MSGPose:基于多语义图卷积与图引导状态空间模型的单目单人三维人体姿态估计

李 俊[1,2*],李 昱[1],陈 黎[1,2]

(1. 武汉科技大学计算机科学与技术学院,湖北武汉 430065;
2. 智能信息处理与实时工业系统湖北省重点实验室,湖北武汉 430065)

摘 要: 单目单人三维人体姿态估计在动作识别与人机交互等领域极具应用价值。然而,受固有深度模糊、严重自遮挡与成像噪声等因素影响,从含有误差的2D观测中鲁棒地恢复3D姿态仍是一大挑战。针对现有基于图卷积神经网络(Graph Convolutional Network, GCN)方法过度依赖单一且静态的物理骨架拓扑,难以充分表达左右对称性等非物理连接语义,以及基于自注意力的方法在长序列下因二次方复杂度导致计算冗余与参数量过大的问题,本文提出MSGPose,一种基于多语义动态分离图卷积(Multi-Semantic Dynamic Separable Graph Convolution,MSDG)与语义图引导的时空双向Mamba(Semantic Graph-guided Mamba block,SGM)的双流并行框架,用于联合建模并提取2D姿态序列的空间与时间特征。具体而言,MSDG模块不仅通过自连接、物理连接与左右对称性先验构建多层级语义图,为各语义分支配置独立权重以避免特征耦合,还引入带权修正矩阵动态缓解固定拓扑的归纳偏置。随后,结合基于K近邻策略构建的稀疏动态时序图卷积,自适应捕获复杂运动下的跨关节与跨帧依赖。同时,为弥补Mamba架构在空间拓扑建模上的不足,SGM模块在双向状态空间扫描前引入多语义图卷积引导,将解剖学结构先验显式注入序列表示中,为后续的状态空间模型提供了一个具备局部几何感知能力的特征空间,从而更高效地进行长程时空依赖建模。在特征融合与优化阶段,通过可学习的自适应权重对两条特征流进行互补融合,并采用三维位置损失与速度损失进行联合训练,以增强预测姿态的时间一致性与稳定性。在Human3.6M数据集上,MSGPose取得了38.9 mm的平均每关节位置误差(Mean Per Joint Position Error,MPJPE),相较于MotionBERT降低了0.3 mm。值得注意的是,MSGPose展现出极佳的参数效率,其参数量(13.3 M)仅约为MotionBERT的31%。同时,在场景更复杂的MPI-INF-3DHP数据集上,MSGPose的MPJPE降至14.5 mm,相较于MotionAGFormer进一步下降了1.7 mm。此外,在无噪声的真实2D标注输入下,MSGPose进一步将Human3.6M上的MPJPE显著降至12.7 mm。这充分验证了多语义先验与双流结构的结合,能有效提升2D到3D的姿态回归能力。

关键词: 三维人体姿态估计;多语义图卷积;状态空间模型;时空建模;语义先验;深度学习
基金项目: 国家自然科学基金(No.62271359)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2026)03-1118-14
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250860

MSGPose: Monocular Single-Person 3D Human Pose Estimation Via Multi-Semantic Graph Convolution and Graph-Guided State Space Models

LI Jun[1,2*], LI Yu[1], CHEN Li[1,2]

(1. *College of Computer Science and Technology*, *Wuhan University of Science and Technology*, *Wuhan*, *Hubei* 430065, *China*;
2. *Hubei Province Key Laboratory of Intelligent Information Processing and Real-Time Industrial System*, *Wuhan*, *Hubei* 430065, *China*)

Abstract: Monocular single-person 3D human pose estimation holds immense value in action recognition and human-computer interaction. However, the lack of depth cues in monocular setups introduces inherent depth ambiguity, while self-occlusion and imaging noise further complicate the task. Robustly recovering 3D poses from erroneous 2D observations remains a major challenge. Existing graph convolutional network (GCN) methods often abstract the human skeleton into a predefined physical graph and mainly rely on this fixed topology. Consequently, they fail to express non-physical semantic relations, such as bilateral symmetry. Conversely, self-attention-based methods excel at long-range modeling. Yet, they suffer from quadratic complexity in long sequences. This complexity leads to increased computational cost and parameter over-

收稿日期:2025-09-28;录用日期:2026-02-28;责任编辑:覃怀银
*通信作者:李俊

head. To address these limitations, we propose MSGPose. This is a dual-stream parallel framework for parameter-efficient monocular 3D human pose estimation. It integrates a multi-semantic dynamic separable graph convolution (MSDG) module and a semantic graph-guided mamba (SGM) module. The framework jointly models and extracts spatial and temporal features from 2D pose sequences. The MSDG module tackles spatial and temporal relations. It constructs dynamic multi-level semantic graphs using three priors: self-connections, physical connections, and anatomical symmetry. MSDG assigns independent learnable weights to each semantic branch, which helps alleviate semantic feature coupling. Additionally, a weighted modification matrix dynamically mitigates the inductive bias of fixed topologies. For temporal dynamics, MSDG employs a sparse dynamic temporal graph convolution. It builds this graph using a K-Nearest Neighbors (K-NN) strategy based on feature similarity. This enables the modeling of cross-joint and inter-frame dependencies during complex movements. The SGM module addresses the spatial topology limitations of standard Mamba architectures. Flattening spatial tokens into 1D sequences may disrupt the natural topology of the human skeleton. To fix this, SGM introduces a multi-semantic graph convolution guidance mechanism. This mechanism operates before the causal 1D convolution and bidirectional state space scanning. This step explicitly injects anatomical structure priors into the sequence representation. It provides the subsequent state space model with a geometry-aware feature space. This enables efficient modeling of long-range spatio-temporal dependencies with linear complexity. During the feature fusion stage, MSGPose employs an adaptive mechanism. Learnable weights complementarily integrate the outputs from the two streams. The framework utilizes joint training optimized by 3D position and velocity losses. The velocity loss limits differences between adjacent frames. This strategy improves the temporal consistency and stability of the predicted poses. Extensive experiments demonstrate the effectiveness of MSGPose. On the Human3.6M dataset, it achieves a mean per joint position error (MPJPE) of 38.9 mm, representing a 0.3 mm improvement over MotionBERT while using only 13.3M parameters (approximately 31% of MotionBERT). On the challenging MPI-INF-3DHP dataset, MSGPose demonstrates strong generalization ability. It achieves an MPJPE of 14.5 mm, representing a 1.7 mm improvement over MotionAGFormer. Using noise-free ground-truth 2D annotations, the MPJPE on Human3.6M drops significantly to 12.7 mm. These results demonstrate the effectiveness of combining multi-semantic priors with a dual-stream architecture. This combination improves the performance of 2D-to-3D pose regression.

Keywords: 3D human pose estimation; multi-semantic graph convolution; state space model; spatio-temporal modeling; semantic prior; deep learning

Foundation Item(s): National Natural Science Foundation of China (No.62271359)

0　引言

单目三维人体姿态估计(3D Human Pose Estimation, 3D HPE)旨在从单目RGB图像或视频序列中恢复人体关键点的三维空间坐标,在动作识别[1]、人机交互及自动驾驶[2]等领域具有重要价值。受深度模糊、遮挡与成像噪声影响,单目条件下的3D姿态回归往往依赖不完整或含误差的2D关键点观测,如何在保证精度的同时控制模型参数规模,是该任务的关键难点。

现有方法大体可分为两类:端到端方法与两阶段方法。端到端方法试图直接从RGB图像中预测3D姿态[3],而两阶段方法则先通过2D姿态检测器从原始图像中提取2D关键点坐标,再利用模型进行2D到3D的映射[4]。受益于近年来2D姿态检测器的快速发展(如CPN[5]、SH[6]、HRNet[7]),大多数工作倾向于采用后者。然而,由于单目数据中固有的深度模糊与自遮挡问题,如何在存在噪声的2D关键点条件下实现鲁棒的3D姿态估计仍面临巨大挑战。为此,现有主流方法主要探索了基于图卷积网络(GCN)的方法和基于Transformer的方法。

GCN方法最早由Kipf等人[8]在半监督节点分类任务中提出,其通过在图结构数据上定义高效的卷积运算,为后续在其他任务中对人体骨骼建模奠定了基础。LCN[9]通过利用图结构建模人体骨骼的物理连接关系,将GCN网络应用于姿态估计任务。文献[10]在提出模型中引入全局节点来捕获全局姿态信息。SemGCN[11]首次在GCN方法中引入语义信息,在物理连接的基础上引入语义边,通过学习额外的邻接关系增强跨关节依赖建模。CD-GCN[12]在GCN方法中提出有向图卷积,并通过条件建模来动态调整边权,使得依赖关系根据输入姿态变化而变化。尽管这些基于GCN的改进方法在人体姿态估计中取得了较为理想的效果,但这类方法通常依赖单一的物理连接图或其扩展。这一设计限制了模型在非相邻关节点之间的信息交互能力,从而难以显式捕获更高层次的人体语义信息。

与此同时,基于Transformer的方法在建模长程依赖方面展现出优势,但其二次复杂度在处理长序列时带来较大计算开销。在此背景下,状态空间模型(State Space Model, SSM)[13]及其改进的Mamba[14]

被提出。作为一种不同于Transformer的新型序列建模范式,SSM通过一组离散时间差分方程刻画序列动态变化,因其可形式化为矩阵与向量运算,系统离散化后能够实现高效训练与推理,从而规避注意力机制带来的计算负担。在此基础上,Mamba模型被提出,作为一种基于SSM的新型深度学习架构,其线性复杂度的计算过程使其在建模长程依赖关系时尤为高效,在处理长序列任务中表现尤为突出。进一步地,已有研究[15]提出了基于半可分矩阵的双向Mamba框架,以弥补单向Mamba在推理当前帧时无法利用未来序列信息的不足。然而,在三维人体姿态估计任务中,Mamba模型本质上是对一维线性序列的建模,其难以刻画人体骨骼的复杂结构。由于缺乏显式的结构先验,非拓扑相邻的关节点之间无法充分进行信息交互,从而导致局部空间信息的缺失。因此,如何在同时保证长程依赖建模效率的同时,引入多语义结构先验,实现时空特征的协同建模,成为需解决的核心问题。

