融合动作描述生成与跨模态语义对齐的骨架动作识别方法

李雨桐[1],马 苗[1,2*],陈建芮[1,2]
(1. 陕西师范大学人工智能与计算机学院,陕西西安 710119;2. 现代教学技术教育部重点实验室,陕西西安 710062)

摘 要: 动作识别旨在通过对人体动作的建模与分析,实现对人类行为的自动识别与理解,广泛应用于智能监控、人机交互、智慧教育等领域. 近年来,自监督骨架动作识别方法因其计算成本低、适应能力强和标注数据依赖性小,逐渐成为动作识别的重要研究方向之一. 然而现有方法多依赖模板提示生成动作概念的解释语句,存在时空结构信息缺失及语义建模能力有限问题,为此本文提出一种跨模态先验辅助的自监督骨架动作识别方法,旨在充分融合骨架结构特征与语义先验知识,实现更具语义理解能力的动作表征. 该方法一方面利用双分支解耦骨架编码器分别建模动作的空间结构与时间信息,结合跨域对比学习策略,从空间、时间及全局视角建立特征对齐与一致性约束,以获得具有丰富时空结构和全局信息的骨架模态特征;另一方面将时序拼接的动作图像和提示指令共同输入视觉语言模型(Vision-Language Model,VLM)生成动作描述,并利用对比语言-图像预训练(Contrastive Language-Image Pre-training,CLIP)模型的文本编码器提取包含动作语义的文本特征,从而弥补单一骨架模态在细粒度语义表示上的不足;在此基础上,通过骨架调制文本的跨模态对比学习策略,在骨架特征引导下利用特征线性调制(Feature-wise Linear Modulation,FiLM)机制动态调控文本语义,实现骨架、文本信息的跨模态语义对齐. 实验结果表明,在 NTU-RGB+D 60 和 NTU-RGB+D 120 数据集上所提方法的识别准确率优于 C^2VL 等 10 余种先进方法,在 PKU-MMD-II 数据集上识别准确率优于 ACA^{2}Net 等 8 种先进方法. 本文方法融合骨架结构信息与语义先验,实现了骨架特征与语言语义的有效互补,为低标注成本的骨架动作识别研究提供了新思路. 未来工作将进一步探索基于领域自适应的微调策略,以提升视觉语言模型的开放集描述能力,并构建在线协同优化框架,实现动作描述生成与识别任务的联合优化,从而增强该方法在实时人机交互与智慧教育等复杂动态场景中的实用性、智能化与可解释性.

关键词: 骨架动作识别;动作描述生成;跨模态语义对齐;视觉语言模型;对比学习;自监督学习
基金项目: 国家自然科学基金(No.62377031)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2025)11-4116-16
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250652

Leveraging Action Description Generation and Cross-Modal Semantic Alignment for Skeleton-Based Action Recognition

LI Yu-tong[1], MA Miao[1,2*], CHEN Jian-rui[1,2]
(1. *School of Artificial Intelligence and Computer Science*, *Shaanxi Normal University*, *Xi'an*, *Shaanxi* 710119, *China*;
2. *Key Laboratory of Modern Teaching Technology* (*Ministry of Education*), *Xi'an*, *Shaanxi* 710062, *China*)

Abstract: Action recognition aims to model and analyze human motions to automatically identify and understand human behaviors, and it has been widely applied in various fields such as intelligent surveillance, human-computer interaction, and smart education. In recent years, self-supervised skeleton-based action recognition has emerged as an important research area due to its low computational cost, strong adaptability, and minimal reliance on labeled samples. However, existing methods often rely on template-based prompts to generate action concept descriptions, which suffer from the lack of spatio-temporal information and limited semantic modeling capability. To address these issues, this paper proposes a cross-modal prior-assisted self-supervised skeleton-based action recognition method, aiming to effectively integrate skeletal structural features with semantic priors to achieve more semantically rich action representations. On one hand, it employs a dual-

收稿日期:2025-07-25;录用日期:2025-11-10;责任编辑:覃怀银
*通讯作者:马苗

branch decoupled skeleton encoder to separately model the spatial structure and temporal dynamics of actions, and integrates a cross-domain contrastive learning strategy to establish feature alignment and consistency constraints from spatial, temporal, and global perspectives, thereby obtaining skeleton-modal features rich in spatio-temporal structure and global context. On the other hand, it feeds temporally concatenated action images along with prompt instructions into a vision-language model to generate action descriptions, and utilizes the text encoder of the contrastive language-image pre-training (CLIP) model to extract text features, thereby supplementing the limited fine-grained semantic representation capability of the skeleton modality. Furthermore, a cross-modal contrastive learning strategy is proposed, where the textual semantics are dynamically modulated under the guidance of skeleton features using a feature-wise linear modulation (FiLM) mechanism, enabling effective semantic alignment between skeleton and text modalities. Experimental results show that the recognition accuracy of the proposed method outperforms more than ten state-of-the-art approaches, including C^{2}VL, on the NTU-RGB+D 60 and NTU-RGB+D 120 datasets, and surpasses eight competitive methods, such as ACA^{2}Net, on the PKU-MMD-II dataset. The proposed method integrates skeletal structural information with semantic priors, achieving effective complementarity between skeleton features and language semantics, and providing a new perspective for skeleton-based action recognition with low annotation cost. In future work, we will further explore domain-adaptive fine-tuning strategies to enhance the open-set description capability of vision-language models, and develop an online collaborative optimization framework to jointly optimize description generation and action recognition, thereby improving the practicality, intelligence, and interpretability of the proposed method in complex dynamic scenarios such as real-time human-computer interaction and smart education.

Key words: skeleton-based action recognition; action description generation; cross-modal semantic alignment; vision-language model; contrastive learning; self-supervised learning

Foundation Item(s): National Natural Science Foundation of China (No.62377031)

1 引言

动作识别作为计算机视觉领域的核心任务之一,在智能监控、人机交互、智慧教育等多个应用场景中发挥着重要作用. 随着Kinect等深度传感器及姿态估计技术的发展,3D骨架数据因其轻量化、结构化,以及对光照、背景与视角变化的鲁棒性,逐渐成为动作识别领域的重要分支[1]. 然而,现有骨架动作识别方法多属有监督学习框架,样本依赖性强、泛化能力有限,因此探索基于自监督学习的骨架动作识别方法成为该领域的重要研究方向.

近年来,自监督骨架动作识别方法主要分为基于重建的方法[2~5]和基于对比学习的方法[6~12]两类. 前者遵循编码器-解码器框架,借助点云着色、掩码建模等技术重构骨架,增强模型特征表示能力;后者构造正负样本对,结合时空增强、多模态协同建模和语义层级建模等策略,引导模型学习具有判别性的骨架特征. 尽管这些方法在骨架表示学习方面进展顺利,但在识别相似或细粒度动作时常存在语义表达能力不足问题,例如"刷牙、喝水"等物体交互动作识别时难以捕捉细粒度语义信息;此外"旋转、裁剪、掩码"等骨架增强操作会破坏原始动作的结构一致性,从而影响识别性能.

为增强骨架模态的语义表达能力,研究者们开始探索跨模态先验辅助的自监督骨架动作识别方法[13,14],通过引入文本或视觉-语言信号作为外部知识指导骨架特征的学习. 例如,Zhu等人[13]提出运动感知的掩码特征重建(Motion-aware Mask Feature Reconstruction, MMFR)方法,利用GPT-3.5 (Generative Pre-trained Transformer 3.5)将动作类别扩展为包含语义信息的自然语言描述以引导骨架特征学习,如"喝水"被扩展为"喝水是指人类通过口腔摄入水分以补充水分或解渴的行为",再利用语义引导掩码重建模块通过概率分布蒸馏将文本特征嵌入骨架;然而因采用模板提示生成动作概念导致动态时空信息关注不足. Chen等人[14]提出的C^{2}VL(Cross-modal Contrastive learning)融合了开放集目标检测模型Grounding DINO(Marrying DINO with Grounded Pre-Training)[15]与视觉语言模型LLaVA(Large Language and Vision Assistant)[16],通过设计引导性问题生成语义描述,并利用模态内自相似性与模态间交叉一致性增强视觉-语言提示特征与骨架特征间的关联度,但未显式地建模提示与骨架间的语义映射.

针对上述问题,本文引入动作描述生成与跨模态语义对齐策略,并提出了一种基于跨模态先验辅助的自监督骨架动作识别方法AC-SAR(leveraging Action description generation and Cross-modal semantic alignment for Skeleton-based Action Recognition). 该方法包括5个主要步骤:(1)在动作描述生成阶段,利用目标检测模型从视频中提取人物区域并拼接成时序动作图像,结合提示指令输入视觉语言模型生成动作描述文本,并通过对比语言-图像预训练(Contrastive Language-Image Pre-training, CLIP)模型[17]的文本编码器提取具有高层语义信息的文本特征;(2)在预训练阶段,基于

骨架编码器[10]提取的骨架特征，引入特征线性调制(Feature-wise Linear Modulation，FiLM)机制修正文本特征，使其能够自适应地聚焦于与骨架动作相关的语义表征空间，并通过跨模态对比学习对齐骨架和文本模态，实现模态间的语义一致性建模；(3)在微调阶段，冻结预训练骨架编码器，仅优化分类器参数；(4)在推理阶段，直接利用骨架编码器进行预测；(5)在多流融合阶段，融合关节、骨骼、运动数据来提升识别精度. 各阶段输入、输出以及目标见表1.

表1 跨模态先验辅助的自监督骨架动作识别流程

阶段	输入	输出	目标
动作描述生成	视频	动作描述文本	获取精细动作语义
预训练	骨架 文本	预训练模型	语义对齐两类特征
微调	骨架 标签	类别概率	提升下游动作识别性能
推理	骨架	类别概率	预测未知动作类别
多流融合	关节流 骨骼流 运动流	多流融合概率	聚合多维度骨架

本文的主要贡献包括以下3个方面.

