基于人体姿态的铁路车站旅客异常行为研究

王俊康[1] 王荣才[2]
WANG Junkang WANG Rongcai

摘 要 日益增长的铁路客流量使得铁路车站成为重要的社会公共场所，目前车站监控智能化分析较为落后，针对旅客摔倒斗殴等异常行为自动化检测需求日益强烈。基于此，文章提出基于人体姿态估计的旅客异常行为检测模型，首先以 OpenPose 识别算法为基础进行改进，将 MobileNet-v3 替代传统模型内 VGG19，并引入 ULSAM 注意力机制，有效减少了计算复杂度和模型参数量，精准地从车站监控图像中提取旅客人体姿态骨骼图。随后基于双流自适应图卷积网络 2s-AGCN 算法，优化图卷积块结构，有效捕捉到人体骨架数据的双流特征，提高了人体行为识别的准确率。模型成功将旅客行为分成站立、行走、奔跑、坐立、蹦跳、挥拳、踢踹、跌倒和举手九种类别，根据不同铁路车站场景，重点检测不同异常行为，具有较好的应用前景。

关键词 行为识别；骨架数据；姿态估计；铁路车站

doi：10.3969/j.issn.1672-9528.2025.05.045

0 引言

随着我国高速铁路事业的不断建设和发展，日益增长的客流量使得铁路车站成为重要的社会公共场所，针对火车站人员复杂、环境特殊等特点，视频监控安防系统在铁路行车指挥、生产组织、客货运输服务、作业监控、抢险救援以及治安防范等方面得到广泛应用[1]，正发挥着不可替代的作用。目前全路视频监控正面临着数据体量巨大，但大多数视频图像数据挖掘不充分等问题，为视频图像的智能化分析过程带来较大挑战。目前大多数车站在实际的车站视频安全作业管理过程，还处于一种被动管理的方式，即视频监控主要用于车站管理人员的实时巡查，以及对于突发事件的视频溯源回放，整体面临视频利用率低、效率不高等问题。另一方面，深度学习技术快速发展，其中通过人体姿态估计进而判断人体行为识别成为当前计算机视觉热点方向，主要可以用来检测摔倒、破坏公物和打架等对公共环境安全造成威胁的异常行为，减少对监控的人工分析过程。基于以上背景，本文设计基于人体姿态估计的旅客异常行为智能检测模型，从而减轻火车站综合监控人员的视频巡查负担。

1. 郑州警察学院图像与网络侦查系 河南郑州 450000
2. 机械工业第六设计研究院有限公司 河南郑州 450042
[基金项目] 2023 年中央高校基本科研业务项目“针对铁路车站监控旅客异常行为实时监测方法研究”（2023TJJBKY013）；2024 年中央高校基本科研业务项目“基于人体姿态识别的手枪射击辅助教学应用研究”（2024TJJBKY022）

1 基于人体姿态估计研究现状及关键技术

1.1 国内外研究现状

人体姿态估计即从图像中准确预测画面内人体头部、双臂和腿部等关键部位位置，从而构成人体骨架图。铁路监控场景下一般涉及人员较多，需同时对多人进行人体关键特征点提取进而行为姿态估计，多人行为估计目前主流研究分为两个方向，一是自顶向下的姿态估计：即计算机先通过目标检测等方法进行人体检测，获得图像中每个单人图像，再进行单人姿态提取，包括 G-RMI、AlphaPose、RMPE、Mask R-CNN 和 CPN[2-4] 等典型方法；二是自底向上的姿态估计，这类方法通常先获得图像中所有人的人体骨骼关键点，再对所有骨骼点进行分组匹配，最终获得每个人的人体骨架，这类代表算法包括 DeepCut、Part Segmentation、OpenPose[5-6] 等。需要根据具体场景进行选择合适的算法，例如 AlphaPose 符合对高精度要求的场景，而 OpenPose 则更适合需要实时处理复杂场景的任务。自顶向下姿态估计的方法计算量取决于视频画面中人员的数量，因为铁路监控具有短时大人流量等特点，自底向上的姿态估计方法能较好地提高算法检测速度，满足实时检测的业务要求，能较好地实现人流量较大场景下的实时监测效果。

基于人体骨架进行行为分类主要根据主干网络的不同分为基于 RNN、CNN 和 GCN 三类。Liu 等人 [7] 构建了一个时空 LSTM 模型，旨在识别三维人体行为，该模型将 RNN 扩展至时空维度。相比之下，Lee 等人 [8] 则提出了一个复合的

