DOI: 10.19650/j.cnki.cjsi.J2311235

基于多任务学习的视频异常检测方法*

常兴亚[1],武云鹤[1],陈东岳[1,2],邓诗卓[1,2]
(1. 东北大学信息科学与工程学院 沈阳 110819; 2. 东北大学佛山研究生创新学院 佛山 528311)

摘 要:针对异常事件位于图像前景的某个局部区域,且背景区域对于异常检测存在干扰的问题,提出了一种多任务异常检测双流模型,模型架构包含未来帧预测网络和光流重构网络。首先利用前景检测算法获取自然图像和光流图像的目标区域,再将选取的区域送入到编码-解码网络完成未来帧预测和运动重构,对运动特征和表观特征进行提取,最后,使用深度概率网络给出的概率值作为判断异常的决策,并与重构损失及预测损失相结合来判断视频的异常性。本文针对大型场景的 3 个视频监控数据集(UCSD 行人数据集、Avenue、Shanghai Tech)对本文提出的模型进行了异常性评估,所提出的方法在 3 个数据集上的 AUC 值分别为 97.4%,86.4%,73.4%。与现有工作相比,本文的模型架构简洁且易于训练,异常检测结果更加准确。

关键词: 异常检测;未来帧预测;运动重构;深度概率估计;多任务学习
中图分类号: TP391.41 TH701 **文献标识码**: A **国家标准学科分类代码**: 510.99

Video anomaly detection method based on multi task learning

Chang Xingya[1], Wu Yunhe[1], Chen Dongyue[1,2], Deng Shizhuo[1,2]
(*1. College of Information Science and Engineering, Northeastern University, Shenyang 110819, China*;
2. Foshan Graduate School of Innovation, Northeastern University, Foshan 528311, China)

Abstract: To address the problem of anomalous events occurring in a specific local region of the foreground in an image, with the background region posing interference for anomaly detection, proposes a dual-stream multi-task anomaly detection model. The model architecture consists of a future frame prediction network and an optical flow reconstruction network. Firstly, the optical flow information of the video frame image is extracted by the deep optical flow network, and the foreground detection algorithm is used to obtain the foreground object region of the natural image and the optical flow image. Secondly, the encoding-decoding network is used to complete the future frame prediction and motion reconstruction, and the motion features and apparent features are extracted. Finally, the deep probability network is used to give the probability as the decision to judge the anomaly, and it is combined with the reconstruction loss and the prediction loss to determine the anomalous nature of the images. In this article, the anomalousness of the proposed model is evaluated on three video surveillance datasets (UCSD pedestrian dataset, Avenue, Shanghai Tech) of large scenes, and the proposed method achieves AUC values of 97.4%, 86.4% and 73.4% on the three datasets, respectively. Compared with existing works, the proposed model architecture is simple and easy to train, and the anomaly detection results are more accurate.

Keywords: anomaly detection; future frame prediction; motion reconstruction; deep probability estimation; multi-task learning

0 引 言

视频异常检测是面向监控视频片段中异常目标、行为或事件的自动识别,所谓异常事件,是指有别于场景内正常目标与事件的小概率事件。尽管正常与异常的区分可以视为两个无交集的类别,但异常事件检测在以下 3 个方面有别于传统的有监督二分类问题。1)异常样本在定义上并不具有统一且确定的概率分布,所有有别于正常类的样本都被视为异常;2)考虑到异常样本出现概率极低,因此大多数异常检测模型在训练中只有正常样本,缺少异常样本,因此通常表现为正常类别的单类学习问

收稿日期:2023-03-29 Received Date: 2023-03-29
*基金项目:广东省基础与应用基础研究基金(2021B1515120064)项目资助

题;3)即便存在少量异常样本,考虑到异常样本概率分布的不确定性,其训练集与测试集也不符合独立同分布假设,无法使用传统有监督学习方法来解决。

视频异常检测的核心问题是异常性估计问题,目前主要分为误差法与概率法两种代表性方法。基于误差的视频异常检测模型通常使用帧重构误差或帧预测误差来定量估计视频异常性。文献[1-3]以短时视频帧序列作为自编码器的输入,以重构当前帧或预测未来帧作为网络学习任务,最后使用输入和输出图像之间的像素误差值作为评估视频帧异常性的依据。但此类方法对于异常目标的表观特征或短时运动特征在像素尺度上的异常性十分敏感,因此对于小目标异常检测效果较好,但很难实现更大时空尺度上的异常检测基于概率的异常检测方法通常将异常事件定义为小概率事件,利用正常事件的数据来构建概率模型,并将样本从属于正常类别的后验概率作为异常性度量。与误差法模型以视频帧自身作为监督信号不同,基于概率的异常检测方法通常不具备明确的监督信号,只能通过对概率分布模型的无监督学习实现异常检测。传统的概率估计算法,如单类支持向量机[4]、孤立森林[5]、混合高斯模型[6]等大多只适用于低维数据的概率估计。部分工作尝试将深度特征网络与传统概率估计方法相结合,例如文献[7]使用编码-解码网络提取视频帧特征并利用单类支持向量机来构建正常类别在高维空间中的单类包络曲面。然而支持向量机损失函数优化属于离散规划问题,无法通过梯度下降法实现局部特征和包络模型的端对端学习,因此只能采用特征网络和概率估计算法分离的二阶段法,限制了深度网络在特征学习方面的优势。针对这一问题,文献[8]利用基于“自编码器+高斯混合模型”框架学习视频正常事件高维特征的概率模型。虽然在原理上看该框架可以实现端对端学习,但高斯混合模型优化在高维空间中容易陷入局部最小,且高斯分量数量的自适应选择也是一个棘手的问题,因此该模型在实验结果上并未表现出明显的优势。

