arXiv:2601.00625v1 [cs.CV] 2 Jan 2026

RePose：面向康复的实时三维人体姿态估计与生物力学分析框架

Junxiao Xue[1†]、Pavel Smirnov[1],Ziao Li[2],Yunyun Shi[3],Shi Chen[3],Xinyi Yin[4]
Xiaohan Yue[5], Lei Wang[6], Yiduo Wang[4],Feng Lin[7,8],Yijia Chen[1],Xiao Ma[1]
Xiaoran Yan[1], Qing Zhang[1],Fengjian Xue[3],Xuecheng Wu[3], §

[1]之江实验室， [2]东北大学， [3]西安交通大学[4]郑州大学， [5]大连民族大学， [6]南京航空航天大学[7]福耀科技大学， [8]湘湖实验室

[§]通讯作者，[†]项目负责人

摘要

我们提出了一种用于康复训练的实时三维人体姿态估计与动作分析方法，称为 RePose。该方法能够实时监测和评估患者在康复过程中的动作，并提供即时反馈与指导，帮助患者正确进行康复训练。首先，我们提出了一套统一流程，利用来自多台摄像机的 RGB 视频输入，端到端地实时进行人体姿态估计和动作分析，可应用于康复训练领域。该处理流程有助于监测并纠正患者的动作，从而帮助他们恢复肌力和运动功能。其次，我们针对存在多人干扰的医疗康复场景，提出了一种快速跟踪方法，单帧跟踪耗时不到 1 毫秒。此外，我们对 SmoothNet 进行了改进，用于实时姿态估计，有效降低姿态估计误差并还原患者的真实运动状态，使动作在视觉上更加平滑。最后，我们使用 Unity 平台实时监测和评估患者在康复过程中的动作，并展示肌肉应力情况，以辅助患者进行康复训练。

日期： 2026年1月5日**通讯作者：**
xuecwu@gmail.com

1 引言

无论是术后康复，还是治疗各种肌肉骨骼疾病，参与物理治疗和康复计划通常都至关重要。然而，让临床医生参与每一次康复训练并不现实，在经济上也不合理 [21]。为解决这一问题，现代医疗系统开始采用相关技术和方法，以提供远程监测和康复服务。远程监测利用传感器、移动设备和互联网，使医生能够实时监测患者的生理指标和康复进展，从而促进远程指导与干预。这些方法可以缩短患者在医院的停留时间，并提供更加便捷且经济高效的康复方案。

尽管人们开发了各种支持物理治疗的工具和设备，例如机器人辅助系统[22], 虚拟现实和游戏界面 [7], 以及基于Kinect的方法 [16], , 但要实现能够自动监测与评估患者表现的多功能、稳健系统，仍存在局限。Liao等人 [18]

提出了一种用于评估身体康复训练的深度学习框架，该框架通过多个抽象层次的时空建模来评估训练效果。然而，该方法需要使用昂贵的光学动作捕捉系统采集数据。Swakshar 等人 [4] 提出了一种基于图卷积网络的方法，用于评估身体康复训练。该方法能够对人体关节进行自注意力处理，但仍依赖 RGB-D 相机作为输入数据。针对帕金森病运动严重程度的评估，Lu 等人 [20] 设计了一种基于姿势的评估系统。Gu 等人 [9] 推出了一种用于居家物理治疗的交互式计算机视觉系统，可跟踪人体运动并进行分析，但无法评估特定的肌肉用力程度。尽管已有进展，开发多功能、稳健的自动监测与评估系统仍面临挑战。为了帮助患者在居家物理治疗中获得指导和反馈，理想的支持系统需要克服两项难题：首先，系统必须能够实时准确地分析患者训练时的动作，这意味着系统需要具备高精度的姿态估计能力，以准确捕捉关节角度和位置。其次，系统还需要及时提供具体的反馈和指导，包括根据患者表现实时提出建议并进行调整，同时向患者展示正确的姿势和动作示范。这有助于患者纠正错误的姿势和动作方式，确保他们在家中获得有效的训练和康复效果。

为克服上述挑战，本文提出一种用于康复训练的实时姿态识别方法，可提供实时反馈和指导。该方法基于 RGB 视频输入，能够准确估计患者的三维身体位置，从而在运动过程中进行动作分析与评估 [43]。该方法可广泛应用于各种环境，例如通过监控摄像头分析公共场所中的人体行为，或通过智能手机摄像头辅助个人康复训练。借助计算机视觉和深度学习技术，我们的方法具有以下优势。首先，它能够实时监测和评估患者的康复行为，并提供即时反馈和指导，帮助患者正确完成训练。这对康复过程的有效性至关重要，因为它可以帮助患者纠正错误动作、调整姿势并提高训练效率。其次，在实验中我们发现，与健康个体相比，康复场景中患者的关节活动范围相对较小，导致重建的人体模型出现明显抖动。为解决这一问题，我们采用 SmoothNet [45] 来减少抖动。该滤波器可以使重建的人体模型更加平滑、逼真，从而提高姿势识别的准确率和可靠性。此外，在有多名参与者（如家庭成员、医生等）的康复场景中，模型只跟踪并检测特定康复患者的行为和动作。这有助于集中分析和评估该患者的康复进展，并提供个性化的康复计划和支持。最后，我们利用 Unity 平台直观展示肌肉和骨骼受力。通过创建虚拟环境并以三维形式呈现，我们可以将康复过程中肌肉和骨骼受到的力可视化，从而提升康复治疗体验的效果和质量。

为克服上述挑战，本文提出了一种用于康复训练的实时姿势识别方法，可提供即时反馈和指导。该方法基于RGB图像或视频输入，能够准确估计患者的三维身体位置，从而分析和评估运动过程中的动作。该方法可广泛应用于各种环境中，例如通过监控摄像头分析公共场所中的人体行为，或通过智能手机摄像头辅助个人康复训练。借助计算机视觉和深度学习技术 [37, 40]，，我们的方法具有以下优势。首先，它能够实时监测和评估患者的康复行为，提供即时反馈和指导，帮助患者正确完成动作。这对康复过程的效果至关重要，有助于患者及时纠正错误、调整姿势并提升训练效果。其次，我们将 SmoothNet 集成到处理流程中以减少抖动。它能够降低数据中的高频分量，使重建的人体模型更加稳定，从而提高姿势识别的准确率和可靠性。第三，在有多名参与者（如家庭成员、医生等）的康复场景中，模型仅跟踪并检测康复患者的行为和动作。这有助于集中精力分析和评估该患者的康复进展，并提供个性化的康复计划和支持 [42]。最后，我们利用 Unity 平台直观地展示肌肉和骨骼应力。通过创建虚拟环境并对其进行三维呈现，我们能够直观地展示康复过程中作用于肌肉骨骼系统的力，从而提升治疗体验的效果和质量。

总之，本文的主要贡献如下：

- 我们提出了 RePose，这是一种实时人体姿态估计处理流程，在 Linux 系统的多人医疗康复场景中可保持约 30 帧率，并且也能够在 Windows 系统上运行。
- 在存在多名参与者的情况下，我们的模型只跟踪并检测正在接受康复治疗的特定个体的动作和行为。此外，我们使用 SmoothNet 减少姿态估计误差导致的抖动。

- 借助 Unity 平台，我们能够直观地实时监测和评估康复行为，并展示肌肉骨骼应力状况，以帮助患者进行康复训练和治疗。

本文的组织结构如下：第 2 节概述相关工作。第 3 节介绍所提出的实时康复运动姿态识别框架和肌肉骨骼应力分析。第 4 节展示所提出的 RePose 框架在公共数据集上的验证结果。第 5 节提出若干有前景的未来研究方向。最后，第 6 节总结全文。

2 相关工作

2.1 人体姿态估计

人体姿态估计是指定位人体身体部位并构建人体表征（如骨骼结构）的过程， [50], 其输入数据包括图像或视频。人体姿态估计（HPE）提供有关人体运动的定量信息，医疗专业人员可利用这些信息诊断复杂病症、制定康复训练计划并指导物理治疗。

