

中图法分类号: TP391.41 文献标识码: A 文章编号: 1006-8961(2025)10-3377-15

论文引用格式: Jia D, He D K, Han X F, Yang L, Cheng S and Liu Y Q. 2025. Monocular vision 3D human posture estimation network based on multi-joint feature fusion. Journal of Image and Graphics, 30(10):3377-3391(贾迪, 何德堃, 韩雪峰, 杨柳, 程硕, 刘宇琪. 2025. 融合多关节特征的单目视觉三维人体姿态估计. 中国图象图形学报, 30(10):3377-3391)[DOI:10.11834/jig.240672]

# 融合多关节特征的单目视觉三维人体姿态估计

贾迪<sup>1</sup>, 何德堃<sup>2\*</sup>, 韩雪峰<sup>1</sup>, 杨柳<sup>2</sup>, 程硕<sup>2</sup>, 刘宇琪<sup>2</sup>

1. 辽宁工程技术大学鄂尔多斯研究院, 鄂尔多斯 017000; 2. 辽宁工程技术大学电子与信息工程学院, 葫芦岛 125105

**摘要:** 目的 针对目前三维人体姿态估计方法未能有效处理时间序列冗余, 难以捕获人体关节上微小变化的问题, 提出一种融合多关节特征的单目视觉三维人体姿态估计网络。方法 在关节运动特征提取模块中, 采用多分支操作提取关节在时间维度上的运动特征, 并将不同特征融合形成具有高度表达力的特征表示。关节特征融合模块整合了不同关节组和中间帧的全局信息, 通过矩阵内积的方式表达不同关节组在高纬度空间的相对位置及相互联系, 得到中间3D姿态的初估值。关节约束模块引入中间帧的2D关节点空间位置关系作为隐式约束, 与中间帧3D姿态初估值融合, 减少不合理的姿态输出, 提高最终3D姿态估计的准确性。结果 实验结果表明, 与MHFormer方法相比, 本文方法在Human3.6M数据集上的平均关节位置误差(mean per joint position error, MPJPE)结果为29.0 mm, 误差降低4.9%, 对于复杂动作, 如SittingDown和WalkDog, 误差降低了7.7%和8.2%。在MPI-INF-3DHP数据集上, MPJPE指标降低36.2%, 曲线下面积(area under the curve, AUC)指标提升12.9%, 正确关节点百分比(percentage of correct keypoints, PCK)指标提升3%。实验结果体现出在面对复杂动作问题时, 网络利用各分支提取了不同的关节时序运动特征, 将不同关节组的位置信息进行融合交互, 结合当前帧的关节姿态信息加以约束, 取得更高的精度。在HumanEva数据集上的实验结果表明了本文方法适用不同数据集, 消融实验进一步验证了各个模块的有效性。结论 本文网络有效地融合了人体多关节特征, 可以更好地提高单目视觉三维人体姿态估计的准确性, 且具备较高的泛化性。

**关键词:** 三维人体姿态估计; 人体拓扑结构; 多分支网络; 特征融合; 姿态约束

## Monocular vision 3D human posture estimation network based on multi-joint feature fusion

Jia Di<sup>1</sup>, He Dekun<sup>2\*</sup>, Han Xuefeng<sup>1</sup>, Yang Liu<sup>2</sup>, Cheng Shuo<sup>2</sup>, Liu Yuqi<sup>2</sup>

1. Ordos Institute of Liaoning Technical University, Ordos 017000, China;

2. School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China

**Abstract: Objective** Despite the widespread use of monocular cameras for capturing photos and videos in various scenarios, they still encounter challenging issues such as human occlusion and loss of depth information. The goal is to achieve stability and accuracy in this process. Conventional approaches have attempted to estimate the 3D human position

收稿日期: 2024-11-18; 修回日期: 2025-02-07; 预印本日期: 2025-02-14

\* 通信作者: 何德堃 lntu\_hedekun@163.com

**基金项目:** 国家自然科学基金项目(61601213); 辽宁省教育厅重点项目(LJ212410147003); 辽宁工程技术大学鄂尔多斯研究院校地科技合作培育项目(YJY-XD-2023-003)

**Supported by:** National Natural Science Foundation of China (61601213); Key Project of the Education Department of Liaoning Province (LJ212410147003)

solely from images, and they rely on manual feature extraction to reconstruct the stance. However, these methods suffer from low accuracy because they lack depth information and encounter issues with occlusion. The lack of depth information and the presence of human occlusion in monocular camera images still hinder the accurate estimation of the 3D position of an individual. Owing to the improved accuracy of 2D joint point detectors, the work may be divided into two stages. Initially, the task entails employing human 2D joint point detectors on video or RGB images to estimate the location of 2D joint points in the image space. In the second phase, the 2D joint points are transformed into 3D space by leveraging their positions. This work focuses on the second stage of the research, which involves using 2D joint point data to rebuild the 3D position of the human body. Recurrent neural network (RNN) has been utilized for human posture estimation with the successful application of deep learning in various other fields. RNNs can extract human postural variables that change in space and time. Moreover, RNNs perform more efficiently than other approaches. Nevertheless, the high computational cost of RNNs makes them impractical for predicting extended sequences. Consequently, Temporal convolutional networks (TCNs) are employed to estimate human posture. The TCN model has a reduced number of parameters and achieves a lower level of inaccuracy. However, TCN faces limitations in effectively combining spatial and temporal features and not fully utilizing information from 2D joints. Consequently, this study presents an estimation network for 3D human poses that integrates multi-joint features. **Method** The joint motion feature extraction module comprises four parallel branches, each extracting distinct joint motion features in the temporal dimension. By fusing these diverse features, the module forms a highly expressive feature representation that enhances the capability of the network to capture subtle motion changes. In estimating 3D human poses, the joint feature fusion module independently extracts temporally significant features for each joint set and utilizes matrix inner products to describe the relative spatial relationships between different joint groups. This method establishes overall inter-group connections, which ensures the accuracy and coherence of the estimated poses. The joint restraint module enhances the accuracy of the final predicted 3D pose and avoids the occurrence of unrealistic poses by using the spatial information of the 2D joints in the intermediate frames as global information and applying 2D-3D implicit constraints. **Result** In this study, we assess the accuracy of three public datasets to compare the proposed method with state-of-the-art models. Metric evaluation is based on the capability of the model to reconstruct 3D bit-postures. To assess the suitability of the proposed method, we conducted experiments on three separate datasets using the same network structure. The experimental findings on the Human3.6M dataset indicate that the current method yields a mean per joint position error (MPJPE) value of 29.0 mm. This value represents a 4.9% reduction in error when compared with the MHFormer method, which demonstrates the capability of the method to attain noteworthy outcomes and substantiates its efficacy. For complex actions such as SittingDown and WalkDog, the errors are reduced by 7.7% and 8.2%. The proposed method exhibits substantial improvements over the MHFormer approach on the 3DHP dataset in terms of evaluation metrics. Specifically, it obtains a reduction of 36.2% in the MPJPE, an increase of 12.9% in the area under curve (AUC) metric, and a noteworthy uplift of 3% in the percentage of correct keypoints (PCK) score. The experimental results demonstrate that, for complex actions, the network utilizes different branches to extract diverse temporal motion features of joints. This step enables the inference of joint positions based on contextual information and enhances spatiotemporal information interaction between joint groups using the positional information of other joints. By integrating the joint information from the current frame, the network constrains the estimated poses to be more realistic, which achieves superior spatiotemporal modeling. Despite these advancements, when benchmarked against state-of-the-art methods on the HumanEva dataset, our method maintains a strong competitive edge in the estimation of 3D human poses. Across every dataset tested, the present method consistently yields favorable outcomes, which substantiates its superior generalizability and applicability to diverse human pose datasets. Moreover, this study conducts thorough ablation experiments to validate the effectiveness of each module. The experiments demonstrate that each component indeed contributes to enhancing the overall accuracy of the network. Ablative studies on the length of input time series reveal that a longer input sequence corresponds to better performance of our method, which leads to increasingly accurate estimates of 3D human poses. **Conclusion** In this study, the experimental results and visual representations indicate that the proposed estimation model for human poses, which incorporates three integral modules, excels at building temporal and spatial dependencies between joints compared with other methodologies. This model facilitates a comprehensive representation of the joint space of the human body, particularly

effective when addressing the issue of self-occlusion. In scenarios where body parts are occluded, our method effectively utilizes the interdependencies within the context to alleviate the impact of occlusions on human pose estimation, which boosts the precision of the predictions. Even in situations where depth information is lacking, our approach consistently estimates human poses accurately by leveraging the spatial positional relationships among various joints.

