

DOI: 10.11830/ISSN.1000-5013.202409023



基于改进图卷积网络 and 人体骨架的 扶梯场景危险行为识别

何建海¹, 郑力新¹, 臧佳明¹, 庄琼云², 潘书万¹

(1. 华侨大学 工学院, 福建 泉州 362021;

2. 黎明职业大学 信息与电子工程学院, 福建 泉州 362000)

摘要: 为了使图神经网络从前后相邻帧中获取缺失的人体骨架信息, 解决自动扶梯狭长环境的遮挡问题和相似人体骨架动作准确识别问题, 提出一种注意力引导的多尺度层次边缘聚合时序图卷积网络 (AMHGCN)。首先, 对时序卷积网络加入不同扩张率的多尺度特征, 延展出的 7 个分支可增强网络对时间域的特征提取能力; 其次, 在多尺度特征时序卷积网络后面加入层次边缘卷积, 使局部特征向全局特征扩张; 最后, 在每个时空图卷积块中, 加入空间通道注意力机制, 强化网络对空间、通道信息的处理, 使 AMHGCN 在分类过程中更加关注不同行为的细节特征, 提高分类的准确率。在 NTU RGB+D 数据集和扶梯危险行为数据集上, 对 AMHGCN 进行评估。结果表明: 相较于基线方法 STGCN++, AMHGCN 在 NTU RGB+D 数据集和扶梯危险行为数据集上的识别准确率均有较大的提高。

关键词: 扶梯; 图神经网络; 危险行为识别; 多尺度特征; 层次边缘卷积

中图分类号: TP 391.4; TU 229

文献标志码: A

文章编号: 1000-5013(2025)03-0308-11

Dangerous Behavior Recognition in Escalator Scene Based on Improved Graph Convolutional Network and Human Skeleton

HE Jianhai¹, ZHENG Lixin¹, ZANG Jiaming¹,
ZHUANG Qiongyun², PAN Shuwan¹

(1. College of Engineering, Huaqiao University, Quanzhou 362021, China;

2. College of Information and Electronic Engineering, Liming Vocational University, Quanzhou 362000, China)

Abstract: To address the occlusion problem in the narrow environment of escalators and the accurate recognition of similar human skeleton actions, a novel method called attention-guided multi-scale hierarchical edge aggregation sequential graph convolutional network (AMHGCN) is proposed to enable the graph neural network to capture missing human skeleton information from adjacent frames. Firstly, multi-scale features with different dilation rates are added to the temporal convolutional network, the extended seven branches can enhance the network's ability to extract features in the time domain. Secondly, hierarchical edge convolution is added after the multi-scale feature temporal convolutional network to expand local features to global features. Finally, a spatial channel attention mechanism is incorporated into each spatiotemporal graph convolutional block to strengthen the network's processing of spatial and channel information, making AMHGCN pays more atten-

收稿日期: 2024-09-29

通信作者: 郑力新(1967-), 男, 教授, 博士, 主要从事图像分析、机器视觉和深度学习方法的研究。E-mail: zlx@hqu.edu.cn.

基金项目: 福建省科技计划重点项目(2020Y0039); 黎明职业大学 2022 年度校级一般课题(自然科学类)(LZ 202211)

tion to the detailed features of different behaviors in the classification process and improves the classification accuracy. The evaluation of AMHGCN is conducted on the NTU RGB+D dataset and the escalator dangerous behavior dataset. The results show that compared to the baseline method STGCN++, AMHGCN achieves a significant improment in recognition accuracy on both the NTU RGB+D dataset and the escalator dangerous behavior dataset.

Keywords: escalator; graph neural network; dangerous behavior identification; multi-scale feature; hierarchical edge convolution

自动扶梯已成为商场、医院、车站等公共场所常见的载客设备^[1]，这些设备在提高人流运输效率的同时，也带来了安全隐患。在实际应用中，由于急速奔跑、逆向行走、儿童玩耍及应急救援不及时等原因，自动扶梯容易引发乘客坠落、碰撞、挤压等事故^[2-4]。这些事故不仅会导致严重的人身伤害，还可能引起大范围的恐慌和混乱，进一步增加危险性^[5]。因此，建立及时、可靠的危险行为识别系统，对于保障自动扶梯的安全使用至关重要。

行为识别通过分析人的肢体姿态与运动轨迹，对人的动作进行分类，先前的一些方法大多基于 RGB 视频进行研究。李伟达等^[6]基于目标识别方法对扶梯危险行为进行识别，用 YOLOV5 结合轻量级算法设计了实时的报警系统。然而，RGB 视频数据易受光照变换的影响，导致颜色信息的丢失，同时，由于环境遮挡视频帧会缺失部分人体关节，无法显示完整的人体特征，大大降低识别的准确率。文献^[7-8]结合特征匹配的多目标跟踪算法，提高了遮挡频繁的自动扶梯场景中目标关联的准确性，但对于缺失的关键节点数据和视频帧数据无法进行关联。在基于人体骨架的行为识别领域中，林志鸿等^[9]基于 YOLOPOSE 简化模型参数量，设计多任务解耦姿态网络，提高了自动扶梯场景中人体骨架形式的关键节点的识别精度和模型的推理速度，但并未解决对危险行为和正常行为的分类。Yan 等^[10]首次提出时空图卷积网络(STGCN)，将图卷积网络(GCN)应用到行为识别领域，提取人体骨架数据中蕴藏的时空运动信息。Zhang 等^[11]整合人类骨架的空间相邻边缘和时间相邻边缘表示人类骨骼图中的一条边。Shi 等^[12]提出一种新的双流自适应图卷积网络，并对一阶关节点信息和二阶骨骼信息进行建模。Wen 等^[13]基于模型的图卷积来编码分层空间结构，同时采用可变的时间密集块来利用人体骨骼序列中不同范围的局部时间信息。Li 等^[14]提出一种新的基于骨骼的动作识别时空图路由器，能自适应地学习物理分离骨骼关节的内在高阶连接关系。上述基于图卷积的方法提高了关节点和骨骼的衔接，将人体骨架数据采用不同的描述方法形成新的数据，这些方法虽然提高了精度，但都没有解决遮挡导致的数据丢失问题，以及相似动作导致人体骨架出现相同的空间结构误判问题。

