激光与光电子学进展
Laser & Optoelectronics Progress
ISSN 1006-4125,CN 31-1690/TN

《激光与光电子学进展》网络首发论文

题目：基于深度学习的视频异常行为检测综述
作者：彭嘉丽，赵英亮，王黎明
收稿日期：2020-06-19
网络首发日期：2020-09-17
引用格式：彭嘉丽，赵英亮，王黎明．基于深度学习的视频异常行为检测综述[J/OL]．激光与光电子学进展.
https://kns.cnki.net/kcms/detail/31.1690.TN.20200916.1142.002.html

网络首发时间：2020-09-17 10:24:40
网络首发地址：https://kns.cnki.net/kcms/detail/31.1690.TN.20200916.1142.002.html

基于深度学习的视频异常行为检测综述

彭嘉丽，赵英亮*，王黎明

中北大学信息探测与处理山西省重点实验室，山西 太原 030051

摘 要 视频异常行为检测对保障公共安全至关重要，该文对基于深度学习的异常行为检测算法进行分类与总结。首先介绍了异常行为检测的整体流程。然后，根据神经网络训练的方式，从有监督、弱监督和无监督三个方面论述了深度学习在异常行为检测领域的发展与应用，同时分析了不同训练方式的优缺点。最后，简要介绍了常用数据集以及性能评估准则，汇总了不同算法性能，并展望了未来发展方向。

关键词 深度学习；视频异常行为检测；有监督；弱监督；无监督

中图分类号 TP183 **文献标志码** A

An Overview of Video Anomaly Behavior Detection Based on Deep Learning

Peng Jiali, Zhao Yingliang*, Wang Liming

Shanxi Key Laboratory of Signal Capturing and Processing, North University of China, Taiyuan, Shanxi 030051, China

Abstract Video anomaly behavior detection plays an important role in protecting public security. This paper classifies and summarizes the anomaly behavior detection methods based on deep learning. Firstly, the whole process of anomaly behavior detection is introduced. Then, according to the different training mode of convolutional neural networks, the development and application of anomaly behavior detection based on deep learning is discussed from three aspects of supervised, weakly supervised and unsupervised, meanwhile, the advantages and disadvantages of different training methods are analyzed. Finally, the common datasets and performance evaluation criteria are briefly introduced, and the performance of different algorithms are summarized, the future development direction is prospected as well.

Key words Deep Learning; Video Abnormal Behavior Detection; Supervised; Weakly Supervised; Unsupervised

OCIS codes 100.3008; 100.4996; 110.4155

1 引 言

随着监控摄像头在日常生活中的普及，监控视频数据呈爆炸式增长态势。传统的人工异常事件检测不仅耗费大量人力资源，而且由于疲劳工作或侥幸心理，人工检测往往容易漏检异常。国内外研究人员对监控视频异常行为检测算法进行了大量研究，如何实时且精准地检测和定位异常已成为图像处理、机器视觉等领域的研究热点。

基金项目：电子测试技术国防重点实验室基金(6142001180410)、山西省青年科技研究基金(201901D211250)、山西省高等学校优秀成果奖（科学技术）培育项目

E-mail：1069713632@qq.com；*E-mail：zhaoyl18@nuc.edu.cn ；

异常行为检测的难点主要在于：1)正异常行为都没有明确地定义；2)不同的场景下异常的定义有所不同，异常行为检测系统难以泛化；3)行为种类繁多，无法穷举以及 4)异常行为发生概率远低于正常行为，正负样本不均衡，难以学习足够的异常行为特征。

现有的异常行为检测综述中，文[1]将异常检测分为基于高斯混合模型、隐马尔可夫模型、光流法和时空技术等传统方法，分类不恰当，且并未囊括所有传统检测方法。文[2]总结了半监督及无监督的端到端异常行为检测方法，忽略了定义异常的有监督方法，不够全面。文献[3]综述了 2019 年以前的基于深度学习的异常检测算法，但其所综述的文章较少，不够详细。文[4]依据算法的发展阶段、模型类型以及异常判别标准对异常检测技术进行了较全面的 3 级分类，但缺乏对算法的深入分析。

针对上述文献中存在的分类不恰当，综述不全面、不详细、不深入的问题，本文从有监督、弱监督和无监督三个方面对基于深度学习的视频异常行为检测技术进行全面而深入的综述。第二节对异常行为检测的整体流程进行了概述；第三节从神经网络训练的方式对不同算法进行了归纳和比较；第四节介绍了常用数据集以及性能评估准则；第五节总结了全文并对未来研究趋势进行了思考与展望。

2 异常行为检测概述

视频异常行为检测与定位是指利用正异常行为特征表示之间的差异性自动检测及定位异常行为，在安防领域具有重要应用，通常由前景提取与运动目标检测、特征提取、分类及异常检测三部分组成。如图 1 所示，首先预处理输入视频序列，分离冗余的背景，提取运动前景，然后根据行为特征进行分类并检测异常。

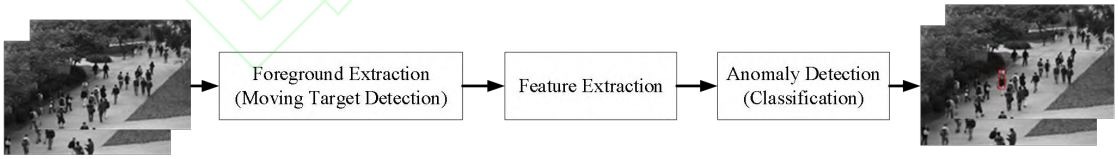

图1 视频异常行为检测流程图

Fig.1 Flow chart of video anomaly detection

传统的运动目标检测方法主要有帧差法、背景减除法以及光流法，随着深度学习在目标检测领域的发展[4]，目标检测网络被广泛用于前景提取及异常检测，如罗凡波等人利用 YOLO v3 目标检测网络检测行人持棍、持枪、持刀以及面部遮挡等异常[5]。

