doi:10.3969/j.issn.1001-4616.2024.02.014

基于前景对象检测和回归的视频异常检测方法

肖　剑[1],刘天元[2],吴祥[1],吉根林[1]

(1.南京师范大学计算机与电子信息学院/人工智能学院,江苏 南京 210023)
(2.香港理工大学工业及系统工程学系,香港 999077)

[**摘要**]　视频异常检测在智能安防领域具有广泛的应用. 基于生成模型的方法以其强大的生成能力受到学术界广泛关注. 然而,这类方法通常涉及较多的参数,且往往依赖于大量的训练数据,这限制了其在实际应用场景中的适用性. 本文提出了一种基于前景对象检测和回归的视频异常检测方法(FODR-VAD). 首先,利用目标检测器检测前景对象并构建以对象为中心的时空立方体. 其次,采用随机乱序的方法构造伪异常数据. 最后,将单分类视频异常检测问题转换为回归任务,在有监督学习范式下优化特征表示. 在模型训练参数量小于 1 M,使用不到一半训练集的前提下,所提出的方法在 UCSD Ped2、CUHK Avenue 和 ShanghaiTech 数据集上的 Micro-AUC 分别是 99.09%、88.16% 和 78.47%. 结果表明,所提出方法在保证较高异常检测能力的同时,可显著降低对训练数据的需求量.

[**关键词**]　视频异常检测,伪异常,监督学习,回归,时空立方体
[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1001-4616(2024)02-0117-12

Foreground Object Detection and Regression-based Video Anomaly Detection Method

Xiao Jian[1],Liu Tianyuan[2],Wu Xiang[1],Ji Genlin[1]

(1.School of Computer and Electronic Information/Artificial Intelligence,Nanjing Normal University,Nanjing 210023,China)
(2.Department of Industrial and Systems Engineering,The Hong Kong Polytechnic University,Hongkong 999077,China)

Abstract:Video anomaly detection finds wide applications in the field of intelligent security. Methods based on generative models have garnered extensive attention in academia due to their powerful generative capabilities. However, such methods typically involve a large number of parameters and often rely on a vast amount of training data,limiting their applicability in real-world scenarios. This paper proposes a video anomaly detection method based on foreground object detection and regression(FODR-VAD). Firstly,foreground objects are detected using an object detector,and spatiotemporal cubes centered around these objects are constructed. Secondly,pseudo-anomalous data is created using a random shuffling approach. Finally,the video anomaly detection problem is transformed into a regression task,optimizing feature representation under the supervised learning paradigm. With the model parameter count less than 1 million and using less than half of the training set,the proposed method achieves Micro-AUC scores of 99.09%,88.16%,and 78.47% on the UCSD Ped2,CUHK Avenue,and ShanghaiTech datasets,respectively. The results demonstrate that the proposed method significantly reduces the requirement for training data while ensuring high anomaly detection capability.

Key words:video anomaly detection,pseudo anomaly,supervised learning,regression,spatio-temporal cube

视频异常检测是监控视频中的一项关键任务,具有重要的现实研究意义和应用价值. 异常事件通常被定义为在给定环境下不被期望出现的事件. 例如,汽车行驶在高速公路上是一个正常事件,而行驶在人行道上是一个异常事件. 与正常事件相比,异常事件具有发生频率低的特点,这导致其收集难度大,标注

收稿日期:2023-09-26.
基金项目:国家自然科学基金项目(41971343).
通讯作者:吉根林,博士,教授,博士生导师,研究方向:大数据分析与挖掘技术. E-mail:glji@njnu.edu.cn

成本高. 异常数据的难以获得,导致研究者无法直接利用传统的有监督学习范式来解决视频异常检测问题[1]. 考虑到监控视频中存在着丰富的正常事件,因此视频异常检测有这样一种问题设置,单分类视频异常检测:训练集仅包含正常样本数据. 因此,许多文献选择在训练阶段使模型拟合正常样本,在测试阶段将反响强烈的数据判别为异常,从而解决这一问题[2]. 其中,最为流行的为帧重构方法[3-6]与帧预测方法[7-9]. 这些方法利用自编码器或生成对抗网络对数据进行建模,模型的输入和输出都是整个视频帧或经过裁剪的视频帧.

尽管这些方法已取得一定进展,但在实际应用中仍存在两个主要问题. 首先,模型训练参数量较大. 模型参数的增多意味着在训练过程中需要更多的计算资源和时间来处理这些参数,从而增加能源消耗和计算成本. 在实际应用中,计算资源通常是有限的,尤其是在嵌入式设备和移动设备等场景下. 因此,在确保模型性能的同时,尽量减少模型参数的数量显得尤为重要. 其次,模型对标注的训练数据需求量较大. 异常具有环境依赖性,因此每次将模型部署到新的场景时,都需要重新收集正常数据以训练模型. 尽管正常数据相对容易获取,但在大规模数据需求下,收集且正确标注正常数据仍是一项相当困难的任务.

针对上述问题,本文提出了一种基于前景对象检测和回归的视频异常检测方法(foreground object detection and regression-based video anomaly detection method,FODR-VAD). 首先,利用目标检测器检测前景对象并构建时空立方体[10],即以目标对象为中心的经裁剪后的视频帧序列. 然后,通过构造伪异常数据,将单分类视频异常检测问题转换为回归任务,以在有监督学习范式下优化特征表示,降低模型训练难度. 目标检测器的应用减小了模型输入数据的大小,降低了模型输入端的参数量. 将问题转换为回归任务进一步减少了模型输出端的参数量. 时空立方体的构造使模型更加聚焦于前景对象本身,去除了背景冗余信息,增加了训练数据集的冗余性. 结合回归任务的简单性与模型训练参数量的减少,可实现模型对正常训练数据的需求量的降低.

本文方法面临的关键问题是:构造出的伪异常数据能否代替真正的异常数据作为正常数据的负例,继而用于回归任务? 对此,我们进行如下讨论. 正常与异常是一对相对的概念,任何形式的非正常都可归为异常. 因此,非正常的伪异常数据天然可作为正常数据的负例. 在图像异常检测领域,Munawar 等[11]的工作也支持这一观点. 对其实验结果进一步分析可知,训练阶段所使用的伪异常数据,即使与测试集中真正的异常数据有所区别,仍有助于模型训练获得对正常样本的紧凑表示,继而间接帮助模型检测真正的异常.