扶梯行人识别的难题，一是自动扶梯狭长的空间导致行人前后左右的遮挡，当骨架序列的人体关节点被遮挡或某些视频帧丢失时，一些图卷积网络方法^[15-18]的性能可能会显著恶化；二是相似动作容易出现误判，如逆行和正常行走这两个动作在某些帧具有相同的空间结构特征，识别时容易出现误判。基于此，本文提出一种注意力引导的多尺度层次边缘聚合时序图卷积网络(AMHGCN)。

1 注意力引导的多尺度层次边缘聚合时序图卷积网络

1.1 模型结构设计

为了使图神经网络从前后相邻帧中获取缺失的信息，并且更加关注每一帧的动作细节，设计了 3 个模块：多尺度特征时序卷积网络(MFTCN)、层次边缘卷积(HEConv)、空间通道注意力机制(SCAM)。注意力引导的多尺度层次边缘聚合时序图卷积网络框架，如图 1 所示。

首先，将每一帧中人体骨骼的所有关节和中心关节之间的坐标差作为相对坐标；然后，计算两个连续相邻帧之间的关节坐标差，将其作为时间差，以表示动作在时间域中的变化；最后，将相对坐标和时间差进行拼接，得到一个新的输入数据(坐标序列)并输入网络。

对输入的每一帧人体骨架关键节点坐标序列，先通过时间空间子块进行多尺度特征融合与层次边缘聚合，帮助模型在时间和空间维度上收集信息量最大特征，再由全连接层和 softmax 算子给出最终的行为预测。

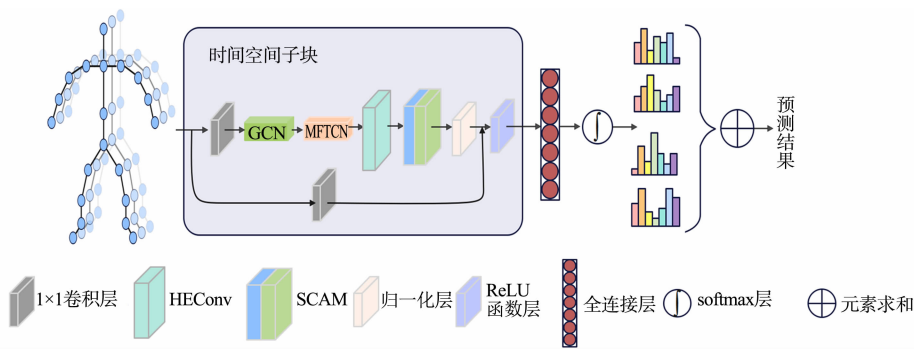


图 1 注意力引导的多尺度层次边缘聚合时序图卷积网络框架
Fig. 1 Framework of attention-guided multi-scale hierarchical edge aggregation sequential graph convolutional network

1.2 多尺度特征时序卷积网络

为解决不完整骨架数据引起的网络性能下降问题,并使整体网络更加关注全局信息,参照文献[19-20],提出一种多尺度特征时序卷积网络,它可以强化网络对长时空运动信息的捕捉,更好地感知行为的前后逻辑,减少某些片段帧的关键节点数据的缺失。多尺度特征时序卷积网络由 7 个分支组成:包括 1 个 1×1 卷积(Conv)分支、1 个最大池化(MaxPool)分支,以及 5 个不同扩张率(d)的一维时间卷积分支。将输入特征从通道维度平均分成 7 个组,分别输入多尺度特征时序卷积的分支中。这种设计有效地结合了多种特征提取方法,使网络能够更加全面地捕捉和表征输入数据中的时空特征。多尺度特征时序卷积网络的整体结构,如图 2 所示。

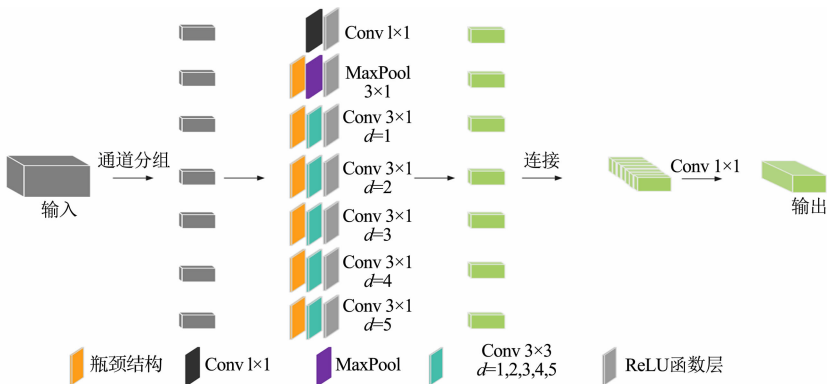


图 2 多尺度特征时序卷积网络的整体结构
Fig. 2 Overall structure of multi-scale feature sequential convolutional network

为了减少模型计算成本,在 5 个不同扩张率的时间卷积分支和最大池化分支前加入 1×1 卷积形式的瓶颈结构(Bottleneck)。同时,为了避免梯度饱和现象发生,在每个多尺度特征时序卷积分支的后面加入 ReLU 函数。7 个分支的结果最终通过直接连接结合,7 个分支输出连接后,通过 1×1 卷积将通道维数还原。

5 个一维时间卷积分支的内核大小(K)均为 3,扩张率为 1~5,经扩张后的卷积核有效尺寸(K')及输出大小(o)分别为

$$K'=K+(K-1)(d-1), \tag{1}$$

$$o=\left\lceil \frac{b+2p-K-(K-1)(d-1)}{s} \right\rceil +1. \tag{2}$$

式(1)、(2)中: s 为卷积时的采样间隔; p 为零填充的大小,用于减缓图像边缘信息的丢失; b 为初始输入的大小。

多尺度特征时序卷积网络引入因果卷积操作,能够有效地捕捉时序数据的依赖关系来处理时序性问题。序列问题可转化为:根据输入系列 x_1, x_2, \dots, x_e 和卷积核 w_1, w_2, \dots, w_h 去预测输出结果 y_1, y_2, \dots, y_e 。因果卷积的计算公式为

$$y_e = \sum_{g=1}^h w_g \cdot x_{e-g+1} \quad (3)$$

原本大小为 $m \times n$ 的卷积核要对两个维度的特征进行处理, 但当 $n=1$ 时, 卷积核只会处理 1 个维度的特征。因此, 通过这种方式只对 1 个时间维度进行特征提取。多尺度特征可以通过设置不同的膨胀系数来实现。当膨胀卷积具有不同的膨胀系数时, 卷积核的感受野会随之变化, 从而提取不同的特征。膨胀卷积引入空洞扩展感受野, 在不增加参数和计算量的情况下, 可有效捕捉多尺度和长距离依赖信息, 并保持特征图的空间分辨率。

