Computer Science and Application 计算机科学与应用, 2024, 14(3), 96-107
Published Online March 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.143061

Hans汉斯

基于DirtNet与惯性测量单元的人体姿态估计

罗　胜，张元正*，叶润泽，朱锦乐，张博文

温州大学计算机与人工智能学院，浙江　温州

收稿日期：2024年2月20日；录用日期：2024年3月20日；发布日期：2024年3月27日

摘　要

仅使用少量的惯性测量单元(IMU, Inertial Measurement Unit)进行人体姿态估计是一种非侵入性且经济的人体姿态估计方法，该方法主要面临的挑战是从带有噪声的IMU信号中精确估计人体姿态。为此，对人体姿态估计问题提出了一种仅使用6个IMU精确估计人体姿态的方法。1) 提出了一种双重信息保留注意力Transformer网络(DirtNet, Dual information retention transformer Network)，它能够有效保留历史信息并通过注意整个序列的信息来获得更好的结果。2) 通过对加速度进行积分了获得了近似变化速度，并将其作为额外的输入通道以提高了人体姿态估计的精确度。3) 使用均匀滤波过滤和白噪声模拟的方法对合成的加速度进行了数据增强，以此来拟合真实的IMU数据并得到更好的训练结果。与之前的研究相比，改进后的方法有效提高了姿态估计的精确度。

关键词

人体姿态估计，惯性测量单元，SMPL，骨架模型，实时，DirtNet

Human Pose Estimation Based on DirtNet and Inertial Measurement Units

Sheng Luo, Yuanzheng Zhang*, Runze Ye, Jinle Zhu, Bowen Zhang

School of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou Zhejiang

Received: Feb. 20th, 2024; accepted: Mar. 20th, 2024; published: Mar. 27th, 2024

Abstract

Using a small number of inertial measurement units (IMUs) for human pose estimation is a non-intrusive and cost-effective method. However, accurately estimating human pose from noisy IMU signals poses a significant challenge. To address this challenge, a method that utilizes only six IMUs for precise human pose estimation is proposed. 1) A dual information retention attention

*通讯作者。

文章引用: 罗胜, 张元正, 叶润泽, 朱锦乐, 张博文. 基于 DirtNet 与惯性测量单元的人体姿态估计[J]. 计算机科学与应用, 2024, 14(3): 96-107. DOI: 10.12677/csa.2024.143061

Transformer network, called DirtNet, is introduced. This network effectively preserves historical information and leverages attention over the entire sequence to achieve better results. 2) The approximate velocity is obtained by integrating the acceleration, and it is used as an additional input channel to improve the accuracy of human pose estimation. 3) A data augmentation technique is applied by filtering the synthesized acceleration using uniform filtering and simulating white noise. This approach helps to fit the real IMU data and achieve better training results. Compared to previous research, the improved method significantly enhances the accuracy of pose estimation. By combining the strengths of DirtNet, leveraging historical information, incorporating velocity as an input, and applying data augmentation techniques, this method provides more precise human pose estimation results.

Keywords

Human Pose Estimation, Inertial Measurement Units, SMPL, Skeleton Model, Real-Time, DirtNet

1. 引言

人体的动作捕捉在游戏(人物模型的动作、体感游戏)、体育(运动姿势的矫正)、医学(病人的复建、不良姿势的矫正)、VR/AR 和电影制作等各种应用中发挥着重要作用。目前比较流行的是基于视觉人体姿态估计方法。其中一种是通过多个摄像机与深度学习来进行人体姿态估计，如使用了 RGB 相机[1] [2]和深度相机[3]进行了人体姿态估计，这种方法可以达到较高的精确度，但无法解决遮挡的问题，不适合在有较多遮挡物的室内使用。另一种是通过在人体上安装光学标记并通过摄像头记录光学信号以实时捕捉人体的动作的方法，比如 Vicon 就使用这种方法进行人体姿态估计并达到较高的精确度。但是基于光学标记的方法需要庞大且昂贵的基础设施，测试者只能在室内进行运动，同时也无法解决遮挡的问题。目前来说基于视觉的方法都需要昂贵的设备以及合适的场所，并不适合消费者级使用。

与基于视觉的姿态估计方法相比，IMU 安装于人体且独立于环境，因此不会受到环境遮挡的影响，可以在各种各样的环境中使用。同时由于 IMU 的价格较低，比较适合消费者级的用户进行使用。基于 IMU 的人体姿态估计方法的缺点是当所使用的 IMU 数量较少时，进行姿态估计时会产生较大的误差。但是随着神经网络的发展，目前的基于单纯基于 IMU 的高精度人体姿态估计已经有了一定的发展。有的研究者使用了双向递归神经网络[4]并使用 6 个 IMU 对人体姿态进行了估计，也有的通过将姿态估计分为三个部分提高了姿态估计的精确度[5]，有的通过使用简单递归单元(SRU, Simple Recurrent Unit) [6]进行快速的人体姿态估计[7]。以上方法能够通过仅使用 6 个 IMU 进行较为准确的人体姿态估计，但所使用的网络都是基于 RNN 及其变体，没有充分利用 IMU 信息。因此，针对仅使用 IMU 进行人体姿态估计任务提出了一种新的网络结构。