目标检测。 为了全面理解图像，不仅要对不同图像进行分类，还需要估计每幅图像中物体的类别和位置。通用目标检测旨在定位并分类任意给定图像中的物体，主要分为两类：一类遵循传统的目标检测流程，先生成区域建议，再将每个建议分类为不同的物体类别；另一类则将目标检测视为回归或分类问题，采用统一框架直接生成最终结果（包括类别和位置） [49]。基于区域建议的方法，如 Fast R-CNN [8], Faster R-CNN [28], 特征金字塔网络（FPN） [19],和 Mask R-CNN [12]，会先生成区域建议，然后根据这些建议对物体进行分类和定位。另一方面，基于回归／分类的方法采用了不同的目标检测方法。它们将目标检测视为回归或分类任务，旨在直接预测物体的类别和位置，而不依赖区域建议。基于回归／分类的方法包括反卷积单次检测器（DSSD） [6], 深度监督目标检测器（DSOD） [29], YOLOv5 [39], 和 YOLOv8 [33]。这些方法通常使用深度神经网络从输入图像中提取特征，并运用各种技术（如锚框）来预测物体的类别和边界框坐标。基于区域建议的方法以显著增加计算开销为代价实现较高精度，而基于回归／分类的方法则优先考虑实时处理，但会牺牲一定的准确率。

二维人体姿态估计。 二维人体姿态估计（2D HPE）旨在从图像或视频序列中估计人体关键点的二维空间位置。根据图像中受试者的数量，可将其分为两种不同的方法：单人姿态估计和多人姿态估计 [41]。单人姿态估计是二维人体姿态估计中的基础研究领域，无需考虑目标之间的遮挡问题。一般而言，单人姿态估计可采用回归方法或基于热图的方法。回归方法采用端到端框架，学习从输入图像到人体关节位置或人体模型参数的映射。通过训练深度神经网络，回归方法能够直接从图像中预测人体关节的精确位置。例如，Toshev 等人的 DeepPose 方法 [34] 基于深度神经网络（DNNs）解决人体关节回归问题。Sun 等人 [32]提出了“组合式姿态回归”，这是一种基于 ResNet-50、具有结构感知能力的回归方法，它采用重参数化姿态表示，以骨骼取代关节。这些方法的优势在于能够同时估计多个关节的位置，并处理人体体型和姿势的变化。另一种常见方法是基于热图的方法。该方法旨在预测身体部位和关节的大致位置，并通过它们在热图上的表示进行监督 [38]。

具体而言，基于热图的人体姿态估计方法会在每个关节所在位置放置二维高斯核，以此估计二维热图，而非直接估计人体关节的二维坐标。热图是与图像尺寸相同的二维矩阵，其中每个元素表示对应身体部位或关节的置信度。通过训练深度神经网络，基于热图的方法能够根据输入图像生成热图，并从中提取关键点的位置信息。例如，Newell 等人 [23]提出了“堆叠沙漏”编码器-解码器网络，该网络通过中间监督执行迭代式自底向上和自顶向下处理。Sun 等人 [31] 提出了高分辨率网络（HRNet），通过并行连接多分辨率子网络并反复进行多尺度融合，学习可靠的高分辨率表征，从而获得更准确的关键点热图预测。

三维人体姿态估计。 三维人体姿态估计（3D HPE）的目标是预测人体关节在三维空间中的位置。三维人体姿态估计旨在根据给定的输入（如图像或视频序列）准确估计人体关节的三维坐标。这一过程需要理解不同身体部位之间的空间关系，并推断它们在三维世界中的位置。精确估计三维人体姿态可以提供有关人体力学的宝贵信息，这些信息可用于动作分析、康复和动画等各种应用。三维人体姿态估计方法通常可分为两类：基于模型的方法和无模型方法。基于模型的方法利用人体解剖学和运动学方面的先验知识来重建三维姿势，通常通过将参数化人体模型拟合到输入数据，并优化其参数使其与观测到的关节位置对齐来实现。另一方面，无模型方法采用深度学习技术，直接根据输入数据回归三维关节位置，而不显式依赖预定义的人体模型。根据视角数量和场景中的人数，三维人体姿态估计方法可进一步分为三类：单视角单人三维人体姿态估计、单视角多人三维人体姿态估计和多视角三维人体姿态估计。本文主要研究多视角和无模型方法。Iskakov 等人 [24] 提出了一种基于可学习三角测量方法的多视角三维人体姿态估计方案。Zhang 等人 [47] 提出了 AdaFuse，该方法在完成标定后，根据所有相机参数，将多视角图像中的二维关键点热图聚合到三维结构模型中。Tu 等人 [35] 提出了 VoxelPose，该方法在三维体素空间中聚合来自各个摄像机视角的特征。他们设计了长方体提议网络和姿态回归网络，分别用于定位所有个体并估计其三维姿态。Zhang 等人 [46] 提出了多视角姿态变换器（MvP），这是一种直接从多视角图像回归三维姿态、无需依赖任何中间任务的模型。为降低 VoxelPose 的计算成本，Ye 等人提出了 Faster VoxelPose [44], ，该方法将特征体重新投影到三个二维坐标平面上，并分别估计 X、Y 和 Z 坐标，从而将 VoxelPose 的速度提高了十倍。

2.2 肌肉骨骼运动

人体运动由肌肉骨骼系统驱动并提供支撑。该系统包括骨骼、骨骼肌、关节和软结缔组织，共同构成复杂的多体系统 [36]。为解决肌肉骨骼模型中的运动问题，通常会采用正向解和逆向解。正运动学方法将肌电图（EMG）测量值 [10, 27] 或从控制器 [11, 26] 获取的神经指令作为模型输入，以预测人体的运动状态。这种方法可用于估算肌肉活动的强度和时机，进而推断身体的姿势和运动。然而，正运动学方法的计算效率相对较低，难以应用于三维步态分析等复杂运动研究。相比之下，逆运动学方法将测得的关节运动学和外力作为模型输入，以估算关节力矩和肌肉力 [25, 30]。该方法通过反向计算肌肉力矩和关节力来推断肌肉激活模式和力输出，对于研究人体运动控制和康复治疗具有重要价值。

2.3 康复系统

目前已开发出各种家用康复设备，包括机器人、可穿戴设备和基于 Kinect 的康复应用。机器人辅助康复 [5] 设备能够借助机器人的力量与精确性，协助患者进行康复训练。这些设备通常配备多关节和传感器，以提供个性化康复治疗。例如，Lokomat、Hocoma 等下肢康复机器人运用体外治疗原理，通过支撑患者的下肢运动，促进下肢功能障碍的康复。机器人辅助康复设备能够提供高度精确的运动控制，从而精准监测