**Key words:** 3D human pose estimation; human topology structure; multi-branch network; feature fusion; pose constraints

## 0 引言

三维人体姿态估计是计算机视觉领域的重要研究内容之一,主要利用图像或视频中提供的信息,重现人体在三维立体空间中的位置和动态变化。该技术应用领域广泛,如虚拟现实、动作识别以及运动分析等。传统方法(Belagianni等,2014)常依赖关节移动性假设,通过手动选择参数,从人体的2D投影中估计其3D姿态。由于缺少单目相机的深度信息,导致传统方法无法精确估计人体的3D姿态。随着深度学习技术的发展,采用卷积神经网络(Pavlakos等,2017)可以直接从RGB图像中估计人体3D姿态,无需中间监督过程,与传统方法相比取得了更准确的结果。由于直接从图像中预测人体3D姿态需要较高的计算成本,Tekin等人(2017)将该任务分解为两个步骤:1)采用2D姿态检测器,检测图像空间中的人体2D关节点位置;2)利用2D和3D人体结构间的对应关系,将2D关节点坐标映射到三维空间中。受益于现有2D关键点检测准确度的提升,近年来2D到3D的相关方法多数优于端到端的方法,为此本文重点研究该任务中的第2个步骤,目的是提高3D人体姿态预测的准确性。

在上述第2个步骤中,一些方法(Sun等,2017)使用少量的帧数甚至单帧信息估计人体3D姿态,这种方法忽略了输入序列的时间信息。一些方法利用时间信息解决深度模糊问题,有的方法直接对全部序列的姿态进行建模,如Hossain和Little(2018)、李书杰等人(2022)使用一种序列到序列的长短期记忆网络(long short-term memory, LSTM),但输入与输出姿态序列较长,结果会发生较大偏移。Lin和Lee(2019)利用矩阵分解来处理LSTM模型中存在的偏移问题,但该方法计算复杂度高。Zheng等人(2021)构建PoseFormer模型顺序捕获序列中所有关节点之间时间和空间的依赖性,会产生较大计算负担。Zhang等人(2023)同时估计所有输入帧的3D姿

态,建立了不同关节的全局依赖关系,但是缺少对局部运动的关注,并且模型计算量很大。但是相邻帧之间的动作变化较小,对输入进行全局建模会增大模型计算量。

因此,一些方法使用不同策略,对时间序列进行采样,如Liu等人(2020a)通过使用不同尺度的图注意力时空卷积,只关注动作变化大的输入,缺少对关节整体运动的关注。Wang等人(2020)使用UGCN(U-shaped graph convolutional network)来捕捉人体姿态在短时间和长时间的运动变化。Chen等人(2022)将姿态估计任务分解为预测骨骼长度和骨骼方向来提升时域卷积网络(temporal convolutional network, TCN)的性能。Shan等人(2022)为了解决时序序列冗余,在编码阶段对输入的部分序列进行遮掩操作,但相邻姿态之间高度相似,因此需要高的遮掩比,会降低模型的性能。Diaz-Arias和Shin(2024)提出一种ConvFormer架构,该架构通过引入动态多头卷积自注意力机制,在不同尺度上提取序列的运动特征,形成更丰富的特征表示。这些方法减轻了时间冗余的影响,但是各关节间的特征没有充分融合。

为提高不同关节之间的相互联系,减少不合理姿态的生成,Yan等人(2018)在人体姿态序列上构建一组时空图卷积,在相邻的关节间进行卷积,但缺少考虑不相邻关节间的关联性。此外,Liu等人(2020b)对中间层张量进行加权,强调了不同帧关节的时间相关性,忽略了中间帧的空间信息。Zeng等人(2020)提出一种分割重组网络,将人体分割成多个关节组,并在低纬度空间中将它们重新组合,但是关节组之间融合不充分。Lin等人(2021)利用Transformer编码器建模人体顶点—顶点和顶点—关节间的相互联系,通过自注意力机制捕捉人体关节的局部和全局依赖关系。Hu等人(2021)使用骨骼有向图表示法表示人体骨架的层次特性。Li等人(2022)提出一种MHFormer网络结构,使用空间变换编码器预测多个人体姿态,再将多个姿态聚合成最

终的3D姿态。Xue等人(2022)将具有相似运动模式的关节分组并计算组内关节运动的时间相关性。Li等人(2024)将人体关节的图结构对应到多层感知机(multilayer perceptron, MLP)模型中,融合关节的局部和全局空间特征,但未利用时序信息。这些方法缺少对中间帧2D关节信息的充分使用。薛峰等人(2024)利用2D骨骼的长度和方向信息估计不同关节旋转角对应的三维人体坐标。

为此,本文提出一种融合多关节特征的三维人体姿态估计网络。在关节运动特征提取模块,针对时间序列冗余,设计多分支架构跨步卷积,分间隔提取关节的运动特征,降低对冗余序列的重复使用。其次,针对变化较小的动作,每个分支捕获关节运动的不同特征,并将这些特征融合,增强网络对关节微小变化的敏感性。在关节特征融合模块,针对关节局部信息和全局信息融合不充分的问题,使用关节特征融合模块,每组关节融合其他组关节位置信息和关节的全局位置信息,建立关节局部信息和全局信息间的联系,减轻遮挡对网络影响。最后,在关节约束模块,针对预测出不合理姿态的问题,提取中间帧2D关节姿态的全局信息和约束信息,将约束信息与初始的预估值进行融合,进一步提高预测的准确性。

本文工作的主要贡献如下:1)针对输入时序序列冗余和人体微小动作变化难以捕捉的问题,提出一种融合多关节特征的网络,通过关节约束模块,进一步提升模型预测精度;2)提出一种关节运动特征提取模块,该模块采用了多分支跨步卷积并行提取特征的方法,通过结合不同分支的特征,丰富了关节运动变化的特征表示,增强捕捉小幅度运动变化的能力;3)提出一种关节特征融合模块结构,根据人体的拓扑结构,融合不同关节的姿态特征,建立不同关节间的联系,同时融合关节全局信息,实现对人体关节的整体姿态进行建模。

## 1 本文方法

### 1.1 网络整体架构

该网络共有两个分支,在第1个分支中,将中间帧的2D关节输入到关约束模块,获取中间帧关节整体的空间位置信息,并将其分别作为约束信息和全局信息,输入到网络的第2个分支中。在第2个分支中,将人体关节的二维坐标序列作为输入,将

其表示为  $K = \{k^j\}$ ,  $k^j = (x^j, y^j) \in \mathbb{R}^2$ ,  $j = 1, 2, \dots, J$ ,  $t = 1, 2, \dots, T$ 。其中,  $T$ 代表输入序列的长度,  $J$ 代表人体关节的序号。一些方法(Pavlo等, 2019)将关节的像素坐标作为输入,将其他关节相对根关节的位置坐标作为输出,未考虑输入和输出分布的一致性,降低了全局运动的鲁棒性。为解决该问题,对输入进行信息编码,通过使用其他关节的坐标减去根关节的坐标,对输入进行位置信息编码,即

$$K_p = \{k^j - k^0\}_{j=1}^J \quad (1)$$

式中,  $k^0$ 表示根关节的2D坐标,  $k^j$ 表示所有关节的2D坐标。位置信息编码使网络仅捕获与人体姿势相关的有意义空间信息,有利于增强网络对全局运动的鲁棒性。使用二维关节序列估计中间帧3D姿态,其他关节姿态在时域连续性方面相对于中间帧姿态可起到辅助作用。为加强表示中间帧与其他帧姿势之间的相对位置关系,对输入进行时间信息编码,可以表示为

$$K_T = \left\{ k_t - k_{\frac{T}{2}} \right\}_{t=1}^T \quad (2)$$

式中,  $k_{\frac{T}{2}}$ 表示中间帧的2D关节坐标,  $k_t$ 表示所有帧的2D关节坐标。时间信息编码可使网络更专注于中间帧的周围关节位置变化,而不是每个姿势的绝对位置。当运动范围较小的局部运动发生时,这些变化将被放大,有助于3D姿态的高精度建模。关节相对关系图如图1(a)所示,红色虚线表示中间帧关节在时间维度上的运动变化,褐色虚线表示根关节相对其他关节的空间位置关系。将原始输入、位置信息编码和时间信息编码进行拼接操作,使模型可以同时学习不同类型的特征,该过程可以表示为

$$K_e = f_{\text{Concat}}(K, K_p, K_T) \quad (3)$$

式中,  $K_e$ 表示最终的信息增强编码,将人体关节进行分组,结果如图1(b)所示。

$$\begin{aligned} K^0 &= \{\text{hip, spine, throat, neck, head}\} \\ K^1 &= \{\text{right\_hip, right\_knee, right\_feet}\} \\ K^2 &= \{\text{left\_hip, left\_knee, left\_feet}\} \\ K^3 &= \{\text{right\_shoulder, right\_elbow, right\_wrist}\} \\ K^4 &= \{\text{left\_shoulder, left\_elbow, left\_wrist}\} \end{aligned} \quad (4)$$

根据人体的动态链结构将身体分为5部分,分别是躯干、右腿、左腿、右臂和左臂,表示为  $K^i =$

$\{k_i^j\}_{j=1}^{J_i}$ , 其中,  $i$  表示关节组数,  $J_i$  表示第  $i$  组中的关节点数量。将分组后的各关节组并行输入到关节运动特征提取模块。在该模块中, 建模每个分组内关节运动轨迹的时序特征, 并将得到的时序特征并行输入到关节特征融合模块, 融合当前组特征、非当前组特征和全局信息, 在该模块中显式地学习不同关节组之间的相对空间位置关系, 使所有分组产生联系, 并将融合后的特征与约束信息相加得到最终输出, 表示为  $P = \{p_i^j\}$ ,  $p_i^j = (X_i^j, Y_i^j, Z_i^j) \in \mathbb{R}^3$ , 整体网络架构如图2所示。

### 1.2 关节运动特征提取模块

传统时间卷积模型由多个时间卷积块串联组成, 在提取关节运动特征中发挥着重要作用(Pavlo等, 2019), 但该结构在提取关节时序变化方面能力较弱。为使网络更好地学习人体姿态运动轨迹在时间维度上的表示, 从有限的输入序列中提取人体关节在时间维度上更为丰富的运动特征, 关节运动

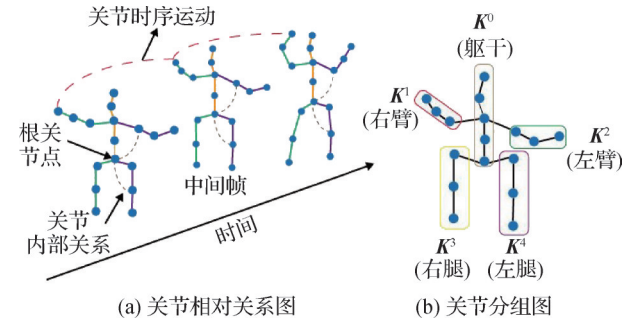


图1 关节关系图和分组图

Fig. 1 Joint relation map and group map

((a) joint relation map; (b) joint group map)

