计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(12): 3028−3041, 2025

融合运动领域知识与自适应时空 Transformer 的人体骨架行为识别

梁成武[1,2] 蒋松琪[1,2] 杨 杰[1,2,3] 朱培旺[1,2] 帖 云[5] 高 磊[6] 胡 伟[1,2,4] 郭文博[1]

[1](河南城建学院电气与控制工程学院 河南平顶山 467041)

[2](三峡大学电气与新能源学院 湖北宜昌 443002)

[3](国网湖南省电力有限公司怀化供电分公司 湖南怀化 418006)

[4](国网安徽省电力有限公司铜陵供电分公司 安徽铜陵 244151)

[5](郑州大学电气与信息工程学院 郑州 450001)

[6](多伦多都会大学电气计算机和生物医学工程系 加拿大多伦多 M5B2K3)

(liangchengwu0615@126.com)

Skeleton-Based Human Action Recognition with Fusion of Motion Domain Knowledge and Adaptive Spatio-Temporal Transformer

Liang Chengwu[1,2], Jiang Songqi[1,2], Yang Jie[1,2,3], Zhu Peiwang[1,2], Tie Yun[5], Gao Lei[6], Hu Wei[1,2,4], and Guo Wenbo[1]

[1](*School of Electrical and Control Engineering*, *Henan University of Urban Construction*, *Pingdingshan*, *Henan* 467041)

[2](*College of Electrical Engineering and New Energy*, *China Three Gorges University*, *Yichang*, *Hubei* 443002)

[3](*State Grid Hunan Electric Power Co.*, *Ltd. Huaihua Power Supply Branch*, *Huaihua*, *Hunan* 418006)

[4](*State Grid Anhui Electric Power Co.*, *Ltd. Tongling Power Supply Branch*, *Tongling*, *Anhui* 244151)

[5](*School of Electrical and Information Engineering*, *Zhengzhou University*, *Zhengzhou* 450001)

[6](*Department of Electrical*, *Computer and Biomedical Engineering*, *Metropolitan University of Toronto*, *Toronto*, *Canada M5B2K3*)

Abstract Existing methods for skeleton-based human action recognition always ignore motion domain knowledge, resulting in the lack of interpretability of logical decision-making that human can understand. In this paper, we propose a novel skeleton-based human action recognition method by fusion of domain knowledge and adaptive spatio-temporal Transformer, to improve recognition performance and interpretability. Firstly, inspired by the short-term motion knowledge, a temporal multi-branch structure is designed to learn and capture the characteristics of short-term sub-acitons. Secondly, a dynamic information fusion module is proposed to learn the weight vectors of different temporal branches, and then fuse multiscale short-term motion features. Finally, to learn the relationship between different sub-actions and facilitate the motion information interaction between skeleton joints, a multiscale temporal convolution feature fusion module is proposed to capture the long-term motion correlations, by integrating the domain knowledge of the long-term motion. Experimental evaluations are conducted on four large action datasets, including NTU RGB+D, NTU RGB+D 120, FineGym, and InHARD. The experimental results show that the recognition performance of the proposed method is superior to several data-driven methods, effectively improving the modelling ability of short-term motion feature learning and information interaction between skeleton joints, with the interpretability.

Key words skeleton-based action recognition; domain knowledge; temporal segment; sub-aciton; Transformer; self-

收稿日期: 2024-06-21; **修回日期**: 2025-04-22

基金项目: 国家自然科学基金项目(62176086, 62505076, U1804152); 河南省科技攻关计划项目(242102211055, 242102210032); 河南省自然科学基金项目(242300421405); 河南省软科学计划项目(252400410618)

This work was supported by the National Natural Science Foundation of China (62176086, 62505076, U1804152), the Science and Technology Program of Henan Province (242102211055, 242102210032), the Natural Science Foundation of Henan Province (242300421405), and the Soft Science Research Program of Henan Province (252400410618).

attention mechanism

摘　要　现有人体骨架行为识别方法往往忽略运动学领域知识，造成模型内在人类可理解的逻辑决策解释性不足. 基于此，提出一种融合领域知识与自适应时空片段 Transformer 的骨架行为识别方法，以提高骨架行为识别模型的性能和可解释性. 首先，受短时运动领域知识启发，设计多时间分支结构用于学习和捕捉多时间尺度的短时子动作特征. 其次，提出一种动态信息融合模块，学习不同时间分支的权重向量进而动态融合多时间分支、多尺度短时运动特征. 最后，融合长时运动领域知识，提出多尺度时间卷积特征融合模块捕捉长时运动关联，用于学习不同子动作片段之间的关系并促进不同骨架关节点间的运动信息交互. 在 4 个数据集上进行评估与实验，包括人体日常行为数据集 NTU RGB+D 和 NTU RGB+D 120、体育行为数据集 FineGym，以及工业场景行为数据集 InHARD. 结果表明所提方法的行为识别性能，优于包括基准 Transformer 方法在内的多个先进方法，可有效提升骨架序列短时运动特征学习和关节点之间信息交互的建模能力，并具有一定可解释性.

关键词　骨架行为识别；领域知识；时间片段；子动作；Transformer；自注意力机制

中图法分类号　TP391

DOI: 10.7544/issn1000-1239.202440558　　**CSTR:** 32373.14.issn1000-1239.202440558

人体行为识别作为计算机视觉领域的研究热点，在视频监控、自动驾驶以及人机交互等多个领域具有广泛应用[1-3]. 人体行为识别技术可以赋予具身智能系统理解和适应人类工作生活环境的能力，提高智能体对真实世界的理解能力. 鉴于骨架序列数据的紧凑性和对背景与光照条件变化的鲁棒性，基于骨架序列的人体行为识别任务逐渐引起研究者的广泛关注. 随着深度学习的迅速发展，目前基于骨架序列的人体行为识别方法主要分为基于图卷积神经网络（graph convolutional neural network, GCN）[4-8] 和 Transformer 的方法[9-11]. Yan 等人[4] 利用骨架序列的人体内在拓扑结构和时间关系构建时空图，并通过空间图卷积和时间卷积逐步学习更高层次的时空运动特征. 针对 GCN 难以有效捕捉长时特征和全局依赖关系的局限性，石跃祥等人[11] 提出协作卷积 Transformer 模型，通过自注意力机制提升模型的全局特征建模能力. 人体运动领域知识和骨架行为识别密切相关，主要包括人体运动的动态规律、行为习惯模式和时空上下文先验知识等. 尽管以上方法在人体骨架识别任务中取得一定进展，但往往忽略人体短时子动作、长时运动特征学习在模型设计中的重要性，缺乏人体运动领域知识的运用，导致骨架序列数据特征的挖掘学习和可解释性不足.

由短时运动领域知识可知，人体行为可以由多个具有运动学语义的基本短时子动作组成，因此对短时骨架子动作的显式建模有助于学习和分类完整的人体行为. 具体而言，骨架序列中短时间范围（相邻帧）中的不同关节点通常呈现高度相关性. 以“跳远”动作为例，前一帧的手臂与后一帧的腿部动作会协同运动并存在一定相关性. 此外，不同人体行为的子动作也存在持续时间、运动速率等变化规律的差异，对不同的人体行为需要进行针对性设计. 受上述人体运动领域知识启发，行为识别研究者们对骨架行为识别方法进行一系列改进[12-15]. 其中 Wang 等人[13] 提出一种时间差分网络，采用先建模短时运动信息，后建模长期运动信息的策略，显著提升了人体行为动作的时空特征建模能力. 基于人体行为可由若干子动作组成这一领域知识，Qiu 等人[15] 提出一种固定长度的时间分段注意力机制对短时运动特征进行显式建模，但未能考虑到不同人体行为的特征差异性，难以适用于不同行为的短时子动作建模.

短时子动作特征的建模是学习和识别人体行为的基础，与之不同，长时运动特征的建模可以学习不同子动作间的关联以及不同骨架关节点间的长时相关性. Chen 等人[7] 采用多尺度时间卷积模块学习不同尺度的时间特征，忽略了不同骨架关节点之间的长时信息交互. 考虑到现有方法缺乏对关节点间的长时依赖建模，Duan 等人[16] 在时间特征建模的过程中通过引入全局骨架特征来增强不同关节点之间的信息交互，但缺少对不同局部子动作之间关联的建模.

基于上述分析，本文提出一种融合运动领域知识与自适应时间片段注意力的 Transformer 模型用于人体骨架行为识别. 主要贡献如下：

1）受短时运动领域知识启发，提出自适应时间

分段注意力(adaptive temporal segment attention, ATSA)模块学习不同行为的短时子动作特征. ATSA 模块通过多时间分支结构对不同时间尺度的局部人体子运动特征进行建模，并通过动态信息融合模块整合多时间分支、多尺度子动作特征.

2)融合长时运动领域知识，提出多尺度时间卷积特征融合模块 MTCF(multiscale temporal convolution with skeleton feature fusion)学习长时行为片段间的运动特征和不同子动作间的关联. MTCF 采用不同感受野的 2D 时间卷积进行长时特征建模，以学习人体行为子动作间的关联，同时引入骨架特征融合机制增加人体骨架不同关节点间的信息交互.