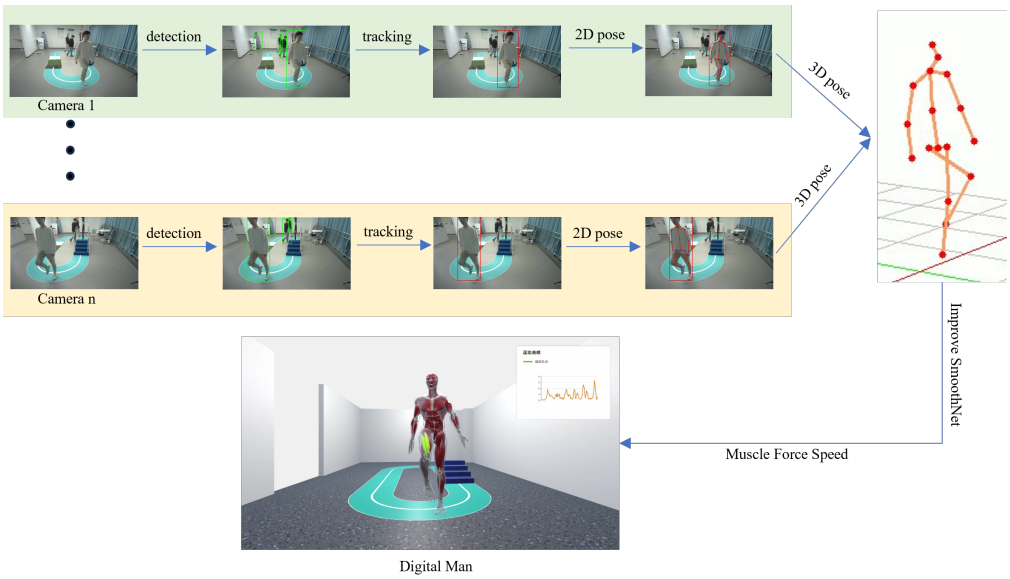


图 1 所提出的 RePose 多摄像头实时数字人处理流程概览。红框表示跟踪框。每台摄像机都会生成二维姿态并合成为三维姿态，随后生成数字人。

并纠正患者的动作，进而促进肌力和运动功能的恢复。然而，这些机器人 [2, 15] 成本高、体积大，且只能提供特定类型的康复训练。可穿戴设备直接佩戴在身体上，用于监测并辅助患者的运动。这些设备小巧便携，患者可以随时随地佩戴，以进行康复训练。不过，可穿戴设备通常只能监测和支持特定部位或类型的运动。对于全身运动或涉及多关节的康复，可能需要额外的设备或方法。基于 Kinect 的康复应用利用 Microsoft Kinect传感器进行康复训练。基于 Kinect 的应用无需患者与设备直接接触，并可通过摄像机识别和跟踪患者的运动，从而支持多种康复训练。然而，由于摄像机技术的局限性，这类应用可能无法提供高度精确的运动跟踪与控制，并且需要购买 Kinect专用硬件。

3 方法

解决方案框架。 我们的方法从多路视频流中重建三维虚拟数字人体模型。在进行三维人体姿态估计之前，需要对用于采集的摄像机进行内参和外参标定，以获取二维图像坐标系与三维空间坐标系之间的映射关系。通过标定，可以将检测到的人体二维坐标转换为三维坐标。随后，多摄像头系统采集多路同步视频流，并对每路视频流的每一帧执行以下处理步骤。采用 YOLO-v5 [39] 目标检测模型估计视频中人物的位置并输出边界框。之后，使用跟踪算法跟踪需要估计姿势的人物。对于多视角三维姿态估计，我们采用 [14]提出的方法。我们使用跟踪框裁剪出感兴趣人物的图像，并估计每帧中 17个人体关节的位置。基于同一时间戳下多帧图像中检测到的 17 个关节的二维坐标，使用可学习线性三角测量方法估计相应关节在三维空间中的位置。接下来，我们使用 SmoothNet 对三维关节位置进行滤波，并将其输入数字人部分进行肌肉骨骼计算。处理流程如图 1所示。

3.1 人体姿态估计模型

目标检测。 在目标检测阶段，使用OpenCV库读取摄像头的视频流，并使用预训练YOLOv5模型检测人员。该过程简述如下：首先，骨干网络从输入图像中提取特征；然后，特征融合模块输出不同尺度的特征图；

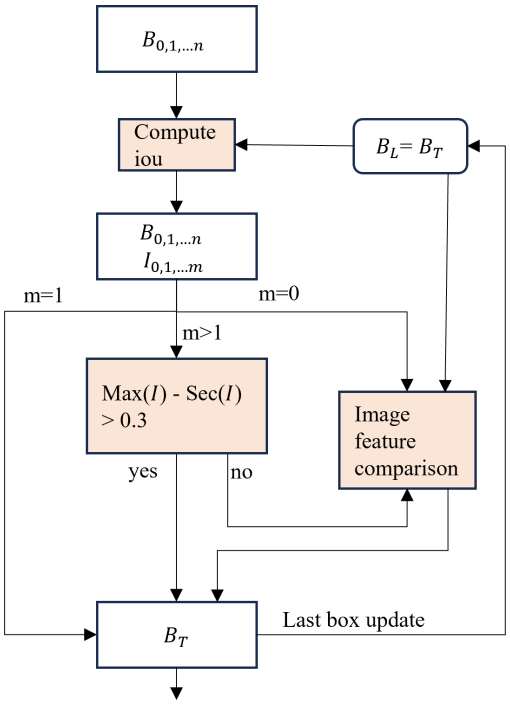


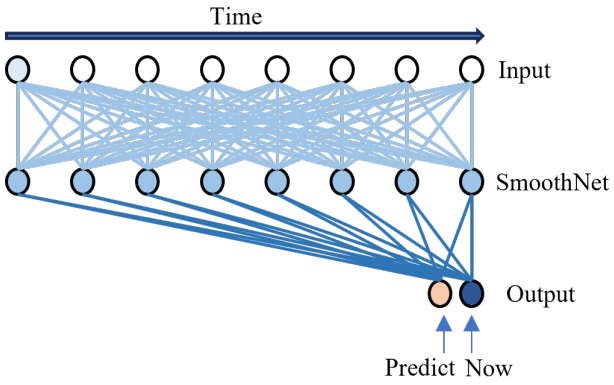
图2 所提出的用于复杂多人康复场景的分层决策逻辑实时单人跟踪算法流程图。

最后，预测模块根据特征图输出目标边界框和目标类别的回归参数。通过设置非极大值抑制（NMS）提取识别出的人体目标。

目标跟踪。 在实际康复场景中，常常会出现单人或多人同时在场的情况。这里，单人指需要进行姿势估计的患者；多人则可能包括患者、家庭成员和医务人员。在此类场景中，需要跟踪患者，以便更好地监测并协助康复过程。当场景中只有一人时，即 n=1，直接进行实时单人跟踪。然而，如果检测到多个人体目标，生成 n 个边界框 $Bboxs = [B_0, B_1...B_n]$，即 n>1，则需要采用跟踪算法，专门针对患者进行康复控制和评估。

如上图 2 所示，我们采用基于交并比（IoU）的相似度计算来衡量两个边界框的重叠程度。重叠区域越大，IoU 值就越大。具体而言，首先将上一帧的目标框 B_L 保存为跟踪目标，然后计算其与当前帧中所有边界框 Bboxs 的 IoU。如果 IoU 大于某个阈值（默认值为 0.2），则将 B_i视为候选框，候选框的数量为 m（m<= n）。如果 m=1，则将当前候选框视为最佳跟踪目标 B_T。如果 m>1，则对候选框进行排序，并计算 IoU 值最高的边界框之间的差值；如果差值大于或等于 0.3，则将 IoU 值最大的边界框视为最佳跟踪目标 B_T。否则（或当 m=0时），这表明仅凭 IoU 无法准确定位跟踪框 B_T，因此采用其他图像特征提取方法，估计当前帧中检测到的边界框与上一帧跟踪边界框之间的相似度。具体而言，计算 n 个候选边界框的颜色特征，并对 B_L采用相同的方法。然后，逐一估计每个候选框 $Bboxs$ 与 B_L 之间的特征相似度，并选择特征相似度最高的边界框作为跟踪目标 B_T。

尽管其他现有跟踪方案可以提供更高的准确率，但需要更多计算资源。然而，在实际医疗康复场景中，患者与周围人群通常不会发生重叠，因此 m 通常等于 1。对于这种情况，我们的跟踪算法无需进行特征计算，只需比较交并比（IoU），因此处理速度快且准确率高。如



Fi图 3 用于实时时序精炼和运动预测的改进版 SmoothNet 架构。

表 1，我们的跟踪算法的平均推理时间不到 1 毫秒。

二维姿态估计。 我们的姿态估计算法采用了 [14] 中提出的方法。首先，我们使用边界框 Bbox 裁剪患者 I_c 的图像。随后，二维姿态估计骨干网络 ResNet152 将该图像作为输入，以估计关节热图 $H_{c,\ j} = h_\theta(I_c)_{\ j}$，其中 $H_{c,\ j}$ 表示图像 I_c 中关节 j 的热图，h_θ 表示一个 1×1 卷积核，$j \in 1,\ \ldots.,\ J$，且 J 为关节数量。为了根据关节热图估计二维位置，我们首先沿热图的空间域计算 softmax，具体如下：

$$H'_{c,j} = \frac{\exp(\alpha H_{c,j})}{\sum_{r_x=1}^{W} \sum_{r_y=1}^{H} \exp(\alpha H_{c,j}(r))} \tag{1}$$

