DOI:10.19651/j.cnki.emt.2518483

基于跨模态协同感知的双流融合动作识别模型*

刘 罡[1,2,3] 李小雨[2] 吴 烨[2] 郑泽林[2]
(1.无锡学院集成电路科学与工程学院 无锡 214105;2.南京信息工程大学电子与信息工程学院 南京 210044;
3.无锡学院江苏省集成电路可靠性技术及检测系统工程研究中心 无锡 214105)

摘 要:针对现有动作识别算法中时空特征融合不充分及丰富的骨架信息未能得到充分利用等问题,本文提出一种基于跨模态协同感知的双流融合动作识别模型。首先,本文提出一种双流融合模型,通过融合 RGB 视频流和骨架流,获取两个模块的全局信息,实现优势互补;然后,提出时空交互注意力模块,实现了时空特征的深度协同与动态互补,动态增强相关时空区域的注意力权重;最后,设计出一种多模态特征融合模块,将通过 RGB 视频流和骨架流的输出进行特征融合增强,通过自适应权重分配与跨模态交互,充分挖掘 RGB 视觉外观与人体骨骼运动间的互补信息,从而提升动作识别准确率。多组实验结果表明,该双流融合动作识别模型在 NTU RGB+D 和 NTU RGB+D 120 数据集上实现了高精度的动作识别,分别获得 97.2%和 92.3%的准确率,与基线方法 MMTM 相比,精度分别提高了 3.6%和 3.2%。通过结果表明,该模型可以充分提取利用人体骨架信息,同时充分融合时空特征,提高对动作识别的准确率。

关键词:时空注意力机制;特征融合;动作识别;多模态
中图分类号:TP391.41;TN914 **文献标识码**:A **国家标准学科分类代码**:520.20

Dual-stream fusion action recognition model based on cross-modal co-sensing

Liu Gang[1,2,3] Li Xiaoyu[2] Wu Ye[2] Zheng Zelin[2]
(1. School of Integrated Circuit Science and Engineering, Wuxi University, Wuxi 214105, China;
2. School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China;
3. Jiangsu Province Engineering Research Center of Integrated Circuit Reliability Technology and Testing System, Wuxi 214105, China)

Abstract: Aiming at the problems of insufficient spatio-temporal feature fusion and failure to fully utilize the rich skeleton information in the existing action recognition algorithms, this paper proposes a dual-stream fusion action recognition model based on cross-modal synergetic perception. Firstly, this paper proposes a dual-stream fusion model, which obtains the global information of the two modules by fusing the RGB video stream and the skeleton stream, realizing the complementary advantages; proposes a spatio-temporal interaction and attention enhancement module, which realizes the in-depth synergistic and dynamic complementarity of spatio-temporal features and dynamically enhances the attention weight of the relevant spatio-temporal region; and finally, designs a Multimodal Feature Fusion Module. Feature Fusion Module, which will be enhanced by feature fusion through the outputs of RGB video streams and skeleton streams, and fully exploits the complementary information between RGB visual appearance and human skeleton motion through adaptive weight assignment and cross-modal interaction, so as to improve the accuracy of action recognition. The results of multiple sets of experiments show that this CC-DFARM achieves high accuracy on the NTU RGB+D and NTU RGB+D 120 datasets of action recognition, obtaining 97.2% and 92.3% accuracies, respectively, and improving the accuracy by 3.6% and 3.2% compared to the baseline method MMTM. The results show that the model can fully extract and utilize the human skeleton information, and at the same time fully integrate the spatio-temporal features to improve the accuracy of action recognition.

Keywords: spatio-temporal attention mechanism; feature fusion; action recognition; multimodality

0 引 言

随着人工智能技术的快速发展,动作识别作为计算机视觉领域的核心课题,在智能监控[1]、人机交互[2]、运动分析[3]、自动视频跟踪[4]等场景中展现出重要价值。传统方法多基于单一模态(如 RGB 视频或骨骼序列)进行动作表

收稿日期:2025-03-31
*基金项目:国家自然科学基金(62204172)项目资助

征，然而真实场景中人体动作具有显著的时空关联性与多模态互补特性。例如，RGB 视频可捕捉表观细节但易受背景干扰，骨骼数据虽能准确描述运动学特征却丢失了物体交互信息（如持握工具的姿态）。如何建立跨模态时空协同机制，实现多源信息的动态融合与高效推理，已成为提升动作识别系统实用性的关键挑战。

近年来，双流网络架构通过并行处理时空特征，在动作识别任务中取得显著进展。Yan 等[5]开创性地将图卷积网络（GCN）引入人体动作分析任务，其研究通过设计基于骨骼节点距离的动态采样机制，构建了具有空间适应性的图卷积操作。Li 等[6]创新性地引入注意力驱动机制改进图卷积架构，设计出具有动作结构感知能力的图卷积网络（AS-GCN）。该方法通过动态注意力权重分配策略，自适应聚焦关键关节的空间拓扑关系与时间演化模式，有效解决了传统图卷积在长程依赖建模中的感受野受限问题。Shi 等[7]开发了关节-骨骼异构特征协同的双流学习架构。该方法通过设计具有拓扑动态感知机制的图卷积模块，实现关节几何关系与骨骼运动模式的双流特征交互：一方面利用关节坐标刻画人体姿态的静态结构，另一方面通过骨骼向量编码肢体运动的动态约束。该框架突破了传统固定拓扑图卷积的建模局限，在提升时空特征判别力的同时保持参数效率。Wang 等[8]对比了 RGB 数据和骨架数据在人体动作识别的方法，并简要介绍了提取骨架的姿态估计算法。Tasnim 等[9]将边缘卷积机制引入时空图神经网络架构，设计出具有动态拓扑感知能力的卷积运算模块。该方法通过联合建模空间域关节关联与时序维运动连续性，构建帧内局部结构约束与帧间全局演化关联的双重特征交互机制，最终形成动态边缘卷积网络（Dynamic Edge-ConvNet），有效增强了复杂动作模式下时空特征的判别性表达。Joze 等[10]提出了一个多模块传输模块可以直接添加到特征层的不同层次，实现慢速模态融合，利用压缩和激励操作重新校准每个 CNN 流中的信道特征。与其他中间融合方法不同，该模型可用于不同空间维度卷积层的特征模态融合。

然而，上述动作识别算法仍面临许多问题。首先是时空建模割裂，多数研究采用“空间卷积＋时序池化”的串行结构导致运动细节在分离处理过程中丢失，难以捕捉短时微动作，其次是跨模态协同低效，RGB 与骨骼等异构模态的时空特征尺度差异显著，直接级联或加权融合易引发特征冲突，尤其在复杂场景下易产生错误注意力聚焦，基于单模态的图卷积网络虽能通过局部关节建模提升计算效率，却难以捕捉跨关节的全局时空依赖（如 NTU 数据集中的“两人交互”动作）。