3)在 4 个数据集 NTU RGB+D[17], NTU RGB+D 120[18], FineGym[19], InHARD[20] 上进行性能评估，分别取得 93.2%(Cross-Subject)、89.7%(Cross-Subject)、95.0% 和 88.3% 的 Top1 识别准确率. 结果表明本文方法通过融合长短时运动领域知识，可有效增强模型的短时运动特征建模能力和长时信息交互能力，进而提高骨架行为识别性能和可解释性.

1 本文方法

现有人体行为识别模型在捕捉长短时运动特征以及可解释性方面存在局限性，模型内在人类可理解的逻辑决策解释性不足. 受短时运动领域知识启发，设计 ATSA 用于学习和捕捉多时间尺度的短时子动作特征. 为学习骨架运动不同时间分支的权重向量及其动态融合，提出动态信息融合(dyamic information fusion, DIF)模块进行多时间尺度分支短时运动特征融合. 为捕捉长时骨架运动关联，提出 MTCF 学习不同子动作片段之间的关系.

如图 1 所示，首先，线性嵌入层将骨架序列中的关节点坐标信息映射为高维特征向量序列. 然后，输入特征序列被送入L层堆叠的编码器模块进行时空特征建模，其中每一层分别由 ATSA 和 MTCF 组成. 最后，得到的输出特征经全局平均池化层、线性层和 Softmax 层后输出预测行为类别.

具体地，图 1 右下侧 ATSA 模块中实线和虚线分别表示不同时间尺度分支所对应帧内和帧间的关节点信息交互. ATSA 模块通过学习相邻帧中不同关节点的相关性来建模局部子动作特征. 图 1 右上侧 MTCF 模块中实线和虚线代表骨架特征融合过程和多尺度时间卷积. MTCF 模块采用多尺度时间卷积来学习不同子动作之间的关联.

综上，本文提出一种融合运动领域知识与自适应时间片段注意力的 Transformer 方法用于人体骨架

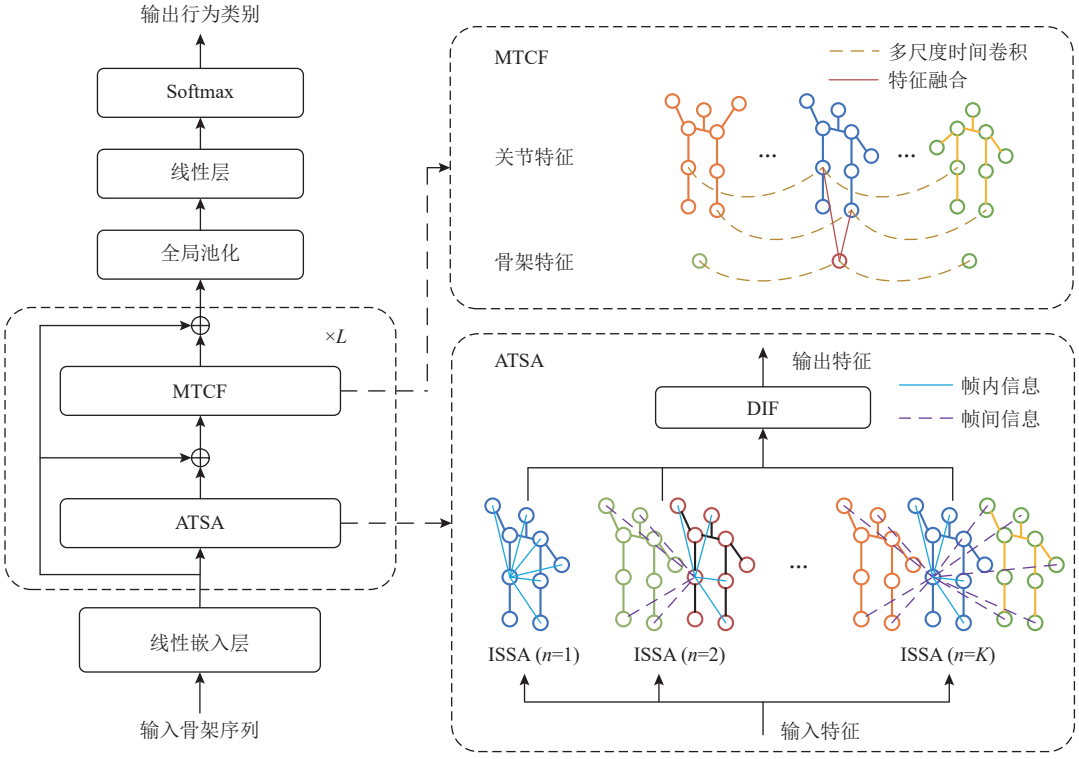


Fig. 1 Fusion of motion domain knowledge and adaptive spatio-temporal Transformer model

图 1 融合运动领域知识与自适应时空 Transformer 模型

行为识别.

1.1 线性嵌入层

Transformer 模型中嵌入层的目的是将离散化的输入数据转换为高维密集的连续向量. 这些向量进一步与位置编码相结合, 共同构成了模型处理的输入序列. 本文方法中的线性嵌入层用于将输入骨架序列编码为输入特征向量序列.

不同于传统视觉 Transformer[21] 将图像裁剪为若干固定大小的图像块, 然后进行线性映射后编码得到输入特征向量序列的方法, 骨架序列仅由少量的人体关节点组成, 因此可直接对每个关节点进行线性映射得到输入特征向量矩阵:

$$\tilde{\boldsymbol{Z}}_0 = Conv2D_{(1\times1)}(\boldsymbol{X}_{\text{seq}}) \in \mathbb{R}^{T\times V\times C}, \tag{1}$$

其中 $\tilde{\boldsymbol{Z}}_0$ 表示经过线性映射的特征向量矩阵. T, V, C 分别表示帧数、关节点数和特征通道数. $Conv2D_{(1\times1)}$ 表示卷积核大小为 1×1 的 $2D$ 卷积.

经线性映射得到的特征向量序列未包含人体骨架序列的空间位置、时间顺序信息以及固有属性信息, 需要引入位置编码向量对该信息进行补充. 本文与 Transformer[22] 一致, 采用不同频率的正弦和余弦函数来生成位置编码向量 $\boldsymbol{PE}$, 其中

$$\boldsymbol{PE}(p,2i) = \sin\left(p/10\,000^{2i/C}\right), \tag{2}$$

$$\boldsymbol{PE}(p,2i+1) = \cos\left(p/10\,000^{2i/C}\right), \tag{3}$$

其中 p 表示每个特征向量在输入序列中所处的位置, $i\in\{0,1,\cdots,C/2\}$ 表示位置编码向量的具体索引. 然后, 将特征向量序列和位置编码向量相加得到输入特征矩阵 $\boldsymbol{Z}_{\text{in}}\in\mathbb{R}^{T\times V\times C}$:

$$\boldsymbol{Z}_{\text{in}} = \tilde{\boldsymbol{Z}}_{\text{in}} + \boldsymbol{PE}. \tag{4}$$

经线性嵌入层后, 输入特征矩阵 $\boldsymbol{Z}_{\text{in}}$ 送入 L 层堆叠的编码器模块进行特征建模.

1.2 段间自注意力机制

基于人体运动领域知识可知, 完整的人体行为可视为若干个子动作在时间维度上的语义组合, 因此对人体子动作特性的学习是对人体行为分析理解的基础和关键. 为使模型能有效学习局部时间片段窗口内的子动作特征, 段间自注意力机制(intra-segment self-attention, ISSA)对固定长度时间片段内的短时运动特征进行显式建模.

如图 2 所示, 相较传统时空自注意力机制[23] 对时空特征进行全局建模, ISSA 通过将自注意力机制操作限制在局部时间片段内, 以此达到对子动作特征显示建模的效果. 首先输入特征 $\boldsymbol{Z}_{\text{in}}$ 沿时间维度被分为 T_n 个时间长度相等且不重叠的时间片段:

$$\boldsymbol{Z}_{\text{in}}\in\mathbb{R}^{T\times V\times C}\rightarrow\mathbb{R}^{T_n\times n\times V\times C}\rightarrow\mathbb{R}^{T_n\times V_n\times C}, \tag{5}$$

其中 n 表示每个时间片段的长度, $V_n = V\times n$ 表示每个时间片段内的关节点总数.