为解决现有方法难以兼顾多语义结构先验与高效长程依赖建模的难题,本文提出了一种混合时空架构MSGPose。该方法由多语义动态分离图卷积(MSDG)与语义图引导的时空双向Mamba(SGM)组成,用于联合建模2D姿态序列的空间与时间特征。具体而言,MSDG不同于传统GCN仅依赖单一骨骼拓扑的建模方式,而是显式引入多语义结构先验:自连接图、物理连接图与关节对称图三类互补语义。为充分刻画不同语义下的关节依赖关系,本文在多分支结构中为各语义图配置独立的权重矩阵并行学习,使每一分支能够专注于对应的结构模式;同时,引入带权修正矩阵以在训练过程中对邻接关系进行自适应调整,从而动态缓解固定拓扑带来的归纳偏置。与此同时,SGM在状态空间模型(SSM)之前引入多语义图卷积引导,将人体的物理连接关系与关节对称先验显式注入序列表示之中,使得后续序列建模在更符合骨骼拓扑的特征空间内进行。基于此,SSM能够在长序列范围内持续整合跨时间与跨关节的上下文信息,从而更充分地刻画动作过程中的长程时空依赖。通过MSDG与SGM的协同作用,MSGPose在保持SSM优势表征能力的同时,有效弥补其对显式空间结构建模的不足,进而为三维人体姿态估计提供一种兼顾结构先验注入与语义表征强化的模型设计范式。

1 相关工作

1.1 图卷积网络

基于图卷积网络(GCN)的三维人体姿态估计方法通常将人体骨架抽象为图结构:骨骼关节点作为图节点,关节间的物理连接作为边,从而在图域中显式引入关节点之间的拓扑约束。此类方法的突出优势在于能够自然编码骨骼结构先验,并通过图卷积算子对局部关节依赖进行建模与传播。以SemGCN[11]为代表的开创性工作首次系统地将人体骨架建模为图并引入图卷积运算,验证了骨架图先验在姿态表示学习中的有效性。随后,GroupGCN[16]进一步通过将节点划分为若干组并在组内执行卷积,以及为不同组件配置独立权重矩阵与聚合核,提升了空间关系建模的表达灵活性。然而,这类方法多依赖基于人体解剖学的相邻物理骨骼图,其拓扑往往呈现单一且静态的特点,从而在刻画跨关节、跨部位的全局依赖时受到一定限制。为缓解固定拓扑带来的归纳偏置,Modulated-GCN[17]通过引入权重调制与关联调制机制,使图结构在特征交互过程中具备自适应调整能力,进而捕获更广泛的关节关联并增强全局表达。与此同时,GLA-GCN[18]提出了全局-局部自适应图卷积框架,通过自适应图学习将全局关系与局部骨骼结构进行融合,并结合分层下采样设计对时间维度进行压缩,以提升序列建模的效率与可扩展性。

综上所述,现有的GCN-based方法在利用骨骼先验进行姿态估计方面展现了巨大潜力,但仍面临两个关键挑战:(1)过度依赖单一且静态的物理骨架图,限制了对多样化人体语义关系的表达;(2)缺乏对序列中关节点动态变化的拓扑建模能力,难以适应跨时序的人体动作变化。针对上述不足,本文提出了多语义动态分离图卷积(MSDG),能够同时编码多种显式语义先验(如自连接、物理骨骼、对称性),并允许拓扑结构在时间维度上动态演化,从而实现更稳定且可解释的图建模范式。

1.2 状态空间模型

最近,状态空间模型(SSM)在长序列建模领域受到广泛关注,尤其在多模态场景下展现出良好的序列表示能力。结构化状态空间序列模型S4[19]通过引入基于HiPPO矩阵的隐状态参数化、离散化方法以及卷积化来实现,使得SSM能够在较长依赖范围内保持稳定的动态建模能力。在此基础上,Mamba[14]进一步提出参数选择性机制与并行扫描策略,使模型在处理长序列时能够保持更具伸缩性的计算与内存开销,从而推动SSM向更大规模序列建模任务的应用扩展。随着SSM向视觉任务的迁移,Vision Mamba[20]通过双向扫描增强特征的上下文聚合能力,形成适用于视觉表征学习的通用骨干;VMamba[21]则提出二维选择性扫描(SS2D),沿多路径建模局部与全局依赖,以缓解二维视觉数据的非序列结构带来的建模障碍。PoseMamba[22]将Mamba引入姿态序列建模并取得了有竞争力的效果。尽管如此,现有Mamba体系仍主要基

于一维序列展开进行扫描,其隐含的邻域关系由序列化方式决定,这与人体骨骼所固有的图拓扑结构并不一致,因而难以显式表达关节之间的物理连接与对称等结构先验。针对这一问题,本文提出语义图引导的Mamba(SGM):在Mamba之前引入图卷积增强,以聚合关节点间的物理连接与关节对称等显式语义关系,从而为后续序列建模提供结构感知的输入特征,弥补纯序列扫描在空间拓扑建模上的不足,并进一步强化跨时序的人体姿态表达能力。

2　本文算法

2.1　相关算法介绍

2.1.1　图卷积网络

图卷积网络(GCN)被广泛应用于建模结构化的数据,其在三维人体姿态估计中的核心思想是通过关节点之间的信息传递实现特征的聚合与更新。人体骨骼数据通常可被建模为一个无向图 $\boldsymbol{G}=(V,E)$,其中 V 和 E 分别表示关节点集合和关节之间的骨骼连接关系。在经典的GCN中,图卷积操作可形式化为

$$\mathrm{GCN}(\boldsymbol{x})=\sigma\left(\mathrm{Norm}\left(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{x}\boldsymbol{W}_1\right)\right) \tag{1}$$

其中,$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}$ 表示在邻接矩阵 $\boldsymbol{A}$ 上加入自连接后的增强邻接矩阵;$\tilde{\boldsymbol{D}}_{ii}=\sum_j\tilde{\boldsymbol{A}}_{ij}$ 为度矩阵,通过对 $\tilde{\boldsymbol{A}}$ 的每一行元素求和得到;$\boldsymbol{X}\in\mathbb{R}^{B\times T\times V\times C}$ 为输入特征;$\boldsymbol{W}_1$ 为可学习的权重矩阵,用于特征映射;$\sigma(\cdot)$ 是ReLU激活函数;Norm(·)为批归一化。

2.1.2　状态空间模型

近年来,大量研究表明状态空间模型(SSM)在长序列数据建模中展现出强大的能力。SSM假设一个动态系统在任意时刻 t 的状态,已包含了决定其未来演化的全部必要信息。其基本思想是通过隐含状态 $\boldsymbol{h}(t)\in\mathbb{R}^N$ 的压缩表示来编码之前序列的信息,从而将一维输入序列 $x(t)\in\mathbb{R}^L$ 映射到输出序列 $y(t)\in\mathbb{R}^L$。经典的SSM可以形式化为一组线性常微分方程:

$$\boldsymbol{h}'(t)=\boldsymbol{A}\boldsymbol{h}(t)+\boldsymbol{B}\boldsymbol{x}(t),\boldsymbol{y}(t)=\boldsymbol{C}\boldsymbol{h}(t) \tag{2}$$

其中,$\boldsymbol{A}\in\mathbb{R}^{N\times N}$、$\boldsymbol{B}\in\mathbb{R}^{N\times 1}$、$\boldsymbol{C}\in\mathbb{R}^{1\times N}$ 为可学习参数矩阵。在此基础上,结构化状态空间序列模型(S4)通过零阶保持方法对上述返程进行离散化,并引入步长 Δ 来求解。其离散化形式可表示为

$$\boldsymbol{h}_k=\bar{\boldsymbol{A}}\boldsymbol{h}_{k-1}+\bar{\boldsymbol{B}}x_k,y_k=\boldsymbol{C}\boldsymbol{h}_k \tag{3}$$

其中,$x_k\in\mathbb{R}^{L\times D}$ 为输入序列的第 k 个切片;$\bar{\boldsymbol{A}}$、$\bar{\boldsymbol{B}}$ 为离散化后的参数。然而,传统的线性时不变SSM在表达复杂动态依赖时存在局限性。为此,Mamba在S4的基础上提出了输入依赖的参数化机制,并设计了硬件感知的并行扫描算法,能够在保证计算效率的同时实现递归式序列建模,从而显著提升了在长序列场景下的适用性。

2.2　网络框架

本文提出的MSGPose框架主要由三部分组成:关节位置嵌入、堆叠的MSGPose模块和回归头,模型的结构概览如图1所示。首先,关节位置嵌入将输入的二维姿态序列映射到高维特征空间,并结合可学习的位置编码以增强时序与空间表达。随后,堆叠的MSGPose模块通过双流并行架构联合建模,其中一条分支利用语义图引导的时空双向Mamba(SGM)捕获时空依赖关系,另一条分支通过多语义动态分离图卷积(MSDG)融合人体的多种先验语义关系,从而学习到更精细的空间依赖。最终,回归头将MSGPose模块的输出特征进行投影,直接回归得到精确的三维姿态坐标。模型训练时,参考之前的工作[23],采用位置损失与速度损失来联合优化目标:

$$L=L_{3\mathrm{D}}+\lambda L_{\mathrm{v}} \tag{4}$$

$$L_{3\mathrm{D}}=\sum_{t=1}^{T}\sum_{j=1}^{J}\left\| P_{3\mathrm{D}}(t,j)-\hat{P}_{3\mathrm{D}}(t,j)\right\| \tag{5}$$

$$L_{\mathrm{v}}=\sum_{t=2}^{T}\sum_{j=1}^{J}\left\|\Delta_{P_{3\mathrm{D}}(t,j)}-\Delta_{\hat{P}_{3\mathrm{D}}(t,j)}\right\| \tag{6}$$