(1)提出动作描述生成策略，利用视觉语言模型，以时序动作图像为输入并结合提示指令，生成描述动作变化的自然语言文本，为跨模态学习提供动态语义先验和引导.

(2)设计骨架调制的语义对齐策略，引入FiLM机制，通过骨架特征调制文本表示实现文本语义重构，并结合跨模态对比学习实现骨架与文本的对齐，提升骨架动作的精细表征能力.

(3)提出的基于跨模态先验辅助的自监督骨架动作识别方法AC-SAR在NTU-RGB+D 60、NTU-RGB+D 120和PKU-MMD-II数据集上的实验验证了其有效性和优越性.

2 相关工作

2.1 视觉语言模型

视觉语言模型(Vision-Language Model，VLM)通过结合语言、图像或视频模态实现复杂理解和生成任务，是一种同时处理文本和视觉信息的跨模态大模型，在描述生成、视觉问答、图文匹配等任务中表现出色[18,19]. 例如，里程碑工作CLIP[17]构建了视觉和文本编码器，通过对比学习将图像和文本嵌入至统一空间. Li等人[20]提出BLIP(Bootstrapping Language-Image Pre-training)模型，通过自引导机制优化噪声图文数据，在视觉-语言理解和生成任务上表现优异. Liu等人[16]构建了LLaVA模型，具备强大的视觉理解和对话能力，但未涉及视频数据处理. 为此，Lin等人[21]基于多模态指令数据微调LLaVA提出的Video-LLaVA模型，可同时处理图像和视频，通过统一视觉表示增强跨模态理解和推理. Li等人[22]提出多模态视频理解基准数据集MVBench (Multi-modal Video understanding Benchmark)，并研发了VideoChat2模型进一步提升时序理解性能. Fei等人[23]提出Video-CCAM(enhancing Video-language understanding with Causal Cross-Attention Masks)模型，通过因果交叉注意力机制增强视频-语言的理解能力. Lu等人[24]提出基于混合专家(Mixture-of-Experts，MoE)的视觉语言模型DeepSeek-VL2，在文档、表格、科学文献理解任务中性能突出. Lu等人[25]提出的Ovis(Open vision)模型采用结构化嵌入对齐策略提升视觉与文本表示一致性，与同等参数规模模型相比，在多模态基准测试中表现优异.

本文在Ovis模型基础上生成动作描述文本，为自监督骨架动作识别提供丰富的语义线索.

2.2 自监督骨架动作识别

当前自监督骨架动作识别主要分为基于骨架重建[2~5]和对比学习[6~12]两种方式. 基于骨架重建的方法通过编码器-解码器结构，借助点云着色、掩码建模等技术重建骨架，以增强模型特征表示能力. Yang等人[2]提出基于点云着色的骨架表征方法，将未标注骨架序列转化为时空着色点云，并设计自编码结构进行特征提取与学习. Mao等人[3]提出掩码运动预测框架MAMP (Masked Motion Predictors)，通过运动强度先验引导骨架掩码，利用掩码后的骨架序列进行时序运动预测和动作表征学习. Xu等人[4]引入教师-学生架构，采用教师编码器生成全局上下文特征作为预测目标，学生模型则在运动感知的时序掩码下进行蒸馏学习. Cao等人[5]提出结合运动和方向强度的罗盘感知掩码策略，设计自适应对齐模块，引导学生编码器从教师模型中捕捉动作上下文. 然而，这类方法在细粒度动作表征方面仍存在局限.

相比之下，基于对比学习的方法通过构建正负样本对驱动模型学习判别性骨架表示，核心技术策略包括三类：(1)骨架数据增强，目的是通过旋转、裁剪、缩放等扰动提升骨架数据多样性；(2)多模态特征协同，目的是联合关节坐标、骨骼向量和关节运动数据建模；(3)特征建模，目的是基于时空解耦或语义层次建模等方式提取具有判别性的动作表征. 现有研究通常采用上述一种或多种策略增强动作识别性能. 例如，Mao等人[6]提出跨模态互蒸馏框架，结合对比学习与教师-学生蒸馏机制，实现不同模态间的双向知识传递. Lin等人[7]自动提取动作区域，引入自适应数据增强与语义感知特征池化策略，差异化建模动态与静态骨架区域.

Zhang 等人[8]设计渐进式分层数据增强策略与非对称损失函数,在特征空间中进行定向聚类以促进层级间的一致性学习. Sun 等人[9]通过早期融合实现模态联合建模,提出模态内和模态间一致性约束以缓解模态融合带来的偏置问题. Wu 等人[10]提出时空解耦对比学习方法 SCD-Net(Spatiotemporal Clues Disentanglement Network),通过时空结构感知的掩码策略增加训练多样性,引入双路径解耦模块分别建模时间与空间线索,并构建全局锚点以对比学习的方式提升与时间、空间的交互表达. Weng 等人[11]基于特征去相关思想构建统一的骨架密集表征框架,在时间、空间和实例维度进行特征去相关,降低冗余信息干扰,并通过密集时空编码器捕捉精细动作变化. 此外,Zhang 等人[12]提出的 PCM³(Prompted Contrast with Masked Motion Modeling)框架将对比学习和掩码预测协同整合,通过掩码预测提供对比学习新视角,同时利用对比学习捕获的高层语义信息反向指导掩码预测.

本文拟借鉴 SCD-Net[10]的时空解耦思想,结合结构感知掩码与双分支建模策略,实现对骨架动作中时间与空间线索的精细表征与准确捕捉.

3　AC-SAR 方法

跨模态先验辅助的自监督骨架动作识别将语义提示融入自监督骨架动作识别,为构建具有语义理解能力的动作表示提供了新思路. Zhu 等人[13]提出的 MMFR 方法将动作类别文本提示作为语义引导,借助语义蒸馏与多粒度对比策略对齐视觉与语言信息,以提升动作语义理解能力. Chen 等人[14]利用 Grounding DINO[15]和 LLaVA[16]预训练视觉语言模型生成知识提示来引导骨架特征学习,并结合模态内/模态间软目标与自蒸馏机制,从带噪声的骨架-视觉-语言数据中学习鲁棒的动作表征. 然而,这些方法对动作时空信息关注不足,语义建模能力有限,尤其在细粒度动作识别方面性能有待提升. 为此,我们提出一种基于跨模态先验辅助的自监督骨架动作识别方法,即融合动作描述生成与跨模态语义对齐的骨架动作识别方法 AC-SAR.

3.1　基本框架

给定一组人体骨架序列 X 及对应的视频序列 V,利用视觉语言模型为视觉模态 V 生成动作描述 T,形成骨架-文本对(X, T). 在预训练阶段,一方面引导模型从骨架序列中学习具有判别性的空间、时间与全局特征表示;另一方面,以文本为监督信号,通过对比学习对齐骨架与文本模态的嵌入空间,以支持下游的骨架动作识别任务.

图 1 为 AC-SAR 方法的详细框架结构,包括骨架特征建模、基于 VLM 的动作描述生成、骨架调制文本的跨模态对比学习三个部分.

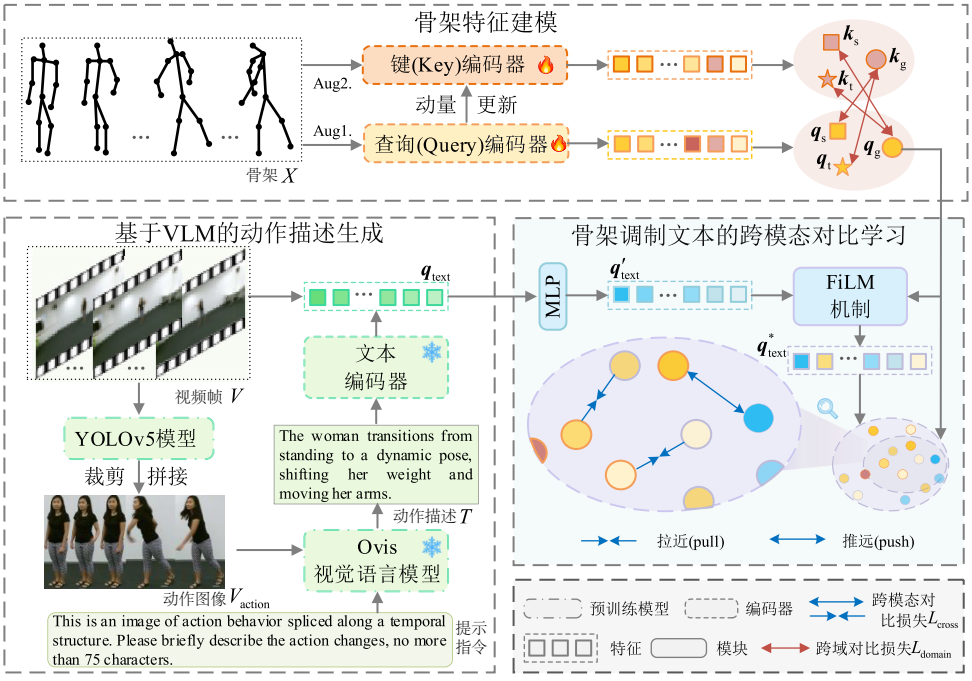


图 1　AC-SAR 方法的详细框架结构

3.2 骨架特征建模

受文献[10]启发,本节利用双分支解耦编码器和跨域对比学习策略挖掘骨架序列中的时空结构和全局信息,完成骨架特征建模,主要步骤如下.

(1)数据增强

对输入$\boldsymbol{X}\in\mathbf{R}^{C\times L_{\mathrm{skl}}\times J}$(其中$C$为通道数,即关节点三维坐标,$L_{\mathrm{skl}}$为序列长度,$J$为关节数量),施加随机遮挡、时序扰动等常规增强操作,同时采用结构化约束的时空掩码策略,生成增强的骨架序列$\boldsymbol{X}_1$和$\boldsymbol{X}_2$. 其中,结构化约束的掩码策略包括空间与时序两个方面:空间掩码基于骨架物理结构,对选定关节及其相邻区域响应最高的k个点进行掩码;时间掩码将输入序列划分为t_{cube}个等长立方体,随机选择n个立方体进行掩码.