特征提取对于异常检测至关重要，正异常行为特征区分度越高，检测精度就越高。特征

提取的方法主要有(1)传统的基于手工构建的特征，通过人为定义的低级视觉特征表征行为，如梯度方向直方图(HOG)可表征静态图像中的人体形状和轮廓信息，光流能够描述相邻帧之间像素点灰度值的变化，常用于表征运动信息[6]；轨迹(trajectory)用于描述运动目标轨迹；然而手工制作的特征无法表征较复杂的行为，且所提取的特征比较单一，这就导致基于手工构建的特征泛化能力通常比较弱；(2)基于深度学习提取特征，其优势在于能自动地从海量数据集中学习数据本身的分布规律，提取更加鲁棒的高级语义特征，对场景拥挤的情况不敏感，现已渐渐取代了传统方法。如文[8]利用三维卷积神经网络[9]（C3D）提取 HOG 时空特征，提高了对人群行为的表征能力。

传统的异常行为检测在提取手工构建的特征之后还需训练分类器以检测异常。常用的分类器有：1)基于有监督训练的分类器如：支持向量机(SVM)、随机森林[10]、朴素贝叶斯分类器等；2)半监督分类器如：多实例支持向量机(MISVM)、稀疏字典等；3)无监督方法训练的一类支持向量机(OC SVM)、高斯分类器等基于聚类的分类器。而基于深度学习提取特征之后既可使用分类器分类正异常行为，也可直接端到端地用神经网络实现异常检测。

综上，传统的异常行为检测方法存在人工参与多，不够客观且严重依赖场景等问题。基于深度学习的异常行为检测技术其泛化能力强，易于场景迁移，能识别更多的行为类型，已成为近几年的研究热点，基于深度学习端到端地检测异常更是未来发展方向。因此，本文下一节将详细阐述深度学习在异常行为检测领域的研究现状。

3 深度学习在异常行为检测领域的发展及应用

在异常行为检测领域，深度学习由于其出色的特征提取效果以及强大的数据拟合能力，达到了较高的检测精度，成为主流研究方法。按照训练神经网络的数据类型及其标签类型可将基于深度学习的异常行为检测分为有监督、弱监督以及无监督三类，如图 2 所示。

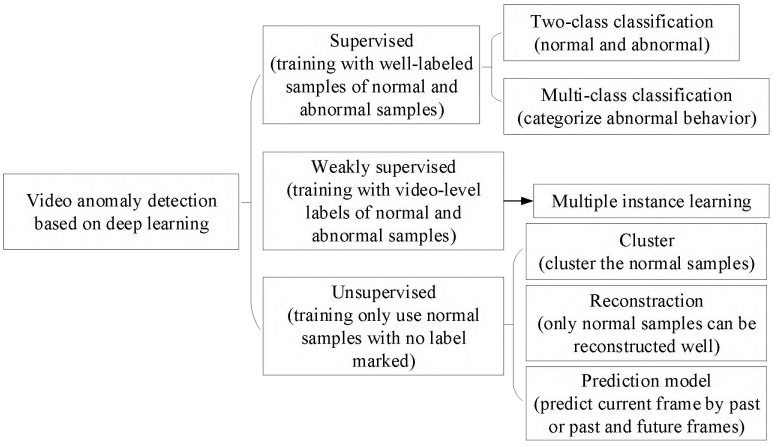

图2 基于深度学习的异常行为检测分类框图

Fig.2 Block diagram of classification on abnormal behavior detection based on deep learning

3.1 有监督异常行为检测

有监督方法定义异常行为，将异常检测视作二分类或多分类问题，即用详细标记的正异常行为样本训练神经网络，提取正异常之间更具区分性的特征。二分类即对正、异常行为进行分类，多分类在检测异常的基础上进一步识别具体的异常行为。

3.1.1 二分类异常行为检测

二分类异常行为检测将正常行为和异常行为视作不同的类别，利用有监督学习将行为归为正常或异常类。由于视频集成了时间和空间信息，因此异常行为检测需要使用能同时提取时空特征的神经网络，如三维卷积神经网络(3D CNN)[11]、循环神经网络和双流网络模型。

与 2D 卷积相比，3D 卷积多了一个时间维度，在行为识别任务中表现优异。文[12]将视频帧划分为大小相同且互不重叠的子区域以实现异常人群定位，然后将子区域输入改进的 C3D 模型提取行为特征并输出正异常分类概率。文[13]将 YOLO 提取的前景人体作为 3D CNN 的输入提取行为的时空特征进而分类正、异常行为。维数的上升使超参数量暴增，计算代价过高，难以满足实时性要求。为了减小计算量，伪三维残差网络[14]和 R(2+1)D[15]网络中提到的将 3×3×3 的 3 维卷积核拆分成 1×3×3 的空间卷积核和 3×1×1 的时间卷积核为异常行为检测提供了新思路，与 3D 卷积相比，(2+1)D 卷积网络在检测精度提高的同时大大减少了运算量。

循环神经网络具有处理时间序列的能力，长短期记忆网络(LSTM)作为循环神经网络的变体在长时序建模的同时解决了梯度弥散和梯度爆炸问题。文[16]将对光照和背景变化不敏感的三通道矫正光流运动历史图输入 3D 卷积核获取短时序特征，结合 LSTM 进一步提取长时序信息。3D-LSTM 的网络结构克服了 3D CNN 只能对短时序运动信息建模的缺点，提高了算法的准确度，在 UMN 数据集上的识别准确率达到 99%，但计算相当复杂，牺牲了实时性。完全卷积神经网络（FCN）可以实现像素级分类与定位，文[17]提出利用 FCN 提取图像特征，再以时间为轴线输入到 LSTM 中提取行为的语义特征，最后通过上采样直接输出异常区域标记，实现异常行为的精准定位。

