ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 1,Jan. 2024,pp. 204－214
DOI:10. 16337/j. 1004-9037. 2024. 01. 018

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

# 基于时空依赖关系和特征融合的弱监督视频异常检测

柳德云，李　莹，周　震，吉根林

（南京师范大学计算机与电子信息学院/人工智能学院，南京 210023）

**摘　要：**弱监督视频异常检测由于抗干扰性强、数据标注要求低，成为视频异常事件检测研究的热点。在现有的工作中，大多数弱监督视频异常检测方法认为各个视频段独立同分布，单独判断每个视频段是否异常，忽略了视频段间的时空依赖关系。为此，提出了一种基于时空依赖关系和特征融合的弱监督视频异常检测方法，在保留视频段原始特征的同时，使用视频段之间的索引距离和特征相似程度拟合视频段的时间和空间依赖关系，构建视频段的关系特征。通过融合原始特征和关系特征，更好地表达视频的动态特性和时序关系。在UCF-Crime和ShanghaiTech两个基准数据集上进行了大量实验，实验结果表明所提方法的AUC指标优于其他方法，AUC值分别达到了80.1%和94.6%。

**关键词：**视频异常事件检测；时空依赖关系；特征融合；图卷积神经网络；注意力机制

**中图分类号：**TP391　　**文献标志码：**A

# Weakly Supervised Video Anomaly Detection Based on Spatio-Temporal Dependence and Feature Fusion

LIU Deyun, LI Ying, ZHOU Zhen, JI Genlin

(School of Computer and Electronic Information/Artificial Intelligence, Nanjing Normal University, Nanjing 210023, China)

**Abstract:** Weakly supervised video anomaly detection has become a hot spot in video anomaly detection research due to its strong anti-interference and low data labeling requirements. In the existing methods, most of the weakly supervised video anomaly detection methods assume that the clips in each video distribute independently, and determine whether it is abnormal for each video clip independently, ignoring the temporal and spatial information between video clips. To alleviate these problems, this paper proposes a weakly supervised anomaly detection method based on spatio-temporal dependence and feature fusion. Retaining the original characteristics of video clips, this method uses the distance of index and the similarity of features between video clips to fit the time dependence and the spatial dependencies of video, which builds the relationship characteristics of video clips. By fusing the original features and relationship features, the dynamic characteristics and temporal relationship of videos can be better expressed. Extensive experiments on two benchmark datasets, UCF-Crime and ShanghaiTech, demonstrate that the proposed method outperforms other methods with the AUC values reaching 80.1% and 94.6%, respectively.

**Key words:** video anomaly event detection; spatio-temporal dependence; feature fusion; graph convolutional neural network; attention mechanism

**基金项目：**国家自然科学基金 (41971343, 62102186)。
**收稿日期：**2022-09-09；**修订日期：**2023-02-14

## 引　言

视频异常检测即检测视频中不符合预期的事件或正常行为概念的模式[1]。随着监控视频数量与日俱增,视频异常检测得到广泛关注。然而,由于监控场景的多样性,异常事件定义的主观不确定性,异常事件检测任务存在一定挑战[2]。为了规避异常事件的复杂定义,研究者往往使用仅包含正常事件的视频训练检测模型,使模型学习正常事件的模式,将不同于正常模式的事件视为异常。这种思路虽然可以拟合训练集中正常事件的模式,但由于没有异常视频参与训练,训练集中未曾出现过的正常事件也会因无法重构被误报为异常。因此,这类视频异常检测方法在复杂或恶劣工作环境下,往往具有较高的误报率[3]。

使用正常视频和异常视频同时参与训练,可以提高系统对异常事件的认知能力。但这种完全监督的训练模式,需要使用大量具有视频帧级别标注的监控视频数据。人工对海量监控视频数据进行标注需要花费大量时间和精力,且具有一定主观性。因此,仅存在大量视频级别标注的监控视频数据,无法支撑研究者训练完全监督的视频异常检测模型。如何利用仅有视频级别标注的数据,实现视频异常事件的检测和准确定位,成为目前视频异常事件检测领域的难点。

弱监督视频异常检测由于抗干扰性强、数据标注要求低等特点,成为解决上述问题的有效途径。在现有的方法[4-6]中,研究者往往将视频划分为定长视频段,单独对每个视频段进行评分,根据异常分数和阈值的大小关系,判断视频段是否存在异常。然而,对每个视频段单独进行异常评分,忽略了视频段间的时空依赖关系[7]。这类方法虽然可以检测外观较为明显的异常事件,如车祸、火灾等。但如果视频中的事件是与正常事件差异较小的异常事件,这类方法往往会漏检。如图1所示,对在人行道使用滑板异常进行检测时,由于慢速滑行和步行的外观差异很小,仅凭单个视频段的特征,模型无法区分正常事件和异常事件。

图1　人行道使用滑板的异常事件

Fig.1　Abnormal events of using skateboards on sidewalks

针对以上问题,本文提出了一种基于时空依赖关系和特征融合的弱监督视频异常检测方法(Weakly supervised anomaly detection method based on spatio-temporal dependence and feature fusion, SDFF)。在保留视频段原始特征的同时,使用视频段之间的索引距离拟合视频段的时间依赖关系,使用视频段之间特征相似程度拟合视频段的空间依赖关系,构建视频段的关系特征。通过融合原始特征和关系特征,更好地表达视频的动态特性和时序关系。本文的主要贡献如下:

(1)提出了一种视频段时空依赖关系提取方法(Spatiotemporal dependency extraction method, SDE),通过注意力机制和时空依赖关系图对视频段的时间依赖关系和空间依赖关系进行建模,发掘视频段间的关系特征。

(2)提出了一种基于时空依赖关系和特征融合的弱监督视频异常检测方法SDFF,使用视频段原始特征和关系特征进行异常检测,改善了现有方法中视频段时空依赖关系利用不充分的问题。

## 1　相关工作

传统的视频异常检测方法使用手工特征进行异常检测。Basharat等[8]分别将检测主体的运动模式和轮廓尺寸变化情况作为变量,使用多变量高斯混合模型检测主体运动轨迹异常。这种手工提取特征的方式,只能检测某类特定异常,泛化性较差。随着深度学习方法的出现,越来越多的视频异常检测方法[4-6,9-16]使用预训练的卷积神经网络提取视频特征,使用深度学习模型进行异常检测,以弥补手工提取

特征的主观性强、泛化性差的缺陷。Liu等[9]使用帧预测的方式检测视频中的异常事件。李欣璐等[10]将光流特征与HOG特征融合,并使用卷积自编码器进行重构视频帧,根据重构视频帧与真实帧的误差判断是否存在异常事件。但由于没有异常视频参与训练,训练集中未曾出现过的正常事件也会因无法重构被误报为异常。为了增强模型对异常事件的认识能力,同时降低数据标注带来的巨大成本,研究者开始使用弱监督的方式训练模型。

### 1.1 弱监督视频异常检测

弱监督视频异常检测是视频异常检测领域的重要分支。Sultani等[4]首次将视频划分为定长的视频片段,将异常视频片段集称为正包,正常视频片段集合称为负包。使用多示例排序损失扩大正包和负包异常分数的差异,根据片段的异常分数,判断是否存在异常事件。此方法为视频异常检测领域提供了新的解决方案,但也带来了新的问题。多示例排序损失只扩大了正包和负包中最值片段的分数差异,忽略了正包的包内损失,多次迭代后,正包中的视频片段区分度下降。因此,Sultani等[4]的检测模型存在很高的误报率,即大量正常视频段被误判为异常。为了解决此问题,Zhang等[5]提出新型排序损失,缩小负包中最大值和最小值片段的距离,增大正包中最大值和最小值片段的距离。该方法既关注包间距离,又关注包内距离。然而,在不同时长的视频中异常片段并非仅有一个,仅考虑最值片段忽略了视频时长与其所包含异常片段数量的关系。为了解决此问题,Wan等[6]使用动态排序损失进行视频异常检测,根据视频时长动态选择多个最值异常分数,计算排序损失。虽然动态排序损失考虑了视频长度与异常片段数量的关系,但和Sultani等[4]相同,忽略了正包的包内损失,使正常段产生虚假的高分,在检测时被误报为异常,因此仍需要进一步完善。Zaheer等[11]引入基于自我推理的聚类损失,新损失的引入增大了视频片段的分离性,但没有考虑视频段之间的相关性。

以上方法虽然在逐步加强异常视频段和正常视频段的可分离性,但都基于视频段间独立同分布的假设,忽略了视频段间的依赖关系。接下来部分学者注意到视频段之间关系的重要性,Zhu等[12]在时序模型中添加注意力网络,获得视频段全局关系,然而该方法只注重了视频的空间关系,并未考虑视频片段间的时间关系。Ma等[13]提出一种多重注意力框架,使用AAM模块关注异常视频帧,并使用判别异常注意力模块DAAM和生成异常注意力模块GAAM增强AAM的判别效果,多重注意力的加入提升了模型对异常的判别能力。Tian等[14]提出RTFM方法,使用MTN捕捉片段特征之间的长期和短期时间相关性,同时扩大异常和正常视频特征之间的差异性,获得了良好的效果,体现了关注视频段之间的依赖关系对提升模型异常检测能力的重要性。

### 1.2 图卷积神经网络

近年来,图卷积神经网络(Graph convolutional neural network, GCN)广泛应用于目标检测领域[15-17]。研究者将实际问题看作图中节点之间的连接和信息传播问题,对节点之间的依赖关系进行建模,取得了优异的效果。随后,研究者开始在视频异常检测领域使用GCN发掘视频段关系。Zhong等[18]使用GCN对异常视频中被误判的正常片段进行去噪,生成视频段伪标签,并使用伪标签进行模型训练。该方法尽管在训练阶段捕获了视频长距离的时间相关性,但去噪过程会导致部分异常信息丢失。周航等[19]使用图卷积神经网络发掘视频的依赖关系,但直接使用视频段的特征构建时空图,会具有较弱的特征可分离性,同时忽略了视频原始特征对检测结果的影响。Wu等[20]提出了一种基于图卷积神经网络的多分支视频异常检测方法。该方法分别捕获长期依赖关系、局部位置关系和预测分数的相似关系,以期完善视频段的特征表示。然而,该方法扩大了模型的计算代价,同时忽略了视频段自身的信息。本文使用注意力机制和图卷积神经网络发现视频段间的依赖关系特征,并将其与视频段原始特征融合,更好地表达视频的动态特性和时序关系,使用虚高损失弥补了动态多示例损失对异常视频中正常视频片段的约束缺失,实现对视频异常事件的精准检测。

