGB图像估计三维人体姿态，连续的视频帧可以提供时间信息来提高三维人体姿态估计的准确性和鲁棒性。文献[24]引入了由LSTM单元组成的序列到序列网络，并在训练期间施加时间平滑性约束，以确保序列的时间一致性。然而缺乏空间构型约束，生成的三维人体姿态依旧可能存在物理上的结构错误。空间依赖性和时间一致性应当同样受到关注，文献[25]在时间网络中加入了解剖学约束，文献[26]在图卷积网络中添加了人体结构先验知识，文献[27]通过骨骼方向和骨骼长度对人体结构进行约束。然而，现有方法主要依靠循环或卷积运算对这些时间信息进行建模，限制了捕捉人体运动全局关系的能力。文献[28]提出了一种运动姿态和形状网络（motion pose and shape network，MPS-Net），以有效地捕捉运动中的人，从视频中估计准确和时间连贯的三维人体姿态和形状。不同关节的运动具有明显的差异性，文献[29]提出了混合时空编码器（mixed spatio-temporal encoder），对每个关节在时序运动上进行建模，并学习关节间的空间关系，以提取到更好的时空信息。

1.1.5 基于SMPL模型的单人三维姿态估计

SMPL模型[6]是一种参数化的人体模型，该方法可以进行任意的人体建模和动画驱动，模拟人的肌

肉在肢体运动过程中的凸起和凹陷，可以避免人体在运动过程中的表面失真，精准地刻画人的肌肉拉伸以及收缩运动的形貌，如图2。在三维姿态估计中SMPL模型[6]也得到了广泛的运用，文献[30]在一个端到端的框架中引入SMPL模型[6]，预测SMPL模型[6]的参数，生成三维人体网格，最后投影三维网格；文献[31]使用基于区域卷积神经网络（region-convolutional neural networks，R-CNN）[32]的网络模型，并引入了SMPL模型[6]参数估计分支作为表示；文献[33]引入一个自监督的人体恢复网格提升了模型的泛化性。然而直接回归SMPL模型[6]会丢失人体部分细节特别是一些高频信息；文献[34]改用图卷积神经网络（graph-convolutional neural networks，G-CNN）仅回归SMPL模型[6]的各个坐标；文献[35]结合了基于回归和基于优化的方法来进行3D人体的姿态和形状估计；文献[36]直接预测每个顶点对应的一维热力图来代替直接回归对应的三维人体相关参数。然而当分辨率降低时，以上的模型可能会失效。文献[37]提出一种基于分辨率感知结构的自我监督网络RSC-Net，能够使用单个模型学习不同分辨率的三维体型和姿势；文献[38]利用特征金字塔从高分辨率特征中提取网格对齐数据反馈给参数进行修正。

图2 SMPL模型

Fig.2 SMPL model

1.1.6 自顶向下的多人三维姿态估计

自顶向下的方法，通常依赖高性能的人体检测方法和单人姿态估计方法，文献[39]在检测出的每个人体边界框中对人体姿态进行定位，再使用一种姿态建议网络进行优化。然而文献[39]在固定数据集中表现良好，对于野外数据集的泛化性较为一般，文献[40]在文献[39]的基础上增加了数据增强模块，提高了模型的泛化能力。随着图像中人体数量的增加，计算复杂度和推理时间可能会变多，特别是在拥挤的场景中。文献[41]依靠图像级别的语义信息，来进行姿态估计，然后利用身体形状、外观参数和使用匈牙利匹配方法解决时间分配问题。以上方法没有考虑到检测出的边界框估计深度可能与实际深度的顺序不一致，预测的人体可能被放置在重叠的位置。文献[42]引入了一种低分辨率的基于锚的表示方法，通过去除模糊锚点来解决重叠问题，再利用每个检测框的相对坐标确定深度顺序。此外，由于自顶向下的方法首先检测到每个人的边界框，场景中的全局信息可能会被忽略。文献[43]引入一种分层多人序数关系的监督形式来解决自顶向下方法缺乏全局视角的问题。

1.1.7 自底向上的多人三维姿态估计

自底向上的方法具有线性计算和时间复杂度，与自顶向下的方法相比，自底向上方法的挑战主要在于如何将不同人体的关键点分类。文献[44]提出了具有可微分阶段的多任务深度神经网络（Muby-Net），它使用肢体评分模块估计被检测关节的候选运动学连接，再使用骨骼分组模块将肢体组装成骨骼。文献[45]使用单级多人姿势机对每个人体定义唯一的身份识别根关节点，利用分层结构化姿势表示将关节点与根关节点组合，解决不同关节点与根关节点距离不一致问题。文献[46]开发了一种基于距离的启发式算法，用于在多人环境中连接关节。具体来说，从检测到的即置信度最高的关节开始，根据三维欧氏距离选择最近的关节连接剩余的关节。由于不使用人体检测，自底向上的方法会受到尺度变化的影响，文献[47]将自顶向下和自底向上的方法结合，提出了一种新型双分支框架，自顶向下分支负责检测图像中的所有人，自底向上分支融入自顶向下分支中的检测信息，负责融合归一化的图像块，解决了由于检测误差引起的尺度变化问题。在处理多人交互产生的遮挡问题中，文献[48]对绝对根节点地图中每个人的远近进行排序，从近到远进行计算，避免重叠。而文献[49]利用遮挡鲁棒姿势图（occlusion-robust pose-maps，ORPM），将不同人的同一关节标定在一张定位图上，并借助二维姿态的信息圈定每个人的位置。最后利用冗余策略生成无法在定位图中标定的遮挡关节点。文献[50]利用二维姿态作为先验知识结合全局背景推断遮挡关节来重建完整的三维姿态。单目图像进行三维人体姿势估计时，往往需要大量带标记数据集。文献[51]利用一些简单的先验知识，在不用任何标注的情况下，通过交叉、变换等操作在三维空间中生成新的三维骨架。文献[52]将单人的三维骨骼随机放置在一个三维网格中，通过生物力学专家提供的关节角度，限制合理的骨骼范围，人工合成包含未知的目标分布的多人三维场景。