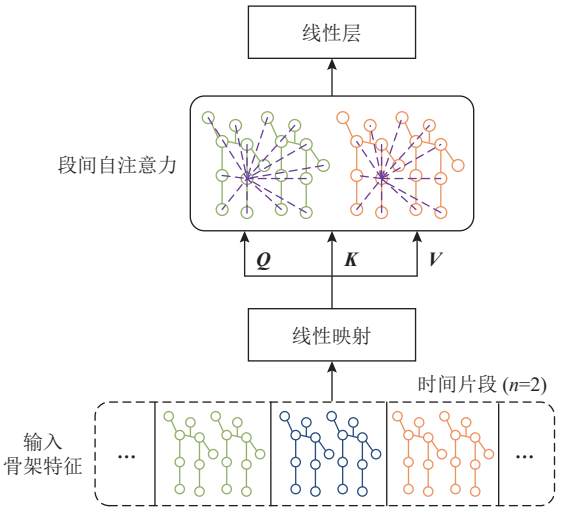


Fig. 2 Inter-segment self-attention mechanism

图 2 段间自注意力机制

经过时间分段后, ISSA 对每个时间片段分别进行自注意力机制操作. 自注意力机制通过查询($\boldsymbol{Q}$)与键-值对($\boldsymbol{K}$-$\boldsymbol{V}$)之间的交互实现对时间片段内不同关节点信息的动态聚合. $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ 分别通过对输入特征 $\boldsymbol{Z}_{\text{in}}$ 进行线性映射得到:

$$\boldsymbol{Q} = \boldsymbol{W}_{\boldsymbol{Q}}\boldsymbol{Z}_{\text{in}},\ \boldsymbol{K} = \boldsymbol{W}_{\boldsymbol{K}}\boldsymbol{Z}_{\text{in}},\ \boldsymbol{V} = \boldsymbol{W}_{\boldsymbol{V}}\boldsymbol{Z}_{\text{in}}, \tag{6}$$

其中 $\boldsymbol{W}_{\boldsymbol{Q}}$, $\boldsymbol{W}_{\boldsymbol{K}}$, $\boldsymbol{W}_{\boldsymbol{V}}$ 分别对应 $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ 的线性映射矩阵. 通过计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的相关性可得到注意力权重矩阵, 然后将其与 $\boldsymbol{V}$ 相乘即可得到输出特征:

$$\boldsymbol{D} = \tanh(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{C}), \tag{7}$$

$$\boldsymbol{Z}_{\text{attn}} = \boldsymbol{V}(\alpha\times\boldsymbol{D}+\boldsymbol{P}), \tag{8}$$

其中 $\boldsymbol{Z}_{\text{attn}}$ 表示自注意力机制操作的输出特征. tanh 表示双曲正切函数, $\sqrt{C}$ 用于提高训练过程中的梯度稳定性. $\boldsymbol{D}$ 表示通过计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 之间的相似度得到的注意力权重矩阵. 为引入人体骨架内在拓扑关系, 增加可学习的参数矩阵 $\boldsymbol{P}$ 用于表示人体固有铰链的拓扑连接关系, 该矩阵在模型训练过程中迭代更新参数. α 为可学习的矩阵权重参数, 用于平衡 $\boldsymbol{D}$ 和 $\boldsymbol{P}$ 的关系.

为提高特征多样性和模型性能, ISSA 同样采用多头注意力机制. 多头注意力机制的计算方法是将输入特征分为多组后, 分别进行自注意力机制计算, 最后通过拼接得到输出特征 $\boldsymbol{Z}_{\text{attn}}$:

$$\boldsymbol{Z}_{\text{attn}} = Concat\left(\boldsymbol{Z}^1_{\text{attn}},\boldsymbol{Z}^2_{\text{attn}},\cdots,\boldsymbol{Z}^h_{\text{attn}}\right), \tag{9}$$

其中 $Concat$ 表示沿特征通道维度进行矩阵拼接, $\boldsymbol{Z}^h_{\text{attn}}$ 表示第 h 组注意力输出特征. 与 Transformer 编码器模

块类似, ISSA 最后采用全连接层用于提升网络的拟合能力:

$$\boldsymbol{Z}_{\mathrm{ISSA}} = Conv2D_{(1\times1)}(\boldsymbol{Z}_{\mathrm{attn}}), \tag{10}$$

其中$\boldsymbol{Z}_{\mathrm{ISSA}}$表示 ISSA 的输出特征. $Conv2D_{(1\times1)}$表示 1×1卷积操作, 用于实现全连接层的功能.

ISSA 通过将自注意力机制限制在固定长度的时间片段内, 从而实现对人体行为子动作特征的显式学习和捕捉. 然而, 由人体运动领域知识可知, 在不同人体行为类别之间和相同行为类别的不同时间维度, 子动作的持续时间和动态规律均存在显著差异, 这导致单一固定时间长度的 ISSA 难以适用于不同人体行为样本的子动作特征建模.

1.3 ATSA 模块

不同骨架行为样本之间, 运动速率和运动阶段存在较大差异. 因此, 需要针对每个特定骨架行为样本选择合适长度的时间片段进行短时运动特征建模. 受短时运动领域知识的进一步启发, 提出 ATSA, 用于学习不同人体行为的短时子动作特征. ATSA 采用多分支结构对不同长度的时间片段进行局部特征建模, 并通过动态信息融合模块 DIF 整合不同时间尺度的短时运动特征.

如图 3 所示, 首先输入特征$\boldsymbol{Z}_{\mathrm{in}}$沿通道维度被分为$K$组特征:

$$\boldsymbol{Z}_{\mathrm{in}}^{1}, \boldsymbol{Z}_{\mathrm{in}}^{2}, \cdots, \boldsymbol{Z}_{\mathrm{in}}^{i}, \cdots, \boldsymbol{Z}_{\mathrm{in}}^{K} = Split(\boldsymbol{Z}_{\mathrm{in}}), \tag{11}$$

其中 *Split* 表示沿通道维度对输入特征进行分割操作. 每一组特征$\boldsymbol{Z}_{\mathrm{in}}^{i} \in \mathbb{R}^{T\times V\times C/K}$被输入不同的分支以分别进行短时运动特征建模:

$$\boldsymbol{Z}_{\mathrm{ISSA}}^{i} = ISSA_{n=i}\left(\boldsymbol{Z}_{\mathrm{in}}^{i}\right), \tag{12}$$

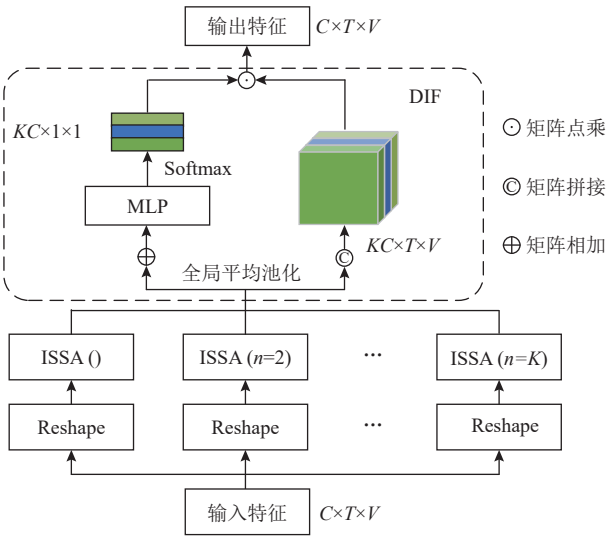


Fig. 3 ATSA module structure

图 3 ATSA 模块结构

其中$\boldsymbol{Z}_{\mathrm{ISSA}}^{i}$表示第$i$个分支的输出特征. $ISSA_{n=i}$表示采用时间片段长度为i的段间自注意力机制处理过程. 经过不同分支的 ISSA 建模后, 模型能够学习到不同时间尺度的人体行为子动作特征. 此外值得注意的是, 由于各个分支的特征通道维度仅为输入特征的$1/K$, ATSA 模块的注意力机制计算量降低为单分支的$1/K^2$, 从而有效减少了模型计算开销.

为有效适应不同人体行为的子动作差异性, 受文献 [24] 启发提出动态信息融合(dynamic information fusion, DIF)模块用于对不同时间尺度的子动作特征进行融合. DIF 由特征选择和特征融合 2 个阶段组成, 特征选择阶段基于全局表征计算每组特征的权重向量, 特征融合阶段根据每组特征的权重进行多尺度特征融合.

首先, 在特征选择阶段, 对K个分支的不同尺度的输出特征进行相加, 再送入全局平均池化层(global average pooling, *GAP*). 经由全局平均池化处理后的特征可以视为一种全局表征, 具有数据紧凑的特点, 并且可极大降低线性层的计算开销. 然后, 全局表征经过多层感知器层(multi-layer perceptron, *MLP*)和 Softmax 层计算得到每组特征各自的权重向量:

$$\boldsymbol{W} = MLP\left(GAP\left(\sum_{i=1}^{K}\boldsymbol{Z}_{\mathrm{ISSA}}^{i}\right)\right) \in \mathbb{R}^{KC\times1\times1}, \tag{13}$$

$$\boldsymbol{W}_1, \boldsymbol{W}_2, \cdots, \boldsymbol{W}_K = Split(\boldsymbol{W}) \in \mathbb{R}^{C\times1\times1}, \tag{14}$$

其中$\boldsymbol{W}_i$表示第i个分支的权重向量.

