中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)03-0615-26

论文引用格式:Wang Y, Zhou J G, Yan J and Guan J H. 2025. Survey of anomaly detection methods in surveillance videos based on deep learning. Journal of Image and Graphics, 30(3):0615-0640(汪洋, 周脚根, 严俊, 关佶红. 2025. 基于深度学习的监控视频异常检测方法综述. 中国图象图形学报, 30(3):0615-0640)[DOI:10.11834/jig.240329]

基于深度学习的监控视频异常检测方法综述

汪洋[1,2],周脚根[2],严俊[1],关佶红[1*]

1. 同济大学计算机科学与技术学院,上海　201804; 2. 淮阴师范学院地理科学与规划学院,淮安　223001

摘　要: 利用监控视频监测异常在社会治理中具有至关重要的地位,因此视频异常检测一直是计算机视觉领域备受关注且具有挑战性的议题。鉴于此,以深度学习的视角,对当前关键的视频异常检测方法进行了分类和综述。首先,全面介绍了视频异常的定义,包括异常的划定和类型分类;随后,分析了目前全监督、弱监督、无监督等方面的深度学习方法在视频异常检测领域的进展,探讨了各自的优缺点,特别针对结合大模型的最新研究进展进行了探讨;接着,详细介绍了常见和最新的数据集,并对它们的特点进行了比较分析和截图展示;最后,介绍了多种异常判定和性能评估标准,对各算法的性能表现进行了对比分析。根据这些信息,本文展望了未来数据集、评估标准以及方法研究的可能发展方向,特别强调了大模型在视频异常检测中的新机遇。综上,本文对于深化读者对视频异常检测领域的理解,以及指导未来的研究方向具有积极意义。

关键词: 视频异常检测;深度学习;数据集;大模型;监督学习;弱监督学习;无监督学习;多模态

Survey of anomaly detection methods in surveillance videos based on deep learning

Wang Yang[1,2], Zhou Jiaogen[2], Yan Jun[1], Guan Jihong[1*]

1. School of Computer Science and Technology National, Tongji University, Shanghai　201804, China;

2. School of Geography and Planning, Huaiyin Normal University, Huaian　223001, China

Abstract: Video anomaly detection plays a crucial role in social governance by utilizing surveillance footage, making it a crucial and challenging topic within the field of computer vision. This paper presents a detailed classification and review of current key video anomaly detection methods from a deep learning perspective, analyzing existing technical challenges and future development trends. First, the paper provides a comprehensive introduction to the definition of video anomalies, including the delineation of anomalies and video anomalies, the five types of video anomalies (intuitive anomalies, action change anomalies, trajectory change anomalies, group change anomalies, and spatiotemporal anomalies), and the three characteristics of anomaly detection (abstraction, uncertainty, and sparsity). The paper then reviews the development trends in video anomaly detection research from 2008 to the present based on the digital bibliography & library project (DBLP) database and provides a detailed analysis of the progress of fully supervised, weakly supervised, and unsupervised deep learning methods in the field of video anomaly detection. The core innovations, structures, and advantages and disadvantages of each method are discussed, particularly focusing on the latest research advancements involving large mod-

收稿日期:2024-06-13;**修回日期**:2024-09-12;**预印本日期**:2024-09-19

*通信作者:关佶红　jhguan@tongji.edu.cn

基金项目:国家重点研发计划资助(2021YFC3300304);国家自然科学基金项目(62172300, 62372326)

Supported by: National Key R&D Program of China(2021YFC3300304); National Natural Science Foundation of China(62172300, 62372326)

els. For instance, some studies address the challenge of applying virtual anomaly video datasets to real-world scenarios by designing anomaly prompts that guide mapping networks to generate unseen anomalies in real-world settings. Additionally, some works have designed dual-branch model structures based on multimodal large model frameworks. One branch uses the contrastive language-image pre-training (CLIP) visual encoding module for coarse-grained binary classification, while the other branch aligns textual features of anomaly category labels with visual encoding features for fine-grained anomaly classification, surpassing the current state-of-the-art performance in video anomaly detection. Furthermore, research has explored the potential of using GPT-4V, a powerful large visual language model, to tackle general anomaly detection tasks, examining its applications in multimodal and multidomain anomaly detection tasks, including image, video, point cloud, and time-series data across various fields such as industry, healthcare, logic, video, 3D anomaly detection, and localization. The introduction of large models presents new opportunities and challenges for video anomaly detection. Moreover, the paper introduces 10 commonly used and latest datasets, providing a comparative analysis of their characteristics and presenting detailed content through figures, along with corresponding download links. These datasets play a crucial role in video anomaly detection research, and this paper offers their comprehensive evaluation. The paper also introduces four anomaly determination standards (frame-based, pixel-based, and trajectory-based) and three performance evaluation standards (area under the receiver operating characteristic curve (AUC), equal error rate (EER), and average precision (AP)), and conducts a comparative analysis of the performance of various algorithms. The strengths and weaknesses of current video anomaly detection algorithms are summarized, and suggestions for improvement are proposed. Based on this information, datasets may have become a bottleneck in the development of current methods. In complex real-world scenarios, research methods based solely on simple scenes may not effectively address anomaly issues in the real world. Future datasets will aim to better reflect real-world anomalies, such as collecting data from the remote sensing field, improving the quality of existing image and video data through models, and collecting multi-camera, multidimensional annotated data, to detect more diverse and challenging anomaly events and effectively promote research development. Additionally, in terms of evaluation standards, common evaluation methods primarily rely on calculating the true false positive rates and computing the area under the receiver operating characteristic curve. However, in practical applications, some methods may achieve high AUC but exhibit a high false alarm rate due to the direct influence of different anomaly determination methods on the true and false positive rates. Adopting different anomaly determination methods may result in models achieving high AUC performance while generating high false alarm rates. Therefore, this paper proposes the need to design an evaluation system that simultaneously considers AUC performance and false alarm rates to comprehensively evaluate methods. Finally, the outlook of the paper emphasizes new opportunities presented by large models in video anomaly detection. The emergence of large models in recent years has substantially improved the performance of deep learning-based methods on commonly used video anomaly detection datasets. This field has accumulated a solid academic research foundation. Therefore, future research should not only focus on improving anomaly detection performance but also consider the application of this field to practical problems to address existing challenges. Future research should aim to design more fine-grained and general models, leveraging the rich prior knowledge of large models to gradually develop video anomaly detection models that can distinguish specific types of anomalies. With the powerful multimodal information understanding capabilities of large models, video anomaly detection models will evolve toward a more general direction, ultimately blurring the boundaries between supervised, weakly supervised, and unsupervised learning methods. Overall, this paper substantially enhances readers' understanding of the field of video anomaly detection and provides valuable references and guidance for future research directions. Through a systematic review and analysis of existing research, this paper offers crucial insights for further development of the video anomaly detection field.

Key words: video anomaly detection; deep learning; dataset; large model; supervised learning; weakly supervised learning; unsupervised learning; multimodal

0 引 言

当前，社会公共安全已成为一个日益突出的全球性挑战。近年来，全球发生了多起严重的暴力恐怖事件，如 2011 年挪威奥斯陆和于特岛的袭击事件、2014 年昆明火车站事件、2016 年法国尼斯事件等，这些事件凸显了有效视频监控在公共安全中的必要性，以及构建处理突发公共安全事件的快速响应机制的重要性。

随着“天网”等视频监控系统的完善（郎江涛，2017），监控设备数量和产生的视频数据激增，如何有效处理和分析海量视频数据成为挑战。传统视频监控系统多依赖人力进行检视，但长时间的检视不仅耗费大量人力资源，还容易因人体疲劳导致检视效率和准确性的下降。此外，网络视频也已成为大众生活娱乐的重要组成部分。各网络平台每天都在增加数以万计的新视频。然而，人工审核无法跟上互联网视频的发展速度，导致视频违规展示问题频繁出现。以上问题都凸显出急需更有效的方法来处理庞大的视频数据流。因此，自动、智能的视频异常检测逐渐成为计算机视觉（computer vision，CV）领域的新兴研究热点。

视频异常检测技术是一项根据视频内容自动检测其中是否存在异常的技术。通常，这项技术需要自动检测视频中的异常，并对异常位置进行定位。视频中的异常内容通常包括打斗、抢劫、破坏公物和违反法规等特定行为，以及可疑遗留物、武器等异常物品（Sultani 等，2018）。确定异常概念时，需要根据当前环境的实际情况来制定与问题相关的异常定义。在学术和日常生活中，通常使用例外、偏差、新奇、噪声和离群值等形容词来描述异常这个概念。

异常通常具有以下 3 个特性：1）抽象性，即模糊的意义和丰富的内涵；2）异常本身通常具有极大的不确定性和时空相对性；3）异常的发生具有随机性和稀疏性。 根据这 3 个特性，本文认为视频异常取决于上下文和主题，可以定义为在不寻常的地点、不寻常的时间进行的活动，或在外观和运动上与正常有根本差异的活动。因此，针对视频异常，本文将其分为以下 5 种类型，对应分类的视频异常示例如图 1 所示，图 1 中图像的左上角数字代表对应的异常分类。

图 1　不同类型的视频异常示例

Fig. 1　Examples of different types of video abnormalities

1）直观异常。通常指在环境中出现意料之外的物体，如在高速路上出现猪，或者在人行道上出现车辆。

2）动作变化异常。通常指环境中短时间内发生的不正常动作，如在图书馆里小孩打闹，或者敲打商品柜台玻璃。这类异常通常需要一段时间的持续发生才能成立。

3）轨迹变化异常。通常指环境中物体出现不符合规则的运动轨迹，如汽车在高速上连续变道，或者陌生人在军事基地附近徘徊。这类异常需要较长的视频片段来确认。

4）群体变化异常。通常指物体间的相对运动异常，如人群突然四散或聚集。

5）时空异常。通常指环境中物体在时空维度上产生的异常，如地铁站遗留的行李箱，或金店闭店后出现人等情况。这类异常通常在不同的时间和空间上具有不同定义。

当前，视频异常检测面临多重挑战，包括环境条件（如光照变化、物体的阴影效应、物体遮挡、复杂背景等）、人群密度、数据质量差、行为复杂性、拍摄角度、时空变化、样本极度不平衡等。这些挑战使得视频异常检测成为计算机视觉领域中最具挑战性和价值的任务之一。此外，在对现有视频异常检测综述的研究中，主要发现以下几个问题：

1）王志国和章毓晋（2020）、杨帆等人（2021）和吉根林等人（2024）对方法的发展展望的分析也较为简略。

2）Nayak 等人（2021）、Ren 等人（2021）和 Pang 等人（2021）偏向对理论推导的分析，缺乏传统方法、数据集和指标等相关细节的介绍和分析。

当前，深度学习在视频异常检测任务中表现出色，尤其是随着大模型在深度学习领域的兴起，为这一任务带来了更广阔的发展空间。因此本文

针对以上工作的不足，以深度学习为角度对视频异常检测的相关工作进行综述，并着重关注了大模型在该领域的新机遇，现有工作概览如图2所示。

本文综述对比以上工作有以下几点优势：1）本文旨在全面综述视频异常检测领域的现有研究工作，并着重关注深度学习的最新发展趋势，尤其是与大模型相关的进展。2）全面介绍常用异常判定和模型性能评估指标。3）全面介绍代表性数据集，并分析数据集优势劣势，且对比现有方法在数据集上的效果。4）以通俗易懂的方式完整地对该领域方法、数据集、评价指标、未来发展等问题进行综述，有利于新接触该领域的研究人员清晰快速了解本领域的整体情况。

本文首先按照不同监督方式对现有视频异常检测方法进行分类，并逐类概述、分析各类型算法模型的优缺点。然后，介绍该领域的常用和最新公共数据集、评估标准，并汇总比较不同算法在这些数据集上的检测效果。最后，单独讨论了大模型在视频异常检测中的发展，以及现有的技术挑战和未来发展趋势。

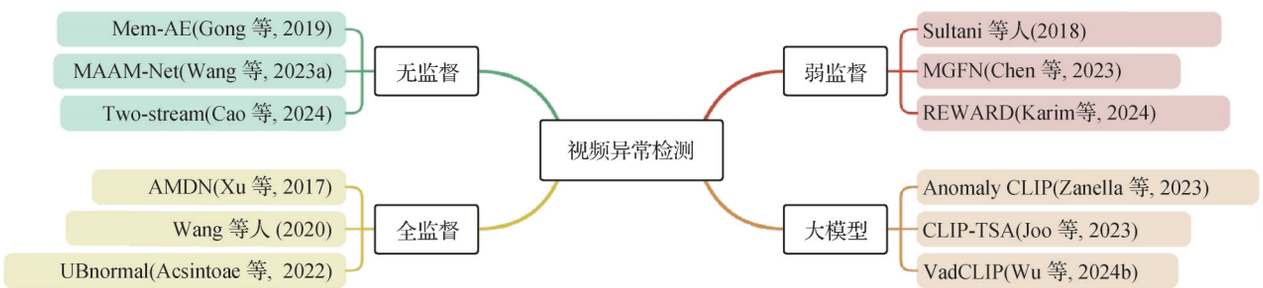

图2　基于深度学习的视频异常检测方法概览

Fig. 2　Overview of deep learning based video anomaly detection methods

1　研究问题

视频异常检测的定义可简单概括为识别与正常事件存在显著差异的事件。基于距离的视频异常检测方法（Saligrama 和 Chen，2012；Ionescu 等，2019b；Lee 等，2015）很好地体现了这一定义，它们通过在向量空间中度量事件之间的距离来评估事件的属性。此外，尽管不同类型的异常事件检测方法在形式上有所差异，但它们通常都依赖于正常和异常特征或者分布之间的差异这一基本概念。由于异常样本往往稀缺，异常事件检测以无监督、半监督和弱监督学习等方法为主。这些方法通常基于一些共同的假设，后续各小节将对此进行详细的阐述和分析。

1.1　异常检测

异常检测（anomaly detection）中，将样本分为正常样本（normal）和异常样本（anomaly）。正常样本被认为是处于事件分布的主要区域内，而在该区域之外的样本被视为异常样本。如图3所示，给定样本x，函数$f_\theta(x)$将样本映射到二维空间中（为了形象，采用二维空间），函数或者模型$M\left(f_\theta(x)\right)$用于判断样本是正常还是异常。具体为

$$M(f_\theta(x))=\begin{cases}\text{normal} & D\left(f_\theta(x),P\right)\leqslant\partial\\ \text{anomaly} & D\left(f_\theta(x),P\right)>\partial\end{cases}\tag{1}$$