1.2 多目三维人体姿态估计

在单目环境下，遮挡是一个具有挑战性的问题。在多目环境中，一个视图中的遮挡部分可能会在其他视图中可见，这个问题可以得到解决，如图3。然而多目环境又产生了新的挑战——如何匹配多个视角中的人物。文献[53]使用二维姿态注释作为监督，提出了一种新颖的弱监督编码器-解码器框架，来学习人体姿势的几何感知三维表示。具体地说，首先将源图像和目标图像映射成二维骨架图，然后训练编码器-解码器从源骨架合成目标骨架。文献[54]在多路匹配算法中加入了时间信息。文献[55]基于体素表达方式，提出了一种方法可以直接在三维空间进行推理，无需在二维图像上进行任何硬决策。文献[56]利用动态匹配模块生成所有二维姿态对与相应的三维姿态，再从三维姿态中筛选正确结果。文献[57]提出了回环约束，确保正确地匹配二维姿态。对极几何是多视角匹配最常用的技术之一，文献[58]在每个视角中检测出关节点热图，再根据相机参数使用对极几何进行视角匹配。然而在视角发生变化时，需要重新对模型训练，文献[59]提出了一种预训练的多视角融合模型，将模型分解成两个子模型，其中较大的模型被所有摄像机共享，另外一个轻量化模型则负责在相机姿态发生变化时，使用少量训练图像进行微调，再通过部署元学习框架对模型进行训练，提高多视角融合的泛化能力。然而在拥挤环境下对极几何仍然可能失效，文献[60]提出了一种足部匹配方法。首先在多个视图中找到脚的最佳匹配，然后利用人体运动链将脚对应扩展到其他关节。文献[61]在没有三维标注的情况下可以自动获取人体姿态的三维标注，并用于微调预训练的网络。文献[62]利用可见视图中的特征来增强遮挡视图中的特征，通过热图的稀疏性来解决两个视图之间的点对应关系。

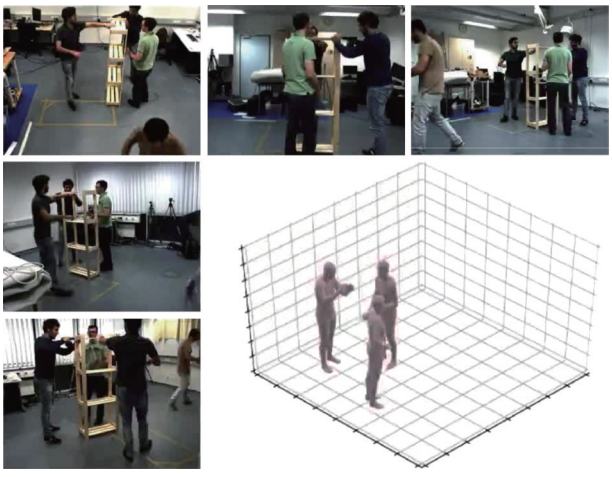

图3 多视点多人三维姿态估计

Fig.3 3D poses estimation of multiple people from multiple views

多目三维人体姿态估计中，模型的推理时间也是考虑的重点。在对所有视图进行二维姿态匹配时的计算复杂度会随着摄像机数量的增加而激增，文献[63]采用迭代处理策略，按照时间顺序获取视频帧，并迭代地逐帧输入，使得计算代价与相机的个数成线性关系。文献[64]将每个视图的图像编码为一个统一的潜在表示，从而将特征图从摄像机视角中分离出来。作为一个轻量级的规范融合，这些二维表示被提升到三维姿势使用基于GPU的直接线性变换来加速处理。

自大规模运动捕捉数据集的引入以来，在三维姿态估计方面基于学习的方法，特别是深度学习的方法发展势头越来越迅猛。由于其表征学习能力，深度学习模型已经实现了前所未有的高精度。尽管它们取得了成功，但深度学习模型需要大量的数据进行训练，而且数据的收集受到很大限制。文献[65]利用大型动作捕捉数据集AMASS[66]来训练基于视频的人体姿态和形态的生成对抗网络模型，来解决训练数据不足的问题。为了减少对带标记数据集的依赖，各种带监督的方法被提出。文献[67]利用投影多视图一致性创建了一个新的半监督学习框架(multiview-consistent semi-supervised learning，MCSS)，MCSS使用来自未注记、未校准的人体运动多视图视频中姿势信息的相似性作为额外的弱监督信号来指导三维人体姿势回归。文献[68]使用多视图一致性实现弱监督训练。文献[53]从多视角的图片信息中学习几何表示，仅使用二维姿态注释作为监督。文献[69]提出了一种将多个权重共享神经网络的输出混合的自监督方法，利用多视图一致性约束将观察到的二维姿势分解为底层三维姿势和相机旋转，可以从未标记的多视图数据中学习单个图像，进行三维估计姿态。然而，这些带监督的方法除了需要二维真值之外，还需要各种形式的附加监督或多视图设置中的相机参数，相比利用数据增强复杂了许多。文献[70]提出了一种利用二维姿态和对极几何来推理出三维姿态的方法，该方法从多视角图片估计二维姿态，随后利用对极几何去获取三维姿态用于训练三维姿态估计。然而该方法依赖于预先定义的规则，如关

节角度限制和运动学约束，限制了生成数据的多样性，使得生成的模型难以推广到更具挑战性的野外场景。为了解决这一问题，文献[71]提出了一种自动数据增强框架，该框架可以在训练中不断学习训练结果，并反馈出相应强度的数据增强，将训练姿态增强到更大的多样性，从而提高训练后的模型泛化能力。

2 基于其他输入的三维人体姿态估计

单眼RGB相机是三维人体姿态估计最常用的输入设备，然而其无法简单地获取深度信息。引入惯性测量单元、RGB-D摄像机等其他输入设备能很好地克服这一问题。这促进了关于其他输入设备的三维人体姿态估计的研究。

RGB-D图像也被称为深度图像，是指将从图像采集器到场景中各点的距离作为像素值的图像，它直接反映了物体可见表面的几何形状。在人体姿态估计中，RGB-D图像能清晰地显示人体各个部位的深度信息。使用RGB-D图像作为输入，文献[72]同时重建详细的人体几何形状、人体非刚性运动和人体内部形状。文献[73]通过捕获全局空间和局部空间的上下文信息作为关节的局部回归量，以集成的方式预测关节位置，增强泛化能力。文献[74]将基于学习的三维人体恢复与非刚性人体融合相结合，生成精确的稀疏部分扫描。文献[75]利用入射光准确地估计局部表面几何形状和反照率，使用光度学约束作为自我监督，实现详细的表面几何和高分辨率纹理估计。

