DOI: 10.13382/j.jemi.B2408024

基于双阶时空交叉卷积 Transformer 的三维人体姿态估计方法*

邹 宇[1,2] 周先春[1] 潘志庚[1] 蔡创新[1]
(1. 南京信息工程大学人工智能学院 南京 210044;2. 江苏开放大学科学技术处 南京 210036)

摘 要:在三维人体姿态估计领域,尽管基于 Transformer 的方法已取得显著进展,但在处理大规模关节亲和力矩阵,尤其是动态视频序列分析时,依然面临显著的计算挑战。为此,提出一种双阶时空交叉卷积 Transformer 模型(dual-stage spatio-temporal convolutional transformer, DSTCFormer),旨在高效融合时空信息并提升三维姿态估计的精度与鲁棒性。通过并行时空路径设计、卷积强化注意力机制及结构驱动的位置编码技术,解决现有方法在处理长时序视频时计算效率低、时空特征建模不足的问题。在空间路径中,提出卷积位置嵌入模块,通过局部邻域卷积显式建模人体骨骼拓扑结构;在时间路径中,设计轴向特异性自注意力机制,捕捉跨帧关节运动轨迹。引入卷积多尺度注意力(convolutional multi-scale attention, CMSA)模块,结合深度可分离卷积与特征变换层,实现多尺度时空特征的交叉融合。此外,通过逐步细化关节间时空依赖关系,降低计算开销。实验表明,在 Human3.6M 数据集上,DSTCFormer 以 243 帧输入取得 P1 协议下 40.1 mm 的平均关节位置误差,较 PoseFormer、MixSTE 和 STCFormer 分别降低 4.1、0.8 和 0.4 mm;在 MPI-INF-3DHP 数据集上,PCK@150 mm 和曲线下面积(AUC)分别达到 99.1%和 85.2%,较基准模型提升 0.8 mm 误差优势。提出的方法为三维人体姿态估计提供了高效的理论框架,并为虚拟现实、人机交互等应用奠定了技术基础。

关键词:人体姿态估计;Transformer;时空信息关联;关节点间运动
中图分类号:TN60;TP29 **文献标识码**:A **国家标准学科分类代码**:510.40

3D human pose estimation with dual-stage spatio-temporal convolutional transformer

Zou Yu[1,2] Zhou Xianchun[1] Pan Zhigeng[1] Cai Chuangxin[1]
(1. School of Aritificial Intelligence, Nanjing University of Information Science & Technology, Nanjing 210044, China;
2. Science and Technology Office, Jiangsu Open University, Nanjing 210036, China)

Abstract: In recent years, transformer-based methods have achieved remarkable progress in the field of 3D human pose estimation. However, current approaches are still confronted with two major challenges. First, the computational inefficiency arises from the quadratic complexity of global self-attention when processing large-scale joint affinity matrices in dynamic video sequences. This issue significantly hampers the real-time performance of the models. Second, the suboptimal spatiotemporal feature fusion restricts the model's ability to capture fine-grained motion patterns and structural dependencies between joints, leading to less accurate pose estimation results. To tackle these limitations, this paper proposes a novel architecture named the dual-stage spatio-temporal convolutional transformer (DSTCFormer). The key innovation of DSTCFormer lies in its decoupling of spatiotemporal feature learning into parallel spatial and temporal pathways. Specifically, the convolutional multi-scale attention (CMSA) module is introduced to hierarchically aggregate local and global correlations through convolution-enhanced multi-head attention. In the spatial pathway, convolutional position embeddings are utilized to encode skeletal topology, enabling the model to focus on intra-frame joint relationships. Meanwhile, the temporal pathway captures inter-frame motion coherence via axial-specific self-attention. Moreover, a cross-stage fusion mechanism is designed to integrate multi-scale spatiotemporal features through depthwise separable convolutions and feature transformation layers, which ensures efficient computation and robust feature representation. Extensive experiments conducted on the Human3.6M and MPI-

收稿日期:2024-12-08 Received Date: 2024-12-08
*基金项目:国家自然科学基金(62072150)、江苏省产学研合作项目(BY20230641)资助

INF-3DHP datasets demonstrate the superiority of DSTCFormer. Under Protocol 1 (P1), DSTCFormer achieves a state-of-the-art Mean Per Joint Position Error (MPJPE) of 40.1mm on Human3.6M with 243 input frames, outperforming PoseFormer (44.3 mm), MixSTE (40.9 mm), and STCFormer (40.5 mm). On the MPI-INF-3DHP dataset, it attains a percentage of correct keypoints at 150 mm (PCK@150 mm) of 99.1% and an area under curve (AUC) of 85.2%, surpassing existing methods by 0.4% and 1.3%, respectively. In summary, the proposed method not only advances the theoretical frameworks for spatiotemporal modeling but also offers practical implications for real-time applications, paving the way for more efficient and accurate 3D human pose estimation in various scenarios.

Keywords: human pose estimation; Transformer; spatiotemporal information association; articular interpoint movement

0 引 言

近年来,三维人体姿态估计(3D human pose estimation, 3DHPE)技术在动作识别[1-2]、虚拟现实[3-4]和人机交互[5-6]等多领域的广泛应用而备受关注。其核心任务在于从单视角输入数据(如图像或视频序列)中准确预测人体关节在三维空间中的位置。尤其是针对基于二维姿态数据的三维姿态估计[7-11],研究者们已经取得了显著的进展,旨在克服单帧图像深度信息不足的挑战。

在基于视频的单目三维姿态估计中,一种常见的策略是采用"两步法"[12-13],即先利用二维姿态检测器提取关键点的二维坐标[14],然后通过坐标提升方法回归得到三维关节位置[15]。尽管这类方法取得了显著的成就,但深度估计的固有模糊性仍然给准确的三维姿态回归带来了挑战。为了应对这些挑战,研究者们探索了从传统卷积网络到图卷积网络再到 Transformer 自注意力等多种方法。

早期的三维人体姿态估计方法主要依赖于卷积神经网络(convolutional neural network, CNN)[16-17]、全连接层(fully connected layers, FC)[18]以及多层感知机(multilayer perceptron, MLP)[19]等传统深度网络,通过在视频序列中整合空间和时间信息以提升姿态估计的精度。一些研究还探索了无监督或弱监督的学习策略[20],以在缺乏大量标注数据的情况下实现良好的性能。然而,这些方法在处理长时间序列或复杂动态场景时,由于缺乏对时空动态的精细建模能力,往往难以保持高精度。

随着图卷积网络(graph convolutional network, GCN)的发展,研究者们开始利用其捕捉人体骨骼的结构信息,从而显著提升了三维姿态估计的精度。例如,Graphormer[21]通过融合局部和全局特征以及整合多视角信息,实现了准确的三维姿态回归。Jiang 等[22]创新性地提出了一种概率三角化模块,嵌入到多视角三维人体姿态估计中以增强模型对未校准场景的泛化能力。Liu 等[23]则提出了精细时间金字塔压缩与放大模型(refined temporal pyramidal compression-and-amplification, RTPCA),通过时间金字塔压缩与放大以及 XLR 模块,显著提升了长时序特征的学习效果。然而,GCN 在处理长视频序列时计算复杂度会显著增加,并且在建模复杂的时间依赖关系方面仍然存在一定的局限性。为了更好地捕捉复杂动作的时空依赖,越来越多的研究者转向了 Transformer 架构[24]。

Transformer 模型凭借其强大的自注意力机制,能够有效地建模帧内关节间的空间关联以及跨帧的时间依赖,因此被广泛应用于三维人体姿态估计。这些方法通过整合多尺度特征和注意力机制,在提升姿态估计精度方面取得了显著进展。例如,PoseFormer 模型[25]凭借级联的 Transformer 层有效地捕捉了帧间的空间关系,代表了利用单一时空自注意力进行建模的方法;为了更有效地处理时空信息,研究者们提出了时空分解方法,如 P-STMO 模型[26]通过时空交叉注意力模块在通道维度上分离时空特征的学习;StridedFormer[27]、CrossFormer[28]、MHFormer[29]、MixSTE[30]等也在时空注意力分解和多尺度特征提取方面进行了探索。此外,为了更好地处理遮挡和动作变化,研究者们还提出了时空交叉解耦方法,如 STCFormer[31]利用空间-时间交叉注意力;3D-LFM[32]则巧妙利用了 Transformer 的排列等价,增强对遮挡的处理能力,并能泛化至未知类别;李功浩等[33]则提结合了 Transformer 与语义图卷积的模型,利用离线二维姿态检测来获得骨盆中心坐标,从而提升三维预测的可靠性。