本文针对上述问题，提出一种基于跨模态协同感知的双流融合动作识别模型。为了评估该模型与上述方法相比带来的有效性，实验在两个大型数据集 NTU RGB＋D[11]和 NTU RGB＋D 120 数据集上进行，通过结果表明，本文提出模型可以充分提取利用人体骨架信息，同时充分融合时空特征，提高对动作识别的准确率，在人体动作识别方面的准确率具有较大优势。

1　相关工作

1.1　基于 CNN 的共现特征学习

卷积神经网络（CNN）作为深度学习领域的基石模型，其核心设计思想源于对生物视觉系统的仿生学启发。通过局部感知野（Local Receptive Fields）、权值共享（Weight Sharing）和层次化特征抽象（Hierarchical Feature Extraction）三大机制，CNN 在计算机视觉任务中展现出独特的优势。相较于循环神经网络（RNN）等序列模型，CNN 的并行化架构使其能更高效地捕获空间-时间联合表征，从而在图像、视频等多模态数据分析中占据主导地位。通过研究卷积计算，我们可以将它分解为两个步骤如图 1 所示。首先，a 部分是从每个输入通道的空间域中独立 2D 卷积，其中特征是从 3×3 邻域局部聚合而来。b 部分是跨通道的元素聚合，输出是从所有的输入通道全局聚合特征而得出。共现特征本质上描述了数据中空间、通道或时序维度上具有统计关联性的模式组合，例如图像中相邻纹理的共生规律、视频中动作片段间的因果关联等。CNN 的核心优势在于将传统卷积运算分解为空间局部聚合与通道全局整合两个阶段，从而实现对共现特征的显式学习。

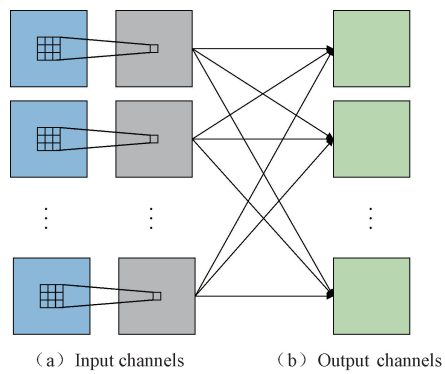

图 1　3×3 卷积分解成两 a、b 两步

Fig. 1　3×3 convolutional decomposition into two a and b steps

早期研究，Du 等[12]普遍采用关节坐标通道化编码方案，即将人体关节点的二维/三维空间坐标直接映射为输入特征图的通道维度。这种设计虽然能够通过卷积操作捕捉局部相邻关节的共现模式，但其本质受限于卷积核的局部感受野特性，难以建模跨肢体远端关节的长期功能关联。为此，聚集全球共现功能是非常重要的，并导致更好的动作识别性能。它可以通过将联合维度放入 CNN 输入的通道中来轻松实现。

1.2　骨架运动

针对骨架序列动作识别的时空特征解耦问题，关节的时间运动是识别潜在动作的关键线索。虽然时间演化模式

可以通过 CNN 隐式学习,但显式运动表征策略更可取。基于骨架数据共现特征学习的分层聚合动作识别与检测,引入骨架运动的表示,并将其显式地馈送到网络中。

首先,定义骨架序列为时变图结构,对于帧数为 t 的人体骨架,如式(1)所示。

$$S_t = \{J_t^i\}_{i=1}^N, J_t^i \in \mathbb{R}^3(x,y,z) \tag{1}$$

其中,N 为关节的数量,J 是 3D 关节坐标点,骨架运动是连续的两个关键帧的每个关节时间差,如式(2)所示。

$$M^t = S^{t+1} - S^t = \{J_1^{t+1} - J_1^t, J_2^{t+1} - J_2^t, \cdots, J_N^{t+1} - J_N^t\} \tag{2}$$

为了融合两个关键帧带来的两个源信息,在网络的后续层中实现跨通道连接它们的特征图,如图 2 所示,原始骨架坐标 S 和骨架运动 M 以双流特征被独立的输入到网络中。

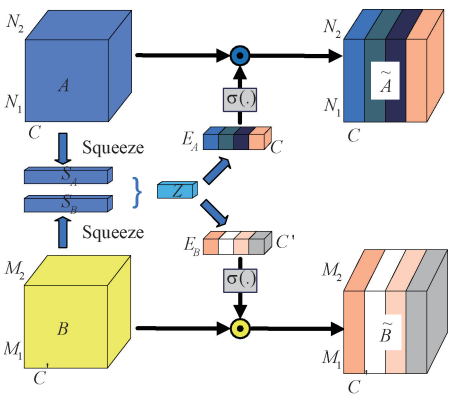

图 2　双模态 MMTM 架构

Fig. 2　Architecture of MMTM for two modalities

1.3　双模态 MMTM 架构

在多模式传输模型(multimodal transfer module,MMTM)架构中,首先讨论两个不相交的 CNN 流 CNN1 和 CNN2 之间融合的最简单情况。设 $A \in \mathbb{R}^{N_1 \times \cdots \times N_K \times C}$,$B \in \mathbb{R}^{M_1 \times \cdots \times M_L \times C'}$ 分别表示 CNN1 和 CNN2 的给定层的特征。其中,N_i 和 M_i 表示空间维度,C 和 C' 分别表示 CNN1 和 CNN2 中相应特征的通道数。MMTM 接收特征 A 和 B 作为输入,学习全局多模态嵌入,并使用该嵌入来重新校准输入特征。具体通过多模态挤压和激励过程完成。

卷积神经网络中,常规卷积层生成的特征图受限于局部感受野范围,难以捕获全局语义信息。首先通过全局均值池化操作将二维空间特征压缩为一维通道特征向量,如式(3)、(4)所示,将输入特征 A 和 B 挤压成 S_A 和 S_B。

$$S_A(c) = \frac{1}{\prod_{i=1}^{K} N_i} \sum_{n_1,\cdots,n_k} A(n_1,\cdots,n_K,c) \tag{3}$$

$$S_B(c) = \frac{1}{\prod_{i=1}^{L} M_i} \sum_{m_1,\cdots,m_L} A(m_1,\cdots,m_L,c) \tag{4}$$

这种特征压缩策略具有两个关键优势:其一,通过降维处理有效整合全局空间上下文信息;其二,其空间无关性特性使其能适配不同分辨率的多源数据融合场景。此方法虽采用基础池化方法,但该框架具备良好的扩展性,可兼容最大池化、注意力池化等进阶特征聚合方式。

多模态激励单元的功能是产生激励信号 E_A 和 E_B,其可用于通过简单的选通机制来重新效准输入特征 A 和 B,如式(5)、(6)所示。

$$\tilde{A} = 2 \times \sigma(E_A) \odot A \tag{5}$$

$$\tilde{B} = 2 \times \sigma(E_B) \odot B \tag{6}$$

其中,$\sigma(\cdot)$ 是 Sigmoid 函数,$\odot$ 是乘积运算,从而允许抑制或激励每个流中的不同滤波器。MMTM 的权重会被正则化以便将 E_A 和 E_B 的接近度控制位 0,增加 E_A 的正则化权重会让门控信号更接近单位向量,从而限制了门控对特征 A 的影响。