其中,$L_{3\mathrm{D}}$ 用于衡量预测的三维关节点位置与真实值之间的误差,而 L_{v} 则用于约束相邻时间帧之间关节点的运动速度,从而提升时间一致性与姿态平滑性。通过这种端到端的优化方式,模型能够在捕获全局时序信息与局部结构依赖的同时,生成更加稳定和精确的三维人体姿态预测。

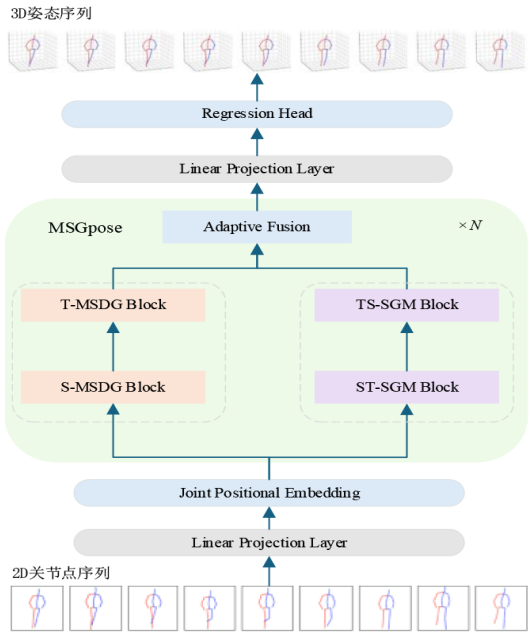


图1　本文所提出MSGPose网络结构图

Figure 1　Architecture of the proposed MSGPose network

2.3 多语义动态分离图卷积

在三维人体姿态估计任务中，图卷积网络（GCN）能够利用人体骨架的图结构对关节间依赖关系进行建模，因而在空间关系建模方面表现出良好性能。经典的空间图卷积操作可表示为

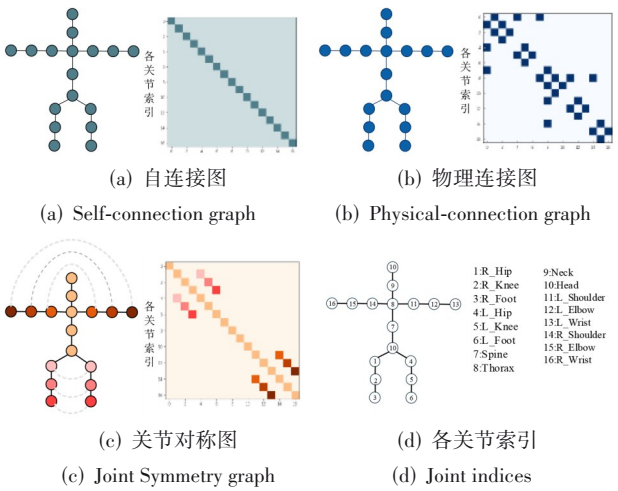
$$Z=\sigma\left(\hat{\boldsymbol{A}}_s X\boldsymbol{W}+\hat{\boldsymbol{A}}_i X\boldsymbol{W}+\hat{\boldsymbol{A}}_o X\boldsymbol{W}\right) \tag{7}$$

其中，$\hat{\boldsymbol{A}}_s$、$\hat{\boldsymbol{A}}_i$、$\hat{\boldsymbol{A}}_o$代表自连接物理入边与出边的归一化邻接矩阵。该方法能够刻画局部骨架拓扑，但其表达能力往往受限于单一且预定义的骨架连接方式：一方面，难以显式区分节点自身信息与邻域信息的重要性；另一方面，难以刻画对姿态回归同样关键的语义多样性（如关节物理连接与左右对称性）。事实上，人体结构除包含固定的解剖连接外，还蕴含以脊柱为中心的全局对称几何先验。如何将多种结构先验有效引入图结构，并允许其在训练过程中自适应调整，是亟须解决的问题。为此，不同于传统的自连接、入边、出边构成物理骨骼相邻的邻接矩阵，本文构建基于人体先验的多语义动态图结构，其连接方式如图2所示，定义如下：

$$\boldsymbol{A}=\{A_{\text{self}},A_{\text{phy}},A_{\text{sym}}\} \tag{8}$$

其中，A_{self}为自连接图，提供基础的节点自我表征；A_{phy}为物理连接图，用于捕获由骨骼连接决定的局部空间上下文；A_{sym}为关节对称图，用于显式建模人体左右两侧的双边对称性（例如左/右手腕、左/右肘等）。该对称先验在单侧肢体遮挡或观测模糊导致的歧义情形下尤为关键。进一步地，本文为每一类语义分支引入可学习的分支权重与动态修正机制，使得各语义连接在训练过程中能够根据数据自适应地进行强度调节，从而获得更稳健的结构建模。

然而，在经典GCN架构中通常使用同一组特征变换权重$\boldsymbol{W}$来聚合不同拓扑下的邻域信息，导致不同语义分支在参数层面被迫共享，从而引发语义特征的耦合与混叠，甚至带来多重共线性问题。针对这一不足，本文提出空间多语义动态分离图卷积（S-MSDG），以实现不同语义的并行独立建模，如图3所示。



(a) 自连接图 (b) 物理连接图

(a) Self-connection graph (b) Physical-connection graph

(c) 关节对称图 (d) 各关节索引

(c) Joint Symmetry graph (d) Joint indices

图2 三种骨架图的连接方式

Figure 2 Connectivity of the three proposed skeletal graphs

具体地，给定输入特征$\boldsymbol{X}\in\mathbb{R}^{B\times C_{\text{in}}\times T\times J}$，为了能够在多个语义图上进行并行的独立处理，同时又保留一个信息丰富的原始旁路，首先通过逐点卷积$\boldsymbol{W}_{\text{p}}\in\mathbb{R}^{4C_{\text{in}}\times C_{\text{in}}\times 1\times 1}$对输入特征$\boldsymbol{X}$在通道做线性变换，公式如下：

$$F_{\text{expanded}}=\sigma\left(\boldsymbol{W}_{\text{p}}\left(\text{Norm}\left(\boldsymbol{X}\right)\right)\right)\in\mathbb{R}^{B\times 4C_{\text{in}}\times T\times J} \tag{9}$$

其中，$\sigma(\cdot)$为激活函数；Norm$(\cdot)$为层归一化操作；B为批次大小；C_{in}为输入通道数；T为时间序列长度；J为关节点数量。F_{expanded}在通道维度划分为四组$\{F_1,F_2,F_3,F_{\text{static}}\}$，其中$F_{\text{static}}$用于保留原始静态信息，$F_k\in\mathbb{R}^{B\times C_{\text{in}}\times T\times J}\left(k\in\{1,2,3\}\right)$分别对应三类语义分支。对第$k$个语义分支，采用其对应的归一化动态邻接矩阵$\hat{\boldsymbol{A}}_k$与动态修正矩阵$\boldsymbol{A}_D$进行卷积更新：

$$Y_k=\left(\hat{\boldsymbol{A}}_k+\lambda_k\boldsymbol{A}_D\right)F_k\in\mathbb{R}^{B\times C_{\text{in}}\times T\times J} \tag{10}$$

其中，$\hat{\boldsymbol{A}}_k\in\mathbb{R}^{J\times J}$是第$k$个人体语义先验对应的归一化动态邻接矩阵；$\boldsymbol{A}_D\in\mathbb{R}^{J\times J}$为每组动态修正矩阵；$\lambda_k$为可学习权重。静态分支$F_{\text{static}}$不与邻接矩阵相乘，以避免过度平滑并保留原始表征。最后，将三路语义输出与静态分支在通道维拼接，并通过逐点卷积完成融合：

$$\boldsymbol{Z}=\boldsymbol{W}_f\left(\text{Concat}\left[Y_1,Y_2,Y_3,F_{\text{static}}\right]\right)\in\mathbb{R}^{B\times C_{\text{out}}\times T\times J} \tag{11}$$

其中，逐点卷积核$\boldsymbol{W}_f\in\mathbb{R}^{C_{\text{out}}\times 4C_{\text{in}}\times 1\times 1}$。从计算角度看，GCN的主要开销来自特征与邻接矩阵之间的批量矩阵乘法。本文采用爱因斯坦求和约定（Einstein summation，einsum）对该过程进行张量化，并通过repeat$(\cdot)$在批次维度对邻接矩阵进行复制以支持并行计算：

$$\tilde{A}=\text{repeat}(\boldsymbol{A})=\{\tilde{\boldsymbol{A}}_1,\tilde{\boldsymbol{A}}_2,\cdots,\tilde{\boldsymbol{A}}_B|\tilde{\boldsymbol{A}}_i=\boldsymbol{A}\} \tag{12}$$

$$\tilde{Y}=\text{einsum}(\text{“}BVUk,BCTVk\rightarrow BCTUk\text{”},\tilde{\boldsymbol{A}},\boldsymbol{X}) \tag{13}$$

其中，V和U是图的维度；k是语义子图的数量，B、C、T分别是批次、通道、帧数；repeat$(\boldsymbol{A})$表示将$\boldsymbol{A}$在批次维度复制B倍。

在多语义空间特征增强后，本文进一步引入了动态时序图卷积模块（T-MSDG），以建模T帧序列的长程时间依赖，如图4所示。不同于Transformer的二次复杂度，T-GCN通过在时间维度上构建稀疏的内容感知图，实现高效的时序建模。

