基于骨架的人体动作识别技术研究进展

刘宝龙[1,2], 周森[3], 董建锋[1)*], 谢满德[3], 周胜利[2,4], 郑天一[1], 张三元[5], 叶修梓[6], 王勋[1]

[1)] (浙江工商大学计算机科学与技术学院 杭州 310018)
[2)] (基于大数据架构的公安信息化应用公安部重点实验室 杭州 310053)
[3)] (浙江工商大学信息与电子工程学院 杭州 310018)
[4)] (浙江警察学院计算机与信息安全系 杭州 310053)
[5)] (浙江大学计算机科学与技术学院 杭州 310013)
[6)] (温州大学大数据与信息技术研究院 温州 325035)
(djf@zjgsu.edu.cn)

摘 要: 近年来, 随着深度学习技术的发展, 已有很多新颖的基于骨架的人体动作识别算法被提出, 极大地推动了该领域的发展. 对基于骨架的人体动作识别领域的主要数据集和算法进行全面、细致的总结. 首先对 NTU, Kinetics-Skeleton 和 SYSU 3DHOI 等骨架相关的数据集进行回顾; 然后将基于骨架的人体动作识别算法归纳为基于监督学习的、基于半监督学习的和基于无监督学习的 3 大类, 并对分属不同类别的算法进行介绍和比较; 最后分析和总结得出该领域当前面临过度依赖大数据、大算力和大模型等挑战, 并针对性地提出缓解以上挑战的 3 点未来发展方向: 高精度骨架数据集建设、细粒度骨架动作识别和数据有效学习的骨架动作识别.

关键词: 动作识别; 骨架特征提取; 深度学习; 图卷积网络
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023.19640

Research Progress in Skeleton-Based Human Action Recognition

Liu Baolong[1,2], Zhou Sen[3], Dong Jianfeng[1)*], Xie Mande[3], Zhou Shengli[2,4], Zheng Tianyi[1], Zhang Sanyuan[5], Ye Xiuzi[6], and Wang Xun[1]

[1)] (*School of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou* 310018)
[2)] (*Key Laboratory of Public Security Informatization Application Based on Big Data Architecture, Ministry of Public Security, Hangzhou* 310053)
[3)] (*School of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou* 310018)
[4)] (*Department of Computer and Information Security, Zhejiang Police College, Hangzhou* 310053)
[5)] (*College of Computer Science and Technology, Zhejiang University, Hangzhou* 310013)
[6)] (*Institute of Big Data and Information Technology, Wenzhou University, Wenzhou* 325035)

Abstract: In recent years, with the development of deep learning technology, many novel skeleton-based human action recognition algorithms have been proposed, which has greatly promoted the development of this field. This paper aims to give a comprehensive and detailed summary of the main datasets and algorithms in the skeleton-based human action recognition field. Firstly, the main skeleton-related datasets such

收稿日期: 2022-03-24; 修回日期: 2022-08-04. **基金项目**: 国家自然科学基金(61976188, 61972352); 基于大数据架构的公安信息化应用公安部重点实验室开放课题(2021DSJSYS001); 浙江工商大学“数字+”学科建设管理项目(SZJ2022C012); 浙江省重点研发计划(2021C03150); 浙江省基础公益技术研究计划(LGF21F020010); 浙江省省属高校基本科研业务费专项资金; 公安部科技计划(2022LL16). **刘宝龙**(1990—), 男, 博士, 讲师, CCF 会员, 主要研究方向为计算机视觉、动作识别; **周森**(1997—), 男, 硕士研究生, CCF 学生会员, 主要研究方向为动作识别; **董建锋**(1991—), 男, 博士, 研究员, 硕士生导师, CCF 会员, 论文通信作者, 主要研究方向为多媒体理解、计算机视觉; **谢满德**(1977—), 男, 博士, 教授, 硕士生导师, CCF 会员, 主要研究方向为边缘计算、网络安全; **周胜利**(1982—), 男, 博士, 高级工程师, 硕士生导师, 主要研究方向为网络空间安全; **郑天一**(1998—), 男, 硕士研究生, 主要研究方向为动作识别; **张三元**(1963—), 男, 博士, 教授, 博士生导师, 主要研究方向为数字媒体、图像处理; **叶修梓**(1966—), 男, 博士, 教授, “长江学者奖励计划”特聘教授, 主要研究方向为计算机图形学; **王勋**(1967—), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究方向为移动图形计算、计算机视觉.

as NTU, Kinetics-Skeleton, and SYSU 3DHOI are reviewed. Secondly, the skeleton-based human action recognition algorithms are summarized into three categories, i.e., supervised learning-based, semi-supervised learning-based, and unsupervised learning-based, the main algorithms of each category are further introduced and compared. Finally, challenges that the field is currently facing, i.e., over-reliance on big data, large computing power, and large models, are concluded, and three future development directions are proposed to alleviate the above challenges: high-precision skeleton dataset construction, fine-grained skeleton-based action recognition, and skeleton-based action recognition with data-efficient learning.

Key words: action recognition; skeleton feature extraction; deep learning; graph convolutional network

近年来, 随着深度学习技术的蓬勃发展, 智能化已经渗透到人们日常生活的方方面面, 为人们的生活带来了诸多便利. 计算机视觉是人工智能领域一个关键的研究方向, 其肩负着让机器“看懂”世界的使命. 作为计算机视觉领域的研究热点之一, 基于视觉技术的人体动作识别近年来受到广泛的关注, 其应用场景十分广阔, 在人机交互、医疗康复、视频监控、智能体育等方面都发挥着重要的作用.

基于视觉技术的动作识别任务指通过设计计算机算法来识别出视频中人体动作的类别信息[1]. 从算法输入的数据模态视角来看, 当前主流的动作识别算法可以分为基于 RGB 视频的动作识别算法[2-6]、基于深度图序列的动作识别算法[7-9]、基于骨架序列的动作识别算法[10-11]和基于多模态数据的动作识别算法[12-13]. 其中, 基于 RGB 相机获取的视频数据在进行视频特征提取时可能存在背景噪声大、数据复杂度高、运动模糊和光照变化等问题, 因此基于 RGB 视频的动作识别算法面临着视频特征表征鲁棒性较差的挑战. 基于深度相机硬件获取的深度图去除了 RGB 图像的颜色和纹理信息, 能较好地缓解 RGB 视频数据背景复杂和光照变化等问题, 但数据复杂度仍然较大. 由于数据复杂度较高, 基于 RGB 视频和深度图序列的动作识别模型往往需要较大的参数量, 不仅训练时间长, 而且对硬件要求也很高, 因此基于 RGB 视频和深度图连续帧的动作识别算法速率往往较低, 可扩展性不够好. 随着 Microsoft Kinect[14-15]等深度图像传感器的普及, 人体姿态估计算法[16-18]的发展越来越成熟, 对人体骨架数据的获取也变得越来越方便. 与传统 RGB 视频或者深度图序列相比, 利用骨架数据进行人体动作识别具有诸多优势: (1) 骨架数据具有抽象性高、复杂性低和鲁棒性好的特点, 不易受背景、尺度、视角、光照等因素的影响[19]; (2) 骨架数据更加贴合人体动作的实际物理意义, 可以更好地表征人体运动的过程. 上述优势使得基于骨架的人体动作识别算法具有运行速率快、鲁棒性好及可扩展性强的特点, 吸引了越来越多研究人员的关注, 近年有大量围绕该方向的文献陆续发表.

1 骨架动作识别概述

1.1 问题定义与算法分类

人体可以被由关节连接的刚体段构成的铰接系统表示, 这种铰接系统即骨架[20]. 骨架描述了人体关节点在空间的坐标位置和相对关系, 是人体姿态和动作的高抽象性表现形式(如图 1 所示). 目前, 骨架通过深度图像传感器、惯性传感器或人体姿态估计算法获取. 由于使用的人体姿态估计算法不同, 骨架的关节点个数可能存在一定的差异. 早期的人体姿态估计工作[21]用 20 个关节点进行人体骨架建模; 当前主流的人体姿态估计工作[17]则增加了手部以及颈部 5 个关节点, 用 25 个关节

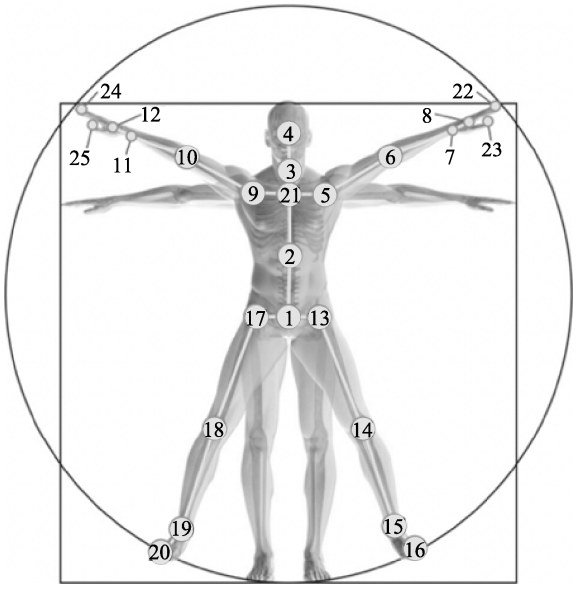


图 1 人体骨架关节点示意图[14]

点对人体骨架进行更加精细化的建模.

基于骨架的动作识别算法框架如图 2 所示. 首先通过人体姿态估计算法提取视频片段中的骨架表征，然后将获取到的骨架表征输入到深度神经网络中进行特征学习和提取，最后基于提取到的动作特征得到动作的类别信息. 从算法的训练方式角度看，基于骨架的动作识别算法可以归类为基于监督学习的骨架动作识别算法、基于半监督学习的骨架动作识别算法和基于无监督学习的骨架动作识别算法 3 大类.

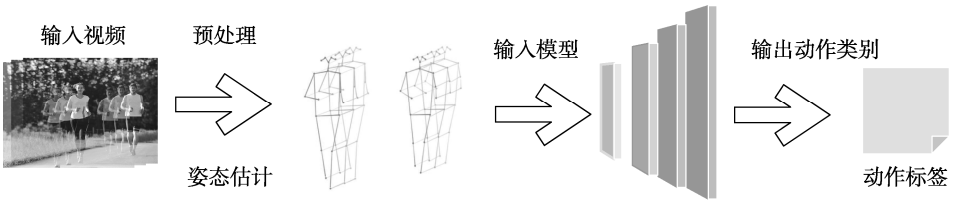


图 2　基于骨架的动作识别算法框架

在基于监督学习的骨架动作识别算法方面，早期的算法通过人工设计的特征提取器进行动作特征提取，如通过几何变换探索关节点的空间关系[22]，利用傅里叶变换[23]、马尔可夫链[24-25]进行骨架特征建模. 这种人工特定的特征提取器通常针对特定数据集而设计，其可以在目标数据集上取得不错的效果，但此类算法也往往存在泛化能力差的问题，一旦数据集更换或者实景部署测试，识别准确率则会急剧下降. 近年来，随着深度学习技术的迅速发展，循环神经网络(recurrent neural network, RNN)[26-27]被应用于骨架动作识别任务中，其对时间序列数据有着天然的建模优势，因此很自然地被用于进行基于骨架的动作识别研究；同时，随着卷积神经网络(convolutional neural network, CNN)[28]在图像分类任务中大放异彩，研究人员也考虑将 CNN 应用于骨架动作识别任务. 基于 CNN 的骨架动作识别算法先将骨架数据转换为 RGB 伪图像，再将 RGB 伪图像输入 CNN 中进行特征提取并计算类别信息. 虽然 RNN 对时序特征的提取优势明显，但人体动作的发生不仅是时序上的变化，还有空间上肢体位置的改变，因此也需要算法对骨架空间特征及时地理解，而 RNN 却不善于对空间关系进行建模. CNN 可以较好地对动作空间关系进行建模，在一定程度上缓解基于 RNN 的骨架动作识别算法空间特征提取受限的难题，但 CNN 时序建模能力不足的问题依然存在. 为了解决上述问题，图卷积网络(graph convolutional network, GCN)[10]应运而生. 基于 GCN 的骨架动作识别算法同时对骨架序列进行时间维度和空间维度的统一建模，可以提取到更加高级和稳定的动作语义特征，因此该类算法近年来得到了蓬勃的发展，在相关数据集上的准确率也取得了很好的效果，但其过于依赖复杂模型的设置，导致模型过大、效率低下. 基于 Transformer 的骨架动作识别算法完全由注意力机制构成，模型轻量效率较高，可以很好地缓解基于 GCN 的骨架动作识别算法中的问题. 虽然基于监督的骨架动作识别算法可以取得较高的识别准确率，但往往也需要大量有标签数据作为基础，而标签数据的获取十分消耗人力和物力[29-31]. 因此，研究人员以半监督的方式进行骨架动作识别模型的训练.

半监督学习范式的本质是利用有标签数据和无标签数据在嵌入空间上的特征分布具有一致性和连续性[32]，缓解模型对有标签数据的需求. 基于半监督学习的骨架动作识别算法在编码器-解码器网络(如图 3 所示)中添加分类器，执行分类任务，并利用有标签数据和无标签数据在嵌入空间中的邻域一致性，使得编码器学习到更具代表性的骨架特征. 目前，基于半监督学习的骨架动作识别算法的学习方式有对抗学习[33]、主动学习[34]、一次学习[35]等；然而这些算法仍然需要有标签数据的辅助. 因此，研究人员以无监督学习的方式进行骨架动作识别模型的训练，进一步降低对标签数据的依赖.

无监督学习范式的本质是从大量无标签数据中学习到特征表征. 目前，基于无监督学习的骨架动作识别算法通常以编码器-解码器网络[29-30]为基础，先将骨架序列输入到编码器中提取特征，再将提取到的特征输入到解码器中重新生成骨架序列，通过计算输入的骨架序列与生成的骨架序列之间的位置损失，优化编码器学习到的骨架特征. 大多数算法也会设置一些特定的辅助任务，如帧序排列[36]、对比学习[36-37]、视角一致[38]等，来达到学习到更高级的动作语义特征的目的.