式中，函数$D(\cdot)$用于度量样本间距离，P表示正常样本分布中心坐标，∂是一个预先设定或者可学习的门限值（参数）。当$f_\theta(x)$与P的距离$\leqslant\partial$时认为x为正常样本，反之为异常样本。实际上，还可以用测试样本分布和正常样本分布的差异来判断是否为异常样本。这时，KL散度（Kullback-Leibler divergence）经常用于度量两个分布之间的差异。

图3展示了样本集在二维空间上的分布情况。从图中可以看出，正常样本（绿色小圆点）聚集在绿色圈内，而异常样本（红色三角形）则与正常样本的分布不同，它们分散在绿色圈外。因此，在判断新样本的类别归属时，一种做法是设定一个距离正常样本分布中心点的阈值，然后根据新样本与中心点的距离来作出判断。

1.2　视频异常检测

在视频异常检测（video anomaly detection，VAD）领域，无监督、半监督以及弱监督学习方法占据了主导地位。这主要是由于视频中的异常情况通常被视为小概率事件，导致在构建数据集时难以收集和完

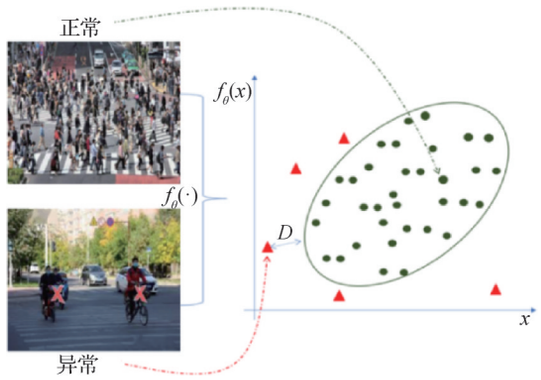

图 3　异常事件判定概念图

Fig. 3　Conceptual diagram for determining abnormal events

全覆盖所有异常情况。此外，现有视频数据主要来源于监控视频，这些视频通常具有时间长、内容庞大等特点，使得标注视频样本的成本变得非常高。因此，现有方法通过减少对大量标注数据的依赖，降低了数据收集和处理的难度，并针对性地提出了一些假设或说明，使得视频异常检测的研究更加可行。具体来说，目前主要有如下 3 类视频异常检测：

1）基于无监督学习和半监督学习的视频异常检测（Gong 等，2019；胡海洋 等，2020；Wang 等，2023b；Kommanduri 和 Ghorai，2024）（简称无监督、半监督视频异常检测）。无监督和半监督视频异常检测通常基于一个共同的假设，即现实中的视频数据中异常样本是罕见的，绝大多数样本都是正常的。这一假设使得这两种方法都能够利用大量未标记的正常样本进行训练，从而减少了对标注异常样本的依赖，降低了标注成本。文献中也有将这两类 VAD 方法都称做无监督方法。在这两种方法中，模型的训练目标都是学习重建或预测正常样本，而不是直接学习异常样本的特征。因此，当模型遇到异常样本时，由于这些样本与正常样本在特征上的差异，模型将无法有效地进行样本重建或预测，从而产生较大重建或者预测误差，由此识别出异常样本。这些方法的一个关键优势是能够缓解异常样本难以收集的问题。通过假设正常样本占主导地位，无监督和半监督视频异常检测能够利用大量的未标记正常样本进行训练，从而提高了视频异常检测的效率和可行性。

2）基于弱监督学习的视频异常检测（Chen 等，2023；AlMarri 等，2024；朱新瑞 等，2024）（简称弱监督视频异常检测）。在弱监督学习框架下，通常只提供视频级标签，表示视频是否包含异常，而没有详细的帧级标注。这种弱标签方法认为视频级标记数据中仍然包含足够的信息来推断异常事件。因此，该类方法需要对弱标签数据标签具体位置这一不确定问题进行建模，并利用多实例学习、自监督学习或自注意力机制等技术，有效地捕获异常事件。

3）基于全监督学习的视频异常检测（Sabokrou 等，2018a；Xu 等，2017）（简称全监督视频异常检测）。全监督视频异常检测依赖于包含帧级甚至像素级标注的数据集，需要耗费大量资源进行标注。这类方法与常见的全监督分类方法相似，这里不作进一步阐述。

2　视频异常检测方法分类与简述

Xiang 和 Gong（2008）早在 2008 年就开展了视频异常检测问题的研究。他们提出了基于动态贝叶斯网络和谱聚类的视频异常识别框架。随后，这一研究问题吸引了越来越多研究人员的关注。根据对计算机领域最大的数据库 DBLP（digital bibliography & library project）的检索数据的统计分析，2008 年到 2023 年期间有关视频异常检测的论文发表呈现显著增长趋势，如图 4 所示，并且从 2020 年开始呈现快速增长趋势，到 2023 年全年论文发表量达 181 篇。这一趋势反映了目前视频异常检测研究的活跃度和热度。对 2008 年—2023 年的视频异常检测代表性工作进行整理分析发现，视频异常检测工作主要包括 3 种类型：传统方法、深度学习方法和传统方法+深度学习方法。2017 年之前，视频异常检测研究主要集中在传统方法上，这一时期的相关研究数量相对较少，之后的研究主要以深度学习或深度学习结合传统方法为主。

2.1　传统方法

传统方法主要包含两类研究方法，第 1 类是通过手工特征提取视频特征，一般需要用到的技术有梯度直方图（histogram of gradient，HOG）（Saligrama 和 Chen，2012）、时空梯度（Lu 等，2013a）、前景掩膜（Antić 和 Ommer，2011）、动态纹理（Mahadevan 等，2010）等；完成特征提取之后下一步通常使用聚类方法（Ionescu 等，2019b；Lee 等，2015）、最近邻方法（Saligrama 和 Chen，2012）、深度高斯模糊（Feng 等，2017）等方法根据样本在投影空间之间的距离度量来对正常与异常进行判别。第 2 类方法是在手工特

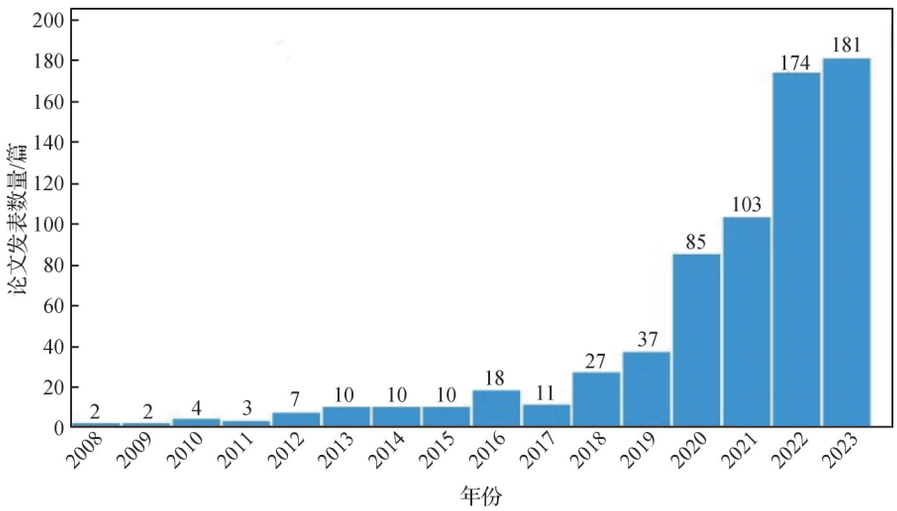

图4　相关论文发表数量增长趋势

Fig. 4　Trends in the number of relevant papers published

征提取后利用特征重构的方式对正常与异常之间的判别，主要涉及的方法是稀疏重构（Cong等，2013；Zhu等，2014；Ren等，2015；Li等，2015a；Yuan等，2018），通过训练模型学习正常样本的稀疏表达能力，因此当模型遇到异常样本时会产生较大的重构误差。上述两类方法的简要结构如图5（a）所示，传统方法的优势有可解释性强、在数据量较少的情况下依旧有效；劣势为不能端到端运行、泛化性较差和计算量随着数据量的增大而增大。

2.2　深度学习方法

视频数据通常由多模态数据构成，具有图像、时间、声音等维度，而深度学习方法通常为黑箱方法，擅长对多模态数据进行特征抽取以及学习其中复杂结构和关系。深度学习方法通常通过预训练的卷积神经网络（Smeureanu等，2017；Hinami等，2017；Sabokrou等，2017；Chong和Tay，2017）或根据特定任务结构（Hasan等，2016；Liu等，2018；Chang等，2020）（例如：自编码器的重构编码方法）的神经网络提取样本特征，最后通过分类或重建误差的方式区分正常与异常。在众多具有特色的工作中，有的通过加入目标检测网络来增强异常出现的可解释性；有的通过加入人体骨架识别网络的方式增强人体动作异常识别效果（Zhong等，2019；Morais等，2019；Markovitz等，2020）；有的设计存储模块（Park等，2020；Cai等，2021），存储构成正常样本的特定特征，使得异常样本的重构误差增大；也有使用神经网络提取特征结合传统方法进行研究（Wang等，2020；Wang等，2018；Ionescu等，2019a；Fan等，2020）。当前，基于深度学习的方法已经在众多主流数据集上展现出最佳性能，简要的方法结构如图5（b）所示，对比传统方法通常能够具备更深更复杂的模型结构，并且能够实现端到端的学习。此外，目前结合大规模模型技术的深度学习方法，它们大幅提升了现有模型性能的上限，进一步确立了深度学习方法显著的优越性。这一发展不仅在视频异常检测的研究领域证实了深度学习方法的领先地位，而且进一步凸显了其在处理高复杂度数据集时的强大能力和应用前景。但是，现有方法依然有泛化性弱、依赖大量数据、可解释差等问题。综上可知深度学习在视频异常检测问题上表现出比传统方法更大的潜力。因此，本文以深度学习方法的研究视角对视频异常检测工作进行综述并对未来的研究发展提出相关见解，也为未来加入该领域的研究者提供一个能够快速了解该领域的途径。

3　深度学习方法

视频异常检测通常仅使用无标签数据进行模型训练，基于无监督假设，识别超出训练集分布的数据以检测异常。在某些情况下能够获取部分异常和正常样本数据。因此，基于全监督、弱监督的方法也成为视频异常检测方法研究的一部分。本节对处理不同监督数据的异常检测方法及其发展进行了详细描述。此外，还介绍了结合大模型技术的视频异常检

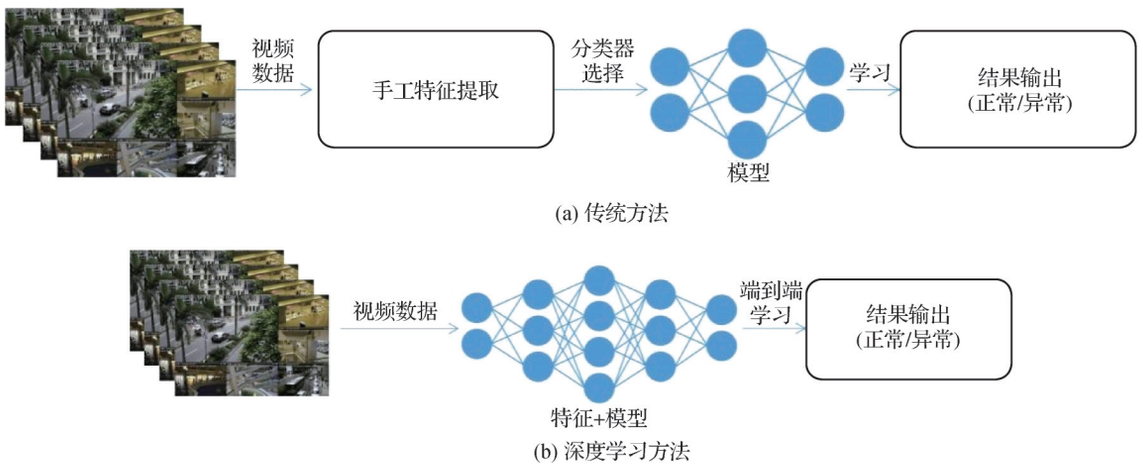

图 5　传统方法与深度学习方法框架比较

Fig. 5　Comparison of framework between traditional method and deep learning method

((a) traditional method; (b) deep learning method)

测方法的最新发展。

3.1　基于全监督的方法

全监督视频异常检测依赖于帧或像素级别的精确标注数据进行模型训练，并输出帧级别甚至像素级别的检测结果。这些方法的优势在于能够利用详尽的标注信息来精确指导模型学习正常、异常样本的特征，从而实现较为准确、稳定的异常检测效果，其大致流程如图 6 所示。

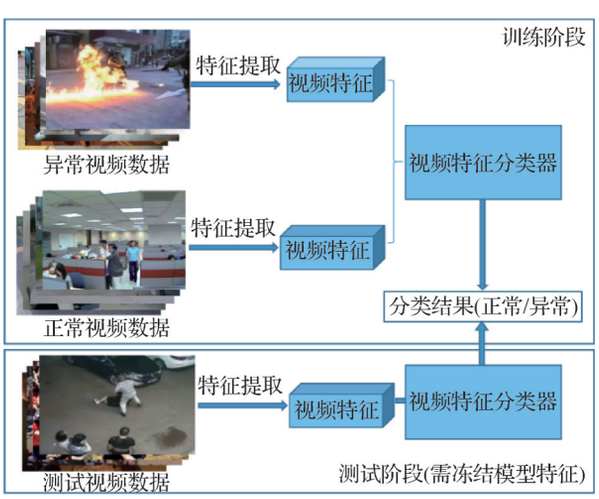

图 6　全监督视频异常检测基本流程

Fig. 6　Fully supervised video anomaly detection workflow

图 6 展示了从数据输入、模型训练到最终测试的整个过程。下面从解决不同问题的角度来介绍全监督视频异常检测的主要方法。

针对老人和患者等的异常动作的检测问题，Xu 等人(2017)提出一种基于深度神经网络的特征表示自动学习方法，即外观与运动深度网络(appearance and motion deepnet，AMDN)。为了更好地融合外观和运动信息，作者首先引入了一种新的融合方法，称为堆叠式去噪自动编码器。该方法分别学习外观和运动特征，然后将这两种特征融合在一起。接下来，基于融合的特征，使用多个单类支持向量机模型来预测每个输入的异常得分。最后，提出了一种新的后期融合策略，将计算得到的得分结合在一起，用于检测异常事件。此外，Tran 和 Hogg(2017)根据赢者通吃自编码器(winner-take-all autoencoders，WTA-AE)(Makhzani 和 Frey，2015)，提出一种基于该策略的卷积自动编码器的视频异常检测算法。其主要创新之处在于：1)卷积自动编码器提取的运动特征编码作为一类支持向量机的输入；2)训练过程加入空间赢家通吃步骤，引入高度稀疏性更利于区分正负样本。Sabokrou 等人(2018b)受生成对抗网络(generative adversarial network，GAN)(Goodfellow 等，2014)的启发，提出一种端到端的竞争和协作算法。该算法涉及两个网络：一个负责像素级检测，另一个是分块级别的检测。经过对抗性的自监督训练后，这两个网络能够更好地协同工作，检测给定测试视频中的异常，并对异常区域进行精准的定位和分割。此外，人体异常动作检测还可以借助毫米波雷达(元志安 等，2021)、融合骨架检测(Amsaprabhaa 等，2023)或注意力网络(Vidya 和 Selvakumar，2024)实现。这些方法通过利用外部设备、预训练模型以及注意力网络提取更丰富的目标外形和动态特征，并