具体而言，空间增强后的特征表示为$\boldsymbol{Z}\in\mathbb{R}^{B\times C\times T\times J}$。为在时间维度上构建自适应的时序邻接关系，本文将张量$\boldsymbol{Z}$沿关节维展开并重排为$\tilde{\boldsymbol{Z}}\in\mathbb{R}^{(B\times J)\times T\times C}$，即将每个关节在整个序列上的轨迹视为一个独立样本。基于$\tilde{\boldsymbol{Z}}$，首先计算任意两帧特征

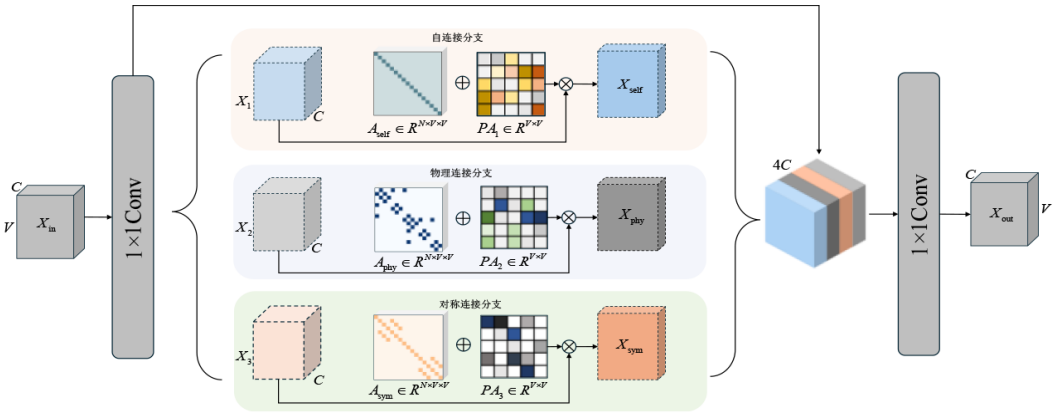


图3　多语义动态分离图卷积(S-MSDG)结构图

Figure 3　Structure of the multi-semantic dynamic separation graph convolution (S-MSDG)

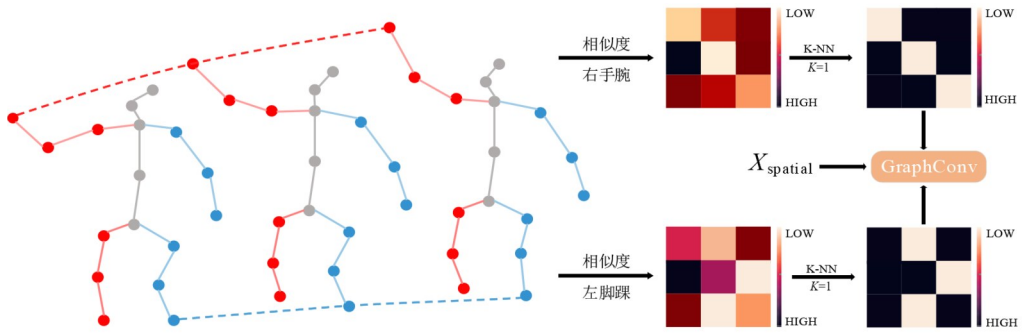


图4　动态时序图卷积

Figure 4　Dynamic temporal graph convolution

的相似度矩阵 $\boldsymbol{S}=\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}\in\mathbb{R}^{(BJ)\times T\times T}$，随后采用 K 近邻(K-Nearest Neighbors，K-NN)策略对 $\boldsymbol{S}$ 进行稀疏化：对每个时刻 t_i，仅保留其相似度最高的 k 个邻近帧，从而得到时序邻接矩阵 $\boldsymbol{A}_t\in\mathbb{R}^{(BJ)\times T\times T}$，记为

$$\boldsymbol{A}_t=\mathrm{KNN}_{k=2}(\boldsymbol{S}) \tag{14}$$

其中，k 为邻居数。该构图方式能够根据输入序列的动态变化自适应生成时序连接，使得每一帧特征可聚合其 k 个最相关的帧信息，从而提升模型面对复杂运动情况的鲁棒性。最后，本文在所构建的时序图上执行标准图卷积以完成时序特征更新：

$$\mathrm{GCN}_t\left(\tilde{\boldsymbol{Z}}\right)=\sigma\left(\mathrm{Norm}\left(\boldsymbol{A}_t\tilde{\boldsymbol{Z}}\boldsymbol{W}_t\right)\right) \tag{15}$$

其中，$\boldsymbol{W}_t$ 为可学习的线性变换矩阵；$\sigma(\cdot)$ 表示非线性激活函数；Norm(·)为归一化操作。上述更新后的特征随后可按关节维重排回 $\mathbb{R}^{B\times C\times T\times J}$ 以与后续模块对接。

2.4　基于语义图引导的 Mamba 模块

为了弥补 Mamba 在视觉任务中对空间结构建模能力的不足，Vision Mamba[20] 通过构建双向 token 扫描序列，将选择性状态空间模型(SSM)的建模对象由单向时间序列扩展为双向序列，从而增强模型对视觉场景的适应性。给定第 l 层 Mamba 的输入 $\boldsymbol{X}_{l-1}\in R^{B\times T\times J\times d}$，其中 d 代表特征维度。为解耦并分别建模时域与空域依赖，本文将 $\boldsymbol{X}_{l-1}$ 以两种 token 排序方式重排为长度 $L=TJ$ 的序列表示：一种为按 (T,J) 顺序展开的空间序列 $\boldsymbol{X}_{l-1}^{\mathrm{spa}}\in R^{B\times(L=TJ)\times d}$，用于刻画各帧内部的关节空间关系；另一种为按 (J,T) 顺序展开的时间序列 $\boldsymbol{X}_{l-1}^{\mathrm{tmp}}\in R^{B\times(L=JT)\times d}$，用于刻画单个关节随时间的动态变化。两路序列随后分别经由线性映射将状态维度由 d 提升到 d'，得到主干特征 $\tilde{\boldsymbol{X}}\in\mathbb{R}^{B\times L\times d'}$ 和门控单元 $\boldsymbol{Z}\in\mathbb{R}^{B\times L\times d'}$：

$$\tilde{\boldsymbol{X}}=\boldsymbol{X}_{l-1}\boldsymbol{W}_x,\boldsymbol{Z}=\boldsymbol{X}_{l-1}\boldsymbol{W}_z\in\mathbb{R}^{B\times L\times d'} \tag{16}$$

其中，$\boldsymbol{W}_x$ 与 $\boldsymbol{W}_z\in\mathbb{R}^{d\times d'}$。随后主干特征 $\tilde{\boldsymbol{X}}$ 经过一个包含向前路径与向后路径的双向 SSM 模块，两条路径的计算过程形式一致但方向相反。

$$\tilde{\boldsymbol{X}}_{\mathrm{f}}=\mathrm{SSM}_{\mathrm{f}}\left(\sigma\left(\mathrm{Conv1d}_{\mathrm{f}}\left(\tilde{\boldsymbol{X}}\right)\right)\right)\in\mathbb{R}^{B\times L\times d'} \tag{17}$$

$$\tilde{\boldsymbol{X}}_{\mathrm{b}}=\mathrm{SSM}_{\mathrm{b}}\left(\sigma\left(\mathrm{Conv1d}_{\mathrm{b}}\left(\mathrm{flip}\left(\tilde{\boldsymbol{X}}\right)\right)\right)\right)\in\mathbb{R}^{B\times L\times d'} \tag{18}$$

其中，f 和 b 代表前向传播与反向传播过程；$\tilde{\boldsymbol{X}}\in\mathbb{R}^{B\times L\times d'}$；$\sigma(\cdot)$ 为 SiLU 激活函数；flip(·)表示将特征在 L 维度上翻转；Conv1d(·)为因果 1d 卷积，用于在 token 维度 L 上进行局部混合(输入/输出张量形状保

持为 $\mathbb{R}^{B\times L\times d'}$)；SSM(·)则为 Mamba 架构的核心选择性状态空间扫描算子。最后，将前向扫描输出 $\tilde{\boldsymbol{X}}_{\mathrm{f}}$ 与经翻转恢复原序的后向扫描输出 $\tilde{\boldsymbol{X}}_{\mathrm{b}}$ 融合，并通过门控分支 Z 进行调制，得到第 l 层的输出：

$$\boldsymbol{X}_l^m=\boldsymbol{X}_{l-1}+\left(\sigma(\boldsymbol{Z})\odot\tilde{\boldsymbol{X}}_{\mathrm{f}}+\sigma(\boldsymbol{Z})\odot\tilde{\boldsymbol{X}}_{\mathrm{b}}\right)\boldsymbol{W}_{\mathrm{out}} \tag{19}$$

其中，$\boldsymbol{W}_{\mathrm{out}}\in\mathbb{R}^{d'\times d}$ 为输出线性映射；$\boldsymbol{X}_{l-1}\in\mathbb{R}^{b\times l\times d}$；⊙代表 Hadamard(逐元素)乘法。$\sigma(\boldsymbol{Z})$、$\tilde{\boldsymbol{X}}_{\mathrm{f}}$、$\tilde{\boldsymbol{X}}_{\mathrm{b}}$ 形状一致均为 $\mathbb{R}^{B\times L\times d'}$，从而经 $\boldsymbol{W}_{\mathrm{out}}$ 投影回到 $\mathbb{R}^{B\times L\times d}$ 以与残差项相加。通过该双向门控 SSM，模型能够在序列层面聚合来自过去与未来的上下文信息，进而生成信息更为充分的特征表示。