视频帧高维数据的特征学习是视频异常检测领域的另一个挑战性问题。由于缺少异常训练样本提供的监督信号,大多数基于重构和基于概率的模型[7,9-10]使用卷积编码解码网络作为基础框架网络,从而将输入的正常视频帧作为网络学习的自监督信号,利用这种方法训练得到的特征网络只对正常样本有良好的表征能力,对于异常样本的表征效果难以预测;另有少数基于概率模型构建的方法[11]采用在大数据集上的预训练的检测或分类网络作为基础网络结构,虽然总体泛化能力较好,但没有针对当前任务的训练样本集进行微调,所以检测精度还有很大的上升空间。

除了异常性评价与特征学习方法外,区域与尺度选择也是视频异常检测无法回避的一个难题。监控视频中的异常目标与事件通常只占视频帧面积的一小部分,且具有不确定的尺度多样性。误差法虽然理论上可以覆盖视频帧的各个局部区域且无需人为构建多尺度检测框架,但其实际检测效果主要集中于像素级的小尺度异常性。概率估计法对于局部区域的选择更加灵活,通常采用需要将视频帧预先划分为一系列均匀分布的子图像块,并将每个子图像块视为一个样本进行概率模型估计,但是此类方法[12]的尺度单一性会降低模型对不同尺度异常目标的检测能力,而如果盲目采用多尺度框架又会导致概率模型学习难度加大,计算量激增。

基于上述分析,本文针对无异常样本参与训练的视频异常检测问题,提出了一种预测,重构及概率法融合的多任务异常检测算法,主要贡献包括如下 3 个方面:1)利用前景检测算法提取视频帧内运动目标的表观和运动的前景区域,减少视频场景中背景因素对异常检测的影响。并训练相应的卷积自编码器提取正常样本的表观特征和运动特征;2)将训练好的卷积自编码器的编码层权重参数共享给深度单类估计网络,通过概率损失与重构损失和预测损失融合,实现深度特征网络与概率推断网络的端对端学习;3)将基于概率的异常性与基于重构误差和预测误差的异常性相结合,构建具有更好的鲁棒性与适应性的综合异常性评分方法。实验结果表明,本文提出的方法在多个具有挑战性的视频监控数据集上取得了令人满意的性能。

1 相关工作

视频异常检测方法可以分类为误差法和概率法。误差法又可以细分为重构法和预测法。重构法是以输入图像与重构图像之间的误差作为网络训练的损失函数实现对正常样本的重构,并最终以重构误差作为当前输入图像异常性评价的依据。文献[1,7]使用自动编码器完成对输入图像或图像序列的重构。然而这种思路存在两方面的隐患。1)在泛化性能方面,过度的拟合可能会导致测试样本中的正常样本也被误判为异常样本;2)像素级重构网络更关注局部表观细节的重构,对于局部细节正常而整体模式异常的目标检测能力较弱。此外,考虑到正常与异常目标运动模式存在显著差异,因此以多帧图像序列作为输入来预测下一帧图像,这是一个较为合理的假设,文献[3]将视频帧序列作为 U-net 网络的输入来预测未来视频帧。还存在工作[13]将单帧图像和光流图像输入到编码解码网络,预测下一帧图像,并结合生成对抗网络来增强细节信息。然而这类方法只是对于短时间链的表观信息有较好的表达,没有充分的融入运动信息辅助异常性的判断。

基于概率估计的异常检测方法一般需要学习正常样本在特征空间中的概率分布,在测试阶段将概率值较低的样本评定为异常样本。Xu 等[7]利用 3 个堆叠式去噪自编码器重构,学习空间特征、时间特征以及时空融合特征,每一类特征都使用单类支持向量机预测每个流的异常概率值。然而该模型采用两阶段方式对特征提取网络和概率推断网络分别进行学习,无法充分发挥深度网络端对端学习在特征工程方面的优势。也有部分工作尝试构建特征表达与概率推理的端对端学习框架。Zong 等[10]也采用了混合高斯模型的假设,但不采用解耦和期望值最大化算法,而是利用单独的参数估计网络学习高斯混合模型参数,以端到端的方式同时对深度自编码器网络和高斯混合模型的参数进行学习;Fan 等[8]采用变分自编码捕获正常样本的特征去拟合高斯分量的参数,并使用样本与高斯混合模型中某一个高斯分量的关联性作为异常检测的依据。相比于重构法只能依据像素级重构图进行异常评价,基于概率的异常性评价指标更加接近异常性的原始定义,特征学习具有更好的适应性、可操控性和泛化性。但相应增加的巨大计算负担决定了不同检测尺度的数量不能太多,通常很难覆盖接近像素级的较小尺度,因此对于图像细节异常的检测能力往往弱于重构法。

