计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(8): 2037−2058, 2026

基于人体骨架的视频异常检测方法综述

何 锴[1] 文嘉俊[2,3] 方鸿铭[1] 赖志辉[1,3] 沈琳琳[2,3] 徐 勇[4]
[1]（深圳大学计算机与软件学院 广东深圳 518060）
[2]（深圳大学人工智能学院 广东深圳 518060）
[3]（广东省智能信息处理重点实验室（深圳大学） 广东深圳 518060）
[4]（哈尔滨工业大学（深圳）计算机科学与技术学院 广东深圳 518055）
（2021280111@email.szu.edu.cn）

Survey on Video Anomaly Detection Methods Based on Human Skeleton

He Kai[1], Wen Jiajun[2,3], Fang Hongming[1], Lai Zhihui[1,3], Shen Linlin[2,3], and Xu Yong[4]
[1]（*College of Computer Science and Software Engineering*, *Shenzhen University*, *Shenzhen*, *Guangdong* 518060）
[2]（*School of Artificial Intelligence*, *Shenzhen University*, *Shenzhen*, *Guangdong* 518060）
[3]（*Guangdong Provincial Key Laboratory of Intelligent Information Processing* (*Shenzhen University*), *Shenzhen*, *Guangdong* 518060）
[4]（*School of Computer Science and Technology*, *Harbin Institute of Technology* (*Shenzhen*), *Shenzhen*, *Guangdong* 518055）

Abstract Video anomaly detection is of great practical value in the fields of intelligent video surveillance, public security, etc. With the advancement of deep learning, conventional pixel-based video anomaly detection methods gradually expose their vulnerability to noises, high computation cost, while skeleton-based approaches have emerged as a research hotspot due to their robustness and efficiency. This study provides a detailed explanation of the fundamental concepts, learning paradigms, and methodological workflows of human skeleton-based video anomaly detection methods, and systematically reviews and classifies the representative skeleton-based studies published in the recent five years into different categories, including prediction-based method, reconstruction-based method, hybrid reconstruction and prediction-based method, hybrid reconstruction and clustering-based method, as well as unclassified method. An in depth analysis is given to explore the underlying principles and innovations of each type of method. Moreover, this study summarizes existing benchmark datasets and evaluation metrics for skeleton-based video anomaly detection, as well as gives a comparative performance analysis of the mainstream methods across the benchmarks. Currently, skeleton-based anomaly detection methods face significant challenges in skeleton feature extraction, modeling and understanding of multi-person interaction, etc. Therefore, this study further proposes scientifically grounded future directions with regard to robust skeleton feature extraction algorithm, multimodal information integration and heterogeneous data fusion, etc., aiming at exploring the positive factors for constructing robust video anomaly detection framework, as well as analyzing the existing problems and applicable research directions of these series of video anomaly detection methods in different aspects, including adaptation ability,

收稿日期：2025-09-21；**修回日期**：2026-04-24
基金项目：国家自然科学基金项目（62476175）；广东省科技厅资助项目（2024A1515011816）；广东省智能信息处理重点实验室项目（2023B1212060076）；广东省自然科学基金项目（2026A1515011189, 2024A1515011637）；深圳市科技计划项目（JCYJ20240813142206009）；深圳市俊唐晨启科技有限公司研究项目（JTC2026R001）
This work was supported by the National Natural Science Foundation of China (62476175), the Foundation of the Department of Science and Technology of Guangdong Province (2024A1515011816), the Foundation of the Guangdong Provincial Key Laboratory (2023B1212060076), the Natural Science Foundation of Guangdong Province (2026A1515011189, 2024A1515011637), the Shenzhen Science and Technology Program (JCYJ20240813142206009), and the Foundation of the Shenzhen JTC Aurora Technology Co., Ltd. (JTC2026R001).
通信作者：文嘉俊（wenjiajun@szu.edu.cn）

generalization ability, running efficiency, etc.

Key words video anomaly detection; deep learning; human skeleton analysis; pattern recognition; spatio-temporal feature extraction

摘 要 视频异常检测在智能监控、公共安全等领域具有重要应用价值。随着深度学习的发展,传统基于像素的视频异常检测方法逐渐暴露出对噪声敏感、计算代价高的局限性,而基于人体骨架的方法因其鲁棒性和高效性成为了研究热点。详细阐述基于人体骨架的视频异常检测方法的基本概念、学习范式和方法流程,系统地回顾近5年基于人体骨架的视频异常检测代表性工作并把它们划分为基于预测、基于重构、基于重构和预测相结合、基于重构和聚类相结合以及其他方法,并对每种方法的原理和创新点进行深入分析。此外,总结了现有基于人体骨架的视频异常检测基准数据集与评估指标,综合评估了主流方法在基准数据集上的性能表现,并进行对比分析。当前基于人体骨架的视频异常检测方法在骨架特征提取、多人体交互建模与理解等方面仍面临着重大挑战,对此进一步从鲁棒骨架特征提取算法、多模态信息集成与异构数据融合等方面提出科学展望,旨于探索构建鲁棒视频异常检测框架的积极因素,分析该系列方法在适应性、泛化性、实时性等方面存在的问题及其可行研究方向。

关键词 视频异常检测;深度学习;人体骨架分析;模式识别;时空特征提取

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202550684 **CSTR:** 32373.14.issn1000-1239.202550684

异常数据模式是模式识别的重要研究对象。而异常检测指的是识别与常规模式或预期模式显著不同的数据或事件[1]。在视频分析领域,异常通常是指视频片段中那些不符合正常行为或活动的事件[1-2],如视频中发生的安全事件(打斗、偷窃等)、突发的环境变化(火灾、溢水等),甚至是不合规的行为或非预期活动[1-2]。监控视频中与人活动相关的视频段及其分析最具有信息量及研究价值。对于现实生活中的监控视频,人们最感兴趣的异常绝大部分为视频中人员行为的异常。近年来,随着智能监控、公共安全等领域对行为分析需求的日益增加,与人活动相关的视频异常检测方法逐步崭露头角,具有广泛应用前景[3]。

传统视频异常检测方法主要依赖人工判决。然而,海量视频数据给人工分析带来巨大挑战。通过人工回看监控录像来发现异常的方式,不仅消耗大量人力资源,而且还会因主观因素导致漏检或出错。因此,开发一种不需要依赖大量人力且能够从监控视频中自动分析异常事件的技术显得尤为重要[4]。

随着深度学习的发展,研究人员开始引入深度神经网络模型实现视频异常检测。视频异常检测旨在从大量视频数据中自动识别出不符合常规模式的行为或事件[5]。大多数传统研究以基于像素的视频异常检测为主,旨于通过提取视频帧的图像特征,结合背景信息,并融合光流、颜色、纹理等特征进行建模,从而识别正常模式帧和异常模式帧,实现异常事件的检测[6]。相比基于人工判决的视频异常事件检测,基于像素的视频异常检测大大提高了检测效率[4]。

然而,研究人员也指出基于像素的视频异常检测方法的局限性:1)对噪声和背景变化敏感[7]。像素级的异常检测方法容易受到场景中噪声、光照变化、背景杂波和人群遮挡的影响,从而导致误检。2)计算成本高[8]。例如使用光流建模处理大量视频数据需要较大的计算资源,且训练过程复杂。3)缺乏语义理解[9]。像素级分析主要关注视频的视觉特征,无法理解视频中复杂的语义信息,如人员行为、运动特征及人与人之间的交互。4)公平性与隐私保护问题[2,10]。现实中视频异常检测数据大多来源于公共场所的监控摄像头。这些数据包含不同性别、年龄和种族的人群。像素级方法主要处理目标的视觉特征,因此可能存在对某些个体或群体的不合理分析,误将其行为标记为异常。此外,视频数据可能包含个人面部特征、行为模式和位置等敏感信息。基于像素的方法通过提取视频中的人物图像来检测异常事件,存在侵犯个人隐私的隐患。

为解决以上问题,研究人员提出基于人体骨架的视频异常检测方法。这类方法剔除了无关的背景干扰以及与人体行为无关的异常,仅保留了人体骨架,并通过捕捉人体骨架关键点,如头部、四肢、躯干等位置的空间与时间变化,利用人体动作和姿势的变化来分析和识别异常行为[7]。基于人体骨架的视频异常检测方法有以下优势:1)与基于像素的异

常检测方法不同，基于骨架的方法能够从更高层次行为语义上对视频内容进行理解，有效避免光照变化和背景干扰等因素的影响，具有更强的鲁棒性。2）基于骨架的异常检测能够高效提取更简洁的特征。稀疏的骨架表示大大减少特征分析的数据量，有效降低计算复杂度，使得检测过程更加高效。3）骨架特征具有紧凑、结构良好和语义丰富的特点，能够有效描述人员行为和运动，非常适用于复杂行为的异常检测问题，同时可实现对个人隐私的有效保护[10]。

人体骨架作为一种高效且鲁棒的特征表示方法，已成为面向人员活动分析的视频异常检测研究热点。本文旨于系统地回顾与总结基于人体骨架的视频异常检测方法，探讨该领域的研究现状、关键技术和未来发展趋势。本文贡献主要体现在3个方面：

1）从基本概念、学习范式和方法流程多个方面系统阐述基于人体骨架的视频异常检测的基本内涵，并对基于人体骨架的视频异常检测方法进行分类，系统介绍各类方法的基本思想、代表性工作创新点及其优劣势。

2）全面总结现有基于人体骨架的视频异常检测数据集与评估指标，综合评估主流方法在标准数据集上的性能表现，并对比分析。

3）进一步探讨基于人体骨架的视频异常检测方法的局限性与应用场景限制，并从鲁棒的骨架特征提取算法、多模态信息集成与异构数据融合等5个方面提出未来展望。

1 基于人体骨架的视频异常检测概述

根据文献[1−4]，基于人体骨架的视频异常检测流程如图1所示。具体执行步骤如下：

1）把模型训练和验证所需的视频数据按帧提取[6]，并对每一帧采用姿态估计算法提取骨架信息，并获得骨架运动轨迹[11-13]。

2）对获得的骨架信息实行建模预处理[7]，然后提取特征[14]，为模型对异常事件的学习与识别提供可靠的特征表示。

3）所提取的骨架特征用于模型训练，构建异常检测学习模型[15]，为异常模式识别做好准备。

4）在测试推理阶段，把测试样本特征输入已训练完成的模型。模型通过对视频中时空特征的分析，判断视频是否包含异常事件。此过程涉及异常分数的计算以及异常检测阈值等参数的设定[10]。

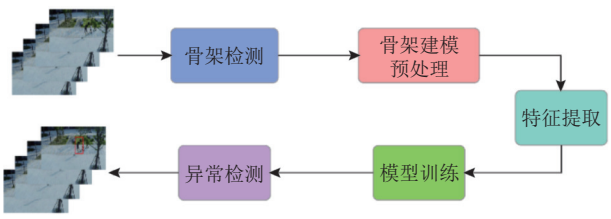


Fig. 1 The general process of video anomaly detection based on human skeleton

图1 基于人体骨架的视频异常检测一般流程

1.1 异常模式基本类型

根据视频中异常模式的规模大小和行为语义特征，现有研究[1,3]把视频存在的异常模式划分为局部异常和全局异常2种基本类型。

1）局部异常

局部异常是指在视频的某一时刻或短时间内，人体骨架中某些局部区域的异常行为[1]。这类异常通常表现为个体或小范围的动作异常，是由个别或少数人员行为不符合常规的行为模式或预期所产生[4]。例如，个体出现极不自然的动作、错误的姿势或行为模式，导致其与周围环境及其他人的行为显著不同，典型案例为人走路时突然摔倒，或在图书馆里乱舞等。

2）全局异常

全局异常指的是视频整体范围内，人员行为或整个场景的动态变化与常规行为或模式显著不同[1]。全局参考指标既包括单个个体的整体速度和位移，也涉及多个人体骨架之间的交互与关联程度[4]。全局异常通常表现在多个区域或多个时刻的不正常行为，如群体中不协调的动作、大规模异常事件等[14]。单个个体的典型案例如1个学生在教室奔跑；多个个体的典型案，例如在公共场所中突然出现打斗或骚乱、奔逃等群体性事件。

1.2 异常检测学习范式

人体骨架及其运动为视频中人员行为的分析提供了最简约化描述，指代了视频中最重要的感兴趣时序性中层语义。根据现有研究，基于人体骨架的视频异常检测学习范式主要分为4种方法[1-5,10]。

1）基于人体骨架的监督型视频异常检测方法

基于人体骨架的监督型视频异常检测方法（简称有监督学习）是一种基于数据标注的学习方法。监督模型通过学习一组标注为“正常”或“异常”的训练数据，建立输入数据与输出类别的映射关系[1]。在基于人体骨架的视频异常检测中，有监督学习的目标是通过学习包含正常和异常标签的视频数据，训练出能够区分正常行为与异常行为的识别模型。例

如，研究人员通过使用标注过的人体骨架数据对模型进行训练，可区分正常行走行为和异常摔倒行为。有监督学习方法的优势在于其能够直接利用标注数据进行学习，从而在训练过程中确立正常与异常行为的特征空间，使模型在已知异常类型的情况下具有较高的检测精度[4]。然而，由于异常事件的稀缺性和多样性，收集异常视频数据并进行标注耗时长、成本高[7]。此外，有监督学习方法对已学习的异常行为具有较强依赖，对未标注的新型异常行为，模型可能无法准确识别，导致其泛化能力低，从而限制其在实际中的应用。因此，有监督学习方法在视频异常检测中较少被采用。