1.3 层次边缘卷积

为了更有效地捕获图结构数据中的局部和全局信息, 构造层次边缘卷积在边缘卷积(EdgeConv)的基础上引入分层结构。在计算每一层特征时, 基于前一层的特征, 通过叠卷积层逐步提取更高层次的边缘信息。每一层的特征在卷积过程中通过相邻节点的连接信息得到更新, 使局部特征向全局特征扩展。

层次边缘卷积能够聚合邻居节点的特征, 并逐层传递和更新, 计算节点对之间的特征差异来构建边缘特征, 从而捕获节点的局部结构信息。多尺度卷积的多通道特征增加了特征冗余, 而层次边缘卷积的层次化聚合在一定程度上能够平滑和稳定节点特征。

通过堆叠 4 层边缘卷积, 并在每一层的层次边缘卷积输出后, 引入空间注意力机制(SAM), 计算每个节点的空间重要性, 自动选择对任务更为关键的节点位置。这使模型能够更关注图结构显著或关键的节点信息, 而忽略可能无关或次要的节点特征。

通过 SAM 对图结构中每个节点分配一个权重值, 将卷积后的特征与空间注意力权重逐元素相乘, 使关键节点的特征在权重加权后得到放大, 次要节点的特征被抑制。将 SAM 嵌入层次边缘卷积中, 该融合使模型既能捕捉骨架结构的层次信息, 又能有效筛选出重要特征, 有助于提升人体行为动作识别的准确性。层次边缘卷积的整体结构, 如图 3 所示。图 3 中: $H^{(1)} \sim H^{(4)}$ 为第 1~4 层的特征。

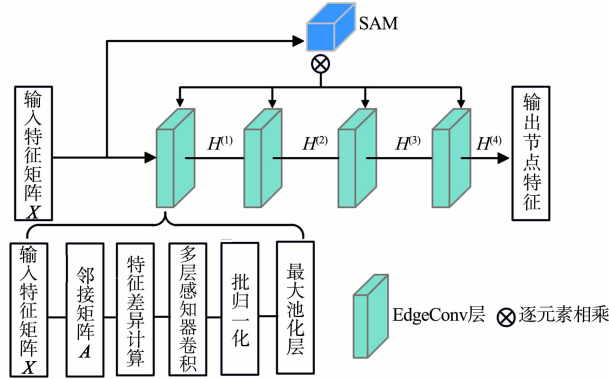


图 3 层次边缘卷积的整体结构

Fig. 3 Overall structure of hierarchical edge convolution

令输入特征矩阵 $X \in \mathbf{R}^{N \times C \times T \times V}$ (N, C, T, V 分别表示批量大小、通道数、帧数和关节数)。赋值给初始特征 $H^{(0)}$, 则第 l 层的特征表示为 $H^{(l)}$, 在每一层的边缘卷积中, 每个节点的特征更新是与其 k 个近邻节点的特征差值聚合。边缘卷积公式为

$$\text{EdgeConv}(H^{(l-1)}, \mathbf{A}) = \sum_{j \in k(i)} \varphi(H_i^{(l-1)}, H_j^{(l-1)}) \quad (4)$$

式(4)中: $k(i)$ 为节点 i 的邻居集合; $\varphi(H_i^{(l-1)}, H_j^{(l-1)})$ 为节点 i 和邻居节点 j 的特征差。

网络总共采用 4 层层次边缘卷积公式以递归形式表示, 从 $H^{(0)}$ 开始, 通过叠加 4 层得到 $H^{(4)}$, 即

$$H^{(4)} = \alpha^{(4)} \otimes [\sigma(\mathbf{W}^{(4)} \cdot (H^{(3)} + \text{EdgeConv}(H^{(3)}, \mathbf{A})) + b^{(4)})], \quad (5)$$

$$\alpha^{(4)} = \sigma(\text{SAM}(H^{(4)})) \quad (6)$$

式(5)、(6)中: σ 为激活函数; $\mathbf{W}^{(4)}$ 为权重矩阵; $b^{(4)}$ 为第 4 层的偏置; $\alpha^{(4)}$ 为空间注意力权重, 对输入特征矩阵 $X \in \mathbf{R}^{N \times C \times T \times V}$ 通过卷积层将每个节点的特征映射到 0~1, 可得到空间注意力权重。

1.4 注意力机制增强模块

为了让模型在分类过程中更关注不同行为的细节特征, 提高分类的准确性, 引入注意力机制, 提出

数据驱动的可迁移的空间通道注意力增强模块,此模块可被插入图卷积网络的任一卷积层之间。

1.4.1 空间注意力增强模块 空间注意力机制使模型更关注重要关节区域,空间注意力模块根据输入特征矩阵 $\mathbf{X} \in \mathbf{R}^{N \times C \times T \times V}$ 生成一个空间掩码矩阵 $\mathbf{M}_s \in \mathbf{R}^{N \times C \times T \times V}$,输入特征矩阵与空间掩码矩阵进行逐元素相乘,将信息聚合。

输入特征矩阵 $\mathbf{X} \in \mathbf{R}^{N \times C \times T \times V}$ 先分别进行通道维度的平均池化和通道最大池化,并进行降维,得到一个形状为 $\mathbf{X} \in \mathbf{R}^{N \times T \times V}$ 的三维张量。平均池化计算输入特征矩阵通道维度上的平均值($\mathbf{X}_{\text{ave},s}$),有

$$\mathbf{X}_{\text{ave},s} = \text{AvgPools}(\mathbf{X}). \quad (7)$$

最大池化计算输入特征矩阵通道维度上的最大值($\mathbf{X}_{\text{max},s}$),有

$$\mathbf{X}_{\text{max},s} = \text{MaxPools}(\mathbf{X}). \quad (8)$$

将平均值特征矩阵和最大值特征矩阵在通道维度上进行拼接,形成一个形状为 $\mathbf{X} \in \mathbf{R}^{N \times 2 \times T \times V}$ 的张量(\mathbf{X}_c),即

$$\mathbf{X}_c = \text{Connect}(\mathbf{X}_{\text{ave},s}, \mathbf{X}_{\text{max},s}). \quad (9)$$

