E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)08-2822-13

论文引用格式:Wu Z Z, Chen X, Xu T, Nian F D, Wang X F and Li T. 2025. Dynamic multi-granularity graph convolutional networks for skeleton-based action recognition. Journal of Image and Graphics, 30(8):2822-2834(吴志泽, 陈鑫, 徐童, 年福东, 王晓峰, 李腾. 2025. 基于动态多粒度图卷积网络的人体骨架行为识别. 中国图象图形学报, 30(8):2822-2834)[DOI:10.11834/jig.240352]

基于动态多粒度图卷积网络的人体骨架行为识别

吴志泽[1],陈鑫[1],徐童[2],年福东[1],王晓峰[1],李腾[3*]

1. 合肥大学人工智能与大数据学院,合肥　230601; 2. 中国科学技术大学计算机学院,合肥　230027;
3. 安徽大学人工智能学院,合肥　230601

摘　要:**目的**　基于图卷积网络的方法在人体骨架行为识别任务中越来越受欢迎,并取得了显著进展。传统图卷积在远距离节点信息交互方面的局限,导致在捕获骨架中非自然连接节点信息时表现不佳,同时现有致力于复杂空间建模的方法,也面临着特征冗余和参数量显著增加的问题。为此,提出一种基于动态多粒度图卷积网络的人体骨架行为识别方法。**方法**　本文根据人体关节点的不同组合方式重构骨架图,设计3种不同粒度的图结构,从而更好地捕获骨架图中的非自然连接节点信息。为了应对特征冗余和参数量增大的难题,引入了空间重组卷积模块,该模块通过分离—重建操作将信息丰富与匮乏的特征进行交叉重构,有效减少了空间维度特征的冗余。在特征融合阶段,根据3种粒度的图结构引出了全新的六流融合方式,利用它们的互补信息以提高模型的整体性能。**结果**　与基线方法CTR-GCN(channel-wise topology refinement graph convolution network)相比,所提方法在基准数据集NTU-RGB+D、NTU-RGB+D 120和Northwestern-UCLA上分别得到了0.6%、0.7%和0.7%的提升。**结论**　动态多粒度图卷积网络结合多粒度图结构和空间—通道重组卷积,是一种新的时空建模方法,通过扩大图卷积网络的感受野并显著减少时空建模过程中的特征冗余,提高了模型捕捉复杂人体动作的能力和准确性。

关键词:图卷积;骨架行为识别;多粒度;特征冗余;重组卷积

Dynamic multi-granularity graph convolutional networks for skeleton-based action recognition

Wu Zhize[1], Chen Xin[1], Xu Tong[2], Nian Fudong[1], Wang XiaoFeng[1], Li Teng[3*]

1. School of Artificial Intelligence and Big Data, Hefei University, Hefei　230601, China;
2. School of Computer Science and Technology, University of Science and Technology of China, Hefei　230027, China;
3. School of Artificial Intelligence, Anhui University, Hefei　230601, China

Abstract: **Objective** In recent years, methods based on graph convolutional networks (GCNs) have become increasingly popular in human skeleton-based action recognition, which resulted in significant strides in this challenging domain. These advances are primarily attributed to the ability of GCNs to model spatial and temporal dependencies inherent in human skeletal data. However, traditional graph convolutions exhibit notable limitations, particularly in capturing interaction informa-

收稿日期:2024-06-24;**修回日期**:2025-01-15;**预印本日期**:2025-01-23

*通信作者:李腾　liteng@ahu.edu.cn

基金项目:国家自然科学基金项目(62406095);安徽省自然科学基金项目(2308085MF213);安徽省重点研发计划资助(2022K07020011);合肥市自然科学基金项目(HZR2447);安徽省高校科学研究创新团队项目(2022AH010095, 2024AH010030)

Supported by: National Natural Science Foundation of China (62406095); Natural Science Foundation of Anhui Province, China (2308085MF213); Key R&D Program of Anhui Province, China (2022K07020011); Hefei Natural Science Foundation (HZR2447); Innovation Team of the Anhui Higher Education Institutions of China (2022AH010095, 2024AH010030)

tion between distant nodes. This shortcoming leads to suboptimal performance in recognizing non-natural connections within the skeleton graph, which is a crucial aspect for accurately modeling complex human actions. Traditional GCNs are adept at processing locally connected nodes, but their efficacy diminishes as the distance between nodes increases. This concern is common in the context of human skeletons, where actions often involve coordinated movements of body parts that are not directly connected. For instance, actions involving simultaneous hand and foot movements necessitate an understanding of long-range dependencies. The inability of conventional GCNs to effectively capture these dependencies results in a limited understanding of the overall action, which reduces recognition accuracy. Moreover, existing approaches that attempt to model complex spatial relationships usually encounter significant issues related to feature redundancy and an exponential increase in parameter count. Although these methods are sophisticated, they tend to generate a large number of redundant features, which not only increase computational complexity but also hamper the overall efficiency of the model. **Method** A novel multi-granularity graph structure called the dynamic multi-granularity graph convolutional network (DMG-GCN) is proposed for skeleton graph construction to address the aforementioned challenges. This approach involves designing three different granularity graph structures, with each of them being tailored to capture distinct aspects of the skeletal data. By combining various human body joint points in innovative ways, these multi-granularity graphs enable the model to capture interaction information between non-naturally connected nodes more effectively. This hierarchical representation allows for a more nuanced understanding of the spatial relationships within the skeleton graph. Based on the multi-granularity graph structure, a dynamic adjacency matrix is introduced in spatial modeling. Unlike static adjacency matrices, which remain fixed regardless of the specific action being performed, the dynamic adjacency matrix adapts depending on the current spatial configuration of the nodes. This adaptability ensures a more accurate representation of the semantic relationships between nodes, which leads to improved recognition performance. In addition to the dynamic adjacency matrix, a spatial reorganization convolution module is proposed to mitigate feature redundancy and growing parameter volume. This module operates by cross-reconstructing information-rich and -poor features through separation-reconstruction operations. The module effectively distinguishes and reorganizes these features. Thus, it reduces spatial dimension feature redundancy, which enhances the efficiency and performance of the model. During the feature fusion stage, a new six-stream fusion method is introduced, which leverages the complementary information derived from the three-granularity graph structures. This method integrates the diverse insights provided by each granularity level, which leads to a more comprehensive understanding of the skeletal data. The integration of these streams guarantees that the model captures the full spectrum of spatial and temporal dependencies, which significantly improves overall performance. **Result** The efficacy of the proposed approach is confirmed by its performance on benchmark datasets. Compared with the baseline method CTR-GCN, the proposed method achieves improvements of 0.6%, 0.7%, and 0.7% on the NTU-RGB+D, NTU-RGB+D 120, and Northwestern-UCLA datasets, respectively. These improvements are seemingly modest, but they represent significant advancements in the highly competitive field of human skeleton-based action recognition. The ablation studies further validate the effectiveness of the multi-granularity graph structure and spatial channel reconstruction convolution within the proposed architecture. These studies highlight the individual contributions of each component, which demonstrates how the multi-granularity approach enhances the ability of the model to capture complex interactions while the spatial reorganization convolution reduces redundancy and improves efficiency. In addition, comparative visualizations underscore the superiority of the dynamic adjacency matrix over conventional adjacency matrices. These visualizations reveal how the dynamic matrix more effectively captures semantically informative connections between nodes, which facilitates a deeper understanding of complex actions. **Conclusion** Our DMG-GCN represents a significant advancement in spatiotemporal modeling for human skeleton-based action recognition. By integrating a multi-granularity graph structure with spatial channel reconstruction convolution, this approach expands the receptive field of GCNs and substantially reduces feature redundancy. The dynamic adjacency matrix further enhances the capability of the model to capture intricate semantic relationships, which leads to more accurate and nuanced action recognition. The proposed DMG-GCN not only addresses the limitations of traditional GCNs but also sets a new benchmark for future research in the field. Its innovative approach to handling long-distance node interactions and reducing feature redundancy lays the foundation for developing more advanced and efficient models. As human skeleton-based action recognition continues to evolve, the principles and techniques introduced by

DMG-GCN are likely to inspire further advancements. Such innovations will drive the field toward even greater accuracy and applicability in real-world scenarios.

