计算机科学
COMPUTER SCIENCE

# 基于双重动态记忆网络的弱监督视频异常检测

周文浩, 胡宏涛, 陈旭, 赵春晖

引用本文

周文浩, 胡宏涛, 陈旭, 赵春晖. 基于双重动态记忆网络的弱监督视频异常检测[J]. 计算机科学, 2024, 51(1): 243-251.

ZHOU Wenhao, HU Hongtao, CHEN Xu, ZHAO Chunhui. Weakly Supervised Video Anomaly Detection Based on Dual Dynamic Memory Network [J]. Computer Science, 2024, 51(1): 243-251.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300134

# 基于双重动态记忆网络的弱监督视频异常检测

**周文浩　胡宏涛　陈　旭　赵春晖**
浙江大学控制科学与工程学院　杭州 310027
(zhouwenhao@zju.edu.cn)

**摘　要**　视频异常检测需从整段视频中识别帧级别的异常行为。弱监督方法使用正常与异常视频,辅以视频级别标签训练模型,相比无监督视方法展现出了更优越的性能。然而,目前的弱监督视频异常检测方法无法记录视频长期模态,且部分方法为了获得更优的检测效果,利用了未来帧的信息,导致无法在线应用。为此,文中首次提出了一种基于双重动态记忆网络的弱监督视频异常检测方法,通过设计包含两个记忆模块的记忆网络来分别记录视频中长期的正常和异常模态。为了实现视频特征和记忆项的协同更新,采用读操作基于记忆模块中的记忆项对视频帧的特征进行增强,采用写操作基于视频帧特征对记忆项的内容进行更新,同时记忆项的数量在训练的过程中会动态调整从而适应不同视频监控场景的需求。在训练时,设计模态分离损失增加记忆项之间的区分度。在测试时,仅需要记忆项而不需要未来视频帧的参与,从而实现准确的在线检测。在两个公开的弱监督视频异常检测数据集上的实验结果表明,所提方法优于所有在线应用的方法,相比只能离线应用的方法也具有很强的竞争力。

**关键词**:视频异常检测;弱监督学习;记忆网络;多示例学习;深度学习
**中图分类号**　TP183

## Weakly Supervised Video Anomaly Detection Based on Dual Dynamic Memory Network

ZHOU Wenhao,HU Hongtao,CHEN Xu and ZHAO Chunhui
School of Control Science and Engineering,Zhejiang University,Hangzhou 310027,China

**Abstract**　Video anomaly detection aims to identify frame-level abnormal behaviors from the video. The weakly supervised methods use both normal and abnormal video supplemented by the video-level labels for training, which show better performance than the unsupervised methods. However, the current weakly supervised video anomaly detection methods cannot record the long-term mode of the video. At the same time, some methods use the information of future frames to achieve better detection results, which makes it impossible to apply online. For this reason, a weakly supervised video anomaly detection method based on dual dynamic memory network is proposed for the first time in this paper. The memory network containing two memory modules is designed to record the normal and abnormal modes of video in the long term respectively. In order to realize the collaborative update of video features and memory items, the read operation is used to enhance the features of video frames based on the memory items in the memory module, and the write operation is used to update the contents of memory items based on the features of video frames. At the same time, the number of memory items will be dynamically adjusted during the training process to meet the needs of different video monitoring scenarios. In training, a modality separation loss is proposed to increase the discrimination between memory items. During the test, only memory items are needed without the participation of future video frames, so that accurate online detection can be achieved. Experimental results on two public weakly supervised video anomaly detection datasets show that the proposed method is superior to all online application methods, and also has strong competitiveness compared with offline application methods.

**Keywords**　Video anomaly detection, Weakly supervised learning, Memory network, Multiple instance learning, Deep learning

## 1　引言

视频监控系统关乎着人民和社会的安全,是国家稳定运作不可或缺的一部分。随着我国经济和网络信息技术的迅速发展,视频监控系统在我国得到了广泛普及,并在交通调控、工业巡检和公共安全等领域发挥着重大的作用[1]。

基于机器视觉的方法[2-4]近年来被广泛应用在异常检测任务中。而视频异常检测任务旨在寻找视频中与普通事件不同的事件。在现实中,异常事件在视频中发生的频率较低,给视频帧打上标签的成本高昂,因此许多无监督异常检测方

到稿日期:2023-03-16　返修日期:2023-09-22
基金项目:国家自然科学基金杰出青年基金(62125306);NSFC——浙江两化融合联合基金(U1709211)
This work was supported by the National Natural Science Foundation of China(62125306) and NSFC—Zhejiang Joint Fund for the Integration of Industrialization and Informatization(U1709211).
通信作者:赵春晖(chhzhao@zju.edu.cn)

法[5-9]被提出。无监督视频异常检测方法假设在训练时只有正常视频的数据集,因此对训练集中未出现的正常事件具有高误报率,在复杂场景下异常检测能力较弱。

为了进一步解决复杂场景下的视频异常检测问题,近年来的一些研究致力于用弱监督的方法[10-16]进行异常检测,在视频级别标签监督下使用正常和异常视频进行训练。相比无监督方法,弱监督视频检测具有更优异的性能,且由于采用了视频级别而非帧级别的标签,其标注成本很低。目前弱监督视频异常检测方法多采用基于时序特征融合的方式,通过时序网络或关系网络来融合相邻帧或者全局视频的特征得到视频片段的特征,然后在多示例学习(Multiple Instance Learning,MIL)的框架下进行训练。然而,基于时序特征融合的方法存在如下的问题:(1)在视频内部融合了视频短期的特征,缺乏对视频间的长期正常和异常模态的表征能力;(2)一般需要融合视频中未来一段时间视频帧的特征,因此其产生的异常检测结果具有一定的滞后性,无法实现实时的在线检测。针对上述问题,本文提出了一种基于双重动态记忆网络的弱监督视频异常检测方法。与一般的记忆网络[17-19]仅记录正常模态不同,所提出的双重记忆网络分别记录正常和异常事件的长期模态,通过样本重采样获得置信正常和异常样本。基于采样得到的样本采用读写操作在增强视频帧特征的同时更新记忆网络中的记忆项。为了使记忆网络能够自适应不同的视频监控场景,本文设计了记忆项动态更新策略,在训练过程中根据训练数据动态调整记忆网络所需的记忆项数量,从而更精准地记录不同场景的长期模态。在应用时只需将测试帧与预先训练好的记忆项进行融合而无需未来帧的参与,从而实现准确的实时在线检测。本文的主要贡献如下:

(1)针对目前弱监督视频检测方法无法记录视频长期模态的问题,提出了一种基于双重动态记忆网络的弱监督视频异常检测方法,视频片段特征通过记忆网络中存储的视频长期正常和异常模态的记忆项得到增强。

(2)提出了记忆网络中记忆项的动态更新策略,使得模型能够自动调整记忆项的数量从而适应不同的视频监控场景。

(3)提出了模态分离损失用于训练记忆网络,使得记忆项之间更具有区分度,从而记录更丰富的视频模态。

## 2 相关工作

### 2.1 视频异常检测

随着深度学习的发展,基于深度学习的视频异常检测成为了当下的研究热点。通常可将常见的视频异常检测方法根据应用场景的不同分为基于无监督学习和基于弱监督学习的视频异常检测方法,前者用于训练数据中缺乏异常样本而仅含有正常视频的场景,后者用于含有异常视频但是缺乏帧级别标签的场景。主流的无监督视频异常检测算法通常采用自编码器(Autoencoder,AE)模型或生成对抗网络(Generative Adversarial Networks,GAN)[20]来拟合正常视频的分布,通过重构或预测的方式来判断当前帧是否为异常[21]。

近年来,弱监督异常检测算法在不同领域展现出了优异的性能[10-11]。在视频异常检测领域,弱监督算法利用具有视频级别标签的异常视频,相比无监督视频异常检测算法展现出了更优越的性能。Sultani 等[12]第一次提出基于多示例学习的视频异常检测方法,将视频视作包,将视频中的片段视为示例,然后通过排序损失函数自动学习一个排序模型来为异常视频帧预测更高的异常分数。为了利用视频上下文的信息增强特征的表达,许多方法采用了时序特征融合[22]的方式,如 Zhong 等[14]提出一种图卷积网络(Graph Convolutional Network,GCN)来捕获不同样本间的特征相似性和时序一致性,基于分类器的异常分数建立高置信度片段和低置信度片段之间的关系。Wu 等[23]提出一个包含 3 个并行分支的神经网络来捕获视频片段之间的不同关系并进行特征的聚合。Purwanto 等[15]使用拓展的时间关系网络(Temporal Relational Network,TRN)[24]从视频中提取多尺度特征并进行多尺度分区和内积运算。然而,这些方法只能记录视频内部短期的信息,不能有效地对视频长期的正常和异常模态进行建模。且由于基于时序特征融合的视频异常检测方法融合了视频的全局信息,利用了未来帧的特征,因此在实际应用时,其不能用于在线检测,获得的异常结果存在滞后性。

### 2.2 记忆网络

记忆网络[25]指具有存储功能的神经网络,可以被读取和写入,在训练时记录数据中长期的特征,在推理时与长期记忆成分相结合得到最终的结果。记忆网络最初用于文本问答场景,可以利用记忆项来保存数据集中的场景信息,从而实现长期的记忆存储。但是该模型需要给网络的每一层赋予监督信号,因此难以基于反向传播算法进行训练。为了解决这个问题,Sukhbaatar 等[26]提出了一个连续性的记忆网络。该网络可以端到端(End-To-End)进行训练,从而适用于更多的任务。记忆网络目前在视频异常检测中的应用仅限于无监督场景,Gong 等[17]推出记忆增强自编码器,来解决自编码器对异常事件也具有较强重构能力的问题。给定一张输入图片,该方法不会直接将其编码输入解码器,而是将其作为查询项来检索记忆模块中最相关的记忆项,然后将这些记忆项进行聚合并送入解码器进行图像的重构。Park 等[18]认为原型特征不足以表示正常数据的多种模态。为此,他们提出了用于无监督异常检测的记忆模块,其中记忆模块中的单个记忆项对应正常模态的原型特征;并提出特征紧密损失和特征分离损失对特征和记忆项进行约束。然而,仅具有单一的记忆模块不足以记录视频中包含的隐藏信息,并且帧级别标签的缺失导致不能准确记录视频帧的类别信息,因此基于记忆网络的视频异常检测方法难以应用于弱监督场景。同时,已有的记忆网络中记忆项的数量都由人工设定并在训练时保持不变,记录的信息有限,不能动态适应不同的视频异常检测场景。

## 3 基于双重动态记忆网络的弱监督视频异常检测

本文提出的基于双重动态记忆网络的弱监督视频异常检测方法的整体框架如图 1 所示,包含特征提取、样本重采样、记忆网络和损失函数等。首先,模型的输入为一个正常视频及一个异常视频,采用预训练好的三维卷积网络 I3D[27]分别提取正常和异常视频中每个片段的特征。然后通过一个预分类器获得每个片段的异常分数,根据异常分数进行重采样得到置信样本对应的特征。接着将采样得到的置信特征送入记忆网络,记忆网络包含正常和异常记忆模块。之后,对记忆网络进行读写操作,得到增强后的特征并动态更新记忆网络中

的记忆项，基于增强特征采用另一个分类器获得增强的异常分数。最后，在训练阶段采用分类损失函数以及模态分离损失函数进行模型学习。在推理阶段，将测试样本对应的增强异常分数作为最终的异常分类结果。

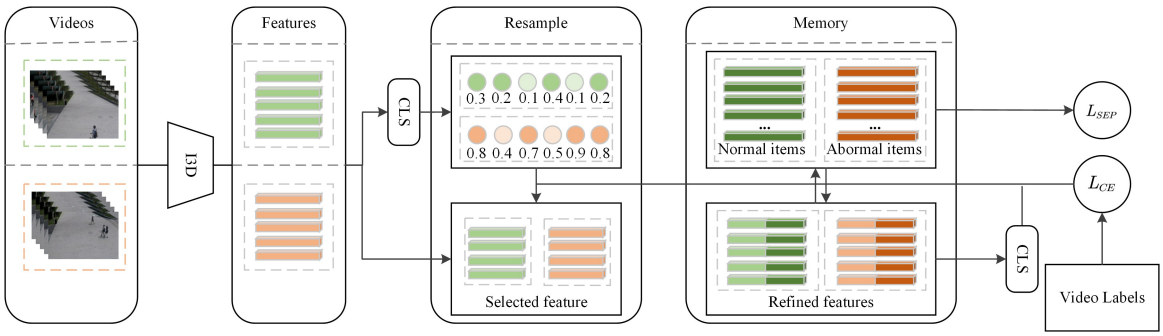



图 1 所提方法的整体框架图

Fig. 1 Overll framwork of the proposed method

### 3.1 特征提取

在弱监督的设定下，训练时有 $M$ 个训练视频 $D=\{V_i\}_1^M$，同时有每个训练视频的标签 $Y=\{y_i\}_1^M$。其中 $y_i=0$ 表示视频 $V_i$ 是一个正常视频，即所有视频帧都是正常的；$y_i=1$ 表示视频 $V_i$ 是一个异常视频，即这段视频中至少含有一小段异常片段。每轮训练同时输入一个正常视频和一个异常视频，将每个视频分割为许多不重叠的小片段，每个小片段包含 16 个视频帧，分别记作 $I_{i,t}^N$ 和 $I_{i,t}^A$，其中下标 $i$ 和 $t$ 代表第 $i$ 个视频中的第 $t$ 个小片段，上标 $N$ 和 $A$ 分别代表该视频为正常视频和异常视频。对每个视频采样得到固定数目的 $T$ 个小片段，输入到三维卷积 I3D 网络 $f_{\mathrm{I3D}}$ 中得到每个小片段的特征。