然而，Mamba 中的因果 1d 卷积虽可高效处理一维线性序列，但其缺乏对人体骨骼拓扑的显式建模，这一不足对关节序列尤为关键。例如，在常见的关节展开序列中，索引上相邻的左脚踝(索引为 6)与脊柱(索引为 7)在人体物理连接上可能相距甚远，而人体物理连接上直接相邻的胸部(索引为 8)与左肩(索引为 11)在序列中可能并不相邻。该现象会导致模型的归纳偏置与人体骨骼的内在拓扑结构不匹配，使得模型难以有效聚合“物理近邻但序列远邻”的关节信息，进而迫使模型训练时从数据中重新学习本应作为先验的连接关系，显著增加学习难度与计算开销。

为弥补上述结构信息鸿沟，本文提出基于语义图引导的 Mamba 模块(SGM)，在特征输入 SSM 之前引入多语义图卷积引导(multi-Semantic-graph-guided Graph Convolutional Network，SGCN)以增强关节特征聚合。相较于 Mamba 在长程依赖建模方面的优势，SGCN 通过与输入无关的权重矩阵对人体关节连接关系进行显式刻画，从而将骨骼拓扑的先验知识注入特征交互过程。特别地，多语义图学习能够将人体骨骼的多层次语义(如物理连接、对称性等)进行显式编码，使模型在捕获复杂局部几何关系时更具优势。如图 5(b)所示，给定融合物理连接图与关节点对称图的邻接矩阵 $\boldsymbol{G}=\{V,E\}$，其中 V 是包含 J 个关节点的集合，E 为边的集合，则 SGCN 可形式化为

$$\mathrm{SGCN}(x)=\sigma\left(x+\mathrm{BN}\left(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\dot{\boldsymbol{G}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}x\boldsymbol{W}_1+x\boldsymbol{W}_2\right)\right) \tag{20}$$

其中，$\dot{\boldsymbol{G}}=\boldsymbol{G}+\boldsymbol{I}_N$ 代表多语义邻接矩阵与自连接矩阵之和；$\boldsymbol{D}$ 是度矩阵，其定义为 $\tilde{D}_{ii}=\sum_j\dot{G}_{ij}$；$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 为可学习参数；BN(·)为批归一化；$\sigma(\cdot)$ 为 ReLU 激活函数。最终，本文将 SGCN 模型插入 Mamba 中因果 1d 卷积之前，用以在序列扫描前对关节特征进行结构化聚合，如图 5 所示，整体公式可表示为

$$\tilde{\boldsymbol{X}}_{\mathrm{f}}=\mathrm{SSM}_{\mathrm{f}}\left(\sigma\left(\mathrm{Conv1d}_{\mathrm{f}}\left(\mathrm{SGCN}\left(\mathrm{LN}\left(\tilde{\boldsymbol{X}}\right)\right)\right)\right)\right) \tag{21}$$

$$\tilde{\boldsymbol{X}}_{\mathrm{b}}=\mathrm{SSM}_{\mathrm{b}}\left(\sigma\left(\mathrm{Conv1d}_{\mathrm{b}}\left(\mathrm{SGCN}\left(\mathrm{LN}\left(\mathrm{flip}\left(\tilde{\boldsymbol{X}}\right)\right)\right)\right)\right)\right) \tag{22}$$

其中，$\tilde{\boldsymbol{X}}_{\mathrm{f}}$ 与 $\tilde{\boldsymbol{X}}_{\mathrm{b}}\in\mathbb{R}^{B\times L\times d'}$；LN(·)为层归一化。各算子在 token 维度 L 上进行处理时均保持张量形状一致(输入/输出均为 $\mathbb{R}^{B\times L\times d'}$)。

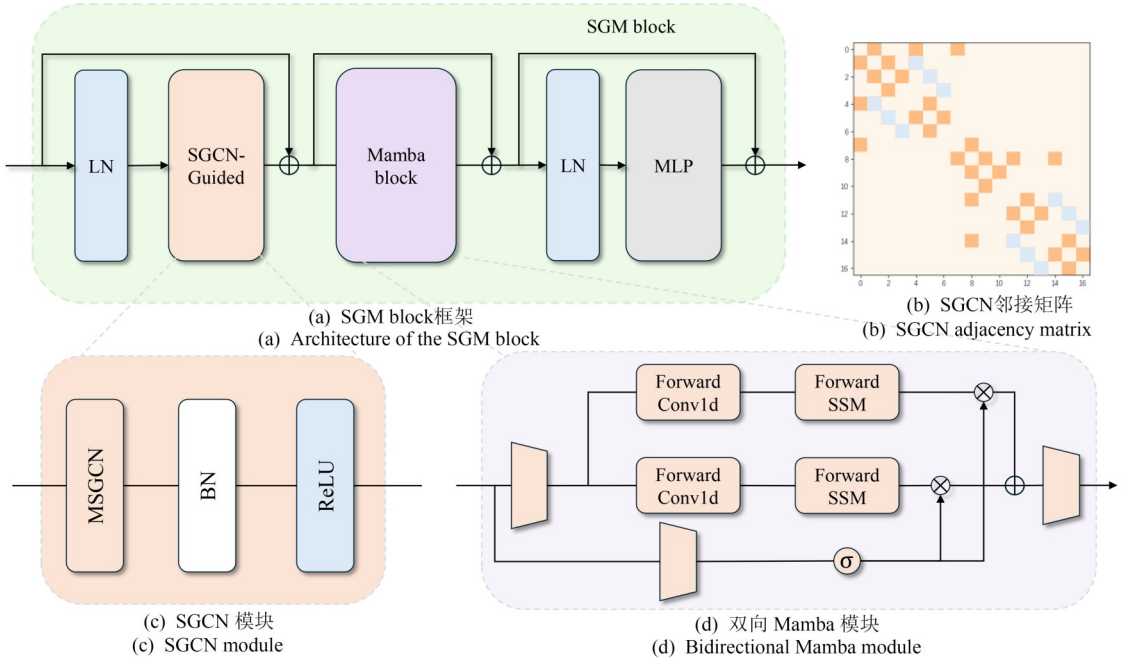


图 5 多语义图引导的时空双向 Mamba(SGM)结构图

Figure 5 Architecture of the multi-semantic graph-guided spatio-temporal bidirectional mamba (SGM)

3　实验结果与分析

3.1　数据集与实验环境

Human3.6M是目前3D人体姿态估计(3D HPE)领域广泛使用的基准数据集。该数据集包含了由11位受试者演绎15种日常活动所构成的360万张图像。为确保评估的公正性并检验模型的泛化能力,本研究遵循了该领域的标准评估协议:采用受试者S1、S5、S6、S7、S8的数据对模型进行训练,并保留受试者S9、S11的数据作为独立的测试集,用以进行最终的性能验证。输入的二维姿态来自SH[6]的预测结果或直接采用真实标注的2D姿势序列。评估指标包括平均每关节位置误差(Mean Per Joint Position Error,MPJPE),即预测的三维关节坐标与真实值之间的欧氏距离平均值;以及通过刚性变换后对预测姿态与真实姿态进行最优对齐后,再测量它们之间的普氏对齐平均每关节位置误差(Procrustes-aligned Mean Per Joint Position Error,P-MPJPE)。

MPI-INF-3DHP是一个比Human3.6M数据集更具有挑战性的数据集,该数据集的特点在于其场景的复杂性与多样性,它不仅包含受控的室内环境,也涵盖了更具挑战性的室外场景。在评估阶段,本文使用真实的2D姿势序列以便与先前的工作直接进行对比。本文遵循前者的工作,采用三项核心指标:平均每关节位置误差(MPJPE)、正确关键点百分比(Percentage of Correct Keypoints,PCK)以及曲线下面积(Area Under the Curve,AUC)。

所有模型训练与评估均在PyTorch框架下基于单张NVIDIA RTX 4090 GPU完成。针对训练配置,本文采用AdamW[24]优化器进行90个周期的迭代优化,批处理大小设为8,权重衰减设为0.01,并固定随机种子数为3 407。初始学习率设定为1×10^{-3},并采用每个周期0.99的衰减因子的指数学习率衰减策略。在数据处理与相机参数方面,本文遵循文献[23,25]的标准设置:对于Human3.6M数据集,训练与测试均应用了随机水平翻转数据增强,并严格执行根相对(root-relative)归一化处理。特别地,对于使用真实2D标注(GT-2D)的实验,直接通过3D坐标的正交投影构建输入;而在MPI-INF-3DHP数据集上,则依据文献[23,25]直接采用真实2D坐标。评估阶段,除了常规的MPJPE指标外,还实施了基于普氏分析的刚性对齐策略(即P-MPJPE),并在推理时引入翻转测试时增强以进一步验证模型的稳健性。

3.2　实验结果与分析

3.2.1　Human3.6M上的实验结果分析

为验证所提MSGPose模型的有效性,本文在Human3.6M数据集上与近年来的主流方法进行了对比。为确保公平性,所有实验未使用额外数据对模型进行预训练。表1与表2展示了15个动作的MPJPE与P-MPJPE性能比较。本文方法($T=243$)的MPJPE与P-MPJPE误差分别为38.9 mm与32.2 mm,优于基于纯Transformer的方法。例如,相比MotionBERT[23]与MixSTE[26],MPJPE误差分别降低了0.3 mm与2.0 mm,充分证明了语义图卷积在捕获局部拓扑结构方面的优势。为了进一步验证方法的有效性,本文与2024—2025年的最新方法进行了对比,尽管PoseAnchor[27]通过优化根节点提升了定位能力,但受限于MixSTE骨干,其MPJPE仍停留在40.3 mm,相比之下,MSGPose在精度上领先1.4 mm,且参数量仅为MixSTE的40%,展现了更优的性能与效率平衡。相较于引入运动学先验的KTPFormer[28],MSGPose凭借Mamba对长序列时序依赖的强大建模能力,将误差进一步降低了1.2 mm。

表1　本文网络与对比方法在Human3.6M数据集上基于MPJPE评价指标的对比结果　单位:mm

Table 1　Performance comparison of different methods on Human3.6M dataset in terms of MPJPE　unit:mm