2)基于人体骨架的无监督型视频异常检测方法

基于人体骨架的无监督型视频异常检测方法(简称无监督学习)并不依赖标注数据，而是通过分析未标注数据，识别出其中的潜在结构和模式[1]。在基于人体骨架的视频异常检测中，无监督学习主要通过分析视频中的行为模式，发现与常规行为模式显著不同的异常事件[4]。无监督学习的最大优势是无需标注数据，适用于标注成本高或数据缺乏的场景。因此无监督学习逐渐成为视频异常检测的主流学习范式，典型方法包括基于聚类、自编码器等的视频异常检测方法[10]。由于缺乏标注信息，采用无监督学习的模型需要从大量未标注数据中学习正常行为模式。如果数据中包含噪声或异常样本，模型有可能会将这些噪声或异常行为误认为是正常模式的一部分，从而导致检测性能下降。因此，为保证无监督学习方法的性能，研究人员需要为模型训练提供质量较好的数据[5]。

3)基于人体骨架的弱监督型视频异常检测方法

基于人体骨架的弱监督型视频异常检测方法(简称弱监督学习)是介于有监督与无监督学习之间的学习范式，利用少量标注信息(如不完全或不精确标签)进行训练[4]。在基于人体骨架的视频异常检测中，弱监督学习不要求对每个视频帧或每个人体骨架进行标注，而是利用视频的部分标注信息指导模型学习。例如，可通过标注少数包含异常行为的视频来训练模型，从而识别其他视频中的异常行为。弱监督学习减少了模型对大量标注数据的依赖，显著降低数据标注的工作量，并在不完全标注的情况下提高模型的检测性能[1]。然而，标注信息的稀疏性会影响模型训练效果，导致模型对异常行为识别不够准确。因此，需要为弱监督学习方法设计更有效的推断与学习机制[16]，以确保模型能够充分利用有限的标注信息，防止未标注数据在模型训练过程中产生误导。

4)基于人体骨架的自监督型视频异常检测方法

基于人体骨架的自监督型视频异常检测方法(简称自监督学习)本质上属于无监督学习的1个分支，是一种通过数据本身生成监督信号的学习方法。自监督学习模型无需外部标签，而是通过挖掘隐含在数据中的内在结构进行学习[4]。在基于人体骨架的视频异常检测中，自监督学习方法通过设计合适的预设任务(简称预任务)，使模型能够自我生成训练信号，进而学习数据中的异常模式。例如，在没有标签的视频数据中，模型通过时间序列预测或对比学习，可获得视频帧的时序关系，进而对异常运动或姿势进行检测[17]。自监督学习的设计和训练通常更加复杂，且对数据质量要求较高。如果数据中存在噪声或异常样本，模型会在学习过程中受到干扰，导致其对正常行为模式的学习效果下降。此外，自监督学习方法在确定预任务和生成监督信号时，需要充分考虑数据的特性和异常检测任务的需求，否则会导致模型学习到无效或错误的特征，进而影响异常检测的性能[1]。因此，研究人员需要精心设计合适的自监督任务，确保模型能够从数据中学习到有效的特征，并且能够准确识别异常行为。

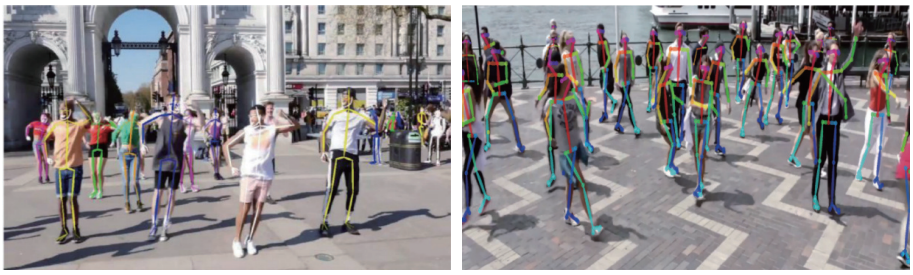
通过对以上4种学习范式的分析可以看出，每种学习范式都有其独特的优势和局限性。在实际视频异常检测任务中，选择合适的学习范式需要综合考虑数据的标注情况、异常行为的类型和特点以及模型的检测精度和泛化能力等诸多因素。

2 基于人体骨架的视频异常检测方法

本节系统地阐述基于人体骨架的视频异常检测方法流程，包括骨架检测、骨架建模预处理和异常检测方法三大模块。

2.1 骨架检测

实现基于人体骨架的视频异常检测，首先要获取视频帧中每个个体的骨架信息，即骨架检测。骨架检测是通过姿态估计算法实现。姿态估计的主要任务是从图像或视频中提取出人体各个关键点的位置，包括头部、手臂、腿部等关键点信息。根据现有文献，AlphaPose[11]，OpenPose[12]，HRNet[13]是当前应用最广泛的3种姿态估计算法。图2展示了AlphaPose[11]和OpenPose[12]在不同数据集上的可视化输出。此外，近年来以扩散模型为主的SkeletonDiffusion[18-19]方法



(a) AlphaPose (b) OpenPose

Fig. 2 Visualization outputs of AlphaPose[11] and OpenPose[12] on different datasets

图 2 AlphaPose[11] 和 OpenPose[12] 在不同数据集上的可视化输出

和以 Transformer 方法为主的 PoseFormer[20-22] 也逐渐引起更多的关注，下面将详细分析姿态估计算法的性能表现及优缺点。

1）AlphaPose

AlphaPose[11] 提出了对称积分关键点回归（symmetric integral keypoints regression, SIKR）方法，能够精确且快速地对关键点进行定位，解决传统热图方法中存在的量化误差问题。此外，AlphaPose 还引入参数化姿态非极大值抑制（non-maximum suppression, NMS）来消除冗余的人体检测结果，并通过姿态感知身份嵌入实现姿态估计与跟踪的联合优化。

然而，AlphaPose 也存在一些缺点[11]。首先，计算复杂度相对较高，尤其是在处理大规模数据集时，其训练和推理时间显著增加。其次，AlphaPose 的姿态感知身份嵌入模块会受到背景噪声影响，导致身份识别的准确性下降。此外，AlphaPose 依赖于外部人体检测器，如果检测器性能不佳，会导致 AlphaPose 姿态估计失败。

2）OpenPose

OpenPose[12] 是一种非参数化基于部分亲和场（part affinity fields, PAFs）学习身体部位与个体间的关联，并迭代预测 PAFs 和关键点置信度图，结合贪婪算法实现高效的多人姿态估计。它在 MPII 和 COCO 等数据集上表现出色，尤其擅长处理多人场景，并且扩展了脚部关键点检测功能，进一步提升了姿态估计的完整性和准确性[12]。

尽管 OpenPose 在多人姿态估计方面表现优异，但其 PAFs 方法在处理人体部位重叠或严重遮挡时可能出现关联错误。此外，在处理高分辨率图像时计算资源消耗较大。同时，OpenPose 的训练需要大量标注数据，增加了数据准备的难度和成本。

3）HRNet

HRNet[13] 通过并行连接多分辨率子网络，持续保持高分辨率特征表示，并运用多尺度融合来生成丰富的特征，从而提升姿态估计的准确性。它在 COCO 和 MPII 等数据集上表现出色，尤其在关键点相似度计算和检测准确性方面有显著提升。

然而，HRNet 也存在不足之处[13]。首先，HRNet 网络结构较为复杂，计算复杂度较高。其次，HRNet 在处理多人场景时，会受到人体检测器性能制约。此外，HRNet 多尺度融合机制会引入一定噪声，导致关键点定位的精度下降。

4）SkeletonDiffusion

SkeletonDiffusion[18-19] 是将扩散模型应用于骨骼关键点估计的技术框架。其核心思想是通过逐步去噪过程生成或优化人体骨骼关键点坐标，利用扩散模型的概率建模能力学习姿态分布规律。通过学习大量姿态数据的分布，SkeletonDiffusion 可从噪声中逐步生成符合约束的骨骼结构，缓解传统方法在遮挡、复杂姿态下的关键点定位模糊问题。

然而，由于扩散过程通常需要多步迭代[18-19]，SkeletonDiffusion 会面临计算成本高、实时性差的问题，并难以满足低延迟场景需求。此外，SkeletonDiffusion 对训练数据质量和数量有较高要求。如果缺乏大规模高质量标注的姿态数据，该模型可能生成不符合人体物理约束的“伪姿态”。

5）PoseFormer

PoseFormer[20-22] 是基于 Transformer 架构的 3 维人体姿态估计网络，其核心思想是利用 ViT（vision Transformer）的自注意力机制来捕捉姿态数据中的长距离空间和时间依赖关系，以克服传统卷积网络感受野受限问题。这类方法通常采用“2 维提升至 3 维”的策略，即首先使用 2 维姿态估计算法获取图像中的 2 维关键点，然后将这些关键点序列作为 Transformer 模型输入来预测对应 3 维姿态。

然而 PoseFormer 的这种 2 维提升至 3 维的范式，意味着它需要一个高质量的 2 维姿态估计器来保证输入的稳定性[20-22]。如果前置 2 维姿态估计器在复

杂场景(如遮挡、模糊、极端视角、光照不足等)中产生不准确的2维关键点，那么PoseFormer的3维预测也会直接受到影响，不断累积错误。

综上所述，作为当前主流的姿态估计算法，AlphaPose[11]，OpenPose[12]，HRNet[13]，SkeletonDiffusion[18-19]，PoseFormer[20-22]各有其独特的性能表现和优缺点。AlphaPose在全身姿态估计和跟踪方面表现出色，尤其在处理复杂场景和多人交互时具有优势，但其计算复杂度较高且对背景噪声敏感。OpenPose在多人2维姿态估计方面具有较高的准确性和实时性，适合实时应用，但在复杂场景中会消耗较大资源且会出现关联失误。HRNet则在高分辨率特征表示和关键点检测精度方面表现出色，适用于高精度姿态估计场景，但其网络结构复杂且计算复杂度较高。尽管SkeletonDiffusion通过扩散模型提升了复杂姿态和遮挡下的关键点定位能力，其计算成本高、实时性差，且要求大量高质量训练数据。PoseFormer作为Transformer的3维姿态估计网络，虽然擅长捕捉长距离时空依赖关系，但是其性能高度依赖前置2维估计，且计算成本和数据需求较大。在选择姿态估计算法时，应根据具体的应用场景和需求进行权衡，以获得最优的骨架检测效果。

2.2 骨架建模预处理

运动的骨架不仅包含丰富的时空信息，而且可有效表示个体的运动模式和行为特征。把骨架信息输入异常检测模型之前，需要对其进行建模预处理，从而让模型更好地学习骨架特征。现有基于骨架的异常检测方法，通常将骨架建模预处理分为2类：基于全局和局部组件分解的预处理[7]以及基于层级图的骨架建模预处理[14]。

2.2.1 基于全局和局部组件分解的预处理

基于全局和局部组件分解的预处理方法将骨架运动分解为全局身体运动和局部姿态2个子过程，从而实现对正常行为模式的有效建模。该方法允许模型捕捉全局运动(整个身体在场景中的动态变化)和局部姿态(骨架各个关节部位的动作细节)。

设$f_t=\{\boldsymbol{f}_t^i\}=\left\{(x_t^i,y_t^i)^{\mathrm{T}}_{i=1,2,\cdots,k}\right\}$($k$为骨架关节数目)为时刻$t$骨架关节位置坐标向量集合。根据2维向量分解定理[7]，得出：

$$\boldsymbol{f}_t^i=\boldsymbol{f}_t^{\mathrm{g}}+\boldsymbol{f}_t^{\mathrm{l},i}, \tag{1}$$

其中$\boldsymbol{f}_t^{\mathrm{g}}$表示时刻$t$骨架的全局组件信息，代表当前骨架中心在原始图像帧内的绝对位置。$\boldsymbol{f}_t^{\mathrm{l},i}$表示时刻$t$骨架的局部组件信息，其值为骨架各个关节的绝对坐标向量减去全局分量后的剩余部分，代表骨架关节相对于骨架中心的2维坐标向量，如图3所示。

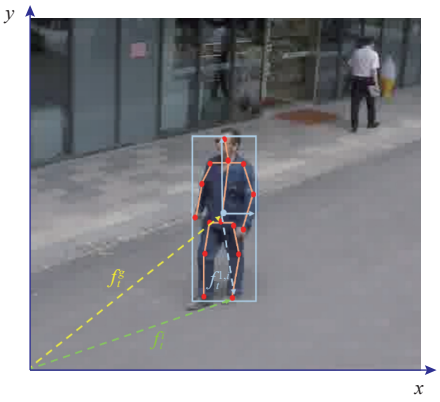


Fig. 3 Schematic diagram of global and local components decomposition method

图3 全局和局部组件分解法示意图

由于图像仅包含2维信息，缺少深度信息，单独的x和y坐标难以代表场景中的空间位置。为了弥补这个差距，研究人员根据骨架边界框大小与骨架位置在场景中的深度关系，联合骨架边界框中心坐标、宽度和高度所形成四元组向量$\boldsymbol{f}^{\mathrm{g}}=(x^{\mathrm{g}},y^{\mathrm{g}},w,h)^{\mathrm{T}}$来增强全局组件，并通过归一化获得局部组件$\boldsymbol{f}^{\mathrm{l},i}=(x^{\mathrm{l},i},y^{\mathrm{l},i})^{\mathrm{T}}$，具体步骤如下。