主要的改进点是：

1) 提出了一种新的具有双重信息保留注意力模块的网络结构，更好地建模了长距离依赖，提高了姿态估计的准确度。

2) 使用均匀滤波过滤和白噪声模拟的方法对合成加速度进行了数据增强以更好地拟合真实数据。

3) 使用近似变化速度作为额外参数输入并降低了姿态估计的误差。

之后将在第 2 节介绍了各类人体姿态估计的相关工作。第 3 节介绍了运动学模型和数据集。第 4 节介绍了所使用的网络结构。第 5 节给出了网络在 DIP-IMU 数据集上的定量和定性评估结果，并进行了消融实验，以证明该方法的有效性。第 6 节总结了目前方法的局限性并说明未来可能的发展方向。

2. 相关工作

人体的姿态估计目前有三种主要的方法。

2.1. 基于视觉的方法

基于视觉的方法

基于视觉的人体姿态估计是目前最流行的方法，通过相机捕捉人体运动并通过深度学习进行人体姿态估计。例如，利用视频帧之间丰富的时间特征来辅助关键点识别，编码关键点时空上下文以提供足够的搜索空间，然后通过姿态校正网络进行处理，以有效地细化姿态估计[8]。通过设计了一种简单有效的半监督训练策略以利用未标记的视频数据，使用基于 2D 关键点上的扩展时间卷积的全卷积模型快速预测电影中的 3D 姿态[9]。通过提出了一种具有双分支解码器的独特编码器－解码器结构，解决 2D 关节位置的可变不确定性，并提供了新的大规模真实感合成数据集[10]。也有研究者提出了一种用于估计 3D 人体姿态的快速、统一的端到端模型，该模型将各阶段最佳做法进行结合，之后有人将其应用于艺术体操、训练和舞蹈的评估和评分[11]。

2.2. 基于视觉与惯性传感器融合的方法

通过将摄像机与惯性传感器进行融合使用，可以使姿态估计达到较高的精度。例如，将每个图像中的 2D 姿势检测与每个人配备的相应的 IMU 相关联，然后使用连续优化框架来优化统计身体模型姿势[12]。通过提出了使用多视图图像和连接到人体肢体的一些 IMU 来估计 3D 人体姿势[13]。此方法首先从两个信号中检测二维姿态，然后将其提升到三维空间。通过使用多通道 3D 卷积神经网络从视觉占用中学习姿势嵌入，并从离散体积概率视觉外壳中的 MVV 中学习语义 2D 姿势估计[14]。

2.3. 基于纯惯性传感器的方法

目前基于纯惯性传感器的人体姿态估计主要通过姿态估计优化方法和网络模型来进行优化。

姿态估计优化方法是从姿态估计的流程中进行优化，目前国内外对基于 IMU 的人体姿态估计研究已经有了一定的进展。在早期就有人使用了 17 个惯性传感器测量人体运动时各关节的旋转数据[15]，以此对人体的姿态进行较为准确的估计，但过多的传感器会限制人体的动作，设置的过程也不方便，同时过多的 IMU 会增加成本。有研究者为人体研究提供了一个可靠的运动学模型[16]，之后在姿态估计方面，有人做出了一项开创性的工作，通过一种基于迭代优化的方法使用 6 个 IMU 估计人体姿态[17]，但它必须以离线方式操作，这使得实时应用变得不可行。而另外的研究者通过利用双向递归神经网络(也使用 6 个 IMU)来直接学习 IMU 测量到人体关节旋转之间的映射，相较于之前的工作提高了精度但仍然有可优化的地方。最近也有同样使用双向递归神经网络的方法，同时通过将姿态估计分为多个阶段，即从 IMU 测量信息分步估计叶关节位置、全关节位置、全关节旋转，通过这种逐步估计的方法实现了更高精度的姿态估计并估计了穿戴者的整体平移。

网络模型的优化是从使用的神经网络上进行优化，改进神经网络的性能，使其更适合人体姿态估计任务。目前人体姿态估计都使用递归神经网络(RNN, Recurrent Neural Network)和双向递归神经网络作为网络模型，保留时间信息以提高精确度。RNN 通过传递当前时间步的隐藏状态，来对序列数据进行建模。作为一种数据处理模型，RNN 的隐藏状态是通过逐步迭代计算得到的，每个时间步只能看到当前位置之

前的信息，对于人体的运动这种有长距离依赖的数据并不能很好地捕捉其依赖关系，同时比较容易产生梯度消失和梯度爆炸地问题。Transformer 通过对输入乘以三个矩阵得到三个新的矩阵，并使用这三个矩阵计算自注意力，该方法可以通过全局信息提高了序列建模的能力，并通过堆叠多个这样的多头注意力模块和前馈神经网络模块提升了性能[18]。保留网络(RetNet, Retentive Network)与 Transformer 同样堆叠了多个相同的模块，但其注意力模块在保留了 Transformer 中的注意力矩阵的同时加入了类似 RNN 的状态信息参数，提高了推理时的性能[19]。因此相较于 RNN 结构，可以关注全局信息并进行序列建模的 Transformer 结构更加适合应用于人体姿态估计任务中。因此在 RetNet 网络和 Transformer 网络的基础上提出了双重信息保留 Transformer 网络，能够更进一步提高姿态估计的精确度。

3. IMU 校准和数据集预处理

本节将介绍进行姿态估计前所需的 IMU 校准和数据集预处理工作。在 2.1 节介绍所使用的运动学模型，在 2.2 节介绍了 IMU 的校准方法，在 2.3 节介绍所使用的数据集以及数据集预处理方法。