门控信号必须基于相同的输入表示将不同的校准权重应用于不同的模态。我们通过首先从压缩信号预测联合表示 $Z \in \mathbb{R}^{C_Z}$ 来实现这一点,如式(7)所示。

$$Z = W[S_A, S_B] + b \tag{7}$$

然后通过两个独立的全连接层预测每个模态的激励信号,如式(8)所示。

$$E_A = W_A Z + b_A, E_B = W_B Z + b_B \tag{8}$$

以这种方式学习联合表示允许一个模态的特征重新校准另一模态的特征。在动作识别中,当动作在 RGB 相机中模糊并且在骨骼模态中更明显时,MMTM 跨模态重新校准在 RGB 流中提供更有效的处理。如图 2 总结了所提出的 MMTM 的总体架构。

2　算法研究

2.1　模型改进

本文是以 MMTM 为基线网络做算法改进。MMTM 框架基本设计是通过多模态传输模块将不同模态的数据融合,但它只在有限的层次上进行模态间的信息传递,不同模态之间的信息共享是关键,尤其在动作识别任务中。针对此问题,本文提出跨模态协同感知的双流融合架构,通过三级融合机制实现多层次特征交互。首先,设计多模态联合嵌入模块(dual stream fusion module framework,DSFM),将骨架数据的关节拓扑结构编码为可学习的图卷积特征,与 RGB 视频流进行通道维度对齐,为后续跨模态交互奠定基础。在 MMTM 框架中,其中骨骼流选取的为 HCN[13] 网络,HCN 网络通常侧重于全局上下文的建模,但对于视频数据或动作识别等任务,时序信息至关重要。HCN 本身可能没有专门的机制来捕捉跨时间步的动态变化,同时 HCN 模型在跨模态融合时,依赖于显式的模态共享表示或交叉注意力机制,虽然这能够有效结合不同模态的信息,但这种融合方式仍然可能受到每个模态信息的质量和数据对齐问题的限制。针对此问题,本文在骨骼流 HCN 网络特征提

取阶段提出时空交互增强注意力模块(temporospatial interactive attentive module,TIAM),该模块通过时空交叉注意力机制建立跨模态动态关联,在空间维度上,通过骨骼关键点热力图引导 RGB 特征聚焦于人体核心区域,在时间维度上,利用光流特征构建时序注意力权重,强化骨架序列中肢体运动的关键相位特征。这种双向注意力机制使得模型能够自适应增强具有判别性的时空区域,抑制背景噪声干扰,有效提升了模型在复杂环境中的表现和准确性。在 MMTM 框架中,原作者采取 RGB 视频流和 HCN 骨架流的双流模态数据融合,RGB 视频流通常包含丰富的时空信息,但由于动作识别任务中对细粒度时序特征的要求,单一的视频流可能难以提供足够的时空细节。另一方面,骨架流(通常由关节点组成的骨架序列)虽然能有效表示人体的运动模式,但缺少环境背景信息和一些细节。同时,尽管 MMTM 设计了模态传输模块来实现跨模态信息流动,但这种信息流动可能在某些任务中仍不够高效。尤其是对细粒度动作的识别来说,模态之间的交互需要更深入、更高效的处理方式。因此,针对这一问题,本文设计出多模态特征融合模块(multimodal feature fusion module,MFFM),将通过 RGB 视频流和骨架流的输出最后进行特征融合增强,通过自适应权重分配与跨模态交互,充分挖掘 RGB 视觉外观与人体骨骼运动间的互补信息,从而提升动作识别准确率。

2.2　算法模型

基于跨模态协同感知的双流融合动作识别模型(dual-stream fusion action recognition model based on cross-modal co-sensing,CC-DFARM)的整体结构如图 3 所示,本文在 MMTM 框架上进行优化设计,其骨干网络采用 DSFM