为了评估 FODR-VAD 的性能,在 UCSD Ped2[12]、CUHK Avenue[13]和 ShanghaiTech[14]数据集上进行了实验. 实验结果表明,在模型训练参数量小于 1 M 的前提下,FODR-VAD 在取得较高异常检测能力的同时,显著降低了模型训练过程中对正常训练数据的需求量.

1 相关工作

由于异常样本数据的缺乏,许多文献在解决单分类视频异常检测问题时,使用生成模型拟合正常样本数据,测试时根据给定样本对模型的响应程度判别异常. 现有两类非常流行的方法:帧重构方法[3-6]与帧预测方法[7-9]. 这些方法假设正常数据可以被很好地重构或预测,而异常数据则不能. 帧重构方法学习一个模型来重构正常的训练数据,并利用重构误差来识别异常. 帧预测方法可看作是一种特殊的帧重构方法,它学习一个以一系列连续帧作为输入、预测输出下一帧的模型,利用预测帧与真实帧的差异程度来区分异常.

Hasan 等[3]提出基于帧重构的方法,利用全卷积自编码器对正常视频帧进行建模,有效地捕获了视频中的表观信息. 但基于重构的方法没能显式地建模运动信息,因此 Liu 等[7]提出基于未来帧预测的方法,结合光流约束,通过生成对抗网络对视频中的表观信息与运动信息进行建模. 但是,由于生成模型强大的泛化能力,这些方法在测试时面对异常输入样本也可以进行很好的重构[4-5,9],继而影响模型对异常的检测效果. 为了解决模型对异常样本的生成问题,一系列基于记忆模块[5,9]、伪异常[4,6]的方法被提出. 此外,Tang 等[15]和 Liu 等[16]将帧重构方法与帧预测方法进行结合,实现了更好的结果. 然而,这些方法都建立在像素级别的生成上,模型训练参数量较大,极其消耗计算资源,以至于无法满足现实应用场景的需要.

不同于上述方法,直接对单一类别的正常样本数据进行建模,本文认为可以通过构造不属于正常样本

数据的伪异常数据,将单分类视频异常检测问题转换为有监督的回归任务,继而利用判别模型予以解决. 回归思想用于视频异常检测,在其他问题设置下已有先例. Sultani 等[17]首次提出弱监督视频异常检测,采用多示例学习的方法来予以解决,将异常检测问题视为回归任务,为每一个实例预测一个异常分数. Landi 等[18]将回归技术用于有监督视频异常检测,通过回归网络直接为每一个输入数据预测一个异常分数. Pang 等[19]利用自训练序数回归解决无监督视频异常检测问题,对视频帧的异常分数进行迭代优化求解. 而本文则通过构造伪异常数据的方式,将回归技术用于单分类视频异常检测问题.

2　本文方法

2.1　算法整体框架

本文提出的基于前景对象检测和回归的视频异常检测方法 FODR-VAD,由 4 个部分组成:时空立方体构造、图像级伪异常构造、特征提取网络 $f(\cdot)$ 和回归网络 $f_c(\cdot)$. 其中,图像级伪异常构造仅用在模型的训练阶段. 算法的整体框架如图 1 所示.

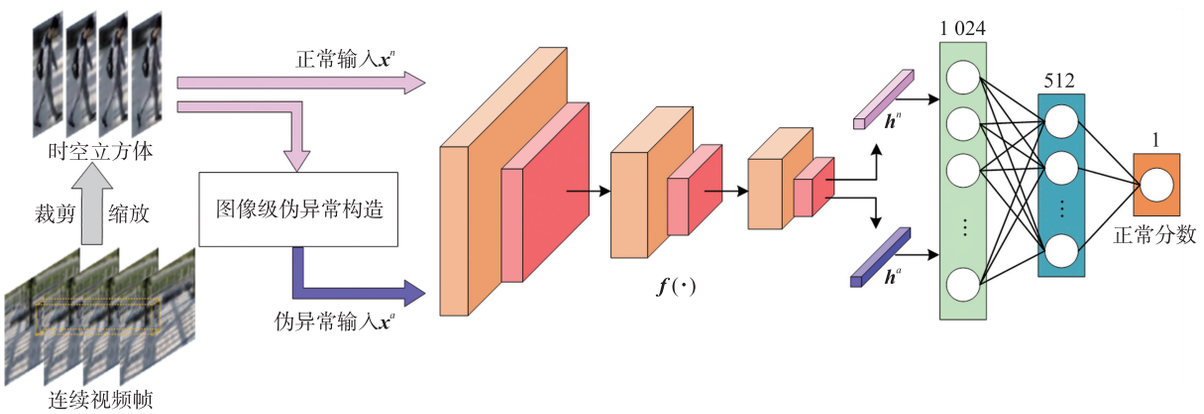

图 1　算法整体框架

Fig. 1　Overall framework of algorithm

进行时空立方体的构造时,首先使用预先训练的目标检测器对训练集和测试集中的每帧进行前景对象检测,得到一系列边界框. 继而对在帧 i 中检测到的每个对象,该帧及其相邻帧 $\{i-t,\cdots,i-1,i,i+1,\cdots,i+t\}$ 按照该对象的边界框进行裁剪得到一系列图像补丁,将这些图像补丁缩放到固定的大小,并在时间维度上进行堆叠,得到以对象为中心的时空立方体. 由测试集得到的时空立方体,包含正常时空立方体与异常时空立方体,构成测试数据集合 X^{test}. 由训练集得到的正常时空立方体,对它们经图像级伪异常构造(见图 2)可得到伪异常时空立方体. 这些正常时空立方体与伪异常时空立方体一起,共同组成训练数据集合 $X^{\text{train}}=\{x_1,x_2,\cdots,x_k\}$. 为训练数据集合 X^{train} 赋予训练标签,得到标签集合 $Y=\{y_1,y_2,\cdots,y_k\}$,其中,当 x_i 为正常时空立方体时,对应的 $y_i=1$,否则 $y_i=0$. X^{train} 与 Y 一起,共同支撑模型进行有监督的回归训练. 一个深度神经网络 $f(\cdot)$,用于对时空立方体进行特征提取. 这些特征随后输入到回归网络 $f_c(\cdot)$ 得到正常分数.

