计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(8):250−256 [doi: 10.15888/j.cnki.csa.009602]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多层空间特征融合的三维人体姿态估计①

梁桉源, 肖学中

(南京邮电大学 计算机学院, 南京 210023)
通信作者: 梁桉源, E-mail: 1228153458@qq.com

摘　要: 在三维人体姿态估计任务当中, 人体关节之间的连接关系形成了一种复杂的拓扑结构, 利用图卷积网络对该结构进行建模, 可以有效捕捉局部关节间的联系; 尽管不相邻关节之间没有直接的物理连接, 但由于人体的运动和姿态受到生物力学约束以及人体关节之间的协同作用, 利用 Transformer 编码器建立关节之间的上下文关系, 可以更好地推断出人体姿态; 在大模型的背景下, 如何在保证模型性能的同时, 降低参数量, 也显得尤为重要. 针对上述问题, 设计了一个基于图卷积和 Transformer 的多层空间特征融合网络模型 (MLSFFN), 在使用相对少量的参数基础上, 有效地融合了局部和全局空间特征. 实验结果表明, 本文提出的方法在仅需 2.1M 参数量的情况下, 在 Human3.6M 数据集上达到了 49.9 mm 的平均每关节误差 (MPJPE). 此外, 模型在 MPI-INF-3DHP 数据集上也展示出了较强的泛化能力.

关键词: 多层空间特征融合; 三维人体姿态估计; 图卷积网络; Transformer; 轻量型

引用格式: 梁桉源,肖学中.基于多层空间特征融合的三维人体姿态估计.计算机系统应用,2024,33(8):250–256. http://www.c-s-a.org.cn/1003-3254/9602.html

3D Human Pose Estimation Based on Multi-layer Spatial Feature Fusion

LIANG An-Yuan, XIAO Xue-Zhong

(School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: In the task of 3D human pose estimation, the complex topology formed by the connection relationship between human joints presents a challenge. Effective capture of the connections between local joints is possible through modeling this structure with a graph convolutional network. Although non-adjacent joints lack direct physical connections, Transformer encoders establish contextual relationships between joints, which is crucial for better human posture inference due to the biomechanical constraints influencing human motion and pose, as well as the synergistic interaction of human joints. Balancing model performance with a reduction in the number of parameters is of particular importance for large-scale models. To tackle these challenges, a multi-layer spatial feature fusion network model (MLSFFN) based on graph convolution and Transformer is designed. This model proficiently fuses local and global spatial features with a relatively minimal parameter set. Experimental results demonstrate that the proposed method achieves a mean point per joint error (MPJPE) of 49.9 mm on the Human3.6M dataset with only 2.1M parameters. Moreover, the model demonstrates a robust generalization capability.

Key words: multi-layer spatial feature fusion; 3D human pose estimation; graph convolutional network (GCN); Transformer; lightweight

① 收稿时间: 2024-02-26; 修改时间: 2024-03-28; 采用时间: 2024-04-10; csa 在线出版时间: 2024-06-28
CNKI 网络首发时间: 2024-07-02

人体姿态估计 (human pose estimation, HPE) 在计算机视觉领域一直是研究热点, 尤其在三维人体姿态估计方面, 对人机交互、康复治疗以及增强现实等应用领域产生深远影响. 在过去几年中, 基于深度学习的三维人体姿态估计方法取得了显著的进展, 但是实现对三维人体姿态的精确估计仍然是一个具有挑战性的问题.

基于单帧视频的三维人体姿态估计的关键在于空间特征的提取及处理, 其效率与准确度直接决定了模型的性能表现. 当前, 图卷积网络 (graph convolutional network, GCN) 与 Transformer 模型成为处理此类问题的两种主流技术, 它们分别通过图结构和自注意力机制进行空间特征的深度挖掘. 然而, 源自频谱卷积的原始 GCN 存在一个重要问题, 即权重共享卷积. 在原始的 GCN 模型中, 每个节点都通过相同的转换矩阵进行转换, 然后邻近特征将被聚合, 用以传输到下一层[1]. 但这种权重共享方法并不能有效捕捉到人体关节的复杂信息, 因为人体运动的灵活性和速度会随着关节的变化而变化. 此外, 其中还存在着 3D HPE 中关节间非局部依赖性的问题[2–5]. Transformer 中的自注意力机制能够捕捉特征的全局信息. 在 3D HPE 中, 自注意力机制能够帮助关节与其他关节建立联系, 从而获得全局依赖性. 这种全局依赖性适合模拟关节间的相似性关系, 因为它能够随着输入姿态的变化而进行相应的调整.然而, 纯粹的 Transformer 模型也有其局限性, 它可能会忽视人体骨骼中的物理信息[6].