取得了更好的检测效果，在解决人体动作异常检测问题上展现了显著的潜力。

针对拥挤场景下的视频异常检测问题，Sabokrou等人(2017)提出一种基于三维图块的级联分类器方法。该方法以“局部”和“全局”描述符为基础，通过引入时空上下文来描述视频特征。局部描述基于当前区域与邻域的关系，而全局描述则由稀疏自动编码器提供支持。通过训练局部和全局描述，生成两个参考模型作为单类分类器，随后将它们组合成级联分类器。该级联分类器首先利用速度更快的局部分类器对“许多”正常的图块进行早期识别，然后由全局分类器“仔细”检查剩余的图块。此外，文中还提出了一种技术，即利用小图块学习判别能力，利用大图块推断异常，以提高整体异常检测性能。

为了进一步提高全监督任务的异常检测性能，Ionescu等人(2017)提出一种无需训练序列的视频异常检测方法，方法的核心是Unmasking(Koppel等，2007)技术。作者对该技术进行了改进，通过迭代地训练二分类器来区分两个连续的视频序列。在每一步，最具判别性的特征都被去除，最终获得的分类器在迭代过程中表现出较高的异常检测准确率。Luo等人(2017a)提出一种时间相干稀疏编码(temporally-coherent sparse coding，TSC)，用于对相似的相邻帧进行编码，并采用特殊类型的堆叠式递归神经网络映射稀疏编码，通过利用堆叠循环神经网络(stacked recurrent neural network，SRNN)(Dong等，2016)同时学习所有参数，避免了TSC中琐碎的超参数选择。此外，借助较浅的SRNN，可以在一次前向传递内推断重构系数，降低了稀疏系数学习的计算量，提高了异常检测的速度。作者还构建了一个庞大的数据集，无论从数据量还是场景多样性来看，这个数据集都比当时所有异常检测数据集总和还大。此外，Acsintoae等人(2022)提出了UBnormal数据集，这是一个使用Cinema4D生成的全监督数据集，它包含用于视频异常检测的多个虚拟场景，并且提出了一种新的范式——将视频异常检测作为一个有监督的开集分类问题进行框架化。在此方法中，正常和异常事件在训练时都可用，但在测试推理阶段发生的异常属于一组不同的异常类型(类别)，即训练时是一种异常类别，测试时又是另一种类别。

在全监督学习框架下，异常事件的明确定义使得收集正常和异常样本以训练模型相对简单，构成了该类方法的一个主要优势。然而，全监督方法在现实应用中仍存在局限性。异常数据通常非常稀缺，导致深度监督分类器的性能难以达到最优。此外，异常的多样性也会影响训练效果，使得训练变得困难。由于需要大量的标注数据，全监督方法在实际应用中面临挑战，尤其是在视频数据的标注过程中，通常耗时巨大且成本高昂。这限制了全监督方法在广泛场景中的应用。因此，尽管全监督方法在视频异常检测领域仍有研究价值，但针对基准数据集比较的研究工作在2018年后鲜有出现，目前已不是视频异常检测领域的主流研究方向。相反，弱监督、半监督和无监督学习方法由于对标注数据的依赖较少，成为视频异常检测领域更受欢迎的研究方向。这些方法通过减少对大量标注数据的依赖，降低了数据收集和处理的难度，使得视频异常检测技术更加实用、高效，并有望解决实际应用中的关键问题。

3.2 基于弱监督的方法

在视频异常检测领域，弱监督方法提出之前，视频标注通常采用逐帧或像素级别的标注方式，这两种方式都需要消耗大量的人力和物力。相比之下，弱监督方法仅需要对单个视频进行标注，将其中的异常视频标记为异常，非异常视频标记为正常，这种方式大大减少了标注的工作量并提高了标注效率，并且可以实现帧级别的视频异常检测。针对这种弱标记数据设计的方法称为弱监督视频异常检测(weakly supervised video anomaly detection，WVAD)方法。这种方法通常基于多实例学习(multiple instance learning，MIL)框架实现，基本工作流程如图7所示。

Sultani等人(2018)提出了首个大规模弱监督视频异常检测数据集，创新性地采用多示例学习设计了弱监督视频异常检测方法，成为目前该领域的主流实现方式。在该方法中，作者将正常视频和异常视频作为包，将视频固定切分为32个片段作为MIL中的实例。基于MIL，他们构建了深度多实例排名框架来学习区分正常和异常。此外，通过引入稀疏性和时间平滑性约束优化排序损失函数，使模型能够输出更稳定、更准确的异常打分曲线。此后，出现了一系列的针对基于MIL设计的弱监督视频异常检测的改进方法，它们主要是从以下3个方向进行改进：

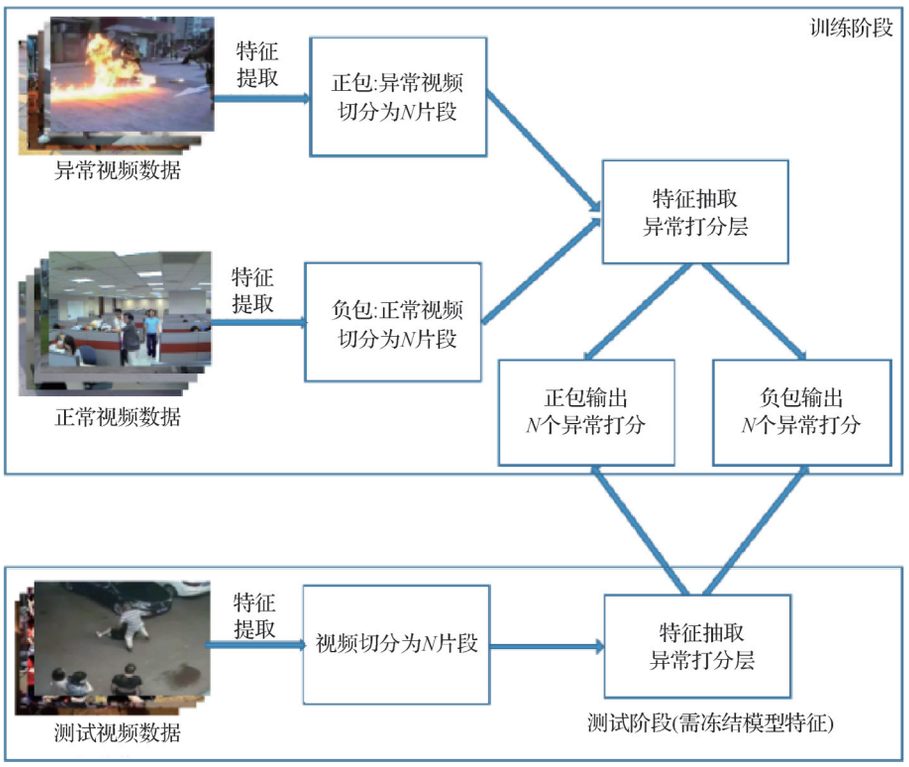

图 7　弱监督视频异常检测基本流程

Fig. 7　Weakly supervised video anomaly detection workflow

1)优化损失函数。Zhang 等人(2019)根据正包的最低分和最高分的差异,提出一种新的内包损失(inner bag loss, IBL)来调整弱监督问题的函数空间,以提高正常、异常特征空间的区分度。Wan 等人(2020)引入动态多实例学习损失和中心损失,以更好地学习异常和正常的区分特征。动态多实例学习损失旨在增大异常实例与正常实例之间的类内距离,而中心损失则用于减小正常实例的类内距离,使正常与异常实例之间的距离更为显著。此外, Majhi 等人(2024)提出一种自校正损失函数,该函数可以动态地从视频级标签计算伪时间标注,以有效优化异常检测网络对人物与空间线索表征的理解能力。

2)优化时空理解能力。Liu 和 Ma(2019a)指出背景偏差会影响异常事件检测的性能,因此提出一个端到端可训练的异常区域学习引导框架,采用新的区域损失来明确驱动网络学习异常区域。Landi 等人(2019)探讨了从时空角度考虑视频异常检测,提出一种添加了时空通道标签的方法,证实其效果优于仅仅使用全帧视频片段的方法。针对视频运动信息,Zhu 和 Newsam(2019)提出一种时间增强网络,用于学习运动感知特征,该特征将时间上下文引入多实例学习排名模型中,可以显著提高性能。此外,针对视频特征中的时间属性,Tian 等人(2021)提出鲁棒时间特征幅度学习(robust temporal feature magnitude learning, RTFM)方法。该方法通过训练一个特征幅度学习函数来有效识别正样本,显著提升了 MIL 方法对异常视频中负样本的鲁棒性。然而,有学者指出 RTFM 使用特征量级表示异常程度,常忽略场景变化影响,会导致性能抖动。因此,Chen 等人(2023)提出一种新的扫视和聚焦网络,能整合时空信息,实现更精确的异常检测。此外,弱监督时态判别方法(Huang 等, 2024)引入了 Transformer 时态特征聚合器以增强视频片段间时间关系建模,以及自引导判别特征编码器用于提取判别特征以进一步区分正常和异常片段。

3)优化弱标签使用策略。针对弱标签中没有标签具体位置信息的问题,其中一种解决方式是生成伪标签。Feng 等人(2021)提出基于多实例的伪标签生成器,利用多实例伪标签生成器采用稀疏连续采样策略来产生更细粒度的片段级伪标签,目的在于提取特征的同时能自动关注于帧中的异常区域,提高异常检测效果。Zhang 等人(2023)提出了两阶

段自训练方法，自我生成伪标签，并利用这些标签自提炼异常评分，提高了模型性能。另一方法是在训练中定义可信样本。例如，在RTFM（robust temporal feature magnitude learning）（Tian等，2021）中，通过统计正样本的比例，在训练过程中将预测为正样本中特征幅度排在前三名的正样本视为可信正样本，并用于计算损失，以更充分地利用训练数据。但是，由于不同数据集的阳性实例比例不同，Wang等人（2024b）进一步提出一种动态实例选择策略，能在训练过程中自适应选择一定比例的可信预测实例参与损失计算，减少弱标记的不确定性。此外，针对正常与异常视频数据不平衡问题，He等人（2024）提出一种通过对异常视频进行对抗和聚焦训练的新方法。该方法包括两个模块：一个是基于数据的对抗训练模块，通过基于潜在空间的对抗样本生成来进行数据增强；另一个是基于模型的聚焦训练模块，以及关注异常视频的成本敏感损失。

除了基于MIL实现弱监督视频异常检测的方法外，还有多种创新性的方法相继提出。Zhong等人（2019）提出一种具有矫正噪声标签的图卷积神经网络（graph convolutional network，GCN），用于解决监督学习任务中的噪声标签问题。该方法能够剔除标签噪声，并将全监督动作分类器应用到弱监督异常检测中，实现最大限度地利用成熟的分类器。同时，Liu等人（2019b）提出一种边际学习嵌入预测方法，通过学习更紧凑的正态数据分布来提高异常事件检测的性能，尤其对于以前从未观察到的异常情况也有良好的适应性。Wu等人（2020）提出了当前规模最大的弱监督视频异常检测数据集，这是一个同时包含视频和音频的大规模多场景数据集，即XD-Violence。此外，他们还设计了一种弱监督异常检测方法，以减轻批次效应和标签噪声问题。Ramachandra和Jones（2020）引入孪生卷积神经网络（siamese convolutional neural network，SCNN），用于学习视频块之间的距离度量函数，进而识别异常事件。Wan等人（2020）将视频异常检测看做弱监督下的视频片段异常评分的回归问题，并设计了异常回归网络，利用动态多实例学习损失和中心损失来更好地学习异常和正常的区分特征。Liu等人（2022）采用无监督深度自动编码器训练正常视频的时空模式原型，然后利用正常和异常视频训练回归模块，目标是使异常视频平均得分高于正常视频最高得分。最后，异常视频中得分低于平均得分的片段被视为正常视频，并用于微调自动编码器。无监督自动编码器与弱监督回归模型合作，提取正常片段的原型特征，从而更容易区分学习到的正常和异常事件特征。Karim等人（2024）认为现有方法大多依赖于特别的特征聚合技术和使用度量学习损失，从而限制了模型实时检测异常的能力，提出了一种实时端到端训练的弱监督视频异常检测方法。该方法可以直接从原始数据中自动学习有效的特征，端到端地实现训练和测试过程。上述这些研究为视频异常检测提供了多样的思路和技术手段，为该领域的进一步发展做出了重要贡献。

弱监督是监督、半监督与无监督的一种折中方案，通过为视频异常检测提供便捷的视频级的标注，降低了标注代价，实现了较好的检测效果。基于MIL的方法（Zhang等，2019；Wan等，2020；Tian等，2021；Feng等，2021；Sun等，2017）是最常见的弱监督方法。总之，尽管弱监督方法只有视频级标注，但仍能像监督方法一样定位异常信息，达到片段级甚至帧级的异常检测水平，并且性能出色。相对于监督方法，弱监督方法更适用于开放场景，展现出极大的发展潜力。然而，现有弱监督方法仍有局限性，如对弱标签信息利用不足（Tian等，2021）、参数量较大（Wang等，2024b），以及缺乏细粒度视频异常检测研究等。

3.3 基于无监督的方法

早期的神经网络方法（LeCun等，1998）将远离整体行为的样本标记为异常，并根据具体情况调整阈值。深度学习兴起后，大多数无监督方法以变分自编码器（variational auto-encoder，VAE）（LeCun等，1998）为基础框架，通过重建误差来识别异常，其基本工作流程如图8所示。

Chong和Tay（2017）提出一种端到端的自编码器，取代了传统的无监督学习方法（如稀疏编码方法），无需手工提取先验知识，能够更有效地抽取特征。Wang等人（2018）提出了对自编码器的改进，采用跳过连接的卷积变分自编码器（skip convolutional variational auto-encoder，SC-VAE）和堆叠全连接变分自编码器（stacked fully connected variational auto-encoder，SF-VAE）的组合网络S2-VAE（skip and stacked convolutional variational auto-encoder）。SF-VAE负责获得类似高斯混合的模型来拟合实际数