1)第1步。利用关节位置坐标集合计算骨架的平均中心坐标，即骨架的全局坐标[7]：

$$x^{\mathrm{g}}=\frac{\max(x^i)+\min(x^i)}{2},\quad y^{\mathrm{g}}=\frac{\max(y^i)+\min(y^i)}{2}, \tag{2}$$

其中x^{g}和y^{g}分别为骨架平均中心的横坐标和纵坐标，$(x^{\mathrm{g}},y^{\mathrm{g}})$为骨架的全局坐标。

2)第2步。利用关节位置坐标集合计算出骨架边界框的宽和高[7]：

$$w=\max(x^i)-\min(x^i),\quad h=\max(y^i)-\min(y^i), \tag{3}$$

其中w和h是骨架边界框的宽和高。

3)第3步。利用第2步求出的w和h计算出归一化的骨架局部组件，得到式(4)[7]。

$$x^{\mathrm{l},i}=\frac{x^i-x^{\mathrm{g}}}{w},\quad y^{\mathrm{l},i}=\frac{y^i-y^{\mathrm{g}}}{h}, \tag{4}$$

其中$x^{\mathrm{l},i}$和$y^{\mathrm{l},i}$分别为骨架关节点相对骨架中心的归一化横坐标和纵坐标，$\boldsymbol{f}^{\mathrm{l},i}=(x^{\mathrm{l},i},y^{\mathrm{l},i})^{\mathrm{T}}$为骨架的归一化局部组件。

基于全局和局部组件分解方法的优势在于该模型能显式地利用人体运动的自然分解形式，从而在复杂场景下获得较高的异常识别精度和较好的鲁棒性[7]。通过分解全局和局部组件，能够有效分离背景运动和个体姿态变化，减少背景噪声和光照变化对

异常检测的影响[7]，此外，该方法采用了结构化和富含语义的骨架特征，模型可解释性强。在计算效率方面，局部组件的归一化处理对特征表示进行了简化，有效降低模型计算复杂度，提高检测效率。然而，基于全局和局部组件分解的方法仍存在局限性：1）局部信息丢失[10]。全局组件主要关注整体运动，会丢失一些局部姿态的细节信息。而使用单纯的局部组件会忽略全局运动的上下文信息。2）适用场景有限。在多人交互场景下，仅依赖全局和局部组件分解难以捕捉个体间的交互关系[8]，导致模型对复杂异常行为的检测能力不足，模型性能在人体遮挡严重时会大大下降。因此，基于全局和局部组件分解的方法适用于单人或少量个体的场景。尤其是当背景变化较大或光照条件复杂时，该方法能够有效提取个体的运动特征。对于简单的异常行为（如个体突然摔倒、快速奔跑等），基于全局和局部组件分解的方法能够提供较为准确的检测结果。

2.2.2 基于层级图的骨架建模预处理

基于层级图骨架建模预处理是通过构建多级图表示来解决视频异常检测问题，一般包括高阶图表示和低阶图表示，前者用于编码个体间的交互，后者用于编码个体局部姿态。假设1个骨架的所有关节点数量为N，那么原始骨架坐标集合可表示为$f_n=\{(x_n,y_n)\}$，$n\in\{1,2,\cdots,N\}$。基于层级图骨架建模一般需要骨架的2维全局坐标和局部坐标信息。但是对于骨架全局坐标的计算，无需选取当前骨架关节的所有点，仅需选取骨架的4个核心关节点然后求平均即可得到全局坐标，如图4和式(5)[14]所示。

$$\hat{x}^{\mathrm{g}}=\frac{1}{4}\sum_{n\in k}x_n,\quad \hat{y}^{\mathrm{g}}=\frac{1}{4}\sum_{n\in k}y_n,\quad k\in\{5,6,11,12\}\ , \tag{5}$$

其中$\hat{x}^{\mathrm{g}}$和$\hat{y}^{\mathrm{g}}$分别为当前骨架全局坐标下的横坐标和纵坐标。由此可得当前骨架的全局坐标为$\hat{f}^{\mathrm{g}}=(\hat{x}^{\mathrm{g}},\hat{y}^{\mathrm{g}})$。

当前骨架的局部坐标是利用骨架边界框高度对原始骨架的每个关节坐标归一化得到，见式(6)[14]。

$$\hat{x}_n^{\mathrm{l}}=\frac{x_n}{h/2},\quad \hat{y}_n^{\mathrm{l}}=\frac{y_n}{h},\quad n\in\{1,2,\cdots,N\}\ , \tag{6}$$

其中h是骨架的边界框高度，$\hat{x}_n^{\mathrm{l}}$和$\hat{y}_n^{\mathrm{l}}$分别为骨架局部坐标下的横坐标和纵坐标，那么骨架所有点的局部坐标集合为$\hat{f}_n^{\mathrm{l}}=\{(\hat{x}_n^{\mathrm{l}},\hat{y}_n^{\mathrm{l}})\}$，$n\in\{1,2,\cdots,N\}$。

接下来，定义高阶图和低阶图表示。高阶图由所有骨架的全局坐标节点组成。每个节点代表1个个体（骨架），用于编码个体间的相对位置和个体间的交互。低阶图表示每个个体的骨架姿态细节。每个节点代表所属骨架的关节，用于编码个体的局部

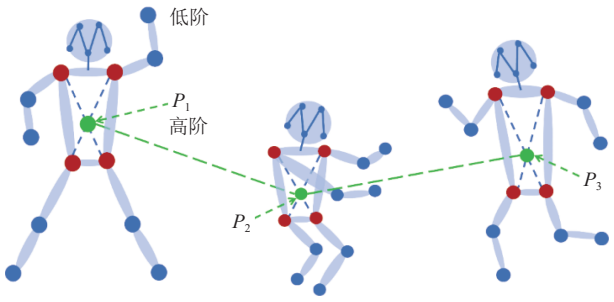


Fig. 4 Schematic diagram of hierarchical skeleton modeling

图4 层级图骨架建模示意图

行为。对于时刻t，高阶图和低阶图的通用定义为[14]

$$G_t=(V(G_t),E(G_t))\ , \tag{7}$$

其中$V(G_t)$代表图G_t的节点坐标集合，$E(G_t)$代表图G_t的边集合。若图中任意2个节点i和j相连，那么$e_t^{i,j}=1$，否则$e_t^{i,j}=0$。此外，高阶图和低阶图表示还引入邻接矩阵$\boldsymbol{A}_t$，其中每个元素$a_t^{i,j}$按式(8)[14]计算。

$$a_t^{i,j}=\begin{cases}\dfrac{1}{\left(V_i(G_t)-V_j(G_t)\right)^2}, & \left(V_i(G_t)-V_j(G_t)\right)^2\neq 0\ ;\\ 0, & \left(V_i(G_t)-V_j(G_t)\right)^2=0\ 。\end{cases} \tag{8}$$

式(8)采用节点间的欧氏距离平方的倒数表示邻接矩阵，以反映节点间的相对位置。这种设计方式相当于一种基于特征差异的连接权重分配。它强调特征差异较大的节点之间的连接性。在异常检测的上下文中，这种基于特征的邻接矩阵能够有效帮助模型识别出与正常行为模式显著不同的异常行为[23]。最后，邻接矩阵$\boldsymbol{A}_t$的每个元素$a_t^{i,j}$乘以边权重$e_t^{i,j}$，得到最终的高阶图和低阶图的邻接矩阵通用表示，如式(9)[23]所示。

$$A_t^{i,j}=a_t^{i,j}\otimes e_t^{i,j}\ , \tag{9}$$

其中$\otimes$表示矩阵元素对应位置相乘操作，$A_t^{i,j}$表示边权重$e_t^{i,j}$作用下的结果。

基于层级图骨架建模方法的优点是能够同时捕捉个体内和个体间的空间关系[14]，加强模型对行为的理解能力和模型的鲁棒性。这种优势在模型处理群体异常事件时尤为突出。高阶图和低阶图的构建，能够使模型同时捕捉个体局部行为和个体间的交互关系，提高模型对复杂异常行为的检测能力[14]。此外，层级图具备一定的灵活性，可以根据不同场景调整图的层次和节点连接方式，以适应各种复杂场景[24]。然而，基于层级图骨架建模的方法存在2个方面不足：1）计算复杂度高。构建和处理多级图需要较高的计算资源。在处理大规模数据集时，计算效率会

成为瓶颈。2)检测性能依赖骨架提取精度。基于层级图骨架建模方法对骨架提取的精度要求较高。如果骨架提取存在误差,会影响图的构建以及后续的异常检测[10]。因此,基于层级图的骨架建模方法适用于多人交互场景,尤其是异常行为涉及个体间的复杂交互[23],如群体打斗、追逐等。对于场景复杂、个体数量较多的视频数据,该方法能够有效捕捉个体间的交互关系,提高异常检测准确性。

综上,基于全局和局部组件分解的方法与基于层级图的骨架建模方法各具优势。在实际应用中应根据具体场景特点进行方法选择或方法组合,以实现更准确和鲁棒的视频异常检测。

2.3 基于人体骨架的异常检测方法

本节回顾近5年基于人体骨架的异常检测代表性工作。根据对异常建模方式的不同,异常检测方法可划分为基于预测、基于重构、基于重构和预测相结合、基于重构和聚类相结合以及其他方法。

2.3.1 基于预测的方法

基于预测的方法是目前基于人体骨架的异常检测最为常用的方法。在有标注视频骨架数据的情况下(比如有监督),模型可通过接收骨架特征输入,最终得到1个[0, 1]区间的概率值,并结合阈值判断是否出现异常模式。

假设x代表视频帧或者视频片段中的人体骨架,将其输入到预测模型中获得预测概率值y,如式(10)所示。

$$y = h_1(x)\,, \tag{10}$$

其中h_1代表预测模型。当y大于预定义阈值时,判定当前视频出现异常模式。

如果是在没有标注视频骨架数据的情况下,基于预测的方法通常利用视频数据的时序特性预测骨架运动轨迹或行为,并根据预测误差判断异常是否发生。与基于重构的方法不同,基于预测的方法更注重动态建模。通过对比预测骨架与实际行为,当差异较大时可判定异常发生。

假设视频中连续t帧的人体骨架为$\{x_1, x_2, \cdots, x_t\}$,将其输入预测模型,得到下一帧预测的人体骨架为

$$\hat{x}_{t+1} = h_2(x_1, x_2, \cdots, x_t)\,, \tag{11}$$

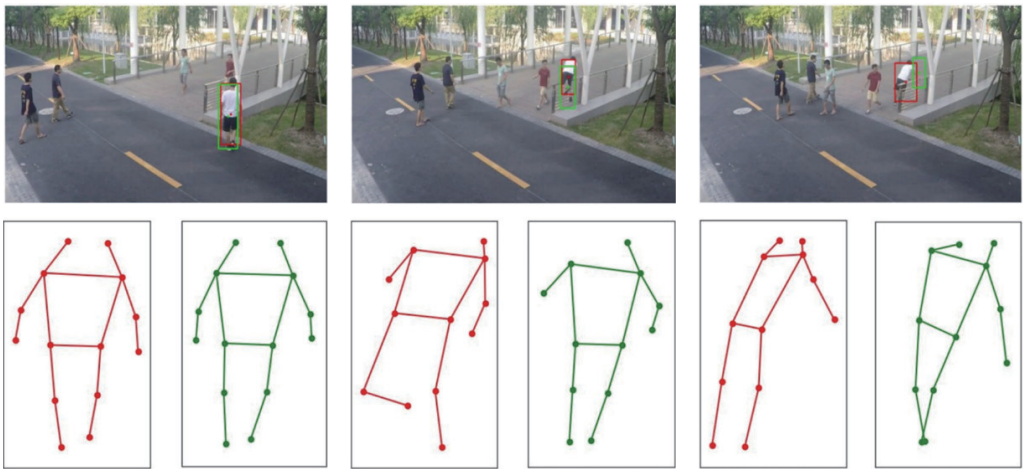
其中h_2代表预测模型。预测模型训练的目的是获得尽可能与真实x_{t+1}保持一致的预测值$\hat{x}_{t+1}$。在模型推理阶段,通过计算预测结果$\hat{x}_{t+1}$与真实x_{t+1}之间的误差:

$$\varepsilon_1 = l_1(x_{t+1}, \hat{x}_{t+1})\,, \tag{12}$$

可判断异常是否发生。式(12)中的l_1表示预测结果$\hat{x}_{t+1}$与真实x_{t+1}的预测误差函数。

图5展示了正常模式和异常模式下通过预测模型输出得到的骨架和真实骨架的示例。在图5中,第1行展示的是对应的视频帧;第2行的每个子图中,左侧骨架表示真实骨架,右侧骨架则为模型预测的骨架。图5(a)对应1个正常事件的视频帧,其预测骨架与真实骨架在形态和尺度上高度相似,表明模型在此情形下的预测误差较小。相比之下,图5(b)(c)中的预测骨架与真实骨架在形态和大小上存在明显差异,反映出较大的预测误差,从而揭示了潜在的异常模式。

基于预测的方法一般包括有监督、无监督和自监督学习范式。早期的研究偏向于有监督学习,即在有标注视频骨架数据的情况下进行异常判定。Yu等人[25]通过构建时空骨架图并采用时空图卷积网络



(a) 正常骨架 (b) 异常骨架1 (c) 异常骨架2

Fig. 5 Comparison of predicted skeleton and truth skeleton under normal and abnormal patterns[15]

图5 正常与异常模式下的预测骨架与真实骨架对比[15]