2　本文方法

2.1　前景目标区域图生成

视频异常检测有别于图像异常检测的一个特点在于,事件的异常性既可能来源于表观模式也可能来源于运动模式,因此模型的输入需要同时包含视频的表观与运动信息。针对这一特点,本文模型采用了计算机视觉领域广泛使用的双流框架,如图 1 所示,上层的表观流采用 RGB 图像作为输入,下层的运动流采用光流图像作为输入。考虑到传统的光流算法计算量较大,本文采用了一种端对端深度光流网络 FlowNet[14]实现光流图像的生成。FlowNet 采用大数据训练得到,对真实场景视频数据具有较强的泛化能力,且算法实时性高。设时间上连续的训练视频帧数据为 $I=\{I_t\mid t=1,2,\cdots,T\}$,$I_t\in\Re^{(h_0\times w_0\times c_0)}$ 表示宽为 w_0,高为 h_0,通道数为 c_0 的第 t 帧视频彩色图像。利用 FlowNet 将其转化为光流序列 $F=\{f_t\mid t=1,\cdots,T-1\}$,$f_t\in\mathbb{R}^{(h_0\times w_0\times c_0)}$。最后将表观流数据 I 和光流数据 F 作为整个双流异常检测模型的原始输入。

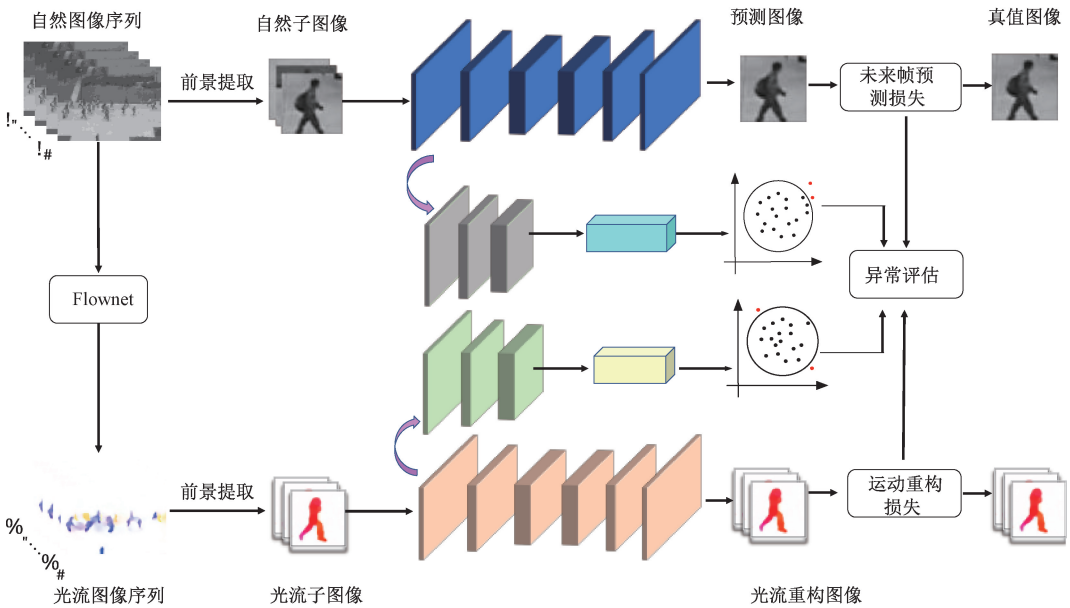


图 1　异常检测网络框架图

Fig. 1　Framework of anomaly detection network

根据前文分析,异常和正常目标通常只占据视频帧面积的一小部分。如果将整帧图像作为特征网络的输入,则背景的正常特征会覆盖和干扰局部的异常特征,模型的检测敏感性将会大大下降。如果将视频帧切分为多个相互交叠的多尺度区域进行分块检测,则需要为每一个局部区域建立训练一个检测模型,不但计算量激增,而且由于每一个局部模型的训练样本有限,会降低检测模型的泛化性能。针对这一问题,本文引入了文献[15]的工作作为视频异常检测的预处理模块,通过目标检测算法和运动检测算法提取连续帧的子图像块,再将该区域对应的图像块作为异常检测的输入。这样既解决了局部目标定位问题,而且还排除了背景区域的干扰。

当前视频帧 I_t 生成了 K 个目标框，记 $r_i, i=1,\cdots,k_t$。利用目标框分别从表观流和光流的当前帧图像序列中选取子图像块并调整到设定大 $h_r \times w_r \times 4$，记为 $x_{t,i}^l \in \Re^{h_r \times w_r \times c_l}$ $l=a$ 或 m 表示表观流，m 表示运动流。

2.2 表观-运动网络

本文模型在双流分支上分别构建了一个轻量化卷积编码解码网络，记作 $S_l=\{E_l, D_l\}$，E_l 表示编码网络，D_l 表示解码网络，$l=a$ 或 m，a 表示表观流，m 表示运动流。采用常见的卷积自编码器架构。如图 2 所示，编码网络 E_l 由 5 个卷积层和最大池化层组成；为完成子图像特征提取，前 3 个卷积层采用 3×3 大小的卷积核；考虑下游单类估计任务的输入数据量不宜过多特性，本文将第 4、5 个卷积核大小设计为 1×1，对每个像素点，在不同的通道上进行线性组合，在保留原有平面结构基础上对通道数进行降维；各卷积层从前到后包含的卷积核个数分别为 32、64、128、64、16；每个卷积层都采用 ReLU 激活函数；最大池化层的局部池化范围为 2×2，步长为 2。卷积层后接全连接层[4 096,128]降维。解码网络 D_l 由反卷积层和上采样层组成，其结构与编码器 E_l 镜像对称。当前输入 $x_{t,i}^l$ 在对应的编码解码网络 S_l 的隐层输出记为 $z_{t,i}^l = E_l(x_{t,i}^l)$，$z_{t,i}^l$ 是 128 维度。当 $l=a$ 时，$\hat{x}_{t+1,i}^a = D_a(z_{t,i}^a) = S_l(x_{t,i}^a)$，$\hat{x}_{t+1,i}^l$ 是预测的下一帧图像，则损失函数为：

