2024年3月
第43卷 第3期

DOI:10.19652/j.cnki.femt.2305600

融合 Transformer 和语义图卷积的三维人体姿态估计方法*

李功浩　贾振堂
(上海电力大学电子与信息工程学院 上海 201306)

摘　要:为了进一步提升从单目二维人体姿态预测三维人体姿态的方法性能,提出一种融合 Transformer 和语义图卷积的三维人体姿态估计模型,模型由 4 个部分组成,Transformer 编码网络、语义图卷积编码网络、姿态坐标预测模块和姿态坐标错误回归模块。首先,Transformer 编码网络对关节特征进行全局特征编码,以增强人体姿态的全局关联性。其次,语义图卷积编码网络专注于局部关节特征提取,以加强局部关节特征之间的关联性。接下来,姿态坐标预测模块和姿态坐标错误回归模块将关节全局和局部编码特征融合,以增强对三维姿态的准确建模能力。通过在 Human3.6M 数据集上进行实验表明,方法在估计性能方面取得了较好的改进,以真实的二维人体姿态作为输入,在 MPJPE 和 PA-MPJPE 值分别为 32.7 和 25.9 mm,与实验对照方法相比,性能分别提升了 3.82%和 1.14%。

关键词:三维人体姿态;语义图卷积;Transformer
中图分类号: TP391.41　　**文献标识码**:A　　**国家标准学科分类代码**: 520.20

3D human pose estimation method fusing Transformer and semantic graph convolution

Li Gonghao　Jia Zhentang
(College of Electronics and Information Engineering, Shanghai University of Electric Power, Shanghai 201306, China)

Abstract: In order to enhance 3D human pose prediction from monocular 2D poses, we propose a model that combines Transformer and semantic graph convolution. The model consists of four components: Transformer encoding network, semantic graph convolutional encoding network, pose coordinate prediction module, and pose coordinate error regression module. The Transformer network captures global joint features to improve posture relevance, while the Semantic Graph Convolutional Encoding Network focuses on local joint feature extraction to enhance correlations. The pose prediction and error regression modules fuse global and local joint features, improving 3D pose accuracy. Experimental results on Human3.6M dataset show significant improvements, achieving MPJPE and PA-MPJPE values of 32.7 and 25.9 mm, respectively, representing a 3.82% and 1.14% improvement over the control method.

Keywords: 3D human pose estimation; semantic graph convolution; Transformer

0　引　言

三维人体姿态估计是计算机视觉领域的一个研究热点,涉及从图像或视频中准确估计人体的三维姿态信息。这项技术在许多应用领域都有广泛应用,例如运动分析[1]、人机交互[2]、虚拟现实[3]和增强现实[4]等。

当前三维人体姿态估计方法大部分基于深度神经网络。现有的三维人体姿态估计方法按照模型处理阶段分为两类。一类是单阶段直接从图像中预测三维姿态。另一类是将三维姿态估计的任务分解为两个独立的阶段:首先,采用基于深度卷积神经网络[5](convolutional neural network,CNN)的二维姿态估计方法[6-8],通过该方法预测图像空间中的关节位置。随后,从二维姿态中学习三维人体姿态。针对两阶段三维姿态估计任务,文献[9]利用全

收稿日期:2023-09-27
* **基金项目**:国家自然科学基金(62105196)项目资助

连接网络(fully connected network,FCN)训练了一组成对的二维-三维数据,其中包括地面真实的二维位置和相应的三维地面真实数据。该研究直接从二维关节信息中推断出三维姿态,与直接从图像中预测的二维姿态作为输入相比,该方法取得了更为卓越的结果。由于人体骨架天然地呈现为一个图结构,学者们对将图神经网络引入三维人体姿态估计表达了浓厚的兴趣。文献[10]提出了一种通用的表达方式,其中既包括图卷积网络[11](graph convolutional network,GCN),也包括 FCN。该方法成功解决了 GCN 在用于估计三维姿态时受限的表示能力问题。文献[12]结合 GCN 和 CNN 的特性,提出了一种语义图卷积网络(semantic graph convolutional network,SemGCN)。该方法在二维人体骨架基础上实对三维人体骨架的回归,将 GCN 对人体静态骨架转变为可以学习的动态骨架。文献[13]分析了图卷积中关节特征聚合和转换方式,提出了一种先对关节特征进行非共享变换,然后结合描述人体骨架的矩阵实现对关节特征的聚合。该方法解决了 GCN 仅通过共享权重来建模相邻节点之间关系的局限性,并关注不同关节的重要性。文献[14]提出多尺度的关节特征编解码结构来学习不同尺度特征,解决了 GCN 提取单一尺度关节特征问题。文献[15]提出自调节图卷积 UNet 网络,通过可学习矩阵自动调节人体骨架关系,并结合 UNet[16]多尺度关节点池化和去池化网络,实现了对关节多尺度特征提取。文献[17]提出预加权调制密集图卷积网络,解决了 GCN 对人体三维姿态建模存在过平滑和未区分关节点与相邻关节点重要性的问题。文献[18]受自注意力机制启发,从全局的角度对人体姿态进行建模,提出 JF(Jointformer)方法。该方法在不使用 GCN 方法的情况下实现了三维人体姿态的估计,并且展现出较好的估计效果,为二维人体姿态到三维人体姿态估计方法提供了新的思路。