2.2　图像级伪异常构造

由于非正常即可视为异常,所以为得到伪异常数据,有两种思路可以选择. 一种是在图像级别对训练集中的正常样本数据(以时空立方体为数据单位)做任意变换,另一种是在特征级别对由正常样本数据得到的正常特征做随机扰动. 本文实验选择在图像级别构造伪异常数据(进一步讨论见第 4 部分). 由于视频数据带有时空属性,所以从时间和空间两个维度分别对其做任意变换. 受自监督学习中辅助任务的启发[20],选择随机乱序的方式,如图 2 所示. 具体来说,于空间维度上的随机乱序,对每一个时空立方体,首先将其中的每一帧分解为 $n\times n$ 个大小相等的补丁(图中以 $n=2$ 为例),然后对各补丁在时间维度上顺序保持不变的前提下进行随机乱序. 于时间维度上的随机乱序,对时空立方体中的每一帧在时间维度上进行随机排列即可. 特别地,对于静止的对象,由于在时间维度上对其进行乱序没有意义,因而只对其进行空间维度上的乱序.

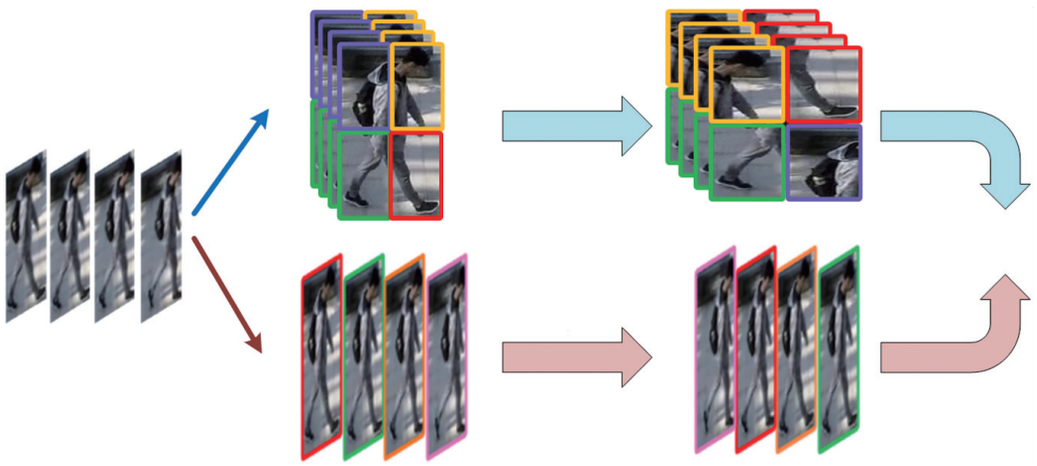

图 2　图像级伪异常构造

Fig. 2　Image-level pseudo anomaly generation

2.3　网络模型

由于回归任务的简单性,所以可以设计训练参数量较少、复杂度较低的网络模型,以增加模型的易训练性和部署性,进而提高模型的实用性和适用范围. FODR-VAD 网络模型由特征提取网络 $f(\cdot)$ 与回归网络 $f_c(\cdot)$ 组成,模型训练参数量大小为 0.979 M. 特征提取网络 $f(\cdot)$ 由 3D 卷积与 2D 卷积两部分组成. 3D 卷积部分共有 6 组 3D 卷积块,每组 3D 卷积块都由 3×3×3 卷积核大小的 3D 卷积层、实例归一化层和 ReLU 激活层组成,且每两组 3D 卷积块后跟一层 3D 池化层. 6 个 3D 卷积层的输出通道数依次为 32,32,64,64,64,64. 2D 卷积部分由输出通道数为 64、卷积核大小为 2×2 的 2D 卷积层,实例归一化层,ReLU 激活层和 1 层 2×2 大小的最大池化层构成. 回归网络 $f_c(\cdot)$ 依次由 1 024×512 的全连接层,ReLU 激活层,512×1 的全连接层和 Sigmoid 激活层构成. 由于 Sigmoid 激活函数可将任意实数映射到(0,1),因此经由回归网络 $f_c(\cdot)$ 得到的正常分数为 0 到 1 之间的数值.

2.4　回归训练

回归作为机器学习中的基本方法,通常用于预测连续型输出变量的值. 其目标是建立输入特征与输出之间的函数关系. 在单分类视频异常检测问题中,我们将异常程度视为输出变量,通过训练数据集来学习这种关系. 形式化定义如下:给定由 K 个时空立方体组成的训练数据集合 $X^{\text{train}}=\{x_1,x_2,\cdots,x_k\}$ 及对应的标签集合 $Y=\{y_1,y_2,\cdots,y_k\}$,目标是学习得到一个正常评分函数 $\phi:X\to(0,1)$. ϕ 由特征提取网络 $f(\cdot)$ 与回归网络 $f_c(\cdot)$ 组成. 若 x_i 为正常时空立方体(即 $y_i=1$),则使 $\phi(x_i)$ 趋近于 1,代表正常程度越高;否则,使 $\phi(x_i)$ 趋近于 0,代表异常程度越高.

首先,从 X^{train} 与 Y 中采样数据标签对 $\{x_i,y_i\}$:

$$\begin{cases} x^n=x_i, & \text{if } y_i=1 \\ x^a=x_i, & \text{if } y_i=0 \end{cases}, \tag{1}$$

式中,利用特征提取网络 $f(\cdot)$ 提取其特征:

$$\begin{cases} h^n=f(x^n) \\ h^a=f(x^a) \end{cases}. \tag{2}$$

最后,将提取的特征输入回归网络 $f_c(\cdot)$,得到对应特征的正常分数. 为使其值趋近于对应标签,可采用交叉熵损失函数 Loss 予以约束:

$$Loss=\frac{1}{N}\sum_i -[y_i\cdot(p_i)+(1-y_i)\cdot(1-p_i)], \tag{3}$$

式中,p_i 表示 $f_c(\cdot)$ 预测输出的正常分数.

注意,虽然此处使用的是分类任务中常用的损失函数,但本文依旧是将异常检测问题视作回归任务,而非分类任务. 这是由于模型最后输出的正常分数为连续值,而非离散值(类别种类).

具体的回归训练算法如算法 1 所示.