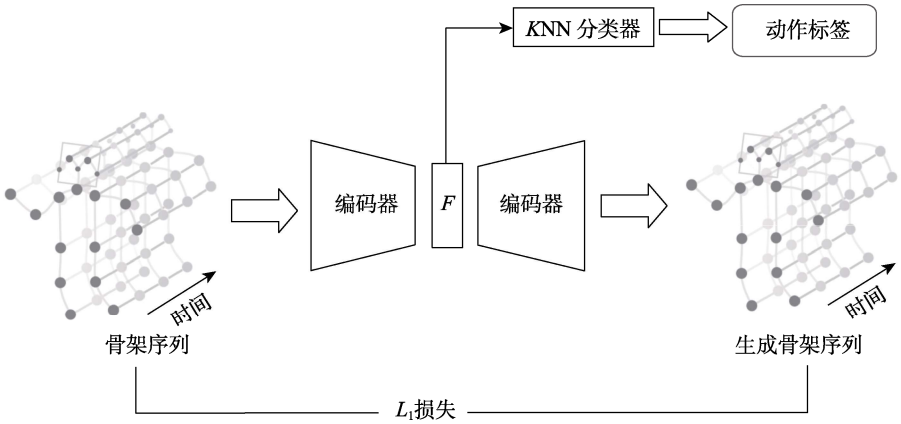


图 3 编码器-解码器网络

1.2 相关综述工作简介

当前，国内外关于人体动作识别的综述文献集中在基于视频的动作识别算法方面[1,39-42]. Wu 等[39]根据数据集属性特点对人体动作数据集进行划分，并以划分的数据集类别为基础，进行基于深度学习的动作识别算法介绍和归纳，但并未涉及目前使用较为广泛的骨架数据模态；朱煜等[1]对基于深度学习的动作识别算法进行归纳总结，同样也未涉及骨架动作识别的相关算法；胡建芳等[40]系统地介绍了基于深度图和骨架的动作识别算法，包括传统算法和深度学习算法，但对骨架相关的动作识别算法总结较少，时间上只总结到 2018 年；黄晴晴等[41]侧重于对基于 RGB 图像和深度图这 2 个数据模态的动作识别算法进行归纳总结，同样对骨架相关的动作识别算法总结较少，时间上只总结到 2018 年; Zhu 等[42]十分全面地对基于视频的动作识别算法进行分析，包括算法的细节处理以及评估设置，较为不足的是未涉及骨架数据模态的相关算法.

近年来，也有部分综述文献专注在基于骨架的动作识别方向[19,43-44]. Wang 等[43]对基于 Kinect 硬件设备的 10 种动作识别算法进行分析，但对骨架相关的动作识别算法很少，也未根据算法类型进行细致的划分；Ren 等[19]和 Xing 等[44]系统地对基于骨架的动作识别算法进行总结，包含基于 RNN 的骨架动作识别算法、基于 CNN 的骨架动作识别算法和基于 GCN 的骨架动作识别算法，但相关算法最新只总结到 2020 年，并且都不包含贴合未来发展趋势的基于半监督和基于无监督学习的骨架动作识别算法.

目前，大多数动作识别综述文献依然将关注点放在基于视频的动作识别算法总结方面，基于骨架数据模态的动作识别综述相对稀少，致使该领域的总结不够全面[1,42]；另外，近 2 年的一些重要的新工作、新创意在当前已经发表的文献中没有进行总结[19,43]，截至目前最新的骨架动作识别综述文献中也只总结到 2020 年，并未包含基于半监督和基于无监督的骨架动作识别算法(如表 1 所示). 虽然近年来相关领域取得了较大进展，但也面临着过度依赖大数据、大算力和大模型等挑战，如何应对这些挑战并促进相关算法和技术落地应用也需要进一步讨论.

为了应对上述问题，本文聚焦基于深度学习的骨架人体动作识别技术领域，旨在对该领域发展现状、面临挑战和未来发展方向进行总结与探讨. 本文不仅介绍已经发表一段时间的经典工作，更注重对该领域提出不久的新工作、新进展进行回

表 1 包含骨架数据模态的综述工作对比

综述文献	骨架数据集介绍		深度学习算法介绍				
	出版年	算法个数	最新算法提出年	算法个数	监督	半监督	无监督
胡建芳等[40]	2016	12	2018	7	✓	×	×
黄晴晴等[41]	2016	4	2018	3	✓	×	×
Ren 等[19]	2019	2	2019	16	✓	×	×
Xing 等[44]	2019	2	2020	26	✓	×	×
本文	2021	15	2022	69	✓	✓	✓

注. 骨架数据集个数只统计包含骨架数据模态的数据集，算法个数只统计属于深度学习技术的基于骨架的动作识别算法.

顾，时间上总结到本文撰写完成之时，即 2022 年 5 月．图 4 所示为主流监督学习算法的年历表概览，图 5 所示为主流半监督学习和无监督学习算法的年历表概览．

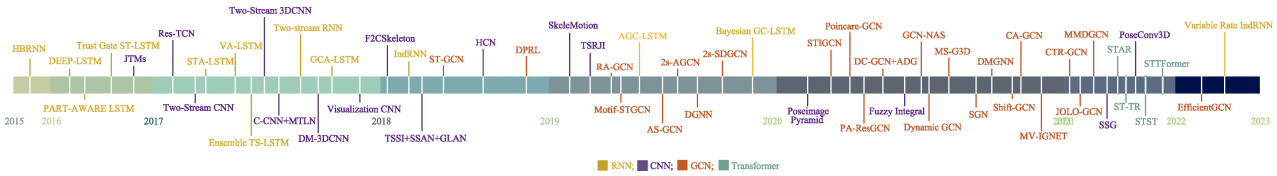


图 4　基于监督学习的骨架动作识别算法年历表概览

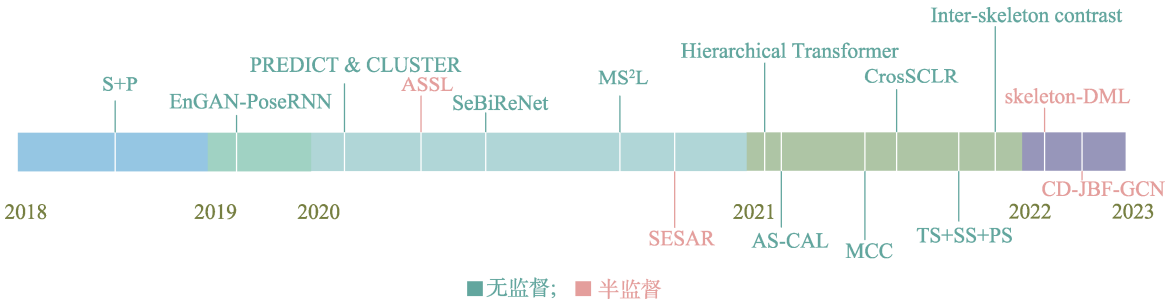


图 5　基于半监督学习和无监督学习的骨架动作识别算法年历表概览

2　数据集简介

在数据驱动的深度学习时代，数据在算法的性能表现上发挥着十分重要的作用．随着 Kinect 等深度相机的普及，国内外研究人员提出了一系列人体骨架动作行为数据集．本节归纳总结了 15 个包含骨架数据模态的公开数据集，如表 2 所示；同时对使用较为广泛的 MSR Daily Activity 3D[23]，Northwestern-UCLA[45]，SYSU 3DHOI[46]，NTU[14,47] 和 Kinetics-Skeleton[10]数据集进行回顾．

2.1　MSR Daily Activity 3D 数据集

MSR Daily Activity 3D 数据集[23]由 Kinect 相机捕获，记录了 10 个受试者的 16 种日常行为数据(如喝水、吃饭等)，每种行为分别在站姿和坐姿状态下拍摄一次，一共包含 320 个样本，每个样本同时包含 RGB 图像、深度图和骨架 3 种模态的数据．该数据集采用交叉受试者评估设置，其中，5 个受试者样本作为训练集，另外 5 个受试者样本作为测试集．由于该数据集在真实背景下拍摄，且大部分样本涉及人体与周围物体的交互，因此采集到的 3D 骨架噪声较多，具有一定的挑战性．

2.2　Northwestern-UCLA 数据集

Northwestern-UCLA 数据集[45]由 3 个 Kinect 相机从 3 个视角同时捕获受试者的日常行为数据(如丢东西、走动、坐下等)，包含 RGB 图像、深度图和骨架 3 种数据模态．该数据集记录了 10 个受试者的 10 种日常行为，包含 1 475 个样本；部分人体行为包含与周围物体对象的交互，且部分行为之间具有较高的相似性，因此较难区分，具有一定的挑战性；采用交叉受试者评估设置，一半受试者样本用于训练，另一半用于测试．

2.3　SYSU 3DHOI 数据集

SYSU 3DHOI 数据集[46]是由中山大学收集的专注于人体与物体交互的数据集，包含 RGB 图像、深度图和骨架 3 种数据模态．该数据集记录了 40 个受试者的 12 种“人-物”交互行为(如喝水、打电话等)，包含 480 个视频样本．在每个行为中，受试者仅使用手机、椅子、书包、钱包、扫把以及拖把中的一种物体进行动作演示．因此，该数据集的挑战在于一些动作涉及的交互对象物体外观比较相似，容易产生混淆．该数据集有 2 种评估设置：(1) 基于样本出发．一半样本用于训练，另一半样本用于测试；(2) 基于受试者出发．一半受试者样本用于训练，另一半用于测试．

2.4　NTU 数据集

NTU RGB+D 数据集是 Shahroudy 等[14]于 2016 年提出的，使用 Microsoft Kinect v2 传感器进行数据采集，共包含 56 880 个视频样本．该数据集的样本如图 6 所示，每个样本包含深度图、3D 骨架、RGB 帧和 IR 序列 4 种数据模态．其中，3D 骨架由 25 个人体关节点的 3D 位置信息构成．该数据集安排 40 个受试者进行数据采集，包含 60 个动作大类(40 个日常动作、9 个健康相关动作、11 个交互动作)；提出交叉受试者评估设置(cross-subject，X-Sub)

和交叉视角评估设置(cross-view, X-View)对算法进行评估. 其中, X-Sub 将 40 个受试者分为 2 类, 训练集和测试集各 20 个受试者, 训练集有 40320 个样本, 测试集有 16560 个样本; X-View 将 1 号摄像机拍出的数据作为测试集, 2 号和 3 号摄像机拍出的数据作为训练集, 训练集有 37920 个样本, 测试集有 18960 个样本. 之后, 该团队提出 NTU RGB+D 的扩展版本 NTU RGB+D 120[47], 将动作类别由 60 类扩展到 120 类, 组织 106 个受试者完成数据集的采集, 共得到 114480 个视频样本, 视角也由 80 个扩展至了 155 个. 该数据集中, X-Sub 将 106 个受试者分为 2 类, 训练集和测试集各 53 个受试者; 而 X-View 改为交叉设置评估设置(cross-setup, X-Set), 选取所有具有偶数集合编号的样本进行训练, 选取具有奇数集合编号的样本进行测试.

2.5 Kinetics-Skeleton 数据集

Kinetics-Skeleton 是基于大规模动作数据集 Kinetics-400[48]提出的. Kinetics-400 数据集于 2017 年发布, 从不同的 YouTube 视频中收集而来, 包含 306245 个视频片段, 涵盖 400 个动作类别, 每个类别至少包含 400 个视频片段, 每个片段大概持续 10 s; 该数据集中涵盖的动作类别较为广泛, 包括人与物的交互(如演奏乐器等)和人与人的交互(如握手、拥抱等), 由于其是从不同的 YouTube 视频中收集而来, 并不包含骨架这种数据模态. Yan 等[10]使用 OpenPose[16]人体姿态估计算法对 Kinetics-400 数据集进行人体骨架数据提取, 得到 Kinetics-Skeleton 数据集; 其 2D 骨架数据包含 18 个人体关节点的位置信息和置信度, 在面对多人动作时, 该数据集选取平均置信度最大的 2 个人体骨架表示该动作. Kinetics-Skeleton 数据集的训练集包含 240000 个骨架序列, 测试集包含 20000 个骨架序列.

3 基于监督学习的骨架动作识别算法

按照应用的深度神经网络类型的不同, 本文将基于监督学习的骨架动作识别算法按照基于 RNN 的骨架动作识别算法、基于 CNN 的骨架动作识别算法、基于 GCN 的骨架动作识别算法和基于 Transformer 的骨架动作识别算法 4 大类进行介绍.