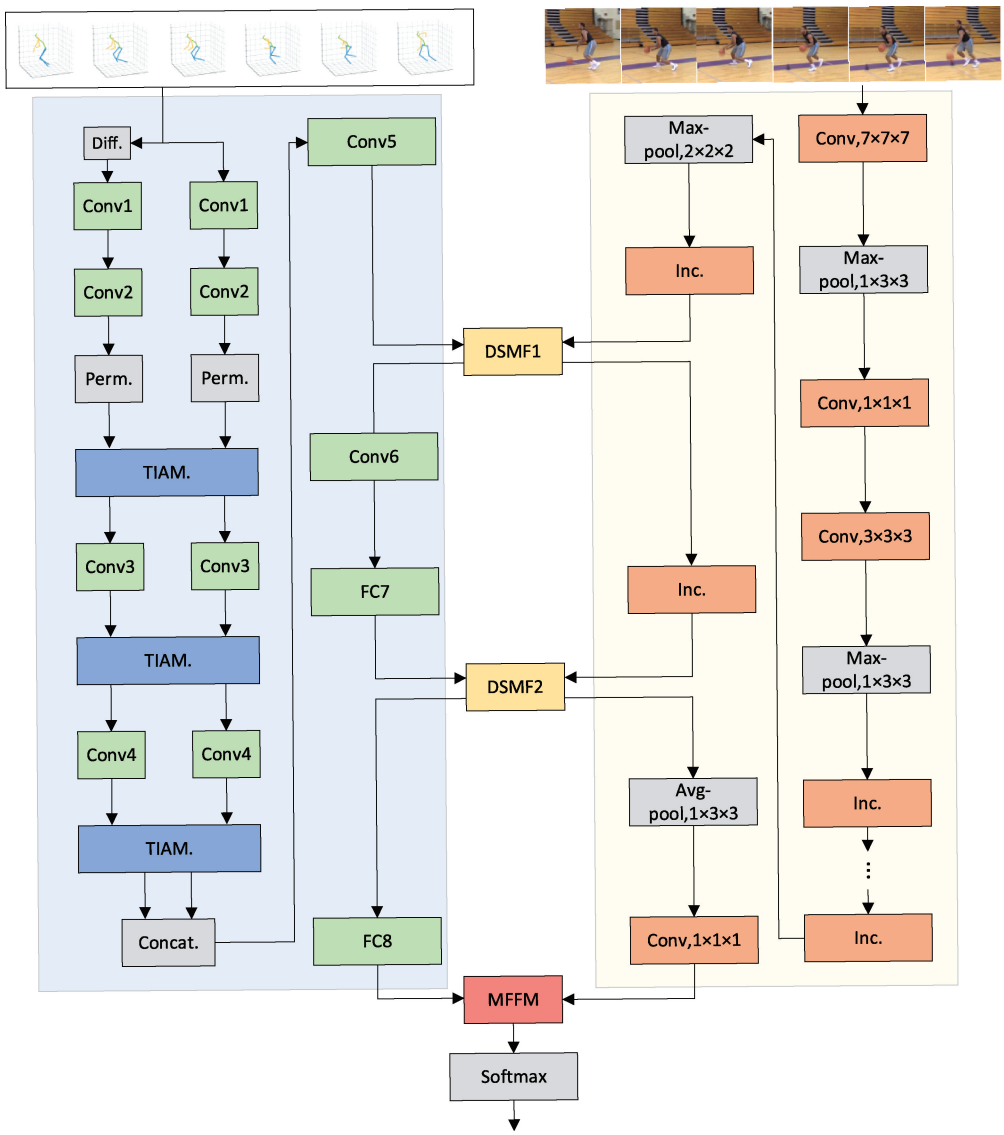

图 3　CC-DFARM 网络框架

Fig. 3　CC-DFARM network framework

模块将用于视频数据的 I3D[14] 模型与用于骨架流的经过本文改进的 HCN 模型融合。此设计的目的在于解决单模态表征瓶颈,HCN 仅依赖骨架数据,而骨架信息易受传感器噪声、遮挡或视角变化影响,单一模态难以全面表征复杂动作的时空语义。在本网络中,部署了两个 DSFM 模块,一个在两个网络中的中间级别,另一个用于高级功能重新校准。第 1 个 DSFM 模块每个块有 64 个通道,把它插入到 I3D 中倒数第 2 个 Inception 层和改进后的 HCN 的 conv5 之间。第 2 个 DSFM 每个数据块有 256 个通道,把它插入到 I3D 的最后一个 Inception 层和改进后的 HCN 中的 FC7 层之间。DSFM 在骨干网络阶段融合,迫使两类特征在深层编码时即相互校正,形成更具判别性的联合表征。同时实现了多尺度时空协同,通过动态特征校准,抑制了模态冲突与噪声传播。其次,本文提出时空交互增强注意力模块,实现了时空特征的深度协同与动态互补,动态增强相关时空区域的注意力权重。解决了传统动作识别模型中时空特征融合浅层化、注意力权重固化等问题。最后,设计一种多模态特征融合模块,通过 RGB 视频流和骨架流的输出进行特征融合增强,提升动作识别准确率。

2.3　双流融合模块(DSFM)

针对现有算法单模态表征瓶颈,单一模态难以全面表征复杂动作的时空语义等问题,本文设计双流融合模块 DSFM,如图 4 所示。

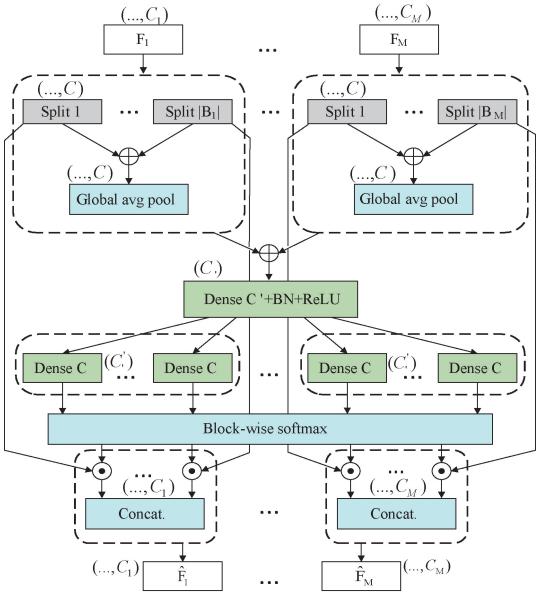

图 4　双流融合模块框架

Fig. 4　Dual stream fusion module framework

首先,设定模态数为 M,本文中为 2,分别为 RGB 视频流和骨架流,采用动态特征优化机制实现跨模态信息交互。模态 $m \in \{1,2,\cdots,M\}$ 的特征映射为 $F_m \in \mathbb{R}^{N_1\times N_2\times\cdots\times N_K\times C_m}$。此处,$K$ 是模态 m 的空间维度的数量,C_m 是模态 m 中的通道的数量。DSFM 模块通过获取特征图 $\{F_1,\cdots,F_M\}$ 并生成由对应的每个通道通过注意力模块优化后的特征图 $\{\hat{F}_1,\cdots,\hat{F}_M\}$。该模块分为 3 个递进阶段。

1)通道分治策略

我们首先对各模态特征实施通道维度剖分,生成等通道特征子块集合。其中每个块中的通道数为 C。我们将属于模态 m 的特征块的集合表示为 B_m,其中 $|B_m|=\lceil C_m/C\rceil, m\in\{1,\cdots,M\}, B_m^i$ 是 B_m 中的第 i 个特征块,$i\in\{1,\cdots,|B_m|\}$。

2)跨模态特征聚合

通过双路特征融合构建全局上下文表征,首先对跨模态特征 B_m 实施逐点加和运算生成混合特征 S_m,经空间维度全局均值压缩后获得模态特征描述符融合操作时学习多模态全局上下文的关键步骤,多模态全局上下文用于生成每个通道的块式注意。将属于模态 m 的块加入到共享表示 D_m 中,如式(9)所示。

$$D_m(c)=\frac{1}{\prod_{i=1}^{K}N_i}\sum_{(n_1,\cdots,n_K)}S_m(n_1,n_2,\cdots,n_K,c) \tag{9}$$

每个通道描述符都是一个长度为 C 的特征向量,它概括了一种模式内的特征块。为融合跨模态信息,对全部模态描述符执行逐元素叠加,来形成多模态信道描述符 G。然后,通过压缩比为 r 的维度变换层(含全连接、批归一化及 ReLU 激活)来建模通道间依赖关系。该变换将 G 映射到联合表示 $Z\in\mathbb{R}^{C'}, C'=\lfloor C/r\rfloor$,这有助于复杂模型的推广。其中 Z 的表示如式(10)所示。

$$Z=W_ZG+b_Z \tag{10}$$

其中,$W_Z\in\mathbb{R}^{C'}, b_Z\in\mathbb{R}^{C'}$。

3)动态特征强化

跨模态联合描述符 Z 承载着全局上下文语义信息。针对特定特征子块 B_m^i,通过线性投影与归一化处理实现注意力权重学习,首先将联合描述符映射为中间特征 U_m^i,经 Softmax 函数计算得到通道注意力分布,如式(11)、(12)所示。

$$U_m^i=W_m^iZ+b_m^i \tag{11}$$

$$A_m^i=\frac{\exp(U_m^i)}{\sum_{k}^{M}\sum_{j}^{|B_k|}\exp(U_k^j)} \tag{12}$$

其中,投影矩阵 $W_m^i\in\mathbb{R}^{C\times C'}$ 和偏置项 $b_m^i\in\mathbb{R}^C$ 构成可学习参数。为缓解多块竞争导致的特征抑制现象,设计平衡系数 $\lambda\in[0,1]$ 实现注意力强度调节,通过凸组合形式生成优化特征 $\hat{B}i_m$,如式(13)所示。

$$\hat{B}_m^i=[\lambda+(1-\lambda)\times A_m^i]\odot B_m^i \tag{13}$$

最终沿通道维度整合各子块,得到重构增强后的模态特征 $\hat{F}_m$,如式(14)所示。

$$\hat{F}_m=[\hat{B}_m^1,\hat{B}_m^2,\cdots,\hat{B}_m^{|B_m|}] \tag{14}$$

2.4　时空交互注意力模块(TIAM)

