2024 年 第 24 卷 第 19 期
2024, 24(19): 07941-15

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2305413
引用格式:张琳, 陈兆波, 马晓轩, 等. 无监督和弱监督视频异常检测方法回顾与前瞻[J]. 科学技术与工程, 2024, 24(19): 7941-7955.
Zhang Lin, Chen Zhaobo, Ma Xiaoxuan, et al. Review of unsupervised and weakly supervised video anomaly detection methods[J]. Science Technology and Engineering, 2024, 24(19): 7941-7955.

自动化技术、计算机技术

无监督和弱监督视频异常检测方法回顾与前瞻

张琳[1], 陈兆波[1*], 马晓轩[1], 张凡博[2]
(1. 北京建筑大学电气与信息工程学院, 北京 102616; 2. 交通银行软件开发中心, 北京 100031)

摘 要 随着监控技术的不断发展,监控摄像头已经被广泛部署到各种场景中。手动检测视频异常情况已经变得不可能。因此,作为智能监控系统核心的视频异常检测技术正在受到广泛关注和研究。随着深度学习的发展,视频异常检测领域取得了显著的成就,并涌现出许多新的异常检测方法。梳理了应用在不同数据类型上的无监督和弱监督视频异常检测学习方法,分析现有方法的贡献,并比较不同模型的性能。此外,还整理了一些常用的和新发布的数据集,并总结了未来工作要面临的挑战和发展趋势。

关键词 视频异常检测; 无监督; 弱监督; 数据集; 视频监控
中图法分类号 TP391.4; **文献标志码** A

Review of Unsupervised and Weakly Supervised Video Anomaly Detection Methods

ZHANG Lin[1], CHEN Zhao-bo[1*], MA Xiao-xuan[1], ZHANG Fan-bo[2]
(1. School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 102616, China;
2. Bank of Communications Software Development Center, Beijing 100031, China)

[**Abstract**] With the continuous development of monitoring technology, surveillance cameras have been widely deployed in various scenarios. Manual detection of video abnormality has become impossible. Therefore, video anomaly detection technology, as the core of intelligent surveillance systems, is receiving extensive attention and research. With the development of deep learning, the field of video anomaly detection has made significant achievements and has emerged many new anomaly detection methods. Unsupervised and weakly supervised video anomaly detection learning methods applied to various data types were sorted out, the contributions of existing methods were analyzed, and the performance of different models was compared. In addition, some commonly used and newly released datasets have also been compiled, and the challenges and development trends that future work will face have been summarized.
[**Keywords**] video anomaly detection; unsupervised; weakly supervised; dataset; video surveillance

视频已经成为当今社会中不可或缺的信息载体,广泛应用于安防、娱乐、教育、医疗等众多领域。随着智慧城市和数字社会的迅速发展,监控摄像头在各种场景中被广泛部署,产生了海量监控视频数据。这为各类场景下的异常检测带来了巨大的工作量和技术挑战。传统的人工检测耗时耗力,高度依赖人的视觉和注意力。然而,视觉疲劳和精神疲惫可能导致误检和漏检,无法应对大规模视频流的处理需求。视频异常检测(video abnormal detection, VAD)旨在自动分析监控视频数据,识别并准确定位可能存在的异常事件[1]。通过这种方法,可以显著降低异常检测成本,提高检测准确性和效率。作为智能监控系统的核心技术之一,VAD 在公共安全、工业制造、医疗保健、城市管理等领域具有广阔的应用前景[2]。因此,VAD 受到了学术界的广泛关注。作为一个热门的研究领域,VAD 任务每年都会有大量学术论文发表。针对不同的场景都有对应的 VAD 方法,特定场景训练的方法在其他场景中进

收稿日期: 2023-07-18; **修订日期**: 2024-03-13
基金项目: 北京市教育科学“十三五”规划重点课题(CHAA19081)
第一作者: 张琳(1975—),女,汉族,内蒙古乌兰察布人,博士,教授,硕士研究生导师。研究方向:智能信息处理,图像处理,人工智能。E-mail:1442115289@qq.com。
* **通信作者**: 陈兆波(1997—),男,汉族,湖南永州人,硕士研究生。研究方向:图像处理,视频异常检测。E-mail:chenzhaobo97@163.com。

行检测时,性能表现较差。因此,有必要对 VAD 方法进行系统的比较和分类,总结其优缺点,以展示当前的发展趋势和未来发展方向。传统的 VAD 算法通常需要从视频中提取手工特征,并构建特征空间进行异常检测。其中常见的手工特征包括轨迹[3]、梯度方向直方图(histogram of oriented gradients, HOG)[4]、光流直方图(histogram of optical flow, HOF)[5-6]和光流场[7-8]等低级视觉特征。然而,异常场景种类繁多且变化多样,低级的视觉特征难以捕获到有效的特征,导致检测性能不佳且严重依赖先验知识。因此,传统的 VAD 发展进展缓慢。

近年来,随着深度学习的快速发展,尤其是卷积神经网络在处理复杂数据方面的优越性能,深度学习在各种任务中得到广泛应用,如语音识别[9]、图像识别[10]、推荐系统[11]等。这也为 VAD 带来新的发展机遇。越来越多的研究员将基于深度学习的模型应用于 VAD 任务,并取得了显著成就。卷积神经网络(convolutional neural network, CNN)能够端到端的提取视频特征,减少对复杂先验知识的依赖。与传统的手工特征方法相比,深度学习视频异常检测方法在捕获多维空间语义特征和时间上下文特征方面表现更为出色。此外,随着图形处理单元(graphics processing unit, GPU)算力的迭代更新,大规模数据模型的训练变得更加可行。

VAD 是备受关注的研究领域,过去几年里有多项研究[12-13]致力于对 VAD 技术进行系统分类和调研。例如,为了研究 VAD 的发展特点和发展趋势,相关研究员根据方法的特点对相关文献进行梳理归类。例如,Kiran 等[14]从无监督和半监督的角度对 VAD 方法进行分类,包括重构学习、深度生成模型和预测模型等。Ramachandra 等[15]回顾了现有模型方法,并在多个基准数据集上进行了比较。Santhosh 等[16]侧重于道路交通场景的 VAD 方法,涵盖了无监督、半监督和基于监督的方法。然而,这些研究未涉及现实世界多场景和跨场景的方法研究。Ramzan 等[17]介绍了监控视频中的暴力检测技术,将其分为基于传统机器学习、支持向量机(support vector machine, SVM)和深度学习的方法,并讨论了相关的视频特征和数据集。Pang 等[2]根据通用特征提取、常态表示学习和端到端异常分数学习等原则对 11 种建模方法进行了分类,并强调了异常检测面临的复杂性和未解决的挑战。何平等[18]将 VAD 方法分为基于重构、预测、分类和回归四类方法,并对比了各种方法的性能。申栩林等[19]根据场景密度以及行为发生的对象,将基于生成对抗网络(generative adversarial network,GAN)的 VAD 方法分为个体和群体异常两种情况,并对基于重构和预测方法的个体和群体异常检测进行对比,详细阐述了各个方法的贡献和优缺点。然而,随着深度学习技术的迅速发展,VAD 方法也呈爆发式增长,针对不同应用场景开发出了对应的 VAD 方法,且不同的检测方法的检测效果也存在差异。现有的综述已无法涵盖该领域的最新研究成果和技术趋势。因此,需要对现有的 VAD 法进行梳理归类,以便在实际应用中便捷地选取对应的算法,现从无监督和弱监督两个方面详细阐述和分析 VAD 现有方法的优缺点,收集现有的 VAD 基准数据集和主要的性能评价指标,并探讨 VAD 领域面临的挑战和未来发展趋势。

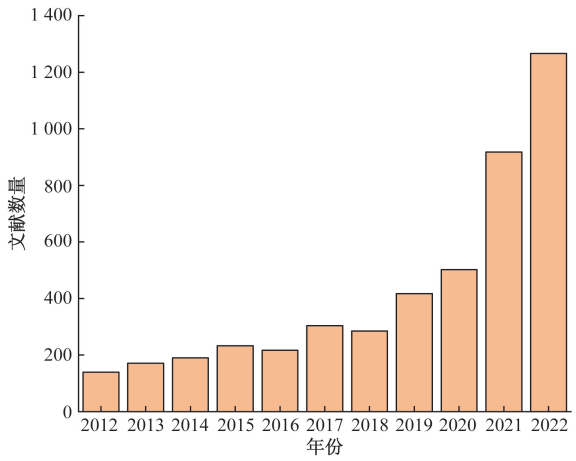
1 异常定义

在一般情况下,异常被定义为与正常状态不同的情况。视频异常可以理解为与预期外观或运动行为不符的事件,或者预期外观或运动行为在不正常的位置或时间发生。然而,视频中异常的定义与特定场景相关,一个场景中的异常活动在另一个场景中可能是正常的[20]。例如,在草坪上行走被视为异常,而在人行道上行走则是正常的。因此,模型的训练需要足够多的正常视频数据来描述各种场景中的正常活动。然而,由于现实世界的多样性和复杂性,几乎不可能收集到所有场景中的异常活动,而标注所有可能的正常事件也需要高昂的成本。