(2)双分支解耦编码器

增强后的骨架序列$\boldsymbol{X}_1$和$\boldsymbol{X}_2$分别输入查询编码器$f_{\mathrm{q}}(\cdot)$和键编码器$f_{\mathrm{k}}(\cdot)$,生成解耦的空间、时间表示,其中查询编码器(Query Encoder)用于处理当前输入数据生成表示,键编码器(Key Encoder)用于生成稳定的目标表示. 以查询编码器为例,通过3层的CTR-GCN基础块(CTR-GCN Basic Block)[26]提取初始特征,得到统一表示$\boldsymbol{Y}$,然后对其执行时空解耦操作,生成空间特征$\boldsymbol{Z}_{\mathrm{s}}$和时序特征$\boldsymbol{Z}_{\mathrm{t}}$. 时空解耦操作具体过程如下:首先将表示$\boldsymbol{Y}\in\mathbf{R}^{C_1\times L_{\mathrm{skl}}\times J}$分别变换为$\boldsymbol{Y}_{\mathrm{s}}\in\mathbf{R}^{J\times(C_1L_{\mathrm{skl}})}$和$\boldsymbol{Y}_{\mathrm{t}}\in\mathbf{R}^{L_{\mathrm{skl}}\times(C_1J)}$;然后通过两层线性层映射为结构化的空间嵌入(Embedding)$\boldsymbol{y}_{\mathrm{s}}\in\mathbf{R}^{J\times C_2}$和时间嵌入$\boldsymbol{y}_{\mathrm{t}}\in\mathbf{R}^{L_{\mathrm{skl}}\times C_2}$;最终通过1层Transformer特征细化模块获得空间特征$\boldsymbol{Z}_{\mathrm{s}}\in\mathbf{R}^{C_2}$和时序特征$\boldsymbol{Z}_{\mathrm{t}}\in\mathbf{R}^{C_2}$. 双分支下的编码器结构相同但参数更新方式不同,借鉴MoCo v2(Momentum Contrastive learning v2)[27]的动量更新机制实现参数更新:

$$\theta_{\mathrm{k}}\leftarrow\alpha\theta_{\mathrm{k}}+(1-\alpha)\theta_{\mathrm{q}} \tag{1}$$

其中,$\alpha\in[0,1]$是动量系数,通过该系数使键编码器保留部分先前的参数$\alpha\theta_{\mathrm{k}}$,同时结合查询编码器当前的参数$(1-\alpha)\theta_{\mathrm{q}}$,使键编码器能在保持相对稳定的同时逐渐整合查询编码器学习到的新信息.

(3)跨域对比损失

基于解耦后的空间、时间表示,引入全局视角作为对比中介,定义新的跨域损失. 通过映射函数F_{s}、F_{t}、F_{g}得到查询编码器在空间、时间和全局视角下的骨架表示$\boldsymbol{q}_{\mathrm{s}}$、$\boldsymbol{q}_{\mathrm{t}}$、$\boldsymbol{q}_{\mathrm{g}}$,即

$$\boldsymbol{q}_{\mathrm{s}}=F_{\mathrm{s}}(\boldsymbol{Z}_{\mathrm{s}}),\boldsymbol{q}_{\mathrm{t}}=F_{\mathrm{t}}(\boldsymbol{Z}_{\mathrm{t}}),\boldsymbol{q}_{\mathrm{g}}=F_{\mathrm{g}}\left([\boldsymbol{Z}_{\mathrm{s}},\boldsymbol{Z}_{\mathrm{t}}]\right) \tag{2}$$

其中,$F_{\mathrm{s}}(\cdot)$、$F_{\mathrm{t}}(\cdot)$和$F_{\mathrm{g}}(\cdot)$分别为处理空间、时间和全局表示的映射函数,均由2层线性层构成;$[\cdot,\cdot]$为拼接操作. 通过映射操作,获得输出通道为C_3的骨架表示,即$\boldsymbol{q}_{\mathrm{s}},\boldsymbol{q}_{\mathrm{t}},\boldsymbol{q}_{\mathrm{g}}\in\mathbf{R}^{C_3}$. 与查询编码器的处理过程相同,键编码器输出为$\boldsymbol{k}_{\mathrm{s}}$、$\boldsymbol{k}_{\mathrm{t}}$、$\boldsymbol{k}_{\mathrm{g}}$. 跨域损失函数$L_{\mathrm{domain}}$定义为

$$L_{\mathrm{domain}}\triangleq L(\boldsymbol{q}_{\mathrm{g}},\boldsymbol{k}_{\mathrm{s}})+L(\boldsymbol{q}_{\mathrm{g}},\boldsymbol{k}_{\mathrm{t}})+L(\boldsymbol{q}_{\mathrm{s}},\boldsymbol{k}_{\mathrm{g}})+L(\boldsymbol{q}_{\mathrm{t}},\boldsymbol{k}_{\mathrm{g}}) \tag{3}$$

其中,对于任意样本对$\boldsymbol{u}$和$\boldsymbol{v}$,$L(\boldsymbol{u},\boldsymbol{v})$用于评估$\boldsymbol{u}$和$\boldsymbol{v}$之间相关性,其目标是最小化查询编码器和键编码器中正样本对间的距离,同时最大化与其他特征的距离. 为此,采用基于InfoNCE(Information Noise Contrastive Estimation)[28]的对比损失,即

$$L(\boldsymbol{u},\boldsymbol{v})=-\log\frac{h(\boldsymbol{u},\boldsymbol{v})}{h(\boldsymbol{u},\boldsymbol{v})+\sum_{l=1}^{M}h(\boldsymbol{u},\boldsymbol{m}_l)} \tag{4}$$

其中,$h(\boldsymbol{u},\boldsymbol{v})=\exp(\boldsymbol{u}\cdot\boldsymbol{v}/\tau_{\mathrm{domain}})$是指数相似度度量函数,需采用L2范式归一化$\boldsymbol{u}$和$\boldsymbol{v}$,以构建稳定的骨架对比学习空间;$\tau_{\mathrm{domain}}$为控制相似度分布的温度超参数;$h(\boldsymbol{u},\boldsymbol{v})$为正样本对间的相似度,同一骨架序列的查询特征与键特征被视为正样本对;构建记忆库$\boldsymbol{Q}=\{\boldsymbol{m}_l\}_{l=1}^{M}$,是一个先进先出队列,用于存储来自不同骨架序列的键特征$\boldsymbol{m}$,M为记忆库大小,$\boldsymbol{u}$与$\boldsymbol{Q}$中的其他特征$\boldsymbol{m}_l$为负样本对. 通过最小化对比损失,引导模型学习到期望的特征表示. 为展示骨架特征建模的实现过程,以$L(\boldsymbol{q}_{\mathrm{g}},\boldsymbol{k}_{\mathrm{s}})$为例将上述跨域对比学习流程形式化,如算法1所示.

算法1 跨域对比学习

输入: 骨架序列$\boldsymbol{X}$,队列容量M,温度系数τ_{domain},批次大小N
输出: 损失$L(\boldsymbol{q}_{\mathrm{g}},\boldsymbol{k}_{\mathrm{s}})$,更新后的编码器参数
1. 随机初始化查询编码器f_{q},并将参数复制给键编码器f_{k};
2. 初始化空间分支记忆队列$\boldsymbol{Q}_{\mathrm{s}}$,并设指针$p_{\mathrm{s}}\leftarrow 0$;
3. 循环每个小批量样本$\boldsymbol{X}$:
  3.1 生成两个增强序列$\boldsymbol{X}_1$和$\boldsymbol{X}_2$;
  3.2 将序列$\boldsymbol{X}_1$输入f_{q}和F_{g},提取查询分支全局表示$\boldsymbol{q}_{\mathrm{g}}$;
    将序列$\boldsymbol{X}_2$输入f_{k}和F_{s},提取键分支空间表示$\boldsymbol{k}_{\mathrm{s}}$;
  3.3 对$\boldsymbol{q}_{\mathrm{g}}$和$\boldsymbol{k}_{\mathrm{s}}$进行L2范数归一化;
  3.4 计算正样本相似度:$l_{\mathrm{gs}}^{+}\leftarrow\boldsymbol{q}_{\mathrm{g}}{}^{\mathrm{T}}\boldsymbol{k}_{\mathrm{s}}$;
  3.5 计算负样本相似度:$l_{\mathrm{gs}}^{-}\leftarrow\boldsymbol{q}_{\mathrm{g}}{}^{\mathrm{T}}\boldsymbol{Q}_{\mathrm{s}}$;
  3.6 构造对比损失:$L\leftarrow\mathrm{CrossEntropy}([l_{\mathrm{gs}}^{+},l_{\mathrm{gs}}^{-}]/\tau_{\mathrm{domain}})$;
  3.7 通过反向传播更新查询编码器f_{q};
  3.8 按照式(1)动量更新键编码器f_{k}参数:
  3.9 更新队列
    将当前批次键特征$\boldsymbol{k}_{\mathrm{s}}$入队:$\boldsymbol{Q}_{\mathrm{s}}[:,p_{\mathrm{s}}:p_{\mathrm{s}}+N]\leftarrow\boldsymbol{k}_{\mathrm{s}}^{\mathrm{T}}$;
    移动指针:$p_{\mathrm{s}}\leftarrow(p_{\mathrm{s}}+N)\bmod M$;//遵循先进先出原则
    保证队列始终存储最近M个$\boldsymbol{k}_{\mathrm{s}}$特征;
4. 结束循环

3.3 基于VLM的动作描述生成

下面引入VLM生成动作描述文本,补充运动模式和与物体交互等细节信息,以提供丰富的语义线索,步骤如下.