为了解决传统动作识别模型中时空特征融合浅层化、注意力权重固化等问题，本文设计时空交互注意力模块TIAM，如图 5 所示。本方法的核心在于构建双时序特征的 Gram 矩阵交互机制，通过矩阵运算同步推导时空注意力权重。具体地，定义空间视角依赖度 Y_2 与时间风格关联度 Y_1。Y_2 通过语义一致性约束增强目标表征鲁棒性，有效抑制场景误检，类似地，Y_1 量化特征流间的风格漂移程度，缓解背景特征干扰。

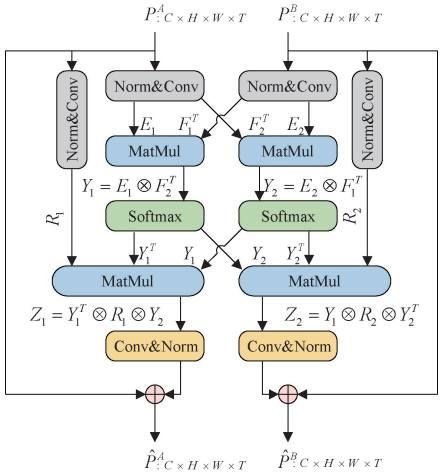

图 5　时空交互注意力模块

Fig. 5　Temporospatial interactive attentive module

首先，针对输入特征对 $\{P^A, P^B\} \in \mathbb{R}^{C\times H\times W\times T}$，执行以下层次化处理。将四维张量维度 $\mathbb{R}^{C\times H\times W\times T}$ 重构成 $\mathbb{R}^{C\times HW\times T}$，采用批标准化与 1×1 卷积核的线性变换进行特征校准，抑制梯度异常。其中，R_1 和 R_2 共享参数，E_1 和 E_2 不共享参数，具体如公式(15)所示。

$$\begin{cases} R_i = Conv^{(1)}(BN(P')) \\ E_i = Conv_i^{(1)}(BN(P')) \\ F_i = E_i^T \end{cases} \tag{15}$$

其中，$(i,t) \in \{(1,A),(2,B)\}$，下标 i 区分特征来源。接着，通过嵌入式高斯核计算跨模态注意力权重，如式(16)所示。

$$Y_i = \mathrm{Softmax}(E_i \otimes F_j^{\mathrm{T}}) \tag{16}$$

其中，$(i,j) \in \{(1,2),(2,1)\}$，$T_1$ 表示时序风格关联度，量化跨特征通道的语义一致性，反映时间维度的风格迁移特性。T_2 表示空间视角依存度，评估跨空间位置的几何相关性，捕捉视角变化的鲁棒表征。然后，通过双路注意力权重实现特征重建，将 T_1 和 T_2 在时间和空间上加权重建 R_i，具体如式(17)所示。

$$\begin{cases} Z_1 = Y_1^{\mathrm{T}} \otimes R_k \otimes Y_2 \\ Z_2 = Y_1 \otimes R_k \otimes Y_2^{\mathrm{T}} \end{cases} \tag{17}$$

其中，$\{Y_1, Y_2\}$ 中依据特征类型动态选择转置操作，实现时空维度的自适应对齐。重建后的特征 Z_k 经反卷积与批标准化恢复原始维度，最终通过残差连接增强特征表示，得到输出 $\hat{P}'$，具体如式(18)所示。

$$\hat{P}t = BN(Conv^{(1)}(Z_k)) \oplus P' \tag{18}$$

2.5　时空交互注意力模块与 HCN 的集成

针对层次化卷积网络 HCN 的特征学习特性，本文提出将时空交互注意力模块 TIAM 嵌入至网络的高层次特征聚合阶段，如图 6 所示。传统 HCN 框架通过双阶段架构实现动作表征学习：Stage 1 聚焦于关节级别的局部时空模式挖掘，而 Stage 2 通过特征变换层将局部特征升维至全局共现特征空间，从而捕捉跨关节的长程依赖关系与动作语义关联。

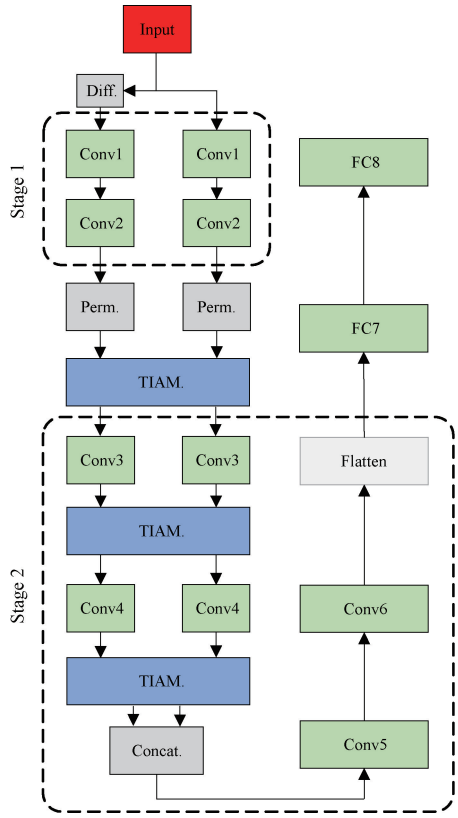

图 6　时空交互注意力模块与 HCN 的集成

Fig. 6　Integration of the spatio-temporal interactive attention module with HCN

理论分析表明，时空交互注意力机制的有效性与特征抽象层级密切相关：在低层关节特征(Stage 1)中，关节运动模式受局部噪声干扰较大(如肢体遮挡、传感器抖动)，过早引入注意力模块可能导致模型过度关注低信噪比区域；而在经过特征变换层后的高层次特征(Stage 2)中，全局共现特征已具备较强的语义结构化表达，此时通过 TIAM 模块可动态强化关键时空节点的权重分配，同时抑制冗余或冲突特征。基于此，本文将 TIAM 模块置于 HCN 的 Transition Layer 之后，构建“全局特征增强-时空权重自适应”的双向优化机制。

2.6　多模态特征融合模块(MFFM)