## 2 视频异常检测方法

### 2.1 处理流程

视频异常检测(SDFF)的处理流程如图2所示。在训练阶段,首先以16帧为步长将视频划分为不重叠的视频段,送入在Kinetics数据集[21]预训练的I3D模型提取特征,生成特征矩阵。然后使用一维卷积神经网络将视频段特征矩阵降维,得到视频原始特征矩阵。使用本文提出的SDE方法发掘视频段间的依赖关系,生成关系特征矩阵。其次融合原始特征矩阵和关系特征矩阵形成最终的特征矩阵。最后使用全连接神经网络将融合特征映射为异常分数,并根据损失函数迭代训练模型。在测试阶段,将测试视频划分片段并提取特征后,输入到训练完成的检测模型中,生成对应的视频段异常分数,若某视频段的异常分数超过异常阈值,则认为该视频片段中所有视频帧均为异常。在SDFF的处理流程中,使用SDE方法提取关系特征为关键部分,而注意力机制模块和时空依赖图模块为此部分的核心。

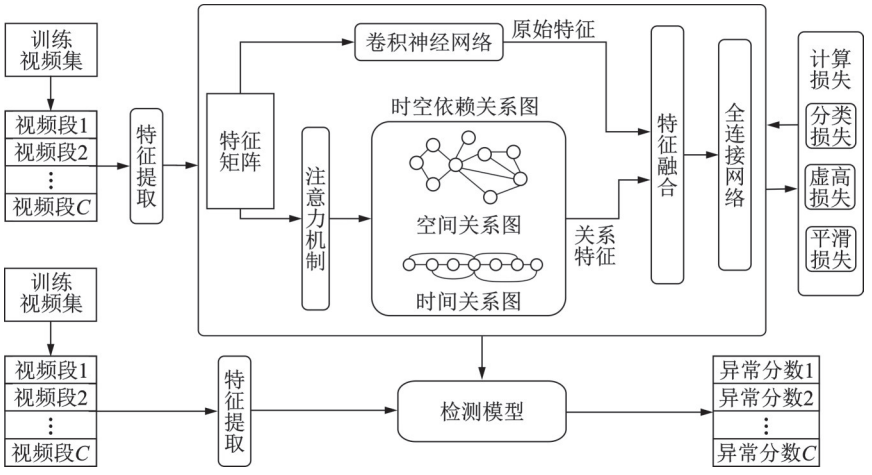



图2 SDFF处理流程

Fig.2 Process flow of SDFF

### 2.2 注意力机制

传统卷积神经网络的感受野只在卷积核大小的范围内,无法获得特征间的全局关系,直接进行卷积降维将丢失视频片段的全局信息。SDFF方法使用注意力机制,提取视频段的关键特征,保留关键信息,为下一步关系特征的提取奠定基础。注意力机制的结构如图3所示,其中"⊗"表示矩阵乘法,"⊕"表示矩阵加法。

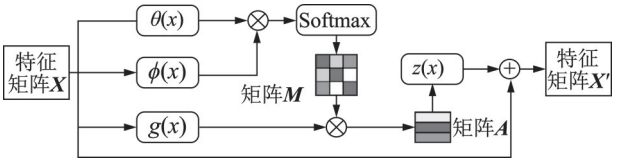



图3 注意力机制结构

Fig.3 Structure of attention mechanism

首先,对特征矩阵$\boldsymbol{X}$分别进行$\theta(x)$,$\phi(x)$两个一维卷积操作,将特征维度压缩到原来的1/8。然后,通过矩阵乘法计算当前视频段与其他视频段的特征关联性,计算方法为

$$f\left(x_i,x_j\right)=\boldsymbol{\theta}(x_i)^{\mathrm{T}}\cdot\boldsymbol{\phi}(x_j) \tag{1}$$

式中:$x_i$、$x_j$是视频段的特征;$i$、$j$表示视频段的索引号;$\theta(x)=w_\theta\cdot x$,$\phi(x)=w_\phi\cdot x$。

然后,使用Softmax函数按行对关联性分数进行归一化处理,得到关联性矩阵$\boldsymbol{M}$。$\boldsymbol{M}_{i,j}$表示视频段$i$与视频段$j$的关联强度。其次,将整个特征矩阵$\boldsymbol{X}$送入$g(x)$进行一维卷积操作并与关联性矩阵$\boldsymbol{M}$进行

矩阵乘法，得到注意力矩阵$\boldsymbol{A}$，计算方式如式(2)所示，其中$g(x)=w_g \cdot x$。

$$\boldsymbol{A}=\boldsymbol{M} \cdot g(\boldsymbol{X}) \tag{2}$$

最后，将注意力矩阵$\boldsymbol{A}$送入$z(x)$进行一维卷积操作并与原始特征$\boldsymbol{X}$相加获得带有注意力的特征矩阵$\boldsymbol{X}^{`}$，计算方式如式(3)所示，其中$z(x)=w_z \cdot x$。

$$\boldsymbol{X}'=z(\boldsymbol{A})+\boldsymbol{X} \tag{3}$$

### 2.3 时空依赖关系图

时空依赖关系图由时间关系图和空间关系图两个子图构成。分别通过图卷积神经网络挖掘时间维度的依赖关系和空间维度的依赖关系，最终形成视频段的关系特征。

时间关系图：视频段的时间依赖关系取决于视频段之间的时间间隔，视频段间隔时间越小，其相关性越高。SDFF使用视频段索引之间的欧式距离$d(i,j)$来拟合这种关系，构建时间关系图$G^{\mathrm{t}}$。距离计算公式$d(i,j)$为

$$d(i,j)=|i-j| \tag{4}$$

式中：$i$、$j$为视频段的索引号，$i$、$j \in \{0,1,\cdots,C\}$，$C$为视频划分的视频段数。时间关系图的邻接矩阵$\boldsymbol{A}^{\mathrm{t}}$的元素$A_{ij}^{\mathrm{t}}$定义为

$$A_{ij}^{\mathrm{t}}=\frac{\exp\left(-d(i,j)\right)}{\sum_{j=0}^{C}\exp\left(-d(i,j)\right)} \tag{5}$$