大多数现有的研究往往只依赖于其中一种技术, 忽略了 GCN 与 Transformer 间可能存在的互补优势. 此外, 如何有效自适应局部和全局空间特征的权重也是一个重要问题. 局部和全局空间特征在模型处理过程中扮演着不同的角色, 缺乏适应性的权重调整可能会导致模型处理这两种特征的能力不足, 进一步影响模型性能. 而且 Transformer 编码器的参数量呈指数级增长, 对模型的计算效率和存储需求也提出了严峻的挑战. 因此, 如何在保证模型性能的前提下, 更加合理地设计模型结构以减少参数量, 成为当前研究的一个重要任务.

为了克服上述这些问题, 同时充分利用人体骨骼中关节的局部和非局部关系, 我们提出了一种新的架构. 这种架构将 GCN 和 Transformer 编码器两个独立的模块交替使用, 混合并利用局部和全局依赖性的信息. 在这种架构下, GCN 模块和 Transformer 编码器可以交互地进行信息处理, 使得模型在处理关节间的复杂关系时, 既保留了局部的精细信息, 又捕捉到了全局的依赖性, 从而提高了模型的性能和准确度.

本文的主要贡献包括: 提出了一种新颖的结构设计, 将 GCN 和 Transformer 编码器设计成 6 层结构, 交替混合提取局部和全局空间特征后, 使用 SE 模块进一步突出重要的空间特征, 在使用相对少量参数的同时提升了模型的效果. 并且在 Human3.6M 数据集中进行了大量实验, 结果表明, 本文提出的模型 MLSFFN 在各项指标上均优于与之比较的 7 个模型, 并在 MPI-INF-3DHP 数据集上进一步验证其泛化性.

1 相关工作

三维人体姿态估计是计算机视觉领域的一个重要研究方向, 它的目标是从二维图像或视频中估计出人体的三维关键点信息. 这个问题具有很大的挑战性. 因为人体姿态的变化非常复杂, 而且图像的投影会导致深度信息的丢失. 本文研究的内容是单视角单人三维人体姿态估计, 根据是否使用二维姿态结果作为中间表示, 可以分为直接估计法和二维提升到三维两种方法. 本节回顾了基于深度学习的单视角单人三维人体姿态估计的一些相关工作.

1.1 基于直接估计的方法

直接估计法即直接从二维图像中推断出三维人体姿态, 而不需要中间估计 2D 姿态表示. 早期的工作应用了回归范式来直接估计三维人体姿态. Li 等人[7]最初将端到端的方法与深度神经网络相结合, 通过关节检测器和关节回归器的结合进行 3D HPE. 与早期的研究相比, 现在有许多单阶段方法将热图表示应用到 3D HPE 中. 例如, Pavlakos 等人[8]提出了一种单阶段方法, 该方法在体素空间中预测三维热图, 并提出了一个粗到细的预测方案以减少大的三维热图代价.

1.2 基于二维提升到三维的方法

随着深度学习的发展, 二维人体姿态估计已经相对成熟. 使用现有的二维人体姿态估计模型来估计二维姿态, 然后将二维姿态作为中间表示, 进一步提升到三维姿态已经成为当下一种流行的三维人体姿态估计方法. 得益于先进的二维姿态检测器的卓越性能, 二维提升到三维的方法通常优于直接估计方法. GAT[9]使用自注意力机制学习每个节点的权重以聚合邻域信息. aGCN[10]与 GAT 的工作方式相同, 通过自注意力机制