时空滑动 LSTM 框架，即 TS-LSTM，融合了短期、中期及长期的 TS-LSTM 网络，以全面捕捉并分析跨不同时间尺度的运动动态。Li 等人 [9] 开发了一个端到端的框架，该框架依赖于卷积共现特征学习技术，专门用于骨骼行为识别任务。相较于 RNN 和 CNN，GCN 专门用于处理图结构数据，人体骨骼可以被视为非欧几里得的图数据，因此 GCN 在基于骨架的行为识别中具有先天的优势，GCN 能够直接处理图结构数据，捕捉关节之间的空间关系以及时间序列上的依赖性，但 GCN 的计算复杂度整体相对较高，Yan 等人 [10] 开创性地设计了 ST-GCN（时空图卷积网络）用于骨架行为识别，Shi 等人 [11] 提出了双流自适应 GCN（2s-AGCN），该模型能够借助反向传播机制自动调整图的拓扑结构，无需人工预先定义，随后 Saini 等人 [12] 在 MS-AAGCN 赛算法中，采用数据驱动的图结构学习，并融入空间、时间及通道注意力模块，以此提升模型对关键关节、关键帧以及关键特征的敏感度。

1.2 传统 OpenPose 算法原理及问题

面对铁路车站场景复杂人员众多等问题，OpenPose 自底向上的算法逻辑具有更好的处理效果，能够从静态图片或者视频中检测和估计人体关键点的位置，并将其表示为关键点和骨骼的集合，算法核心在于识别人体各个关节的位置，并通过 PAF 来建立连接。每个关节被表示为一个关键点，每对相邻关节通过一个向量表示连接关系，对于输入的图像，网络会为每一个人体关节生成一个热图，PAF 被用于表示关节之间的连接，任务是帮助网络判断哪些关节应该连接在一起，以及这些关节之间的空间关系。传统 OpenPose 算法能在普通的机器上运行，但其性能和速度受到硬件的限制，特别是在处理高清视频流或进行多人姿态估计时，可能会对计算资源要求较高，导致运行速度较慢，在多人姿态估计中，若多个体之间的距离较近，系统可能会出现误识别或者连接错误的问题。

1.3 ST-GCN 原理及问题

GCN 在人体行为分类中应用广泛，其中 ST-GCN 作为针对骨骼视频数据的行为识别标志性算法，同时捕捉时空关系，有效提升图数据上的预测任务表现。核心思想是将人体骨架表示为图结构，并通过图卷积网络来捕捉骨骼点之间的空间和时间依赖关系，人体骨架被表示为图结构，其中骨骼点被视为图的节点，骨骼点之间的关系被视为图的边，在人体骨架时空图中，节点集为 $V=\{V_i|t=1,2,\cdots,T;\ i=1,2,\cdots,N\}$，对于 V 中的每一个节点 V_i 都有一个对应的特征向量 $\boldsymbol{F}_{vi}$，这个向量中保存了 V_{ti} 的坐标。在同一帧中的关节点会根据人体的自然连接关系连接起来，不同帧里的同一个关节点之间会被连接起来。其网络结构图如图 1 所示，将构造好的时空图作为输入，在 ST-GCN 中经过了多层的时空卷积运算之后，ST-GCN 会提取出高阶的特征，从而在最后一层输出动作的分类结果。

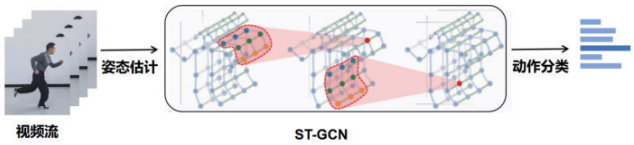

图 1 ST-GCN 网络结构原理图

相比于传统的时间序列模型或图神经网络，ST-GCN 能够在处理复杂时空数据时表现出更强的建模能力，但由于 ST-GCN 同时处理图卷积和时间卷积，计算量较大，在处理大规模图数据时，尤其是节点数量和时间步长较多时，训练和推理过程可能需要大量的计算资源；另一方面 ST-GCN 假设图的结构在整个时间序列中是固定的，对于动态变化的图结构不太适用，在处理动态图时，性能会受到一定影响。