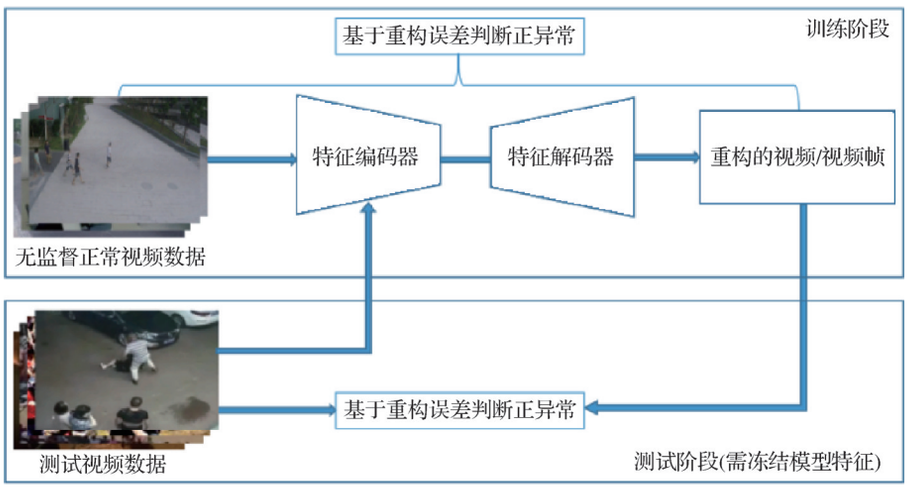

图 8　无监督视频异常检测基本工作流程

Fig. 8　Unsupervised video anomaly detection workflow

据分布，而 SC-VAE 则利用 CNN（convolutional neural network）、AE（auto-encoder）和跳跃连接的深度自编码网络，以获取更丰富的表征特征。两者的结合进一步提升了异常检测性能。为了提高自编码器对视频数据中运动和外观信息的提取能力，Ye 等人（2019）和 Yu 等人（2020）进行了进一步的改进。Ye 等人（2019）提出了利用深度神经网络（deep neural network，DNN）设计的预测编码模块，以增强模型对运动信息的利用，并参考 U-Net（Ronneberger 等，2015）设计了误差细化模块，从而提高了正常和异常之间的重建误差差异。Yu 等人（2020）设计了外观和运动互补的定位感兴趣块，通过擦除图块以产生不完整事件，构建了一种视觉完形填空的过程，使模型能够学到更强的全局构建能力。后续内容将依据该问题的技术发展线路进行介绍。

在上述模型结构研究的基础上，后续研究逐渐转向关注模型对视频中的外观、运动和时空等信息的理解。Rodrigues 等人（2020）和 Georgescu 等人（2021）从多个时间尺度对异常进行研究，并指出单个和预定义的时间尺度不足以捕获随着不同时间持续变化的宽范围异常。Ye 等人（2019）提出一种多时间尺度模型，可捕获不同时间尺度的时间动态。该多层次结构模型在给定输入姿势轨迹的情况下，生成不同时间尺寸的异常事件预测。Georgescu 等人（2021）以三维卷积神经网络为基础，联合学习多个代理任务，包括区分向前/向后移动的对象、区分连续/间歇帧中的对象和重建特定于对象的外观信息。通过这种方法产生了多时间尺度的异常特定信息，用于检测异常。另外，Luo 等人（2017b）提出一种利用卷积长短期记忆网络（convolutional long short term memory，ConvLSTM）（Shi 等，2015 ）的方法 ConvLSTM-AE，该方法能够记忆与运动信息相对应的所有过去帧来实现异常检测。与三维卷积自动编码器（Georgescu 等，2021）相比，ConvLSTM-AE 方法在对正常事件的外观变化和运动变化进行编码和特征提取方面表现更为优越。

此外，Liu 等人（2018）发现并非所有异常样本在 VAE 中都产生足够大的重建误差。因此，他们在 U-Net（Ronneberger 等，2015）的基础上提出一种方法，通过预测未来帧与实际帧之间的差异来检测异常事件，从而识别不符合预测的异常事件。在此基础之上，Nguyen 和 Meunier（2019）将卷积自编码器（convolutional autoencoder，Conv-AE）和 U-Net 结合，设计了一种基于外观—运动对应的视频异常未来帧预测方法。作者在输入层整合了一个改进的 Inception（Szegedy 等，2016）模块，并最终在输出阶段提出了一种基于图块的方案。该方案通过估计帧的正常性评分，有效降低了输出中噪声的可能性。后续工作（Zhao 等，2017）进一步利用视频时空特征增强模型对未来帧预测能力。该工作提出一种时空自动编码器（spatio-temporal autoencoder，STAE），利用三维卷积从视频的空间和时间两个维度提取特征的同时，引入一种用于生成未来帧的减权预测损失，从而增强了视频中的运动特征学习能力。此外，Wang 等人（2023b）提出一种基于物体间时空关系的视频异常检测方法。该方法利用注意力机制增强了模型

对视频中各类物体之间时空关系的理解，从而提高了模型对正常样本未来帧的预测能力。

尽管增强模型对视频外观和时空等信息的关注能够提升模型性能，但有研究指出，自动编码器由于强大的泛化能力，可能导致某些异常帧被重建或预测得较好，从而引发假阳性问题，导致漏检（Gong等，2019）。针对这一问题，Gong等人（2019）提出在自动编码器中引入记忆模块，以限制模型的泛化能力。记忆扩充编码器（memory-augmented deep auto-encoder，MemAE）利用编码器生成编码，然后将其用做查询以检索最相关的内存项进行重建。在训练阶段，更新存储正常数据的原型；在测试阶段，测试样本结合最相似的存储编码特征进行预测。因此，预测结果将趋向于接近正常样本，从而增强了正常和异常的预测差异，提高了模型的性能。Park等人（2020）对MemAE进行改进，引入分离性损失和紧密性损失。分离性损失促使存储中正常原型之间的差异增大，强化存储的多样性；紧密性损失使样本与存储中最相似的正常原型关联性最大化，强化存储的代表性。Tao等人（2024）利用类似理论设计了内存模块，用于捕捉正常模式，以及伪标签生成模块和一个用于负学习的异常事件生成模块，用于获得与内存模块匹配的正常事件表示和帮助模型在严格的无监督设置下取得更好性能。Wang等人（2023a）提出一种专注于外观和运动信息的记忆增强网络（memory-augmented appearance-motion network，MAAM-Net）。该网络利用基于边际的潜在损失，扩大了记忆模块中模式间的边际距离，从而强化了记忆能力。

在基于VAE或U-Net结构的视频异常检测方法之外，Lee等人（2018）创新性地提出一种基于时空生成对抗网络的视频异常检测方法。首先，作者利用双向ConvLSTM的时空特性，设计了时空生成器，并相应地设计了时空鉴别器，用于确认生成特征是否符合正常样本的时空特征。通过这两个网络进行对抗训练，使网络有效地学习编码样本正常的时空特征模式。最后，生成的结果与学到的正常模式相比，如果偏差超过阈值，则被检测为异常。Sabokrou等人（2018b）、Ravanbakhsh等人（2019）、Vu等人（2019）、Zaheer等人（2020）和Barbalau等人（2023）对基于生成对抗网络的视频异常检测进行了进一步研究。他们的研究认为，由于对抗性训练的参与导致模型不稳定或产生过高的假阳性结果。因此，提出将鉴别器从识别真实和虚假数据的任务转变为区分重建质量的好坏。为此，建议将生成器训练成生成数据的正态分布表示，这对任务更为适用。同时，建议利用当前生成器来准备高质量的重构训练样本，同时利用同一生成器的旧状态获取质量较差的示例或伪异常，使鉴别器能够检测在异常输入的重建中经常出现的微小失真。此外，结合去噪自动编码器（Lu等，2013b）与条件生成对抗网络（Mirza和Osindero，2014），充分利用去噪自动编码器强大的表示学习能力和基于条件生成对抗网络的层次化表示生成，以获得更稳定的模型和更好的模型效果。

由于无监督训练数据信息量有限，一些研究者建议利用预训练模型引入外部信息以进行异常检测。Morais等人（2019）以人体姿态信息为出发点，提出一种利用人体骨架特征的视频异常检测方法。通过设计循环神经网络（recurrent neural network，RNN），创建两个分支以分别处理全局和局部的人体骨架特征，并通过每个步骤进行跨分支消息交互。这一方法通过解耦的特征相互作用，能够准确地从监控视频中识别与人相关的异常事件。Markovitz等人（2020）则提取不同的人体姿态特征并进行聚类，每个动作对应一个群集。这为数据提供了一种“词袋”表示，其中每个动作都可由一组基本动作词的相似性表示。通过相似性对比的方式来判定是否存在异常。Barbalau等人（2023）提出一种引入外部知识的自监督多任务学习框架，利用YOLO（you only live once）框架引入目标位置、人体骨架等信息，并同时引入了二维和三维卷积视觉转换器模块。作者比较了多种外部信息引入配置，并找到最佳配置以实现当时最佳的异常检测性能。Yan等人（2024）认为视频帧包含面部信息和人类目标等隐私敏感信息，提出一种隐私保护的异常检测框架，引入图像分割掩码来保护人类目标的隐私。同时，通过引入上下文信息的对象检测来提高异常检测性能。

除了上述研究角度之外，Lu等人（2020）首次探索了小样本学习与无监督视频异常检测任务的结合。作者将元学习（Finn等，2017）与自编码器相结合，形成了自己的框架，在学习鉴别新异常能力时只需少量无监督数据。在Lu等人（2020）基础上，Lyu等人（2021）和Guo等人（2024）进行了优化，Lyu等人（2021）向原模型结构中添加了可微存储模块，并引

入元学习策略到异常检测中，进一步提升了在少样本场景下的异常检测模型性能。Guo等人(2024)提出一个自适应视频异常检测框架解决小样本问题，在预训练阶段合成异常样本，并设计基于自监督的预测任务来预训练域不变模块；在适应阶段，通过对抗训练方法减少分布偏移，将预训练模型适应到目标域的少样本中提高模型性能。此外，为解决在现实场景中数据标记困难和建模困难问题，Wu等人(2024c)提出一种能够自适应地选择合适的网络架构的动态网络方法。该方法利用生成的时空伪异常数据和正常数据作为网络的输入，设计了混合异常动态卷积(hybrid anomaly dynamic convolution，HADConv)以从多样的异常中自适应提取特征。最终，该方法在4个公共数据集上实现了显著的性能提升。Cao等人(2024)根据测试事件与来自训练数据的正态性知识之间的一致性来检测异常事件，提出一种基于上下文恢复和知识检索的新型双流框架，这两种流可以相互补充，可以充分利用运动信息来预测未来帧。

无监督方法无需明确定义异常和正常，因此通常被设计为单类分类(one-class classification)任务，旨在检测不符合常见数据场景的情况。在实际研究中，一般认为可以相对容易地获取大量的视频数据和充足的计算资源，而这其中大部分都是正常数据。因此，与有标注训练数据的方法相比，无监督方法的研究更具实现的便利性。目前，许多无监督方法通过引入预训练模型并融入外部知识，取得与有监督方法不相上下的性能。然而，由于无监督数据不定义异常的具体特征与判断规则，这导致在实际应用中模型更容易出现错检和漏检情况，并且具有通用性、可移植性较差和依赖外部信息等局限性。

3.4 结合大模型技术的视频异常检测方法

随着各类大模型迅速发展，可以清晰地看到多模态信息处理能力正在逐步融入预训练大模型体系中。大规模自监督预训练最初源于自然语言处理领域，近年来逐渐扩展到包括图像、音频、视频等多模态数据处理任务。例如，Google发布的文本到图像扩散模型Imagen(Saharia等，2022)，OpenAI提出的CLIP(contrastive language-image pre-training)(Radford等，2021)文本图像匹配模型，以及Salesforce提出的BLIP(bootstrapping language-image pre-training)(Li等，2022)和BLIP2(Li等，2023)多模态大模型等等。在视频异常检测领域，目前已有一些结合多模态大模型的探索性工作，这些工作使该领域呈现出新的活力。

在人工智能(artificial intelligence，AI)大模型中，Prompt能够更好地帮助模型理解输入的意图，并作出相应的响应。受到Prompt的启发，Liu等人(2023)提出一种新颖的基于Prompt的特征映射方法PFMF(prompt-based feature mapping framework)，用于视频异常检测。通过设计异常Prompt引导的映射网络，该方法致力于生成真实场景中未曾见过的具有无限类型的异常，解决了将虚拟异常视频异常检测数据集应用于真实场景的挑战，对模型实际应用具有重要价值。Joo等人(2023)建议利用CLIP中的ViT(vision Transformer)编码视觉特征。与该领域中的传统C3D(convolutional 3D)或I3D(inflated 3D network)特征相比，这项新技术能更有效地提取判别性表征。此外，作者还引入了时间自注意力，以引导模型关注关键片段，最终实现模型性能大幅超越现有最先进的方法。

此外，还有一些工作直接基于多模态大模型框架设计视频异常检测方法。Zanella等人(2023)将CLIP与视频异常检测结合，并有针对性的方法设计。作者引入正态原型驱动的大语言模型特征空间的变换策略，以更有效地生成不同异常类型的文本提示特征表示，更好地区分不同种类的异常。此外，通过利用相邻帧之间的短期关系和片段之间的长期依赖关系进行特征增强，更好地利用了视频的时间信息。Wu等人(2024b)提出基于CLIP的双分支结构，其中一个分支利用CLIP视觉编码模块对视频进行粗粒度的二元分类；另一个分支则充分利用异常类别标签的语言特征与视觉编码特征对齐实现细粒度的异常分类，使其在视频异常检测性能上超越了当前最佳水平。Zanella等人(2024)提出一种基于语言的视频异常检测方法LAVAD(language-based VAD)(Liu等，2024)，这是一种新颖的、无需训练的范式，仅依赖预训练的大型语言模型(large language model，LLM)和现有的视觉—语言模型LLAVA的能力进行异常检测。首先基于LLAVA为测试视频的每一帧生成文本描述。通过文本场景描述，设计了一种提示机制，以解锁LLM在时间聚合和异常评分估计方面的能力，使LLM成为一个有效且通用的视频异常检测器。