首先,将拼接后的特征矩阵 $\mathbf{X}_c \in \mathbf{R}^{N \times 2 \times T \times V}$ 输入 7×7 的二维卷积层中,得到一个形状为 $\mathbf{X}_c \in \mathbf{R}^{N \times 1 \times T \times V}$ 的张量。然后,通过 Sigmoid 激活函数进行归一化。归一化后的 $\mathbf{X}_c \in \mathbf{R}^{N \times 1 \times T \times V}$ 经过通道维度重塑(Reshape)生成空间掩码矩阵 $\mathbf{M}_s \in \mathbf{R}^{N \times C \times T \times V}$ 乘以 $\mathbf{X} \in \mathbf{R}^{N \times C \times T \times V}$,归一化后的注意力图与输入特征矩阵逐元素相乘,将信息聚合。最后,经过一个卷积核大小为 1×1 的二维卷积层 C_s 进行通道数削减,再经过一个归一化(BN)层处理后,可得该模块的最终输出 $\mathbf{Y}_s \in \mathbf{R}^{N \times C/2 \times T \times V}$,即

$$\mathbf{Y}_s = C_s(\mathbf{X} \times \mathbf{M}_s). \quad (10)$$

1.4.2 通道注意力增强模块 通道注意力机制(CAM)主要用来处理通道信息,对骨架序列中每个通道进行不同程度的关注,根据输入特征矩阵 $\mathbf{X} \in \mathbf{R}^{N \times C \times T \times V}$ 生成一个通道掩码矩阵 $\mathbf{M}_c \in \mathbf{R}^{N \times C \times 1 \times 1}$,两者进行逐元素相乘,将信息聚合。

首先,输入特征矩阵 $\mathbf{X} \in \mathbf{R}^{N \times C \times T \times V}$ 先分别进行空间维度的平均池化和最大池化,对不同通道间的输入张量进行平均,并进行降维,得到一个相对于通道轴的全局时空张量。平均池化对空间维度的所有值取平均值($\mathbf{X}_{\text{ave},c}$),有

$$\mathbf{X}_{\text{ave},c} = \text{AvgPools}(\mathbf{X}). \quad (11)$$

最大池化计算输入特征矩阵空间维度上的最大值 $\mathbf{X}_{\text{max},c}$,有

$$\mathbf{X}_{\text{max},c} = \text{MaxPools}(\mathbf{X}). \quad (12)$$

然后,将池化后的平均值特征矩阵和最大值特征矩阵输入全连接层 FC1,FC2 中,结果分别用矩阵 $\mathbf{W}_1, \mathbf{W}_2$ 表示,FC1,FC2 之间用 ReLU 函数连接,再输入 Sigmoid 激活函数中,得到形状为 $\mathbf{X} \in \mathbf{R}^{N \times C}$ 的二维张量,即

$$\mathbf{X}_a = \sigma(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \cdot \mathbf{X}_{\text{ave},c})), \quad (13)$$

$$\mathbf{X}_m = \sigma(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{X}_1 \cdot \mathbf{X}_{\text{max},c})). \quad (14)$$

全连接层 FC1 的输入、输出通道分别为 C 和 C/r (r 为通道缩减比例),而全连接层 FC2 与之相反。将 $\mathbf{X}_a, \mathbf{X}_m$ 进行逐元素相加,得到拼接后的特征矩阵 \mathbf{X}_c ,对 \mathbf{X}_c 经过空间维度重塑,得到通道掩码矩阵 $\mathbf{M}_c \in \mathbf{R}^{N \times C \times 1 \times 1}, \mathbf{M}_c \in \mathbf{R}^{N \times C \times 1 \times 1}$ 与输入特征矩阵 $\mathbf{X} \in \mathbf{R}^{N \times C \times T \times V}$ 逐元素相乘,将信息聚合。

最后,再经过一个卷积核大小为 1×1 的二维卷积层 C_s 进行通道数削减,最后,经一个 BN 层处理后,得到该模块的最终输出 $\mathbf{Y}_s \in \mathbf{R}^{N \times C/2 \times T \times V}$,即

$$\mathbf{Y}_s = C_s(\mathbf{X} \cdot \mathbf{M}_s). \quad (15)$$

将空间注意力增强模块与通道注意力增强模块的输出进行通道维度上的合并,两个模块串连在多尺度特征时序卷积网络的后面。

2 实验结果与分析

2.1 实验环境

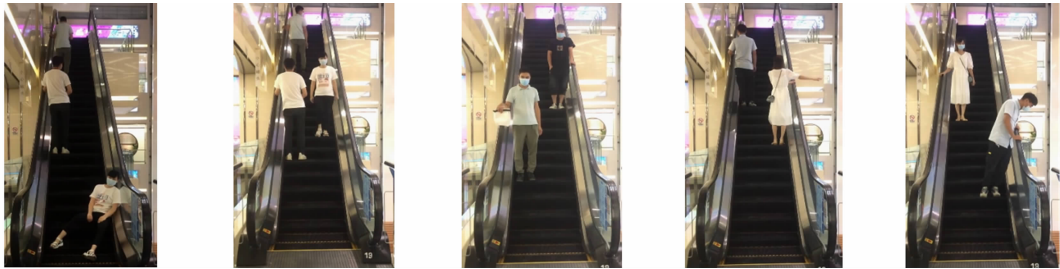
训练时将初始学习率设置为 0.1,批处理大小设置为 128,并使用余弦退火学习调度器(LR)对每个

模型训练 80 个周期。将优化器的动量设置为 0.9, 为了防止过拟合, 权重衰减设置为 5×10^{-4} , 并使用 Nesterov 动量法。

实验在 3 台 NVIDIA GeForce GTX TITAN Xp 型 GPU 上进行, 使用 Python 3.8 和 Pytorch 深度学习框架。实验全部采用以上设备及参数, 实验条件相同。