2 无监督视频异常检测(UVAD)

为了全面了解 VAD 方法的发展过程,通过参考数据库和搜索引擎统计了过去 10 年中与视频异常检测相关的学术论文数量,并将统计结果如图 1 所示。从图 1 中可以看出,相关出版物的数量呈稳步增长的趋势。

现有的 VAD 任务主要可以分为无监督视频异常检测方法(unsupervised video anomaly detection, UVAD)和弱监督视频异常检测方法(weakly supervised video anomaly detection, WSVAD)两类。在无监督视频异常检测方法中训练集仅包含正常事件活动,而测试集包含训练集中的正常事件和不在训练集中的异常事件。UVAD 方法通过比较异常样本与正常样本之间的差异来区分测试集中的正常和异常事件。在训练过程中,UVAD 方法学习了正常事件的时空特征边界,将边界之外的测试样本视为异常。其基本逻辑是,如果无法用已学习到的大量正常事件样本来表示某个样本,那么该样本就被认



数据来源于 SpringerLink 数据库

图 1　关于视频异常检测主题出版物的统计数据

Fig. 1　Statistical data on video anomaly detection publications

为是异常的。如图 2 所示,UVAD 中两阶段的视频异常检测方法。通过正常视频数据训练深度学习网络模型,在测试阶段计算测试数据的异常分数。公式为

$$e = D[m(x_{\text{train}}), x_{\text{test}}] \tag{1}$$

式(1)中:x_{train}为训练样本;x_{test}为测试样本;D 为测试样本 x_{test}和训练后的深度学习模型 m 之间的误差。常用的误差计算方法有预测结果的均方误差(mean square error, MSE)和空间特征中的 L2 距离等[21-22]。在基于重构[23-24]和基于预测[25-26]的深度学习方法中明确要求训练数据都要是正常数据,且每个测试视频数据中必须含有异常数据。在计算异常分数时,通常将异常数据归一化到[0, 1]的相对值,其中分数越大表示偏差越大,异常越明显。

传统的 UVAD 方法通常依赖于手工特征,过度依赖人的先验知识,导致检测结果的代表性较差。然而,近年来随着深度学习在计算机视觉领域的应用,最新的 UVAD 方法更倾向于使用自动编码器(autoencoder, AE)[23,27]、生成对抗网络(generative adversarial network, GAN)[28-29]和视觉转换器(visual transformer, ViT)[30-32]。根据输入数据的结构,UVAD 方法可以分为帧级和对象级两种。

2.1　帧级方法

在深度学习中,卷积神经网络可以直接提取潜在特征进行表示。帧级方法中常用完整的 RGB(red, green, blue)帧、帧序列和光流帧作为输入。现有的帧级方法可分为单流模型和双流模型。单流模型以原始视频帧为输入,引入 3D 卷积[33]和 U-net[25]等表征学习器,通过重构输入帧或者预测未来帧来学习视频序列的时空特征。双流模型则通过分别学习视频帧的外观和运动信息,探索原始视频帧的时空分离特征,并利用时空相关性[27]和时空一致性[34]进行异常检测。

2.1.1　单流模型

Xu 等[35]提出了基于自动编码器的方法,通过学习视频的外观和运动信息,并将其结合用于异常检测。Hu 等[36]构建了深度增量慢特征分析网络,提取全局高级特征以检测全局异常。Hasan 等[23]利用卷积自动编码器(convolutional autoencoder, ConvAE)学习时空局部特征,通过重构误差进行异常识别。Luo 等[37]提出了一个 ConvLSTM-AE 模型,结合了卷积神经网络和长短期记忆网络(long short term memory, LSTM)的优势,对视频帧进行编码,并识别出异常事件的外观和运动信息。为了削弱 AE 模型对异常样本的泛化能力,Gong 等[24]在编码器和解码器之间引入了内存记忆网络,构建内存增强自动编码器(memory-enhanced autoencoder, memAE)用来学习正常样本的表示。随后,Park 等[22]引入了基于注意力的内存寻址机制,并提出新的内存更新方法,以确保模型能更好地学习和表示正常事件。

Liu 等[25]提出了一种未来帧预测的方法,利用 GAN 网络自适应地学习视频的动态特征。Luo 等[21]在未来帧预测框架中引入元学习增加模型的泛化能力,更快地适应新的学习场景。同时确定了未来帧预测 UVAD 网络的设计规则。Chen 等[38]构建了双向预测框架(Bi-Pre),通过前向和后向预测子网络预测真实帧,并使用滑动窗口方案计算异常分数,更加关注于预测前景目标。Cai 等[28]采用多路径的 ConvGRU 记忆时间特征,并充分利用时间约

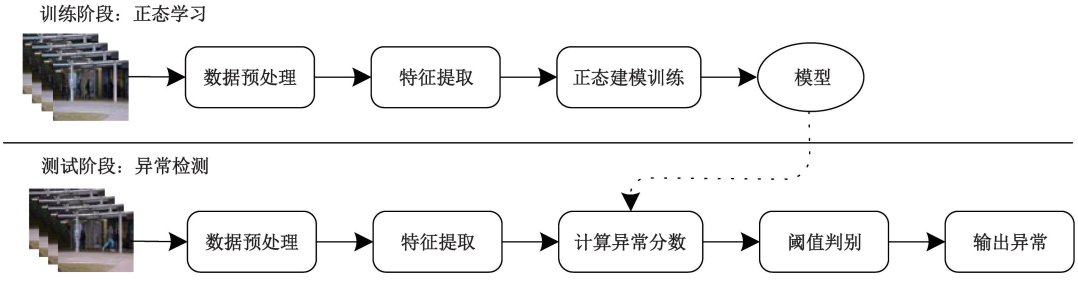


图 2　两阶段视频异常检测方法

Fig. 2　Two-stage video anomaly detection method

束信息进行预测。随着 GAN 网络的应用,Yu 等[39]使用 GAN 来学习数据的正态性,提出了对抗性事件预测(adversarial event prediction,AEP)网络,以抑制对过去事件的表征学习,并强制学习预测未来事件,从而探索其相关性。Zhao 等[40]使用时空长短期记忆(spatiotemporal long short term memory,ST-LSTM)网络提取和记忆外观和运动变化,同时受 GAN 网络的启发,引入一个鉴别器与 ST-LSTM 进行对抗学习,以增强模型对连续视频帧之间时空相关性的学习。刘慧等[41]通过在 SRGAN(super-resolution generative adversarial network)网络中引入 Dropout 层,增加网络的鲁棒性和稳定性,提高重构视频帧质量,增强检测效果。

为了解决深度重构方法中正则性分差较小的问题,Ye 等[42]提出的 AnoPCN 模型将重构分解为预测和细化问题,引入误差细化模块来重构预测误差并细化预测编码模块生成的粗略预测,将重构和预测方法统一到端到端框架中。此外,Liu 等[43]和 Tang 等[44]也进行了类似的工作。

受基于稀疏编码异常检测[45]方法的启发,Luo 等[46]提出了时间相干稀疏编码方法,使用重构过程中学习到的相似系数对相邻帧进行编码,并通过顺序迭代软阈值算法(sequential iterative soft-thresholding algorithm,SIATA)优化稀疏系数。Deepak 等[47]将残差连接[48]引入到自动编码器中,提出了残差时空自动编码器(residual spatiotemporal autoencoder,R-STAE),可以端到端地进行训练,解决了训练过程中的梯度消失问题。

2.1.2 双流模型

双流模型是指具有两个数据来源的深度学习模型。视频中的异常情况通常可分为外观异常和运动异常,即空间异常和时间异常。由于单流模型在提取和处理时空信息的过程中可能存在信息丢失的问题[49],因此提出双流网络分别对视频的空间和时间信息进行建模,从而更好地提取和利用视频中的时空信息。此外,视频中时空信息的一致性和相关性近年来也备受关注。如表 1 所示为近年来双流模型在 UVAD 领域所取得的一些成果。

随着时空自动编码器(spatiotemporal autoencoder,STAE)的应用,Zhao 等[40]首次将 STAE 方法用于 VAD 任务,该方法用 3D 卷积从视频的时间和空间维度中提取特征,使用两个解码器输出重构帧和预测帧。除了常规的重构损失外,还引入了一种权重递减预测损失,用于生成未来帧,从而降低正常事件的重构误差。Li 等[50]提出双流 DSTAE 框架,以 RGB 帧和光流帧为输入,采用联合重建误差策略来融合时间流和空间流,从而增强相邻帧间连续性信息时空特征的提取能力。Gunale 等[51]使用 CNN