多模态大模型不仅在有监督性能上表现出色，更在少样本、零样本已经开放词汇场景中展现出强大的潜力。Kim等人(2023)主张对训练数据集中未包含的异常种类进行提前定义。通过结合多模态大模型，探索将这些文本描述与未标记的视频数据集一起使用的潜力。通过大型语言模型获取符合数据集场景的文本描述，然后结合这些文本描述以形成不同数据集的文本描述组合。接着，使用CLIP视觉语言模型计算输入帧和文本描述之间的余弦相似度来检测异常帧。最后，对相似性度量方法进行了改进，以进一步提升模型性能，使其优于现有的无监督方法。Wu等人(2024a)利用预先训练好的大型模型来检测和分类已见和未见的异常，提出将开放词汇视频异常检测问题分解为无类别检测和类别特定分类(class-agnostic detection and class-specific classification)两个相辅相成的任务并共同优化这两个任务。此外，在模型中设计了一个语义知识注入模块，以引入来自大型语言模型ERNIE Bot(enhanced language representation with informative entities)(Sun等，2020)的语义知识用于检测任务，并设计了一个新颖的异常合成模块，在大型视觉生成模型DALL-E mini(draw & all image maker enhanced mini)(Ramesh等，2022)的帮助下生成伪未见异常视频用于分类任务。这些语义知识和合成异常扩展了该模型在检测和分类各种已见和未见异常方面的能力。Cao等人(2023)探索了以通用方式使用一种强大的大型视觉语言模型GPT-4V(generative pre-trained Transformer 4 vision)解决异常检测任务的可能性。该研究考察了GPT-4V在多模态、多域异常检测任务中的应用，包括图像、视频、点云和时间序列数据，覆盖工业、医疗、逻辑、视频、3D异常检测和定位等多个应用领域。在实验中，作者采用了多种附加提示，如类别信息、人类专业知识和参考图像，以增强GPT-4V的异常检测性能。实验结果表明，GPT-4V在单/零样本异常检测方面表现出色，充分展示了大模型在异常检测领域的强大潜力。

目前的视频异常检测研究充分展示了结合大型模型在提升性能和实现细粒度异常检测方面的显著效果。此外，它在处理单/零样本任务时表现出超越现有方法的潜力，并能够较好的实现。因此，我们相信结合大型模型的视频异常检测工作有望实现更精确、更通用的视频异常检测。

4 常用数据集介绍

在人工智能时代，数据集的发展成为技术进步的关键因素。数据集能够为问题的明确定义提供具体的范围，并为研究提供公平的评估和比较方式。目前视频异常检测领域共有10个常用数据集，包括2个仅带有视频级标注(弱标签)和8个带有帧级或更细粒度的标注，其中，有4个是2020年以来发布的数据集。相比之前的数据集，它们提供了更丰富的标记形式或者针对特定问题提供了更高质量的数据。图9展示了各个数据集中的正常和异常视频截图，异常部分用红色方框标记。表1总结了各数据集的要点。

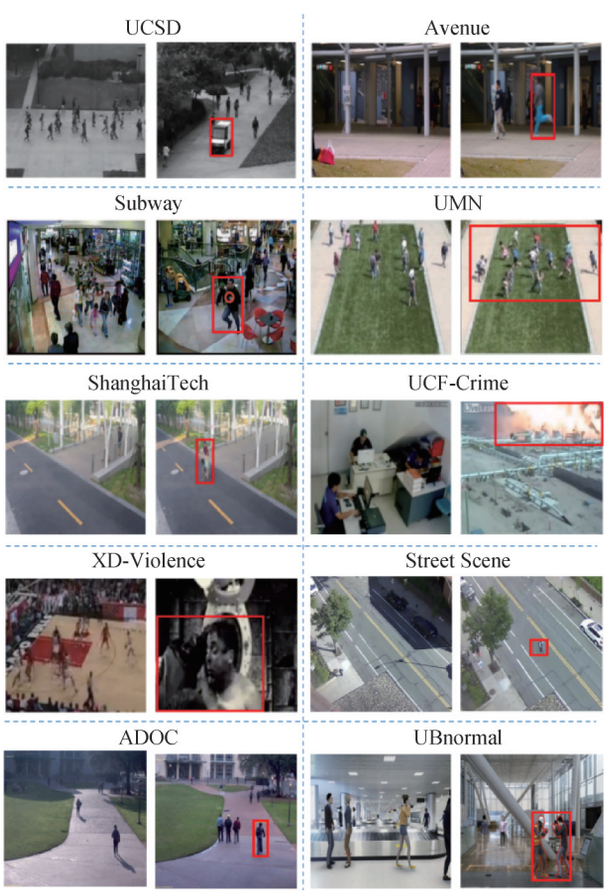

图9 各数据集中的正常和异常视频截图

Fig. 9 Screenshots of normal and abnormal videos from each dataset

4.1 数据集介绍(2020年之前)

UCSD (University of California, San Diego)(Mahadevan等，2010)数据集中包含Ped1和Ped2两

表 1　现有数据集对比表

Table 1　Existing dataset comparison tables

数据集	视频数	时长	标注	异常种类
UCSD(Mahadevan 等,2010)	98	10 min	像素级	5
Avenue(Lu 等,2013a)	15	30 min	像素级	3
Subway(Adam 等,2008)	2	2.32 h	帧级	8
UMN(University of Minnesota,2011)	11	5 min	帧级	1
ShanghaiTech(Luo 等,2017b)	437	3.6 h	像素级	11
UCF-Crime(Sultani 等,2010)	1 900	128 h	视频级	13
XD-Violence(Wu 等,2020)	4 754	217 h	视频级、音频	6
Street Scene(Ramachandra 和 Jones,2020)	81	4 h	帧级、打框	17
ADOC(Mantini 等,2021)	1	24 h	帧级、打框	25
UBnormal(Acsintoae 等,2022)	29	2.2 h	像素级	22

个数据集,分别包括校园内竖直和水平方向上正常行走的人行道上固定监控视角的视频。其中 Ped1 中存在标记错误,但在后期被改正(Vu 等,2019)。该数据集包含 98 个视频片段,共有 18 560 帧,总时间长度为 10 min。数据集的标记方式为帧级标注(对每帧是否异常进行了标注)以及像素级标注(对每帧中每个像素点是否属于异常区域进行了标注)。数据集中的训练集只包含人在人行道的正常视频,测试集中包含正常和异常行为,异常行为有人行道上骑自行车、开车、轮椅、滑板和踩草坪等异常行为。数据下载链接:http://www.svcl.ucsd.edu/projects/anomaly/UCSD_Anomaly_Dataset.tar.gz。

Avenue(Lu 等,2013a)数据集是在公共场所利用固定摄像头拍摄的数据集,共包含 21 个视频片段,共有 30 652 帧,总时间长度为 30 min。数据集的标注方式有帧级别和像素级别标注,并存在一些镜头抖动以及标注遗漏等问题。其中所包含的异常行为有行人奔跑、抛掷物体、异常物体、徘徊等。数据下载链接:http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/Avenue_Dataset.zip。

Subway(Adam 等,2008)数据集是固定角度拍摄设备在地铁出入口拍摄的数据集,包含出口入口两个视频,共 125 475 帧。数据集的标注方式只有帧级别标注,其中标注为异常行为主要有行人错误的行走方向、逃票和徘徊等。数据下载链接:https://vision.eecs.yorku.ca/research/anomalous-behaviour-data/。

UMN(Unehran 等,2009)数据集包含 3 个场景,分别是在室内、广场和草坪 3 种场景下用固定摄像机拍摄的,共包括 11 个视频,总时长为 5 min,共有 3 855 帧。数据集的标注方式只有帧级别标注,其中被标记为异常的事件包括人群突然四散等。数据下载链接:http://people.ece.umn.edu/users/parhi/.DATA/OCT/DME/UMNDataset.mat。

ShanghaiTech(Luo 等,2017b)数据集包含 13 个不同光照条件和拍摄角度的场景,这也是近几年常用数据集中最有挑战的数据集之一。数据集共有 437 个视频,共包括 317 398 帧。数据集的标记方式为帧级标注以及像素级标注(对每帧中每个像素点是否属于异常区域进行了标注)。数据集中包含的异常事件有翻越围栏和在人行道骑车、跑步和滑滑板等。数据下载链接:https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Zhang_Single-Image_Crowd_Counting_CVPR_2016_paper.html。

UCF-Crime (University of Central Florida crime dataset)(Sultani 等,2008)数据集包括 1 900 个未经过剪辑处理的各类监控视频,共 128 h,13 820 600 帧,涵盖 13 种异常类型,包括有虐待、抓捕、纵火、突袭、盗窃和斗殴抢劫等。该数据集对视频进行了视频级标注(针对视频中是否包含异常帧进行标注,包含任意一帧的为异常视频,反之为正常视频),是弱监督任务常用数据集之一。该数据集虽然数据量大、异常种类多,但是也有许多不足之处,例如同一个视频

内容重复、视频尺寸不一和视频色彩不一等。数据下载链接：https://aimagelab.ing.unimore.it/imagelab/page.asp?IdPage=30。

4.2 新数据集介绍（2020年之后）

近几年来，异常事件检测工作不仅提出了新颖的方法，数据集方面也有进一步的发展。在研究过程中，某些工作认为之前常用的数据集只有单一的图像标注或不满足任务需要而提出了更适合研究问题的数据集。

XD-Violence（Wu等，2020）数据集是一个大规模多场景数据集，总持续时间为217 h，包含4 754个含音频信号的未剪辑视频，涵盖6种常见的异常类型，该数据集只对视频进行视频级标注（弱标签）。该数据集是目前视频异常检测最大的数据集，其中包含的异常行为有车祸、虐待和枪击等。数据下载链接：https://roc-ng.github.io/XD-Violence/。

Street Scene（Ramachandra和Jones，2020）数据集是由静态摄像机在一条有自行车道与行人道的双车道场景记录下的，共有81个视频17种异常类型，包含203 257个视频帧。该数据集目前只提供了帧级别标注，其中异常事件包括人行道逆行、乱停车和游荡的人等。数据下载链接：https://www.merl.com/demos/video-anomaly-detection。

ADOC（anomaly detection in object-centric video）（Mantini等，2021）数据集是由固定摄像头在校园内记录24 h构造而成。ADOC数据集包括25种异常类型，并且是拥有边界框注释的最大数据集，数据集中存在的异常事件有人行道骑自行车、踩草坪、吸烟、开高尔夫车和遛狗等。数据下载链接：http://qil.uh.edu/main/datasets/。

UBnormal（ubnormal video anomaly detection dataset）（Acsintoae等，2022）数据集是由Cinema4D软件生成的虚拟数据集，涵盖了29种场景，如街道、火车站和办公室。正常视频与异常视频的比例接近1∶1，正常动作包括行走、边走边发短信、站立、坐着以及与他人交谈等；而异常事件则包括跑步、跌倒、打架、跳舞、偷窃、火灾、在车道外行驶和跳跃等22种类型。数据集还覆盖多个对象类别，并确保了异常事件执行者的多样性。该数据集中的异常经过像素级别的注释，提供了分割掩码和对象标签，是目前最大且标签最丰富的虚拟视频异常检测数据集。数据下载链接：http://suo.nz/2Rix5f。

5 视频异常检测性能评估

5.1 异常判定标准

在视频异常检测领域主要的异常判定标准有帧级和像素级标准。此外，还有相对不常见的基于区域和基于轨迹两种标准。

1）基于帧级的标准。在帧级标准下，当帧的异常打分大于阈值时，该帧被视为异常帧。这种方式不考虑异常的具体位置，无法做到精确的异常目标定位。

2）基于像素级的标准。对于像素级标准，当帧中一定数量像素点的重构误差大于阈值时，该帧被认为是异常帧。这种方法能够定位异常的具体位置。通常当被判断为异常的像素覆盖超过40%的异常标注像素时，认为该预测是真阳性，否则是假阳性。

3）基于区域的标准（Ramachandra和Jones，2020）。模型检测为异常的区域是指检测值大于给定阈值的像素相连接的区域。当检测为异常区域与真实异常区域的交并比（intersection over union，IoU）大于设定阈值时，认为是真阳性；其他情况认为是假阳性。该评价标准的作者还指出，检测为异常的区域可能同时跨越多个真实异常区域，导致真阳性计数不平衡。为缓解这一问题，作者建议将较低的阈值设为0.1，实验证明这是合适的。

4）基于轨迹的标准（Ramachandra和Jones，2020）。在连续的异常帧中，异常区域通常高度相关。对于发生在多帧上的异常，重要的是在至少一些帧中检测到异常区域，但在轨迹中的每一帧中检测到该区域通常并不重要。考虑到异常轨迹何时开始和结束的不确定性，以及在异常活动被严重遮挡几帧的情况下，这一点尤为正确。在街景数据集中，每个异常区域都有唯一的轨迹编号，用于标识异常区域所属的事件。如果被检测为异常区域的帧连续10帧（原文中的假设）则认为检测正确。如果其中一帧的异常检测区域与真实异常区域之间的IoU大于0.1，则认为检测到了该异常。

5.2 算法评估指标

在VAD领域，主要有以下3种方式来对模型的性能进行评估，本节将进行简要介绍。

1）AUC（area under curve）。是ROC（receiver

operating characteristic)曲线下的面积。ROC曲线以假阳率(false positive rate, FPR)为横坐标、真阳率(true positive rate, TPR)为纵坐标绘制。ROC曲线上的每个点代表分类器在特定阈值下的TPR和FPR。因此,AUC反映了分类器对正负样本的区分能力。AUC的取值范围在0.5～1之间,数值越接近1表明分类器的性能越优越。AUC的计算方法是对ROC曲线下每个小矩形的面积进行累加。假设有N个小矩形,则AUC的计算式为

$$AUC = \sum_{i=1}^{N} \frac{(TPR[i] + TPR[i-1])}{2} \times (FPR[i-1] - FPR[i]) \tag{2}$$

式中,$TPR[i]$和$FPR[i]$分别表示第i个点的TPR和FPR。ROC曲线通过绘制FPR和TPR之间的关系,展示了分类器在不同阈值下的性能变化情况。

2)等错误率(equal error rate,EER)。定义为当假阳率(FPR)和假阴率(false negative rate,FNR)相等时,错误分类视频帧所占的百分比。计算EER的具体步骤为,计算假阴率($FNR = 1 - TPR$),在ROC曲线上找到FPR与FNR相等的点,即

$$EER = FPR, \text{when } FPR = FNR \tag{3}$$

总的来说,EER是通过分析系统在不同决策阈值下的表现,通过ROC曲线找到假阳率和假阴率相等的点来确定的。EER越低,表示异常检测性能越好。

3)平均精度(average precision,AP)。是基于查准率—召回率曲线(precision-recall curve)下面积计算得出。AP取值范围在0～1之间。AP越接近1,表示模型性能越好,反之越差。具体来说,对于每个类别,可以根据模型预测的置信度对预测结果进行排序,然后计算不同召回率下的精确度,最后计算这些精确度的平均值,得到该类别的AP值。

5.3 算法效果对比

本文根据不同的监督方式对各类视频异常检测方法进行了性能比较分析,同时针对结合大模型的方法进行了详细的性能比较分析。

如表2所示,在6个数据集上比较了5个全监督方法的性能表现。由表2可知,帧级任务性能通常较佳,而像素级任务的性能明显低于帧级任务。这主要是因为视频数据具有时空性,异常区域的位置不固定,因此需要更丰富的标注或特征提取方法来有效定位像素级的异常。但是,随着无监督和弱监督方法的迅速发展,需要复杂标注信息的全监督方法的进展逐渐趋缓。

