利用可选择多尺度图卷积网络的骨架行为识别

曹 毅* 李 杰 叶培涛 王彦雯 吕贤海
(江南大学机械工程学院 无锡 214122)
(江南大学江苏省食品先进制造装备技术重点实验室 无锡 214122)

摘 要：针对目前骨架行为识别方法忽视骨架关节点多尺度依赖关系和无法合理利用卷积核进行时间建模的问题，该文提出了一种可选择多尺度图卷积网络(SMS-GCN)的行为识别模型。首先，介绍了人体骨架图的构建原理和通道拓扑细化图卷积网络的结构；其次，构建成对关节邻接矩阵和多关节邻接矩阵以生成多尺度通道拓扑细化邻接矩阵，并引入图卷积网络，进一步提出多尺度图卷积(MS-GC)模块，以期实现对骨架关节点的多尺度依赖关系的建模；然后，基于多尺度时序卷积和可选择大核网络，提出可选择多尺度时序卷积(SMS-TC)模块，以期实现对有用的时间上下文特征的充分提取，同时结合MS-GC和SMS-TC模块，进而提出可选择多尺度图卷积网络模型并在多支流数据输入下进行训练；最后，在NTU-RGB+D和NTU-RGB+D 120数据集上进行大量实验，实验结果表明，该模型能够捕获更多的关节特征和学习有用的时间信息，具有优异的准确率和泛化能力。

关键词：骨架行为识别；图卷积网络；多尺度通道拓扑细化邻接矩阵；可选择多尺度时序卷积；可选择多尺度图卷积网络

中图分类号：TN911.73; TP391.41 **文献标识码**：A **文章编号**：1009-5896(2025)03-0839-11

DOI: 10.11999/JEIT240702

1 引言

骨架行为识别技术在人机交互[1]、智慧医疗[2]等领域应用广泛，是计算机视觉方向的热门研究分支之一。基于骨架数据的行为识别方法相较于基于深度数据和图像数据的方法，其对相机视角变化、背景遮挡、光照等因素的影响更具鲁棒性[3,4]，因此受到众多国内外学者的广泛关注和研究。

骨架序列包含了大量的关节和时间信息，基于循环神经网络的方法将骨架数据处理为向量序列作为输入，文献[5]通过将复杂网络编码和长短时记忆神经网络(Long Short-Term Memory, LSTM)结合来提高行为识别的准确率。文献[6]提出了一种循环注意力机制，并关注动作序列中的每一帧关节信息以实现全局上下文信息的提取。基于卷积神经网络的方法则将骨架数据处理为伪图像作为输入，文献[7]将卷积和注意力机制相结合来提取骨架序列的局部和全局判别特征。文献[8]提出了一种端到端的语义引导神经网络，通过引入高级语义提高卷积的提取能力以实现对关节依赖关系的建模。但基于循环神经网络的模型在空间域上建模能力弱，而基于卷积神经网络的模型提取时间信息能力不足，为此，基于图卷积神经网络的模型将骨架序列处理为骨架拓扑图的形式作为输入，从而实现对骨架节点的依赖关系进行建模。文献[9]提出了一种双流自适应图卷积网络模型，根据输入特征自适应地提取成对关节间的潜在依赖关系，并结合时序卷积实现时空特征的提取。文献[10]提出了一种通道拓扑细化图卷积模型，通过探索各个通道内的成对关节点间的特定依赖关系来建立通道拓扑图，并结合多尺度时序卷积实现时空特征的提取。文献[11]通过采用图注意力机制构造了空域自适应邻接矩阵，增强了对人体空域特征的提取能力。文献[12]设计了一种时间运动激励模块以突出运动敏感特征，并结合多尺度时间卷积丰富了时间特征的表示能力。文献[13]将原始特征图解耦为空间和时间两个部分，通过使用判别性特征细化模块提升了对模糊动作的识别性能。文献[14]提出了一种通道拓扑自适应图卷积模块，能够学习和细化关节的拓扑关系，提高了模型的性能。

尽管如此，目前基于图卷积神经网络的骨架行为识别模型仍存在以下不足：(1)对于“写字”这类具有细微变化特征的动作，两个手部关节的建模可以捕获局部的运动细节。而对于“梳头”这类复杂的动作，头、拇指、指尖和手等多个关节往往更能表示整体的动作特征变化，现有模型对于关节间的空间建模仅局限于一对关节之间的依赖关系建模，未能充分表示整体的动作特征，单个关节与多

收稿日期：2024-08-12；改回日期：2025-02-17；网络出版：2025-02-24
*通信作者： 曹毅 caoyi@jiangnan.edu.cn
基金项目：国家自然科学基金(51375209)，江苏省“六大人才高峰”计划(ZBZZ-012)，高等学校学科创新引智计划(B18027)
Foundation Items: The National Natural Science Foundation of China (51375209), The Six Talent Peaks Project in Jiangsu Province (ZBZZ-012), The Programme of Introducing Talents of Discipline to Universities (B18027)

关节之间的依赖关系建模同样重要，因此缺乏对骨架关节点的多尺度依赖关系的建模，无法捕获更多代表性关节特征；(2)现有模型仅采用多个膨胀卷积进行简单的特征融合，其无法合理利用时序卷积核进行时间上下文信息的建模，故无法有效学习有用的时间信息。

针对上述问题，论文提出了一种利用可选择多尺度图卷积网络的骨架行为识别模型。首先，介绍了人体骨架图的构建原理和通道拓扑细化图卷积网络的结构；其次，构建成对关节邻接矩阵和多关节邻接矩阵以生成多尺度通道拓扑细化邻接矩阵，并引入图卷积网络，进而构造多尺度图卷积(Multi-Scale Graph Convolution, MS-GC)模块来建立人体骨架关节点的多尺度依赖关系；然后，基于多尺度时序卷积和可选择大核网络，提出可选择多尺度时序卷积(Selective Multi-Scale Temporal Convolution, SMS-TC)模块来选择合适的时序卷积核以提取长短时间动作信息，同时结合MS-GC模块和SMS-TC模块，进一步提出可选择多尺度图卷积网络(Selective Multi-Scale Graph Convolutional Network, SMS-GCN)模型并在多支流数据输入下进行训练；最后，在NTU-RGB+D和NTU-RGB+D 120两个大型数据集上开展实验研究并评估模型的有效性、优越性和泛化性能。

本文的研究贡献为：(1)构建了多尺度通道拓扑细化邻接矩阵，对成对关节和单个关节与多关节进行多尺度依赖关系建模，以期实现对更多代表性的关节信息的有效关注；(2)提出了可选择多尺度时序卷积，对不同时间长短的特征采用合适的时序卷积核进行提取，以期实现对有用的时间上下文信息的充分提取。

2 相关工作

2.1 人体骨架图的构建

人体骨架图可完整地表示一个骨架序列，设其包含T个时间帧和V个人体关节点。人体骨架图$\boldsymbol{G}$的定义为$\boldsymbol{G}=(\boldsymbol{X},\ \boldsymbol{O})$，其同时包含所有人体关节点的状态信息$\boldsymbol{X}$和邻接矩阵$\boldsymbol{O}$。其中，邻接矩阵$\boldsymbol{O}$可表述人体关节点之间的连接关系和相关性大小，其也是骨架图必不可少的组成元素。

2.2 通道拓扑细化图卷积网络

基于人体骨架图，通道拓扑细化图卷积网络(Channel-wise Topology Refinement Graph Convolutional Network, CTR-GCN)[10]在空间域上针对各通道内的独特相关性进行建模，得到通道拓扑细化邻接矩阵$\boldsymbol{R}\in R^{C'\times V\times V}$：

$$\boldsymbol{R}=\mathcal{R}(\boldsymbol{Q}_1,\boldsymbol{A})=\alpha\boldsymbol{Q}_1+\boldsymbol{A} \tag{1}$$