算法 1　FODR-VAD 回归训练过程

输入:训练数据集合 X^{train},标签集合 Y,训练轮次 E,批大小 N;网络模型 $f(\cdot)$ 与 $f_c(\cdot)$;
输出:训练好的 $f(\cdot)$ 与 $f_c(\cdot)$ 的模型参数 Θ.

```
for epoch = 1 to E do
    sample a mini-batch {x_i, y_i}_{i=1}^N from X^train, Y
    if y_i = 1
    then h^n = f(x^n) / * 提取正常特征 * /
    else h^a = f(x^a) / * 提取伪异常特征 * /
    end if
    p_i = f_c(h^{n,a}) / * 得到对应特征的正常分数 * /
    compute Loss by Eq. 3 and update f(·), f_c(·) through gradient descent to minimize Loss / * 计算交叉熵损失并更新网络参数 Θ * /
end for
return Θ
```

2.5　正常分数

在测试阶段,对测试数据集合 X^{test} 中的每一个时空立方体,经由 $f(\cdot)$ 提取特征后送入回归网络 $f_c(\cdot)$ 即可得到对应的正常分数. 对于帧的正常分数,选择该帧中所有时空立方体的正常分数的最小值作为该帧的正常分数. 对这些分数首先进行最大最小归一化处理:

$$p_i=\frac{p_i-\min(p_i)}{\max(p_i)-\min(p_i)}, \tag{4}$$

将结果映射到[0,1]之间. 其中 $\min(p_i)$ 和 $\max(p_i)$ 分别是正常分数 p_i 在所有帧上的最小值与最大值. 再使用高斯函数进行平滑处理:

$$f(x)=\frac{1}{\sqrt{2\pi}\cdot\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{5}$$

得到最终的帧级正常分数. 其中 σ 本文选用 20,μ 选用 24. 由 1 减去帧级正常分数即可得相应的异常分数.

3　实验与分析

3.1　实验设置

为了全面的验证本文方法 FODR-VAD 的性能,在三个不同规模的数据集上进行了实验. UCSD Ped2[12](Ped2)是一个小规模的人行道场景数据集,异常事件包括自行车、汽车和滑板. 该数据集中的目标尺寸相对较小. CUHK Avenue[13](Ave)是一个中小规模的校园大道场景数据集,异常事件包括跑步、丢书包和推自行车等. 相对于 Ped2,该数据集中的目标尺寸稍大. ShanghaiTech[14](SHT)是一个大规模的校园多场景数据集,共有 130 个异常事件,包括跑步、闲逛和翻越栅栏等. 在这个数据集中,目标尺寸较大.

选用 PyTorch 搭建网络模型,并使用 Adam 随机梯度下降法来优化模型参数,其中学习率设为 0.000 1,其它参数使用默认值. 训练轮次设为 1,批尺寸设为 64. 选用 YOLOv3[21] 进行目标检测(亦可使用其他目标检测器),针对误检情况,利用置信分数阈值来予以过滤,对 Ped2,Ave 和 SHT,阈值分别设置为 0.5、0.8 和 0.8. 时空立方体的高度和宽度统一设为 64,时间维度上对 Ped2,Ave 和 SHT 分别设 t 为 3、3 和 4. 以概率方式生成伪异常数据,对正常时空立方体,保持正常概率为 0.1. 具体来说,0.1 概率保持不变,0.9 概率进行空间和时间维度上的等概率随机变换.

3.2　实验结果与分析

为了评估模型性能,本文采用帧级 AUC(ROC 曲线下的面积)作为指标,该数值越高,表明模型性能越好. 如 Georgescu 等[22] 所述,现有文献有两种计算帧级 AUC 的方式:1)Micro-AUC,在测试集中将所有视频的帧逐帧拼接,然后计算 AUC;2)Macro-AUC,分别对测试集中的每个视频计算 AUC,然后取平均值. 通常情况下,同一模型在相同数据集上计算 AUC,Macro-AUC 要高于 Micro-AUC. 由于实验结果分析发现帧级

Macro-AUC 对于衡量模型异常检测性能存在不合理性(见 3.2.4 小节),因此本文选择帧级 Micro-AUC 作为评价标准,以公平地比较不同的方法.

3.2.1　训练过程分析

实验表明,本文方法 FODR-VAD 在保持较高的异常检测性能的同时,成功降低了模型训练参数量以及对训练数据的需求. 数据集经目标检测器预处理后,在批大小为 64 的设置下,模型在 Ped2、Ave 和 SHT 上训练一轮分别需要 429、1 490 和 2 265 个批次. 而 FODR-VAD 在其上,仅经过少量批次(分别在第 147、155 和 233 批次达到最高 AUC)的训练就可以收敛,且 AUC 表现出很高的水平(分别为 99.09%、88.16%和 78.47%). FODR-VAD 在三个数据集上的收敛情况以 AUC 曲线图和训练损失曲线图呈现,如图 3 所示. 这表明 FODR-VAD 无需在整个数据集上进行训练即可获得出色的性能表现,这是目前其他方法所无法达到的. 为了进一步比较和探究模型的收敛速度(收敛越快,意味着所需训练数据量越少),计算整理了 FODR-VAD 与部分其他方法所需的训练轮次与模型训练参数量,数据如表 1 所示. 横向对比可知,其他方法在 Ped2 与 Ave 上需要训练 60 或 80 个轮次,而在 SHT 上相对较少. 这可能是因为 SHT 数据集的数据量远大于 Ped2 和 Ave 数据集,其中包含更多冗余数据,使得模型无需经过那么多轮次训练就能够收敛. 然而,与其他方法相比,FODR-VAD 方法竟然在这三个数据集上仅需要部分数据集即可完成训练,即所需训练轮次小于 1. 这与其他方法形成了明显的对比. FODR-VAD 之所以能够快速收敛,有两个主要原因. 首先,使用目标检测器构造时空立方体后,不同场景下的相同行为被转换成了没有场景的相同行为,从而极大程度上增加了数据集中的数据冗余度. 这意味着模型在训练过程中可以从多个类似的数据样本中获取信息,而无需遍历全部数据. 其次,FODR-VAD 相较于其他方法而言,模型训练参数量小(小于 1 M),进行的是