表 1 双流 UVAD 方法

Table 1 Dual-stream UVAD method

方法	骨干网络	优缺点
STM-AE[26]	AE,GAN	优点:引入时间-空间记忆模块,记录正常视频特征,通过对抗学习探索相关性 缺点:计算成本高
AMAE[27]	AE	优点:基于通道注意力的时空解码器有效融合时空特征 缺点:训练成本较高,依赖光流
MSM-NET[28]	GAN	优点:多路径记忆增强时间特征,避免使用光流信息,降低计算成本 缺点:对数据质量依赖性较高
DGGAN[29]	GAN	优点:双生成器结构,同时生成伪异常帧和正常帧 缺点:忽略时序信息的影响
STAE[33]	AE	优点:利用 3D 卷积学习时空特征 缺点:双解码器需要的计算成本较高
AMMC-Net[34]	AE	优点:探索时空特征的一致性 缺点:缺少对特征比较方法的分析
DSTAE[50]	AE, ConvLSTM	优点:引入光流增强帧间信息提取能力 缺点:忽略了背景信息的影响
HOME-FAST[51]	AlexNet	优点:使用光流估计方法捕捉运动信息 缺点:仅用一个二分类器,无法捕捉复杂多样的异常
闫善武等[52]	AE	优点:使用双流网络结合行人的时空信息有效地捕捉视频中的动态变化 缺点:训练成本较高,背景噪声干扰
AMC[53]	ConvAE,U-net	优点:共享编码器学习外观和运动的对应关系 缺点:单帧输入没考虑视频帧的时序信息
CDDAE[54]	AE	优点:使用两个 AE 分别学习时间和空间特征 缺点:聚类训练成本较高

和光流估计分别提取外观特征和运动特征,以更好地处理快速运动和遮挡等复杂情况。闫善武等[52]引入改进 AE,结合时空双流网络,实现了对连续帧差图和视频帧的预测,并通过融合后的最终帧进行异常检测。除了时空分离的研究外,还有学者致力于探索时空特征的一致性,如 Nguyen 等[53]提出一种双解码器的 AE 模型,其中一个解码器用于重构输入帧,另一个用于预测对应光流帧,以学习输入视频帧的外观和运动之间的对应关系。Chang 等[54]用两个 AE 分别提取视频帧的时间和空间特征,并构建 RGB 差分来模拟生成光流,以捕捉时间特征。随后,采用 K 均值聚类强制提取紧凑的时空特征用于异常检测。此外,Liu 等[26]将内存增强引入到双流自动编码器中,用来记录正常视频帧的外观和运动特征。并通过与鉴别器进行对抗学习探索时空特征间的相关性。另外,Qi 等[29]提出一个双生成器框架,同时生成伪异常帧和正常帧进行对抗学习,并引入二阶通道注意力机制增强模型对异常分布的学习能力。

2.2　对象级方法

随着拍摄设备的更新,高清视频数据包含的信息越来越丰富。这也极大地推动了目标检测模型[55-58]性能的快速提升,能够从复杂的背景信息中准确提取前景目标。近年来,异常检测研究员发现相邻几帧的背景变化很小,但又包含大量不相关信息,因此,依靠先进的目标检测方法,提出了对象级 UVAD[59]。该方法在预训练阶段使用目标检测模型提取感兴趣的前景目标,并忽略冗余的背景噪声,专注于前景目标的时空信息。现有的研究表明对象级 UVAD 方法的性能要优于帧级方法,尤其在多场景数据集[25]上的表现更为突出。如表 2 所示,总结了对象级方法的检测器、分类依据和贡献。

Hinami 等[60]提出了一种集成检测和描述视频异常的方法,利用多任务学习 CNN 获取异常相关的语义信息,并将其嵌入到与环境相关的异常检测器中,生成人类可理解的事件描述,实现了异常检测与事件描述的联合优化,从而提高准确性。Amraee 等[61]利用 HOG-LBP 和 HOF 来描述区域的外观和运动特征,然后使用两个不同的 SVM 模型来检测外观异常和运动异常。Ionescu 等[62]提出的 OC-AE 模型首先对选定对象的运动和外观信息进行编码,将训练数据中的正常样本进行聚类,然后使用一对多的异常事件分类器来区分出异常样本。

异常事件的定义与场景有很大的关系,所以 Sun 等[63]提出了一种基于场景感知的上下文推理 UVAD 方法,利用视觉特征中的上下文信息弥合了视觉上下文与异常事件含义之间的语义鸿沟,使得异常检测不受目标遮挡和跟踪失败的影响。同时,还引入了基于图的深度高斯混合模型进行场景聚类,以便推理时空上下文关系。为了充分学习时间上下文信息,Yu 等[64]通过视觉完形填空测试来提高模型对高层语义和时序上下文信息的利用能力,精确全面地定位视频中的活动区域,推理已删除的补丁来恢复原始视频,从而实现异常检测。Georgescu 等[59]将 OC-AE[62]方法推广到 Background-Agnostic 框架,该框架通过实例分割只关注前景对象,适合于在正常事件定义一致而背景不同的场景下检测异常事件。通过生成与场景无关的异常样本对二分类器进行对抗学习,从而降低了对真实异常样本的需求。Liu 等[43]提出了 HF^2-VAD 混合框架,无缝集成了重构和预测两种视频异常检测范式。该框架使用多级内存增强模块记录正常样本的光流重

表 2　对象级 UVAD 方法

Table 2　Object-level UVAD method

年份	方法	检测器	分类依据	贡献
2017	LDGK[60]	Fast-RNN	异常分数	联合 CNN 和环境异常检测器来检测视频异常
2018	OC-SVM[61]	HOG-LBP	异常分数	结合 HOG-LBP 和 HOF 来描述提取区域的外观和运动,一分类 SVM 判断异常
2019	OC-AE[62]	SSD	一对多二分类	以对象为中心的 AE 编码外观和运动,使用分类器判别异常
2020	SACR[63]	RPN	分类异常分数	基于场景感知的上下文推理方法,利用视觉特征中的上下文信息进行视频异常检测
2020	VEC[64]	R-CNN	重构误差	提出视频事件完成方法,利用高级语义和时间上下文信息进行异常检测
2021	Background-Agnostic[59]	YOLOv3	二分类	使用一组 AE 提取前景对象的运动和外观特征,通过生成伪异常样本进行对抗学习训练二分类器区分异常样本
2021	HF^2-VAD[43]	Cascade R-CNN	预测误差	无缝集成重构和预测框架进行视频异常检测
2022	HSNBM[65]	Cascade R-CNN	预测误差	提出一个前景背景分层绑定模型,从全局到局部捕捉异常现象
2022	AU-Net[66]	YOLOv3	预测误差	提出基于注意力的 U-Net 网络,实现对象级异常检测和定位
2023	OSIN[67]	Faster R-CNN	预测误差	提出以对象为中心的场景推理网络,以对象-场景交互进行建模

构信息,并利用条件变分自动编码器捕获原始视频帧和重构光流帧之间的相关性,以预测未来帧并检测异常。Bao 等[65]提出了分层场景正常性绑定(hierarchical scene normality-binding modeling, HSNBM)模型。该模型从全局到局部地解析视频场景,并进行场景重构和物体预测。同时,利用前景和背景之间的关系进行未来帧预测。Fang 等[66]提出了一种基于注意力的 U-Net 网络,能够在无监督的情况下实现视频异常检测和定位,同时提高合成图像的质量并减少误报。Liu 等[67]使用一个三流结构学习全局场景和特定的局部对象,同时利用场景内存记忆网络来探索对象和场景之间的交互,实现异常检测和定位,然而模型没有考虑视频中的时空关系和上下文信息,限制了模型在复杂场景中的检测效率。

3 弱监督视频异常检测

弱监督视频异常检测在训练阶段使用的标注数据相对较少,只需要视频级标签或局部标注。与需要精确标注每个视频帧的方法相比,WSVAD 方法更加灵活,能够从具有噪声标注的数据中学习到异常。Sultani 等[68]首先提出使用弱语义视频级标签来训练模型,并发布收集了 13 类现实世界犯罪行为的 UCF-Crime 数据集,该数据集提供了训练集的视频级标注,为 WSVAD 的研究奠定了基础。在后续的研究中,Wu 等[69]收集的 XDViolence 数据集将音频引入到数据集中,形成了第一个音频视频数据集。这一进展将异常检测从单模态视频理解扩展到多模态的信号处理。为了详细的介绍现有的 WSVAD模型,本文将现有 WSVAD 方法分为单模态的多实例学习方法[70-72]和多模态方法[73-75]两类。

3.1 多实例学习方法

多实例学习(multiple instance learning,MIL)框架[68]通常将视频分割成固定长度的片段。将每个片段视为一个实例,而视频中的所有片段形成一个具有相同视频级标签的包。每个实例都有一个异常得分,用于表示该实例是否包含异常。然后使用 C3D[76]、I3D[77]和时间段网络(time slot network, TSN)[78]等预训练特征提取器来提取时空特征。最后,使用异常得分回归器来确保异常实例的异常得分高于正常实例的得分,并以此来定位和识别异常。如表 3 所示,总结了现有的多实例学习方法的特征提取器、决策依据和贡献。

