山西大学学报(自然科学版)49(2):220—231,2026
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2025094

基于自适应通道特征交互融合图卷积网络的骨骼行为识别

施宇航[1],陈琳琳[1,2*],郭峰[3],何强[1,2]
(1.北京建筑大学 理学院,北京 102616;
2.北京建筑大学 大数据建模理论与技术研究所,北京 102616;
3.奇安信科技集团,北京 100044)

摘 要:本研究提出了一种基于图卷积网络(Graph Convolutional Networks, GCN)的骨骼行为识别方法,针对传统时空图卷积框架在处理时空特征时存在的统一化处理及忽视通道间交互的问题,提出了一种有效的改进方案。模型通过融合多种拓扑矩阵,增强了空间信息的互补表达,同时引入了通道交互注意力模块,通过捕捉时空维度中的帧间动态信息和人体结构特征,建模不同通道间的交互关系,提升特征表达能力。此外,模型设计的时间自适应特征融合模块(Temporal Adaptive Feature Fusion, TAF)通过自适应选择不同网络层中的扩张率和卷积核大小,解决了上下文聚合和初始特征集成的问题。TAF模块分别关注初始特征和时间维度的信息,进行有效的特征融合,成功整合了初始特征与高维时间特征,从而显著提高了时空特征提取的能力。在NW-UCLA数据集上,所提出的方法相比基准模型CTR-GCN(Channel-wise Topology Refinement Graph Convolution Network)提升了2.1%的识别精度,相较最新方法Info-GCN提高了0.7%。在NTU RGB+D 120和NTU RGB+D数据集的不同划分方式下,模型分别比基础模型识别准确率提高了0.7%、0.8%及0.5%、0.6%,并在各项评价指标上均超过了现有最新方法。实验结果表明,所提出的模型在时空特征提取和骨骼行为识别任务中均表现出显著的性能优势。

关键词:骨骼行为识别;通道注意力机制;时空特征融合;通道交互
中图分类号:TP391.4 **文献标志码**:A **文章编号**:0253-2395(2026)02-0220-12

Adaptive Channel Feature Interactive Fusion Network for Skeleton-based Action Recognition

SHI Yuhang[1], CHEN Linlin[1,2*], GUO Feng[3], HE Qiang[1,2]
(1. *School of Science, Beijing University of Civil Engineering and Architecture, Beijing* 102616*, China;*
2. *Institute of Big Data Modeling Theory and Technology, Beijing University of Civil Engineering and Architecture, Beijing* 102616*, China;*
3. *Qi An Xin Technology Group Inc., Beijing* 100044*, China*)

Abstract: This study proposed a novel skeleton-based action recognition method utilizing Graph Convolutional Networks (GCN), which addressed the limitations of conventional spatiotemporal graph convolution frameworks that uniformly process spatiotemporal features while neglecting inter-channel interactions. Specifically, the proposed model enhanced the complementary representation of spatial information through the fusion of multiple topological matrices coupled with the introduction of a Channel Interaction Attention (CIA) module. The CIA module was designed to capture dynamic frame-level information and human structural features

收稿日期:2025-07-26;**修回日期**:2025-09-10
基金项目:国家自然科学基金(12301581);北京市自然科学基金(4252033);北京市教育委员会科学研究计划项目(KM202210016002);北京建筑大学基本科研业务费资助(X25039);北京建筑大学硕士研究生创新项目(PG2025172)
作者简介:施宇航(1998—),男,河南周口人,硕士研究生,研究方向为计算机视觉、行为识别。E-mail:ansyh0512@126.com
* **通信作者**:陈琳琳 (CHEN Linlin), E-mail: chenlinlin@bucea.edu.cn
引文格式:施宇航,陈琳琳,郭峰,等.基于自适应通道特征交互融合图卷积网络的骨骼行为识别[J].山西大学学报(自然科学版),2026,49(2):220-231. DOI:10.13451/j.sxu.ns.2025094.

across spatiotemporal dimensions, effectively modeling inter-channel relationships and thereby improving skeletal data representation. Furthermore, a Temporal Adaptive Feature Fusion (TAF) module was incorporated to adaptively select varying dilation rates and kernel sizes across network layers. This module replaced traditional residual connections between initial features and temporal module outputs, effectively addressing context aggregation and initial feature integration challenges. The TAF module separately processed initial features and temporal information, enabling efficient feature fusion and successful integration of initial features with high-dimensional temporal features, which significantly enhanced spatiotemporal feature extraction. Experimental results demonstrated that on the NW-UCLA dataset, the proposed method achieved 2.1% higher recognition accuracy than the baseline model CTR-GCN (Channel-wise Topology Refinement Graph Convolution Network) and 0.7% improvement over state-of-the-art methods Info-GCN. For the NTU RGB+D 120 and NTU RGB+D datasets under different splits, the model showed consistent performance gains of 0.7%, 0.8% and 0.5%, 0.6%, respectively, surpassing all existing methods across evaluation metrics. These results confirmed the model's superior performance in both spatiotemporal feature extraction and skeleton-based action recognition tasks.

Key words: skeleton-based action recognition; channel attention mechanism; spatiotemporal feature fusion; channel interaction

0 引言

互联网的发展产生大量的视频数据,其中以人为主题的数据居多,所以人类的行为识别能够帮助人类学习和理解人类动作。作为计算机视觉领域的热点之一,人类的动作识别可以用在很多领域,例如视频监控、体育仲裁[1-2]、智能医疗[3]等方面。

早期基于深度学习的方法直接将人体关节坐标视为坐标向量序列[4]或伪图像[5],并将它们输入到卷积神经网络(Convolutional Neural Network, CNN)或循环神经网络(Recurrent Neural Network, RNN)中。这种表示忽略了关节之间内在的图结构关系,为了解决这个问题,Yan等[6]提出时空图卷积网络,利用图卷积网络捕获帧间关键空间关系时间维度上应用时间卷积捕获每个节点的帧间关系。随后出现许多变体,Li等[7]提出的AS-GCN(Actional-structural Graph Convolution Network)和Shi等[8]提出的2s-AGCN(Two-stream Adaptive Graph Convolutional Network),主张从数据中自适应学习空间关节之间的关系。Chen等[9]提出的CTR-GCN(Channel-wise Topology Refinement Graph Convolution Network)网络将三种按照图划分的共享邻接矩阵在通道维度上嵌入到动态空间中。然而,这些方法忽略了时间维度的建模,并且现有的工作在提取时间维度特征工作中仅依赖于固定大小的卷积核。

为了捕获关节之间的关系并增强帧内更具判别性的关节的空间信息,利用空间注意力机制[10]为每个关节生成空间注意力图,应用时间注意力机制[11]为每帧生成时间注意力图。不同通道表示不同的运动特征[9],关节之间的相关性随着动作的变化而变化,因此,探索不同通道中运动的重要性可以丰富时空特征的信息,帮助网络区分出复杂的动作。Ma等[12]提出的CBAM(Convolutional Block Attention Module)方法将通道和空间注意力按顺序结合起来,但只是单独地处理每个维度,其他的维度被全局平均处理为单个的标量。