这里，$H'_{c,\ j}$ 表示应用 softmax 操作后的热图，W 和 H 分别表示热图的宽度和高度，$H_{c,\ j} = h_\theta(I_c)_j$ 表示坐标 r 处的热图值，而 α 是用于调整热图的超参数。每个关节在二维空间中的中心位置通过对热图 $H'_{c,\ j}$ 应用 soft-argmax 操作计算得出：

$$x_{c,j} = \sum_{r_x=1}^{W} \sum_{r_y=1}^{H} r \cdot (H'_{c,j}(r)), \tag{2}$$

其中，$x_{c,\ j} \in R^{J2}$ 表示热力图 $H'_{c,\ j}$ 中与关节点对应的二维坐标 r，从而获得每个视角中所有关节点的二维坐标。

三维姿态估计。 获得图像中的二维关节位置 $x_{c,\ j}$ 后，我们使用线性代数三角测量推断关节的三维位置 y_j。为便于计算，我们将三维坐标的求解问题转化为求解关于齐次三维坐标向量的超定方程组：

$$(w_j \circ A_j) y_j = 0. \tag{3}$$

这里，y_j 表示待求的三维坐标向量，$A_j \in R^{2C4}$ 是由投影矩阵 P_C 和二维关节坐标 $x_{c,\ j}$ 构成的矩阵。w_j 由卷积网络 q_ϕ 估计，其输入来自 ResNet152 网络的输出，其中 ∘ 表示哈达玛积。最后，使用奇异值分解（SVD）求解关节的三维位置 y_j。

Modify SmoothNet。 为了获得更平滑的三维姿态估计结果并实现更高的帧率（FPS），我们对用于实时三维姿态估计的 Modify SmoothNet [45] 算法进行了改进。原始 SmoothNet 算法仅适用于视频数据，不仅需要当前帧和前序帧的时序信息，还需要未来帧的信息。然而，在实时处理中只能获取历史数据。此外，我们还旨在通过对帧间的估计姿势进行插值来增强输出的视觉连贯性，从而提高帧率。模型架构沿用了原始 SmoothNet，但对输入和输出数据以及训练



Fi图 4 用于肌肉应力量化和可视化的肌肉骨骼数字人模型。

损失函数。图 3 展示了 Modify SmoothNet 算法的示意图。在 Modify 模型中，我们输入当前帧及其前八帧的三维姿态，使模型能够输出两帧的姿势，即当前帧和预测的中间帧。这种设计使帧率提高一倍，从而实现更流畅的可视化。

此外，除了使用位置损失外，我们还引入了速度损失，以确保输出的姿势更加平滑。由于我们仅输出两个关节点，即当前帧和预测帧，因此采用速度损失。训练损失（Loss）函数计算如下：

$$L_{total} = \alpha L_{pos} + (1-\alpha)L_{acc}, \tag{4}$$

$$L_{pos} = \frac{1}{T \times C} \sum_{t=0}^{T} \sum_{i=0}^{C} |G_{t,i} - Y_{t,i}|, \tag{5}$$

$$L_{vel} = \frac{1}{T \times C} \sum_{t=0}^{T} \sum_{i=0}^{C} |G_{t,i} - G_{t-1,i}||Y_{t,i} - Y_{t-1,i}|. \tag{6}$$

此处， α 表示两种损失之间的比例系数，T 和 C 分别表示输出时间帧数和输出关节数，$G_{t,\ i}$ 和$Y_{t,\ i}$ 分别表示关节位置真实值和估计关节位置，而$|G_{t,\ i} - G_{t-1,\ i}|$以及$|Y_{t,\ i} - Y_{t-1,\ i}|$表示关节 i 在时间帧 t 的速度。

3.2 计算肌肉力

肌肉应力量化计算。 上述模型得出的世界坐标基于右手坐标系，而 Unity 通常采用左手坐标系。因此，我们首先使用罗德里格旋转公式转换三维坐标系，使原点和坐标轴与 Unity 的坐标系对齐。随后，我们采用 FABRIK（前向和后向到达逆运动学）算法 [1] 实现逆运动学。将接收到的坐标数据作为目标点输入后，我们通过 FABRIK 算法驱动数字人体模型运动；该算法根据目标点位置和约束条件计算模型关节的位置和姿势。我们使用从 Unity 库获取的原始数字人体模型，如图 4所示。通过逐帧调整数字人体模型的姿势，确保其与目标点对齐。

FABRIK 算法基于迭代求解方法，通过多次

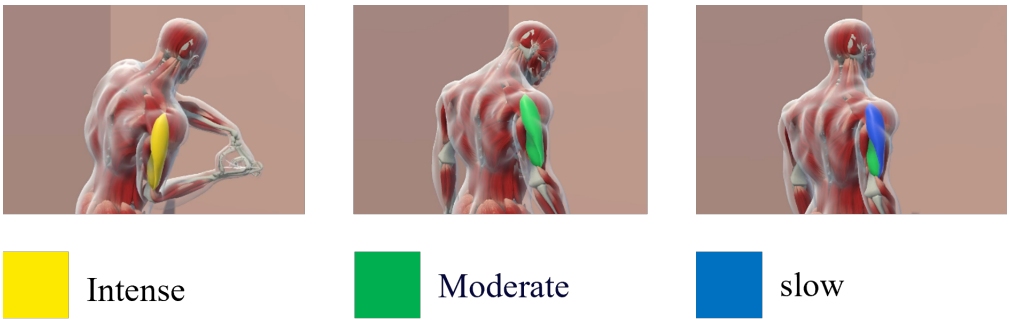


图 5 康复训练期间用于肌肉应力可视化的彩色编码热力图表示。

iter次迭代来逼近目标点。输入包括目标位置 t 以及各关节位置之间的距离

$$d_i = |p_{i+1} - p_i|, \tag{7}$$

其中， p_i 表示关节位置坐标，i=1，.....，n-1。通过迭代算法计算最终关节位置 p_i。

Ass假设 p_1 为根关节，根关节与目标位置之间的距离为

$$dist = |p_1 - t|. \tag{8}$$

如果 $dist > d_1 + d_2 + \cdots + d_{n-1}$，则目标位置不可达。在这种情况下，目标位置与关节位置 r_i 之间的距离使用以下公式确定：

$$r_i = |t - p_i|, \tag{9}$$

$$\lambda_i = \frac{d_i}{r_i}. \tag{10}$$

新的关节位置 P_{i+1} 确定如下：

$$p_{i+1} = (1 - l_i)p_i + l_{it}. \tag{11}$$

如果 $dist \le d_1 + d_2 + \cdots + d_{n-1}$，则目标位置可达。假设根关节 p_1 的初始位置为 b，我们首先检查末端执行器 p_n 与目标 t 之间的距离是否大于容差值 tol

$$dif_A = |p_n - t|. \tag{12}$$

如果 $dif_A > tol$，我们通过将末端执行器 p_n 设置到目标位置 t 来执行前向到达，即 $p_n = t$。目标位置与关节位置 r_i 之间的距离计算如下：

$$r_i = |p_{i+1} - p_i|, \tag{13}$$

$$\lambda_i = \frac{d_i}{r_i}. \tag{14}$$

新的关节位置 p_i 确定如下：

$$p_i = (1 - l_i)p_{i+1} + l_i p_i. \tag{15}$$

接下来，将根关节 p_1 设为初始位置 $p_1 = b$，进行反向伸展。新关节位置 r_i 之间的距离计算如下：

$$r_i = |p_{i+1} - p_i|, \tag{16}$$

$$\lambda_i = \frac{d_i}{r_i}. \tag{17}$$

最终关节位置确定如下：

$$p_{i+1} = (1 - l_i)p_i + l_i p_{i+1}. \tag{18}$$

在算法执行过程中，FABRIK 首先设置数字人体模型的初始关节位置。随后，它通过迭代调整关节位置，使数字人体模型的末端（如手或脚）尽可能靠近目标点。同时，FABRIK 算法遵循关节约束，确保数字人体模型的姿势保持在生理极限范围内。这种方法能够精确控制数字人体模型，使其根据接收到的目标数据做出动作。

