第 38 卷　第 6 期
2023 年 12 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 38　No. 6
Dec. 2023

DOI:10.13364/j.issn.1672-6510.20230056
数字出版日期：2023-07-21；数字出版网址：http://kns.cnki.net/kcms2/detail/12.1355.N.20230720.1432.002.html

基于图卷积神经网络的人体骨架动作识别研究进展

杨巨成[1]，张泉钰[1]，王　波[2]，王　媛[1]，陈亚瑞[1]，赵婷婷[1]
（1. 天津科技大学人工智能学院，天津 300457；2. 思腾合力（天津）科技有限公司，天津 301799）

摘　要：基于人体骨架的动作识别是实现计算机视觉智能的重要分支。本文对基于图卷积神经网络的人体骨架动作识别技术进行研究并分析，对基于频谱图卷积和空域图卷积的研究现状进行综述，并从邻接矩阵和输入特征两个角度详述了图卷积模型在人体骨架动作识别领域的研究进展。此外，对现有的基于图卷积神经网络的人体骨架动作识别算法进行了分析比较，最后展望了图卷积神经网络在人体骨架动作识别领域的未来发展方向。

关键词：图理论；图神经网络；图卷积神经网络；基于骨架的动作识别；时空域融合
中图分类号：TP391.4　　**文献标志码**：A　　**文章编号**：1672-6510（2023）06-0001-11

Review of Human Skeleton Action Recognition Based on Graph Convolution Neural Network

YANG Jucheng[1]，ZHANG Quanyu[1]，WANG Bo[2]，WANG Yuan[1]，CHEN Yarui[1]，ZHAO Tingting[1]
（1. College of Artificial Intelligence，Tianjin University of Science & Technology，Tianjin 300457，China；
2. Sitonholy（Tianjin）Technology Co.，Ltd.，Tianjin 301799，China）

Abstract：Action recognition based on human skeleton is an important branch to achieve computer vision intelligence. In this article we first review the research on human skeleton action recognition based on graph convolutional neural networks and analyze the key techniques. Then，we outline the research progress of graph convolution approaches based on spectral convolution and space domain convolution，and detail the research progress of graph convolution models in the field of human skeleton action recognition from two perspectives of adjacency matrix and input features. Furthermore，we analyze and compare the existing algorithms for human skeleton action recognition based on graph convolution neural networks. Finally，we look forward to the future development direction of graph convolution neural networks in the field of human skeleton action recognition.

Key words：graph theory；graph neural network；graph convolutional neural network；skeleton-based action recognition；spatiotemporal fusion

人体动作识别旨在识别动作信息中人体动作的具体类别，是实现计算机视觉智能的重要分支。近年来，随着计算机科学技术的不断发展，人体动作识别[1-2]已经成为计算机视觉领域中备受关注的研究课题之一，其在各个领域具有很大的应用前景和商业价值，如视频监控[3]、视频检索[4]、人机交互等[5]。智能化视频监控是智慧工厂衍生出的新型管理模式，在节省人力、物力的前提下，可有效加强对人员的直观管理；加入人体动作识别技术的视频检索任务能够大幅提高建立索引的效率及搜索效果；人体动作识别作为人机交互的关键技术，可以将其应用到智能驾驶中，如疲劳检测等[6]。因此，开展人体动作识别的研究具有重大的理论意义和较高的应用价值。

随着视频采集传感器以及骨架提取算法的发展，用于人体动作识别任务的 2D/3D 人体骨架数据集先后被提出。骨架数据比其他模态数据（如 RGB 图像、

收稿日期：2023-03-15；修回日期：2023-06-11
基金项目：国家自然科学基金项目（61976156）
作者简介：杨巨成（1980—），男，湖北人，教授，jcyang@tust.edu.cn

光流等)包含更丰富的信息,它既包含关节点之间的空间信息,又包含关节点之间的时间信息。骨架数据本质上可以看作拓扑图,属于非欧氏空间数据,这一模态与人体动作更加密切,属于图结构。因此许多研究人员使用骨架数据进行人体动作识别任务。

在早期,卷积神经网络(convolutional neural network, CNN)[7-8]和递归神经网络(recurrent neural network, RNN)被应用在基于骨架的人体动作识别任务中,并取得了一定进展。这两种技术经常结合起来用于骨架动作识别,CNN 用于对原始 RGB 图像建模空间特征,RNN 的递归结构用于建模时间特征。与此同时,基于 CNN/RNN 的方法也面临一些问题,如模型大小、识别速度、遮挡情况以及视角变化等。近几年,研究人员把卷积的思想应用在图拓扑结构上,提出图卷积网络(graph convolutional networks, GCN)。GCN 与 CNN/RNN 有所不同,GCN 专注于处理图结构数据之间的依赖关系,擅于对非欧氏空间数据的建模。在基于骨架的人体动作识别任务中,GCN 具有在时间维度和空间维度同时建模的能力。此外,骨架数据不易受人体外观衣着、背景光照和拍摄视角变化等因素的影响,且能很好地避免噪声干扰,所以基于图卷积网络的人体骨架动作识别技术成了新的研究方向,受到学者们的高度关注。

目前已有许多学者对人体动作识别领域进行了综述,已有文献[9]侧重于从深度学习方法的角度进行总结,分为 CNN、RNN 和 GCN。本文聚焦图卷积神经网络在人体骨架动作识别领域的应用,首先对人体骨架动作识别任务和图卷积神经网络分别进行介绍,根据处理信号的不同(空域和频谱)介绍现有图卷积模型;对比分析不同的图卷积神经网络模型在基于骨架人体动作识别数据集上的表现;最后,文章针对现有基于图卷积神经网络的人体骨架动作识别技术,并对未来的发展方向进行展望。

1　基于骨架的人体动作识别

在人体动作识别领域,研究人员已经研究了多种模态,例如 RGB 图像、光流模态、骨架模态等。与其他模态相比,骨架模态是一种具有关节和骨骼的人体拓扑表示,当面对复杂环境以及涉及光照变化、视角变化、运动速度变化时,该模态更稳定。因此,基于骨架的人体动作识别是目前研究的重点。