本文受到 JF 和 SemGCN 方法的启发,将局部关节特征和全局关节特征中综合考虑建模三维人体姿态,提出一种融合 Transformer 和语义图卷积的三维人体姿态估计模型。首先,对离线的二维姿态检测器[7]预测的二维人体关节点坐标进行高维度特征向量嵌入。然后对关节嵌入特征进行全局和局部建模。其中包括对关节坐标嵌入特征向量进行全局信息提取的 Transformer 编码网络以及对关节特征进行姿态局部特征提取的语义图卷积编码网络。最后将 Transformer 全局信息编码和语义图卷积局部特征信息编码融合输入姿态坐标预测和姿态错误回归模块,实现三维人体姿态估计。为进一步提升姿态估计性能,本文结合了三维人体姿态的预测结果和离线二维人体姿态检测器[7]的输出,以进行三维姿态微调,进一步提出一种融合 Transformer 和语义图卷积的微调模型。为验证方法的有效性,在 Human3.6 M[19]数据集进行对比实验,实验结果表明,本文模型性能相较于文中实验对照方法取得较好的改进。

1 融合 Transformer 和语义图卷积的三维人体姿态估计模型

本文提出了一种将 Transformer 与语义图卷积相融合的三维人体姿态估计模型。该方法使用离线二维姿态检测模型从图像中获取二维人体姿态的估计结果[7],并将这些结果应用于在相机坐标系下以骨盆关节点为中心的三维人体姿态的预测。首先,本文使用基础模型直接从单帧的二维人体姿态中学习回归三维人体姿态。然后,在微调模型中,利用基础模型回归的三维姿态和二维姿态检测器预测的二维姿态结果[7]进一步微调三维人体姿态。实验证明,本文提出的融合 Transformer 和语义图卷积的三维人体估计模型在单帧三维人体姿态估计任务中表现出良好的性能。模型结构如图 1 所示。

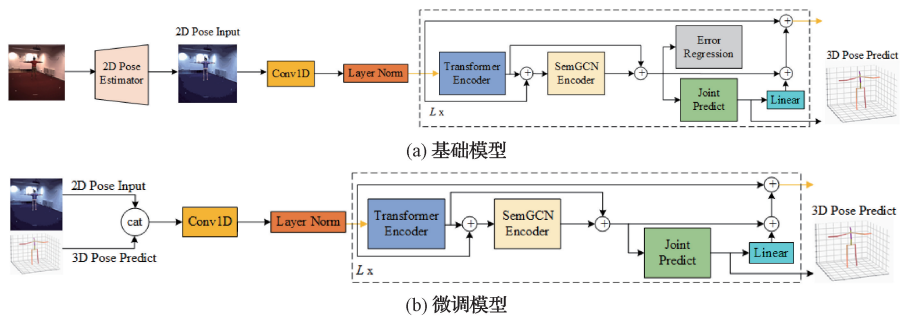

图 1 融合 Transformer 和语义图卷积的三维人体姿态估计模型结构

1.1 语义图卷积编码网络

给定一个人体骨架结构 $S=\{J,E\}$,关节点集合用字母 J 表示,边的集合用字母 E 表示。其中字母 E 中的连接关系可以用邻接矩阵 $\boldsymbol{A}\in\{0,1\}^{N\times N}$ 表示,N 表示关节点的数量。每个关节点都具有一个 C 维的属性向量 $\boldsymbol{h}_i\in\boldsymbol{R}^C$。这些关节点属性组成了描述关节点属性的特征矩阵 $\boldsymbol{H}^l\in\boldsymbol{R}^{C\times N}$。本文引入文献[12]中 SemGCN 公式:

$$\boldsymbol{H}^{(l+1)}=\sigma(\boldsymbol{W}\boldsymbol{H}^{(l)}\rho_i(\boldsymbol{M}\odot\boldsymbol{A}')) \tag{1}$$

$$\boldsymbol{A}'=\boldsymbol{D}^{-1/2}(\boldsymbol{A}+\boldsymbol{I})\boldsymbol{D}^{1/2} \tag{2}$$

式中:$\boldsymbol{A}'$ 是 $\boldsymbol{A}$ 经过 $\boldsymbol{A}+\boldsymbol{I}$ 自连接之后再经过 $\boldsymbol{A}+\boldsymbol{I}$ 的度矩阵 $\boldsymbol{D}$ 归一化后的一阶邻接矩阵;$\boldsymbol{M}\in\boldsymbol{R}^{K\times K}$ 是可学习加权矩阵;ρ_i 是用于对节点 i 的输入矩阵进行 Softmax 非线性