特征提取模块采用多分支的架构, 每个分支提取关节不同的运动特征。该模块的输入是经信息增强分组后的关节组, 使用卷积操作的优势是可以依据卷积核的大小, 提取关节在不同时间上的运动变化特征, 保持关节运动变化的一致性。由于输入序列存在冗余, 相邻帧之间人体的动作变化较小, 因此使用平均池化操作, 该操作可以加强模块

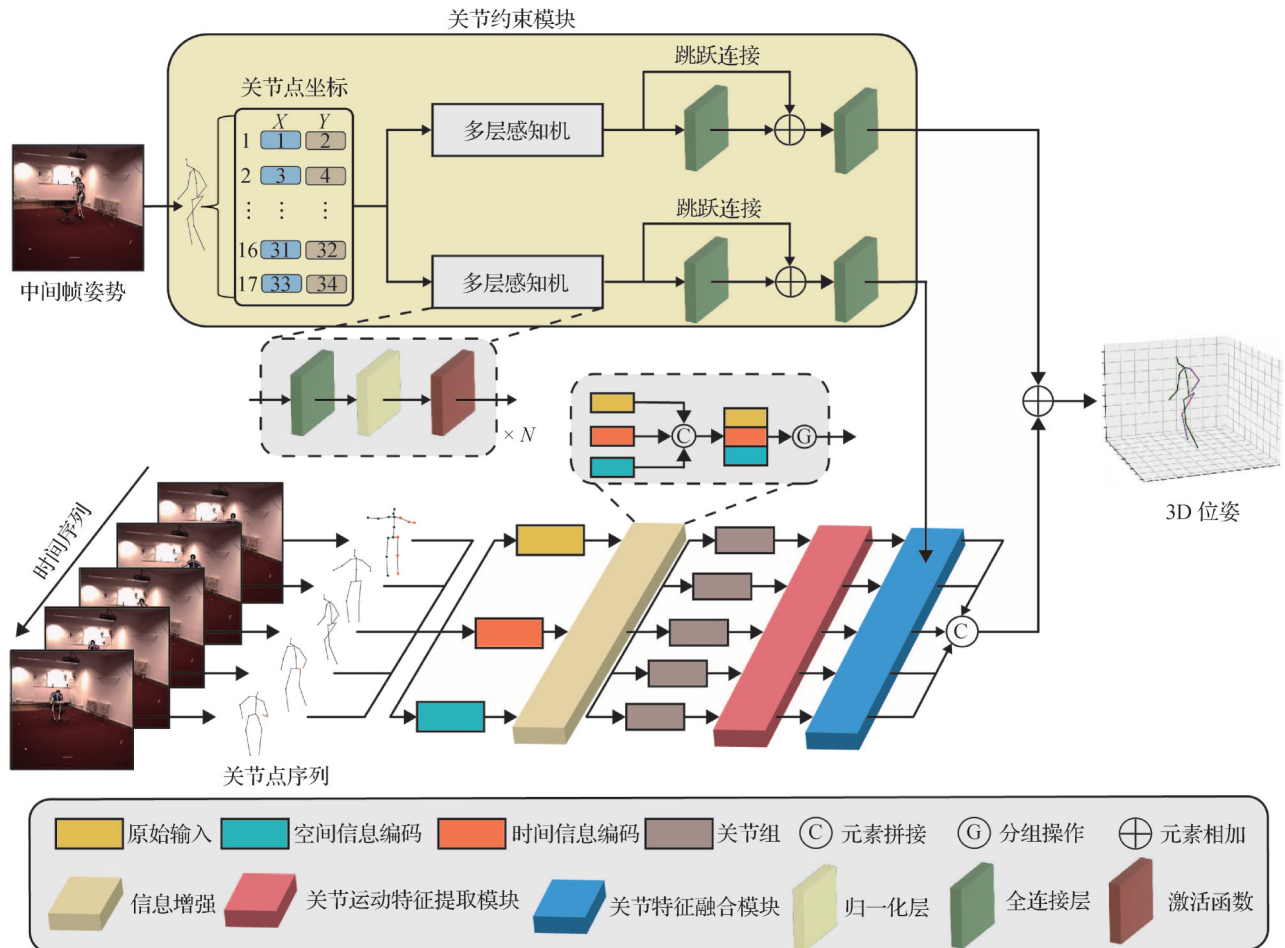


图2 整体网络框架

Fig. 2 Global network framework

对关节动作微小变化的捕捉能力。最后将3个分支得到的特征进行融合,增强最终特征对关节运动变化的表征能力,关节运动特征提取模块的整体架构如图3所示。

在关节运动特征提取模块中,第1个分支中的输入经过 $1 \times 1$ 卷积,将关节点的空间关系转化为向量空间中的几何关系,更好地表达了不同关节间的运动相似性和差异性,该过程表示为

$$f_1 = Bn(Conv(K^i, W_{1 \times 1})) \quad (5)$$

经平均池化层获取相邻时间关节变化的空间聚合信息,更好地捕捉关节随时间的细微变化,该过程表示为

$$b_1 = Bn(Avg(f_1, S_{1 \times 3})) \quad (6)$$

式(5)和式(6)中, $K^i$ 表示输入的关节组, $f_1$ 表示输入经过卷积和归一化操作后的特征, $b_1$ 表示第1个分支的特征输出结果, $Avg(\cdot)$ 表示平均池化操作, $S_{1 \times 3}$ 代表池化窗口大小为 $1 \times 3$ ,步长为3。

第2个分支为传统的时间卷积操作,可表示为

$$b_2 = Bn(Conv(K^i, W_{1 \times 3})) \quad (7)$$

式中, $b_2$ 表示第2个分支的特征输出结果。第3个分支中, $1 \times 3$ 卷积的作用是逐步缩小序列长度并聚合不同时间同一关节的变化信息,该过程表示为

$$f_2 = Bn(Conv(K^i, W_{1 \times 1})) \quad (8)$$

$$b_3 = Bn(Conv(f_2, W_{1 \times 3})) \quad (9)$$

式(8)和式(9)中, $f_2$ 表示输入经过卷积和归一化操

作后的特征, $b_3$ 表示第3个分支的特征输出结果。 $Conv(\cdot)$ 表示卷积操作, $Bn(\cdot)$ 表示归一化操作, $W_{1 \times 1}$ 代表卷积核的大小为 $1 \times 1$ , $W_{1 \times 3}$ 表示卷积核的大小为 $1 \times 3$ 。

将以上3个分支得到的特征进行融合,该过程可以表示为

$$b_4 = b_1 \oplus b_2 \oplus b_3 \quad (10)$$

式中, $b_4$ 表示融合后的特征, $\oplus$ 表示相加操作。将融合后的特征进行多次跨步卷积,具体为

$$F_B = Conv(Conv(b_4, W_{1 \times 3}), W_{1 \times 1}) \quad (11)$$

式中, $F_B$ 表示得到的最终特征,分层次地提取特征,可使模型更深的层能够根据前期特征,学习人体姿势变化更为复杂、时间范围更长的特征,让模型更好地学习姿态序列在时间维度中的变化规律。

### 1.3 关节特征融合模块

关节分组后,每个组均分别提取到具有时间意义的关节特征,但是各组间的关系相对独立,在推断人体3D姿态时,当前组关节与其他组关节的空间相对位置未知,不利于保持整体身体姿势的一致性。理想情况下,网络应保证每组关节间的连续性。因此,本文提出如下的关节特征融合模块,可将其他组的关节信息及全局信息融入到当前组中,其架构如图4所示。

首先,将其他组的关节特征在通道维度上拼接,经全连接层融合这4组关节特征,建立非当前组关节间的空间位置联系,该过程可表示为

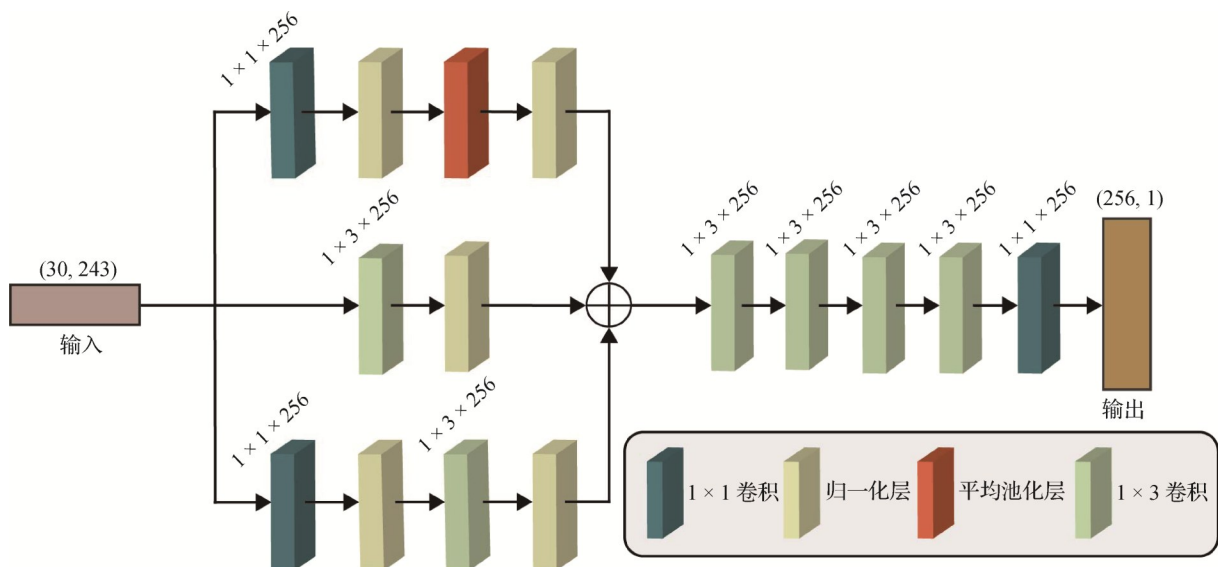


图3 关节运动特征提取模块

Fig. 3 Joint motion feature extraction module

$$f_{m1} = \text{Ln}\left(f_{\text{Concat}}\left(F_B^2, F_B^3, F_B^4, F_B^5\right)\right) \quad (12)$$