随着深度学习的不断发展,使用卷积神经网络(CNN)、递归神经网络(RNN)和图卷积网络(GCN)进行基于骨架的人体动作识别的深度学习方法应运而生。

1.1　基于 CNN 的人体骨架动作识别

卷积神经网络(CNN)已经应用在基于骨架的人体动作识别任务中,主要采用两种方式进行人体动作识别。第一种是将原始关键点坐标数据信息(x、y、z坐标)看作图像的 RGB 通道,把原始骨架数据映射成三通道数据,使用 CNN 模型进行分类。Li 等[10]提出一种与数据集无关且平移尺度不变的映射方法,将三维骨架数据映射成彩色图像,平移尺度不变的映射方法能保证图片尺度具有不变性。为了充分提取骨架信息,Shi 等[11]提出双流 CNN 模型用于人体骨架动作识别,同时考虑关节点和骨骼边的特征,使用非对称卷积块减轻骨骼序列变形的负面影响。

第二种是在骨架数据上自定义动作信息和时间信息等特征,将其映射成彩色图像,使用 CNN 模型进行分类。Li 等[12]基于骨架数据自定义距离特征,并将其映射成彩色纹理图像,使用 AlexNet 网络进行训练。为了更好地引入时间信息,Wang 等[13]提出了关节轨迹图,通过颜色编码将关节轨迹的时空信息从主、侧、俯三视图转化为 3 个纹理图像进行动作识别。

在骨架人体动作识别任务中,基于 CNN 的方法通过简单映射将骨架信息表示为图像,导致只有卷积核内的相邻关节才会被学习共现特征,与所有关节相关的一些潜在相关性会被忽略。

1.2　基于 RNN 的人体骨架动作识别

递归神经网络(RNN)能够传递并提取长时间的序列信息,人体骨架动作数据作为时间序列数据,关节点坐标随着时间的推移而变化,因此 RNN 及其变体也被应用于人体骨架动作识别任务中。Shahroudy 等[14]将人体骨架分成 5 部分,把骨架中的关节点坐标排列为长向量,输入长短期记忆网络(long short term memory, LSTM)[15]进行动作识别。一些人体动作只需要几个关节的移动或者旋转就可以完成,基于此,Song 等[16]只考虑骨架信息中的重要帧和关键节点,提出端到端的基于 RNN 的时空注意力网络模型。但由于 RNN 模型本身的空间建模能力较弱,所以基于 RNN 的方法通常无法获得具有竞争性的结果。

1.3　基于 GCN 的人体骨架动作识别

近几年,图卷积技术得到显著发展,且人体骨架数据具有非欧氏数据特性,关节点代表人体关节,可以天然作为图结构的输入。相较于 CNN/RNN,图卷积网络(GCN)可以同时在空间域和时间域进行建

模，所以基于 GCN 的人体骨架动作识别技术受到越来越多学者的重视。

GCN 通过卷积不仅能够学习关节点自身以及相邻关节的特征，还能够捕获动作变化的时空特征用于动作识别任务。本文主要对基于 GCN 的人体骨架识别技术从频域和空域两个角度进行综述。

2　图卷积神经网络

研究人员成功地将卷积应用到图结构数据上，提出图卷积神经网络（GCN），其核心思想是利用边的信息对节点信息进行聚合，从而生成新的节点表示。图卷积神经网络[17]能够直接对具有图拓扑结构的数据进行学习，善于对非欧氏空间数据进行建模。随着 GCN 的发展，其卷积方式主要分为基于频域卷积和基于空域卷积两种方式。

2.1　基于频域的图卷积神经网络

对于经典卷积来说，卷积核的形状是固定的，这意味着其感受野中心节点必须要有固定数量的邻域才能使用卷积核，但图上节点的邻域节点是不确定的，除此之外，中心节点的邻域节点也是没有顺序的，所以经典卷积难以应用在图结构数据上。但在频域进行卷积时，只需要在每个频域分量上进行放大或者缩小，不需要考虑空域上存在的问题。基于频域的图卷积神经网络是利用图信号理论，使用傅里叶变换将图信号转换为频域信号，然后在频域上执行卷积操作，最后再利用傅里叶逆变换恢复到图信号所在的空域，基于频域的图卷积示意图如图 1 所示。基于频域的方法主要有第一代 GCN[18]、切比雪夫网络（Chebyshev network，ChebyNet）[19]、一阶近似 ChebyNet 等[20]。由于第一代 GCN 需要对拉普拉斯矩阵进行特征值分解，导致计算开销非常大，所以实际应用较少。针对第一代 GCN 的不足，David 等[19]提出利用切比雪夫多项式近似卷积核的 ChebyNet，ChebyNet 不需要对拉普拉斯矩阵进行分解，这使得计算效率大大提高。为了使 ChebyNet 有更好的局部连接性，Kipf 等[20]使用一阶切比雪夫公式进行简化，切比雪夫一阶近似为

$$\boldsymbol{H}^{(l+1)}=\sigma\left(\tilde{\boldsymbol{D}}^{-1/2}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-1/2}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}\right) \tag{1}$$

其中：$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}_N$ 是具有自连接无向图的邻接矩阵，$\boldsymbol{I}_N$ 是单位矩阵；$\tilde{\boldsymbol{D}}$ 是 $\tilde{\boldsymbol{A}}$ 的度矩阵，$\tilde{\boldsymbol{D}}_{ii}=\sum_j \tilde{\boldsymbol{A}}_{ij}$；$\boldsymbol{W}$ 代表第 l 层神经网络可训练的权重矩阵；$\sigma(\cdot)$ 代表非线性激活函数；$\boldsymbol{H}^{(l)}$ 表示第 l 层特征，对于输入层来说，$\boldsymbol{H}=\boldsymbol{X}$。