然而,尽管上述基于 Transformer 的方法大大增强了对长时间序列和全局依赖的建模能力,但依旧面临两方面主要挑战:1)二维至三维的时空相关性学习困难,直接在高维特征上执行全局注意力,可能会加大模型优化难度;2)计算复杂度偏高,随帧数或关节点数的增加,自注意力机制需构建更大规模的亲和矩阵,计算和内存开销呈二次方增长,制约了模型在长序列或实时场景中的应用潜力。

为克服上述局限,本文提出了一种新颖的名为双阶时空交叉 Transformer 架构(dual-stage spatio-temporal convolutional transformer, DSTCFormer)的三维人体姿态估计方法,专门针对在连续帧中融合空间与时间信息的痛点设计,旨在更好的实现从二维到三维的姿态转换。其核心思想是在并行的空间和时间路径中分别关注关节的空间连接性和时间上的运动轨迹,并通过卷积多尺度自注意力(convolutional multi-scale attention, CMSA)模块进行有效整合。时间路径侧重于捕捉不同帧间相同关节运动的连贯性,而空间路径则专注于单帧内各关节间的

相互关系。此外,为了有效地建模人体骨骼的结构信息,在 DSTCFormer 中引入了一种位置嵌入机制,通过在局部关节邻域施加卷积核来捕捉相对位置特征,增强模型对局部骨骼动态的感知能力。

本文首先,提出了 DSTCFormer,一种新颖的双阶段时空 Transformer 架构,该架构通过融合结构强化位置编码与多重 DSTC 模块堆叠,显著提升了三维姿态估计的性能。其次,设计了 DSTC 模块,采用创新的空间—时间注意力分解策略,旨在提升三维人体姿态估计的效能与经济性,在长序列时空建模中取得了更优的效果与可扩展性。在 Human3.6M 和 MPI-INF-3DHP 数据集上进行广泛验证,与现有领先技术相比,DSTCFormer 在精度与泛化能力方 Z 面均表现突出,并通过进一步的消融实验量化了卷积注意力交叉融合与位置嵌入对模型性能的贡献。

1　相关工作

1.1　单目三维人体姿态估计

单目三维人体姿态估计领域涉及将单视角的输入数据,如静态图像或二维坐标,用于在三维空间中准确重构人体关节的位置。在此领域,单阶段与双阶段策略是两种主流技术路线。单阶段方法[34]直接将输入图像映射到三维姿态,通常依赖卷积神经网络提取特征并通过回归网络进行预测,有时也会结合二维关键点检测。双阶段方法则通常先检测二维关键点[35-36],然后再通过三维姿态估计网络完成姿态重建,在处理复杂姿态时往往能取得更高的精度。本文的研究重点为基于视频序列的三维姿态估计,这些方法通常在单目方法的基础上,通过融合时间信息来提升性能。

1.2　基于视频序列的三维姿态估计

为解决三维坐标回归中深度模糊性的问题,近期研究开始探索融合相邻帧的时间信息。例如,Pavllo 等[13]提出了一种时间全卷积网络(temporal fully-convolutional network, TCN),通过在时间维度上进行卷积操作来建模局部时空上下文。Chen 等[36]将姿态估计分解为骨骼长度和方向的预测。Liu 等[37]在 TCN 的基础上引入注意力机制,以识别关键帧和关键姿态。Xue 等[38]对具有相似运动模式的关节进行分组,并计算其内部时间相关性。还有一些研究采用时空卷积网络同时捕捉关节间的空间和时间相关性。这些方法利用卷积操作在一定程度上提升了性能,但对于捕捉长距离时间依赖和全局时空关系仍存在局限性。随着 Transformer 在序列建模方面的成功,研究者们开始探索将其应用于视频序列的三维人体姿态估计。

1.3　基于 Transformer 的时空三维姿态估计

基于 Transformer 的架构因其强大的序列建模能力和自注意力机制,在三维人体姿态估计领域取得了显著进展。最初的研究主要集中在直接应用 Transformer 的自注意力机制来同时建模帧内空间关联和跨帧时间依赖。Zheng 等[25]运用基于 Transformer 的架构捕捉关节间的相互作用及时序依赖。为了降低计算复杂度并更好地分离空间和时间特征的学习,研究者们提出了时空分解方法。Zeng 等[35]提出了一种新的时间感知动态卷积方法,可根据骨架的物理拓扑及其特征进行动态调整。MHFormer[29]则提出使用空间 Transformer 编码器生成多种姿态假设,通过不同时间 Transformer 模块构建多级全局依赖关系。此外,StridedFormer[27]和 CrossFormer[28]通过结合一维时空和一维空间卷积,有效地强化了时空局部性的捕捉,也可以看作是在局部层面进行时空信息分解和交互的尝试。近年来,研究者们开始关注更精细化的时空信息交互和解耦。上述提到的 PoseFormer[25]可以通过级联 Transformer 层有效地捕捉帧间的空间关系,虽然没有明确的时空分解,但其结构可以看作是在不同层级上进行时空信息交互。STCFormer[31]利用空间时间交叉注意力来有效提升对人体遮挡和动作变化的鲁棒性。

尽管上述方法在三维人体姿态估计领域取得了显著进展,但在处理大规模关节亲和力矩阵,尤其是在长时序动态视频序列分析时,依然面临显著的计算挑战,并且时空特征的融合仍然有待进一步优化。本文提出的 DSTCFormer 模型正是针对现有方法的不足,创新性地采用了双阶时空交叉卷积 Transformer 架构。

2　双阶时空交叉的三维人体姿态估计的网络设计

本文设计的 DSTCFormer 通过同步处理时空信息,并对增强的时空数据进行建模。其核心在于将关节特征分为两部分,分别通过 CMSA 模块捕捉时间轨迹相关性和空间关节连接性。通过堆叠多个 DSTC 模块,构建了 DSTCFormer 网络,以实现高精度的三维人体姿态估计,流程如图 1 所示。

通过空间卷积和时间卷积模块分别提取视频数据的空间和时间特征,利用空间注意力和时间注意力机制对特征进行加权,突出关键信息。特征经过融合后,通过 MLP 作进一步处理。同时,框架还引入了可变形卷积(deformable convolution, DC)操作,结合时间到空间(tempo-spatial correlation, TS)和空间到时间(spatio-temporal correlation, ST)相关性特征转换,以捕捉复杂的人体姿态时空结构。通过融合空间和时间的深度特征信息,提高三维人体姿态估计的准确性和鲁棒性,在处理动

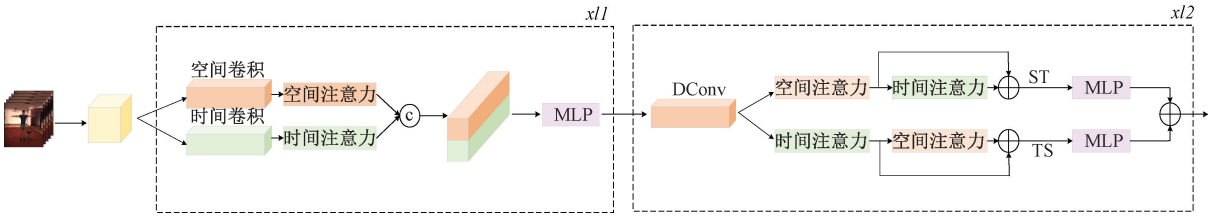


图1　框架流程图

Fig. 1　Overall framework process

态视频序列时,能够更为有效地捕捉人体关节的运动轨迹和姿态变化。

2.1　基本理论

本文针对基于Transformer网络的模型进行了深入优化。Transformer网络[24]作为一种先进的多元表征学习架构,主要融合了多头自注意力(multi-head self-attention, MSA)与MLP两大核心模块,旨在提升模型的表达能力。

MSA 在模块中输入序列 $\boldsymbol{Q} \in \boldsymbol{R}^{N\times D}$ 首先经过线性变换,分别衍生出查询(query)序列 $\boldsymbol{Q} \in \boldsymbol{R}^{N\times D}$、键(key) $\boldsymbol{K} \in \boldsymbol{R}^{N\times D}$ 和值(value) $\boldsymbol{V} \in \boldsymbol{R}^{N\times D}$ 3组向量。在此过程中,序列长度由符号 N 表示,而向量维度则由 D 来指代。缩放点积注意力(scaled sot-product attention, SDPA)的计算方式如下所示:

$$MSA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Softmax(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{D})\cdot\boldsymbol{V} \tag{1}$$