Key words: graph convolution; skeleton-based action recognition; multi-granularity; feature redundancy; reconstruction convolution

0 引 言

行为识别是通过接收视频或图像数据(王帅琛等,2022)作为输入并对其进行分类的任务,是计算机视觉领域的重要研究方向之一。相关行为识别以及姿态识别的方法也广泛应用于辅助驾驶(李少凡等,2023)和异常行为识别检测(周航 等,2021;郝亚洲等,2016)。近年来,随着深度学习技术的发展,基于RGB视频的方法(Simonyan和Zisserman,2014)以及基于点云信息的行为识别方法(尤凯军 等,2024)取得了显著进展。然而,上述方法由于数据模态的原因受到环境噪声的强烈影响,如背景颜色、光线亮度和衣物都可能影响识别效果,无法鲁棒地识别人类行为。因此,使用骨架模态的方法(卢健 等,2023)因受噪声的影响较小而受到关注。人体骨架行为识别已涌现许多优秀工作,目前主流的方法为提取空间与时间不同维度的特征,采用时间—空间分离卷积(Yan等,2018)的操作。空间维度采用图卷积神经网络(graph convolutional network, GCN)架构,时间维度采用多尺度时间卷积网络架构(Liu等,2020),最后聚合空间与时间的特征以提取时空特征。联合时间与空间进行建模提升了识别精度,但这种方法仍然面临一些挑战。

首先,从空间建模的角度来看,物理骨架图是根据人体的内在连通性预定义的。因此,在图卷积神经网络中缺乏灵活性,无法捕捉非自然连接节点以及远距离节点之间的联系。同样无法处理不同样本之间的各种联合关系,特别是当它们执行不同的动作时。例如,“手指”和“头部”之间的关系对于区分“挥手”和“触头”非常重要。

此外,从时间建模的角度看,现有的工作局限性在于使用固定长度的窗口进行卷积,无法有效捕捉到全局和局部的时间关系。传统的时间卷积模块处理长期依赖关系可能存在困难。研究者试图通过使用3D图卷积(Wei等,2020)或基于注意力的多流模型(Liu等,2022)缓解这一问题。虽然这些方法提高了识别精度,但注意力机制的引入导致了高昂的计算成本与特征冗余,严重限制实际应用。

为了克服远距离以及非自然连接节点信息交互匮乏的挑战,本文构建了一种动态多粒度图结构。具体而言,将骨架图根据人体的上下结构自然地划分为不同粒度的图层,再与通道拓扑细化模块相结合,以学习非自然连接节点之间的联系。在训练的过程中,由于输入网络的是不同粒度的骨架图层,网络学习到的是层与层之间的联系,更适合提取语义信息,利于捕获非自然连接节点、距离较远节点之间的关系。为了应对时间模块处理全局信息能力较弱和特征冗余的难题,本文将Li等人(2023)提出的空间—通道重组卷积模块(spatial channel reconstruction convolution, ScConv)重构后引入动态多粒度图卷积架构。具体而言,空间维度引入空间重组卷积模块降低通道细化模块带来的特征冗余;时间维度上构建含有通道重组卷积模块的多尺度时间建模,重组卷积的引入提高了模型全局建模能力,以实现“轻量化注意力机制”的效果。

基于上述分析,本文通过融合基于多粒度图结构的多粒度空间卷积模块、空间—通道重组卷积模块以及改进的多尺度时间卷积模块,构成动态多粒度图卷积网络(dynamic multi-granularity graph convolution network, DMG-GCN)。为了验证所提架构在人体骨架行为识别任务中的有效性,在3个流行数据集NTU-RGB+D(Shahroudy 等,2016)、NTU-RGB+D 120(Liu 等,2020)和NorthWestern-UCLA(Wang等,2014)上进行验证。实验结果表明,本文架构在相应指标下优于对比模型。

本文的主要贡献如下:1)提出一种空间建模模块—多粒度空间卷积模块,将提出的多粒度图结构与通道拓扑细化建模结合。该模块通过生成动态邻接矩阵扩大了感受野,从而使网络能够学习物理上非自然连接节点间的信息交互。2)对空间—通道重组卷积模块进行了重构。空间、时间建模完毕后,加入空间—通道重组卷积的特定模块,以降低时空建模带来的特征冗余。3)将多粒度空间卷积模块、空

间—通道重组卷积模块以及改进的多尺度时间卷积模块结合，构成动态多粒度图卷积网络架构，该模型在骨架行为识别的3个基准数据集上实现了显著的性能提升。

1 相关工作

1.1 图卷积神经网络

图卷积神经网络是一种能够有效处理非网格结构数据的神经网络，通过将卷积操作扩展到图结构数据，实现对节点表示的更新。马帅等人(2022)提出图卷积神经网络的关键思想是通过聚合节点的邻居信息更新节点的表示，从而考虑了节点之间的连接关系，使网络能够捕捉到节点之间的信息交互。然而，在传统的图卷积操作中，节点之间的连接关系仅限于局部邻居节点，无法捕捉到远距离节点之间的信息交互。针对上述问题，本文提出多粒度重建单元，通过多粒度邻接矩阵的迭代更新以学习远距离节点的特征表示。

1.2 骨架行为识别

相较于传统的基于RGB视频的行为识别，骨架模态由于受环境因素(如光照、遮挡和姿态)影响较小而得到广泛应用。

早期的骨架行为识别工作(Xia等，2012)通过使用深度相机采集的人体骨架数据，以直方图的形式描述人体关节的运动模式，并利用这些直方图进行行为分类。Yan等人(2018)摒弃了传统的手工制作方法，提出一种全新的架构时空图卷积网络(spatial temporal graph convolutional network，ST-GCN)，该架构使用端到端的方法学习空间和时间特征。由于ST-GCN架构中对于时间卷积网络的设计过于单一，导致时空图卷积缺乏多尺度上下文聚合能力。针对上述问题，Liu等人(2020)提出的多尺度解耦聚合图卷积(disentangling and unifying graph convolution，MS-G3D)解耦了不同邻域中节点的重要性，实现了时间维度有效长期建模。该架构出现之后，时间维度建模趋于稳定，但空间建模中对于特征的不同通道维度均采用了统一的处理方式。

基于上述问题，Chen等人(2021a)提出一种新的通道拓扑优化图卷积架构(channel-wise topology refinement graph convolutional network，CTR-GCN)，动态学习不同的拓扑，并有效地聚合不同通道中的关节特征。通道拓扑优化图卷积架构作为本文的基线模型，仍继续沿用时空图卷积(Yan等，2018)的固定图拓扑结构，并未考虑到包含语义信息的节点交互。针对这一问题，本文提出多粒度图结构，将原始骨架节点按照人体形态学分为不同粒度以学习含有语义信息的特征表示。

1.3 空间—通道重组卷积

空间—通道重组卷积(ScConv)的设计构想由Li等人(2023)提出，其目的是减少特征冗余并促进代表性特征的学习。整体架构由两部分组成，分别是空间重组卷积单元(spatial reconstruction convolution union，SRU)以及通道重组卷积单元(channel reconstruction convoluton union，CRU)。该架构作为目标检测领域内一种即插即用的卷积模块代替普通卷积，以达到降低特征冗余、提升特征表征能力的效果。

空间重组卷积单元利用了分离和重组操作。分离操作即在特征图中将信息丰富与信息匮乏的内容进行提取和筛选。首先利用组归一化层中的比例因子评估不同特征图的信息内容，得到空间权重 $\boldsymbol{W}_\gamma$，具体为

$$\boldsymbol{W}_\gamma = \{\omega_i\} = \frac{\gamma_i}{\sum_{j=1}^{C}\gamma_j}, i,j = 1,2,\cdots,C \tag{1}$$