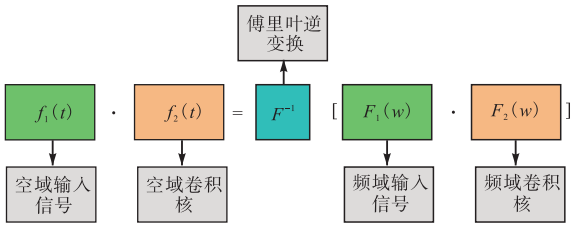

图 1　频域图卷积过程

Fig. 1　Process of spectral graph convolution

2.2　基于空域的图卷积神经网络

基于空域的图卷积神经网络不再依靠图谱理论[21]，其从空间角度出发，不断聚合中心节点的邻居节点信息以更新中心节点信息。该方法摆脱了对拉普拉斯矩阵的依赖，使图卷积网络能够应用于有向图。基于空域的图卷积神经网络可以看作 CNN 的拓展，其本质和 CNN 相同，都是对节点信息进行加权求和，从而达到卷积的目的。基于空域的图卷积操作大致可以分为 3 个步骤：第一，图中的每个节点将自身特征信息传递给邻居节点；第二，图中的每个节点能够聚合周围节点的特征信息，实现特征信息更新；第三，将更新后的节点做线性与非线性变换提升模型的表达能力。基于空域的方法有扩散卷积神经网络（diffusion convolution neural network，DCNN）[22]、图采样聚合模型（graph sample and aggregate，GraphSAGE）[23]、图注意卷积神经网络（graph attention networks，GAT）等[24]。DCNN 通过设定节点之间的转移概率控制节点之间的信息传递，GraphSAGE 对图中每个节点的邻居节点进行随机采样实现特征信息聚合，GAT 通过计算图中相邻节点之间的注意力系数实现更新节点特征信息。它们分别从信息传递、采样方式、注意力角度定义图拓扑结构的卷积方式。

2.3　两种方式对比

基于频域的图卷积方法不适用于有向图，这是因为傅里叶变换是以拉普拉斯矩阵的特征向量为基底，而拉普拉斯矩阵的对称性使其只对无向图有意义。第一代 GCN 由于需要对拉普拉斯矩阵进行特征分解，导致时间复杂度很高。因此，基于频域的图卷积方法存在灵活性不强、泛化能力弱、计算开销大等问题。基于空域的图卷积方法直接在空间上定义卷积操作，在降低了复杂度的同时还增强了泛化能力。人体动作本身具有的空间复杂性和时间差异性，促使计算复杂度低、泛化能力好的空域图卷积网络模型成为基于人体骨架动作识别任务的主流方法。

3　基于图卷积的人体骨架动作识别

基于骨架的动作识别任务的本质是对给定的一段动作信息进行分类，其不仅包含了空间信息，还包含时间信息，这增加了识别任务的难度。目前，基于图卷积神经网络的模型大多应用在基于骨架的人体动作识别数据集上。根据第 2 小节对图卷积神经网络的介绍，图卷积定义为

$$\boldsymbol{X}^{(l+1)}=\sigma\left(\hat{\boldsymbol{A}}\boldsymbol{X}^{(l)}\boldsymbol{W}\right) \tag{2}$$

由式(2)可知：卷积过程主要涉及邻接矩阵 $\boldsymbol{A}$、输入特征 $\boldsymbol{X}$，其中 $\boldsymbol{W}$ 为可训练参数，因此图卷积神经网络设计主要体现在前两个方面。在基于骨架的人体动作识别中，网络的设计变化如图 2 所示。

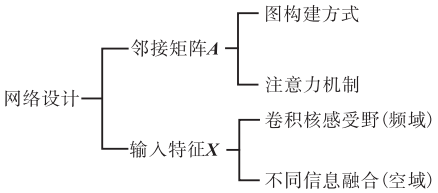

图 2　图卷积神经网络设计

Fig. 2　Design of graph convolution neural network

本节按照网络设计不同，从频域和空域两个角度对人体骨架动作识别任务进行讨论。

3.1　频域图卷积算法

目前，在基于骨架的人体动作识别任务中，ChebyNet[19]及一阶近似 ChebyNet[20]被广泛使用。本小节将按照图构建方式、注意力机制以及卷积核的感受野对算法进行分类讨论。

3.1.1　图构建方式

人体动作通常涉及多个关节，一些动作则需要两个相距较远的关节协作完成，但相距较远的关节在骨架图中并不构成边，无法进行信息传递，所以研究人员通常在骨架数据中手动添加一些边缓解此类问题。

通常情况下，GCN 的输入是原始骨架数据的邻接矩阵，但 Tang 等[25]提出的深度递进强化学习(deep progressive reinforcement learning，DPRL)网络为了获取距离较远关节之间的潜在关系，在物理连接的基础上手动添加边作为外部连接。图 3[25]中以“拍手”动作为例，其中实线代表物理连接，虚线代表外部连接。

Gao 等[26]提出了基于骨架的广义 GCN 网络，把节点的外部连接划分为强边和弱边，利用高阶切比雪夫多项式卷积核进一步增强了对动作的感知。以上两种构图方式是人工进行划分的，缺乏灵活性，未来可以尝试构造自适应邻接矩阵提高模型识别精度。

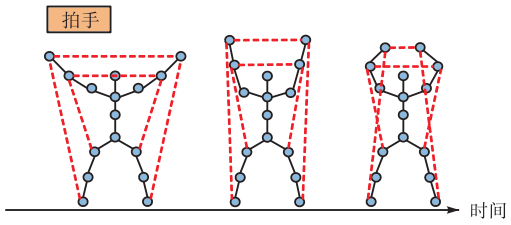

图 3　DPRL 网络骨架图构建

Fig. 3　Skeleton graph construction of DPRL network

3.1.2　注意力机制

注意力机制的应用十分广泛，其有效性已经在各类任务中得到体现。在骨架人体动作识别任务中，大多数算法考虑了骨架数据中的所有帧，并没有将注意力集中在重要的帧上。为了寻找一段动作信息中包含信息量最大的帧，Tang 等[25]提出的 DPRL 算法使用注意力机制在全部帧中挑选出了更重要的帧，丢弃信息量小的帧。此方法既能减少网络模型的参数量，又达到具有竞争性的识别效果。

3.1.3　卷积核的感受野

在图卷积的过程中，卷积核的感受野由其大小决定，不同的大小对最后的结果可能会产生一定影响。对于骨架数据来说，其邻接矩阵属于稀疏矩阵，使用比较小的卷积核的时候可能无法表示其特征，所以研究人员尝试扩大卷积核感受野提升模型精度。

Peng 等[27]最先提出把神经架构搜索(NAS)[28-29]与图卷积网络结合进行动作识别任务。一阶切比雪夫多项式不能很好地捕捉到高阶信息，为了捕获高阶节点关系，引入了高阶切比雪夫多项式扩大图卷积的感受野。然而，若各层图卷积都采用高阶多项式会使计算量成倍增加，同时并不确定是否每一层卷积都需要扩大感受野，故引入了 NAS 技术自动为不同层的卷积核生成图矩阵(embedding matrix，EM)，省去了人工调参的工作量。但随着卷积核感受野的扩大，图卷积神经网络出现过平滑现象，可以引入残差连接、添加非线性激活函数等方法缓解此类问题。