3.1. 运动学模型

人体姿态估计的运动学模型 SMPL 骨架模型。SMPL 是一种蒙皮多人线性模型，在有着较高的准确度的同时与现有的图像管道相兼容。SMPL 骨骼将身体姿态、动态身体软组织和其他的模板结合，然后通过混合蒙皮进行转换。其定义为：

$$M(\theta)=W(T,J,\theta,\mathcal{W}) \tag{1}$$

其中 T 是静止姿势中的模板网格，J 是 24 个身体关节，θ 是根据关节角度的姿势参数，$\mathcal{W}$ 是线性混合蒙皮函数。由于 IMU 数据不包含人类形态特征，所以数据集中的所有肢体长度都相等，即姿态估计都在同一体型下进行。

3.2. 传感器校准

使用的 6 个 IMU 分别固定在人体的腰部、左小腿、右小腿、左前臂、右前臂、头部这六个位置。由于每个传感器收集的数据都在其局部坐标系中，而人体姿态估计需要使用运动学模型 SMPL 坐标系下的数据，因此需要对传感器的数据进行校准。

每个 IMU 数据测量的加速度与传感器坐标系 F^S 有关，其基矩阵为 B^S。IMU 测量的方向数据与惯性坐标系 F^I 相关，其基矩阵为 B^I。令姿态估计所使用的人体运动学模型 SMPL 模型的坐标系为 F^M，其基矩阵为 B^M。使得腰部的 IMU 坐标系与 SMPL 坐标系一致，IMU 的方向 O 从 F^I 变为了 F^M，可得：

$$B^M = B^I O \tag{2}$$

让穿戴 IMU 的测试人员保持 T 姿势静止几秒，在 T 姿势中可以方便得知与 F^M 坐标系相关的 SMPL 骨骼方向的数据 R_M^{bone}，之后读取与坐标系 F^S 有关的平均加速度读数 a_S^{IMU} 以及与坐标系 F^I 有关的方向数据 R_I^{IMU}。并且由于穿戴的 IMU 与人体骨骼之间的方向必定存在误差，将误差设为 R_I^{offset}，由此可得在 F^I 坐标系中的人体骨骼方向：

$$R_I^{bone} = R_I^{IMU} R_I^{offset} \tag{3}$$

由于人体的骨骼方向在两个坐标系下是等效的，由此可得：

$$B^M R_M^{bone} = B^I R_I^{bone} \tag{4}$$

将(2) (3) (4)三个式子相结合可以得到在 F^M 坐标系中的方向数据：

$$R_M^{bone} = O^{-1} R_I^{IMU} R_I^{offset} \tag{5}$$

对于加速度数据的转换，首先假设 IMU 在安装后的位置是固定的，不会产生相对移动，所以可以得到在 F^I 坐标系下的骨向量和

IMU 中的加速度是一致的，即：

$$a_I^{bone} = a_I^{IMU} \tag{6}$$

同时可以将 IMU 局部坐标系 F^S 中的加速度转化为全局惯性坐标系 F^I 下的加速度：

$$a_I^{IMU} = a_S^{IMU} R_I^{IMU} \tag{7}$$

由于传感器的误差以及方向的不确定性，全局坐标系 F^M 中的加速度会存在一个偏移，加上偏移之后再由加速度在两个坐标系下的等效关系可得：

$$B^M \left(a_M^{bone} + a_M^{offset} \right) = B^I a_I^{bone} \tag{8}$$

由于测试人员在测试中保持了 T 姿势的静止状态，因此 a_M^{bone} 为 0，将(2) (5) (6) (7)这 4 个式子结合可得：

$$a_M^{offset} = O^{-1} a_S^{IMU} R_I^{IMU} \tag{9}$$

由于 a_M^{offset} 变为了一个已知量，因此可以计算在全局坐标系 F^M 下骨向量的加速度为：

$$a_M^{bone} = O^{-1} a_S^{IMU} R_I^{IMU} - a_M^{offset} \tag{10}$$

由此就得到了全局坐标系 F^M 下的方向 R_M^{bone} 和加速度数据 a_M^{bone} 。

3.3. 训练数据

用 IMU 进行人体姿态估计的网络需要的输入包含加速度和方向，以及人体运动时各关节的旋转的数据来进行训练。目前达到要求的数据集有 DIP (该数据集包括 10 名佩戴 17 个 IMU 的受试者进行约 90 分钟运动的 IMU 测量和姿势参数)和 TotalCapture [20] (包括 5 名佩戴 13 个 IMU 的受试者进行约 50 分钟运动的 IMU 测量、姿势参数和全局平移)。目前所拥有的数据集无法训练出具有足够泛化能力的模型，因此需要通过使用合成 IMU 数据的方法来获取足够的数据。

合成 IMU 数据的原始数据集为 AMASS 运动数据集[21]，该数据集是现有动作捕捉数据集的集合，包括在 300 多个实验个体上收集和执行的 40 多小时不同类型运动的姿势数据参数。AMASS 数据集中的每个子数据集包含 SMPL 模型的每个关节的旋转和全身位移参数，但其中并没有 IMU 数据，因此需要通过放置虚拟 IMU 的方法来获得 IMU 数据。SMPL 是一个具有 6890 个网格顶点的模型，选取其中与人体配置 IMU 位置相近的顶点设置虚拟 IMU 直接获得相应的速度和旋转数据并通过计算获得加速度，加速度的计算公式为：

$$a_i(t) = \frac{x_i(t-n) + x_i(t+n) - 2x_i t}{(n\Delta t)^2}, \quad i = 1, 2, \cdots, 6 \tag{11}$$

其中表示第 t 帧中第 i 个传感器的加速度测量值，Δt 为指定两个连续帧之间的间隔。模型以 60 帧/秒的速度对所有数据进行采样，因此 Δt 的值为约为 0.0167。在计算的过程中，n = 3 或 4 时接近与真实的加速度值，最终选择 n = 4 以获取更平滑的加速度数据。