针对不同动作持续时间有所不同的问题，文[18]采用双流 C3D 网络模型分别处理两种帧长的时空兴趣块，融合不同时间尺度的判别结果能更充分地学习行为的时空特征，同时提出结合非均匀细胞分割和光流的前景提取方法，提取时空兴趣块以便精准定位异常。

3.1.2 多分类异常行为识别

多分类异常行为检测将异常行为具体分类以实现检测并识别异常类别的目的。由于异常事件发生率低，可用于训练的样本较少，文[19]利用迁移学习训练 C3D 网络模型提取行为特征，并对持械、打架斗殴、挟持和抢劫这四类危险行为进行分类，平均识别率达到了 83.2%。LSTM 能够记忆长时间的时序信息，对某些延续时间比较长的行为具有更好的识别效果。文[20]将 YOLO 提取的特征向量输入 LSTM 模型，这种 CNN+LSTM 的结构充分利用了视频的时空融合特征，实现了监狱场景下的实时暴力行为、非法越线和异常奔跑行为识别，每秒可处理 63 帧，识别精度达到 87%。

双流网络旨在将外观信息和运动信息分流提取[21]。文[22]将原图及其帧差图输入双流 VGG 网络，再利用支持向量机实现分类，对拥挤和不拥挤场景下暴力行为的检测精度分别为 93.25%和 95.90%，延迟约 1 秒。在文[22]的基础上，文[23]空间流采用 C3D 网络对连续的视频帧提取时空特征，时间流采用堆叠的光流帧作为输入，由双流分类概率线性加权融合得最终分类结果。3D 卷积以及光流图包含了更多的运动信息，使检测精度得以提升，拥挤和非拥挤场景下的暴力行为识别准确率分别达到了 96%和 99%，但同时带来的复杂计算，造成了算法的延时。

除神经网络外，时空流融合策略对双流模型的性能也有很大影响。在双流模型中，时间流识别精度普遍高于空间流[24]。文[25]研究了在不同层融合时空特征对检测精度的影响。与异常得分融合相比，在全连接层融合时空流特征的识别准确率最高，对打架、抢劫、晕倒以及砸东西 4 类异常识别准确率分别达到了 87.9%、88.7%、88.9%、87.5%。文[26]创造性地将 2D 和 3D 卷积所有可能的融合策略嵌入到一个概率空间中，将融合策略作为一个优化问题，由网络对各种融合策略进行评估。

由于利用了充足的先验信息进行训练，有监督方法的识别和定位精度普遍较高，在现实生活中被广泛使用。但它只能检测预先定义好的异常，只适合在已知所有可能出现的异常种类的特定场景下使用，同时繁琐的人工标注限制了有监督方法的发展，弱监督或无监督异常检测成为最近的研究热点。

3.2 弱监督异常行为检测

弱监督方法仅给出训练样本视频级的正常或者异常标签，即在训练时只知道一段视频中有没有异常事件，而不知道异常事件的具体种类及时间位置，在测试时检测出异常及其持续时间。

弱标签下的异常行为检测是典型的多实例学习问题，文[27]使用 C3D 提取视频片段特征，结合 3 层全连接神经网络预测异常得分，在 UCF-Crime 数据集上检测精度为 75.4%。运动信息对异常行为检测十分关键，针对上述多实例学习排序损失忽略了潜在时间结构的问题，Zhu 等人利用注意力模块加强网络对运动特征的学习，并通过实验验证，引入注意力模块的确对提高异常检测精度有帮助[28]。该算法对 UCF-Crime 数据集中的逮捕、攻击和打架这三类异常行为识别较好，但从整体来看，平均 AUC 值只有 72.1%，并没有超过文[27]，不过模型参数更少，因此检测速度比之高(400+fps VS 300+fps)。文[29]提出利用图卷积网络矫正异常视频中正常片段的噪声标签，校正后的标签用于训练动作分类器检测异常，在 UCF-Crime 数据集上帧级标准 AUC 值达到了 82.12%。视听结合能有效提高模型检测性能，文[30]使用多模态信息作为输入，同时提取视频和音频特征实现在线异常检测，在 UCF-Crime 数据集上帧级 AUC 值达到了 82.44%。

将目标检测网络应用于异常行为检测任务中，使模型从对象的角度分析目标行为，可显著提高检测速度及模型泛化能力。文[31]首先利用目标检测网络提取视频前景目标，然后，使用大尺度光流直方图描述符描述对象行为，最后，利用多实例支持向量机(MISVM)分类正异常行为。该系统在不需要完全标记的条件下获得了较高的识别精度，在人群遮挡严重的情况下仍然鲁棒，UMN 数据集上 AUC 值达到了 98.9%。开集(openset)是指使用训练集中未出现过的异常类型测试模型性能，使网络更加泛化。文[32]设计了一种基于开集的边缘学习嵌入预测(MLEP)框架，结合二维卷积编码器和 Conv-LSTM，能有效地区分正异常行为。以视频级标签训练网络时，在 Avenue 数据集上取得 91.3%的平均 AUC 值，以帧级标签训练时 AUC 值达到 92.8%。

同时使用正常和异常数据训练能提高网络学习能力，使正异常行为特征间距最大化。相比有监督方法，弱标签数据集的制作更简便，因此，弱监督异常行为检测更易操作和泛化，却增大了误检和漏检的概率。随着监控视频数据的暴增，即使是视频级的标签处理也会变得繁琐，因此，学者们转而研究更加智能化的无监督方法。

3.3 无监督异常行为检测

无监督方法无需任何标签信息，基于“非正即异”思想，假设异常行为是罕见且无规律的。通过学习大量正常行为的特征表示，将那些不符合正常特征分布的样本检测为异常，具体可分为基于聚类判别、基于重构判别和基于预测模型三种方法。