Sultani 等[68]使用预训练的 C3D 特征捕捉视频的时空信息,并采用三层全连接网络作为回归器来预测异常分数。为增强 MIL 排名损失函数,引入了稀疏性和时间平滑约束。Zhu 等[79]提出增强时空网络学习有效的运动感知特征,并将注意力模块和时间信息的上下文关系与 MIL 排序相结合,提高检测性能。Zaheer 等[80]提出一种自我推理训练方法,使用时空视频特征的二进制聚类来生成伪标签,来监督 MIL 模型,减少噪声干扰。Zaheer 等[81]提出随机批次训练策略减少批次间的相关性,还引入正常区域异常得分抑制机制,并提出聚类距离损失函数鼓励模型生成不同的正常和和异常聚类。Tian 等[82]通过训练特征幅值学习函数识别正实例的鲁棒时间特征幅值(robust temporal feature magnitude learning,RTFM)学习,还采用扩张卷积和自注意力机制捕捉长短程时间依赖,提高特征学习准确性。为解决 WSVAD 无法对长期上下文信息进行建模的

表 3 多实例学习方法

Table 3 Multi-instance learning method

年份	方法	特征	分类依据	贡献
2018	MIR[68]	C3D	MIL 排序	使用视频级标签来监督 MIL 回归模型计算帧级异常分数
2019	MAF[79]	VGG16,C3D,I3D	MIL 排序	提出时空增强网络学习 MIL 排名模型的运动感知特征
2020	SRF[80]	C3D	MIL 排序	使用聚类算法生成二进制伪标签进行训练
2020	CLAWS[81]	C3D	MIL 排序	提出聚类距离损失函数鼓励模型生成不同的正常和异常聚类
2021	RTFM[82]	C3D I3D	MIL 排序	提出 RTFM 来采用扩张卷积和自注意力捕捉长短程时间依赖,更准确地学习特征
2021	WSAL[83]	I3D	MIL 排序	有效利用时空上下文信息实现弱监督视频异常定位,使用增强策略以消除噪声干扰
2021	MIST[70]	C3D I3D	MIL 排序	提出了一种新颖的多实例自训练框架
2022	MSL[30]	C3D I3D VideoSwin-RGB	MIL 排序	使用自训练模块生成伪标签,此外还使用了一个 transformer 模块来捕获视频中的长期依赖关系
2022	WAGCN[84]	I3D	异常分数	将自适应图卷积网络应用于 WSVAD
2022	DTED[71]	C3D	MIL 排序	提出深度时间编码方法学习时间特征,使用联合损失进行优化
2022	STA[72]	C3D I3D	MIL 排序	提出循环交叉注意力探索时空表征之间的关系
2023	NTCN-ML[85]	I3D + TCN	MIL 排序	提取时间表征构建时间序列优化多实例学习过程
2023	HSN[86]	I3D	MIL 排序	提出一种人景网络(HSN),以自校正损失函数进行优化

问题,Lü 等[83]提出的弱监督定位方法,融合时间和空间上下文进行异常检测,使用高阶上下文编码模型来测量时间动态变化。此外,还收集了一个用于交通异常检测的数据集(TAD)。Cao 等[84]提出弱监督自适应图卷积网络(weakly supervised adaptive graph convolutional network,WAGCN),通过构建特征相似性和时间差异的全局图和图学习层来增强时间特征。Shao 等[85]提出 NTCN-ML 模型,提取时间表示构建时间序列优化多实例学习过程,平衡稀有正负实例之间的特征关联。Majhi 等[86]提出一种人类场景网络,通过解耦设计有效地学习人类和场景中心异常的局部和全局判别表示。并使用自校正损失函数增强了类别之间的可分离性,使复杂异常场景分析更为细致。周文浩等[87]提出了一种基于双重动态记忆网络的弱监督视频异常检测方法,通过动态可变的记忆网络记录视频中长期的正常和异常模态,可以根据场景的需求自适应调整记忆项数目和采用模态分离损失,实现准确的在线检测异常情况。

3.2　多模态学习方法

多模态 WSVAD 模型利用光流、音频和视频等多模态信息,挖掘与异常相关的有效线索,以增强视频片段的特征表示。由于多模态数据采集的非常困难,现有的模型大多基于 XD-Violence 数据集[69],专注于视频和音频信息的融合,用于检测暴力行为。如表 4 所示,展示了多模态学习方法的数据类型和创新点。

Wu 等[69]在 2020 年发布了第一个多模态视频异常检测数据集 XDViolence,此外还提出了一个并行三分支网络 HL-Net,以捕获视频片段之间的关系并整合特征。实验证明多模态数据对 WSVAD 有着积极影响。与 HL-Net 相比,Pang 等[73]专注于探索视频和音频的融合方法,利用双线性池化方法整合特征,并鼓励它们相互学习以生成更具代表性的特征。Pu 等[74]探索了音频引导的注意力网络,使用跨模态感知局部唤醒网络增强从音频到视频的特征,捕获高级语义特征。Xiao 等[75]研究了光流特征在多模态暴力检测中的作用,提出双分支光流感知的多模态融合网络,并提供了 3 种融合策略:输入融合、基于注意力的中途融合和基于光流感知的分数融合。为了解决现有多模态数据集不足和模型融合效率底下的问题。Shang 等[88]提出了相互蒸馏方法,将大型数据集中的有效信息转移到小型数据集中。此外,还提出多模态注意力融合网络来融合特征,以获取更有区分性的表征。然而,现有模型大多在特征层面进行融合,无法充分利用多模态信息的互补性。因此,Wei 等[89]提出多模态监督-注意增强融合框架,细化视频标签并生成伪标签来优化模型的预测结果,从而促进不同模态之间的隐式对齐。为了实现实时暴力检测,Liang 等[90]利用交叉注意力图卷积模块提取跨模态时空特征,再用双向门循环单元(bidirectional gate recurrent unit,Bi-GRU)模块捕获时间上下文特征,从而在很小的参数量和推理时间的基础上实现异常检测。为解决相似运动模式的难以区分问题,付荣华等[91]提出一种结合 RGB 模态和骨架模态的方法,有效加强了特征对应关系,并利用时间自注意力捕获全局信息,实现了对相似运动模式下的异常行为更准确的检测。

4　性能比较

如表 5 所示,展示了收集的现有方法在公开数据集[25,68,92-93]上的表现,现有的 UVAD 方法主要应用 AUC 值作为评价 VAD 模型的指标。使用频率表明 UCSD Ped2[92]、CUHK Avenue[93] 和 Shanghai Tech[25]已经成为 UVAD 方法评估的常用数据集基准,未来的 UVAD 工作应该优先考虑在这 3 个数据集上测试比较所提方法的性能。

表 4　多模态学习方法

Table 4　Multimodal learning methods

年份	方法	数据类型	贡献
2020	HL-Net[69]	视频 + 音频	收集 XDViolence 暴力检测数据集并提出用于多模态异常检测的三分支神经网络模型
2021	FVAI[73]	视频 + 音频	提出了一个包含 3 个模块的神经网络来融合视频和音频信息的方法
2022	AGAN[74]	视频 + 音频	使用跨模态交互从时间维度上增强视频和音频特征,使用时间卷积层获取高置信度的暴力分数
2022	OFAB[75]	视频 + 音频 + 光流	利用注意力对不同类型特征进行融合,并提出了 3 种不同的融合策略
2022	MSAF[88]	视频 + 音频 视频 + 光流	提出多模态标签细化以将视频级地面真实性细化为伪剪辑级标签,并将多模态信息与多模态监督-注意融合网络隐式对齐
2022	MD[89]	视频 + 音频 + 流	用多模态注意力融合网络将视频与音频和流特征进行融合
2023	VioNets[90]	视频 + 音频 + 光流	提出一个包含交叉注意力图卷积网络和双向门循环单元的模型,实现实时暴力检测

表6 展示了 WSVAD 方法在 ShanghaiTech weaky[95]、UCF-Crime[68]和 XD-Violence[69]数据集上的性能表现。WSVAD 模型需要使用预训练特征提取器获取特征表示,从表6中可以看出,常用的特征提取器有 C3D 和 I3D,且同一个模型依靠 I3D 提取特征器的性能要比依靠 C3D 的好。因此在未来的 WSVAD 模型比较中应当使用常用的 C3D 和 I3D 特征提取器,以证明性能的提升是来自于模型的设计,而不是更强大的特征提取器。

表5 UVAD 方法性能比较

Table 5 UVAD method performance comparison

年份	方法	AUC/%			
		Ped1	Ped2	Avenue	ShanghaiTech
2015	AMDN[35]	92.1	90.8	—	—
2016	Conv-AE[23]	75.0	85.0	80.0	—
2017	ConvLSTM-AE[37]	75.5	88.1	77.0	—
2017	LDGK[60]	—	92.2	—	—
2018	FFP[25]	83.1	95.4	85.1	72.8
2019	AnoPCN[42]	—	96.8	86.2	73.6
2019	memAE[24]	—	94.1	83.3	71.2
2019	sRNN-AE[46]	—	92.2	83.5	—
2019	AMC[53]	—	96.2	86.9	—
2019	OC-AE[62]	—	97.8	90.4	84.9
2020	SACR[63]	—	—	89.6	74.7
2020	VEC[64]	—	97.3	89.6	74.8
2020	CDDAE[54]	—	96.5	86.0	73.3
2020	MNAD[22]	—	97.0	88.5	70.5
2020	PARAD[44]	82.6	96.2	83.7	71.5
2020	Bi-Pre[38]	89.0	96.6	87.8	—
2021	HF^2-VAD[43]	—	99.3	91.1	76.2
2021	ConvGRU[28]	—	96.8	87.3	74.2
2021	AEP[39]	97.92	97.31	90.2	—
2021	R-STAE[47]	—	83.0	82.0	—
2021	F^2PN[21]	84.3	96.2	85.7	73.0
2021	DSTAE[50]	97.6	97.2	96.3	—
2021	AMMC-Net[34]	—	96.6	86.6	73.7
2022	STM-AE[26]	—	98.1	89.8	73.8
2022	AMAE[27]	—	97.4	88.2	73.6
2022	ST-LSTM[40]	—	96.7	87.8	73.1
2022	HSNBM[65]	—	95.2	91.6	76.5
2022	AU-Net[66]	—	—	—	71.0
2023	OSIN[67]	—	98.3	91.7	79.6
2023	DGGAN[29]	85.7	97.9	86.2	—
2023	DMAD[94]	—	99.7	92.8	78.8