式中，γ_i为可训练参数，测量每个批次和通道的空间像素方差。再进一步将经 $\boldsymbol{W}_\gamma$ 重新加权的特征映射的权值通过sigmoid函数映射到(0,1)范围，并通过阈值Gate函数进行门控，具体为

$$W = Gate\left(f_{\text{sigmoid}}\left(\boldsymbol{W}_\gamma\left(GN(\boldsymbol{X})\right)\right)\right) \tag{2}$$

式中，实验中阈值设置为0.5，将阈值以上的权重设置为1，得到信息权重 $\boldsymbol{W}_1$，将阈值以下的权重设置为0，得到非信息权重 $\boldsymbol{W}_2$。将原始特征 $\boldsymbol{X}$ 分别与 $\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 相乘，得到特征 $\boldsymbol{X}_1^{\omega}$ 与特征 $\boldsymbol{X}_2^{\omega}$，重建操作的目的是利用交叉重构运算，将信息丰富的特征与信息较少的特征相加，生成信息更丰富的特征。具体是将特征 $\boldsymbol{X}_1^{\omega}$ 切分为两个部分 $\boldsymbol{X}_{11}^{\omega}$ 和 $\boldsymbol{X}_{12}^{\omega}$，特征 $\boldsymbol{X}_2^{\omega}$ 也做同样操作得到 $\boldsymbol{X}_{11}^{\omega}$ 和 $\boldsymbol{X}_{12}^{\omega}$，将切分完的特征进行交叉重构得到 $\boldsymbol{X}^{\omega 1}$ 和 $\boldsymbol{X}^{\omega 2}$。最终，将交叉重构得到的特征 $\boldsymbol{X}^{\omega 1}$ 和 $\boldsymbol{X}^{\omega 2}$ 进行拼接，得到空间重组特征映射 $\boldsymbol{X}^{\omega}$，具体为

$$\begin{cases} \boldsymbol{X}_1^{\omega} = \boldsymbol{W}_1 \otimes \boldsymbol{X} \\ \boldsymbol{X}_2^{\omega} = \boldsymbol{W}_2 \otimes \boldsymbol{X} \\ \boldsymbol{X}_1^{\omega} = \boldsymbol{X}_{11}^{\omega} \oplus \boldsymbol{X}_{22}^{\omega} \\ \boldsymbol{X}_2^{\omega} = \boldsymbol{X}_{21}^{\omega} \oplus \boldsymbol{X}_{12}^{\omega} \\ \boldsymbol{X}_1^{\omega} \cup \boldsymbol{X}_2^{\omega} = \boldsymbol{X}^{\omega} \end{cases} \tag{3}$$

此外，通道重组卷积单元使用“分割—变换—消融”策略。分割操作通过对通道的维度按照分割比α进行分割，αC作为上级通道特征提取通道维度信息丰富的特征，$(1-\alpha)C$作为下级特征提取通道维度匮乏的特征，通过逐点卷积变换维度。消融模块通过将上下全局通道描述子$\boldsymbol{S}_1$、$\boldsymbol{S}_2$叠加在一起，并使用通道软注意运算生成特征重要性向量$\boldsymbol{\beta}_1$，$\boldsymbol{\beta}_2$。在特征重要性向量$\boldsymbol{\beta}_1$、$\boldsymbol{\beta}_2$的指导下，将上特征$\boldsymbol{Y}_1$与下特征$\boldsymbol{Y}_2$按通道方向合并，得到通道细化特征$\boldsymbol{Y}$，具体为

$$\boldsymbol{Y} = \boldsymbol{\beta}_1 \boldsymbol{Y}_1 + \boldsymbol{\beta}_2 \boldsymbol{Y}_2 \tag{4}$$

在人体骨架识别任务中，为了应对模型复杂度过高导致的特征冗余难题，本文对空间—通道重组卷积进行了重构并融入多粒度图卷积网络中，具体架构见2.2小节。

1.4 分解图表示

由于骨架图拓扑结构的单一性，一些研究成果致力于重新定义人体骨架图的层次结构。例如Huang等人(2020)提出部分级图卷积网络(part level graph convolutonal network, PL-GCN)以一种数据驱动的方式学习身体部位的划分方式。该结构提出两个部分级块，即部分关系(part releation, PR)块和部分注意力(part attention, PA)块，它们通过图池化操作和图解池操作两种可微分图池化操作(Ying等，2018)来实现。针对上述问题，Zhu等人(2023)提出多层时空激励网络(multi-level spatial temporal graph network, ML-STGNet)，在空间建模中，设计了多层次图卷积网络和空间数据驱动激励模块，分别将人体骨骼的学习解耦为一般图和单个图。多层时空激励网络利用关节级、部分级和身体级图形全面建模人体的层次关系。

此外，传统的骨架图缺乏远距离节点的连通性，并且现有的工作也未能生成有语义信息的邻接矩阵。针对上述挑战，Lee等人(2023)提出分层分解图卷积网络(hierarchically decomposed graph convolutional network, HD-GCN)，该架构采用一种新的层次分解图。具体来说，该方法将每个关节有效地分解为若干个集合，提取主要结构上的相邻和远边，并利用它们构建包含这些边缘的分层分解图。Trivedi和Sarvadevabhatla(2022)提出基于“部分流”的架构，该架构使用基于部分节点的流，避免以整体方式处理输入骨架。这种选择可以实现更丰富和专用的表示，特别是对于由一小部分局部关节(手、腿)主导的动作。

综上所述，虽然现有方法在骨架图的处理上使用“部分流”或“分层”操作，增强了语义关联较强节点之间的信息交互，但仍遗失了一部分关节信息。针对这一挑战，本文将原始骨架图的3种邻接矩阵与多粒度邻接矩阵堆叠，形成动态邻接矩阵，详细设计见2.4.1小节。

2 方 法

2.1 前置知识

定义1 人体骨架可表示为图拓扑结构。骨架图拓扑结构公式化为$\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$，其中$\boldsymbol{V} = \{v^1, v^2, \cdots, v^N\}$代表节点集合，$N$表示关节点的数目，$\boldsymbol{E}$代表边集。邻接矩阵$\boldsymbol{A} \in \mathbf{R}^{N \times N}$的值代表各个边的邻接情况，邻接矩阵元素$a_{ij}$取1或0的值，指示$a_i$和$a_j$节点是否相邻。

给定一个骨架序列，节点特征$\boldsymbol{X}$的表征过程为

$$\boldsymbol{X} = \{x_{t,n} | 1 \leqslant t \leqslant T, 1 \leqslant n \leqslant N; N, T \in Z\} \tag{5}$$

式中，$x_{t,n}$表示节点v^n在帧t处的节点特征。T为骨架序列的总帧数，N表示人体关节点的数目，N在实验中一般设置为25。

定义2 图卷积神经网络已成功用于对人类骨骼动作序列进行建模。在这些工作中，骨架图$\boldsymbol{G} = (N, T)$表示为具有N个关节和T帧的骨架序列图，表征了T帧内人体骨架的运动过程。

在该图中，顶点集可表征为$\boldsymbol{N} = \{V^{ti} | t = 1, \cdots, T;\ i = 1, \cdots, N\}$，包含骨架序列每一帧的所有关节。边集$\boldsymbol{T}^S$由两个子集组成。每个子集包含每个帧中两个骨架节点的连接，表示为$\boldsymbol{T}^S = \{V^{ti}V^{tj} | (i,j) \in N\}$，其中$N$表示人体中存在连接的关节集。第$t$帧处的第$i$个关节点的多粒度图卷积输出的更新规则为

$$f_{\text{out}}(V_{\text{ti}}) = \sum_{V_{\text{tj}} \in \boldsymbol{B}(V_{\text{ti}})} \frac{1}{Z\left(l\left(V_{\text{tj}}\right)\right)} f_{\text{in}}\left(V_{\text{tj}}\right) \boldsymbol{\omega}\left[\left(l\left(V_{\text{tj}}\right)\right)\right] \tag{6}$$