表2 全监督方法在6个常用数据集上的AUC对比

Table 2 AUC comparison of supervised methods in six common datasets

/%

方法	监督方式	Ped1		Ped2		Avenue		Subway Exit	Subway Entrance	UMN
		像素级	帧级	像素级	帧级	像素级	帧级	帧级	帧级	帧级
Tran和Hogg(2017)	有监督	**68.7**	91.9	**89.3**	**96.6**	–	**82.1**	–	–	–
AMDN(Xu等,2017)	有监督	67.2	92.1	–	90.8	–	–	–	–	–
Sabokrou等人(2017)	有监督	–	**93.2**	–	93.9	–	–	–	–	**99.6**
Unmasking(Ionescu等,2017)	有监督	52.4	68.4	–	82.2	**93**	80.6	**70.6**	**85.7**	95.1
Stacked-RNN(Luo等,2017a)	有监督	–	–	–	92.2	–	81.7	–	–	–

注:加粗字体表示各列最优结果。"–"表示缺少实验。

如表3所示,对27个无监督方法在5个数据集上的性能进行了比较。从表3可以得知,无监督方法在这个领域已经取得了显著的进展,其性能超越了相同数据集上的全监督方法,并在多个数据集上实现了超过90% AUC的出色性能。尤其是在引入外部信息的Georgescu等人(2021)方法和SSMTL++(self-supervised multi-task learning)方法在多个数据集上表现最佳。然而,引入外部信息需要异常对象相关的先验信息,这可能降低模型的通用性。因此,方法研究的多元化发展显得尤为重要。

如表4所示,对20个弱监督方法在5个数据集上的性能进行了比较。弱监督方法在与无监督方法

表3　无监督方法在5个常用数据集的AUC对比

Table 3　AUC comparison of unsupervised methods in five common datasets

/%

方法	监督方式	Ped1		Ped2		Avenue		UMN	ShanghaiTech
		像素级	帧级	像素级	帧级	像素级	帧级	帧级	帧级
Online GNG(Sun等,2017)	无监督	65.1	93.8	–	94.1	–	–	99.7	–
DeepAppearance(Smeureanu等,2017)	无监督	–	–	–	–	**93.5**	84.6	97.1	–
STAE(Zhao等,2017)	无监督	–	92.3	–	91.2	–	80.9	–	–
Chong 和Tay(2017)	无监督	–	89.9	–	87.4	–	80.3	–	–
ConvLSTM-AE(Luo等,2017b)	无监督	–	75.5	–	88.1	–	77.0	–	–
Liu 等人(2018)	无监督	–	83.1	–	95.4	–	85.1	–	72.8
AVID(Sabokrou等,2018b)	无监督	–	–	–	–	–	–	99.6	–
S2-VAE(Wang等,2018)	无监督	**94.3**	–	–	–	–	87.6	99.5	–
STAN(Lee等,2018)	无监督	–	82.1	–	96.5	–	87.2	–	–
Ravanbakhsh 等人(2019)	无监督	70.8	**96.8**	–	95.5	–	–	98.9	–
Vu 等人(2019)	无监督	66.6	82.3	**97.2**	99.2	52.8	71.5	–	–
Mem-AE(Gong等,2019)	无监督	–	–	–	94.1	–	83.0	–	71.2
Appearance-Motion(Nguyen和Meunier,2019)	无监督	–	–	–	96.2	–	86.9	–	–
AnoPCN(Ye等,2019)	无监督	–	–	–	96.8	–	86.2	–	73.6
MNAD(Park等,2020)	无监督	–	–	–	97.0	–	88.5	–	70.5
Few-Shot(Lu等,2020)	无监督	–	86.3	–	96.2	–	85.8	–	77.9
VEC(Yu等,2020)	无监督	–	–	–	97.3	–	89.6	–	74.8
Georgescu 等人(2021)	无监督	–	–	–	**99.8**	–	92.8	–	**90.2**
MPN(Lyu等,2021)	无监督	–	85.1	–	96.9	–	89.5	–	73.8
STR-VAD(Wang等,2023b)	无监督	–	–	–	98.4	–	86.1	–	73.2
MAAM-Net(Wang等,2023a)	无监督	–	–	–	97.7	–	90.0	–	71.3
SSMTL++(Barbalau等,2023)	无监督	–	–	–	–	–	**93.7**	–	83.8
DSS-Net(Wu等,2024c)	无监督	–	85.6	–	97.2	–	90.6	99.3	75.5
MGAN-CL(Li等,2024b)	无监督	–	–	–	96.5	–	87.1	–	73.6
Two-stream(Cao等,2024)	无监督	–	–	–	97.1	–	90.8	99.2	83.7
Tao 等人(2024)	无监督	–	87.8	–	99.5	–	89.7	**99.9**	–
Ada-VAD (Guo等,2024)	无监督	–	–	–	99.2	–	90.0	–	77.1

注:加粗字体表示各列最优结果。“–”表示缺少实验。

相同的数据集上表现出色,在弱监督数据集上仍有提升的空间。这是因为弱监督数据集通常设计用于开放世界异常检测,具有多变的场景和更为复杂的异常。弱监督方法仅少量视频级数据即可检测指定异常,通过添加新的标记数据以适应新场景和任务,相较于无监督方法,更具可解释性和可控性。

如表5所示,比较了7种结合大模型技术的方法在4个数据集上的性能。在ShanghaiTech数据集中,结合大模型技术的无监督方法(Kim等,2023)比性能最好的先前无监督方法(Georgescu 等,2021)提升了7.6%,并超越了大多数现有的弱监督方法。结合大模型技术的弱监督方法(Joo 等,2023;Wu 等,2024b)在各弱监督数据集上都达到了最佳性能。此外,这类方法在开放词汇(Wu等,2024a)和无需训练场景(Zanella等,2024)中也表现良好。以上实验结果表明,采用大模型技术的方法在无监督和弱监督

表4　弱监督方法在5个常用数据集的AUC/AP对比

Table 4　AUC/AP comparison of weakly methods in five common datasets

/%

方法	监督方式	AUC						AP
		Ped1		Ped2		ShanghaiTech	UCF-Crime	XD-Violence
		像素级	帧级	像素级	帧级	帧级	帧级	帧级
Sultani 等人(2018)	弱监督	–	–	–	–	–	75.4	–
GCN-Anomaly(Zhong等,2019)	弱监督	–	–	–	93.2	–	82.1	–
MLEP(Liu等,2019b)	弱监督	–	–	–	–	–	76.8	–
IBL(Zhang等,2019)	弱监督	–	–	–	–	–	78.7	–
Motion-Aware(Zhu和Newsam,2019)	弱监督	–	–	–	–	–	79.0	–
Background-Bias(Liu和Ma,2019a)	弱监督	–	–	–	–	–	82.0	–
Siamese(Ramachandra和Jones,2020)	弱监督	**80.0**	**86.0**	**93.0**	94.0	–	–	–
ARNET(Wan等,2020)	弱监督	–	–	–	–	91.2	–	–
XD-Violence(Wu等,2020)	弱监督	–	–	–	–	–	–	78.6
MIST(Feng等,2021)	弱监督	–	–	–	–	94.8	82.3	–
CLAWS(Chang等,2020)	弱监督	–	–	–	–	89.7	83.0	–
RTFM(Tian等,2021)	弱监督	–	–	–	**98.6**	97.2	84.3	77.81
CNL(Liu等,2022)	弱监督	–	–	–	–	88.2	83.1	–
Huang等人(2024)	弱监督	–	–	–	–	98.1	84.0	74.6
Zhang 等人(2023)	弱监督	–	–	–	–	–	86.2	**81.4**
MGFN(Chen等,2023)	弱监督	–	–	–	–	96.4*	86.7	80.1
Light-WSVAD(Wang等,2024b)	弱监督	–	–	–	–	95.9	84.7	72.5*
REWARD(Karim等,2024)	弱监督	–	–	–	–	95.1*	**86.9**	77.7
HSN(Majhi等,2024)	弱监督	–	–	–	–	96.2	85.3	–
AFT(He等,2024)	弱监督	–	–	–	–	**98.2**	85.1	80.1

注:加粗字体表示各列最优结果。“–”表示缺少实验。“*”表示实验结果为本文复现。

数据集上均取得了显著性能提升,甚至在开放词汇和无需训练的场景下也能有效识别视频异常,展现出巨大的发展潜力。

6　总结与展望

本文首先详细阐述视频异常的概念和类别。接着从深度学习的角度,对当前关键的视频异常检测方法进行全面综述,着重介绍结合大模型技术的最新视频异常检测方法的发展。此外,对现有的数据集和评估方法进行介绍,分析它们之间的差异和优劣,为读者在选择数据集和评估方法时提供了参考。最终,在表2—表5中对介绍的经典工作在各类数据集上的性能进行比较和分析。特别是重点介绍了结合大模型技术的方法,这类方法刷新了多个数据集的性能记录,并且在任意形式监督的数据集下,凭借多模态大模型的强大先验知识,实现了细粒度的异常检测。

6.1　数据集的展望

现有的常用数据集主要存在场景简单、镜头固定和标注单一等问题。以UMN数据集为例,S2-VAE方法的AUC为99.5%,其性能表现已基本达到饱和。此外,在Ped1、Ped2、Avenue和ShanghaiTech等常用数据集上,也有方法的AUC超过90%,表示

表5　结合大模型技术的方法在4个常用数据集的AUC/AP对比

Table 5　Methods incorporating large model technology in AUC /AP comparison on 4 common datasets

/%

方法	训练方式	基于的大模型技术	AUC			AP
			Avenue	ShanghaiTech	UCF-Crime	XD-Violence
			帧级	帧级	帧级	帧级
PFMF(Liu等,2023)	无监督	Prompt	**93.6**	85.0	–	–
CLIP-TSA(Joo等,2023)	弱监督	CLIP	–	**98.3**	87.6	–
Anomaly CLIP(Zanella等,2023)	弱监督	CLIP	–	98.1	86.4	78.5
Kim 等人(2023)	无监督	CLIP	–	97.8	85.8	–
VadCLIP(Wu等,2024b)	弱监督	CLIP	–	98.0*	**88.0**	**84.5**
OVVAD(Wu等,2024a)	弱监督	CLIP+ERNIE Bot+ DALL-E mini	–	–	86.4	66.5
LAVAD(Zanella等,2024)	无需训练	LLAVA	–	81.2*	80.3	62.0

注:加粗字体表示各列最优结果。"–"表示缺少实验。"*"表示实验结果为本文复现。

数据集可能已经成为当前方法发展的瓶颈。因此,在复杂的实际场景中,仅以简单场景为背景的研究方法可能无法有效解决真实世界中的异常问题。因此,为了更好地促进研究的发展,未来的数据集将朝着更好地反映真实世界异常的目标发展,例如收集利用遥感领域数据(Hong等,2023)、通过模型提高现有图像视频数据质量(李晨玉 等,2024)以及收集多镜头、多维度标注数据等,以实现对更多样化、更具挑战性的异常事件的检测。此外,大模型时代下的数据集也需要向包含多模态标注发展,有充足的数据支持,才有望实现更通用的视频异常检测方法。

6.2　评估方法的展望

常见的评估方式主要依赖于计算真阳性率和假阳性率,并计算ROC曲线下的面积AUC。然而,在实际应用中,一些方法尽管具有较高的AUC,但却已经具有较高虚假报警概率。因为,真阳性率和假阳性率直接受不同的异常判别方式影响,采用不同的异常判别方式可能使模型在取得高AUC性能的同时产生高虚假报警概率。因此,需要设计一种评价体系,该体系可以同时关注AUC性能和虚假报警概率,以更全面地对方法进行评估。

6.3　模型的展望

近年来,尤其是大模型的涌现,基于深度学习的方法在视频异常检测的常用数据集上取得了显著的性能提升。该领域已经积累了充分的学术研究基础。因此,本文认为未来的研究不应仅关注在检测异常的性能上,而是需要考虑该领域在实际问题中的应用,以解决其中的难题。

1)未来的发展趋势之一是设计更细粒度通用的模型,结合大模型的丰富先验知识,逐步能够设计可以区分具体异常种类的视频异常检测模型。基于大模型强大的多模态信息理解能力,视频异常检测模型将朝着更通用的方向发展(Wu等,2024a;Zanella等,2024),有监督、弱监督和无监督等学习方式的边界将变得模糊。

2)面向边缘计算的轻量视频异常检测:在边沿计算场景下,需要将视频异常检测部署到边缘设备,如监控摄像头,因此确保模型的轻量化至关重要(Wang等,2024b)。未来的研究工作之一是如何在保持模型轻量的同时,进一步提高其检测性能,以推动该领域技术在实际应用中的广泛落地。

3)在视频异常检测中,不应该仅局限于关注与人相关的异常事件,未来应进一步将现有方法衍生至动物行为图像视频异常(Zhou等,2024)、遥感图像/视频异常(Li等,2024a)和工业生产图像/视频异常等领域中。

4)考虑到异常是无法穷举的,并且可能随着政策规则的改变而变化,本文认为研究具有在线学习或终身学习能力的模型(Lyu等,2021)是该领域发展的必要方向。

5)高性能视频异常预测:预测异常事件的发生对于预防或减轻潜在的损害至关重要。现有研究主

要集中在检测已发生的异常，但预测未来异常事件的能力允许采取预防措施，从而在保护生命财产安全、有效遏制违法犯罪行为中起到更有效的作用（Wang等，2024a）。为此，应该考虑借鉴并整合相关领域（特别是时序数据预测领域）和现有视频异常预测的最新研究成果，加速视频异常预测技术的发展。

参考文献（References）

Acsintoae A, Florescu A, Georgescu M I, Mare T, Sumedrea P, Ionescu R T, Khan F S and Shah M. 2022. UBnormal: new benchmark for supervised open-set video anomaly detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 20111-20121 [DOI: 10.1109/CVPR52688.2022.01951]

Adam A, Rivlin E, Shimshoni I and Reinitz D. 2008. Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(3): 555-560 [DOI: 10.1109/TPAMI.2007.70825]

AlMarri S, Zaheer M Z and Nandakumar K. 2024. A multi-head approach with shuffled segments for weakly-supervised video anomaly detection//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 132-142 [DOI: 10.1109/WACVW60836.2024.00022]

Amsaprabhaa M, Nancy J Y and Khanna N H. 2023. Multimodal spatio-temporal skeletal kinematic gait feature fusion for vision-based fall detection. Expert Systems with Applications, 212: #118681 [DOI: 10.1016/j.eswa.2022.118681]