在MSA中,输入序列首先被分割成 H 个部分,每个部分对应一个注意力头。每个头独立地计算其注意力输出,然后将这些输出组合,以捕获不同子空间的信息。

MLP 由两个线性层组成,用于非线性变换和特征操作,替代传统的卷积操作:

$$MLP(x) = \delta(xW_1 + b_1) + W_2 + b_2 \tag{2}$$

式中:σ 表示高斯误差线性单元(Gaussian error linear unit,GELU)激活函数,$W_1 \in \boldsymbol{R}^{D\times D_m}$ 和 $W_2 \in \boldsymbol{R}^{D_m\times D}$ 分别表示两个线性层的权重,而 $b_1 \in \boldsymbol{R}^{D_m}$ 和 $b_2 \in \boldsymbol{R}^{D}$ 表示偏置项。在本方法中,通过顺序使用MSA和MLP,快捷连接构建每个Transformer模块:

$$\begin{aligned}&\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V} = CONV(LN(X))\\&Y = MSA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) + X\\&Z = MLP(LN(Y) + Y)\end{aligned} \tag{3}$$

其中,CONV表示输入 X 的线性投影,LN 为层归一化。输出 Z 作为后续处理模块的输入,该流程在网络中逐层传递,直至最终模块完成所有计算。

2.2　总体结构

图2(a)为网络总体结构;图2(b)为最小组成单元结构卷积多头自注意力模块(convolution multi-head self-attention modules,CMSA);图2(c)为整体框架的第1阶网络结构时空卷积(convolutional spatio-temporal, CST);图2(d)整个框架的第2阶网络结构卷积多阶Transformer(convolutional multi-stage transformer,CMST)。

1)关节点嵌入

在DSTCFormer模型中,为了有效地编码人体骨骼的结构信息并增强模型对局部关节动态的感知能力,本文提出了一种新的位置嵌入机制。不同于传统的绝对位置编码或简单的线性嵌入,该方法通过局部卷积操作显式地建模了关节在人体骨骼结构中的相对位置关系。

实现方式如下:对于输入的每一帧的关节点特征 $F \in \boldsymbol{R}^{N\times C}$,其中,$N$ 为关节数量,C 为特征维度。首先,为每个关节定义其局部邻域,该局部邻域的定义可基于人体骨骼的自然连接性,例如,对于每个关节,将其直接相连的关节以及自身视为其邻域。因此,对于第 i 个关节,能够提取其邻域内所有关节的特征。

接下来,对每个关节及其邻域内的特征施加一个共享的二维卷积核 $W \in \boldsymbol{R}^{k\times K\times C\times C'}$,其中,$K$ 是卷积核的大小(设置为3×3),C 为输入特征的通道数,C' 为输出特征的通道数。该卷积操作将沿空间维度进行,即在每个关节的局部邻域内做特征的聚合和变换。为了引入非线性,本文在卷积层之后应用GELU激活函数。该方法与传统的绝对位置编码或单纯的线性嵌入相比,具有两点明显优势。

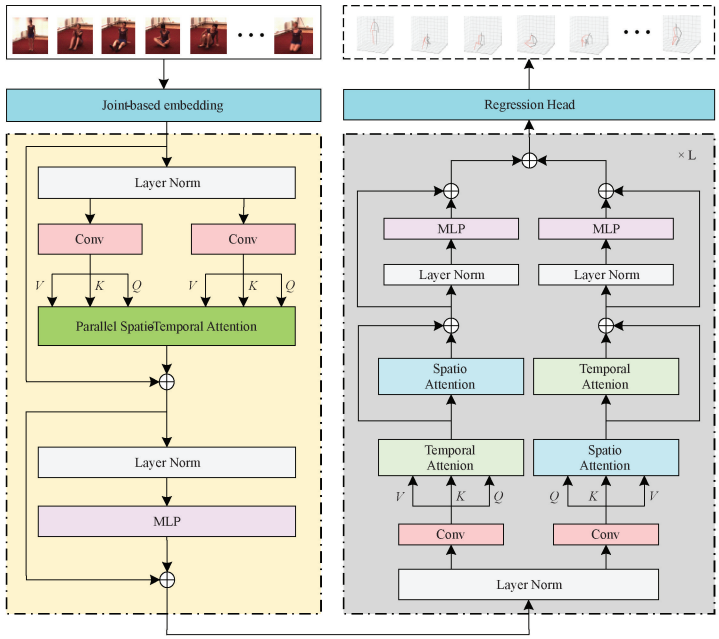
(1)通过卷积局部感受野显式地建模了人体骨骼结构的局部空间相关性,有助于模型更精确地捕捉不同关节点之间的动态关系与运动规律。

(2)采用非线性卷积位置嵌入能够更有效地泛化到未见过的骨骼姿态,大幅提高模型在复杂遮挡与极端动作场景中的鲁棒性。

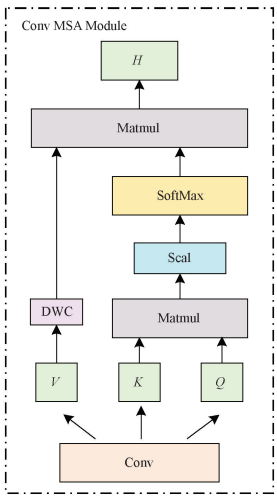
此外,通过消融实验进一步验证了该位置嵌入模块在DSTCFormer整体性能提升中发挥了重要作用,表现出明确的创新性优势。

2)CST模块

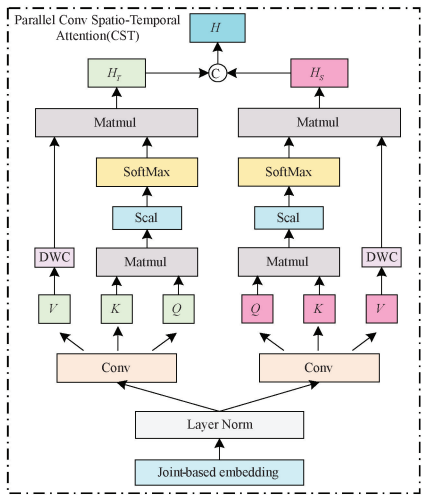
CST模块源自式(3)所示的变换模块,如图2(c)所示,其创新之处在于用卷积时空交叉注意力机制替代了原始的MSA层。此外,CST模块通过部署卷积层以取代传统的全连接层,显著提升了模型对局部空间结构的描



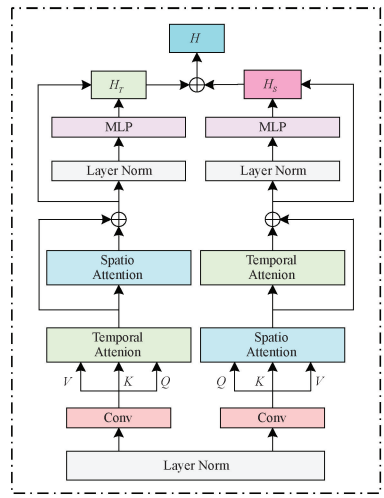
(a) 网络总体结构
(a) Overall network structure



(b) CMSA模块
(b) CMSA module



(c) 第1阶段 CST 模块
(c) Stage 1 CST module



(d) 第2阶段CMST模块
(d) Stage 2 CMST module

图 2　总体结构和各模块

Fig. 2　Overall structure and each module

述能力。

本文对 CMST 模块进行了实验性验证，如图 2(d)所示，并与基准模型进行了对比分析。实验结果，CMST 模块在多项任务中均显著提升了性能，证明了其在处理复杂时空依赖性问题上的高效性。此发现为深度学习领域在时空建模方面的深入研究与探索开辟了新的视角。通过引入创新的时空混合注意力机制，CMST 模块有效增强了模型的时空表征能力，为一系列难题的解决提供了新的方案与策略。

3) 回归头

基于 CMST 模块，本文设计并实现了一种线性回归模型，旨在准确预测目标的三维空间坐标 $P_{3D} \in \boldsymbol{R}_{T\times N\times 3}$。该回归模型作为顶层结构，直接接收 CMST 模块处理后的特征图，进而预测出相应的三维坐标。