归一化；符号⊙表示矩阵元素对应相乘；$\boldsymbol{W} \in \boldsymbol{R}^{C\times C}$ 是可学习的线性变换参数矩阵；σ 是非线性激活函数 ReLU。由于人体骨架模型中存在一些远端关节，例如腕、踝和头部等。它们位于人体骨架的末端，与其他关节只有一个一级相邻关节连接。普通的 SemGCN 仅关注关节点的一阶邻域，导致对这些远端关节在空间中进行准确定位变得相当困难。实际上，这些远端关节通常是导致姿态建模误差的主要来源之一。在文献[20]的启发下，本文考虑了描述人体骨架的自连接邻接归一化矩阵 $\boldsymbol{A}'$，同时也考虑了以躯干为中心的人体对称结构和人体的运动学约束。首先利用一阶邻接矩阵 $\boldsymbol{A}'$（first order adjacency matrix，FO）自身。接下来确定人体骨架对称节点在一阶邻接矩阵 $\boldsymbol{A}'$ 中的位置，将对称节点的位置元素值设为 1，非对称节点的位置元素值设为 0，从而获得了人体骨架对称矩阵（symmetry matrix，SYM），表示为 $\boldsymbol{A}_{sym} \in \{0,1\}^{N\times N}$。最后，在 $\boldsymbol{A}'$ 的基础上，考虑人体骨架各远端节点，将一阶邻接矩阵中远端节点的位置元素值设为 0，非远端节点的元素值保持不变，从而构建了非远端节点的一阶邻接矩阵 $\boldsymbol{A}_1$。接着，进一步考虑了二阶邻接矩阵中远端节点的位置元素值保持不变，非远端节点的元素值设为 0，构建了远端节点的二阶邻接矩阵 $\boldsymbol{A}_2$，将 $\boldsymbol{A}_1$ 和 $\boldsymbol{A}_2$ 求和得到一阶二阶邻域联合矩阵 $\boldsymbol{A}_3$（first-order second-order neighborhood union matrix，FSO）。通过构建一阶二阶邻域联合矩阵明确编码了远端关节二级连接，而对于其余节点，仅通过一阶连接进行建模，以减小远端关节建模可能带来的误差。针对以上一阶邻接矩阵，对称矩阵和一阶二阶邻域联合矩阵分别对关节特征表示进行语义图卷积。最后，进行一维批次标准化、应用非线性激活（ReLU），然后将 3 种不同的语义特征拼接并融合。依次经过一维卷积、一维批次标准化、非线性激活和随机失活实现高级关节语义特征提取。本文在语义图卷积编码网络部分堆叠 L 层语义图卷积用于多次提取关节特征，获取更高级的节点语义特征表示，其结构如图 2 所示。

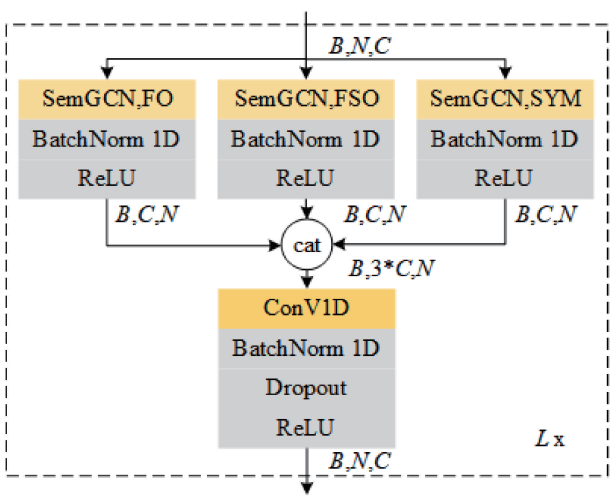

图 2　语义图卷积编码网络

1.2　融合 Transformer 和语义图卷积的三维人体姿态估计基础模型

本文提出的融合 Transformer 和语义图卷积的三维人体估计基础模型直接从二维人体姿态预测三维人体姿态，其模型结构如图 1(a)所示。在得到了离线二维人体姿态检测器[7]输出的关节点坐标集合 $x \in R^{J\times 2}$ 之后，首先将这些关节点坐标进行高维度特征向量嵌入。本文参考 JF 的方法，使用一维卷积对关节点坐标进行高维度特征嵌入得到一个表示为 $f_e(x):R^{J\times 2} \rightarrow R^{J\times c}$ 的结果，并随后对其进行层级标准化。接下来，对这些关节特征进行了全局和局部特征编码，最终用于三维人体姿态的预测。在编码过程中，保持关节点坐标的高维特征嵌入尺寸不变，即嵌入尺寸为 $c = 64$，并采用中间监督的方法逐层建立 L 层模型，以进行三维人体姿态的估计。首先，本文引用文献[21]Transformer 的编码网络，对关节嵌入特征进行全局特征编码，表示为 $f_g(f_e):R^{J\times c} \rightarrow R^{J\times c}$。接下来，使用图 2 语义图卷积编码网络，对关节的全局编码信息以及初始关节嵌入特征进行局部语义信息提取，表示为 $f_l(f_e + f_g):R^{J\times c} \rightarrow R^{J\times c}$。最后，将全局和局部编码信息融合，分别输入到三维坐标预测模块（图 3(a)）和坐标错误回归模块（图 3(b)），以实现三维人体姿态估计。具体而言，三维坐标预测表示为 $f_{pred}(f_g + f_l):R^{J\times c} \rightarrow R^{J\times 3}$，而姿态坐标错误回归表示为 $f_{err}(f_g + f_l):R^{J\times c} \rightarrow R^{J\times 3}$。为了建立三维人体预测模块和编码网络之间的层级关联，将第 L 层的预测关节坐标进行了关节特征嵌入，表示为 $f_{e\_pred}(f_{pred}):R^{J\times 3} \rightarrow R^{J\times c}$，然后将其与全局和局部编码特征以及初始关节特征嵌入进行融合，用于第 $L+1$ 层的输入关节坐标嵌入特征表示，表示为 $f_e = (f_e + f_g + f_l + f_{e\_pred})$，从而进行第 $L+1$ 层三维人体姿态预测。基础模型基于中间学习建立 4 层三维人体姿态估计网络，即 $L = 4$，同时语义图卷积编码网络堆叠层数也为 4。