(spatial temporal graph convolutional network, ST-GCN)学习正常与异常动作的判别特征，利用 Softmax 层生成的概率分布提取异常得分，从而通过监督分类实现骨架动作异常检测。Qiu 等人[26]以 MobileNet 作为骨干网络进行特征提取，并将传统卷积操作替换为残差连接模块，从而有效提升模型的特征表达能力。该方法最终采用轻量级 K 近邻(K-nearest neighbor, KNN)分类器对异常行为进行判别。

由于视频骨架数据的标注成本高且数据缺乏，无监督范式和自监督范式逐渐成为主流。Luo 等人[27]将时空图卷积网络(ST-GCN)应用于基于人体骨架的视频异常检测，提出无监督模型 Normal Graph，并通过构建“正常图”来描述正常数据中关节的图连接，从而使异常事件关节成为图中的离群点。Normal Graph 模型通过对骨架的学习获得预测误差，进而计算出异常分数。异常分数越高表示该行为属于异常的概率更高。Zeng 等人[14]提出层次时空图卷积神经网络 HSTGCNN 进行异常检测。该模型属于无监督模型，使用层级图对骨架进行建模，并采用高阶图和低阶图分别表示个体间的交互和单个个体的姿态。为提升 HSTGCNN 模型在不同场景下的适应性，研究人员按照视频帧中人的数量、骨架边界框大小等指标对输入视频进行聚类划分。如果场景人群密集且骨架较小，HSTGCNN 模型在高阶图误差计算时分配较大权重系数；反之，HSTGCNN 模型在低阶图误差计算时分配较大权重系数。最后该模型通过加权高阶图和低阶图 2 个分支的预测结果获得最终异常分数。Huang 等人[23]使用 Transformer 架构设计出无监督模型 STGformer。该工作使用层级图来表示骨架特征，以实现个体间交互和个体内部关节相关性的建模。为降低模型计算复杂度，他们将 Transformer 模型拆成时间 Transformer 模块和空间 Transformer 模块，分别用于提取骨架姿势的空间特征和时间特征。为提高模型判别能力，研究人员在这 2 个 Transformer 模块之间嵌入图卷积网络模块。最后，该框架通过计算预测骨架与真实骨架之间的误差来确定异常分数。Pang 等人[28]提出无监督双分支生成模型。该模型在人体骨架视频异常检测工作基础上，增加基于像素的视频异常检测分支，实现多模态异常检测，其中像素分支和骨架分支都含有预测模块。此外，模型通过引入多头共注意力(multi-head co-attention, MCA)模块融合 2 个分支特征，使模型在推理过程中综合 2 类特征的优势以增强视频异常检测性能。最后，模型通过计算像素分支和骨架分支的异常分数并加权得到汇总分数。Rodrigues 等人[29]提出使用多重时间尺度模型对未来帧和过去帧进行预测，构建出自监督模型 Multi-Pred，并认为异常行为可在不同时间尺度上发生。如果只在一个固定时间尺度上进行异常检测，很难捕捉不同时间尺度上发生的大范围异常。为此，他们改进了传统编码器-解码器网络，使改进网络的中间层在不同时间尺度上预测过去和未来视频帧，从而捕获不同时间尺度上的动态变化。模型最后综合所有时间尺度的预测误差来计算异常分数。Fan 等人[30]提出使用门控循环单元(gated recurrent unit, GRU)和前馈网络(feedforward network, FFN)对人体运动进行预测，并设计出自监督模型 GRU-FFN。该模型使用双重损失函数确保关键点的范围和速度预测的精度，最后根据动作关键点的预测误差实现异常检测。Jain 等人[9]提出使用条件变分自编码器(conditional variational autoencoder, CVAE)预测未来骨架姿态，即 PoseCVAE。该自监督模型首先将人体姿态作为输入，然后通过 CVAE 学习正常活动的条件后验分布，从而预测未来的姿态。为提高模型对未来姿态的预测能力，模型采用一种自监督和无监督混合训练策略，最后通过计算预测姿态与真实姿态之间的误差来生成异常分数。Kanu-Asiegbu 等人[31]提出把双向轨迹预测器与姿态约束结合起来进行异常检测，提出自监督模型 BiPOCO。该工作在模型训练中引入 3 种与姿态相关的损失函数来提高模型对正常和异常姿态的区分能力，最后通过计算预测姿态与真实姿态之间的误差确定异常分数。Pang 等人[15]将 Transformer 框架应用于基于人体骨架的视频异常检测任务，提出自监督骨架 Transformer 的 SkT 模型。首先，骨架运动轨迹在输入模型之前进行全局和局部组件分解。然后，SkT 模型结合多头自注意力(multi-head self-attention, MSA)模块和时间卷积层对骨架的时空信息进行特征提取。最后，模型通过对未来帧的全局和局部骨架姿势分量的预测计算异常分数。Stergiou 等人[32]结合时空 Transformer 模型提出自监督整体表示学习方法，即通过多任务学习来预测过去、现在和未来的轨迹段以实现异常检测。最后，模型通过比较预测轨迹和真实轨迹获得异常分数。为解决骨架数据中存在的时空依赖性、局部区域特异性以及动作多样性这三大关键挑战。

Karami 等人[33]提出 GiCiSAD 模型。该自监督模型融合图注意力预测模块、图级拼图模块和图条件扩散模型进行异常检测。为综合考虑多个生成样本的异常分数，GiCiSAD 模型采用了多种聚合策略，包

括取平均值、中位数、最大距离和最小距离。研究人员通过实验发现最小距离策略在检测性能上表现最佳，因此将其作为有效的异常分数评估方法，以确定视频帧的异常程度。其优点是基于预测的方法能够充分利用视频数据的时序特性[29]，具有较好的动态行为建模性能，能够有效感知异常行为的突发变化。此外，这类方法灵活性较强，可通过调整预测模型的结构和参数来适应不同数据特性和异常检测需求[15]。其缺点是预测模型性能在很大程度上依赖数据的质量和数量。如果数据中存在噪声或异常样本较多，会导致预测误差增加，从而影响异常检测的准确性[29]。此外，基于预测的方法在处理复杂场景和多人体交互时会遇到人群遮挡、极端姿态等系列挑战[14]。这些场景下的行为模式更加复杂多样，难以通过单一的时间序列实现有效预测和准确建模。

基于预测的方法适用于那些异常行为具有明显时间动态特征且与正常行为存在较大差异的场景，例如场景中出现个体突然摔倒、快速奔跑等突发异常行为[29]。在单人场景或异常行为较为简单的场景中，这类方法能够发挥不错的效果[15]。

2.3.2 基于重构的方法

基于重构的方法通过学习正常行为的骨架，以数据重构的方式评估视频中骨架是否发生异常。这类方法的核心思想是：首先训练一个能对正常行为骨架进行重构的模型，然后通过比较重构骨架与实际骨架之间的误差来判断行为是否异常。如果重构误差较大，则说明异常发生。

假设x代表视频中的人体骨架，f代表重构模型，l_2代表计算x与$f(x)$之间重构误差的函数，ε_2代表重构误差。基于重构的异常检测方法通过极小化式(13)进行重构：

$$\varepsilon_2 = l_2(x, f(x))\text{。} \tag{13}$$

重构模型的学习范式一般为无监督和自监督。Fan等人[34]首次将循环一致生成对抗网络(cycle-consistent generative adversarial network, CycleGAN)应用于视频异常检测。该模型采用无监督学习范式，通过循环一致性损失训练生成器和判别器，旨于最小化视频帧和姿态估计图的重构误差，最后通过计算重构视频帧与原始视频帧之间的峰值信噪比(peak signal-to-noise ratio, PSNR)确定异常分数。异常分数越高表示该行为属于异常的概率越高。Yu等人[35]提出自监督运动先验规则学习器MoPRL模型进行异常检测。该工作首先使用人体运动嵌入器对姿态运动进行时空建模，即采用先验概率分布拟合骨架的位移分布，同时把先验概率分布作为缩放因子应用于骨架特征学习。然后MoPRL利用时空Transformer模型学习骨架的时空特征以生成重构的姿态。最后，模型通过计算重构姿态和原始姿态之间的差异来获得异常分数。其优点是基于重构的方法能够直接学习正常行为的特征表示，通过重构误差来衡量行为的异常程度。这种方法能够避免直接对异常行为建模，降低异常检测的复杂性[34]。此外，重构模型通常具有较好的可解释性，使研究人员通过分析重构误差的分布和变化实现对异常检测结果的理解[35]。其缺点是基于重构方法的性能在很大程度上依赖于重构模型的准确性。如果重构模型无法准确地学习正常行为的特征表示，会导致重构误差较大，进而影响异常检测的性能[10]。此外，基于重构的方法在架构设计上，可能面临“过度泛化”的问题[10]，即模型不仅能很好地重构正常模式，而且能很好地重构异常模式，导致模型无法准确识别异常骨架，从而出现漏检。

基于重构的方法适用于正常行为特征稳定且具有规律性的场景，例如在监控视频中对正常行走、站立等行为进行异常检测[35]。在数据量充足且正常行为模式明确的场景中，这类方法能够发挥较好的效果。

2.3.3 基于重构和预测相结合的方法

基于重构的方法和基于预测的方法各有优势。近年来，研究人员提出基于重构和预测相结合的异常检测方法。此类方法从多个角度对骨架进行分析，有效提升异常检测准确率和鲁棒性。在该异常检测框架下，基于重构的部分通过重构正常行为骨架实现异常检测，基于预测的部分通过对比预测骨架运动与实际骨架运动进行分析。该方法可在多个层面捕捉视频异常，且对复杂多变的行为模式具有较强适应性。

早期基于重构和预测相结合的骨架异常检测研究集中在有监督和弱监督学习范式。Hristov[36]设计了1DCNN-LSTM模型，通过1维的卷积神经网络(convolutional neural network, CNN)来提取和重建骨架序列的局部时空特征，并且使用长短期记忆网络(long short-term memory, LSTM)推理出逐帧骨架的异常得分，结合端到端模型的监督训练判断每个骨架是否发生异常。Morais等人[7]首次提出使用全局和局部组件分解法对骨架运动进行建模。该方法将骨架运动分解为全局运动和局部姿势2个子过程，并设计了消息传递编码器和解码器循环网络MPED-RNN来处理这2个子过程；然后使用2个解码器输

出重构骨架序列和预测骨架序列；最后通过融合2个序列的输出结果计算异常分数。MPED-RNN模型采用弱监督视频级标签进行训练[7,37]，使其具备良好的异常检测效果。

近年来，无监督与自监督学习范式因其无需异常标签、适应开放场景等优势，已成为当前骨架异常检测研究的热点。在基于全局和局部组件分解的骨架建模基础上，Li等人[38]设计了无监督模型STGCAE-LSTM，使用时空图卷积自编码器（spatial-temporal graph convolutional autoencoder, ST-GCAE）和LSTM联合对输入数据进行重构以及对未来序列进行预测，然后通过重构误差和预测误差的融合实现异常检测。Luo等人[39]进一步改进基于ST-GCAE的单编码器双解码器网络，设计出无监督模型Mem-STGCAE，他们在检测架构中引入记忆模块记录正常模式，并通过检索最相关的模式来增强自编码器的性能，最后通过结合重构和预测误差生成异常分数。

鉴于生成模型在特征学习的有效性，研究人员尝试把生成模型运用在基于重构和预测相结合的视频异常检测模型中。Li等人[8]提出结合自编码器和梯度惩罚的无监督生成对抗网络MemWGAN-GP模型进行视频异常检测。在该模型架构中，判别器拟合真实样本和生成样本分布之间的Wasserstein距离，而生成器最小化Wasserstein距离，使真实样本和生成样本的分布更接近。模型还引入记忆模块以克服自编码器强泛化能力对异常样本重构和预测所带来的影响，最后通过比较重构和预测的特征序列与其真值的差异对异常进行识别。Flaborea等人[40]首次将扩散模型应用在基于人体骨架的视频异常检测。设计了一个无监督多模态运动条件扩散模型生成骨架姿态。该框架首先把当前骨架序列输入到图卷积编码器得到相应的潜在表示。这个潜在表示可输入到解码器重构骨架序列，也可输入到扩散模型中指导骨架序列的反扩散去噪过程，以实现骨架序列的预测。最后，模型通过聚合重构误差和预测误差获得异常分数。Xue等人[41]改进了Flaborea等人[40]的工作，提出无监督模型MCDiffusion。该模型使用层级图对骨架建模以捕捉个体间的动态交互，并在扩散模型中设计一组特定的掩码机制，用于掩蔽将要重构和预测的骨架序列，最后计算原始骨架序列和由掩蔽生成的骨架序列之间的误差以生成异常分数。Noghre等人[42]提出基于双流图的自监督视频异常检测方法TSGAD。该模型结合Pishgu模型[43]预测骨架的运动轨迹，使用图注意力变分自编码器（graph attentive variational autoencoder, GA-VAE）重构输入的骨架姿势，最后计算姿势和运动轨迹的异常分数并融合起来获得最终的异常分数。