通过将合成的加速度和真实的加速度进行对比，可以发现合成的加速度数据与真实的加速度数据在噪声分布上存在差异，原因是用 IMU 测量得出的加速度存在着漂移与白噪声。如图 1 两类的加速度由于

漂移和噪声的问题难以拟合，最终会导致在训练出的模型在进行测试时产生额外误差。对真实的加速度数据和合成的加速度数据进行了滤波，并添加正负值为 0.2 的均匀噪声来模拟加速度测量中的白噪声，对合成的数据进行了数据增强。图 2 为滤波后的加速度图像，经过滤波后的合成加速度与真实加速度具有更好的拟合度，能够在训练出更好的模型。

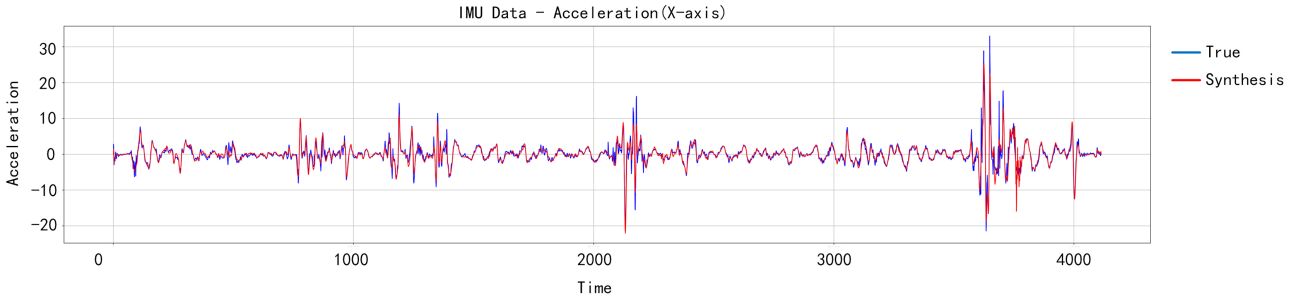

Figure 1. No filtered synthetic acceleration and true acceleration
图 1. 无滤波合成加速度和真实加速度

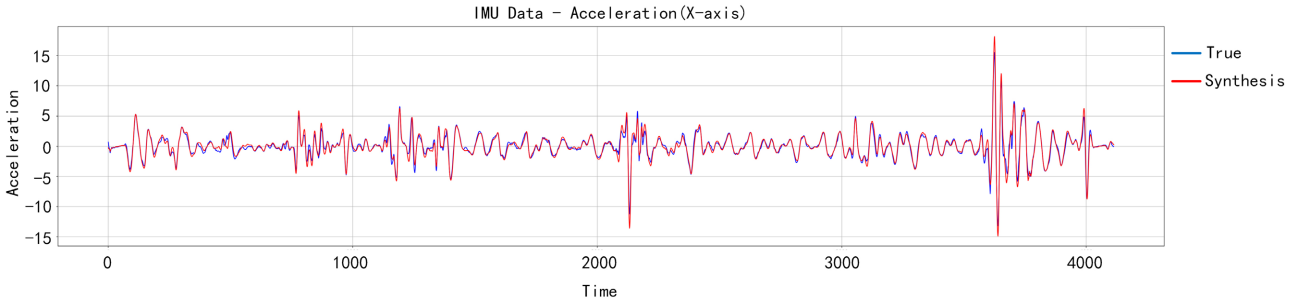

Figure 2. Filtered synthetic acceleration and true acceleration
图 2. 滤波后合成加速度与真实加速度

4. 方法

在此节介绍所使用的姿态估计方法，在 4.1 介绍系统输入与数据标准化的方法，在 4.2 介绍变化速度 Δv 的获取，在 4.3 介绍网络结构，在 4.4 介绍训练细节。

4.1. 系统输入与数据标准化

网络的目标是通过六个 IMU 还原人体姿态，即对全身其他 15 个关节的旋转进行估计。将设置在腰部的 IMU 称为根节点，将其他节点称为叶节点。系统的输入为 6 个 IMU 数据中与人体运动相关的加速度数据为 $a=\left[a_{leaf},\cdots,a_{root}\right]\in\mathbf{R}^{18}$ 并进行标准化得到：

$$a_{leaf}=R_{root}^{-1}\left(a_{leaf}-a_{root}\right)/30 \tag{12}$$

$$a_{root}=R_{root}^{-1}a_{root} \tag{13}$$

标准化的目的是将加速度数据变为与根节点相关，比例因子 30 为经验所得，是为了使加速度适应网络输入并调整加速度数据大小。

IMU 的方向数据为 $r=\left[r_{leaf},\cdots,r_{root}\right]\in\mathbf{R}^{54}$ 并进行标准化得到：

$$R_{leaf}=R_{root}^{-1}R_{root} \tag{14}$$

对方向数据进行标准化同样是为了将方向数据变为与根节点相关。

4.2. 对加速度积分获得变化速度 Δv

仅凭 IMU 的加速度和旋转数据对人体全身 15 个关节的映射是相对具有挑战性的。因为当加速度和旋转数据相同时，稀疏的 IMU 数据会具有不确定性，对关节旋转的估计会产生偏差。例如当人在站着和坐着时，imu 的数据几乎是相同的。因此，需要利用全局信息来判断当时的姿态，以及输入更多的信息提高估计的精度。