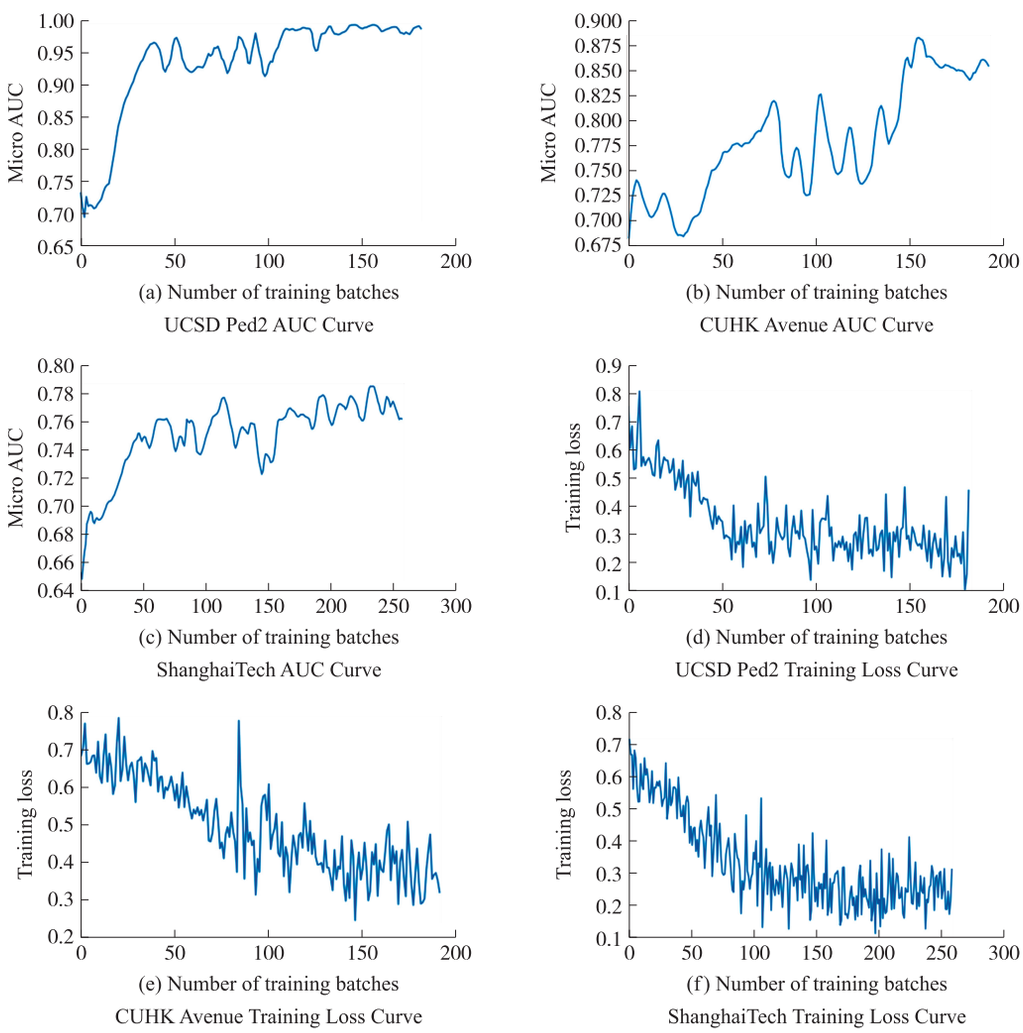

图 3　FODR-VAD 在三个数据集上的训练损失曲线和 AUC 曲线

Fig. 3　Training loss curve and AUC curve of FODR-VAD on three datasets

简单的有监督回归任务,训练难度较低. 相比于其他复杂的方法,FODR-VAD 的模型结构和训练目标相对简单明确. 这种优势使得模型能够更快速地学习数据中的关键特征和模式,从而实现快速收敛、降低对正常标注数据的需求量.

表 1　训练轮次和模型训练参数量的比较

Table 1　Comparison of training epochs and model training parameter counts

Method	Ped2 epochs	Ave epochs	SHT epochs	Params M
DLAN-AC[8]	80	80	20	—
HF^2-VAD[16]	80	80	50	129.459
LLSH[23]	—	60	60	4.718
STEAL-Net[4]	60	60	60	5.980
MNAD[9]	60	60	10	15.651
SLO-VAD[24]	>50	>50	>50	2.727
Ours	**<0.35**	**<0.15**	**<0.15**	**0.979**

3.2.2　测试结果分析

为验证 FODR-VAD 方法的有效性,将其与近年来的其他先进方法在三个数据集上进行了比较,结果如表 2 所示. 最高 AUC 加粗加下划线显示,次高 AUC 加粗显示. FODR-VAD 方法的实验结果的 ROC 曲线图如图 4 所示.

表 2　本文方法 FODR-VAD 与其他方法的比较

Table 2　Comparison between FODR-VAD and other methods

Method	Object Detector	Ped2 AUC/%	Ave AUC/%	SHT AUC/%
Frame-Pred[7]	—	95.40	85.10	72.80
Mem-AE[5]	—	94.10	83.30	71.20
MNAD-Rec[9]	—	90.20	82.80	69.80
MNAD-Pred[9]	—	97.00	88.50	70.50
IntegradAE[15]	—	96.30	85.10	73.00
STEAL-Net[4]	—	98.40	87.10	73.70
DLAN-AC[8]	—	97.60	**89.90**	74.70
AMMC[1]	—	96.60	86.60	73.70
LLSH[23]	—	—	87.40	77.60
HF2-VAD[16]	√	**99.30**	**91.10**	76.20
SLO-VAD[24]	√	97.60	90.90	**78.80**
Ours	√	**99.09**	88.16	**78.47**

根据表 2 的结果可以看出,FODR-VAD 在 SHT 数据集上的性能均优于其他方法. 在 Ave 数据集上,FODR-VAD 的性能接近于先进方法. 在 Ped2 数据集上,FODR-VAD 的性能虽然略逊于最优方法,但仍然具有 99.09%的高性能表现. 这说明了本文方法的有效性. 与本文方法 FODR-VAD 相比,HF^2-VAD 方法与 SLO-VAD 方法也使用了目标检测器构建时空立方体,这三种方法在数据预处理方面存在共同之处. 两方法除了使用目标检测器之外,HF^2-VAD 还在两阶段框架下使用了光流、多层记忆模块和变分自编码器等精心设计的技术. 而 SLO-VAD 则在训练过程中依据任务的难度,以多阶段、多任务的方式训练模型. 相比之下,FODR-VAD 方法显得更加简单高效,直接通过回归预测对数据进行处理,并使用预测的正常分数来识别异常事件. 这表明本文方法 FODR-VAD 具有简单、有效的特点.