Wang等人[44]设计了自监督双条件扩散框架DCMD，在基于骨架的视频异常检测中引入“条件运动”和“条件嵌入”2种条件，以联合指导扩散模型的反向去噪过程，同时兼顾运动的显式几何约束与隐式语义先验。为增强正常模式和异常模式的区分度，该框架在损失函数中引入联合关联差异正则化，有效结合高斯核时间关联和自注意力全局关联。DCMD框架包含1个重建分支，可在时空上分离自编码器，重构历史运动序列；还包含1个预测分支，通过扩散模型对目标运动进行预测。该框架最后结合重构误差和预测误差计算异常分数。其优点是基于重构和预测相结合的方法能够综合利用重构误差和预测误差衡量行为的异常程度，以提高异常检测的准确性和鲁棒性[7]。通过结合重构和预测的优点，可在一定程度上克服单一方法局限性，让模型更好地适应复杂场景和多人体交互的异常检测需求[40]。此外，这种方法具有较强灵活性，可通过调整重构和预测模型的结构以及参数适应不同数据特性和异常检测需求[42]。其缺点是基于重构和预测相结合的异常检测模型结构相对复杂，模型训练和推理需要消耗更多计算资源和时间[8]。此外，由于涉及重构和预测2个过程，模型的参数优化和调试更加困难。这要求模型在重构误差和预测误差之间进行合理权衡，才能达到最佳异常检测效果[10]。

基于重构和预测相结合的方法适用于异常行为复杂且难以通过单一方法有效检测的场景[40]，例如多人交互场景中的异常行为检测。在数据量充足且计算资源允许的情况下，这类方法能够发挥较好的效果[38]。

2.3.4 基于重构和聚类相结合的方法

基于重构和聚类相结合的视频异常检测方法不仅可根据正常行为的重建实现异常检测，而且可通过聚类策略对正常行为模式进行分析。通过结合重构方法和聚类分析的优势，所构建模型能够准确捕捉视频中的异常，特别在复杂多变行为模式下表现尤为突出。

目前主流基于重构和聚类相结合的异常检测方法聚焦在无监督学习范式。Markovitz等人[45]首次提出无监督时空图卷积自编码器（ST-GCAE），并将其结合深度聚类层以实现基于骨架的异常检测。该方法首先把骨架表示成时空图输入并嵌入到ST-GCAE

模型。随后把骨架特征嵌入到潜在的特征空间进行聚类分析。模型微调阶段通过最小化重构损失和聚类损失组成的多目标损失函数，对编码器和聚类的潜在特征输出进行优化，最后使用 Dirichlet 过程混合分布模型计算每个样本的正常分数。Yan 等人[46]进一步改进 Markovitz 等人[45]的工作，提出一种无监督结合记忆模块和 ST-GCAE 的骨架异常检测模型。该模型采用自编码器、聚类策略和记忆模块，能够实现伪标签和网络参数的实时更新。为衡量样本特征与记忆模块质心特征的相似性，Yan 等人[46]设计一种新的损失函数，并通过 Dirichlet 过程混合模型对聚类结果进行评分。Chen 等人[24]提出一种无监督多尺度时空注意力图卷积网络，通过多尺度时空图捕获人体骨架运动信息。空间尺度包括关节、部位和肢体这 3 个层次。时间尺度涵盖不同时间步长。多尺度时空注意力图卷积异常检测模型包括特征提取编码器、面向编码器优化的重构解码器，以及异常分数生成的聚类层。在训练阶段，解码器通过最小化重构误差来优化编码器。在测试阶段，解码器不参与计算，仅通过聚类层生成异常评分。为提高聚类的稳定性，Javed 等人[47]设计了一种结合空间图卷积算子和无解码器的无监督深度聚类架构，并首次在模型训练时采用兴趣帧选择策略替代传统滑动窗口策略，从而大幅降低时间复杂度，减少训练时间。模型最终通过计算重构误差和最小化互信息的下界来获得异常分数。

除了主流的无监督异常检测学习范式，研究人员也尝试了其他学习范式，如弱监督学习。Cui 等人[37]改进了 ST-GCAE 模型，提出一种基于原型生成的图卷积网络。该模型使用骨架图作为输入，通过移位 ST-GCAE 模块提取特征，然后根据原型生成模块捕获正常原型分布。该模型的训练采用半监督学习（隶属于弱监督学习范式），在预训练阶段，模型在原型重构损失和生成损失的共同引导下逐步优化 GCAE 和原型生成模块；在微调阶段，模型使用多目标损失函数（包括重构损失、原型生成损失和聚类损失）对整个网络进行微调。最后，该方法采用深度聚类计算异常分数。该方法优点是基于重构和聚类相结合的方法能够综合利用重构误差和聚类分析结果衡量行为的异常程度，从而提高异常检测准确率和鲁棒性[45]。对正常行为进行聚类分析能在一定程度上捕捉隐藏在正常行为中的多样性和复杂性，从而使模型更好地适应复杂场景和多人体交互的异常检测需求[46]。这种方法通常具有较强可扩展性，可以通过改进聚类算法和调整重构模型结构来适应不同数据特性和异常检测需求[24]。其缺点是基于重构和聚类相结合方法的模型结构相对复杂，其训练和推理过程需要较多的计算资源和时间[47]。由于此类模型在特征学习中涉及到重构和聚类 2 个过程，模型参数优化和调试会更加困难。然而，通过对重构误差和聚类分析结果进行合理权衡，可以达到较满意的异常检测效果[37]。此外，受到数据分布和不同聚类算法的影响[45]，聚类分析结果会在一定程度上影响异常检测性能。

基于重构和聚类相结合的方法适用于正常行为复杂多样的场景，例如在人群密集公共场所对异常行为进行检测[45]。在数据量充足且计算资源允许情况下，这类方法能够发挥较好效果。

2.3.5 其他方法

近年来，研究人员还通过引入流模型、自回归模型等其他新的策略进一步提升异常检测的精度及其适应性。这些模型在复杂和动态场景下，展现出较强的鲁棒性。

Hirschorn 等人[48]将标准化流（normalizing flows，NF）模型引入基于人体骨架的视频异常检测。他们设计了时空图标准化流 STG-NF 模型。该模型能够学习骨架数据的分布，然后通过计算输入样本的似然度来生成异常分数，并且在有监督和无监督学习下均展现强劲性能。Xiao 等人[49]结合人体运动学特征和标准化流模型进行异常检测。该无监督方法首先提取骨架数据中的步态、脚部位移和颈部位移等运动学特征来实现人体运动建模，然后使用标准化流模型估计骨架时间序列的数据密度，最后根据估计的密度生成异常分数。Li 等人[50]提出一种结合空间图卷积网络和全连接层的自训练回归模型。该模型采用无监督学习范式，其异常分数学习模块可在模型训练过程中通过迭代优化异常检测性能。为分析骨架运动在潜在特征空间中的不同表现，Flaborea 等人[51]提出无监督学习框架 COSKAD。该模型通过图卷积网络对骨架运动进行编码，把骨架运动特征嵌入到各种潜在特征空间（欧几里得空间、球面空间和双曲空间）进行分析，最后通过计算样本嵌入与样本分布中心的距离得出异常分数。

Yao 等人[52]在此基础上提出一种无监督在线学习策略，使模型参数可通过流数据持续更新，以适应场景的实时变化。其优点是通过引入流模型等新思想，异常检测方法能够很好地适应复杂场景和多人体交互的需求，并在精度和适应性上获得进一步

提升。例如，标准化流模型能够学习骨架数据的分布，有效处理人体骨架的时空序列，提高异常检测精度[48]；在线学习策略能够使模型参数通过流数据持续更新，提高异常检测的适应性和实时性[52]。其缺点是尽管以上方法新颖而有成效，这些方法时间复杂度较高，在训练和推理过程需要更多计算资源和时间。因此这类方法的泛化性仍需在更复杂数据集和实际环境中验证[51]。

以上方法适用于对异常检测精度和适应性要求较高的场景，例如复杂及多人体交互场景下的异常行为检测。在数据量充足且计算资源允许的情况下，这些方法能够发挥良好的效果。

为更清晰地呈现各类基于人体骨架的异常检测方法的技术脉络与内在特性，本文进一步从方法论角度对上述5类方法进行系统归纳，如表1所示。该表总结了每类方法的代表工作、基本原理、学习范式、优缺点及适用场景，旨于总结技术演进路线，并为后续研究的方法选型等方面提供有效参考。

Table 1 Classification and Summary of Anomaly Detection Methods Based on Human Skeleton

表1 基于人体骨架的异常检测方法的分类与总结

方法类别	代表工作	基本原理	学习范式	优点	缺点	适用场景
基于预测	HSTGCNN[14]，STGformer[23]	有标注数据：直接预测骨架的异常概率；无标注数据：预测未来骨架序列，通过预测误差检测异常	有监督、无监督、自监督	动态行为建模性能较好，灵活性强	噪声敏感，在复杂场景以及多人交互面临挑战	异常行为具有明显时间动态特征且与正常行为存在较大差异的场景
基于重构	CycleGAN[34]，MoPRL[35]	学习重构正常骨架序列，通过重构误差检测异常	无监督、自监督	异常检测复杂性低，可解释性强	依赖于重构模型的准确性，可能面临“过度泛化”问题	正常行为特征稳定且具有规律性的场景
基于重构和预测相结合	MoCoDAD[40]，DCMD[44]	结合重构和预测任务，综合2种误差检测异常	有监督、无监督、弱监督、自监督	同时具备重构和预测方法的优点，灵活性强	模型结构复杂，计算开销大，参数优化调试困难	异常行为复杂且难以通过单一方法有效检测的场景
基于重构和聚类相结合	GEPC[45]，SKELBA[47]	结合重构与特征聚类任务，通过特征分布检测异常	无监督、弱监督	同时具备重构和聚类方法的优点，鲁棒性和可扩展性强	模型结构复杂，计算开销大，参数优化调试困难，受数据分布影响	正常行为复杂多样的场景
其他	STG-NF[48]，COSKAD[51]	引入流模型等新技术检测异常	有监督、无监督	理论新颖，在复杂场景和多人体交互适应性强	尚不成熟，计算开销大，泛化性有待实际验证	对异常检测精度和适应性要求较高的场景

3 数据集

数据的多样性及数据集的质量对基于人体骨架的视频异常检测模型的训练和评估至关重要。CUHK Avenue，ShanghaiTech，UBnormal，UCF-Crime，IITB-Corridor数据集的提出为人体行为分析和异常检测提供了丰富的人体姿态数据和行为模式[29,53-56]。本节回顾用于测试基于人体骨架的异常检测方法的主流数据集。表2详细列出各个数据集的具体信息。每个数据集中关于正常样本和异常样本的示例见图6。

Table 2 Basic Information of Each Dataset

表2 各个数据集的基本信息

数据集	总视频数	训练视频数	测试视频数	异常活动数	异常帧总数	相机视图数	异常活动类型
CUHK Avenue[53]	37	16	21	47	3 820	1	跑步、抛掷物体、逗留等
ShanghaiTech[54]	437	330	107	130	17 090	13	推搡、追逐、打斗、跳跃等
UBnormal[55]	543	268	211	660	89 015	29	训练集：跌倒、跳舞、爬行等 测试集：跑步、躺下、拖拽等
UCF-Crime[56]	1 900	1 610	290	950			打斗、盗窃、抢劫、枪击等
IITB-Corridor[29]	358	208	150		108 278	1	抗议、骑自行车、打斗、追逐等

1）CUHK Avenue数据集

CUHK Avenue（简称Avenue）数据集[53]来源于香港中文大学，是一个专为视频异常检测任务设计的公开数据集。该数据集主要记录了监控场景下的地铁车站、公共场所等高人流密度区域视频。Avenue数据集的每个视频长度通常为1~2 min，训练集和测试集分别包含16个视频序列和21个视频序列，且训练总帧数达15 328，测试总帧数达15 324。每个视频片段在不同摄像机角度下拍摄，包含正常和异常2类行为，其中异常事件包括跑步、抛掷物体、逗留等。研究人员在原有Avenue数据集上删除与人无关的异常事件视频，引入人体骨架数据，整合成HR（human-



Fig. 6 Examples of normal samples and abnormal samples in each dataset[29,53-56]

图 6 各个数据集的正常样本和异常样本示例[29,53-56]

related)版本的 Avenue 数据集 HR-Avenue。HR-Avenue 数据集包含了高密度人群及复杂行为的场景，是对原始 Avenue 数据集的扩展，旨于测试并推动基于人体骨架的视频异常检测模型性能的提升。HR-Avenue 数据集提供精细的人体骨架标注信息，以支持高精度的异常行为检测模型的测试。

2)ShanghaiTech 数据集

ShanghaiTech 是由上海科技大学研究团队发布的视频异常检测数据集[54]。该数据集提供了人群密集场景下复杂多样的异常事件，如推搡、追逐、打斗、跳跃等，包含 330 个训练视频和 107 个测试视频，且训练帧总数达 274 515，测试帧总数达 42 883。与 Avenue 数据集类似，研究人员剔除了 ShanghaiTech 数据集中与人无关的异常视频帧，扩展成 HR-ShanghaiTech 数据集。该数据集大大增加异常检测任务的挑战性，支持高精度的异常检测测试。

3)UBnormal 数据集

UBnormal 是一个高质量的视频异常检测数据集[55]，面向多人体行为分析异常检测任务，可为探索复杂环境下的异常行为检测提供有效支持。UBnormal 数据集包含来自多个真实场景的视频数据，尤其专注于室内公共区域和办公环境下的异常行为识别。该数据集涵盖了各种常见的正常行为(如走路、交谈等)，其中测试集中的异常类型与训练集中的异常类型不同，满足开放集设置要求。UBnormal 数据集共有 268 个训练视频、64 个验证视频和 211 个测试视频，且训练帧总数为 116 087，测试帧总数为 92 640。训练集中的异常类型包括跌倒、跳舞、爬行等，测试集中的异常类型包括跑步、躺下、拖拽等。HR-UBnormal 数据集在 UBnormal 数据集的基础上进行扩展，除了传统的 RGB 视频数据和标注的异常行为数据外，还为每个视频提供人体骨架标注信息。