学习邻接节点的权重. 区别在于 aGCN 使用不同的激活函数和转换矩阵. 尽管 GAT 和 aGCN 通过聚合邻接节点来提升效果, 但有限的感受野仍然是一个具有挑战性的问题. 为了获得全局感受野, Zhao 等人[11]提出了语义图卷积, 使用 non-local 模块[12]来学习二维关节之间的关系. Lin 等人[13]修改了标准的 Transformer 编码器并调整了编码器层的维度. 然而, 这种交互忽略了上述的相邻关节的局部空间特征信息. 由于人体姿态可以表示为图结构, 其中关节是节点, 骨骼是边, 因此图卷积网络 (GCN) 已经被广泛应用到二维提升到三维的姿态估计任务当中. Ge 等人[14]使用堆叠沙漏网络从图像中提取特征, 这些特征被重塑为图结构. 通过图卷积网络提取的特征预测三维网格, 然后用于预测 3D 姿态. Xu 等人[3]提出了图沙漏网络并采用 SE Block[15]来融合从图沙漏网络的不同层次提取的特征. STCFormer[16]引入了时空交叉注意力模块, 通过串联多个 STC 模块, 并融入了一种新型的结构增强位置嵌入 (SPE), 以深化模型对人体结构的理解. 这种结构增强位置嵌入通过对邻近关节进行时空卷积, 精准捕获关节的局部结构特征, 并通过部分感知嵌入明确每个关节所属的具体部位. GraFormer[6]为了建模非邻近节点间隐含的高阶连接关系, 引入了 ChebGConv 块以在非邻近节点间交换信息, 从而获得更大的感受野. AMPose[17]通过两个独立的模块串联混合局部和全局依赖的信息来捕获人体骨骼中的全局和局部信息.

在本文中, 我们对 AMPose 基线模型进行了优化, 将图卷积和 Transformer 编码器设计成多层来交叉融合局部和全局空间特征, 并用 SE Block 进一步突出重要的空间特征, 从而用少量的参数基础上, 能够获得与原模型三维人体姿态估计接近的估计效果.

2 基于多层空间特征融合的模型

如图 1 所示, 本文提出的模型 MLSFFN 的网络结构主要包括 GCN 模块、Transformer 编码器和 SE 模块 3 部分. 首先用成熟的二维姿态估计器估计出输入图片的二维关键点坐标$X \in R^{N\times 2}$, 然后通过 Patch Embedding 将映射到高维向量空间$X' \in R^{N\times C}$, 其中是 N 关节的数量, C 是通道维度. 然后我们将网络模型中学习得到的位置矩阵$E_{\text{pos}} \in R^{N\times C}$嵌入到$X'$中, 得到嵌入特征向量$X^{\ell}$. 接下来, X^{ℓ}被均匀分成 6 部分: X_1^{ℓ}、X_2^{ℓ}、X_3^{ℓ}、X_4^{ℓ}、X_5^{ℓ}、X_6^{ℓ}. 它们都具有相同的维度$C' = C/6$, 并根据X_i^{ℓ}中的i送入对应的第i层的模块. 这些模块从多个语义层次构建人体关节和身体空间特征. 然后将每一层的输出向量Y^{ℓ}拼接起来获得向量$Y^{\ell} \in R^{N\times C}$, 再将向量$Y_i^{\ell}$输入 SE 模块进行权重自适应, 突出重要的空间特征. 最后输入到多层感知机中, 再使用线性层回归头来估计出三维坐标$Y \in R^{N\times 3}$. 图卷积网络模块、Transformer 编码器和 SE 模块的结构将在以下章节介绍.