表 2 包含骨架的人体动作数据集

数据集	出版年	视频数	类别数	个体数	视角	关节点数
MSR-Action3D[49]	2010	567	20	10	1	20
CAD-60[50]	2011	60	12	4		15
MSR Daily Activity 3D[23]	2012	320	16	10	1	20
UT-Kinect[24]	2012	200	10	10	4	20
CAD-120[51]	2013	120	20	4		15
Multiview 3D Event[52]	2013	3815	8	8	3	20
Northwestern-UCLA[45]	2014	1475	10	10	3	21
UTD-MHAD[53]	2015	861	27	8	1	20
UWA3D Multiview Ⅱ[54]	2015	1075	30	10	5	20
NTU RGB+D[14]	2016	56880	60	40	80	25
SYSU 3DHOI[46]	2017	480	12	40	1	25
Kinetics-Skeleton[10]	2018	260000	400			18
NTU RGB+D 120[47]	2019	114480	120	106	155	25
NTU60-X[55]	2021	56148	60	40	80	118
NTU120-X[55]	2021	113821	120	106	155	118

图 6 NTU RGB+D 数据集样本示例[14]

3.1　基于 RNN 的骨架动作识别算法

RNN 对于长时间间隔语义信息具有较好的建模能力，已被广泛应用于自然语言、视频分析等领域．描述人体动作的骨架序列本身是一种沿时间轴展开的关键点序列，因此 RNN 很自然地被用于进行基于骨架的动作识别研究．虽然传统意义上 RNN 已经可以较好地对时序特征进行建模，但传统 RNN 存在梯度消失的问题，随着时间长度的增加，传统 RNN 对于时间间隔较长的骨架特征的保留能力会变弱．长短时记忆网络 (long short-term memory, LSTM)[27]可以有效地缓解传统 RNN 梯度消失的问题，使得模型能够对序列数据的长期依赖进行较好的建模．因此，当前基于 RNN 的骨架动作识别算法大多数基于 LSTM，本文着重对基于 LSTM 相关的骨架动作识别算法进行探讨．在深度独立 RNN(independently RNN, IndRNN)[56-57]提出之后，基于 RNN 的骨架动作识别算法得到进一步的发展，这是因为 IndRNN 将 RNN 内的神经元解耦，使得神经元相互独立，可解释性更强，并且 IndRNN 可以较好地缓解梯度消失的问题，因此可以构建更深、更长的 RNN 来捕获动作语义特征．

本文将基于 RNN 的骨架动作识别算法分为 5 个方面: (1) 将 RNN 引入到骨架动作识别任务中; (2) 考虑 RNN 空间建模能力较弱，介绍空间结构建模算法; (3) 介绍注意力机制相关算法; (4) 考虑视角差异对模型识别率影响较大，介绍骨架视角一致性算法; (5) 考虑目前一些较新的工作将 RNN 与 GCN 结合使用，介绍 RNN 与 GCN 相结合的算法．

3.1.1　RNN 的引入

早期的工作重点关注如何对骨架序列的时序特征进行建模，同时减少模型参数量防止过拟合[14,58-59]．Du 等[58]提出端到端的层次双向 RNN (hierarchical bidirectional RNN, HBRNN)，其结构如图 7 所示. HBRNN 使用双向 RNN 提取关节点过去和未来的语义特征；同时，先将人体分为 5 个部件输入到 HBRNN 中进行特征提取，再通过层次结构逐层将人体部件的特征进行融合，以提取到更高层次的人体语义特征．

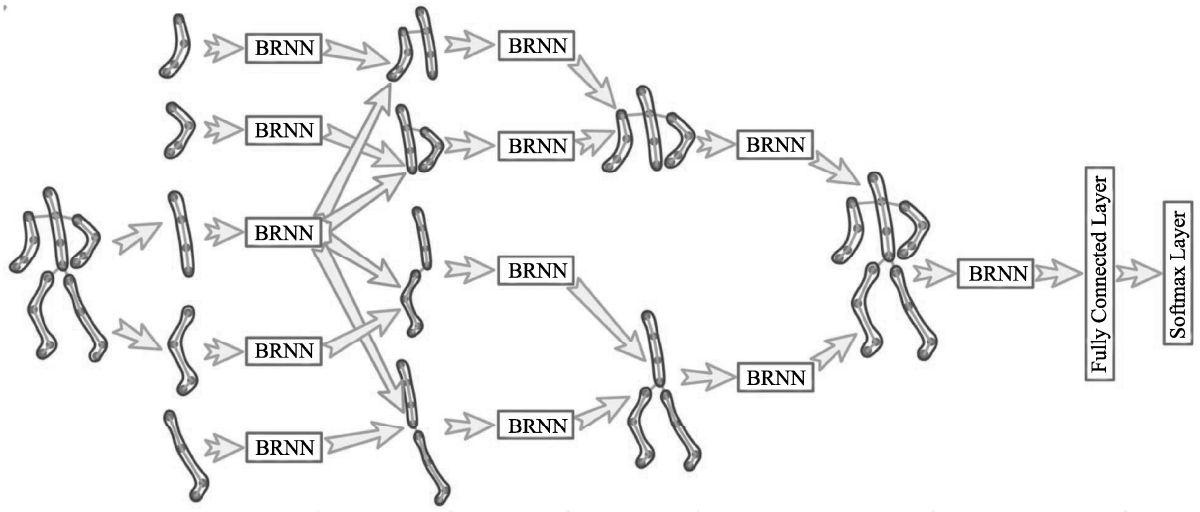


图 7　HBRNN 结构[58]

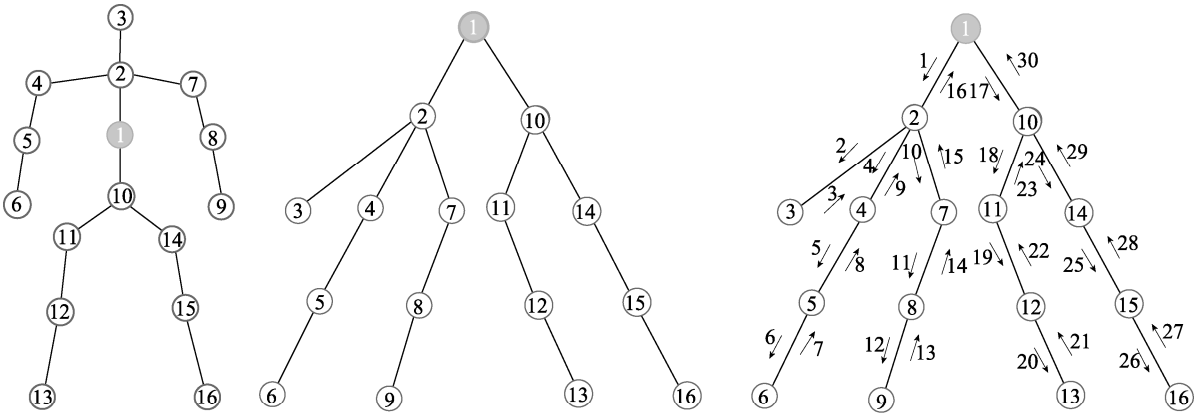
Zhu 等[59]提出端到端的全连接深度 LSTM，以全连接的方式探索关节点间的共生关系；对传统的 Dropout 机制[60]进行改进，允许网络对 LSTM 内部的神经单元进行丢弃，使得训练过程更加高效．Shahroudy 等[14]提出部件感知的 LSTM，首先将人体骨架分成 5 个部件，然后每个部件作为一路独立的信号送入到神经网络中进行时序特征学习，最后网络的输出层由事先划分的 5 个部件共享；通过 5 个身体部件共享同一个输出层这种机制，可以将不同身体部件的特征进行整合，得到整个身体的上下文信息，同时也减少了模型的参数量，防止过拟合．

3.1.2　空间关节点的建模算法

对于骨架动作识别问题来说，空间特征的建模也十分重要，空间特征的建模考虑空间内关节点之间的链接顺序、相对位置关系等．其中，链接顺序是指将空间内人体关节点按一定的顺序链接起来，文献[14,59]仅仅按人体关节点的编号进行链接，没有考虑人体关节点的链接顺序．因此，将时间特征和空间特征进行联合建模，探索时空的一致性关系，逐渐成为相关工作关注的重点问题[61-62]．

Liu 等[61]提出时空 LSTM，将 LSTM 扩展到时空域，探索时间特征和空间特征之间的隐藏关系；

受图结构的启发，提出双向遍历的树状结构(如图 8 所示)，更好地考虑了空间内关节点的链接顺序. 除了直接将时间和空间进行统一建模外，也有一些算法通过双流网络的设置，先对时间特征和空间特征分开建模，再进行双流特征的融合. Wang 等[62]提出双流 RNN 来建模时间特征和空间特征，提出层次 RNN 来建模时间特征，思路与 HBRNN[58] 类似. 在空间建模方面提出新的关节点链接顺序，选择躯干的中心节点作为起始点遍历到左臂，再到尾端，最后返回到右臂，并通过遍历正向和反向 2 个方向的关节点链接顺序来保证图中的空间结构关系. 在关节点的空间关系建模方面，双向遍历的树状结构可以更好地考虑骨架的空间结构以及关节点之间的交互方式.



a. 链式结构 b. 树状结构 c. 双向遍历树状结构

图 8 关节点的遍历方式[61]

3.1.3 基于注意力机制的算法

Song 等[63]和 Liu 等[64]发现，不同关节点或不同的骨架帧对于动作被识别成功的贡献度是不一样的，即每一帧中关节点的重要性不同，不同骨架帧也有不同的重要性，因此将工作的重点放在模型的注意力机制方面.

Song 等[63]提出一个端到端的时空注意力 LSTM，通过高效地提取时间特征和空间特征来建模时间和空间的变化. 网络分为空间注意力模块和时间注意力模块. 其中，空间注意力模块通过关节点选择门控单元来确定贡献度较大的关节点；时间注意力模块通过帧选择门控单元来挑选比较重要的骨架帧. 虽然该时空注意力机制本质上是较为简单地对关节点和骨架帧增加一个贡献度权重，但这是对骨架动作识别注意力机制的一次有效尝试. 而 Liu 等[64]提出基于全局上下文注意力机制的 LSTM，以全局语义信息为指导，在保持骨架序列建模能力的同时也对重要的关节点进行了关注；提出递归注意力机制，将初始化的注意力表征重新输入到全局语义存储单元中进行迭代，来对全局语义进行渐进式改进.

注意力机制的引入使得网络对重要关节点或骨架帧的计算权重加大，在骨架动作识别任务上取得了显著的性能提升，这是因为大部分人体动作是由少数重要的关节点的运动来完成的.

3.1.4 骨架视角的一致性算法

对于 3D 骨架数据，还需要考虑的关键因素是视角变化. 在不同的视角下观测到的骨架外观差异很大，如何保证视角不变的一致性或者选择最佳的视角对骨架进行观测，是一个值得深入研究的问题.

Lee 等[65]采用人为定义视角的方式，通过预处理(缩放、旋转、平移)使得骨架序列在坐标系中的方向对齐，骨架序列具有方向一致性；考虑不同时间步长的 LSTM 可以建模骨架序列的多种动态特征，提出集成时间滑动的 LSTM，可以同时捕捉短、中、长 3 种时间依赖. Zhang 等[66]提出一种视角自适应方案，可以在动作发生时自动调节观测骨架的视角，并设计一种端到端的视角自适应 LSTM，能够自适应地寻找出最合适的观测视角.

3.1.5 RNN 与 GCN 结合算法

RNN 的优势是在对时间特征的建模方面，其对于空间特征的建模效果并不理想，因此考虑引入 GCN 来弥补 RNN 空间建模能力不足的问题. Si 等[67]首次将 LSTM 与 GCN 相结合，提出注意力增强图卷积的 LSTM(attention enhanced graph convolu-

tional LSTM, AGC-LSTM). 将 GCN 融入到 LSTM 中，在更好地提取动作特征的同时也探索了时空的一致性关系. 与 AGC-LSTM[67]不同，Zhao 等[68]先通过 GCN 提取空间特征，再用 LSTM 提取时间动态特征，最后在贝叶斯框架下将整个模型扩展为概率模型，以便于网络更好地捕捉到数据集中动作样本的随机性，提高模型的泛化能力.

表 3 总结了在 NTU-60 数据集上取得准确率前 10 名 RNN 相关算法. 可以看出, RNN 和 GCN 相结合能够较大地提高模型的识别准确率. 类似地，Li 等[69]利用 RNN 与 CNN 相结合的方式提升模型的识别准确率，这种融合不同网络结构的优势的方式在视频动作识别任务中也得到了很好的应用. 例如, Muhammad 等[70]将基于双向 LSTM 的注意力机制与膨胀 CNN 相结合; He 等[71]利用空间和时间 CNN 提取动作的空间特征和短时特征，并通过密集连接的双向 LSTM 提取动作的长时特征.

表 3　NTU-60 数据集上准确率前 10 名 RNN 相关算法

算法	出版年	X-Sub/%	X-View/%
AGC-LSTM[67]	2019	89.20	95.00
dense-IndRNN-aug[56]	2018	86.70	93.97
Variable Rate IndRNN[57]	2022	84.32	89.71
Bayesian GC-LSTM[68]	2019	81.80	89.00
VA-LSTM[66]	2017	79.40	87.60
Ensemble TS-LSTM[65]	2017	74.60	81.25
GCA-LSTM[64]	2017	74.40	82.80
STA-LSTM[63]	2017	73.40	81.20
Two-stream RNN[62]	2017	71.30	79.50
Trust Gate ST-LSTM[61]	2016	69.20	77.70

3.2　基于 CNN 的骨架动作识别算法

基于 RNN 的骨架动作识别算法对人体动作空间特征的建模通过对空间内人体关节点的链接顺序的调整来完成，该机制虽然可以提取一定的人体动作空间特征，但比较简单，对人体动作空间特征的建模能力相对较弱. 同时，基于 RNN 的骨架动作识别算法也无法很好地考虑动作发生时骨架关节点之间的协同性，对关节点空间特征的聚合能力考虑不够. 而 CNN 天然具有较好的空间结构特征聚合能力，已经在图像分类任务上取得了十分优秀的效果. 因此，通过 CNN 对骨架的空间结构进行建模，也成为解决骨架动作识别任务的研究方向之一.

本文将基于 CNN 的骨架动作识别算法分为 2 个方面: (1) 将 2D CNN 引入到骨架动作识别任务中，针对将骨架序列转换为 2D 数组难以凸显出不同关节点之间的差异的问题，对骨架序列转换为 RGB 伪图像的算法进行回顾; (2) 探索了目前较新的 3D CNN 如何应用在骨架动作识别任务中. 如图 9 所示，基于 2D CNN 和 3D CNN 的骨架动作识别的区别在于骨架序列的表征方式不同，以及表征之后提取特征的算法不同.