IMU是测量物体三轴姿态角以及加速度的装置。人体姿态估计中惯性传感器的使用能够正确地估计那些在所有视角下都被遮挡的节点，如图4。文献[76]将单个手持相机和一组连接到身体四肢的惯性测量单元相结合进行姿态估计。文献[77]通过融合IMU数据和多视图图像来估计人类在三维空间中的姿势。然而相机位置发生变化时，该方法需要对模型重新调参。文献[78]利用一种几何方法将多视角下的摄像机和可穿戴式的IMU进行融合，使得摄像机的位置发生变化时，不需要对模型进行调整，只需知道相机参数即可。另外，这个方法也可以应用到没有传感器的场景，仅对多个摄像机的特征进行融合。

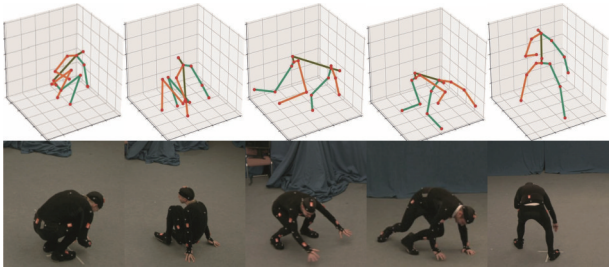

图4 利用IMU辅助的三维姿态估计

Fig.4 3D pose estimation assisted by IMU

3 数据集与评价指标

3.1 三维姿态估计数据集

基于深度学习的人体姿态估计研究需要依赖大量数据来训练模型，数据样本量越大，越多样性，越有利于训练鲁棒的人体姿态估计模型。为三维人体姿态估计数据集获取准确的三维注释是一项具有挑战性的任务，需要像动作捕捉设备和可穿戴的惯性测量单元这样的运动捕捉系统。由于这一需求，许多三维姿势数据集是在受限的环境中创建的。表1列出了几种广泛使用的基于深度学习的三维姿态估计的国际标准数据集，介绍了数据集的样本数量、数据集来源以及适用类型。

表1 三维姿态估计数据集

Table 1 3D human pose estimation datasets

数据集	样本数	样本采集	适用范围	视角
MPI-INF-3DHP[79]	1 300 000	室内/室外	单人	单目
GTA-IM[80]	1 000 000	虚拟场景	单人	单目
NBA2K[81]	27 000	虚拟场景	单人	单目
AMASS[66]	9 000 000	室内/室外	单人/多人	单目
3DPW[82]	51 000	室内/室外	单人/多人	单目
HumanEva[83]	40 000	室内	单人	单目/多目
Human3.6M[5]	3 600 000	室内	单人	单目/多目
CMU Panoptic[84]	1 500 000	室内	单人/多人	单目/多目

MPI-INF-3DHP数据集[79]是一个三维人体姿势估计数据集，由受约束的室内和复杂的室外场景组成。它记录了8名演员在14个摄像机视图内执行的8项活动。它由从14个摄像头捕获大于130万帧的图片组成。除了一个人的室内视频外，他们还提供MATLAB代码，通过混合分段的前景人类外观来生成一个多人数据集MuCo-3DHP。通过提供的身体部分分割，研究人员还可以使用额外的纹理数据来交换衣服和背景。

GTA-IM数据集[80]是一个GTA室内活动数据集。由侠盗猎车手(GTA)电子游戏虚拟引擎从侠盗猎车手(GTA)电子游戏中收集。它包含100万个

1 920×1 080分辨率的RGB-D帧,具有带标注的98个三维人体姿态关节点,涵盖了各种动作,包括坐姿、走路、爬坡和开门。每个场景都包含多个设置,例如客厅、卧室和厨房,这些设置强调人与场景的交互。

NBA2K数据集[81]包含一些NBA运动员的人体网格和纹理数据,每一个运动员有大约1 000个不同的动作。对于每个人体网格,还提供了包含脸、手指等35个关键点的三维姿态和其对应的彩色图片和相机参数。数据集包含27个真实球星,但作者没有权限公开这些包含NBA运动员的数据,因此又构建了包含28个虚拟运动员的合成数据集并重新训练了整个框架,合成的运动员有着同样的几何和视觉质量。

AMASS数据集[66]是一个大型开源三维运动捕捉数据集,包含40 h的运动数据,344个主题,超过11 000个动作。这个数据集将15个不同的基于光学标记的人体运动捕捉数据集与SMPL模型[6]统一为人体骨架和表面网格的标准拟合表示。在这个丰富的数据集中,每个身体关节有3个旋转自由度,这些自由度用指数坐标参数化。

3DPW数据集[82]是在自然环境中用一台手持相机拍摄的。该方法利用附着在被试肢体上的IMU图像,利用视频惯性姿态估计三维标注。这个数据集由60个视频片段、超过51 000帧组成,其中包括在城市里散步、上楼梯、喝咖啡或坐公共汽车等日常活动。3DPW数据集[82]包含了大量的三维注释,包括二维/三维姿态注释、三维身体扫描和SMPL模型[6]参数。然而,在一些拥挤的场景中,3DPW数据集[82]只提供目标人的标签。

HumanEva数据集[83]由HumanEva-I和HumanEva-II两个子集构成。HumanEva-I数据集包含与三维身体姿势同步的7视图视频序列(4个灰度和3个颜色)。在3 m×2 m的捕捉区域内,有4名受试者身上执行步行、慢跑、手势、投球和接球、拳击6种常见动作。HumanEva-II是HumanEva-I测试数据集的扩展,包含两个执行动作组合的受试者。

Human3.6M数据集[5]是在室内实验室中收集的,它包含5名女性和6名男性穿着普通的衣服进行的17项日常活动,包括讨论、吸烟、拍照、通话等。它包含360万张三维人体姿势图像和来自4个不同视角的相应图像。主要拍摄设备包括4台数码摄像机、1台飞行时间传感器、10台同步工作的运动摄像机。拍摄区域约为4 m×3 m。提供的标签包括三维关节位置、关节角度、人物边界框以及每个演员的三维激光扫描。

CMU Panoptic数据集[84]是一个大规模的多视图和多人三维姿态数据集。使用包含480个VGA摄像机视图、31个高清视图、10个RGB-D传感器和基于硬件的同步系统进行无标记运动捕捉的。它包含65个片段(5.5 h)的社交互动和150万的三维关键点。标注包括三维关键点、云点、光流等。

3.2 三维姿态估计评价指标

平均关节位置误差(mean per joint position error,MPJPE),由预测关节点与对应实际关节点的欧氏距离决定。MPJPE通常被称为Protocol #1,它还有两个变体P-MPJPE(Protocol #2)和N-MPJPE(Protocol #3),P-MPJPE是先进行旋转等处理向实际值对齐再进行MPJPE,N-MPJPE仅在规模上进行对齐,用于半监督实验。