3.3.1 基于聚类判别的异常行为检测

基于聚类判别的异常行为检测通过拟合正常样本空间并对正常样本进行聚类，然后将远离正常类聚类中心的样本识别为异常。常用的聚类算法有一类分类器[33]或高斯混合模型等。

一类分类器用于寻找一个囊括正常行为特征的超平面，将不在圈内的样本判断为异常行为。Xu 等人提出的 AMDN[34]将 RGB 图、光流图及其融合图像分别输入三个相同的叠层去噪自编码器提取特征，融合三个一类分类器的异常判断结果检测异常。但这不是端到端的网络，丢失了实时性。文[35]利用稀疏自编码器对 AlexNet 提取的特征进行特征降维，再输入一类分类器检测异常，然而 AlexNet 容易压缩视频中的时间信息，可能会丢失运动相关性。

高斯混合模型(GMM)是多个高斯分布函数的线性组合，理论上 GMM 可以拟合任意类型的分布。Fan 和 Li 等人在双流结构的基础上利用编码器提取测试样本特征，与拟合正常样本特征空间分布的多元混合高斯模型(GMM)比对，不属于任何高斯分量的即是异常行为。此外，针对同一目标在监控视频帧中，由于相对位置的变化引起面积和速度的相对变化的问题，设计了一种多尺度分块结构，有效地解决了由相机视角引起的透视问题[36,37]。

视频帧分块卷积不仅容易分割目标，而且相当耗时。文[38]直接利用完全卷积网络实现异常定位，可以大约 370fps 的速度处理任意尺寸的视频帧。首先将测试样本输入预训练的前两层 AlexNet 提取特征，然后第一个高斯分类器 G1 对特征分类，规则如式(1)，其中，d 指的是正常样本高斯分布和测试样本分布之间的欧式距离。可疑区域再输入一个稀疏自编码器中提取更具区分性的特征，接着第二个高斯分类器对可疑区域分类，级联检测结构增强了模型性能。

$$G_1=\begin{cases}\text{Normal} & \text{if } d\le\beta\\ \text{Suspicious} & \text{if } \beta<d<\alpha\\ \text{Abnormal} & \text{if } d\ge\alpha\end{cases} \tag{1}$$

基于聚类的方法由于缺乏异常行为的先验信息，无法提取正异常行为间有区分性的关键特征。同时，由于正常行为样本类间差异大，需要大量正常行为数据进行训练以拟合尽可能多种类的正常行为，否则容易造成正常的误检。

3.3.2 基于重构判别的异常行为检测

重构，实际上是对输入的帧通过编码提取特征，再将特征解码为重构图像的操作。基于重构的方法假设仅在正常数据上学习的模型不能准确地表示和重构异常，以重构误差作为异常得分以检测异常。

卷积自编码器常用于重构输入图像，文[39]使用卷积自编码器(CAE)对原图、边缘图以及光流图的重构误差作为异常评分。然而二维卷积无法捕获时域信息。文[40]使用 3D 卷积网络对视频片段进行编码，获取低维表示，再使用 3D 反卷积网络进行解码，在解码过程中使用了双路模型，一路重构过去的行为，一路预测下一时刻的行为，以此增强运动特征。文[41,42]基于双流结构，时间流采用 Conv-LSTM 自编码器重构短期光流序列，空间流采用卷积自编码网络重构梯度图，采用贝叶斯融合方法融合双流重构误差值，若融合后的重构误差超过阈值即检测为异常。

同时提取视频外观和运动特征使网络对遮挡问题更加鲁棒，Chong 等人利用端到端的时空自编码器重构输入帧，2D 卷积核提取空间特征，LSTM 获取空间特征的时间演化信息。尽管该模型检测速度达到 140fps，但并不能很好的定位异常行为的终止时间，可能会产生更多的假警报[43]。文[44]采用 3D 编码+LSTM+3D 反卷积解码的结构重构图像，同时在编码器和解码器之间增加跳层连接，使得重构图像更加完整，提高了检测精度。

文[45]设计出由两个变分自编码网络(VAE)级联而成的 S^2-VAE，首先用 SF-VAE 生成正常样本的类 GMM 分布，从而在测试阶段滤除明显的正常样本，然后根据 SC-VAE 的重构误差检测异常，结合聚类判别和重构判别提供了更精确的检测结果。生成对抗网络(GAN)由生成器和鉴别器组成，通过二者的对抗训练，生成器产生的重构误差越来越小，鉴别器判断能力也得到提高[46]。文[47]以 VAE 作为 GAN 的生成器，采用双流结构分别重构原图以及密集光流场，利用重构误差以及对抗损失优化网络，该网络异常检测和定位不够精准并且检测结果有延时，这可能与密集光流计算量大有关。

U-Net 的跳层连接结构将输入帧和输出帧的共同特征直接跨层传输，在减少参数量的同时提高了网络对运动信息的学习能力。文[48,49]利用跨通道的 U-Net 作为 GAN 的生成器，一路由 RGB 原图生成光流图，另一路由相应的光流图生成 RGB 图，通过生成图与真实图之间的局部差异定位异常区域。GANomaly[50]采用“编码-解码-再编码”的结构，在常规 GAN 网络的基础上再加一个再编码器，利用重构损失、再编码损失以及对抗损失共同优化网络检测异常。

由于神经网络具有强大的学习能力，异常行为也有可能被很好的重构，因此，基于重构判别的方法容易造成异常的漏检。同时，由于正常行为也是无法穷尽的，新出现的正常行为容易被误检。