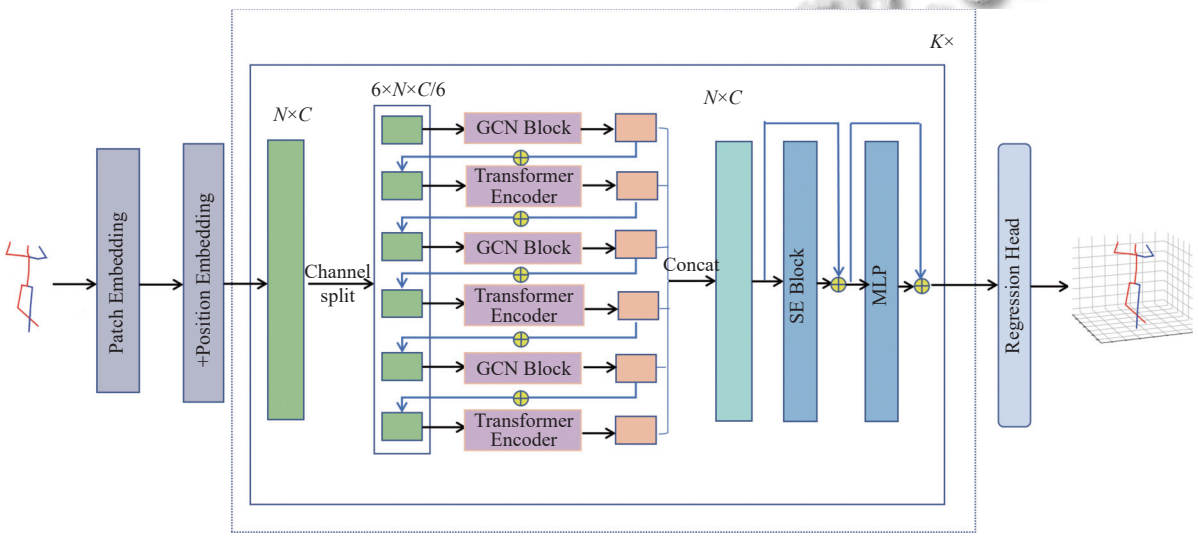

图 1 MLSFFN 的网络结构

2.1 GCN 模块

人体骨架的图结构可以表示为$G = (V, A)$, 其中 V 是 N 个关节点的集合, $A \in \{0,1\}^{N\times N}$是邻接矩阵, 代表关节之间的连接关系. 给定输入$X_i^{\ell} \in R^{N\times C'}$以通过

GCN Block 将相邻关节的空间特征聚合, 当$i-1$时, 具体公式如下:

$$\mathrm{GCN}(X_i^{\ell}) = \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}X_i^{\ell}W \tag{1}$$

$$Y_i^{\ell} = X_i^{\ell} + \mathrm{GCN}(\sigma(\mathrm{GCN}(X_i^{\ell}))) \tag{2}$$

当$i\in\{3,5\}$时, 先将第 i 层的输入向量与第$i-1$层的输出向量$\tilde{X}_i^{\ell}$相加获得向量$\tilde{X}_i^{\ell}$, 即 $\tilde{X}_i^{\ell} = X_i^{\ell} + Y_{i-1}^{\ell}$, 然后再将$\tilde{X}_i^{\ell}$输入到 GCN Block 中, 具体公式如下:

$$\mathrm{GCN}(\tilde{X}_i^{l}) = \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}\tilde{X}_i^{l}W \tag{3}$$

$$Y_i^{\ell} = \tilde{X}_i^{l} + \mathrm{GCN}(\sigma(\mathrm{GCN}(\tilde{X}_i^{l}))) \tag{4}$$

其中, $\tilde{A}=A+I$, $\tilde{D}$是关节点的度的对角矩阵, W 是权重矩阵, Y_i^{ℓ}表示第 i 层对应的 GCN Block 的输出, σ表示 GELU 激活函数.

2.2 Transformer 编码器

缩放的点积注意力: 缩放的点积注意力通常被定义为一个自注意力函数, 这个函数将 Query (Q) 矩阵, Key (K) 矩阵, Value (V) 矩阵作为输入. 其中$Q,K,V\in R^{N_j\times N_d}$, N_j表示关节的数量, N_d表示通道的数量. 为了避免 Q 和 K 的乘积过大, 引入 $\sqrt{N_d}$作为归一化因子. 自注意力的计算公式如下:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{N}}\right)V \tag{5}$$

当输入特征$\mathbf{Z}\in R^{N_j\times N_d}$经过可学习的变换矩阵$W^Q,W^K,W^V\in R^{N_d\times N_d}$的变换后, 可以得到 Q、K、V, 也就是:

$$Q = ZW^Q, K = ZW^K, V = ZW^V \tag{6}$$

多头自注意力 (multi-head self attention, MSA) 层: 多头自注意力是多个自注意力函数的拼接. 每个头部独立地处理部分输入特征. 多个头的拼接结果会通过W^{out}进行变换. 多头自注意力公式如下:

$$MSA(Q,K,V) = Concat(head_1,\cdots,head_{N_h})W^{\mathrm{out}} \tag{7}$$

其中, $head_{in} = Attention(Q_{in},K_{in},V_{in}), in\in[1,\cdots,N_h]$, 而$W^{\mathrm{out}}\in R^{N_d\times N_d}$是可训练的参数矩阵, N_h是注意力头的数量.