$$\boldsymbol{X}_{i,t}^{*}=f_{\mathrm{I3D}}(I_{i,t}^{*};\boldsymbol{\phi}_{\mathrm{I3D}}) \tag{1}$$

其中，$*\in\{N,A\}$ 代表正常和异常，$\boldsymbol{\phi}_{\mathrm{I3D}}$ 代表 I3D 网络的模型参数。$\boldsymbol{X}_{i,t}^{*}\in\mathbb{R}^{2048}$ 融合了包含 16 帧的小片段短期内的时空信息，相比 2D 卷积提取的特征具有更强的表达能力。

### 3.2 样本重采样

基于 I3D 网络输出的正常视频特征 $\boldsymbol{X}_i^N=\{\boldsymbol{X}_{i,t}^N\}_{t=1}^T$ 和异常视频特征 $\boldsymbol{X}_i^A=\{\boldsymbol{X}_{i,t}^A\}_{t=1}^T$，构建由全连接层和 Sigmoid 激活函数组成的分类网络 $f_{\mathrm{cls}}$，分别输出正常视频和异常视频中每个小片段对应的 0～1 之间的异常分数，记作 $S_i^N=\{S_{i,t}^N\}_{t=1}^T$ 和 $S_i^A=\{S_{i,t}^A\}_{t=1}^T$。

$$S_{i,t}^{*}=f_{\mathrm{cls}}(\boldsymbol{X}_{i,t}^{*};\boldsymbol{\phi}_{\mathrm{cls}}) \tag{2}$$

其中，$*\in\{N,A\}$，$\boldsymbol{\phi}_{\mathrm{cls}}$ 表示分类网络 $f_{\mathrm{cls}}$ 的参数，异常分数越接近于 1 代表该小段视频越可能为异常视频。

由于在训练时无法获得异常视频内部准确的帧级别的标签，因此无法将所有的视频片段用于训练。基于多示例学习的思想，将一个视频视作一个包，将视频中的多个小片段视作包中的多个实例。异常视频中含有至少一个异常小片段，本文称之为正包；正常视频中全部都是正常小片段，称作负包。多示例学习旨在使正包中的异常小片段与负包中的正常小片段的异常分数差距尽可能大，这一般通过迭代优化的方式来实现，采用分类器的分类结果构建实例的标签，再重新训练分类器。我们观察到分类器的分类结果可以反映样本被分类准确的置信程度，因此根据异常视频中小片段的异常分数 $S_i^A$ 挑选出 $T_A$ 个异常分数最接近于 1 的样本，作为很可能为异常的置信正样本特征集合 $F_i^A$。

$$F_i^A=\{\boldsymbol{X}_{i,t}^A\mid t\in\arg\min_t(|S_{i,t}^A-1|;T_A)\} \tag{3}$$

其中，$\arg\min_t(*;k)$ 代表让表达式 $*$ 最小的 $k$ 个值所对应的索引。通过采样得到异常分数最高的样本，可以筛选出异常视频中的置信异常样本作为模型的正样本，从而避免由于异常视频中潜在的正常片段而出现大量错误标签。

我们发现，当正常样本数量和异常样本数量个数达到均衡时，最终模型的异常分类性能最佳。因此，为了平衡正负样本的数量即异常片段和正常片段的数量，我们同时对正常视频进行重采样操作。与异常视频中采样得到尽可能置信的样本不同，由于正常视频中的标签是绝对准确的，即所有视频帧都为正常，为了避免过多的容易区分的负样本主导模型训练，我们根据正常视频的异常分数 $S_i^N$，从正常视频中采样得到 $T_N$ 个难以分类的样本，即异常分数接近于 1 的负样本特征集合 $F_i^N$。

$$F_i^N=\{\boldsymbol{X}_{i,t}^N\mid t\in\arg\min_t(|S_{i,t}^N-1|;T_N)\} \tag{4}$$

虽然对于正常视频和异常视频都是采样得到异常分数接近于 1 的样本，但是对于异常视频是采样得到更可能为异常的置信样本作为正样本，而对于正常视频是采样得到更难以区分的困难样本作为负样本来增强模型的拟合能力。

### 3.3 双重动态记忆网络

基于筛选出的正样本集合 $F_i^A$ 和负样本集合 $F_i^N$，设计双重动态记忆网络来记录所有视频中出现的事件的长期模态，并以此来加强样本特征的表达。与现有的只记录单一模态的记忆网络不同，所提网络同时包含正常记忆模块和异常记忆模块来分别记录正常和异常事件的模态，对正常和异常事件进行更细致的建模，从而能够更好地对其进行区分。每个记忆模块都由记忆项组成，记忆项是从样本集合 $F_i^*$ 中挑选的具有代表性模态的样本。在训练初期，记忆模块中包含的模态较少，表征能力较弱，每个难以被表示的新样本都很有可能作为新的模态被加入到记忆项中，所以记忆项的存储会快速增加。在达到一定的训练轮数后，记忆模块已经具有较好的表示能力，新的样本难以加入到记忆模块中，记忆项数量趋于稳定，这样的动态变化可以更好地适应当前检测的场景。

#### 3.3.1 记忆网络读取

记忆网络读取操作示意图如图 2 所示。设 $\boldsymbol{X}_{i,t}^{*}\in\mathbb{R}^{2048}$ 为正样本集合 $F_i^A$ 和负样本集合 $F_i^N$ 中的任一样本特征，其中 $*\in\{N,A\}$ 指明该样本是正样本或者负样本。每个记忆模块 $P^*$ 包含 $K$ 个记忆项 $\{\boldsymbol{P}_i^*\}_{i=1}^K$，其中 $*\in\{N,A\}$，$\boldsymbol{P}_i^*\in\mathbb{R}^{2048}$ 与