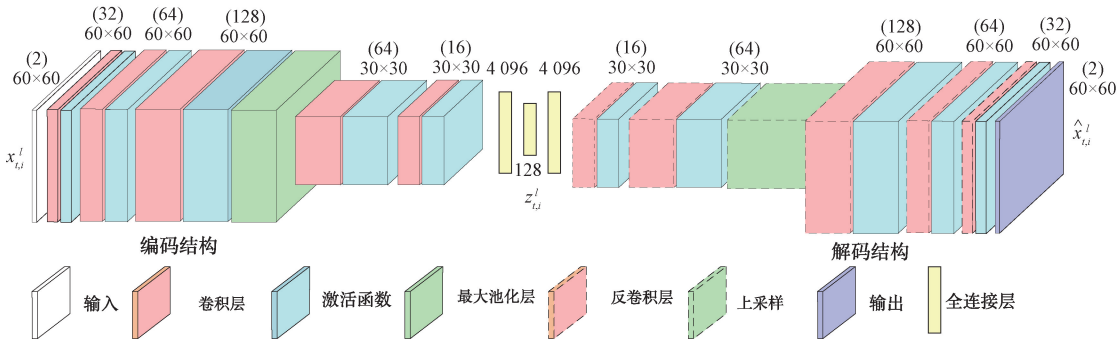


图 2 光流编码解码网络的框架图

Fig. 2 The framework of flow encoder-decoder network

$$\mathcal{L}_a := \sum_{i=1}^{k_t} \|x_{t+1,i}^a - \hat{x}_{t+1,i}^a\|^2 + \frac{\lambda_a}{2}\|\mathcal{W}_a\|^2 \tag{1}$$

当 $l=m$ 时，经由解码网络 D_m 得到的输出可以写为：$\hat{x}_{t,i}^m = D_m(z_{t,i}^m) = S_l(x_{t,i}^m)$，$\hat{x}_{t,i}^m$ 是与输入图像块 x_t^m 具有相同分辨率和相同特征通道数的重构图像块。根据重构任务的定义，则编码解码网络 S_l 的损失函数为：

$$\mathcal{L}_m := \sum_{i=1}^{k_t} \|x_{t,i}^m - \hat{x}_{t,i}^m\|^2 + \frac{\lambda_m}{2}\|\mathcal{W}_m\|^2 \tag{2}$$

其中，$\mathcal{W}_l$ 是编码解码网络 S_l 所有可学习的权值。

$$\mathcal{L}_{con} = \mathcal{L}_m + \mathcal{L}_a \tag{3}$$

基于式(3)的损失函数，采用 Adam 优化器可以对两个编码解码网络分别进行训练。一方面可以为模型提供表观特征 $z_{t,i}^a$ 与运动特征 $z_{t,i}^m$；另一方面，可以为当前图像块 $x_{t,i}^l$ 的异常性评价提供依据。

2.3 深度单类估计网络

支持向量数据描术(support vector data description，SVDD)算法[16]可以视为单类支持向量机的一个变种算法，通过将单类数据映射到高维空间的超球中实现低维空间中的样本分布包络的学习，借助于软间隔的概念，SVDD 允许部分支撑样本侵入到超球范围外，则相应地优化问题可以描述为：

$$\min_{R,\xi} R^2 + \frac{1}{vn}\sum_i \xi_i$$
$$\text{s.t. } \|\phi(x_i) - c\|^2 \leqslant R^2 + \xi_i,\ \xi_i \geqslant 0, \forall_i \tag{4}$$

其中，$\phi(x_i)$ 是样本 x_i 在高维空间 ϕ 上的映射，R 是高维空间 ϕ 上包围所有样本的超球半径，ξ_i 为样本 x_i 的松弛变量，表示样本 x_i 超出超球的程度。整个优化问题的目标是在高维空间 ϕ 中寻找一个尽可能紧致的超球使其能够包裹绝大多数训练样本，超参数 $v \in (0,1]$ 可以控制部分支撑样本超出超球的程度。在异常检测问题中，这个超球可以视为正常和异常样本的分界面。在测试过程中，落在超球范围外的样本将被判为异常。

然而传统 SVDD 算法在应用于监控视频数据过程中遇到了两个巨大的挑战。首先是样本原始维度过高，由于原始 SVDD 算法的高维空间 $\phi(x_i)$ 实际上是通过手动构造或选择核函数 $K=\phi^T\phi$ 来实现的，无法直接通过神经网络对高维数据进行特征学习。其次是 SVDD 算法与 SVM 算法的优化问题本质相同，属于训练样本的离散组合优化问题，在大数据集上学习的计算开销过大，也无法在深度学习框架上实现基于梯度的端对端学习。针对上述问题，Ruff 等[17]提出了 Deep-SVDD。该模型首次使用深度卷积网络来显式地表示原始 SVDD 算法中隐藏在手