然后, 在特征融合阶段, 将每一组 ISSA 模块的输出特征与各自的权重向量相乘, 并进行求和得到最终的输出特征:

$$\boldsymbol{Z}_{\mathrm{DIF}} = \sum_{i=1}^{K}\boldsymbol{W}_i \odot \boldsymbol{Z}_{\mathrm{ISSA}}^{i}. \tag{15}$$

根据输入特征的全局信息来动态计算每一组输入特征的权重向量, 动态信息融合模块能够对不同时间尺度的特征进行动态融合, 使得模型能更好地处理不同类别和不同变化速率的人体行为样本.

此外, 为进一步提升模型训练过程中的稳定性, ATSA 模块在最终阶段引入残差连接:

$$\boldsymbol{Z}_{\mathrm{ATSA}} = DIF\left(\boldsymbol{Z}_{\mathrm{ISSA}}^{1}, \boldsymbol{Z}_{\mathrm{ISSA}}^{2}, \cdots, \boldsymbol{Z}_{\mathrm{ISSA}}^{i}, \cdots, \boldsymbol{Z}_{\mathrm{ISSA}}^{K}\right) + \boldsymbol{Z}_{\mathrm{in}}, \tag{16}$$

其中$\boldsymbol{Z}_{\mathrm{ATSA}}$表示 ATSA 模块的输出特征.

ATSA 通过多分支结构分别学习不同时间尺度的运动特征, 然后对其进行动态融合, 可有效学习和捕捉不同人体行为的子动作特征.

1.4 MTCF 模块

学习人体行为的子动作特征是行为识别的基础,

同时学习不同子动作间的关联与不同骨架关节点间的长时相关性同样重要. 受长时运动领域知识指导，提出 MTCF. 如图 4 所示，MTCF 模块首先采用多尺度时间卷积进行时间特征建模，对 ATSA 模块的短时子动作建模进行补充，然后引入一种骨架特征融合机制来加强关节点特征和骨架特征之间的动态交互.

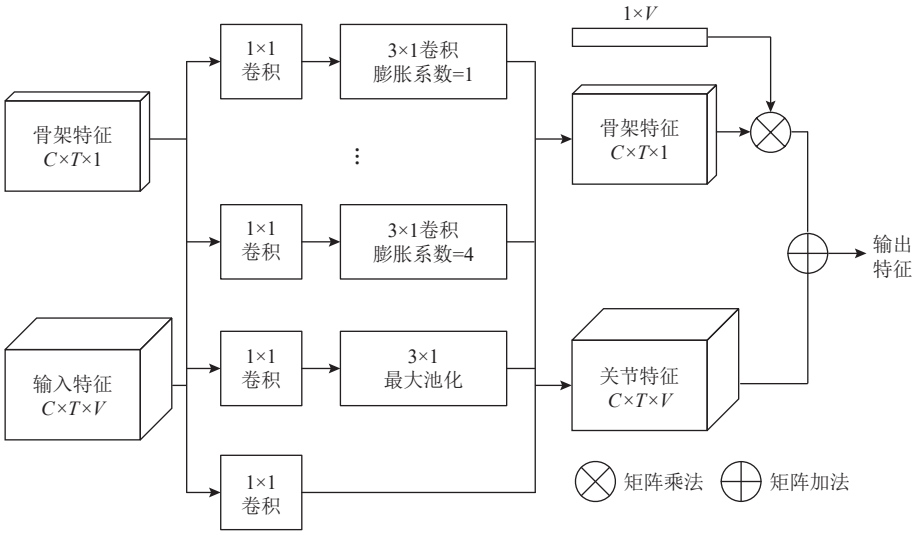


Fig. 4 MTCF module structure

图 4 MTCF 模块结构

首先，给定输入特征 $\boldsymbol{X}_{\text{in}} \in \mathbb{R}^{C\times T\times V}$，对 $\boldsymbol{X}_{\text{in}}$ 在关节点维度进行平均池化操作可以得到包含全局信息的骨架特征：

$$\boldsymbol{X}_{\text{S}} = \frac{1}{V}\sum_{i=1}^{V}\boldsymbol{X}_i, \tag{17}$$

其中 $\boldsymbol{X}_{\text{S}}$ 表示骨架特征，$\boldsymbol{X}_i \in \mathbb{R}^{C\times T}$ 表示第 i 个关节点特征，V 表示关节点数量.

然后，骨架特征和输入特征分别经由 1×1 卷积降低通道维度并分为 6 个分支，进行多尺度时间特征学习. 多尺度时间卷积(multiscale temporal convolution, MTC)由 6 个分支组成：1×1卷积、3×1最大池化以及 4 个膨胀系数不同的3×1卷积. 1×1卷积可以通过调整权重来实现特征的重新组合和增强；3×1最大池化用于捕捉时间序列中的显著帧或运动模式；4 个3×1卷积的感受野逐渐增大，可以捕获从短期到长期的不同时间尺度的信息. 以上分支共同作用，使模型能够学习到多尺度的时序依赖关系，提高了模型对时间信息的处理能力. MTC 首先将输入特征沿通道维度均分为 6 组，然后分别由不同分支处理，最后将所有分支的输出特征沿通道维度进行拼接，并采用1×1卷积进行特征融合得到 MTC 的输出特征. MTC 计算过程为：

$$\boldsymbol{X}_{\text{MTC}} = Conv2D_{(1\times1)}(Concat(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_6)), \tag{18}$$

其中 $\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_6$ 为 6 个分支各自的输出特征，$Concat$ 表示沿通道维度进行拼接操作，$Conv2D_{(1\times1)}$ 表示 1×1 卷积操作.

考虑到多尺度时间卷积仅对每个骨架关节点各自的时间特征进行建模，缺少对骨架关节点之间的长时相关性进行学习. 因此经过多尺度时间特征建模后，关节特征通过一组可学习的动态权重向量与骨架特征进行融合得到输出特征：

$$\boldsymbol{X}_{\text{out}} = MTC(\boldsymbol{X}_{\text{in}}) + MTC(\boldsymbol{X}_{\text{S}})\boldsymbol{W}_j, \tag{19}$$

其中 $\boldsymbol{X}_{\text{out}}$ 表示 MTCF 模块的输出特征. $\boldsymbol{W}_j$ 表示可学习的权重向量，用于动态融合骨架和关节特征.

通过多尺度时间建模和骨架特征融合，MTCF 模块能够有效学习不同子动作特征之间的关联，同时加强不同关节点之间的信息交互.

综上所述，本文提出的融合运动领域知识与自适应时间片段注意力的 Transformer 方法，包括线性嵌入层、ATSA 和 MTCF，最后经全局平均池化层、线性层和 Softmax 层输出预测类别. 该方法可有效提升骨架序列短时运动特征学习和关节点之间信息交互的建模能力，用于人体骨架行为识别.

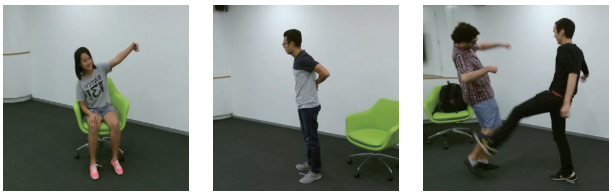
2 性能评估与实验分析

2.1 数据集

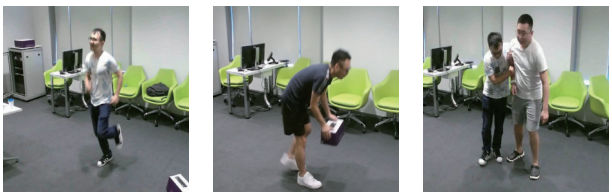
为验证本文方法的有效性和在不同环境的泛化

性, 在 4 个数据集上进行评估与实验, 包括人体日常行为数据集 NTU RGB+D[17]、数据集 NTU RGB+D 120[18]、体育行为数据集 FineGym[19]、工业场景行为数据集 InHARD[20].

1)NTU RGB+D 是室内场景拍摄下的人体日常生活行为数据集, 如图 5 所示包含日常行为、健康相关以及双人交互 3 种场景的行为类别. 数据集包含 60 类行为, 56 880 个样本, 每个样本提供 RGB 视频、深度图序列、3D 骨架序列和红外视频 4 种数据模态. NTU RGB+D 提供 Xsub(Cross-Subject)和 Xview(Cross-View)两种标准来划分训练集和测试集. NTU RGB+D 120 数据集为 NTU RGB+D 的扩展版, 共包含 120 个类别, 114 480 个样本. 与 NTU RGB+D 数据集相比, 此数据集增加了跑步、胸痛以及搀扶等 60 类运动、医疗健康等不同场景的人类行为模式. NTU RGB+D 120 提供 Xsub 和 Xset(Cross-Setup)2 种标准来划分训练集和测试集.



(a) NTU RGB+D



(b) NTU RGB+D 120

Fig. 5 Example of NTU RGB+D and NTU RGB+D 120 dataset samples

图 5 NTU RGB+D 和 NTU RGB+D 120 数据集样本示例

2)FineGym 是一个基于细粒度动作理解的数据集, 主要针对体育运动中体操领域行为识别分析. 如图 6 所示, 数据集包含跳马、自由体操、平衡木以及高低杠 4 种体操项目中的 99 类行为, 32 697 个样本. FineGym 中的样本在背景信息具有高度相似性, 可避免模型利用行为无关信息进行识别.