空间关系图：不同的视频段中可能存在相同的背景或者相似的运动模式，而异常事件与常规情况有着明显差异。因此SDFF使用特征点乘相似程度来表示视频段间的空间关系。空间关系图的邻接矩阵$\boldsymbol{A}^{\mathrm{sim}}$的元素$A_{ij}^{\mathrm{sim}}$定义为

$$A_{ij}^{\mathrm{sim}}=\frac{\exp\left(\boldsymbol{x}_i^{\mathrm{T}} \cdot \boldsymbol{x}_j\right)}{\sum_{j=0}^{C}\exp\left(\boldsymbol{x}_i^{\mathrm{T}} \cdot \boldsymbol{x}_j\right)} \tag{6}$$

最后，SDFF使用图卷积神经网络对时间关系图和空间关系图进行卷积操作，发掘视频段间的时空依赖关系。图卷积神经网络的传播公式为

$$\boldsymbol{H}^{(l+1)}=\sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{D}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}\right) \tag{7}$$

式中：$\boldsymbol{H}^{(l)} \in \mathbf{R}^{C \times \mathrm{D}}$为第$l$层的状态；$D$为特征维度；$\sigma$为激活函数，这里使用ReLU函数；矩阵$\tilde{\boldsymbol{A}}$为带有自连接的邻接矩阵，是邻接矩阵$\boldsymbol{A}$和节点自身关系$\boldsymbol{I}_N$的和，即$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}_N$；$\tilde{\boldsymbol{D}}=\sum_j \tilde{\boldsymbol{A}}_{ij}$为矩阵$\tilde{\boldsymbol{A}}$的对角矩阵；$\boldsymbol{H}^{(0)}=\boldsymbol{X}^{`}$，这里$\boldsymbol{X}^{`}$是注意力机制输出的特征；$\boldsymbol{W}^{(l)}$为第$l$层待训练的参数。

### 2.4 损失函数

SDFF的损失函数分别为：分类损失$L_{\mathrm{c}}$、虚高损失$L_{\mathrm{f}}$和平滑损失$L_{\mathrm{s}}$。

分类损失：为了使检测模型更好地区分正常视频段和异常视频段，参考Wan等[6]的动态多示例损失，SDFF采用相似的分类损失来扩大正常视频段和异常视频段的异常分数差异。分类损失的计算为

$$L_{\mathrm{c}}=\frac{1}{k}\sum_{s_i \in \boldsymbol{S}}\left[-y\ln s_i+(1-y)\ln(1-s_i)\right] \tag{8}$$

式中：$k=\lceil C/\alpha \rceil$，$k$表示视频$v$有$k$个异常视频段，$\alpha$为超参数，$\alpha \in [1,C]$；$s_i$表示视频$v$第$i$个视频段的异

常分数;$y$为视频$v$的标签;集合$S$为视频$v$中前$k$个最大的异常分数,具体定义如式(9)所示。其中$\mathrm{sub}_i$表示视频段中第$i$大的异常分数。

$$S=\left\{\mathrm{sub}_i \mid i=1,2,\cdots,k\right\} \tag{9}$$

虚高损失:为了改进动态多示例损失导致的异常分数虚高现象,加强对异常视频中的正常视频段的约束。SDFF提出虚高损失函数$L_\mathrm{f}$为

$$L_\mathrm{f}=\frac{1}{(c-k)}\sum_{s_i\notin S,\,y=1}s_i \tag{10}$$

平滑损失:为了使正常视频中各个视频段的异常分数更加平滑,SDFF采用平滑损失函数$L_\mathrm{s}$,定义如式(11)所示。其中,$\hat{s}$表示视频$v$异常分数的平均值。

$$L_\mathrm{s}=\sum_{i=1,\,y=0}^{c}(s_i-\hat{s})^2 \tag{11}$$

本文总损失函数如式(12)所示。其中$\lambda_1$、$\lambda_2$、$\lambda_3$为超参数,用于调整各损失的权重关系。综上所述,视频异常检测模型的训练过程如算法1所示。

$$\mathcal{L}=\lambda_1 L_\mathrm{c}+\lambda_2 L_\mathrm{f}+\lambda_3 L_\mathrm{s} \tag{12}$$

**算法1** 视频异常检测模型的训练算法

输入:训练视频集,标签集$Y$,学习率$p$,损失权重$\lambda_1,\lambda_2,\lambda_3$

输出:模型参数$W$

(1) 将视频以16帧为步长划分为片段

(2) WHILE模型的参数没有收敛DO

(3) 选取一组正常和异常视频,划分为片段

(4) 提取特征,通过模型得到异常分数集合$S_i$

(5) 根据$S_i$计算$L_\mathrm{c}$

(6) IF $y_i=1$ THEN

(7) 根据$S_i$计算$L_\mathrm{f}$

(8) ELSE IF $y_i==0$ THEN

(9) 根据$S_i$计算$L_\mathrm{s}$

(10) END IF

(11) $\mathcal{L}\leftarrow\lambda_1 L_\mathrm{c}+\lambda_2 L_\mathrm{f}+\lambda_3 L_\mathrm{s}$