3.2　空域图卷积算法

Yan 等[30]创造性地提出时空图卷积网络模型(spatial temporal graph convolutional networks，ST-GCN)，第一个把基于空域图卷积模型应用在人体动作识别领域，通过对动态骨骼建模，把骨架数据以 2D 或 3D 网格的方式展现。在此之后，越来越多的研

究者在 ST-GCN 模型的基础上进行改进或提出新的基线，把骨架数据应用在图卷积网络中完成动作识别任务。本小节从图构建方式、注意力机制、不同信息融合这 3 个方面对改进模型进行分类讨论。

3.2.1　图构建方式

Yan 等[30]提出的时空图卷积网络模型（ST-GCN）是第一个基于空域图卷积的动态骨架建模方法。该方法首先利用 OpenPose[31]姿态估计算法对实验所用的数据集进行姿态估计，使人工设计的工作量大大减少。如图 4[30]所示，图上每一个节点都对应人体的关节，图中有两种类型的边，一种为符合关节自然连接的空间边，另一种为跨越连续时间步长连接相同关节的时间边。

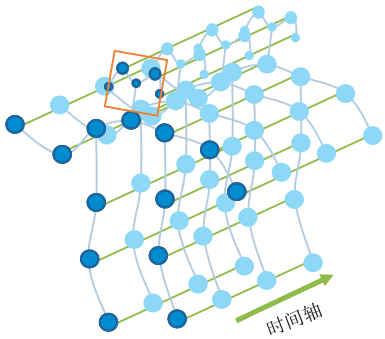

图 4　ST-GCN 骨架图构建

Fig. 4　Skeleton graph construction of ST-GCN

在该图的基础上，多个时空卷积层能够将图拓扑信息在空间和时间维度进行融合。ST-GCN 网络模型首先对输入矩阵在时间和空间维度进行归一化，接着进入 ST-GCN 模块，通过交替使用图卷积和时域卷积完成时空卷积，最后使用全局池化操作和全连接层对动作进行识别分类。图卷积学习空间中相邻节点的局部特征，时序卷积的作用是为关节叠加时序特征。此外，Yan 等[30]还提出了注意力机制，根据各个关节点在不同动作中的重要程度不同，将关节点赋予不同的权重。为了进一步提升模型性能，根据动作识别的特点，将图上节点划分为向心点、离心点和根节点 3 类，对此使用 3 个不同的卷积核对其进行卷积操作。该模型可以从骨架数据中“自主”学习空间和时间特征，使其具有很强的表达能力。

ST-GCN 网络模型也存在一些不足：

（1）ST-GCN 中使用的骨架图是预先人为定义好的，表示人体的物理结构。模型所用的卷积核只能提取局部特征，所以不能保证只通过学习相邻节点的信息，达到最优的识别结果。例如，“跑步”和“跳远”对两条腿之间的关系有较强依赖，但 ST-GCN 无法捕获两条腿之间的关系，因为在人为定义的骨架数据中它们相距很远。

（2）神经网络往往具有多个层级结构，且每个层级包含的语义信息不同。在 ST-GCN 模型中，所有图卷积层中图拓扑结构固定不变，这导致其无法学到多尺度信息。

（3）单一的骨架信息输入可能不利于网络模型对不同动作的感知，无法对具有细微运动差异的动作进行分类。

与 Yan 等[30]相比，Si 等[32]认为动作是由人体各个部分协调完成，例如跑步不仅需要双腿运动，还需要双臂摆动维持身体平衡，所以其在空域建模的时候把骨架图分成 K 个部分，分块学习人体结构的空间特征。此方法简化了关节点数量，提升了网络推理速度。Xiong 等[33]认为人类的动作可以简化为人类肢体运动，因此手臂和腿部包含丰富的动作特征信息，其手动将手臂和腿部构造出额外的边，使得在图卷积的过程中能够聚合更多的特征信息，避免了使用图卷积在进行长距离传递时信息丢失的问题。Chi 等[34]认为学习骨架图的内在拓扑信息十分重要，因此其提出编码器模块推理骨架图的内在拓扑信息。编码器的架构由用于空间建模的基于自注意力的图卷积模块和用于时间建模的多尺度时间卷积模块组成。与人工手动设置相比，此方法更加灵活，针对不同动作会推断出不同的内在拓扑信息，有利于提升动作识别的准确度。Shi 等[35]用有向无环图（图 5）表示骨架数据，方向由顶点与根节点之间的距离确定，通过引入骨骼方向，将骨骼的运动表示为同一骨骼在连续帧中的向量差，该方法进一步挖掘了骨骼、关节与动作识别之间的关系。

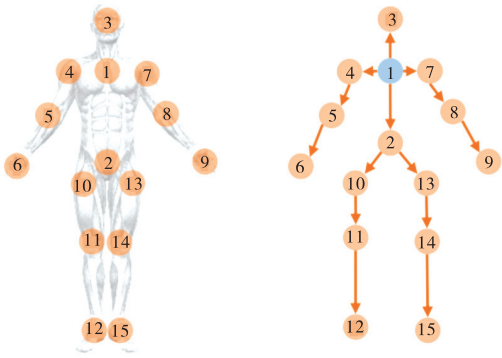

图 5　骨骼数据的有向无环图

Fig. 5　Directed acyclic graph of skeletal data

Li 等[36]更加关注不相邻关节点之间潜在的依赖

关系，提出了行为结构图卷积网络（action structural graph convolutional networks，AS-GCN），用一种由编码器和解码器组成的“动作连接”推理模块捕获节点之间的潜在联系，其提出“结构连接”模块获得节点长距离的连接信息。这种自适应学习的方式，能够捕获与动作行为有用的高阶信息，提高模型灵活性。此外，AS-GCN 不仅可以识别动作，还可以通过多任务学习进行姿势预测，在同一个框架下进行多任务学习有效提高了模型的效率。

3.2.2　注意力机制

近几年来，注意力机制被大量应用于深度学习模型，其不仅能够提升模型精度，而且具有参数少、即插即用的特点。在一段动作信息中，通常一些能表达动作的关节显得尤为重要，也有研究提出引入注意力机制让模型更关注关键节点的时空变化，这一思想成功地提升了模型的识别精度。2s-AGCN 网络模型[37]基于注意力机制提出了自适应图卷积层，其公式表示为

$$f_{\text{out}}=\sum_{k}^{K_v}\boldsymbol{W}_k f_{\text{in}}\left(\boldsymbol{A}_k+\boldsymbol{B}_k+\boldsymbol{C}_k\right) \tag{3}$$