3.2.1　2D CNN

CNN 最初被应用到骨架动作识别任务中时，研究人员只是用 CNN 模块对已有工作中的部分模块进行简单替换. Du 等[72]先将骨架序列编码为 2D 矩阵，骨架序列的时间动态被编码为列，帧内的空间结构被编码为行，再将其送入 CNN 中进行特征提取; Kim 等[73]用 CNN 代替之前的 RNN 操作，同时引入残差机制提出残差时序 CNN; Li 等[74]提出双

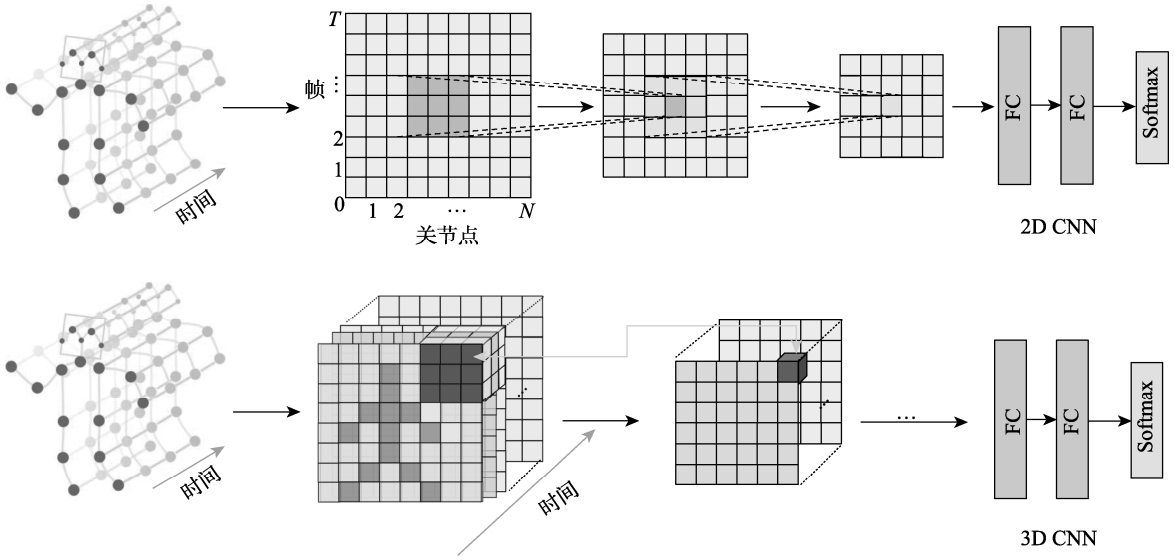


图 9　2D CNN 和 3D CNN 骨架动作识别流程

流 CNN 和一个骨架转换模块，该模块类似于注意力机制，可以自动选取比较重要的关节点；Li 等[75]在后续的工作中提出端到端的层级卷积一致特征学习网络，引入全局空间聚合方案，先用骨架序列的坐标维数作为卷积通道，再将关节点数转换为卷积通道，可以提取到空间结构上距离很远的非物理连接的关节点之间的关系.

上述算法只是对骨架序列进行简单地表征，没有捕捉运动过程中关节点间的差异. 因此，研究人员提出将骨架序列转换为 RGB 伪图像的思路，在较好地捕捉骨架空间结构特征的同时，还能对骨架时序特征的动态变化进行建模. Wang 等[76]首次将 3D 骨架的时空信息通过关节点轨迹图映射到 3 个正交平面中，即将骨架序列转换为 3 幅 RGB 伪图像；同时，注意到在大尺度动作发生时个别相关关节点会有更大幅度的空间变化，因此通过 RGB 伪图像色彩上的饱和程度与明亮程度来捕捉运动变化的强弱. 该算法是对骨架序列转换为 RGB 伪图像的一次有效尝试，但对人体动作空间结构的建模以及时序动态的考虑仍然不足.

一些研究将关注点转向对关节点空间结构的建模方面. 文献[77-79]参照关节点的设置来获取人体关节点的相对坐标，参照关节点指通过设置一些在人体运动过程中较为稳定、变化较小的关节点 (如左肩、右肩、左臀、右臀). Ke 等[77]将骨架序列转换为 3 个片段，片段根据人体坐标的 3 个维度对骨架序列进行划分，代表整个骨架的时序信息；之后将每个片段根据参照关节点转换成 4 幅 RGB 伪图像(如图 10 所示)；最后将生成的 RGB 伪图像输入到 CNN 中进行特征提取. 与 Ke 等[77]的思路类似，Le 等[79]通过参照关节点的设置将骨架序列转换为 RGB 伪图像，不同的是在计算关节点位置特征的同时还加入关节点的速度特征.

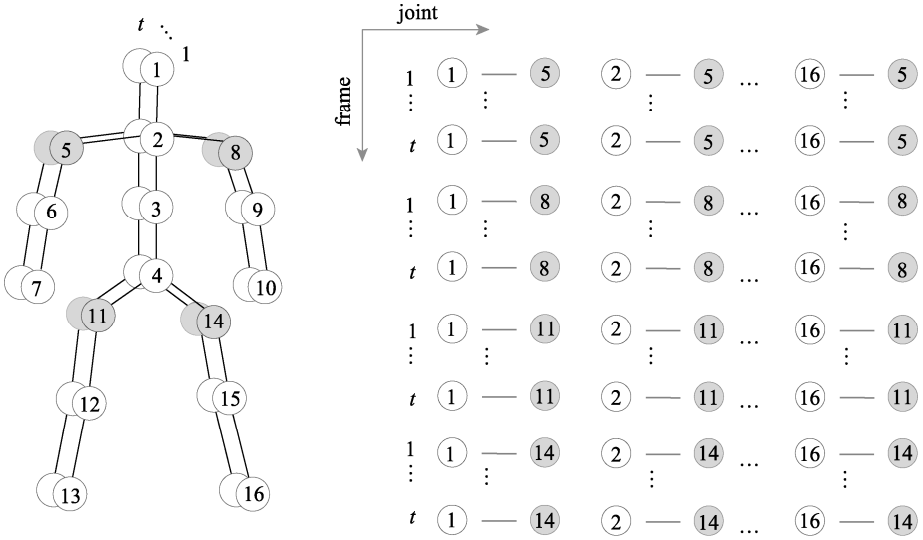


图 10 骨架序列生成片段[77]

还有一些研究从关节点的链接顺序角度出发考虑对人体动作的空间结构关系进行建模，通过树状遍历方式保留人体骨架的语义信息与空间结构. Yang 等[80]提出树状骨架图像，按深度优先树状遍历的顺序链接关节点；Caetano 等[78]提出树状参照关节点图像，将树状结构遍历与参照关节点相结合：前者通过树状遍历关节点保留了人体的空间关系，后者通过参照关节点整合了关节点间不同的空间关系.

也有一些研究通过一些特定的设置(如视角不变的预处理、新的骨架表征等)，将骨架序列转换为 RGB 伪图像. Liu 等[81]提出增强骨架可视化的算法，首先对骨架序列进行视角不变的预处理，通过躯干关节生成视角变换矩阵，用视角变换矩阵对骨架序列进行同步变换，消除人体关节点视角变化的影响；然后对转换后的骨架序列的 5 个维度设定 10 种排列，生成 RGB 伪图像. Caetano 等[82]提出一种骨架图像表示算法 Skelemotion，直接计算关节点变化的位移和方向来编码时间动态特征(如图 11 所示)；通过树状遍历的方式将关节点链式连接，保留人体关节点的空间结构信息. Banerjee 等[83]认为，已有的算法难以捕捉到骨架序列不同维度的特征，其基于角度信息和运动学信息提出 4 种互补的骨架序列特征表示，分别为距离特征、角度特

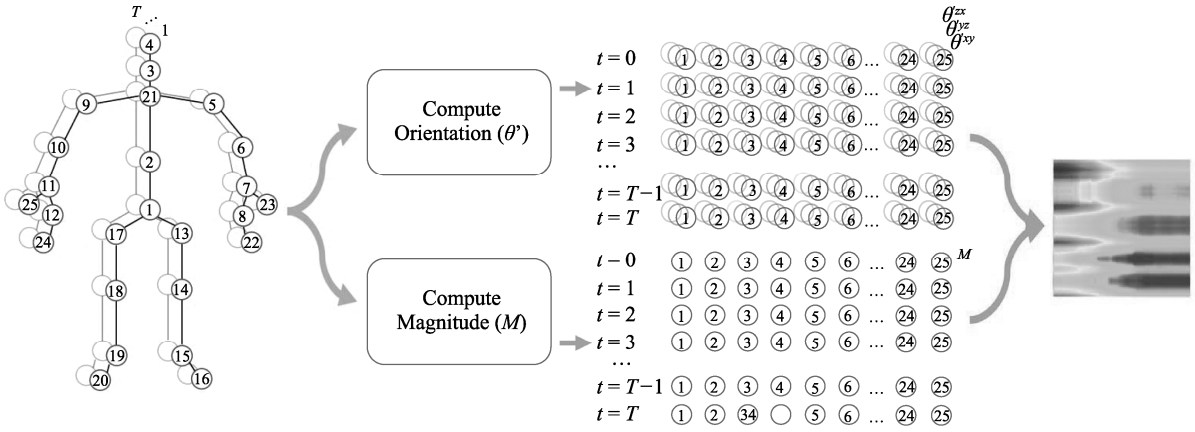


图 11　Skelemotion 表示流程[82]

征、距离向量特征和角度向量特征，利用单通道的灰度图像对这 4 种特征进行编码.

3.2.2　3D CNN

由于可以同时进行空间和时间维度的联合特征学习[84]，越来越多的研究将 3D CNN 应用在视频动作识别中[85-87]，并取得了较好的效果；骨架动作识别领域也有相关尝试. Liu 等[88]将 3D CNN 应用到骨架动作识别任务中，提出双流 3D CNN，先将关节点的坐标特征从时间维度和空间维度分别进行编码，再用双流 3D CNN 进行高维特征提取，进而基于高维特征进行动作类型识别；Ruiz 等[89]通过距离矩阵(用于衡量不同关节点之间距离的矩阵)对每一帧的骨架进行表征，再通过时空 3D CNN 对按时间顺序堆叠的距离矩阵进行特征提取；Lin 等[90]通过计算 2 个关节点之间的归一化距离(欧几里得距离除以最短路径距离)对每一帧的骨架进行表征，这种表征方式具有相似动作转换不变性的优点(即相似动作通过归一化距离表征之后差异较小)；Ding 等[91]将骨架序列转换为基于关节点的正方形网格和基于刚体段的正方形网格(刚体段指人体的骨骼)，分别构建人体各部件的内在和外在依赖，通过双流 3D CNN 来联合学习骨架序列的时空特征.

在较新的 3D CNN 骨架动作识别工作中，Duan 等[92]提出 PoseConv3D 网络，首先将骨架序列转换为 3D 热力图，然后通过 3D CNN[86]对 3D 热力图进行特征提取. 将骨架序列表征为 3D 热图可以更加高效地学习到人体动作的时空特征，同时模型对于骨架数据中的噪声也更加鲁棒，泛化能力更好. 本文总结了在 NTU-60 数据集上取得准确率前 10 名 CNN 相关算法，如表 4 所示. 可以看出，PoseConv3D 算法取得了最好的识别准确率.

表 4　NTU-60 数据集上准确率前 10 名 CNN 相关算法

算法	出版年	X-Sub/%	X-View/%
PoseConv3D[92]	2021	93.70	96.60
SSG+attention+view[91]	2021	90.20	95.70
HCN[75]	2018	86.60	91.10
Banerjee 等[83]	2020	84.22	89.71
Poseimage Pyramid[90]	2020	84.00	90.50
Two-Stream CNN[74]	2017	83.20	89.30
TSSI+SSAN+GLAN[80]	2018	82.40	89.10
DM-3DCNN[89]	2017	82.00	89.50
Visualization CNN[81]	2017	80.03	87.21
F2CSkeleton[79]	2018	79.60	84.60

3.3　基于 GCN 的骨架动作识别算法

经过多年发展，基于 RNN 和 CNN 的骨架动作识别算法在相关数据集上的准确率进入了瓶颈期，难以取得进一步的突破. 原因有 2 个方面：(1) 基于 RNN 的骨架动作识别算法对空间结构特征的建模能力不足；(2) 虽然基于 CNN 的骨架动作识别算法可以缓解空间结构特征建模难的问题，但将骨架序列转换成 RGB 伪图像本身就存在一部分的信息丢失，并且 CNN 也无法较好地对时序特征进行建模. 因此，寻求进行人体骨架动作时间和空间同时统一建模的思路是进一步提高算法准确率的重要方向之一.

由于人体骨架中关节点之间的关系比较适合用图拓扑结构来进行建模，研究人员开始将注意力转向基于 GCN 的骨架动作识别算法. 图卷积是在图结构上进行卷积运算，所以基于 GCN 的骨架动作识别算法的关键在于如何将卷积技术应用到人体骨架图拓扑结构中. 当前，按照实现方式的不同，人体骨架动作识别中图卷积可以被分为空域图卷积[10]和频域图卷积[93]这 2 大类.

Yan 等[10]首先将人体骨架序列在时间和空间上进行统一建模, 设计人体骨架时空图(如图 12 所示), 实现了时空 GCN(spatial temporal GCN, ST-GCN). 该网络在空间维度上将人体关节点作为图的顶点, 关节点间的物理连接作为图的边; 在时间维度上, 将不同帧中的人体骨架同一位置的关节点进行连接, 基于上述空间和时间维度机制构建人体骨架的时空图. ST-GCN 的提出让其他研究人员看到了 GCN 在解决人体骨架动作识别课题上的优势.

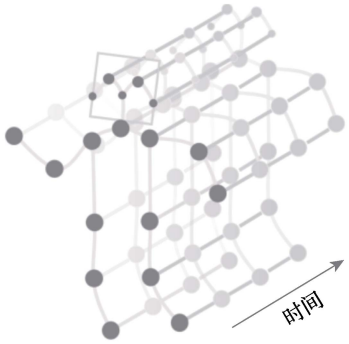


图 12 人体骨架时空图[10]

ST-GCN[10]的提出在骨架动作识别领域产生了较大的影响, 后续许多基于 GCN 的骨架动作识别研究都是围绕对 ST-GCN[10]的改进而展开. 因此, 本文将基于 GCN 的骨架动作识别算法分为 4 个方面: (1) 由于 ST-GCN[10]图拓扑结构固定, 并且对于非物理连接考虑不够充分, 因此回顾研究人员是如何对这一不足进行改进的; (2) 在网络结构优化方面, 总结残差网络的应用以及网络多流融合在效率方面提升的问题; (3) 讨论一些通过引入多视角、多尺度来探索更多的图卷积交互方式的工作; (4) 整理部分通过考虑提取人体自然语义来获得更高维、更抽象的人体动作特征工作.

3.3.1 图拓扑结构的工作

ST-GCN[10]基于人体关节点的物理连接提取骨架的空间特征, 因此未充分考虑关节点的非物理连接情况, 如在拍手这一动作中, 物理连接指手腕关节点与手肘关节点之间的连接关系, 非物理连接则是左手关节点与右手关节点之间的连接关系. 针对上述问题, 研究人员陆续提出一些创新性的改进思路.