式中$\mathcal{R}(\cdot)$表示拓扑细化函数，$\boldsymbol{Q}_1\in R^{C'\times V\times V}$能够反映各通道内成对关节点间的相关性，其中$C'$表示输出的通道数。$\boldsymbol{A}\in R^{V\times V}$是共享的先验拓扑，$\alpha$是可训练参数，用于调整细化强度。

CTR-GCN在时间域上采用多尺度时序卷积网络提取行为信息中的时间特征，并通过不同的膨胀系数对各分支的时间上下文信息进行融合。其计算公式为

$$\boldsymbol{X}_{\text{out}}=\text{Concat}(\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{X}_3,\boldsymbol{X}_4) \tag{2}$$

式中，$\boldsymbol{X}_1$，$\boldsymbol{X}_2$，$\boldsymbol{X}_3$和$\boldsymbol{X}_4$为不同分支提取到的时间特征，Concat$(\cdot)$表示拼接操作。

值得指出的是：(1)CTR-GCN通过对各通道内的拓扑图进行独特的相关性建模，其虽能增强空间域建模的灵活性，但这种建模仍仅是针对成对关节点之间的依赖关系，其不能建立单个关节点相对于更多关节点的相互依赖关系，故无法有效捕获更多具有代表性的关节信息；(2)在时间域上，CTR-GCN通过采用多个膨胀卷积替代单个固定内核大小的一维卷积[15]，其虽能增大时间特征提取时的感受野，但其忽略了不同的动作持续时间不同这一关键问题，混入了大量的无用信息，仍无法充分关注有用的时间特征信息。

3 可选择多尺度图卷积网络

3.1 多尺度图卷积模块

如上文所述，现有行为识别模型虽能提取成对关节的单一尺度依赖关系，但其仍忽略了更多代表性的关节，故无法有效提取关节点的多尺度依赖关系。针对上述问题，基于通道拓扑细化图卷积模块对不同尺度的关节特征进行建模，构造多尺度通道拓扑细化邻接矩阵，并引入图卷积网络[10]，进一步提出多尺度图卷积(MS-GC)模块，如图1所示。MS-GC模块在建立不同通道中成对关节依赖关系的同时，能够学习单个关节相对于更多关节的相关性，以丰富骨架数据的空间上下文，从而实现对重要关节信息的有效提取。

为清楚地描述多尺度图卷积模块，其具体工作原理如下：

步骤1 骨架特征输入$\boldsymbol{X}_{\text{in}}$。其尺寸为$C\times T\times V$，分别表示骨架特征的输入通道数、帧数和关节数；

步骤2 生成成对关节邻接矩阵$\boldsymbol{Q}_1$和多关节邻接矩阵$\boldsymbol{Q}_2$。首先，通过缩减比率为r的$1\times k$卷积对输入特征$\boldsymbol{X}_{\text{in}}$进行通道降维，并聚合$k$个关节的特征。同理，通过缩减比率为$r$的$1\times 1$卷积对输入特征$\boldsymbol{X}_{\text{in}}$进行通道降维，并学习单个关节的特征。其次，利用时间池化(Temporal Pooling, TP)操作和

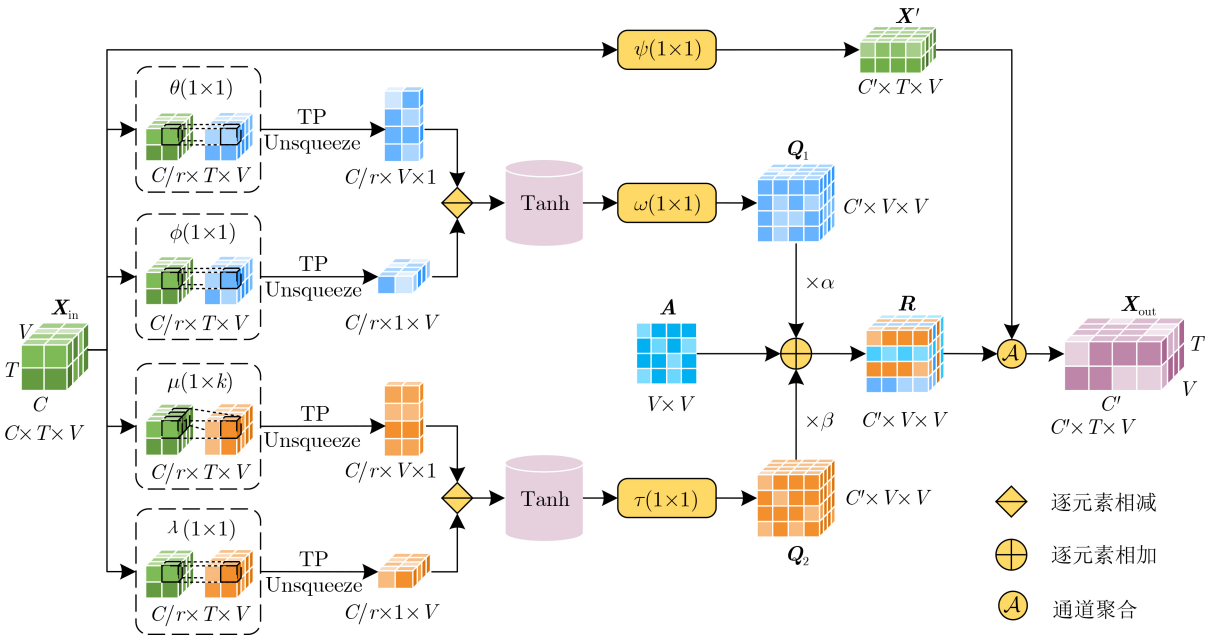


图 1 多尺度图卷积模块结构示意图

unsqueeze函数将特征维度转换为$C/r\times V\times1$和$C/r\times1\times V$，在保留通道维度的同时，通过减法运算得到差分拓扑以分别获取各通道内成对关节的相关性和单个关节相对于多个关节的相关性。最后，经过tanh激活函数计算和1×1卷积升维，得到成对关节邻接矩阵$\boldsymbol{Q}_1$和多关节邻接矩阵$\boldsymbol{Q}_2$

$$\left.\begin{aligned}\boldsymbol{Q}_1&=\omega(\tanh(\mathrm{TP}(\theta(\boldsymbol{X}_{\mathrm{in}}))-\mathrm{TP}(\phi(\boldsymbol{X}_{\mathrm{in}}))))\\\boldsymbol{Q}_2&=\tau(\tanh(\mathrm{TP}(\mu(\boldsymbol{X}_{\mathrm{in}}))-\mathrm{TP}(\lambda(\boldsymbol{X}_{\mathrm{in}}))))\end{aligned}\right\}\tag{3}$$

式中，$\boldsymbol{Q}_1$和$\boldsymbol{Q}_2$的尺寸为$C'\times V\times V$，其中C'表示输出通道数。$\omega(\cdot)$和$\tau(\cdot)$表示1×1的卷积运算，TP表示时间池化操作，$\theta(\cdot)$, $\phi(\cdot)$, $\mu(\cdot)$和$\lambda(\cdot)$表示1×1的卷积运算；