样本特征 $\boldsymbol{X}_{i,t}^{*}$ 维度相同，$K$ 的取值会随着训练过程发生变化。在训练过程中，正样本特征 $\boldsymbol{X}_{i,t}^{\mathrm{A}}$ 只会与异常记忆模块 $P^{\mathrm{A}}$ 进行交互，同理，负样本特征 $\boldsymbol{X}_{i,t}^{\mathrm{N}}$ 只会与正常记忆模块 $P^{\mathrm{N}}$ 进行交互。

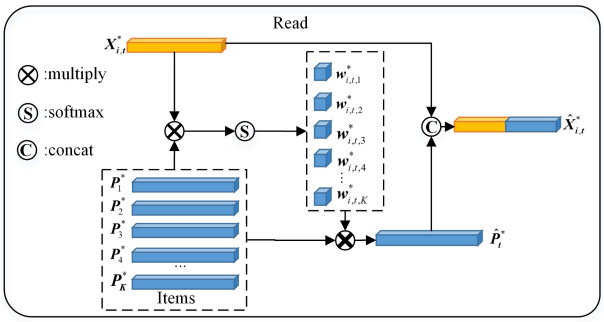



图 2 记忆网络读取操作示意图

Fig. 2 Read operation of memory network

记忆模块中的每个记忆项代表正常或者异常事件的模态，通过读取记忆项的内容，可以将样本特征 $\boldsymbol{X}_{i,t}^{*}$ 与相应事件的记忆项进行融合，进一步加强特征对相应事件的表达程度，从而更好地判断事件是否异常。特征融合已经被广泛应用于视频分类领域以改善特征[28]。为了对记忆项进行读取，对于样本特征 $\boldsymbol{X}_{i,t}^{*}$，首先计算 $\boldsymbol{X}_{i,t}^{*}$ 与对应记忆模块 $\boldsymbol{P}^{*}$ 中的记忆项 $\boldsymbol{P}_{k}^{*}$ 的相似性。

$$s_{i,t,k}^{*}=\frac{\boldsymbol{X}_{i,t}^{*}\cdot(\boldsymbol{P}_{k}^{*})^{\mathrm{T}}}{\|\boldsymbol{X}_{i,t}^{*}\|*\|\boldsymbol{P}_{k}^{*}\|} \tag{5}$$

其中，$s_{i,t,k}^{*}$ 代表样本特征 $\boldsymbol{X}_{i,t}^{*}$ 与记忆项 $\boldsymbol{P}_{k}^{*}$ 的余弦相似度，$\|*\|$ 代表向量的二范数，$(\boldsymbol{P}_{k}^{*})^{\mathrm{T}}$ 代表 $\boldsymbol{P}_{k}^{*}$ 的转置。对 $\boldsymbol{X}_{i,t}^{*}$ 和所有记忆项 $\{\boldsymbol{P}_{k}^{*}\}_{k=1}^{K}$ 的相似度 $\{s_{i,t,k}^{*}\}_{k=1}^{K}$ 进行 Softmax 操作得到 $\boldsymbol{X}_{i,t}^{*}$ 对于记忆项 $\boldsymbol{P}_{k}^{*}$ 的权重系数。

$$w_{i,t,k}^{*}=\frac{\exp(s_{i,t,k}^{*})}{\sum_{k=1}^{K}\exp(s_{i,t,k}^{*})} \tag{6}$$

权重系数可以反映当前样本特征 $\boldsymbol{X}_{i,t}^{*}$ 与记忆模块记录的记忆项的匹配程度。由于每个记忆模块只会记录单一的模态，通过查找样本特征所对应的记忆项进行加权组合，得到新的样本特征，可以更准确地反映样本所属的模态。新特征的计算式如下：

$$\hat{\boldsymbol{P}}_{t}^{*}=\sum_{k=1}^{K}w_{i,t,k}^{*}\boldsymbol{P}_{k}^{*} \tag{7}$$

为了避免记忆项的数目过多导致模态混淆，采用阈值筛选的方式去除权重系数过低的样本。具体为，设定阈值 $\lambda$，将权重系数低于阈值的项置为 0 得到新的权重系数，如式(8)所示：

$$\hat{w}_{i,t,k}^{*}=\begin{cases}w_{i,t,k}^{*}, & \text{if } w_{i,t,k}^{*}>\lambda\\ 0, & \text{otherwise}\end{cases} \tag{8}$$

通过对权重系数进行稀疏化操作，可避免过多的无关记忆项的干扰。将新的权重系数 $\hat{w}_{i,t,k}^{*}$ 采用式(6)所示的 Softmax 操作进行重归一化，$\hat{\boldsymbol{w}}_{i,t}^{*}$ 代表样本特征 $\boldsymbol{X}_{i,t}^{*}$ 对于所有记忆项的权重向量。将归一化后的权重系数 $\hat{\boldsymbol{w}}_{i,t}^{*}$ 应用于式(7)即可得到改进后的新特征 $\hat{\boldsymbol{P}}_{t}^{*}$。

最后，为了保留原来样本特征 $\boldsymbol{X}_{i,t}^{*}$ 的信息，将 $\boldsymbol{X}_{i,t}^{*}$ 与新特征 $\hat{\boldsymbol{P}}_{t}^{*}$ 进行拼接得到最终的样本增强特征 $\hat{\boldsymbol{X}}_{i,t}^{*}$。将样本增强特征输入到另一个分类器 $f_{\mathrm{cls}}'$ 中得到增强的异常分数 $S_{i,t}'^{*}$：

$$S_{i,t}'^{*}=f_{\mathrm{cls}}'(\hat{\boldsymbol{X}}_{i,t}^{*};\boldsymbol{\phi}_{\mathrm{cls}}') \tag{9}$$

其中 $\boldsymbol{\phi}_{\mathrm{cls}}'$ 代表该分类器的参数。由于增强特征由原有特征和新特征拼接而成，因此分类器 $f_{\mathrm{cls}}'$ 的输入维度为第一个分类器 $f_{\mathrm{cls}}$ 输入维度的两倍，输出仍为 0～1 之间的异常分数值。