为了将通过 RGB 视频流和骨架流的输出进行特征融合增强,本文提出多模态特征融合模块 MFFM,如图 7 所示。

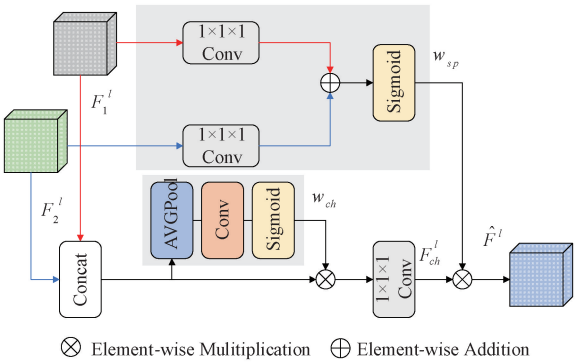

图 7　多模态特征融合模块

Fig. 7　Multimodal feature fusion module

首先,将层级特征对 $F_1^l \in \mathbb{R}^{C\times H\times W\times D}$ 和 $F_2^l \in \mathbb{R}^{C\times H\times W\times D}$ 沿通道维度进行深度融合,构建联合特征张量 $F^l \in \mathbb{R}^{2C\times H\times W\times D}$,具体如式(19)所示。

$$F^l = \mathrm{Concat}([F_1^l; F_2^l]) \tag{19}$$

该操作有效保留双流特征的互补性信息,为后续特征筛选提供丰富输入。为避免传统 1×1×1 卷积导致的信息损失,提出上下文感知的通道压缩策略。MFFM 中的通道缩减不是简单地使用 1×1×1 卷积,而是由全局通道信息 w_{ch} 引导的。w_{ch} 是通过三维全局均值池化(AVGPool)捕获跨时空统计量,经卷积层与 Sigmoid 激活生成通道注意力权重,具体如式(20)所示。

$$w_{ch} = \mathrm{Sigmoid}(\mathrm{Conv}_1(\mathrm{AVGPool}(F^l))) \tag{20}$$

接着,利用注意力权重对拼接特征实施软选择,抑制冗余通道同时增强关键特征响应,通过全局通道信息校准融合特征,使用 1×1×1 卷积层,该信道信息将引导卷积层保留重要的特征 $F_{ch}^l \in \mathbb{R}^{C\times H\times W\times D}$,如式(21)所示。

$$F_{ch}^l = \mathrm{Conv}_1(w_{ch} \otimes F^l) \tag{21}$$

为了对局部特征图之间的空间依赖性进行建模,全局空间信息 w_{sp} 由 1×1×1 卷积层(Conv1)和来自特征图 F_1^l 和 F_2^l 的 Sigmoid 激活捕获,如式(22)所示。这些信息用于校准特征图,并强调突出的空间区域,最终得到输出特征 $\hat{F}^l = w_{sp} \otimes F_{ch}^l$,如式(23)所示。

$$w_{sp} = \mathrm{Sigmoid}(\mathrm{Conv}_1(F_1^l) \oplus \mathrm{Conv}_1(F_2^l)) \tag{22}$$

$$\hat{F}^l = w_{sp} \otimes F_{ch}^l \tag{23}$$

3　实验设计与结果分析

3.1　数据集

NTU RGB-D 和 NTU RGB-D 120 是当前基于骨架动作识别领域最具代表性的两大基准数据集。NTU RGB-D 于 2016 年发布,共包含 56 880 个 3D 骨架序列,涵盖 40 名受试者执行的 60 类动作类别。其动作内容既包含单人日常行为(如挥手、跌倒、行走等),也涵盖双人交互动作(如拥抱、递物、击掌等)。数据采集通过微软 Kinect v2 深度传感器完成,同时记录了 RGB 视频、深度图、红外序列和 3D 骨架坐标信息。该数据集提出两种标准化评估协议:跨对象(Cross-Subject)将 40 名受试者按 20∶20 划分为训练集与测试集;跨视角(Cross-View),采用 3 台不同视角摄像头(侧视、前视、斜视)的数据进行划分,其中两个视角数据用于训练,剩余视角用于测试,有效验证算法对视角变化的鲁棒性。

作为 NTU RGB-D 的扩展版本,NTU RGB-D 120 用于 3D 动作识别。它由 114 480 个骨架序列组成,跨越 120 个动作类,由 106 名参与者在 32 个不同的相机设置中执行。该数据集包括两个评估协议:交叉受试者(CS),其中一半用于训练,其余一半用于测试,以及交叉设置(CE),其中来自奇数 ID 设置的序列用于训练,而来自奇数 ID 设置的序列用于训练。ID 为偶数的设置被保留用于测试。

3.2　实验设置

本文的所有实验均基于 PyTorch 框架上开展,具体环境配置如表 1 所示。

表 1　实验环境

Table 1　Experimental environment

配置名称	版本参数
GPU	NVIDIA GeForce RTX 4060ti
CPU	Intel Core i5-12490F@3 GHz
深度学习框架	PyTorch1.11.0
编程语言	Python3.9
CUDA	Cuda11.4

本文实验通过在 NTU RGB-D 与 NTU RGB+D 120 数据集上进行模型训练和评估。实验中的参数配置如下:首先,本实验设置了 50 轮的训练,确保模型能够充分学习数据特征并达到最佳性能。在优化策略选择上,本实验选取随机梯度下降(SGD),将初始学习率定为 0.1,以避免梯度爆炸或梯度消失问题,确保训练稳定进行,采用余弦退火率策略调整学习率。为了提高计算效率和增强模型的泛化能力,本文选择了批量大小为 32,这样可以兼顾训练速度和稳定性。

3.3　实验结果分析

1)各模型实验对比

为了验证本文提出的跨模态协同感知的双流融合动作识别模型在动作识别准确度上的显著提升,本研究选择在 NTU RGB+D 60 与 NTU RGB+D 120 两个大规模动作识别数据集上做各模型实验仿真,结果如表 2 所示。

CC-DFARM 在跨主体(cross-subject, CS)和跨视角

表 2　各模型在 NTU 60、NTU120 数据集上的实验结果

Table 2　Experimental results of each model on NTU 60, NTU 120 datasets

Model	NTU 60		NTU 120	
	CS/%	CV/%	C-sub/%	C-set/%
PCRP	53.9	63.5	41.7	45.1
ASCAL	58.5	64.8	48.6	49.2
ISC	76.3	85.2	67.1	67.9
CrosSCLR	75.2	78.8	67.9	66.7
Shift-GCN	89.7	96.0	85.3	86.6
BlockGCN	93.1	97.0	90.1	91.0
SAN-GCN	92.1	96.2	88.7	90.1
Ta-CNN	90.7	95.1	85.7	87.3
GSTLN	91.9	96.6	88.1	89.3
InfoGCN	93.0	97.1	89.8	91.2
STEP CAT-Fromer	93.2	97.3	90	91.2
LA-GCN	93.5	97.2	90.7	91.8
MMTM	89.4	93.6	86.8	89.1
CC-DFARM	**93.8**	**97.2**	**90.9**	**92.3**

(Cross-View, CV)任务中分别取得 93.8%与 97.2%的准确率,较基线模型 MMTM 提升 4.4%与 3.6%,且超越当前最优模型 LA-GCN 与 GSTLN。在更具挑战性的 NTU 120 中,CC-DFARM 以 90.9%的 C-sub 和 92.3%的 C-set 准确率刷新性能记录,相较 MMTM 提升 4.1%与 3.2%,显著优于其他模型如 InfoGCN 与 STEP CAT-Former。