步骤3　构造多尺度通道拓扑细化邻接矩阵。为建立成对关节和单个关节相对于多关节的多尺度依赖关系，根据拓扑细化函数$\mathcal{R}(\cdot)$，基于成对关节邻接矩阵$\boldsymbol{Q}_1$和多关节邻接矩阵$\boldsymbol{Q}_2$对共享的先验拓扑$\boldsymbol{A}\in R^{V\times V}$进行细化，得到多尺度通道拓扑细化邻接矩阵$\boldsymbol{R}\in R^{C'\times V\times V}$：

$$\boldsymbol{R}=\mathcal{R}(\boldsymbol{Q}_1,\boldsymbol{A},\boldsymbol{Q}_2)=\alpha\boldsymbol{Q}_1+\boldsymbol{A}+\beta\boldsymbol{Q}_2\tag{4}$$

式中α和β是可训练参数，用于调整细化强度；

步骤4　构建多尺度图卷积模块并提取空间特征。首先，对输入特征$\boldsymbol{X}_{\mathrm{in}}\in R^{C\times T\times V}$采用1×1的卷积运算进行特征转换以生成高阶特征$\boldsymbol{X}'\in R^{C'\times T\times V}$，接着对第$c$ ($c\in\{1,2,\cdots,C'\}$)个通道的多尺度通道拓扑细化邻接矩阵$\boldsymbol{R}_c$和高阶特征$\boldsymbol{X}'_c$进行聚合以实现图卷积操作，通过将所有通道拓扑特征的输出串联得到最终的输出$\boldsymbol{X}_{\mathrm{out}}\in R^{C'\times T\times V}$，计算公式为

$$\left.\begin{aligned}\boldsymbol{X}'&=\psi(\boldsymbol{X}_{\mathrm{in}})\\\boldsymbol{X}_{\mathrm{out}}&=\mathcal{A}(\boldsymbol{X}',\boldsymbol{R})=[\boldsymbol{R}_1\boldsymbol{X}'_1\parallel\boldsymbol{R}_2\boldsymbol{X}'_2\parallel\cdots\\&\quad\parallel\boldsymbol{R}_{C'}\boldsymbol{X}'_{C'}]\end{aligned}\right\}\tag{5}$$

式中，$\psi(\cdot)$表示1×1的卷积运算，$\mathcal{A}(\cdot)$表示通道聚合操作，$\parallel$表示串联操作。

对比式(4)和式(1)可知，MS-GC模块在通道拓扑细化邻接矩阵的基础上进一步构造多关节邻接矩阵$\boldsymbol{Q}_2$，进而构建多尺度通道拓扑细化邻接矩阵，其实现了成对关节和单个关节与多关节的多尺度依赖关系的建模，增强了对多尺度依赖关系的提取能力，故实现了对重要关节的充分关注。

3.2 可选择多尺度时序卷积模块

必须指出的是，由于“站起来”和“坐下”这类动作特征其具有相似的空间结构特征[16]，基于MS-GC模块进行空间建模时无法体现一段时间帧内的动态变化，故阻碍了模型对相似特征的提取，因此，需要对骨架数据进行时间建模以学习时间信息的动态变化。现有的基于图卷积网络的模型大多采用多尺度时序卷积[10,17]，其仅进行简单的特征融合，故无法选择合适的卷积核提取多尺度的时间上下文信息。针对上述问题，为实现对有用的时间上下文信息的充分提取，基于多尺度时序卷积[10]和可选择大核网络[18]，提出了可选择多尺度时序卷积(SMS-TC)模块，如图2所示，其具体实现可细分为以下步骤：

步骤1　特征输入$\boldsymbol{X}_{\mathrm{in}}$。其尺寸为$C'\times T\times V$，分别表示特征的通道数、输入帧数和关节数；

步骤2　特征提取。SMS-TC模块包含4个并行分支，利用1×1卷积将输入特征$\boldsymbol{X}_{\mathrm{in}}$的通道数降至

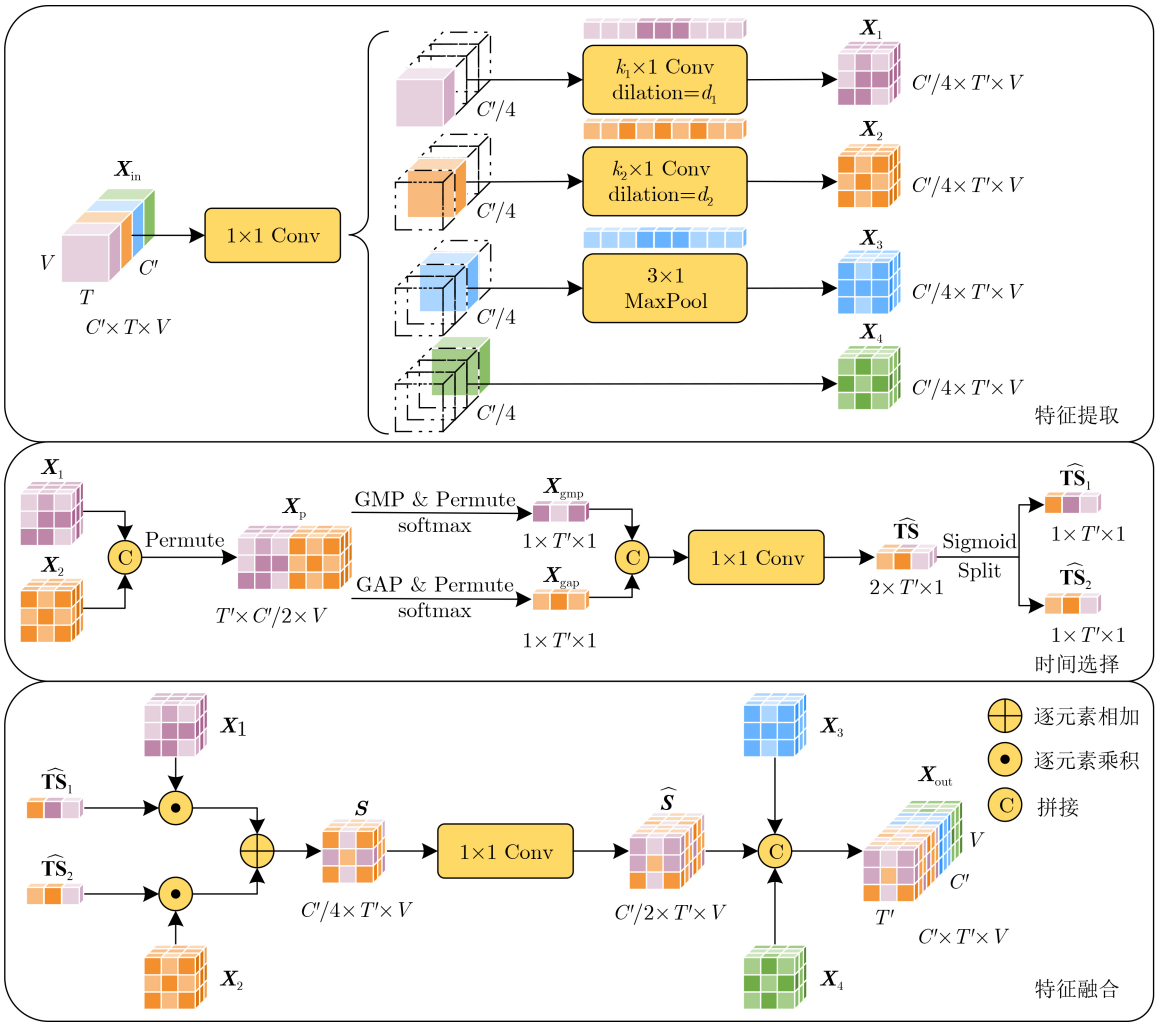


图 2 可选择多尺度时序卷积模块结构示意图

$C'/4$，其中两个卷积分支采用不同膨胀系数的时序卷积进行不同尺度的时间上下文信息的特征提取，最大池化分支实现对关键帧的提取，最终得到4个分支的输出$\boldsymbol{X}_1$，$\boldsymbol{X}_2$，$\boldsymbol{X}_3$和$\boldsymbol{X}_4$，计算公式为

$$\left.\begin{aligned}\boldsymbol{X}_1&=\boldsymbol{W}_{d_1}(\boldsymbol{W}_{1\times1}\boldsymbol{X}_{\text{in}})\\ \boldsymbol{X}_2&=\boldsymbol{W}_{d_2}(\boldsymbol{W}_{1\times1}\boldsymbol{X}_{\text{in}})\\ \boldsymbol{X}_3&=\text{Maxpool}(\boldsymbol{W}_{1\times1}\boldsymbol{X}_{\text{in}})\\ \boldsymbol{X}_4&=\boldsymbol{W}_{1\times1}\boldsymbol{X}_{\text{in}}\end{aligned}\right\}\tag{6}$$