关键点的正确率(percentage of correct keypoints,PCK)用来衡量身体关节定位的准确性。如果目标关节点落在实际关节点预设像素阈值范围内,则认为是定位正确的。PCKh@0.5则是对PCK的一个轻微的修改。采用测试人员头部长度的50%作为匹配阈值。通过改变阈值百分比,可以生成AUC(area under curve),以进一步评估不同的姿态估计算法的能力。

4 模型对比结果

本章展示部分模型在Human3.6M数据集[5]、Campus数据集[85]和Shelf[86]数据集上的结果,如表2、表3。Human3.6M数据集共有11个子数据集,利用第1、5、6、7、8子集作为训练集,第9、11子集作为测试集。表4使用Campus数据集和Shelf数据集作为测试集。

从表2、表3和表4的模型对比结果可以看出,三维姿态估计方法的性能在Human3.6M数据集[5]上提升得很快。单目三维姿态估计中,由于二维提升到三维方法使用了高性能的二维姿态估计器,导致使用直接估计法的模型[8-9,27]精度普遍低于使用二维提升到三维方法的模型[16,25-26,34-35,38,51]精度。其中文献[25-26]利用了时间信息,使得模型性能在二维提升到三维方法中较为突出。

多目三维姿态估计由于多视角视图输入,遮挡和深度模糊问题可以得到有效缓解,模型[53,58-59,64]的精度一般高于单目三维姿态估计方法。随着各类监督学习和数据增强手段[67,69-70]的引入,数据集不足

表2 三维姿态估计在Human3.6M数据集上的Protocol #1结果

Table 2 Protocol #1 result of 3D human pose estimation on Human3.6M 单位：mm

模型	Direction	Discussion	Eat	Greet	Phone Call	Pose	Purchase	Sit
[79]	99.0	100.1	86.1	101.8	101.3	96.7	94.9	125.3
[79](GT)	57.5	68.6	59.6	67.3	78.1	56.9	69.1	100.0
[15]	69.5	80.2	78.2	87.0	100.8	76.0	69.6	104.7
[14]	71.6	66.6	74.7	79.1	70.1	67.6	89.3	90.7
[14](GT)	53.3	46.8	58.6	61.2	56.0	58.1	48.9	55.6
[19]	62.8	69.2	79.6	78.8	80.8	72.5	73.9	96.1
[9]	67.4	72.0	66.7	69.1	72.0	77.0	65.0	68.3
[17]	54.2	61.4	60.2	61.2	79.4	63.1	81.6	70.1
[49]	58.2	67.3	65.7	75.8	62.2	64.6	82.0	93.0
[52]	55.8	61.4	58.4	71.9	67.6	65.2	67.7	86.7
[40]	55.9	60.0	64.5	56.3	67.4	71.8	55.1	55.3
[16]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1
[16](GT)	37.7	44.4	40.3	42.1	48.2	54.9	44.4	42.1
[44]	49.4	46.6	51.2	51.8	60.3	55.7	56.1	81.9
[8]	48.5	54.4	54.4	52.0	59.4	65.3	49.9	52.9
[61]	41.2	49.2	42.8	43.4	55.6	46.9	40.3	63.7
[20]	44.7	48.9	47.0	49.0	56.4	67.7	48.7	47.0
[25]	44.8	50.4	44.7	49.0	52.9	43.5	45.5	63.1
[24]	44.2	46.7	52.3	49.3	59.9	59.4	47.5	46.2
[24](GT)	35.2	40.8	37.2	37.4	43.2	44.0	38.9	35.6
[26]	44.6	47.4	45.6	48.8	50.8	59.0	47.2	43.9
[53]	41.1	44.2	44.9	45.9	46.5	39.3	41.6	54.8
[87]	37.4	43.5	42.7	42.7	46.6	59.7	41.3	45.1
[27]	41.4	43.5	40.1	42.9	46.6	51.9	41.7	42.3
[22]	36.3	42.8	39.5	40.0	43.9	48.8	36.7	44.0
[18]	34.4	42.4	36.6	42.1	38.2	39.8	34.7	40.2
[29]	36.7	39.0	36.5	39.4	40.2	44.9	39.8	36.9
[29](GT)	21.6	22.0	20.4	21.0	20.8	24.3	24.7	21.9
[42]	32.5	31.5	41.5	36.7	36.3	31.9	33.2	36.5
[42] (GT)	31.0	30.6	39.9	35.5	34.8	30.2	32.1	35.0
[11]	31.2	35.7	31.4	33.6	35.0	37.5	37.2	30.9
[58]	28.9	32.5	26.6	28.1	28.3	29.3	28.0	36.8
[58](GT)	24.0	26.7	23.2	24.3	24.8	22.8	24.1	28.6
[64]	27.3	32.1	25.0	26.5	29.3	35.4	28.8	31.6
[77]	18.7	20.7	22.5	24.5	28.3	40.1	22.7	23.1
[62]	17.8	19.5	17.6	20.7	19.3	16.8	18.9	20.2
模型	Sit Down	Smoke	Take Phone	Wait	Walk	Walk Dog	Walk Together	Avg
[79]	158.3	100.2	112.5	99.6	83.4	109.6	95.8	104.3
[79](GT)	117.5	69.4	82.4	68.0	55.2	76.5	61.4	72.5
[15]	113.9	89.7	102.7	98.5	79.2	82.4	77.2	87.3
[14]	195.6	83.5	93.3	71.2	55.7	85.9	62.5	83.8
[14](GT)	73.4	60.3	76.1	62.2	35.8	61.9	51.1	57.3
[19]	106.9	88.0	86.9	70.7	71.9	76.5	73.2	79.2
[9]	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.5
[17]	107.3	69.3	78.3	70.3	51.8	74.3	63.2	69.7
[49]	68.8	84.5	65.1	57.6	62.2	72.0	63.6	69.5
[52]	84.3	68.3	78.9	67.9	51.8	77.9	55.2	67.9
[40]	84.8	90.7	67.9	57.5	47.8	63.3	54.6	63.5
[16]	74.0	94.6	62.3	59.1	59.1	65.1	52.4	62.9
[16](GT)	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
[44]	94.0	64.4	68.6	61.2	47.8	66.3	48.7	60.0
[8]	65.8	71.1	56.6	52.9	60.9	44.7	47.8	56.2
[61]	97.6	119.9	52.1	42.7	51.9	41.8	39.4	55.2
[20]	63.0	78.1	51.1	50.1	54.5	40.1	43.0	52.6
[25]	87.3	51.7	61.4	48.5	37.6	52.2	41.9	52.1
[24]	59.9	65.6	55.8	50.4	52.3	43.5	45.1	51.9
[24](GT)	42.3	44.6	39.7	39.7	40.2	32.8	35.5	39.2
[26]	57.9	61.9	49.7	46.6	51.3	37.1	39.4	48.8
[53]	73.2	46.2	48.7	42.1	35.8	46.6	38.5	46.3
[87]	52.7	60.2	45.8	43.1	47.7	33.7	37.1	45.6
[27]	53.9	60.2	45.4	41.7	46.0	31.5	32.7	44.1
[22]	51.0	63.1	44.3	40.6	44.4	34.9	36.7	43.4
[18]	45.6	60.8	39.0	42.6	42.0	29.8	31.7	39.9
[29]	47.9	54.8	39.6	37.8	39.3	29.7	30.6	39.8
[29](GT)	26.9	24.9	21.2	21.5	20.8	14.7	15.7	21.6
[42]	44.4	36.7	38.7	31.2	25.6	37.1	30.5	35.2
[42] (GT)	43.8	35.7	37.6	30.1	24.6	35.7	29.3	34.0
[11]	42.5	41.3	34.6	36.5	32.0	27.7	28.9	34.4
[58]	42.0	30.5	35.6	30.0	28.3	30.0	30.5	31.0
[58](GT)	32.1	26.9	31.0	25.6	25.0	28.1	24.4	26.1
[64]	36.4	31.7	31.2	29.9	26.9	33.7	30.4	30.2
[77]	26.0	39.9	33.8	22.9	35.0	20.9	21.3	26.9
[62]	25.7	20.1	19.2	20.5	17.2	20.5	17.3	19.5