3)InHARD 为工业生产环境下的人体行为识别数据集, 主要用于涉及人机协作工业环境中的人类行为识别与分析任务. 如图 7 所示, 数据集包含 14 个工业行为类别, 5 303 个行为样本. 每个样本提供 RGB 视频和骨架模态, 以及 3 种不同的视角. InHARD 根据执行主体和视角将样本划分为训练集和测试集.



Fig. 6 Example of FineGym dataset samples

图 6 FineGym 数据集样本示例

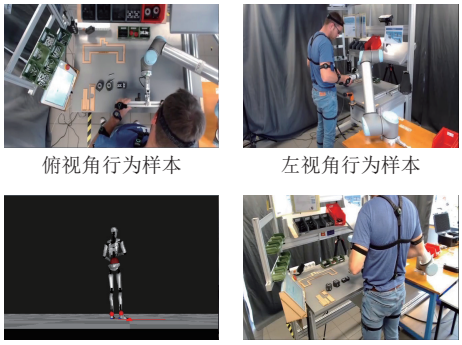


Fig. 7 Example of InHARD dataset samples

图 7 InHARD 数据集样本示例

2.2 实验设置

实验环境配置为 Intel Xeon Platinum 8160Ts CPU, GPU 采用 4 张显存为 24 GB 的 NVIDIA GeForce RTX 3090Ti, 深度学习框架采用 Pytorch1.5. 所有消融和对比实验中, 模型均采用随机梯度下降优化器进行训练, 训练轮数为 120, 初始学习率设置为 0, 批次大小设置为 32. 实验中的模型参数和计算量采用 MMAction2[25] 工具箱计算得到.

2.3 消融实验与分析

为验证提出方法中各个模块的有效性、编码器模块层数以及多模态数据融合等对模型性能提升的影响, 首先需要对提出的模型进行消融实验[26-27]. 鉴于 NTU RGB+D 数据集涵盖多种场景下的人体行为种类, 并且样本数量较为充足, 能够较为全面地反映典型场景中的人体行为特征[7,28]. 因此, 为便于性能评估和消融实验对比, 与文献 [7] 一致, 本文在 NTU RGB+D 数据集的 Xsub 和 Xview 设置下进行消融实验与分析.

2.3.1 ATSA 模块的消融实验

ATSA 通过融合短时运动领域知识, 能根据不同的行为样本针对性学习和捕捉子动作特征. ATSA 模

块由多个不同时间分段长度的 ISSA 模块组成，首先对时间分段长度的影响进行实验验证. 表 1 所示为 ISSA 模块的消融实验，其中 ISSA(n=1)表示时间片段长度为 1 的段间自注意力机制，即空间自注意力机制.

Table 1 Ablation Experiment of ISSA Module

表 1 ISSA 模块的消融实验 %

方法	Xsub	Xview
ISSA（n=1）	89.0	93.8
ISSA（n=2）	89.8	94.2
ISSA（n=3）	90.3	95.0
ISSA（n=4）	88.5	94.9
ISSA（n=5）	87.6	94.9
ISSA（n=6）	87.1	94.7

由表 1 可以发现，单独采用 ISSA(n=1)模块的情况下，模型在 NTU RGB+D 数据集中 2 种设置下的识别准确率分别仅为 89.0% 和 93.8%. 随时间分段长度 n 逐渐增加，模型的识别准确率逐渐提升，ISSA(n=3)模块相较于 ISSA(n=1)模块分别提升 1.3 个百分点和 1.2 个百分点. 此外可以发现，当时间分段长度继续增加后(n>3)，模型识别准确率逐渐下降，说明对长时间分段进行显式建模会影响模型的短时特征建模能力. 以上实验结果表明基于短时运动领域知识设计的 ISSA，通过对相邻骨架帧进行显式建模，能有效对人体行为中的子动作特征进行学习，进而提升模型识别准确率.

表 2 中给出不同信息融合策略的识别性能对比，其中对平均融合中所有分支的输出特征进行相加后取平均得到融合特征，线性融合中所有分支的输出特征经过线性层得到融合特征. 由表 2 可以看出，在 NTU RGB+D 数据集中 2 种设置下，ATSA 采用平均融合策略时的 Top1 识别准确率分别为 90.8% 和 95.6%. 当采用 DIF 模块，模型的识别准确率相对于线性融合策略提升 0.4 个百分点和 0.5 个百分点，验证了 DIF 模块的有效性. 这充分说明对于不同的人体行为，每个分支输出特征的重要性不同，因此动态融合表征可以更好地学习不同尺度的短时运动特征.

Table 2 Recognition Performance Comparison of Different Information Fusion Strategies

表 2 不同信息融合策略的识别性能对比 %

方法	融合策略	Xsub	Xview
ATSA（1，2，3 分支）	平均融合	90.8	95.6
	线性融合	90.6	95.3
	DIF	91.0	95.8

消融实验结果表明，ATSA 模块相较于 ISSA 对短时运动领域知识的融合更加充分，同时说明针对不同行为类别进行自适应时间分段建模的重要性.

2.3.2 MTCF 模块的消融实验

受长时运动领域知识启发，MTCF 模块采用多个不同感受野的 2D 卷积进行长时特征建模，并引入特征融合机制增强不同关节点间的信息交互. 表 3 为 MTCF 模块的消融实验，其中多尺度时间卷积(MTC)的具体介绍见 2.4 节，4 分支表示 1×1卷积、3×1最大池化以及 2 个膨胀系数从 1~2 的 3×1卷积；6 分支表示 1×1卷积、3×1最大池化以及 4 个 3×1卷积.

Table 3 Ablation Experiment of MTCF Module

表 3 MTCF 模块的消融实验 %

方法	Xsub	Xview
MTC（4 分支）	90.2	94.2
MTC（6 分支）	90.5	94.9
MTCF（4 分支+固定权重）	90.7	95.4
MTCF（4 分支+动态权重）	90.8	95.4
MTCF（6 分支+动态权重）	91.0	95.8

由表 3 可以发现，采用 MTC(4 分支)模块的情况下，模型在 NTU RGB+D 数据集中 2 种设置下的识别准确率分别仅为 90.2% 和 94.2%. 将 MTC 模块的分支增加到 6 后，模型的识别准确率分别提升 0.3 个百分点和 0.7 个百分点，表明更丰富的多尺度时间特征学习有利于提升模型识别准确率.

相较于每个关节点单独进行长时特征建模的 MTC 模块，MTCF 模块通过采用权重向量对骨架特征和节点特征进行融合，以增强模型的长时信息建模能力. 由表 3 可以发现，采用固定权重融合全局特征的识别准确率分别为 90.7% 和 95.4%，采用动态权重融合全局特征的识别准确率分别为 90.8% 和 95.4%，相对于 MTC(4 分支)模块分别提升 0.6 个百分点和 1.2 个百分点. 这说明 MTCF 模块，不管是固定权重还是动态权重策略融合方式，均能有效提升模型识别准确率，同时验证了增强关节点长时信息交互的重要性.

更进一步，将 MTCF 模块同样扩展为 6 分支后，识别准确率进一步提升至 91.0% 和 95.8%. 这表明长时运动领域知识指导的 MTCF 模块可通过融合骨架全局特征来有效增强不同关节点之间的信息交互能力，从而提升模型识别准确率.

2.3.3 多模态数据的消融实验

多流模态融合策略被广泛应用于现有骨架行为识别方法. 不同数据模态有各自的特点，通过融合多模态数据可显著提高识别性能. 与主流方法[7]一致，本文分别使用关节点、骨骼、关节点运动以及骨骼运动模态进行训练，然后按 2∶2∶1∶1 的比例对模型的推理输出进行加权求和得到最终结果. 骨骼模态数据基于相邻关节点的差值计算得到，关节点运动和骨骼运动模态数据基于相邻帧差值得到. 表 4 为本文方法采用 4 种单模态以及 4 种模态融合的情况下在 NTU RGB+D 数据集的消融实验结果.

Table 4 Ablation Experiment of Multi-Modal Data

表 4 多模态数据的消融实验 %

方法	Xsub	Xview
关节点模态	91.0	95.8
骨骼模态	91.2	95.2
关节点运动模态	88.8	94.4
骨骼运动模态	88.3	92.9
4 种模态融合	93.2	97.1

由表 4 可以看出，任意单模态数据的识别准确率均低于多模态数据融合的结果. 在单一关节点模态下，本文方法在 Xsub 和 Xview 设置下的识别准确率分别为 91.0% 和 95.8%. 采用 4 种模态数据融合的情况下，本文方法在 Xsub 和 Xview 设置下的识别准确率分别为 93.2% 和 97.1%，相比于单一关节点模态分别提升 2.2 个百分点和 1.3 个百分点. 此外可以发现，尽管关节点运动和骨骼运动 2 种模态各自的识别效果不佳，但通过多模态数据融合后仍可带来一定识别准确率提升，验证了多模态信息融合的有效性.