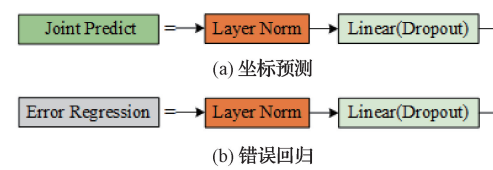

图 3　姿态坐标预测和姿态错误回归模块

1.3　融合 Transformer 和语义图卷积的三维人体姿态估计微调模型

基础模型通过中间监督学习预测了三维坐标以及坐标错误回归部分，从而获得了较好的三维人体姿态估计结果。本文进一步提出了融合 Transformer 和语义图卷积的微调模型，用于微调基础模型估计的人体姿态，其模型结构如图 1(b)所示。将基础模型学习到的模型参数传递给微调模型，然后将其二维人体姿态输入和最后一层预测的三维人体姿态拼接，作为微调模型的输入。接着，借鉴

JF模型的方法，对高维度关节特征进行嵌入，且嵌入尺寸 $c=160$，得到 $f_{e\_r}(x):R^{J\times 5}\rightarrow R^{J\times c}$ 并进行层级标准化。随后，对这些关节高维度嵌入特征进行全局和局部信息编码。在第 L 层对关节特征编码过程中，本文首先引入文献[21]中 Transformer 的编码网络对关节嵌入特征进行全局特征编码，得到 $f_{g\_r}(f_{g\_r}):R^{J\times c}\rightarrow R^{J\times c}$，然后使用图2语义图卷积编码网络对关节的全局编码信息以及初始关节特征嵌入进行局部语义信息提取，表示为 $f_{l\_r}(f_{e\_r}+f_{g\_r}):R^{J\times c}\rightarrow R^{J\times c}$。接下来将关节的全局编码信息和局部语义信息进行融合，然后进行三维坐标的预测，其模块结构如图3(a)所示，结果表示为 $f_{pred\_r}(f_{g\_r}+f_{l\_r}):R^{J\times c}\rightarrow R^{J\times 3}$。为了建立三维人体预测模块和编码网络之间的层级关联，对第L层的预测关节坐标进行了关节特征嵌入，表示为 $f_{e\_pred\_r}(f_{pred\_r}):R^{J\times 3}\rightarrow R^{J\times c}$，然后将其与全局和局部编码特征以及初始关节特征嵌入进行融合，用于第 $L+1$ 层的输入关节坐标嵌入特征表示，表示为 $f_{e\_r}(f_{e\_r}+f_{g\_r}+f_{l\_r}+f_{e\_pred\_r})$，进而开展第 $L+1$ 层的三维人体姿态预测。微调模型基于中间学习建立两层三维人体姿态估计网络，即 $L=2$，且语义图卷积堆叠层数为1。

1.4 损失函数

本文的模型训练学习分为两个部分：一个是基础模型的损失，另一个是微调模型的损失。总体而言，这两部分的损失函数都用于度量模型预测与真实人体17个关节点之间的差异，即 $n=17$。

基础模型总损失 L_{base} 包括了第 L 层中间监督学习的三维人体坐标预测损失 L_{p_loss} 和坐标错误回归损失 L_{e_loss}。具体来说，关节坐标预测损失是预测值 Y_i 和真值误差 Y'_i 之间的均方误差。而关节坐标错误回归损失是预测坐标错误 E_i 和预测值及真实值之间的差值绝对值的均方误差，其中差值绝对值表示为 $|Y_i-Y'_i|$。坐标预测损失和错误回归损失分别如下：

$$L_{p\_loss}=\frac{1}{n}\sum_{i=1}^{n}(Y_i-Y'_i)^2 \tag{3}$$

$$L_{e\_loss}=\frac{1}{n}\sum_{i=1}^{n}(E_i-|Y_i-Y'_i|)^2 \tag{4}$$

基础模型总损失 L_{base} 是基于中间监督学习构建的坐标预测损失 L_{p_loss} 和坐标错误回归损失 L_{e_loss} 求和均值，然后对所有层级求和并取平均，其公式表示如下：