Wen 等[94]将物理连接的关节点分为关节点本身、父节点和子节点 3 类, 通过建立有向图来建模物理连接的关节点; 在非物理连接的建模中提出加权邻接矩阵机制, 对空间距离越近的 2 个关节点分配更大的权值; 提出可变长时间块(不同时间步长的卷积核)来建模关节点的时序动态特征. Li 等[11]提出运动结构 GCN(actional-structural GCN, AS-GCN), 包含运动连接和结构连接 2 个子网络. 其中, 运动连接子网络设计了编码器-解码器网络来捕捉每个关节点与其他所有关节点之间的关系特征; 结构连接子网络本质上是对 ST-GCN[10]算法中的邻域进行扩展, 使得更多的人体关节点线索信息能得以利用.

ST-GCN[10]的图拓扑结构是手工设置的, 即该网络不同层上的图拓扑结构是固定的; 考虑人体动作的发生过程是一个动态变化过程, 这种固定的图拓扑结构可能不是最好的. 因此, Shi 等[95]提出双流自适应 GCN(two-stream adaptive GCN, 2s-AGCN), 其亮点是所包含的图拓扑结构不仅能够以端到端的方式进行统一学习, 也能做到单独学习. 之后, Shi 等[96]对 ST-GCN[10]与 2s-AGCN[95]进行改进, 用有向无环图表征人体骨架, 提出有向图神经网络, 用于提取关节点、骨骼以及这两者之间交互关系的特征信息; 为了更好地学习到动态的图拓扑结构, 对图拓扑结构增加了自适应机制, 进一步改善了关节点之间的交互受限于人体关节点物理连接的问题. Peng 等[97]采用神经网络架构搜索的方式生成动态图拓扑结构, 并设置了多个动态图模块来丰富搜索空间. Ye 等[98]提出动态 GCN, 充分利用 GCN 的图拓扑结构学习能力和 CNN 特征提取能力. 动态 GCN 由一种轻量级的上下文编码网络(context-encoding network, CeN)构成, CeN 可以全局地学习上下文丰富的动态图拓扑结构, 为不同的输入样本以及不同深度的图卷积层构建动态图拓扑结构.

传统 GCN 的图拓扑结构对于网络的通道来说是固定的, 但考虑传统 CNN 在不同的通道上设置了不同的卷积核来获取更加全面的特征信息, 研究人员对如何在不同的通道上应用不同的图拓扑结构进行探索. 受到 CNN 中解耦聚合机制的启发, Cheng 等[99]提出解耦图卷积来增强 GCN 的建模能力, 在不同的通道上设置了独立可训练的邻接矩阵来解决图拓扑结构在通道上固定的问题. Cheng 等[100]提出移位 GCN, 由新的移位图操作和轻量点卷积组成; 在空间和时间层面分别提出移位图卷积操作, 空间层面提出的全局移位图卷积是对当前图是完全连接的图进行移位, 时间层面提出的自适应移位图卷积是对于每个通道都学习出一个时间偏移参数. Chen 等[101]提出通道拓扑细化 GCN

(channel-wise topology refinement GCN, CTR- GCN), 可以动态、高效地对通道图拓扑结构进行建模. 与已有算法不同, CTR-GCN 不是对不同通道进行独立的学习, 而是借由网络学习到的共享拓扑结构作为所有通道的先验知识, 使用每个通道中关节点之间特定的关系对其进行改进.

3.3.2　网络结构的高效化

一方面, 传统 GCN 网络层数一般较深, 参数量较大, 导致网络训练和推理的效率较低; 另一方面, 多流网络融合效率也比较低. 多流网络是指将多个不同的骨架表征(如人体关节点的位置特征、骨骼特征)并行输入到 GCN 中, 从而达到获取更加全面的人体动作语义特征的目的.

残差网络强大的特征提取能力以及对网络过拟合的优化引起广泛的关注, 相关研究人员将残差思想应用到 GCN 中. Wu 等[102]提出基于空间残差和密集连接的双流 GCN, 包含空间残差层和密集连接块 2 个模块. 其中, 空间残差层将空间图卷积与时空图卷积进行跨域连接, 可以提取更精确、更丰富的人体动作特征; 密集连接块使用密集连接来充分利用全局特征信息, 用少量的计算获取丰富的特征图. Huang 等[103]将 CNN 的“拆分-转换-融合”策略应用到图卷积中, 对输入特征进行一个多路径(空间路径、时间路径和残差路径)的处理, 同时论证了这种增大转换的操作集的方法(多路径处理的思路)比增大卷积核的接收域更加有效.

目前, 多流网络的特征融合往往在最后的分类之前完成, 会造成模型计算量较大, 导致训练推理效率低下. Song 等[104]提出早期融合多输入分支的思想, 在模型的前几个阶段就将多流分支进行融合, 可以大大减少模型的复杂度, 并使得训练更加容易拟合; 在此基础上, Song 等[105]又提出了高效 GCN, 相比之前的工作[104], 多分支融合的思路不变, 以一组固定的缩放系数对网络宽度、深度进行均匀缩放, 在保持识别准确率的同时使网络更加高效.

3.3.3　多尺度和多视角

传统 GCN 的图拓扑结构一般是对整个骨架的图拓扑结构进行考量, 导致交互方式较为单一, 忽略了局部身体部件之间的交互. 目前, 研究人员提出多尺度和多视角的概念为骨架提供丰富的交互信息, 提取到更加全面、高级的人体动作特征.

在多尺度方面, 针对传统的 AS-GCN 算法[11]通过邻接矩阵的高阶幂来扩大关节点的邻域, 而这种幂指数方法会存在偏权重, 即距离较近的关节点分配的权重较大, 但实际上图结构上距离很远的点也有可能有很强的关联性的问题. Liu 等[106]提出多尺度聚合(如图 13 所示), 通过移除远近邻域的冗余依赖解决网络的偏权重问题. Li 等[107]提出动态多尺度图神经网络, 通过多尺度图(第 1 个尺度是由所有关节点组成, 第 2 个尺度是将部分节点以相邻原则组合到一起, 第 3 个尺度是在第 2 尺度基础上再进行组合)综合建模人体内部关系, 进行人体动作特征学习; 同时, 提出多尺度图计算单元, 用于提取单个尺度上的特征以及融合跨尺度上的特征.

与多尺度图的思路相似, 多视角通过探索多个视角模态之间的交互, 提取更加全面、丰富的人体特征信息. Wang 等[108]提出多视角交互图网络, 可以统一构建、学习、推断多层次的空间骨架上下文语义, 并且利用不同的骨架图拓扑结构(人体全局语义、人体部件语义和关节点局部语义)作为多视角来协作生成互补的人体动作特征.

3.3.4　人体自然语义的工作

人体自然语义是指不同的人体关节点或不同的骨架帧在人体动作发生时扮演的角色是大不相同的(如手部可能跟抓取相关的动作关联性较大).

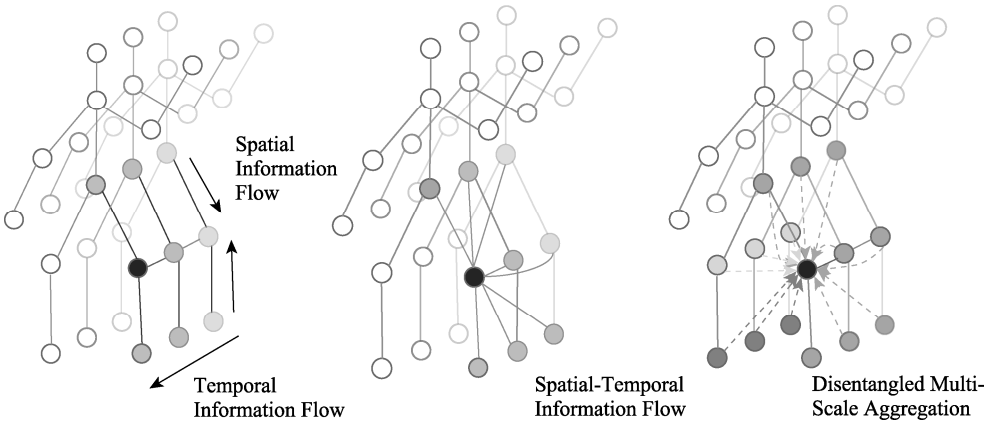


图 13　多尺度解耦聚合图卷积[106]

然而，目前大多数图卷积算法只是从骨架数据本身去考虑如何更好地提取对任务有效的特征，但人体的语义信息其实更加符合人体运动的物理意义. 如何将语义信息融入到以数据为主导的模型中，部分研究人员提出了创新性的思路.

Zhang 等[109]提出语义引导神经网络，将人体关节点类型(如手、脚)和帧索引顺序(如第 1 帧、第 3 帧)的高级人体语义引入 GCN 中. 此外，图卷积操作往往是局部操作，无法完全关联全局的人体关节点，未深入考虑关节点的上下文语义信息. Zhang 等[110]将上下文语义融入图卷积，提出一种上下文感知的 GCN(context aware GCN, CA-GCN)，除了局部图卷积(关节点的邻域)的计算外，通过整合所有其他顶点的信息来考虑每个顶点的上下文，高版本的 CA-GCN 将中心关节点作为接收者，被聚合的关节点称为发送者，以提取更加高级的上下文语义特征.

3.3.5 其他图卷积算法

目前，基于 GCN 算法的一个普遍缺点是仅靠稀疏的骨架信息不足以完全描述人体的动作，导致模型不能很好地识别出一些变化很小的细微动作. Cai 等[111]提出一种将骨架信息和关节中心的轻量化信息联合应用于双流 GCN 的新框架，以相对稀疏的格式将每个关节点周围的视觉信息表示为关节点对齐光流补充信息，关节点对齐光流补充信息包含了丰富的局部细微运动，使得网络对细微动作的识别效果显著提高. 此外，骨架数据是非欧几里得数据，需为骨架序列数据寻找一个合适的特征嵌入空间. Peng 等[112]在黎曼流形上提出更加高效的图卷积来学习多维的结构化嵌入，探索更好的嵌入空间建模骨架序列.本文总结了在 NTU-60 和 NTU-120 数据集上取得准确率的前 10 名 GCN 相关算法，如表 5 和表 6 所示.

表 5 NTU-60 数据集上准确率前 10 名 GCN 相关算法

算法	出版年	X-Sub/%	X-View/%
JOLO-GCN[111]	2021	93.90	98.10
CTR-GCN[101]	2021	92.40	96.80
EfficientGCN-B4[105]	2022	91.70	95.70
MS-G3D[106]	2020	91.50	96.20
Dynamic GCN[98]	2020	91.50	96.00
PA-ResGCN-B19[104]	2020	90.90	96.00
DC-GCN+ADG[99]	2020	90.80	96.60
MMDGCN[113]	2021	90.80	96.50
4s Shift-GCN[100]	2020	90.70	96.50
STIGCN[103]	2020	90.10	96.10

表 6 NTU-120 数据集上准确率前 10 名 GCN 相关算法

算法	出版年	X-Sub/%	X-Set/%
CTR-GCN[101]	2021	88.90	90.60
EfficientGCN-B4[105]	2022	88.30	89.10
JOLO-GCN[111]	2021	87.60	89.70
Dynamic GCN[98]	2020	87.30	88.60
PA-ResGCN-B19[104]	2020	87.30	88.30
MS-G3D[106]	2020	86.90	88.40
MMDGCN[113]	2021	86.80	88.00
DC-GCN+ADG[99]	2020	86.50	88.10
4s Shift-GCN[100]	2020	85.90	87.60
MV-IGNet[108]	2020	83.90	85.60

3.4 基于 Transformer 的骨架动作识别算法

在 Transformer[114]被提出之后，研究人员将其应用于各种计算机视觉任务中，并取得了令人惊叹的结果. Transformer 抛弃了传统的 RNN 和 CNN 的网络结构，整个网络完全由注意力机制构成，与 RNN 和 CNN 相比，可以大幅减少模型的计算量，使得模型变得轻量.

由于 Transformer 完全由注意力机制构成，因此网络可以直接计算任意人体关节点之间的联系，缓解 ST-GCN[11]只能捕捉局部特征的弊端. Plizzari 等[115]提出时空 Transformer 网络(spatial-temporal transformer network, ST-TR)，对时间和空间同时进行注意力机制的建模. 对于空间维度，提出空间自注意模块捕捉同一帧下不同关节点的空间特征；对于时间维度，提出时间自注意模块捕捉同一关节点在不同帧的时间特征. ST-TR[115]只是对 Transformer 应用到骨架动作识别任务中的简单尝试，没有很好地考虑骨架关节点之间的物理连接(人体关节点之间的物理连接只有 24 个，但网络计算任意 2 个关节点之间的联系，共需要 625 次计算)，造成了大量的冗余计算. Shi 等[116]提出一种基于稀疏 Transformer 的骨架动作识别，对空间维度提出稀疏注意力模块，关注来自邻域的关节点信息，并根据身体各部分的逻辑关系人为定义一些关节点之间的联系，通过这种方式来减少无效关节点之间相似度的冗余计算.