2.4 不同行为的识别准确率分析

为直观观察本文方法对不同行为的识别情况，如图 8 中给出 NTU RGB+D 120 数据集上的归一化混淆矩阵. 混淆矩阵的横坐标和纵坐标分别表示预测类别和真实类别，颜色越深表示识别准确率越高.

如图 8 所示，本文方法对 NTU RGB+D 120 中的前 60 类行为的识别准确率较高，第 70~80 类的部分行为识别准确率低于 60%. 其中第 73 类“订书”的识别准确率最低仅为 41%，易混淆为第 76 类“剪纸”，主要原因在于骨架序列数据缺乏关键的物体交互信息，导致难以准确识别人物交互行为. 第 71 类“做 OK 手势”和第 72 类“做胜利手势”准确识别率分别为 57% 和 56%，互为易混淆类别. 主要原因为这 2 类行为的主要区别仅取决于手指姿态的细微变化，而 NTU RGB+D 120 数据集提供的手部关节点较少，导致模型无法获取足够的鉴别信息用于分类决策. 此外，对于需要人体多个关节协同运动的第 80 类“深蹲”的识别准确率达到 98%，充分说明融合长时运动领域知识后，可提升模型对不同关节点之间长时依赖关系的捕捉能力.

2.5 模型输出特征可视化分析

为进一步评估基于自适应时间片段注意力的 Transformer 模型的输出特征对行为识别性能的影响，以及人体运动领域知识的重要性，图 9 对本文方法在 NTU RGB+D 数据集中最后一层的不同 ISSA 分支输出特征进行可视化分析. 图 9 中横轴表示人体骨架关节点个数，纵轴表示时间索引号，颜色越深表示输出特征的激活值越大.

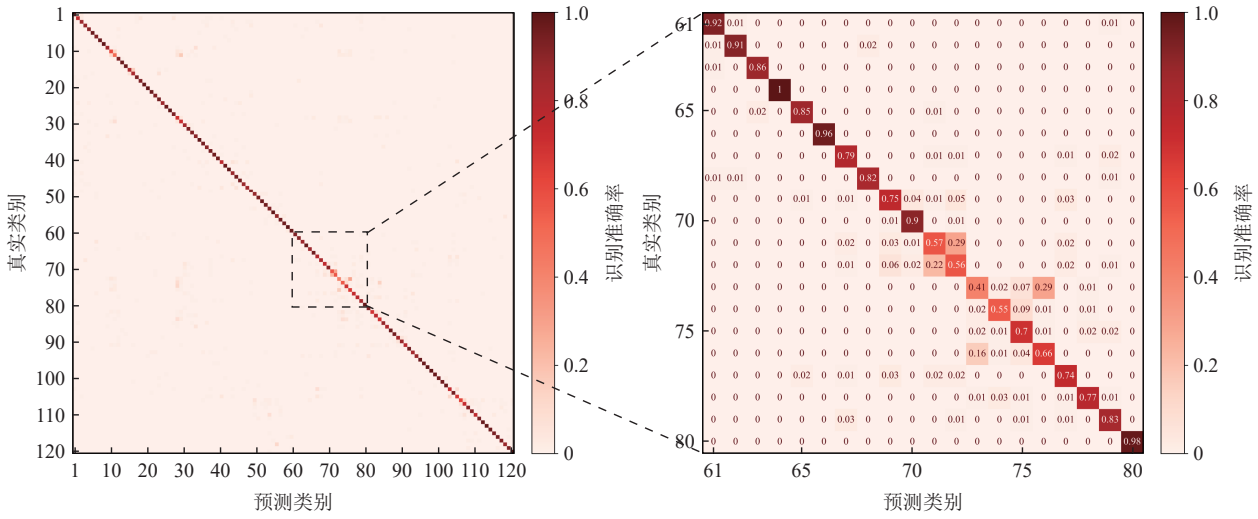


Fig. 8 The normalized confusion matrix on NTU RGB + D 120 dataset

图 8 NTU RGB+D 120 数据集上的归一化混淆矩阵

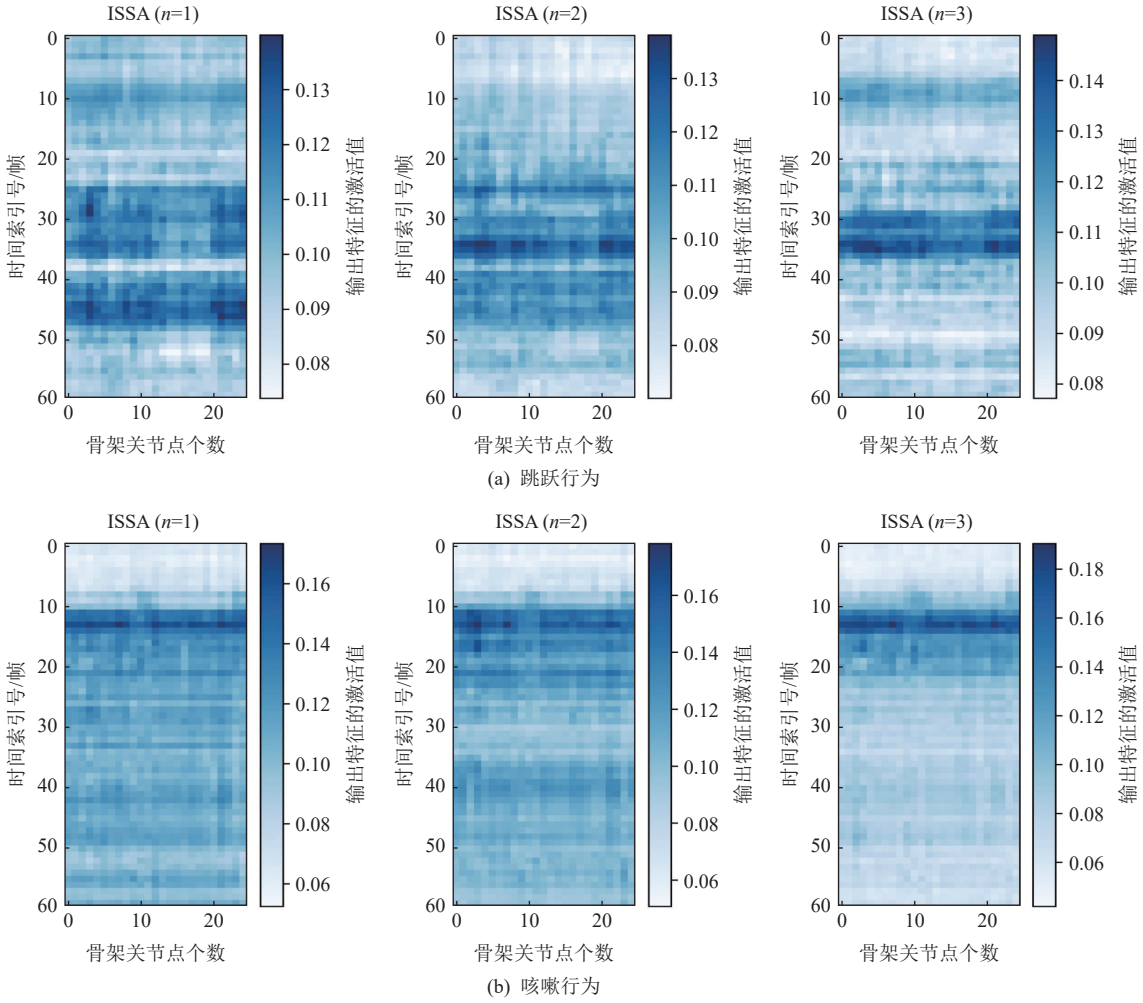


Fig. 9 Visualization of output features of different ISSA branches on NTU RGB+D dataset

图 9 NTU RGB+D 数据集中不同 ISSA 分支输出特征的可视化

首先对比 2 种不同的行为可以发现，跳跃行为的输出特征激活时间区域主要集中在中期，而咳嗽行为的输出特征激活时间区域主要集中在早期，这 2 种行为各自的时间动态规律一致. 该现象表明通过融合人体运动领域知识，本文方法能够聚焦与行为识别高度相关的骨架帧，减少无关信息的干扰，从而提升模型的鲁棒性. 此外可以发现，不同 ISSA 分支关注的时间区域不同，能够学习不同尺度的短时运动特征信息. 对于跳跃行为而言，ISSA(n=1)的输出特征时间区域主要分布在第 25~50 帧，对应行为过程中的整体运动阶段. ISSA(n=3)的输出特征激活时间区域则主要集中在第 30~35 帧，对应过程中的关键阶段. 这充分说明基于短时运动领域知识设计的 ATSA，能够通过多分支结构显式学习不同行为的人体子动作特征，从而有效捕捉人体行为的关键运动信息.

2.6 与其他先进方法对比

为验证本文方法的有效性和泛化性，在 NTU RGB+D, NTU RGB+D 120, FineGym, InHARD 数据集上与其他先进代表方法进行对比. 本文在 4 个数据集上均采用多模态融合策略[7]获得最终的识别结果. 实验结果分别如表 5、表 6、表 7 所示.