针对上述现有模型存在的问题,本文提出自适应通道特征交互融合(Adaptive Channel Feature Interactive Fusion, ACFIF)网络。首先,更大的感受野提取多尺度时空拓扑的自适应特征,使用注意力融合机制,可以高效地聚合时空尺度特征,解决上下文聚合和初始特征融合的问题。具体来说,将经过空间拓扑图卷积(Spatial Topology Graph Convolution, STGC)模块,提取空间拓扑特征经过设计的通道交互注意力模块后输入到时空自适应特征融合模块中。邻接矩阵定义的拓扑对于确定图结构中关节关系的表示能力很重要,融合多种计算方式的邻接矩阵,可以使空间信息很好的互补并在自适应中得到较好的平衡。通道交互注意力(Channel Interaction Attention, CIA)模块,将时间和空间信息融合到跨维度交互的通道注意力中。CIA由通道空间交互(Space Channel-wise Interaction, SCI)、时间通道交互(Time Channel Interaction, TCI)两个部分组成,分别捕捉身体的整体关键位置结构和时间维度的动态变化信息,增加对时空通道的敏感程度。时间自适应

融合(Time Adaptive Fusion, TAF)模块由时间自适应(Temporal Adaptive, TA)替代初始特征,时间模块输出的残差连接的注意力特征融合(Attention Feature Fusion, AFF)机制构成。两部分分别关注初始特征和时间维度的信息,有效地融合了初始特征和高维时间特征,解决上下文聚合和初始特征集成的问题。

总体而言,本文的主要贡献包括3个方面:(1)本文优化了传统的多尺度时间卷积,使其自适应性更合理,并且能够融合初始特征,从而获得更大的感受野和局部全局的上下文信息。(2)CIA模块将时空信息嵌入到通道注意力中,进而允许对动作在通道级别进行更加细致化的判别特征,自适应利用通道维度重新校准时空维度得到特征。(3)对于空间图拓扑结构,拓扑结构的互补是必要的,将三种邻接矩阵$\boldsymbol{A}_{\mathrm{s}}$、$\boldsymbol{A}_{\mathrm{p}}$、$\boldsymbol{A}_{\mathrm{l}}$进行组合构建新的图拓扑,能显著提高性能并解决自适应和搜索空间过大的困难。

1 基于骨骼行为识别的相关工作

1.1 骨骼行为识别

基于骨骼的动作识别旨在根据关节点的时间序列推断动作类型。早期的深度学习方法通过卷积神经网络[13]或者循环神经网络[14]进行建模,但因无法有效捕捉骨骼的拓扑结构,识别性能受到限制。Duan等[15]提出的PoseC3D在CNN的基础上进行改进,将关节点热图堆叠成3D体积,保留骨骼的空间和时间特征,但训练开销也随之增加。鉴于人体关节骨骼是天然的图结构,图卷积网络(Graph Convolutional Networks, GCN)得以高效地识别并提取人体骨骼结构的拓扑结构[8-9]信息。Yan等[6]提出的ST-GCN (Spatio-temporal Graph Convolutional Network)首次使用GCN方法并采用三种分区策略,以时空骨骼数据结合时间卷积网络进行时间特征提取。在此基础上,Liu等[16]提出了MS-G3D (Multi-scale Gated 3D ConvNets),引入跨时间点的边缘连接,改善了拓扑时空建模。Chi等[17]提出一种结合新颖的学习目标和编码方法的动作识别学习框架InfoGCN,将信息嵌入人体动作的潜在表示中。Ke等[18]提出的STF(Spatio Temporal Focus)为基于骨骼动作识别提供了灵活的框架用于时空梯度的学习。上述方法大多聚焦于使用相同的时间卷积核和扩张率进行空间特征提取,而忽略了时间特征的建模。

1.2 行为识别中的注意力机制

最初出现是RNN和LSTM(Long Short-term Memory)结合的一种端到端时空注意力模型[19],模拟骨关节和时空注意之间的差异。Qiu等[20]提出了时空元组自注意力网络STT-Former来捕捉连续帧中不同关节之间的依赖关系。Song等[21]提出了时空关节注意模块EfficientGCN-B4,该模块可以在时空序列中找到关键关节,从而更好地实现高效的拓扑建模。Zhou等[22]提出了一种基于CBAM[12]的图注意模块2s-GATCN,其可用于计算任意两个关节之间的语义相关性。为了改善多尺度建模,帮助网络关注信息量最多的特征,注意力被集成到图卷积网络中。Hu等[23]设计了一个通道注意力模块SENet(Squeeze-and-Excitation Network),将全局的时空信息压缩到一个单元,不考虑空间或者时间的联合相关性。Wang等[24]提出了ACTION-Net(Spatiotemporal, Channel and Motion Excitation Network)模型,在两个全连接层之间插入了一个卷积层,用于时间信息中的通道方向特征。随后出现利用时间帧注意力[11,25]来增强时间依赖关系的建模能力。这些注意力机制独立考虑每个维度,对其他所有维度进行全局平均。

2 自适应通道特征融合网络

如图1(a)所示,输入的初始信息首先经过特征融合框架处理,然后通过全局平均池化和全连接层变换得到各个流的识别准确率,再经过动态加权[26]得到最终的结果。每个ACFIF模块由STGC,CIA和TAF三部分构成,结构如图1(c)所示。图1中X为特征图,C为通道数,T为帧数目,N与V表示关节数,$M(\cdot)$表示映射函数。

2.1 空间拓扑图卷积模块

总结之前方法中邻接矩阵的构建方法,主要参考ST-GCN[6]和变体2s-AGCN[8]和CTR-GCN[9],将邻接矩阵分为三种类型:物理连接

的A_p,可以学习的A_l,表现相似性的A_s。

A_p是人体物理连接的预定义的邻接矩阵,训练过程中保持不变。A_l是全局的可学习的矩阵,体现关节对之间是否连接以及连接的强度。A_s是两个顶点之间的高斯相似度矩阵,主要依靠数据训练得到。与A_p相比,A_l和A_s能适应不同的输入样本,自动捕获全局的图信息,但是会存在搜索空间较大,具有复杂结构的拓扑优化过程的混乱的问题。所以将三种邻接矩阵相加能得到更好的效果。