式中，f_{in}和f_{out}表示特征图的输入输出。$\boldsymbol{B}(V_{\text{ti}}) =$

$\{V_{ti} | d(V_{ti}, V_{tj}) \leqslant 1\}$表示$V_{ti}$节点小于等于1的邻居集。$l(\cdot)$是为$\boldsymbol{B}(V_{ti})$的每个顶点分配从距离为1到$K$的标签。

根据时空图卷积网络(Yan等,2018)中提出的空间配置划分策略,K设置为3,即将$\boldsymbol{B}(V_{ti})$分为3个子集:顶点本身、向心子集和离心子集,$\omega(\cdot)$是1×1卷积的可学习权重,Z为归一化函数。

2.2 动态多粒度图卷积网络总体架构

提出的动态多粒度图卷积网络(dynamic multi-granularity graph convolution network, DMG- GCN)采用CTR-GCN作为基准模型,共有10层堆叠的基本图卷积块,如图1所示。每个基本块的输出通道为64,64,64,64,128,128,128,256,256,256。每个图卷积块包含空间多粒度卷积模块、空间—通道重组卷积模块以及改进的多尺度时间卷积模块。经过详细的实验验证(消融实验见3.4.3小节),空间—通道重组卷积的最佳优化方式如下:首先,利用空间重构单元修正空间多粒度图卷积模块;其次,利用通道重构单元修正多尺度时间卷积模块。

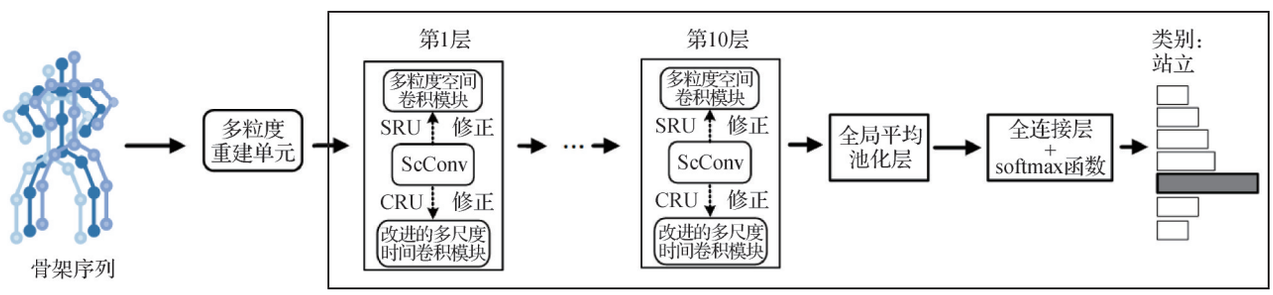


图1 动态多粒度图卷网络总体架构

Fig. 1 The overall architecture of dynamic multi-granularity graph convolution network

动态多粒度图卷积网络的前向传播过程如下:首先对输入的骨架序列经过多粒度重建单元的预处理,得到动态邻接矩阵,接着将动态邻接矩阵送入多粒度图卷积网络中。遍历10个基本图卷积块学习到不同粒度的特征后,输入到全局平均池化层压缩特征映射。最终由softmax函数与全连接层对样本进行评分。

2.3 多粒度图结构

当使用时空图卷积网络(Yan等,2018)预定义的人体骨架结构和3种基本空间分配划分策略时,人体的非物理连接节点信息受限。受到多层时空激励网络架构(Zhu等,2023)中分层图结构的启发,本文设计了多粒度图结构,如图2所示,由3种粒度的骨架结构来捕获不同层级的信息。第1种方法将人体部位划分为6个部分,对应人体的头部、躯干、手臂、指、腿部、脚。此为第1种分层方式,称做“全身关节—粗粒度”,如图2(a)所示。第2种分层方式则注重于上半部分关节细粒度划分,与粗粒度表示的

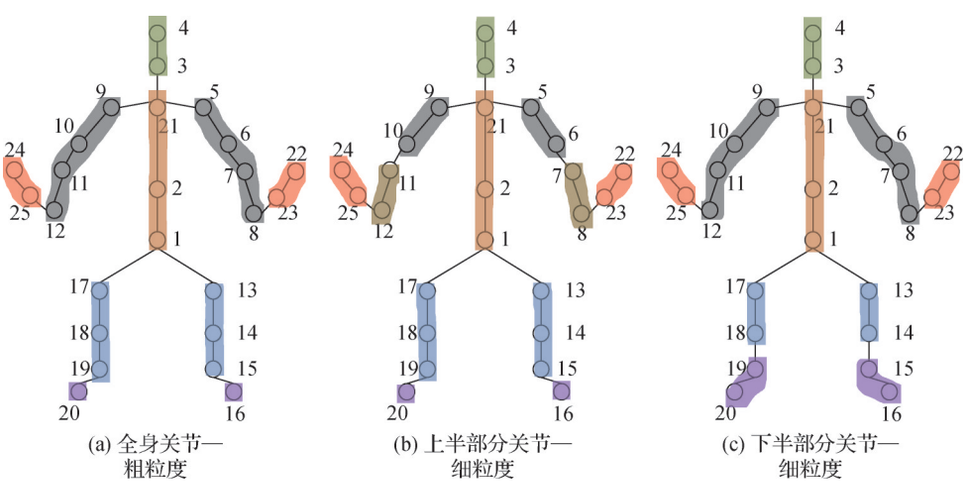


图2 多粒度图结构的定义

Fig. 2 The definition of a multi-granularity graph structure ((a) coarse granularity representation of the human body; (b) fine granularity representation of the upper half of the human body; (c) fine granularity representation of the lower half of the human body)

区别在于对手臂划分为大臂和小臂，设计意图在于更关注人体上半部分的动作，称做“上半部分关节—细粒度”，如图2(b)所示。第3种为下半部分关节细粒度划分，与粗粒度图表示区别在于，将腿部细化为大腿和小腿，对应更关注人体下半部分的动作表示，称做“下半部分关节—细粒度”，如图2(c)所示。

2.4 基于多粒度图结构的空间建模

2.4.1 多粒度重建单元

针对远距离节点信息交互不足以及“分层”或“部分流”操作导致骨架节点信息遗失的挑战，本文提出多粒度重建单元，结构如图3所示。该模块的设计思路如下：由于基线模型中通道拓扑优化图卷积网络(Chen等，2021a)继续延用时空图卷积网络(Yan等，2018)的思路，将骨架图经过预处理分别得到3个矩阵，对应自身矩阵$\boldsymbol{A}_1$、向心矩阵$\boldsymbol{A}_2$和离心邻接矩阵$\boldsymbol{A}_3$。结合本文提出的由多粒度图结构生成的多粒度邻接矩阵$\boldsymbol{A}_4$，将4个邻接矩阵进行堆叠操作，生成动态邻接阶矩阵$\boldsymbol{A}'$，具体为

$$\boldsymbol{A}' \leftarrow \boldsymbol{A}_1 \cup \boldsymbol{A}_2 \cup \boldsymbol{A}_3 \cup \boldsymbol{A}_4 \tag{7}$$