5 基准数据集

在视频异常检测任务的研究中,研究者们为了应对不同场景和异常情况的应用需求,已经创建了多个公开的数据集供研究使用。这些公共数据集为模型测试提供了公正的标准,同时也反映了相关领域的趋势和研究热点。表7展示了现有 VAD 数据集的统计结果,并对它们的属性进行了比较。为了满足不断发展的 VAD 研究需求,研究者们提出了包含不同注释的数据集[25,68,96-97],这些数据集反映了 VAD 研究从无监督到弱监督[68]、从单一场景[92-93]到复杂现实世界[25]的发展变化。

表6 WSVAD 方法性能比较

Table 6 WSVAD method performance comparison

年份(年)	方法	特征	AUC/%		
			Shanghai Tech	UCF-Crime	XD-Violence
2020	SRF[80]	C3D	84.16	79.54	—
2020	CLAWS[81]	C3D	89.67	83.03	—
2021	RTFM[82]	C3D	91.51	83.28	75.89
		I3D	97.21	84.30	77.81
2021	MIST[70]	C3D	93.13	81.40	—
		I3D	94.83	82.30	—
2022	MSL[30]	C3D	94.81	82.85	75.53
		I3D	96.08	85.30	78.28
		VideoSwin-RGB	97.32	85.62	78.59
2022	WAGCN[84]	I3D	96.05	84.67	—
2022	DTED[71]	C3D	87.42	79.49	—
2022	STA[72]	C3D	88.70	81.60	—
		I3D	90.20	83.00	—
2023	NTCN-ML[85]	I3D + TCN	95.30	85.10	—
2023	HSN[86]	I3D	96.22	85.30	—

表7 VAD 数据集

Table 7 VAD datasets

年份	数据集	视频			帧			分辨率	场景	异常事件
		总数	训练	测试	总数	训练	测试			
2008	Subway Entrance[96]	—	—	—	144 250	76 543	67 797	512×384	1	19
2008	Subway Exite[96]	—	—	—	64 901	22 500	42 401	512×384	1	14
2009	UMN[7]	—	—	—	7 741	—	—	240×320	3	11
2010	UCSD Ped1[92]	70	34	36	14 000	6 800	7 200	238×158	1	40
2010	UCSD Ped2[92]	28	16	12	4 560	2 550	2 010	360×240	1	12
2013	CUHK Avenue[93]	37	16	21	30 652	15 328	15 324	360×640	1	47
2018	ShanghaiTech[25]	437	330	107	317 398	274 515	42 883	856×480	13	130
2018	UCF Crime[68]	1 900	1 610	290	13 741 393	12 631 211	1 110 182	240×320	—	950
2020	Street Scene[98]	81	46	35	203 257	56 847	146 410	1 280×720	3	205
2020	XD-Violence[69]	4 754	3 954	800	—	—	—	1 920×1 080	10	4
2020	ADOC[97]	—	—	—	259 123	—	—	1 920×1 080	1	721
2022	UBnormal[99]	40	27	13	236 902	116 087	92 640	1 280×720	—	660

(1) Subway[96]:这是一个较早的数据集,包含两个独立的子数据集 Entrance 和 Exit,记录了地铁入口和出口的行人活动。其中异常事件包括逃票、徘徊和行走方向错误等,并且每个片段都有标注类别。由于标记工作和异常事件定义的不明确性,这个数据集存在一定的质量和可靠性问题,因此现有的大多数研究不再使用它进行模型评估。

(2) UMN[7]:这个数据集包含 11 个短视频,分别在草地、室内大厅和公园 3 个场景拍摄。场景是人为设置的,模拟了人群突然疏散和逃离的异常行为,而不是来自真实的自然拍摄。这些异常现象是人为想象的,忽略了现实世界中异常现象的稀少性和多样性。和 Subway[96] 数据集一样,由于数据集的可靠性存在一定的问题,UMN 最近也已经被研究人员放弃使用。

(3) UCSD[92]:数据集包含 Ped1 和 Ped2 两个子数据集,是大学校园中安装在高处的固定摄像头俯瞰人行道拍摄收集而来,具有简单但真实的场景,是目前应用最广泛的 VAD 数据集之一。它们体现了 VAD 在公共安全领域的应用价值。其中,Ped1 数据集的视角垂直于道路,移动物体的大小会随空间位置的变化而变化。Ped2 数据集由平行于道路方向的相机拍摄而来。数据集将场景中的人行走定义为正常事件,其他行为和物体被视为异常事件,如骑自行车、滑板和驾驶汽车等。由于该数据集具有经典的场景和易于理解的异常事件,已经成为现有研究中广泛使用的数据集之一,其帧级 AUC 已经高达 99.7%[94]。然而,单一简单场景下的数据集限制了 VAD 相关技术的研究和发展。因此,发展跨场景、多模态和大规模的复杂数据集已经成为 VAD 数据集发展的必然趋势。

(4) CUHK Avenue[93]:数据集包含 47 个异常事件,包含外观异常和运动异常,例如投掷物品、游荡和跑步。人的大小可能会因为相机的位置和角度而改变,且该数据集提供像素级和帧级的空间标注。与 UCSD[92] 数据集一样是现有研究中广泛使用的数据集之一。

(5) ShanghaiTech[25]:鉴于 UCSD[96] 和 CUHK Avenue[97] 数据集中仅仅考虑了单一场景中的异常事件,与现实世界中多场景、多视角的异常事件需求不匹配。因此,上海科技大学团队提出了 ShanghaiTech 数据集,它包含 13 个场景 130 个异常事件的,是目前最大的无监督视频异常检测基准数据集。该数据集将异常行为定义为与正常行走有区别的行为,如骑自行车、追逐和争吵等。此外,数据集还提供了异常事件的像素级地面实况注释。

(6) UCF Crime[68]:数据集包含 1 900 个未经剪辑的真实世界监控视频,是第一个弱监督视频异常检测(WSVAD)数据集。其中包含 950 个异常事件,可以分为 13 个类,包含盗窃、抢劫、交通事故等现实世界常见的异常事件。与其他 VAD 数据集不同,UCF Crime 数据集的训练集和测试集都包含异常事件视频,并且每个视频都有视频级标签,其中 0 表示正常,1 表示异常。WSVAD 数据集的异常事件预定义与特定的场景有关,在检测异常事件时更能反应真实世界发生异常时的情况,具有更好的应用潜力。

(7) ADOC[97]:数据集通过大学校园内的监控摄像头拍摄而来,连续记录了 24 h 1 080 p 的视频,涵盖了多个时间段和场景,具有较高的真实性和代表性。根据不同场景,将视频数据从低频到高频进行注释,共注释了 721 个异常事件,例如举牌子、人群聚集等。

(8) Street Scene[98]:数据集由静态 USB 摄像头俯视拍摄一条有自行车道和人行道的双车道街景场景,35 个测试集中包含 205 个异常事件,可分为 17 个不同类型的异常,如横穿马路、闲逛和非法停放的汽车等。该数据集的视频分辨率高,捕捉到的场景具有多种多样没有人为干预的活动,均是自然发生的。每个异常事件都被标记为边界框的轨迹,在每一帧中可以标记多个异常。

(9) XD-Violence[69]:数据集是第一个音频视频数据集,将异常检测从单模态视频理解扩展到多模态的信号处理。XD-Violence 专注于暴力行为的检测,如枪击、爆炸、打架、车祸、虐待和骚乱等。为了扩大数据集规模,除了真实世界的监控视频外,还包括一些电影片段。每一个视频都有视频级标签,且标记了测试视频中每个暴力事件的开始帧和结束帧,提供了帧级标注。

(10) UBnormal[99]:是第一个包含虚拟场景的 VAD 数据集,这得益于合成数据在视频开发中的应用。由于是合成的数据,可以对视频数据进行像素级标注。因此 UBnormal 是用于监督开放集视频异常检测的基准。它能够公平地比较开放集和封闭集模型,并缓解现实数据集中缺乏训练异常数据的问题。