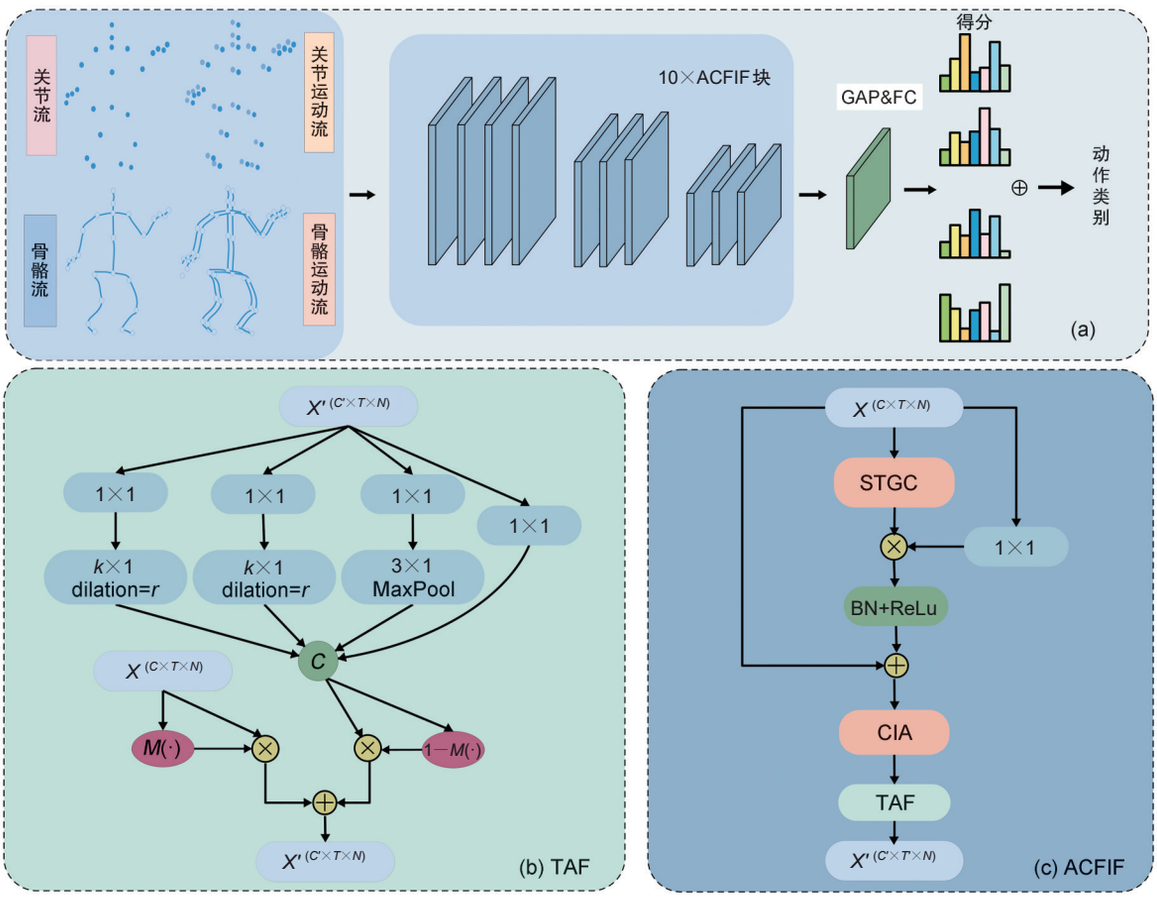


图 1 自适应通道特征交互融合网络模型图

(a)网络整体结构;(b)时间自适应融合模块;(c)特征交互融合网络结构

Fig. 1 Model diagram of adaptive channel feature interactive fusion network

(a) the overall structure of the network; (b) time adaptive fusion module; (c) feature interactive fusion network structure

如图2(a)所示,采用动态拓扑构建空间注意力图,将骨骼数据同时输入到两个并行的分支,每个分支由1×1卷积和时间池化层组成,构建注意力特征。将特征图和改进之后的邻接矩阵$A'=(A_p+A_l+A_s)$相加得到最终的拓扑A_{ct},我们使用注意力特征融合代替普通的残差连接,来更好地聚合空间和时间尺度的信息。

$$A_{ct}=\alpha Q(X_{in})+\sum_{k}^{K_v}(A_p+A_l+A_s), \quad (1)$$

其中α是可学习的参数,A_p和A_l都是$V\times V$邻接矩阵,每个通道都相同。K_v为根据ST-GCN提出的关节的三种划分策略。A_p是包含C个通道特定的邻接矩阵。Q是具体通道的拓扑关系,定义为:

$$Q(X_{in})=\sigma(P_t(\phi(X_{in}))-P_t(\varphi(X_{in}))), \quad (2)$$

其中σ、ϕ、φ都是1×1的卷积,P_t是时间池化。在获得通道拓扑A_{ct}之后,将初始骨骼特征输入到卷积中并将结果和A_{ct}相乘来聚合空间维度的信息得到:

$$X_{out}=A_{ct}\otimes(\theta(X_{in})), \quad (3)$$

其中θ是1×1的卷积,$\otimes$为矩阵的乘法运算。

2.2 通道交互注意力

为了解决联合相关建模中忽略时空维度和通道维度的相互作用的问题,提出空间通道交

互(SCI)、时间通道交互(TCI)两部分构成的通道交互注意(CIA)模块,结构如图2(b)所示。

2.2.1 通道空间交互原理

经过CSI网络之后,可以获得身体整体的关键结构,激发对空间信息敏感的通道,自适应地调整不同通道中关节点的重要性。其结构如图2(c)所示,输入的特征$X\in R^{C\times T\times V}$经过平均池化使网络关注于通道和空间维度之间的相互关系,同时也可以降低计算成本。

$$X_{\mathrm{tp}}=\frac{1}{T}\sum_{j=1}^{T}X[:,j,:],\ X_{\mathrm{tp}}\in R^{C\times V},\tag{4}$$

其中X_{tp}表示时间池化后的特征,T表示帧数。

采用卷积核大小为V的一维卷积层,得到包含帧内所有关节的全局感受野,有助于提取全局结构特征,也减少了通道数量和计算量。输出结果经过ReLU非线性处理之后,用一维的卷积将通道维度扩展为原始维度,并将张量X重塑后送入激活函数中获得注意力掩码。

$$M_{\mathrm{SCI}}=\mathrm{Sigmoid}(\mathrm{conv}_{\mathrm{e}}\otimes\mathrm{relu}(\mathrm{conv}_{\mathrm{s}}\otimes X_{\mathrm{tp}}))\\ M_{\mathrm{SCI}}\in R^{C\times 1\times T}。\tag{5}$$

最后通过原始输入特征X和注意力掩码在通道维度上的乘积获得空间敏感的通道和关键节点,加入残差连接保留原始的信息。

$$F_{\mathrm{SCI}}=X\otimes M_{\mathrm{SCI}}+X,\ F_{\mathrm{SCI}}\in R^{C\times T\times V}。\tag{6}$$