以动态邻接矩阵作为迭代更新的基本单元，扩大感受野的同时保留了原始节点的邻接信息。

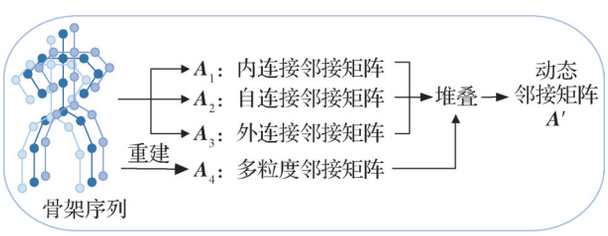


图3 多粒度重建单元结构

Fig. 3 The structure of multi-granularity reconstruction unit

2.4.2 多粒度图卷积单元

多粒度图卷积单元(multi-granularity graph convolution union)作为多粒度空间卷积的基本单元，架构如图4(a)所示。首先输入特征$\boldsymbol{S}^{T\times N\times C}$分别输入3个分支，其中两个分支送入1×1卷积变换维度，作为特征变化的主路，而第3个支的作用是保留原始特征。由于多粒度图卷积模块为空间建模的基本单元，目的在于提取空间维度特征，所以采用时间维度池化消除时间维度T，避免不同时间帧对空间特征提取带来的影响。接下来将得到的输出送入成对相减函数，获取到两个特征的差异性。将相减后的特征作为下一阶段的输入，目的是为了增加网络的鲁棒性。输出送入Tanh激活函数得到输出特征$\boldsymbol{Q}$。为了与原始特征进行加权求和，需要经过1×1的卷积变换维度后进行加权求和，具体为

$$\boldsymbol{R} \leftarrow \boldsymbol{A}' + \boldsymbol{\alpha Q} \tag{8}$$

式中，$\boldsymbol{A}'$为由多粒度重建单元生成的动态邻接矩阵，$\boldsymbol{\alpha}$为可训练参数。由动态邻接阶矩阵$\boldsymbol{A}'$进行修正后的特征向量$\boldsymbol{R}$包含了通道级别、多粒度特征的融合特征$\boldsymbol{R}$再与原始特征$\boldsymbol{S}$进行残差操作，得到最终输出特征$\boldsymbol{S}_0^{T\times N\times C}$。

2.4.3 多粒度空间卷积模块

多粒度空间卷积模块(multi-granularity spatial convolution module)的主体部分由多粒度图卷积单元构成，具体结构如图4(b)所示。输入特征有5个分支，其中4个分支输入多粒度图卷积单元，分别对应自连接邻接矩阵$\boldsymbol{A}_1$、向心邻接矩阵$\boldsymbol{A}_2$、离心邻接矩阵$\boldsymbol{A}_3$和多粒度邻接矩阵$\boldsymbol{A}_4$。将4个图卷积单元的输入相加，经过一个批量归一化层，加速网络收敛速度，激活函数选择ReLU(rectified linear unit)，此时的输出再与第5个分支也就是原始空间特征相加，同样起到残差的作用。由于多粒度图卷积单元通过通道压缩和聚合获取通道维度的重要特征，通道维度存在的特征冗余较少，但在空间维度仍然存在冗余，故在该模块的末尾利用空间重组卷积模块降低空间维度的冗余。

2.5 改进的多尺度时间卷积模块

所提出的改进的多尺度时间卷积单元模块相较于基线模型CTR-GCN中的多尺度时间卷积模块，主要改进在于结构优化和特征提取效率的提升，如图4(c)所示。该架构借鉴了通道拓扑优化图卷积网络的时间建模思想，通过删去两个膨胀分支，减少网络冗余，同时最大程度保留多尺度特征，避免了过多分支导致的网络难以训练的问题。此外，该模块通过引入不同卷积核大小和膨胀率的卷积模块，学习不同感受野的特征，增强了模型对时空信息的捕捉能力。在特征输出之前，融入通道重组卷积模块进一步提升了特征提取的效率并降低通道维度的冗余。

3 实 验

在3个流行数据集NTU-RGB+D、NTU-RGB+D 120和Northwestern-UCLA上评估提出的动态多粒度图卷积网络。消融实验的设计是为了验证动态多粒度

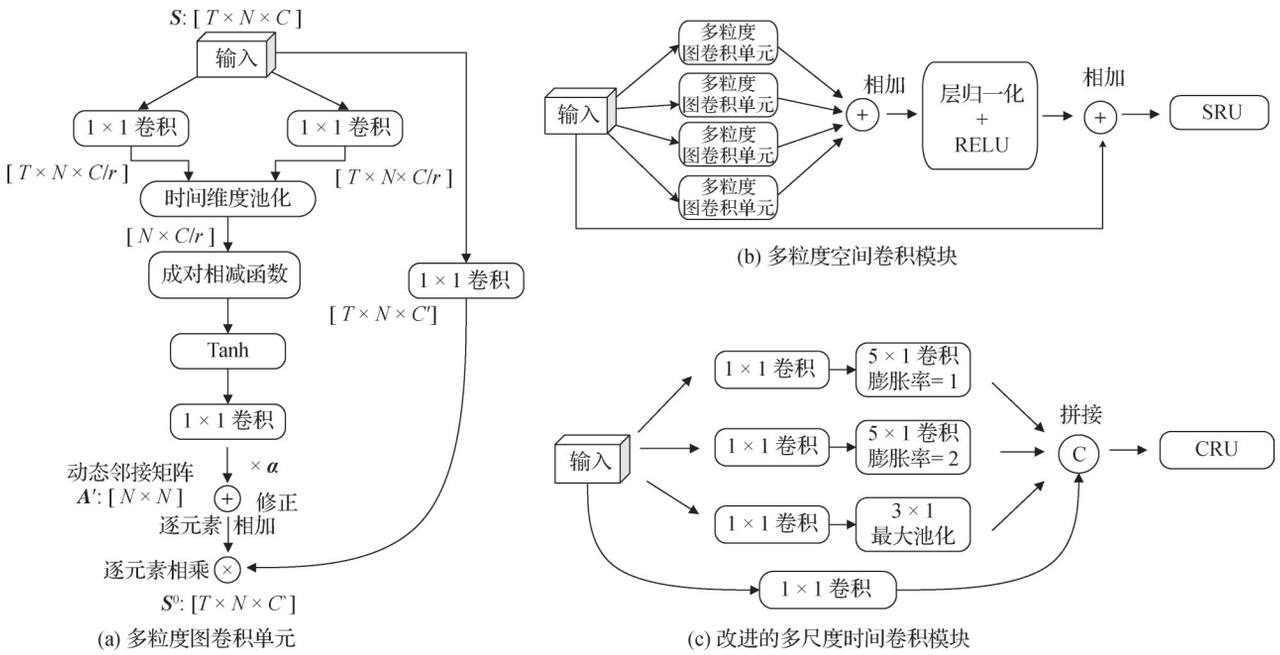


图4　空间建模与时间建模模块图

Fig. 4　Module diagram of spatial and temporal modeling ((a) multi-granularity graph convolution unit; (b) multi-granularity spatial convolution module; (c) improved multi scale temporal convolution module)

图卷积网络中所提出每个模块的有效性。本文分别与最新工作进行精度、参数量和浮点数计算的对比。

消融实验首先验证在多粒度图卷积度单元中使用不同粒度的图结构对于实验精度的影响，其次验证空间—通道重组卷积不同模块在空间、时间建模后的不同组合方式对实验精度的影响。结合最佳粒度图结构与空间—通道重组卷积的最优组合方式，模型达到了最佳性能。

3.1　数据集

3.1.1　NTU-RGB+D

NTU-RGB+D是用于骨骼动作识别的大型数据集，包含RGB图像、深度图像和骨骼关节信息，由新加坡南洋理工大学收集，包含56 880个骨骼动作样本，由40个不同参与者执行，包含60个不同的动作类别，涵盖各种日常生活中的动作，如走路、跑步和握手等。NTU-RGB+D使用两个评价指标，分别为跨主题(X-Sub)和跨视图(X-View)。跨主题(X-Sub)中，40个受试者的测试行为中有20个用于模型的训练，其余20个用于验证。跨视图(X-View)中，3个摄像机视图中的两个视图数据用于训练，另一个用于验证。