### 3.3.2 记忆网络写入

除了对样本特征进行改进外，还需要在训练的过程中动态更新记忆模块内部的记忆项，这是通过记忆网络写入操作完成的。记忆网络写入操作示意图如图 3 所示。

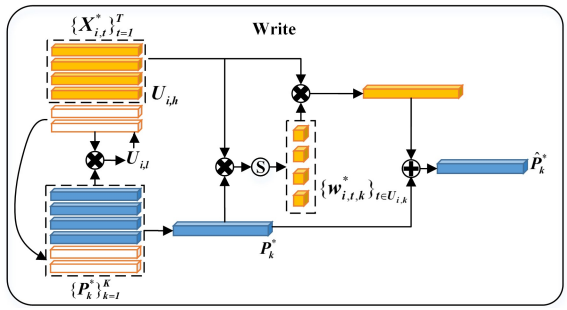



图 3 记忆网络写入操作示意图

Fig. 3 Write operation of memory network

记忆项的更新有两个步骤。首先，对于样本集合 $F_{i}^{*}$ 中的任一样本 $\boldsymbol{X}_{i,t}^{*}$，根据式(5)计算样本与记忆模块中所有记忆项的相似度 $\{s_{i,t,k}^{*}\}_{k=1}^{K}$，取相似性的最大值作为样本 $\boldsymbol{X}_{i,t}^{*}$ 的最大相似度 $v_{i,t}^{*}=\max(\{s_{i,t,k}^{*}\}_{k=1}^{K})$。设定相似度阈值 $\tau$，将样本集合分为 $U_{i,l}=\{t\in\arg_{t}(v_{i,t}^{*}\leqslant\tau)\}$ 和 $U_{i,h}=\{t\in\arg_{t}(v_{i,t}^{*}>\tau)\}$ 两部分，其中 $t\in\arg_{t}(*)$ 代表符合 $*$ 表达式对应的索引，$U_{i,l}$ 代表所有最大相似度低于阈值的样本索引。这些样本由于与所有记忆项的相似度都较低，可以理解为当前没有记忆项可以代表这些样本所蕴含的模态，因此直接将这些样本作为新的记忆项直接加入到记忆模块中，使得记忆模块中的记忆项能够涵盖所有的模态。$U_{i,h}$ 代表所有最大相似度高于阈值的样本索引，说明存在记忆项可以表征这些样本的模态，所以不重复添加这些样本到记忆模块中，而是用于记忆项的更新。同时，数据集中的异常样本越多，可能包含的模态也越多，能够添加到记忆模块的动态记忆项的数量也会相应地增加，记忆模块就具有更强的的特征表示能力。

对于待更新的记忆项 $\boldsymbol{P}_{k}^{*}$，从 $U_{i,h}$ 中选取样本集合 $U_{i,k}$，$U_{i,k}$ 中的样本 $\boldsymbol{X}_{i,t}^{*}$ 与记忆项 $\boldsymbol{P}_{k}^{*}$ 的相似度在相似度集合 $\{s_{i,t,k}^{*}\}_{k=1}^{K}$ 中最大，即这些样本认为 $\boldsymbol{P}_{k}^{*}$ 与自己最相似，因此可以用 $U_{i,k}$ 中的样本特征来更新 $\boldsymbol{P}_{k}^{*}$：

$$\hat{\boldsymbol{P}}_{k}^{*}=\boldsymbol{P}_{k}^{*}+\sum_{t\in U_{i,k}}w_{i,t,k}^{*}\boldsymbol{X}_{i,t}^{*} \tag{10}$$

其中，$w_{i,t,k}^{*}$ 的计算参照式(5)和式(6)，在样本特征的维度进行 Softmax 操作。

$$w_{i,t,k}^{*}=\frac{\exp\left(\dfrac{\boldsymbol{P}_{k}^{*}*(\boldsymbol{X}_{i,t}^{*})^{\mathrm{T}}}{\|\boldsymbol{P}_{k}^{*}\|*\|\boldsymbol{X}_{i,t}^{*}\|}\right)}{\sum_{t\in U_{i,k}}\exp\left(\dfrac{\boldsymbol{P}_{k}^{*}*(\boldsymbol{X}_{i,t}^{*})^{\mathrm{T}}}{\|\boldsymbol{P}_{k}^{*}\|*\|\boldsymbol{X}_{i,t}^{*}\|}\right)} \tag{11}$$

通过对样本特征进行加权组合并与原来的记忆项相加，可以在保留原有记忆项内容的同时根据新的样本特征对其进行动态更新，从而使得该记忆项更具有代表性。

## 3.4 算法训练及推理

算法训练的损失函数包括分类损失 $L_{\mathrm{CE}}$ 以及模态分离损失 $L_{\mathrm{SEP}}$。

$$L_{\text{total}}=\lambda_{\text{CE}}L_{\text{CE}}+\lambda_{\text{SEP}}L_{\text{SEP}} \tag{12}$$

通过权重系数$\lambda_{\text{CE}}$和$\lambda_{\text{SEP}}$控制两者的比重。

3.4.1 分类损失

基于式(2)得到的异常分数$S_{i,t}^{*}$和基于式(9)得到的增强异常分数$S_{i,t}^{\prime *}$，将正常视频中困难样本的标签设为0，异常视频中置信样本的标签设为1，分别构建二元交叉熵分类损失函数，进行相加即得到综合的分类损失$L_{\text{CE}}$。

3.4.2 模态分离损失

为了增强记忆模块中的记忆项之间的区分度，使其能更好地表示不同的模态，分别构建了记忆模块内部的模态分离损失$L_{\text{SEP}}^{\text{intra}}$和记忆模块间的模态分离损失$L_{\text{SEP}}^{\text{inter}}$。为了使记忆模块内部的记忆项尽可能分离，在记忆模块内部计算每个记忆项与其他记忆项的相似度，使得相似度尽可能小，由此得到记忆模块内部的模态分类损失$L_{\text{SEP}}^{\text{intra}}$。

首先计算$\boldsymbol{X}_{i,t}^{*}$与对应记忆模块$\boldsymbol{P}^{*}$中的记忆项$P_{k}^{*}$的相似性：

$$L_{\text{SEP}}^{\text{intra}}=\sum_{i=1}^{N}\sum_{k}\sum_{k'\neq k}\frac{\boldsymbol{P}_{k}^{*}*(\boldsymbol{P}_{k'}^{*})^{\mathrm{T}}}{\|\boldsymbol{P}_{k}^{*}\|*\|\boldsymbol{P}_{k'}^{*}\|} \tag{13}$$

同时，为了使正常记忆项与异常记忆项尽可能分离，在正常记忆模块与异常记忆模块间计算正常记忆项与异常记忆项的相似度，使得该相似度同样尽可能小，得到记忆模块间的模态分离损失$L_{\text{SEP}}^{\text{inter}}$：

$$L_{\text{SEP}}^{\text{inter}}=\sum_{i=1}^{N}\sum_{k}\sum_{k'}\frac{\boldsymbol{P}_{k}^{N}*(\boldsymbol{P}_{k'}^{A})^{T}}{\|\boldsymbol{P}_{k}^{N}\|*\|\boldsymbol{P}_{k'}^{A}\|} \tag{14}$$