3.3.3 基于预测模型的异常行为检测

预测模型假设正常行为是有规律且可预测的，而异常行为是不可预测的，通过预测误差即可检测异常行为，具体可分为单向预测和双向预测。单向预测将当前输入帧建模为过去 t 帧的函数，从而达到预测正常视频帧的目的。如 Liu 等使用 U-Net 作为 GAN 的生成器预测下一帧，并通过像素强度损失、梯度损失、光流损失和对抗损失优化网络，测试阶段，根据预测帧与真实帧之间的峰值信噪比(PSNR)作为异常得分检测和定位异常[51]。Li 等人提出了一种具有多尺度信息的时空 U 网络用于预测下一帧，在提取空间信息的 U-Net 中加入建模运动信息的 Conv-LSTM，利用 RGB 差分代替光流损失优化生成器，减少了损失计算时间。同时由于异常目标出现在视频帧的边缘时容易发生漏检，设计了一种新的异常分数计算函数，能更加精准的定位异常的起止时间[52]。

针对单向预测模型未能充分利用时间信息的缺点，文[53]提出一种双向预测框架，即根据目标帧的前 *t* 帧和后 *t* 帧对其进行双重预测，结合两种预测帧的交叉均方误差和预测帧与真实帧的均方误差作为损失函数，以峰值信噪比作为异常得分。同时，提出了抑制预测图噪声的滑动窗口方案，将注意力集中于前景目标，提高了模型鲁棒性。该模型结构简单，参数少，泛化能力强，但只能在异常行为发生后检测异常，无法实时预警。

人类行为固有的不可预测性注定了基于预测模型的异常行为检测只能在理想状态下使用，无法投入实际应用。基于无监督的异常行为检测只利用正常样本训练的方式使网络无法针对异常行为的误检进行优化，并且无监督方法中基于理想情况的假设限制了其实用性。表 1 比较了有监督、弱监督和无监督异常行为检测方法的优缺点，这三类方法在识别精度、实用性和人工参与程度方面依次递减，而在智能化和泛化能力方面依次递增。

表1 有监督、弱监督和无监督异常行为检测优缺点对比

Table 1 Comparison of advantage and disadvantage on supervised, weakly supervised and unsupervised anomaly behavior detection

Methods	advantage	weakness
supervised	relatively less training samples is needed; the most accurate; easy to understand and apply.	accurate label is time-consuming; only predefined anomaly can be detected; hard to generalize.
weakly supervised	relatively accurate with weak label;	split the difference;

	lower false alarm; relatively easy to use.	relatively lower accuracy.
unsupervised	training without any label; only normal data needed; robust and easy to generalize.	higher false positive rate; poor positioning accuracy; unable to classify anomaly behavior.

4 常用数据集与评估准则

4.1 常用数据集

目前视频异常行为检测领域常用的公共数据集主要有UCSD、Avenue、UMN、UCF-Crime数据集，表2汇总了其视频数量、总时长、标签情况以及异常种类。

UCSD数据集共2个子集(Ped1和Ped2)，分别记录了校园内垂直和水平方向人行道上固定视角的监控视频。训练视频中只有正常行为，即在人行道上正常行走，而测试视频包含正常行为和骑自行车、滑滑板、轮椅、开汽车和草地上行走等异常行为，并对异常区域进行了像素级掩码标注。

Avenue数据集场景固定，异常定义为一些奇怪的动作如“奔跑”、“投掷物体”、“游荡”、“异常物体”和“错误方向”等，以帧级标签标记。其训练视频中的少量无标记的异常和测试视频中的相机抖动给检测带来了挑战，另外，训练数据中的正常模式比较简单，出现新的正常行为时易误检。

UMN数据集分别在草坪、室内和广场三种场景下以固定摄像机拍摄，分辨率为320*240。正常行为包括正常的闲逛、人群聚集谈话，而异常定义为人群单方向跑动、四散逃逸。该数据集主要针对群体异常识别，只提供帧级标注。

UCF-Crime数据集由只有视频级标签的1900个未剪辑的长时监控视频组成，涵盖真实场景中的13种危害公共安全的异常事件，包括虐待、逮捕、纵火、攻击、事故、入室盗窃、爆炸、打架、抢劫、枪击、偷盗、入店行窃、破坏公物。同时，由于对异常进行了具体分类，该数据集还可用于行为识别任务。

表2 常用异常检测数据集对比表

Table2 Comparison table of commonly used anomaly detection datasets

Dataset	Videos	Length	Annotation	Anomaly categories
UCSD	98	10min	Pixel-level	Biker, skater, wheelchair, car, walking on the grass
Avenue	37	30min	Frame-level	Run, throw, abnormal object
UMN	5	5min	Frame-level	Group escape
UCF-Crime	1900	128hours	Video-level	Abuse, arrest, arson, assault, accident, accident, burglary, fighting, robbery

4.2 性能评估准则

在异常行为检测领域，通过对异常分数或异常概率取不同阈值绘制的接收者操作特征曲线(ROC)定性地评估和比较算法性能，通常用识别精确度(ACC)或者接收者操作特征曲线(ROC)下的面积 AUC 和等错误率(EER)定量地评价模型性能。ROC 曲线以伪阳性率(FPR)为横坐标，真阳性率(TPR)为纵坐标，伪阳性率指在所有实际为负例的样本中预测为正例的概率,真阳性率即在所有实际为正例的样本中预测为正例的概率。因此 ROC 曲线越接近左上角，EER 值越小，AUC 值越大，模型性能越好。

异常行为检测需要检测异常发生的时间和空间位置，通常从帧级和像素级两个层次评价检测效果。在帧级准则中，只要某帧中有一个像素被检测为异常，则该帧视为异常帧，不考虑对异常区域的定位是否准确。而像素级准则还考虑了空间定位精度，只有当检测到的异常像素覆盖了至少 40%的真实异常标记时，才认为出现异常。

4.3 算法性能对比

表 3 汇总了所综述的异常行为检测方法在 UCSD、Avenue、Subway 和 UMN 数据集上不同评估准则下的 AUC 值，其中最小 EER 值和最大 AUC 值加粗显示。