通过空间和通道维度的交互可以保留对空间敏感的通道信息,自适应地调节关节重要性,得到交互后的输出F_{SCI}。

2.2.2 时间通道交互原理

类似于SCI,TCI主要是利用时间动态信息区分对时间敏感的通道和序列帧,其结构如图2(d)所示。输入特征X经过平均池化,对空间信息进行概括处理,然后经过一个卷积核大小为K的一维卷积层,捕获t帧帧间信息。根据不同的数据集将t设置为超参数,以获得合适的感受野。然后经过时间通道维度的相互作用,自适应调整帧的重要性,最终得到输出特征:

$$X_{\mathrm{sp}}=\frac{1}{V}\sum_{j=1}^{V}X[:,:,j],\ X_{\mathrm{sp}}\in R^{C\times T},\tag{7}$$

$$M_{\mathrm{TCI}}=\mathrm{Sigmoid}(\mathrm{Conv}_{\mathrm{e}}\otimes\mathrm{Relu}(\mathrm{Conv}_{\mathrm{t}}\otimes X_{\mathrm{tp}}))\\ M_{\mathrm{TCI}}\in R^{T\times C\times 1},\tag{8}$$

$$F_{\mathrm{TCI}}=X\otimes M_{\mathrm{TCI}}+X,\ F_{\mathrm{TCI}}\in R^{C\times T\times V}。\tag{9}$$

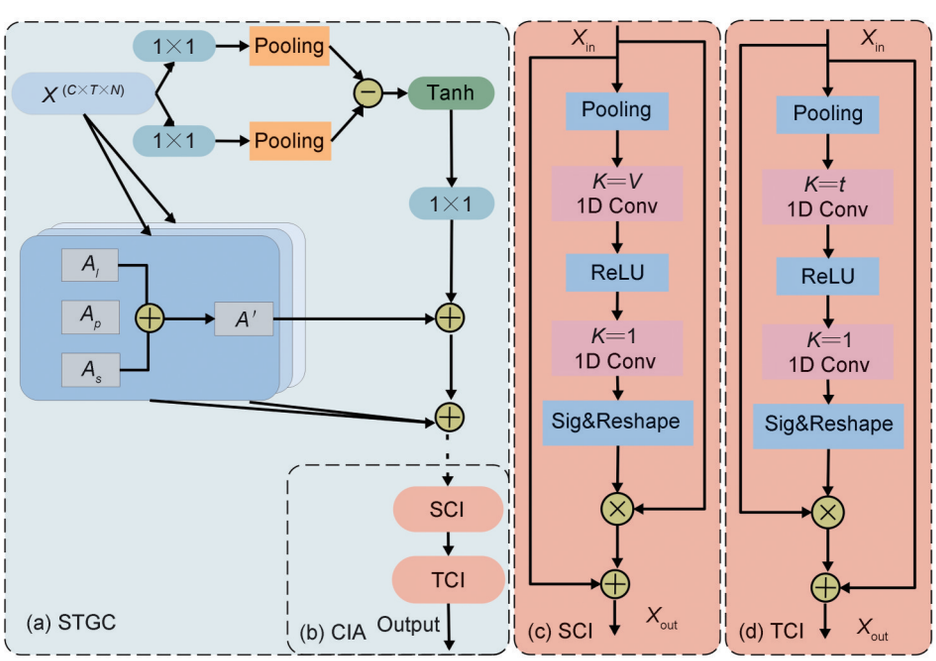


图2 空间拓扑与通道交互模块原理图

Fig. 2 Schematic diagram of spatial topology and channel interaction module

2.3 自适应特征融合模块

时间自适应特征融合模块由TA和AFF两部分组成。第一部分可以动态调整不同网络层的卷积核大小和扩张率。如图1(b)所示,该模块在传统多尺度时间卷积的基础上改进,包含四个分支,每个分支采用1×1卷积降低通道维

数。左边两个分支是自适应函数的核心。通过引入简单的注意机制,可以动态调整卷积核的大小和扩张率。卷积核大小(K)和扩张率(r)可以根据输出通道的不同维度动态调整大小。受到注意力机制[27]的启发,使用以下具体方法公式:

$$t=\left|\frac{\log(C^L,2)+b}{\gamma}\right|, \tag{10}$$

其中C^L是网络L层输出的通道维度,γ和b是映射函数的参数,分别设置为2和1。四个不同规模的分支通过聚合函数得到X_1。

第二个部分使用注意力特征融合模块,沿着通道维度聚合不同尺度和不同维度的上下文信息。使用AFF[28]模块融合不同分支的特征,使用初始的骨骼数据和多尺度聚合特征两个分支。分别关注初始特征和时间维度的信息,有效的融合初始特征和高维时间特征,解决上下文聚合和初始特征融合的问题,提高模型的有效性。

$$X'=X\otimes M+X_1\otimes(1-M), \tag{11}$$

其中X是输入的残差连接,X_1是多尺度卷积的输出,映射函数M[28]的具体表示为:

$$M=\mathrm{Sigmoid}(L(X\uplus X_1)\oplus G(X\uplus X_1)), \tag{12}$$

其中$L(\cdot)$和$G(\cdot)$分别是局部和全局通道上下文信息,其中$\uplus$表示时间特征积分运算,$\oplus$表示将不同的特征乘以元素的操作。局部上下文信息在注意力模块被加入全局上下文信息中。输出特征和X_1进行初步的特征融合,经过激活函数处理之后输出在0~1之间。对X和X_1取加权平均值,并将融合权重减去1,通过学习训练,确定权重。

3 实验

在本节中,首先对于实验所使用的三个数据集和实验配置进行介绍,然后将本研究提出的模型和相关的及最新的方法进行比较,最后通过消融实验来验证本文所提方法的有效性。

3.1 数据集

NW-UCLA数据集是使用三个Kinetic相机从多个视角进行拍摄的。该数据集包含1 494个视频片段,其中包含10个动作类别,用来训练的数据是来自前两个Kinetic摄像头,剩余的一个Kinetic摄像头数据用来进行测试。

NTU-RGB+D数据集中包含60个动作类别,共计有56 880个样本,其中40个类别是日常发生的行为动作,9个类别是和人体健康相关的动作,剩余的11个类别是双人之间的互动动作。利用Kinetic V2传感器获取,使用了三个不同的摄像角度进行捕获。本文使用两个标准来划分训练集和测试集:跨主题(X-sub),交叉视图(X-view)。

NTU-RGB+D 120数据集是现有最大的人体运动3D数据集。该集合添加了57 367个骨骼序列和60个动作类别对NTU RGB+D数据集进行扩展。数据集是利用32个摄像机采集106名志愿者的120动作,总计114 480个骨骼序列。使用两个评估标准:跨主题(X-sub)和交叉设置(X-set)对训练集和测试集进行划分。

3.2 实验参数配置

所有的实验都是在PyTorch框架下进行的。模型使用随机梯度下降(Stochastic Gradient Descent,SGD)优化器训练75个迭代周期,动量设置为0.9,并且在三个数据集训练模型时,我们在前5个迭代周期中采用了热身策略。在NTU RGB+D和NTU RGB+D 120两个数据集的权重衰减设置为0.000 4,批量大小设置为64,NW-UCLA的衰减权重是0.000 1,批量大小为16。本文的实验中采用交叉熵损失,学习率设置为0.1,然后在第35,55,70个迭代周期把学习率缩减至原来的$\frac{1}{10}$。