(12) 根据损失函数修正参数矩阵$\boldsymbol{W}$

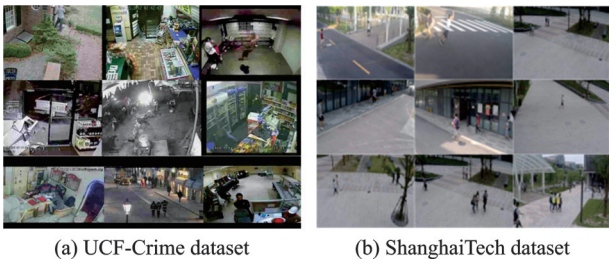
(13) END WHILE

(14) RETURN参数$\boldsymbol{W}$

## 3 实验与结果分析

### 3.1 数据集与评价指标

UCF-Crime数据集[4]是视频异常检测领域公开数据集之一,由长时间未经剪辑的真实世界监控视频组成,包含虐待、纵火、爆炸、打架等现实世界中的13种异常情况,共计1 900个视频。数据集部分视频帧如图4(a)所示。和以往工作[4]相同,本文将其划分为1 610个



(a) UCF-Crime dataset (b) ShanghaiTech dataset

图4 数据集部分视频帧展示

Fig.4 Partial video frame display of datasets

训练视频和290个测试视频。

ShanghaiTech数据集[22]是视频异常检测领域的公开数据集之一。包含非法行驶机动车、使用滑板、抢夺、打架等异常情况,共计13个场景437个视频。数据集部分视频帧如图4(b)所示。和以往工作相同,本文依旧采用Zhong等[18]的数据划分方式,将数据集分为238个训练视频和199个测试视频。

评价指标:和以往工作相同[4-6,11-13,18-20],本文使用阈值为0.5的接收者操作特征曲线下面积(Area under the curve, AUC)作为评价指标。在视频异常检测任务中AUC越高,模型的检测性能越好。

### 3.2 实验设置

实验环境:实验服务器配置为Intel(R) Xeon(R) Silver 4210R CPU @ 2.40 GHz,服务器内存187 GB,GPU采用GeForce RTX3080显存10 GB,采用Ubuntu 18.04系统。实验编程环境为Python 3.6.8,Pytorch版本为1.7.1,CUDA版本11.2。

参数设置:将输入视频帧大小重新调整为224像素×224像素,以16帧为步长将视频划分为非重叠的视频段,使用在Kinetics数据集[21]预训练的I3D模型提取视频段特征,一方面将特征送入卷积神经网络降维,构造原始特征。网络输入输出单元数分别是1 024和64。另一方面,特征输入到注意力机制中添加全局注意力,然后输入到图卷积神经网络生成关系特征,图卷积神经网络输入输出单元数分别是1 024和64。异常检测部分网络结构为全连接神经网络,输入输出单元数分别是128和1,最后一层使用Sigmoid激活,其他层使用ReLU激活。在每层之间使用60%的Dropout,以提高模型泛化能力。在模型参数优化上,本文使用Adam优化器,并设置初始学习率为0.01。分类损失、虚高损失和平滑损失的权重对应的超参数$\lambda_1$、$\lambda_2$、$\lambda_3$分别设置为5、20和15。分类损失和虚高损失的超参数$\alpha$值选择为4以获得最佳性能。

### 3.3 实验结果分析

为了验证SDFF的有效性,在UCF-Crime数据集[4]和ShanghaiTech数据集[22]将其与其他方法进行对比,实验结果分别如表1、2所示。其中"*"表示基准方法,"—"表示使用非标准特征提取网络。

由如表1可知,SDFF相比基准方法(Wan等[6]) AUC指标高出1.05%;相比Zhang等[5]的方法AUC指标高出1.35%。为了分析特征类型对实验结果的影响,本文使用C3D特征做了相关实验,在使用I3D特征时,本文方法高于Sultani等[4]方法4.6%,当使用C3D特征时,本文方法仍高出3.5%;相比Zaheer等[11]的方法AUC指标高出0.53%。以上结果表明,SDFF方法有效弥补前期工作中时空依赖关系的缺失问题。相比同样使用注意力机制的Zhu等[12]的方法AUC指标高出1.01%,表明SDFF使用SDE方法提取的关系特征相比单纯使用注意力机制可以更好地描述视频段间的时空依赖关系。

**表1 在UCF-Crime数据集上AUC结果**

**Table 1 AUC results on UCF-Crime dataset**

| Method | Feature | AUC/% |
|---|---|---|
| Sultani[4] | C3D RGB | 75.41 |
| Hasan[23] | — | 50.60 |
| Wan[6]* | I3D RGB&I3D Flow | 78.96 |
| Zhang[5] | C3D RGB | 78.66 |
| Zaheer[11] | — | 79.54 |
| Zhu[12] | PWC Flow | 79.00 |
| $SDFF_{C3D}$(ours) | C3D RGB | 78.91 |
| $SDFF_{I3D}$(ours) | I3D RGB | 80.01 |