注：GT表示使用二维姿态标注。

表3 三维姿态估计在Human3.6M数据集上的Protocol #2结果

Table 3 Protocol #2 result of 3D human pose estimation on Human3.6M 单位:mm

模型	Direction	Discussion	Eat	Greet	Phone Call	Pose	Purchase	Sit
[14]	89.9	97.6	90.0	107.9	107.3	93.6	136.1	133.1
[19]	90.1	88.2	85.7	95.6	103.9	92.4	90.4	117.9
[39]	76.2	80.2	75.8	83.3	92.2	79.0	71.7	105.9
[15]	66.1	77.9	72.6	84.7	99.7	74.8	65.3	93.4
[42]	51.5	56.8	51.2	52.2	55.2	47.7	50.9	63.3
[42](GT)	50.5	55.7	50.1	51.7	53.9	46.8	50.0	61.9
[16]	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6
[24]	36.9	37.9	42.8	40.3	46.8	46.7	37.7	36.5
[53]	36.9	39.3	40.5	41.2	42.0	34.9	38.0	51.2
[20]	33.6	38.1	37.6	38.5	43.4	48.8	36.0	35.7
[26]	35.7	37.8	36.9	40.7	39.6	45.2	37.4	34.5
[77]	26.8	32.0	25.6	52.1	33.3	42.3	25.8	25.9
[25]	28.0	30.7	39.1	34.4	37.1	28.9	31.2	39.3
[87]	37.4	43.5	42.7	42.7	46.6	59.7	41.3	45.1
[27]	32.6	35.1	32.8	35.4	36.3	40.4	32.4	32.3
[22]	30.5	34.9	32.0	32.2	35.0	37.8	28.6	32.6
[18]	29.1	34.9	29.9	32.6	31.2	32.3	27.0	33.3
[29]	28.0	30.9	28.6	30.7	30.4	34.6	28.6	28.1
模型	Sit Down	Smoke	Take Phone	Wait	Walk	Walk Dog	Walk Together	Avg
[14]	240.1	106.7	139.2	106.2	87.0	114.1	90.6	115.9
[19]	136.4	98.5	103.0	94.4	86.0	90.6	89.5	97.5
[39]	127.1	88.0	105.7	83.7	64.9	86.6	84.0	87.7
[15]	103.1	85.0	98.5	98.8	78.1	80.1	74.8	83.5
[42]	69.9	54.2	57.4	50.4	42.5	57.5	47.7	54.4
[42](GT)	68.0	52.5	55.9	49.9	41.8	56.1	46.9	53.3
[16]	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
[24]	48.9	52.6	45.6	39.6	43.5	35.2	38.5	42.0
[53]	67.5	42.1	42.5	37.5	30.6	40.2	34.2	41.6
[20]	51.1	63.1	41.0	38.6	40.9	30.3	34.1	40.7
[26]	46.9	50.1	40.5	36.1	41.0	29.6	33.2	39.8
[77]	40.5	76.6	39.1	54.5	35.9	25.1	24.2	37.5
[25]	60.6	39.3	44.8	31.1	25.3	37.8	28.4	36.3
[87]	52.7	60.2	45.8	43.1	47.7	33.7	37.1	36.2
[27]	42.7	49.0	36.8	32.4	36.0	24.9	26.5	35.0
[22]	40.8	52.0	35.0	31.9	35.6	26.6	28.5	34.6
[18]	37.6	45.9	32.2	31.5	34.5	22.9	25.9	32.1
[29]	37.1	47.3	30.5	29.7	30.5	21.6	20.0	30.6

注:GT表示使用二维姿态标注。

表4 三维姿态估计在Campus数据集和Shelf数据集上的结果

Table 4 Result of 3D human pose estimation on Campus and Shelf 单位:mm

模型	Campus			Shelf		
	Actor 1	Actor 2	Actor 3	Actor 1	Actor 2	Actor 3
[1]	91.8	92.7	93.2	99.7	92.8	97.7
[57]	97.6	93.3	98.0	98.8	94.1	97.8
[56]	97.1	94.8	97.4	98.8	96.2	97.2
[63]	97.1	94.1	98.6	99.6	93.2	97.5
[55]	97.6	93.8	98.8	99.3	94.1	97.6

注:使用正确关键点所占百分比(percentage of correctly estimated parts,PCP)作为评价指标。

问题正在被逐渐解决,模型的跨数据集泛化性也在逐步提升。

5 问题与展望

近年来,三维人体姿态估计算法已取得显著的成果,但仍然存在许多问题与挑战:

(1)从二维映射到三维产生的深度模糊性和不适定性问题。二维人体关键点估计的微小误差可能会在三维空间中产生重大影响,从数据输入的角度提升估计模型效果是一个不错的选择。例如文献[87]从二维关键点优化入手,利用可靠的二维输入,提升了模型性能。文献[76-78]可穿戴传感器的加入,使得三维关键点在遮挡条件下也能被很好捕捉。