动设定的核函数 K 内部的高维映射函数 ϕ，其总体损失函数可以写为：

$$\min_{R,\mathcal{W}}\left[R^2+\sum_i^n\frac{\max\{0,\|\phi(x_i;\mathcal{W})-c\|^2-R^2\}}{vn}+\frac{\lambda}{2}\|\mathcal{W}\|^2\right] \tag{5}$$

其中，映射 $\phi(x_i;\mathcal{W})$ 由一个权值为 $\mathcal{W}$ 的卷积神经网络完成，$\mathcal{W}^2$ 为网络正则化项。式(5)虽然在形式上将深度网络与 SVDD 算法结合在了一起，但从实际训练角度看，SVDD 目标函数的离散优化本质并没有得到解决。而网络权值 $\mathcal{W}$ 作为一组新的待优化变量的加入，不但使得原始 SVDD 的凸问题变成了非凸问题，还会导致特征空间的坍缩(显然当所有样本 x_i 均映射到高维空间 ϕ 中的同一个点 $\boldsymbol{c}$ 时，超球半径达到最小值 0)。为了解决上述问题，Deep-SVDD 算法提出了一个简化版的优化问题模型。结合本文模型，则网络损失函数可以写为：

$$\min_{\mathcal{W}}\frac{1}{n}\sum_{i=1}^{k_t}\|\phi(x_{t,i}^l;\mathcal{W}_l)-c_l\|^2+\frac{\lambda}{2}\|\mathcal{W}_l\|^2 \tag{6}$$

式(6)给出的损失函数以放弃支撑向量概念为代价，将非凸的离散组合优化问题转化为了凸的连续优化问题，此外，简化版的 Deep-SVDD 模型还通过固定中心点 $\boldsymbol{c}_l$ 以及去除特征映射网络 $\mathcal{W}_l$ 等小技巧避免了特征空间 ϕ 的坍缩，$l=a$ 或 m。

本文中设计双流的 Deep-SVDD 的特征映射网络 ϕ_a 和 ϕ_m 结构与表观流编码网络 E_a 和光流编码网络 E_m 是相同的。将经过重构损失函数学习后的表观流编码网络 E_a 和光流编码网络 E_m 分别复制一份，作为双流的 Deep-SVDD 的特征映射网络 ϕ_a 和 ϕ_m 的初始化权值，并采用简化版 Deep-SVDD 算法对两个映射网络进行训练。在特征映射网络收敛后，将训练用的正类样本在特征映射网络的输出记为 $\phi_l(x_{t,i}^l;\mathcal{W}_l)$，则对应的超球中心为：

$$c_l=\sum_{t=1}^{T}\sum_{i=1}^{k_t}\phi_l(x_{t,i}^l;\mathcal{W}_l) \tag{7}$$

原理上看，在高维映射空间中，距离超球中心越远的样本，其从属于正常样本的概率越小，意味着其异常性越大。

2.4　异常性评价方式

如前所述，重构误差，预测误差以及概率的异常性判据各有优劣，三者的融合有望提高模型检测不同类型与尺度的异常性的泛化能力。当前输入视频帧 I_t 在网络分支 $l=m$ 上产生的基于重构误差的异常性评分定义为 SRE_t^m，定义如式(8)。分支 $l=a$ 上产生的基于预测误差的异常性评分定义为 SRE_t^a，定义如式(8)所示。

$$SRE_t^m=\max_i\|x_{t,i}^l-\hat{x}_{t,i}^l\|^2,\quad i=1,\cdots,k_t \tag{8}$$

$$SRE_{t+1}^a=\max_i\|x_{t+1,i}^l-\hat{x}_{t+1,i}^l\|^2,\ i=1,\cdots,k_t \tag{9}$$

SRE 分数的意义是在当前帧内检测到的所有前进该区域块中选择误差平方和最大的一项作为该视频帧的异常性评分。之所以采用最大值而非平均值是因为异常视频帧中大部分区域都是正常的，只有少数目标才是异常的，取平均值会加大正常背景对异常前景的干扰，而去最大值操作更有利于提取局部异常。

与之相对的，视频帧 I_t 在 l 分支的 Deep-SVDD 异常性评分记为 SDV_t^l，计算如下：

$$SDV_t^l=\max_i\|\phi_l(x_{t,i}^l;\mathcal{W}_l)-c_l\|^2,\ i=1,\cdots,k_t \tag{10}$$

SDV_t^l 评分反映了测试样本与正常样本类中心在特征空间的距离，由于 Deep-SVDD 算法能够使原始数据映射到高维空间的超球中，因此距离中心越远就意味着越可能在超球外，异常性概率也越高。综合 2 个分支上的 4 个异常性判据，则视频帧 I_t 最终的综合异常性评分 SA_t 计算如下：

$$SA_t=\eta_a[(1-\beta_a)C_{re}^a(SRE_t^a)+\beta_aC_{dv}^a(SDV_t^a)]+\cdots+\eta_m[(1-\beta_m)C_{re}^m(SRE_t^m)+\beta_mC_{dv}^m(SDV_t^m)] \tag{11}$$

$$C_{re}^l=\frac{SRE_t^l-\min(SRE_t^l)}{\max(SRE_t^l)-\min(SRE_t^l)} \tag{12}$$

$$C_{dv}^l=\frac{1}{1+\exp[-(\|\phi_l(x_{t,i}^l;\mathcal{W}_l)-\boldsymbol{c}_l\|)]} \tag{13}$$