为提升模型的预测精度，采用最小化估计的三维位置坐标 P_{3D} 与实际三维坐标 P_{3D} 之间均方误差(mean squared error，MSE)作为损失函数，通过不断减少预测坐标与真实坐标间的误差，对整个网络架构进行优化。在

此过程中,模型参数通过高效的优化策略进行迭代更新,确保了对三维位置坐标的高精度估计。

2.3　并行卷积时空注意力

CST 方法旨在高效模拟关节间的时空依赖性,同时避免完全时空注意力机制带来的高计算成本。参考文献[27,39]的分组上下文化策略,本文提出了一种新的通道处理技术,将通道划分为多个并行组,并对每组执行特征上下文操作,从而在多通道上同步捕获空间与时间上下文信息。与文献[13,27,40]的轴向卷积方法不同,CST 采用轴特异性的多头自注意力机制来提取空间及时间上下文,展现出更强大的相关性学习能力。具体而言,输入嵌入 $\boldsymbol{X} \in \boldsymbol{R}^{T\times N\times C}$ 先被转换成查询 $\boldsymbol{Q} \in \boldsymbol{R}^{T\times N\times C}$、键 $\boldsymbol{K} \in \boldsymbol{R}^{T\times N\times C}$ 和值 $\boldsymbol{V} \in \boldsymbol{R}^{T\times N\times C}$ 序列,并按通道维度平均分为时间组 $\{\boldsymbol{Q}_T,\boldsymbol{K}_T,\boldsymbol{V}_T\}$ 和空间组 $\{\boldsymbol{Q}_S,\boldsymbol{K}_S,\boldsymbol{V}_S\}$ 两部分。在此基础上,分别在两个专门的自注意力模块中,对时间和空间相关性进行计算。

时间相关性涉及在连续帧中单一路径上关节点间的内在联系。为精准捕捉并表达此类联系,本文通过式(1),在时间维度对关节间注意力作有效计算。据此,所提出的时间注意力输出度量可被表述为:

$$\boldsymbol{H}_T = MSA_T(\boldsymbol{Q}_T,\boldsymbol{K}_T,\boldsymbol{V}_T) \tag{4}$$

通过对关节间的时序依赖性进行建模,实现了对运动连续性的深入理解。

空间相关性描述了单帧内关节间的相互联系,即在同一帧中,不同身体部位之间存在固有的内在联系。借鉴于时间注意力机制,本文在空间维度对特定轴向注意力作有效计算。据此,空间注意力的输出可形式化为:

$$\boldsymbol{H}_S = MSA_S(\boldsymbol{Q}_S,\boldsymbol{K}_S,\boldsymbol{V}_S) \tag{5}$$

上述两个相关性模块采取并行处理机制,依据自注意力策略对上下文特征进行深入挖掘。通过专门的轴向视角计算标记间的相互关系,实现时间和空间注意力的互补与融合。在此基础上,将两个注意层输出的特征在通道维度上进行有效串联,以增强表示能力,其公式为:

$$\boldsymbol{H} = cat(\boldsymbol{H}_T,\boldsymbol{H}_S) \tag{6}$$

其中,连接由 *cat* 函数完成。由此产生的 CST 系列感受野类似于空间和时间轴的十字交叉模式。通过堆叠多个 CST 块,可以近似实现完整的时空注意力。

2.4　并行卷积混合时空注意力

CMST 模块旨在高效模拟关节间的时空依赖关系,规避了完全时空注意力机制所导致的额外计算成本。本文提出的基于帧的上下文策略[40]将通道划分为多个并行组,通过在各组上施加特征上下文操作,实现了在多个通道中同步捕获空间与时间信息的能力。在 CMST 框架下,采用针对特定轴向的多头自注意力机制,以并行方式学习时空上下文,在相关性学习方面展现出更卓越的效能。首先,本文将输入嵌入 $\boldsymbol{X} \in \boldsymbol{R}^{T\times N\times C}$ 映射至查询 $\boldsymbol{Q} \in \boldsymbol{R}^{T\times N\times C}$、键值 $\boldsymbol{K} \in \boldsymbol{R}^{T\times N\times C}$ 和数值 $\boldsymbol{V} \in \boldsymbol{R}^{T\times N\times C}$,进而沿着通道维度平均分配为两组。为便于阐述,将特征矩阵区分为时间组 $\{\boldsymbol{Q}_T,\boldsymbol{K}_T,\boldsymbol{V}_T\}$ 和空间组 $\{\boldsymbol{Q}_S,\boldsymbol{K}_S,\boldsymbol{V}_S\}$。随后,分别在两个自注意力模块中,针对时间和空间相关性进行计算。

通过在分组输入通道中分别采用自注意力机制,CMST 模块有效实现了对时空交叉信息的同时捕获,从而显著提升了模型的处理效能。此方法不仅增强了模型处理大规模数据及建模时空关联的能力,而且借助多头自注意力机制的引入,进一步提升了模型对多样化数据特征的抽取及其表达力。总体而言,CMST 模块为处理复杂的时空依赖问题提供了一种创新性方案,具备卓越的计算效率与性能表现。

空时相关性(tempo-spatial,T-S)的关节点在不同帧间的一致性路径联系得到有效体现。通过应用多头自注意力模型,如式(1)所示,计算了关节点随时间变化的注意力亲和度,进而形成了精确的时间注意力度量,如式(7)所示。

$$\boldsymbol{H}_{TS} = MSA_{TS}(\boldsymbol{Q}_{TS},\boldsymbol{K}_{TS},\boldsymbol{V}_{TS}) \tag{7}$$

通过 MSA_{TS} 能够精确捕捉关节间交互作用的时效性特征。不同的注意力头分别聚焦于输入序列的各异区域,有效识别并提取各自独特的时间依赖性。

综合各注意力头的输出,本文构建了一个全面的时间相关性图谱,显著提高了动态场景下的时间相关性建模效率,在处理时间序列分析的各类任务中起到关键作用。此外,本文所采用的自注意力机制因其能明确揭示特定时间点关节间的相互关系,使其具备了卓越的可解释性。不仅有助于深入洞察模型的决策逻辑,也为进一步增强模型性能奠定了基础。

时空相关性(spatio-temporal, S-T)涉及单帧内关节点间的互联性,揭示了基于人体骨骼结构本质的帧内部位间内在联系。在此基础上,借鉴时间注意力机制,本文创新性地提出了一种新型结构多头空间-时间自注意力模块(multi-head spatio-temporal attention, MSA_{ST}),作为专用于时空维度的轴向特定 MSA,其公式为:

$$\boldsymbol{H}_{ST} = MSA_{ST}(\boldsymbol{Q}_{ST},\boldsymbol{K}_{ST},\boldsymbol{V}_{ST}) \tag{8}$$

通过 MSA_{ST} 能够捕获单个帧中关节之间的空间相关性,构建了时空拓扑关联网络,有效提取单帧图像中关节节点间的耦合特征,为三维人体姿态估计提供了可量化的表征基础。

CMST 模块通过融合时间与空间注意力机制,有效构建了关节在时序与空间维度上的相互依赖关系。该建模方法使得本文能够更全面地理解动态场景中的人体运动,不仅显著提升了模型的表现力,而且丰富了身体动态解析的信息维度。

两个相关模块采用自注意力机制,并行协同处理上下文特征,通过在不同轴向上的上下文计算,揭示了特征间的亲和性,以实现相互补充。本文在通道维度上将两个注意力层作输出合并,获得全面的特征表示:

$$\boldsymbol{H} = cat(\boldsymbol{H}_{ST}, \boldsymbol{H}_{ST}) \tag{9}$$

在进行拼接操作时,结果作为一个具有时空轴交错模式的感受野,存在于交叉时空序列中,通过不同 CMST 模块的堆叠允许逼近完整的时空注意力。

3　实验

本文分别在 Human3.6M[41] 和 MPI-INF-3DHP[42] 两个广泛认可的数据集上进行测试评估,检验了 DSTCFormer 在不同动作捕捉场景下的应用效果,验证了其性能的优越性。

3.1　数据集和评估指标

Human3.6M 数据是目前最广泛使用的室内三维人体姿态估计基准数据集,包含了 11 名受试者执行的 15 种典型动作,总共生成了 360 万帧视频。本文根据标准协议,采用 *S*1、*S*5、*S*6、*S*7 和 *S*8 子集进行训练,使用 *S*9 和 *S*11 子集进行评估测试。同时,基于两种协议[13,25]测量了平均关节位置误差(mean per joint position error, MPJPE):协议 1(P1)通过髋关节对齐,计算估计姿态与真实值之间的 MPJPE(mm);协议 2(P2)则通过刚性变换来计算真实值与估计姿态之间的 P-MPJPE。