肌肉力速度。肌肉力量的计算通常涉及测量速度变化。当肌肉受到外力或进行自身收缩时，会产生运动，进而导致相关骨骼关节处的速度发生变化。通过测量速度变化，我们可以推断肌肉力量的状态。首先，需要确定肌肉网格蒙皮与骨骼关节之间的关联。例如，根据解剖学数据，肱三头肌位于手臂后侧，作用于肩关节和前臂，并与肱二头肌拮抗。利用 Unity 软件，可以识别与肱三头肌相关的骨骼关节，并根据这些关节的速度推断肌肉产生力量的情况。随后，计算骨骼关节的运动速度。每次帧更新时，将当前帧的三维向量减去上一帧的三维向量，再除以帧时间差（0.02f），即可得到表示骨骼节点速度的三维向量，定义如下：

$$V = \frac{(x_n, y_n, z_n) - (x_{n-1}, y_{n-1}, z_{n-1})}{0.02f}. \tag{19}$$

使用协程可以持续计算关节的速度向量。由于协程在后台运行，因此能够持续执行计算和更新，而不会在逐帧执行期间阻塞主线程。最后，根据骨骼运动速度的变化实时调整肌肉的颜色。通过计算速度向量的绝对值，并对其在 x、y 和 z 轴上的分量进行加权，来估算肌肉的受力强度。该强度分为三个级别：强、中和弱，阈值定义为“速度”。当“速度”大于或等于 0.2 时，强度级别为强；当“速度”小于 0.2 且大于 0.08 时，级别为中；当“速度”小于或等于 0.08 时，级别为弱。这些强度级别分别通过黄色、绿色和蓝色显示在相应的肌肉网格蒙皮上。

4 实验

4.1 数据集与评估指标

我们采用最常用的评估协议。协议1为平均每关节位置误差（MPJPE），用于衡量真实值与估计关节之间的平均欧几里得距离，单位为毫米。MPJPE 的公式定义如下：

$$E_{\text{MPJPE}}(f, S) = \frac{1}{N_S} \sum_{i=1}^{N_S} \left\| P_{f,S}^{(f)}(i) - P_{\text{gt},S}^{(f)}(i) \right\|_2, \tag{20}$$

其中，f 表示一帧，S 表示对应的骨架。 $P^{(f)}{}_{f,S}(i)$ 表示关节 i 的估计位置，而 $P^{(f)}{}_{gt,S}(i)$ 表示对应的真实位置。我们考虑所有关节， $N_S = 17$。MPJPE 是对所有帧取平均得到的。协议2是将预测的三维姿态通过平移、旋转和尺度与真实值对齐后计算得到的 MPJPE（P-MPJPE）。

表1 实时流水线推理速度。“Detection 中的 Prep”指目标检测前的图像变换，“3D Pose 中的 Prep”指三维片段处理前的图像裁剪，“FPS 中的 W/O 或 W”表示是否通过 SmoothNet 将帧率加倍。

摄像头数量	视频	检测		跟踪	三维姿态			总计 (ms)	FPS	
		Prep	YOLO		预处理	2D	3D		不含 SN	W SN
4	离线	9.1	2.1	2.5	9.3	6.0	1.7	30.7	33	66
2	离线	4.4	1.3	1.9	4.8	2.9	1.4	16.7	60	120
4	在线	25.0	2.7	2.1	13.2	5.7	4.9	53.6	19	38
2	在线	22.5	2.0	1.4	7.0	3.8	4.0	40.7	25	50

表2 在Human3.6M数据集的协议1下，与当前最先进方法的定量比较。*表示数据集提供的真实边界框，†表示使用我们提出的检测与跟踪算法得到的边界框。

方法	VideoPose3D* [24]	Anatomy3D* [3]	MHFormer* [17]	PoseFormerV2* [48]	Pipeline4
Dir.	45.16	41.49	39.17	41.26	25.61
Dis.	46.67	43.80	43.06	45.45	33.44
Eat	43.31	39.77	40.07	41.47	35.51
Gre.	45.61	43.12	40.94	44.04	31.47
打电话	48.10	46.18	44.95	46.67	63.73
Pho.	55.13	52.54	51.16	53.75	28.62
Pose	44.62	42.16	40.63	42.63	25.44
Pur.	44.28	41.85	41.29	42.64	29.68
Sit	57.30	54.15	53.56	55.19	137.46
坐下	65.79	60.70	60.37	64.62	111.77
Sm.	47.09	45.49	43.71	45.71	46.28
Wait	43.99	41.56	42.12	42.89	26.70
WD.	48.96	46.04	43.84	45.76	30.67
Walk	32.80	31.41	29.76	32.34	27.72
WT.	33.89	32.45	30.59	32.89	26.24
Avg.	46.85	44.18	42.95	45.15	45.36

为了验证我们建模框架的准确率和实时性能，我们在 Human3.6 数据集 [13]上进行了验证测试。Human3.6 数据集使用 4 台高分辨率逐行扫描摄像机，以 50 Hz 采集视频数据。该数据集包含 11 种不同的人体动作类别，涵盖日常生活中的各种动作，例如行走、跑步、举重等。每种动作类别都包含若干不同的人体运动实例，这些实例从不同角度和视角拍摄而成。遵循以往的研究，我们使用五名受试者（S1、S5、S6、S7、S8）训练单个模型，并在两名受试者（S9 和 S11）上进行测试。

4.2 实现细节

我们的处理流程基于 PyTorch 实现。训练二维姿势估计网络时，我们遵循 [14] 中规定的设置。对于 YOLO-v5m，我们采用官方提供的预训练数据。对于 SmoothNet，我们遵循 [45]，中详细说明的训练配置，并使用 Human3.6 数据集进行训练。模型部署方面，处理流程推理在 Linux 服务器上进行，而可视化则在使用 Unity 平台的 Windows 服务器上完成。值得注意的是，对于 YOLO 和二维模型组件，我们将其转换为 16 位浮点类型，以减小模型权重，并使用 TensorRT 进行模型加速。

表1 记录了我们的模型所使用的不同摄像机数量，以及在视频测试和实时场景中的推理速度。显然，我们的处理流程实现了令人称赞的实时性能。“Num cam”指所用摄像头的数量，“video”表示输入数据是“离线还是在线”，用于表明是否进行实时

表3 在 Human3.6M 数据集上，依据协议2与当前最先进方法进行定量比较。

方法	VideoPose3D* [24]	Anatomy3D* [3]	MHFormer* [17]	PoseFormerV2* [48]	处理流程† （我们的方法）
Dir.	34.12	33.01	31.36	32.34	23.38
Dis.	36.12	35.29	34.91	35.87	31.27
Eat	34.43	32.63	32.81	33.78	37.66
Gre.	37.22	35.37	33.85	35.81	23.93
打电话	36.39	35.85	35.32	35.97	47.75
Pho.	42.20	40.43	39.63	41.13	25.59
Pose	34.42	32.86	31.92	33.17	23.35
Pur.	33.59	32.50	32.22	32.71	26.43
Sit	45.03	42.33	43.56	44.26	58.61
坐下	52.53	49.69	48.96	51.93	50.29
Sm.	37.37	36.88	36.36	37.42	39.72
Wait	33.80	32.47	32.58	32.84	24.73
WD.	37.75	36.13	34.42	35.58	27.81
Walk	25.64	25.02	23.85	25.23	25.22
WT.	27.80	26.34	25.10	26.56	23.93
Avg.	36.53	35.12	34.48	35.64	32.64

处理。“Prep in detection”指目标检测前的图像变换，“prep in 3d pose”指三维片段处理前的图像裁剪，“W/O or W in Fps”表示是否通过 SmoothNet 将帧率加倍。离线处理比在线处理更快的原因在于，相机读取需要耗费大量时间和计算资源。