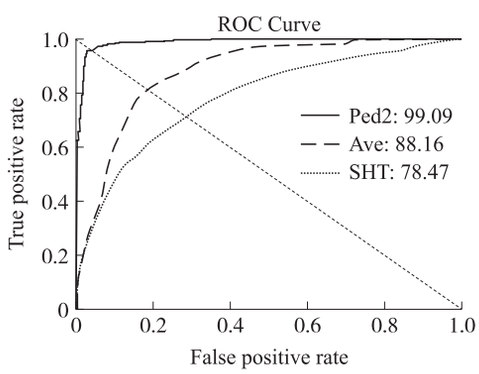

图 4　ROC 曲线图

Fig. 4　ROC curve

3.2.3　泛化性分析

为了评估模型的泛化能力,使用一个数据集进行训练,然后在另一个数据集上进行测试,实验结果如表 3 所示. 括号内的数值表示在跨数据集训练时,相比于原数据集训练模型测试结果的变化情况. 实验结果表明,当 FODR-VAD 方法在 Ped2 和 Ave 数据集上进行测试时,其 AUC 分数较低,显示出较差的泛化能力. 这样的结果,可能是由于数据集之间特征差异过大所致. 尽管通过构造时空立方体的方式去除了部分背景冗杂信息,但由于三个数据集的相机拍摄角度和拍摄距离存在巨大差异,使得相同动作在不同数据集上表现出明显的不同,进而影响了模型的异常检测能力. 然而,需要注意的是,在对 SHT 数据集进行测试时,FODR-VAD 表现出了较高的 AUC,甚至超过了一些在 SHT 数据集上训练的其他方法的性能. 特别是在 Ave 数据集上训练的 FODR-VAD,其 AUC 分数达到了 77.03%,超过了 HF^2-VAD 等方法,显示出一定的泛化能力. 另外,通过比较括号内的性能变化幅度,可以看出模型在目标较小数据集上训练且在目标较大数据集上测试时,性能下降较小,相较之下,在目标较大数据集上训练并在目标较小数据集上测试时,性能下降明显. 这表明模型在处理具有差异特征的数据集时可能存在挑战,但在目标尺寸稍大的情况下也表现出较好的泛化能力.

表 3　不同测试集下的 AUC 分数(%)

Table 3　AUC scores(%) on different test datasets

		Test		
		Ped2	Ave	SHT
Train	Ped2	99.09	79.93(-8.23)	72.43(-6.04)
	Ave	77.11(-21.98)	88.16	77.03(-1.44)
	SHT	79.18(-19.91)	83.03(-5.13)	78.47

3.2.4　可视化分析

由于 SHT 比 Ped2 和 Ave 更加具有挑战性,所以在 SHT 上对 FODR-VAD 的异常检测结果进行可视化分析. 如图 5 所示,在上方的视频帧画面中,红色箭头指向异常所发生区域,蓝色方框表示由目标检测器检测得到的边界框;在下方的异常分数曲线示意图中,横坐标表示视频帧,纵坐标表示异常分数,粉色部分表示异常出现的时间段,蓝色曲线表示经由本文方法得到的异常分数. 图 5(a)、(b)、(c)为在单个视频上具有较低 AUC 值的样例展示,图 5(d)、(e)、(f)为在单个视频上具有较高 AUC 值的样例展示.

对图 5 分析可知:

(1)异常检测性能依赖于目标检测器的表现. 当异常主体在视频画面中占幅较大且不被遮挡时,目标检测器不易出错,这时模型往往表现出较好的异常检测性能;而当异常主体被严重遮挡,在视频画面中占幅较小,或对于目标检测器属于未知类别时,异常主体往往会被漏检,继而导致模型对异常的漏检.

(2)在评估异常检测性能时,不应使用 Macro-AUC 作为评价指标. 如图 5(e)所示,异常发生在整个视频段中,模型预测的异常分数具有很大的波动性,尤其在视频前半部分存在很低的异常分数. 直观上,模型在该视频中未表现出良好的异常检测性能,但计算得到的 AUC 值仍为 1. 由此可见,对单个视频计算 AUC 值无法准确反映模型的异常检测性能,因此不应将 Macro-AUC 作为衡量模型异常检测性能的评价指标.

3.3　消融实验

为了探究空间随机拼图和时间随机乱序两种伪异常数据构造方法对模型性能的影响,通过变化不同的构造方法做了消融实验. 在正常概率 P 值为 0.1 的情况下,消融实验结果如表 4 所示.

观察表 4 数据可知,在构造伪异常数据时,对于 Ped2 和 SHT 数据集,时间维度上的随机乱序比空间维度上的随机拼图更加有效,可以显著提升模型的性能;而对于 Ave 数据集,空间维度上的随机拼图则更加有效. 这可能是由于在 Ped2 和 SHT 数据集中,更多的是与时间因素密切相关的异常,例如滑滑板、骑自行车和追逐打闹等活动. 因此,通过在时间维度上进行随机乱序,可以更好地模拟这些与时间相关的异常情况,从而提高模型的性能. 而在 Ave 数据集中,则存在着相当一部分与空间因素密切相关的异常,例如行人对镜头的遮挡行为等. 因此,在空间维度上进行随机拼图可以更好地模拟这种空间相关的异常情况,进而提升模型的性能. 这表明根据不同数据集中异常类型的特点,选择合适的伪异常数据构造方法可以

显著提升模型的性能. 另外,表 4 也显示了当同时考虑时间维度和空间维度时,可以进一步提升模型的性能表现.