由表 5 可以看出，在 NTU RGB+D 数据集的 2 种设置下，本文方法的 Top1 识别准确率分别达到 93.2% 和 97.1%. 在 NTU RGB+D 120 数据集上的 2 种设置下，相较于现有的 Transformer 方法 TranSkeleton[34]，本文方法分别将 Top1 识别准确率分别提升 0.3 个百分点和 0.9 个百分点. 对比基于 GCN 的 FR-Head[31], HD-GCN[32]，基于 Transformer 的 STSA-Net[15], TranSkeleton[34] 和基于 MLP 的 SiT-MLP[35] 人体行为识别方法，本文方法的 Top1 识别准确率在 NTU RGB+D 和 NTU RGB+D 120 这 2 个数据集上均具有可比性，充分说明融合人体运动领域知识的重要性. 此外，对比最新基于 GCN 的 DEGCN[33] 方法，本文方法在识别准确率相近的情况下，模型参数降低 55%，表明了本文方法的有效性.

Table 5 Recognition Accuracy Comparison of Different Methods on NTU RGB+D and NTU RGB+D 120 Datasets

表 5 不同方法在 NTU RGB+D 和 NTU RGB+D 120 数据集的识别准确率对比

类别	方法	参数量	计算量/GFLOPs	NTU RGB+D		NTU RGB+D 120	
				Xsub/%	Xview/%	Xsub/%	Xset/%
GCN	ST-GCN[4]	3.1×10^6		81.5	88.3	70.7	73.2
	2s-AGCN[5]	3.5×10^6	3.9	88.5	95.1	79.2	81.5
	Shift-GCN[6]	0.7×10^6	2.5	90.7	96.5	85.9	87.6
	CTR-GCN[7]	1.4×10^6	1.8	92.4	96.8	88.9	90.6
	AS-GCN[26]			86.8	94.2	78.3	79.8
	MS-G3D[27]	2.8×10^6	5.2	91.5	96.2	86.9	88.4
	InfoGCN[28]	1.5×10^6	1.7	92.7	96.9	89.4	90.6
	HC-GCN[29]			91.2	96.7	87.6	89.0
	MSS-GCN[30]			92.7	96.9	88.9	90.6
	FR-Head[31]	2.0×10^6		92.8	96.8	89.5	90.9
	HD-GCN[32]	1.7×10^6	1.6	93.0	97.0	89.8	91.2
	DEGCN[33]	5.6×10^6		93.6	97.4	91.0	92.1
Transformer	ST-TR[9]	12.1×10^6	259.4	89.9	96.1	82.7	84.7
	DSTA[10]	4.1×10^6	64.7	91.5	96.4	86.6	89.0
	STSA-Net[15]	1.9×10^6		92.7	96.7	88.5	90.7
	TranSkeleton[34]	2.2×10^6	9.2	92.8	97.0	89.4	90.5
MLP	SiT-MLP[35]	0.6×10^6	0.7	92.3	96.8	89.0	90.5
Transformer	本文方法	2.5×10^6	5.2	93.2	97.1	89.7	91.4

注：模型参数量、计算量参考自文献 [15, 33, 35].

Table 6 Recognition Accuracy Comparison of Different Methods on FineGym Dataset

表 6 不同方法在 FineGym 数据集的识别准确率对比 %

类别	方法	Top1	平均 Top1
GCN	ST-GCN[4]	36.4	25.2
	MDR-GCN[36]	94.5	
CNN	ActionVLAD[37]	69.5	50.1
	I3D[38]	74.8	63.2
	TSN[39]	86.0	76.4
	TRN[40]	87.4	79.8
	TRNms[40]	87.8	80.2
	TSM[41]	88.4	81.2
Transformer	RSANet[42]		86.4
	RGBSformer[43]		86.7
	Deformable[44]		90.3
Transformer	本文方法	95.0	93.1

Table 7 Recognition Accuracy Comparison of Different Methods on InHARD Dataset

表 7 不同方法在 InHARD 数据集的识别准确率对比 %

类别	方法	数据模态	Top1
CNN	TSN[39]	RGB	84.4
	TSM[41]	RGB	88.1
	I3D[38]	RGB	82.1
	Non Local[45]	RGB	86.3
Transformer	Timesformer[21]	RGB	85.2
	VideoSwin[46]	RGB	87.6
Transformer	本文方法	骨架	88.3>

相比于 GCN, Transformer 结构的引入往往会进一步增加算法复杂度. 对比 GCN 与 Transformer 两类架构模型可以发现, 基于 Transformer 架构的方法和基于 GCN 的方法在模型参数量区别不明显, 但在计算量上整体明显高于基于 GCN 的方法. 本文方法采用时间分段的方式进行自注意力机制计算, 相较于全局注意力机制能够显著减少模型参数和计算量, 同时可对短时子动作特征进行显示建模. 此外, 本文方法的模型参数和模型计算量相较于其他 Transformer 类方法均有一定下降, 在识别性能和计算开销之间达到较好的平衡.

由表 6 可以看出, 本文方法在 FineGym 数据集的 Top1 和平均 Top1 识别准确率分别为 95.0% 和 93.1%.

对比最新的 MDR-GCN[36] 方法，本文方法的 Top1 识别准确率提升 0.5 个百分点. 对比最新的 3D Deformable Transformer[44] 的方法，本文方法的平均 Top1 识别准确率提升 2.8 个百分点，充分说明了本文方法的有效性. 由于 FineGym 中的行为样本在外观和背景信息上较为类似，并且大多数行为类别具有相同的子动作，因此本文方法通过融合短时运动领域知识对人体子动作特征进行显式建模，能够有效学习不同行为与子动作之间的关联，从而提升模型识别准确率. 为进一步评估本文方法在不同场景的表现和泛化能力，在工业生产人体行为识别数据集 InHARD 上与其他先进代表方法进行对比. InHARD 为工业生产人体行为识别数据集，样本数量较少，可评估模型在不同领域的泛化能力. 由表 7 可以看出，本文方法在 InHARD 数据集中 Top1 识别准确率为 88.3%，优于 TSM[41] 和 VideoSwin[46]. 此外，本文方法采用骨架模态进行人体行为识别，在取得更高识别性能的情况下保持较低的计算量和模型参数. 结果表明，本文方法能够很好地泛化至工业生产领域.

3 结束语

现有人体骨架行为识别方法对人体运动学领域知识利用不足，并缺少对短时子动作特征的显式建模，导致局部特征没有得到充分利用，造成内在人类可理解的逻辑决策解释性不足. 为此本文提出融合运动领域知识和自适应时间片段注意力的 Transformer 模型用于人体骨架行为识别. 首先，受短时运动领域知识启发，提出 ATSA 学习短时子动作特征. ATSA 通过多分支结构对不同时间尺度的局部运动特征进行建模，并通过动态特征融合整合多个分支的短时特征信息. 然后，融合长时运动领域知识，提出 MTCF 用于学习子动作之间的关联和增强关节点的长时信息交互.

4 个公开数据集的性能评估和实验结果表明，本文方法可有效提升对骨架序列局部动作特征的建模能力，增加不同关节点之间的信息交互能力，从而提高人体骨架行为识别的识别准确率.

作者贡献声明：梁成武提出了方法思路和实验方案；蒋松琪、杨杰执行主要实验；朱培旺、帖云、胡伟负责论文修订、补充和实验分析；梁成武、杨杰撰写论文；高磊、郭文博提出指导意见并修改论文.

参 考 文 献

[1] Pei Lishen, Liu Shaobo, Zhao Xuezhuan. Review of human behavior recognition research[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(2): 305−322 (in Chinese)
(裴利沈, 刘少博, 赵雪专. 人体行为识别研究综述[J]. 计算机科学与探索, 2022, 16(2): 305−322)

[2] Li Hongjun, Ding Yupeng, Li Chaobo, et al. Action recognition of temporal segment network based on feature fusion[J]. Journal of Computer Research and Development, 2020, 57(1): 145−158 (in Chinese)
(李洪均, 丁宇鹏, 李超波, 等. 基于特征融合时序分割网络的行为识别研究[J]. 计算机研究与发展, 2020, 57(1): 145−158)

[3] Lu Jian, Li Xuanfeng, Zhao Bo, et al. A review of skeleton-based human action recognition[J]. Journal of Image and Graphics, 2023, 28(12): 3651−3669 (in Chinese)
(卢健, 李萱峰, 赵博, 等. 骨骼信息的人体行为识别综述[J]. 中国图象图形学报, 2023, 28(12): 3651−3669)

[4] Yan Sijie, Xiong Yuanjun, Lin Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]// Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 7444−7452

[5] Shi Lei, Zhang Yifan, Cheng Jian, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]// Proc of the 28th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 12026−12035

[6] Cheng Ke, Zhang Yifan, He Xiangyu, et al. Skeleton-based action recognition with shift graph convolutional network[C]//Proc of the 29th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 183−192