(1)目标行为图像生成

对于输入的视频序列V,采用统一的处理流程构建

动作图像序列 V_{action}. 具体地,针对每段视频,使用预训练的YOLOv5(You Only Look Once v5)[29]目标检测模型检测并裁剪出目标人物,形成仅包含目标个体的帧序列,并从中均匀采样5帧图像,按时间顺序拼接为一张具有显式时序结构的RGB图像,作为后续文本生成的输入.

(2)动作描述生成

利用Ovis[25]视觉语言模型生成动作图像序列 V_{action} 的描述文本,形成动作描述文本序列 T. Ovis[25]是一种具备结构化嵌入对齐能力的多模态大语言模型,能够结合输入图像和文本指令生成高质量的动作描述. 具体地,将拼接后的RGB图像与合适的提示指令联合输入Ovis模型,采用指令式问答引导模型理解图像行为并生成对应的自然语言描述. 无需依赖手工标注文本,本步骤通过"图像+指令"推理生成高质量描述文本,为跨模态对比学习提供语义补充.

(3)文本编码

为构建统一的文本模态表示,对动作描述文本序列 T 进行统一编码,提取用于骨架调制与跨模态对齐的高层次语义特征 $\boldsymbol{q}_{\text{text}} \in \mathbf{R}^{L_{\text{text}} \times D}$,其中 L_{text} 为序列长度,D 为文本特征维度. 具体地,采用CLIP模型[17]提供的预训练文本编码器(ViT-L/32)对动作描述文本进行嵌入映射,得到固定维度的特征向量. 该文本特征不仅保留了原始动作中的语义信息,还具备良好的跨模态适应性,为后续基于FiLM的骨架调制过程和跨模态对比学习提供关键语义支撑.

3.4　骨架调制文本的跨模态对比学习

下面引入基于FiLM机制[30]的跨模态对比学习策略,通过查询全局骨架特征调制文本特征,在共享语义空间内优化两模态间的一致性.

(1)文本特征处理与模态调制

通过一层线性投影层将 $\boldsymbol{q}_{\text{text}} \in \mathbf{R}^{L_{\text{text}} \times D}$ 每个标记(token)的 D 维特征压缩至一维,得到 $\boldsymbol{q}_{\text{text\_proj}} \in \mathbf{R}^{L_{\text{text}}}$,以聚焦文本描述中的核心词汇信息;并通过线性映射模块将词嵌入转化为骨架调制所需的特征维度 $\boldsymbol{q}'_{\text{text}} \in \mathbf{R}^{C_3}$,定义为:

$$\boldsymbol{q}'_{\text{text}} = \text{MLP}(\boldsymbol{q}_{\text{text\_proj}}) \tag{5}$$

其中,MLP(·)由两层线性层与中间的非线性激活函数ReLU构成.

引入FiLM机制[30],将骨架特征 $\boldsymbol{q}_{\text{g}}$ 作为条件输入调制文本特征,使得文本特征能适应骨架特征的语义空间. FiLM的核心思想是基于条件输入进行仿射变换实现对目标特征的自适应调控. 在一般形式下,FiLM针对输入特征 $\boldsymbol{x}$ 对目标特征 $\boldsymbol{F}$ 进行逐通道级调制,其计算方式为

$$\text{FiLM}(\boldsymbol{F}_c|\gamma_c,\beta_c) = \gamma_c \odot \boldsymbol{F}_c + \beta_c \tag{6}$$

其中,$\boldsymbol{F}_c$ 表示第 c 层特征;⊙表示逐元素乘法;γ_c 和 β_c 分别控制特征的缩放与平移:

$$\gamma_c = f_c(\boldsymbol{x}), \beta_c = h_c(\boldsymbol{x}) \tag{7}$$

其中,$f_c(\cdot)$ 和 $h_c(\cdot)$ 可以为任意函数,如线性层或卷积模块. 该机制自适应地调整特征分布,使模型在不同任务条件下灵活地重构表征空间. 将上述通用形式具体化到骨架模态和文本模态的对齐过程:给定骨架特征 $\boldsymbol{q}_{\text{g}}$,利用映射函数生成调制因子 $\gamma(\boldsymbol{q}_{\text{g}})$ 和 $\beta(\boldsymbol{q}_{\text{g}})$,并作用于文本特征 $\boldsymbol{q}'_{\text{text}}$,定义为

$$\boldsymbol{q}^{*}_{\text{text}} = \gamma(\boldsymbol{q}_{\text{g}}) \odot \boldsymbol{q}'_{\text{text}} + \beta(\boldsymbol{q}_{\text{g}}) \tag{8}$$

其中,⊙表示逐元素乘法,$\gamma(\cdot)$ 和 $\beta(\cdot)$ 分别由输入维度和输出维度均为 C_3 的线性层实现.

(2)特征归一化与相似度计算

为构建稳定的对比学习空间,调制后的文本特征与骨架特征均进行L2范式的归一化,定义为

$$\hat{\boldsymbol{q}}_{\text{text}} = \frac{\boldsymbol{q}^{*}_{\text{text}}}{\left\| \boldsymbol{q}^{*}_{\text{text}} \right\|}, \hat{\boldsymbol{q}}_{\text{g}} = \frac{\boldsymbol{q}_{\text{g}}}{\left\| \boldsymbol{q}_{\text{g}} \right\|} \tag{9}$$

构建模态间的相似度矩阵,定义为

$$\boldsymbol{S} = \hat{\boldsymbol{q}}_{\text{text}} \cdot \hat{\boldsymbol{q}}_{g}^{\text{T}} \tag{10}$$

(3)双向跨模态对比学习设计

采用对称式的InfoNCE[28]损失函数,分别以文本为锚和骨架为锚,计算跨模态匹配能力,定义为

$$\begin{aligned} L_{\text{text-skl}} &= \text{CrossEntropy}\left(\frac{\boldsymbol{S}}{\tau_{\text{cross}}}, \boldsymbol{I}\right) \\ L_{\text{skl-text}} &= \text{CrossEntropy}\left(\frac{\boldsymbol{S}^{\text{T}}}{\tau_{\text{cross}}}, \boldsymbol{I}\right) \end{aligned} \tag{11}$$

其中,τ_{cross} 为跨模态对比学习中的温度超参数;$\boldsymbol{I}$ 是对角标签向量,其对角线元素为骨架文本的正样本配对;CrossEntropy(·)为交叉熵损失函数. 损失 L_{cross} 为两者平均:

$$L_{\text{cross}} = \frac{1}{2}(L_{\text{text-skl}} + L_{\text{skl-text}}) \tag{12}$$

该损失引导模型在语义空间中拉近匹配模态对之间的距离,同时推远不匹配模态对,从而实现跨模态的特征对齐.

3.5　训练目标

为综合考虑骨架模态内部的跨域对比学习与跨模态之间的语义对齐目标,本文加权融合跨域对比损失 L_{domain} 与跨模态对比损失 L_{cross},构建最终优化目标函数. 具体地,定义最终损失 L_{total} 为

$$L_{\text{total}} = L_{\text{domain}} + \lambda L_{\text{cross}} \tag{13}$$

其中,权重系数 λ 控制跨模态对比损失在训练过程中的贡献程度,以实现骨架特征建模与模态间语义对齐的协同优化.

3.6 多流 AC-SAR

引入具有结构信息的骨骼数据和反映时序动态的关节运动数据增强模型对动作的建模能力. 其中,骨骼数据(Bone)由相邻关节点构成的骨骼边表示,通过相邻节点间的三维坐标差向量建模;关节运动数据(Motion)通过计算各关节点在相邻帧之间的三维坐标差,刻画其随时间变化的运动特征.

下面构建包含关节数据流(Joint Stream)、骨骼数据流(Bone Stream)和关节运动数据流(Motion Stream)的三流识别架构3s-AC-SAR,如图2所示. 每个流分别建模其所代表的数据,捕捉人体动作在不同维度的信息. 在推理阶段,采用决策级融合策略(Decision-level Fusion),将三个流的预测结果加权整合,最终获得更为鲁棒的动作识别结果.

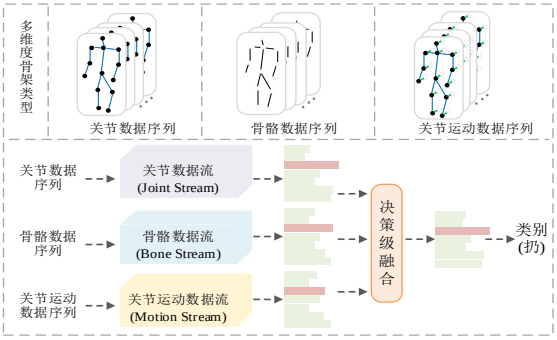


图2 多流架构3s-AC-SAR

4 数据集

4.1 NTU-RGB+D 60数据集

NTU-RGB+D 60(NTU-60)是一个多模态、多视角的动作识别基准数据集[31],包含RGB视频、深度图序列、3D骨架数据和红外视频4种同步采集的模态,采集设备为3部Kinect v2摄像机. 该数据集涉及40名受试者的60类动作,56 880个原始样本经筛选后,共有56 578个有效数据. 每个骨架数据最多包含两个人的动作,每个人体由25个关节点的三维坐标表征. 该数据集的两种评估方式为:(1)Cross-Subject(x-sub):按照受试者ID将数据划分为具有40 091个样本的训练集和16 487个样本的测试集;(2)Cross-View(x-view):使用2号和3号摄像机视角的37 646个样本作为训练集,1号摄像机视角的18 932个样本作为测试集.

4.2 NTU-RGB+D 120数据集

NTU-RGB+D 120(NTU-120)数据集[32]为NTU-60的扩展版本,涉及106名受试者的120类动作,原始样本为114 480个,经筛选后共113 945个可用样本. 该数据集延续了NTU-60的数据规范. 该数据集的两种评估方式为:(1)Cross-Subject(x-sub):按照106名受试者ID平均分配训练集和测试集,分别包含63 026和50 919个样本;(2)Cross-Setup(x-set):依据32种编号奇偶性划分,训练集和测试分别包含54 468和59 477个样本.