4.3 主要结果

我们在 Human3.6 数据集上测试并报告了处理流程的准确率。尽管我们的主要创新并非姿态估计，但处理流程仍展现出令人瞩目的性能。如表2 和 3所示，*表示使用了 Human3.6 提供的真实检测框。“处理流程”表示采用端到端方法进行推理，而 † 表示使用我们自主开发的目标检测与跟踪算法（由 TensorRT 加速），以及第 4.2 节所述的轻量级 16 位浮点模型。即使不使用 Human3.6 的真实检测框，我们的处理流程在协议1标准下仍展现出与当前先进模型相当的性能。在协议2的评估标准下，我们的处理流程甚至取得了最佳结果。这表明，我们的处理流程在确保推理速度的同时保持了模型准确率，从而验证了我们方法的有效性和实用性。

4.4 可视化

为展示我们处理流程的有效性和多功能性，我们在图6中展示了其在真实场景中的应用。原始图像的初始阶段仅显示第一台摄像机的视角。经过目标检测、目标跟踪和二维姿态估计模块后，我们获得了二维关键点。随后，将来自多台摄像机的这些二维关键点输入三维姿态估计和改进的SmoothNet模块进行处理，我们获得了三维关键点。最终，通过肌肉应力量化计算和肌肉力速度计算，我们得到了所显示的数字人。在实际应用中，数字人的肌肉变化可以通过颜色和运动轨迹进行动态实时显示。该功能可帮助患者了解自己在发力时是否调动了正确的肌肉，或是否通过其他肌肉进行代偿。因此，我们的处理流程为康复辅助增添了价值。

在图 7a 中，我们展示了所提出处理流程的实时性能。对于容易被误解的动作，例如交叉腿（如图 7b 所示），以及存在遮挡的情况（如图 7c 所示），我们的处理流程均展现出稳健的性能。这些真实场景测试结果表明，所提出的处理流程不仅具有很高的准确性，而且实用且通用。

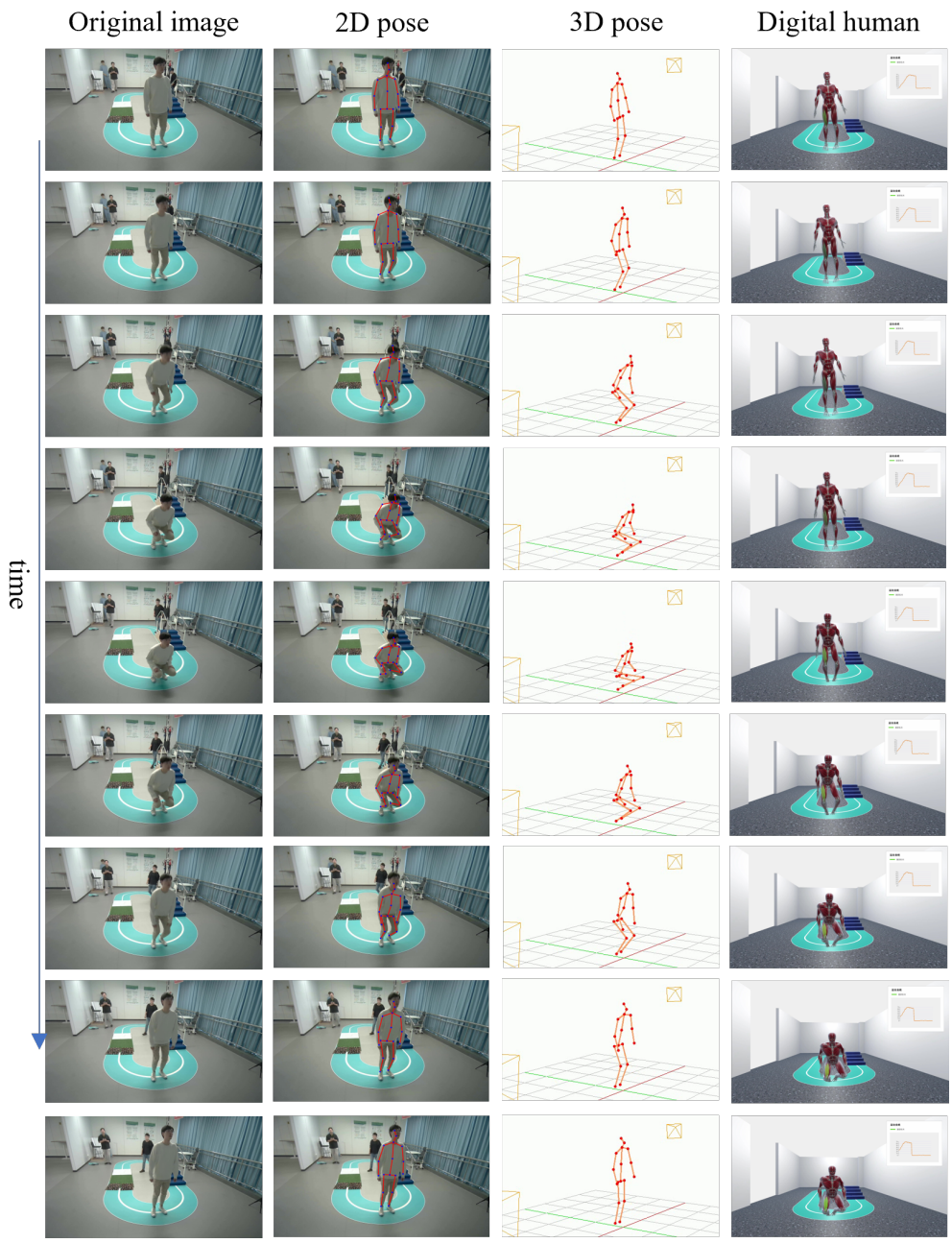


图6 输入图像（仅使用1台摄像机）、二维姿态、处理流程生成的三维姿态，以及以颜色表示受力肌肉的数字人的可视化。

5 未来工作

尽管我们的处理流程在真实场景中取得了令人满意的成果，但仍有一些方面需要改进。未来的研究将主要聚焦于：

遮挡问题。 如图 8 所示，医生或家庭成员协助训练时，遮挡不可避免，这可能严重影响基于图像的姿态估计。我们计划通过多摄像机协作来解决这一问题，由处理流程自动选择视野未被遮挡的摄像机视角进行姿态估计。