综合的模态分离损失为两者之和：

$$L_{\text{SEP}}=\lambda^{\text{intra}}L_{\text{SEP}}^{\text{intra}}+\lambda^{\text{inter}}L_{\text{SEP}}^{\text{inter}} \tag{15}$$

此处通过权重系数$\lambda^{\text{intra}}$和$\lambda^{\text{inter}}$控制两者的影响程度。因为我们希望记忆模块间的相似度更小，所以$\lambda^{\text{inter}}$的值设置为大于$\lambda^{\text{intra}}$，从而使得记忆模块间的损失函数占据主导地位。

在推理时，采用增强异常分数$S_{i,t}^{\prime}$作为最终的异常检测结果。

## 4 实验结果与分析

### 4.1 数据集

为了验证所提方法的性能，在两个公开的弱监督视频异常检测数据集ShangHaiTech[29]和UCF-Crime[12]上进行了实验验证。ShangHaiTech数据集中共有238个训练视频和199个测试视频，其中训练视频包含175个正常视频和63个异常视频，测试视频包含155个测试视频和44个异常视频，测试视频由130个不同的异常事件组成，如车辆、滑板、跳跃和跑步等。UCF-Crime的训练集包含1 610个视频，其中含有正常视频800个，异常视频810个；测试集包含290个视频，其中含有正常视频150个，异常视频140个。该数据集涵盖了13个现实世界中的犯罪行为，包括偷盗、逮捕、纵火等。

### 4.2 实验设定

每次训练从每个视频中采样32个视频片段，每个片段包含16帧。如果视频的长度少于32个视频片段，则进行重复采样补满至32帧。用于特征提取的I3D网络仅采用RGB分支，采用3D版本的ResNet50作为主干网络，在Kinetics数据集[27]上进行预训练。样本的采样个数$T_A=10$，$T_N=10$，记忆网络中的初始正常记忆项和异常记忆项的个数均设置为10。权重系数阈值$\lambda=0.02$，相似度阈值$\tau$在ShangHaiTech数据集上设为2，在UCF-Crime数据集上设为2.5。损失函数中分类损失的系数$\lambda_{\text{CE}}=1$，模态分离损失系数$\lambda_{\text{SEP}}=0.1$。模态分离损失中模态内损失函数系数$\lambda^{\text{intra}}=0.1$，模态间损失函数系数$\lambda^{\text{inter}}=1$。整体模型由Adam优化器训练，学习率为$1\times10^{4}$，权重衰减为$5\times10^{-4}$。采用Batch Size为10在2080TI GPU上训练100轮次。

### 4.3 评估标准

本文遵循弱监督视频异常检测通用的评估设定，采用接收者操作特征(Receiver Operating Characteristic)曲线下的面积(Area Under the Curve，AUC) AUC-ROC来评估模型在ShangHaiTech和UCF-Crime数据集上的性能。AUC-ROC越大，表示分类器的性能越好。实际计算时，可将分类器输出的异常分数设置为阈值，得到阶梯形区域即可计算面积。

### 4.4 超参数分析

为了更好地对算法进行说明，本节对文中涉及的超参数进行实验论证分析，主要包括正负样本采样个数$T_A$和$T_N$、权重系数阈值$\lambda$、相似度阈值$\tau$，以及损失函数权重系数$\lambda_{\text{CE}}$，$\lambda_{\text{SEP}}$，$\lambda^{\text{intra}}$，$\lambda^{\text{inter}}$。实验结果如表1－表5所列。

表1 UCF-Crime数据集上采样个数超参数实验结果

Table 1 Experimental results of hyperparameter sampling on UCF-Crime dataset

| 采样个数($T_A=T_N$) | 设定 | *AUC-ROC*/% |
|---|---|---|
| 6 | 在线应用 | 82.86 |
| 8 | 在线应用 | 83.07 |
| **10** | 在线应用 | **83.15** |
| 12 | 在线应用 | 83.11 |
| 14 | 在线应用 | 82.98 |

表2 UCF-Crime数据集上权重系数阈值超参数实验

Table 2 Experimental results of weight coefficient threshold hyperparameters on UCF-Crime dataset

| 权重系数阈值$\lambda$ | 设定 | *AUC-ROC*/% |
|---|---|---|
| 0.01 | 在线应用 | 83.13 |
| **0.02** | 在线应用 | **83.15** |
| 0.03 | 在线应用 | 82.85 |
| 0.04 | 在线应用 | 82.66 |

表3 UCF-Crime数据集上相似度阈值超参数实验结果

Table 3 xperimental results of similarity threshold hyperparameters on UCF-Crime dataset

| 相似度阈值$\tau$ | 设定 | *AUC-ROC*/% |
|---|---|---|
| 1.5 | 在线应用 | 83.14 |
| 2.0 | 在线应用 | 83.08 |
| **2.5** | 在线应用 | **83.15** |
| 3.0 | 在线应用 | 82.78 |
| 3.5 | 在线应用 | 82.45 |

表4 UCF-Crime数据集上总损失函数权重系数超参数实验结果

Table 4 Experimental results of total Loss function weight coefficient hyperparameter on UCF-Crime dataset

| 分类损失系数$\lambda_{\text{CE}}$ | 模态分离损失系数$\lambda_{\text{SEP}}$ | 设定 | *AUC-ROC*/% |
|---|---|---|---|
| 1 | 0.05 | 在线应用 | 83.05 |
| 1 | **0.10** | 在线应用 | **83.15** |
| 1 | 0.15 | 在线应用 | 83.04 |
| 1 | 0.20 | 在线应用 | 82.88 |

表 5 UCF-Crime 数据集上模态分离损失函数权重系数超参数实验结果

Table 5 Experimental results of modal separation loss weight coefficient hyperparameter on UCF-Crime dataset

| 模态内损失函数系数$\lambda^{intra}$ | 模态间损失函数系数$\lambda^{inter}$ | 设定 | *AUC-ROC*/% |
|---|---|---|---|
| 0.05 | 1 | 在线应用 | 83.12 |
| 0.10 | **1** | 在线应用 | **83.15** |
| 0.15 | 1 | 在线应用 | 83.09 |
| 0.20 | 1 | 在线应用 | 83.09 |