在本文中, 当$i\in\{2,4,6\}$时, 先将第 i 层的输入向量X_i^{ℓ}与第 i–1 层的输出向量Y_{i-1}^{ℓ}相加获得向量$\tilde{X}_i^{t}$, 即$\tilde{X}_i^{\ell} = X_i^{\ell} + Y_{i-1}^{\ell}$, 然后再将$\tilde{X}_i^{\ell}$输入到 Transformer Encoder 中, 具体公式如下:

$$head_t = Softmax\left(\frac{Q^{\ell}K^{\ell^{\mathrm{T}}}}{\sqrt{C'}}\right)V^{\ell},\ t\in\{1,\cdots,h\} \tag{8}$$

$$MSA(\tilde{X}_i^{\ell}) = (head_1,head_2,\cdots,head_h)W^{\mathrm{out}} \tag{9}$$

$$\tilde{Y}_i^{\ell} = \tilde{X}_i^{\ell} + LN(MSA(\tilde{X}_i^{\ell})) \tag{10}$$

其中, h 是注意力头的数量, $Q^{\ell},K^{\ell},V^{\ell}$ 分别是 Query、Key、Value 矩阵, 它们是通过$\tilde{X}_i^{\ell}$线性变换得到的, $\tilde{Y}_i^{\ell}$表示第 i 层对应的 Transformer Encoder 的输出.

2.3 SE 模块

如图 2 所示, SE 模块具体包括全局池化层、全连接层、ReLU 激活函数以及 Scale 操作. 将每一层的输出向量Y_i^{ℓ}拼接起来后获得的向量$Y^{\ell}\in R^{N\times C}$到 SE 模块, 在引入少量的参数的基础上, 显式建模卷积特征通道之间的相互依赖关系, 提高网络的表征能力, 通过权重自适应有选择性地突出人体建模中的重要的特征, 抑制不太重要的特征, 最后输出与输入向量Y^{ℓ}同等维度的向量.

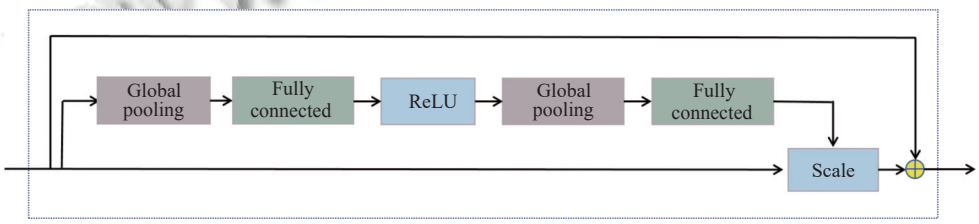

图 2 SE 模块

3 实验

本节描述了实验中所用到的数据集、评估指标、消融实验等. 在 Human3.6M 数据集上与其他基线方法进行了对比, 并通过消融实验证明了本文设计出的模型结构的有效性. 最后, 在 MPI-INF-3DHP 数据集上进行测试, 进一步证明该方法的泛化能力.

3.1 数据集

本文提出的模型在两个公共数据集上进行了测试:

Human3.6M 和 MPI-INF-3DHP.

Human3.6M: 是三维人体姿态估计中最常见的数据集之一, 由 360 万张图像组成, 通过 4 台高分辨率逐行扫描相机以每秒 50 帧的速率捕获视频数据. 该数据集包括 17 个场景, 其中 11 名专业演员 (6 男 5 女) 参与各种活动, 包括讨论、吸烟、拍照、打电话等. 在本文中, 将对象 1、5、6、7、8 归入训练集, 而将对象 9 和 11 划分为测试集.

MPI-INF-3DHP: 是一个具有挑战性的数据集, 图像包括了室内和室外环境. 该数据集包含 130 多万帧图像, 由 14 个摄像角度记录 8 位参与者的 8 类活动. 在本文中通过在 MPI-INF-3DHP 的测试集上进行测试来进一步评估模型的泛化能力.

3.2 评估指标

对于 Human3.6M 数据集, 通常采用 MPJPE (mean per joint position error) 和 P-MPJPE (procrustes MPJPE) 作为评价指标. MPJPE 指估计的关节和地面真值之间的平均欧氏距离, 单位为 mm, MPJPE 也称为 Protocol #1. P-MPJPE 是先进行刚性变换后, 再计算 MPJPE. 通常将 P-MPJPE 称为 Protocol #2. 而对于 MPI-INF-3DHP 数据集, 通常采用正确关键点的百分比 (PCK)、AUC、Outdoor、GS、noGS 作为评价指标.