3.3 识别准确对比实验

为了验证我们提出模型的有效性并且使结果能够相对公平,对比模型选择之前的多流融合模型。我们利用关节、骨骼、关节运动和骨骼运动四个流进行了融合实验,在NW-UCLA,NTU RGB+D,NTU RGB+D 120数据集上和先进的方法进行比较,得到的结果如表1—表3所示。

根据表1结果可知,模型ACFIF在NW-UCLA数据集上的识别准确率相较于最新模型(Info-GCN)提升0.7%,与基础模型(CTR-GCN)相比提升2.1%,由此能看出该模型的有效性。

在NTU RGB+D数据中,模型ACFIF的识

别精度在X-sub和X-view划分标准下的精度分别比最好的模型高0.4%和0.1%,相较于基础模型分别提升0.5%和0.6%。在X-view标准下精度提升较少的问题说明模型对不同角度采集的动作数据进行动作识别方向有一定的改进空间。

在NTU RGB+D 120数据集中,模型ACFIF在X-sub和X-set划分标准下的识别精度分别比最好的模型高0.4%和0.5%,相较于基础模型分别提升0.7%和0.8%,达到最好的效果。

表1 在NW-UCLA数据集上识别准确率百分比的方法比较

Table 1 Comparison of methods for identification accuracy percentage on the NW-UCLA dataset

方法	准确率/%
Lie Group[29]	74.2
Glimpse Clouds[30]	87.6
AGC-LSTM[31]	93.3
Shift-GCN[32]	94.6
CTR-GCN[9]	95.1
Ta-GCN[33]	96.1
Info-GCN[34]	96.5
ours	97.2

表2 在NTU RGB+D数据集上识别准确率百分比的方法比较

Table 2 Comparison of methods for identification accuracy percentage on the NTURGB+D dataset

方法	X-Sub的准确率/%	X-View的准确率/%
ST-LSTM[35]	69.2	77.7
ST-GCN[6]	81.5	88.3
2s-AGCN[8]	88.5	95.1
Shift-GCN[32]	90.7	96.5
Dynamic GCN[36]	91.5	96.2
MST-GCN[37]	91.5	96.6
CTR-GCN[9]	92.4	96.4
EfficientGCN-B4[21]	91.7	95.7
STF[18]	92.5	96.9
Info-GCN[34]	92.3	96.5
STF-Net[38]	91.1	96.5
RSA-Net[39]	91.8	96.8
Ours	92.9	97.0

综上所述,与类似的模型对比ACFIF模型达到最好的效果。由于ACFIF考虑到时空信息和通道信息的融合,不仅可以有效地针对骨骼特征进行高效建模,并且在大型的数据集上有更好的兼容性。

表3 在NTU RGB+D 120数据集上识别准确率百分比的方法比较

Table 3 Comparison of methods for identification accuracy percentage on the NTU RGB+D120 dataset

方法	X-Sub的准确率/%	X-Set的准确率/%
ST-LSTM[35]	55.7	57.9
ST-GCN[6]	70.7	73.2
2s-AGCN[8]	82.9	84.9
Shift-GCN[32]	85.9	87.6
Dynamic GCN[36]	87.3	88.6
MST-GCN[37]	87.5	88.8
CTR-GCN[9]	88.9	90.3
EfficientGCN-B4[21]	88.3	89.1
STF[18]	88.9	89.9
Info-GCN[34]	89.2	90.6
STF-Net[38]	86.5	88.2
RSA-Net[39]	88.4	89.7
Ours	89.6	91.1

3.4 消融对比实验

3.4.1 不同模态消融实验

为了验证基于四个流(关节bone、骨骼joint、关节运动Bone-motion、骨骼运动Joint-motion)融合模型的效果,将模型ACFIF的实验效果和基础模型进行比较,如表4所示。

表4 在NTU-RGB+D数据集上,不同模态识别准确率百分比的比较

Table 4 Comparison of different modal recognition accuracy percentage on the NTU-RGB+D dataset

Stream	CTR-GCN X-sub 准确率/%	Our X-sub 准确率/%	CTR-GCN X-view 准确率/%	Our X-view 准确率/%
bone	90.74	90.65	94.93	95.19
joint	90.06	90.49	94.37	95.69
Bone-motion	87.55	88.16	92.19	92.56
Joint-motion	87.88	88.56	92.92	93.24

模型ACFIF在骨骼流的准确率下降不到0.1%,但是剩余三个流上的准确率分别提升0.43%,0.35%,0.68%,在交叉视图上四个流分别提升0.26%,1.32%,0.37%,0.32%。验证了基于骨骼的多模态动作识别模型的有效性。

3.4.2 模块的消融实验

为了验证网络的有效性,以CTR-GCN网络作为基准模型,分别对不同模块进行消融实验,结果如表5所示。

在NTU-RGB+D 60数据集上进行实验,

观察到模型ACFIF分别去除TA、AFF和CIA三个模块在NTU-RGB+D数据集X-sub划分标准下的识别准确率相较于基准模型分别提升了0.1%,0.2%,0.3%;在X-view的划分标准下的准确率分别提升了0.3%,0.2%和0.3%。模型提出的改进相较于基线模型的识别效果有明显的提升,改进模块的有效性得到验证。

表5 模型ACFIF去除不同模块后识别准确率百分比与基准模型比较

Table 5 Comparison of recognition accuracy percentage with benchmark model after removing different modules of ACFIF

方法	X-sub准确率/%	X-view准确率/%
baseline	92.4	96.4
w/o TA	92.5(+0.1)	96.7(+0.3)
w/o TFF	92.6(+0.2)	96.6(+0.2)
w/o CIA	92.7(+0.3)	96.7(+0.3)
Ours	92.9(+0.5)	97.0(+0.6)

注:表中w/o表示with/without。

3.4.3 自适应动态变化实验

对于多尺度的时间卷积模块,验证动态调整卷积核和膨胀率的大小对模型的有效性。网络中前五层的通道数是64,六到八层通道数是128,九层和十层的通道数是256。根据输出通道的变化,对卷积核和膨胀率的大小进行动态调整,在NTU RGB+D数据集X-view下关节单流识别准确率的结果,如表6所示。

表6 NTU-RGB+D数据集X-view划分标准下卷积核和膨胀率有效性验证

Table 6 Validation of convolution kernel and expansion rate under X-view partition standard of NTU-RGB+D dataset

方法		通道数			准确率/%
		64	128	256	
卷积核	原始大小	5	5	5	95.3
	调整后大小	3	5	5	95.5
膨胀率	原始大小	2	2	2	95.3
	调整后大小	2	2	3	95.6
TA		—	—	—	96.7

注:—表示大小动态变化,不固定。

模型ACFIF分别改变卷积核和膨胀率之后,识别精度分别提升了0.2%,0.3%。TA模块对卷积核和膨胀率同时动态调整后模型的识别精度为96.7%,结果比不调整提升1.4%。如果再增加另外三个分支之后,会带来更大的提升。