通过表 3 算法性能对比可以得出以下结论：1)从整体来看，有监督方法异常检测精度最高，其次是弱监督方法。2)基于无监督方法的异常检测研究较多。3)使用 3D 卷积或 LSTM 建模时序信息的网络性能普遍比只使用 2D 卷积的要好，这说明运动信息的提取对行为识别精度的提高尤为重要。4)除 S^2-VAE 外，无监督异常检测在像素级准则下识别精度普遍偏低，无监督的方法对异常的定位能力较差。5)目前没有出现在任何场景下检测精度都最优的算法，基于深度学习的异常行为检测算法仍有很大的提升空间。

表3 UCSD Ped1、Ped2和Avenue数据集上帧级和像素级EER(%)和AUC值(%)对比表

Table3 Comparison table of Frame level and pixel level EER(%) and AUC values (%) on UCSD Ped1、Ped2 and Avenue datasets

Methods	UCSD Ped1				UCSD Ped2				Avenue	
	Frame-level		Pixel-level		Frame-level		Pixel-level		Frame-level	
	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC
DSTCNN [12]	--	**99.74**	--	--	--	**99.94**	--	--	--	--
LDA-Net [13]	--	--	--	--	5.63	97.87	12.91	92.96	--	--
FCN+LSTM [17]	--	--	--	--	--	**--**	**6.6**	**98.2**	--	--
Two-stream C3D [18]	**6.29**	96.73	**9.22**	**95.27**	**5.59**	96.37	11.80	93.51	--	--
MISVM [31]	22	--	--	--	16	**--**	--	**--**	**21**	84.5

MLEP [32]	--	--	--	--	--	--	--	--	24.8	**92.8**
AMDN [34]	16.0	92.1	40.1	67.2	17.0	90.8	--	--	--	--
OC SVM [35]	--	--	--	--	10.6	93	17.3	88	--	--
GMFC-VAE [36]	11.3	94.9	36.3	71.4	12.6	92.2	19.2	78.2	22.7	83.4
MGFC-AAE [37]	20	85	--	72.6	16	91.6	--	88	22.3	84.2
CAE [39]	--	89.5	--	--	--	54.7	--	--	--	75.4
ST AE [40]	15.3	92.3	--	--	16.7	91.2	--	--	24.4	80.9
LSTM AE [43]	12.5	89.9	--	--	12.0	87.4	--	--	--	80.3
3D-LSTM AE [44]	15.9	90.9	--	--	15.8	93.6	--	--	20.7	81.8
S^2-VAE [45]	14.3	--	--	94.25	11.54	95.77	14.28	90.83	--	87.6
GAN [48]	8	97.4	35	70.3	14	93.5	--	--	--	--
Predict-GAN [51]	--	83.1	--	--	--	95.4	--	--	--	84.9
ST U-Net [52]	22.3	83.82	--	--	8.7	96.56	--	--	--	84.59
Bi-Prediction [53]	--	89.0	--	--	--	96.6	--	--	21.5	87.8

5 总结与展望

本文概述了现有的异常行为检测方法及其常用数据集和评估准则，综述了深度学习在异常行为检测领域的发展及应用，从原理角度对其进行了分类，最后，通过性能对比分析总结了各类别优缺点。在异常检测领域，无监督方法由于其无需人工标记，泛化能力强成为近几年的学术研究热点，然而，在实际应用中出于对检测及定位精度的考虑，有监督方法仍占据更多市场、落地性更强。

除了改进神经网络以外，异常行为检测算法还可从以下方面提高模型性能。对于无监督方法需要训练大量正常样本，而有监督和弱监督中正异常样本不均衡的问题，基于迁移学习方法训练的方法可有效改善网络性能，防止过拟合。目前大部分的异常检测算法基于闭集测试，即测试集中出现的所有行为类别均被训练过，测试集的检测结果只作为调参标准，而基于开集训练的模型泛化能力更强。因此，在测试集中加入训练集中没有过的类别将是新的研究方向。此外，基于深度学习的目标检测网络提取前景、滤除冗余背景信息，注意力机制为有区分性的特征加大权重，都能进一步提高检测精度。

参考文献

[1] Afiq A A, Zakariya M A, Saad A A, et al. A review on classifying abnormal behavior in crowd scene[J]. Journal of Visual Communication and Image Representation, 2019, 58(1047-3203): 285-303.

[2] 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.

[3] Hu Z P, Zhang L, Li S F, et al. Review of abnormal behavior detection and location for intelligent video surveillance systems[J]. Journal of Yanshan University, 2019, 43(01): 1-12.
胡正平,张乐,李淑芳,等.视频监控系统异常目标检测与定位综述[J].燕山大学学报,2019,43(01):1-12.

[4] Wang Z G, Zhang Y J. Anomaly detection in surveillance videos: a survey[J]. Journal of Tsinghua University (Science and Technology), 2020, 60(6): 518-529.
王志国,章毓晋.监控视频异常检测:综述[J].清华大学学报(自然科学版),2020,60(06):518-529.

[5] Duan Z J, Li S B, Hu J J, et al. Object detection of deep learning method and mainstream framework: an overview[J]. Laser & Optoelectronics Progress, 2020, 57(12): 120005.
段仲静,李少波,胡建军,等. 深度学习目标检测方法及其主流框架综述[J]. 激光与光电子学进展, 2020, 57(12): 120005

[6] Luo F B, Wang P, Liang S Y, et al. Crowd abnormal behavior recognition based on deep learning and sparse optical flow[J]. Computer Engineering, 2020, 46(4): 287-293,300.
罗凡波,王平,梁思源,等. 基于深度学习与稀疏光流的人群异常行为识别[J]. 计算机工程, 2020, 46(4): 287-293,300.

[7] Hu X M, Yu J, Deng C Y, et al. Abnormal crowd behavior detection and location based on spatial-temporal cube[J]. Geomatics and Information Science of Wuhan University, 2019, 44(10): 1530-1537.
胡学敏,余进,邓重阳,等.基于时空立方体的人群异常行为检测与定位[J].武汉大学学报(信息科学版),2019,44(10):1530-1537.