2.2 实验数据集

1) 扶梯危险行为数据集。采集 2 个商场的扶梯监控视频数据及手机拍摄的视频数据, 用剪辑软件把拍摄的视频剪辑出 2 500 个视频(每秒 30 帧), 视频的长度为 2~4 s。将 2 500 个视频数据集归类为 5 种不安全行为: 摔倒、逆行、错误携带行李、手臂探出扶梯、身体探出扶梯。剪辑好的视频需进行标签分类, 给每种危险行为贴上对应的标签。最后, 将数据集划分为训练集和测试集, 250 个危险行为视频为测试集, 剩余 2 250 个危险行为视频为训练集。选取每个样本的最大帧数为 120 帧, 对于小于 120 帧的样本, 再次计算样本, 直至达到 120 帧, 共计 30 万帧。扶梯危险行为数据集部分场景, 如图 4 所示。



(a) 摔倒 (b) 逆行 (c) 错误携带行李 (d) 手臂探出扶梯 (e) 身体探出扶梯

图 4 扶梯危险行为数据集部分场景

Fig. 4 Partial scene of escalator dangerous behavior dataset

2) NTU RGB+D 数据集。NTU RGB+D 数据集^[21]由 40 位志愿者在有限的环境里拍摄, 共收集 56 880 个样本, 涵盖 60 种动作。前 50 种动作由单人完成, 后 10 种动作由双人完成, 共计 400 万帧。每个骨架序列有 25 个关节, 分别用三维坐标 (x, y, z) 表示。此数据集在划分训练集和测试集时, 采用交叉个体评估与交叉视角评估两个基准。交叉个体评估按照人物序号进行划分, 其中, 20 个人物的数据用于训练, 其余人物的数据用于测试。交叉视角评估按相机序号进行划分, 其中, 2 台相机的数据用于训练, 1 台相机的数据用于测试。

2.3 评价指标

准确率(η_a)是评估分类模型性能的重要指标之一, 是指模型在测试集上正确预测的样本数量占全部样本数量的比例。

准确率的计算公式为

$$\eta_a = \frac{t_p}{t_p + f_p} \tag{16}$$

式(16)中: t_p 为正确预测的样本数量; f_p 为错误预测的样本数量。

召回率(η_r)是衡量模型在所有实际为正类的样本中有多少被正确预测, 其计算公式为

$$\eta_r = \frac{t_p}{t_p + f_n} \tag{17}$$

式(17)中: f_n 为将正类样本错误预测为负类的样本数。

2.4 公共数据集消融实验

以 STGCN++^[22]为基线网络, 将 AMHGCN 的 3 个模块(STGCN++_MFTCN, STGCN++_HEConv, STGCN++_SAM_CAM)进行消融对比实验。实验数据均以交叉个体评估为基准进行比较。采用多数据流的方式进行评估, 关节点流(J 流)的输入是原始骨架坐标, 骨骼流(B 流)的输入是骨骼中相邻关节坐标的差分, 关节点运动流(J-M 流)的输入是关节流每一帧的时间差分, 骨骼运动流(B-M 流)的输入是骨骼流每一帧的时间差分。

NTU RGB+D 数据集上的消融实验结果, 如表 1 所示。表 1 中: δ_j 、 δ_B 、 δ_{J-M} 、 δ_{B-M} 分别为关节点流、骨骼流、关节点运动流和骨骼运动流的识别准确率。

表 1 NTU RGB+D 数据集上的消融实验结果

Tab. 1 Ablation experiments results on NTU RGB+D dataset

网络	$\delta_J/\%$	$\delta_B/\%$	$\delta_{J-M}/\%$	$\delta_{B-M}/\%$
STGCN++	89.30	92.30	84.00	88.80
STGCN++_MFTCN	90.42	92.32	86.60	89.23
STGCN++_HEConv	90.53	91.02	85.65	88.02
STGCN++_SAM_CAM	90.72	91.81	86.49	88.62
AMHGCN	91.75	92.45	86.58	88.25

由表 1 可得以下 4 个结论。

1) 加入 MFTCN 后,相较于基线网络 STGCN++,J 流、B 流、J-M 流、B-M 流的识别准确率分别提高 1.12%、0.02%、2.60%、0.43%,说明 MFTCN 的加入增强了对时间域的数据处理,在不同层级的网络中提取特征,可以捕获更广泛的上、下文信息。较深层次的特征通常对于全局语义信息具有较好的把握,而较浅层次的特征可以提供更多局部细节信息。通过多尺度卷积,模型可以在不同层级上同时利用全局和局部信息,从而更好地理解图像内容。

2) 加入 HEConv 后,相较于 STGCN++,J 流、J-M 流的识别准确率分别提高 1.23%、1.65%,说明 HEConv 能更好地捕捉关节之间的关系,逐步捕获从局部到全局的空间结构信息,尤其是在多层次的情况下,可以捕获节点之间的长程依赖关系,同时也让模型在关节点细节上有更强的识别能力,这对具有相似人体动作的类别具有较强的区分效果。

3) 加入空间通道注意力机制后,相较于 STGCN++,J 流、J-M 流的识别准确率分别提高 1.42%、2.49%,因为在人体动作中,并非所有关节对每一个动作都同样重要,空间注意力机制可以动态调整网络的关注点,使其更加聚焦于对当前动作重要的关节点,从而提高模型的识别效果。不同的动作类型可能对应不同的重要通道,通道注意力机制可以根据动作特性自适应地分配通道权重,使模型能够灵活应对不同动作。

4) 相较于 STGCN++,AMHGCN 的 J 流、B 流、J-M 流的识别准确率分别提高 2.45%、0.15%、2.58%。

2.5 扶梯数据集消融实验

在扶梯危险行为数据集上,评估注意力引导的多尺度层次边缘聚合时序图卷积网络。将扶梯危险行为数据集的数据格式处理成 NTU RGB+D 数据集的格式,再进行训练和模型评估。首先,截取视频文件中的每一帧,用 Faster R-CNN 对自动扶梯上的行人进行跟踪,并框出行人的身体范围。然后,使用自顶向下的 HRNet 人体姿态估计器识别人体的 25 个关键节点的 x 、 y 坐标,再区分每一视频帧中每个人的人体 25 个关键节点的 x 、 y 坐标,并将同一个人所有帧的 x 、 y 坐标存储在一起。最后,把 2 500 个视频中提取到的所有 x 、 y 坐标数据格式转换成 NTU RGB+D 数据集的格式。