过 IMU 还原人体姿态，预测 15 个关节的旋转时，除了加速度和旋转信息，速度信息也可以提高姿态估计的精度。虽然 IMU 无法直接获得速度，但可以通过累计在很小时间内的加速度，相当于将其积分到 Δv，以此来获得更丰富的信息：

$$a\Delta t = \Delta v \tag{15}$$

令 Δt 的值为 0.5，由于使用 60 帧/秒的速度采集数据，因此相当于累加 30 个加速度值，最后和之前加速度的标准化一样除以一个缩放因子 15 来适应网络输入并调整 Δv 的大小以获得更好的输出，比例因子 15 同样为经验所得。

4.3. 网络结构

姿态估计网络结构的总体框架如图 3 所示，将设置在身体各部位的 6 个 IMU 所收集到的加速度数据 $a \in \mathbf{R}^{18}$ 和旋转数据 $r \in \mathbf{R}^{54}$ 进行标准化并合并。之后将通过累计 0.5 秒内的加速度数据得到近似的 $\Delta v \in \mathbf{R}^{18}$ 并将其也合并到 IMU 数据中，最终输入到线性层的 IMU 数据为 $i \in \mathbf{R}^{90}$。将进行处理后的 IMU 数据输入到一个线性层中，第一个线性层的作用是从输入的 IMU 数据中提取当前动作的特征，其隐藏单元的维度设置为 256。

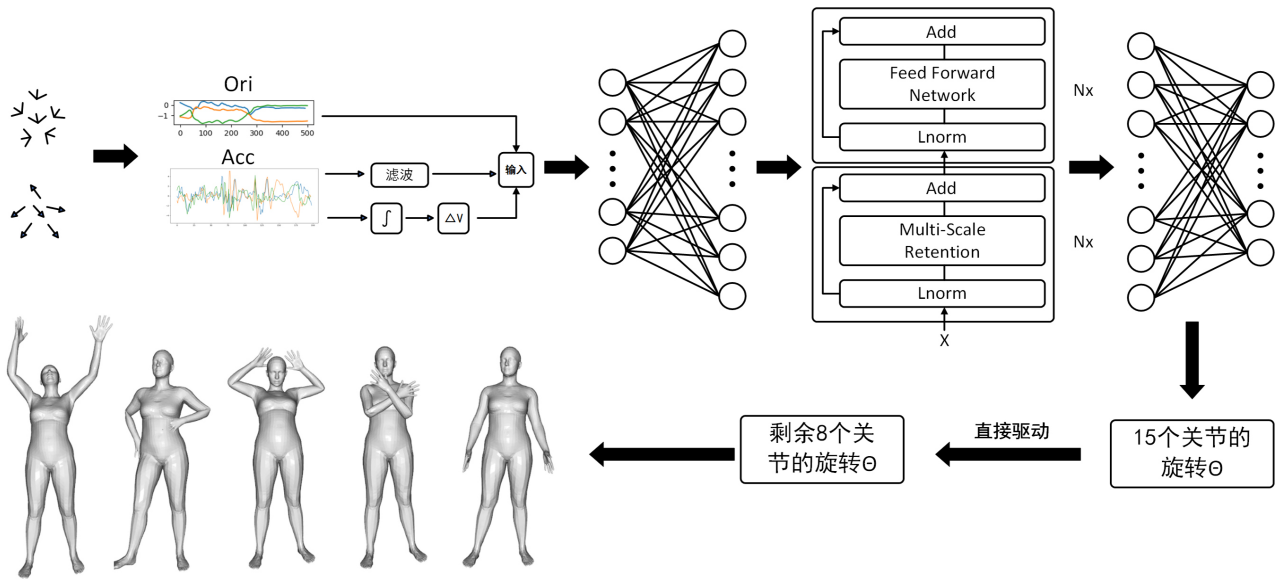

Figure 3. Overall network architecture
图 3. 网络整体框架

经过特征提取后的数据被输入到网络中，网络的整体结构如图 3 中所示为堆叠了多个相同模块的网络。其中双重信息保留注意力模块的结构如图 4 所示。

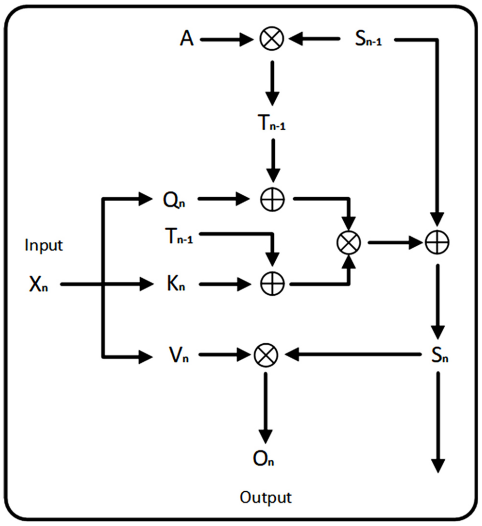

Figure 4. Bidirectional information preservation attention module
图 4. 双向信息保留注意力模块

其中 $Q_n \in \mathbf{R}^{1\times d}$、$K_n \in \mathbf{R}^{1\times d}$、$V_n \in \mathbf{R}^{1\times d}$ 为当前时间步的自注意力计算向量，由当前时间步的输入 $X_n \in \mathbf{R}^{1\times d}$ 与三个矩阵相乘获得。$S_{n-1} \in \mathbf{R}^{d\times d}$ 为上一个时间步的状态信息，$A \in \mathbf{R}^{1\times d}$ 与其相乘得到一个状态参数 $T_{n-1} \in \mathbf{R}^{1\times d}$。根据公式(17)可以计算得到当前时间步的状态参数 $S_n \in \mathbf{R}^{d\times d}$，并将当前时间步的状态信息传递给下一个时间步。通过这种改进后的双重信息保留注意力模块，可以在更好地保留时间信息的同时考虑全局信息，从而提高姿态估计的准确度。