2 实验过程

2.1 异常行为检测总体框架

本文总体框架分为人体骨骼关键点检测、异常行为识别模块和异常行为数据集建立三大部分，其项目总体框架研究流程思路图如图 2 所示。在展开正式检测前，需对视频进行逐帧图像读取，利用人体骨骼关键点检测部分，提取出视频图像中人体骨骼关键点，形成旅客骨架图；其次建立异常行为数据集，建立旅客异常行为识别模型所需的样本库，最后通过异常行为识别模块将监控实时人体骨骼图进行分析，输出最终结果。

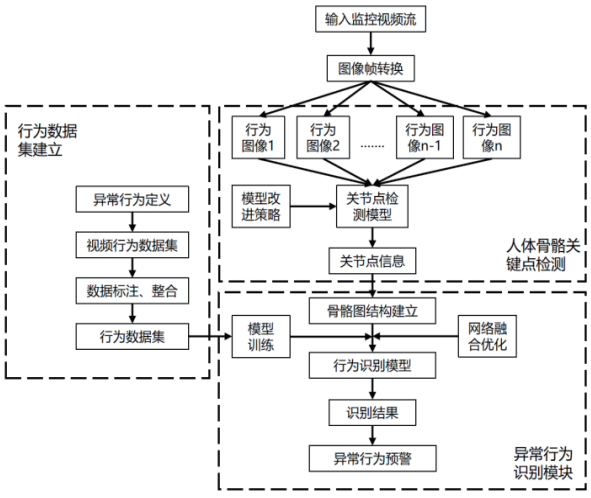

图 2 异常行为检测总体图

人体骨骼关键点检测：主要从车站监控图像中提取旅客人体姿态骨骼图，以 OpenPose 识别算法为基础搭建针对车站监控场景的旅客骨骼关键点检测模型，为进一步提升 OpenPose 的运行速度和检测精度，通过将 OpenPose 特征提取阶段的主干网络，换成基于深度可分离卷积的

MobileNet-V3 网络，引入超轻量级子空间 ULSAM 注意力机制，并对特征细化阶段冗余的卷积结构进行合并，精简计算模型，从而在保证准确率的同时可以实现多人体骨骼关键点检测。异常行为识别模块：该模块核心为行为识别模型的建立，模型处理对象为人体骨骼行为图，本文采用基于图卷积神经网络的双流网络 2s-AGCN，采用经典的双流网络结构，分别处理关节数据和骨骼数据，对图卷积块的结构进行优化，不断调整优化模型参数，将人体行为分为站立、行走、奔跑、坐立、蹦跳、挥拳、踢踹、跌倒和举手九种类别，根据不同铁路车站场景，重点检测不同异常行为。

2.2 基于改进的 OpenPose 人体姿态骨架提取方法

针对铁路车站多人实时检测的要求，系统要求骨架识别网络快速准确，传统 OpenPose 特征提取骨干网络采用 VGG19，VGG19 的网络结构包含了大量的卷积层和全连接层，导致其参数量巨大，本文选择 MobileNet-V3 替代 VGG19，MobileNet-V3 通过结合神经架构搜索（NAS）、高效的深度可分离卷积、轻量化的网络设计以及优化的激活函数等手段，在计算效率、推理速度和模型精度之间取得了较好的平衡。模型首先保留倒残差结构与深度可分离卷积的设计，使得网络能够高效地提取特征，同时保持较低的计算成本。在倒残差块中，使用逐层的瓶颈结构和逐通道卷积操作，以减少计算量和内存占用；采用神经架构搜索技术，通过自动化的搜索过程来优化网络结构，与传统的 ReLU 激活函数不同，MobileNet-V3 引入 H-Swish 激活函数，获得更平滑的非线性变换，从而进一步提高了模型的表达能力减少了计算开销，H-Swish 激活函数用公式表示为：

$$\text{H-Swish}[x] = x \cdot \frac{\text{RELU6}(x+3)}{6} \tag{1}$$

MobileNet-V3 轻量级网络在特征提取环节，采用了 SE 注意力机制，该机制激励阶段采用全连接层，增加模型的参数规模与计算复杂度，导致训练和推理过程中的时间成本上升，本文引入超轻量级子空间 ULSAM 注意力机制替换原有结构，参数减少的同时，准确率得到提高。ULSAM 可以更高效地进行关键点的定位，另外 ULSAM 自适应调整注意力，动态聚焦于关节区域（如肩膀、膝盖、肘部等），而抑制背景、无关区域或遮挡区域的影响。这样，模型不仅能够提高关键点的准确性，也能提升在复杂背景下的鲁棒性。