视频处理。 Cu 目前，我们的处理流程会分别处理每一帧，然后应用 SmoothNet 滤波器。但这 s

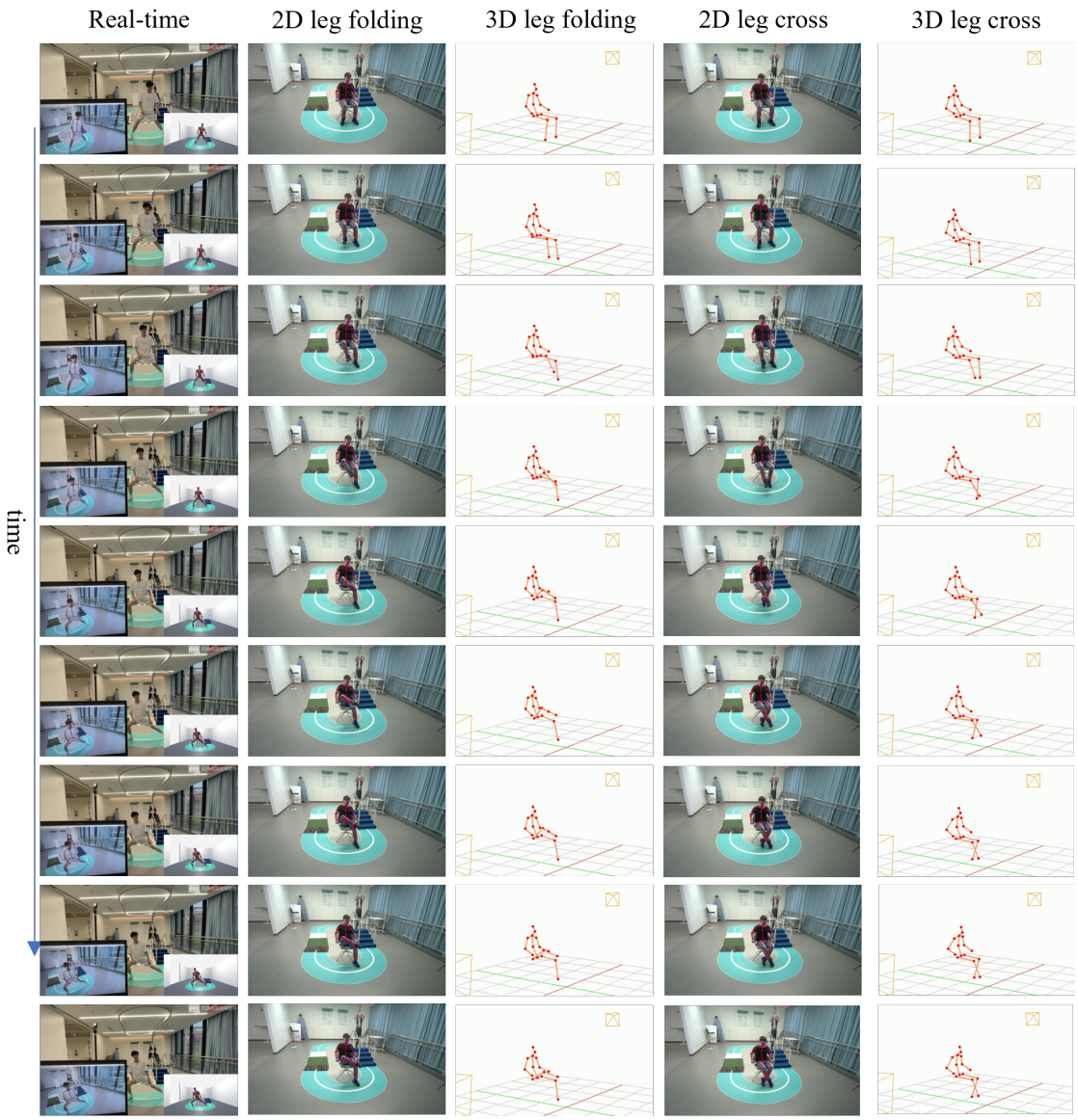


图 7 实时性能可视化。腿部折叠和交叉案例中的二维姿态和三维姿态可视化。

并未充分利用视频中的时序信息。因此，我们将探索如何在不影响推理速度的情况下提取时序特征。

患者特异性姿态先验。 由于患者与健康个体的姿势存在显著差异，有关患者行为的先验信息非常重要。因此，我们将重点开展针对患者的数据收集和训练，以改进我们的处理流程。

总的来说，我们的处理流程在实时应用中表现出色，但我们相信，进一步研究这些方向将提升其在医疗康复场景中的准确率、效率和适用性。

6 结论

我们提出了一种用于康复训练的端到端实时三维人体姿势识别与动作分析框架。该综合处理流程整合了目标检测、高速跟踪、二维/三维姿态估计，以及通过改进版 SmoothNet 进行的时序精炼，最终实现肌肉应力量化和数字人



图 8 遮挡情况的可视化 不同摄像头生成的三维关节点和数字人 e.

可视化。所提出的方法不仅在 Human3.6M 数据集上达到了最先进的准确率，还展现出卓越的计算效率，适用于实时部署。它为患者和临床医生提供了强大的工具，可用于监测肌肉参与情况，并通过即时反馈确保运动安全。除了技术贡献之外，这项研究还具有重大的社会影响：通过促进便捷的居家康复，它有望降低医疗费用，并减轻老龄化社会中临床资源的负担。未来的工作将着重于进一步提升系统在移动设备上的便携性，并开展广泛的临床验证，以满足更多样的康复需求。

参考文献

[1] Andreas Aristidou 和 Joan Lasenby。FABRIK：一种用于逆运动学问题的快速迭代求解器。《图形模型》，73（5）：243–260，2011。[2] Mohamed Bouri、Yves Stauffer、Carl Schmitt、Yves Allemand、Stany Gnemmi、Reymond Clavel、Patrick Metrailler 和 Roland Brodard。步态训练器：一种用于步行康复的机器人系统。载于《2006年IEEE机器人与仿生学国际会议》，第1616–1621页。IEEE，2006。[3] Tianlang Chen、Chen Fang、Xiaohui Shen、Yiheng Zhu、Zhili Chen 和 Jiebo Luo。通过基于骨骼的姿态分解实现具备解剖结构感知能力的三维人体姿态估计。《IEEE视频技术电路与系统汇刊》，32（1）：198–209，2021。[4] Swakshar Deb、Md Fokhrul Islam、Shafin Rahman 和 Sejuti Rahman。用于身体康复训练评估的图卷积网络。《IEEE神经系统与康复工程汇刊》，30：410–419，2022。[5] I˜naki D´ıaz、Jorge Juan Gil、Emilio S´anchez 等。下肢机器人康复：文献综述与挑战。《机器人学杂志》，2011，2011。[6] Cheng-Yang Fu、Wei Liu、Ananth Ranga、Ambrish Tyagi 和 Alexander C Berg。DSSD：反卷积单次检测器。arXiv预印本 arXiv:1701.06659，2017。[7] Lynne V Gauthier、Chelsea Kane、Alexandra Borstad、Nancy Strahl、Gitendra Uswatte、Edward Taub、David Morris、Alli Hall、Melissa Arakelian 和 Victor Mark。门诊卒中电子游戏康复（VIGOR）：一项针对慢性上肢偏瘫康复的居家游戏化限制诱导运动疗法多中心比较效果试验方案。《BMC神经病学》，17（1）：1–18，2017。[8] Ross Girshick。Fast R-CNN。载于《IEEE国际计算机视觉会议论文集》，第1440–1448页，2015。[9] Yiwen Gu、Shreya Pandit、Elham Saraee、Timothy Nordahl、Terry Ellis 和 Margrit Betke。借助交互式计算机视觉系统开展居家物理治疗。载于《IEEE/CVF国际计算机视觉会议研讨会论文集》，第0–0页，2019。[10] Jianqiao Guo、Hongshi Huang、Yuanyuan Yu、Zixuan Liang、Jorge Ambr´osio、Zhihua Zhao、Gexue Ren 和 Yingfang Ao。使用质量可变多体建模公式对肌肉缠绕和质量流动进行建模。《多体系统动力学》，49：315–336，2020。[11] Jianqiao Guo、Wei Guo 和 Gexue Ren。在躯干柔性多体模型中纳入腹内压，并研究静态举升任务中的脊柱减载效应。《生物力学与生物力学建模》，20（4）：1599–1626，2021。[12] Kaiming He、Georgia Gkioxari、Piotr Doll´ar 和 Ross Girshick。Mask R-CNN。载于《IEEE国际计算机视觉会议论文集》，第2961–2969页，2017。[13] Catalin Ionescu、Dragos Papava、Vlad Olaru 和 Cristian Sminchisescu。Human3.6M：用于自然环境中三维人体感知的大规模数据集与预测方法。《IEEE模式分析与机器智能汇刊》，36（7）：1325–1339，2013。[14] Karim Iskakov、Egor Burkov、Victor Lempitsky 和 Yury Malkov。人体姿态的可学习三角测量。载于《IEEE/CVF国际计算机视觉会议论文集》，第7718–7727页，2019。[15] Sa˘so Jezernik、Gery Colombo、Thierry Keller、Hansruedi Frueh 和 Manfred Morari。机器人矫形器 Lokomat：一种康复与研究工具。《神经调控：神经接口技术》，6（2）：108–115，2003。[16] Ravi Komatireddy、Anang Chokshi、Jeanna Basnett、Michael Casale、Daniel Goble 和 Tiffany Shubert。三维运动控制摄像机所提供康复训练的质量与数量：一项试点研究。《国际物理医学与康复杂志》，2（4），2014。[17] Wenhao Li、Hong Liu、Hao Tang、Pichao Wang 和 Luc Van Gool。MHFormer：用于三维人体姿态估计的多假设变换器。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第13147–13156页，2022。[18] Yalin Liao、Aleksandar Vakanski 和 Min Xian。用于评估身体康复训练的深度学习框架。《IEEE神经系统与康复工程汇刊》，28（2）：468–477，2020。doi：10.1109/TNSRE.2020.2966249。[19] Tsung-Yi Lin、Piotr Doll´ar、Ross Girshick、Kaiming He、Bharath Hariharan 和 Serge Belongie。用于目标检测的特征金字塔网络。载于《IEEE计算机视觉与模式识别会议论文集》，第2117–2125页，2017。