最后将提取过时间信息的数据输入到线性层中得到 15 个关节的 6d 旋转表示 Pose $\in \mathbf{R}^{90}$。最后还原的关节数为 15，是因为部分关节的数据可以直接从 IMU 中获得，而 15 个关节中的其中 8 个关节可以直接驱动剩下的 8 个关节旋转，因此得到 15 个关节的旋转就可以完整表示人体的姿态。

在实验过程中，可以发现提出的网络模型虽然减小了姿态估计的角度误差，但会导致较大的抖动误差，因此选择在姿态估计时添加了一个实时低通滤波以减少抖动误差：

$$\hat{\theta} = 0.8\tilde{\theta}(t-1) + 0.2\tilde{\theta}(t) \tag{16}$$

根据经验将参数选为 0.8 和 0.2，该参数可以在减少对角度误差的影响并减少抖动误差。

4.4. 训练细节

模型训练分为预训练与微调两个部分，将经过处理的 AMASS 数据作为预训练的训练集以此来获得人体运动的先验知识。之后在微调使用 DIP 中 S01-S08 的数据作为训练集进行微调。

所有的训练都是在 12 vCPU Intel(R) Xeon(R) Platinum 8255C CPU 和 RTX 3080 显卡上进行的，软件使用 pytorch 1.10.0 和 cuda 11.3 进行训练。模型使用 Adam [22]作为优化器，训练时的学习率为起始学习率为 0.001 的余弦退火学习率，训练时的批次被设置为 256，每个实验训练 200 轮。

5. 实验比较

在本章进行实验结果的对比，在 4.1 节介绍模型的评价指标并进行实验结果比对，在 4.2 节进行消融实验并展示结果对比，在 4.3 节进行了实验结果的可视化对比。在 4.4 节描述了该模型的局限性与未来展望。

5.1. 在线与离线的结果对比

本节对在线和离线的实验结果进行对比，实验的评价指标为：1) SIP 误差(SIP Err)：测量上臂和大腿的在全局坐标系下的平均旋转误差，单位为度；2) 角度误差(Ang Err)：全身关节点的全局平均旋转误差。3) 位置误差(Pos Err)：根关节(脊柱)对齐的所有估计关节的平均欧几里得距离误差。4) 网格误差(Mesh Err)：测量与根关节(脊椎)对齐的估计身体网格的所有顶点的平均欧几里德距离误差。5) 抖动误差(Jitter Err)：测量预测运动中所有身体关节的平均抖动。

Table 1. Offline estimation experimental results
表 1. 离线估计实验结果

方法	SIP Err (deg)	AngErr (deg)	Pos Err (cm)	Mesh Err (cm)	JitErr (10^2 m/s^3)
SIP	21.02 (±9.61)	8.77 (±4.83)	6.66 (±3.33)	7.71 (±3.80)	3.86 (±6.32)
DIP	16.36 (±8.60)	14.41 (±7.90)	6.98 (±3.89)	8.56 (±4.65)	23.37 (±23.84)
TransPose	13.97 (±6.77)	**7.62 (±4.01)**	4.90 (±2.75)	5.83 (±3.21)	**1.19 (±1.76)**
Ours	**12.14 (±7.00)**	8.61 (±4.94)	**4.41 (±2.69)**	**5.08 (±3.05)**	2.08 (±2.89)

Table 2. Online estimation experimental results
表 2. 在线估计实验结果

方法	SIP Err (deg)	AngErr (deg)	Pos Err (cm)	Mesh Err (cm)	JitErr (10^2m/s^3)
DIP	17.10 (±9.59)	15.16 (±8.53)	7.33 (±4.23)	8.96 (±5.01)	30.13 (±28.76)
TransPose	16.68 (**±8.68**)	**8.85 (±4.82)**	5.95 (±3.65)	7.09 (±4.24)	**6.11 (±7.92)**
Ours	**14.68** (±9.13)	9.63 (±5.77)	**5.48 (±3.56)**	**6.40 (±4.08)**	8.21 (±18.53)

表 1 和表 2 分别为模型对人体的姿态进行离线估计和在线估计的定量比较，可以得出模型在 SIP 误差，位置误差和网格误差方面得到了更小的误差结果。由此可以得出该网络在上臂与大腿方面的姿态估计方面有较好的准确率提升，而造成这种准确率提升的原因是改进的双重信息保留注意力模块同时使用了自注意力与双重状态信息传递，同时考虑了全局信息并有效利用了历史信息。而所使用的近似 Δv 拟合以及合成数据集的数据增强，使得模型有了更多的输入信息与更贴近真实数据的合成数据集，这两种方法也进一步提升了模型的性能。

5.2. 消融实验

本节进行消融实验以验证该网络模型的有效性。

Table 3. Ablation experimental results
表 3. 消融实验结果

方法	SIP Err (deg)	AngErr (deg)	Pos Err (cm)	Mesh Err (cm)	JitErr (10^2m/s^3)
正常方法	12.14 (±7.00**)**	8.61 (±4.94)	4.41 (±2.69)	5.08 (±3.05)	2.08 (±2.89)
无加速度滤波	12.93 (±7.08)	8.79 (±4.99)	4.46(±2.70)	5.15 (±3.07)	2.10 (±3.02)
无Δv	12.57 (±7.19)	9.00 (±5.13)	4.57(±2.74)	5.29 (±3.16)	2.13 (±3.31)
无低通滤波	12.04 (±6.92)	8.06 (±4.53)	4.23(±2.57)	4.80 (±2.88)	4.73 (±10.49)