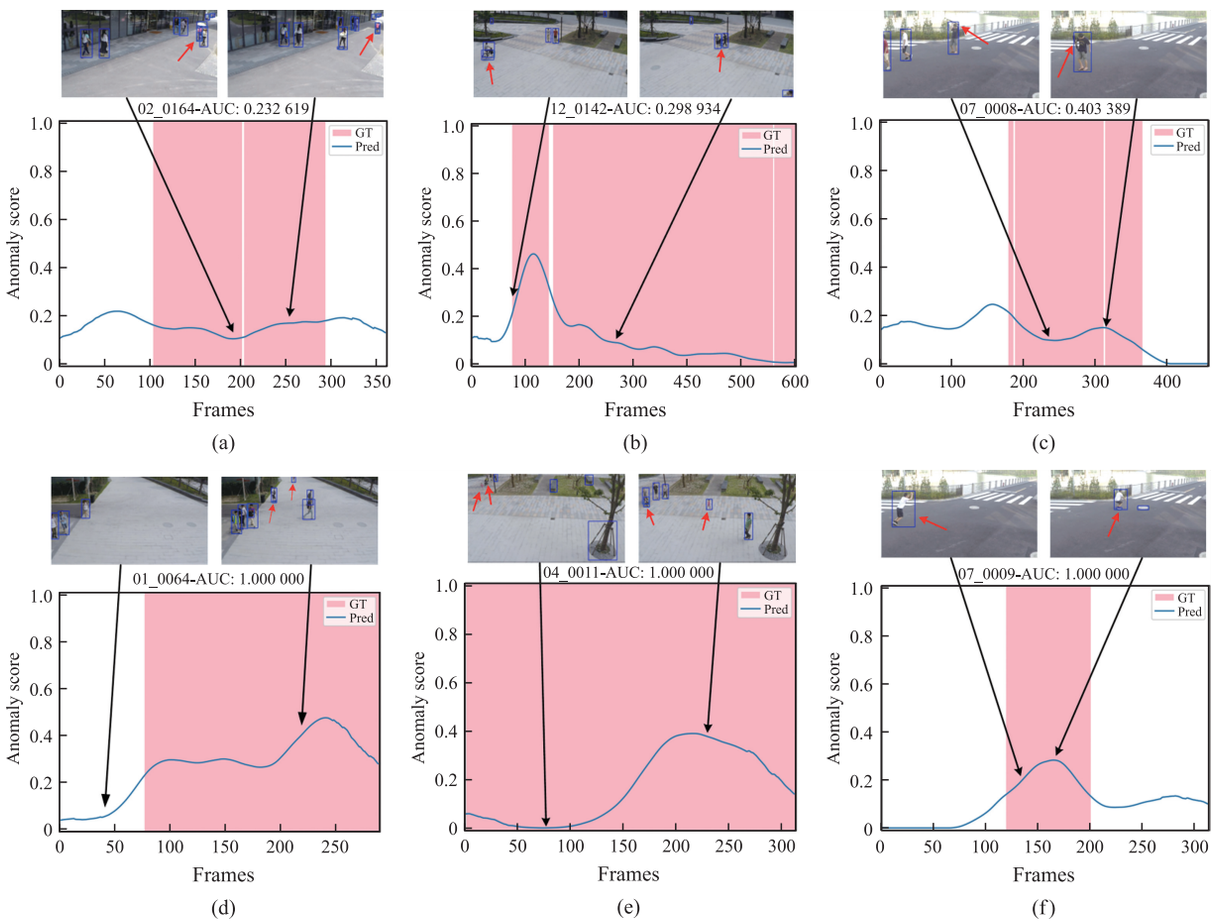

图 5　SHT 上预测结果的可视化

Fig. 5　Visualization of the predicted results on SHT

表 4　不同伪异常构造方法下的 AUC 分数(%)

Table 4　AUC scores(%) under different pseudo anomaly generation methods

	Ped2	Ave	SHT
空间维度	95.01	85.30	70.25
时间维度	96.26	82.89	77.38
时空维度	99.09	88.16	78.47

为了探究在训练过程中正常数据所占比例的不同(正常概率 P 值取值的不同)是否对算法性能存在影响,进行了在时空维度条件下的消融实验. 实验数据在每种概率条件下做 3 次并求平均,结果如表 5 所示,相应的折线图如图 6 所示.

表 5　不同正常概率(P 值)条件下的 AUC 分数(%)

Table 5　AUC scores(%) under different normal probability(P-value) conditions

P	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Ped2	96.83	98.74	98.40	98.87	98.54	98.63	98.67	96.75	96.83	96.34
Ave	85.67	87.58	87.15	86.01	86.40	86.62	83.70	83.78	84.74	83.36
SHT	75.57	78.07	77.79	76.78	77.87	76.04	76.41	76.82	76.59	76.30

根据图 6 所示,从 0.1 开始随着 P 值的增加,模型在 Ped2 和 SHT 数据集上的 AUC 分数没有出现明显波动,整体呈现出趋于平稳的趋势. 在 Ave 数据集上,AUC 分数仅呈轻微下降的趋势. 这表明本文方法对于 P 值的选择不太敏感. 但当 P 值取 0.05 时,在三个数据集上的 AUC 分数都有明显下降. 因此,从节省数

据资源的角度考虑,本文最终选择将 P 值设定为 0.1. 从数据增强的角度来看,当 P 值取 0.1 时,意味着在一次训练过程中,只有十分之一的数据来自真实场景,而余下的九分之一是通过算法生成的数据. 此外,从表 2 中可以看出,相比于其他方法,本文方法的收敛速度非常快. 综合来看,这意味着其他方法的优异性能的获得依赖于大量标注的正常数据,而本文方法只需要少量标注的正常数据就能展现出优秀的异常检测性能. 这表明相比于其他方法,本文方法具有节约资源、适用性强的特点,更能够满足现实应用场景的要求.

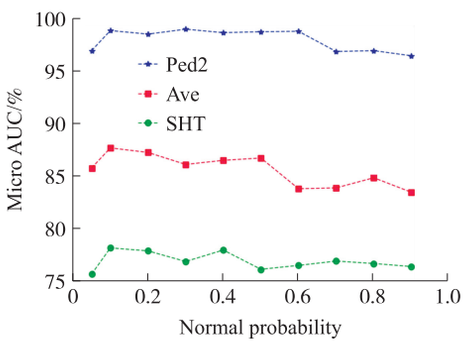

图 6　不同正常概率(*P* 值)条件下的 AUC 分数折线图

Fig. 6　A line graph of AUC scores under different normal probability(*P*-value) conditions

4　讨论及未来工作

未来的工作将就以下三个方面展开讨论:

(1)伪异常数据构造方式的设计. 通过消融实验的结果可以得知,选择合适的伪异常数据构造方式对于显著提升模型异常检测性能十分关键. 目前本文选择在图像级别通过时间维度或空间维度的乱序操作来构造伪异常数据,但可能还存在更好的构造方法. 除了在图像级别设计更巧妙的构造方式外,还可以考虑在特征级别构造伪异常数据. 在图像异常检测领域,Liu 等[25]认为在图像级别构造的伪异常数据与真正的异常之间存在很大差异,这可能会导致无法对正常数据建立紧凑的分布表示,遂提出了在特征级别对正常特征添加高斯噪声来构造伪异常特征的方式,并将其用于图像异常检测且取得了显著的成效. 受此启发,本文进行了初步实验,尝试对提取出的正常特征 h^n 进行加噪处理,但效果并不理想,在 Ped2 数据集上的 AUC 甚至不到 70%. 这说明对于具有时序特点的视频数据,为获得有效的伪异常特征,需要做进一步的探索,不能直接迁移图像数据中的方法.

(2)负例的选择. 一个显然的事实是,通过随机排列方式构造的伪异常数据在异常程度上存在显著差异. 在对比学习领域,Cai 等[26]针对对比实例识别研究了哪些难度范围内的负例对于模型学习到有用的特征最重要. 研究发现,对于下游任务而言,仅仅最困难的 5%的负例是必要且充分的,而最简单的 95%的负例是不必要且不充分的. 此外,最困难的 0.1%的负例不仅是不必要的,而且对模型学习是有害的. 受此启发,并考虑到伪异常数据在异常程度上的差异,本文推测对于构造的伪异常数据,是否也存在类似的情况,即只有一小部分关键数据对于提升模型的异常检测性能至关重要. 如果能够通过某种机制筛选出这些关键数据,并用于模型训练,就可以进一步提升模型的收敛速度和资源利用率.