式中，输出特征$\boldsymbol{X}_1$，$\boldsymbol{X}_2$，$\boldsymbol{X}_3$和$\boldsymbol{X}_4$的尺寸为$C'/4\times T'\times V$，其中T'表示特征的输出帧数。$\boldsymbol{W}_{d_1}$和$\boldsymbol{W}_{d_2}$分别为膨胀系数为d_1和d_2的卷积权重，卷积核的尺寸分别为$k_1\times1$和$k_2\times1$，$\boldsymbol{W}_{1\times1}$为1×1的卷积权重，Maxpool表示最大池化操作；

步骤3 时间选择。为增强网络专注于有用的时间上下文信息来识别长短时间动作特征的能力，采用时间选择机制从不同尺度的时序卷积提取到的特征中选择合适的特征图，其构建过程可细分为以下步骤：

首先，拼接不同尺度的卷积核获得的特征$\boldsymbol{X}_1$和$\boldsymbol{X}_2$，再使用permute函数将维度转换至$T'\times C'/2\times V$，得到特征$\boldsymbol{X}_{\text{p}}$

$$\boldsymbol{X}_{\text{p}}=\mathcal{P}(\text{Concat}(\boldsymbol{X}_1,\boldsymbol{X}_2))\tag{7}$$

式中，$\mathcal{P}(\cdot)$为permute函数，Concat$(\cdot)$表示拼接操作。

接着，对于特征$\boldsymbol{X}_{\text{p}}$，采用全局最大池化(Global Max Pooling, GMP)操作提取行为信息的关键时间特征，为减少全局最大池化过程中的部分细节特征丢失，利用全局平均池化(Global Average Pooling, GAP)操作提取全局时间上下文特征，在提取显著特征的同时保留部分细节信息，使用permute函数进行维度转换并经过softmax函数计算，得到全局最大池化特征$\boldsymbol{X}_{\text{gmp}}\in R^{1\times T'\times1}$和全局平均池化特征$\boldsymbol{X}_{\text{gap}}\in R^{1\times T'\times1}$：

$$\left.\begin{aligned}\boldsymbol{X}_{\text{gmp}}&=\delta(\mathcal{P}(\text{GMP}(\boldsymbol{X}_{\text{p}})))\\ \boldsymbol{X}_{\text{gap}}&=\delta(\mathcal{P}(\text{GAP}(\boldsymbol{X}_{\text{p}})))\end{aligned}\right\}\tag{8}$$

式中，δ表示softmax函数，GMP和GAP分别表示全局最大池化和全局平均池化操作。

最后，为允许特征$\boldsymbol{X}_{\text{gmp}}$和$\boldsymbol{X}_{\text{gap}}$之间的信息交互，将两个特征拼接并使用1×1卷积将池化特征转换为时间注意力图$\widehat{\mathbf{TS}} \in R^{2\times T'\times 1}$，再经过sigmoid函数计算和分离操作得到两个相应的时间选择掩模$\widehat{\mathbf{TS}}_1 \in R^{1\times T'\times 1}$和$\widehat{\mathbf{TS}}_2 \in R^{1\times T'\times 1}$，计算公式为

$$\left.\begin{aligned}\widehat{\mathbf{TS}} &= \boldsymbol{W}_{1\times 1}(\text{Concat}(\boldsymbol{X}_{\text{gmp}}, \boldsymbol{X}_{\text{gap}}))\\ \widehat{\mathbf{TS}}_1, \widehat{\mathbf{TS}}_2 &= \text{Split}(\sigma(\widehat{\mathbf{TS}}))\end{aligned}\right\} \tag{9}$$

式中，Split(·)表示分离函数，σ为sigmoid激活函数；

步骤4　特征融合。将提取到的不同尺度的输出特征$\boldsymbol{X}_1$和$\boldsymbol{X}_2$与其相应的时间选择掩模$\widehat{\mathbf{TS}}_1$和$\widehat{\mathbf{TS}}_2$进行加权得到特征$\boldsymbol{S} \in R^{C'/4\times T'\times V}$，再使用1×1卷积进行融合并升维至$C'/2\times T'\times V$。最终，将升维后的特征$\widehat{\boldsymbol{S}}$和特征提取阶段得到的特征$\boldsymbol{X}_3$和$\boldsymbol{X}_4$进行特征拼接得到最终输出$\boldsymbol{X}_{\text{out}} \in R^{C'\times T'\times V}$，计算公式为

$$\left.\begin{aligned}\boldsymbol{S} &= \sum_{i=1}^{2}(\widehat{\mathbf{TS}}_i \odot \boldsymbol{X}_i)\\ \widehat{\boldsymbol{S}} &= \boldsymbol{W}_{1\times 1}\boldsymbol{S}\\ \boldsymbol{X}_{\text{out}} &= \text{Concat}(\widehat{\boldsymbol{S}}, \boldsymbol{X}_3, \boldsymbol{X}_4)\end{aligned}\right\} \tag{10}$$

式中，⊙表示逐元素乘积操作。

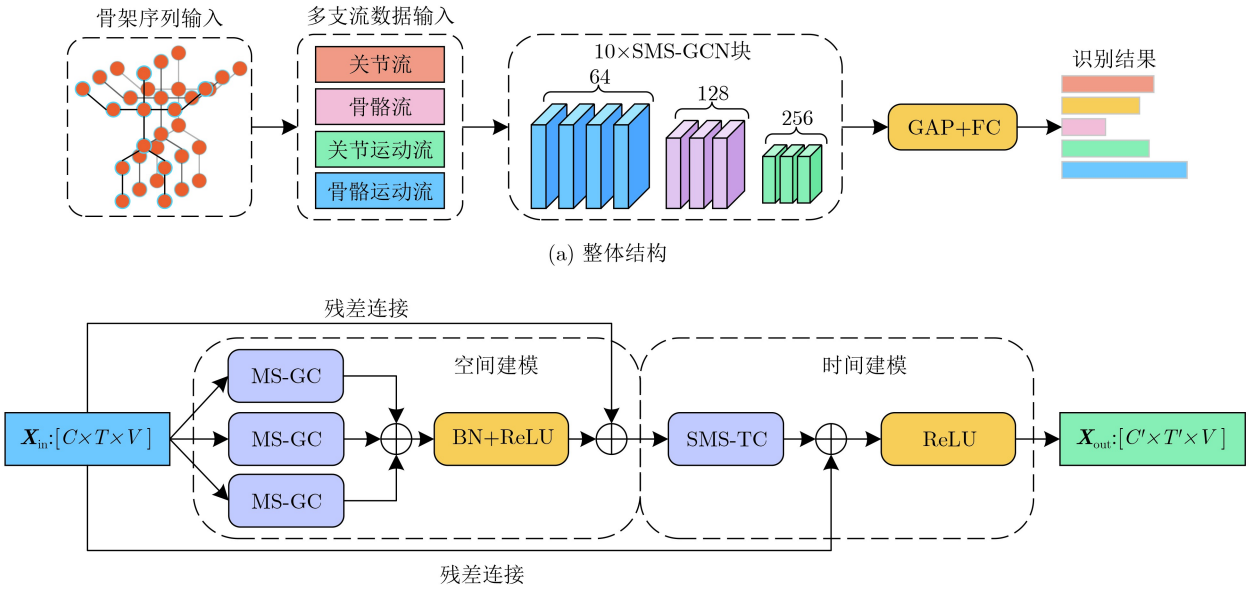
对比式(10)和式(2)可知，SMS-TC模块在多尺度时序卷积的基础上利用时间选择机制选取合适的时序卷积，相较于直接对各分支提取的时间特征进行直接融合的方式而言，其可使网络关注于有用的时间上下文信息，从而减少特征信息的冗余，提高模型的识别准确率。