3.1.2　NTU-RGB+D 120

NTU-RGB+D 120比NTU-RGB+D更大，包含更多的骨骼动作样本，相较NTU-RGB+D 60增加57 367个新动作样本，包含120多个类别的114 480个骨骼动作样本，由106个不同受试者进行。NTU-RGB+D 120使用两个评价指标，分别为跨主题(X-Sub)与交叉设置(X-Set)。跨主题(X-Sub)中，106个受试者的动作中有53个用于训练，其余53个用于验证。交叉设置(X-Set)的32个设置中，设置id为偶数的数据用于训练，id为奇数的数据用于验证。

3.1.3　Northwestern-UCLA

Northwestern-UCLA数据集于美国加州大学洛杉矶分校收集。该数据集包含由3个Kinect摄像头同时捕获的RGB、深度和人体骨骼数据。其中包括10个动作类别：一只手捡起、两只手捡起、扔垃圾、四处走动、坐下、站起来、穿、脱、扔和搬运。该数据集包含10个类别的1 494个视频片段。每个动作都是通过3个不同视角的Kinect摄像头捕捉，并由10个对象执行。3个摄像机视图中的两个用于训练，另一个用于验证。

3.2　实验设置

本实验均在PyTorch深度学习框架上进行验证，使用的加速器均为NVIDIA GeForce RTX 3090 GPU。基准模型采用CTR-GCN(channel-wise tpology refinement graph convolution network)(Chen等，2021a)架构，模型使用动量为0.9的SGD(stochastic gradient descent)进行训练，在NTU-RGB+D、NTU-RGB+D 120数据集上均训练65个轮次，初始学习率

同为0.1,学习率衰减率为0.1。NTU-RGB+D、NTU-RGB+D 120学习率衰减发生的步数在{35,55}、{35,55}。对于前5次,使用热身策略使训练更稳定。对于NTU-RGB+D和NTU-RGB+D 120数据集,权值衰减设为0.000 4,批次大小为64,每个样本大小也调整为64帧。在多流融合与最新方法相比,本文首先以双流自适应图卷积网络(Shi等,2019a)的传统四流融合(4s)进行实验,四流即关节流、骨骼流、关节速度流、骨骼速度流。此外,结合提出的多粒度图结构,将3种多粒度图结构对应的关节流、骨骼流进行融合,进而提出全新的六流融合(6s)方式。

3.3 与最新方法的对比

本文统计了自时空图卷积网络(Yan等,2018)以来的经典模型在3个数据集上的精度,与所提出的动态多粒度图卷积网络进行对比,其他模型数据均在本地进行测试,实验结果如表1所示。

表1 本文方法在NTU-RGB+D 60、NTU-RGB+D 120和Northwestern-UCLA数据集上与最新方法的精度、参数量、浮点数运算对比

Table 1 Comparison of accuracy, parameter quantity, and floating-point operation between the proposed method and the latest method on the NTU-RGB+D 60, NTU-RGB+D 120, Northwestern UCLA datasets

方法	来源	NTU-RGB+D 60 精度/%		NTU-RGB+D 120 精度/%		North-western-UCLA/%	参数量/M	浮点数运算/G
		X-Sub	X-View	X-Sub	X-Set			
ST-GCN(Yan等,2018)	AAAI18	81.5	88.3	70.7	73.2	–	3.1	16.3
PL-GCN(Huang等,2020)	AAAI20	89.2	95.0	–	–	–	3.5	15.2
MS-G3D (Liu等,2020)	CVPR20	91.5	96.2	86.9	88.4	–	2.8	48.80
CTR-GCN(Chen等,2021a)	ICCV21	92.4	96.8	88.9	90.6	96.5	5.80	7.90
MST-GCN(Chen等,2021b)	AAAI21	91.5	96.6	87.5	88.8	–	12.0	–
PSUMNET(Trivedi和Sarvadevabhatla,2022)	ECCV22	92.9	96.7	89.4	90.6	96.7	2.8	2.70
HD-GCN(Lee等,2023)	ICCV2023	92.9	96.8	89.5	90.8	96.9	10.08	9.60
ML-STGNet(Zhu等,2023)	TIP23	91.9	96.2	88.6	90.0	96.8	11.52	–
Efficient-B4(Song等,2023)	TPAMI23	92.1	96.1	88.7	88.9	96.8	2.0	15.20
动态多粒度图卷积网络(4s)	JIG25	92.8	96.9	89.4	90.8	97.0	7.1	6.8
动态多粒度图卷积网络(6s)	JIG25	**93.0**	**97.0**	**89.6**	**91.0**	**97.2**	10.7	10.2

注:加粗字体表示各列最优结果。"–"表示不存在该数据。AAAI: AAAI Conference on Artificial Intelligence; CVPR: IEEE/CVF Conference on Computer Vision and Pattern Recognition; ICCV: IEEE/CVF International Conference on Computer Vision; ECCV: European Conference on Computer Vision; TIP: IEEE Transactions on Image Processing; TPAMI: IEEE Transactions on Pattern Analysis and Machine Intelligence; JIG: Journal of Image and Graphics。

所提出的架构相较于基线模型在3个数据集上最高分别有0.6%、0.7%和0.7%的提升。此外,本文提出的动态多粒度图卷积网络架构与ML-STGNet(multilevel spatial-temporal excited graph network)(Zhu等,2023)、Efficient-B4(Song等,2023)、HD-GCN(hierarchically decomposed graph convolutional networks)(Lee等,2023)在精度与参数量、浮点数运算的对比中,也能获得较有竞争力的表现。

3.4 消融实验

在消融实验的设置中,首先将致力于探究多粒度图卷积单元与通道重组卷积模块之间的相融性。在3.4.1小节中,通过可视化的方法证明动态图卷积网络以及动态邻接矩阵的有效性,在3.4.2小节中,将探究在多粒度图卷积单元中不同粒度图结构对实验带来的影响;在3.4.3小节中,将探究通道重组卷积在时空建模的最优组合方式。

为探究所提出的多粒度图结构与通道重组卷积模块之间的相融性,以CTR-GCN为基准架构,分别加入多粒度图结构与通道重组卷积模块,最终融合两个模块得到所提出的动态多粒度图卷积网络架构。实验在NTU-RGB+D的X-SUB评价标准下测试模型的关节流的结果,消融实验结果如表2所示。结果表明所提多粒度图结构与通道重组卷积模块具有相融性。

3.4.1 所提方法与基线模型的可视化比较

为验证所提动态多粒度图卷积网络的有效性,

表 2　在 NTU-RGB+D 数据集上 DMG-GCN 架构模块消融的关节流精度对比

Table 2　Comparison of joint flow accuracy of different module ablation in DMG-GCN architecture on the NTU-RGB+D dataset

方法	X-Sub 精度/%	参数量/M
基线模型	90.1	1.46
+多粒度图结构	90.2	1.68
+空间—通道重组卷积	90.5	1.62
+多粒度图结构+空间—通道重组卷积	**90.7**	**1.78**

注：加粗字体为模型最优组合。

实验对比了所提方法与基线模型在 NTU-RGB+D 数据集的准确率，并挑选 15 个具有代表性的动作类别进行准确率的可视化分析，结果如图 5 所示。可见，动态多粒度图卷积网络在大多数动作类别上的准确率都明显高于基线模型，在捕捉动作细节和识别复杂动作方面具有显著优势。对 2.4.1 小节中提到的动态邻接矩阵与基本邻接矩阵进行可视化对比，结果如图 6 所示。在基本邻接矩阵 $\boldsymbol{A}$ 在“踢东西”这一行为中学习到的节点之间的联系，如图 6(a)所示，颜色越深对应骨架节点之间联系越密切，图中用红色框标出的部分为权重较大的值，框选部分从左至右分别为左肘和左腕、右肘和右腕、右脚踝和右脚、左脚踝和左脚等存在物理连接的节点。动态邻接矩阵 $\boldsymbol{A}'$，除了学习到以上 4 种节点信息关联外，还包含同时学习到了左脚踝、左膝盖、左脚与右肩、右肘的信息交互；右膝盖、右脚踝；右脚与左肩、左肘等非节点连接节点的信息交互，如图 6(b)所示。现实生活中在进行“踢东西”这一行为时，左手和右脚、右手和左脚也存在相互协调的作用。