4.3 PKU-MMD-II数据集

PKU-MMD[33]是一个涵盖RGB视频、深度图、红外视频和骨架数据4种模态的动作分析数据集,包含PKU-MMD I和II两个版本. PKU-II因视角变化增大更具挑战性,涉及13名受试者的41个动作,在3个摄像机视角下的1 009个短视频序列对应生成6 952个原始样本,筛选后有效样本为6 945个. 遵循跨受试者(x-sub)评估方式,数据集划分为包含5 335个样本的训练集和1 610个样本的测试集.

5 实验

5.1 实验设置

所有实验均在NVIDIA GeForce RTX 5080 GPU的PyTorch框架下实现,实验参数设置如下.

(1)动作描述生成阶段. 属于离线操作,不增加预训练和下游任务训练的开销. 具体地,采用YOLOv5-m作为目标检测器,检测前将输入图像调整并填充至640×640,置信度阈值设为0.25;使用参数量为70亿(7 Billion)的视觉语言模型Ovis[25]作为图像描述生成器.

(2)预训练阶段. 采用带动量的SGD优化器[34],批次大小设为64,初始学习率为0.01,总训练周期为450. 学习率采用线性衰减策略,在第350个周期时衰减为0.001.

(3)下游模型评估阶段. 我们采用多种评估方式,包括线性评估以及以采用1%、5%和10%标注样本进行训练的半监督评估,其中半监督评估的标注样本采样策略遵循随机采样,与文献[6]和文献[10]保持一致. 线性评估中,在NTU-60和NTU-120数据集上,批次大小为1 024,初始学习率为2;在PKU-II数据集上,批次大小为256,初始学习率为0.032,权重衰减设置为0.001. 对于半监督评估,批次大小设置为64,初始学习率为0.01. 训练过程采用动态学习率调整策略和早停机制,若模型连续3个周期准确率无提升,则减小学习率,若连续5个周期准确率无提升,则提前终止训练.

超参数设置. 数据增强前,每个骨架随机采样64帧[8,10],数据增强时,旋转、翻转、错切、空间掩码和时间掩码执行概率均为50%;空间掩码策略中,k=8;时间掩码策略中,t_{cube}=16,n=6. 对于骨架编码器,参考MoCo v2框架[27]构建查询编码器和相应的键编码器进行对比学习,且结构相同. 网络优化时,α=0.999,τ_{domain}=0.2;在NTU-60和NTU-120上,M=8 192,在PKU-II上,M=2 048. 跨模态对比学习时,τ_{cross}=0.15,λ=0.5. 骨架编码器及文本特征映射时,C=3,C_1=64,C_2=2 048,C_3=128,

L_{text}=77,D=512.

5.2　视觉语言模型动作描述生成的对比分析

下面比较三种视觉语言模型 Video-CCAM[23]、DeepSeek-VL[24]和 Ovis[25]在动作描述生成中的表现.三种模型参数量均为70亿,采用统一指令"这是一个沿时间结构拼接的动作图像,请描述其中的动作变化",限定生成文本长度不超过75字符.

在 NTU-60 x-sub 上,测试 Video-CCAM、DeepSeek-VL和Ovis三个模型生成描述文本对方法性能的影响,识别准确率分别为86.6%、86.9%和87.3%,Ovis性能最优.进一步地,分别选取单人动作、双人交互行为以及高相似度的成对动作样本,对比分析它们的语义信息捕捉能力和动作描述准确性,结果如图3所示.

(1)在图3(a)的"刷牙"动作识别中,Video-CCAM出现"拿着麦克风讲话"的幻觉描述,未能识别动作;而DeepSeek-VL和Ovis均识别到"刷牙"行为,前者着重描述手部与牙刷动作,后者则捕捉到动作变化且表达清晰.

(2)在图3(b)的"握手"交互动作识别中,Video-CCAM描述结果与动作无关;DeepSeek-VL较完整地指出"握手"过程;Ovis简洁地描述"握手"行为,语义提取能力较强.

(3)在图3(c)和图3(d)易混淆的"脱夹克"与"穿夹克"动作识别中,Video-CCAM将"脱夹克"误判为"穿夹克";DeepSeek-VL语义模糊,识别失败;Ovis则在两类动作中均给出准确、清晰的简要描述,动作识别能力强.

5.3　与其他先进方法的比较

下面系统评估所提方法在自监督与半监督条件下相较于现有先进方法的性能表现,其中标注"*"表示基于与本文相同的实验配置复现所得结果;标注"§"表示其为跨模态先验辅助的自监督骨架动作识别方法.

5.3.1　自监督骨架动作识别

表2为AC-SAR方法在NTU-60、NTU-120和PKU-II数据集上与多种先进自监督骨架动作识别方法的对比结果,其中加粗数据为最优结果.

在NTU-60和NTU-120数据集上,AC-SAR整体表现最优.在NTU-60数据集上的x-sub和x-view准确率分别为87.3%和91.8%,在NTU-120数据集上的x-sub和x-setup准确率分别为79.3%和80.4%.

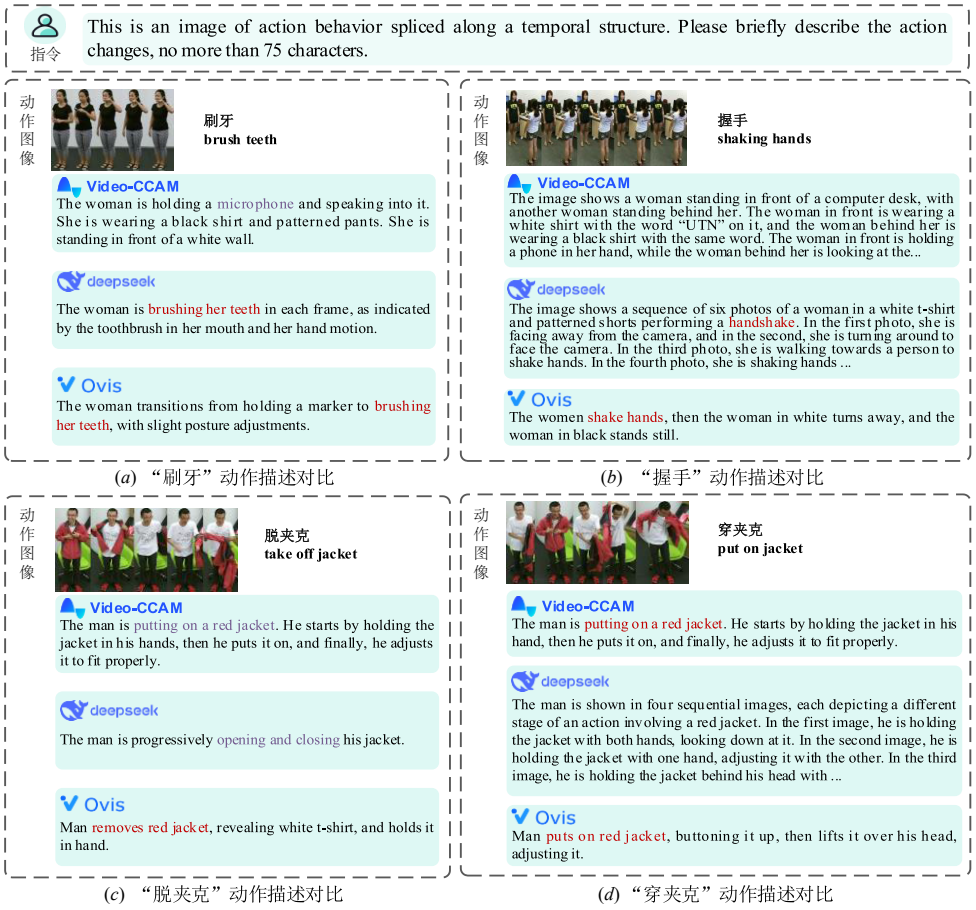


图3　不同视觉语言模型动作描述生成效果对比

表2 AC-SAR方法与其他自监督方法的对比结果 单位:%

方法		编码器	NTU-60		NTU-120		PKU-II
			x-sub	x-view	x-sub	x-setup	x-sub
骨架重建	H-Transformer[35](ICME 2021)	Transformer	69.3	72.8	—	—	—
	Colorization[2](ICCV 2021)	GCN	75.2	83.1	—	—	—
	GL-Transformer[36](ECCV 2022)	Transformer	76.3	83.8	66.0	68.7	—
	MAMP[3](ICCV 2023)	Transformer	84.9	89.1	78.6	79.1	53.8
	MMFR[13]/§(TCSVT 2024)	Transformer	84.9	89.5	77.1	78.8	54.4
	ACA^{2}Net[5](TCSVT 2025)	Transformer	86.0	89.6	79.1	79.8	53.7
对比学习	CrosSCLR[37](CVPR 2021)	GCN	72.9	79.9	—	—	21.2
	AimCLR[38](AAAI 2022)	GCN	74.3	79.7	63.4	63.4	38.5
	CMD[6](ECCV 2022)	GRU	79.8	86.9	70.3	71.5	43.0
	ActCLR[7](CVPR 2023)	GCN	80.9	86.7	69.0	70.5	—
	HiCLR[8](AAAI 2023)	Transformer	78.8	83.1	67.3	69.9	—
	UmURL[9](ACM MM 2023)	Transformer	82.3	89.8	73.5	74.3	52.1
	PCM^3[12](ACM MM 2023)	GRU	83.9	90.4	76.5	77.5	51.5
	Skeleton-logoCLR[39](TCSVT 2024)	GCN	82.4	87.2	72.8	73.5	54.7
	KTCL[40](TMM 2024)	Transformer	82.4	89.4	74.4	74.5	**55.5**
	MID-ECL[41](TIP 2024)	GRU	83.9	90.3	75.7	77.2	—
	SCD-Net[10](AAAI 2024)	GCN&Transformer	86.6	91.7	76.9	80.1	54.0
	SCD-Net*(AAAI 2024)	GCN&Transformer	84.7	90.9	78.8	79.7	49.0
	C^{2}VL[14]/§(TMM 2025)	GCN	84.4	89.8	76.0	78.7	52.6
	USDRL[11](AAAI 2025)	Transformer	85.2	91.7	76.6	78.1	54.4
	AC-SAR(本文方法)	GCN&Transformer	**87.3**	**91.8**	**79.3**	**80.4**	54.0

与仅利用骨架自监督学习的方法对比,以基线SCD-Net[10]为例,在相同实验设置下,AC-SAR在两个数据集的所有评估方式上,识别准确率均有明显提升.所提方法引入动作语义描述并通过跨模态对比学习引导骨架特征学习语义相关表示,从而突破单一骨架自监督的语义局限.