式中,  $f_{m1}$  表示非当前组关节融合后的特征,  $F_B^2, F_B^3, F_B^4, F_B^5$  表示4组关节组的特征,  $f_{\text{Concat}}(\cdot)$  表示拼接操作,  $\text{Ln}(\cdot)$  表示全连接层。

将当前关节组与非当前关节组的特征共同输入特征融合模块中进行融合, 得到当前关节组与非当前关节组之间的空间位置关系, 具体为

$$f_{m2} = \text{FFM}\left(F_B^1, f_{m1}\right) \quad (13)$$

再将融合后的特征与全局信息同时输入特征融合模块, 得到当前组关节在整体姿态中的高维空间位置, 再输入到全连接层, 该过程表示为

$$F_J = \text{Ln}\left(\text{FFM}\left(f_g, f_{m2}\right)\right) \quad (14)$$

式(13)和式(14)中,  $F_B^1$  表示当前组关节特征,  $f_{m2}$  表示当前组与非当前组融合后的特征,  $f_g$  表示全局信息,  $F_J$  表示关节特征融合模块最终的输出特征,  $\text{FFM}(\cdot)$  表示特征融合操作。特征融合模块如图5

所示, 输入为包含关节位置信息的特征, 空间位置矩阵的计算过程可以表述为

$$\begin{cases} y_1 = \text{Lr}\left(\text{Bn}\left(\text{Ln}\left(l_1\right)\right)\right) \\ s_1 = l_1 \oplus \text{Lr}\left(\text{Bn}\left(\text{Ln}\left(y_1\right)\right)\right) \end{cases} \quad (15)$$

信息补充矩阵的计算过程可以表示为

$$\begin{cases} y_2 = \text{Lr}\left(\text{Bn}\left(\text{Ln}\left(l_2\right)\right)\right) \\ s_2 = \text{Ln}\left(l_2 \oplus \text{Lr}\left(\text{Bn}\left(\text{Ln}\left(y_2\right)\right)\right)\right) \end{cases} \quad (16)$$

式(15)和式(16)中,  $l_1$  和  $l_2$  表示输入,  $s_1$  表示空间位置矩阵,  $s_2$  表示信息补充矩阵,  $\text{Lr}(\cdot)$  表示 LeakyReLU (leaky rectified linear unit) 激活函数。

通过矩阵相乘操作提高信息间的互补性, 同时可以更好地获得权重来调整输入关节组在高纬度空间中的位置, 显式表达特征间相对空间位置关系, 该过程可表示为

$$S = \text{Ln}\left(s_1 \odot s_2\right) \quad (17)$$

式中,  $S$  表示特征矩阵,  $\odot$  表示矩阵相乘。

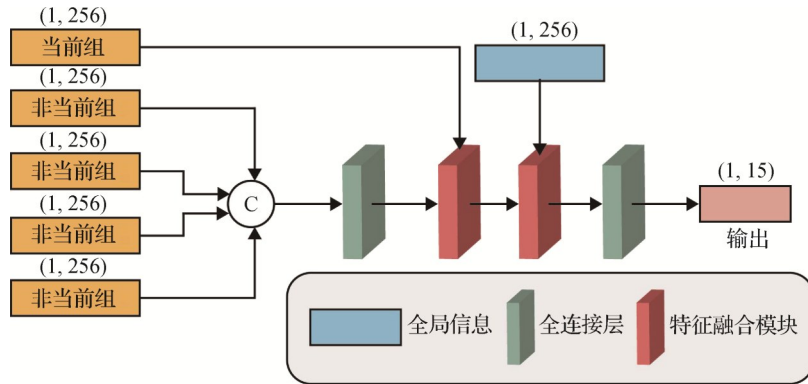


图4 关节特征融合模块

Fig. 4 Joint feature fusion module

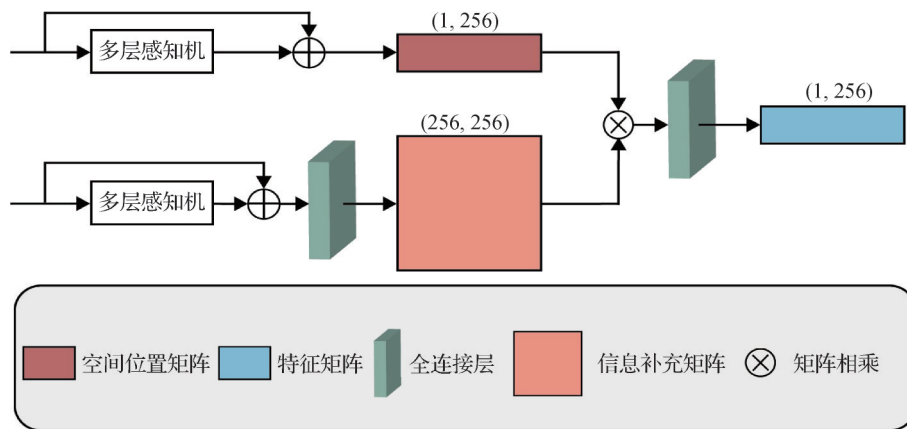


图5 特征融合模块

Fig. 5 Feature fusion module

### 1.4 关节约束模块

大多方法通过输入 2D 关节点序列估计中间帧 3D 姿态,主要关注 3D 姿态假设生成,忽略了中间帧的 2D 姿态与 3D 姿态之间的对应关系。本模块的输入为中间帧 2D 关节点信息(表示为  $K_c \in \mathbf{R}^2$ ),将每个关节点通过线性变换映射到高维空间,这些高维向量包含整体姿态的全局关系,该过程表示为

$$f_g = \delta(Bn(Ln(K_c))) \quad (18)$$

将这些信息作为全局信息输入到关节特征融合模块中,指导关节特征融合模块中的关节组建立更好的全局位置关系。其中关节的 2D 姿态还包含关节间的位置关系,一些关节间存在间接联系(如肩部和膝盖),一些关节间则存在直接联系(如肘部和手腕),各关节间的关联关系对整体姿态起着约束作用。因此通过注意力机制,建立可学习的参数矩阵,自适应地建立各关节之间的关联关系,该过程表示为

$$f_r = Ln(r_1 \oplus Ln(r_1)) \quad (19)$$

式(18)和式(19)中, $\delta(\cdot)$ 为 ReLU 激活函数, $f_r$ 表示约束信息,最后将其作为约束信息与初估值融合,生成更合理的 3D 姿态。

### 1.5 损失函数

损失函数具体表示为预测关节点与真值关节点二者坐标的均方误差,具体为

$$L_{MPJPE} = \frac{1}{T \cdot J} \sum_{t=1}^T \sum_{j=1}^J \|p_t^j - \tilde{p}_t^j\|_2 \quad (20)$$

式中, $T$ 表示全部的输入帧数, $J$ 表示关节点数, $p_t^j$ 表示 3D 姿态的真值, $\tilde{p}_t^j$ 表示网络预测的 3D 姿态坐标。

## 2 实验

### 2.1 数据集与评价指标

本文采用 3 个公共数据集进行训练和评估,为保证与相关方法实验结果比较的一致性,在所有数据集的训练和测试集的划分上均保持一致。

Human3.6M(Ionescu 等,2014)是人体 3D 姿态估计中最大、使用最广的室内数据集,共有 11 个对象,分为 15 个动作,如讨论、拍照、坐立、打电话等,含有 360 万帧视频图像。本文选取 5 个对象(S1,S5,S6,S7,S8)作为训练集,2 个对象(S9,S11)作为测试集。

HumanEva(Sigal 等,2010)数据集较小,从 3 个 60 Hz 的摄像机视图中记录了 3 个对象和 3 个动作

(步行、慢跑、箱子)。本文选取 3 个对象(S1,S2,S3)中的 2 个动作(步行、慢跑)。本实验在 Human3.6M 和 HumanEva 数据集上使用的评估指标为:指标 1 为平均关节位置误差(mean per joint position error, MPJPE),表示所有时间序列上预测关节点与真值关节点二者坐标欧氏距离的平均值;指标 2 为重建平均关节位置误差(pose-processing MPJPE, P-MPJPE),表示经过刚性变换和根关节对齐后关节位置的平均误差。两个指标单位均为毫米(mm)。

MPI-INF-3DHP(Mehta 等,2017a)是一个大规模数据集,含有 130 万帧视频图像,由绿屏、非绿屏和室外 3 个场景组成。通过使用 14 个摄像头,数据集记录了 8 名参与者完成 8 项训练活动和 7 项评估活动。MPI-INF-3DHP 将 2D 人体姿态估计中广泛使用的正确关节点百分比(percentage of correct keypoints, PCK)和曲线下面积(area under the curve, AUC)扩展到了 3D 人体姿态估计。因此,使用 MPJPE、3D PCK 和 AUC 作为评价指标,3D PCK 表示预测正确关节点所占的百分比,当预测关节点与真实关节点的距离小于阈值(通常设置为 150 mm),则为预测正确。AUC 表示 ROC 曲线(receiver operating characteristic curve)下与坐标轴围成的面积,具体为

$$AUC = \int_0^1 ROC(x) dx \quad (21)$$

$$PCK = \frac{\sum_i \wp\left(\frac{d_i}{d} \leq T_k\right)}{\sum_i 1} \quad (22)$$

式中, $T_k$ 表示阈值, $i$ 表示关节点编号, $d_i$ 表示第  $i$  个关节点的预测值和真值的欧氏距离, $d$ 表示一个人体的尺度因子, $\wp(\cdot)$ 表示 sigmoid 函数。