3.5 模型有效性验证实验

在NTU-RGB+D的交叉主题(X-sub)划分标准下,将物理连接的邻接矩阵$\boldsymbol{A}_{\mathrm{p}}$,可学习的邻接矩阵$\boldsymbol{A}_{\mathrm{l}}$和通道相似度的邻接矩阵$\boldsymbol{A}_{\mathrm{s}}$分别删除,进行实验验证,结果如表7所示。

表7 NTU-RGB+D数据集X-sub划分标准下去除不同邻接矩阵的识别准确率

Table 7 The recognition accuracy of NTU-RGB+D dataset divided by different adjacency matrices by X-sub partition standard

数据集	邻接矩阵方案			准确率/%
	A_p	A_l	A_s	
NTU X-sub	✓		✓	92.4
	✓	✓		92.6
		✓	✓	92.6
	✓	✓	✓	92.9

注:✓表示使用此邻接矩阵,空白表示未使用。

使用融合邻接矩阵,模型的识别准确率达到92.9%,当分别除去矩阵$\boldsymbol{A}_{\mathrm{s}}$、$\boldsymbol{A}_{\mathrm{p}}$、$\boldsymbol{A}_{\mathrm{l}}$时,性能分别下降了0.3%,0.3%,0.5%。验证使用三种邻接矩阵是高效且能够互补的,也证实了融合矩阵方案的合理性。

为了验证模型的稳定性,对NTU RGB+D、NTU RGB+D 120数据集X-Sub下不同的模态分别进行3次实验,得到识别准确率并计算出平均精确度和对应方差。方差都小于0.05,可知网络模型是稳定的,本文选取精确度最高的作为最终结果,如表8所示。

3.6 可视化实验

为了更直观地展示模型对骨骼动作的识别效果,本研究选取多个动作序列进行可视化分析。通过解析关键帧中骨骼的动态变化,可以清晰地观察到模型对不同动作特征的捕捉能力。

喝水动作中,模型能够准确识别手部向嘴部移动的关键轨迹,同时捕捉到身体姿态的细微变化,体现了对时空特征的细腻建模。坐下动作中,模型不仅能够检测到身体下蹲的整体运动趋势,还能注意到腿部关节的弯曲角度变化以及上半身的平衡调整。这种对局部关节动态的敏感性,验证了通道交互注意力模块在增强空间信息表达上的有效性。选取两个动作序列可视化实验,分别截取三帧图像,结果如图3和图4所示。

表 8 模型稳定性验证表

Table 8 Stability verification of model

数据集	模态	第一次实验	第二次实验	第三次实验	平均值	方差
NTU RGB+D 120	Bone	86.85%	87.06%	86.75%	86.89%	0.017
NTU RGB+D 120	Joint	85.13%	85.23%	85.55%	85.30%	0.032
NTU RGB+D	Bone-motion	87.90%	88.16%	87.89%	87.98%	0.016
NTU RGB+D	Joint-motion	88.22%	88.56%	88.26%	88.35%	0.023

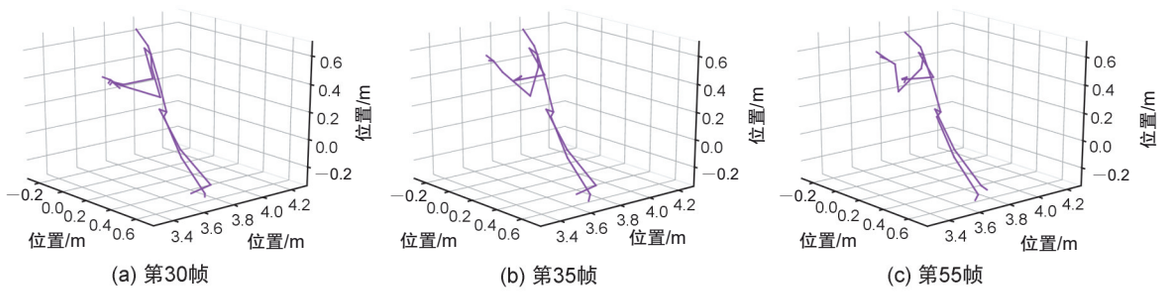


图 3 喝水动作序列解析图

Fig. 3 Analysis diagram of action sequence of drinking water

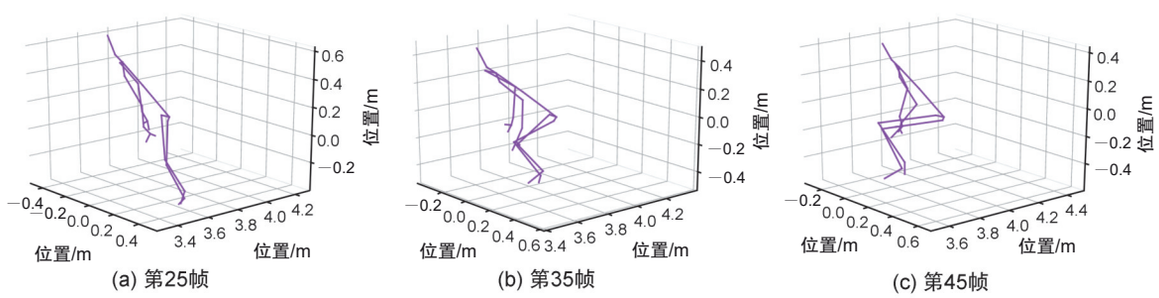


图 4 坐下动作序列解析图

Fig. 4 Analysis diagram of the sequence of sitting down actions

在举双手动作中,模型清晰地捕捉到了双臂上举的对称性和时间同步性;而在跳跃动作中,模型则突出了腿部发力与身体腾空的动态关联。这种对不同动作模式的区分能力,充分体现了模型在多尺度时空特征融合上的优势。多帧连续动作可视化如图5所示。