MPI-INF-3DHP 数据也是新近推出的常用数据集,包括绿幕、非绿幕和户外等 3 种不同场景。该数据集由 14 个摄像机记录的 8 名演员的活动视频构成,训练集包含 8 种活动,测试集包含另外 7 种活动。依据先前研究[29-31],所使用的评估指标包括 MPJPE、150 mm 标准的关键点正确率(percentage of correct keypoints, PCK)以及曲线下面积(area under the curve, AUC)。

3.2　实验细节

本文在搭载 RTX 4090Ti GPU 的服务器上,采用 PyTorch 模型进行实验。实验输入包括两个二维姿态序列:1)用级联金字塔网络(cascaded pyramid network, CPN)预训练模型预测得到的二维姿态;2)真实的二维姿态(Ground-truth 数据)。模型训练过程采用最小 batch size 为 128,共执行 20 个训练周期。网络参数的优化通过 Adam 优化器实现,学习率设定为 0.001。在模型中,注意力块 L 的叠加次数、隐藏嵌入通道 C 的维度以及头部 H 的个数均被视为自由调整的参数。

3.3　Human3.6M 数据集上的性能分析

本文在 Human3.6M 数据集上与多种前沿方法进行了性能对比,表 1 为在以 CPN 估计的二维姿态作为输入,在 P1 与 P2 协议下,不同采样帧数 T 对应的误差性能指标。表 2 为在采用真实二维姿态作为输入的条件下,DSTCFormer 与现有顶尖模型的性能对比。该实验设计旨在消除二维姿态估计中潜在的噪声影响,并确立二维姿态向三维姿态转换的理论性能上限。每列中的最优结果标为加粗,次优结果标为下划线,以示区分。

观察表 1 和 2 可以发现,DSTCFormer 在两种输入设定下均取得了优异的性能,充分证明了其鲁棒性与有效性。

从表 1 可以看出,基于 CPN 估计的二维姿态输入下,DSTCFormer 在 P1 协议下实现了平均 MPJPE 为 40.1 mm,相较于时空分解方法(如 PoseFormer,其平均误差为 44.3 mm)有明显改善。其充分表明,所提出的双阶段并行处理策略能更充分地利用时空信息进行特征交互。在更严格的 P2 协议下,DSTCFormer 同样取得 31.6 mm 的 P-MPJPE,优于 PoseFormer 的 34.5 mm。进一步,与采用单阶段时空交叉策略的方法(如 STCFormer 和 MixSTE)对比,DSTCFormer 在 P1 协议下分别达到了 40.1 mm,对比 STCFormer 的 40.5 mm 和 MixSTE 的 40.9 mm,展现出略优的性能;在 P2 协议下,其 31.6 mm 同样优于 STCFormer 的 31.8 mm 和 MixSTE 的 32.6 mm。上述对比数据证明了卷积强化的双阶段时空交互机制在提高精度的同时,具备更强的泛化能力,且不易出现过拟合现象。

当以真实的二维姿态作为输入时(表 2),所有参照模型的 P1 协议误差均展现出明显的下降;同时,性能的变化趋势显示出较高的一致性。在此研究中,DSTCFormer 模型的 P1 协议平均 MPJPE 为 21.0 mm,相较于 STCFormer 的 21.3 mm 和 MixSTE 的 21.5 mm,分别取得了 0.3 和 0.5 mm 的性能提升。上述结果,进一步验证了 DSTCFormer 在不同输入质量下的鲁棒性和有效性,并表明即使在理想的二维姿态输入下,本文的模型也能更好地进行三维姿态的推断。

此外,所引入的 CMST 模块在整个网络中起到了关键作用。实验结果表明,在 P1 协议的 15 个测试类别中,DSTCFormer 在 11 个类别中取得了最低误差,在其他 4 个类别中为次低误差,但是与 PoseFormer, MixSTE 和 STCFormer 等竞争模型相比,平均误差分别减少 4.1、0.8 和 0.4 mm;在 SitD 遮挡数据集上,虽然性能略逊于 MixSTE,但相对于 STCFormer 仍有 0.3 mm 的改进。在 P2 协议下,DSTCFormer 在 15 个测试类别中有 12 个类别表现最佳、2 个类别次佳,相较于 STCFormer,平均误差也提升了 0.2 mm。以上数据充分证明了 DSTCFormer 在时空信息建模与特征融合方面的卓越性能,以及其在复杂场景下的鲁棒性和泛化能力。

无论是在存在噪声的 CPN 估计输入,还是在理想的

真实二维姿态输入条件下,DSTCFormer 均展现出对比现有时空解耦和交叉方法更为显著的精度优势,进一步验证了本文提出的时空解耦架构、卷积-注意力交叉融合机制以及人体结构驱动的位置编码技术的有效性。

3.4　MPI-INF-3DHP 数据集上的性能分析

为了验证三维姿态估计模型的效果,本文深入评估了模型在包含复杂背景的 MPI-INF-3DHP 数据集上的性能优势。与先前研究[29-31]保持一致,本文采用相同的测试标准,并使用真实的二维姿态作为输入。鉴于视频序列的长度限制,本文设定了9、27 或 81 帧作为输入。表3为 DSTCFormer 与其他模型的误差性能比较。与基于 Human3.6 M 的评估方法相似,在 P1 协议下,当帧数 T=81 时,DSTCFormer 的 PCK 值达到 99.1%,AUC 值为 85.2%,P1 误差为 22.3 mm。与 STCFormer 相比,分别提升了 0.4%、1.3%和降低了 0.8 mm,表现出显著的改进。特别地,DSTCFormer 展现了更强的泛化能力。在 Human3.6 M 数据集上,根据 P1 协议,相较于 MixSTE,其性能大幅提升了 32.6 mm,显著证明了本文模型在处理复杂数据时的有效性。

表 1　基于 Human3.6M 数据集的不同方法 CPN 二维姿态输入 P1 与 P2 协议的评估对比

Table 1　Evaluation of different methods of CPN 2D pose input P1 & P2 protocols based on Human3.6M dataset

P1	Dir.	Dis.	Eat.	Gre.	Phone	Photo	Pose	Purch.	Sit.	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
文献[43](T=243)	40.8	44.5	41.4	42.7	46.3	55.6	41.8	41.9	53.7	60.8	45.0	41.5	44.8	30.8	31.9	44.2
PoseFormer[25](T=81)	41.5	44.8	39.8	42.5	46.5	51.6	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.2
CrossFormer[28](T=81)	40.7	44.1	40.8	41.5	45.8	52.8	41.2	40.8	55.3	61.9	44.9	41.8	44.6	29.2	31.1	43.8
MHFormer[29](T=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	42.9
P-STMO[26](T=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
MixSTE[30](T=81)	39.8	43.0	38.6	40.1	43.4	50.6	40.6	41.4	52.2	56.7	43.8	40.8	43.9	29.4	30.3	42.3
MixSTE[30](T=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
STCFomer[31](T=81)	40.6	43.0	38.3	40.2	43.5	52.6	40.3	40.1	51.8	57.7	42.8	39.8	42.3	28.0	29.5	42.0
STCFormer[31](T=243)	38.4	41.2	36.8	38.0	42.7	50.5	38.7	38.2	52.5	56.8	41.8	38.4	40.2	**26.2**	27.7	40.5
DSTCFomer(T=27)	42.1	43.8	41.6	42.0	45.1	53.2	41.1	41.6	52.9	59.4	43.9	40.9	44.3	29.9	30.8	43.5
DSTCFomer(T=81)	41.7	42.3	38.3	40.4	43.5	50.6	40.7	39.2	52.1	56.8	42.9	39.5	41.9	28.2	29.3	41.8
DSTCFomer(T=243)	38.1	**40.4**	**36.5**	**37.5**	**42.1**	50.4	**38.4**	**38.0**	**51.4**	55.7	**41.1**	**38.3**	**40.0**	26.7	**27.1**	**40.1**
P2	Dir.	Dis.	Eat.	Gre.	Phone	Photo	Pose	Purch.	Sit.	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
文献[43](T=243)	32.5	36.2	33.2	35.3	35.6	42.1	32.6	31.9	42.6	47.9	36.6	32.1	34.8	24.2	25.8	34.9
PoseFormer[25](T=81)	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	34.8	24.2	25.8	36.1
CrossFormer[28](T=81)	31.4	34.6	32.6	33.7	34.3	39.7	31.6	31.0	44.3	49.3	35.9	31.3	34.4	23.4	25.5	34.2
MHFormer[29](T=351)	31.5	34.9	32.8	33.6	35.3	39.6	32.0	32.2	43.5	48.7	36.4	32.6	34.3	23.9	25.1	34.4
P-STMO[26](T=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
MixSTE[30](T=81)	32.0	34.2	31.7	33.7	34.4	39.2	32.0	31.8	42.9	46.9	35.5	32.0	34.4	23.6	25.2	34.0
MixSTE[30](T=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
STCFormer[31](T=81)	30.4	33.8	31.1	31.7	33.5	39.5	30.8	30.0	41.8	45.8	34.3	30.1	32.8	21.9	23.4	32.7
STCFormer[31](T=243)	29.3	33.0	30.7	30.6	32.7	38.2	29.7	28.8	42.2	45.0	33.3	29.4	31.5	**20.9**	22.3	31.8
DSTCFomer(T=27)	31.1	34.2	32.9	32.8	34.2	39.8	31.4	31.0	42.3	46.6	35.1	30.7	33.5	22.9	24.0	33.5
DSTCFomer(T=81)	30.0	33.3	31.1	30.8	32.9	38.0	30.4	29.6	41.8	44.4	33.9	29.9	44.4	32.1	22.7	33.7
DSTCFomer(T=243)	**29.0**	**33.0**	30.5	**30.1**	**32.3**	**37.7**	**29.5**	**28.5**	**41.6**	45.2	**33.1**	**29.1**	**31.3**	20.9	**22.1**	**31.6**