### 2.2 实验环境

实验使用的操作系统为 Ubuntu18.04,深度学习框架为 PyTorch1.10.1,cuda10.2,选用 Lion 优化器(Chen 等,2023)。硬件上使用 RTX 1080Ti GPU 进行训练和推理。在超参数的设置上,训练轮数为 90 轮,初始学习率设定为  $1 \times 10^{-4}$ ,学习衰减率为 0.95,批次大小为 256。

### 2.3 实验结果与分析

本文在 3 个数据集上的实验结果如表 1—表 4 所示,表中所有对比方法结果均与原文保持一致。为验证本文结果的可靠性,在 3 个数据集上重复 5 次实验,每次实验结果接近。如图 6 所示,在

Human3.6M 数据集上的结果平均值为 29.04 mm、方差为 3.71 mm;在 HumanEva 数据集上的结果平均值为 20.42 mm、方差为 12.36 mm;在 MPI-INF-3DHP 数据集上的结果平均值为 37.01 mm、方差为 3.49 mm。

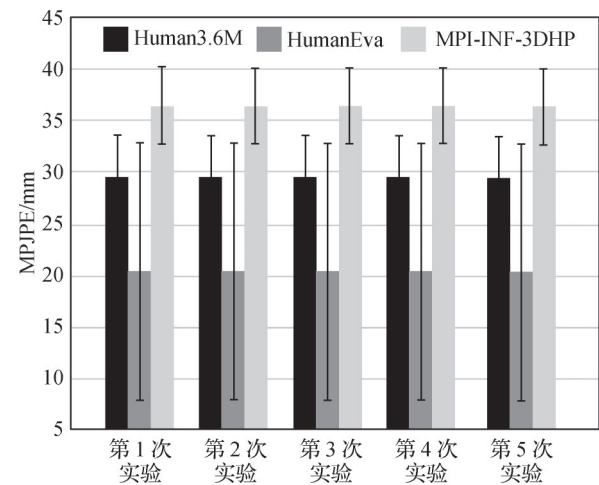


图 6 平均值和方差  
Fig. 6 Mean and variance plots

2.3.1 定量分析

在 Human3.6M 数据集的实验结果如表 1 所示。本文方法与现有方法相比,在 15 项运动中 8 个动作在指标上获得最优值,1 个动作在指标上达到次优,尤其在出现自遮挡和小幅度动作时(如 Sitting 和 WalkDog),本文方法相比次优方法在 Sitting 动作上

的误差降低 2.3%、在 WalkDog 上误差降低 3.9%,优于 P-STMO 序列遮蔽策略和 ConvFormer 的动态多头注意力机制。由于注意力机制重点关注运动变化较大的姿态,导致其忽视小幅度运动的变化,而本文的关节运动特征提取模块的各分支平等地提取了关节运动不同的时序信息,保留了小幅度运动变化的特征,最后融合不同分支提取的关节变化特征,增强特征表达能力,因此在面对动作小幅度变化问题时,网络可以更好地捕捉小幅度运动变化。跨步卷积的设计可以利用被遮挡帧的相邻帧信息,推断被遮挡关节的位置。一些研究工作(李书杰等,2022)和 MHFormer 在关节间的信息交互不足,本文关节特征融合模块聚合了非当前组关节特征,通过关节特征融合模块增强当前组关节与不同关节组间的时空信息交互,可以使每个关节组特征关联其他关节组信息。当出现遮挡问题时,可以利用未被遮挡的关节特征预测被遮挡关节的信息。关节分组方法(Zeng 等,2020;Xue 等,2022)只关注关节的局部运动,缺少对关节全局信息的交互,本文中的关节运动特征提取模块可以提取当前帧 2D 关节点相对位置关系和关节间的拓扑关系特征,获得关节的全局特征和约束特征。将全局特征输入到关节特征融合模块中,建立局部关节和全局姿态的联系,通过约束信息与初估值融合约束预估姿态生成更符合实际的姿态,因此本文方法可以获得较好的评价指标值。

表 1 在 Human3.6M 数据集上以 2D 姿态的真实值作为输入的 MPJPE 结果  
Table 1 Results of MPJPE on Human3.6M dataset with the ground truth of 2D poses as input

| 方法                | Dir.        | Disc.       | Eat.        | Greet.      | Phone       | Photo       | Pose        | Purch.      | Sit         | SitD.       | Smoke       | Wait        | WalkD.      | Walk        | WalkT.      | 平均          |
|-------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| PoseFormer        | 30.0        | 33.6        | 29.9        | 31.0        | 30.2        | 33.3        | 34.8        | 31.4        | 37.8        | 38.6        | 31.7        | 31.5        | 29.0        | 23.3        | 23.1        | 31.3        |
| Ray3D(Zhang,2022) | 31.2        | 35.7        | 31.4        | 33.6        | 35.0        | 37.5        | 37.2        | 30.9        | 42.5        | 41.3        | 34.6        | 36.5        | 32.0        | 27.7        | 28.9        | 34.4        |
| MHFormer          | <u>27.7</u> | 32.1        | 29.1        | 28.9        | 30.0        | 33.9        | 33.0        | 31.2        | 37.0        | 39.3        | 30.0        | 31.0        | 29.4        | 22.2        | 23.0        | 30.5        |
| P-STMO            | 28.5        | <u>30.1</u> | 28.6        | <b>27.9</b> | <u>29.8</u> | <u>33.2</u> | <u>31.3</u> | <u>27.8</u> | 36.0        | 37.4        | <b>29.7</b> | <u>29.5</u> | <u>28.1</u> | <b>21.0</b> | <b>21.0</b> | <u>29.3</u> |
| Li 等人(2023a)      | 32.9        | 38.3        | 28.3        | 33.8        | 34.9        | 38.7        | 37.2        | 30.7        | <u>34.5</u> | 39.7        | 33.9        | 34.7        | 34.3        | 26.1        | 28.9        | 33.8        |
| Kang 等人(2023)     | 30.7        | 35.6        | <u>26.8</u> | 32.0        | 32.6        | 36.7        | 36.0        | 30.0        | 35.1        | <u>36.5</u> | 32.2        | 33.6        | 33.9        | 26.5        | 27.5        | 32.4        |
| Zhang 等人(2023)    | <b>27.5</b> | 31.1        | 30.1        | 29.5        | 30.4        | 34.6        | 32.9        | 30.2        | 37.1        | 38.9        | 30.8        | 30.2        | 29.6        | 22.2        | 22.9        | 30.5        |
| Chen 等人(2024)     | 35.7        | 39.4        | 35.4        | 37.8        | 41.3        | 48.1        | 39.0        | 37.9        | 49.8        | 52.8        | 41.1        | 39.0        | 39.4        | 27.3        | 27.2        | 39.5        |
| ConvFormer        | 28.9        | 31.8        | 28.0        | <u>28.2</u> | <b>29.5</b> | <b>33.0</b> | 32.9        | 30.1        | 36.8        | 37.4        | <u>29.8</u> | 29.6        | 28.2        | <u>21.7</u> | <u>21.5</u> | 29.8        |
| 本文                | 28.2        | <b>28.2</b> | <b>26.6</b> | 28.3        | 29.9        | 34.7        | <b>30.6</b> | <b>26.6</b> | <b>33.7</b> | <b>36.3</b> | <u>29.8</u> | <b>29.4</b> | <b>27.0</b> | 22.6        | 22.9        | <b>29.0</b> |

注:加粗、下划线字体表示各列最优、次优结果。

表2和表3给出了使用人体2D姿态的检测值作为输入,以MPJPE和P-MPJPE的值作为评价指标的结果。采用人体2D姿态检测器级联金字塔网络(cascaded pyramid network, CPN)检测后,得到人体2D姿态检测值,与人体2D姿态真值相比,两者之间存在误差,且检测值中存在偏差较大的噪声点,导致预测与真实值间的误差增大。由结果可见,本文方法与对比方法相比,在MPJPE和P-MPJPE两个指标上均获得了最好结果,分别为45.0 mm和35.4 mm。在MPJPE指标上有7个动作达到最优、4个动作达到次优。在P-MPJPE指标上有6个动作达到最优、5个

动作达到次优,说明本文方法即便使用2D姿态检测值作为输入,仍然可以取得较好的结果,证明了本文方法的鲁棒性,因此本文方法可以与现有人体2D姿态检测器较好地搭配使用。

在HumanEva数据集的测试中,网络输入为经2D姿态检测器检测的结果,采用P-MPJPE作为评价指标,结果如表4所示。可见,本文方法在两名受试者(S2、S3)的Jog动作上指标达到最优。在一名受试者S2的Walk动作上指标达到了次优,验证了本文方法在不同数据集上,以2D姿态检测值作为输入的条件下依然能够保持较高的鲁棒性。