3.3　可选择多尺度图卷积网络模型

为实现对更多代表性关节信息和有用的时间上下文信息的有效提取，结合上述MS-GC模块和SMS-TC模块，提出一种利用可选择多尺度图卷积网络(SMS-GCN)的骨架行为识别模型，模型的整体结构和SMS-GCN基本块的结构如图3(a)和3(b)所示，其实现原理如下。

首先，将骨架序列处理为多支流数据，包括关节流、骨骼流、关节运动流和骨骼运动流。将骨架序列中的关节点位置坐标作为关节流信息，骨骼流信息可表示为相邻关节点间的关节流信息的差分，关节运动流信息表示为相邻帧间的关节流信息的差分，骨骼运动流信息表示为相邻帧间的骨骼流信息的差分。其次，将处理后的数据输入到网络中，该网络主要包含10个SMS-GCN基本块，每个基本块的输出通道数配置为：前4个通道数为64，中间3个通道数为128，最后3个通道数为256。接着，在网络末端经过全局平均池化(Global Average Pooling, GAP)层统一特征的输出大小。最后，经过全连接(Fully Connected, FC)层得到最终的识别结果。

SMS-GCN基本块包含空间建模、时间建模和残差连接3个部分。在空间建模的过程中，通过并行部署3个MS-GC模块以捕获多尺度依赖关系，实现对代表性关节信息的充分关注。在时间建模的过程中，利用SMS-TC模块选择合适的时序卷积核以实现对有用的时间上下文信息的有效提取。利用残差连接稳定模型的训练效果并加速收敛。图中BN



(a) 整体结构

(b) SMS-GCN块

图 3 SMS-GCN结构示意图

为批归一化(Batch Normalization, BN)层，ReLU为激活函数。

文献[9]采用双流融合策略，将训练得到的关节流和骨骼流的输出结果进行加权融合，但其仅描述骨架序列的空间位置信息，忽略了关节和骨骼的位置随时间变化的速度信息。为增强对骨架序列的表征能力，文献[10]和文献[19]等均采用多流融合策略，在双流的基础上增加了包含速度信息的关节运动流和骨骼运动流信息，其在行为识别任务上效果显著。因此，可选择多尺度图卷积网络模型同样采用此策略，分别将关节流、骨骼流、关节运动流和骨骼运动流4个不同模态的数据作为模型的输入，基于可选择多尺度图卷积网络模型进行网络训练，将训练得到的各支流输出结果加权融合，并得到最终的人体行为识别结果。

4 实验

4.1 实验设计

为验证可选择多尺度图卷积网络的识别准确率和泛化能力，分别基于NTU-RGB+D[20]和NTU-RGB+D 120[21]两个大型公开数据集开展骨架行为识别的研究。

4.1.1 数据集

NTU-RGB+D是广泛应用于行为识别的公开数据集，其包含56 880个动作样本和60类动作，针对40个对象采用3个深度传感相机采集样本，每个样本包含25个关节的3D坐标。数据集依据跨视角(Cross View, CV)和跨对象(Cross Subject, CS)两种基准进行评估。

NTU-RGB+D 120相较于NTU-RGB+D数据集包含了更多的动作类别、样本和对象，其包含的动作类别增加至120类，动作样本增加至114 480个，采集的对象增加至106个。数据集依据跨设置(Cross Setup, CSet)和跨对象(Cross Subject, CSub)两种基准进行评估。

4.1.2 实验细节

实验平台配置为：单张RTX 4090显卡， Pytorch深度学习框架。实验的优化策略为随机梯度下降算法， Nesterov动量为0.9，选择交叉熵函数作为反向传播的损失函数，初始学习率为0.1。对于NTU-RGB+D和NTU-RGB+D 120数据集，将权重衰减设置为0.0004，批量大小设为64，每个样本大小调整至64帧，训练轮数为65次，衰减周期设在第35和第55个训练轮数。

4.2 消融实验

4.2.1 多尺度图卷积核尺寸对比实验

多尺度图卷积模块MS-GC通过使用1×1卷积和1×k卷积来建立单个关节与多个关节的依赖关系，从而实现对代表性关节的充分关注。为验证不同大小的卷积核尺寸k对模型空间建模性能的影响，在CV基准下，以关节支流数据作为输入，基于NTU-RGB+D数据集进行实验研究，实验结果如表1所示。

由表1可知：(1)当卷积核尺寸k设置为3时，模型具有最高的Top-1准确率95.09%，Top-5准确率也取得了不错的效果；(2)在增大卷积核的尺寸时，感受野虽不断增大，但其引起了特征的冗余，导致识别准确率逐渐下降。因此后续均采用尺寸为3的卷积核提取多个关节的特征并进行空间建模。

4.2.2 可选择多尺度时序卷积核尺寸和膨胀系数对比实验

可选择多尺度时序卷积模块SMS-TC中分别采用膨胀系数为d_1的$k_1\times1$卷积和膨胀系数为d_2的$k_2\times1$卷积进行时间特征提取，其既能够提取不同尺度的时间上下文信息，又能选择合适的卷积核捕获不同时间的动作特征。由于时域感受野是影响SMS-TC模块性能的重要因素之一，感受野的大小决定了SMS-TC模块捕获时间上下文的能力。为验证不同的卷积核尺寸k_1和k_2与膨胀系数d_1和d_2对模型时间建模性能的影响，论文提出了两种卷积核选择方案。

方案1 采用传统卷积：通过调整卷积核的尺寸调整感受野的大小，使用小尺寸卷积核提取短时间动作特征，使用大尺寸卷积核提取长时间动作特征。在CV基准下，保证膨胀系数d_1和d_2一致，以关节支流数据作为输入，基于NTU-RGB+D数据集进行实验，实验结果如表2所示。

由表2可知：(1)当卷积核尺寸k_1和k_2分别设置为1和11时，模型的准确率最高，达到95.04%；(2)对比模型5，模型15的性能虽达到了不错的效果，但其参数量提升了0.13M，增加了计算成本。

方案2 采用膨胀卷积：通过调整卷积核的膨胀率调整感受野的大小，使用小膨胀率的卷积核提取短时间动作特征，使用大膨胀率的卷积核提取长时间动作特征。在CV基准下，保证卷积核尺寸k_1和k_2一致，以关节支流数据作为输入，基于NTU-RGB+D数据集进行实验，实验结果如表3所示。

表 1 不同卷积核尺寸的模型准确率对比(%)

模型	Top-1	Top-5
SMS-GCN(k=3)	**95.09**	99.41
SMS-GCN(k=5)	95.00	**99.43**
SMS-GCN(k=7)	94.99	99.40
SMS-GCN(k=9)	94.93	99.36