Method	Reference	T	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
GLA-GCN[18]	ICCV'23	243	41.3	44.3	40.8	41.8	45.9	54.1	42.1	51.5	57.8	62.9	45.0	42.8	45.9	29.4	29.9	44.4
MHFormer[30]	CVPR'22	351	39.2	43.1	40.1	40.9	44.9	51.2	40.6	41.3	53.5	60.3	43.7	41.1	43.8	29.8	30.6	43.0
FTCM[31]	TCSVT'24	351	39.0	41.1	38.6	41.6	45.6	50.4	41.4	37.8	52.9	65.4	42.4	41.2	42.9	31.7	32.9	43.0
P-STMO[32]	ECCV'22	243	38.9	42.7	40.4	41.1	45.6	49.7	40.9	39.9	55.5	59.4	44.9	42.2	42.7	29.4	29.4	42.8
HDFormer[33]	IJCAI'23	96	38.1	43.1	39.3	39.4	44.3	49.1	41.3	40.8	53.1	62.1	43.3	41.8	43.1	31.0	29.7	42.6
MixSTE[26]	CVPR'22	243	37.6	40.9	37.3	39.7	42.3	49.9	40.1	39.8	51.7	55.0	42.1	39.8	41.0	27.9	27.9	40.9
D3DP[34]	ICCV'23	243	37.3	39.5	35.6	37.8	**41.3**	**48.2**	39.1	37.6	**49.9**	**52.8**	41.2	39.2	39.4	27.2	27.1	39.5
STCFormer[35]	CVPR'23	243	39.6	41.6	37.4	38.8	43.1	51.1	39.1	39.7	51.4	57.4	41.8	38.5	40.7	27.1	28.6	41.0
MotionBERT[23]	ICCV'23	243	**36.3**	**38.7**	38.6	**33.6**	42.1	50.1	**36.2**	35.7	50.1	56.6	41.3	37.4	37.7	25.6	**26.5**	39.2
PoseRetNet[29]	ECCV'24	243	36.9	40.1	38.7	38.3	42.9	48.6	38.2	40.0	52.5	55.4	42.3	38.7	39.7	**26.2**	27.8	40.4
KTPFormer[28]	CVPR'24	243	37.3	39.2	**35.9**	37.6	42.5	**48.2**	38.6	39.0	51.4	55.9	41.6	39.0	40.0	27.0	27.4	40.1
PoseAnchor[27]	ICCV'25	243	37.8	39.8	36.6	38.8	42.6	48.6	39.4	38.1	50.5	55.1	42.0	39.1	40.4	27.8	27.8	40.3
MSGPose	Ours	243	37.8	40.2	37.5	33.8	41.5	49.3	37.8	**34.2**	50.4	53.1	**41.0**	**37.2**	**36.6**	27.5	27.1	**38.9**

注:T为序列帧数;加粗数据代表最优结果;下划线代表次优结果,下同。

表2 本文网络与对比方法在Human3.6M数据集上基于P-MPJPE评价指标的对比结果 单位:mm

Table 2 Performance comparison of different methods on the Human3.6M dataset in terms of P-MPJPE unit:mm

Method	Reference	T	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
FTCM[31]	TCSVT'24	351	32.7	35.5	32.5	35.4	59.9	41.6	33.0	31.9	45.1	50.1	36.3	33.5	35.1	23.9	25.0	35.2
STE[36]	TMM'22	351	32.7	35.5	32.5	35.4	35.9	41.6	33.0	31.9	45.1	50.1	36.3	33.5	35.1	23.9	25.0	35.2
MHFormer[30]	CVPR'22	351	31.5	34.9	32.8	33.6	35.3	39.6	32.0	32.3	43.5	48.7	36.4	32.6	34.3	23.9	25.1	34.4
P-STMO[32]	ECCV'22	243	31.3	35.2	32.9	33.9	35.4	39.3	32.5	31.5	44.6	48.2	36.3	32.9	34.4	23.8	23.9	34.4
HDFormer[33]	IJCAI'23	96	29.6	33.8	31.7	31.3	33.7	37.7	30.6	31.0	41.4	47.6	35.0	30.9	33.7	25.3	23.6	33.1
MixSTE[26]	CVPR'22	243	30.8	33.1	30.3	31.8	33.1	39.1	31.1	30.5	42.5	44.5	34.0	30.8	32.7	22.1	22.9	32.6
DUE[37]	MM'22	300	30.3	34.6	**29.6**	31.7	**31.6**	38.9	31.8	31.9	**39.2**	**42.8**	**32.1**	32.6	31.4	25.1	23.8	32.5
STCFormer[35]	CVPR'23	243	**29.5**	33.2	30.3	31.0	33.0	38.0	30.4	29.4	41.8	45.2	33.6	**29.5**	31.6	**21.3**	**22.6**	32.0
MotionBERT[23]	ICCV'23	243	30.8	32.8	32.4	28.7	34.3	38.9	30.1	30.0	42.5	49.7	36.0	30.8	**22.0**	31.7	23.0	32.9
PoseRetNet[29]	ECCV'24	243	30.8	33.1	31.3	31.8	33.4	37.7	30.1	30.5	43.4	45.5	34.3	30.3	31.5	21.4	22.7	32.5
KTPFormer[28]	CVPR'24	243	30.1	32.3	**29.6**	30.8	32.3	**37.3**	**30.0**	30.2	41.0	45.3	33.6	29.9	31.4	21.5	**22.6**	**31.9**
PoseAnchor[27]	ICCV'25	243	30.6	**32.1**	29.9	31.5	33.5	38.2	30.6	**29.2**	41.3	44.0	34.2	30.0	32.1	21.8	22.8	32.1
MSGPose	Ours	243	31.3	33.4	31.8	**28.2**	34.5	39.6	31.8	29.8	42.2	46.4	35.7	30.6	31.0	23.1	23.4	32.2

此外,相比于采用Retentive机制的PoseRetNet[29],MSGPose依托Mamba架构在长程依赖建模上的优势,并结合多语义图卷积对局部结构细节的显式增强,从而获得了更为显著的性能提升。

此外,为进一步评估模型在2D到3D的提升能力,并规避上游2D姿态检测器可能带来的噪声,本文在输入为无噪声的真实2D姿态序列条件下进行了对比实验。如表3所示,MSGPose在MPJPE上达到了12.7 mm,对比MotionAGFormer[38]在MPJPE上显著降低了4.6 mm。这表明,当输入为高保真度的关节点坐标时,模型能够更好地利用人体结构先验(如双边对称性)在几何空间中的表达,进而获得了最佳的提升效果,这一实验结果有力证明了MSGPose在2D到3D映射任务中的有效性。

表3 本文网络与对比方法在Human3.6M数据集上基于真实2D姿态序列作为输入的MPJPE对比结果

Table 3 Performance comparison on the Human3.6M dataset in terms of MPJPE, taking ground-truth 2D poses as input

Method	Reference	T	Params/M	MPJPE/mm
MHFormer[30]	CVPR'22	351	30.9	30.5
P-STMO[32]	ECCV'22	243	6.2	29.3
FTCM[31]	TCSVT'24	243	4.49	25.1
STCFormer[35]	CVPR'23	243	18.9	22.0
MixSTE[26]	CVPR'22	243	33.8	21.6
HDFormer[33]	IJCAI'23	96	3.7	21.6
MotionBERT[23]	ICCV'23	243	42.4	17.8
MotionAGFormer[38]	WACV'24	243	19.0	16.2
KTPFormer[28]	CVPR'24	243	33.7	18.1
PoseRetNet[29]	ECCV'24	243	25.2	21.5
MSGPose	Ours	243	13.3	**12.7**

3.2.2 MPI-INF-3DHP数据集上的实验结果分析

为了进一步验证MSGPose模型的性能,本文在MPI-INF-3DHP数据集上与多种主流算法进行了对比实验。考虑到该数据集姿势序列的特殊性,本文选择视频帧长度为81进行对比实验。如表4所示,MSGPose的PCK、AUC和MPJPE分别取得了98.0 mm、82.3 mm与14.5 mm。相较于MotionAGFormer[38],MPJPE进一步下降了1.7 mm。该结果表明,MSGPose模型具有良好的泛化能力,能够在更复杂的场景中保持较优的表现。

表4 在MPI-INF-3DHP数据集上的对比结果 单位:mm

Table 4 Performance comparison on the MPI-INF-3DHP dataset unit:mm

Method	Reference	T	PCK ↑	AUC ↑	MPJPE ↓
MixSTE[26]	CVPR'22	27	94.4	66.5	54.9
HDFormer[33]	IJCAI'23	96	98.7	72.9	37.2
P-STMO[32]	ECCV'22	81	95.4	75.8	32.2
FTCM[31]	TCSVT'24	81	97.9	79.8	31.2
PoseFormerV2[25]	CVPR'23	81	97.9	78.8	27.8
STCFormer[35]	CVPR'23	81	98.7	83.9	23.1
GLA-GCN[18]	ICCV'23	81	98.5	79.1	27.7
PoseRetNet[29]	ECCV'24	81	99.1	84.4	22.2
MotionAGFormer[38]	WACV'24	81	98.2	85.3	16.2
PoseAnchor[27]	ICCV'25	81	**99.3**	**88.1**	17.2
MSGPose	Ours	81	98.0	82.3	**14.5**

3.2.3 资源开销与推理效率分析

为进一步评估模型规模与资源开销,表5对比了MSGPose与代表性方法在参数量(Params)、计算量(MACs)及精度(MPJPE)上的表现。从表5可见,MSGPose在MPJPE上取得最优结果(38.9 mm),同时参

数量仅为 13.3 M,约为 MotionBERT 的 31%。在计算量方面,MSGPose 的 MACs 为 37.4 G,显著低于 MixSTE(139 G)与 MotionBERT(174.8 G)。上述结果表明,所提出的多语义结构建模能够在较小模型规模与更低计算量下实现更优的姿态回归精度,体现出更好的精度-规模权衡。