2.3 人体行为分类设计

近年来，图卷积网络（GCN）在骨架行为识别中取得了显著进展，本文采用基于双流自适应图卷积网络 2s-AGCN（two-stream adaptive graph convolutional network）算法，该算法基于图卷积神经网络（GCN）框架，利用人体骨架数据的时空特性，将人体的骨架结构视作图结构，对骨架节点和骨架骨连线进行建模，从而提取丰富的空间特征。通过时间流和空间流的双流设计，2s-AGCN 能有效捕捉到骨架数据的动态变化和静态结构特征，取得了较好的表现，算法示意图如图 3 所示。

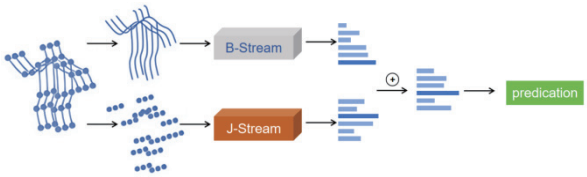

图 3 2s-AGCN 算法原理图

现有的基于 GCN 的方法中，图的拓扑结构是手动设置的，并且在所有层和输入样本上都是固定的，这对于动作识别任务中的分层 GCN 和不同的样本并不是最优的，本文选用一种自适应图卷积网络（AGCN），该网络能够以端到端的方式学习不同 GCN 层和骨架样本的图的拓扑结构，以增加模型构建图的灵活性，并提高了适应各种数据样本的通用性。自适应图卷积层包含 3 种类型的图：$\boldsymbol{A}_k$、$\boldsymbol{B}_k$ 和 $\boldsymbol{C}_k$，$\boldsymbol{A}_k$ 表示人体的物理结构，与原始归一化 $N\times N$ 邻接矩阵相同；$\boldsymbol{B}_k$ 是一个 $N\times N$ 的邻接矩阵，其元素在训练过程中与其他参数一起进行参数化和优化，其值没有进行约束，主要根据训练数据学习图；$\boldsymbol{C}_k$ 是一个数据依赖图，为每个样本学习唯一的图，为了判断两个顶点之间是否存在连接以及连接的强度，采用归一化的嵌入高斯函数来计算两个顶点的相似度。其相关计算公式为：

$$\boldsymbol{f}_{\text{out}} = \sum_k^{K_v} \boldsymbol{W}_k \boldsymbol{f}_{\text{in}} (\boldsymbol{A}_k + \boldsymbol{B}_k + \boldsymbol{C}_k) \tag{2}$$

2s-AGCN 采用了经典的双流网络结构，分别处理关节数据和骨骼数据，捕捉空间和时间信息。这种设计可以充分利用视觉的不同维度信息，提高识别精度。双流网络结构包括关节流（J 流，关节数据计算特征图）和骨骼流（B 流，根据关节数据计算骨骼数据），然后将两个流的 softmax 得分相加，得到融合得分并预测动作标签。2s-AGCN 采用动态图融合策略，根据帧间的关节运动信息动态更新图结构，使得模型在处理非刚体变换时表现出色，通过自适应地调整不同关节的重要性，使模型能够更加聚焦于动作的关键部位，但对于某些复杂动作，模型的识别精度仍有待提高，其次模型的训练时间较长，限制了其在实际场景中的应用，本文在原有基础上主要进行以下两方面改进，引入注意力机制：在自适应图卷积层中引入注意力机制，通过自适应地调整不同关节的重要性，使模型能够更加聚焦于动作的关键部位，有助于提高模型的识别精度；优化图卷积块结构：对图卷积块的结构进行优化，减少不必要的计算量，提高模型的训练效率，

可以减少图卷积层的数量，并在卷积层之间添加 Dropout 层以防止过拟合。

3 实验结果分析

3.1 骨架提取网络实验结果

实验首先针对骨架提取网络改进效果进行对比，详细对比 OpenPose 网络与改进后 OpenPose 网络，实验采用 MSCOCO 数据集，其涵盖 80 类物体，包含多张带有多人标注的图像，包含头部、四肢、躯干等部位，主要针对 2D 关键点，提供了丰富的姿态数据，适用于多人的关键点检测任务。主要评价指标有 PCKh@0.5、mAP、AP50 和速度，PCKh@0.5：衡量在给定阈值内不同关节点定位精度的指标；mAP：计算所有关键点的平均精度，综合评估模型表现；AP50：预测关键点正确率大于 50% 的平均值，结果如表 1 所示。