上述 2 种算法都是以相同的思路对时间和空间进行建模，结果往往不是最优的. Zhang 等[117]提出基于骨架动作识别的时空专用 Transformer，在空间层面，提出空间 Transformer，对每一帧骨架单独进行建模，并通过自注意力机制计算帧内关节点的关系；在时间层面，提出定向时间 Transformer，由于自注意力机制难以区分序列顺序，提出定向

感知策略，通过应用方向掩码来强制模型识别序列顺序．已有的算法无法捕获帧间不同类型关节点之间的相关性，但这对于人体动作识别是十分重要的，如在跳远这个动作中，前一帧的手臂与下一帧的腿相关性很高．Qiu 等[118]提出时空元组 Transformer，在连续帧中建立不同关节点之间的联系(一个骨架序列被分割为几个不重叠的部分，每部分被称为一个“元组”)；该网络中包含时空元组注意力模块和帧间特征组合模块，时空元组注意力模块用于捕获一个时空元组内不同关节点之间的关系；帧间特征组合模块用于建立不同元组之间的关系，增强对相似动作的区分能力(如跳高、跳远)．

本文总结了在 NTU-60 和 NTU-120 数据集上 Transformer 相关算法的识别准确率，如表 7 和表 8 所示．可以看出，虽然基于 Transformer 的骨架动作识别算法的网络结构较为简单，但也能取得较高的识别准确率．因此，如何将 Transformer 与骨架动作识别任务更好地结合，是下一步研究需要思考的重点问题．

表 7　NTU-60 数据集上 Transformer 相关算法准确率

算法	出版年	X-Sub/%	X-View/%
STTFormer[118]	2022	92.30	96.50
STST[117]	2021	91.90	96.80
ST-TR[115]	2021	88.70	95.60
STAR-128[116]	2021	83.40	89.00

表 8　NTU-120 数据集上 Transformer 相关算法准确率

算法	出版年	X-Sub/%	X-Set/%
STTFormer[118]	2022	88.30	89.20
ST-TR[115]	2021	81.90	84.10
STAR-128[116]	2021	78.30	80.20

4　基于半监督学习的骨架动作识别算法

基于监督学习的骨架动作识别算法通常需要大量有标签数据作为基础，而标签数据的标注十分消耗人力、物力[29-31]．因此，研究人员尝试以半监督学习的方式来进行骨架动作识别模型的训练．

Si 等[33]提出对抗无监督学习的方式实现半监督学习的骨架动作识别，通过对抗学习缓解半监督学习到的特征与无监督学习到的特征不对齐的问题；提出邻域一致性策略，即每个样本在特征空间的多个近邻都要有相似的预测结果，通过这种策略使得模型学习到有区分度的特征．

基于半监督学习的骨架动作识别算法中，有标签数据采用随机选取和均匀选取；但是，随机选取的样本可能并不能很好地代表这个类，均匀选取往往会导致模型泛化能力较差．为了对动作识别模型设置合理的有标签样本的选取策略，Li 等[34]提出主动学习来实现稀疏的半监督骨架动作识别，用主动学习的方法来为无标签数据打标签，机器自身迭代出需要打标签的样本；同时，提出新的样本选取策略，要求所选注释样本既接近聚类的中心，又与被注释的样本足够远．类似地，Memmesheimer 等[35]提出深度度量学习解决基于骨架的一次学习动作识别问题，将一次学习转换为度量学习问题，在每个动作类别中都放入一个有标签的样本，将其作为锚点，采用度量学习的思路，寻找距离锚点嵌入距离最远的正样本对以及嵌入距离最近的负样本对，从而优化样本聚类的过程．

Tu 等[119]的研究重点是对编码器网络结构和半监督学习方式进行设计，其首先提出一种关系驱动关节点和骨骼信息融合的 GCN 作为编码器，探索关节点和骨骼之间的运动转换；然后通过编码器-解码器网络进行帧预测的辅助任务自监督地训练无标签数据；最后将训练过的编码器用于少量有标签数据的监督分类任务，从而达到半监督学习的设定．

目前，半监督骨架动作识别的关键是对有标签样本的选取策略，以及有标签数据和无标签数据在嵌入空间上分布一致性的问题．未来，研究更加有效的有标签数据选取算法以及样本的邻域一致性，将成为半监督骨架动作识别工作的重点．本文总结了在 NTU-60 数据集上半监督学习相关算法的识别准确率，如表 9 所示，其中，半监督设定为有标签数据占比 10%．

表 9　NTU-60 数据集上半监督学习相关算法准确率

算法	出版年	X-Sub/%	X-View/%
CD-JBF-GCN[119]	2022	71.70	78.00
ASSL[33]	2020	64.30	69.80
SESAR-KJS[34]	2020	62.90	-

5　基于无监督学习的骨架动作识别算法

基于监督学习和半监督学习的骨架动作识别算法已经取得了较好的效果，但仍需要有标签的数据．近年来，基于无监督学习的骨架动作识别算法因不依赖任何标签数据而受到了广泛的关注[29-30,120-122]．

为了防止读者混淆产生歧义，本文将自监督学习的算法归为无监督学习. 本文将基于无监督学习的骨架动作识别算法分为 2 个方面: (1) 编码器-解码器网络中各种各样辅助任务的设置; (2) 由于目前对比学习在无监督学习中取得了很好的效果，探索对比学习与骨架动作识别的结合方式.

5.1 辅助任务的设计

各种辅助任务的设计可以为编码器-解码器网络探索出更多的伪监督信号，以丰富骨架数据自身的交互. Kundu 等[121]在编码器中加入对抗生成网络来区分真、假姿态，使得编码器学习到一个连续的嵌入子空间，更好地建模人体动作的动力学特征. Zheng 等[29]用长时动态特征进行无监督骨架动作识别的特征学习，在解码器之前加入样本特征的高斯分布作为输入，用于鉴别器区分原始序列与生成序列，使解码器修复的姿态更加真实; 同时，编码器使用双向门控循环单元，而解码器为单向门控循环单元，通过这种方式来弱化解码器，使编码器提取到更好的特征.

为了使不同类型的样本聚类效果更好、界限更加明确，Su 等[30]提出预测和聚类的无监督骨架动作识别模型，并提出固定权重和固定状态 2 种解码器思路. 其中，固定权重可以弱化解码器，而固定状态可以缓解梯度消失; 文中提出自组织聚类的算法进行聚类，使无监督学习的算法得到进一步发展. 还有一些工作在网络中加入多种辅助任务. Lin 等[36]提出多辅助任务的无监督骨架动作识别模型，将多辅助任务并行输入，解决动作识别网络因过多关注关节点的位置信息而忽略人体动作的时空信息的问题; 设置了姿态预测、帧序排列和对比学习 3 个辅助任务. 其中，姿态预测是常规的思路，帧序排列将帧序分块打乱后进行重新排序来学习时序动态特征，对比学习进一步正则化样本的嵌入空间.

一些传统的辅助任务较为简单，无法精确地提取到动作特征(如骨架序列的重构、预测等). Su 等[120]通过动作的一致性和连续性来学习无监督的 3D 骨架特征，设置了空间动作一致性和时间动作连续性 2 个任务. 其中，空间动作一致性任务通过设置采样动作序列的正片段(改变骨架序列的播放速度，本质上是调整帧采样的时间间隔)和负片段(破坏骨架序列的时序顺序)，在训练过程中使得正片段在嵌入空间中与原始片段靠近，负片段与原始片段远离，学习动作的一致性特征; 时间动作连续性通过对缺失区间帧进行插值和补全来恢复整个动作，使动作看起来更加连贯自然.

已有的无监督学习算法通过原始骨架序列与生成骨架序列之间的损失来优化编码器，这种思路是逐帧的，并没有将人体动作建模为一个整体. Yang 等[122]提出骨架云着色用于无监督的 3D 骨架动作表征学习，将骨架序列叠加到一起表示为 3D 时空骨架云，并根据原始骨架序列中的时间和空间顺序为云中的每个关节点着色; 利用着色的骨架点云设计基于点云的编码器-解码器网络，计算原始骨架云与生成骨架云的距离，优化编码器学习到的人体动作特征.

5.2 对比学习的使用

在基于无监督学习的骨架动作识别算法中，对比学习思路基本上是借鉴一些经典的对比学习算法，如 MoCo[123]，SimCLR[124]. Rao 等[37]提出基于增强骨架和对比动作学习的无监督骨架动作识别网络(augmented skeleton based contrastive action learning, AS-CAL)，借鉴 MoCo[123]对比学习的思路提高同一骨架序列的不同增强实例之间的一致性; 但 AS-CAL 是一种骨架间的对比学习，没有为骨架数据探索出更加丰富的表征方式. Thoker 等[125]提出一种骨架内的对比学习，以交叉对比的方式从多个不同的骨架表征中学习，输入的骨架表征有骨架序列、骨架图和骨架伪图像.

对比学习的问题在于只使用数据增强产生一对正样本，其他相似的样本被视作负样本，但负样本也有可能有很高的相似性，这种设置不利于样本的聚类. Li 等[126]提出针对无监督 3D 骨架动作表征的跨视角对比学习框架，如图 14 所示，其创新点在于将对比语义作为跨视角的一致性知识，通过高置信度知识挖掘来寻找最相似的样本，并用该算法从互补视角挖掘出高置信度的正样本，使特征嵌入在多个视角上保持一致，从而达到跨视角的一致性.

5.3 其他算法

Nie 等[38]在无监督骨架动作识别任务中对骨架的视角进行探索，提出将视角和姿态解耦合的无监督 3D 人体姿态特征学习，首次将人体姿态与视角解耦合，可以在保留固有的姿态信息的同时适应视角的变化. 目前，大多数无监督骨架动作识别算法过于关注对骨架序列时序信息的探索[30,36]，如遮掩帧的重构任务通常可以通过前一帧和后一帧的插值计算得到，往往会忽略对骨架序列空间结构的建模. Cheng 等[127]提出层次 Transformer 解决无监督的骨架动作识别问题，层次 Transformer 的

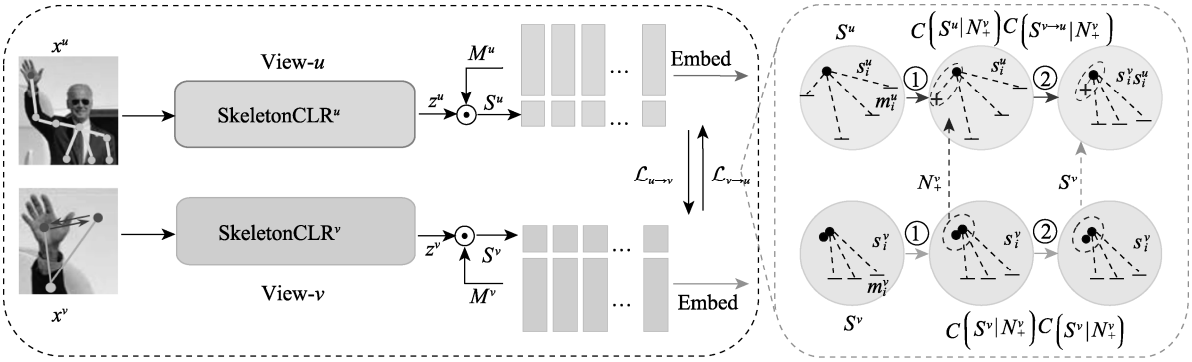


图 14　跨视角对比学习框架[126]

网络结构与 HBRNN[58]类似，通过这种逐层将人体部件进行连接的结构对骨架的空间特征进行建模；该算法在遮掩帧重构的基础上加入动态特征预测(预测骨架中每一个关节点的运动方向)，使得网络可以更好地捕捉骨架序列中的时间特征和空间特征．本文总结了在 NTU-60 数据集上无监督学习相关算法的识别准确率，如表 10 所示．

表 10　NTU-60 数据集上无监督学习相关算法准确率

算法	出版年	X-Sub/%	X-View/%
3s-CrosSCLR[126]	2021	77.80	83.40
Inter-skeleton contrast[125]	2021	76.30	85.20
TS+SS+PS[122]	2021	75.20	83.10
SeBiReNet[38]	2020		79.71
Hierarchical Transformer[127]	2021	69.30	72.80
EnGAN-PoseRNN[121]	2019	68.60	77.80
AS-CAL[37]	2021	58.50	64.80
MS^{2}L[36]	2020	52.55	
P&C FW-AEC[30]	2020	50.70	76.10
LongT GAN[29]	2018	39.10	48.10

6　总结与展望

从研究活跃度来看，目前大部分研究聚焦于监督学习的骨架动作识别方向．在基于监督学习的算法中，基于 GCN 的骨架动作识别算法在识别准确率方面明显优于基于 RNN 和基于 CNN 的算法(如图 15 所示)，这是因为基于 GCN 的算法充分考虑人体的自然结构，并采用人体动作在时间和空间上进行统一建模的策略．虽然基于 GCN 的算法在相关数据集上取得了较高的准确率，但模型的网络结构通常复杂度较高．虽然基于 Transformer 的骨架动作识别算法数量较少，但从图 15 可以看出，其识别准确率与基于 GCN 算法的识别准确率差距较小，原因是基于 Transformer 的骨架动作识别网络完全由注意力机制构成，可以更好地捕捉关节点与关节点之间、骨架序列帧与帧之间的关系．

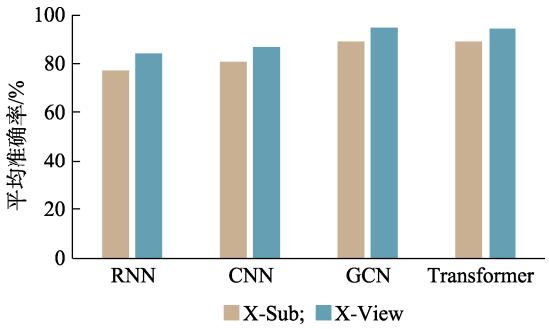


图 15　NTU-60 数据集上 4 种算法的平均准确率对比

图 16 所示为对本文总结的基于监督学习、半监督学习和无监督学习的骨架动作识别算法的数量的统计结果．可以看出，基于监督学习的算法数量(56 个)远远多于基于半监督(4 个)和无监督学习(11 个)的算法数量．

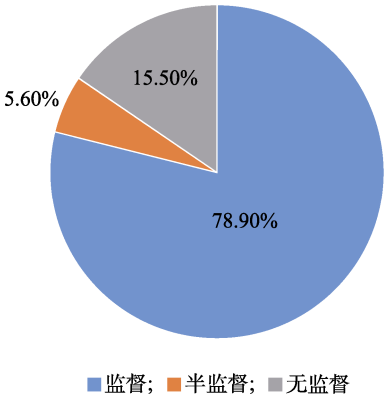


图 16　3 类算法数量统计

图 17 所示为基于监督、半监督和无监督学习的相关算法按出版年在 NTU-60 数据集上获得的平均准确率发展态势统计图．其中，单个算法准确率按 X-Sub, X-View 取平均计算，半监督选取 10%

有标签数据. 可以看出, 在公开数据集上, 基于监督学习的骨架动作识别模型能够取得更高的识别准确率. 尽管如此, 基于监督学习的算法通常需要大量的有标签数据, 数据标注需要耗费大量人力, 而人体骨架的标注难度也更大. 同时, 虽然目前基于半监督、无监督学习的算法与监督学习算法在准确率方面仍有一定的差距(如图 17 所示), 但仍可以较好地缓解基于监督学习的算法对大量有标签数据严重依赖的缺陷. 从发展态势来看, 这种准确率差距正在逐步缩小. 因此本文认为, 基于半监督学习和无监督学习的数据高效的骨架动作识别算法是未来较有发展前景、值得进一步投入的重要发展方向.