[7] Chen Yuxin, Zhang Ziqi, Yuan Chunfeng, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]//Proc of the 18th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 13359−13368

[8] 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 (in Chinese)
(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)

[9] Plizzari C, Cannici M, Matteucci M. Skeleton-based action recognition via spatial and temporal transformer networks[J]. Computer Vision and Image Understanding, 2021, 208/209: 103219−103228

[10] Shi Lei, Zhang Yifan, Cheng Jian, et al. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition[C]// Proc of the 15th Asian Conf on Computer Vision. Berlin: Springer, 2020: 38−53

[11] Shi Yuexiang, Zhu Maoqing. Collaborative convolutional Transformer network based on skeleton action recognition[J]. Journal of Electronics and Information Technology, 2019, 45(4): 1485−1493 (in Chinese)
(石跃祥, 朱茂清. 基于骨架动作识别的协作卷积 Transformer 网

络[J]. 电子与信息学报, 2019, 45(4): 1485−1493)

[12] Zhang Shiwen, Guo Sheng, Wang Liming, et al. Knowledge integration networks for action recognition[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 12862−12869

[13] Wang Liming, Tong Zhan, Ji Bin, et al. TDN: Temporal difference networks for efficient action recognition[C]//Proc of the 30th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 1895−1904

[14] Liang Chengwu, Hu Wei, Yang Jie, et al. Integration of spatiotemporal domain knowledge and data-driven skeleton behavior recognition[J]. Computer Engineering and Applications, 2025, 61(5): 165−176 (in Chinese)
(梁成武, 胡伟, 杨杰, 等. 融合时空领域知识与数据驱动的骨架行为识别[J]. 计算机工程与应用, 2025, 61(5): 165−176)

[15] Qiu Helei, Hou Biao, Ren Bo, et al. Spatio-temporal segments attention for skeleton-based action recognition[J]. Neurocomputing, 2023, 518: 30−38

[16] Duan Haodong, Wang Jiaqi, Chen Kai, et al. DG-STGCN: Dynamic spatial-temporal modeling for skeleton-based action recognition[J]. arXiv preprint, arXiv: 2210.05895, 2022

[17] Shahroudy A, Liu Jun, Ng T T, et al. NTU RGB+D: A large scale dataset for 3D human activity analysis[C]//Proc of the 25th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 1010−1019

[18] 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, 2019, 42(10): 2684−2701

[19] Shao Dian, Zhao Yue, Dai Bo, et al. FineGym: A hierarchical video dataset for fine-grained action understanding[C]//Proc of the 29th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 2616−2625

[20] Dallel M, Havard V, Baudry D, et al. Inhard-industrial human action recognition dataset in the context of industrial collaborative robotics[C]//Proc of the 1st IEEE Int Conf on Human-Machine Systems (ICHMS). Piscataway, NJ: IEEE, 2020: 1−6

[21] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J] arXiv preprint, arXiv: 2010.11929, 2021

[22] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proc of the 31st Int Conf on Neural Information Processing Systems. Long Beach, CA: Curran Associates Inc, 2017: 6000−6010

[23] Bertasius G, Wang Heng, Torresani L. Is space-time attention all you need for video understanding?[C/OL]//Proc of the 38th Int Conf on Machine Learning. 2021[2025-04-10]. https://proceedings.mlr.press/v139/bertasius21a.html

[24] Li Xiang, Wang Wenhai, Hu Xiaolin, et al. Selective kernel networks[C]//Proc of the 28th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 510−519

[25] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark[CP/OL]. 2020[2025-04-10]. https://github.com/open-mmlab/mmaction2

[26] Li Maosen, Chen Siheng, Chen Xu, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C]//Proc of the 28th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 3595−3603

[27] Liu Ziyu, Zhang Hongwen, Chen Zhenghao, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]//Proc of the 29th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 143−152

[28] Chi H, Ha M H, Chi S, et al. InfoGCN: Representation learning for human skeleton-based action recognition[C]//Proc of the 31st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 20186−20196

[29] Cao Yi, Xia Yu, Gao Qingyuan, et al. Based on super connection diagram convolution network skeleton behavior recognition method[J/OL]. Journal of Jilin University: Engineering Science, 2025,[2025-04-12]. https://doi.org/10.13229/j.cnki.jdxbgxb.20230440 (in Chinese)
(曹毅, 夏宇, 高清源, 等. 基于超连接图卷积网络的骨架行为识别方法[J/OL]. 吉林大学学报: 工学版, 2025, [2025-04-12]. https://doi.org/10.13229/j.cnki.jdxbgxb.20230440)

[30] Jang S, Lee H, Kim W J, et al. Multi-scale structural graph convolutional network for skeleton-based action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(8): 7244−7258

[31] Zhou Huanyu, Liu Qingjie, Wang Yunhong. Learning discriminative representations for skeleton based action recognition[C]//Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 10608−10617

[32] Lee J, Lee M, Lee D, et al. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 10444−10453

[33] Myung W, Su Nan, Xue Jinhao, et al. DEGCN: Deformable graph convolutional networks for skeleton-based action recognition[J]. IEEE Transactions on Image Processing, 2024, 33: 2477−2490

[34] Liu Haowei, Liu Yongcheng, Chen Yuxin, et al. TranSkeleton: Hierarchical spatia-temporal Transformer for skeleton-based action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 4137−4148

[35] Zhang Shaojie, Yin Jianqi, Dang Yonghao, et al. SiT-MLP: A simple MLP with point-wise topology feature learning for skeleton-based action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(9): 8122−8134

[36] Liu Shenlan, Ding Yuning, Zhang Jingro, et al. Multidimensional refinement graph convolutional network with robust decouple loss for fine-grained skeleton-based action recognition[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(4): 7615−7626

[37] Girdhar R, Ramanan D, Gupta A, et al. ActionVLAD: Learning spatio-temporal aggregation for action classification[C]//Proc of the 26th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 971−980

[38] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset[C]//Proc of the 26th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 6299−6308

[39] Wang Liming, Xiong Yuanjun, Wang Zhe, et al. Temporal segment networks: Towards good practices for deep action recognition[C]// Proc of the 11th European Conf on Computer Vision. Berlin: Springer, 2016: 20−36

[40] Zhou B, Andonian A, Oliva A, et al. Temporal relational reasoning in videos[C]//Proc of the 13th European Conf on Computer Vision. Berlin: Springer, 2018: 803−818

[41] Lin Ji, Gan Chuang, Han Song. TSM: Temporal shift module for efficient video understanding[C]//Proc of the 17th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 7083−7093

[42] Kim M, Kwon H, Wang Chunyu, et al. Relational self-attention: What' s missing in attention for video under-standing[C]//Proc of the 35th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2021: 8046−8059

[43] Shi Jing, Zhang Yuanyuan, Wang Weihang, et al. A novel two-stream Transformer-based framework for multi-modality human action recognition[J]. Applied Sciences, 2023, 13(4): 2058

[44] Kim S, Ahn D, Ko B C. Cross-modal learning with 3D deformable attention for action recognition[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 10265−10275

[45] Wang Xiaolong, Girshick R, Gupta A, et al. Non-local neural networks[C]//Proc of the 27th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7794−7803

[46] Liu Ze, Ning Jia, Cao Yue, et al. Video swin Transformer[C]//Proc of the 31st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 3202−3211



Liang Chengwu, born in 1982. PhD, professor, master supervisor. Member of CCF. His main research interests include computer vision, machine learning, and multimedia processing.

梁成武, 1982 年生. 博士, 教授, 硕士生导师, CCF 会员. 主要研究方向为计算机视觉、机器学习、多媒体处理.



Jiang Songqi, born in 1999. Master candidate. His main research interests include target detection and human action recognition.

蒋松琪, 1999 年生. 硕士研究生. 主要研究方向为目标检测、人体动作识别.



Yang Jie, born in 1998. Master. His main research interests include computer vision and multi-modality information fusion.

杨　杰, 1998 年生. 硕士. 主要研究方向为计算机视觉、多模态信息融合.



Zhu Peiwang, born in 2000. Master candidate. His main research interest includes computer vision.

朱培旺, 2000 年生. 硕士研究生. 主要研究方向为计算机视觉.



Tie Yun, born in 1974. PhD, professor, PhD supervisor. His main research interests include intelligent multimedia, and virtual reality and interaction.

帖　云, 1974 年生. 博士, 教授, 博士生导师. 主要研究方向为智能多媒体、虚拟现实及交互.



Gao Lei, born in 1983. PhD, associate professor. His main research interests include computer vision and multimedia processing.

高　磊, 1983 年生. 博士, 副教授. 主要研究方向为计算机视觉、多媒体处理.



Hu Wei, born in 1999. Master. His main research interest includes smart grid.

胡　伟, 1999 年生. 硕士. 主要研究方向为智能电网.



Guo Wenbo, born in 1994. PhD, lecturer. His main research interests include computer vision and image processing.

郭文博, 1994 年生. 博士, 讲师. 主要研究方向为计算机视觉、图像处理.