表 1 骨架提取网络改进效果对比表

指标	OpenPose	OpenPose (MobileNetV3+ULSAM)
PCKh@0.5	0.76	0.79
mAP	0.65	0.68
AP50	0.85	0.88
速度	19.2 帧 /s	31.5 帧 /s

结合实验数据可以看出，改进模型性能均有提升，PCKh@0.5 指标提升 3.9%，mAP 指标提升 4.6%，AP50 提高了约 3.5%，这些提升反映了在多人姿态估计中，模型对关键点的预测精度有所增加，由于 MobileNet-V3 减少了模型的参数量和计算量，改进模型的推理速度相较基准模型显著提升，在相同硬件平台下，改进模型的处理时间减少明显，使得模型在多人实时应用场景中更加高效。另一方面：引入 ULSAM 后，模型在复杂场景下表现出更强的鲁棒性，MSCOCO 数据集中，一些场景中的人物姿态交叉或部分遮挡时，ULSAM 能够有效抑制背景干扰，提升关键点的检测精度。其次模型在细节部位（如头部、肩膀、膝盖等）检测效果均有提升，这说明 ULSAM 在增强局部区域注意力方面具有较好的效果，尤其在这些关键点的定位上得到了更高的精度。

3.2 人体行为分类实验结果

行为分类实验采用 NTU RGB+D 120 数据集，该数据集是 NTU RGB+D 数据集的扩展，包含 120 个动作类别和 114 480 个样本。为验证模型检测效果，本文采用 ST-GCN、2s-AGCN 与改进 2s-AGCN 三种模型进行对比数据对比，根据数据集 X-Sub 和 X-Set 两种分类，分别对比其准确率、模型参数和处理帧数效果，其数据结果如表 2 所示。

表 2 不同模型行为分类效果对比表

指标	ST-GCN	2s-AGCN	改进 2s-AGCN
准确率 /%（X-Sub）	72.8	81.2	82.5
准确率 /%（X-Set）	75.6	82.6	85.4
参数 /10^6	3.1	6.9	6.2
处理帧数	15.3 帧 /s	9.6 帧 /s	11.9 帧 /s

观察实验结果可知，在 NTU RGB+D 120 数据集上，与传统 ST-STG 网络相比，2s-AGCN 在准确率上均有显著提升，这主要得益于 2s-AGCN 在邻接矩阵和融合方式上的改进，随之模型参数上也相应提升，模型处理速度也相应下降。在 2s-AGCN 基础上，本文引入注意力机制，在一定程度上准确度有所提升，另一方面优化图卷积块结构，降低模型参数，提升模型处理速度。

图 4 为模型识别的旅客行为分类结果，模型将人体行为分成站立（stand）、行走（walk）、奔跑（run）、坐立（sit）、蹦跳（jump）、挥拳（punch）、踢踹（kick）、跌倒（fall）、举手（wave）九类行为进行预测，可以通过实验结果直观看出行为效果，模型在视频输入模式下，首先针对视频连续帧进行处理，模型具备实时追踪图像中人体并全面提取所有检测对象骨骼关键点的能力，在多人场景下模型尽可能识别个体骨骼关键点的基础上，进一步优化匹配过程，确保相邻的人体骨骼关键点被正确联结，减少了误匹配的发生，从而在精确度上相较于过往研究取得了进步。

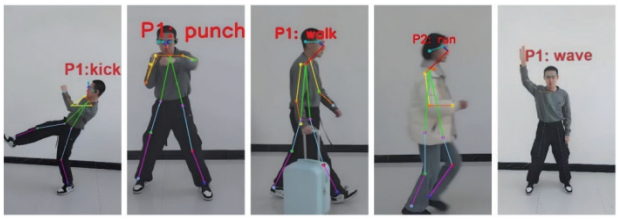

图 4 部分行为分类结果图

图 5 为本文检测效果图，将待检测视频输入模型中，首先基于改进 OpenPose 算法提取画面内所有人体骨架图，后将骨架图序列批量进行预处理后导入 2s-AGCN 模型，分别提取骨架和关键点数据进行处理，最终预测画面内所有人的行为。