(2)缺少可供深度学习训练的带标注数据集数据。目前大多三维人体姿态数据集都是在室内环境或合成场景中捕捉采集的,无法完全模拟真实室外环境,导致训练的姿态估计模型泛化能力较差。数据增强是解决缺少数据集最直接的手段[70-71],除了数

据增强、半监督和弱监督等学习方法，文献[53,67-69]可以有效降低网络模型训练对三维人体姿态数据的需求。

（3）人体姿态结构的复杂性。灵活的身体构造、表示复杂的关节点间关系和高自由度肢体，这可能会导致自我闭塞或罕见、复杂的姿势。文献[68]等方法转换角度，从人体外观入手，解决人体复杂性的问题。

（4）实际应用困难。速度是产品落地中需要重点考虑的问题。目前大部分研究都是在GPU做到接近实时的水平，然而很多应用场景需要在端设备上实现具体应用，例如在手机上实现实时高效的居家运动姿态检测。

目前三维人体姿态估计的研究大多集中在以单模态输入为基础，然而单一模态的信息局限性限制了输入数据的精度和信息多样性。多模态输入利用其多种类信息的独特优势，结合多模态特征融合，可以使得采集到的数据更加精准和多样，为模型学习提供更多有价值的信息。从IMU和深度摄像机的使用可以看出，多模态输入恰好弥补了三维人体姿态估计的模型训练对精确输入数据的要求。在未来的研究中，基于多模态的三维人体姿态估计是一个值得研究的部分。

6 小结

三维人体姿态估计作为近年来计算机视觉的研究热点，在运动分析[1]、虚拟现实[2]、医疗辅助[3]、电影制作[4]等领域都取得了广泛的应用。本文对近年来基于深度学习的三维人体姿态估计算法进行了回顾，并对相关方法进行了分析与对比；最后探讨了三维人体姿态估计目前所面临的挑战以及未来发展趋势。

参考文献：

[1] MULTI-PERSON BRIDGEMAN L, VOLINO M, GUILLEMAUT J Y, et al. Multi-person 3D pose estimation and tracking in sports[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16-21, 2019. Piscataway: IEEE, 2019: 2487-2496.

[2] ZHANG H, SCIUTTO C, AGRAWALA M, et al. Vid2player: controllable video sprites that behave and appear like professional tennis players[J]. ACM Transactions on Graphics, 2021, 40(3): 1-16.

[3] CHEN W, JIANG Z, GUO H, et al. Fall detection based on key points of human-skeleton using openpose[J]. Symmetry, 2020, 12(5): 744.

[4] WILLETT N S, SHIN H V, JIN Z, et al. Pose2Pose: pose selection and transfer for 2D character animation[C]//Proceedings of the 25th International Conference on Intelligent User Interfaces, Cagliari, Mar 17-20, 2020. New York: ACM, 2020: 88-99.

[5] IONESCU C, PAPAVA D, OLARU V, et al. Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36(7): 1325-1339.

[6] LOPER M, MAHMOOD N, ROMERO J, et al. SMPL: a skinned multi-person linear model[J]. ACM Transactions on Graphics, 2015, 34(6): 1-16.

[7] SUN X, SHANG J, LIANG S, et al. Compositional human pose regression[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision, Hawaii, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 2602-2611.

[8] PAVLAKOS G, ZHOU X, DANIILIDIS K. Ordinal depth supervision for 3D human pose estimation[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018. Washington: IEEE Computer Society, 2018: 7307-7316.

[9] PAVLAKOS G, ZHOU X, DERPANIS K G, et al. Coarse-to-fine volumetric prediction for singleimage 3D human pose[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 7025-7034.

[10] WANG Z, NIE X, QU X, et al. Distribution-aware single-stage models for multi-person 3D pose estimation[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, Jun 21-24, 2022. Piscataway: IEEE, 2022: 13096-13105.

[11] ZHAN Y, LI F, WENG R, et al. Ray3D: ray-based 3D human pose estimation for monocular absolute 3D localization[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, Jun 21-24, 2022. Piscataway: IEEE, 2022: 13116-13125.

[12] SUN X, XIAO B, WEI F, et al. Integral human pose regression[C]//LNCS 11210: Proceedings of the 15th European Conference on Computer Vision, Munich, Sep 8-14, 2018. Cham: Springer, 2018: 536-553.

[13] LI J, BIAN S, ZENG A, et al. Human pose regression with residual log-likelihood estimation[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Oct 11-17, 2021. Piscataway: IEEE, 2021: 11025-11034.

[14] CHEN C H, RAMANAN D. 3D human pose estimation= 2D pose estimation+matching[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recog-

nition, Hawaii, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 7035-7043.

[15] MORENO-NOGUER F. 3D human pose estimation from a single image via distance matrix regression[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 2823-2832.

[16] MARTINEZ J, HOSSAIN R, ROMERO J, et al. A simple yet effective baseline for 3D human pose estimation[C]// Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Oct 22-29, 2017. Washington: IEEE Computer Society, 2017: 2640-2649.

[17] TEKIN B, MÁRQUEZ-NEILA P, SALZMANN M, et al. Learning to fuse 2D and 3D image cues for monocular body pose estimation[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Oct 22-29, 2017. Washington: IEEE Computer Society, 2017: 3941-3950.

[18] ZHOU K, HAN X, JIANG N, et al. Hemlets pose: learning part-centric heatmap triplets for accurate 3D human pose estimation[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 2344-2353.

[19] NIE B X, WEI P, ZHU S C. Monocular 3D human pose estimation by predicting depth on joints[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Oct 22-29, 2017. Washington: IEEE Computer Society, 2017: 3447-3455.

[20] WANG J, HUANG S, WANG X, et al. Not all parts are created equal: 3D pose estimation by modeling bi-directional dependencies of body parts[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 7771-7780.

[21] NIE Q, LIU Z, LIU Y. Unsupervised 3D human pose representation with viewpoint and pose disentanglement [C]//LNCS 12364: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 102-118.

[22] MA X, SU J, WANG C, et al. Context modeling in 3D human pose estimation: a unified perspective[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 6238-6247.

[23] YU T, ZHENG Z, ZHONG Y, et al. Simulcap: single-view human performance capture with cloth simulation[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16-21, 2019. Piscataway: IEEE, 2019: 5504-5514.

[24] HOSSAIN M R I, LITTLE J J. Exploiting temporal information for 3D human pose estimation[C]//LNCS 11214: Proceedings of the 15th European Conference on Computer Vision, Munich, Sep 8-14, 2018. Cham: Springer, 2018: 69-86.