表2 在Human3.6M数据集上以经过CPN检测的2D姿态值作为输入的MPJPE结果  
Table 2 Results of MPJPE on Human3.6M dataset with the 2D poses detected by CPN as input

| 方法                     | /mm         |             |             |             |             |             |             |             |             |             |             |             |             |             |             |             |
|------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                        | Dir.        | Disc.       | Eat.        | Greet.      | Phone       | Photo       | Pose        | Purch.      | Sit         | SitD.       | Smoke       | Wait        | WalkD.      | Walk        | WalkT.      | 平均          |
| VideoPose3D            | 45.2        | 46.7        | 43.3        | 45.6        | 48.1        | <u>55.1</u> | 44.6        | 44.3        | 57.3        | 65.8        | 47.1        | 44.0        | 49.0        | 32.8        | 33.9        | <u>46.8</u> |
| MHFormer<br>(Li等,2022) | 43.7        | 49.3        | 45.5        | 47.9        | 50.5        | 56.0        | 46.3        | 44.1        | 55.9        | <u>59.0</u> | 48.4        | 45.7        | 51.2        | 37.1        | 39.1        | 48.0        |
| Ray3D(Zhang,2022)      | 44.7        | 48.7        | 48.7        | 48.4        | 51.0        | 59.9        | 46.8        | 46.9        | 58.7        | <b>31.7</b> | 50.2        | 46.4        | 51.5        | 38.6        | 41.8        | 49.7        |
| HTNet(Cai等,2023)       | 44.0        | 49.0        | 45.2        | 47.8        | 50.3        | 56.6        | 47.2        | 45.0        | 56.3        | 65.8        | 48.7        | 46.0        | 52.4        | 39.1        | 40.2        | 48.9        |
| Kang等人(2023)           | <u>42.2</u> | 47.3        | 44.6        | 47.6        | 49.7        | 56.1        | 45.3        | 43.8        | <u>55.3</u> | 59.4        | <u>45.5</u> | 44.7        | 51.0        | 36.8        | 38.3        | 47.4        |
| Einfal等人(2023)         | 44.5        | <u>45.1</u> | <b>40.3</b> | <u>44.6</u> | <b>46.3</b> | <b>50.7</b> | <u>44.4</u> | <u>43.7</u> | <b>54.6</b> | 62.3        | <b>44.9</b> | <b>43.1</b> | <u>47.0</u> | <u>32.3</u> | <b>31.9</b> | <b>45.0</b> |
| 本文                     | <b>41.0</b> | <b>44.2</b> | <u>42.4</u> | <b>42.6</b> | <u>46.8</u> | 57.2        | <b>42.5</b> | <b>41.4</b> | 55.4        | 61.9        | 46.2        | <u>43.2</u> | <b>45.6</b> | <b>31.5</b> | <u>32.7</u> | <b>45.0</b> |

注:加粗、下划线字体表示各列最优、次优结果。

表3 在Human3.6M数据集上以经过CPN检测的2D姿态值作为输入的P-MPJPE结果  
Table 3 Results of P-MPJPE on Human3.6M dataset with the 2D poses detected by CPN as input

| 方法                     | /mm         |             |             |             |             |             |             |             |             |             |             |             |             |             |             |             |
|------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                        | Dir.        | Disc.       | Eat.        | Greet.      | Phone       | Photo       | Pose        | Purch.      | Sit         | SitD.       | Smoke       | Wait        | WalkD.      | Walk        | WalkT.      | 平均          |
| VideoPose3D            | 34.1        | <u>36.1</u> | 34.4        | 37.2        | 36.4        | 42.2        | 34.4        | 33.6        | 45.0        | 52.5        | 37.4        | 33.8        | 37.8        | 25.6        | 27.3        | 36.5        |
| Chen等人(2022)           | <u>33.1</u> | <b>35.3</b> | 33.4        | <u>35.9</u> | <u>36.1</u> | 41.7        | <b>32.8</b> | <u>33.3</u> | <b>42.6</b> | 49.4        | 37.0        | <b>32.7</b> | <u>36.5</u> | <u>25.5</u> | 27.9        | <u>35.6</u> |
| MHFormer<br>(Li等,2022) | 35.1        | 38.2        | 36.5        | 39.8        | 39.8        | 43.5        | 35.7        | 34.0        | 45.6        | <u>47.6</u> | 39.8        | 35.41       | 41.1        | 30.0        | 33.4        | 38.4        |
| HTNet(Cai等,2023)       | 35.1        | 38.6        | 36.6        | 39.4        | 43.8        | <b>36.7</b> | 34.7        | 45.3        | 52.8        | <b>39.9</b> | <b>35.4</b> | 35.4        | 41.8        | 31.2        | 33.5        | 39.0        |
| Kang等人(2023)           | 34.7        | 37.8        | 35.4        | 39.8        | 39.4        | 44.1        | 36.2        | 34.2        | 44.5        | 48.6        | 39.9        | 34.9        | 41.1        | 29.5        | 32.5        | 38.2        |
| Einfal等人(2023)         | 36.4        | 36.5        | <b>33.0</b> | 37.1        | 36.4        | <u>40.1</u> | 34.8        | 34.3        | 45.0        | 50.1        | <u>36.9</u> | 34.2        | 37.9        | 26.5        | <b>25.8</b> | 36.3        |
| 本文                     | <b>32.9</b> | 36.2        | 33.8        | <b>35.3</b> | <b>35.7</b> | 42.7        | <u>33.3</u> | <b>31.7</b> | <u>43.5</u> | 48.5        | 37.0        | <u>33.3</u> | <b>35.5</b> | <b>24.7</b> | <u>25.9</u> | <b>35.4</b> |

注:加粗、下划线字体表示各列最优、次优结果。

为了在3DHP数据集上验证本文方法估计三维人体姿态的准确性,采用PCK、AUC和MPJPE值作为评价指标,结果如表5所示。可见,本文方法得到的PCK指标为96.7%,AUC指标为72.7%,MPJPE指标为37.0 mm,相比次优结果,在MPJPE上降低了

12.9%,在AUC上提升4.1%,结果验证了本文方法具有较高的鲁棒性与泛化能力。

2.3.2 可视化定性比较

图7(a)给出了本文方法与Li等人(2024)方法在Human3.6M数据集上识别17个关节点的可视化

表 4 HumanEva数据集的结果  
Table 4 The results of HumanEva dataset

|                        |             |             |             |             |             |             | /mm         |
|------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| 方法                     | Walk        |             |             | Jog         |             |             | 平均          |
|                        | S1          | S2          | S3          | S1          | S2          | S3          |             |
| PoseFormer             | 16.3        | 11.0        | 47.1        | 25.0        | 15.2        | 15.1        | 21.6        |
| MixSTE (Zhang 等, 2022) | <b>12.7</b> | 10.9        | <b>17.6</b> | <u>22.6</u> | 15.8        | 17.0        | <b>16.1</b> |
| Li 等人 (2023b)          | <u>14.0</u> | <b>10.0</b> | <u>32.8</u> | <b>19.5</b> | <u>13.6</u> | <u>14.2</u> | <u>18.9</u> |
| 本文                     | 14.8        | <u>10.5</u> | 46.7        | 23.4        | <b>13.5</b> | <b>14.0</b> | 20.4        |

注:加粗、下划线字体表示各列最优、次优结果。

表 5 MPI-INF-3DHP数据集的结果  
Table 5 The results of MPI-INF-3DHP dataset

| 方法                             | PCK<br>/% ↑ | AUC<br>/% ↑ | MPJPE<br>/mm ↓ |
|--------------------------------|-------------|-------------|----------------|
| Hu 等人 (2021)                   | <b>97.9</b> | 69.5        | <u>42.5</u>    |
| CrossFormer (Hassanin 等, 2022) | 89.1        | 57.5        | 76.3           |
| MHFormer (Li 等, 2022)          | 93.8        | 63.3        | 58.0           |
| MixSTE (Zhang 等, 2022)         | 94.4        | 66.5        | 54.9           |
| Einfal 等人 (2023)               | 95.4        | 67.6        | 46.9           |
| ConvFormer                     | 96.4        | <u>69.8</u> | 53.6           |
| 本文                             | <u>96.7</u> | <b>72.7</b> | <b>37.0</b>    |

注:加粗、下划线字体表示各列最优、次优结果。“↑”和“↓”分别表示值越高越好和值越低越好。

结果。图像中实线代表了三维人体姿态的预测值,其中黑色实线表示人体的左侧,红色实线表示人体的右侧。虚线代表了人体姿态的真实值,其中绿色虚线表示人体的左侧,蓝色虚线表示人体的右侧。本文选取了4个代表性的动作进行展示,分别是打电话、指路、坐和遛狗。4个动作中均出现了人体被遮挡的情况,例如在图像第2行的指路动作中,受试者做出了侧身动作,人体的右侧手肘部分被自身遮挡严重,从可视化结果来看,本文方法预测的手肘倾斜角度与真实值最为贴近,其余方法都存在较大误差。在图像第4行的遛狗动作中,部分关节在空间中的位置相近,并且随时间变化的运动幅度较小,从可视化结果来看,本文方法的预测值与真实值更为贴近,实现了较为准确的人体姿态建模。

图7(b)展示了本文方法在3DHP数据集上可视化的结果,从结果可以看出,在受试者展示的不同动作中,面对出现的(如图中第1行人体自遮挡)问题,本文方法都能估计出较为准确的姿态。

对多个不同人体数据集进行严谨的定量评估与定性对比,结果显示,本文方法可以适应多种人体数据集,并在各数据集上均显示出相对于对比方法较为明显的优势。为充分验证本文方法的泛化能力,在真实场景下得到的实验结果如图8所示。可见,本文方法可以较好地应用在多种现实场景中。

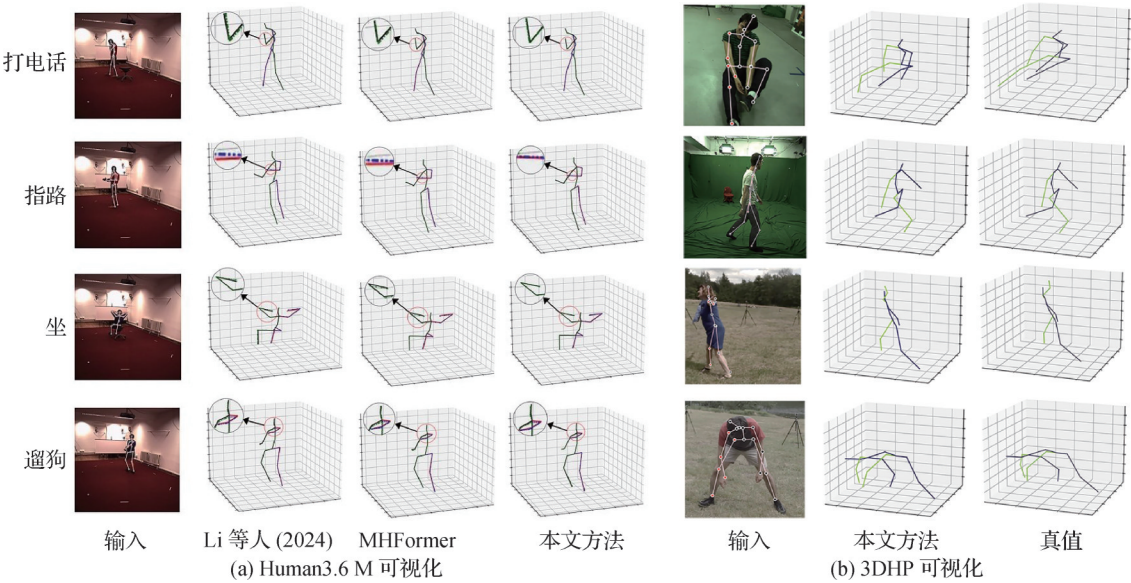


图7 可视化结果

Fig. 7 Visual results ((a) Human3.6M visualization; (b) 3DHP visualization)