4)UCF-Crime 数据集

UCF-Crime 数据集[56]是由佛罗里达大学的研究团队构建，近年来在视频异常检测领域备受关注。该数据集包含 13 种异常活动类别，共 1 900 个视频片段，其中 1 610 个为训练视频，290 个为测试视频，具有奔跑、摔倒等较为常见的异常动作，以及抢劫、斗殴、枪击等严重犯罪行为。UCF-Crime 数据集中的异常行为不仅涵盖了个体异常动作，还包括群体之间的异常交互行为，这为多人体交互场景下的异常检测研究提供重要实验数据。此外，UCF-Crime 数据

集还提供了详细的标注信息，包括异常行为的起始和结束时间，以及异常行为的类型等，这为模型的训练和评估提供了有力支持。鉴于 UCF-Crime 中复杂的异常行为类型和丰富多样的场景，该数据集已成为异常活动识别任务中具有挑战性的基准。

5）IITB-Corridor 数据集

IITB-Corridor（简称 Corridor）数据集[29]是通过单目摄像头获取的异常活动数据集，由印度理工学院孟买分校的研究团队创建。该数据集中的视频是在孟买校园内的走廊场景下拍摄，其中的正常活动主要包括行走和站立，异常活动包括群体抗议、骑自行车、突然奔跑、追逐、打斗等。该数据集提供了帧级别的正常和异常数据标注，其中测试集包含 150 个视频，共 181 567 帧，其中异常帧有 108 278 帧。与现有其他通过单目摄像头拍摄的数据集相比，Corridor 数据集不仅包括单人异常行为，还包括群体级别异常活动，因此更具挑战性，成为了视频异常检测研究的测试基准。为了给基于人体骨架的视频异常检测模型提供更全面测试，研究人员对 Corridor 数据集进行了扩充，增加了人体骨架信息，创建了 HR-Corridor 数据集，从而为异常检测测试提供更精细的标注和更丰富的数据特征。

4 模型性能评估

4.1 评估指标

1）ROC 曲线下的面积

接收者操作特征曲线 ROC（receiver operating characteristic），是基于真阳性率（true positive rate, TPR）和假阳性率（false positive rate, FPR）绘制的曲线。ROC 曲线下包含的面积（area under curve, AUC）常用于衡量分类模型性能，特别适用于不平衡数据的分析。AUC 值的范围在 0~1：如果该数值越接近 1，表示模型分类能力越强；如果数值越接近 0.5，表示模型性能较差，接近随机猜测。

在基于人体骨架的视频异常检测任务中，AUC 主要用于评估模型区分正常行为与异常行为的能力。异常行为的检测可视为一个二分类问题，其中阳性类别为异常行为，阴性类别为正常行为。AUC 综合考虑模型在各种阈值下的性能，有效反映模型的异常检测能力。

2）等错误率

等错误率（equal error rate, EER）表示在二分类问题中，假阳性率与假阴性率（false negative rate, FNR）相等时的错误率。换而言之，EER 是模型区分正常行为与异常行为的“平衡点”，即假阳性和假阴性错误率相等时的数值。

在基于人体骨架的视频异常检测任务中，EER 可衡量模型在处理不平衡数据时的稳定性和可靠性。EER 越低，表示模型在异常行为检测中的精度越高，假阳性与假阴性错误率的差异较小。

在视频异常检测中，除了使用 AUC 和 EER 作为评估指标，一些基于像素的视频异常检测工作还使用了 *F*1-Score 和平均精度（average precision, AP）指标辅助评估模型性能。然而，受限于基于骨架方法的数据形式、学习范式及输出结构，*F*1-Score 和 AP 指标并不适合作为基于人体骨架的视频异常检测任务的主流评估指标。

4.2 方法性能比较

本文收集近 5 年基于人体骨架的视频异常检测方法并进行对比分析。根据实现原理上的不同，这些方法可划分为五大类，包括基于预测的异常检测方法、基于重构的异常检测方法、基于重构和预测相结合的异常检测方法、基于重构和聚类相结合的异常检测方法及其他方法。不同类型方法的占比如图 7 所示。

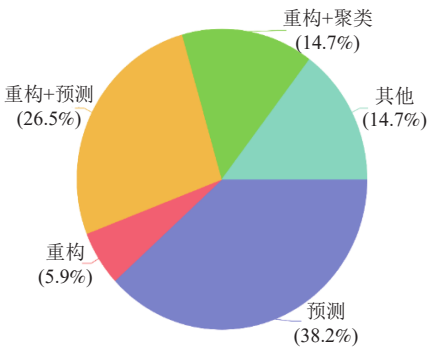


Fig. 7 Proportion of types of video anomaly detection methods based on human skeleton in recent five years

图 7 近 5 年基于人体骨架的视频异常检测方法类型占比

根据图 7，基于预测的方法占比 38.2%，成为主流方法。相比之下，基于重构的方法占比最少，为 5.9%。主要原因是研究人员倾向于将基于重构的方法和基于预测、聚类的方法相结合，利用混合方法的优势来提升模型在视频异常检测上的性能。其他方法是指除了基于预测、基于重构、基于重构和预测相结合、基于重构和聚类相结合之外的视频异常检测方法，如运用标准化流模型对骨架特征分布进行建模[48-49]，或者把骨架特征嵌入到不同的潜在特征空间进行异

常检测[51]。这些方法技术新颖、鲁棒性强，在基于人体骨架的视频异常检测领域展现出不错的效果。

在基于人体骨架的视频异常检测任务中，模型性能的评估不仅依赖其在单一数据集上的表现，还应考虑在不同数据集之间的异同。表 3~8 汇总了多种主流异常检测方法在多个公开数据集上的性能表现，其中，AUC 作为广泛采用的评估指标，贯穿于表 3~7 所有各项对比方法。而 EER 指标由于仅在部分方法中被报告，因此本文在表 8 中单独汇总了可获得的 EER 结果，以便更全面地反映不同方法的性能表现。表格中的数据均来源于参考文献中的实验结果，均在同等实验条件下得到。

Table 3 Comparison of Experimental Results for Frame-Level AUC Values Among Different Models on Avenue Datasets

表 3 不同模型在 Avenue 系列数据集上的帧级别 AUC 值实验结果对比

年份	类别	模型	帧级别 AUC/%	
			Avenue	HR-Avenue
2020	预测	Multi-Pred[29]	82.85	88.33
2021	预测	GRU-FFN[30]	**91.70**	
2021	预测	PoseCVAE[9]	82.10	87.78
2021	预测	Normal Graph[27]	87.30	
2022	预测	STGformer[23]	88.83	89.67
2022	预测	BiPOCO[31]	80.20	87.00
2022	预测	SkT[15]		86.70
2023	预测	HSTGCNN[14]	87.51	88.65
2024	预测	TrajREC[32]		89.40
2024	预测	3D U-Net + SkT[28]		87.00
2025	预测	GiCiSAD[33]		89.60
2022	重构	CycleGAN[34]	87.80	**90.20**
2019	重构+预测	MPED-RNN[7]		86.30
2022	重构+预测	STGCAE-LSTM[38]	83.70	86.30
2023	重构+预测	MoCoDAD[40]		89.00
2023	重构+预测	MemWGAN-GP[8]	83.70	88.50
2024	重构+预测	MCDiffusion[41]		89.70
2025	重构+预测	DCMD[44]		90.00
2023	重构+聚类	MSTA-GCN[24]	87.60	
2023	其他	Self-Trained SGCN[50]	82.00	88.30
2024	其他	COSKAD[51]	85.70	87.80

注：黑体数值表示最优结果。

各个异常检测方法在 Avenue 数据集实验效果如表 3 所示。在该数据集上，基于预测的方法普遍表现较好。例如，Fan 等人[30]提出的 GRU-FFN 模型在 Avenue 数据集上的 AUC 为 91.70%，表现最为突出。紧随其后的是 Huang 等人[23]提出的 STGformer 模型，其 AUC 为 88.83%。在 HR-Avenue 数据集上，各个模型的表现均比其在原有 Avenue 数据集上的效果好。这是因为 HR-Avenue 删除了与人无关的异常事件视频，从而使基于人体骨架的视频异常检测方法更有效识别出异常行为。尽管基于重构的 CycleGAN 模

Table 4 Comparison of Experimental Results for Frame-Level AUC Values Among Different Models on ShanghaiTech Datasets

表 4 不同模型在 ShanghaiTech 系列数据集上的帧级别 AUC 值实验结果对比

年份	类别	模型	帧级别 AUC/%	
			ShanghaiTech	HR-ShanghaiTech
2020	预测	Multi-Pred[29]	76.03	77.04
2021	预测	GRU-FFN[30]	82.60	
2021	预测	PoseCVAE[9]	74.90	75.70
2021	预测	Normal Graph[27]	74.10	76.50
2022	预测	STGformer[23]	82.86	86.97
2022	预测	BiPOCO[31]	73.70	74.90
2022	预测	SkT[15]		77.65
2023	预测	HSTGCNN[14]	81.80	83.40
2024	预测	TrajREC[32]		77.90
2024	预测	3D U-Net + SkT[28]		78.81
2025	预测	GiCiSAD[33]		78.00
2024	重构	MoPRL[35]	83.35	84.34
2019	重构+预测	MPED-RNN[7]	73.40	75.40
2022	重构+预测	STGCAE-LSTM[38]	75.60	77.20
2022	重构+预测	Mem-STGCAE[39]	77.80	78.72
2023	重构+预测	MoCoDAD[40]		77.60
2023	重构+预测	MemWGAN-GP[8]	75.30	78.00
2024	重构+预测	MCDiffusion[41]		78.30
2024	重构+预测	TSGAD[42]	80.67	81.77
2025	重构+预测	DCMD[44]		78.60
2020	重构+聚类	GEPC[45]	76.10	
2021	重构+聚类	DPM-GCN[37]	76.70	
2023	重构+聚类	MC-GCAE[46]	76.50	
2023	重构+聚类	MSTA-GCN[24]	75.90	76.80
2023	重构+聚类	SKELBA[47]	81.00	
2023	其他	STG-NF[48]	**85.90**	**87.40**
2023	其他	HKVAD[49]	72.21	73.15
2023	其他	Self-Trained SGCN[50]	71.70	72.90
2024	其他	COSKAD[51]	74.30	77.10

注：黑体数值表示该列指标排名第 1 的方法。

型[34]在Avenue数据集上只取得87.80%的AUC值，其在HR-Avenue数据集上的AUC值为90.20%。相比其他模型，这个数值是最优的，有效证实CycleGAN模型在基于骨架特征的密集人群异常检测上的应用潜力。

从表4可知，Hirschorn等人[48]提出的STG-NF模型性能最优，其在ShanghaiTech数据集和HR-ShanghaiTech数据集上的AUC值分别为85.90%和87.40%，均达到SOTA水平。这说明通过结合时空图卷积与标准化流技术可在各种异常类型的处理上展现出不错的准确性和鲁棒性。对比各种类型方法，基于预测的方法仍然占主流，其中Huang等人[23]提出的STGformer模型，在ShanghaiTech数据集和HR-ShanghaiTech数据集的AUC值分别为82.86%和86.97%，性能非常突出。STG-NF模型和STGformer模型之所以能取得优秀的性能，在于其网络架构均结合由时空图卷积网络设计的自编码器。在无监督学习范式下，时空图卷积网络能有效提高模型对骨架序列的时空特征提取能力。

根据表5和表6，各个异常检测模型在Corridor数据集和UBnormal数据集的综合表现较差。这是因为Corridor数据集和UBnormal数据集场景复杂多样，涵盖更多不同类型的异常模式，更具挑战性。尽管如此，STGformer模型以及STG-NF模型仍保持着SOTA水平，表现出强劲的优势。对于HR-UBnormal数据集，Wang等人[44]提出的DCMD模型的AUC值为69.00%，达到了目前SOTA水平，而Xue等人[41]提出的MCDiffusion模型与Karami等人[33]提出的GiCiSAD模型性能次之。值得注意的是，这些模型架构都采用了扩散模型进行骨架预测，有效表明扩散模型在提升基于人体骨架的视频异常检测方法的性能方面存在巨大应用潜力。

Table 5 Comparison of Experimental Results for Frame-Level AUC Values Among Different Models on Corridor Datasets

表5 不同模型在Corridor系列数据集上的帧级别AUC值实验结果对比

年份	类别	模型	帧级别AUC/%	
			Corridor	HR-Corridor
2020	预测	Multi-Pred[29]	67.12	
2021	预测	PoseCVAE[9]	67.34	70.60
2022	预测	STGformer[23]	**76.60**	**83.23**
2022	预测	SkT[15]	67.14	
2023	预测	HSTGCNN[14]	70.46	73.92
2024	预测	3D U-Net + SkT[28]	68.04	
2024	重构	MoPRL[35]	71.63	
2019	重构+预测	MPED-RNN[7]	64.27	68.07
2022	重构+预测	Mem-STGCAE[39]	69.30	75.63

注：黑体数值表示最优结果。

Table 6 Comparison of Experimental Results for Frame-Level AUC Values Among Different Models on UBnormal Datasets