**表2 在ShanghaiTech数据集AUC结果**

**Table 2 AUC results on ShanghaiTech dataset**

| Method | Feature | AUC/% |
|---|---|---|
| Sultani[4] | I3D RGB | 85.33 |
| Zhong[18] | TSN RGB | 84.44 |
| Zhong[18] | C3D RGB | 76.44 |
| Zhang[5] | I3D RGB | 82.50 |
| Wan[6]* | I3D RGB &I3D Flow | 91.24 |
| Zhou[19] | I3D RGB | 89.88 |
| Ma[13] | I3D RGB | 85.70 |
| SDFF(ours) | I3D RGB | 94.67 |

由表2结果可知,SDFF相比基准方法(Wan等[6])AUC指标高出3.43%;表明SDFF的关系特征和虚高损失相比基准方法中AR-Net网络和动态多示例损失具有更好的异常检测效果;相比同样基于GCN的Zhong等[18]的方法,AUC指标高出10.23%,相比Ma等[13]的多重注意力方法高出8.97%。这表明SDFF的时空依赖关系提取方法SDE可以更有效地捕获视频段的时空依赖关系。

SDFF的可视化结果分别如图5、6所示。横坐标为视频帧序号,纵坐标为视频帧异常分数;中间深色区域为真实的异常区域,直线和虚线为不同方法的检测结果。

SDFF与Wan等[6]的方法,对于在人行道使用滑板异常事件(ShanghaiTech数据集01_0064视频)的检测结

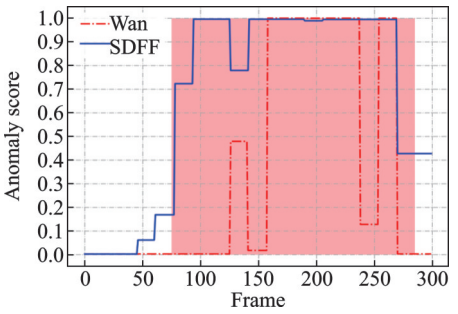



图5 SDFF和Baseline方法视频检测结果对比

Fig.5 Comparison of video detection results between SDFF and Baseline method

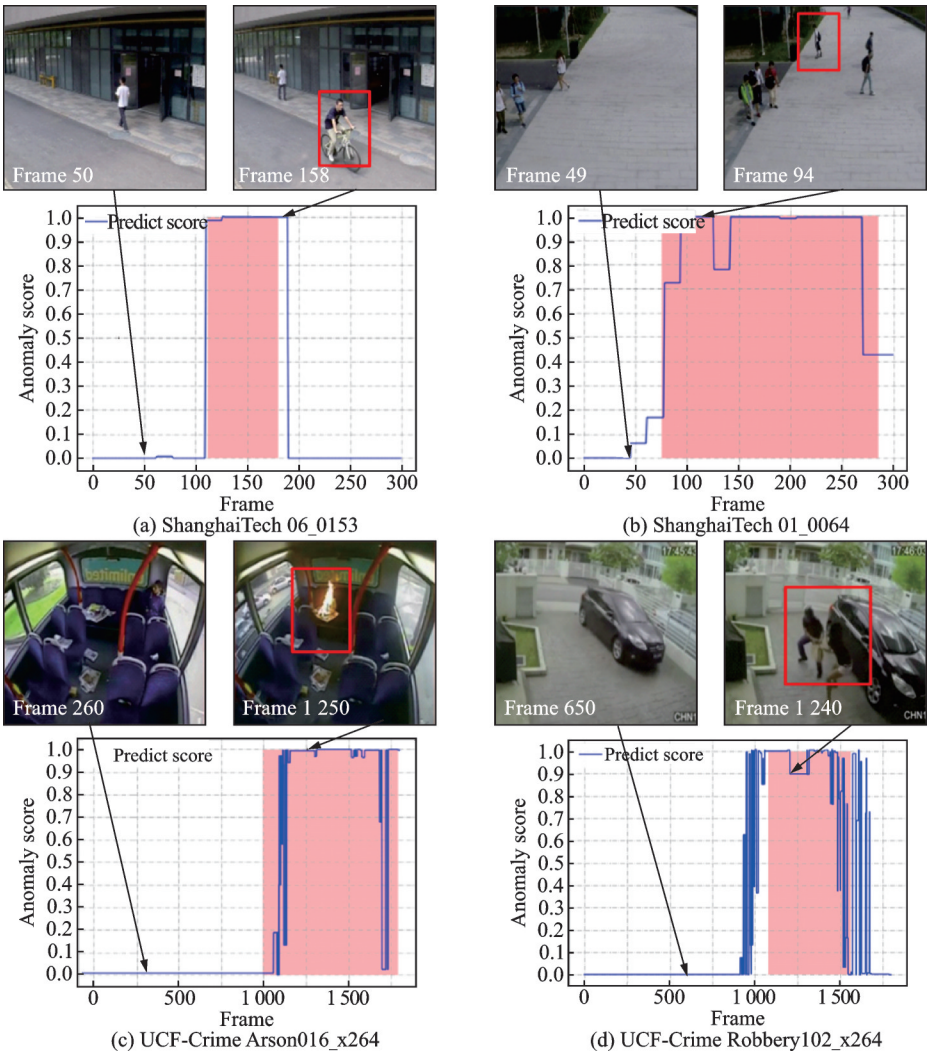



图6 不同测试视频检测结果展示

Fig.6 Display of different test video detection results

果可视化情况,如图5所示。由图5可知,SDFF方法对于滑板等细微异常具有更好的检测能力。

此外,SDFF在不同数据集的可视化检测结果如图6所示。下面以图6(b,c)两个结果为例,进行详细描述,其他可视化结果同例。ShanghaiTech数据集测试视频01_0064的异常检测结果,如图6(b)所示。在49帧时监控画面显示并无异常情况发生,该帧的异常分数接近0;然而在94帧时,如视频画面中心方框所标,有人使用滑板经过路口。这个行为在ShanghaiTech数据集中被定义为异常,该帧的异常分数接近1。UCF-Crime数据集测试视频Arson016_x264异常检测结果,如图6(c)所示。在260帧时,视频画面中乘客在乘坐公交车,并无异常情况,该帧的异常分数接近0;在1 250帧时,如视频画面中心方框所示,公交车座位上报纸正在燃烧,这在UCF-Crime数据集中被定义为纵火异常,该帧的异常分数接近1。从以上实验结果可以看出,SDFF方法对外观不明显的异常事件和外观明显的异常事件都具有良好的检测能力。

同时,图6(d)中在异常事件发生前后,模型检测的结果产生剧烈抖动,主要原因是异常定义的模糊性。即相同的动作,在不同场景下异常结果不同。在抢夺(Robbery)类别下,视频标注认为当抢夺成功后异常结束。而原视频(Robbery102_x264)在抢夺结束后仍在追逐,此行为在打斗(Fighting)类别下仍被认为是异常事件。因此,模型的检测结果出现了剧烈抖动。

### 3.4 消融实验

为了分析模型的结构和超参数对视频异常检测性能的影响。在ShanghaiTech数据集上进行消融实验。