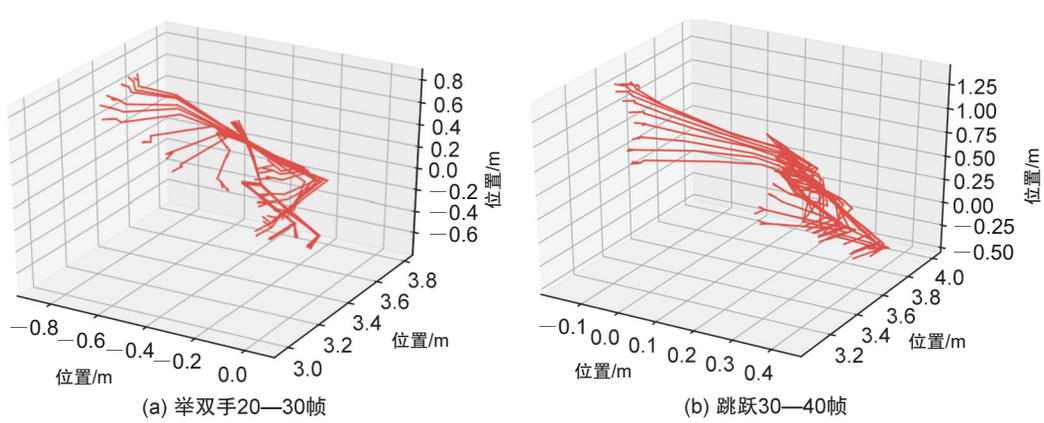


图 5 多帧动作序列可视化

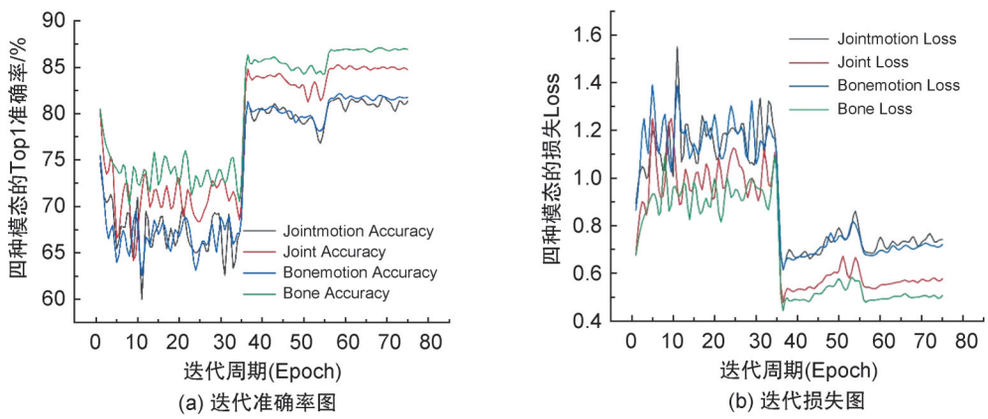
Fig. 5 Visualization of multi-frame action sequence

这些可视化结果不仅验证了模型对单一动作的识别精度,还展示了其对连续动作中时空关联性的建模能力。通过动态调整关节重要性权重,模型能够自适应地关注不同动作中的关键帧

和关键关节点,从而实现对复杂动作的准确解析。

对模型进行评价时,不仅要考虑精度也需要考虑收敛的速度,所以将NTU RGB+D 120数据集X-sub划分标准下四个模态的准确率和损失函数利用折线图直观地体现出来。本文设置的35,55,70分别为学习率变化的迭代节点,效果如图3所示。

根据图6(a)和图3(b)显示,在给定训练权重后,模型在前35次迭代中,识别精度和损失均呈现波动。自第36次迭代开始,损失函数在固定的小范围内波动属于正常现象。与此同时,模型的识别精度在第36次迭代后显著提升,逐渐接近最终的精确度,并在后续迭代中保持小幅波动,逐步趋于稳定。



注:TOP1准确率是预测结果中的第一个预测(概率最高)与实际标签相符的概率。

图6 NTU RGB+D 120数据集X-sub划分标准下的迭代准确率图(a)和迭代损失图(b)

Fig. 6 Iterative accuracy chart (a) and iterative loss chart (b) under the X-sub partition standard of NTURGB+D 120 dataset

4 结论

本文提出了一种自适应通道特征交互融合图卷积网络用于骨骼动作识别。(1)引入了拓扑互补策略,并采用邻接矩阵融合的方式实现多元拓扑信息的有效整合,以克服自适应过程中搜索空间过大的难题;(2)利用通道交互注意力模块,将时间特征和空间特征分别与通道注意力结合,充分捕获不同通道下的关节点结构与帧间动态信息,并自适应地调控关节点和帧间的重要性关系。(3)通过时空信息融合模块,动态调整网络模型中卷积核大小及不同层的膨胀率,沿通道维度聚合多层次上下文信息,从而在时空两方面实现更为细腻的动作表征。在数据集NTU RGB+D 120的X-sub和X-set划分标准下的识别精度分别得到0.7%和0.8%的显著提升。同时网络的检测精度也优于其他主流基于图卷积的骨骼行为识别的方法。

综上所述,本文提出的补充拓扑信息、利用时空信息和通道维度的关系提取重要信息并动态调整参数的方法,在多个公开数据集上均取得了显著的性能提升,不仅验证了所提方法在骨骼动作识别上的有效性与泛化能力,也为进一步研究通道间特征交互与多拓扑融合提供了新的思路。未来将继续优化网络的结构或结合不同方法优化骨骼行为识别的精度。

参考文献:

[1] WEINLAND D, RONFARD R, BOYER E. A Survey of Vision-based Methods for Action Representation, Segmentation and Recognition[J]. *Comput Vis Image Underst*, 2011, **115**(2): 224-241. DOI: 10.1016/j.cviu.2010.10.002.

[2] POPPE R. A Survey on Vision-based Human Action Recognition[J]. *Image Vis Comput*, 2010, **28**(6): 976-990. DOI: 10.1016/j.imavis.2009.11.014.

[3] MOCCIA S, MIGLIORELLI L, CARNIELLI V, *et al*. Preterm Infants' Pose Estimation with Spatio-temporal Features[J]. *IEEE Trans Biomed Eng*, 2020, **67**(8): 2370-2380. DOI: 10.1109/TBME.2019.2961448. [PubMed]

[4] LI S, LI W, COOK C, *et al*. Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN[EB/OL]. (2018-05-22) [2024-07-05]. https://arxiv.org/abs/1803.04831.

[5] KE Q H, BENNAMOUN M, AN S J, *et al*. A New Representation of Skeleton Sequences for 3D Action Recog-

nition[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2017: 4570−4579. DOI: 10.1109/CVPR.2017.486.

[6] YAN S J, XIONG Y J, LIN D H. Spatial Temporal Graph Convolutional Networks for Skeleton-based Action Recognition[J]. *Proc AAAI Conf Artif Intell*, 2018, **32**(1): 7444−7452. DOI: 10.1609/aaai.v32i1.12328.

[7] LI M S, CHEN S H, CHEN X, *et al*. Actional-structural Graph Convolutional Networks for Skeleton-based Action Recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2019: 3590−3598. DOI: 10.1109/CVPR.2019.00371.

[8] SHI L, ZHANG Y, CHENG J, *et al*. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 12018−12027. DOI: 10.1109/CVPR.2019.01230

[9] CHEN Y X, ZHANG Z Q, YUAN C F, *et al*. Channel-wise Topology Refinement Graph Convolution for Skeleton-based Action Recognition[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). New York: IEEE, 2021: 13339−13348. DOI: 10.1109/ICCV48922.2021.01311.

[10] GAO B K, DONG L, BI H B, *et al*. Focus on Temporal Graph Convolutional Networks with Unified Attention for Skeleton-based Action Recognition[J]. *Appl Intell*, 2022, **52**(5): 5608−5616. DOI: 10.1007/s10489-021-02723-6.

[11] QIU H Y, WU Y, DUAN M M, *et al*. GLTA-GCN: Global-local Temporal Attention Graph Convolutional Network for Unsupervised Skeleton-based Action Recognition[C]//2022 IEEE International Conference on Multimedia and Expo (ICME). New York: IEEE, 2022: 1−6. DOI: 10.1109/ICME52920.2022.9859752.