表 2　不同卷积核尺寸的准确率对比

模型	(k_1, k_2)	(d_1, d_2)	Top-1(%)	参数量(M)	时间(s)	模型	(k_1, k_2)	(d_1, d_2)	Top-1(%)	参数量(M)	时间(s)
1	(1, 3)	(1, 1)	94.58	1.77	277	9	(3, 11)	(1, 1)	94.87	1.93	277
2	(1, 5)	(1, 1)	94.79	1.80	275	10	(5, 7)	(1, 1)	94.69	1.90	278
3	(1, 7)	(1, 1)	94.92	1.83	277	11	(5, 9)	(1, 1)	94.82	1.93	277
4	(1, 9)	(1, 1)	94.78	1.87	282	12	(5, 11)	(1, 1)	94.89	1.97	277
5	(1, 11)	(1, 1)	**95.04**	1.90	276	13	(7, 9)	(1, 1)	94.84	1.97	270
6	(3, 5)	(1, 1)	94.92	1.83	287	14	(7, 11)	(1, 1)	94.83	2.00	285
7	(3, 7)	(1, 1)	94.74	1.87	281	15	(9, 11)	(1, 1)	95.03	2.03	272
8	(3, 9)	(1, 1)	94.91	1.90	271						

表 3　不同卷积核尺寸和膨胀系数的准确率对比

模型	(k_1, k_2)	(d_1, d_2)	Top-1(%)	参数量(M)	时间(s)	模型	(k_1, k_2)	(d_1, d_2)	Top-1(%)	参数量(M)	时间(s)
1	(1, 1)	(1, 2)	92.95	1.74	747	7	(9, 9)	(1, 3)	95.07	2.00	761
2	(3, 3)	(1, 2)	94.69	1.80	886	8	(9, 9)	(1, 4)	95.09	2.00	1 000
3	(5, 5)	(1, 2)	94.89	1.87	865	9	(9, 9)	(2, 3)	94.98	2.00	1 466
4	(7, 7)	(1, 2)	94.66	1.93	854	10	(9, 9)	(2, 4)	94.96	2.00	1 490
5	(9, 9)	(1, 2)	**95.09**	2.00	735	11	(9, 9)	(3, 4)	94.85	2.00	1 479
6	(11, 11)	(1, 2)	94.97	2.06	767						

由表3可知：(1)对比模型1, 2, 3, 4, 5, 6，采用9×1的卷积核识别准确率最高，达到95.09%，所需时间(单个周期内训练时间和测试时间的和)也相对较少，采用1×1的卷积核识别准确率最低，其虽容易捕捉局部时间特征，且参数量最小，但会增加长时间动作特征信息的损失；(2)相比于采用9×1的卷积核，11×1的卷积核虽具有更大的感受野，但其不利于对短时间动作特征的提取，且容易引起特征信息的冗余，增加参数量，降低模型性能；(3)对比模型5, 7, 8, 9, 10和11，当时序卷积核的膨胀系数d_1和d_2分别设置为1和2时，性能最优，所需时间也相对较少。且在增加膨胀系数时，感受野虽随之增加，但准确率呈现下降的趋势，因为膨胀系数的增加会导致提取到的时间信息更加稀疏，易造成关键时间帧的丢失；(4)模型8和模型5虽识别准确率相同，但所需时间也有所增加，增加计算成本，降低模型的收敛速度。

对比方案1和方案2，采用膨胀卷积的方案2识别效果优于采用传统卷积的方案1，因此后续均采用膨胀系数分别为1和2的9×1时序卷积进行时间特征提取。

4.2.3　可选择多尺度时序卷积模块结构对比实验

为验证不同结构的SMS-TC模块在时间建模上的有效性，通过删除SMS-TC模块中的部件以验证其识别性能。在CV基准下，以关节支流数据作为输入，基于NTU-RGB+D数据集进行实验，结果如表4所示。

由表4可知，与SMS-GCN模型相比：(1)SMS-GCN(无SMS-TC)删去SMS-TC模块并替换为普通的时序卷积[9]，其参数量增加了1.76M，识别准确率下降了0.66%，验证了SMS-TC模块的有效性，进一步表明SMS-GCN模型可以在使用少量参数的同时提取不同尺度的时间信息；(2)SMS-GCN(无S)删去时间选择机制并采取与CTR-GCN[10]相同的直接融合方式进行时间特征提取，其识别准确率下降了0.13%，验证了时间选择机制的有效性，进一步表明SMS-GCN模型可以有效地捕获时间上下文信息；(3)SMS-GCN(无GMP)删去全局最大池化操作，其性能有所下降，因为仅使用全局平均池化操作会忽略关键时间信息，降低识别准确率；(4)SMS-GCN(无GAP)删去全局平均池化操作，其性能同样有所下降，因为仅使用全局最大池化操作会导致部分细节特征丢失，影响识别效果。

4.2.4　模块有效性验证实验

为验证MS-GC模块和SMS-TC模块的有效性，

表 4　不同结构的模型准确率对比

模型	参数量(M)	Top-1(%)
SMS-GCN	2.00	**95.09**
SMS-GCN(无SMS-TC)	3.76	94.46
SMS-GCN(无S)	1.96	94.97
SMS-GCN(无GMP)	2.00	94.90
SMS-GCN(无GAP)	2.00	94.93

选取CTR-GCN[10]作为基线模型，在CV基准下，以关节支流和骨骼支流数据作为输入，基于NTU-RGB+D数据集进行模型消融实验，结果如表5所示。

由表5可知，与基线模型相比：(1)基于MS-GC模块的模型在关节流和骨骼流上的准确率分别提升了0.17%和0.20%，双流融合后的准确率提高了0.23%，验证了多尺度依赖关系建模的优越性，相比于仅使用成对关节依赖关系建模，单个关节与多关节的依赖关系建模更能捕获具有代表性的多关节信息；(2)基于SMS-TC模块的模型在关节流和骨骼流上的准确率分别提升了0.12%和0.20%，双流融合后的准确率提高了0.22%，验证了可选择多尺度时序卷积模块的有效性，选择合适的卷积核提取时间特征相比直接融合多尺度特征的方式，更有利于对长短时间动作的提取；(3)基于MS-GC模块和SMS-TC模块的SMS-GCN模型在关节流和骨骼流上的准确率分别提升了0.35%和0.26%，双流融合后的准确率提高了0.45%。模型的消融实验充分验证了SMS-GCN模型中各模块的有效性。

4.3 识别准确率对比实验

为评估可选择多尺度图卷积网络模型的优越性能和泛化能力，以关节支流、骨骼支流、关节运动支流和骨骼运动支流数据作为输入，在NTU-RGB+D和NTU-RGB+D 120两个大型数据集上与国内外先进方法进行对比实验，包括基于RNN, CNN和GCN的方法。

4.3.1 基于NTU-RGB+D数据集

为验证SMS-GCN模型的识别性能，在CS和CV基准下，基于NTU-RGB+D数据集开展模型准确率对比实验，对比结果如表6所示。

表 5 添加不同模块的模型准确率对比(%)

模型	关节流	骨骼流	双流
CTR-GCN	94.74	94.70	96.07
CTR-GCN + MS-GC	94.91	94.90	96.30
CTR-GCN + SMS-TC	94.86	94.90	96.29
SMS-GCN	**95.09**	**94.96**	**96.52**

表 6 NTU-RGB+D数据集下模型的准确率对比(%)

模型	CS	CV	模型	CS	CV
CNC-LSTM[5]	83.3	91.8	3D-GCN[23]	89.4	93.3
LAGA-Net[7]	87.1	93.2	ML-STGNet[12]	91.9	96.2
ST-GCN[15]	81.5	88.3	MADT-GCN[19]	90.4	96.5
2s-AGCN[9]	88.5	95.1	SMS-GCN(单流)	89.7	95.1
CTR-GCN[10]	92.4	96.8	SMS-GCN(双流)	91.9	96.5
VN-GAN[22]	92.0	96.7	SMS-GCN(多流)	**92.6**	**96.9**