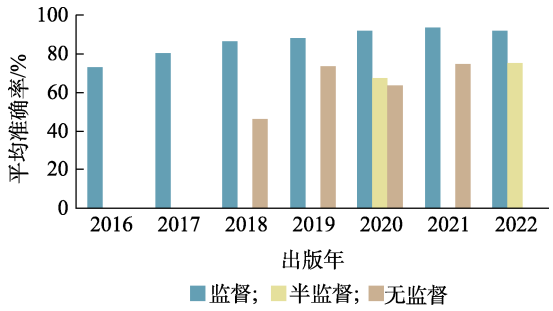


图 17 NTU-60 数据集上算法平均准确率发展态势图

从图 17 可以看出, 虽然基于监督、半监督和无监督的相关算法在相关公开数据集上已经达到了不错的识别效果, 但在实际的社会生产生活中, 当前大部分骨架动作识别依然面临众多挑战. 有 3 个比较突出的问题:

(1) 大规模骨架数据集数量较少, 数据质量需要进一步提高. 首先, 从表 2 可以看出, 目前大规模的骨架数据集只有 NTU 数据集, 骨架数据集的可选择性很小; 其次, 部分数据集由于采集方法年代较为久远、采集设备较为原始等原因, 导致数据集包含的骨架数据准确度不高. 例如, 文献[92]指出, NTU 数据集通过 Kinect 相机收集到的 3D 骨架在部分场景中的关节点精度较低, 已经不及近年来提出的先进人体姿态估计算法[128]得到的 2D 骨架. 公开数据集的数据标注质量决定了在该数据集上训练的算法识别精度上限. 因此, 提出更多大规模、高精度的数据集十分必要.

(2) 模型面向实际应用时鲁棒性不足. 在算法方面, 当前大部分算法只关注在标准数据集上模型的拟合能力与准确率评判, 对模型在现实场景的鲁棒性关注不足; 在数据集方面, 当前大多数骨架数据集的动作类别都是粗粒度的(粗粒度指数据集中的标注的动作类别是动作大类, 如体操、舞蹈, 因此某个动作实际上包含了一些相关度不大的“噪声”动作), 粗粒度动作数据集包含的动作类别模糊性较大, 导致在此类数据集上训练的模型受到干扰动作、背景等噪声的影响, 一旦在真实场景下部署使用, 鲁棒性较差.

(3) 过度依赖大数据、大模型. 目前, 模型对人工标注数据依赖性过高, 而精确的骨架人体动作数据的获取需要专业的硬件设备和软件设置, 十分耗费人力、物力. 同时, 模型的复杂度过高, 常见的计算终端(如手机、智能电视、智能摄像头)的算力不足以支撑大模型的计算, 限制了应用范围.

尽管有诸多挑战, 但随着动作识别相关技术的发展, 上述问题有望缓解. 未来可以从 3 方面进行重点突破:

(1) 高精度骨架数据集建设. 为了克服早期 Kinect 相机收集到的 3D 骨架数据在某些场景下(如遮挡、距离过远)精度不足的问题, 可以考虑进行多视角、多机位高质量 3D 骨架数据的获取. 首先, 在拍摄空间中设置多路(如 6 路)RGB 高速摄像机, 对人体不同角度的图像进行高清晰抓拍; 其次, 利用先进的 2D 人体姿态估计算法[128]计算不同角度人体 2D 关键点; 最后, 基于不同视角的 2D 关键点, 结合相机位姿进行 3D 骨架坐标重构. 为了提高骨架精度、减少背景噪声干扰, 拍摄的场景可以尽可能干净(如使用绿幕). 相关建库流程以及采集的骨架数据示意图如图 18 所示.

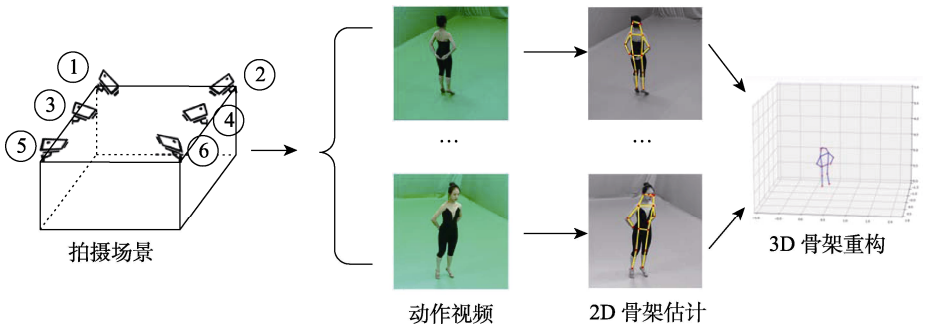


图 18 骨架数据集建库流程

(2) 细粒度骨架动作识别算法及数据集. 要建立细粒度骨架动作数据集. 首先，在设置细粒度的动作类别时，更关注动作的原子性(如跳远动作按时间轴可以分为助跑、起跳和落地 3 个子动作)；其次，采用多层次级别划分的方法对动作的类别进行设定，使得模型能够在动作分类层次的下一级区分其子类. 进一步，研发算法专注于对原子动作的识别，从而减少噪声动作的干扰[129].

(3) 数据有效学习的骨架动作识别. 缓解模型对于大量有标签数据的依赖、提高数据利用率是未来骨架动作识别的一个重要主题. 当前，围绕该主题的方向有：基于无监督学习的骨架动作识别、基于小样本学习的骨架动作识别和基于大规模预训练模型的骨架动作识别等. 其中，基于无监督学习的骨架动作识别通过对基于“编码器-解码器”等范式学习方式的探索，缓解对大量有标签数据的依赖；基于小样本学习的骨架动作识别[130]在仅利用少量有标签样本的前提下，探索如何提高有标签样本的利用率和模型的学习能力；基于大规模预训练模型的骨架动作识别的工作重点在于，将从超大规模数据集训练得到大模型(如 CLIP[131])的知识通过提示学习[132]等方式迁移到下游骨架动作识别子任务，从而较好地完成骨架动作识别任务.

上述 3 个方向都对数据有效学习的骨架动作识别具有较好的参考和指导意义.

7 结　语

本文首先分析骨架数据相比 RGB 数据和深度图数据的优势；然后从监督、半监督和无监督视角，全面、细致地总结了基于骨架的动作识别算法；最后面对该领域目前过度依赖昂贵标注数据、算力消耗过大等挑战，给出下一步的重点发展方向. 今后，动作识别的实际应用将十分广泛，开发出成熟的动作识别算法以及相关衍生产品，让动作识别真正应用到社会生产生活中，是未来共同努力的目标.

参考文献(References):

[1] Zhu Yu, Zhao Jiangkun, Wang Yining, *et al*. A review of human action recognition based on deep learning[J]. Acta Automatica Sinica, 2016, 42(6): 848-857(in Chinese)
(朱煜, 赵江坤, 王逸宁, 等. 基于深度学习的人体行为识别算法综述[J]. 自动化学报, 2016, 42(6): 848-857)

[2] Donahue J, Hendricks L A, Guadarrama S, *et al*. Long-term recurrent convolutional networks for visual recognition and description[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 2625-2634

[3] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 1933-1941

[4] Sharma S, Kiros R, Salakhutdinov R. Action recognition using visual attention[OL]. [2022-03-24]. https://arxiv.org/abs/1511.04119

[5] Tran D, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3D convolutional networks[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 4489-4497

[6] Yang C Y, Xu Y H, Shi J P, *et al*. Temporal pyramid network for action recognition[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 588-597

[7] Yang X D, Zhang C Y, Tian Y L. Recognizing actions using depth motion maps-based histograms of oriented gradients[C] //Proceedings of the 20th ACM International Conference on Multimedia. New York: ACM Press, 2012: 1057-1060

[8] Yang X D, Tian Y L. Super normal vector for activity recognition using depth sequences[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014: 804-811

[9] Chen C, Jafari R, Kehtarnavaz N. Action recognition from depth sequences using depth motion maps-based local binary patterns[C] //Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 1092-1099

[10] Yan S J, Xiong Y J, Lin D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[C] //Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 7444-7452

[11] Li M S, Chen S H, Chen X, *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. Los Alamitos: IEEE Computer Society Press, 2019: 3590-3598

[12] Sanchez-Riera J, Hua K L, Hsiao Y S, *et al*. A comparative study of data fusion for RGB-D based visual recognition[J]. Pattern Recognition Letters, 2016, 73: 1-6

[13] Liu L, Shao L. Learning discriminative representations from RGB-D video data[C] //Proceedings of the 23th International Joint Conference on Artificial Intelligence. New York: ACM Press, 2013: 1489-1496

[14] Shahroudy A, Liu J, 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. Los Alamitos: IEEE Computer Society Press, 2016: 1010-1019

[15] Zhang Z Y. Microsoft kinect sensor and its effect[J]. IEEE MultiMedia, 2012, 19(2): 4-10

[16] Cao Z, Simon T, Wei S E, *et al*. Realtime multi-person 2D pose estimation using part affinity fields[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Los Alamitos: IEEE Computer Society Press, 2017: 1302-1310
[17] Cao Z, Hidalgo G, Simon T, *et al*. OpenPose: realtime multi-person 2D pose estimation using part affinity fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 172-186
[18] Fang H S, Xie S Q, Tai Y W, *et al*. RMPE: regional multi- person pose estimation[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2353-2362
[19] Ren B, Liu M Y, Ding R W, *et al*. A survey on 3D skeleton-based action recognition using learning method[OL]. [2022-03-24]. https://arxiv.org/abs/2002.05907
[20] Zatsiorsky V M. Kinetics of human motion[M]. Champaign: Human Kinetics, 2002
[21] Shotton J, Fitzgibbon A, Cook M, *et al*. Real-time human pose recognition in parts from single depth images[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2011: 1297-1304
[22] Vemulapalli R, Arrate F, Chellappa R. Human action recognition by representing 3D skeletons as points in a lie group[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014: 588-595
[23] Wang J, Liu Z C, Wu Y, *et al*. Mining actionlet ensemble for action recognition with depth cameras[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2012: 1290-1297
[24] Xia L, Chen C C, Aggarwal J K. View invariant human action recognition using histograms of 3D joints[C] //Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2012: 20-27
[25] Wu D, Shao L. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014: 724-731
[26] Elman J L. Finding structure in time[J]. Cognitive Science, 1990, 14(2): 179-211
[27] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780
[28] Albawi S, Mohammed T A, Al-Zawi S. Understanding of a convolutional neural network[C] //Proceedings of the International Conference on Engineering and Technology. Los Alamitos: IEEE Computer Society Press, 2017: 1-6
[29] Zheng N G, Wen J, Liu R S, *et al*. Unsupervised representation learning with long-term dynamics for skeleton based action recognition[C] //Proceedings of the 32th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 2644-2651
[30] Su K, Liu X L, Shlizerman E. PREDICT & CLUSTER: unsupervised skeleton based action recognition[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 9628-9637
[31] Jing L L, Tian Y L. Self-supervised visual feature learning with deep neural networks: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(11): 4037-4058
[32] Chapelle O, Scholkopf B, Zien E. Semi-supervised learning (chapelle, o. *et al*., eds.; 2006)[book reviews][J]. IEEE Transactions on Neural Networks, 2009, 20(3): 542
[33] Si C Y, Nie X C, Wang W, *et al*. Adversarial self-supervised learning for semi-supervised 3D action recognition[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2020: 35-51
[34] Li J Y, Shlizerman E. Sparse semi-supervised action recognition with active learning[OL]. [2022-03-24]. https://arxiv. org/abs/2012.01740
[35] Memmesheimer R, Häring S, Theisen N, *et al*. Skeleton-DML: deep metric learning for skeleton-based one-shot action recognition[C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2022: 837-845
[36] Lin L L, Song S J, Yang W H, *et al*. MS2L: multi-task self-supervised learning for skeleton based action recognition[C] //Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 2490-2498
[37] Rao H C, Xu S H, Hu X P, *et al*. Augmented skeleton based contrastive action learning with momentum LSTM for unsupervised action recognition[J]. Information Sciences, 2021, 569: 90-109
[38] Nie Q, Liu Z W, Liu Y H. Unsupervised 3D human pose representation with viewpoint and pose disentanglement[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2020: 102-118
[39] Wu D, Sharma N, Blumenstein M. Recent advances in video-based human action recognition using deep learning: a review[C] //Proceedings of the International Joint Conference on Neural Networks. Los Alamitos: IEEE Computer Society Press, 2017: 2865-2872
[40] Hu Jianfang, Wang Xionghui, Zheng Weishi, *et al*. RGB-D action recognition: recent advances and future perspectives[J]. Acta Automatica Sinica, 2019, 45(5): 829-840(in Chinese)
(胡建芳, 王熊辉, 郑伟诗, 等. RGB-D 行为识别研究进展及展望[J]. 自动化学报, 2019, 45(5): 829-840)
[41] Huang Qingqing, Zhou Fengyu, Liu Meizhen. Survey of human action recognition algorithms based on video[J]. Application Research of Computers, 2020, 37(11): 3213-3219(in Chinese)
(黄晴晴, 周风余, 刘美珍. 基于视频的人体动作识别算法综述[J]. 计算机应用研究, 2020, 37(11): 3213-3219)
[42] Zhu Y, Li X Y, Liu C H, *et al*. A comprehensive study of deep video action recognition[OL]. [2022-03-24]. https://arxiv.org/abs/2012.06567
[43] Wang L, Huynh D Q, Koniusz P. A comparative review of recent kinect-based action recognition algorithms[J]. IEEE Transactions on Image Processing, 2020, 29(1): 15-28
[44] Xing Y L, Zhu J. Deep learning‐based action recognition with 3D skeleton: a survey[J]. CAAI Transactions on Intelligence Technology, 2021, 6(1): 80-92
[45] Wang J, Nie X H, Xia Y, *et al*. Cross-view action modeling, learning and recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014: 2649-2656