式中：$\boldsymbol{A}_k$ 代表原始骨架数据的邻接矩阵，表示人体关节物理结构；$\boldsymbol{B}_k$ 不仅表示各个关节之间是否存在连接，而且表明连接强度。模型对 $\boldsymbol{B}_k$ 中的元素进行了参数化，与模型共同学习和更新。$\boldsymbol{C}_k$ 矩阵使用归一化的高斯函数，用来计算关节点之间的相似性。采用 3 个矩阵相加的方式构造输出矩阵，每个矩阵代表不同的信息，增加了模型的灵活性。自适应图卷积层结构如图 6[35]所示。

Song 等[38]把骨架划分为 5 个身体部位，更加关注不同身体部分对动作识别的重要性，提出局部注意力机制，局部注意力模块公式表述为

$$f_p=f_{\text{in}}(p)\otimes\delta\left(\theta\left(\text{pool}\left(f_{\text{in}}\right)W\right)W_p\right) \tag{4}$$

$$f_{\text{out}}=\text{Concat}\left(\left\{f_p\mid p=1,2,\cdots,P\right\}\right) \tag{5}$$

式中：pool(·) 代表全局平均池化；$\delta(\cdot)$ 和 $\theta(\cdot)$ 代表 softmax 函数和 ReLU 激活函数；W 和 W_p 都是可学习的参数，W 对所有部分共享，W_p 是具体到每一个部分的注意力权重。式 (5) 是将式 (4) 中每个部分拼接起来后得到的输出 f_{out}。

Qian 等[39]在局部注意力的基础上，又引入身体对称轨迹注意力，提出了 SA-GCN 模型。通过对身体对称轨迹的学习，网络模型获得身体左右部分在动作中协作的信息。通过两种注意力的结合，使模型性能进一步提高。

Hu 等[40]通过引入先验知识的方式提升模型性能。先验引导注意力模块（BPGA）通过选择性丢弃时空骨架关节生成不同的增强骨架序列，对参与动作的特定身体关节进行额外编码，进一步增强提取动作特征的能力。

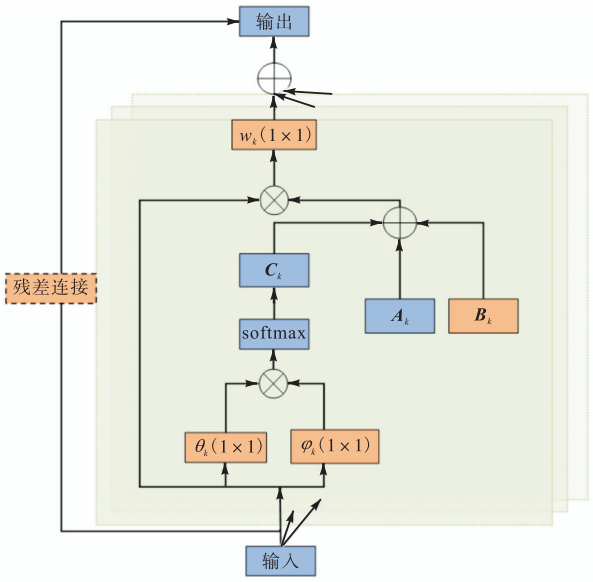

图 6　自适应卷积层

Fig. 6　Adaptive convolutional layer

3.2.3　不同信息融合

ST-GCN 模型没有充分挖掘骨架数据，为了进一步提升识别精度，研究人员尝试通过挖掘更多骨架信息达到更好的效果。2s-AGCN 网络模型是基于 ST-GCN 模型的改进。Shi 等[37]认为骨架数据的二阶信息（即骨骼数据）应该被充分发掘与利用。基于此，研究人员提出了双流架构融合骨架的一阶信息和二阶信息，如图 7[35]所示。

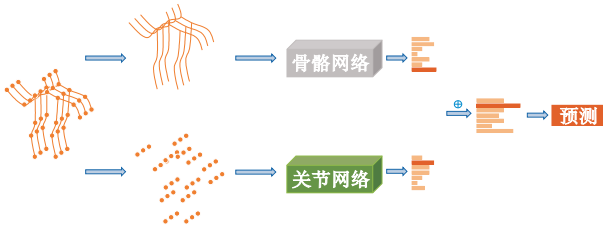

图 7　双流架构

Fig. 7　Two-stream structure

双流网络在图 7 中分别表示为骨骼网络和关节网络。定义靠近骨架重心的关节为源关节，远离重心的关节为目标关节，骨骼数据被表示为从源关节指向目标关节的向量。将两个网络的预测结果进行融合，有效地提升了识别精度。

Song 等[38]引入多输入分支结构、瓶颈结构和残差连接结构，提出了 PA-ResGCN 网络模型。多输入分支结构包括骨架数据中的空间和时间信息，即关节位置、骨骼特征、运动速度 3 个特征分支，在早期进行特征融合时保留信息输入的丰富性。He 等[41]提出的瓶颈结构被应用到模型中，有效减少了参数量，加快模型的推理速度。

Chen 等[42]在 Song 等[38]提出的多分支结构的基础上额外引入姿势信息，提出了 PG-GCN 网络模型。姿势信息本身就包含了动作识别信息和判别线索，利用姿态数据作为更新网络模型的指导信息，提高了模型性能。随着研究者不断对网络模型进行创新，骨架信息以外的信息被引入模型，丰富了输入信息的多样性，为以后的工作提供了新的方向。Cai 等[43]认为仅仅依靠于骨架信息不能够充分得到人体运动特征，导致现有模型无法对有细微运动差异的动作进行分类，因而联合骨架信息和光流对准贴片（JFP）提出了一个新的双流图卷积网络模型。在骨架数据中以每一个节点为中心的小范围内提取出运动光流，把两部分信息进行融合以此提高模型识别精度。

Zhang 等[44]提出了语义信息对于骨架动作识别的重要性，把关节点的类型和帧索引这两种语义信息加入模型中，构建了基于语义引导的图神经网络（SGN）。语义信息使该模型具有更强的可解释性，与其他模型相比，该网络在训练参数较小的情况下，取得了具有竞争性的结果。