表6 不同模型在UBnormal系列数据集上的帧级别AUC值实验结果对比

年份	类别	模型	帧级别AUC/%	
			UBnormal	HR-UBnormal
2024	预测	TrajREC[32]	68.00	68.20
2025	预测	GiCiSAD[33]	68.60	68.80
2023	重构+预测	MoCoDAD[40]	68.30	68.40
2024	重构+预测	MCDiffusion[41]	68.60	68.80
2025	重构+预测	DCMD[44]	69.00	**69.00**
2023	其他	STG-NF[48]	**79.20**	
2023	其他	HKVAD[49]	59.57	
2024	其他	COSKAD[51]		65.50

注：黑体数值表示最优结果。

各个模型在UCF-Crime数据集的性能展示如表7所示。由于UCF-Crime数据集发布时间较其他数据集晚，早期基于人体骨架的视频异常检测工作在UCF-Crime数据集上的测试较少。相对单一重构的模型，基于重构和聚类相结合的方法能够综合利用重构误差和聚类分析结果来衡量行为的异常程度，提高异常检测的准确性和鲁棒性。Javed等人[47]提出的SKELBA模型在AUC值上比仅使用重构方法的MoPRL模型[35]提高了22.07%。这反映重构和聚类混合模型比单一重构模型更有优势。

Table 7 Comparison of Experimental Results for Frame-Level AUC Values Among Different Models on UCF-Crime Datasets

表7 不同模型在UCF-Crime系列数据集上的帧级别AUC值实验结果对比

年份	类别	模型	帧级别AUC/%
2023	重构+聚类	MC-GCAE[46]	75.90
2023	重构+聚类	SKELBA[47]	**86.20**
2024	重构	MoPRL[35]	64.13

注：黑体数值表示最优结果。

在基于人体骨架的视频异常检测领域，EER也是重要评估指标。表8收集了过去工作在不同数

据集上帧级别的 EER 值。各个模型在 ShanghaiTech 数据集的 EER 值比 Avenue 数据集要高。这说明 ShanghaiTech 数据集的场景比 Avenue 数据集的场景更加复杂,在异常检测上更具挑战性。而且,HR 类型数据集的总体 EER 值比对应非 HR 类型的数据集低。这也证实基于人体骨架的视频异常检测方法对只保留人类行为的异常模式具有良好的识别效果。分析 Avenue 系列数据集可知,尽管 Li 等人[50]设计的 Self-Trained SGCN 模型在 HR-Avenue 数据集的 EER 值最低,其在 Avenue 数据集上的 EER 值比 MSTA-GCN 模型[24]要高。这反映在 Avenue 数据集上,Self-Trained SGCN 模型的异常行为检测精度比 MSTA-GCN 模型更低。这一结论也从表 3 中获得有效验证,表 3 也展示了 Self-Trained SGCN 模型的 AUC 值比 MSTA-GCN 模型低。对于 ShanghaiTech 系列数据集,Noghre 等人[42]提出的 TSGAD 模型在 ShanghaiTech 数据集和 HR-ShanghaiTech 数据集均取得了最低的 EER 值,体现了基于重构和预测相结合的模型能够很好地融合重构及预测 2 种策略的优势,在正常和异常行为检测过程中表现更佳。

Table 8 Comparison of Experimental Results for Frame-Level EER Values Among Different Models on Various Datasets

表 8 不同模型在各个数据集上的帧级别 EER 值实验结果对比

年份	类别	模型	帧级别 EER/%			
			Avenue	HR-Avenue	ShanghaiTech	HR-ShanghaiTech
2022	重构+预测	STGCAE-LSTM[38]	20.70	19.80	30.70	29.50
2023	重构+预测	MemWGAN-GP[8]	22.60	21.20	31.70	30.20
2024	重构+预测	TSGAD[42]			**25.00**	**25.00**
2023	重构+聚类	MSTA-GCN[24]	**19.20**		31.10	
2023	其他	Self-Trained SGCN[50]	23.90	**19.60**	34.10	32.70

注:黑体数值表示最优结果。

5 讨 论

虽然基于人体骨架的视频异常检测研究已取得显著进展,但如何根据目前方法存在的主要局限和实际应用面临难点展开深入研究是一个重要课题。

5.1 检测模型局限

1)骨架信息的准确获取与利用

人体骨架的检测质量直接影响异常检测模型的性能。现有 AlphaPose[11]和 OpenPose[12]等姿态估计算法在拥挤等复杂场景中难以提供高质量骨架数据。多人遮挡、快速运动、极端姿态和玻璃镜面反射等因素会导致骨架关键点定位产生误差[10,28]。这种误差对后续骨架重构、预测和异常检测会产生严重影响。尽管有一些改进方法尝试通过多视角或更复杂的姿态估计网络解决此问题,但依然需要模型在精度和计算效率之间进行合理权衡[57]。所以如何在复杂环境中高效且准确地提取人体骨架信息仍是一个难点。

2)复杂多样异常行为的识别

异常行为的多样性和复杂性使视频异常检测任务充满挑战。与常规行为相比,异常行为通常表现出高度变异性,并且在多个时间尺度上发生[29]。某些异常行为可能是短时的,难以通过传统时间序列分析方法捕捉,而另一些异常则表现出逐渐演变的行为模式。这对基于骨架的异常检测模型提出了更高要求。此外,异常行为经常受到场景、人物交互以及非人类物体交互等因素影响[48]。因此单一骨架信息难以表示所有异常模式。如何有效捕捉和区分不同类型的异常行为,是基于人体骨架的异常检测模型设计时首要考虑的问题[17]。

3)多人体交互的建模与理解

在多人物互动场景中,人体骨架的时空关系更加复杂。这要求模型对个体间的交互和群体行为模式有更高的理解能力。例如,在打斗、集会等异常事件中,模型不仅要分析个体的骨架,还需要理解个体间的相对运动关系[14,27]。现有大多数基于骨架的异常检测方法通常假设人物的行为是独立的,并且忽视人物间的互动关系。这导致在多人场景下,基于骨架的异常检测模型容易出现数据丢失或误检[16]。因此,如何在复杂环境中对多人物互动准确建模并捕捉人体间的相互关系,是模型性能提升的关键。

4)跨场景和跨数据集的模型泛化性

目前许多基于人体骨架的视频异常检测方法在特定数据集上取得较好表现,但一般难以在新的场景和数据集取得良好的泛化性。由于不同数据集的

拍摄条件、场景环境及人物行为等差异，模型性能在跨场景和跨数据集的测试中会大幅下降[58]。此外，为了适应不同场景或者数据集，研究人员还需要为异常检测方法设置不同的异常分数阈值。然而，这种参数选择无疑给模型稳定性的保持增加了困难[10,17]。因此，如何设计具有良好泛化能力的模型，以适应不同场景和环境，是模型优化的重要目标。

5）方法效率

基于人体骨架的视频异常检测方法通常需要处理大量时空数据，尤其在长时间视频或高分辨率数据集情况下，需要较高计算资源。尽管近年来深度学习方法在视频异常检测上的性能取得显著提升，其计算成本也随之增加。特别是在安防监控、智能交通等实际应用场景，模型对实时性要求非常高[52]。因此，如何平衡检测精度与实时性，确保模型在有限资源情况下仍能高效运行，是目前需要突破的瓶颈[16]。

5.2 应用场景限制

基于人体骨架的视频异常检测方法仅面向人的行为异常，从而导致这类方法会受到使用场景的限制。在 UCF-Crime 数据集[56] 中的部分异常类别比如爆炸、纵火等异常行为的发生是无法直接通过人体骨架进行识别。此外，在人群较为密集的场景下的视频异常识别受限于骨架检测方法，人群密集时的遮挡问题也会影响到视频异常检测的结果。因此，有研究人员提出了基于人体骨架与像素结合的视频异常检测方法[28]，能够解决部分场景限制的问题。尽管基于人体骨架的异常检测方法存在局限性，但大多数监控视频中关于人的异常行为是最值得关注的异常，因此对这类方法的分析与研究非常必要。

5.3 未来展望

1）鲁棒的骨架特征提取算法

人体骨架的检测质量直接影响异常检测的精度。尽管 AlphaPose[11] 和 OpenPose[12] 等现有姿态估计算法已取得显著进展，复杂场景下骨架特征提取仍面临巨大挑战。未来研究应聚焦于提高姿态估计模型在复杂环境中的精度和鲁棒性，特别是在低分辨率视频和遮挡严重的场景中，如何更好地获得鲁棒的人体骨架特征是关键。此外，基于深度学习的多模态数据融合方法已展现巨大潜力，是未来研究的重要方向。通过结合 RGB 图像、深度信息、红外图像等多种传感器数据，可增强骨架估计的准确性，进一步提升异常检测性能。结合视频序列以及合理的物理一致性约束来实现人体骨架特征提取，也是一个值得探索的方向。

2）强泛化性异常检测方法

异常行为的复杂性和多样性使基于骨架的检测方法在泛化性方面面临极大挑战。若能利用标准化流模型和扩散模型等对正常行为的分布进行有效建模，将有望识别出非显著的异常模式。为提高模型对人体骨架序列的时空特征提取能力，还可以使用时空图卷积网络模型、时空 Transformer 模型等进行特征学习。此外，如何在更细粒度层面识别不同类型的异常行为将是未来研究的一个重要方向。如能提升模型对视频中蕴含的语义信息、上下文信息以及动作相互关系的特征提取能力，极其有益于异常行为辨识。未来研究可探索更加灵活的模型架构，结合多层次、跨尺度的时间信息，并针对不同类型异常行为（如突发性异常、渐变性异常等）设计专门的识别策略。

3）多模态信息集成与异构数据融合

为提高异常检测的准确性和鲁棒性，未来研究可探索来自不同模态和异构数据的特征的有效融合[10]。例如，结合人体骨架、面部表情、语音信息以及环境感知数据（如温度、湿度等）可提供更加丰富的上下文信息，帮助模型更好地理解和预测异常行为。Pang 等人[28] 结合像素外观和骨架信息进行的视频异常检测研究，正是一个多模态融合的典型示例。此外，多模态学习和跨领域学习方法，尤其是在数据稀缺或标签不足的情况下，将为视频异常检测提供更具泛化性的解决方案。

4）模型跨场景适应能力

当前基于人体骨架的视频异常检测方法通常针对特定数据集进行训练和优化，其性能依赖于训练数据的质量和场景的相似性[55]。随着数据集的扩增，如何提高模型跨场景和跨数据集的泛化能力，成为未来研究的重要方向[58]。针对这一问题，领域自适应和迁移学习等方法具有较大的潜力。通过在不同场景和数据集之间进行知识迁移，提升模型在新环境中的适应性，是未来视频异常检测技术的重要发展方向。

5）模型实时性与计算效率

随着实际应用对视频异常检测需求的增加，如何在保证模型准确度的同时，提升模型的实时性和计算效率，成为未来研究的核心问题。为解决上述问题，Yao 等人[52] 率先提出新的在线学习策略，允许模型通过流数据持续对参数进行更新以适应场景的实时变化。这为研究人员解决异常检测模型的实时性问题带来新的思路。研究人员还可通过模型剪枝、

量化、蒸馏等技术来减少模型的计算复杂度,实现低延迟、实时的视频异常检测。此外,边缘计算和分布式计算技术的发展,也为大规模视频异常检测提供更高效解决方案。

6 结 论

本文全面回顾基于人体骨架的视频异常检测技术的研究现状,从基本概念、学习范式、方法流程到数据集评估,系统地总结了该领域的关键技术及未来发展方向。通过对基于骨架的异常检测文献的深入分析,可知现有方法在准确性、鲁棒性以及实时性等方面取得了显著进展,但仍面临骨架信息难以准确获取、异常行为复杂多样、多人体交互建模与理解难度大等局限。未来研究应聚焦于鲁棒的骨架特征提取算法、强泛化性异常检测方法、多模态信息集成与异构数据融合、模型跨场景适应能力以及模型实时性与计算效率。相信随着不断的技术创新和方法优化,基于人体骨架的视频异常检测技术将在智能监控等应用发挥更加重要的作用,为行为分析和社会安全提供更加有力的技术保障。

作者贡献声明:何锴撰写和修改论文;文嘉俊负责论文框架设计和论文修改;方鸿铭负责实验数据收集、整理与数据分析;赖志辉负责理论分析以及论文组织结构的调整与更新;沈琳琳负责对比模型的选择、算法分析及论文修改;徐勇负责论文框架优化、修改和定稿。

参 考 文 献

[1] Chalapathy R, Chawla S. Deep learning for anomaly detection: A survey[J]. arXiv preprint, arXiv: 1901.03407, 2019

[2] Yang Fan, Xiao Bin, Yu Zhiwen. Anomaly detection and modeling of surveillance video[J]. Journal of Computer Research and Development, 2021, 58(12): 2708−2723 (in Chinese)
(杨帆, 肖斌, 於志文. 监控视频的异常检测与建模综述[J]. 计算机研究与发展, 2021, 58(12): 2708−2723)

[3] Wang Zhiguo, Zhang Yujin. Anomaly detection in surveillance videos: A survey[J]. Journal of Tsinghua University: Science and Technology, 2020, 60(6): 518−529 (in Chinese)
(王志国, 章毓晋. 监控视频异常检测: 综述[J]. 清华大学学报: 自然科学版, 2020, 60(6): 518−529)

[4] He Ping, Li Gang, Li Huibin. A survey on deep learning based video anomaly detection[J]. Computer Engineering and Science, 2022, 44(9): 1620−1629 (in Chinese)
(何平, 李刚, 李慧斌. 基于深度学习的视频异常检测方法综述[J]. 计算机工程与科学, 2022, 44(9): 1620−1629)