扶梯危险行为数据集上的消融实验结果,如表 2 所示。表 2 中: t_J 、 t_B 、 t_{J-M} 、 t_{B-M} 分别为关节点流、骨骼流、关节点运动流、骨骼运动流的识别速率。

表 2 扶梯危险行为数据集上的消融实验结果

Tab. 2 Ablation experiments results on escalator dangerous behavior dataset

网络	$\delta_J/\%$	t_J/ms	$\delta_B/\%$	t_B/ms	$\delta_{J-M}/\%$	t_{J-M}/ms	$\delta_{B-M}/\%$	t_{B-M}/ms
STGCN++	81.05	8.0	87.50	13.3	70.56	13.3	77.02	13.3
STGCN++_MFTCN	83.47	14.2	89.52	16.9	71.77	14.2	77.00	15.5
STGCN++_HEConv	82.27	16.8	86.43	15.8	72.53	16.3	76.68	16.8
STGCN++_SAM_CAM	87.90	16.0	84.68	15.6	74.60	15.6	75.81	16.9
AMHGCN	84.92	14.1	90.63	13.8	75.92	10.6	76.32	13.8

由表 2 可得以下 6 个结论。

1) 相较于 STGCN++,加入 MFTCN 之后,J 流、B 流、J-M 流的识别准确率分别提高了 2.42%、2.02%、1.21%;加入 HEConv 后,J 流、J-M 流的准确率分别提高 1.22%、1.97%;加入空间通道注意力机制后,J 流、J-M 流分别提高 6.85%、4.04%。

2) 同时使用 MFTCN, HEConv 和空间通道注意力机制的 AMHGCN 在 J 流、B 流、J-M 流的识别准确率分别提高了 3.87%、3.13%、5.36%, 且 AMHGCN 的对每个视频的识别速度比单独使用一个模块更快。

3) J 流模型只加入空间通道注意力后, 识别准确率有较大提升, 因为在扶梯危险行为数据集中的人员比较密集, 引入空间注意力机制和通道注意力机制可以显著提高 J 流的识别准确率, 这些注意力机制能够更好地聚焦于 J 流中的关键信息, 从而优化模型对人体动作的理解和识别。

4) B 流模型的最高识别准确率为 90.63%, 相较于基线网络提高了 3.13%, 因为在扶梯危险行为数据集上关节点遮挡严重, 而利用骨骼的高阶信息数据有利于识别扶梯上危险行为。源关节是距离骨架的重心近关节点, 目标关节则是距离骨架的重心更远的关节点, 从源关节指向其目标关节的向量就可以表示一个骨骼, 这些骨骼向量都是符合人体自然骨架连接的, 处理的骨骼数据包含着骨骼的方向和长度的二阶信息, 进而可以处理更具特征的角度信息。

5) 在 J-M 流中, AMHGCN 相较于 STGCN++ 提升幅度最大, 提升了 5.36%, 可能是由于加入多尺度特征时序卷积网络使模型能更加捕捉到不同帧关节点的信息, 这对于运动特征有很大提升, 同时, 层次边缘卷积结合空间通道注意力机制可以将关键的关节点信息放大而忽略无关或次要特征, 并且构建边缘特征, 提高识别准确率。

6) 在 B-M 流中, 单独使用 HEConv 和空间通道注意力机制的识别准确率有所下降, 因为这两个模块都对关节点产生过多关注, 导致对于骨骼数据的关注产生偏差, 进而使 AMHGCN 对运动的骨骼数据识别准确率下降。

2.6 AMHGCN 与其他先进网络的比较

在 NTU RGB+D 数据集上, 将 AMHGCN, STGCN++ 与基于人体骨架数据的其他先进网络 (STGCN、AAGCN^[23]、MS-G3D^[24]、CTRGCN^[25]) 进行比较。

NTU RGB+D 数据集上识别准确率的比较结果, 如表 3 所示。

表 3 NTU RGB+D 数据集上识别准确率的比较结果

Tab. 3 Comparison results of recognition accuracy on NTU RGB+D dataset

网络	$\delta_J/\%$	$\delta_B/\%$	$\delta_{J-M}/\%$	$\delta_{B-M}/\%$
STGCN	81.5	—	—	—
AAGCN	88.0	88.4	85.9	86.0
MS-G3D	89.4	90.1	—	—
CTRGCN	90.6	92.7	89.4	90.3
STGCN++	89.3	92.3	84.0	88.8
AMHGCN	91.8	92.5	86.6	88.3

由表 3 可知: 相较于其他先进网络, AMHGCN 在 4 个流的分类准确率都高于 STGCN、AAGCN、MS-G3D, 在 J 流中, AMHGCN 比 CTRGCN 高 1.2%; 在 NTU RGB+D 数据集上, 提出的注意力引导的多尺度层次边缘聚合时序图卷积网络的识别分类效果普遍有一定的提高。

为进一步比较 AMHGCN 在自动扶梯场景中的性能, 将其与其他先进网络 (STGCN、AAGCN、CTRGCN) 在扶梯危险行为数据集上进行比较, 结果如表 4 所示。

表 4 扶梯危险行为数据集上识别准确率的比较结果

Tab. 4 Comparison results of recognition accuracy on escalator dangerous behavior dataset

网络	$\delta_J/\%$	t_J/ms	$\delta_B/\%$	t_B/ms	$\delta_{J-M}/\%$	t_{J-M}/ms	$\delta_{B-M}/\%$	t_{B-M}/ms
STGCN	72.58	6.7	85.48	6.2	67.34	6.7	72.18	6.2
AAGCN	77.42	13.8	86.29	14.2	64.92	14.7	76.21	14.2
CTRGCN	77.42	14.2	87.50	13.7	68.55	13.8	71.77	13.8
AMHGCN	83.87	13.8	90.73	13.8	75.40	10.6	77.42	13.8

由表 4 可知以下 3 个结论。

1) AMHGCN 在扶梯危险行为数据集上对危险行为的识别准确率明显优于其他网络。STGCN 通过结合图卷积网络和时序卷积网络 (TCN) 首次实现人体骨架行为识别中的时空联合建模, 能够高效地