[25] DABRAL R, MUNDHADA A, KUSUPATI U, et al. Learning 3D human pose from structure and motion[C]//LNCS 11213: Proceedings of the 15th European Conference on Computer Vision, Munich, Sep 8-14, 2018. Cham: Springer, 2018: 679-696.

[26] CAI Y, GE L, LIU J, et al. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 2272-2281.

[27] CHEN T, FANG C, SHEN X, et al. Anatomy-aware 3D human pose estimation with bone-based pose decomposition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(1): 198-209.

[28] WEI W L, LIN J C, LIU T L, et al. Capturing humans in motion: temporal-attentive 3D human pose and shape estimation from monocular video[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, Jun 21-24, 2022. Piscataway: IEEE, 2022: 13211-13220.

[29] ZHANG J, TU Z, YANG J, et al. MixSTE: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, Jun 21-24, 2022. Piscataway: IEEE, 2022: 13232-13242.

[30] PAVLAKOS G, ZHU L, ZHOU X, et al. Learning to estimate 3D human pose and shape from a single color image[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018. Washington: IEEE Computer Society, 2018: 459-468.

[31] JIANG W, KOLOTOUROS N, PAVLAKOS G, et al. Coherent reconstruction of multiple humans from a single image [C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 14-19, 2020. Piscataway: IEEE, 2020: 5579-5588.

[32] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition , Columbus, Jun 24-27, 2014. Washington: IEEE Computer Society, 2014: 580-587.

[33] KUNDU J N, RAKESH M, JAMPANI V, et al. Appearance consensus driven self-supervised human mesh recovery[C]// LNCS 12346: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 794-812.

[34] KOLOTOUROS N, PAVLAKOS G, DANIILIDIS K. Convolutional mesh regression for single-image human shape

reconstruction[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition , Long Beach, Jun 16-21, 2019. Piscataway: IEEE, 2019: 4501-4510.

[35] KOLOTOUROS N, PAVLAKOS G, Black M J, et al. Learning to reconstruct 3D human pose and shape via model-fitting in the loop[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 2252-2261.

[36] MOON G, LEE K M. I2L-MeshNet: image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image[C]//LNCS 12352: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 752-768.

[37] XU X, CHEN H, MORENO-NOGUER F, et al. 3D human shape and pose from a single low-resolution image with self-supervised learning[C]//LNCS 12354: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 284-300.

[38] ZHANG H, TIAN Y, ZHOU X, et al. Pymaf: 3D human pose and shape regression with pyramidal mesh alignment feedback loop[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Oct 11-17, 2021. Piscataway: IEEE, 2021: 11446-11456.

[39] ROGEZ G, WEINZAEPFEL P, SCHMID C. LCR-Net: localization-classification-regression for human pose[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 3433-3441.

[40] ROGEZ G, WEINZAEPFEL P, SCHMID C. LCR-Net++: multi-person 2D and 3D pose detection in natural images [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(5): 1146-1161.

[41] ZANFIR A, MARINOIU E, SMINCHISESCU C. Monocular 3D pose and shape estimation of multiple people in natural scenes: the importance of multiple scene constraints [C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018. Washington: IEEE Computer Society, 2018: 2148-2157.

[42] MOON G, CHANG J Y, LEE K M. Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 10133-10142.

[43] WANG C, LI J, LIU W, et al. HMOR: hierarchical multi-person ordinal relations for monocular multi-person 3D pose estimation[C]//LNCS 12348: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 242-259.

[44] ZANFIR A, MARINOIU E, ZANFIR M, et al. Deep network for the integrated 3D sensing of multiple people in natural images[C]//Proceedings of the Annual Conference on Neural Information Processing Systems 2018, Montréal, Dec 3-8, 2018: 8420-8429.

[45] NIE X, FENG J, ZHANG J, et al. Single-stage multi-person pose machines[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 6951-6960.

[46] FABBRI M, LANZI F, CALDERARA S, et al. Compressed volumetric heatmaps for multi-person 3D pose estimation [C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 14-19, 2020. Piscataway: IEEE, 2020: 7204-7213.

[47] CHENG Y, WANG B, YANG B, et al. Monocular 3D multi-person pose estimation by integrating top-down and bottom-up networks[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Piscataway: IEEE, 2021: 7649-7659.

[48] ZHEN J, FANG Q, SUN J, et al. Smap: single-shot multi-person absolute 3D pose estimation[C]//LNCS 12360: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 550-566.

[49] MEHTA D, SOTNYCHENKO O, MUELLER F, et al. Single-shot multi-person 3D pose estimation from monocular RGB[C]//Proceedings of the 2018 International Conference on 3D Vision, Verona, Sep 5-8, 2018. Washington: IEEE Computer Society, 2018: 120-130.

[50] MEHTA D, SOTNYCHENKO O, MUELLER F, et al. XNect: real-time multi-person 3D motion capture with a single RGB camera[J]. ACM Transactions on Graphics, 2020, 39(4): 82.

[51] LI S, KE L, PRATAMA K, et al. Cascaded deep monocular 3D human pose estimation with evolutionary training data [C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 14-19, 2020. Piscataway: IEEE, 2020: 6173-6183.

[52] KUNDU J N, REVANUR A, WAGHMARE G V, et al. Unsupervised cross-modal alignment for multi-person 3D pose estimation[C]//LNCS 12358: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 35-52.

[53] CHEN X, LIN K Y, LIU W, et al. Weakly-supervised discovery of geometry-aware representation for 3D human pose estimation[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16-21, 2019. Piscataway: IEEE, 2019: 10895-10904.

[54] ZHANG Y, AN L, YU T, et al. 4D association graph for

realtime multi-person motion capture using multiple video cameras[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 14-19, 2020. Piscataway: IEEE, 2020: 1324-1333.

[55] TU H, WANG C, ZENG W. VoxelPose: towards multi-camera 3D human pose estimation in wild environment[C]//LNCS 12346: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 197-212.

[56] HUANG C, JIANG S, LI Y, et al. End-to-end dynamic matching network for multi-view multi-person 3D pose estimation[C]//LNCS 12373: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 477-493.

[57] DONG J, JIANG W, HUANG Q, et al. Fast and robust multi-person 3D pose estimation from multiple views[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16-21, 2019. Piscataway: IEEE, 2019: 7792-7801.

[58] QIU H, WANG C, WANG J, et al. Cross view fusion for 3D human pose estimation[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 4342-4351.