与利用跨模态先验辅助自监督学习的MMFR[13]与C^{2}VL[14]方法相比,AC-SAR优势明显.在NTU-60 x-sub数据集上,所提方法比MMFR[13]和C^{2}VL[14]分别提高2.4个百分点和2.9个百分点;而在NTU-120 x-sub数据集上分别提高2.2个百分点和3.3个百分点.原因在于所提方法使用跨模态先验知识时并非简单融合和对齐外部信息,而是结合骨架结构特性调制文本语义信息.

另外,所提方法在PKU-II数据集上x-sub识别准确率为54.0%,优于SCD-Net、C^{2}VL、UmURL、PCM^{3}等11种方法.

综上所述,相较于近年来19种先进方法,所提方法AC-SAR在多个基准数据集上展现出良好的识别能力和泛化性能.

5.3.2 多流自监督骨架动作识别

表3为所提方法在不同数据输入流下的识别准确率.其中,"J-""B-"和"M-"分别表示以关节数据、骨骼数据和关节运动数据为输入的单流模型,"JB-""BM-"和"JM-"则表示双流融合架构,"3s-"表示三流融合架构.

表3 不同数据输入流对AC-SAR的影响 单位:%

方法	NTU-60		NTU-120		PKU-II
	x-sub	x-view	x-sub	x-setup	x-sub
J-AC-SAR	87.3	91.8	79.3	80.4	54.0
B-AC-SAR	85.6	88.8	77.3	78.5	51.6
M-AC-SAR	83.2	87.7	73.3	74.2	47.0
JB-AC-SAR	88.1	92.3	80.4	81.8	56.9
JM-AC-SAR	88.2	93.0	79.8	81.6	58.3
BM-AC-SAR	87.0	91.2	78.9	79.9	55.7
3s-AC-SAR	88.4	93.0	81.0	82.2	60.0

本文采用简单集成策略[8],在双流和三流架构中,各流的softmax概率均以0.5的权重进行加权相加,并取融合后概率最大的类别作为最终预测结果,用于评估准确率.

表3表明,不同模态数据间的组合相较于单流输入均有性能提升.其中,在NTU-60和PKU-II数据集上,JM融合优于JB融合,而在NTU-120上则表现为JB略优.为进一步探究差异,在NTU-60 x-sub类别层面对比"JM-AC-SAR"与"JB-AC-SAR"两种融合方式的性能,如图4所示.

观察图4可知,JM-AC-SAR在“eat meal、wear a shoe”等幅度小且短时动态明显的类别上表现更优,说明运动数据在时序动态建模方面提供了补充信息;而JB-AC-SAR在“pointing to something with finger、put the palms together”等依赖空间结构的动作上表现更好,表明骨骼数据在捕捉关节空间布局方面具有优势.

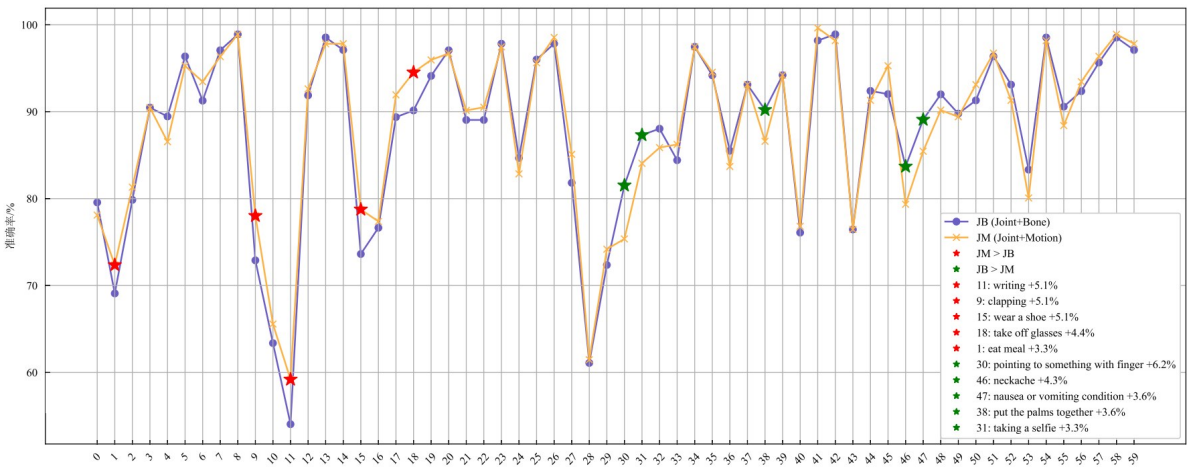


图4　在NTU-60数据集x-sub上JM-AC-SAR与JB-AC-SAR各类别对比结果

为进一步评估所提3s-AC-SAR的性能,表4列出了其与当前先进多流自监督骨架动作识别方法的对比结果,其中加粗数据为最优结果.

根据表4易知,3s-AC-SAR在NTU-60和PKU-II数据集上展现出优越性能.在NTU-60上的x-sub和x-view分别取得了88.4%和93.0%的准确率,整体优于大多数仅依赖骨架自监督学习的方法,其中在x-view上仅比3s-USDRL[11]低0.2个百分点.同时,相比同样引入跨模态先验的3s-C^{2}VL[14](88.3%和92.8%),也实现了小幅超越.在PKU-II上,x-sub达到60.0%,与3s-C^{2}VL性能相当.可见本文提出的利用外部先验知识辅助的跨模态对比学习方式使得AC-SAR方法在多流架构下同样具备良好的特征建模能力.

表4　AC-SAR方法与其他多流自监督方法的对比结果　　单位:%

方法		编码器	NTU-60		NTU-120		PKU-II
			x-sub	x-view	x-sub	x-setup	x-sub
对比学习	3s-CrosSCLR[37](CVPR 2021)	GCN	77.8	83.4	67.9	66.7	21.2
	3s-AimCLR[38](AAAI 2022)	GCN	78.9	83.8	68.2	68.8	39.5
	3s-CMD[6](ECCV 2022)	GRU	84.1	90.9	74.7	76.1	52.6
	3s-ActCLR[7](CVPR 2023)	GCN	84.2	88.8	74.3	75.7	—
	3s-HiCLR[8](AAAI 2023)	Transformer	80.4	85.5	—	—	—
	3s-UmURL[9](ACM MM 2023)	Transformer	84.4	91.4	75.9	77.2	54.3
	3s-PCM^3[12](ACM MM 2023)	GRU	87.4	93.1	80.0	81.2	58.2
	3s-Skeleton-logoCLR[39](TCSVT 2024)	GCN	86.1	89.8	79.8	80.1	57.7
	3s-MID-ECL[41](TIP 2024)	GRU	87.0	92.9	79.4	81.2	—
	3s-SCD-Net[10](AAAI 2024)	GCN&Transformer	87.3	—	—	—	—
	3s-C^{2}VL[14]/§(TMM 2025)	GCN	88.3	92.8	**82.5**	**84.3**	**60.0**
	3s-USDRL[11](AAAI 2025)	Transformer	87.1	**93.2**	79.3	80.6	59.7
	3s-AC-SAR(本文方法)	GCN&Transformer	**88.4**	93.0	81.0	82.2	**60.0**

5.3.3　半监督骨架动作识别

在NTU-60数据集的不同评估方式下,对比AC-SAR与先进的半监督骨架动作识别方法在1%、5%、10%标注比例下监督微调分类器后的性能,结果见表5,其中加粗数据为最优结果.

由表5易知,AC-SAR在多个评估方式中性能优越,尤其在高标注比例(10%)的x-sub上取得82.2%的准确率以及在x-view上获得86.0%的准确率,在所有对比方

表5 AC-SAR方法与其他半监督方法的对比结果 单位:%

方法	x-sub			x-view		
	1%	5%	10%	1%	5%	10%
ISC[42](ACM MM 2021)	35.7	59.6	65.9	38.1	65.7	72.5
Colorization[2](ICCV 2021)	48.3	65.7	71.7	52.5	70.3	78.9
CMD[6](ECCV 2022)	50.6	71.0	75.4	53.0	75.3	80.2
GL-Transformer[35](ECCV 2022)	—	—	68.6	—	—	74.9
HiCLR[8](AAAI 2023)	54.7	—	82.1	53.7	—	84.8
UmURL[9](ACM MM 2023)	58.1	72.5	—	58.3	76.8	—
PCM^3[12](ACM MM 2023)	53.8	—	77.1	53.1	—	82.8
MID-ECL[41](TIP 2024)	55.2	74.9	78.9	54.9	78.6	82.9
SCD-Net[10](AAAI 2024)	**69.1**	—	82.2	**66.8**	—	85.8
SCD-Net*(AAAI 2024)	66.0	78.6	81.4	64.0	**82.1**	84.8
USDRL[11](AAAI 2025)	57.3	—	80.2	60.7	—	84.0
AC-SAR(本文方法)	68.0	**79.3**	**82.2**	64.0	**82.1**	**86.0**

法中表现最优.