表 2　基于 Human3.6M 数据集的不同方法真实二维姿态输入 P1 协议的评估对比

Table 2　Evaluation of different methods of real 2D pose input P1 protocols based on Human3.6M dataset

P1	Dir.	Dis.	Eat.	Gre.	Phone	Photo	Pose	Purch.	Sit.	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
文献[37](T=243)	34.5	37.1	33.6	34.2	32.9	37.1	39.6	35.8	40.7	41.4	33.0	33.8	33.0	26.6	26.9	34.7
PoseFormer[25](T=81)	30.0	33.6	29.9	31.0	30.2	33.3	34.8	31.4	37.8	38.6	31.7	31.5	29.0	23.3	23.1	31.3
文献[43](T=243)	29.5	30.8	28.8	29.1	30.7	35.2	31.7	27.8	34.5	36.0	30.3	29.4	28.9	24.1	24.7	30.1
MHFormer[29](T=351)	27.7	32.1	29.1	28.9	30.0	33.9	33.0	31.2	37.0	39.3	30.0	31.0	29.4	22.2	23.0	30.5
P-STMO[26](T=243)	28.5	30.1	28.6	27.9	29.8	33.2	31.3	27.8	36.0	37.4	29.7	29.5	28.1	21.0	21.0	29.3
StridedFormer[27](T=243)	27.1	29.4	26.5	27.1	28.6	33.0	30.7	26.8	38.2	34.7	29.1	29.8	26.8	19.1	19.8	28.4
CrossFormer[28](T=81)	26.0	30.0	26.8	26.2	28.0	31.0	30.4	29.6	35.4	37.1	28.4	27.3	26.7	20.5	19.9	28.2
MixSTE[30](T=81)	25.6	27.8	24.5	25.7	24.9	29.9	28.6	27.4	29.9	29.0	26.1	25.0	25.2	18.7	19.9	25.9
MixSTE[30](T=243)	21.6	22.0	20.4	21.0	**20.8**	**24.3**	24.7	21.9	**26.9**	**24.9**	**21.2**	21.5	20.8	14.7	15.7	21.5
STCFormer[31](T=81)	25.9	25.9	22.7	24.0	24.6	27.5	27.6	23.1	30.1	31.5	25.1	24.7	23.8	18.4	19.6	25.0
STCFormer[31](T=243)	20.8	21.8	**20.0**	20.6	23.4	25.0	23.6	**19.3**	27.8	26.1	21.6	20.6	19.5	14.3	15.1	21.3
DSTCFormer (T=81)	25.5	25.4	22.4	24.1	24.2	27.3	27.0	23.3	30.0	31.4	24.3	24.2	23.7	18.5	19.3	24.7
DSTCFormer (T=243)	**20.5**	**21.4**	20.1	**20.3**	23.3	25.1	**23.1**	19.4	27.2	25.4	21.4	**20.2**	**19.1**	**14.2**	**14.9**	**21.0**

表 3 在 MPI-INF-3DHP 数据集上,基于 PCK、AUC 和 P1 三个标准,与其他模型进行性能对比。在 PCK 值和 AUC 值越高的同时,P1 值越低,则表示该模型具有更好的回归效果。每列中的最佳性能标为加粗。

表 3　基于 MPI-INF-3DHP 数据集的不同方法性能对比

Table 3　Performance comparison of different methods based on MPI-INF-3DHP dataset

方法	Publication	PCK ↑	AUC ↑	P1/mm ↓
UGCN[44] (T=96)	ECCV'20	86.9	62.1	68.1
Anatomy3D[36] (T=81)	TCSVT'21	87.8	53.8	79.1
PoseFormer[25] (T=9)	ICCV'21	88.6	56.4	77.1
文献[45] (T=96)	ACMMM'21	97.9	69.5	42.5
CrossFormer[28] (T=9)	arXiv'22	89.1	57.5	76.3
PATA[38] (T=243)	TIP'22	90.3	57.8	69.4
MHFormer[29] (T=9)	CVPR'22	93.8	63.3	58.0
MixSTE[30] (T=27)	CVPR'22	94.4	66.5	54.9
文献[46] (T=81)	arXiv'22	95.4	67.6	46.9
P-STMO[26] (T=81)	ECCV'22	97.9	75.8	32.2
STCFormer[31] (T=9)	CVPR'23	98.2	81.5	28.2
STCFormer[31] (T=27)	CVPR'23	98.4	83.4	24.2
STCFormer[31] (T=81)	CVPR'23	98.7	83.9	23.1
DSTCFormer (T=9)		98.4	82.1	27.5
DSTCFormer (T=27)		98.5	84.3	23.4
DSTCFormer (T=81)		99.1	85.2	22.3

3.5　消融实验

为了深入剖析本文提出的 DSTCFormer 模型,本文基于 Human3.6M 数据集进行了严谨的实验研究。在此过程中,本文以基于 CPN 估计的二维姿态作为模型输入,开展了一系列消融实验,以量化各模块对模型性能的具体贡献。

第 1 组实验设置旨在探究 DSTCFormer 在不同输入帧数 T 条件下的表现。表 4 为在 P1 与 P2 两种协议下的详细性能对比,每列中的最佳性能标为加粗。观察发现,输入帧数增加时,模型性能普遍呈单调上升态势。与主流对比方法 STCFormer 相较,DSTCFormer 在 27、81 及 243 帧输入设置下,均展现出更卓越的性能,验证了本模型在处理不同长度视频序列时的卓越适用性。尽管在计算参数上,本模型与对比算法相比并不占优,但在 P1 协议下,以 27、81 和 243 帧为输入时,性能分别提升了 0.82、0.51 与 0.43 mm;同理,在 P2 协议下,输入帧数同样设置为 27、81 和 243 时,性能分别增进 1.29、1.15 与 0.8 mm,表明 DSTCFormer 在精确度上的显著优势。

为了进一步深入评估 DSTCFormer 模块对模型性能的贡献,本文设计了第 2 组消融实验。实验输入采用 CPN 估计的二维姿态,固定输入序列长度为 27 帧,在 Human3.6M 数据集下进行验证,如表 5 所示,最佳性能标为加粗。

表 4　基于 Human3.6M 数据集的 P1 与 P2 协议下不同采样帧数的评估对比

Table 4　Evaluation comparison of different frames in P1 & P2 protocols based on Human3.6M dataset

方法	T/Frames	Parameters/($\times 10^6$)	M/FLOPs	P1	P2
STCFormer[31]	27	4.75	2 173	44.4	34.8
DSTCformer	27	8.046	2 322	**43.5**	**33.5**
STCFormer[31]	81	4.75	6 520	42.3	33.3
DSTCformer	81	8.046	11 300	**41.8**	**32.2**
STCFormer[31]	243	4.75	19 561	40.5	31.8
DSTCformer	243	8.046	35 710	**40.07**	**31.0**