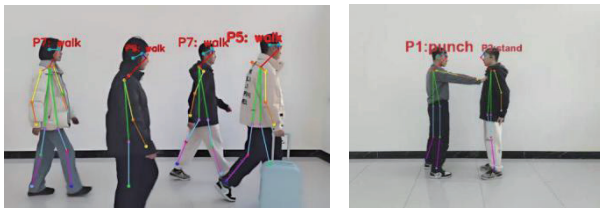

图 5 多人行为检测结果图

因铁路车站空间人员复杂，所以模型尽可能细化人体行

为分类，以适应不同场景需求。例如在车站进站或乘车通道处，侧重检测大规模人群奔跑行为，以预防出现踩踏等危险情况；在进站口或检票口重点检测蹦跳或奔跑等行为可能出现逃票等行为；在候车厅等人群密集区域，当出现踢踹和挥拳等行为可能出现破坏公共物品或斗殴等行为；在扶梯或步梯等处则重点预警跌倒等行为。

4 结论

本文主要从车站监控图像中提取旅客人体姿态骨骼图，通过异常行为识别模块将监控实时人体骨骼图进行分析，输出人体行为分类结果。其中人体骨骼关键点检测，以 OpenPose 识别算法为基础进行改进，通过将 MobileNet-V3 替代传统模型内 VGG19，并引入 ULSAM 注意力机制，实验表明改进后 OpenPose 在 MSCOCO 数据集上的性能得到了显著提升，尤其在多人姿态估计任务中的关键点定位表现优异，MobileNet-V3 的引入有效减小了计算复杂度和模型参数量，使得整体网络运行速度得到提升，ULSAM 注意力机制则显著增强了模型在复杂背景和多人物情况下的鲁棒性，提升了检测精度。成功提取人体骨架图后，采用基于双流自适应图卷积网络 2s-AGCN 算法，优化图卷积块结构，有效捕捉到了人体骨架数据的双流特征，提高了人体行为识别的准确率。模型将旅客行为分成站立、行走、奔跑、坐立、蹦跳、挥拳、踢踹、跌倒和举手九种类别，根据不同铁路车站场景，重点检测不同异常行为，具有较好的应用前景。后续针对不同场景的特点和需求，可对模型进行定制开发和优化，以满足不同人群密度较高场景需求，提升公共安全管理效率。

参考文献：

[1] 佟立本 . 铁道概论 [M]. 北京 : 中国铁道出版社 ,2021.

[2] CHEN Y L, WANG Z C, PENG Y X, et al. Cascaded pyramid network for multi-person pose estimation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7103-7112.

[3] FANG H S , XIE S, TAI Y W, et al. Rmpe: regional multi-person pose estimation[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 2334-2343.

[4] FANG H S, LI J F, TANG H Y, et al. Alphapose: whole-body regional multi-person pose estimation and tracking in real-time[J]. IEEE transactions on pattern analysis and machine intelligence, 2023, 45(6):7157-7173.

[5] TOSHEV A, SZEGEDY C. Deeppose: human pose estimation via deep neural networks[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE, 2014:1653-1660.

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

[7] LIU J, SHAHROUDY A, XU D, et al. Spatio-temporal LSTM with trust gates for 3D human action recognition[C]// Computer Vision–ECCV 2016.Berlin:Springer,2016:816-833.

[8] LEE I, KIM D, KANG S, et al.Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks[C]//2017 IEEE International Conference on Computer Vision (ICCV).Piscataway:IEEE, 2017:1012-1020.

[9] LI C, ZHONG Q, XIE D, et al.Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation[C]//IJCAI'18: Proceedings of the 27th International Joint Conference on Artificial Intelligence. NewYork: ACM, 2018:786-792.

[10] YAN S J, XIONG Y J, LIN D H. Spatial temporal graph convolutional networks for skeleton-bBased action recogni-tion [C]//AAAI'18/IAAI'18/EAAI'18: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. Alto:AAAI Press, 2018: 7444-7452.

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

[12] SAINI R , JHA N K , DAS B ,et al.ULSAM:ultra-lightweight subspace attention module forcompact convolutional neural networks[C]//2020 IEEE Winter Conference on Applications of Computer Vision (WACV).Piscataway: IEEE, 2020: 1616-1625.

【作者简介】

王俊康（1996—），男，湖北孝感人，硕士研究生，助教，研究方向：图像处理。

王荣才（1997—），男，河南郑州人，硕士研究生，助理工程师，研究方向：人工智能。

（收稿日期：2025-01-17 修回日期：2025-05-19）