表 5　本文网络与对比方法在 Human3.6M 数据集上的计算开销对比

Table 5　Computational cost comparison between the proposed network and state-of-the-art methods on the Human3.6M dataset

Method	Reference	T	Params/M	MACs/G	Latency/ms	FPS	MPJPE
MixSTE[26]	CVPR'22	243	33.8	139	9.86	101.3	40.9
MotionBERT[23]	ICCV'23	243	42.4	174.8	17.3	58	39.2
PoseRetNet[29]	ECCV'24	243	25.2	104.5	79.18	12.6	40.4
MSGPose	Ours	243	13.3	37.4	24.8	40.3	38.9

3.3　消融实验

为了验证 MSGPose 模型各个组成模块的有效性与不可或缺性,以及各种不同结构设计对模型的影响,本文对 MSGPose 模型在 Human3.6M 数据集上进行消融实验,并使用 MPJPE 作为性能评估指标,采用 2D 姿态估计检测模型 SH 作为输入。

3.3.1　各模块性能分析

为了分析 MSGPose 模型各模块对网络整体性能的影响,本文以原始 GCN 与注意力双分支结构作为基线模型,并保持其他实验设置不变,结果如表 6 所示。首先,将基线模型中的 GCN 卷积替换为本文提出的 MSDG 模块后,MPJPE 由 40.1 mm 下降至 39.5 mm。这一结果表明,MSDG 通过在多语义分支上分离建模关节关系,有效增强了空间特征的表达能力,相较于单一的 GCN 卷积更具优势。

表 6　MSGPose 各模块及组合方式消融实验结果

Table 6　Results of the ablation study on different components and combination strategies in the proposed MSGPose

Method	MPJPE/mm
Baseline	40.1
Baseline+MSDG	39.5
Baseline+SGM	39.6
Baseline+MSDG+SGM (Sequential)	39.2
Baseline+MSDG+SGM (Parallel)	38.9

随后,在保持 Baseline 的 GCN 卷积操作不变的前提下,将其注意力分支替换为 SGM 模块,MPJPE 降低至 39.6 mm,说明由多语义图卷积引导的状态空间模型能够更充分地捕捉人体姿态序列中的时空依赖关系,从而提升 3D 姿态估计的效果。最后,本文进一步比较了串行结构与并行结构的差异。结果表明,当 MSDG 模块与 SGM 模型以并行方式进行特征提取时,二者能进行互补的自适应融合,从而获得优于串行结构的性能表现。

3.3.2　模型深度与宽度

表 7 展现了 MSGPose 不同的超参数对模型 MPJPE 的影响,其中主要的超参数为模型框架的层数 N,模型输入的隐藏通道维度 D。本文设置了两组超参数,分别在特征维度为 64 与 128 时,逐步提升模型层数来寻找最优性能。

表 7　模型层数与特征维度对网络性能的影响

Table 7　Effect of different numbers of layers and feature dimensions on model performance

层数(N)	特征维度(D)	参数量/M	MPJPE/mm
4	64	0.93	48.6
8	64	1.8	43.5
16	64	3.6	41.3
4	128	3.4	40.1
8	128	6.7	39.3
16	128	13.3	38.9

3.3.3　对多语义图的消融

为验证 MSDG 模块中多语义图的有效性,本文将基线模型中 GCN 卷积分支替换为含有不同语义图的 MSDG 模块来探究对性能的影响,如表 8 所示。实验表明,若单独使用物理图与对称图时,结果并没有显著提升,但当两者结合并行分离图卷积时,结果出现显著变化,MPJPE 从 40.1 mm 降至 39.0 mm。这证明了将多种语义结合的并行分离卷积对关节建模具有重要意义。

接下来为系统性验证 SGM 模块内部架构设计的合理性,本文保持 MSDG 模块分支不变,将注意力分支替换为 SGM 模块,以研究不同的特征流处理方式对模型性能的影响。如表 9 所示,当 MSDG 模块仅使用空间优先 Mamba(ST-Mamba #e)与仅使用时序优先 Mamba(TS-Mamba #g)时,模型性能均出现了小幅的性能退化。这表明捕捉单帧内的空间关节依赖与建模帧间的时序依赖对于精确的 3D 姿态估计而言同等重要。随后,本文探究了融合这两种能力的最佳方式。当采用并行融合的策略时,有着更优异的性能,将误差进一步降低至 39.0 mm。

最终,为了构建完整的 SGM 模块,在时空并行

Mamba之前引入了多语义图卷积(SGCN)作为引导。该模块的加入使模型性能进一步提升,MPJPE降至38.9 mm。这一结果验证了先结构聚合、后序列建模策略的有效性:通过前置图卷积,模型能够显式地向特征空间注入人体结构先验,从而弥补了纯Mamba架构在局部关节依赖建模上的不足。

表8 MSDG各模块对网络性能的影响

Table 8 Effect of individual modules in MSDG on network performance

Method	GCN+TCN	物理图(A_{phy})	对称图(A_{sym})	T-GCN	MPJPE/mm ↓
Baseline	✓				40.1
#a		✓		✓	39.4
#b			✓	✓	39.7
#c		✓	✓	✓	39.0

表9 SGM各模块对网络性能的影响

Table 9 Effect of different modules in SGM on network performance

方法	ST-Mamba	TS-Mamba	Forward	Backward	SGCN	MPJPE/mm ↓
#e	✓		✓			39.6
#f	✓		✓	✓		39.4
#g		✓	✓	✓		39.5
#q1	✓	✓	✓	✓		39.2
#q2	✓	✓	✓	✓		39.0
#z	✓	✓	✓	✓	✓	38.9

3.4 可视化结果

为了更加直观地展现本文提出算法对人体姿态估计的结果,同时进一步评估本文模型的泛化能力,本文在互联网的真实场景视频序列上进行了3D人体姿态估计的可视化结果,如图6所示。这些视频序列涵盖了多种复杂身体运动姿势的运动场景以及身体快速的动态变化。从图6结果中可以看出,即使在视频中对于关节遮挡较为严重的动作,本文方法依然能准确地估计3D姿态,图中黑色虚线圆圈处标记了本文方法与对比方法姿态估计差异较大的区域。

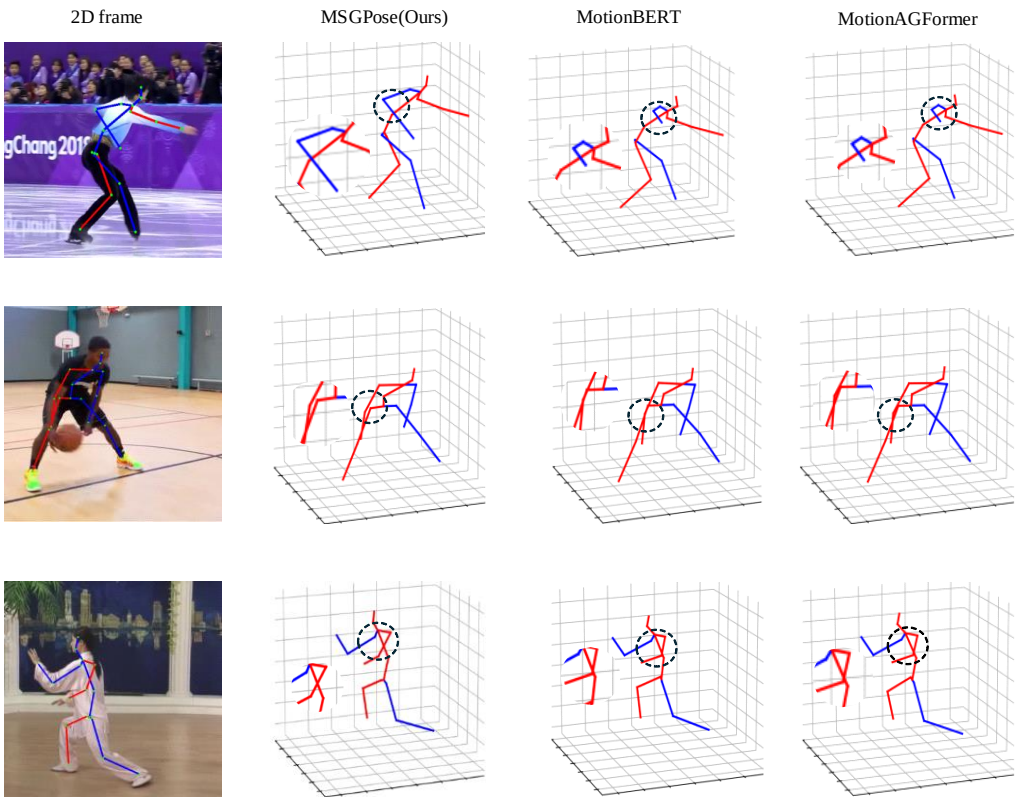


图6 本文方法与对比方法在场景视频上的可视化结果

Figure 6 Visualization results of the proposed method and comparison methods on in-the-wild videos

3.5 失败案例与局限性分析

尽管MSGPose在Human3.6M与MPI-INF-3DHP上取得了较优的结果,但在更具挑战性的真实场景中仍存在一定局限性。特别地,当出现严重遮挡或快速大

幅度运动时,上游2D姿态估计器的关键点定位易受到运动模糊与自遮挡的影响而产生漂移或错检,进而在2D-to-3D提升过程中被放大,导致三维重建出现肢体错连、左右肢体交换或深度漂移等失效现象。如图7所示,在遮挡场景中,部分关节被遮挡导致2D观测不完整,使得模型难以恢复可靠的空间结构;在快动作场景中,末端关节(如腕、踝)在相邻帧间变化剧烈,造成时序特征不稳定,从而引发3D骨架形态畸变。未来工作将进一步探索面向遮挡与快动作的鲁棒建模策略,结合更强的2D检测与时序一致性约束,以及融合更丰富的运动先验以提升复杂场景下的重建稳定性。需要指出的是,本文研究对象为单目单人2D到3D的提升。多人场景除lifting外还需多目标检测与身份关联/跟踪等前端模块以保证个体一致性,本文未对此展开;本文方法可作为单人lifting后端,与现有多人2D检测/跟踪框架组合实现扩展。