在验证本文提出的跨模态协同感知的双流融合动作识别模型的有效性上,通过与无监督单骨架动作识别 PCRP[15]、ASCAL[16]、ISC[17]、CrosSCLR[18]模型对比,本文提出的 CC-DFARM 模型突破单模态信息瓶颈,比单骨骼模型提升约 20%准确率。由此说明,单一数据模态的方法存在识别性能上限,多模态融合能够有效的提高综合识别性能,单一方法与单一模态的技术已经无法应对动作识别更深层次的挑战。在两大数据集上,与 GCN 方法如 Shift-GCN[19]、BlockGCN[20]、SAN-GCN[21]、InfoGCN[22]、STEP CAT-Fromer[23]、LA-GCN[24],与 CNN 方法 Ta-CNN[25]、GSTLN[26]等进行比较,本文提出模型均达到了更高的准确率,证明了本文方法的有效性。

由表 2 对比实验表所知,本文算法 CC-DFARM 相对单一模态模型有较大性能提升,与主流方法 GCN、CNN 方法相比,仍能保持较高的优越性,源于其双流架构与多模态特征融合设计,RGB 流与骨架流的全局信息融合使模型在复杂动作(如遮挡场景)中表现更加具有鲁棒性。同时,模型的动态权重分配显著增强关键时空区域的表征能力。

2)与其他多模态融合动作识别模型性能对比

动作识别模型的分类精度在基准数据集上持续突破,但研究表明,基于单模态的识别方法面临显著的性能瓶颈。学界已通过大量实证证实,跨模态特征融合与异构技术协同能显著提升系统综合效能。传统单模态架构难以应对复杂场景下的细粒度动作解析需求。因此,为了验证本文提出的跨模态协同感知的双流融合动作识别模型在动作识别准确度上优势,将本文算法与其他多模态融合动作识别算法做对比,并介绍所选模型选取与本文算法区别,如表 3 所示。

表 3　与其他多模态融合动作识别模型性能对比

Table 3　Comparison of performance with other multimodal fusion action recognition models

Model	多模态处理方式	NTU 60		NTU 120	
		CS/%	CV/%	C-sub/%	C-set/%
LA-GCN	语言大模型+骨架	93.5	97.2	90.7	91.8
3DA	视频+骨架	92.1	95.8	90.5	91.4
Star-transformer	视频+骨架	92.0	96.5	90.3	92.7
MMNet	视频+骨架	91.4	94.8	88.3	89.2
MMTM	视频+骨架	89.4	93.6	86.8	89.1
DSTSA-GCN	视频+骨架	92.7	97.0	89.1	90.9
CC-DFARM	视频+骨架	**93.8**	**97.2**	**90.9**	**92.3**

由表 3 所示,本文提出的跨模态协同感知的双流融合动作识别模型相较于其他多模态融合动作识别模型,在数据集上的表现更加突出。在与 LA-GCN 中,LA-GCN 使用语言大模型和骨架融合,其局部注意力固化的缺陷,导致仅通过语言提示生成固定区域的关节注意力,同时,该模型架构存在时空割裂缺陷,本文算法提出的 TIAM 时空交互注意力模块可以解决类似问题。3DA[27]、Star-transformer[28]、MMNet[29]、MMTM 和 DSTSA-GCN[30]都是与本文一样采用视频与骨架两种模态融合,但这些多模型动作识别模型都缺少时空特征的深度协同与动态互补,不能实现动态增强相关时空区域的注意力权重,本文算法 CC-DFARM 在 HCN 中集成时空交互注意力模块,通过将动态时空协同模块嵌入分层卷积网络 HCN 框架,实现多尺度时空特征的交互增强。该机制通过自适应权重分配策略,对关键运动区

域的时空上下文信息进行选择性强化。

3.4　消融实验

1)DSMF 的数量

为了评估 DSMF 的有效性及其数量对网络性能的影响,本文对模型进行消融实验。首先,确认消融基准模型为时空交互注意力模块与 HCN 的集成后,并且经过最终 MFFM 多模态特征融合模块的模型。本文将 DSMF 模块分别放在 conv5 卷积层后还有 FC7 全连接层后,消融结果如表 4 所示。

表 4　DSMF 数量消融实验

Table 4　DSMF quantitative ablation experiment

DSFM 数量	NTU 60		NTU 120	
	CS/%	CV/%	C-sub/%	C-set/%
DSMF1	91.3	95.4	88.5	89.2
DSMF2	90.8	94.2	87.3	88.6
CC-DFARM	**93.8**	**97.2**	**90.9**	**92.3**

根据表 4 的实验结果分析,DSMF 模块的引入显著优化了模型性能。数据显示,整合 DSMF 结构的模型各评估指标准确率较基准模型 HCN 均实现显著提升。这种性能增益现象表明,DSMF 架构通过其特有的跨模态特征聚合能力,有效增强了模型的鲁棒性,验证了该模块在提升网络抗干扰能力方面的核心价值。值得注意的是,实验过程中 DSMF 模块始终表现出稳定的性能增益特性,进一步佐证了其结构设计的合理性。其中,DSMF1 在网络的中间部分,接收来自顶部的两个主要输入流,模块位于这些卷积和池化层之后,整合不同特征图的数据。DSMF2 模块处于 FC7 全连接层后,用于从下游特征中提取融合信息,整合和融合不同深度的空间特征信息,这种布局帮助网络有效地利用来自不同层的深度信息。

2)各模块消融实验

为了验证 TIAM 模块和 DSMF 模块给实验带来的准确性,以及 MFFM 多模态特征融合模块对模型的性能提升,本文对 TIAM 模块、DSMF 模块、MFFM 多模态特征融合模块进行消融对比实验,本文的消融实验在 NTU RGB+D 60 和 NTU RGB+D 120 数据集上进行验证,如表 5、6 所示。

表 5　在 NTU 60 数据集上模块消融实验数据表

Table 5　Datasheet of module ablation experiments on the NTU 60 dataset

TIAM	DSMF	MFFM	CS/%	CV/%
			89.4	93.6
	√		90.9	93.9
	√	√	91.3	94.8
√	√		92.7	96.1
√	√	√	**93.8**	**97.2**

表 6　在 NTU 120 数据集上模块消融实验数据表

Table 6　Datasheet of module ablation experiments on the NTU 120 dataset

TIAM	DSMF	MFFM	C-sub/%	C-set/%
			86.8	89.1
	√		88.3	90.2
	√	√	89.6	91.2
√	√		89.8	91.5
√	√	√	**90.9**	**92.3**

根据表 5、6 的消融实验结果分析,在基线网络 MMTM 上,将原有 MMTM 双模态融合架构用本文提出的多模态融合架构 DSMF 替换,新的多模态架构解决了 MMTM 中信息流动不高效,DSMF 架构使得模态之间的信息交互更深入、更高效。通过在模型最后的输出前放置本文提出的 MFFM 模块,使得模型在融合过程中动态地选择重要特征的全局信息来实现,由消融实验数据所得,该模块对整体模型有不错的提升。TIAM 模块可动态强化关键时空节点的权重分配,同时抑制冗余或冲突特征,在加入 TIAM 模块后,构建"全局特征增强-时空权重自适应"的双向优化机制。