6　评价标准

现有的 VAD 方法的性能评估可以从检测精度和运行成本两个方面进行。检测精度是衡量模型区分异常事件的能力,而运行成本则关注模型在有限资源设备上的部署潜力。根据异常检测的数据类型,检测精度可分为三个级别:面向对象检测、面

向时间检测和面向空间定位检测。其中面向对象检测涵盖对象级、区域级和轨道级,侧重于检测特定异常对象的轨迹。面向时间检测就是帧级检测,是模型在不定位异常像素的情况下确定异常事件的时间位置关系。而面向空间定位检测则是定位异常事件的异常像素位置。运行成本的评估标准包括模型参数大小、浮点运算数和推理速度等。

在模型评估过程中,通常是将模型的预测结果与实际标签进行比较。然而,某些模型的预测结果不是离散的0或1,而是介于0~1的连续值。为了评估模型的性能,需要选择一个阈值来进行异常判断。将预测异常分数低于阈值的样本视为正常样本,反之,将高于阈值的视为异常的。为了解决视频异常检测问题面临的挑战,引入混淆矩阵[45],将其作为二元分类问题的解决方案。混淆矩阵中的TP、FN、FP、TN分别表示正确检测异常、异常样本误检为正常、正常样本被误检为异常、正常样本检测为正常。而真阳性(true positive rate,TPR)、假阳性(false positive rate,FPR)、真阴性(true negative rate,TNR)和假阴性(false negative rate,FNR)[100]的定义为

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{1}$$

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \tag{2}$$

$$\mathrm{TNR}=\frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}} \tag{3}$$

$$\mathrm{FNR}=\frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}} \tag{4}$$

可以用来计算接收者操作特征曲线下面积(area under the receiver operating characteristic,AUROC)[101]和平均精度(average precision,AP)[102]。

AUROC[101]:ROC曲线是以FPR为横坐标和TPR为纵坐标,计算多组阈值下的FPR和TPR绘制而成,如图3所示。ROC曲线与横坐标围成区域的面积称为AUC值,常用于评价二分类任务。AUC的值在[0,1],且AUC的值越大表示模型的性能越好。曲线可以很好地展示模型的性能,而且可以帮助选择最佳的阈值。

平均精度(AP)[102]:由于VAD任务数据集中的正负样本分布不平衡,正常样本远多于异常样本,因此,认为由精确率和召回率绘制的精度-召回率曲线下面积更适合用于评估VAD模型性能。其中精确率表示在所有检测结果中,真是异常事件的比例。召回率表示在所有真是异常事件中,被检测出来的比例。此外,还有F_1[103]和错误率等异常检测性能评估指标。

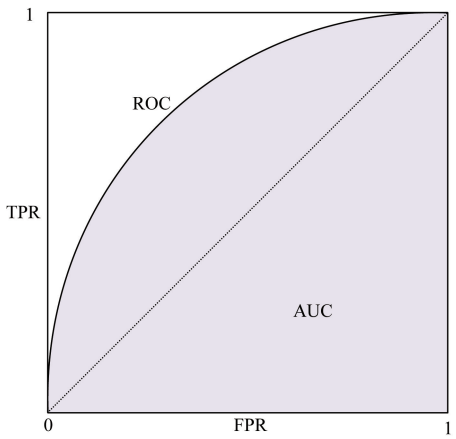


图3 VAD模型性能评价标准ROC曲线示意图

Fig. 3 Schematic diagram of VAD model performance evaluation standard ROC curve

7 挑战与发展趋势

7.1 挑战

7.1.1 数据挑战

VAD领域面临许多的数据挑战。首先,异常数据的模糊性和场景依赖性导致异常的定义没有统一和客观的标准。不同的场景和应用可能有不同的异常定义,甚至同一事件在不同的场景下也可能具有不同的异常属性。其次,异常数据的稀疏性和多样性使得有监督方法难以训练出有效的分类器。实际的视频监控数据中,正常事件或行为占据了绝大多数,而异常事件或行为则非常少见,导致数据不平衡的问题,此外,异常事件或行为是多样性和不确定性的,无法穷举所有可能的情况。最后,监控视频数据存在噪声和隐私问题。由于视频监控数据通常是由摄像头设备在复杂环境中采集得到的,因此可能存在各种噪声和干扰,如光照变化、遮挡、运动模糊等[104]。这些噪声和干扰会影响视频数据中有效信息的提取和表示,降低视频异常检测模型的准确性和鲁棒性。另外,由于视频监控数据涉及人们的面部、身份、行为等隐私信息,很多数据集并不公开或开源,这限制了视频异常检测模型的训练和评估。

7.1.2 模型挑战

VAD模型面临许多挑战。首先,模型泛化能力不足。尽管深度学习在VAD领域取得了显著进展,但现有的模型大多是在单场景的数据集上进行的训练,因此在特定场景下表现良好。然而现实世界是由多个场景组成的,现有模型难以应对多样化的异常行为和复杂的环境。其次,VAD往往要求实时性和准确性兼顾,要想将VAD模型应用到实际检测任务中,需要模型实时响应视频中的异常事件,并

且及时的标注和保存异常片段,降低传输和存储成本。然而现有公开数据集训练的 VAD 模型大多追求检测性能的提高,忽略了模型检测的实时性。另外,VAD 模型的可解释性较差,难以解释其内部的决策过程,为人们提供直观的理解。这在一定程度上限制了模型在某些敏感领域,如医疗、司法等的应用。

7.2　发展趋势

7.2.1　数据发展趋势

随着数据获取能力的增强,VAD 数据集正在向多场景、大规模的现实场景发展。新数据集提供了更真实、更丰富的场景和异常事件,有助于提升模型的泛化能力和鲁棒性。互联网的发展使得在线视频网站上高质量的影视作品越来越多,降低了多场景和多视角视频数据采集的难度。此外,多模态融合在 VAD 研究中的重要性也日益增强,将视频数据与其他传感器数据融合可以提供更全面、准确的异常检测。XDViolence[69] 数据集也证明了多模态数据对 VAD 研究的积极影响。

7.2.2　模型发展趋势

近年来 VAD 模型的发展大多趋于两个方向,首先,追求高精度的大模型,为了学习更多的特征信息,VAD 模型变得越来越大,层数不断加深。而过度强大的神经网络可能会导致泛化异常情况和异常丢失。因此,在未来的研究中应该考虑用巧妙的特征表示抑制模型对异常事件的泛化能力,增强对正常事件的特征表示能力。近年来,图学习[105] 和扩散模型[106] 等强大的生成模型可以提供了更有效的常态学习工具。迁移学习和自适应学习的方法能够将知识和模型参数从一个领域迁移到另一个领域,提高 VAD 模型在新场景下的性能和泛化能力。

其次,面向模型部署的轻量级模型也是发展趋势之一。在实际检测任务中,需要将训练好的 VAD 模型部署到终端设备上。由于现实世界场景和异常的多样性,终端设备的算力有限,现有的大多数 VAD 模型都无法实现轻量级部署和实时在线检测。因此,模型压缩[107]、知识蒸馏[108]、迁移学习[109] 和自适应学习[110] 等方法都将推动轻量级 VAD 模型在应用部署方面的发展。

7.2.3　深度学习的应用

深度学习在 VAD 领域的应用取得了显著的成就,通过深度神经网络可以学习到更丰富和抽象的时空特征表示。为了获得更深层次的特征,CNN 变得越来越复杂,预训练的特征提取器也从 C3D 发展到更强大的 I3D,进一步提升了 VAD 性能。同时,新的视觉表示学习模型,如 Transformer 的应用,也极大地推动了 VAD 往大模型领域的发展。

8　总结

本次调查旨在提供基于深度学习的视频异常检测模型的综合概述,包括无监督和弱监督方法,并针对输入数据类型进行分类分析。本文回顾了现有研究成果,收集并介绍了已有的基准数据集和常用评价标准,并列举了各方法在公开数据集上的表现。通过对现有技术的分析和讨论,总结了当前视频异常检测方法面临的研究挑战以及未来的发展趋势。

参　考　文　献

[1] Liu Y, Yang D, Wang Y, et al. Generalized video anomaly event detection: systematic taxonomy and comparison of deep models[J]. ACM Computing Surveys, 2023, 189: 1-38.

[2] Pang G, Shen C, Cao L, et al. Deep learning for anomaly detection: a review[J]. ACM Computing Surveys, 2021, 54(2): 1-38.

[3] Bera A, Kim S, Manocha D. Realtime anomaly detection using trajectory-level crowd behavior learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. New York: IEEE, 2016: 50-57.

[4] Bertini M, Del Bimbo A, Seidenari L. Multi-scale and real-time non-parametric approach for anomaly detection and localization[J]. Computer Vision and Image Understanding, 2012, 116(3): 320-329.

[5] Cong Y, Yuan J, Liu J. Abnormal event detection in crowded scenes using sparse representation[J]. Pattern Recognition, 2013, 46(7): 1851-1864.

[6] Saligrama V, Chen Z. Video anomaly detection based on local statistical aggregates[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2012: 2112-2119.

[7] Mehran R, Oyama A, Shah M. Abnormal crowd behavior detection using social force model[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2009: 935-942.