4　基于骨架的人体动作识别数据集与模型性能评估

4.1　基于骨架的人体动作识别数据集

随着研究人员对人体动作识别的不断探索，大量与动作识别相关的数据集被创建，用于评估和检测算法的性能。主流的开源数据集（表 1）有 Kinetics[45]、NTU RGB+D[46]、UT-Kinect[47]、SYSU[48]以及Florence 3D。本节将对目前应用最为广泛的 Kinetics 和 NTU RGB + D 数据集进行详细介绍，其余数据集信息见表 1。

表 1　人体动作识别数据集汇总表

Tab. 1　List of human action recognition datasets

数据集	动作类别数	视频数	骨架情况	介绍
Kinetics	400/600/700	254 380/500 000/650 000	每一帧包括 18 个关节，每个关节用 2D 坐标和置信度表示。	涵盖上百种人类动作，每类动作至少包含 600 个视频剪辑，动作类别包括个体动作、人物交互、人人交互。
NTU RGB+D	60/120	56 880/114 480	每一帧包括 25 个关节，每个关节用 3D 坐标表示。	由 40 个 10 至 35 岁的人完成，从 3 个角度进行拍摄，动作类别包括日常动作、与健康相关的动作、双人交互动作。
UT-Kinect	10	200	每一帧包括 20 个关节，每个关节用 2D 坐标和置信度表示。	包括 10 种日常行为动作，共 200 个视频片段，每种动作由 10 名受试者执行 2 次。
SYSU	12	480	每一帧包括 20 个关节，每个关节用 3D 坐标表示。	由 40 人执行的 12 种不同动作，一共 480 个视频片段。
Florence 3D	9	215	每一帧包括 15 个关节，每个关节用 3D 坐标表示。	由 10 人执行的 9 种常见室内动作，每个动作重复执行 2 或 3 次，共 215 个视频片段。

Kinetics 数据集是目前最大的无约束行为识别数据集。该数据集由 Google 公司的 deepmind 团队提供，是从视频网站 YouTube 上剪辑的片段，一段视频对应一个标签，每段视频大约持续 10 s。行为主要有 3 类：单人行为、人与物之间的行为、人与人之间的行为，其分为 Kinetics 400/600/700，不同的数字代表不同的动作类别数。

NTU RGB+D 数据集是目前最大的室内动作捕获三维关节的行为识别数据集，是三维骨架动作识别的标准测试数据集。该数据集是由南洋理工大学的 RoseLab 实验室提出的数据集。采用 Kinect v2 传感器从 3 个不同角度拍摄，由 40 个不同年龄的志愿者在室内采集得到。3D 骨架数据包含每帧 25 个关节点的三维位置，每个片段保证最多有 2 个人的人体骨架。该数据集包含两个划分规则，分别为按人物 ID 划分（X-Sub）与按摄像头视角划分（X-View），其分为 NTU RGB + D 60/120，不同的数字代表不同的动作类别。

4.2　模型性能评估与比较

在人体动作识别任务中，判断一个算法的优劣通常以动作识别准确率为标准。在 Kinetics 数据集中采用 top-1 和 top-5 准确率进行表示。

$$P_{\text{top-1}} = N_1/N \tag{6}$$

$$P_{\text{top-5}} = N_2/N \tag{7}$$

式中：$P_{\text{top-1}}$ 和 $P_{\text{top-5}}$ 分别为 top-1 和 top-5 准确率，N 为总样本数，N_1 为正确分类的样本数，N_2 为正确标签包含在前 5 个分类概率中的个数。在其余数据集中均采用 top-1 准确率进行表示，本文涉及的算法在各个数据集上的性能见表 2。

表 2　基于图卷积人体动作识别算法性能比较

Tab. 2　Performance comparison of human action recognition algorithms based on graph convolution

模型	会议	参数	准确率/%						
			NTU RGB+D		Kinetics		UT-Kinect	SYSU	Florence 3D
			X-Sub	X-View	top-1	top-5			
DPRL[18]	CVPR18	—	83.5	89.8	—	—	98.5	76.9	—
GGCN[19]	CVPR18	—	87.5	94.3	—	—	98.5	77.9	98.4
NAS-GCN[20]	AAAI20	—	89.4	95.7	37.1	60.1	—	—	—
ST-GCN[23]	AAAI18	3.10	81.5	88.3	30.7	52.8	—	—	—
DGNN[26]	CVPR19	26.23	89.9	96.1	36.9	59.6	—	—	—
AS-GCN[27]	CVPR19	6.99	86.8	94.2	34.8	56.5	—	—	—
2s-AGCN[28]	CVPR19	6.94	88.5	95.1	36.1	58.7	—	—	—
ResGCN-N51[29]	ACM MM2020	0.77	89.1	93.5	—	—	—	—	—
ResGCN-B19[29]	ACM MM2020	3.64	90.9	96.0	—	—	—	—	—
JOLO-GCN[31]	WACV2021	10.42	93.8	98.1	38.3	62.3	—	—	—
SGN[32]	CVPR20	1.90	89.0	94.5	—	—	—	90.6	—
Xiong 等[33]	—	—	89.4	96.2	36.9	59.7	—	—	—
Chi 等[34]	CVPR22	—	89.8	91.2	—	—	—	—	—
SA-GCN[39]	ICME22	—	89.0	90.7	—	—	—	—	—
PG-GCN[42]	—	—	91.8	95.8	—	—	—	—	—

由于 ST-GCN 是第一个基于图的动态骨骼建模方法，为后续网络框架提供了指引，所以成为其他行为识别算法的基准。ST-GCN 在 NTU RGB+D 数据集的两个基准 X-Sub 和 X-View 上的识别准确率为 81.5% 和 88.3%。其他算法分别对图构建方式（DGNN）、注意力机制（PA-ResGCN）、融合不同信息（SGN、JOLO-GCN）进行改进，使模型识别性能不断提升。目前，联合骨架数据与光流的动作识别网络（JOLO-GCN）在 NTU RGB+D 数据集和 Kinetics 数据集上取得了最优结果，准确率分别为 93.8% 和 98.1%。其原因是在提取骨架数据的基础上，又引入了局部光流信息，两种角度的信息更有助于计算机理解人体动作。此外，ResGCN-N51 网络在参数量达到最小的同时，也能取得具有竞争力的结果，准确率分别为 89.1% 和 93.5%。