表 5　基于 Human3.6 数据集的模块消融结果

Table 5　Module ablation results based on the Human3.6 dataset

	Frames	Conv	S-T	T-S	P1	P2
STCFormer[31]	27				44.4	34.8
DCFormer	27	√			44.3	34.2
DSCFormer	27	√	√		44.1	34.1
DTCFormer	27	√		√	44.4	34.3
DSTCFormer	27	√	√	√	**43.5**	**33.5**

本文首先考察了 CMSA 模块的作用。DCFormer 模型为基线模型,仅采用了替换后的 CMSA 模块,用于替代原始 Transformer 中的标准多头自注意力模块。从实验结果可以看出,与 STCFormer 相比,仅采用 CMSA 模块后,性能指标 P1 和 P2 分别提高了 0.07 和 0.55 mm。表明本文设计的 CMSA 模块通过其协同的多尺度特征融合机制,能够更有效地捕捉人体关节之间的复杂关系,从而带来性能的提升。

接下来,分别评估了双阶段时空模块中空间路径和时间路径的贡献。DSCFormer 模型代表仅包含空间路径的双阶段时空模块,其专注于在空间维度上进行特征提取和交互。实验结果显示,在引入 S-T 单路径双阶段时空模块后,P1 和 P2 分别显著提升了 0.22 和 0.71 mm。实验结果表明,通过双阶段的结构在空间维度上进行更深层次的特征学习,能够有效地提升模型的性能。DTCFormer 模型则代表仅包含时间路径的双阶段时空模块,其专注于在时间维度上建模关节的运动轨迹。实验结果表明,采用 T-S 单路径双阶段时空模块,尽管 P1 的性能保持稳定,P2 却进一步提升了 0.51 mm。上述结果证明,在时间维度上进行双阶段的建模,对于提升模型在对齐后的姿态预测精度(P2 协议)方面具有积极作用。

最后,本文测试了同时集成 S-T 路径和 T-S 路径双阶段时空模块的效果。当同时集成这两个互补的路径时,模型能够同时从空间和时间两个维度上进行更全面的特征学习和交互。实验结果显示,P1 和 P2 的性能分别达到了 0.82 和 1.29 mm 的显著增幅。以上结果证明

了本文提出的双路径双阶段时空模块的有效性,表明空间和时间信息在模型中进行协同处理能够带来显著的性能提升。

综上所述,消融实验充分验证了本文提出的各关键模块(CMSA、S-T 和 T-S 双阶段模块及其双路径融合)在模型性能提升中的作用,为 DSTCFormer 的整体设计提供了有力支撑。

3.6　定性分析

通过可视化三维人体姿态估计结果,对 DSTCFormer 模型的性能进行验证。所采用的样本随机抽取自 Human3.6M 数据集的评估子集,如图 3 所示。将所提 DSTCFormer 模型与 PoseFormer 及 STCFormer 两种典型时空解耦 Transformer 方法在复杂动作及关节遮挡场景下进行直观比较如图 4 所示。可清晰看出,PoseFormer 模型由于空间与时间信息缺乏有效的交互机制,在关节遮挡明显或动作剧烈时,容易产生明显的误差;而 STCFormer 虽然通过交叉注意力增强了时空交互,但在处理快速动作变化与局部细节时,依然存在较大误差;相比之下,本文提出的 DSTCFormer 模型能够更精细地捕捉到各关节局部细节变化,体现出明显的精度提升与鲁棒性优势,这主要归功于其创新性的双阶段时空交叉卷积注意力机制和高效位置嵌入模块。综合来看,DSTCFormer 模型在处理单目视频时,展现出快速且有效的三维姿态估计能力。

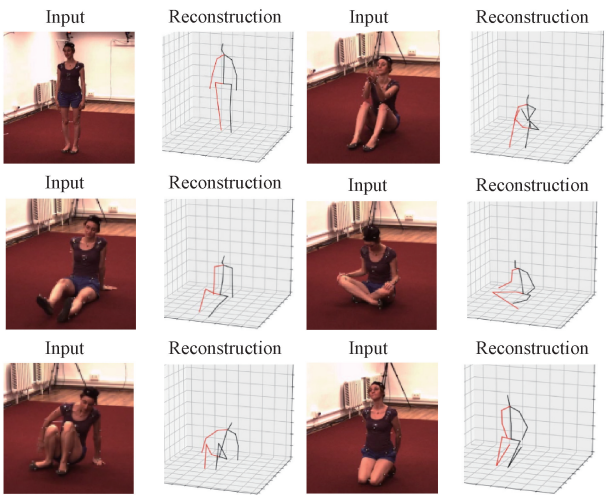


图 3　Human3.6M 数据集可视化结果

Fig. 3　Visualization results of the Human3.6M dataset

4　结　论

本文提出了一种 DSTCFormer 三维人体姿态估计方法。该方法创新性地运用双阶段时空注意力交叉机制,深入挖掘视频序列中三维人体姿态的空间和时间关联性。DSTCFormer 设计了多个 CMST 模块,这些模块能够沿通道维度将关节点特征划分为两个子集,分别对其进行独立的空间与时间交互模拟,该设计使得 CMST 模块在感受野上展现出一种独特的时空交错模式。经过 Human3.6M 和 MPI-INF-3DHP 两个主流基准测试的验证,DSTCFormer 展现出显著的有效性,并且相较于现有顶尖技术,展现出更卓越的泛化能力。

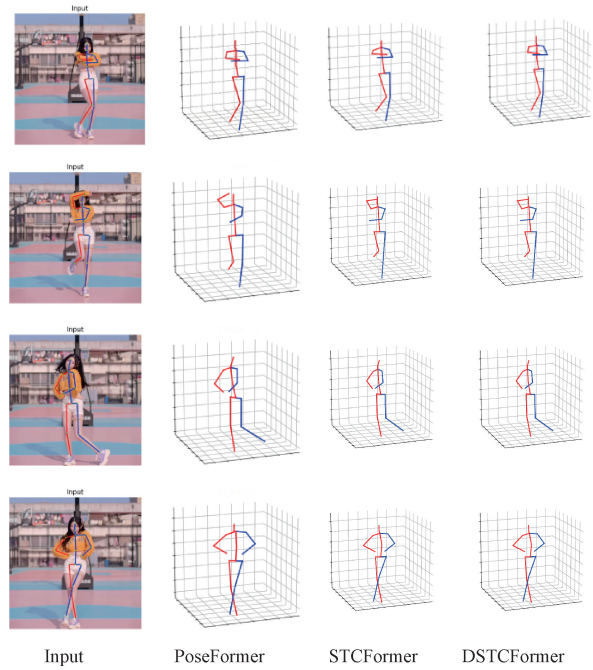


图 4　DSTCFormer 算法与 PoseFormer、STCFormer 算法定性可视化比较

Fig. 4　Qualitative visualization comparison between DSTCFormer and the PoseFormer, STCFormer

参考文献

[1] LIU M, YUAN J. Recognizing human actions as the evolution of pose estimation maps[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 1159-1168.

[2] LIU H, CHEN Y, ZHAO W, et al. Human pose recognition via adaptive distribution encoding for action perception in the self-regulated learning process[J]. Infrared Physics & Technology, 2021, 114: 103660.

[3] LIU L, YANG L, CHEN W, et al. Dual-view 3D human pose estimation without camera parameters for action recognition[J]. IET Image Processing, 2021, 15(14): 3433-3440.

[4] MEHTA D, SRIDHAR S, SOTNYCHENKO O, et al. Vnect: Real-time 3d human pose estimation with a single rgb camera[J]. ACM Transactions on Graphics, 2017,

36(4): 1-14.

[5] KISACANIN B, PAVLOVIC V, HUANG T S. Real-Time Vision for Human-Computer Interaction [M]. Dordrecht: Springer Science & Business Media, 2005.

[6] 周佳裕, 蔡晋辉, 章乐, 等. 接触式交互感知的人体三维坐姿姿态估计[J]. 仪器仪表学报, 2022, 43(11): 132-141.
ZHOU J Y, CAI J H, ZHANG L, et al. Human 3D sitting pose estimation based on contact interaction perception[J]. Chinese Journal of Scientific Instrument, 2022, 43(11): 132-141.

[7] PAVLAKOS G, ZHOU X, DERPANIS K G, et al. Coarse-to-fine volumetric prediction for single-image 3D human pose[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7025-7034.

[8] 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.

[9] 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.