其中，η_a,η_m 分别为表观分支和光流分支的权重，且 $\eta_a+\eta_m=1$，β_a,β_m 分别为两个分支中 Deep-SVDD 异常性判据的权重，$C_{re}^a,C_{dv}^a,C_{re}^m$ 和 C_{dv}^m 为 4 个异常性判据的归一化因子，是为了确保每项评分在(0,1) 范围内。

3　实验验证

3.1　数据集

UCSD 数据集包含两个室外场景：Ped1 和 Ped2，Ped2 图像尺寸为 240×360，帧率为 10 fps。其包含 16 个训练视频和 12 个包含异常事件的测试视频。正常对象为行走的行人，而异常事件为骑自行车、驾驶汽车、玩滑板等。CUHK Avenue 数据集图像尺寸为 360×640，包含 16 个训练视频和 21 个测试视频，共 47 个异常事件，包括投掷物体、徘徊和奔跑等。ShanghaiTech 数据集是一个极具挑战性的异常检测数据集，采集自 13 个场景，包含 330 个训练视频和 107 个测试视频，共 103 个异常事件，包括人行道路上的车辆、扒窃、扭打等，图像尺寸为 480×856。

3.2　模型参数

本文提出的模型基于 Pytorch 框架实现，根据功能将模型分为 3 部分，数据预处理模块、编码解码模块和概率推断模块。光流估计算法采用 FlowNet 网络，并使用在 Flying Chairs 数据集上预训练好的网络及权重参数生成

光流图像。将前景图像尺寸统一为 60×60,特征提取阶段,编码解码网络采用了 Adam 优化算法对权重参数优化,初始 100 个 epoch 的学习率设为 0.001,后 100 个 epoch 的学习率减为 0.000 1,min batch 选取值为 256。本文使用 ROC 曲线下面积 AUC 的值来评估模型,并展示帧级别的检测结果。

3.3 消融实验

为了更好的理解模型架构及参数选择带来的性能提升,本节设计了两个消融实验在数据集 Ped2 上进行测试。1)在结构上:单流网络(单一的表观网络或运动网络)和双流网络结果的对比,实验结果如表 1 所示;2) 异常评估组合方式:测试不同评估方法组合对于异常检测的影响,结果如表 2 所示。

1)如表 1 的实验结果可以看出,在同等条件下(每个网络的评价体系使用重构损失和 Deep SVDD 的结合作为帧级别的异常性得分),在 UCSD Ped2 数据集上,表观单流网络的表现最好的指标为 95.1%,运动单流网络指标为 93.8%,而双流结合网络的指标为 97.4%。从实验结果可以看出视频帧内发生异常时,物体的外观特征或运动模式可能会发生显著变化,因此仅针对表观或运动的单流网络并不能很好地表征视频帧中的异常对象,实验结果表明将表观和运动模式相结合的双流网络异常检测性能最好。

表 1 不同的基础网络框架的异常检测结果

Table 1 Anomaly detection results obtained with different backbones %

方法	Ped 2
表观单流网络	95.1
运动单流网络	93.8
表观+运动双流网络	97.4

2)本文在双流网络基本架构上,测试 4 种不同的异常评价方案,如表 2 所示,分别为表观重构,运动重构,表观特征的单类估计和运动特征的单类估计。不同的组合方式都会对异常事件检测产生影响。如果只使用表观或运动重构损失来作为视频帧的异常性得分,可以达到 93.9%和 92.7%,如果使用运动和表观流的 Deep SVDD 融合得分为 90.4%。如果将 4 种方式的得分以一定权重的融合可以达到 97.4%。

3.4 实验结果

如表 3 所示,本文方法与本领域一些具有代表性的先进算法[8-9,13,18-25]的性能对比,结果证明本文提出的方法在经典视频异常检测数据集上的结果优于绝大多数对比方法。本文方法在 UCSD Ped2 数据上性能指标达到 97.4%,相较于目前代表性工作[24]相比指标提升了 3.3%,与本领域最新模型[25]相比提升了 0.7%。证明本文的方法在 UCSD Ped2 数据集上异常检测达到了本领域先进水平。本文在数据集 Avenue 上性能指标为 86.4%,相比代表性工作[8]的指标提升了 3%,但与本领域最新模型[25]相差 0.7%。但本文采用的深度网络相较于前面的工作更加的简洁且易于训练,且具有检测视频帧内局部异常区域的能力,对于 Avenue 数据集的异常事件检测效果依然保持较强的竞争力。Shanghai Tech 数据集上测试结果为 73.4% 与本领域代表性工作[24]相比指标提升了 2.2%,但与本领域最新模型[25]相差 0.3%,指标基本持平。以上指标证明了我们的方法对于复杂场景人群密集处的检测是有效的。

表 2 不同评价体系组合方式的检测结果

Table 2 Anomaly detection results of different evaluation system combinations %

方法	Ped 2
表观重构	93.9
运动重构损失	92.7
(运动+表观)Deep SVDD	90.4
(运动+表观)重构+(运动+表观)Deep SVDD	97.4

表 3 所提出的方法在 3 个数据集上的帧级别检测结果

Table 3 Frame-level detection results of the proposed method on three data sets %