5　展　望

随着图卷积神经网络在人体动作识别领域研究的不断深入，分析现有模型，其发展方向主要有：

（1）图卷积与图池化[49-50]相结合。在模型中引入池化层可以增大节点间的差异性，减少过拟合的发生，使构建深度图卷积模型成为可能。例如，Ying 等[51]提出的 Diffpool 池化方法应用在图分类任务上提高了模型性能。未来，研究人员可以把图卷积与图池化相结合构建深度网络模型。

（2）多特征融合方法。多特征融合是将不同模态的信息进行融合，其通常优于单一数据特征方法的识别结果。在基于图卷积网络的人体动作识别中，除骨架数据外，可以考虑引入光流、RGB 图像、语义等模态信息，如 JOLO-GCN[43]、SGN 等[44]。这些改进有效地提升了模型精度，证明了多特征融合的必要性，但多特征融合的难点在于如何将多个模态的信息进行有效融合。此外，多模态特征融合增大了网络模型的计算复杂度，所以对于模态的选择和特征融合方法是未来该领域重要的研究方向。

（3）细粒度动作识别。随着网络模型在粗粒度数据集上的识别精度越来越高，动作识别正在由粗粒度向细粒度进行转变，一些细粒度动作识别数据集已经被提出，如 Epic-Kitchens[52]、Jester、FineGym[53]。细粒度动作识别要求模型区分相似动作之间的微小差异，对模型提出了更高的要求。如何区分动作之间微小的差异是未来细粒度动作识别的主要研究方向。

（4）小样本学习。传统深度学习模型进行分类任务时往往需要标注大量的数据作为支撑，小样本学习旨在利用先验知识使模型快速适用于只包含少量带

有监督信息的样本任务中。在动作识别任务中，数据的收集成本昂贵，数据标注耗费大量人力物力，研究小样本学习的意义不言而喻。此外，人体动作种类繁多，构建一个包含所有种类的动作识别数据集将面临巨大困难，故小样本学习研究受到了越来越多的关注。Cao 等[54]通过对齐视频的帧，用最小路径的代价衡量两个视频的相似度，从而实现动作分类。Ma 等[55]同时使用基于秩最大化解耦的空间对齐和基于动态规划思想最优匹配的时序对齐，从空域和时域两个角度进行动作识别。如何充分利用好先验知识，是小样本学习未来的研究方向。

参考文献：

[1] 钱慧芳，易剑平，付云虎. 基于深度学习的人体动作识别综述[J]. 计算机科学与探索，2021，15(3)：438–455.

[2] 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]. Applied intelligence，2021，52(5)：5608–5616.

[3] BAISWARE A，SAYANKAR B，HOOD S. Review on recent advances in human action recognition in video data[C]//IEEE. 9th International Conference on Emerging Trends in Engineering and Technology：Signal and Information Processing(ICETET-SIP-19). New York：IEEE，2019：1–5.

[4] GERONIMO D，KJELLSTROM H. Unsupervised surveillance video retrieval based on human action and appearance[C]//IEEE. International Conference on Pattern Recognition. New York：IEEE，2014：4630–4635.

[5] TRAN T T M，PARKER C，TOMITSCH M. A review of virtual reality studies on autonomous vehicle-pedestrian interaction[J]. IEEE Transactions on human-machine systems，2021，51(6)：641–652.

[6] YANG H，LIU L，MIN W，et al. Driver yawning detection based on subtle facial action recognition[J]. IEEE Transactions on multimedia，2021，23：572–583.

[7] 张顺，龚怡宏，王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用[J]. 计算机学报，2019，42(3)：453–482.

[8] AROHAN A，ACHARYA K，SAMANTA A. A review of convolutional neural networks[C]//IEEE. 2020 International Conference on Emerging Trends in Information Technology and Engineering(IC-ETITE). New York：IEEE，2020：1–5.

[9] REN B，LIU M，DING R，et al. A survey on 3D skeleton-based action recognition using learning method [EB/OL]. (2020–02–14)[2022–11–19]. https://arxiv.org/pdf/2002.05907v1.pdf.

[10] LI B，DAI Y C，CHENG X L，et al. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN[EB/OL]. [2022–11–19]. https://arxiv.org/pdf/1704.05645.pdf.

[11] SHI Y，WU M Q，XU W Y. Two-stream convolutional network for skeleton-based action recognition[EB/OL]. [2022–11–19]. https://arxiv.org/pdf/1805.07694.pdf.

[12] LI C K，HOU Y H，WANG P C，et al. Joint distance maps based action recognition with convolutional neural networks[J]. IEEE Signal processing letters，2017，24(5)：624–628.

[13] WANG P C，LI Z Y，HOU Y H，et al. Action recognition based on joint trajectory maps using convolutional neural networks[J]. Elsevier knowledge based system，2018，158：43–53.

[14] SHAHROUDY A，LIU J，NG T T，et al. NTU RGB+D：a large scale dataset for 3D human activity analysis[C]//IEEE. IEEE Conference on CVPR. New York：IEEE，2016：1010–1019.

[15] SHI X，CHEN Z，WANG H，et al. Convolutional LSTM network：a machine learning approach for precipitation nowcasting[C]//ACM. Proceedings of the 28th International Conference on Neural Information Processing Systems. New York：ACM，2015：802–810.

[16] SONG S J，LAN C L，XING J L，et al. An end-to-end spatiotemporal attention model for human action recognition from skeleton data[C]//IEEE. IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE，2016：4263–4270.

[17] 徐冰冰，岑科廷，黄俊杰，等. 图卷积神经网络综述[J]. 计算机学报，2020，43(5)：756–780.

[18] BRUNA J，ZAREMBA W，SZLAM A，et al. Spectral networks and deep locally connected networks on graphs[C]//ICLR. Proceedings of the 2nd International Conference on Learning Representations. Banff：ICLR，2013：1–14.

[19] HAMMOND D K，VANDERGHEYNST P，GRIBONVAL R. Wavelets on graphs via spectral graph theory[J]. Applied and computational harmonic analysis，2010，30(2)：129–150.

[20] KIPF T N，WELLING M. Semi-supervised classification

with graph convolutional networks[C]//ICLR. 5th International Conference on Learning Representations, Toulon: ICLR, 2016: 1–14.

[21] SHUMAN D I, NARANG S K, FROSSARD P, et al. The emerging field of signal processing on graphs: extending high-dimensional data analysis to networks and other irregular domains[J]. IEEE Signal processing magazine, 2013, 30(3): 83–98.