捕捉人体动作中的复杂空间和时间依赖关系。AAGCN 引入自适应邻接矩阵实现自适应的图卷积操作,这个邻接矩阵不是固定的,而是由网络根据数据自适应地生成和调整的,它允许模型学习到骨架中关节点之间的潜在关系,而不仅仅依赖于预定义的骨架拓扑结构。CTRGCN 不再依赖固定的骨架结构,引入通道级别的拓扑优化机制,允许网络在每个通道上动态调整关节点之间的连接,从而实现更精细的动作建模和更高的识别准确度。

2) 在 J-M 流中,AMHGCN 的识别准确率远高于其他先进模型,这得益于加入了多尺度特征时序卷积网络、层次边缘卷积和空间通道注意力机制,缓解了部分身体被遮挡带来的识别问题,使模型更加关注前、后相邻帧来获取残缺的信息,并且更加关注人体骨架图中的全局信息和细节特征,缓解了相似动作对模型的影响。

3) AMHGCN 采用交叉熵损失函数,用于比较预测类别分布和真实类别分布的差异,衡量模型的性能。同时,引入非线性 ReLU 激活函数能帮助模型更好地捕捉复杂的图结构和时空关系。

AMHGCN 在 NTU RGB+D 数据集上训练 80 个周期,以及在扶梯危险行为数据集上训练 100 个周期后可得函数损失(L),结果如图 5、6 所示。

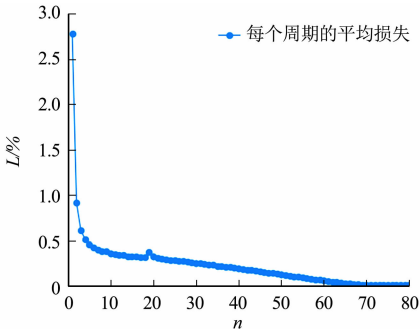


图 5 NTU RGB+D 数据集损失函数图
Fig. 5 Loss function diagram of NTU RGB+D dataset

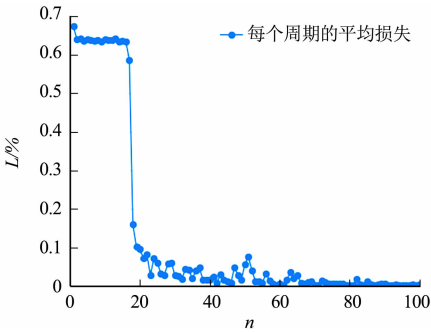


图 6 扶梯危险行为数据集损失函数图
Fig. 6 Loss function diagram of escalator dangerous behavior dataset

2.7 自动扶梯危险行为识别性能比较

在扶梯危险行为识别中漏报危险行为的代价很高,漏报意味着系统未能识别出实际存在的危险行为,将其误判为正常行为。这会导致重大的人员伤害和伤亡,因此,需要评测 AMHGCN 的召回率。

选用识别准确率最高的 B 流进行测试,设置 2 个训练集和 1 个测试集,训练集 1 包含 500 个危险行为视频和 500 个正常行为视频,训练集 2 包含 500 个危险行为视频和 1 000 个正常行为视频,测试集包含 50 个危险行为视频和 50 个正常行为视频。

AMHGCN 自动扶梯危险行为识别性能比较,如表 5 所示。表 5 中: t 为识别速率。

表 5 AMHGCN 自动扶梯危险行为识别性能比较

Tab. 5 Comparison of dangerous behavior recognition performance of escalators in AMHGCN

训练集	$\eta_p / \%$	$\eta_r / \%$	t / ms
训练集 1	97.06	82.00	12.1
训练集 2	98.70	98.00	12.1

由表 5 可知:训练集中危险行为与正常行为之比为 1 : 1 时,准确率较高,但召回率仅有 82.00%,说明 50 个危险行为样本中,有 9 个样本被分为正常行为;危险行为与正常行为之比为 1 : 2 时,召回率为 98.00%,说明 50 个危险行为样本中,仅有 1 个被分为正常行为。

正常行为样本显著增加使模型在学习正常行为的特征时变得更加敏感和准确,因此,对危险行为的区分能力也得到了提升。这种调整在训练时帮助模型更好地区分不同的类别,从而减少了漏报情况,提高了 AMHGCN 的召回率。召回率的提升能够有效防止自动扶梯的运行风险,避免潜在事故的发生,同时,高召回率可以减少人工干预的需要,提高系统自动化水平,从而提升自动扶梯的运行效率。在实时的安全任务中,高召回率可以确保大多数危险行为被识别到,而高准确率可以确保被标记危险行为的样本是真正的危险行为。AMHGCN 模型的召回率和准确率都保持较高水平,可最大限度地减少漏报

和误报,从而提升系统在工程上的实际应用效果。

3 结论

为了使图神经网络从前、后相邻帧中获取缺失的信息,且更加关注每一帧的动作细节,提出一种注意力引导的多尺度层次边缘聚合时序图卷积网络。将时序卷积加入 7 个多尺度分支,增强网络对时间域中前后相邻帧的特征提取能力,以缓解扶梯狭长空间导致的部分关键节点遮挡。加入层次边缘卷积后,对人体骨架的关注由局部特征向全局特征扩展,从而增强识别准确率,通过全局特征精准识别相似的动作。

加入空间通道注意力机制后,强化对不同行为细节特征的关注来提高行为识别准确率。相较于 STGCN++,AMHGCN 在 NTU RGB+D 数据集上 J 流、B 流、J-M 流分别提高了 2.45%、0.15%、2.58%。AMHGCN 在扶梯危险行为数据集上 J 流、B 流、J-M 流分别提高了 3.87%、3.13%、5.36%,在扶梯危险行为数据集上,AMHGCN 的识别准确率均高于其他先进网络。相较于 AAGCN 和 CTRGCN,AMHGCN 的识别速率在 J-M 流有所提升,在其他 3 个流大致相同。对模型召回率进行测试,当正常行为样本多于危险行为样本时,可以得到很高的召回率。

相较于 J 流和 B 流,J-M 流和 B-M 流的识别准确率普遍较低。因此,下一阶段的研究目标考虑以胸部、腹部、臀部为中心,构造 3 种图结构对 J 流、B 流进行训练,并将多个流的识别结果进行加权求和,得到更高的识别准确率。

参考文献:

[1] 舒文华,欧阳惠卿. 自动扶梯乘客行为智能感知和自主安全管理技术标准探讨[J]. 质量与标准化,2021,354(10): 39-43. DOI:10.3969/j.issn.2095-0918.2021.10.015.