方法	Ped 2	Avenue	SH-Tech
GMM-VAE[8]	92.2	**83.4**	–
Conv-AE[9]	85.0	80.0	60.9
HDN[13]	84.3	82.8	–
MDT[18]	81.8	–	–
Detection at 150 fps[19]	–	80.9%	–
ConvLSTM-AE[20]	88.1	77.0	–
MPPCA[21]	69.3	–	–
Stacked RNN[22]	91.2	81.7	67.9
Unmasking[23]	82.2	80.6	–
Memory-AD[24]	**94.1**	83.3	**71.2**
STD[25]	**96.7**	**87.1**	**73.7**
本文	**97.4**	**86.4**.	**73.4**

如图 3 所示的矩形区域表示异常区域。从中可以看出本文方法能够在较低分辨率下,对汽车、自行车、异常运动的行人等异常对象进行有效检测,具体检测到的异常对象表现出了形状、纹理和尺度上的表观异常,以及运动和动作等方面的时域异常,这说明本文方法能够对多

种类型的异常特征进行有效的检测,反映了算法的泛化能力与鲁棒性。同时,尽管本文方法的定量性能指标的统计采用了视频帧作为基本评价对象,但从可视化实验结果上看,本文采用的前景检测模块能够实现局部区域的异常检测,因此可以提供目标级异常的检测结果,有助于提升监控视频异常检测系统的检测分辨率与实用性。

(a) 原始图像 (a) Original image　(b) 真值图像 (b) Truth image　(c) 结果图像 (c) Result image

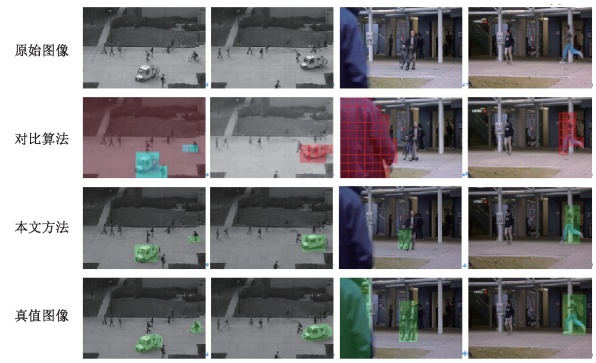
图 3　本文算法在不同数据集上的代表性检测结果

Fig. 3　Typical detection results of the proposed model on different data sets

目前帧级别区域性定位的异常检测工作有[8,13,22-23],本文对比其中前沿工作[8,13]在 UCSDPed2, Avenue 数据集上的检测结果图像。如图 4 所示,本文算法和 Nguyen 等[13]算法都可以检测出汽车和自行车等异常目标,但相比本文算法,上述工作异常定位不够准确。GMM-VAE 算法和本文算法在 UCSD 数据集上检测效果基本一致,在 Avenue 数据上,算法 GMM-VAE 输入为整帧图像,对小区域目标物体检测不敏感,存在正常行人推自行车事件的漏检。本文在目标检测的基础上可以准确的定位训练集里没有出现的异常物体或者错误的运动方向,但目标检测对非全景物体检测效果不佳时,则会错失当前异常帧的判断,相比算法 GMM-VAE,可能会存在半人体出镜时的漏检情况。

3.5　模型参数与算法运行时间分析

本文框架采用 Pytorch 实现。实验是在 Intel(R)Xeon(R)CPU E5-2640 v4 2.40 GHz CPU 上进行的。如表 4 所示,我们的平均运行速度约为 31 fps,其中包含视频帧生成和规则性分数计算。对比目前主流的异常检测方法,包括 Unmasking[23]、AMDN[26]、NCC[27], Frame-Pred[28] 和 Song



(a) 混合深度网络 (a) HDN[12]　(b) 高斯变分自编码器 (b) GMM-VAE[8]　(c) 高斯变分自编码器 (c) GMM-VAE[8]　(d) 高斯变分自编码器 (d) GMM-VAE[8]

图 4　本文算法对比代表性算法的检测结果

Fig. 4　The detection results comparison between the proposed algorithm and other representative algorithms

等[29],其运行速度分别为 0.13、20、24、25 和 35 fps,证明了本文算法在计算速度上具有很强的竞争力。本文方法在浮点运算数(floating point operations, FLOPs)也比较低,为 5.96 G FLOPs,模型参数为 6.1 G,模型的参数量要低于其他算法。上述实验证明我们的模型不管在计算速度上还是模型空间复杂度上都优于大多数竞争模型,而且它也满足视频异常检测任务的实时性要求。

表 4　不同方法在计算时间方面和参数的性能比较

Table 4　Performance comparison of different method in terms of computational time and parameters

方法	CPU/GHz	GPU	FLOPs/M	Param/G	帧率/fps
Unmasking[23]	2.3	–	–	–	20.00
AMDN[26]	2.1	K400	–	–	0.13
NNC[27]	2.3	–	–	–	24.00
Frame-Pred[28]	3.4	TITAN	78	64.200	25.00
Song[29]	–	–	6.829	6.196	35.00
RTFM[30]	–	–	28	186.900	–
本文算法	2.4	2 080 ti	5.96	6.100	31.00

4　结　　论

本文面向视频监控的异常事件检测领域提出了一个基于深度学习的端对端的检测框架。异常事件通常发生于前景的运动区域且异常事件出现的区域位于视频场景的某个局部区域。基于此使用前景检测算法提取前景区域,去除背景区域对异常检测的干扰。对于表观和运动特征分别用编码解码网络提取特征表述量,并利用深度概率估计网络生成高维特征空间的超球面构建概率模型。不同于以往的异常评价体系,本文同时使用概率估

计值、重构误差值和预测误差值作为互补量来联合评估视频帧的异常性。通过大量实验数据表明,本文的方法优于现有的大多数异常检测算法,证实了方法对异常检测的有效性。

参考文献

[1] HASAN M, CHOI J, NEUMANN J. Learning temporal regularity in video sequences[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 733-742.