#### 3.4.1 模型结构分析

SDFF方法的不同结构在ShanghaiTech数据集上的检测结果,如表3所示。其中,GCN表示SDE中时空依赖关系图结构;Atten表示SDE中注意力机制结构;Floss表示损失函数中的虚高损失结构;“√”表示实验中包括该结构;“×”表示实验中未包括该结构。

**表3 不同模型结构在ShanghaiTech数据集上异常检测AUC结果**

**Table 3 AUC results of anomaly detection on ShanghaiTech dataset with different model structures**

| OF | GCN | Atten | Floss | AUC/% |
|---|---|---|---|---|
| × | √ | √ | √ | 92.73 |
| √ | × | √ | √ | 91.60 |
| √ | √ | × | √ | 94.17 |
| √ | √ | √ | × | 94.34 |
| √ | √ | √ | √ | 94.67 |

实验结果表明,本文方法的优势在于视频段原始特征和关系特征的有机融合。其中在关系特征提取过程中,注意力机制和时空依赖关系图模块发挥了重要作用。如表3所示,当停用原始特征(Original future, OF)后,检测结果的AUC指标下降1.94%。当停用时空依赖关系提取方法SDE中的GCN部分后,视频片段间的依赖关系表示不足,检测结果的AUC指标下降了3.07%。在停用时空依赖关系提取方法SDE中的注意力机制(Atten)后,异常视频段的特征和正常视频段的特征差异性减小,影响到后续时空依赖关系图构建的有效性,检测结果的AUC指标下降0.5%。在停用虚高损失Floss后,模型对于异常视频中的正常片段异常分数约束不足,检测结果的AUC指标下降0.33%。综上所述,SDFF的几个关键结构都具有提升视频异常检测结果AUC指标的能力。

#### 3.4.2 超参数对模型性能的影响

分类损失和虚高损失的超参数$\alpha$是异常视频中包含异常片段数量的度量。由式(8)可知,$\alpha$越小,认为异常视频中包含的异常片段越多,可能会导致正常视频段被判断为异常;$\alpha$越大,认为异常视频中包含的异常片段越少,可能会导致异常视频段被认为正常。因此,$\alpha$是一个重要参数,它的取值会对实验结果产生一定影响。

为了验证$\alpha$值选取的有效性,本文在$\alpha$分别为2、4、6、8、16、32情况下,在ShanghaiTech数据集上做了对比试验,异常检测模型的检测结果的AUC指标随着$\alpha$变化的情况如图7所示。随着超参数$\alpha$的增加,模型检测结果的AUC指标增长至顶峰后开始下降。AUC指标在$\alpha$值为4处达到峰值,因此将超参数$\alpha$的值设置为4,以获得最佳效果。

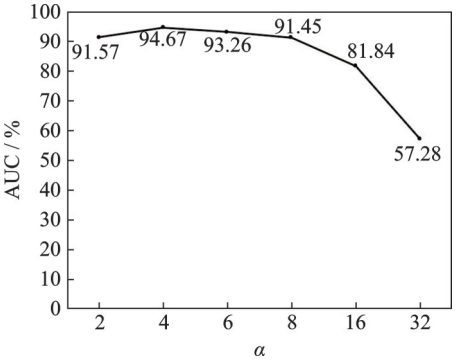



图7 超参数$\alpha$取值和AUC的关系

Fig.7 Relationship between hyper parameter α with AUC

## 4 结束语

本文提出了一种基于时空依赖关系和特征融合的弱监督视频异常检测方法SDFF。此方法在考虑视频段原始特征的同时,还关注到了视频段之间的时间依赖关系和空间依赖关系,构造了时空依赖关系图来拟合视频段依赖关系。此外,引入注意力机制自适应地发掘视频特征中的重要内容,增强特征的表示能力。在两个典型数据集上进行大量实验,结果表明SDFF方法在视频异常检测方面取得了优异的表现。未来的工作包括提高模型处理异常边界的能力和开发在线检测模型。

**References:**

[1] CHANDOLA V, BANERJEE A, KUMAR V. Anomaly detection: A survey[J]. ACM Computing Surveys, 2009, 41(3): 1-58.