[22] ATWOOD J, TOWSLEY D. Diffusion-convolutional neural networks[C]//ACM. Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: ACM, 2016: 2001–2009.

[23] HAMILTON W L, YING R, LESKOVEC J. Inductive representation learning on large graphs[C]//ACM. Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 1025–1035.

[24] VELIKOVI P, CUCURULL G, CASANOVA A, et al. Graph attention networks[C]//ICLR. International Conference on Learning Representations. Vancouver: ICLR, 2018: 1–12.

[25] TANG Y, YI T, LU J, et al. Deep progressive reinforcement learning for skeleton-based action recognition[C]// IEEE. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 5323–5332.

[26] GAO X, HU W, TANG J, et al. Generalized graph convolutional networks for skeleton-based action recognition [EB/OL]. [2022–11–19]. https://arxiv.org/pdf/1811.2013v1.pdf.

[27] PENG W, HONG X P, CHEN H Y, et al. Learning graph convolutional network for skeleton-based human action recognition by neural searching[C]//AAAI. AAAI Conference on Artificial Intelligence. New York: AAAI, 2019: 2326–3095.

[28] LIU C, ZOPH B, NEUMANN M, et al. Progressive neural architecture search[C]//ECCV. European Conference on Computer Vision. Munich: ECCV, 2017: 21–36.

[29] PHAM H, GUAN M Y, ZOPH B, et al. Efficient neural architecture search via parameter sharing[J]. Journal of machine learning research, 2018, 80(12): 52–63.

[30] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//AAAI. Conference on Artificial Intelligence. Louisiana: AAAI, 2018: 6665–7655.

[31] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]// IEEE. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 1302–1310.

[32] SI C, JING Y, WANG W, et al. Skeleton-based action recognition with hierarchical spatial reasoning and temporal stack learning network[J]. Pattern recognition, 2020, 107: 107511.

[33] XIONG X, MIN M, WANG Q, et al. Human skeleton feature optimizer and adaptive structure enhancement graph convolution network for action recognition[J]. IEEE Transactions on circuits and systems for video technology, 2023, 33(1): 342–353.

[34] CHI H G, HA M H, CHI S, et al. InfoGCN: representation learning for human skeleton-based action recognition[C]//IEEE. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2022: 20154–20164.

[35] SHI L, ZHANG Y, CHENG J, et al. Skeleton-based action recognition with directed graph neural networks [C]//IEEE. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2019: 7904–7913.

[36] LI M, CHEN S, CHEN X, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C]//IEEE. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2019: 3590–3598.

[37] SHI L, ZHANG Y, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//IEEE. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2019: 12018–12027.

[38] SONG Y F, ZHANG Z, SHAN C, et al. Stronger, faster and more explainable: a graph convolutional baseline for skeleton-based action recognition[C]//ACM. In Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 1625–1633.

[39] QIAN R, WANG J, WANG J, et al. Structural attention for channel-wise adaptive graph convolution in skeleton-based action recognition[C]//IEEE. 2022 IEEE International Conference on Multimedia and Expo (ICME). New York: IEEE, 2022: 1–6.

[40] HU Q, LIU H, WANG H Q, et al. Body prior guided

graph convolutional neural network for skeleton-based action recognition[C]//IEEE. 2023 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). New York: IEEE, 2023: 1–5.

[41] HE K, ZHANG X, REN S et al. Deep residual learning for image recognition[C]//IEEE. 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). New York: IEEE, 2016: 770–778.

[42] CHEN H, JIANG Y, KO H. Pose-guided graph convolutional networks for skeleton-based action recognition[J]. IEEE Access, 2022, 10: 111725–111731.

[43] CAI J, JIANG N, HAN X, et al. JOLO-GCN: mining joint-centered light-weight information for skeleton-based action recognition[C]//IEEE. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). New York: IEEE, 2021: 2734–2743.

[44] ZHANG P, LAN C, ZENG W, et al. Semantics-guided neural networks for efficient skeleton-based human action recognition[C]//IEEE. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). New York: IEEE, 2020: 1109–1118.

[45] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset[C]// IEEE. 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). New York: IEEE, 2017: 4724–4733.

[46] SHAHROUDY A, LIU J, NG T T, et al. NTU RGB+D: a large scale dataset for 3D human activity analysis[C]// IEEE. 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). New York: IEEE, 2016: 1010–1019.

[47] XIA L, CHEN C C, AGGARWAL J K. View invariant human action recognition using histograms of 3D joints [C]//IEEE. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. New York: IEEE, 2012: 20–27.

[48] HU J F, ZHENG W S, LAI J H, et al. Jointly learning heterogeneous features for RGB-D activity recognition[C]//IEEE. IEEE Transactions on Pattern Analysis & Machine Intelligence. New York: IEEE, 2016: 2186–2200.

[49] CHEUNG M, SHI J, JIANG L Y, et al. Pooling in graph convolutional neural networks[C]//IEEE. 2019 53rd Asilomar Conference on Signals, Systems, and Computers. New York: IEEE, 2020: 462–466.

[50] ALEJANDRO P M, RUIZ L, RIBEIRO A. Graphon pooling in graph neural networks[C]//IEEE. 2020 28th European Signal Processing Conference(EUSIPCO). New York: IEEE, 2021: 860–864.

[51] YING R, YOU J, MORRIS C, et al. Hierarchical graph representation learning with differentiable pooling[C]// ACM. Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM, 2018: 4805–4815.

[52] DAMEN D, DOUGHTY H, FARINELLA G M, et al. Scaling egocentric vision: the EPIC-KITCHENS dataset[C]//ECCV. European Conference on Computer Vision. Munich: ECCV, 2018: 753–771.

[53] SHAO D, ZHAO Y, DAI B, et al. FineGym: a hierarchical video dataset for fine-grained action understanding[C]//IEEE. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). New York: IEEE, 2020: 2613–2622.

[54] CAO K, JI J, CAO Z, et al. Few-shot video classification via temporal alignment[C]//IEEE. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2020: 1061–1062.

[55] MA N, ZHANG H Y, LI X H, et al. Learning spatial-preserved skeleton representations for few-shot action recognition[C]//ECCV. 2022 European Conference on Computer Vision(ECCV). Berlin: Springer, 2022, 13644: 174–191.

责任编辑：郎婧