[2] 于晓升, 许茗, 王莹, 等. 基于卷积变分自编码器的异常事件检测方法[J]. 仪器仪表学报, 2021, 42(5):151-158.
YU X SH, XU M, WANG Y, et al. Anomaly detection method based on convolutional variational auto-encoder[J]. Chinese Journal of Scientific Instrument, 2021, 42(5): 151-158.

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

[4] 付乐天, 李鹏, 高莲. 考虑样本异常值的改进最小二乘支持向量机算法[J]. 仪器仪表学报, 2021, 42(6):179-190.
FU L T, LI P, GAO L. Improved LSSVM algorithm considering sample outliers[J]. Chinese Journal of Scientific Instrument, 2021, 42(6): 179-190.

[5] LIU F T, TING K M, ZHOU Z H. Isolation forest[C]. 2008 Eighth IEEE International Conference on Data Mining, 2008: 413-422.

[6] REYNOLDS D A. Gaussian mixture models[J]. Encyclopedia of Biometrics, 2009, 741: 659-663.

[7] 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: 117-127.

[8] FAN Y, WEN G, LI D. Video anomaly detection and localization via Gaussian mixture fully convolutional variational autoencoder[J]. Computer Vision and Image Understanding, 2020: 102920.

[9] RIBEIRO M, LAZZARETTI A E, LOPES H S. A study of deep convolutional auto-encoders for anomaly detection in videos[J]. Pattern Recognition Letters, 2018: 13-22.

[10] ZONG B, SONG Q, MIN M R. Deep autoencoding gaussian mixture model for unsupervised anomaly detection[C]. International Conference on Learning Representations, 2018.

[11] HINAMI R, MEI T, SATOH S. Joint detection and recounting of abnormal events by learning deep generic knowledge[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 3619-3627.

[12] SABOKROU M, KHALOOEI M, FATHY M. Adversarially learned one-class classifier for novelty detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 3379-3388.

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

[14] DOSOVITSKIY A, FISCHER P, ILG E, et al. Flownet: Learning optical flow with convolutional networks[C]. Proceedings of the IEEE International Conference on Computer Vision, 2015: 2758-2766.

[15] YU G, WANG S, CAI Z, et al. Cloze test helps: Effective video anomaly detection via learning to complete video events[C]. Proceedings of the 28th ACM International Conference on Multimedia, 2020: 583-591.

[16] TAX D M J, DUIN R P W. Support vector data description[J]. Machine Learning, 2004, 54: 45-66.

[17] RUFF L, VANDERMEULEN R, GOERNITZ N. Deep one-class classification[C]. International Conference on Machine Learning, 2018: 4393-4402.

[18] MAHADEVAN V, LI W, BHALODIA V. Anomaly detection in crowded scenes[C]. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010: 1975-1981.

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

[20] LUO W, LIU W, GAO S. Remembering history with convolutional lstm for anomaly detection[C]. 2017 IEEE International Conference on Multimedia and Expo, 2017: 439-444.

[21] KIM J, GRAUMAN K. Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates[C]. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009: 2921-2928.

[22] LUO W, LIU W, GAO S. A revisit of sparse coding based anomaly detection in stacked rnn framework[C].

Proceedings of the IEEE International Conference on Computer Vision, 2017: 341-349.

[23] TUDOR I R, SMEUREANU S, ALEXE B. Unmasking the abnormal events in video[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 2895-2903.

[24] GONG D, LIU L, LE V. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection[C]. Proceedings of the IEEE International Conference on Computer Vision. 2019: 1705-1714.

[25] CHANG Y, TU Z, XIE W, et al. Video anomaly detection with spatio-temporal dissociation[J]. Pattern Recognition, 2022, 122: 108213.

[26] 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: 117-127.

[27] IONESCU R T, SMEUREANU S, POPESCU M, et al. Detecting abnormal events in video using narrowed normality clusters[C]. 2019 IEEE winter Conference on Applications of Computer Vision (WACV), 2019: 1951-1960.

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

[29] SONG J F, ZHAO H L, WEN D Y, et al. Video anomaly detection based on optical flow feature enhanced spatio-temporal feature network FusionNet-LSTM-G[J]. IEEE Access, 2022, 10: 130314-130325.

[30] TIAN Y, PANG G, CHEN Y, et al. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 4975-4986.

作者简介

常兴亚, 2018 年于东北大学获得硕士学位,现为东北大学信息科学与工程学院博士研究生,主要研究方向为图像处理与计算机视觉。

E-mail: 1810338@stu.neu.edu.cn

Chang Xingya received his M. Sc. degree from Northeastern University in 2018. He is currently a Ph. D. candidate in the School of Information Science and Engineering at Northeastern University. His main research interests include image processing and computer vision.

陈东岳(通信作者),分别在 2002 年和 2007 年于复旦大学获得硕士学位和博士学位,现为东北大学教授,主要研究方向为图像处理与计算机视觉。

E-mail: chendongyue@ise.neu.edu.cn

Chen Dongyue (Corresponding author) received his M. Sc. and Ph. D. degrees both from Fudan University in 2002 and 2007, respectively. He is currently a professor at Northeastern University. His main research interests include image processing and computer vision.