[8] Peng Y P, Jiang R Q, Xu L. An algorithm for identifying crowd abnormal behavior based on c3d-grnn model[J]. Measurement & Control Technology, 2020, 39(07): 44-50.
彭月平,蒋镕圻,徐蕾. 基于C3D-GRNN模型的人群异常行为识别算法[J]. 测控技术, 2020, 39(07): 44-50.

[9] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]//International Conference on Computer Vision (ICCV) 2015, December 13-16, 2015, Santiago, Chile. New York: IEEE Computer Society, 2015. 4489-4497.

[10] Yang X X, Li H B, Hu G. An abnormal behavior detection algorithm based on imbalanced deep forest[J]. Journal of China Academy of Electronics and Information Technology, 2019, 14(09): 935-942.
杨欣欣,李慧波,胡罡. 一种基于不平衡类深度森林的异常行为检测算法[J]. 中国电子科学研究院学报, 2019, 14(09): 935-942.

[11] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.

[12] Hu X M, Chen Q, Yang L, et al. Abnormal crowd behavior detection and localization based on deep spatial-temporal convolutional neural network[J]. Application Research of Computers, 2020, 37(03): 891-895.
胡学敏,陈钦,杨丽,等. 基于深度时空卷积神经网络的人群异常行为检测和定位[J]. 计算机应用研究, 2020, 37(03): 891-895.

[13] Gong M G, Zeng H M, Xie Y, et al. Local distinguishability aggrandizing network for human anomaly detection[J]. Neural Networks, 2020, 122(0893-6080): 364-373.

[14] Qiu Z F, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3d residual

networks[C]//IEEE International Conferenceon Computer Vision. October 22-29, 2017, Venice, Italy. New York: IEEE Computer Society, 2017. 5534-5542.

[15] Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. June 18-22, 2018, Salt Lake City. New York: IEEE, 2018. 6450-6459.

[16] Hu X Y, Guan Y P. 3D-LRCN based video abnormal behavior recognition[J]. Journal of Harbin Institute of Technology, 2019, 51(11): 183-193.
胡薰尹,管业鹏. 基于 3D-LCRN 视频异常行为识别方法[J]. 哈尔滨工业大学学报, 2019, 51(11): 183-193.

[17] Wu G L, Guo Z Z, Li L T, et al. Video abnormal detection combine fcn with lstm[J]. Journal of Shanghai Jiaotong University, 2020, 120: 1-8.
武光利,郭振洲,李雷霆,等. 融合FCN和LSTM的视频异常事件检测[J].上海交通大学学报, 2020, 120:1-8.

[18] Zhang L. Anomaly detection and localization in video surveillance by deep neural network[D]. Qinhuangdao: Yanshan University, 2019.
张乐. 深度神经网络视频异常目标检测与定位算法研究[D]. 秦皇岛: 燕山大学, 2019.

[19] Li C Z, Zhang X J, Zhu H T, et al. Research on dangerous behavior identification method based on transfer learning[J]. Science Technology and Engineering, 2019, 19(16): 187-192.
李辰政,张小俊,朱海涛,等. 基于迁移学习的危险行为识别方法研究[J]. 科学技术与工程, 2019, 19(16): 187-192.

[20] Zou Y F. Recognition and research about abnormal behavior of human based on video[D]. Kunming: Yunnan University, 2019.
邹云飞. 基于视频的人体异常行为识别与研究[D]. 昆明: 云南大学, 2019.

[21] Wang Z W, Gao B P. Spatio-temporal fusion convolutional neural network for abnormal behavior recognition[J]. Computer Engineering and Design, 2020, 41(07): 2052-2056.
王泽伟,高丙朋.基于时空融合卷积神经网络的异常行为识别[J].计算机工程与设计,2020,41(07):2052-2056.

[22] Xia Q. Research on crowd abnormal behavior detection in video surveillance[D]. Chengdu: University of Electronic Science and Technology of China, 2019.
夏清. 视频监控中的人群异常行为检测研究[D]. 成都: 电子科技大学, 2019.

[23] Yuan X X. Research on video violence behavior detection algorithm of deep convolutional network[D]. Qinhuangdao: Yanshan University, 2019.
苑鑫鑫. 深度卷积网络视频暴力行为检测算法研究[D]. 秦皇岛: 燕山大学, 2019.

[24] Zhang M. Research on human abnormal behavior detection based on deep learning[D]. Xi'an: Xi'an University of Science and Technology, 2019.
张梦. 基于深度学习的人体异常行为识别研究[D]. 西安: 西安科技大学, 2019.

[25] Gao Y. Fighting behavior detection in surveillance video by two-stream convolutional networks[D]. Xi'an: Xi'an University of Technology, 2018.
高阳. 基于双流卷积神经网络的监控视频中打斗行为识别研究[D]. 西安: 西安理工大学, 2018.

[26] Zhou Y Z, Sun X Y, Luo C, et al. Spatiotemporal fusion in 3d cnns: a probabilistic view[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, June 14-19, 2020, Seattle, Washington, USA.New York: IEEE, 2020. 2004.04981.

[27] Sultani W, Chen C, Shah M, et al. Real-World Anomaly Detection in Surveillance

Videos[C]// IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, California, USA: American Association of Electrical and Electronic Engineers, 2018. 6479-6488.

[28] Zhu Y, Newsam S. Motion-aware feature for improved video anomaly detection[J]. arXiv: Computer Vision and Pattern Recognition, 2019.

[29] Zhong J, Li N, Kong W, et al. Graph convolutional label noise cleaner: train a plug-and-play action classifier for anomaly detection[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 16-20. Long Beach, California, USA: American Association of Electrical and Electronic Engineers, 2019. 1237-1246.