[10] WANG C, WANG Y, LIN Z, et al. Robust estimation of 3d human poses from a single image[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 2361-2368.

[11] YAO W, ZHANG H, SUN Y, et al. STAF: 3D human mesh recovery from video with Spatio-temporal alignment fusion[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(11): 10564-10577.

[12] 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.

[13] 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.

[14] FANG H S, XU Y, WANG W, et al. Learning pose grammar to encode human body configuration for 3d pose estimation[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018: 6821-6828.

[15] CHEN C H, RAMANAN D. 3d human pose estimation = 2d pose estimation + matching[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7035-7043.

[16] CAI Y, GE L, LIU J, et al. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 2272-2281.

[17] 李一凡, 袁龙健, 王瑞. 基于 OpenPose 改进的轻量化人体动作识别模型[J]. 电子测量技术, 2022, 45(1): 89-95.
LI Y F, YUAN L J, WANG R. Improved lightweight human action recognition model based on OpenPose[J]. Electronic Measurement Technology, 2022, 45(1): 89-95.

[18] ZOU Y, PAN Z, ZHOU X, et al. Human pose evaluation based on full-domain convolution and LSTM[J]. Applied Mathematics and Nonlinear Sciences, 2023, 9(1): 1-16.

[19] 张新峰, 范铭, 曹哲宇, 等. 基于几何统计的人体姿态语义描述方法[J]. 电子测量与仪器学报, 2023, 37(8): 52-59.
ZHANG X F, FAN M, CAO ZH Y, et al. Human posture semantic description method based on geometric statistics[J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(8): 52-59.

[20] WANG M, LAI B, HUANG J, et al. Camera-aware proxies for unsupervised person re-identification [C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 2764-2772.

[21] ZHAO W, TIAN Y, YE Q, et al. Graformer: Graph convolution transformer for 3d pose estimation[J]. ArXiv preprint arXiv:2109.08364, 2021.

[22] JIANG B, HU L, XIA S. Probabilistic triangulation for uncalibrated multi-view 3d human pose estimation[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 14850-14860.

[23] LIU H, XIANG W, HE J Y, et al. Refined temporal pyramidal compression-and-amplification transformer for 3D human pose estimation[J]. ArXiv preprint arXiv: 2309.01365, 2023.

[24] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6000-6010.

[25] ZHENG C, ZHU S, MENDIETA M, et al. 3d human pose estimation with spatial and temporal transformers [C]. Proceedings of the IEEE/CVF International Conference

on Computer Vision, 2021: 11656-11665.

[26] SHAN W, LIU Z, ZHANG X, et al. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation[C]. European Conference on Computer Vision, 2022: 461-478.

[27] LI W, LIU H, DING R, et al. Exploiting temporal contexts with strided transformer for 3d human pose estimation[J]. IEEE Transactions on Multimedia, 2022, 25: 1282-1293.

[28] HASSANIN M, KHAMISS A, BENNAMOUN M, et al. Crossformer: Cross spatio-temporal transformer for 3d human pose estimation [J]. ArXiv preprint arXiv: 2203. 13387, 2022.

[29] LI W, LIU H, TANG H, et al. Mhformer: Multi-hypothesis transformer for 3d human pose estimation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 18-24 June, 2022: 13147-13156.

[30] ZHANG J, TU Z, YANG J, et al. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 18-24 June, 2022: 13232-13242.

[31] TANG Z, QIU Z, HAO Y, et al. 3D human pose estimation with spatio-temporal criss-cross attention[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 17-24 June, 2023: 4790-4799.

[32] DABHI M, JENI L A, LUCEY S. 3D-LFM: Lifting Foundation Model[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16-22 June, 2024: 10466-10475.

[33] 李功浩, 贾振堂. 融合 Transformer 和语义图卷积的三维人体姿态估计方法[J]. 国外电子测量技术, 2024, 43(3): 10-17.
LI G H, JIA ZH T. 3D human pose estimation method fusing Transformer and semantic graph convolution[J]. Foreign Electronic Measurement Technology, 2024, 43(3): 10-17.

[34] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]. European Conference on Computer Vision, Glasgow, UK, 23-28 August, 2020: 213-229.

[35] ZENG AI L, SUN X, HUANG F Y, et al. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach[C]. Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, 23-28 August, 2020: 507-523.

[36] CHEN T, FANG C, SHEN X, et al. Anatomy-aware 3d human pose estimation with bone-based pose decomposition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(1): 198-209.

[37] LIU R, SHEN J, WANG H, et al. Attention mechanism exploits temporal contexts: real-time 3D human pose reconstruction[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 5064-5073.

[38] XUE Y, CHEN J, GU X, et al. Boosting monocular 3D human pose estimation with part aware attention[J]. IEEE Transactions on Image Processing, 2022, 31: 4278-4291.

[39] LI S, CHAN A B. 3d human pose estimation from monocular images with deep convolutional neural network[C]. Computer Vision-ACCV 2014: 12th Asian Conference on Computer Vision, 2015: 332-347.

[40] MA X, SU J, WANG C, et al. Context modeling in 3d human pose estimation: A unified perspective[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 6238-6247.

[41] 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.

[42] MEHTA D, RHODIN H, CASAS D, et al. Monocular 3d human pose estimation in the wild using improved CNN supervision[C]. 2017 International Conference on 3D Vision (3DV), 2017: 506-516.

[43] SHAN W, LU H, WANG S, et al. Improving robustness and accuracy via relative information encoding in 3D human pose estimation[C]. Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 2021: 3446-3454.

[44] WANG J, YAN S, XIONG Y, et al. Motion guided 3D pose estimation from videos[C]. Proceedings of the EuropeanS Conference on Computer Vision, 2020: 764-780.

[45] HU W, ZHANG C, ZHAN F, et al. Conditional directed graph convolution for 3D human pose estimation[C]. Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 2021: 602-611.

[46] EINFALT M, LUDWIG K, LIENHART R. Uplift and upsample: Efficient 3D human pose estimation with uplifting transformers[C]. Proceedings of the IEEE/CVF

Winter Conference on Applications of Computer Vision, 2023: 2903-2913.

作者简介

邹宇,2016 年于南京信息工程大学滨江学院获得学士学位,2019 年于南京信息工程大学获得硕士学位,现为南京信息工程大学博士研究生,江苏开放大学助理研究员,主要研究方向为计算机视觉、图像处理和深度学习。

Zou Yu received his B. Sc. degree from Nanjing University of Information Science and Technology Binjiang College in 2016, M. Sc. degree from Nanjing University of Information Science and Technology in 2019. He is now a Ph. D. candidate at Nanjing University of Information Science and Technology, and a research assistant at Jiangsu Open University. His main research interests include computer vision, image processing and deep learning.

周先春(通信作者),2011 年于南京信息工程大学获得博士学位,现为南京信息工程大学硕士生导师,教授,中国电子学会高级会员,主要研究方向为图像处理、深度学习和信号与信息处理。

E-mail:zhouxc2008@ 163. com

Zhou Xianchun (Corresponding author), received Ph. D. from Nanjing University of Information Science and Technology in 2011. Now he is a professor and M. Sc. supervisor at Nanjing University of Information Science and Technology, also a senior member of China Electronics Society. His main research interests include signal and information processing.

潘志庚,1993 年于浙江大学获得博士学位,现为南京信息工程大学人工智能学院院长,博士生导师,教授,中国虚拟现实技术与创新平台副理事长和中国移通联元宇宙产业委员会常务副主任。主要研究方向为虚拟现实、人机交互、元宇宙和智慧教育。

Pan Zhigeng, received Ph. D. from Zhejiang University in 1993. Dean of the School of Artificial Intelligence at Nanjing University of Information Science and Technology, professor and Ph. D. supervisor, vice chairman of the China Virtual Reality Technology and Innovation Platform, and executive deputy director of the Metaverse Industry Committee under the China Mobile Communications Association. His main research interests include virtual reality, human-computer interaction, metaverse and smart education.

蔡创新,2014 年于南京晓庄学院获得学士学位,2018 年于淮阴工学院获得硕士学位,现为南京信息工程大学博士研究生,主要研究方向为计算机视觉、情感分析和多模态情绪识别。

Cai Chuangxin received his B. Sc. degree from Nanjing Xiaozhuang University in 2014, M. Sc. degree from Huaiyin Institute of Technology in 2018. He is now a Ph. D. candidate at Nanjing University of Information Science and Technology. His main research interests include computer vision, sentiment analysis and multimodal emotion recognition.