[12] MA B, WANG X R, ZHANG H, *et al*. CBAM-GAN: Generative Adversarial Networks Based on Convolutional Block Attention Module[M]//Artificial Intelligence and Security. Cham: Springer International Publishing, 2019: 227−236. DOI: 10.1007/978-3-030-24274-9_20.

[13] LIU M Y, LIU H, CHEN C. Enhanced Skeleton Visualization for View Invariant Human Action Recognition [J]. *Pattern Recognit*, 2017, **68**: 346−362. DOI: 10.1016/j.patcog.2017.02.030.

[14] LIU J, WANG G, HU P, *et al*. Global Context-aware Attention LSTM Networks for 3D Action Recognition[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2017: 3671−3680. DOI: 10.1109/CVPR.2017.391.

[15] DUAN H D, ZHAO Y, CHEN K, *et al*. Revisiting Skeleton-based Action Recognition[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2022: 2959−2968. DOI: 10.1109/CVPR52688.2022.00298.

[16] LIU Z, ZHANG H, CHEN Z, *et al*. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition[EB/OL]. (2020−05−19) [2024−07−10]. https://arxiv.org/abs/2003.14111.

[17] CHI H G, HA M H, CHI S, *et al*. InfoGCN: Representation Learning for Human Skeleton-based Action Recognition [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2022: 20154−20164. DOI: 10.1109/CVPR52688.2022.01955.

[18] KE L P, PENG K C, LYU S W. Towards To-a-T Spatio-temporal Focus for Skeleton-based Action Recognition [J]. *Proc AAAI Conf Artif Intell*, 2022, **36**(1): 1131−1139. DOI: 10.1609/aaai.v36i1.19998.

[19] SONG S J, LAN C L, XING J L, *et al*. An End-to-end Spatio-temporal Attention Model for Human Action Recognition from Skeleton Data[J]. *Proc AAAI Conf Artif Intell*, 2017, **31**(1): 4263−4270. DOI: 10.1609/aaai.v31i1.11212

[20] QIU H, HOU B, REN B, *et al*. Spatio-Temporal Tuples Transformer for Skeleton-Based Action Recognition [EB/OL]. (2022−01−08) [2024−07−10]. https://arxiv.org/abs/2201.02849.

[21] SONG Y F, ZHANG Z, SHAN C F, *et al*. Constructing Stronger and Faster Baselines for Skeleton-based Action Recognition[J]. *IEEE Trans Pattern Anal Mach Intell*, 2023, **45**(2): 1474−1488. DOI: 10.1109/TPAMI.2022.3157033.

[22] ZHOU S B, CHEN R R, JIANG X Q, *et al*. 2s-GATCN: Two-stream Graph Attentional Convolutional Networks for Skeleton-based Action Recognition[J]. *Electronics*, 2023, **12**(7): 1711. DOI: 10.3390/electronics12071711.

[23] HU J, SHEN L, ALBANIE S, *et al*. Squeeze-and-Excitation Networks[EB/OL]. (2019−05−16) [2024−01−05]. https://arxiv.org/abs/1709.01507.

[24] WANG Z W, SHE Q, SMOLIC A. ACTION-Net: Multipath Excitation for Action Recognition[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2021: 13209−13218. DOI: 10.1109/cvpr46437.2021.01301.

[25] XIE Y L, ZHANG Y, REN F. Temporal-enhanced Graph Convolution Network for Skeleton-based Action Recognition[J]. *IET Comput Vis*, 2022, **16**(3): 266−279. DOI: 10.1049/cvi2.12086.

[26] WANG S Q, ZHANG Y, ZHAO M, *et al*. Skeleton-

Based Action Recognition via Temporal-Channel Aggregation[EB/OL]. (2022−08−08) [2024−10−25] https://arxiv.org/abs/2205.15936.

[27] WANG Q L, WU B G, ZHU P F, *et al*. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2020: 11531−11539. DOI: 10.1109/cvpr42600.2020.01155.

[28] DAI Y M, GIESEKE F, OEHMCKE S, *et al*. Attentional Feature Fusion[C]//2021 IEEE Winter Conference on Applications of Computer Vision (WACV). New York: IEEE, 2021: 3559−3568. DOI: 10.1109/wacv48630.2021.00360.

[29] VEMULAPALLI R, ARRATE F, CHELLAPPA R. Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2014: 588−595. DOI: 10.1109/CVPR.2014.82.

[30] GAO X H, DU S Y, YANG Y. Glimpse and Focus: Global and Local-scale Graph Convolution Network for Skeleton-based Action Recognition[J]. *Neural Netw*, 2023, **167**: 551−558. DOI: 10.1016/j.neunet.2023.07.051.

[31] SI C Y, CHEN W T, WANG W, *et al*. An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-based Action Recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2019: 1227−1236. DOI: 10.1109/CVPR.2019.00132.

[32] CHENG K, ZHANG Y F, HE X Y, *et al*. Skeleton-based Action Recognition with Shift Graph Convolutional Network[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2020: 180−189. DOI: 10.1109/cvpr42600.2020.00026.

[33] XU K L, YE F F, ZHONG Q Y, *et al*. Topology-aware Convolutional Neural Network for Efficient Skeleton-based Action Recognition[J]. *Proc AAAI Conf Artif Intell*, 2022, **36**(3): 2866−2874. DOI: 10.1609/aaai.v36i3.20191.

[34] HUANG X, ZHOU H, WANG J, *et al*. Graph Contrastive Learning for Skeleton-Based Action Recognition [EB/OL]. (2023−06−10) [2024−10−15]. https://arxiv.org/abs/2301.10900.

[35] LIU J, SHAHROUDY A, XU D, *et al*. Spatio-temporal LSTM with Trust Gates for 3D Human Action Recognition[M]//*Computer Vision-ECCV 2016*. Cham: Springer International Publishing, 2016: 816−833. DOI: 10.1007/978-3-319-46487-9_50.

[36] YE F F, PU S L, ZHONG Q Y, *et al*. Dynamic GCN: Context-enriched Topology Learning for Skeleton-based Action Recognition[C]//Proceedings of the 28th ACM International Conference on Multimedia. Seattle: ACM, 2020: 55−63. DOI: 10.1145/3394171.3413941.

[37] CHEN Z, LI S C, YANG B, *et al*. Multi-scale Spatial Temporal Graph Convolutional Network for Skeleton-based Action Recognition[J]. *Proc AAAI Conf Artif Intell*, 2021, **35**(2): 1113−1122. DOI: 10.1609/aaai.v35i2.16197.

[38] WU L Y, ZHANG C, ZOU Y X. SpatioTemporal Focus for Skeleton-based Action Recognition[J]. *Pattern Recognit*, 2023, **136**: 109231. DOI: 10.1016/j.patcog.2022.109231.

[39] GEDAMU K, JI Y L, GAO L L, *et al*. Relation-mining Self-attention Network for Skeleton-based Human Action Recognition[J]. *Pattern Recognit*, 2023, **139**: 109455. DOI: 10.1016/j.patcog.2023.109455.