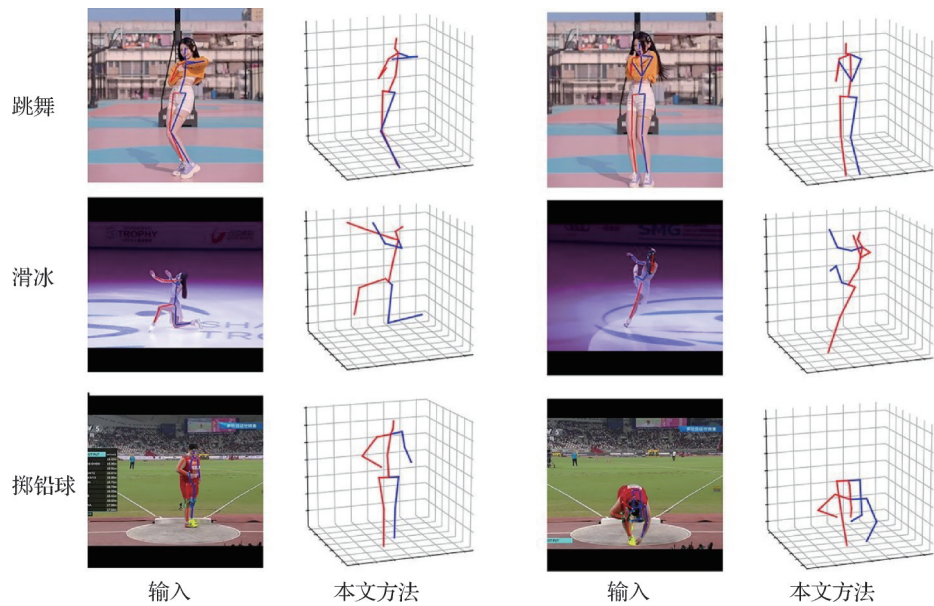


图 8 真实场景可视化  
Fig. 8 Real scene visualization

2.4 收敛速度与优化器对比

为了验证本文方法的收敛性,在 Huaman3. 6M 数据集上将本文方法与其他方法进行对比,结果如图 9 所示。纵轴为式(20)的预测关节点与真值关节点二者坐标的均方误差,横轴为训练时间,单位为小时。与其他方法相比,本文方法(红色曲线)在相同时间,误差下降更快,最终误差更低。在 Human3. 6M 数据集上对 Lion 优化器与 Adam 优化器性能进行比较,结果如图 10 所示,图中横轴表示时间,纵轴表示误差,由图可见,Lion 优化器收敛速度更快。

2.5 模型对比分析

如表 6 所示,与 Einfal 等人(2023)和 P-STMO 方

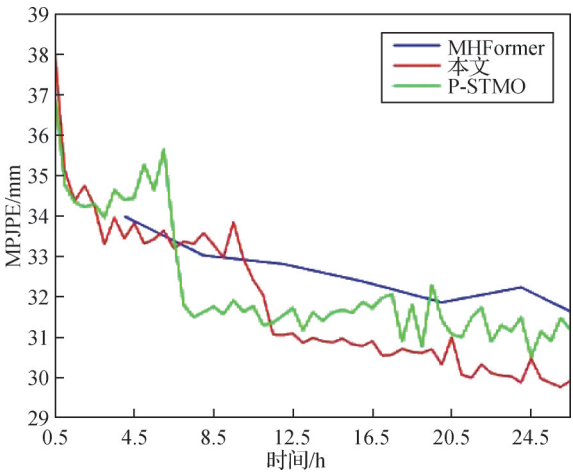


图 9 收敛对比图  
Fig. 9 Convergence comparison map

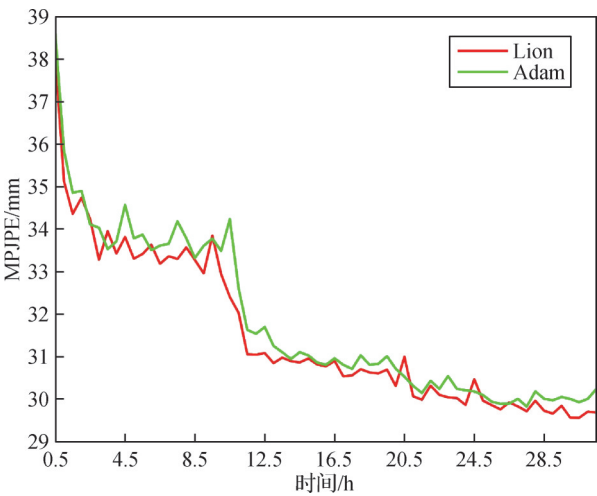


图 10 优化器对比图  
Fig. 10 Optimizer comparison map

法相比,本文方法的模型参数量(Params)和复杂度(FLOPs)有所增加,但训练时长上差别不大。相较于 MHFormer 和 MixSTE 方法,本文方法在模型参数、复杂度和每轮训练时间上均有所降低,且模型预测精度提高。

2.6 消融实验

为验证本文所提出模块的有效性,在 Human3. 6M 数据集上进行消融实验,根据 MPJPE 和 P-MPJPE 两个指标变化评估关节运动特征提取模块、关节特征融合模块和关节约束模块的性能。表 7 给出了本文方法的消融实验结果,模型 M4 相较于模型 M1 增加了关节特征消融模块,在 MPJPE 上降低了 2. 3%,在

P-MPJPE 上降低了 1.2%。由实验结果可见,关节特征融合模块可以有效建立关节整体空间关系并降低误差。模型 M4 相较于模型 M2 增加了关节运动特征模块,在 MPJPE 上降低了 1%、在 P-MPJPE 上降低了 1.2%,由实验结果可见,关节运动特征提取模块可以提取更丰富的特征,增强特征的表达能力,从而降低误差。模型 M5 与模型 M4 相比增加了关节约束模块,在 MPJPE 上降低了 0.6%、在 P-MPJPE 上降低了 1.7%,验证了关节约束模块在高维空间中对关节的约束作用。模型 M5 相较于模型 M1 增加了两个模块,在 MPJPE 上降低了 3%,在 P-MPJPE 上降低了 2.9%。由此可见,单独使用每个模块都可以在不同程度上降低姿态估计的误差,在使用模块的数量增加的时候,姿态估计的误差也随之减小,验证了多模块在联合使用时的有效性。

表 6 模型对比  
Table 6 Model comparison

| 方法              | Params/M     | FLOPs/M      | 训练时长/<br>(epoch/min) |
|-----------------|--------------|--------------|----------------------|
| Einfal 等人(2023) | <u>10.39</u> | <b>966</b>   | <b>14</b>            |
| P-STMO          | <b>6.70</b>  | <u>1 737</u> | <u>37</u>            |
| MHFormer        | 31.52        | 14 160       | 253                  |
| MixSTE          | 33.61        | 138 623      | 384                  |
| 本文              | 30.26        | 9 631        | 62                   |

注:加粗、下划线字体表示各列最优、次优结果。

表 7 消融实验  
Table 7 Ablation experiment

| 模型 | 关节运动特<br>征提取模块 | 关节特征<br>融合模块 | 关节约<br>束模块 | MPJPE       | P-MPJPE     |
|----|----------------|--------------|------------|-------------|-------------|
| M1 | √              |              |            | 29.9        | 23.5        |
| M2 |                | √            |            | 29.5        | 23.5        |
| M3 |                |              | √          | 29.8        | 23.5        |
| M4 | √              | √            |            | 29.2        | 23.2        |
| M5 | √              | √            | √          | <b>29.0</b> | <b>22.8</b> |

注:加粗字体表示各列最低误差。“√”表示实验中加入对应模块。

3 结 论

本文针对三维人体姿态估计中深度信息缺失和

难以捕捉关节细微变化的问题,提出一种融合多关节特征的单目视觉三维人体姿态估计网络。给出一种关节运动特征提取模块,有助于模型在时间维度上提取人体关节的运动特征。通过使用多分支的架构,丰富了特征信息,减轻了遮挡对模型的影响;通过关节特征融合模块建立不同关节在同一空间中的关节位置关系,加强了不同关节之间的联系,从而建立了全局关系;在关节约束模块设计中,充分利用中间帧的关节信息提供隐式约束,减少了歧义姿态,使模型生成合理的三维姿态,从而增强了模型的鲁棒性和精确度。实验验证了本文方法的有效性和鲁棒性,该模型可以应用在人体动作识别、路人行为预测等视觉任务中。下一步工作将重点放在简化模型上,在减少输入序列长度的同时保持模型的预测精度。

参考文献(References)

Belagiannis V, Amin S, Andriluka M, Schiele B, Navab N and Ilic S. 2014. 3D pictorial structures for multiple human pose estimation// Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 1669-1676 [DOI: 10.1109/CVPR.2014.216]

Cai J L, Liu H, Ding R W, Li W H, Wu J B and Ban M J. 2023. HTNet: human topology aware network for 3D human pose estimation//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island, Greece: IEEE: 1-5 [DOI: 10.1109/ICASSP49357.2023.10095949]