[20] Mandy Lu、Kathleen Poston、Adolf Pfefferbaum、Edith V Sullivan、李飞飞、Kilian M Pohl、Juan Carlos Niebles 和 Ehsan Adeli。基于视觉的MDS-UPDRS步态评分估计，用于评估帕金森病的运动严重程度。载于《医学图像计算与计算机辅助干预——MICCAI 2020：第23届国际会议，秘鲁利马，2020年10月4日至8日，论文集，第三部分23》，第637–647页。施普林格，2020。[21] Steven R Machlin、Julia Chevan、William W Yu 和 Marc W Zodet。影响成年人门诊物理治疗疗程利用率和支出的因素。《物理治疗》，91（7）：1018–1029，2011。[22] Pawe l Maciejasz、J¨org Eschweiler、Kurt Gerlach-Hahn、Arne Jansen-Troy 和 Steffen Leonhardt。上肢康复机器人设备综述。《机器人神经工程与康复杂志》，11（1）：1–29，2014。[23] Alejandro Newell、Kaiyu Yang 和 Jia Deng。用于人体姿态估计的堆叠沙漏网络。载于《计算机视觉——ECCV 2016：第14届欧洲计算机视觉会议，荷兰阿姆斯特丹，2016年10月11日至14日，论文集，第VIII部分14》，第483–499页。施普林格，2016。[24] Dario Pavllo、Christoph Feichtenhofer、David Grangier 和 Michael Auli。通过时间卷积和半监督训练进行视频中的三维人体姿态估计。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第7753–7762页，2019。[25] John Rasmussen、Michael Damsgaard 和 Michael Voigt。基于最小/最大准则的肌肉募集——一项比较数值研究。《生物力学杂志》，34（3）：409–415，2001。[26] Reza Sharif Razavian、Naser Mehrabi 和 John McPhee。用于解释自然运动变化的中央模式发生器神经元模型。《计算与非线性动力学杂志》，11（2）：021007，2016。[27] Swithin S Razu 和 Trent M Guess。通过肌电驱动的正向动力学仿真，估算正常、平稳和弹跳步态中的体内关节接触力。《生物医学工程学杂志》，140（7）：071012，2018。[28] Shaoqing Ren、Kaiming He、Ross Girshick 和 Jian Sun。Faster R-CNN：借助区域提议网络实现实时目标检测。《神经信息处理系统进展》，28，2015。[29] Zhiqiang Shen、Zhuang Liu、Jianguo Li、Yu-Gang Jiang、Yurong Chen 和 Xiangyang Xue。DSOD：从零开始学习深度监督目标检测器。载于《IEEE国际计算机视觉会议论文集》，第1919–1927页，2017。[30] MPT Silva 和 JAC Ambr´osio。生物力学系统逆动力学分析中的运动学数据一致性。《多体系统动力学》，8：219–239，2002。[31] Ke Sun、Bin Xiao、Dong Liu 和 Jingdong Wang。用于人体姿态估计的深度高分辨率表征学习。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第5693–5703页，2019。[32] Xiao Sun、Jiaxiang Shang、Shuang Liang 和 Yichen Wei。组合式人体姿态回归。载于《IEEE国际计算机视觉会议论文集》，第2602–2611页，2017。[33] Juan Terven 和 Diana Cordova-Esparza。YOLO综述：从YOLOv1到YOLOv8及未来发展。arXiv预印本 arXiv:2304.00501，2023。[34] Alexander Toshev 和 Christian Szegedy。DeepPose：通过深度神经网络进行人体姿态估计。载于《IEEE计算机视觉与模式识别会议论文集》，第1653–1660页，2014。[35] Hanyue Tu、Chunyu Wang 和 Wenjun Zeng。VoxelPose：迈向野外环境中的多相机三维人体姿态估计。载于《计算机视觉——ECCV 2020：第16届欧洲计算机视觉会议，英国格拉斯哥，2020年8月23日至28日，论文集，第一部分16》，第197–212页。施普林格，2020。[36] Xinyue Wang、Jianqiao Guo 和 Qiang Tian。人体肌肉骨骼多体系统的正逆动力学建模框架。《力学学报》，38（11）：522140，2022。[37] Yifan Wang、Jian Zhao、Zhaoxin Fan、Xin Zhang、Xuecheng Wu、Yudian Zhang、Lei Jin、Xinyue Li、Gang Wang、Mengxi Jia 等。JTD-UAV：用于反无人机系统的多模态大语言模型增强联合跟踪与描述框架。载于《计算机视觉与模式识别会议论文集》，第1633–1644页，2025。[38] Shih-En Wei、Varun Ramakrishna、Takeo Kanade 和 Yaser Sheikh。卷积姿态机。载于《IEEE计算机视觉与模式识别会议论文集》，第4724–4732页，2016。[39] Tian-Hao Wu、Tong-Wen Wang 和 Ya-Qi Liu。基于改进YOLOv5网络的实时车辆与距离检测。载于《2021年第三届世界人工智能研讨会（WSAI）》，第24–28页。IEEE，2021。

[40] Xuecheng Wu, Jiaxing Liu, Danlei Huang, Xiaoyu Li, Yifan Wang, Chen Chen, Liya Ma, Xuezhi Cao, and Junxiao Xue. Vic- bench：使用自由形式中间状态表示对多模态大语言模型的视觉交错思维链能力进行基准测试。*arXiv*预印本 *arXiv:2505.14404*，2025。[41] Xuecheng Wu, Heli Sun, Yifan Wang, Jiayu Nie, Jie Zhang, Yabing Wang, Junxiao Xue, and Liang He. Avf-mae++：通过高效的视听自监督学习扩展情感视频面部掩码自编码器。发表于计算机视觉与模式识别会议论文集，第9142–9153页，2025。[42] Xuecheng Wu, Heli Sun, Junxiao Xue, Jiayu Nie, Xiangyan Kong, Ruofan Zhai, Danlei Huang, and Liang He. 面向短视频情绪分析：大规模数据集与基线。发表于*2025*年国际多媒体检索会议论文集，第1497–1506页，2025。[43] Xuecheng Wu, Junxiao Xue, Xinyi Yin, Yunyun Shi, Liangyu Fu, Danlei Huang, Yifan Wang, Jia Zhang, Jiayu Nie, and Jun Wang. 用于高效情感视频面部分析的可扩展视听掩码自编码器。*Intelligent Computing*，2025。[44] Hang Ye, Wentao Zhu, Chunyu Wang, Rujie Wu, and Yizhou Wang. Faster voxelpose：通过正交投影实现实时三维人体姿态估计。发表于欧洲计算机视觉会议，第142–159页。施普林格，2022。[45] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. S moothnet：一种用于优化视频中人体姿态的即插即用网络。发表于欧洲计算机视觉会议，第625–642页。施普林格，2022。[46] Jianfeng Zhang, Yujun Cai, Shuicheng Yan, Jiashi Feng, et al. 直接进行多视角多人三维姿态估计。神经信息处理系统进展，34：13153–13164，2021。[47] Zhe Zhang, Chunyu Wang, Weichao Qiu, Wenhu Qin, and Wenjun Zeng. Adafuse：用于野外精准人体姿态估计的自适应多视角融合。国际计算机视觉杂志，129：703–718，2021。[48] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2：探索频域以实现高效、稳健的三维人体姿态估计。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第8877–8886页，2023。[49] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. 基于深度学习的目标检测：综述。*IEEE*神经网络与学习系统汇刊，30(11)：3212–3232，2019。[50] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, J u Shen, Nasser Kehtarnavaz, and Mubarak Shah. 基于深度学习的人体姿态估计：综述。*ACM*计算综述，56(1)：1–37，2023。