[8] Antic B, Ommer B. Video parsing for abnormality detection[C]//2011 International Conference on Computer Vision. New York: IEEE, 2011: 2415-2422.

[9] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[10] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[11] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[12] Tran T M, Vu T N, Vo N D, et al. Anomaly analysis in images and videos: a comprehensive review[J]. ACM Computing Surveys, 2022, 55(7): 1-37.

[13] Song Y. Weakly-supervised and unsupervised video anomaly de-

tection[J]. Highlights in Science, Engineering and Technology, 2022, 12: 160-170.

[14] Kiran B R, Thomas D M, Parakkal R. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos[J]. Journal of Imaging, 2018, 4(2): 36.

[15] Ramachandra B, Jones M J, Vatsavai R R. A survey of single-scene video anomaly detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(5): 2293-2312.

[16] Santhosh K K, Dogra D P, Roy P P. Anomaly detection in road traffic using visual surveillance: a survey[J]. ACM Computing Surveys (CSUR), 2020, 53(6): 1-26.

[17] Ramzan M, Abid A, Khan H U, et al. A review on state-of-the-art violence detection techniques[J]. IEEE Access, 2019, 7: 107560-107575.

[18] 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.
He Ping, Li Gang, Li Huibin. Review of video anomaly detection methods based on deep learning[J]. Computer Engineering and Science, 2022, 44(9): 1620-1629.

[19] 申栩林, 李超波, 李洪均. 人群密集度下 GAN 的视频异常行为检测进展[J]. 计算机工程与应用, 2022, 58(7): 21-30.
Shen Xulin, Li Chaobo, Li Hongjun. Progress in GAN video abnormal behavior detection under crowd density[J]. Computer Engineering and Applications, 2022, 58(7): 21-30.

[20] 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.
Zhang Xiaoping, Ji Jiahui, Wang Li, et al. Review of video-based abnormal human behavior recognition and detection methods[J]. Control and Decision, 2022, 37(1): 14-27.

[21] Luo W, Liu W, Lian D, et al. Future frame prediction network for video anomaly detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(11): 7505-7520.

[22] Park H, Noh J, Ham B. Learning memory-guided normality for anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 14372-14381.

[23] Hasan M, Choi J, Neumann J, et al. Learning temporal regularity in video sequences[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 733-742.

[24] Gong D, Liu L, Le V, et al. Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2019: 1705-1714.

[25] Liu W, Luo W, Lian D, et al. Future frame prediction for anomaly detection—a new baseline[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 6536-6545.

[26] Liu Y, Liu J, Zhao M, et al. Learning appearance-motion normality for video anomaly detection[C]//2022 IEEE International Conference on Multimedia and Expo (ICME). New York: IEEE, 2022: 1-6.

[27] Liu Y, Liu J, Lin J, et al. Appearance-motion united auto-encoder framework for video anomaly detection[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2022, 69(5): 2498-2502.

[28] Cai Y, Liu J, Guo Y, et al. Video anomaly detection with multi-scale feature and temporal information fusion[J]. Neurocomputing, 2021, 423: 264-273.

[29] Qi X, Hu Z, Ji G. Improved video anomaly detection with dual generators and channel attention[J]. Applied Sciences, 2023, 13(4): 2284.

[30] Li S, Liu F, Jiao L. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection[C]// Proceedings of the AIAA Conference on Artificial Intelligence. Reston: AIAA, 2022: 1395-1403.

[31] Lee J, Nam W J, Lee S W. Multi-contextual predictions with vision transformer for video anomaly detection[C]//2022 26th International Conference on Pattern Recognition (ICPR). New York: IEEE, 2022: 1012-1018.

[32] 姬晓飞, 赵东阳. 人体检测与异常行为识别联合算法[J]. 科学技术与工程, 2023, 23(8): 3370-3378.
Ji Xiaofei, Zhao Dongyang. Joint algorithm for human body detection and abnormal behavior recognition[J]. Science Technology and Engineering, 2023, 23(8): 3370-3378.

[33] Zhao Y, Deng B, Shen C, et al. Spatio-temporal autoencoder for video anomaly detection[C]//Proceedings of the 25th ACM International Conference on Multimedia. New York: ACM, 2017: 1933-1941.

[34] Cai R, Zhang H, Liu W, et al. Appearance-motion memory consistency network for video anomaly detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Reston: Association for the Advancement of Artificial Intelligence, 2021, 35(2): 938-946.

[35] Xu D, Ricci E, Yan Y, et al. Learning deep representations of appearance and motion for anomalous event detection[J]. arXiv Preprint ArXiv: 1510.01553, 2015.

[36] Hu X, Hu S, Huang Y, et al. Video anomaly detection using deep incremental slow feature analysis network[J]. IET Computer Vision, 2016, 10(4): 258-267.

[37] Luo W, Liu W, Gao S. Remembering history with convolutional lstm for anomaly detection[C] //2017 IEEE International Conference on Multimedia and Expo (ICME). New York: IEEE, 2017: 439-444.

[38] Chen D, Wang P, Yue L, et al. Anomaly detection in surveillance video based on bidirectional prediction[J]. Image and Vision Computing, 2020, 98: 103915.

[39] Yu J, Lee Y, Yow K C, et al. Abnormal event detection and localization *via* adversarial event prediction[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(8): 3572-3586.

[40] Zhao M, Liu Y, Liu J, et al. Exploiting spatial-temporal correlations for video anomaly detection[C]//2022 26th International Conference on Pattern Recognition (ICPR). New York: IEEE, 2022: 1727-1733.

[41] 刘慧, 卢云志, 张雷. 基于 Dropout 改进的 SRGAN 网络 DrSRGAN[J]. 科学技术与工程, 2023, 23(23): 10015-10022.
Liu Hui, Lu Yunzhi, Zhang Lei. Improved SRGAN network DrSRGAN based on Dropout[J]. Science Technology and Engineering, 2023, 23(23): 10015-10022.

[42] Ye M, Peng X, Gan W, et al. Anopcn: Video anomaly detection

via deep predictive coding network[C]//Proceedings of the 27th ACM International Conference on Multimedia. Reston: ACM, 2019: 1805-1813.

[43] Liu Z, Nie Y, Long C, et al. A hybrid video anomaly detection framework *via* memory-augmented flow reconstruction and flow-guided frame prediction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2021: 13588-13597.

[44] Tang Y, Zhao L, Zhang S, et al. Integrating prediction and reconstruction for anomaly detection[J]. Pattern Recognition Letters, 2020, 129: 123-130.

[45] Lu C, Shi J, Jia J. Abnormal event detection at 150 fps in matlab[C]//Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE, 2013: 2720-2727.

[46] Luo W, Liu W, Lian D, et al. Video anomaly detection with sparse coding inspired deep neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(3): 1070-1084.

[47] Deepak K, Chandrakala S, Mohan C K. Residual spatiotemporal autoencoder for unsupervised video anomaly detection[J]. Signal, Image and Video Processing, 2021, 15(1): 215-222.

[48] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE, 2016: 770-778.

[49] Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 6450-6459.

[50] Li T, Chen X, Zhu F, et al. Two-stream deep spatial-temporal auto-encoder for surveillance video abnormal event detection[J]. Neurocomputing, 2021, 439: 256-270.

[51] Gunale K G, Mukherji P. Deep learning with a spatiotemporal descriptor of appearance and motion estimation for video anomaly detection[J]. Journal of Imaging, 2018, 4(6): 79.

[52] 闫善武, 肖洪兵, 王瑜, 等. 融合行人时空信息的视频异常检测[J]. 图学学报, 2023, 44(1): 95-103.
Yan Shanwu, Xiao Hongbing, Wang Yu, et al. Video anomaly detection integrating pedestrian spatio-temporal information[J]. Journal of Graphics, 2023, 44(1): 95-103.

[53] Nguyen T N, Meunier J. Anomaly detection in video sequence with appearance-motion correspondence[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2019: 1273-1283.

[54] Chang Y, Tu Z, Xie W, et al. Clustering driven deep autoencoder for video anomaly detection[C]// Proceedings of Computer Vision-ECCV 2020: 16th European Conference. Berlin:Springer International Publishing, 2020: 329-345.

[55] Girshick R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE, 2015: 1440-1448.

[56] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 779-788.

[57] Xu H, Yao L, Zhang W, et al. Auto-FPN: Automatic network architecture adaptation for object detection beyond classification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2019: 6649-6658.

[58] 赵立新, 邢润哲, 白银光, 等. 深度学习在目标检测的研究综述[J]. 科学技术与工程, 2021, 21(30): 12787-12795.
Zhao Lixin, Xing Runzhe, Bai Yinguang, et al. Review of research on deep learning in target detection[J]. Science Technology and Engineering, 2021, 21(30): 12787-12795.

[59] Georgescu M I, Ionescu R T, Khan F S, et al. A background-agnostic framework with adversarial training for abnormal event detection in video[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(9): 4505-4523.

[60] Hinami R, Mei T, Satoh S. Joint detection and recounting of abnormal events by learning deep generic knowledge[C]//Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE, 2017: 3619-3627.