$$L_{base}=\frac{1}{2L}\sum_{i=1}^{L}(L_{p\_loss}+L_{e\_loss}) \tag{5}$$

微调模型损失用 L_r 描述。其损失包含 L 层三维人体姿态和真实姿态之间的差异，采用均方误差函数建立真值 Y'_i 和预测值 Y_i 之间的误差关系，其第 L 层损失公式表示如下：

$$L_{r\_loss}=\frac{1}{n}\sum_{i=1}^{n}(Y_i-Y'_i)^2 \tag{6}$$

而微调模型损失描述为所有层损失 L_{r_loss} 求和的平均值，其公式表示如下：

$$L_r=\frac{1}{L}\sum_{i=1}^{L}L_{r\_loss} \tag{7}$$

2 实验

2.1 实验环境及实施细节

本文方法是基于 Pytorch[22] 框架构建的，并使用英伟达(NVIDIA GeForce RTX3060)进行实验训练和评估。融合 Transformer 和语义图卷积的三维人体估计模型的超参数设置如下：批量大小为256，基础模型的训练数据迭代次数为50，而微调模型的迭代次数设置为30。模型参数更新优化采用 AdamW[23] 优化器，设置初始学习率为0.001，并使用余弦退火学习率衰减策略[24]。在训练中，本文采用了不同的随机失活率：基础模型的所有预测层使用0.2，微调模型使用0.1，语义图编码网络的输出层使用0.05。模型在实验评估中采用了水平左右翻转[25]的数据增强，以充分验证模型的性能。

2.2 数据集及评价指标

本文实验部分采用受欢迎的三维人体姿态数据集 Human3.6 M[19] 进行实验评估。该数据集由7个专业演员对15个通用动作进行扮演，通过4个不同相机视角拍摄人体三维姿态捕获系统所构建。根据官方数据集分割标准，使用编号为1、5、6、7、8的专业演员作为训练数据，使用编号为9、11的专业演员作为评估数据。数据包括演员扮演动作从不同视角所拍摄的图片以及图片所处视角对应的三维信息和三维空间到图像空间的相互转换参数。

根据 Human3.6 M 数据集评估方法有效性的常见做法，将相机坐标系下三维人体姿态关节点坐标对准其根关节后，使用真实三维关节位置和预测三维关节位置的平均欧氏距离(MPJPE)进行评估，其计算公式如下：

$$E_{MPJPE}=\frac{1}{N_S}\sum_{i=1}^{N_S}\|m_{gt,s}(i)-m_{pre,s}(i)\|_2 \tag{8}$$

式中：E_{MPJPE} 为误差值；S 为骨架结构，本文采用级联金字塔网络(cascaded pyramid network，CPN)[7] 生成的姿态17个关节点的骨架结构；N_S 为对应骨架结构的关节点数；$m_{gt,s}(i)$ 和 $m_{pre,s}(i)$ 分别表示骨架结构 S 第 i 个关节的真实值和预测值。同时，使用文献所提出的刚性变换[10]将预测三维关节位置与真实三维关节位置对齐之后，再计算真实三维关节位置和预测三维关节对齐之后位置的平均欧氏距离(PA-MPJPE)进行评估。

2.3 实验结果与分析

本文方法在三维人体姿态估计数据集 Human3.6 M 上进行评估，在进行评估分析时直接从二维人体姿态中回归三维人体姿态。

首先调查本文提出方法对来自于野外二维人体姿态检测器预测的二维人体姿态作为输入来评估模型性能。其中表1为本文在 Human3.6 M 测试数据上的实验结果。这些结果是基于离线二维人体检测器[7]的人体姿态

预测结果，然后通过本文方法进行了评估，以获得相应的三维人体姿态估计结果。表 1 中数据呈现是相关方法在 Human3.6 M 数据集 15 个动作姿态预测在 MPJPE 指标上的评估结果和对应 15 个动作的平均 MPJPE 指标。本文提出方法在 15 个动作平均 MPJPE 超过基于单视图的三维人体姿态估计 JF 方法 0.8 mm，提升 1.58%。

表 1　不同方法 CPN 二维姿态输入 MPJPE 评估对比

方法	方向	讨论	进食	问候	打电话	拍照	姿势	购物	坐着	坐下	吸烟	等待	遛狗	步行	同行走	平均值
文献[9]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
文献[12]	48.2	60.8	51.8	64.0	64.6	53.6	51.1	67.4	88.7	57.7	73.2	65.6	48.9	64.8	51.9	60.8
文献[10]	46.8	52.3	44.7	50.4	52.9	68.9	49.6	46.4	60.2	78.9	51.2	50.0	54.8	40.4	43.3	52.7
文献[13]	46.3	52.2	47.3	50.7	55.5	67.1	49.2	46.0	60.4	71.1	51.5	50.1	54.5	40.3	43.7	52.4
文献[14]	45.2	49.9	47.5	50.9	54.9	66.1	48.5	46.3	59.7	71.5	51.4	48.6	53.9	39.9	44.1	51.9
文献[18]	45.0	48.8	46.6	49.4	53.2	60.1	47.0	46.7	59.6	67.1	51.2	47.1	53.8	39.4	42.4	50.5
本文	44.5	49.1	45.2	50.0	52.8	58.8	47.1	45.5	58.0	64.3	50.4	47.3	52.6	38.3	41.1	49.7
本文微调	44.8	50.7	45.2	49.3	51.9	57.2	48.2	46.7	57.3	62.6	49.9	47.4	52.6	38.0	40.8	49.5