4　结　　论

针对动作识别中跨模态特征交互不足与时空信息割裂的局限性,本研究从协同感知理论出发,构建了一种双流深度互融的新型识别框架。不同于传统单模态或浅层融合方法,该模型通过跨模态动态协同机制,在特征编码阶段实现了骨骼运动轨迹与 RGB 视觉表征的深度互增强,同时引入时空维度自适应的特征重构策略,有效突破了传统方法对时序依赖性与空间关联性建模的瓶颈。实验验证表明,该框架在多模态动作理解中展现出更强的语义解耦能力,其核心价值在于为异构模态的协同表达提供了可扩展的融合范式,而非局限于特定数据集的性能优化。未来将进一步探索模型在复杂场景下的鲁棒性优化,并研究无监督跨模态对齐方法以降低数据标注成本。本工作从认知协同角度为多模态动作分析开辟了新的研究视角,对行为理解模型的认知可解释性提升具有启发意义。CC-DFARM 在更大规模的 NTU 120 数据集上仍保持领先,表明其良好的泛化性。

参考文献

[1]　ZHENG C, WU W H, CHEN CH, et al. Deep learning-based human pose estimation: A survey[J]. ACM Computing Surveys, 2023, 56(1): 1-37.

[2]　罗浩, 姜伟, 范星, 等. 基于深度学习的行人重识别研究进展[J]. 自动化学报, 2019, 45(11): 2032-2049. LUO H, JIANG W, FAN X, et al. Research progress on pedestrian re-recognition based on deep

learning[J]. Journal of Automation, 2019, 45(11): 2032-2049.

[3] WEI W L, LIN J CH, LIU T L, et al. Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 13211-13220.

[4] DASS S D S, KRISHNASAMY G, PARAMESRAN R, et al. Schatten p-norm based image-to-video adaptation for video action recognition [C]. 2023 International Joint Conference on Neural Networks (IJCNN). IEEE, 2023: 1-8.

[5] YAN S J, XIONG Y J, LIN D H. Spatial temporal graph convolutional networks for skeleton-based action recognition [C]. AAAI Conference on Artificial Intelligence, 2018, 32(1).

[6] LI M S, CHEN S H, CHEN X, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 3595-3603.

[7] SHI L, ZHANG Y F, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 12026-12035.

[8] WANG C L, YAN J J. A comprehensive survey of rgb-based and skeleton-based human action recognition [J]. IEEE Access, 2023, 11: 53880-53898.

[9] TASNIM N, BAEK J H. Dynamic edge convolutional neural network for skeleton-based human action recognition[J]. Sensors, 2023, 23(2): 778.

[10] JOZE H R V, SHABAN A, IUZZOLINO M L, et al. MMTM: Multimodal transfer module for CNN fusion [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 13289-13299.

[11] SHAHROUDY A, LIU J, NG T T, et al. Ntu rgb+ d: A large scale dataset for 3d human activity analysis[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1010-1019.

[12] DU Y, FU Y, WANG L. Skeleton based action recognition with convolutional neural network [C]. 2015 3rd IAPR Asian Conference on Pattern Recognition(ACPR). IEEE, 2015: 579-583.

[13] LI CH, ZHONG Q Y, XIE D, et al. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation[J]. ArXiv preprint arXiv:1804.06055, 2018.

[14] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6299-6308.

[15] XU SH H, RAO H C, HU X P, et al. Prototypical contrast and re verse prediction: Unsupervised skeleton based action recognition [J]. IEEE Transactions on Multimedia, 2021, 25: 624 634.

[16] RAO H C, XU SH H, HU X P, et al. Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition[J]. Information Sciences, 2021, 569: 90-109.

[17] THOKER F M, DOUGHTY H, SNOEK C G M. Skeleton-contrastive 3d action representation learning[C]. 29th ACM International Conference on Multimedia, 2021: 1655-1663.

[18] LI L G, WANG M S, NI B B, et al. 3d human action representation learning via cross-view consistency pursuit [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 4741-4750.

[19] CHENG K, ZHANG Y F, HE X Y, et al. Skeleton-based action recognition with shift graph convolutional network [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 183-192.

[20] ZHOU Y X, YAN X D, CHENG ZH Q, et al. Blockgcn: Redefine topology awareness for skeleton-based action recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 2049-2058.

[21] TIAN H Y, MA X, LI X, et al. Skeleton-based action recognition with select-assemble-normalize graph convolutional networks[J]. IEEE Transactions on Multimedia, 2023, 25: 8527-8538.

[22] HYUNG-GUN CHI, MYOUNG HOON HA, SEUNGGEUN CHI, et al. Infogcn: Representation learning for human skeleton-based action recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 20186-20196.

[23] LONG N H B. Step catformer: Spatial-temporal effective body-part cross attention transformer for skeleton-based action recognition[J]. ArXiv preprint arXiv:2312.03288, 2023.

[24] XIANG W M, LI CH, ZHOU Y X, et al. Generative action description prompts for skeleton-based action recognition[C]. IEEE/CVF International Conference on Computer Vision, 2023: 10276-10285.

[25] XU K L, YE F F, ZHONG Q Y, et al. Topology-aware convolutional neural network for efficient

skeleton-based action recognition [C]. AAAI Conference on Artificial Intelligence, 2022, 36(3): 2866-2874.

[26] DAI M, SUN ZH H, WANG T Y, et al. Global spatio-temporal synergistic topology learning for skeleton-based action recognition [J]. Pattern Recognition, 2023, 140: 109540.

[27] KIM S, AHN D, KO B C. Cross-modal learning with 3D deformable attention for action recognition[C]. IEEE/CVF International Conference on Computer Vision, 2023: 10265-10275.

[28] AHN D, KIM S, HONG H, et al. Star-transformer: A spatio-temporal cross attention transformer for human action recognition [C]. IEEE/CVF Winter Conference on Applications of Computer Vision, 2023: 3330-3339.

[29] BRUCE X B, LIU Y, ZHANG X, et al. Mmnet: A model-based multimodal network for human action recognition in rgb-d videos[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3): 3522-3538.

[30] CUI H, HUANG R J, ZHANG R Y, et al. DSTSA-GCN: Advancing Skeleton-Based Gesture Recognition with Semantic-Aware Spatio-Temporal Topology Modeling [J]. ArXiv preprint arXiv: 2501.12086, 2025.

作者简介

刘罡(通信作者),高级工程师,硕士生导师,主要研究方向为深度学习、移动通信等。

E-mail:oiugang@cwxu.edu.cn

李小雨,硕士研究生,主要研究方向为深度学习、目标检测。

E-mail:1144127574@qq.com

吴烨,硕士研究生,主要研究方向为深度学习、目标检测。

郑泽林,硕士研究生,主要研究方向为深度学习、移动通信。