对正负样本采样个数$T_A$和$T_N$的超参数实验如表 1 所列，当采样个数过少时记忆模块的表示能力不足，当采样个数过多则会导致信息混淆和冗余，说明采样个数过多或过少对模型检测精度都会产生不利影响。对权重系数阈值$\lambda$的超参数实验如表 2 所列，$\lambda$用于剔除与样本相关系数不高的模态，当$\lambda$过低时，样本与模态的相关系数较高，但是能够用来表示样本的记忆项会减少；当$\lambda$过高时，容易造成不同模态混叠，两者都会在一定程度上影响检测精度。对相似度阈值$\tau$的超参数实验如表 3 所列，$\tau$用于控制记忆项数量，$\tau$过大时样本$\boldsymbol{X}_{i,t}^{*}$难以加入到记忆模块中，记忆模块中包含的模态就过少；$\tau$过小也会造成模态冗余。针对损失函数权重系数$\lambda_{CE}$和$\lambda_{SEP}$以及权重系数$\lambda^{intra}$和$\lambda^{inter}$的超参数实验分别如表 4 和表 5 所列，损失函数权重系数用于控制不同损失函数在总体损失函数中的比重，比重越大的损失函数在参数优化过程中的作用越强。从实验结果中可以看出，只有选取合适的损失函数权重才能取得较好的整体模型检测精度。

### 4.5 实验结果

本节将所提方法与经典的弱监督视频异常检测方法进行对比，同时也比较了目前基于时序特征融合的弱监督视频异常检测方法。

在 ShangHaiTech 数据集上，基于时序特征融合的方法除了 MLEP 模型以外都只能离线应用。从表 6 所列的评估结果中可以看出，能够在线应用的方法的 AUC-ROC 都较低，所提方法相比 Sultani 等提出的在线应用方法的检测精度的提升了 10%以上。同时，与只能离线应用的方法相比，所提方法的检测精度高于部分方法。相比精度最高的方法 RTFM，所提方法的检测精度只降低了 0.1%。

表 6 在 ShangHaiTech 数据集上实验对比结果

Table 6 Experimental results comparison on ShangHaiTech dataset

| 模型 | 设定 | *AUC-ROC*/% |
|---|---|---|
| MLEP[13] | 在线应用 | 76.80 |
| Sultani 等[12] | 在线应用 | 86.30 |
| AR-Net[30] | 在线应用 | 91.24 |
| MIST[31] | 在线应用 | 94.83 |
| GCN-Anomaly[14] | 离线应用 | 84.44 |
| Purwanto 等[15] | 离线应用 | 96.85 |
| RTFM[16] | 离线应用 | 97.21 |
| Ours | 在线应用 | **97.10** |

从表 7 的结果可以看出，在 UCF-Crime 数据集上，本文方法检测精度高于所有在线应用的方法，与目前最优的用于在线检测的方法相比检测精度增加了 4.15%。与离线应用的方法相比，本文方法的检测精度超过了部分离线检测方法。相比目前精度最高的 Purwanto 等提出的采用更为复杂特征的方法，所提方法精度下降不到 2%。

表 7 在 UCF-Crime 数据集上实验对比结果

Table 7 Experimental results comparison on UCF-Crime dataset

| 模型 | 设定 | *AUC-ROC*/% |
|---|---|---|
| Sultani 等[12] | 在线应用 | 75.41 |
| TCN-IBL[32] | 在线应用 | 78.66 |
| Motion-Aware[33] | 在线应用 | 79.00 |
| GCN-Anomaly[14] | 离线应用 | 82.12 |
| Wu 等[23] | 离线应用 | 82.44 |
| RTFM[16] | 离线应用 | 84.03 |
| Purwanto 等[15] | 离线应用 | 85.00 |
| Ours | 在线应用 | **83.15** |

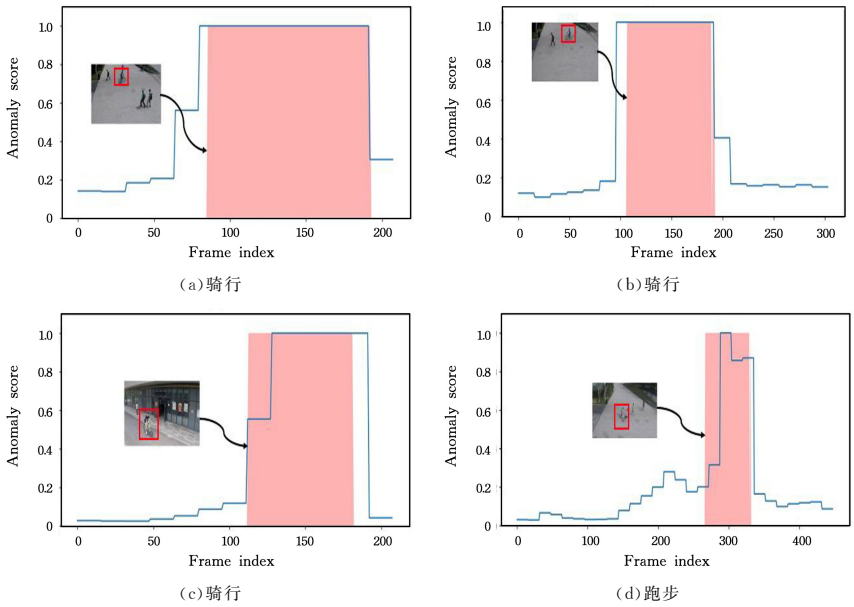


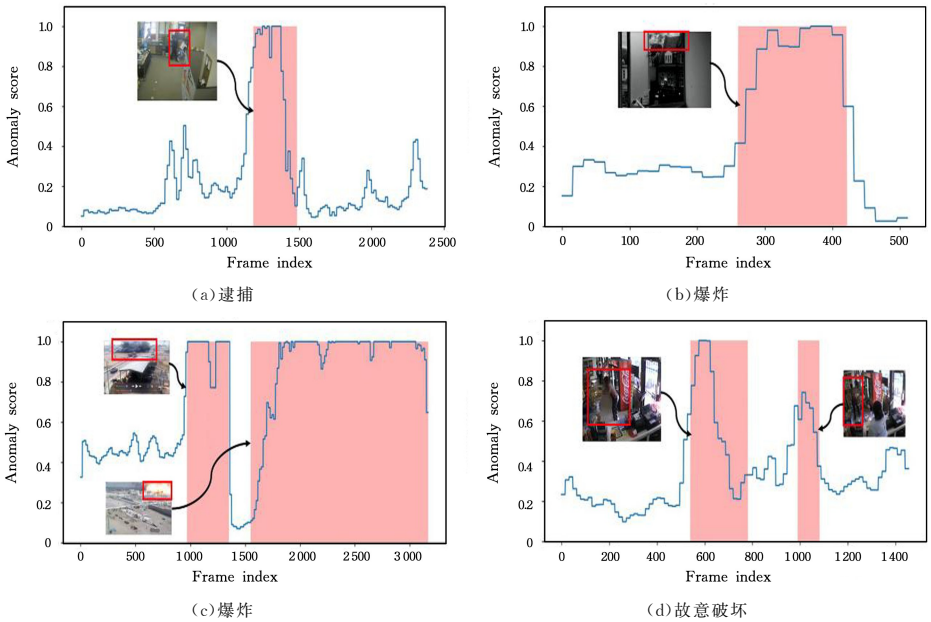

图 4 ShanghaiTech 的可视化结果(电子版为彩图)