Chen T L, Fang C, Shen X H, Zhu Y H, Chen Z L and Luo J B. 2022. Anatomy-aware 3D human pose estimation with bone-based pose decomposition. IEEE Transactions on Circuits and Systems for Video Technology, 32(1): 198-209 [DOI: 10.1109/TCSVT.2021.3057267]

Chen X N, Liang C, Huang D, Real E, Wang K Y, Liu Y, Pham H, Dong X Y, Luong T, Hsieh C J, Lu Y F and Le Q V. 2023. Symbolic discovery of optimization algorithms [EB/OL]. [2025-01-12]. <https://arxiv.org/pdf/2302.06675.pdf>

Chen Z M, Dai J and Pan J J. 2024. A conditional diffusion model for 3D human pose estimation//Proceedings of the 4th International Conference on Consumer Electronics and Computer Engineering. Guangzhou, China: IEEE: 20-24 [DOI: 10.1109/ICCECE61317.2024.10504230]

Diaz-Arias A and Shin D. 2024. ConvFormer: parameter reduction in transformer models for 3D human pose estimation by leveraging dynamic multi-headed convolutional attention. The Visual Computer, 40(4): 2555-2569 [DOI: 10.1007/s00371-023-02936-5]

- Einfalt M, Ludwig K and Lienhart R. 2023. Uplift and upsample: efficient 3D human pose estimation with uplifting Transformers//Proceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 2902-2912 [DOI: 10.1109/WACV56688.2023.00292]
- Hassanin M, Khamiss A, Bennamoun M, Boussaid F and Radwan I. 2022. CrossFormer: cross spatio-temporal transformer for 3D human pose estimation [EB/OL]. [2025-01-12]. <https://arxiv.org/pdf/2203.13387.pdf>
- Hossain M R I and Little J J. 2018. Exploiting temporal information for 3D human pose estimation [EB/OL]. [2025-01-12]. <https://arxiv.org/pdf/1711.08585v4.pdf>
- Hu W B, Zhang C G, Zhan F N, Zhang L and Wong T T. 2021. Conditional directed graph convolution for 3D human pose estimation//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event, China: ACM: 602-611 [DOI: 10.1145/3474085.3475219]
- Ionescu C, Papava D, Olaru V and Sminchisescu C. 2014. Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36 (7): 1325-1339 [DOI: 10.1109/TPAMI.2013.248]
- Kang H B, Wang Y, Liu M Y, Wu D D, Liu P and Yang W M. 2023. Double-chain constraints for 3D human pose estimation in images and videos [EB/OL]. [2025-01-12]. <https://arxiv.org/pdf/2308.05298.pdf>
- Li H, Shi B W, Dai W R, Zheng H W, Wang B T, Sun Y, Guo M, Li C L, Zou J N and Xiong H K. 2023a. Pose-oriented transformer with uncertainty-guided refinement for 2D-to-3D human pose estimation//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 1296-1304 [DOI: 10.1609/aaai.v37i1.25213]
- Li S J, Zhu H S, Wang L and Liu X P. 2022. Dual auto-encoder network for human skeleton motion data optimization. *Journal of Image and Graphics*, 27(4): 1277-1289 (李书杰, 朱海生, 王磊, 刘晓平. 2022. 面向人体骨骼运动数据优化的双自编码器网络. *中国图象图形学报*, 27(4): 1277-1289) [DOI: 10.11834/jig.200780]
- Li W H, Liu H, Ding R W, Liu M Y, Wang P C and Yang W M. 2023b. Exploiting temporal contexts with strided transformer for 3D human pose estimation. *IEEE Transactions on Multimedia*, 25: 1282-1293 [DOI: 10.1109/TMM.2022.3141231]
- Li W H, Liu H, Tang H, Wang P C and Van Gool L. 2022. MHFormer: multi-hypothesis transformer for 3D human pose estimation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13137-13146 [DOI: 10.1109/CVPR52688.2022.01280]
- Li W H, Liu M Y, Liu H, Guo T Y, Wang T, Tang H and Sebe N. 2024. GraphMLP: a graph MLP-like architecture for 3D human pose estimation [EB/OL]. [2025-01-12]. <https://arxiv.org/pdf/2206.06420.pdf>
- Lin J H and Lee G H. 2019. Trajectory space factorization for deep video-based 3D human pose estimation [EB/OL]. [2025-01-12]. <https://arxiv.org/pdf/1908.08289.pdf>
- Lin K, Wang L J and Liu Z C. 2021. End-to-end human pose and mesh reconstruction with transformers//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1954-1963 [DOI: 10.1109/CVPR46437.2021.00199]
- Liu J F, Rojas J, Liang Z J, Li Y H and Guan Y S. 2020a. A graph attention spatio-temporal convolutional network for 3D human pose estimation in video [EB/OL]. [2025-01-12]. <https://arxiv.org/pdf/2003.14179.pdf>
- Liu R X, Shen J, Wang H, Chen C, Cheung S C and Asari V. 2020b. Attention mechanism exploits temporal contexts: real-time 3D human pose reconstruction//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5063-5072 [DOI: 10.1109/CVPR42600.2020.00511]
- Mehta D, Rhodin H, Casas D, Fua P, Sotnychenko O, Xu W P and Theobalt C. 2017a. Monocular 3D human pose estimation in the wild using improved CNN supervision//Proceedings of 2017 International Conference on 3D Vision. Qingdao, China: IEEE: 506-516 [DOI: 10.1109/3DV.2017.00064]
- Pavlakos G, Zhou X W, Derpanis K G and Daniilidis K. 2017. Coarse-to-fine volumetric prediction for single-image 3D human pose//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1263-1272 [DOI: 10.1109/CVPR.2017.139]
- Pavlo D, Feichtenhofer C, Grangier D and Auli M. 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7745-7754 [DOI: 10.1109/CVPR.2019.00794]
- Shan W K, Liu Z H, Zhang X F, Wang S S, Ma S W and Gao W. 2022. P-STMO: pre-trained spatial temporal many-to-one model for 3D human pose estimation//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 461-478 [DOI: 10.1007/978-3-031-20065-6\_27]
- Sigal L, Balan A O and Black M J. 2010. HumanEva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1/2): 4-27 [DOI: 10.1007/s11263-009-0273-6]
- Sun X, Shang J X, Liang S and Wei Y C. 2017. Compositional human pose regression//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2621-2630 [DOI: 10.1109/ICCV.2017.284]
- Tekin B, Márquez-Neila P, Salzmann M and Fua P. 2017. Learning to fuse 2D and 3D image cues for monocular body pose estimation//Proceedings of 2017 IEEE International Conference on Computer

- Vision. Venice, Italy: IEEE: 3961-3970 [DOI: 10.1109/ICCV.2017.425]
- Wang J B, Yan S J, Xiong Y J and Lin D H. 2020. Motion guided 3D pose estimation from videos//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 764-780 [DOI: 10.1007/978-3-030-58601-0\_45]
- Xue F, Bian F L and Li S J. 2024. Attention fusion network for estimation of 3D joint coordinates and rotation of human pose. *Journal of Image and Graphics*, 29(10): 3116-3129 (薛峰, 边福利, 李书杰. 2024. 面向三维人体坐标及旋转角估计的注意力融合网络. *中国图象图形学报*, 29(10): 3116-3129) [DOI: 10.11834/jig.230502]
- Xue Y Z, Chen J S, Gu X M, Ma H M and Ma H B. 2022. Boosting monocular 3D human pose estimation with part aware attention. *IEEE Transactions on Image Processing*, 31: 4278-4291 [DOI: 10.1109/TIP.2022.3182269]
- Yan S J, Xiong Y J and Lin D H. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI: 7444-7452 [DOI: 10.1609/aaai.v32i1.12328]
- Zeng A L, Sun X, Huang F Y, Liu M H, Xu Q and Lin S. 2020. SRNet: improving generalization in 3D human pose estimation with a split-and-recombine approach [EB/OL]. [2025-01-12]. <https://arxiv.org/pdf/2007.09389.pdf>
- Zhan Y, Li F H, Weng R L and Choi W. 2022. Ray3D: ray-based 3D human pose estimation for monocular absolute 3D localization//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13106-13115 [DOI: 10.1109/CVPR52688.2022.01277]
- Zhang J L, Tu Z G, Yang J Y, Chen Y J and Yuan J S. 2022. MixSTE: seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13222-13232 [DOI: 10.1109/CVPR52688.2022.01288]
- Zhang K X, Luan X M, Shah Syed T H and Xiang X Z. 2023. ICRFormer: an improving cos-reweighting transformer for 3D human pose estimation in video//Proceedings of the 35th Chinese Control and Decision Conference. Yichang, China: IEEE: 436-441 [DOI: 10.1109/CCDC58219.2023.10326602]
- Zheng C, Zhu S J, Mendieta M, Yang T J N, Chen C and Ding Z M. 2021. 3D human pose estimation with spatial and temporal transformers//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 11636-11645 [DOI: 10.1109/ICCV48922.2021.01145]

## 作者简介

贾迪,男,教授,主要研究方向为立体匹配与三维重建、摄影测量和位姿估计。E-mail:lntu\_jiadi@163.com

何德堃,通信作者,男,硕士研究生,主要研究方向为三维人体姿态估计。E-mail:lntu\_hedekun@163.com

韩雪峰,男,研究员,主要研究方向为人工智能和智慧矿山。E-mail: 173657676@qq.com

杨柳,男,硕士研究生,主要研究方向为三维人体姿态估计。E-mail:lntu\_yangliu@163.com

程硕,男,硕士研究生,主要研究方向为6D位姿估计。

E-mail:lntu\_cs@163.com

刘宇琪,男,硕士研究生,主要研究方向为单目深度估计。

E-mail:yuqi030204@sina.com