由表6可知，在CS和CV基准下：(1)基于多支流数据输入的SMS-GCN模型相比于基于单流(关节流)和双流的SMS-GCN模型，准确率分别提升了2.9%, 1.8%和0.7, 0.4%，验证了多支流数据输入方法的优越性；(2)SMS-GCN模型以92.6%和96.9%的准确率超越了其他模型，验证了该模型的优秀识别性能；(3)相较于基线模型CTR-GCN, SMS-GCN的准确率分别提升0.2%和0.1%，验证了该模型在性能上的提升；(4)SMS-GCN模型与RNN模型CNC-LSTM和CNN模型LAGA-Net相比，准确率分别提高了9.3%, 5.1%和5.5%, 3.7%，验证了基于图卷积网络的SMS-GCN模型与基于传统深度学习网络的模型相比性能更优；(5)SMS-GCN模型相较于经典的GCN模型ST-GCN和2s-AGCN，准确率分别提升11.1%, 8.6%和4.1%, 1.8%，验证了其在性能上的提升；(6)SMS-GCN模型与最新的GCN模型MADT-GCN和VN-GAN相比，准确率分别提升了2.2%, 0.4%和0.6%, 0.2%，验证了该模型与主流方法相比优势显著。

4.3.2 基于NTU-RGB+D 120数据集

为验证模型在不同数据集上的泛化能力，在CSub和CSet基准下，基于NTU-RGB+D 120数据集进行准确率对比实验，对比结果如表7所示。

由表7可知，在CSub和CSet基准下：(1)相较于单流和双流数据输入的方法，多流SMS-GCN模型的准确率分别提高4.0%, 4.1%和0.5%, 0.7%，验证了多支流数据输入方法的优越性；(2)SMS-GCN模型以89.3%和90.7%的准确率超过了其他模型，验证了该模型的优异泛化性能；(3)SMS-GCN模型与RNN模型GCA-LSTM和CNN模型LAGA-Net相比，准确率分别提高了31.0%, 31.5%和8.3%, 8.5%，验证了基于图卷积网络的SMS-GCN模型与基于传统深度学习网络的模型相比性能更优；(4)SMS-GCN模型相较于基线模型CTR-GCN，准确率分别提升0.4%和0.1%，验证了其在性能上的提升；(5)SMS-GCN模型与最新的GCN模型MADT-GCN和ML-STGNet相比，准确率分别提升了2.8%, 2.5%和0.7%,

表 7 NTU-RGB+D 120数据集下模型的准确率对比(%)

模型	CSub	CSet	模型	CSub	CSet
GCA-LSTM[6]	58.3	59.2	ML-STGNet[12]	88.6	90.0
LAGA-Net[7]	81.0	82.2	MADT-GCN[19]	86.5	88.2
ST-GCN[15]	70.7	73.2	VN-GAN[22]	87.6	89.4
2s-AGCN[9]	82.9	84.9	SMS-GCN(单流)	85.3	86.6
CTR-GCN[10]	88.9	90.6	SMS-GCN(双流)	88.8	90.0
STFE-GCN[11]	84.1	86.3	SMS-GCN(多流)	**89.3**	**90.7**

0.7%，验证了该模型与最新的GCN模型相比优势显著，其泛化性能更强。

5　结束语

针对目前骨架行为识别方法忽视骨架关节点多尺度依赖关系和无法合理利用卷积核进行时间建模的问题，提出了一种利用可选择多尺度图卷积网络的骨架行为识别模型。空域上构造了多尺度通道拓扑细化邻接矩阵并参与图卷积，以构建多尺度图卷积模块来建立成对关节和单个关节与多关节的多尺度依赖关系，实现了对更多代表性关节特征的充分提取。时域上提出了一种可选择多尺度时序卷积模块，通过选择合适的时序卷积核对不同尺度的时间上下文信息进行提取并融合，实现了对有用的时间信息的充分关注。

实验结果表明，所提出的模型相较于国内外主流方法，具有优秀的识别准确率和泛化能力，进一步验证了模型能够捕获更多的代表性关节特征和学习有用的时间上下文信息。然而当前该模型仅针对人体骨架的数据进行训练，缺乏对物体、多人交互动作等因素的有效识别。因此，后续将加入多模态的数据进行多模态融合，以提升模型的鲁棒性，减少行为识别的困难。

参 考 文 献

[1] IODICE F, DE MOMI E, and AJOUDANI A. HRI30: An action recognition dataset for industrial human-robot interaction[C]. Proceedings of the 26th International Conference on Pattern Recognition, Montreal, Canada, 2022: 4941–4947. doi: 10.1109/ICPR56361.2022.9956300.

[2] SARDARI S, SHARIFZADEH S, DANESHKHAH A, *et al*. Artificial intelligence for skeleton-based physical rehabilitation action evaluation: A systematic review[J]. *Computers in Biology and Medicine*, 2023, 158: 106835. doi: 10.1016/j.compbiomed.2023.106835.

[3] SUN Zehua, KE Qiuhong, RAHMANI H, *et al*. Human action recognition from various data modalities: A review[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(3): 3200–3225. doi: 10.1109/TPAMI.2022.3183112.

[4] 曹毅, 吴伟官, 张小勇, 等. 基于自校准机制的时空采样图卷积行为识别模型[J]. 工程科学学报, 2024, 46(3): 480–490. doi: 10.13374/j.issn2095-9389.2022.12.25.002.

CAO Yi, WU Weiguan, ZHANG Xiaoyong, *et al*. Action recognition model based on the spatiotemporal sampling graph convolutional network and self-calibration mechanism[J]. *Chinese Journal of Engineering*, 2024, 46(3): 480–490. doi: 10.13374/j.issn2095-9389.2022.12.25.002.

[5] SHEN Xiangpei and DING Yanrui. Human skeleton representation for 3D action recognition based on complex network coding and LSTM[J]. *Journal of Visual Communication and Image Representation*, 2022, 82: 103386. doi: 10.1016/j.jvcir.2021.103386.

[6] LIU Jun, WANG Gang, HU Ping, *et al*. Global context-aware attention LSTM networks for 3D action recognition[C]. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, 2017: 3671–3680. doi: 10.1109/CVPR.2017.391.

[7] XIA Rongjie, LI Yanshan, and LUO Wenhan. LAGA-Net: Local-and-global attention network for skeleton based action recognition[J]. *IEEE Transactions on Multimedia*, 2022, 24: 2648–2661. doi: 10.1109/TMM.2021.3086758.

[8] ZHANG Pengfei, LAN Cuiling, ZENG Wenjun, *et al*. Semantics-guided neural networks for efficient skeleton-based human action recognition[C]. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 2020: 1109–1118. doi: 10.1109/CVPR42600.2020.00119.

[9] SHI Lei, ZHANG Yifan, CHENG Jian, *et al*. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA, 2019: 12018–12027. doi: 10.1109/CVPR.2019.01230.

[10] CHEN Yuxin, ZHANG Ziqi, YUAN Chunfeng, *et al*. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, 2021: 13339–13348. doi: 10.1109/ICCV48922.2021.01311.

[11] 曹毅, 吴伟官, 李平, 等. 基于时空特征增强图卷积网络的骨架行为识别[J]. 电子与信息学报, 2023, 45(8): 3022–3031. doi: 10.11999/JEIT220749.

CAO Yi, WU Weiguan, LI Ping, *et al*. Skeleton action recognition based on spatio-temporal feature enhanced graph convolutional network[J]. *Journal of Electronics & Information Technology*, 2023, 45(8): 3022–3031. doi: 10.11999/JEIT220749.