为进一步验证模型性能，本文对模型输入二维人体姿态，预测和真实三维人体姿态进行可视化呈现，其结果如图 4 所示，包含模型 JF 和本文模型在 Human3.6 M 数据集子动作拍照和姿势动作的比较。通过对比拍照动作的可视化结果，在输入二维人体姿态内侧右手臂被遮挡的情况下，JF 未能对其三维姿态进行较好预测，而本文方法预测的姿态和真实三维姿态更为接近。在比较姿势动作可视化结果时，输入二维姿态内侧右手臂被遮挡的情况下，JF 将左右手臂完全分离，与本文预测结果和真实姿态相差较大，其手臂明显变短且未能表现出演员所呈现的姿态。通过对比二维检测器输入条件下的可视化结果显示，本文方法能够解决部分人体自遮挡问题，在二维输入姿态呈现遮挡问题时，方法预测结果可以对其输入遮挡姿态进行矫正，恢复三维人体姿态。

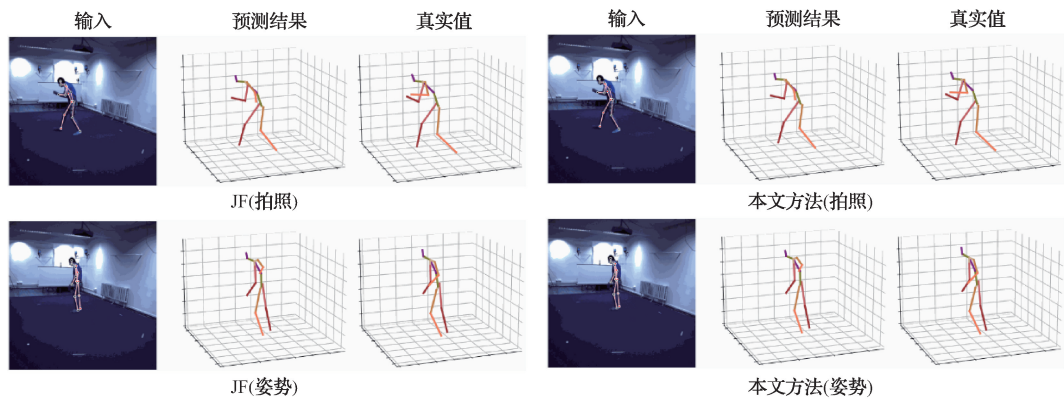

图 4　JF 和本文方法 CPN 二维姿态输入可视化对比

其次，为充分验证本文提出方法的可行性，本文用真实的二维人体姿态作为输入预测三维人体姿态。本文提出方法在 Human3.6 M 数据集 15 个专业动作评估平均 MPJPE 指标超过 JF 方法 1.3 mm，提升 3.82%。其 15 个通用动作在 MPJPE 指标评估结果以及所有动作的平均 MPJPE 如表 2 所示。

为充分验证模型对真实二维姿态输入预测三维人体姿态的性能，本文对来自于真实的二维人体姿态和三维预测结果以及真实三维人体姿态进行可视化结果如图 5 所示，包括 JF 模型和本文提出模型在 Human3.6 M 数据集子动作拍照的可视化比较结果。从 JF 模型可视化结果可以看出预测的三维姿态和真实的三维姿态区别更大相较于本文方法。在 JF 直立拍照动作中，真实左手臂姿态应与三维空间网格水平参考线具备一定夹角，其预测结果几乎和网格水平参考线重合。而本文方法与网格水平参考线呈现一定角度，预测的姿态更接近真实形态。在 JF 弯腰屈膝拍照的动作中，左右手臂手肘应该存在一定交叉可见形态，其呈现的结果是左手小臂将右手手肘遮挡，与真实的三维姿态存在一定的偏差。而本文呈现的预测结果能够更加接近真实姿态，手肘之间呈现交叉可见的效果。通过本文对真实的二维姿态输入预测三维姿态的可视化结果对比显示，本文方法呈现的三维姿态更接近真实形态，姿态估计结果更好。