(3)面向简单场景时模型的选择. 针对不同的应用场景,选择适合的模型很重要. 在消融实验中,当把正常概率 P 值设为 0.05 时,发现模型在面对 Ped2 这样简单的场景时其 AUC 仍然可以稳定地达到 96%以上. 由此可知,对于简单场景,使用简单的模型亦可较好完成异常检测任务,而无需使用那些可以检测复杂异常但却资源耗费严重的模型. 因此,当面向简单场景时,在保证较高性能的前提下,设计训练参数量更小、数据需求量更小的模型是一个值得努力的方向.

5　结论

本文提出了一种基于前景目标检测和回归的视频异常检测方法 FODR-VAD. 通过构造以对象为中心的时空立方体,扩大了训练集冗余度. 通过构造伪异常数据,将单分类视频异常检测问题转换为有监督的

回归任务. 在模型训练参数量小于 1 M 的前提下,使模型在取得较高异常检测能力的同时,显著降低了其在训练过程中对正常训练数据的需求量. 这为设计更符合现实应用的模型提供了重要参考价值. 进一步,通过实验与分析,深入探讨了 FODR-VAD 方法的优劣以及改进的方向.

[参考文献]

[1] CAI R, ZHANG H, LIU W, et al. Appearance-motion memory consistency network for video anomaly detection[C]// Proceedings of the AAAI Conf on Artificial Intelligence. Menlo Park:AAAI,2021,35(2):938-946.

[2] 周航,詹永照,毛启容. 基于时空融合图网络学习的视频异常事件检测[J]. 计算机研究与发展,2021,58(1):48-59.

[3] HASAN M,CHOI J,NEUMANN J,et al. Learning temporal regularity in video sequences[C]//Proceedings of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2016:733-742.

[4] ASTRID M, ZAHEER M Z, LEE S I. Synthetic temporal anomaly guided end-to-end video anomaly detection[C]// Proceedings of the IEEE/CVF Int Conf on Computer Vision. Piscataway,NJ:IEEE,2021:207-214.

[5] GONG D,LIU L,LE V,et al. Memorizing normality to detect anomaly:memory-augmented deep autoencoder for unsupervised anomaly detection[C]//Proceedings of the IEEE/CVF Int Conf on Computer Vision. Piscataway,NJ:IEEE,2019:1705-1714.

[6] ASTRID M,ZAHEER M Z,LEE J Y,et al. Learning not to reconstruct anomalies[DB/OL]. (2021-10-24)[2024-04-08] http://airxiv.org/abs/2110.09742.

[7] LIU W,LUO W,LIAN D,et al. Future frame prediction for anomaly detection-a new baseline[C]//Proceedings of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2018:6536-6545.

[8] YANG Z,WU P,LIU J,et al. Dynamic local aggregation network with adaptive clusterer for anomaly detection[C]//European Conf on Computer Vision. Berlin:Springer,2022:404-421.

[9] PARK H,NOH J,HAM B. Learning memory-guided normality for anomaly detection[C]//Proceedings of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2020:14372-14381.

[10] GEORGESCU M I,BARBALAU A,IONESCU R T,et al. Anomaly detection in video via self-supervised and multi-task learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE. 2021:12742-12752.

[11] MUNAWAR A,VINAYAVEKHIN P,DE MAGISTRIS G. Limiting the reconstruction capability of generative neural network using negative learning[C]//2017 IEEE 27th Int Workshop on Machine Learning for Signal Processing(MLSP). Piscataway,NJ:IEEE,2017:1-6.

[12] MAHADEVAN V,LI W,BHALODIA V,et al. Anomaly detection in crowded scenes[C]//2010 IEEE Computer Society Conf on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2010:1975-1981.

[13] LU C,SHI J,JIA J. Abnormal event detection at 150 fps in matlab[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway,NJ:IEEE. 2013:2720-2727.

[14] LUO W,LIU W,GAO S. A revisit of sparse coding based anomaly detection in stacked rnn framework[C]//Proceedings of the IEEE Int Conf on Computer Vision. Piscataway,NJ:IEEE,2017:341-349.

[15] TANG Y,ZHAO L,ZHANG S,et al. Integrating prediction and reconstruction for anomaly detection[J]. Pattern recognition letters,2020,129(1):123-130.

[16] LIU Z,NIE Y,LONG C,et al. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction[C]//Proceedings of the IEEE/CVF Int Conf on Computer Vision. Piscataway,NJ:IEEE,2021:13588-13597.

[17] SULTANI W,CHEN C,SHAH M. Real-world anomaly detection in surveillance videos[C]//Proceedings of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2018:6479-6488.

[18] LANDI F,SNOEK C G M,CUCCHIARA R. Anomaly locality in video surveillance[J/OL]. arXiv Preprint arXiv:1901.10364,2019.

[19] PANG G, YAN C, SHEN C, et al. Self-trained deep ordinal regression for end-to-end video anomaly detection[C]// Proceedings of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2020:12173-12182.

[20] JING L,TIAN Y. Self-supervised visual feature learning with deep neural networks:a survey[J]. IEEE transactions on pattern analysis and machine intelligence,2020,43(11):4037-4058.

[21] REDMON J,FARHADI A. Yolov3:an incremental improvement[J]. arXiv Preprint arXiv:1804.02767,2018.

[22] GEORGESCU M I,IONESCU R T,KHAN F S,et al. A background-agnostic framework with adversarial training for abnormal event detection in video[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence,Piscataway,NJ:IEEE,2021,44(9):4505-4523.

[23] LU Y,CAO C,ZHANG Y,et al. Learnable locality-sensitive hashing for video anomaly detection[J]. IEEE transactions on circuits and systems for video technology,2022,33(2):963-976.

[24] SHI C,SUN C,WU Y,et al. Video anomaly detection via sequentially learning multiple pretext tasks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway,NJ:IEEE. 2023:10330-10340.

[25] LIU Z,ZHOU Y,XU Y,et al. Simplenet:a simple network for image anomaly detection and localization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE. 2023:20402-20411.

[26] CAI T T,FRANKLE J,SCHWAB D J,et al. Are all negatives created equal in contrastive instance discrimination? [J/OL]. arXiv Preprint arXiv:2010.06682,2020.

[责任编辑:杜忆忱]