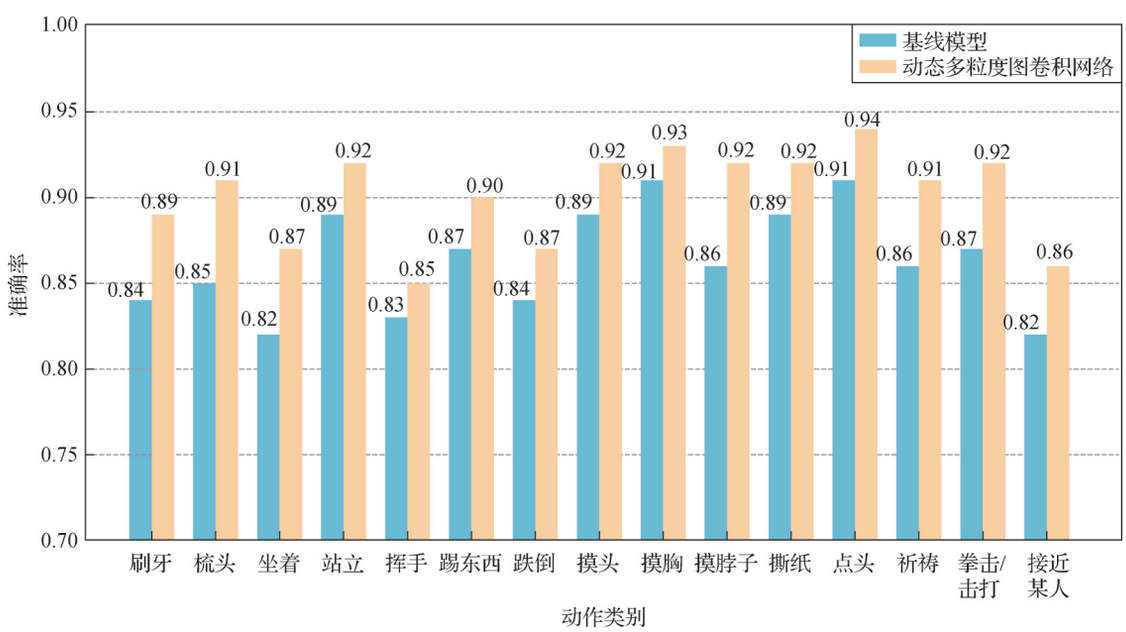
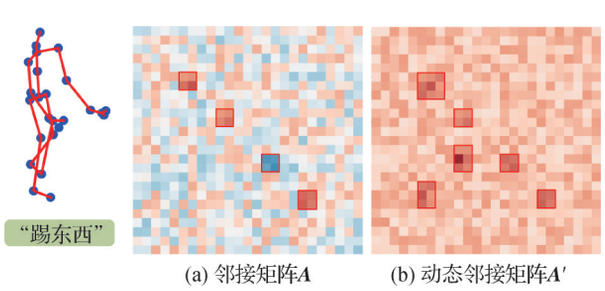


图 5　所提方法在 NTU-RGB+D 60 数据集上与基线模型的精度对比

Fig. 5　The comparison of accuracy between the proposed method and the baseline model on the NTU-RGB+D 60 dataset



(a) 邻接矩阵$\boldsymbol{A}$　(b) 动态邻接矩阵$\boldsymbol{A}'$

图 6　动态邻接矩阵的可视化比较

Fig. 6　Visual comparison of dynamic adjacency matrix ((a) the adjacency matrix $\boldsymbol{A}$; (b) dynamic adjacency matrix $\boldsymbol{A}'$)

3.4.2　多粒度图卷积单元

为探究不同粒度的图结构生成的动态邻接矩阵表征能力强弱，本文设计了消融实验以论证不同粒度的图结构对学习非自然和语义信息密切的特征能力。构建多粒度重建单元时，由于多粒度矩阵 $\boldsymbol{A}_4$ 具有 3 种不同粒度的骨架图结构，生成的动态邻接矩阵也各不相同。通过将多粒度邻接矩阵 $\boldsymbol{A}_4$ 分别替换为全身关节—粗粒度、上半部分关节—细粒度、下半部分关节—细粒度获得细粒度特征表示。不同粒

度骨架图在NTU-RGB+D数据集关节流的精度和参数量如表3所示。实验证明将下半部分关节进行细粒度设置后性能最佳。

表3　在NTU-RGB+D 60数据集上不同粒度骨架图的关节流精度对比

Table 3　Comparison of joint flow accuracy of skeleton maps with different granularities on the NTU-RGB+D 60 dataset

方法	X-Sub精度/%	X-View精度/%	参数量/M
基线模型	90.1	94.6	1.46
+全身关节—粗粒度	90.2	94.9	1.72
+上半部分—细粒度	90.1	94.8	1.64
+下半部分—细粒度	**90.3**	**95.1**	**1.68**

注：加粗字体为最优粒度图结构。

3.4.3　空间—通道重组卷积

针对空间—通道重组卷积在空间、时间建模的不同组合方式进行了详细的消融实验。实验设计了4种组合方式，如图7所示。前两种方式将完整的空间—通道重组卷积模块直接加入。第3种方式在空间建模后添加空间重组卷积，在时间建模后添加通道重组卷积。第4种方式则在空间建模后添加通道重组卷积，在时间建模后添加空间重组卷积模块。

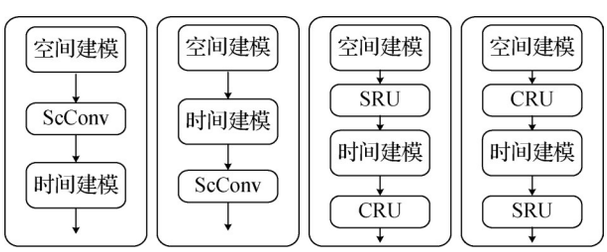


图7　ScConv模块在时空建模的不同组合方式

Fig. 7　Different combinations of ScConv modules between spatio and temporal modeling

最终，统计4种组合方式在NTU-RGB+D数据集的精度和参数量，如表4所示。由表中实验数据可知，空间建模后使用空间重组卷积以及时间建模后使用通道重组卷积效果最佳。

表4　在NTU-RGB+D 60数据集上空间—通道重组卷积的不同组合方式的关节流精度对比

Table 4　Comparison of joint flow accuracy of different combinations of ScConv on the NTU-RGB+D 60 dataset

方法	X-Sub精度/%	X-View精度/%	参数量/M
基线模型	90.1	94.6	1.46
空间建模+ScConv	90.2	94.6	1.74
时间建模+ScConv	90.2	94.5	1.74
空间建模+SRU，时间建模+CRU	**90.5**	**95.0**	**1.62**
空间建模+CRU，时间建模+SRU	90.3	94.8	1.62

注：加粗字体为ScConv的最优组合。

4　结　论

本文提出一种新颖的基于动态多粒度图卷积网络骨架行为识别方法，定义了一组人体骨架多粒度图结构，以捕获更具语义信息的特征，并引入了多粒度重建单元，通过构建动态邻接矩阵在保留原始节点信息的同时学习非自然节点交互信息。在时空建模中，为了缓解空间—通道维度变换所产生的特征冗余，对特定的空间—通道重组卷积模块进行重构。将3种不同粒度图结构的骨骼流、关节流进行融合，形成了一种独特的六流融合方式，以学习多尺度特征表示。实验表明，所提出的方法在NTU-RGB+D、NTURGB+D 120和Northwestern-UCLA等骨架数据集上优于对比的主流方法。未来的研究将专注于将多粒度图结构融入对比学习框架中，并迁移到相关多模态数据集中验证其有效性。