表 2　不同方法真实二维姿态输入 MPJPE 评估对比

方法	方向	讨论	进食	问候	打电话	拍照	姿势	购物	坐着	坐下	吸烟	等待	遛狗	步行	同行走	平均值
文献[9]	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
文献[10]	36.3	38.8	29.7	37.8	34.6	42.5	39.8	32.5	36.2	39.5	34.4	38.4	38.2	31.3	34.2	36.3
文献[12]	37.8	49.4	37.6	40.9	45.1	41.4	40.1	48.3	50.1	42.2	53.5	44.3	40.5	47.3	39.0	43.8
文献[13]	36.8	40.3	33.0	36.3	37.5	45.0	39.7	34.9	40.3	47.7	37.4	38.5	38.6	29.6	32.0	37.8
文献[14]	35.8	38.1	31.0	35.3	35.8	43.2	37.3	31.7	38.4	45.5	35.4	36.7	36.8	27.9	30.7	35.8
文献[18]	31.0	36.6	30.2	33.4	33.5	39.0	37.1	31.3	37.1	40.1	33.8	33.5	35.0	28.7	29.1	34.0
本文	31.6	35.4	28.4	32.9	33.2	36.1	35.5	29.0	33.4	38.9	32.7	33.6	33.5	27.2	29.3	32.7
本文微调	32.6	35.9	28.7	32.3	33.5	34.9	36.4	30.0	33.4	38.1	33.5	34.3	32.8	27.1	29.6	32.9

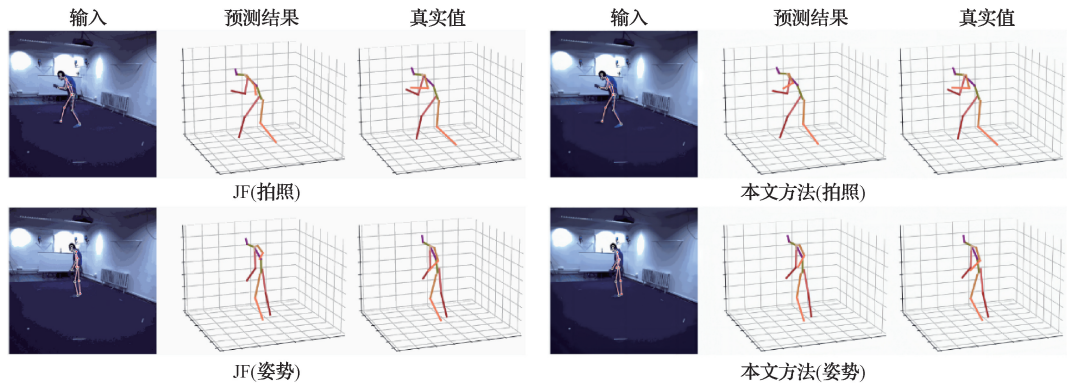

图 5　JF 和本文方法真实二维姿态输入可视化对比

2.4　消融实验

首先，对本文方法和 JF 方法进行实验比较，其实验结果如表 3 所示。在基础阶段，本文方法的参数量相比 JF 多了 0.6×10^6，而在微调阶段则减少了 1.74×10^6。此外，在 MPJPE 和 PA-MPJPE 指标上，本文方法表现更为优越。具体表现在 CPN 的二维姿态输入条件下，基础阶段相较于 JF 在 MPJPE 提升 0.8 mm，PA-MPJPE 提升 0.1 mm。而相较于 JF 微调阶段本文基础模型在 MPJPE 指标上仍然领先 0.4 mm。在真实二维姿态作为输入的条件下比 JF 基础阶段 MPJPE 指标提升 1.3 mm，PA-MPJPE 提升 0.3 mm。相较于 JF 微调阶段 MPJPE 提升 1.5 mm。通过对本文提出的基础模型和 JF 基础和微调模型的实验结果分析，本文融合 Transformer 和语义图卷积的三维人体估计基础模型性能相较于 JF 模型性能要更好。

表 3　本文方法基础模型和 JF 实验比较

CPN 二维姿态输入	MPJPE（平均值）	PA-MPJPE（平均值）	参数量	真实二维姿态输入	MPJPE（平均值）	PA MPJPE（平均值）
JF	50.5	39.4	1.48×10^6	JF	34.0	26.2
本文方法	49.7	39.3	2.08×10^6	本文方法	32.7	25.9
JF(微调)	50.1	39.0	3.82×10^6	JF(微调)	34.2	—

其次，对本文方法中语义图卷积编码网络图堆叠层数在二维姿态检测器的结果作为输入的条件下，基于融合 Transformer 和语义图卷积的三维人体估计基础模型进行消融实验，其实验结果如表 4 所示。从表 4 可以得知在基础阶段语义图卷编码网络堆叠层数为 4 时在 MPJPE 指标的评估结果表现最好，相较于 JF 基础模型在 MPJPE 指标提升了 0.8 mm。

表 4　本文方法语义图卷积编码网络堆叠层数消融试验

堆叠层数	参数量（$\times10^6$）	MPJPE（平均值）	PA-MPJPE（平均值）
一层	1.63	50.5	39.2
两层	1.78	49.8	38.8
三层	1.93	50.3	38.8
四层	2.08	49.7	39.2

3 结 论

本文提出一个融合 Transformer 和语义图卷积的三维人体姿态估计模型，由 Transformer 编码网络、语义图卷积编码网络、姿态坐标预测模块和姿态坐标错误回归模块组成。通过实验证明，与相关基于单帧的三维人体姿态估计方法相比，本文提出的模型表现出更好的性能优势。目前本文模型处理基于单帧的人体二维姿态直接回归预测三维人体姿态表现较好效果，但是与处理视频或者多视图模型相比，性能相差较大，因此探索将本文模型拓展到视频或多视图领域将成为后续的研究工作。

参 考 文 献

[1] 孙文昊，路光达，秦转萍，等. 引入 GAN 与可变形注意力的多维人体运动分析 [J]. 电子测量技术，2023，46(16)：78-88.