[2] 吉根林,许振,李欣璐,等.监控视频中异常事件检测技术研究进展[J].南京航空航天大学学报,2020,52(5): 685-694.
JI Genlin, XU Zhen, LI Xinlu, et al. Progress on abnormal event detection technology in video surveillance[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2020, 52(5): 685-694.

[3] 王志国,章毓晋.监控视频异常检测:综述[J].清华大学学报(自然科版),2020,60(6): 518-529.
WANG Zhiguo, ZHANG Yujin. Anomaly detection in surveillance videos: A survey[J]. Journal of Tsinghua University (Science and Technology), 2020, 60(6): 518-529.

[4] SULTANI W, CHEN C, SHAH M. Real-world anomaly detection in surveillance videos[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 6479-6488.

[5] ZHANG Jianggong, Qing Laiyun, MIAO Jun. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection[C]//Proceedings of 2019 IEEE International Conference on Image Processing (ICIP). Taipei (China): IEEE, 2019: 4030-4034.

[6] WAN Boyang, FANG Yuming, XIA Xue, et al. Weakly supervised video anomaly detection via center-guided discriminative learning[C]//Proceedings of 2020 IEEE International Conference on Multimedia and Expo. London: IEEE, 2020: 1-6.

[7] ZHOU Zhihua, SUN Yuyin, LI Yufeng. Multi-instance learning by treating instances as Non-I.I.D. samples[C]//Proceedings of the 26th Annual International Conference on Machine Learning. Quebec: ACM, 2009: 1-8.

[8] BASHARAT A, GRITAI A, SHAH M. Learning object motion patterns for anomaly detection and improved object detection [C]//Proceedings of 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Anchorage: IEEE Computer Society, 2008: 24-26.

[9] LIU Wen, LUO Weixin, LIAN Dongze, et al. Future frame prediction for anomaly detection: A new baseline[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 6536-6545.

[10] 李欣璐,吉根林,赵斌.基于卷积自编码器分块学习的视频异常事件检测与定位[J].数据采集与理,2021,36(3): 489-497.
LI Xinlu, JI Genlin, ZHAO Bin. Convolutional auto-encoder patch learning based video anomaly event detection and localization[J]. Journal of Data Acquisition and Processing, 2021, 36(3): 489-497.

[11] ZAHEER M Z, MAHMOOD A, SHIN H, et al. A self-reasoning framework for anomaly detection using video-level labels[J]. IEEE Signal Process Letters, 2020, 27: 1705-1709.

[12] ZHU Yi, NEWSAM S D. Motion-aware feature for improved video anomaly detection[C]//Proceedings of the 30th British Machine Vision Conference 2019. Cardiff: BMVA, 2019: 270.

[13] MA Hualin, ZHANG Liyan. Attention-based framework for weakly supervised video anomaly detection[J]. The Journal of Supercomputing, 2022, 78(6): 8409-8429.

[14] TIAN Yu, PANG Guansong, CHEN Yuanhong, et al. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 4955-4966.

[15] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[C]//Proceedings of 5th International Conference on Learning Representations 2017. Toulon: OpenReview.net, 2017: 148226-148236.

[16] GKALELIS N, GOULAS A, GALANOPOULOS D, et al. ObjectGraphs: Using objects and a graph convolutional network for the bottom-up recognition and explanation of events in video[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops. Virtual: IEEE, 2021: 3375-3383.

[17] 马帅,刘建伟,左信.图神经网络综述[J].计算机研究与发展,2022,59(1):47-80.
MA Shuai, LIU Jianwei, ZUO Xin. Survey on graph neural network[J]. Journal of Computer Research and Development, 2022, 59(1): 47-80.

[18] ZHONG Jiaxing, LI Nannan, KONG Weijie, et al. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 1237-1246.

[19] 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.
ZHOU Hang, ZHAN Yongzhao, MAO Qirong. Video anomaly detection based on space-time fusion graph network learning[J]. Journal of Computer Research and Development, 2021, 58(1): 48-59.

[20] WU Peng, LIU Jing, SHI Yujia, et al. Not only look, but also listen: Learning multimodal violence detection under weak supervision[C]//Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020: 322-339.

[21] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE Computer Society, 2017: 4724-4733.

[22] LUO Weixin, LIU Wen, GAO Shenghua. A revisit of sparse coding based anomaly detection in stacked RNN framework[C]//Proceedings of IEEE International Conference on Computer Vision (ICCV 2017). Venice: IEEE Computer Society, 2017: 341-349.

[23] HASAN M, CHOI J, NEUMANN J, et al. Learning temporal regularity in video sequences[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Computer Society, 2016: 733-742.

**作者简介:**

**柳德云**(1997-),男,硕士研究生,研究方向:大数据分析与挖掘,E-mail: bboy-devin@163.com。

**李莹**(1990-),女,博士,副教授,研究方向:计算机视觉、图像检索与排序,E-mail: freeliying08@gmail.com。

**周震**(1996-),男,硕士研究生,研究方向:大数据分析与挖掘,E-mail: 2280153179@qq.com。

**吉根林**(1964-),**通信作者**,男,博士,教授,博士生导师,CCF高级会员,研究方向:大数据分析与挖掘,E-mail: glji@njnu.edu.cn。

(编辑:夏道家)