5.4 消融实验

在NTU-60数据集的x-sub评估方式下进行消融实验,系统分析所提方法中关键策略的有效性. 同时,进一步探讨了超参数设置、文本生成长度与处理方式、VLM图像拼接帧数以及不同跨模态交互机制对方法性能的影响.

5.4.1 策略有效性

下面测试不同语义建模方式与跨模态对比策略组合对方法性能的影响,见表6. 其中,“未解码文本”指直接使用视觉语言模型生成的高维分布式特征参与对比学习;而“解码文本+CLIP编码文本”则先生成自然语言文本,再通过CLIP文本编码器获取语义嵌入,用于显式的语义对齐.

表6 策略有效性

动作描述生成策略		跨模态对比学习策略		NTU-60 (x-sub)
未解码文本	解码文本+CLIP编码文本	未调制文本	骨架调制文本	
×	×	×	×	84.7%
√	×	√	×	85.9%
×	√	√	×	85.8%
√	×	√	√	87.0%
×	√	√	√	87.3%

由表6可知,未引入语义信息与对比学习的基线模型准确率为84.7%. 当使用未解码的文本提示配合无调制策略进行对比学习时,准确率提升至85.9%;文本经解码后通过CLIP编码的性能为85.8%,表明两类语义建模方式对识别均有提升但效果相近. 引入骨架调制文本策略后,准确率分别提升至87.0%和87.3%,体现骨架信息对文本语义引导的增强作用. 因此,语义描述生成策略与骨架调制下的跨模态对比学习机制在建模动作语义方面具有互补优势.

5.4.2 超参数对方法性能的影响

下面测试跨模态对比学习中温度系数$\tau_{\text{cross}}\in\{0.1,0.15,0.2\}$以及权重平衡系数$\lambda\in\{0.4,0.5,0.6\}$对方法性能的影响,见表7.

表7 超参数对方法性能的影响

超参数		NTU-60(x-sub)
τ_{cross}	λ	
0.10	0.5	86.6%
0.20	0.5	86.8%
0.15	0.5	87.3%
0.15	0.4	86.4%
0.15	0.6	86.5%

由表7可知,当τ_{cross}=0.15且λ=0.5时,所提方法取得了最高识别结果. 此处,τ_{cross}控制特征相似度分布的平滑程度,取值过小会使模型过于关注局部差异,难以捕捉跨模态语义关系,反之可能削弱正负样本的区分度,降低对比效果;λ用于平衡跨域对比损失与跨模态对比损失的贡献,取值偏小会导致对模态语义对齐学习不足,而偏大则可能削弱骨架特征建模能力.

5.4.3 文本长度与处理方式对方法性能的影响

下面测试文本生成长度设置与文本处理方式对方法性能的影响,实验结果如表8所示.

由表8可知,将生成文本最大长度从512缩短至75

表8 文本长度与处理方式对方法性能的影响

文本长度	文本处理方式	NTU-60(x-sub)
512	未解码文本提示	86.7%
75	未解码文本提示	87.0%
75	解码+CLIP句嵌入编码	87.0%
75	解码+CLIP词嵌入编码	87.3%

时,未经解码的表示准确率由86.7%提升至87.0%,说明压缩生成内容有助于语义聚焦并提升跨模态对齐效果. 在此基础上,将生成文本解码为自然语言并采用句级嵌入能保持整体语义信息,但性能未进一步提升. 而采用词级嵌入时,准确率提升至87.3%,表明词级建模能引入局部语义线索,辅助模型在跨模态空间中增强类别间的可分性. 因此,合理控制生成长度并结合词级语义表征,能有效提升动作识别性能.

5.4.4　VLM图像拼接帧数设置对方法性能的影响

下面测试不同拼接帧数输入Ovis视觉语言模型对方法性能的影响,见表9.

表9　VLM图像拼接帧数设置对方法性能的影响

帧数	VLM	NTU-60(x-sub)
3	Ovis	86.7%
5		87.3%
8		86.5%

由表9可见,随着拼接帧数的变化,性能具有细微波动. 当拼接帧数从3帧增加到5帧时,准确率由86.7%提升至87.3%;而在拼接8帧时,性能下降至86.5%,低于拼接3帧时的性能. 这说明,适度增加帧数能提供丰富的时间上下文,有助于VLM理解动作序列,但过多的帧数可能会引入一些细节变化,导致输入中出现冗余信息,从而干扰模型的描述生成能力.

5.4.5　不同跨模态交互机制对方法性能的影响

下面测试不同跨模态交互机制对方法性能的影响,如图5所示. 线性融合、门控融合和跨模态注意三种典型的融合方式的性能分别为86.8%、86.4%和86.5%,整体表现相近,且相较于无交互情况性能均有一定提升. 而使用FiLM调制机制时,准确率为87.3%,相较于其他机制均有明显增益. 结果显示,与前述静态融合方式不同,FiLM调制能基于骨架特征动态调整文本表示,在通道和语义层面实现条件化建模,从而增强跨模态表征能力.

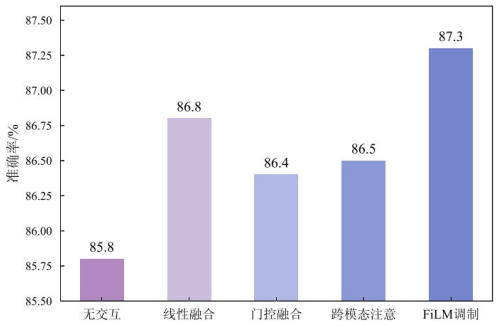


图5　不同跨模态交互机制对方法性能的影响

5.5　可视化分析

5.5.1　t-SNE特征可视化

为定性评估“骨架调制文本”策略中FiLM调制机制对文本特征的影响,在NTU-60 x-sub数据集上提取最后一轮预训练模型生成的特征,并应用t-SNE算法[43]进行可视化分析,对比文本特征调制前后嵌入分布,如图6所示. 随机选取12 000个特征点,在自监督设定下使用k-Means聚类生成伪标签(簇数为60)并着色. 实验结果显示,初始文本特征呈“离散但隐含类别边界”的分布,体现了文本的软语义约束能力,可捕捉动作高层语义关联;骨架特征经双分支解耦建模后,特征簇紧凑、清晰可分,不同动作类别之间分界明显;骨架调制后,文本特征簇更加紧凑且类别间可分性提升. 这表明FiLM调制能够将文本高层语义与骨架物理约束有效融合,提升文本特征的语义表达能力与区分度.

(a) 文本特征 $\boldsymbol{q}'_{\text{text}}$　(b) 骨架特征 $\boldsymbol{q}_{\text{g}}$　(c) 骨架模态调制后文本特征 $\boldsymbol{q}^{*}_{\text{text}}$

图6　骨架调制文本过程的特征t-SNE可视化对比

5.5.2　混淆矩阵

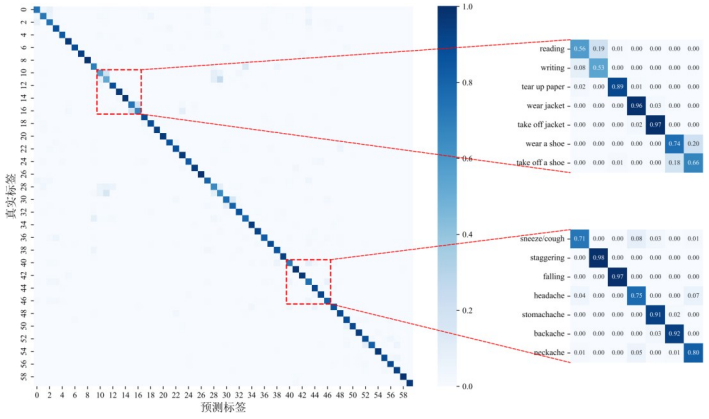
为分析“骨架调制文本”策略在提升相似动作识别性能中的有效性,本文可视化在NTU-60 x-sub数据集上未使用和使用该策略时骨架动作识别结果的混淆矩阵.

使用“骨架调制文本”策略后,所提方法的准确率从85.8%提升至87.3%,表明该策略有效提升了整体识别性能,具体分析如下.

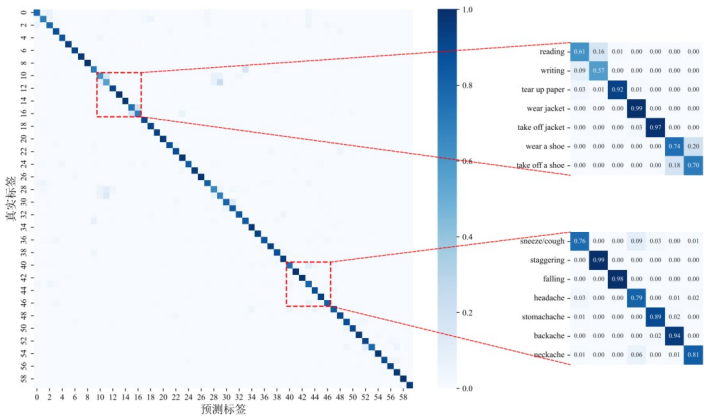
(1)细粒度手部动作的识别情况改善. 如图7(*a*)所示,未使用调制策略时,“reading”与“writing”“headache”与“neckache”等手部细粒度动作存在混淆;引入骨架调制文本策略后,图7(*b*)中这些动作的误判率明显下降,表明该策略进一步增强了骨架与文本语义的对齐效果.

(2)穿脱衣物、鞋子相关类的区分能力增强. 对于运动轨迹相似、以时序区分的动作对照类别(如“wear a shoe”与“take off a shoe”),调制策略提升了分类准确率,表明其能捕获动作的时序演化特征.

图7可视化结果表明,骨架调制文本策略不仅提升了整体准确率,还增强了模型对易混淆类别的区分能力,特别是在细粒度动作、方向性动作及微动作的识别任务上. 该策略有效提升了语义引导下的骨架表征能力,为骨架动作识别任务提供了一种有效且具有泛化性的解决方案.