[2] 孟杰，杨鹏程，杨朝，等. 基于 Mediapipe 的幻影成像装置自然手势交互系统设计 [J]. 国外电子测量技术，2023，42(3)：116-122.

[3] 陆熊，孙东，鄢昱星，等. 基于电磁力控制的非接触式二维力触觉再现系统 [J]. 仪器仪表学报，2022，43(7)：174-180.

[4] 吴国新，左云波，秦文丽，等. 工业室内环境中建立增强现实系统模型研究 [J]. 电子测量与仪器学报，2021，35(5)：196-201.

[5] LECUN Y，BENGIO Y，HINTON G. Deep learning[J]. Nature，2015，521(7553)：436-444.

[6] SUN K，XIAO B，LIU D，et al. Deep high-resolution representation learning for human pose estimation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2019：5693-5703.

[7] CHEN Y，WANG Z，PENG Y，et al. Cascaded pyramid network for multi-person pose estimation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2018：7103-7112.

[8] NEWELL A，YANG K，DENG J. Stacked hourglass networks for human pose estimation[C]. Computer Vision-ECCV 2016：14th European Conference. Springer，2016：483-499.

[9] MARTINEZ J，HOSSAIN R，ROMERO J，et al. A simple yet effective baseline for 3d human pose estimation[C]. Proceedings of the IEEE International Conference on Computer Vision，2017：2640-2649.

[10] CI H，WANG C，MA X，et al. Optimizing network structure for 3d human pose estimation[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision，2019：2262-2271.

[11] KIPF T N，WELLING M. Semi-supervised classification with graph convolutional networks[C]. International Conference on Learning Representations，2017.

[12] ZHAO L，PENG X，TIAN Y，et al. Semantic graph convolutional networks for 3d human pose regression[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2019：3425-3435.

[13] LIU K，DING R，ZOU Z，et al. A comprehensive study of weight sharing in graph networks for 3d human pose estimation[C]. Computer Vision-ECCV 2020：16th European Conference. Springer，2020：318-334.

[14] XU T，TAKANO W. Graph stacked hourglass networks for 3d human pose estimation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2021：16105-16114.

[15] 马金林，崔琦磊，马自萍，等. 自调节图卷积 UNet 的三维人体姿态估计方法 [J/OL]. 北京航空航天大学学报，1-15[2023-12-16] https://doi.org/10.13700/j.bh.1001-5965.2022.0969.

[16] RONNEBERGER O，FISCHER P，BROX T. U-net：Convolutional networks for biomedical image segmentation[C]. Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015：18th International Conference. Springer，2015：234-241.

[17] 马金林，崔琦磊，马自萍，等. 预加权调制密集图卷积网络三维人体姿态估计[J/OL]. 计算机科学与探索，1-17[2023-12-16]http://kns.cnki.net/kcms/detail/11.5602.tp.20230606.1330.002.html.

[18] LUTZ S，BLYTHMAN R，GHOSAL K，et al. Jointformer：Single-frame lifting transformer with error prediction and refinement for 3d human pose estimation[C]. 26th International Conference on Pattern Recognition(ICPR). IEEE，2022：1156-1163.

[19] 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.

[20] LIU J，ROJAS J，LI Y，et al. A graph attention spatio-temporal convolutional network for 3D human pose estimation in video[C]. 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE，2021：3374-3380.

[21] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]. Advances in Neural Information Processing Systems，2017.

[22] PASZKE A，GROSS S，MASSA F，et al. Pytorch：

An imperative style, high-performance deep learning library[C]. Advances in Neural Information Processing Systems, 2019.

[23] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[C]. International Conference on Learning Representations, 2019.

[24] LOSHCHILOV I, HUTTER F. SGDR: Stochastic gradient descent with warm restarts[C]. International Conference on Learning Representations, 2017.

[25] PAVLLO D, FEICHTENHOFER C, GRANGIER D, et al. 3d human pose estimation in video with temporal convolutions and semi-supervised training[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 7753-7762.

作 者 简 介

李功浩，硕士研究生，主要研究方向为三维人体姿态估计。

E-mail：haogongli@163.com

贾振堂（通信作者），博士，副教授，主要研究方向为智能视频监控（涉及多模态深度学习、目标识别、人体姿态分析、立体视觉等）。

E-mail：462458081@qq.com