Antić B and Ommer B. 2011. Video parsing for abnormality detection// Proceedings of 2011 International Conference on Computer Vision. Barcelona, Spain: IEEE: 2415-2422 [DOI: 10.1109/ICCV.2011.6126525]

Barbalau A, Ionescu R T, Georgescu M I, Dueholm J, Ramachandra B, Nasrollahi K, Khan F S, Moeslund T B and Shah M. 2023. SSMTL++: revisiting self-supervised multi-task learning for video anomaly detection. Computer Vision and Image Understanding, 229: #103656 [DOI: 10.1016/J.CVIU.2023.103656]

Cai R C, Zhang H, Liu W, Gao S H and Hao Z F. 2021. Appearance-motion memory consistency network for video anomaly detection// Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtually: AAAI: 938-946 [DOI: 10.1609/AAAI.V35I2.16177]

Cao C Q, Lu Y and Zhang Y N. 2024. Context recovery and knowledge retrieval: a novel two-stream framework for video anomaly detection. IEEE Transactions on Image Processing, 33: 1810-1825 [DOI: 10.1109/TIP.2024.3372466]

Cao Y K, Xu X H, Sun C, Huang X N and Shen W M. 2023. Towards generic anomaly detection and understanding: large-scale visual-linguistic model (GPT-4V) takes the lead [EB/OL]. [2024-05-31]. https://arxiv.org/pdf/2311.02782.pdf

Chang Y P, Tu Z G, Xie W and Yuan J S. 2020. Clustering driven deep autoencoder for video anomaly detection//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 329-345 [DOI: 10.1007/978-3-030-58555-6_20]

Chen Y X, Liu Z Z, Zhang B H, Fok W, Qi X J and Wu Y C. 2023. MGFN: magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 387-395 [DOI: 10.1609/AAAI.V37I1.25112]

Chong Y S and Tay Y H. 2017. Abnormal event detection in videos using spatiotemporal autoencoder//14th International Symposium on Advances in Neural Networks. Hokkaido, Japan: Springer: 189-196 [DOI: 10.1007/978-3-319-59081-3_23]

Cong Y, Yuan J S and Liu J. 2013. Abnormal event detection in crowded scenes using sparse representation. Pattern Recognition, 46(7): 1851-1864 [DOI: 10.1016/J.PATCOG.2012.11.021]

Dong C, Loy C C, He K M and Tang X O. 2016. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2): 295-307 [DOI: 10.1109/TPAMI.2015.2439281]

Fan Y X, Wen G J, Li D R, Qiu S H, Levine M D and Xiao F. 2020. Video anomaly detection and localization via Gaussian mixture fully convolutional variational autoencoder. Computer Vision and Image Understanding, 195: #102920 [DOI: 10.1016/J. CVIU. 2020.102920]

Feng J C, Hong F T and Zheng W S. 2021. MIST: multiple instance self-training framework for video anomaly detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14004-14013 [DOI: 10.1109/CVPR46437.2021.01379]

Feng Y C, Yuan Y and Lu X Q. 2017. Learning deep event models for crowd anomaly detection. Neurocomputing, 219: 548-556 [DOI: 10.1016/J.NEUCOM.2016.09.063]

Finn C, Abbeel P and Levine S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: JMLR.org: 1126-1135

Georgescu M I, Barbalau A, Ionescu R T, Khan F S, Popescu M and Shah M. 2021. Anomaly detection in video via self-supervised and multi-task learning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 12737-12747 [DOI: 10.1109/CVPR46437.2021.01255]

Gong D, Liu L Q, Le V, Saha B, Mansour R M, Venkatesh S and Van Den Hengel A. 2019. Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 1705-1714

[DOI: 10.1109/ICCV.2019.00179]

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets. Advances in Neural Information Processing Systems, Canada: Curran Associates Inc.: 2672-2680

Guo D L, Fu Y and Li S. 2024. Ada-VAD: domain adaptable video anomaly detection//Proceedings of 2024 SIAM International Conference on Data Mining. Houston, USA: SIAM: 634-642 [DOI: 10.1137/1.9781611978032.73]

Hasan M, Choi J, Neumann J, Roy-Chowdhury A K and Davis L S. 2016. Learning temporal regularity in video sequences//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 733-742 [DOI: 10.1109/CVPR.2016.86]

He P, Zhang F, Li G and Li H. 2024. Adversarial and focused training of abnormal videos for weakly-supervised anomaly detection. Pattern Recognition, 147: 110119-110128 [DOI: 10.1016/j.patcog.2023.110119]

Hinami R, Mei T and Satoh S. 2017. Joint detection and recounting of abnormal events by learning deep generic knowledge//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 3639-3647 [DOI: 10.1109/ICCV.2017.391]

Hong D F, Zhang B, Li H, Li Y X, Yao J, Li C Y, Werner M, Chanussot J, Zipf A and Zhu X X. 2023. Cross-city matters: a multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. Remote Sensing of Environment, 299: #113856 [DOI: 10.1016/j.rse.2023.113856]

Hu H Y, Zhang L and Li Z J. 2020. Anomaly detection with autoencoder and one-class SVM. Journal of Image and Graphics, 25(12): 2614-2629 (胡海洋, 张力, 李忠金. 2020. 融合自编码器和one-class SVM的异常事件检测. 中国图象图形学报, 25(12): 2614-2629) [DOI: 10.11834/jig.200042]

Huang C, Liu C L, Wen J, Wu L, Xu Y, Jiang Q P and Wang Y W. 2024. Weakly supervised video anomaly detection via self-guided temporal discriminative transformer. IEEE Transactions on Cybernetics, 54(5): 3197-3210 [DOI: 10.1109/TCYB.2022.3227044]

Ionescu R T, Khan F S, Georgescu M I and Shao L. 2019a. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7834-7843 [DOI: 10.1109/CVPR.2019.00803]

Ionescu R T, Smeureanu S, Alexe B and Popescu M. 2017. Unmasking the abnormal events in video//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2914-2922 [DOI: 10.1109/ICCV.2017.315]

Ionescu R T, Smeureanu S, Popescu M and Alexe B. 2019b. Detecting abnormal events in video using narrowed normality clusters//Proceedings of 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE: 1951-1960 [DOI: 10.1109/WACV.2019.00212]

Ji G L, Qi X S and Wang J Q. 2024. Review of deep learning-based video anomaly detection. Pattern Recognition and Artificial Intelligence, 37(2): 128-143 (吉根林, 戚小莎, 王嘉琦. 2024. 基于深度学习的视频异常检测研究综述. 模式识别与人工智能, 37(2): 128-143) [DOI: 10.16451/j.cnki.issn1003-6059.202402003]

Joo H K, Vo K, Yamazaki K and Le N G. 2023. CLIP-TSA: clip-assisted temporal self-attention for weakly-supervised video anomaly detection//Proceedings of 2023 IEEE International Conference on Image Processing (ICIP). Kuala Lumpur, Malaysia: IEEE: 3230-3234 [DOI: 10.1109/ICIP49359.2023.10222289]

Karim H, Doshi K and Yilmaz Y. 2024. Real-time weakly supervised video anomaly detection//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 6834-6842 [DOI: 10.1109/WACV57701.2024.00670]

Kim J, Yoon S, Choi T and Sull S. 2023. Unsupervised video anomaly detection based on similarity with predefined text descriptions. Sensors, 23(14): #6256 [DOI: 10.3390/S23146256]

Kommanduri R and Ghorai M. 2024. DAST-Net: dense visual attention augmented spatio-temporal network for unsupervised video anomaly detection. Neurocomputing, 579: #127444 [DOI: 10.1016/j.neucom.2024.127444]

Koppel M, Schler J and Bonchek-Dokow E. 2007. Measuring differentiability: unmasking pseudonymous authors. The Journal of Machine Learning Research, 8: 1261-1276

Landi F, Snoek C G M and Cucchiara R. 2019. Anomaly locality in video surveillance [EB/OL]. [2024-05-31]. https://arxiv.org/pdf/1901.10364.pdf

Lang J T. 2017. Bottlenecks and future prospects of the "Skynet" project in the public security system. Science and Technology and Innovation, (9): 45-46 (郎江涛. 2017. 公安系统天网工程瓶颈及未来展望. 科技与创新, (9): 45-46) [DOI: 10.15913/j.cnki.kjycx.2017.09.045]

LeCun Y, Bottou L, Bengio Y and Haffner P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11): 2278-2324 [DOI: 10.1109/5.726791]

Lee D G, Suk H I, Park S K and Lee S W. 2015. Motion influence map for unusual human activity detection and localization in crowded scenes. IEEE Transactions on Circuits and Systems for Video Technology, 25(10): 1612-1623 [DOI: 10.1109/TCSVT.2015.2395752]

Lee S, Kim H G and Ro Y M. 2018. STAN: spatio-temporal adversarial networks for abnormal event detection//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE: 1323-1327 [DOI: 10.1109/ICASSP.2018.8462388]

Li C Y, Hong D F and Zhang B. 2024. Deep unfolding network for hyperspectral anomaly detection. National Remote Sensing Bulletin, 28(1): 69-77 (李晨玉, 洪丹枫, 张兵. 2024. 深度展开网络的

高光谱异常探测. 遥感学报, 28(1): 69-77[DOI: 10.11834/jrs.20233075]

Li C Y, Zhang B, Hong D F, Zhou J, Vivone G, Li S T and Chanussot J. 2024a. CasFormer: cascaded Transformers for fusion-aware computational hyperspectral imaging. Information Fusion, 108: #102408[DOI: 10.1016/j.inffus.2024.102408]

Li D H, Nie X S, Gong R, Lin X M and Yu H. 2024b. Multi-branch GAN-based abnormal events detection via context learning in surveillance videos. IEEE Transactions on Circuits and Systems for Video Technology, 34(5): 3439-3450[DOI: 10.1109/TCSVT.2023.3325451]

Li J N, Li D X, Savarese S and Hoi S. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: 19730-19742

Li J N, Li D X, Xiong C M and Hoi S. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: JMLR.org: 12888-12900

Li N N, Wu X Y, Xu D, Guo H W and Feng W. 2015a. Spatio-temporal context analysis within video volumes for anomalous-event detection and localization. Neurocomputing, 155: 309-319[DOI: 10.1016/J.NEUCOM.2014.12.064]

Liu H T, Li C Y, Li Y H and Lee Y J. 2024. Improved baselines with visual instruction tuning//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 26286-26296[DOI: 10.1109/CVPR52733.2024.02484]

Liu K and Ma H D. 2019a. Exploring background-bias for anomaly detection in surveillance videos//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM: 1490-1499[DOI: 10.1145/3343031.3350998]

Liu W, Luo W X, Li Z X, Zhao P L and Gao S H. 2019b. Margin learning embedded prediction for video anomaly detection with a few anomalies//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: International Joint Conferences on Artificial Intelligence: 3023-3030[DOI: 10.24963/IJCAI.2019/419]

Liu W, Luo W X, Lian D Z and Gao S H. 2018. Future frame prediction for anomaly detection—a new baseline//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6536-6545[DOI: 10.1109/CVPR.2018.00684]

Liu Y, Liu J, Zhao M Y, Li S and Song L. 2022. Collaborative normality learning framework for weakly supervised video anomaly detection. IEEE Transactions on Circuits and Systems II: Express Briefs, 69(5): 2508-2512[DOI: 10.1109/TCSII.2022.3161061]

Liu Z H, Wu X M, Zheng D, Lin K Y and Zheng W S. 2023. Generating anomalies for video anomaly detection with prompt-based feature mapping//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 24500-24510[DOI: 10.1109/CVPR52729.2023.02347]

Lu C W, Shi J P and Jia J Y. 2013a. Abnormal event detection at 150 FPS in MATLAB//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE: 2720-2727[DOI: 10.1109/ICCV.2013.338]

Lu X G, Tsao Y, Matsuda S and Hori C. 2013b. Speech enhancement based on deep denoising autoencoder//Proceedings of the 14th Annual Conference of the International Speech Communication Association. Lyon, France: ISCA: 436-440[DOI: 10.21437/INTERSPEECH.2013-130]

Lu Y W, Yu F, Reddy M K K and Wang Y. 2020. Few-shot scene-adaptive anomaly detection//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 125-141[DOI: 10.1007/978-3-030-58558-7_8]

Luo W X, Liu W and Gao S H. 2017a. A revisit of sparse coding based anomaly detection in stacked RNN framework//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 341-349[DOI: 10.1109/ICCV.2017.45]

Luo W X, Liu W and Gao S H. 2017b. Remembering history with convolutional LSTM for anomaly detection//Proceedings of 2017 IEEE International Conference on Multimedia and Expo (ICME). Hong Kong, China: IEEE: 439-444[DOI: 10.1109/ICME.2017.8019325]

Lyu H, Chen C, Cui Z, Xu C Y, Li Y and Yang J. 2021. Learning normal dynamics in videos with meta prototype network//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 15420-15429[DOI: 10.1109/CVPR46437.2021.01517]

Mahadevan V, Li W X, Bhalodia V and Vasconcelos N. 2010. Anomaly detection in crowded scenes//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE: 1975-1981[DOI: 10.1109/CVPR.2010.5539872]

Majhi S, Dai R, Kong Q, Garattoni L, Francesca G and Brémond F. 2024. Human-scene network: a novel baseline with self-rectifying loss for weakly supervised video anomaly detection. Computer Vision and Image Understanding, 241: #103955[DOI: 10.1016/j.cviu.2024.103955]

Makhzani A and Frey B. 2015. Winner-take-all autoencoders//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2791-2799

Mantini P, Li Z G and Shah K S. 2021. A day on campus—an anomaly detection dataset for events in a single camera//Proceedings of the 15th Asian Conference on Computer Vision. Kyoto, Japan: Springer: 619-635[DOI: 10.1007/978-3-030-69544-6_37]

Markovitz A, Sharir G, Friedman I, Zelnik-Manor L and Avidan S. 2020. Graph embedded pose clustering for anomaly detection//Pro-

ceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10536-10544 [DOI: 10.1109/CVPR42600.2020.01055]

Mirza M and Osindero S. 2014. Conditional generative adversarial nets [EB/OL]. [2024-05-31]. https://arxiv.org/pdf/1411.1784.pdf

Morais R, Le V, Tran T, Saha B, Mansour M and Venkatesh S. 2019. Learning regularity in skeleton trajectories for anomaly detection in videos//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 11988-11996 [DOI: 10.1109/CVPR.2019.01227]

Nayak R, Pati U C and Das S K. 2021. A comprehensive review on deep learning-based methods for video anomaly detection. Image and Vision Computing, 106: #104078 [DOI: 10.1016/J.IMAVIS.2020.104078]