(*a*) 未使用骨架调制文本策略



(*b*) 使用骨架调制文本策略

图7 所提方法骨架动作识别混淆矩阵可视化结果

6 结论

本文针对现有自监督骨架动作识别方法中语义表达和建模能力有限的问题,提出了一种跨模态先验辅助的自监督骨架动作识别方法AC-SAR. 一方面,利用双分支解耦编码器结构和跨域对比学习策略建模具有丰富时空结构和全局信息的骨架特征;另一方面,引入视觉语言模型生成动作描述文本,捕捉与动作时序演化相一致的语义信息,借助CLIP模型的文本编码器提取文本特征;最后,在骨架特征引导下,通过FiLM机制动态调整文本语义,并通过跨模态对比学习促进骨架与文本模态间的协同建模. 在视觉语言模型生成动作描述的对比实验中,本文所采用的视觉语言模型在捕捉关键动作变化和生成精炼描述方面表现出良好稳定性,为骨架调制文本策略提供了具有语义判别力的文本特征. 在NTU-60和NTU-120数据集上的实验结果表明,所提方法的性能优于同样利用跨模态先验辅助自监督学习的MMFR和C^2VL方法.

目前,本文方法主要针对单人或双人场景下的日常动作识别,对于多人复杂交互情境的适应性仍存在不足. 此外,依赖离线VLM生成的动作描述可能存在

语义模糊、错误或冗余等问题,且与识别流程解耦. 未来的研究将尝试通过领域自适应微调提升VLM对多人动作的开放集描述能力;在此基础上,研究在线协同框架,实现动作描述生成与识别在线联合优化,以增强复杂动态场景的实用性.

参考文献

[1] 罗会兰, 曹立京. 基于多维动态拓扑学习图卷积的骨架动作识别[J]. 电子学报, 2024, 52(3): 991-1001.
LUO H L, CAO L J. Multi-dimensional dynamic topology learning graph convolution for skeleton-based action recognition[J]. Acta Electronica Sinica, 2024, 52(3): 991-1001. (in Chinese)

[2] YANG S Y, LIU J, LU S J, et al. Skeleton cloud colorization for unsupervised 3D action representation learning[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2022: 13403-13413.

[3] MAO Y Y, DENG J J, ZHOU W G, et al. Masked motion predictors are strong 3D action representation learners[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2024: 10147-10157.

[4] XU R Z, HUANG L Z, WANG M, et al. Skeleton2vec: A self-supervised learning framework with contextualized target representations for skeleton sequence[EB/OL]. (2024-01-01)[2025-07-22]. https://arXiv.org/abs/2401.00921.

[5] CAO W M, QIAN L X, ZHANG Y C, et al. Asymmetric context-guided adaptive alignment network for skeleton-based action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(6): 5939-5951.

[6] MAO Y Y, ZHOU W G, LU Z B, et al. CMD: Self-supervised 3D action representation learning with cross-modal mutual distillation[C]//Computer Vision – ECCV 2022. Cham: Springer, 2022: 734-752.

[7] LIN L L, ZHANG J H, LIU J Y. Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 2363-2372.

[8] ZHANG J H, LIN L L, LIU J Y. Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations[C]//Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence. New York: ACM, 2023: 3427-3435.

[9] SUN S K, LIU D Z, DONG J F, et al. Unified multi-modal unsupervised representation learning for skeleton-based action understanding[C]//Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 2973-2984.

[10] WU C, WU X J, KITTLER J, et al. SCD-net: Spatiotemporal clues disentanglement network for self-supervised skeleton-based action recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(6): 5949-5957.

[11] WENG W J, WANG H S, WANG J B, et al. USDRL: Unified skeleton-based dense representation learning with multi-grained feature decorrelation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(8): 8332-8340.

[12] ZHANG J H, LIN L L, LIU J Y. Prompted contrast with masked motion modeling: Towards versatile 3D action representation learning[C]//Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 7175-7183.

[13] ZHU X Y, SHU X B, TANG J H. Motion-aware mask feature reconstruction for skeleton-based action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(11): 10718-10731.

[14] CHEN Y, HE T, FU J F, et al. Vision-language meets the skeleton: Progressively distillation with cross-modal knowledge for 3D action representation learning[J]. IEEE Transactions on Multimedia, 2025, 27: 2293-2303.

[15] LIU S L, ZENG Z Y, REN T H, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection[C]//Computer Vision – ECCV 2024. Cham: Springer, 2025: 38-55.

[16] LIU H, LI C, WU Q, et al. Visual instruction tuning[C]//The 37th Conference on Neural Information Processing Systems. New York: Curran Associates Inc, 2023: 34892-34916.

[17] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//The 38th International Conference on Machine Learning. Cambridge: PMLR, 2021, 139: 8748-8763.

[18] ZHANG J Y, HUANG J X, JIN S, et al. Vision-language models for vision tasks: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(8): 5625-5644.

[19] 才华, 易亚希, 付强, 等. 基于跨模态引导和对齐的多模

态预训练方法[J]. 电子学报, 2024, 52(10): 3368-3381.
CAI H, YI Y X, FU Q, et al. Multimodal pretraining with cross-modal guidance and alignment[J]. Acta Electronica Sinica, 2024, 52(10): 3368-3381. (in Chinese)

[20] LI J N, LI D X, XIONG C M, et al. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//The 39th International Conference on Machine Learning. Cambridge: PMLR, 2022, 162: 12888-12900.

[21] LIN B, YE Y, ZHU B, et al. Video-LLaVA: Learning united visual representation by alignment before projection[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2024: 5971-5984.

[22] LI K C, WANG Y L, HE Y N, et al. MVBench: A comprehensive multi-modal video understanding benchmark[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 22195-22206.

[23] FEI J J, LI D, DENG Z D, et al. Video-CCAM: Enhancing video-language understanding with causal cross-attention masks for short and long videos[EB/OL]. (2024-08-26)[2025-07-22]. https://arXiv.org/abs/2408.14023.

[24] LU H Y, LIU W, ZHANG B, et al. DeepSeek-VL: Towards real-world vision-language understanding[EB/OL]. (2024-03-11)[2025-07-22]. https://arXiv.org/abs/2403.05525.

[25] LU S Y, LI Y, CHEN Q G, et al. Ovis: Structural embedding alignment for multimodal large language model[EB/OL]. (2024-06-17)[2025-07-22]. https://arxiv.org/abs/2405.20797.

[26] CHEN Y X, ZHANG Z Q, YUAN C F, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 13339-13348.

[27] CHEN X L, FAN H Q, GIRSHICK R, et al. Improved baselines with momentum contrastive learning [EB/OL]. (2020-03-09)[2025-07-22]. https://arXiv.org/abs/2003.04297.

[28] VAN DEN OORD A, LI Y Z, VINYALS O. Representation learning with contrastive predictive coding[EB/OL]. (2019-01-22)[2025-07-22]. https://arXiv.org/abs/1807.03748.

[29] Ultralytics. ultralytics/yolov5: v7. 0 - YOLOv5 SOTA realtime instance segmentation[EB/OL]. (2022-11-22)[2025-07-22]. https://github.com/ultralytics/yolov5/releases.

[30] PEREZ E, STRUB F, DE VRIES H, et al. FiLM: Visual reasoning with a general conditioning layer[EB/OL]. (2017-12-18)[2025-07-22]. https://arxiv.org/abs/1709.07871.

[31] SHAHROUDY A, LIU J, NG T T, et al. NTU RGB+D: A large scale dataset for 3D human activity analysis[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 1010-1019.

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

[33] LIU C H, HU Y Y, LI Y H, et al. PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding[EB/OL]. (2017-03-28) [2025-07-22]. https://arXiv.org/abs/1703.07475.

[34] QIAN N. On the momentum term in gradient descent learning algorithms[J]. Neural Networks, 1999, 12(1): 145-151.

[35] CHENG Y B, CHEN X P, CHEN J H, et al. Hierarchical transformer: Unsupervised representation learning for skeleton-based human action recognition[C]//2021 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE, 2021: 9428459.

[36] KIM B, CHANG H J, KIM J, et al. Global-local motion transformer for unsupervised skeleton-based action learning[C]//Computer Vision - ECCV 2022. Cham: Springer, 2022: 209-225.

[37] LI L G, WANG M S, NI B B, et al. 3D human action representation learning via cross-view consistency pursuit[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 4739-4748.

[38] GUO T Y, LIU H, CHEN Z, et al. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(1): 762-770.

[39] HU J H, HOU Y H, GUO Z H, et al. Global and local contrastive learning for self-supervised skeleton-based action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(11): 10578-10589.

[40] WANG X H, MU Y D. Localized linear temporal dynamics for self-supervised skeleton action recognition[J]. IEEE Transactions on Multimedia, 2024, 26: 10189-10199.

[41] LIN L L, ZHANG J H, LIU J Y. Mutual information driven equivariant contrastive learning for 3D action representation learning[J]. IEEE Transactions on Image Processing, 2024, 33: 1883-1897.

[42] THOKER F M, DOUGHTY H, SNOEK C G M. Skele-

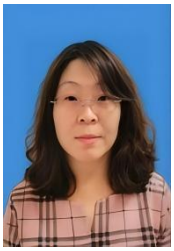
ton-contrastive 3D action representation learning[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 1655-1663.

[43] MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9: 2579-2605.

作者简介

李雨桐　女,1998年10月生,山西长治人.陕西师范大学人工智能与计算机学院博士研究生.主要研究方向为智慧教育、动作识别等.中国电子学会会员编号:E190197993A.
E-mail: liyutongstu@snnu.edu.cn

马　苗　女,1977年4月生,山东聊城人.陕西师范大学人工智能与计算机学院教授.主要研究方向为智慧教育、视频分析等.
E-mail: mmthp@snnu.edu.cn



陈建芮　女,1979年3月生,内蒙古呼和浩特人.陕西师范大学人工智能与计算机学院教授.主要研究方向为多媒体分析、复杂网络等.
E-mail: jianrui_chen@snnu.edu.cn