3.3 实验参数设置

本文提出的模型 MLSFFN 是一种两阶段方法, 即基于二维提升到三维的方法. 首先将图像输入到成熟的二维姿态估计模型中, 本方法采用的二维姿态估计模型是级联金字塔网络 (cascaded pyramid network, CPN). 然后 MLSFFN 以二维关节坐标作为输入生成三维姿态坐标. 本文将模型的深度 K 设置为 3, 编码后的向量通道大小 C 设置为 240, 多头注意力的头数设置为 8. 在本文提出的模型 MLSFFN 中, 将 batch-size 设置为 512, 迭代次数为 50. 学习率最初设置为 0.000 1, 每 5 个 epoch 后学习率以 0.95 倍指数衰减, 优化器采用的 Adam 优化器.

3.4 实验结果对比

从表 1 可以看出, MLSFFN 与基线方法 AMPose 相比, 仅使用了原模型将近 1/9 的参数量和不到 1/10 的浮点运算次数, 就实现了和 AMPose 差不多的 P1 和更优的 P2 值. 这表明 MLSFFN 在保持性能的同时, 显著提高了模型的效率和轻量化程度.

表 1 MLSFFN 与 AMPose 在参数量和浮点运算次数上的对比

方法	参数量 (M)	FLOPs (M)	P1	P2
AMPose[17]	18.3	312.2	49.5	39.4
Ours	2.19	28.5	49.9	39.2

表 2、表 3 展示了本文提出的方法与其他方法在 Human3.6M 数据集上的表现.

表 2 MLSFFN 与其他方法在 Human3.6M 上 Protocol #1 (MPJPE) 和 Protocol #2 (P-MPJPE) 的对比

方法	P1 (CPN)	P2 (CPN)	P1 (GT)
Ci等人[18]	52.7	42.2	45.5
Cai等人[2](refine)	50.6	40.2	38.1
Li等人[19]	49.9	39.2	38.0
Xu等人[3]	51.9	—	35.8
Lutz等人[20]	50.5	—	34.0
Zou等人[5](refine)	49.4	39.1	37.4
AMPose[17]	49.5	39.4	**33.7**
Ours	49.9	39.2	38.0
Ours(refine)	**49.1**	**38.9**	36.7

注: P1和P2分别代表Protocol #1和Protocol#2, CPN表示级联金字塔网络(cascaded pyramid network), GT表示将2D ground truth作为输入.

表 3 MLSFFN 与其他方法在 Human3.6M 上不同动作的 MPJPE 的对比

Method	Dire.	Disuc.	Eat	Greet	Phone	Photo	Pose	Purchu.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Ci等人[18]	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
Cai等人[2](refine)	46.5	**48.8**	47.6	50.9	52.9	61.3	48.3	45.8	59.2	64.4	51.2	48.4	53.5	39.2	41.2	50.6
Li等人[19]	47.8	52.5	47.7	50.5	53.9	60.7	49.5	49.4	60.0	66.3	51.8	48.8	55.2	40.5	42.6	51.8
Xu等人[3]	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
Lutz等人[20]	45.0	49.8	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
Zou等人[5](refine)	45.4	49.2	45.7	49.4	50.4	58.2	47.9	46.0	57.5	**63.0**	49.7	46.6	52.2	**38.9**	40.8	49.4
AMPose[17]	44.9	49.3	45.2	**48.8**	51.3	58.6	47.8	**44.8**	57.1	66.5	49.9	46.4	52.9	39.0	40.6	49.5
Ours	45.3	49.7	46.7	49.5	51.0	56.6	48.4	45.8	57.5	65.3	50.5	47.0	53.4	40.3	41.5	49.9
Ours(refine)	**44.8**	49.5	45.3	49.2	**50.2**	**56.0**	48.2	44.9	**56.9**	63.8	**49.6**	**46.1**	**52.0**	39.4	**40.6**	**49.1**

注: refine 代表着Pose Refinement, 是细化三维关节点位置的常用方法.

从表 2 和表 3 可以看出, 本文提出的方法相比于其他基线方法, 在打电话、拍照等多个动作场景下具有更好的表现. 当使用 CPN 估计得到的二维关键点作为输入时, MLSFFN 在 MPJPE (49.1 mm) 和 P-MPJPE

(38.9 mm) 下都取得了最佳结果. 此外, 使用 ground truth 二维关节作为输入时, MLSFFN 也表现很好, 平均每关节误差达到了 36.7 mm. 为了进一步评估 MLSFFN 的泛化能力, 在 Human3.6M 上训练 MLSFFN, 并在 MPI-INF-3DHP 上进行测试. 如表 4 中所示, MLSFFN 使用少量的参数保持了在 MPI-INF-3DHP 上与 AMPose 相近的性能, 并且在 PCK、AUC、Outdoor 等 3 个指标上优于其他基线模型.