[12] ZHU Yisheng, SHUAI Hui, LIU Guangcan, *et al*. Multilevel spatial-temporal excited graph network for skeleton-based action recognition[J]. *IEEE Transactions on Image Processing*, 2023, 32: 496–508. doi: 10.1109/TIP.2022.3230249.

[13] ZHOU Huanyu, LIU Qingjie, and WANG Yunhong. Learning discriminative representations for skeleton based action recognition[C]. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition

(CVPR), Vancouver, Canada, 2023: 10608–10617. doi: 10.1109/CVPR52729.2023.01022.

[14] WANG Kaixuan, DENG Hongmin, and ZHU Qilin. Lightweight channel-topology based adaptive graph convolutional network for skeleton-based action recognition[J]. *Neurocomputing*, 2023, 560: 126830. doi: 10.1016/j.neucom.2023.126830.

[15] YAN Sijie, XIONG Yuanjun, and LIN Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]. Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, USA, 2018: 7444–7452. doi: 10.1609/aaai.v32i1.12328.

[16] GEDAMU K, JI Yanli, GAO Lingling, *et al*. Relation-mining self-attention network for skeleton-based human action recognition[J]. *Pattern Recognition*, 2023, 139: 109455. doi: 10.1016/j.patcog.2023.109455.

[17] LIU Ziyu, ZHANG Hongwen, CHEN Zhenghao, *et al*. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 2020: 140–149. doi: 10.1109/CVPR42600.2020.00022.

[18] LI Yuxuan, HOU Qibin, ZHENG Zhaohui, *et al*. Large selective kernel network for remote sensing object detection[C]. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023: 16748–16759. doi: 10.1109/ICCV51070.2023.01540.

[19] XIA Yu, GAO Qingyuan, WU Weiguan, *et al*. Skeleton-based action recognition based on multidimensional adaptive dynamic temporal graph convolutional network[J]. *Engineering Applications of Artificial Intelligence*, 2024, 127: 107210. doi: 10.1016/j.engappai.2023.107210.

[20] AMIR S, LIU Jun, NG T T, *et al*. NTU RGB+D: A large scale dataset for 3D human activity analysis[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA: IEEE, 2016: 1010–1019. doi: 10.1109/CVPR.2016.115.

[21] LIU Jun, SHAHROUDY A, PEREZ M, *et al*. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 42(10): 2684–2701. doi: 10.1109/TPAMI.2019.2916873.

[22] PAN Qingzhe, ZHAO Zhifu, XIE Xuemei, *et al*. View-normalized and subject-independent skeleton generation for action recognition[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, 33(12): 7398–7412. doi: 10.1109/TCSVT.2022.3219864.

[23] 曹毅, 刘晨, 盛永健, 等. 基于三维图卷积与注意力增强的行为识别模型[J]. 电子与信息学报, 2021, 43(7): 2071–2078. doi: 10.11999/JEIT200448.

CAO Yi, LIU Chen, SHENG Yongjian, *et al*. Action recognition model based on 3D graph convolution and attention enhanced[J]. *Journal of Electronics & Information Technology*, 2021, 43(7): 2071–2078. doi: 10.11999/JEIT200448.

曹 毅：男，教授，博士，研究方向为机器人机构学、深度学习.
李 杰：男，硕士生，研究方向为深度学习、行为识别.
叶培涛：男，硕士生，研究方向为机器人控制系统、路径规划.
王彦雯：男，硕士生，研究方向为深度学习、声纹识别.
吕贤海：男，硕士生，研究方向为机器人机构学、行为识别.

责任编辑：陈 倩

Skeleton-based Action Recognition with Selective Multi-scale Graph Convolutional Network

CAO Yi LI Jie YE Peitao WANG Yanwen LÜ Xianhai

(*School of Mechanical Engineering, Jiangnan University, Wuxi* 214122, *China*)

(*Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology, Jiangnan University, Wuxi* 214122, *China*)

Abstract:

Objective Human action recognition plays a key role in computer vision and has gained significant attention due to its broad range of applications. Skeleton data, derived from human action samples, is particularly robust to variations in camera viewpoint, illumination, and background occlusion, offering advantages over depth image and video data. Recent advancements in skeleton-based action recognition using Graph Convolutional Networks (GCNs) have demonstrated effective extraction of the topological relationships within skeleton data. However, limitations remain in some current approaches employing GCNs: (1) Many methods focus on the

discriminative dependencies between pairs of joints, failing to effectively capture the multi-scale discriminative dependencies across the entire skeleton. (2) Some temporal modeling methods use dilated convolutions for simple feature fusion, but do not employ convolutional kernels in a manner suitable for effective temporal modeling. To address these challenges, a selective multi-scale GCN is proposed for action recognition, designed to capture more joint features and learn valuable temporal information.

Methods The proposed model consists of two key modules: a multi-scale graph convolution module and a selective multi-scale temporal convolution module. First, the multi-scale graph convolution module serves as the primary spatial modeling component. It generates a multi-scale, channel-wise topology refinement adjacency matrix to enhance the model's ability to learn multi-scale discriminative dependencies of skeleton joints, thereby capturing more joint features. Specifically, the pairwise joint adjacency matrix is used to capture the interactive relationships between joint pairs, enabling the extraction of local motion details. Additionally, the multi-joint adjacency matrix emphasizes the overall action feature changes, improving the model's spatial representation of the skeleton data. Second, the selective multi-scale temporal convolution module is designed to capture valuable temporal contextual information. This module comprises three stages: feature extraction, temporal selection, and feature fusion. In the feature extraction stage, convolution and max-pooling operations are applied to obtain temporal features at different scales. Once the multi-scale temporal features are extracted, the temporal selection stage uses global max and average pooling to select salient features while preserving key details. This results in the generation of temporal selection masks without directly fusing temporal features across scales, thus reducing redundancy. In the feature fusion stage, the output temporal feature is obtained by weighted fusion of the temporal features and the selection masks. Finally, by combining the multi-scale graph convolution module with the selective multi-scale temporal convolution module, the proposed model extracts multi-stream data from skeleton data, generating various prediction scores. These scores are then fused through weighted summation to produce the final prediction outcome.

Results and Discussions Extensive experiments are conducted on two large-scale datasets: NTU-RGB+D and NTU-RGB+D 120, demonstrating the effectiveness and strong generalization performance of the proposed model. When the convolution kernel size in the multi-scale graph convolution module is set to 3, the model performs optimally, capturing more representative joint features (Table 1). The results (Table 4) show that the temporal selection stage is critical within the selective multi-scale temporal convolution module, significantly enhancing the model's ability to extract temporal contextual information. Additionally, ablation studies (Table 5) confirm the effectiveness of each component in the proposed model, highlighting their contributions to improving recognition performance. The results (Tables 6 and 7) demonstrate that the proposed model outperforms state-of-the-art methods, achieving superior recognition accuracy and strong generalization capabilities.

Conclusions This study presents a selective multi-scale GCN model for skeleton-based action recognition. The multi-scale graph convolution module effectively captures the multi-scale discriminative dependencies of skeleton joints, enabling the model to fully extract more joint features. By selecting appropriate temporal convolution kernels, the selective multi-scale temporal convolution module extracts and fuses temporal contextual information, thereby emphasizing useful temporal features. Experimental results on the NTU-RGB+D and NTU-RGB+D 120 datasets demonstrate that the proposed model achieves excellent accuracy and robust generalization performance.

Key words: Skeleton-based action recognition; Graph Convolutional Network (GCN); Multi-scale channel-wise topology refinement adjacency matrix; Selective multi-scale temporal convolution; Selective multi-scale graph convolutional network