基于深度学习的监控视频异常行为检测方法

崔旭龙，张力文，刘　铠，赵哲宇，朱晓程

（齐鲁理工学院，山东 济南 250200）

摘　要　监控视频异常行为检测是智能安防技术持续演进中的关键研究方向，其核心任务在于依托视频数据识别偏离常态行为模式的异常事件。受监控场景复杂、目标遮挡频繁、行为边界模糊及异常样本稀缺等因素影响，传统依赖人工规则和浅层特征的方法已较难适应真实环境中的检测需求。深度学习方法依托对图像外观信息、时序运动信息及时空关联特征的联合建模，为异常行为识别提供了新的技术支撑。本文围绕监控视频异常行为检测中的空间特征学习、时序建模、三维卷积与双流结构、重构预测与弱监督学习、注意力机制与 Transformer 等方法展开分析，有助于进一步梳理该领域的主要技术路径，并为后续模型优化和工程应用提供参考。

关键词　深度学习；监控视频；异常行为检测；时空建模；智能安防

中图分类号：TN948.6　　**文献标志码**：A　　**DOI**：10.3969/j.issn.2097-3365.2026.12.007

0　引言

视频监控已深度嵌入校园、社区、交通枢纽及商业场所等多类公共空间，海量监控画面的持续累积，使异常行为的快速识别逐渐成为安防管理中的现实需求。相较于普通行为分类，异常行为往往具有突发性、低频性和开放性，实际场景中还常伴随光照变化、背景干扰、遮挡重叠及视角偏移等情况，这使传统检测方式很难稳定应对复杂环境。深度学习技术在视觉表征和时序关系建模方面表现出较强的适配性，因而为监控视频异常行为检测提供了更具延展性的研究空间。

1　监控视频异常行为检测的研究基础

1.1　监控视频异常行为检测的任务特点

监控视频异常行为检测与一般行为识别并不相同。一般行为识别多建立在既定类别上，识别目标较明确，模型只需判断目标属于哪一类动作；异常行为检测面对的则是大量正常视频中的少量偏离事件，判断重点不只是动作名称，更在于行为状态是否脱离场景常态[1]。打斗、跌倒、闯入、异常聚集、可疑徘徊等行为在监控画面中往往出现突然、持续时间不固定，且容易受到拍摄角度、背景运动、遮挡重叠和光照波动干扰，致使检测过程带有较强的不确定性。从任务属性看，异常行为检测通常具有四个突出特点：其一，异常样本占比低，正常视频远多于异常视频，模型训练容易向正常模式偏移；其二，异常类别具有开放性，不同场所对异常的判定标准并不一致，校园、车站、商场和工业区域的风险触发条件存在明显差异；其三，异常行为边界不够清晰，很多事件并非在某一帧突然形成，而是由局部异常逐渐发展为完整事件；其四，场景依赖性较强，同一动作放在不同空间环境中，风险含义可能完全不同。

1.2　监控视频异常行为检测的主要任务形式

从研究和应用的结合方式来看，监控视频异常行为检测通常可分为视频级判别、时间定位和时空定位三类任务。视频级判别侧重对整段视频作出总体判断，即识别其中是否存在异常事件。这一形式标注成本相对较低，适宜用于海量视频的初筛，但输出结果较粗，只能说明“有无异常”，难以直接支撑实时预警。时间定位是在视频级判别基础上的进一步细化，其目标是找出异常出现的具体时间片段。对于值守监控、事件回放和报警触发而言，这一任务更具应用意义，因为系统不再停留在整体判断，而是能够指出异常大致发生于哪一时段。若要继续增强检测精度，就需要进入时空定位层面，即在识别异常片段的同时，进一步标出异常目标所在区域或异常发生位置。三类任务之间存在清晰递进关系，视频级判别解决“是否异常”，时间定位解决“何时异常”，时空定位解决“何处异常”。

2　基于深度学习的监控视频异常行为检测核心方法

2.1　基于空间特征学习的异常检测方法

深度学习具有出色的特征提取效果及强大的数据拟合能力，检测精度较高，是异常行为检测领域中的

作者简介：崔旭龙（1998-），男，本科，助教，研究方向：计算机视觉。

主流研究算法[2]。监控视频进入智能识别阶段后，较早被广泛采用的是空间特征学习方法，这一路径通常以单帧图像或短片段关键帧为输入，依托卷积神经网络对人物外观、目标姿态、场景结构和局部异常线索进行编码，进而完成异常判别。其任务起点在于，许多异常行为在空间层面往往会先表现出可观测差异，如人员突然倒地、多人近距离扭打、目标翻越围栏、可疑物体异常滞留，这些变化在画面形态、轮廓分布和区域纹理上都具有一定偏离特征。模型在提取特征时，通常会把边缘信息、局部姿态、人体与背景的相对位置一并纳入表征过程，使原本依赖人工描述的视觉线索转化为高维特征表示，此类方法的优势在于结构清晰、训练过程相对稳定，尤其适宜处理静态异常较显著的场景。