表 4 MLSFFN 与其他方法在 MPI-INF-3DHP 上的性能对比

Method	GS	noGS	Outdoor	PCK	AUC
Zeng等人[21]	—	—	80.3	77.6	43.8
Zou等人[5]	**86.4**	86.0	85.7	86.1	53.7
Xu等人[3]	81.5	81.7	75.2	80.1	45.8
Liu等人[22]	77.6	80.5	80.1	79.3	47.6
Zhao等人[6]	80.1	77.9	74.1	79.0	43.8
AMPose[17]	86.1	**87.5**	**87.4**	**87.0**	**55.2**
Ours	84.5	87.1	83.1	86.4	54.2

注: 粗体代表在单列中数值最优的数, 下划线代表在单列中数值排第二的数字.

3.5 消融实验

为了评估 MLSFFN 在模型设计结构上的有效性, 本文在 Human3.6M 数据集上进行了消融实验, 针对不同设计结构、参数和分层的层数进行了参数量和模型效果的对比. 实验结果如表 5 和表 6 所示.

表 5 在 Human3.6M 上不同结构的性能对比

Structure	Param (M)	MPJPE
G	2.65	50.8
T	2.26	52.5
G - T	3.33	50.4
G | T	2.24	51.4
G->T	2.24	50.7
G->T->G->T->G->T	2.19	50.8
G->T->G->T->G->T+ SE (Ours)	**2.19**	**49.9**

注: G代表GCN模块, T代表Transformer编码器, - 代表串行结构, | 代表并行结构, ->代表本文提出的多层空间特征融合结构.

表 6 在 Human3.6M 上不同层数和参数的性能对比

Layers	k	Param (M)	MPJPE
2	2	1.9	50.6
2	3	2.24	50.5
6	2	1.8	52.1
6	3	2.19	**49.9**
6	4	2.65	50.4
10	2	1.67	51.7
10	3	2.12	50.8

注: 由于模型中分层以后后续会用到多头注意力, 实验中通道数设置的240, 多头注意力的头数设置的8, 所以确保Layers的选择上既是30的因子且是偶数, 实验中选择了2、6、10进行对比.

结果表明, MLSFFN 相比于单个 GCN 模块、单个 Transformer 模块、并行设计、串行设计、多层空间特征融合 (不包括 SE 模块), 本文设计的结构可以在使用的少量参数量的基础上, 模型获得最优的性能. 并且进一步在层数和参数的选择上进行了实验对比, 当层数为 6, 模型深度 k 为 3 时, 模型效果最优.

3.6 定性结果

图 3 展示了本文提出的模型 MLSFFN 在 Human-3.6M 上的定性结果. 采用 Human3.6M 数据集中的视频帧作为输入, 利用本文提出的模型对视频帧中的人体姿态进行估计, 得到了图 3 中的预测结果. 为了验证本文算法的准确性, 将预测结果与三维地面真值 (ground truth) 进行了对比分析. 经过比较, 发现估计值与真实值之间的差异极其微小, 也进一步证明了我们算法的有效性.