表 3 为进行消融实验后的实验结果，可以看出模型所使用的 Δv 与合成数据增强方法都能够有效提高

姿态估计的准确度。通过数据集的滤波，使得真实产生的 IMU 加速度数据与合成所获得的 IMU 加速度数据具有更好的拟合性，通过合成 Δv，并将其加入 IMU 数据中为姿态估计提供了更丰富的信息，可以看到各项误差都在一定程度上减少了。同时对于模型会产生较高的抖动误差的问题，通过使用低通滤波略微增加了其他方面的误差，但大幅降低了预测过程中产生的抖动误差。

5.3. 可视化定性比较

图 5 为在线和离线的定性可视化比较，通过从测试的数据集中选取了一些帧并进行可视化来进行定性的可视化比较。

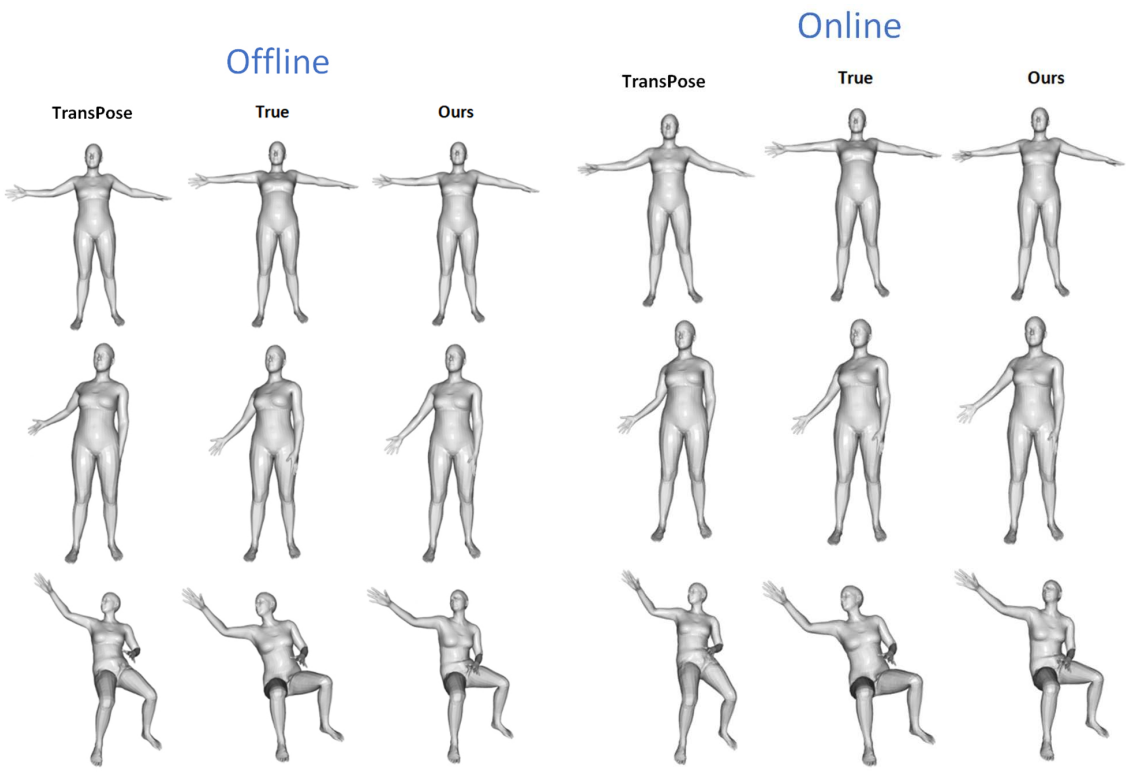

Figure 5. Offline and online visualization comparison
图 5. 离线与在线可视化比较

从图中可以看出，改进后的人体姿态估计模型对于手臂部分和腿部的部分的估计更加接近真实的动作。TransPose 在手轴和膝盖部分的弯曲以及对于手臂的位置估计有着一定的误差，而改进后的人体姿态估计模型对于这些部分的估计更加精确，在姿态估计时消除了这些误差从而更加接近真实的动作。

5.4. 局限性与展望

对于人体姿态估计的任务，改进后的方法取得了一些进步但也受到了一些限制，限制有 IMU 硬件和估计任务本身两个方面。

在 IMU 的测量方面。由于 IMU 在长时间测量时会积累漂移同时由于加速度中噪声的存在，最终会导致姿态的估计产生较大的误差。方向的测量是由 IMU 中的磁力计所进行的，因此方向的测量很容易受到周围磁场的影响而产生不确定性。

在姿态的估计方面。在进行在线姿态估计时改进后的方法会产生较大的抖动误差，且虽然模型所使

用的Δv 对姿态的估计由一定的帮助，但是如果可以确定姿态的初始状态就可以获得人体运动时的速度，对姿态的估计可以产生更大的帮助。

针对以上的问题，可以在未来额外训练一个模型来确定人体姿态的初始状态，之后使用速度数据来进一步提高姿态估计的精确度，同时可以利用人体结构对姿态估计任务产生一定的约束从而提高姿态估计的精确度。