[46] Hu J F, Zheng W S, Lai J H, *et al*. Jointly learning heterogeneous features for RGB-D activity recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 5344-5352

[47] 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

[48] Kay W, Carreira J, Simonyan K, *et al*. The kinetics human action video dataset[OL]. [2022-03-24]. https://arxiv.org/abs/1705. 06950

[49] Li W Q, Zhang Z Y, Liu Z C. Action recognition based on a bag of 3D points[C] //Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2010: 9-14

[50] Sung J, Ponce C, Selman B, *et al*. Human activity detection from RGBD images[C] //Proceedings of the 25th AAAI Conference on Artificial Intelligence Workshops. Palo Alto: AAAI Press, 2011: 47-55

[51] Koppula H S, Gupta R, Saxena A. Learning human activities and object affordances from RGB-D videos[J]. International Journal of Robotics Research, 2013, 32(8): 951-970

[52] Wei P, Zhao Y B, Zheng N N, *et al*. Modeling 4D human-object interactions for event and object recognition[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2013: 3272-3279

[53] Chen C, Jafari R, Kehtarnavaz N. UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor[C] //Proceedings of the IEEE International Conference on Image Processing. Los Alamitos: IEEE Computer Society Press, 2015: 168-172

[54] Rahmani H, Mahmood A, Huynh D, *et al*. Histogram of oriented principal components for cross-view action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(12): 2430-2443

[55] Trivedi N, Thatipelli A, Sarvadevabhatla R K. NTU-X: an enhanced large-scale dataset for improving pose-based recognition of subtle human actions[C] //Proceedings of the 12th Indian Conference on Computer Vision, Graphics and Image Processing. New York: ACM Press, 2021: 1-9

[56] Li S, Li W Q, Cook C, *et al*. Independently recurrent neural network (IndRNN): building a longer and deeper RNN[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 5457-5466

[57] Gao Y B, Li C K, Li S, *et al*. Variable rate independently recurrent neural network (IndRNN) for action recognition[J]. Applied Sciences, 2022, 12(7): 3281

[58] Du Y, Wang W, Wang L. Hierarchical recurrent neural network for skeleton based action recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 1110-1118

[59] Zhu W T, Lan C L, Xing J L, *et al*. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks[C] //Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016: 3697-3703

[60] Baldi P, Sadowski P. Understanding dropout[C] //Proceedings of the 26th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2013: 2814-2822

[61] Liu J, Shahroudy A, Xu D, *et al*. Spatio-temporal LSTM with trust gates for 3D human action recognition[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2016: 816-833

[62] Wang H S, Wang L. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 3633-3642

[63] Song S J, Lan C L, Xing J L, *et al*. An end-to-end spatio-temporal attention model for human action recognition from skeleton data[C] //Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2017: 4263-4270

[64] Liu J, Wang G, Hu P, *et al*. Global context-aware attention LSTM networks for 3D action recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 3671-3680

[65] 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. Los Alamitos: IEEE Computer Society Press, 2017: 1012-1020

[66] Zhang P F, Lan C L, Xing J L, *et al*. View adaptive recurrent neural networks for high performance human action recognition from skeleton data[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2136-2145

[67] Si C Y, Chen W T, Wang W, *et al*. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 1227-1236

[68] Zhao R, Wang K, Su H, *et al*. Bayesian graph convolution LSTM for skeleton based action recognition[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 6881-6891

[69] Li C, Xie C Y, Zhang B C, *et al*. Memory attention networks for skeleton-based action recognition[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(9): 4800-4814

[70] Muhammad K, Mustaqeem, Ullah A, *et al*. Human action recognition using attention based LSTM network with dilated CNN features[J]. Future Generation Computer Systems, 2021, 125: 820-830

[71] He J Y, Wu X, Cheng Z Q, *et al*. DB-LSTM: densely-connected bi-directional LSTM for human action recognition[J]. Neurocomputing, 2021, 444: 319-331

[72] Du Y, Fu Y, Wang L. Skeleton based action recognition with convolutional neural network[C] //Proceedings of the 3th IAPR Asian Conference on Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 579-583

[73] Kim T S, Reiter A. Interpretable 3D human action analysis with

temporal convolutional networks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2017: 1623-1631

[74] Li C, Zhong Q Y, Xie D, *et al*. Skeleton-based action recognition with convolutional neural networks[C] //Proceedings of the IEEE International Conference on Multimedia & Expo Workshops. Los Alamitos: IEEE Computer Society Press, 2017: 597-600

[75] Li C, Zhong Q Y, Xie D, *et al*. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation[OL]. [2022-03-24]. https://arxiv.org/abs/1804.06055

[76] Wang P C, Li Z Y, Hou Y H, *et al*. Action recognition based on joint trajectory maps using convolu tional neural networks[C] //Proceedings of the 24th ACM International Conference on Multimedia. New York: ACM Press, 2016: 102-106

[77] Ke Q H, Bennamoun M, An S J, *et al*. A new representation of skeleton sequences for 3D action recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 4570-4579

[78] Caetano C, Brémond F, Schwartz W R. Skeleton image representation for 3D action recognition based on tree structure and reference joints[C] //Proceedings of the 32nd SIBGRAPI Conference on Graphics, Patterns and Images. Los Alamitos: IEEE Computer Society Press, 2019: 16-23

[79] Le T M, Inoue N, Shinoda K. A fine-to-coarse convolutional neural network for 3D human action recognition[OL]. [2022-03-24]. https://arxiv.org/abs/1805.11790

[80] Yang Z Y, Li Y C, Yang J C, *et al*. Action recognition with spatio–temporal visual attention on skeleton image sequences[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(8): 2405-2415

[81] Liu M Y, Liu H, Chen C. Enhanced skeleton visualization for view invariant human action recognition[J]. Pattern Recognition, 2017, 68: 346-362

[82] Caetano C, Sena J, Brémond F, *et al*. SkeleMotion: a new representation of skeleton joint sequences based on motion information for 3D action recognition[C] //Proceedings of the 16th IEEE International Conference on Advanced Video and Signal Based Surveillance. Los Alamitos: IEEE Computer Society Press, 2019: 1-8

[83] Banerjee A, Singh P K, Sarkar R. Fuzzy integral-based CNN classifier fusion for 3D skeleton action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(6): 2206-2216

[84] Chen C F R, Panda R, Ramakrishnan K, *et al*. Deep analysis of CNN-based spatio-temporal representations for action recognition[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 6161-6171

[85] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 4724-4733

[86] Feichtenhofer C, Fan H Q, Malik J, *et al*. SlowFast networks for video recognition[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 6201-6210

[87] Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 6546-6555

[88] Liu H, Tu J H, Liu M Y. Two-stream 3D convolutional neural network for skeleton-based action recognition[OL]. [2022-03-24]. https://arxiv.org/abs/1705.08106

[89] Ruiz A H, Porzi L, Bulò S R, *et al*. 3D CNNs on distance matrices for human action recognition[C] //Proceedings of the 25th ACM International Conference on Multimedia. New York: ACM Press, 2017: 1087-1095

[90] Lin Z Y, Zhang W, Deng X M, *et al*. Image-based pose representation for action recognition and hand gesture recognition[C] //Proceedings of the 15th IEEE International Conference on Automatic Face and Gesture Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 532-539

[91] Ding W W, Ding C Y, Li G, *et al*. Skeleton-based square grid for human action recognition with 3D convolutional neural network[J]. IEEE Access, 2021, 9: 54078-54089

[92] Duan H D, Zhao Y, Chen K, *et al*. Revisiting skeleton-based action recognition[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 2959-2968

[93] Tang Y S, Tian Y, Lu J W, *et al*. Deep progressive reinforcement learning for skeleton-based action recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 5323-5332

[94] Wen Y H, Gao L, Fu H B, *et al*. Graph CNNs with motif and variable temporal block for skeleton-based action recognition[C] //Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019: 8989-8996

[95] Shi L, Zhang Y F, Cheng J, *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. Los Alamitos: IEEE Computer Society Press, 2019: 12018-12027

[96] Shi L, Zhang Y F, Cheng J, *et al*. Skeleton-based action recognition with directed graph neural networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 7904-7913

[97] Peng W, Hong X P, Chen H Y, *et al*. Learning graph convolutional network for skeleton-based human action recognition by neural searching[C] //Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 2669-2676

[98] Ye F F, Pu S L, Zhong Q Y, *et al*. Dynamic GCN: context- enriched topology learning for skeleton-based action recognition[C] //Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 55-63

[99] Cheng K, Zhang Y F, Cao C Q, *et al*. Decoupling GCN with dropgraph module for skeleton-based action recognition[C] //Proceedings of the European Conference on Computer Vision.

Heidelberg: Springer, 2020: 536-553

[100] Cheng K, Zhang Y F, He X Y, *et al*. Skeleton-based action recognition with shift graph convolutional network[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 180-189

[101] Chen Y X, Zhang Z Q, Yuan C F, *et al*. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 13339-13348

[102] Wu C, Wu X J, Kittler J. Spatial residual layer and dense connection block enhanced spatial temporal graph convolutional network for skeleton-based action recognition[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Los Alamitos: IEEE Computer Society Press, 2019: 1740-1748

[103] Huang Z, Shen X, Tian X M, *et al*. Spatio-temporal inception graph convolutional networks for skeleton-based action recognition[C] //Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 2122-2130

[104] Song Y F, Zhang Z, Shan C F, *et al*. Stronger, faster and more explainable: a graph convolutional baseline for skeleton-based action recognition[C] //Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 1625-1633

[105] Song Y F, Zhang Z, Shan C F, *et al*. Constructing stronger and faster baselines for skeleton-based action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(2): 1474-1488

[106] Liu Z Y, Zhang H W, Chen Z H, *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. Los Alamitos: IEEE Computer Society Press, 2020: 140-149

[107] Li M S, Chen S H, Zhao Y H, *et al*. Dynamic multiscale graph neural networks for 3D skeleton based human motion prediction[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 211-220

[108] Wang M S, Ni B B, Yang X K. Learning multi-view interactional skeleton graph for action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020: 33085614

[109] Zhang P F, Lan C L, Zeng W J, *et al*. Semantics-guided neural networks for efficient skeleton-based human action recognition[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 1109-1118

[110] Zhang X K, Xu C, Tao D C. Context aware graph convolution for skeleton-based action recognition[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 14321-14330

[111] Cai J M, Jiang N J, Han X G, *et al*. JOLO-GCN: mining joint-centered light-weight information for skeleton-based action recognition[C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 2734-2743

[112] Peng W, Shi J G, Xia Z Q, *et al*. Mix dimension in poincaré geometry for 3D skeleton-based action recognition[C] //Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 1432-1440

[113] Xia H L, Gao X K. Multi-scale mixed dense graph convolution network for skeleton-based action recognition[J]. IEEE Access, 2021, 9: 36475-36484

[114] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need[C] //Proceedings of the 31th International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010

[115] 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

[116] Shi F, Lee C H, Qiu L, *et al*. STAR: sparse transformer-based action recognition[OL]. [2022-03-24]. https://arxiv.org/abs/2107.07089

[117] Zhang Y H, Wu B, Li W, *et al*. STST: spatial-temporal specialized transformer for skeleton-based action recognition[C] //Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM Press, 2021: 3229-3237

[118] Qiu H L, Hou B, Ren B, *et al*. Spatio-temporal tuples transformer for skeleton-based action recognition[OL]. [2022-03-24]. https://arxiv.org/abs/2201.02849

[119] Tu Z G, Zhang J X, Li H Y, *et al*. Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition[J]. IEEE Transactions on Multimedia, 2022: 3168137

[120] Su Y K, Lin G S, Wu Q Y. Self-supervised 3D skeleton action representation learning with motion consistency and continuity[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 13308-13318

[121] Kundu J N, Gor M, Uppala P K, *et al*. Unsupervised feature learning of human actions as trajectories in pose embedding manifold[C] //Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 1459-1467

[122] Yang S Y, Liu J, Lu S J, *et al*. Skeleton cloud colorization for unsupervised 3D action representation learning[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 13403-13413

[123] He K M, Fan H Q, Wu Y X, *et al*. Momentum contrast for unsupervised visual representation learning[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 9726-9735

[124] Chen T, Kornblith S, Norouzi M, *et al*. A simple framework for contrastive learning of visual representations[C] //Proceedings of the International Conference on Machine Learning. New York: ACM Press, 2021: 1597-1607

[125] Thoker F M, Doughty H, Snoek C G M. Skeleton-contrastive 3D action representation learning[C] //Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM Press, 2021: 1655-1663

[126] Li L G, Wang M S, Ni B B, *et al*. 3D human action representation learning via cross-view consistency pursuit[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 4739-4748

[127] Cheng Y B, Chen X P, Chen J H, *et al*. Hierarchical transformer: unsupervised representation learning for skeleton-based human action recognition[C] //Proceedings of the IEEE International Conference on Multimedia and Expo. Los Alamitos: IEEE Computer Society Press, 2021: 1-6

[128] Sun K, Xiao B, Liu D, *et al*. Deep high-resolution representation learning for human pose estimation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 5686-5696

[129] Shao D, Zhao Y, Dai B, *et al*. FineGym: a hierarchical video dataset for fine-grained action understanding[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 2613-2622

[130] Thatipelli A, Narayan S, Khan S, *et al*. Spatio-temporal relation modeling for few-shot action recognition[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 19926-19935

[131] Radford A, Kim J W, Hallacy C, *et al*. Learning transferable visual models from natural language supervision[C] //Proceedings of the International Conference on Machine Learning. New York: ACM Press, 2021: 8748-8763

[132] Ju C, Han T D, Zheng K H, *et al*. Prompting visual-language models for efficient video understanding[OL]. [2022-03-24]. https://arxiv.org/abs/2112.04478