[30] Wu P, Liu J, Shi Y J, et al. Not only look, but also listen: learning multimodal violence detection under weak supervision[C]. European Conference on Computer Vision, August 23-28, 2020, Glasgow, United Kingdom.2020.

[31] Hu X, Dai J, Huang Y P, et al. A weakly supervised framework for abnormal behavior detection and localization in crowded scenes[J]. Neurocomputing, 2020, 383(0925-2312): 270-281.

[32] Zhou P P, Ding Q H, Luo H B, et al. Anomaly Detection and Location in Crowded Surveillance Videos[J]. Acta Optica Sinica, 2018, 38(8): 0815007.
周培培,丁庆海,罗海波,等. 视频监控中的人群异常行为检测与定位[J].光学学报, 2018, 38(08): 0815007.

[33] Liu W, Luo W X, Li Z X, et al. Margin learning embedded prediction for video anomaly detection with a few anomalies[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, August 10-16, 2019, Macao, China. California: IJCAI, 2019. 3023-3030.

[34] Xu D, Yan Y, Ricci E, et al. Detecting anomalous events in videos by learning deep representations of appearance and motion[J]. Computer Vision and Image Understanding, 2017, 156(1077-3142): 117-127.

[35] Lei L Y. Video anomaly detection based on deep learning[D]. Hangzhou: Hangzhou Dianzi University, 2018.
雷丽莹. 基于深度学习的视频异常检测[D]. 杭州: 杭州电子科技大学, 2018.

[36] Fan Y X, Wen G J, Li D R, et al. Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder[J]. Computer Vision and Image Understanding, 2020, 195(1077-3142): 102920.

[37] Li N J, Chang F L. Video anomaly detection and localization via multivariate gaussian fully convolution adversarial autoencoder[J]. Neurocomputing, 2019, 369(0925-2312): 92-105.

[38] Sabokrou M, Fayyaz M, Fathy M, et al. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes[J]. Computer Vision and Image Understanding, 2018, 172(1077-3142): 88-97.

[39] Ribeiro M, Lazzaretti A E, Lopes H S, et al. A study of deep convolutional auto-encoders for anomaly detection in videos[J]. Pattern Recognition Letters, 2018, 105(0167-8655): 13-22.

[40] 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, October 23-27, 2017, Mountain View, CA, USA: the University of Augsburg, 2017. 1933-1941.

[41] He D D. Abnormal behavior detections under surveillance video scenes[D]. Wuxi: Jiangnan University, 2018.

何丹丹. 监控视频场景下的异常行为检测研究[D]. 无锡: 江南大学, 2018.

[42] Chen Y, He D D. Spatial-temporal stream anomaly detection based on bayesian fusion[J]. Journal of Electronics & Information Technology, 2019, 41(5): 1137-1144.
陈莹,何丹丹. 基于贝叶斯融合的时空流异常行为检测模型[J]. 电子与信息学报, 2019, 41(05): 1137-1144.

[43] Chong Y S, Tay Y H. Abnormal Event Detection in Videos Using Spatiotemporal Autoencoder[C]//international symposium on neural networks, Jun. 21-23, 2017. Sapporo, Japan: Clark Memorial Student Center, 2017. 189-196.

[44] Yue H C. Abnormal events detection method based on autoencoder[D]. Changchun: Jilin University, 2020.
岳海纯. 基于自动编码器的异常行为检测[D]. 长春: 吉林大学,2020.

[45] Wang T, Qiao M, Lin Z, et al. Generative neural networks for anomaly detection in crowded scenes[J]. IEEE Transactions on Information Forensics & Security, 2019, 14(5): 1390-1399.

[46] Ouyang J, Shi Q W, Wang X X, et al. Pedestrian trajectory prediction based on gan and attention mechanism[J]. Laser & Optoelectronics Progress, 2020, 57(14): 141016.
欧阳俊,史庆伟,王馨心,等. 基于GAN和注意力机制的行人轨迹预测[J]. 激光与光电子学进展, 2020, 57(14): 141016.

[47] Wu H M. A research of unspecified anomaly detection and localization in surveillance videos[D]. Chengdu: University of Electronic Science and Technology of China, 2018.
武慧敏. 视频中的非特定异常事件时空位置检测[D]. 成都: 电子科技大学, 2018.

[48] Ravanbakhsh M, Nabi M, Sangineto E, et al. Abnormal event detection in videos using generative adversarial nets[C]//International Conference on Image Processing. Beijing, China: China National Conventional Center, 2017. 1577-1581.

[49] Ravanbakhsh M, Sangineto E, Nabi M, et al. Training adversarial discriminators for cross-channel abnormal event detection in crowds[C]//Proceedings of 2019 IEEE Winter Conference on Applications of Computer Vision. Waikoloa Village, USA: IEEE, 2019. 1896-1904.

[50] Akcay S, Atapourabarghouei A, Breckon T P, et al. GANomaly: semi-supervised anomaly detection via adversarial training[C]//ACCV 2018: Asian Conference on Computer Vision, Dec. 2-6, 2018. Perth, Australia: Asian Computer Vision Alliance, 2018. 622-637.

[51] Liu W, Luo W, Lian D, et al. Future frame prediction for anomaly detection - a new baseline[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 6536-6545.

[52] Li Y, Cai Y, Liu J, et al. Spatio-temporal unity networking for video anomaly detection[J]. IEEE Access, 2019: 172425-172432.

[53] Chen D Y, Wang P T, Yue L Y, et al. Anomaly detection in surveillance video based on bidirectional prediction[J]. Image and Vision Computing,2020,98(0262-8856):103915.

网络首发：

标题：基于深度学习的视频异常行为检测综述
作者：彭嘉丽,赵英亮,王黎明
收稿日期：2020-06-19
录用日期：2020-08-31
DOI：10.3788/lop58.061014

引用格式：
彭嘉丽,赵英亮,王黎明. 基于深度学习的视频异常行为检测综述[J].激光与光电子学进展,2021,58(06):061014.