6. 结论

对于人体姿态估计任务提出了一种新型的双重信息保留网络，可以在考虑全局信息的同时利用好历史信息，相比于 RNN 适合人体姿态估计的任务。此外，通过对合成数据的数据增强和平均滤波使其更加拟合真实数据，以及通过对加速度进行积分获得了近似变化速度Δv，以上的方法进一步提高了姿态估计的精确度，最终完成了仅通过 6 个 IMU 完成较高精度的人体姿态估计的任务。之后将尝试训练一个确定人体初始状态的模型来获得运动速度以进一步提高姿态估计的精确度，同时尝试完成仅通过 6 个 IMU 达到较高准确率的人体动作识别任务。

参考文献

[1] Cao, Z., Simon, T., Wei, S.E. and Sheikh, Y. (2017) Real-Time Multi-Person 2D Pose Estimation Using Part Affinity Fields. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 1302-1310. https://doi.org/10.1109/CVPR.2017.143

[2] Güler, R.A., Neverova, N. and Kokkinos, I. (2018) Densepose: Dense Human Pose Estimation in the Wild. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7297-7306. https://doi.org/10.1109/CVPR.2018.00762

[3] Wei, X.L., Zhang, P.Z. and Chai, J.X. (2012) Accurate Realtime Full-Body Motion Capture Using a Single Depth Camera. *ACM Transactions on Graphics*, **31**, 1-12. https://doi.org/10.1145/2366145.2366207

[4] Huang, Y.H., *et al*. (2018) Deep Inertial Poser: Learning to Reconstruct Human Pose from Sparse Inertial Measurements in Real Time. *ACM Transactions on Graphics*, **37**, 1-15. https://doi.org/10.1145/3272127.3275108

[5] Yi, X.Y., Zhou, Y.X. and Xu, F. (2021) Transpose: Real-Time 3D Human Translation and Pose Estimation with Six Inertial Sensors. *ACM Transactions on Graphics*, **40**, 1-13. https://doi.org/10.1145/3450626.3459786

[6] Lei, T., *et al*. (2017) Simple Recurrent Units for Highly Parallelizable Recurrence. arXiv: 1709.02755.

[7] Xia, D., Zhu, Y.Q. and Zhang, H. (2022) Faster Deep Inertial Pose Estimation with Six Inertial Sensors. *Sensors*, **22**, Article 7144. https://doi.org/10.3390/s22197144

[8] Liu, Z.G., *et al*. (2021) Deep Dual Consecutive Network for Human Pose Estimation. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 525-534. https://doi.org/10.1109/CVPR46437.2021.00059

[9] Pavllo, D., Feichtenhofer, C., Grangier, D. and Auli, M. (2019) 3D Human Pose Estimation in Video with Temporal Convolutions and Semi-Supervised Training. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 7745-7754. https://doi.org/10.1109/CVPR.2019.00794

[10] Tome, D., Peluse, P., Agapito, L. and Badino, H. (2019) xR-Egopose: Egocentric 3D Human Pose from an HMD Camera. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 7727-7737. https://doi.org/10.1109/ICCV.2019.00782

[11] Nguyen, H.C., *et al*. (2022) Unified End-to-End YOLOv5-HR-TCM Framework for Automatic 2D/3D Human Pose Estimation for Real-Time Applications. *Sensors*, **22**, Article 5419. https://doi.org/10.3390/s22145419

[12] Von Marcard, T., *et al*. (2018) Recovering Accurate 3d Human Pose in the Wild Using IMUs and a Moving Camera. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018, Springer, Cham, 614-631. https://doi.org/10.1007/978-3-030-01249-6_37

[13] Zhang, Z., Wang, C.Y., Qin, W.H. and Zeng, W.J. (2020) Fusing Wearable IMUs with Multi-View Images for Human Pose Estimation: A Geometric Approach. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 2197-2206. https://doi.org/10.1109/CVPR42600.2020.00227

[14] Gilbert, A., *et al*. (2019) Fusing Visual and Inertial Sensors with Semantics for 3D Human Pose Estimation. *International Journal of Computer Vision*, **127**, 381-397. https://doi.org/10.1007/s11263-018-1118-y

[15] Schepers, M., Giuberti, M. and Bellusci, G. (2018) Xsens MVN: Consistent Tracking of Human Motion Using Inertial Sensing. *Xsens Technologies*, **1**, 1-8.

[16] Loper, M., *et al*. (2023) Smpl: A Skinned Multi-Person Linear Model. *Seminal Graphics Papers*: *Pushing the Boundaries*, **2**, 851-866. https://doi.org/10.1145/3596711.3596800

[17] von Marcard, T., Rosenhahn, B., Black, M.J., Pons-Moll, G., *et al*. (2017) Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs. *Computer Graphics Forum*, **36**, 349-360. https://doi.org/10.1111/cgf.13131

[18] Vaswani, A., *et al*. (2017) Attention Is All You Need. arXiv: 1706.03762.

[19] Sun, Y.T., *et al*. (2023) Retentive Network: A Successor to Transformer for Large Language Models. arXiv: 2307.08621.

[20] Trumble, M., *et al*. (2017) Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors. *Proceedings of* 28*th British Machine Vision Conference*, London, 4-7 September 2017, 1-13.

[21] Mahmood, N., *et al*. (2019) Amass: Archive of Motion Capture as Surface Shapes. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 5441-5450. https://doi.org/10.1109/ICCV.2019.00554

[22] Kingma, D.P. and Ba, J. (2014) Adam: A Method for Stochastic Optimization. arXiv: 1412.6980.