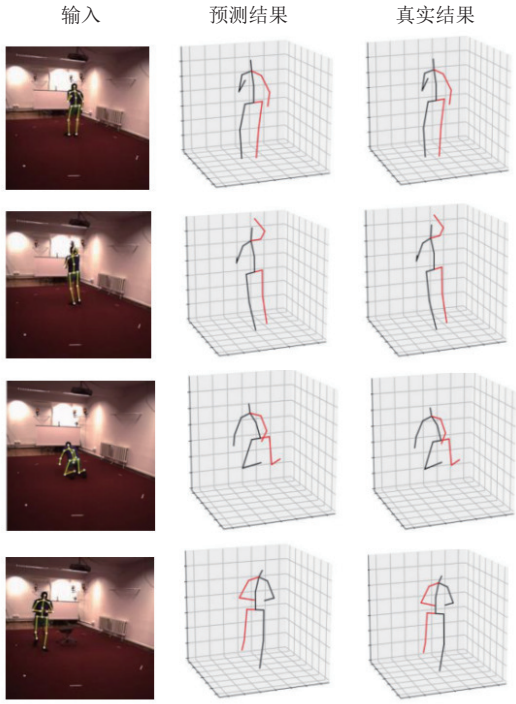

图 3 MLSFFN 在 Human3.6M 上的定性结果

4 总结

本文基于基线模型 AMPose 进行了优化, 提出了一种新颖的架构 MLSFFN. 人体关节之间的关系可以分为两种: 非物理连接和物理连接关系在 MLSFFN 中, 分别由 Transformer 编码器和 GCN 对关节关系模块化. 其中, Transformer 编码器用于处理每个关节与所有其他关节的连接关系, 而 GCN 用于捕获人体相邻关节

的信息. 通过 Transformer 编码器和 GCN 模块进行 6 层空间特征交互融合, 再通过 SE 模块突出其中重要特征的权值, 有效提取了关节间的空间特征信息. 在提升模型性能的同时, 一定程度上降低了参数量. 本文所提出的方法在 Human3.6M 和 MPI-INF-3DHP 数据集上与基线模型进行了对比. 实验结果表明, MLSFFN 在准确性和参数量方面表现出更好的性能.

参考文献

1 Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. Proceedings of the 5th International Conference on Learning Representations. Toulon: ICLR, 2017. 1–14.

2 Cai YJ, Ge LH, Liu J, *et al*. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 2272–2281.

3 Xu TH, Takano W. Graph stacked hourglass networks for 3D human pose estimation. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 16100–16109.

4 Liu KK, Ding RQ, Zou ZM, *et al*. A comprehensive study of weight sharing in graph networks for 3D human pose estimation. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 318–334.

5 Zou ZM, Tang W. Modulated graph convolutional network for 3D human pose estimation. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 11457–11467.

6 Zhao WX, Wang WQ, Tian YJ. GraFormer: Graph-oriented Transformer for 3D pose estimation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 20406–20415.

7 Li SJ, Chan AB. 3D human pose estimation from monocular images with deep convolutional neural network. Proceedings of the 12th Asian Conference on Computer Vision. Singapore: Springer, 2015. 332–347.

8 Pavlakos G, Zhou XW, Derpanis KG, *et al*. Coarse-to-fine volumetric prediction for single-image 3D human pose. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1263–1272.

9 Veličković P, Cucurull G, Casanova A, *et al*. Graph attention networks. arXiv:1710.10903, 2017.

10 Yang JW, Lu JS, Lee S, *et al*. Graph R-CNN for scene graph generation. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 690–706.

11 Zhao L, Peng X, Tian Y, *et al*. Semantic graph convolutional networks for 3D human pose regression. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3420–3430.

12 Wang XL, Girshick R, Gupta A, *et al*. Non-local neural networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7794–7803.

13 Lin K, Wang LJ, Liu ZC. End-to-end human pose and mesh reconstruction with Transformers. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 1954–1963.

14 Ge LH, Ren Z, Li YC, *et al*. 3D hand shape and pose estimation from a single rgb image. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 10825–10834.

15 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.

16 Tang ZH, Qiu ZF, Hao YB, *et al*. 3D human pose estimation with spatio-temporal criss-cross attention. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 4790–4799.

17 Lin HX, Chiu Y, Wu PY. AMPose: Alternately mixed global-local attention model for 3D human pose estimation. Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island: IEEE, 2023. 1–5.

18 Ci H, Wang CY, Ma XX, *et al*. Optimizing network structure for 3D human pose estimation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 2262–2271.

19 Li H, Shi BW, Dai WR, *et al*. Hierarchical graph networks for 3D human pose estimation. Proceedings of the 32nd British Machine Vision Conference 2021. BMVC, 2021. 1–14

20 Lutz S, Blythman R, Ghosal K, *et al*. Jointformer: Single-frame lifting Transformer with error prediction and refinement for 3D human pose estimation. Proceedings of the 26th International Conference on Pattern Recognition (ICPR). Montreal: IEEE, 2022. 1156–1163.

21 Zeng AL, Sun X, Yang L, *et al*. Learning skeletal graph neural networks for hard 3D pose estimation. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 11416–11425.

22 Liu RX, Shen J, Wang H, *et al*. Attention mechanism exploits temporal contexts: Real-time 3D human pose reconstruction. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5063–5072.

(校对责编: 张重毅)