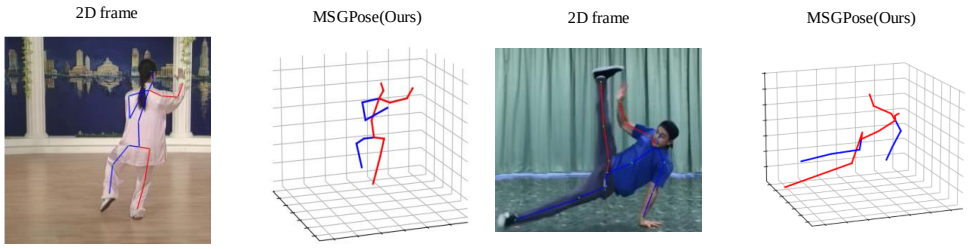


图7　失败案例分析

Figure 7　Analysis of failure cases of our MSGPose

4　结论

本文提出了一种基于多语义动态分离图卷积与语义图引导的时空双向Mamba的双流并行模型(MSGPose)。该模型通过设计MSDG与SGM两个模块,实现了对人体关节点序列的空间与时间特征的并行建模。在MSDG模块中,利用多语义图卷积并行建模人体骨架的局部依赖关系,并通过动态构建的时序相似度矩阵进一步建模跨帧关联;在SGM模块中,则在近年来表现突出的Mamba架构前引入多语义图卷积引导,以注入结构先验并为状态空间模型提供人体先验。最终,两条特征流在融合阶段通过自适应机制实现互补,从而显著增强了模型的表征能力。实验结果表明,MSGPose在以真实2D姿态序列作为输入的Human3.6M数据集与MPI-INF-3DHP数据集上超越了当前的主流算法,验证了所提出框架的有效性。然而值得注意的是,MSGPose的性能对输入的2D序列质量较为敏感,未来的研究可以从优化2D姿态检测器或提升输入数据质量的角度进一步改进系统整体性能。

参考文献

[1] 李雨桐, 马苗, 陈建芮. 融合动作描述生成与跨模态语义对齐的骨架动作识别方法[J]. 电子学报, 2025, 53(11): 4116-4131.
Li Yutong, Ma Miao, Chen Jianrui. Leveraging action description generation and cross-modal semantic alignment for skeleton-based action recognition[J]. Acta Electronica Sinica, 2025, 53(11): 4116-4131. (in Chinese)

[2] Zheng J X, Shi X W, Gorban A, et al. Multi-modal 3D human pose estimation with 2D weak supervision in autonomous driving[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2022: 4477-4486.

[3] Wang K Z, Lin L, Jiang C H, et al. 3D human pose machines with self-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(5): 1069-1082.

[4] Peng S, Hu J W. 3D human pose estimation in video with temporal and spatial transformer[C]//International Conference on Image, Signal Processing, and Pattern Recognition. SPIE, 2023: 136.

[5] Chen Y L, Wang Z C, Peng Y X, et al. Cascaded pyramid network for multi-person pose estimation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7103-7112.

[6] Newell A, Yang K Y, Deng J. Stacked hourglass networks for human pose estimation[M]//Computer Vision - ECCV 2016. ChamSpringer International Publishing2016: 483-499.

[7] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 5686-5696.

[8] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[PP/OL]. V4. arXiv (2017-02-22)[2025-09-28]. https://doi.org/10.48550/arXiv.1609.02907.

[9] Ci H, Wang C Y, Ma X X, et al. Optimizing network structure for 3D human pose estimation[C]//2019 IEEE/CVF In-

ternational Conference on Computer Vision. Piscataway: IEEE, 2019: 2262-2271.

[10] Liu K K, Zou Z M, Tang W. Learning global pose features in graph convolutional networks for 3D human pose estimation[M]//Computer Vision - ACCV 2020. Cham-Springer International Publishing2021: 89-105.

[11] Zhao L, Peng X, Tian Y, et al. Semantic graph convolutional networks for 3D human pose regression[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 3420-3430.

[12] Hu W B, Zhang C G, Zhan F N, et al. Conditional directed graph convolution for 3D human pose estimation[C]// Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 602-611.

[13] Smith A C, Brown E N. Estimating a state-space model from point process observations[J]. Neural Computation, 2003, 15(5): 965-991.

[14] Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces[PP/OL]. V2. arXiv (2024-05-31) [2025-09-28]. https://doi.org/10.48550/arXiv.2312.00752.

[15] Dao T, Gu A, Hwang S, et al. Hydra: Bidirectional state space models through generalized matrix mixers[C]//Advances in Neural Information Processing Systems 37. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 110876-110908.

[16] ZHANG Z. Group graph convolutional networks for 3D human pose estimation[C]//Proceedings of the British Machine Vision Conference. London: BMVA Press, 2022: 1019.

[17] Zou Z M, Tang W. Modulated graph convolutional network for 3D human pose estimation[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 11457-11467.

[18] Yu B X B, Zhang Z, Liu Y X, et al. GLA-GCN: Global-local adaptive graph convolutional network for 3D human pose estimation from monocular video[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 8784-8795.

[19] Gu A, Goel K, Ré C. Efficiently modeling long sequences with structured state spaces[PP/OL]. V3. arXiv (2022-08-05)[2025-09-28]. https://doi.org/10.48550/arXiv.2111.00396.

[20] Zhu L H, Liao B C, Zhang Q, et al. Vision mamba: Efficient visual representation learning with bidirectional state space model[PP/OL]. V3. arXiv (2024-11-14)[2025-09-28]. https://doi.org/10.48550/arXiv.2401.09417.

[21] Jiao J B, Liu Y, Liu Y F, et al. VMamba: Visual state space model[C]//Advances in Neural Information Processing Systems 37. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 103031-103063.

[22] Huang Y L, Liu J S, Xian K, et al. PoseMamba: Monocular 3D human pose estimation with bidirectional global-local spatio-temporal state space model[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(4): 3842-3850.

[23] Zhu W T, Ma X X, Liu Z Y, et al. MotionBERT: A unified perspective on learning human motion representations[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 15039-15053.

[24] Loshchilov I, Hutter F. Decoupled weight decay regularization[PP/OL]. V3. arXiv (2019-01-04) [2025-09-28]. https://doi.org/10.48550/arXiv.1711.05101.

[25] Zhao Q T, Zheng C, Liu M Y, et al. PoseFormerV2: Exploring frequency domain for efficient and robust 3D human pose estimation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 8877-8886.

[26] Zhang J L, Tu Z G, Yang J Y, et al. MixSTE: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 13222-13232.

[27] KIM J H, HAN J, LEE S W. PoseAnchor: Robust root position estimation for 3D human pose estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2025: 7079-7088.

[28] Peng J H, Zhou Y H, Mok P Y. KTPFormer: Kinematics and trajectory prior knowledge-enhanced transformer for 3D human pose estimation[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 1123-1132.

[29] Zheng K L, Lu F X, Lv Y H, et al. 3D human pose estimation via non-causal retentive networks[M]//Computer Vision-ECCV 2024. ChamSpringer Nature Switzerland2024: 111-128.

[30] Li W H, Liu H, Tang H, et al. MHFormer: Multi-hypothesis transformer for 3D human pose estimation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 13137-13146.

[31] Tang Z H, Hao Y B, Li J, et al. FTCM: Frequency-temporal collaborative module for efficient 3D human pose estimation in video[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(2): 911-923.

[32] Shan W K, Liu Z H, Zhang X F, et al. P-STMO: Pre-

trained spatial temporal many-to-one model for3D human pose estimation[C]//Computer Vision-ECCV 2022. Cham: Springer, 2022: 461-478.

[33] Chen H Y, He J Y, Xiang W M, et al. HDFormer: High-order directed transformer for 3D human pose estimation [PP/OL]. V2. arXiv (2023-05-22) [2025-09-28]. https://doi.org/10.48550/arXiv.2302.01825.

[34] Shan W K, Liu Z H, Zhang X F, et al. Diffusion-based 3D human pose estimation with multi-hypothesis aggregation[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 14715-14725.

[35] Tang Z H, Qiu Z F, Hao Y B, et al. 3D human pose estimation with spatio-temporal criss-cross attention[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 4790-4799.

[36] Li W H, Liu H, Ding R W, et al. Exploiting temporal contexts with strided transformer for 3D human pose estimation[J]. IEEE Transactions on Multimedia, 2023, 25: 1282-1293.

[37] Zhang J L, Chen Y J, Tu Z G. Uncertainty-aware 3D human pose estimation from monocular video[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 5102-5113.

[38] Mehraban S, Adeli V, Taati B. MotionAGFormer: Enhancing 3D human pose estimation with a transformer-GCNFormer network[C]//2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2024: 6905-6915.

作者简介

李　俊　男,1978年出生于湖北省黄石市。现为武汉科技大学计算机科学与技术学院副教授。主要研究方向为机器视觉、智能优化算法。

E-mail: lijun@wust.edu.cn

李　昱　男,2002年出生于湖北省鄂州市。现为武汉科技大学计算机科学与技术学院硕士研究生。主要研究方向为三维人体姿态估计。

E-mail: 4realyl@wust.edu.cn

陈　黎　男,1977年出生于湖北省武汉市。现为武汉科技大学计算机科学与技术学院教授、博士生导师。主持国家自然科学基金4项,主要研究方向为多模态大模型、机器视觉、人工智能。

E-mail: chenli@wust.edu.cn