[5] Pang Guansong, Shen Chunhua, Cao Longbing, et al. Deep learning for anomaly detection: A review[J]. ACM Computing Surveys, 2021, 54(2): 1−38

[6] Liu Wen, Luo Weixin, Lian Dongze, et al. Future frame prediction for anomaly detection—A new baseline[C]//Proc of the 2018 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6536−6545

[7] Morais R, Le V, Tran T, et al. Learning regularity in skeleton trajectories for anomaly detection in videos[C]//Proc of the 2019 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 11988−11996

[8] Li Nanjun, Chang Faliang, Liu Chunsheng. Human-related anomalous event detection via memory-augmented wasserstein generative adversarial network with gradient penalty[J]. Pattern Recognition, 2023, 138: 109398

[9] Jain Y, Sharma A K, Velmurugan R, et al. PoseCVAE: Anomalous human activity detection[C]//Proc of the 25th Int Conf on Pattern Recognition. Piscataway, NJ: IEEE, 2021: 2927−2934

[10] Mishra P K, Mihailidis A, Khan S S. Skeletal video anomaly detection using deep learning: Survey, challenges, and future directions[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2024, 8(2): 1073−1085

[11] Fang Haoshu, Li Jiefeng, Tang Hongyang, et al. AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6): 7157−7173

[12] Cao Zhe, 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

[13] Sun Ke, Xiao Bin, Liu Dong, et al. Deep high-resolution representation learning for human pose estimation[C]//Proc of the 2019 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 5686−5696

[14] Zeng Xianlin, Jiang Yalong, Ding Wenrui, et al. A hierarchical spatio-temporal graph convolutional neural network for anomaly detection in videos[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(1): 200−212

[15] Pang Wenfeng, He Qianhua, Li Yanxiong. Predicting skeleton trajectories using a skeleton-Transformer for video anomaly detection[J]. Multimedia Systems, 2022, 28(4): 1481−1494

[16] Zhu Honglei, Wei Pengjuan, Xu Zhigang. Research progress on skeleton-based human abnormal behavior recognition and detection[J]. Control and Decision, 2024, 39(8): 2484−2501 (in Chinese)
(朱红蕾, 卫鹏娟, 徐志刚. 基于骨架的人体异常行为识别与检测研究进展[J]. 控制与决策, 2024, 39(8): 2484−2501)

[17] Zhang Xiaoping, Ji Jiahui, Wang Li, et al. Overview of video based human abnormal behavior recognition and detection methods[J]. Control and Decision, 2022, 37(1): 14−27 (in Chinese)
(张晓平, 纪佳慧, 王力, 等. 基于视频的人体异常行为识别与检测方法综述[J]. 控制与决策, 2022, 37(1): 14−27)

[18] Shan Wenkang, Zhang Yuhuai, Zhang Xinfeng, et al. Diffusion-based hypotheses generation and joint-level hypotheses aggregation for 3D human pose estimation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(11): 10678−10691

[19] Li Qianxing, Kong Dehui, Li Jinghua, et al. Multi-anchor offset representation based coarse-to-fine diffusion model for human pose estimation[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2025, 21(4): 1−21

[20] Zheng Ce, Zhu Sijie, Mendieta M, et al. 3D human pose estimation with spatial and temporal transformers[C]//Proc of the 2021 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 11636−11645

[21] Zhu Wentao, Ma Xiaoxuan, Liu Zhaoyang, et al. MotionBERT: A unified perspective on learning human motion representations[C]// Proc of the 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 15039−15053

[22] Xu Guangsheng, Zhang Guoyi, Ye Lejia, et al. Optimizing local-global dependencies for accurate 3D human pose estimation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(12): 12306−12316

[23] Huang Chao, Liu Yabo, Zhang Zheng, et al. Hierarchical graph embedded pose regularity learning via spatio-temporal transformer for abnormal behavior detection[C]//Proc of the 30th ACM Int Conf on Multimedia. New York: ACM, 2022: 307–315

[24] Chen Xiaoyu, Kan Shichao, Zhang Fanghui, et al. Multiscale spatial temporal attention graph convolution network for skeleton-based anomaly behavior detection[J]. Journal of Visual Communication and Image Representation, 2023, 90: 103707

[25] Yu B X B, Liu Yan, Chan K C C. Skeleton-based detection of abnormalities in human actions using graph convolutional networks[C]//Proc of the 2nd Int Conf on Transdisciplinary AI. Piscataway, NJ: IEEE, 2020: 131−137

[26] Qiu Jiefan, Yan Xinlei, Wang Wei, et al. Skeleton-based abnormal behavior detection using secure partitioned convolutional neural network model[J]. IEEE Journal of Biomedical and Health Informatics, 2022, 26(12): 5829−5840

[27] Luo Weixin, Liu Wen, Gao Shenghua. Normal Graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection[J]. Neurocomputing, 2021, 444: 332−337

[28] Pang Wengfeng, He Qianhua, Li Yanxiong, et al. Detecting video anomalies by jointly utilizing appearance and skeleton information[J]. Expert Systems with Applications, 2024, 246: 123135

[29] Rodrigues R, Bhargava N, Velmurugan R, et al. Multi-timescale trajectory prediction for abnormal human activity detection[C]//Proc of the 2020 IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2020: 2615−2623

[30] Fan Bing, Li Peng, Jin Shu, et al. Anomaly detection based on pose estimation and GRU-FFN[C]//Proc of the 2021 IEEE Sustainable Power and Energy Conf. Piscataway, NJ: IEEE, 2021: 3821−3825

[31] Kanu-Asiegbu A M, Vasudevan R, Du Xiaoxiao. BiPOCO: Bi-directional trajectory prediction with pose constraints for pedestrian anomaly detection[J]. arXiv preprint, arXiv: 2207.02281, 2022

[32] Stergiou A, Weerdt B D, Deligiannis N. Holistic representation learning for multitask trajectory anomaly detection[C]//Proc of the 2024 IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2024: 6715−6725

[33] Karami A, Ho T K K, Armanfard N. Graph-jigsaw conditioned diffusion model for skeleton-based video anomaly detection[C]//Proc of the 2025 IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2025: 4237−4247

[34] Fan Zheyi, Yi Shuhan, Wu Di, et al. Video anomaly detection using CycleGan based on skeleton features[J]. Journal of Visual Communication and Image Representation, 2022, 85: 103508

[35] Yu Shoubin, Zhao Zhongyin, Fang Haoshu, et al. Regularity learning via explicit distribution modeling for skeletal video anomaly detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(8): 6661−6673

[36] Hristov. Real-time abnormal human activity detection using 1DCNN-LSTM for 3D skeleton data[C]//Proc of the 12th National Conf with Int Participation: ELECTRONICA. Piscataway, NJ: IEEE, 2021: 1−4

[37] Cui Tao, Song Wenyu, An Gaoyun, et al. Prototype generation based shift graph convolutional network for semi-supervised anomaly detection[C]//Proc of the 2021 Image and Graphics Technologies and Applications. Berlin: Springer, 2021: 159–169

[38] Li Nanjun, Chang Faliang, Liu Chunsheng. Human-related anomalous event detection via spatial-temporal graph convolutional autoencoder with embedded long short-term memory network[J]. Neurocomputing, 2022, 490: 482−494

[39] Luo Sibo, Wang Shangshang, Wu Yuan, et al. Memory enhanced spatial-temporal graph convolutional autoencoder for human-related video anomaly detection[C]//Proc of the 2022 Pattern Recognition and Computer Vision. Berlin: Springer, 2022: 665–677

[40] Flaborea A, Collorone L, Melendugno G M D D, et al. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection[C]//Proc of the 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 10284−10295

[41] Xue Zhenhua, Hu Ronghuai, Huang Chao, et al. Video anomaly detection via motion completion diffusion for intelligent surveillance system[J]. IEEE Sensors Journal, 2024, 24(21): 35928−35938

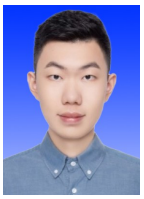
[42] Noghre G A, Pazho A D, Tabkhi H. An exploratory study on human-centric video anomaly detection through variational autoencoders and trajectory prediction[C]//Proc of the 2024 IEEE/CVF Winter Conf on Applications of Computer Vision Workshops. Piscataway, NJ: IEEE, 2024: 995−1004

[43] Noghre G A, Katariya V, Pazho A D, et al. Pishgu: Universal path prediction network architecture for real-time cyber-physical edge systems[C]//Proc of the 14th ACM/IEEE Int Conf on Cyber-Physical Systems (with CPS-IoT Week 2023). New York: ACM, 2023: 88−97

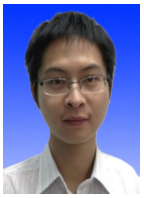
[44] Wang Hongsong, Xu Andi, Ding Pinle, et al. Dual conditioned motion diffusion for pose-based video anomaly detection[C]//Proc of the 39th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2025: 7700−7708

[45] Markovitz A, Sharir G, Friedman I, et al. Graph embedded pose clustering for anomaly detection[C]//Proc of the 2020 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ:

IEEE, 2020: 10536−10544
[46] Yan Mingchao, Xiong Yonghua, She Jinhua. Memory clustering autoencoder method for human action anomaly detection on surveillance camera video[J]. IEEE Sensors Journal, 2023, 23(18): 20715−20728
[47] Javed M H, Yu Zeng, Li Tianrui, et al. Learning anomalous human actions using frames of interest and decoderless deep embedded clustering[J]. International Journal of Machine Learning and Cybernetics, 2023, 14(10): 3575−3589
[48] Hirschorn O, Avidan S. Normalizing flows for human pose anomaly detection[C]//Proc of the 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 13499−13508
[49] Xiao Jian, Liu Tianyuan, Ji Genlin. Human kinematics-inspired skeleton-based video anomaly detection[J]. arXiv preprint, arXiv: 2309.15662, 2023
[50] Li Nanjun, Chang Faliang, Liu Chunsheng. A Self-Trained spatial graph convolutional network for unsupervised human-related anomalous event detection in complex scenes[J]. IEEE Transactions on Cognitive and Developmental Systems, 2023, 15(2): 737−750
[51] Flaborea A, Melendugno G M D D, D'Arrigo S, et al. Contracting skeletal kinematics for human-related video anomaly detection[J]. Pattern Recognition, 2024, 156: 110817
[52] Yao Shanle, Noghre G A, Pazho A D, et al. Evaluating the effectiveness of video anomaly detection in the wild online learning and inference for real-world deployment[C]//Proc of the 2024 IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2024: 4832−4841
[53] Lu Cewu, Shi Jianping, Jia Jiaya. Abnormal event detection at 150 FPS in MATLAB[C]//Proc of the 2013 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2013: 2720−2727
[54] Luo Weixin, Liu Wen, Gao Shenghua. A revisit of sparse coding based anomaly detection in stacked RNN framework[C]//Proc of the 2017 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 341−349
[55] Acsintoae A, Florescu A, Georgescu M I, et al. UBnormal: New benchmark for supervised open-set video anomaly detection[C]//Proc of the 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 20111−20121
[56] Sultani W, Chen Chen, Shah M. Real-world anomaly detection in surveillance videos[C]//Proc of the 2018 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6479−6488
[57] Li Jiangeng, Xie Haizheng. Human abnormal behavior recognition algorithm based on pose estimation[J]. Journal of Beijing University of Technology, 2022, 48(7): 710−720 (in Chinese)
(李建更, 谢海征. 基于姿态估计的人体异常行为识别算法[J]. 北京工业大学学报, 2022, 48(7): 710−720)
[58] Khan S S, Mishra P K, Ye Bing, et al. Empirical thresholding on spatio-temporal autoencoders trained on surveillance videos in a dementia care unit[C]//Proc of the 20th Conf on Robots and Vision. Piscataway, NJ: IEEE, 2023: 265−272



He Kai, born in 2003. Master candidate. Student member of CCF. His main research interests include computer vision, pattern recognition, video analysis, and embodied intelligence.
何 锴, 2003 年生。硕士研究生。CCF 学生会员。主要研究方向为计算机视觉、模式识别、视频分析、具身智能。



Wen Jiajun, born in 1983. PhD, assistant professor, master supervisor. His main research interests include image processing, pattern recognition, and video analysis.
文嘉俊, 1983 年生。博士, 助理教授, 硕士生导师。主要研究方向为图像处理、模式识别、视频分析。



Fang Hongming, born in 1999. Master candidate. His main research interests include computer vision and video analysis.
方鸿铭, 1999 年生。硕士研究生。主要研究方向为计算机视觉、视频分析。



Lai Zhihui, born in 1979. PhD, professor, PhD supervisor. His main research interests include image processing and pattern recognition.
赖志辉, 1979 年生。博士, 教授, 博士生导师。主要研究方向为图像处理、模式识别。



Shen Linlin, born in 1976. PhD, professor, PhD supervisor. Member of CCF. His main research interests include computer vision and multimodal large language model.
沈琳琳, 1976 年生。博士, 教授, 博士生导师。CCF 会员。主要研究方向为计算机视觉、多模态大语言模型。



Xu Yong, born in 1972. PhD, professor, PhD supervisor. Member of CCF. His main research interests include pattern recognition, computer vision, deep learning, and video analysis.
徐 勇, 1972 年生。博士, 教授, 博士生导师。CCF 会员。主要研究方向为模式识别、计算机视觉、深度学习、视频分析。