Nguyen T N and Meunier J. 2019. Anomaly detection in video sequence with appearance-motion correspondence//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 1273-1283 [DOI: 10.1109/ICCV.2019.00136]

Pang G S, Shen C H, Cao L B and van den Hengel A. 2021. Deep learning for anomaly detection: a review. ACM Computing Surveys (CSUR), 54(2): #38 [DOI: 10.1145/3439950]

Park H, Noh J and Ham B. 2020. Learning memory-guided normality for anomaly detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 14360-14369 [DOI: 10.1109/CVPR42600.2020.01438]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Virtual: PMLR: 8748-8763

Ramachandra B and Jones M J. 2020. Street scene: a new dataset and evaluation protocol for video anomaly detection//Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass, USA: IEEE: 2558-2567 [DOI: 10.1109/WACV45572.2020.9093457]

Ramesh A, Dhariwal P, Nichol A, Chu C and Chen M. 2022. Hierarchical text-conditional image generation with CLIP latents [EB/OL]. [2024-05-31]. https://arxiv.org/pdf/2204.06125.pdf

Ravanbakhsh M, Sangineto E, Nabi M and Sebe N. 2019. Training adversarial discriminators for cross-channel abnormal event detection in crowds//Proceedings of 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE: 1896-1904 [DOI: 10.1109/WACV.2019.00206]

Ren H M, Liu W F, Olsen S I, Escalera S and Moeslund T B. 2015. Unsupervised behavior-specific dictionary learning for abnormal event detection//Proceedings of 2015 British Machine Vision Conference 2015. Swansea, UK: BMVA: 28.1-28.13 [DOI: 10.5244/C.29.28]

Ren J, Xia F, Liu Y M and Lee L. 2021. Deep video anomaly detection: opportunities and challenges//Proceedings of 2021 International Conference on Data Mining Workshops (ICDMW). Auckland, New Zealand: IEEE: 959-966 [DOI: 10.1109/ICDMW53433.2021.00125]

Rodrigues R, Bhargava N, Velmurugan R and Chaudhuri S. 2020. Multi-timescale trajectory prediction for abnormal human activity detection//Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass, USA: IEEE: 2615-2623 [DOI: 10.1109/WACV45572.2020.9093633]

Ronneberger O, Fischer P and Brox T. 2015. U-net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Sabokrou M, Fathy M, Moayed Z and Klette R. 2017. Fast and accurate detection and localization of abnormal behavior in crowded scenes. Machine Vision and Applications, 28(8): 965-985 [DOI: 10.1007/S00138-017-0869-8]

Sabokrou M, Khalooei M, Fathy M and Adeli E. 2018a. Adversarially learned one-class classifier for novelty detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3379-3388 [DOI: 10.1109/CVPR.2018.00356]

Sabokrou M, Pourreza M, Fayyaz M, Entezari R, Fathy M, Gall J and Adeli E. 2018b. AVID: adversarial visual irregularity detection//Proceedings of the 14th Asian Conference on Computer Vision. Perth, Australia: Springer: 488-505 [DOI: 10.1007/978-3-030-20876-9_31]

Saharia C, Chan W, Saxena S, Li L, Whang J, Denton E, Ghasemipour S K S, Ayan B K, Mahdavi S S, Gontijo-Lopes R, Salimans T, Ho J, Fleet D J and Norouzi M. 2022. Photorealistic text-to-image diffusion models with deep language understanding//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 36479-36494

Saligrama V and Chen Z. 2012. Video anomaly detection based on local statistical aggregates//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 2112-2119 [DOI: 10.1109/CVPR.2012.6247917]

Shi X J, Chen Z R, Wang H, Yeung D Y, Wong W K and Woo W C. 2015. Convolutional LSTM network: a machine learning approach for precipitation nowcasting//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 802-810

Smeureanu S, Ionescu R T, Popescu M and Alexe B. 2017. Deep appearance features for abnormal behavior detection in video//Proceedings of the 19th International Conference on Image Analysis and Processing. Catania, Italy: Springer: 779-789 [DOI: 10.1007/

978-3-319-68548-9_70]

Sultani W, Chen C and Shah M. 2018. Real-world anomaly detection in surveillance videos//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6479-6488 [DOI: 10.1109/CVPR.2018.00678]

Sun Q R, Liu H and Harada T. 2017. Online growing neural gas for anomaly detection in changing surveillance scenes. Pattern Recognition, 64: 187-201 [DOI: 10.1016/J.PATCOG.2016.09.016]

Sun Y, Wang S H, Li Y K, Feng S K, Tian H, Wu H and Wang H F. 2020. ERNIE 2.0: a continual pre-training framework for language understanding//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 8968-8975 [DOI: 10.1609/AAAI.V34I05.6428]

Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z. 2016. Rethinking the inception architecture for computer vision//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2818-2826 [DOI: 10.1109/CVPR.2016.308]

Tao Y R, Hu Y S and Chen Z Z. 2024. Memory-guided representation matching for unsupervised video anomaly detection. Journal of Visual Communication and Image Representation, 101: #104185 [DOI: 10.1016/j.jvcir.2024.104185]

Tian Y, Pang G S, Chen Y H, Singh R, Verjans J W and Carneiro G. 2021. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 4955-4966 [DOI: 10.1109/ICCV48922.2021.00493]

Tran H T M and Hogg D. 2017. Anomaly detection using a convolutional winner-take-all autoencoder//Proceedings of the 28th British Machine Vision Conference. London, UK: BMVA: 1-12

Mehran R, Oyama A and Shah M. 2009. Abnormal crowd behavior detection using social force model//Proceeding of 2009 IEEE Conference on Computer Vision and Rattern Recognition. Miami, Florida, USA: IEEE: 20-25 [DOI: 10.1109/CVPR.2009.5206641]

Vidya M Q M and Selvakumar S. 2024. An effective framework of human abnormal behaviour recognition and tracking using multiscale dilated assisted residual attention network. Expert Systems with Applications, 247: #123264 [DOI: 10.1016/j.eswa.2024.123264]

Vu H, Nguyen T D, Le T, Luo W and Phung D. 2019. Robust anomaly detection in videos using multilevel representations//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI: 5216-5223 [DOI: 10.1609/AAAI. V33I01. 33015216]

Wan B Y, Fang Y M, Xia X and Mei J J. 2020. Weakly supervised video anomaly detection via center-guided discriminative learning//Proceedings of 2020 IEEE International Conference on Multimedia and Expo (ICME). London, UK: IEEE: 1-6 [DOI: 10.1109/ICME46284.2020.9102722]

Wang L, Tian J W, Zhou S P, Shi H Y and Hua G. 2023a. Memory-augmented appearance-motion network for video anomaly detection. Pattern Recognition, 138: #109335 [DOI: 10.1016/J. PATCOG. 2023.109335]

Wang S Q, Zeng Y J, Liu Q, Zhu C Z, Zhu E and Yin J P. 2018. Detecting abnormality without knowing normality: a two-stage approach for unsupervised video abnormal event detection//Proceedings of the 26th ACM International Conference on Multimedia. Seoul, Korea (South): ACM: 636-644 [DOI: 10.1145/3240508. 3240615]

Wang Y, Liu T Y, Zhou J G and Guan J H. 2023b. Video anomaly detection based on spatio-temporal relationships among objects. Neurocomputing, 532: 141-151 [DOI: 10.1016/J.NEUCOM.2023. 02.027]

Wang Y, Xu J, Zhou J G and Guan J H. 2024a. Video anomaly prediction: problem, dataset and method//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Korea (South): IEEE: 3870-3874 [DOI: 10.1109/ICASSP48485.2024.10448187]

Wang Y, Zhou J G and Guan J H. 2024b. A lightweight video anomaly detection model with weak supervision and adaptive instance selection [EB/OL]. [2024-05-31]. https://arxiv.org/pdf/2310.05330.pdf

Wang Z G and Zhang Y J. 2020. Anomaly detection in surveillance videos: a survey. Journal of Tsinghua University (Science and Technology), 60(6): 518-529 (王志国, 章毓晋. 2020. 监控视频异常检测: 综述. 清华大学学报(自然科学版), 60(6): 518-529) [DOI: 10.16511/j.cnki.qhdxxb.2020.22.008]

Wang Z M, Zou Y X and Zhang Z M. 2020. Cluster attention contrast for video anomaly detection//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 2463-2471 [DOI: 10.1145/3394171.3413529]

Wu P, Liu J, Shi Y J, Sun Y J, Shao F T, Wu Z Y and Yang Z W. 2020. Not only look, but also listen: learning multimodal violence detection under weak supervision//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 322-339 [DOI: 10.1007/978-3-030-58577-8_20]

Wu P, Zhou X R, Pang G S, Sun Y J, Liu J, Wang P and Zhang Y N. 2024a. Open-vocabulary video anomaly detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 18297-18307 [DOI: 10.1109/CVPR52733.2024.01732]

Wu P, Zhou X R, Pang G S, Zhou L R, Yan Q S, Wang P and Zhang Y N. 2024b. VadCLIP: adapting vision-language models for weakly supervised video anomaly detection//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 6074-6082 [DOI: 10.1609/AAAI.V38I6.28423]

Wu P H, Wang W Q, Chang F L, Liu C S and Wang B. 2024c. DSS-net: dynamic self-supervised network for video anomaly detection. IEEE Transactions on Multimedia, 26: 2124-2136 [DOI: 10.1109/TMM.2023.3292596]

Xiang T and Gong S G. 2008. Video behavior profiling for anomaly detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(5): 893-908 [DOI: 10.1109/TPAMI.2007.70731]

Xu D, Yan Y, Ricci E and Sebe N. 2017. Detecting anomalous events in videos by learning deep representations of appearance and motion. Computer Vision and Image Understanding, 156: 117-127 [DOI: 10.1016/j.cviu.2016.10.010]

Yan J W, Yang Y X and Naqvi S M. 2024. Object detection oriented privacy-preserving frame-level video anomaly detection//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul, Korea (South): IEEE: 7640-7644 [DOI: 10.1109/ICASSP48485.2024.10447842]

Yang F, Xiao B and Yu Z W. 2021. Anomaly detection and modeling of surveillance video. Journal of Computer Research and Development, 58(12): 2708-2723 (杨帆, 肖斌, 於志文. 2021. 监控视频的异常检测与建模综述. 计算机研究与发展, 58(12): 2708-2723) [DOI: 10.7544/issn1000-1239.2021.20200638]

Ye M C, Peng X J, Gan W H, Wu W and Qiao Y. 2019. AnoPCN: video anomaly detection via deep predictive coding network//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM: 1805-1813 [DOI: 10.1145/3343031.3350899]

Yu G, Wang S Q, Cai Z P, Zhu E, Xu C F, Yin J P and Kloft M. 2020. Cloze test helps: effective video anomaly detection via learning to complete video events//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 583-591 [DOI: 10.1145/3394171.3413973]

Yuan Y, Feng Y C and Lu X Q. 2018. Structured dictionary learning for abnormal event detection in crowded scenes. Pattern Recognition, 73: 99-110 [DOI: 10.1016/J.PATCOG.2017.08.001]

Yuan Z A, Zhou X Y, Liu X P, Lu D W, Deng B and Ma Y X. 2021. Human fall detection method using millimeter-wave radar based on RDSNet. Journal of Radars, 10(4): 656-664 (元志安, 周笑宇, 刘心溥, 卢大威, 邓彬, 马燕新. 2021. 基于RDSNet的毫米波雷达人体跌倒检测方法. 雷达学报, 10(4): 656-664) [DOI: 10.12000/JR21015]

Zaheer M Z, Lee J H, Astrid M and Lee S I. 2020. Old is gold: redefining the adversarially learned one-class classifier training paradigm//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 14171-14181 [DOI: 10.1109/CVPR42600.2020.01419]

Zanella L, Liberatori B, Menapace W, Poiesi F, Wang Y M and Ricci E. 2023. Delving into CLIP latent space for video anomaly recognition [EB/OL]. [2024-05-31]. https://arxiv.org/pdf/2310.02835.pdf

Zanella L, Menapace W, Mancini M, Wang Y M and Ricci E. 2024. Harnessing large language models for training-free video anomaly detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 18527-18536 [DOI: 10.1109/CVPR52733.2024.01753]

Zhang C, Li G R, Qi Y K, Wang S H, Qing L, Huang Q M and Yang M H. 2023. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 16271-16280 [DOI: 10.1109/CVPR52729.2023.01561]

Zhang J G, Qing L and Miao J. 2019. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection//Proceedings of 2019 IEEE International Conference on Image Processing (ICIP). Taipei, China: IEEE: 4030-4034 [DOI: 10.1109/ICIP.2019.8803657]

Zhao Y R, Deng B, Shen C, Liu Y, Lu H T and Hua X S. 2017. Spatio-temporal autoencoder for video anomaly detection//Proceedings of the 25th ACM International Conference on Multimedia. Mountain View, USA: ACM: 1933-1941 [DOI: 10.1145/3123266.3123451]

Zhong J X, Li N N, Kong W J, Liu S, Li T H and Li G. 2019. Graph convolutional label noise cleaner: train a plug-and-play action classifier for anomaly detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1237-1246 [DOI: 10.1109/CVPR.2019.00133]

Zhou Q, Li W Z, Jiang L H, Wang G L, Zhou G Y, Zhang S H and Zhao H. 2024. PAD: a dataset and benchmark for pose-agnostic anomaly detection//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 44558-44571

Zhu X B, Liu J, Wang J Q, Li C S and Lu H Q. 2014. Sparse representation for robust abnormality detection in crowded scenes. Pattern Recognition, 47(5): 1791-1799 [DOI: 10.1016/J.PATCOG.2013.11.018]

Zhu X R, Qian X Y, Shi Y Z, Tao X D and Li Z Y. 2024. Video anomaly detection with long-and-short-term time series correlations. Journal of Image and Graphics, 29(7): 1998-2010 (朱新瑞, 钱小燕, 施俞洲, 陶旭东, 李智昱. 2024. 长短期时间序列关联的视频异常事件检测. 中国图象图形学报, 29(7): 1998-2010) [DOI: 10.11834/jig230406]

Zhu Y and Newsam S. 2019. Motion-aware feature for improved video anomaly detection [EB/OL]. [2024-05-31]. https://arxiv.org/pdf/1907.10211.pdf

作者简介

汪洋,男,博士,主要研究方向为视频异常检测和目标检测。E-mail: tongji_wangyang@tongji.edu.cn

关佶红,通信作者,女,教授,主要研究方向为计算机视觉和人工智能。E-mail: jhguan@tongji.edu.cn

周脚根,男,教授,主要研究方向为目标检测、异常分析和大数据处理。E-mail: zhoujg@hytc.edu.cn

严俊,男,博士研究生,主要研究方向为人工智能安全和异常分析。E-mail: yanjun@tongji.edu.cn