空间特征学习存在一个明显边界，即模型更擅长识别“异常状态已经形成”的画面，却难以准确把握行为从正常向异常过渡的动态过程。换句话说，若只依靠单帧外观判断，模型看到的是某一时刻的结果，而不是事件演化链条本身，像追逐、推搡升级、缓慢聚集后突然冲突等连续性行为，就容易因时序信息缺失而被弱化。

2.2 基于时序建模的异常检测方法

异常行为并非始终依靠某一帧图像就能清晰识别，很多风险事件本质上属于动态过程，前后帧之间的动作演变往往比单帧外观更具判别价值[3]。基于这一认识，时序建模方法逐渐成为异常行为检测的重要技术方向，此类方法通常会把连续视频片段看作时间序列，把每一时刻提取到的特征依次送入循环结构之中，由模型学习动作变化的顺序关系、状态转移规律及长短期依赖信息。常见网络包括RNN、LSTM和GRU，其中LSTM与GRU因门控机制更适合处理较长时段的视频动态，因而在异常检测研究中更常出现。

时序建模方法的价值在于它能够把异常行为从“静态结果识别”推进为“动态轨迹分析”，比如人员突然奔跑，单帧画面未必足以说明风险，但若结合前后数秒画面，模型便能感知目标速度变化、方向切换及与周边人群的互动关系；又如跌倒行为，前一阶段可能只是身体重心倾斜，真正的异常信号往往形成于连续姿态失衡与落地停滞的组合过程中，若缺少时间序列分析，模型很难稳定区分正常弯腰与真实跌倒。也正因如此，时序建模在监控场景中往往承担“识别行为演化过程”的任务，能够增强对持续性异常、渐进式异常和突发性异常的辨识能力[4]。

2.3 基于三维卷积与双流结构的异常检测方法

异常行为并不是“空间特征+时间特征”的简单叠加，而是两者在连续视频流中的同步耦合结果。基于这一逻辑，三维卷积和双流结构逐渐成为监控视频异常检测中的关键方法。三维卷积网络在传统二维卷积的基础上引入时间维度，卷积核不再只在图像平面内滑动，还会沿着时间轴同步提取片段级动态信息。模型输入通常为连续若干帧画面，其输出特征既包含目标外观，又包含局部动作变化趋势，因而更适合描述奔跑、追逐、冲突、跌倒这类具有明显动态形态的异常事件。

三维卷积的优势在于时空特征能够在同一编码框架内被统一提取，模型不需要先独立分析空间内容，再单独叠加时间信息，而是直接从片段中学习事件的整体演化模式。对于异常检测而言，这一结构更接近视频本身的表达方式，也更容易捕捉短时突变动作带来的风险信号。不过，三维卷积对算力和存储资源的要求相对较高，片段长度、采样间隔和网络深度一旦设定不当，训练成本就会明显上升，边缘部署场景中的实时响应也可能受到影响。

双流结构则是另一条被广泛采用的路径。它通常设置两条并行通道，一条负责处理RGB图像中的外观信息，另一条负责处理光流等运动信息，随后在特征融合阶段形成统一判别结果，这样做的直接优势是把“画面里有什么”和“画面在怎么动”分开建模，再在更高层完成关联分析。对于监控视频而言，光流特征往往能够更敏感地反映局部运动异常，比如短时间内的人群方向紊乱、目标快速靠近、局部肢体剧烈摆动等，RGB分支则更适合保留场景布局、人体轮廓和目标语义。两路信息融合之后，模型对动态异常的识别边界会更清晰，对复杂背景下的误判也会形成一定抑制。

2.4 基于重构预测与弱监督学习的异常检测方法

监控视频异常行为检测面临的一个长期难题在于异常事件本身数量有限，且精细标注成本很高。要把每段视频中的异常起止时刻、异常区域和异常目标全部逐帧标出，不仅工作量巨大，还容易出现人为标注偏差。因此，研究者逐渐把关注点转向重构预测方法与弱监督学习方法，希望在标注不足条件下依旧保持较强的检测能力。重构预测方法的基本思路是先学习正常行为分布，再利用模型对正常样本的重建能力或