参考文献（References）

Chen Y X, Zhang Z Q, Yuan C F, Li B, Deng Y and Hu W M. 2021a. Channel-wise topology refinement graph convolution for skeleton-based action recognition//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 13339-13348 [DOI: 10.1109/ICCV48922.2021.01311]

Chen Z, Li S C, Yang B, Li Q H and Liu H. 2021b. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition//Proceedings of 2021 AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press: 1113-1122 [DOI: 10.1609/aaai.v35i2.16197]

Hao Y Z, Zheng Q H, Chen Y P and Yan C X. 2016. Recognition of abnormal behavior based on data of public opinion on the web. Journal of Computer Research and Development, 53(3): 611-620 (郝亚洲, 郑庆华, 陈艳平, 闫彩霞. 2016. 面向网络舆情数据的异常行为识别. 计算机研究与发展, 53(3): 611-620) [DOI: 10.7544/issn1000-1239.2016.20150746]

Huang L J, Huang Y, Ouyang W L and Wang L. 2020. Part-level graph convolutional network for skeleton-based action recognition//Proceedings of 2020 AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press: 11045-11052 [DOI: 10.1609/aaai.v34i07.6759]

Ma S, Liu J W, Zuo X. Survey on graph neural network. Journal of Computer Research and Development, 2022, 59(1): 47-80 (马帅, 刘建伟, 左信. 图神经网络综述. 计算机研究与发展, 2022, 59(1): 47-80) [DOI: 10.7544/issn1000-1239.20201055]

Lee J, Lee M, Lee D and Lee S. 2023. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 10410-10419 [DOI: 10.1109/ICCV51070.2023.00958]

Li J F, Wen Y and He L H. 2023. ScConv: spatial and channel reconstruction convolution for feature redundancy//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 6153-6162 [DOI: 10.1109/CVPR52729.2023.00596]

Li S F, Gao S B and Zhang Y Y. 2023. Pose-guided instance-aware learning for driver distraction recognition. Journal of Image and Graphics, 28(11): 3550-3561 (李少凡, 高尚兵, 张莹莹. 2023. 用于驾驶员分心行为识别的姿态引导实例感知学习. 中国图象图形学报, 28(11): 3550-3561) [DOI: 10.11834/jig.220835]

Liu J, Shahroudy A, Perez M, Wang G, Duan L Y and Kot A C. 2020. NTU RGB+D 120: a large-scale benchmark for 3D human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(10): 2684-2701. [DOI: 10.1109/TPAMI.2019.2916873]

Liu Y N, Zhang H, Xu D and He K J. 2022. Graph transformer network with temporal kernel attention for skeleton-based action recognition. Knowledge-Based Systems, 240: #108146 [DOI: 10.1016/j.knosys.2022.108146]

Liu Z Y, Zhang H W, Chen Z H, Wang Z Y and Ouyang W L. 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 140-149 [DOI: 10.1109/CVPR42600.2020.00022]

Lu J, Li X F, Zhao B and Zhou J. 2023. A review of skeleton-based human action recognition. Journal of Image and Graphics, 28(12): 3651-3669 (卢健, 李萱峰, 赵博, 周健. 2023. 骨骼信息的人体行为识别综述. 中国图象图形学报, 28(12): 3651-3669) [DOI: 10.11834/jig.230046]

Shahroudy A, Liu J, Ng T T and Wang G. 2016. NTU RGB+D: a large scale dataset for 3D human activity analysis//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 1010-1019 [DOI: 10.1109/CVPR.2016.115]

Shi L, Zhang Y F, Cheng J and Lu H Q. 2019a. Two-stream adaptive graph convolutional networks for skeleton-based action recognition//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 12018-12027 [DOI: 10.1109/CVPR.2019.01230]

Simonyan K and Zisserman A. 2014. Two-stream convolutional networks for action recognition in videos//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 568-576

Song Y F, Zhang Z, Shan C F and Wang L. 2023. Constructing stronger and faster baselines for skeleton-based action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2): 1474-1488 [DOI: 10.1109/TPAMI.2022.3157033]

Trivedi N and Sarvadevabhatla R K. 2022. PSUMNet: unified modality part streams are all you need for efficient pose-based action recognition//Proceedings of 2022 European Conference on Computer Vision. Tel Aviv, Israel: Springer: 211-227 [DOI: 10.1007/978-3-031-25072-9_14]

Wang J, Nie X H, Xia Y, Wu Y and Zhu S C. 2014. Cross-view action modeling, learning, and recognition//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 2649-2656 [DOI: 10.1109/CVPR.2014.339]

Wang S C, Huang Q, Zhang Y F, Li X, Nie Y Q and Luo G C. 2022. Review of action recognition based on multimodal data. Journal of Image and Graphics, 27(11): 3139-3159 (王帅琛, 黄倩, 张云飞, 李兴, 聂云清, 雒国萃. 2022. 多模态数据的行为识别综述. 中国图象图形学报, 27(11): 3139-3159) [DOI: 10.11834/jig.210786]

Wei X, Yu R X and Sun J. 2020. View-GCN: view-based graph convolutional network for 3D shape analysis//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1847-1856 [DOI: 10.1109/CVPR42600.2020.00192]

Xia L, Chen C C and Aggarwal J K. 2012. View invariant human action recognition using histograms of 3D joints//Proceedings of 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Providence, USA: IEEE: 20-27 [DOI: 10.1109/CVPRW.2012.6239233]

Yan S J, Xiong Y J and Lin D H. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition//Proceedings of the 32nd AAAI Conference on Artificial Intelligence and 30th Innovative Applications of Artificial Intelligence Conference and 8th AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans, USA: AAAI Press: 7444-7452 [DOI: 10.1609/aaai.v32i1.12328]

Ying R, You J X, Morris C, Ren X, Hamilton W L and Leskovec J. 2018. Hierarchical graph representation learning with differentiable pooling//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc.: 4805-4815

You K J, Hou Z J, Liang J Z, Zhong Z K and Shi H Y. 2024. Point

cloud human behavior recognition based on coordinate transformation and spatiotemporal information injection. Journal of Image and Graphics, 29(4): 1056-1069 (尤凯军, 侯振杰, 梁久祯, 钟卓锟, 施海勇. 2024. 结合坐标转换和时空信息注入的点云人体行为识别. 中国图象图形学报, 29(4): 1056-1069) [DOI: 10.11834/jig.230215]

Zhou H, Zhan Y Z and Mao Q R. 2021. Video anomaly detection based on space-time fusion graph network learning. Journal of Computer Research and Development, 58(1): 48-59 (周航, 詹永照, 毛启容. 2021. 基于时空融合图网络学习的视频异常事件检测. 计算机研究与发展, 58(1): 48-59) [DOI: 10.7544/issn1000-1239202120200264]

Zhu Y S, Shuai H, Liu G C and Liu Q S. 2023. Multilevel spatial-temporal excited graph network for skeleton-based action recognition. IEEE Transactions on Image Processing, 32: 496-508 [DOI: 10.1109/TIP.2022.3230249]

作者简介

吴志泽,男,教授,主要研究方向为深度学习驱动的视频、图像处理与理解。E-mail: wuzz@hfuu.edu.cn

李腾,通信作者,男,教授,主要研究方向为人工智能与计算机视觉。E-mail:liteng@ahu.edu.cn

陈鑫,男,硕士研究生,主要研究方向为人体行为识别。E-mail:chenxin@stu.hfuu.edu.cn

徐童,男,教授,主要研究方向为数据挖掘与社交媒体分析。E-mail:tongxu@ustc.edu.cn

年福东,男,副教授,主要研究方向为人工智能与计算机视觉。E-mail:nianfd@hfuu.edu.cn

王晓峰,男,教授,主要研究方向为人工智能与计算机视觉。E-mail:xfwang@hfuu.edu.cn