[2] 蒋儒浩. 自动扶梯综合性能检测仪研制[D]. 合肥:合肥工业大学,2019. DOI:10.27101/d.cnki.ghfgu.2019.000056.

[3] 付春平. 自动扶梯几起安全事故的共性分析与探讨[J]. 科技与创新,2023,217(1):82-84,89. DOI:10.15913/j.cnki.kjycx.2023.01.023.

[4] 张栓柱. 基于事故树的商场电梯事故分析[J]. 消防界(电子版),2022,8(21):21-23. DOI:10.16859/j.cnki.cn12-9204/tu.2022.21.040.

[5] 解云蕾. 自动扶梯安全探讨[J]. 中国科技信息,2022(3):64-66. DOI:10.3969/j.issn.1001-8972.2022.03.021.

[6] 李伟达,叶靓玲,郑力新,等. 面向扶梯不安全行为的改进型深度学习检测算法[J]. 华侨大学学报(自然科学版),2022,43(1):119-126. DOI:10.11830/ISSN.1000-5013.202105059.

[7] 叶靓玲,李伟达,郑力新,等. 结合目标检测与特征匹配的多目标跟踪算法[J]. 华侨大学学报(自然科学版),2021,42(5):661-669. DOI:10.11830/ISSN.1000-5013.202105018.

[8] YE Liangling,LI Weida,ZHENG Lixin,*et al.* Lightweight and deep appearance embedding for multiple object tracking[J]. IET Computer Vision,2022,16(6):489-503. DOI:10.1049/cvi2.12106.

[9] 林志鸿,郑力新,曾远跃. 采用空间依赖的 MTDPN 扶梯危险行为的姿态估计[J]. 华侨大学学报(自然科学版),2023,44(6):751-758. DOI:10.11830/ISSN.1000-5013.202305020.

[10] YAN Sijie,XIONG Yuanjun,LIN Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans:AAAI Press,2018: 7444-7452. DOI:10.48550/arXiv.1801.07455.

[11] ZHANG Xikun,XU Chang,TIAN Xinmei,*et al.* Graph edge convolutional neural networks for skeleton-based action recognition[J]. IEEE Transactions on Neural Networks and Learning Systems,2019,31(8):3047-3060. DOI: 10.48550/arXiv.1805.06184.

[12] SHI Lei,ZHANG Yifan,CHENG Jian,*et al.* Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach:IEEE Press,2019:12026-12035. DOI:10.48550/arXiv.1805.07694.

[13] WEN Yuhui,GAO Lin,FU Hongbo,*et al.* Graph CNNs with motif and variable temporal block for skeleton-based action recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu:AAAI Press,

- 2019;8989-8996. DOI;10.1609/aaai.v33i01.33018989.
- [14] LI Bin, LI Xi, ZHANG Zhongfei, *et al.* Spatio-temporal graph routing for skeleton-based action recognition[C]// Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu; AAAI Press, 2019; 8561-8568. DOI; 10.1609/aaai.v33i01.33018561.
- [15] LI Maosen, CHEN Siheng, CHEN Xu, *et al.* Actional-structural graph convolutional networks for skeleton based action recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach; IEEE Press, 2019; 3595-3603. DOI; 10.1109/CVPR.2019.01230.
- [16] SHI Lei, ZHANG Yifan, CHENG Jian, *et al.* Skeleton-based action recognition with directed graph neural networks [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach; IEEE Press, 2019; 7912-7921. DOI; 10.1109/CVPR.2019.00810.
- [17] LEE I, KIM D, KANG S, *et al.* Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks[C]// Proceedings of the IEEE International Conference on Computer Vision, Venice; IEEE Press, 2017; 1012-1020. DOI; 10.1109/ICCV.2017.115.
- [18] CHEN Yuxin, ZHANG Ziqi, YUAN Chunfeng, *et al.* Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal; IEEE Press, 2021; 13359-13368. DOI; 10.48550/arXiv.2107.12213.
- [19] 袁正中, 李灿东. 复杂网络控制核心的进一步分析[J]. 闽南师范大学学报(自然科学版), 2023, 36(2): 27-34. DOI: 10.16007/j.cnki.issn2095-7122.2023.02.008.
- [20] LIU Ziyu, ZHANG Hongwen, CHEN Zhenghao, *et al.* Disentangling and unifying graph convolutions for skeleton-based action recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle; IEEE Press, 2020; 143-152. DOI; 10.1109/CVPR42600.2020.00022.
- [21] SHAHROUDY A, LIU Jun, NG T T, *et al.* Ntu rgb+d: A large scale dataset for 3d human activity analysis[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas; IEEE Press, 2016; 1010-1019. DOI; 10.1109/CVPR.2016.115.
- [22] DUAN Haodong, WANG Jiaqi, CHEN Kai, *et al.* Pyskl: Towards good practices for skeleton action recognition [C]// Proceedings of the 30th ACM International Conference on Multimedia, Nicosia; ACM Press, 2022; 7351-7354. DOI; 10.48550/arXiv.2205.09443.
- [23] SHI Lei, ZHANG Yifan, CHENG Jian, *et al.* Skeleton-based action recognition with multi-stream adaptive graph convolutional networks[C]// IEEE Transactions on Image Processing. [S. l.]; IEEE Press, 2020; 9532-9545. DOI; 10.1109/TIP.2020.3028207.
- [24] LIU Ziyu, ZHANG Hongwen, CHEN Zhenghao, *et al.* Disentangling and unifying graph convolutions for skeleton-based action recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle; IEEE Press, 2020; 143-152. DOI; 10.1109/CVPR42600.2020.00022.
- [25] CHEN Yuxin, ZHANG Ziqi, YUAN Chunfeng, *et al.* Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal; IEEE Press, 2021; 13359-13368. DOI; 10.48550/arXiv.2107.12213.

(责任编辑: 钱筠 英文审校: 陈婧)