[59] XIE R, WANG C, WANG Y. MetaFuse: a pre-trained fusion model for human pose estimation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 14-19, 2020. Piscataway: IEEE, 2020: 13686-13695.

[60] CHEN H, GUO P, LI P, et al. Multi-person 3D pose estimation in crowded scenes based on multi-view geometry[C]//LNCS 12348: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 541-557.

[61] PAVLAKOS G, ZHOU X, DERPANIS K G, et al. Harvesting multiple views for marker-less 3D human pose annotations[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 6988-6997.

[62] ZHANG Z, WANG C, QIU W, et al. AdaFuse: adaptive multiview fusion for accurate human pose estimation in the wild[J]. International Journal of Computer Vision, 2021, 129(3): 703-718.

[63] CHEN L, AI H, CHEN R, et al. Cross-view tracking for multi-human 3D pose estimation at over 100 FPS[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 14-19, 2020. Piscataway: IEEE, 2020: 3279-3288.

[64] REMELLI E, HAN S, HONARI S, et al. Lightweight multi-view 3D pose estimation through camera-disentangled representation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 14-19, 2020. Piscataway: IEEE, 2020: 6040-6049.

[65] KOCABAS M, ATHANASIOU N, BLACK M J. VIBE: video inference for human body pose and shape estimation [C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 14-19, 2020. Piscataway: IEEE, 2020: 5253-5263.

[66] MAHMOOD N, GHORBANI N, TROJE N F, et al. AMASS: archive of motion capture as surface shapes[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 5442-5451.

[67] MITRA R, GUNDAVARAPU N B, SHARMA A, et al. Multiview-consistent semi-supervised learning for 3D human pose estimation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 14-19, 2020. Piscataway: IEEE, 2020: 6907-6916.

[68] IQBAL U, MOLCHANOV P, KAUTZ J. Weakly-supervised 3D human pose learning via multi-view images in the wild [C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 14-19, 2020. Piscataway: IEEE, 2020: 5243-5252.

[69] WANDT B, RUDOLPH M, ZELL P, et al. CanonPose: self-supervised monocular 3D human pose estimation in the wild[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 13294-13304.

[70] KOCABAS M, KARAGOZ S, AKBAS E. Self-supervised learning of 3D human pose using multi-view geometry[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16-21, 2019. Piscataway: IEEE, 2019: 1077-1086.

[71] GONG K, ZHANG J, FENG J. PoseAug: a differentiable pose augmentation framework for 3D human pose estimation[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 8575-8584.

[72] 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]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018. Washington: IEEE Computer Society, 2018: 7287-7296.

[73] XIONG F, ZHANG B, XIAO Y, et al. A2J: anchor-to-joint regression network for 3D articulated pose estimation from a single depth image[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 793-802.

[74] LI Z, YU T, PAN C, et al. Robust 3D self-portraits in

seconds[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 14-19, 2020. Piscataway: IEEE, 2020: 1344-1353.
[75] ZHI T, LASSNER C, TUNG T, et al. TexMesh: reconstructing detailed human texture and geometry from RGB-D Video[C]//LNCS 12355: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 492-509.
[76] VON MARCARD T, HENSCHEL R, BLACK M J, et al. Recovering accurate 3D human pose in the wild using imus and a moving camera[C]//LNCS 11214: Proceedings of the 15th European Conference on Computer Vision, Munich, Sep 8-14, 2018. Cham: Springer, 2018: 601-617.
[77] HUANG F, ZENG A, LIU M, et al. DeepFuse: an imu-aware network for real-time 3D human pose estimation from multi-view image[C]//Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, Mar 1-5, 2020. Piscataway: IEEE, 2020: 418-427.
[78] ZHANG Z, WANG C, QIN W, et al. Fusing wearable IMUs with multi-view images for human pose estimation: a geometric approach[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 14-19, 2020. Piscataway: IEEE, 2020: 2200-2209.
[79] MEHTA D, RHODIN H, CASAS D, et al. Monocular 3D human pose estimation in the wild using improved CNN supervision[C]//Proceedings of the 2017 International Conference on 3D Vision, Qingdao, Oct 10-12, 2017. Washington: IEEE Computer Society, 2017: 506-516.
[80] CAO Z, GAO H, MANGALAM K, et al. Long-term human motion prediction with scene context[C]//LNCS 12346: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 387-404.
[81] ZHU L, REMATAS K, CURLESS B, et al. Reconstructing NBA players[C]//LNCS 12350: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 177-194.
[82] REZAZADEH F, KOWSAR R, RAFIEE H, et al. Fermentation of Soybean meal improves growth performance and immune response of abruptly weaned Holstein calves during cold weather[J]. Animal Feed Science and Technology, 2019, 254: 114206.
[83] SIGAL L, BALAN A O, BLACK M J. HumanEva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion[J]. International Journal of Computer Vision, 2010, 87(1): 4-27.
[84] JOO H, SIMON T, LI X L, et al. Panoptic studio: a massively multiview system for social interaction capture [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(1): 190-204.
[85] LIU W, LUO W X, LIAN D Z, et al. Future frame prediction for anomaly detection—a new baseline[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018. Washington: IEEE Computer Society, 2018: 6536-6545.
[86] BELAGIANNIS V, AMIN S, ANDRILUKA M, et al. 3D pictorial structures for multiple human pose estimation[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, Jun 24-27, 2014. Washington: IEEE Computer Society, 2014: 1669-1676.
[87] XU J, YU Z, NI B, et al. Deep kinematics analysis for monocular 3D human pose estimation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 14-19, 2020. Piscataway: IEEE, 2020: 899-908.

王仕宸（1997—），男，新疆乌鲁木齐人，硕士研究生，主要研究方向为计算机视觉、人体姿态估计、边缘计算。
WANG Shichen, born in 1997, M.S. candidate. His research interests include computer vision, human pose estimation and edge computing.

黄凯（1996—），男，湖南常德人，硕士研究生，主要研究方向为计算机视觉、边缘计算。
HUANG Kai, born in 1996, M.S. candidate. His research interests include computer vision and edge computing.

陈志刚（1964—），男，湖南益阳人，博士，教授，主要研究方向为计算机网络与分布式计算、智能无线网络与协同计算等。
CHEN Zhigang, born in 1964, Ph.D., professor. His research interests include computer networks and distributed computing, intelligent wireless networking and collaborative computing, etc.

张文东（1975—），男，甘肃武威人，博士，副教授，主要研究方向为物联网、群智感知、移动边缘计算。
ZHANG Wendong, born in 1975, Ph.D., associate professor. His research interests include Internet of things, crowd-sensing and mobile edge computing.