未来状态预测能力进行异常识别。换句话说，模型若长期只接触正常视频，它便会逐渐掌握正常场景中的运动规律和空间结构。

这一逻辑与传统分类方法明显不同，它并不要求模型提前见过所有异常类型，而是依托“正常学习”间接发现未知异常。对于开放场景下的监控任务而言，这种思路具有较强适配性，因为现实中的异常往往无法被完整列举，模型若一味依赖已知异常标签，面对新型风险事件时反而容易失效。重构类方法通常更关注输入画面与输出画面之间的差异，预测类方法则更强调下一时刻或后续片段的生成偏差，两者本质上都在利用偏离程度完成异常判别[5]。在重构预测类异常检测中，异常得分通常可依托输入帧与重构帧之间的差异进行计算，其表达式可写为：

$$S(x_t)=\|x_t-\hat{x}_t\|_2^2 \quad (1)$$

式（1）中，x_t 表示第 t 时刻的原始视频帧，$\hat{x}_t$ 表示模型生成的重构帧，$S(x_t)$ 表示对应时刻的异常得分。该值越大，说明当前帧与正常模式之间的偏离程度越明显，进而更可能被判定为异常片段。若将这一得分沿时间轴连续统计，还可进一步生成视频片段级异常变化曲线，为后续异常定位提供依据。

弱监督学习则从另一个角度回应了标注成本问题。它通常不要求逐帧精细标签，而是借助视频级标签完成训练，即只告诉模型某段视频中“是否包含异常”，至于异常具体出现在哪一时刻，则由模型在训练过程中自行挖掘。为了完成这一任务，研究中较多采用多实例学习框架，把一段长视频拆分为多个片段，再依照片段得分排序、注意力加权或伪标签更新，逐步逼近真正异常区域。这类方法的现实意义较强，一方面降低了数据集构建难度，另一方面也更接近工程环境中的数据获取方式。

2.5 基于注意力机制与 Transformer 的异常检测方法

伴随监控场景复杂度持续上升，传统卷积与循环结构在长距离依赖建模方面的限制逐渐显现。很多异常行为并非集中出现在单一目标或短时间窗口内，而是分散在长时段上下文之中，前后线索之间甚至相隔较远，比如局部人群先出现缓慢聚集，随后某一区域突然扰动扩大，若模型只依赖局部时序片段，很难把这些分散信号串联起来[6]。基于这一背景，注意力机制和 Transformer 结构开始被大量引入异常行为检测任务。注意力机制的核心作用在于让模型从复杂输入中主动聚焦更具判别价值的区域、目标或时间片段，把有限计算资源优先分配给关键异常线索，从而减轻背景噪声对最终判断的干扰。

在监控视频中，关键异常往往只占据局部区域或短暂时间窗口，若模型对整段视频平均处理，异常信号很容易被大量正常内容淹没。注意力机制的加入，使模型能够在特征编码阶段生成权重分布，把疑似异常区域、异常动作发生时段或目标交互密集区域突出出来，由此增强异常表征的集中度。相比之下，Transformer 的价值更体现在全局建模能力上，它依托自注意力机制直接建立远距离片段之间的关联，不再像传统循环网络那样严格依赖顺序传播，因而在长视频分析中更容易捕捉跨时间的上下文关系。

3 结束语

监控视频异常行为检测的研究价值已不再停留在单一技术识别层面，更深一层的意义在于让安防系统从被动留痕逐步转向主动感知与前置研判。深度学习的引入，使异常识别由依赖固定规则的静态判断转入面向复杂场景的动态分析，模型关注点也由表层动作捕捉延展到时空关系、环境语义与风险演化过程，这一变化既增强了监控数据的利用深度，又推动了视频智能分析向实战化方向靠近。未来，监控视频异常行为检测仍将围绕复杂场景适配、多源信息协同、轻量部署与智能预警联动持续深化，进而为智慧安防体系提供更强的支撑。

参考文献：

[1] 曾婷,黄东军.智能视频监控系统异常行为检测算法研究综述[J].计算机测量与控制,2021,29(07):1-6,20.
[2] 彭嘉丽,赵英亮,王黎明.基于深度学习的视频异常行为检测研究[J].激光与光电子学进展,2021,58(06):51-61.
[3] 徐涛,田崇阳,刘才华.基于深度学习的人群异常行为检测综述[J].计算机科学,2021,48(09):125-134.
[4] 张扬,刘泫梓,孙文婷.基于深度学习的视频监控异常行为检测技术研究[J].中国新通信,2025,27(17):29-31.
[5] 汪洋,周脚根,严俊,等.基于深度学习的监控视频异常检测方法综述[J].中国图象图形学报,2025,30(03):615-640.
[6] 郑凯东,江怡.基于改进C3D的视频监控异常行为检测算法[J].信息技术与信息化,2024(06):131-134.