[61] Amraee S, Vafaei A, Jamshidi K, et al. Abnormal event detection in crowded scenes using one-class SVM[J]. Signal, Image and Video Processing, 2018, 12: 1115-1123.

[62] Ionescu R T, Khan F S, Georgescu M I, et al. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 7842-7851.

[63] Sun C, Jia Y, Hu Y, et al. Scene-aware context reasoning for unsupervised abnormal event detection in videos[C]//Proceedings of the 28th ACM International Conference on Multimedia. Reston: ACM, 2020: 184-192.

[64] Yu G, Wang S, Cai Z, et al. Cloze test helps: Effective video anomaly detection via learning to complete video events[C]//Proceedings of the 28th ACM International Conference on Multimedia. Reston: ACM, 2020: 583-591.

[65] Bao Q, Liu F, Liu Y, et al. Hierarchical scene normality-binding modeling for anomaly detection in surveillance videos[C]//Proceedings of the 30th ACM International Conference on Multimedia. Reston: ACM, 2022: 6103-6112.

[66] Fang J, Zhang X, Yang B, et al. An attention-based U-Net network for anomaly detection in crowded scenes[C]//2022 14th International Conference on Computer Research and Development (ICCRD). New York: IEEE, 2022: 202-206.

[67] Liu Y, Guo Z, Liu J, et al. Osin: Object-centric scene inference network for unsupervised video anomaly detection[J]. IEEE Signal Processing Letters, 2023, 30: 359-363.

[68] Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 6479-6488.

[69] Wu P, Liu J, Shi Y, et al. Not only look, but also listen: Learning multimodal violence detection under weak supervision[C]//16th European Conference on Computer Vision. Berlin:Springer International Publishing, 2020: 322-339.

[70] Feng J C, Hong F T, Zheng W S. Mist: Multiple instance self-training framework for video anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 14009-14018.

[71] Kamoona A M, Gostar A K, Bab-Hadiashar A, et al. Multiple in-

stance-based video anomaly detection using deep temporal encoding-decoding[J]. Expert Systems with Applications, 2023, 214: 119079.

[72] Liu Y, Liu J, Zhu X, et al. Learning task-specific representation for video anomaly detection with spatial-temporal attention[C]// ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York: IEEE, 2022: 2190-2194.

[73] Pang W F, He Q H, Hu Y, et al. Violence detection in videos based on fusing visual and audio information [C]//ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York: IEEE, 2021: 2260-2264.

[74] Pu Y, Wu X. Audio-guided attention network for weakly supervised violence detection[C]//2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE). New York: IEEE, 2022: 219-223.

[75] Xiao Y, Gao G, Wang L, et al. Optical flow-aware-based multimodal fusion network for violence detection[J]. Entropy, 2022, 24(7): 939.

[76] 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. New York: IEEE, 2015: 4489-4497.

[77] 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. New York:IEEE, 2017: 6299-6308.

[78] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]//European Conference on Computer Vision. Berlin: Springer, 2016: 20-36.

[79] Zhu Y, Newsam S. Motion-aware featu4re for improved video anomaly detection[J]. ArXiv Preprint ArXiv: 1907.10211, 2019.

[80] Zaheer M Z, Mahmood A, Shin H, et al. A self-reasoning framework for anomaly detection using video-level labels[J]. IEEE Signal Processing Letters, 2020, 27: 1705-1709.

[81] Zaheer M Z, Mahmood A, Astrid M, et al. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection[C]//16th European Conference on Computer Vision. Berlin: Springer International Publishing, 2020: 358-376.

[82] Tian Y, Pang G, Chen Y, et al. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2021: 4975-4986.

[83] Lü H, Zhou C, Cui Z, et al. Localizing anomalies from weakly-labeled videos[J]. IEEE Iransactions on Image Processing, 2021, 30: 4505-4515.

[84] Cao C, Zhang X, Zhang S, et al. Adaptive graph convolutional networks for weakly supervised anomaly detection in videos[J]. IEEE Signal Processing Letters, 2022, 29: 2497-2501.

[85] Shao W, Xiao R, Rajapaksha P, et al. Video anomaly detection with NTCN-ML: a novel TCN for multi-instance learning[J]. Pattern Recognition, 2023, 143: 109765.

[86] Majhi S, Dai R, Kong Q, et al. Human-scene network: A novel baseline with self-rectifying loss for weakly supervised video anomaly detection[J]. Computer Vision and Image Understanding, 2024, 241: 103955.

[87] 周文浩, 胡宏涛, 陈旭, 等. 基于双重动态记忆网络的弱监督视频异常检测[J]. 计算机科学, 2024, 51(1): 243-251.
Zhou Wenhao, Hu Hongtao, Chen Xu, et al. Weakly supervised video anomaly detection based on dual dynamic memory network [J]. Computer Science, 2024, 51(1): 243-251.

[88] Shang Y, Wu X, Liu R. Multimodal violent video recognition based on mutual distillation[C]//5th Chinese Conference Pattern Recognition and Computer Vision. Berlin: Springer Nature Switzerland, 2022: 623-637.

[89] Wei D, Liu Y, Zhu X, et al. MSAF: Multimodal supervise-attention enhanced fusion for video anomaly detection[J]. IEEE Signal Processing Letters, 2022, 29: 2178-2182.

[90] Liang W, Xu X, Fu X. VioNets: efficient multi-modal fusion method based on bidirectional gate recurrent unit and cross-attention graph convolutional network for video violence detection[J]. Journal of Electronic Imaging, 2023, 32(2): 023031.

[91] 付荣华, 刘成明, 刘合星, 等. 骨架引导的多模态视频异常行为检测方法[J]. 郑州大学学报(理学版), 2024, 56(1): 16-24.
Fu Ronghua, Liu Chengming, Liu Hexing, et al. Skeleton-guided multi-modal video abnormal behavior detection method[J]. Journal of Zhengzhou University (Science Edition), 2024, 56(1): 16-24.

[92] Mahadevan V, Li W, Bhalodia V, et al. Anomaly detection in crowded scenes[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010: 1975-1981.

[93] Lu C, Shi J, Jia J. Abnormal event detection at 150 FPS in matlab [C]//Proceedings of the IEEE International Conference on Computer Vision. Sydney: IEEE, 2013: 2720-2727.

[94] Liu W, Chang H, Ma B, et al. Diversity-measurable anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2023: 12147-12156.

[95] Zhong J X, Li N, Kong W, et al. Graph convolutional label noise cleaner: train a plug-and-play action classifier for anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York:IEEE, 2019: 1237-1246.

[96] Adam A, Rivlin E, Shimshoni I, et al. Robust real-time unusual event detection using multiple fixed-location monitors[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(3): 555-560.

[97] Pranav M, Li Z, Shan K S. A day on campus—an anomaly detection dataset for events in a single camera[C]//Proceedings of the Asian Conference on Computer Vision. Berlin: Spring, 2020: 619-635.

[98] Ramachandra B, Jones M. Street scene: a new dataset and evaluation protocol for video anomaly detection[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. New York: IEEE, 2020: 2569-2578.

[99] Acsintoae A, Florescu A, Georgescu M I, et al. Ubnormal: new benchmark for supervised open-set video anomaly detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and

Pattern Recognition. New York: IEEE, 2022: 20143-20153.

[100] Zweig M H, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine [J]. Clinical Chemistry, 1993, 39(4): 561-577.

[101] Swets J A. Measuring the accuracy of diagnostic systems[J]. Science, 1988, 240(4857): 1285-1293.

[102] Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos[C]// IEEE International Conference on Computer Vision. New York: IEEE Computer Society, 2003: 1470.

[103] Georgescu M I, Barbalau A, Ionescu R T, et al. Anomaly detection in video *via* self-supervised and multi-task learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 12742-12752.

[104] 薛培康, 杜红棉, 王玮, 等. 基于 AM5708 的智能多目标跟踪监控系统设计[J]. 科学技术与工程, 2021, 21(4): 1471-1476.
Xue Peikang, Du Hongmian, Wang Wei, et al. Design of intelligent multi-target tracking and monitoring system based on AM5708[J]. Science Technology and Engineering, 2021, 21(4): 1471-1476.

[105] Huang C, Liu Y, Zhang Z, et al. Hierarchical graph embedded pose regularity learning *via* spatio-temporal transformer for abnormal behavior detection[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: IEEE, 2022: 307-315.

[106] Croitoru F A, Hondru V, Ionescu R T, et al. Diffusion models in vision: a survey[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2023, 45(9): 10850-10869.

[107] Deng L, Li G, Han S, et al. Model compression and hardware acceleration for neural networks: a comprehensive survey[J]. Proceedings of the IEEE, 2020, 108(4): 485-532.

[108] Gou J, Yu B, Maybank S J, et al. Knowledge distillation: a survey[J]. International Journal of Computer Vision, 2021, 129: 1789-1819.

[109] Zhou K, Liu Z, Qiao Y, et al. Domain generalization: a survey [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(4): 4396-4415.

[110] Long M, Wang J, Ding G, et al. Transfer joint matching for unsupervised domain adaptation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2014: 1410-1417.