Fig. 4 Visualization results of ShanghaiTech

在两个公开数据集上的实验结果表明，本文提出的可在线应用的模型不仅优于目前所有的在线应用模型，且相比只能离线应用的模型，所提出的模型在精度上仍然十分有竞争力。

### 4.6 可视化结果

所提方法在 ShangHaiTech 和 UCF-Crime 数据集上的可视化结果分别如图 4 和图 5 所示。图中的异常分数越高表明出现异常的可能性越大，红色区域代表真实标签下出现异常的视频帧的范围，视频帧中出现异常的区域用红色框标明。从图中可以看出，本文方法在不同的数据集上都能准确地捕捉视频中出现的异常，并且对于不同的异常类型如骑行、逮捕、爆炸等都能够有效进行检测。当一段视频中出现多个异常时，如图 5(c)和图 5(d)所示，所提方法也能够准确地检测出不同时间段的异常。这说明所提方法对于不同的视频异常检测场景都具备有效性。



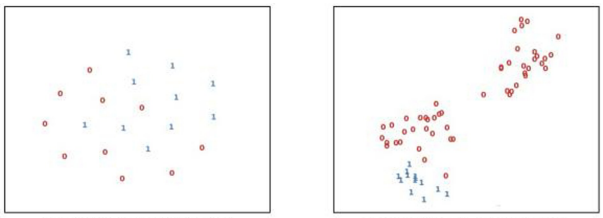

(a)逮捕 (b)爆炸

(c)爆炸 (d)故意破坏

图 5 UCF-Crime 的可视化结果(电子版为彩图)

Fig. 5 Visualization results of UCF-Crime

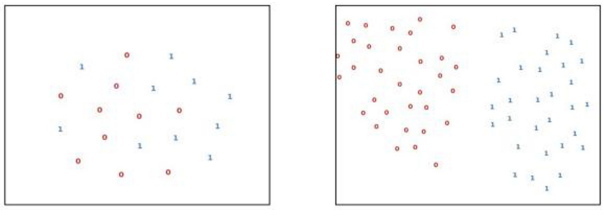
为了进一步说明所提记忆网络的有效性，在 ShangHaiTech和 UCF-Crime 数据集上对记忆网络中的记忆项采用 t-SNE[34] 进行降维，在二维平面进行可视化，分别如图 6 和图 7 所示。其中红色的 0 代表正常记忆项，蓝色的 1 代表异常记忆项。



(a)初始记忆项特征降维图 (b)训练后记忆项特征降维图

图 6 ShangHaiTech 记忆项特征降维图(电子版为彩图)

Fig. 6 Diagram of memory item feature dimension reduction of ShangHaiTech



(a)初始记忆项特征降维图 (b)训练后记忆项特征降维图

图 7 UCF-Crime 记忆项特征降维图(电子版为彩图)

Fig. 7 Diagram of memory item feature dimension reduction of UCF-Crime

从图中可以看出，训练前的初始记忆项是随机初始化的，导致正常和异常的记忆项杂糅在一起。而在训练后，通过记忆网络的更新操作能够使得正常记忆项和异常记忆项区分开来，并且基于所提出的动态增加记忆项个数的策略，记忆项的数量能够根据不同数据集的类型改变，从而更好地适应该数据集。

### 4.7 消融实验

为进一步说明所提方法中各个组件的重要性，在 ShangHaiTech(简称为 SHT)和 UCF-Crime(简称为 UCF)数据集上对所采用的记忆网络中的各个组件进行消融实验，结果如表 8 所列。其中重采用网络是对样本进行一个粗分类，记忆网络指目前仅含有一个记忆模块的记忆网络；双重指在原有记忆网络的基础上，增加本文所提出的同时具有正常和异常记忆模块的双重记忆模块；动态指在更新记忆网络时采用本文所提出的动态更新记忆项数量的策略。

表 8 消融实验结果

Table 8 Ablation experiment results

(%)

| 重采样网络 | 记忆网络 | 双重 | 动态 | *SHT* | *UCF* |
|---|---|---|---|---|---|
| × | × | × | × | 71.35 | 60.23 |
| √ | × | × | × | 86.13 | 75.84 |
| √ | √ | × | × | 93.27 | 80.21 |
| √ | × | √ | × | 95.74 | 82.36 |
| √ | × | √ | √ | 97.10 | 83.15 |

从表 8 中可以看出，在实验设定的基准之上(Shang-

HaiTech-86.13%,UCF-Crime-75.84%)去掉重采样模块后,模型的检测精度大幅度下降,在 ShangHaiTech 数据集上的下降幅度超过 14%,在 UCF-Crime 数据集上也有 15%以上的衰减,这说明模型在重采样阶段已经取得了较好的分类效果,能够为后续的记忆网络训练提供较为准确的正常和异常样本集合,如果去掉重采样模块,记忆网络的训练效果会受到较大的影响。加入单一的记忆模块可以使模型的检测精度明显提升,在 ShangHaiTech 数据集上的提升超过 7%,在 UCF-Crime 数据集上也有 4%以上的提升,说明采用记忆模块记录视频的长期模态有助于异常的检测。而将一般的单一记忆模块更改为所提出的双重记忆模块后,在单一记忆模块的基础上检测精度也有了明显的提升,两个数据集的提升均在 2%以上,说明了采用双重记忆网络分别纪录正常视频和异常视频长期模态的必要性。在双重记忆网络的基础上增加动态的记忆项即可动态调整记忆项的数量,模型的检测精度得到了进一步的提升,在两个数据集上的综合提升达 1%左右,说明动态调整记忆项的数量有助于适应不同的异常检测场景,从而提升检测精度。

**结束语** 本文提出了一种基于双重动态记忆网络的弱监督视频异常检测方法,对正常视频和异常视频分别构造了记忆项动态可变的记忆网络,从而能够记录视频中长期的正常和异常模态,并通过记忆网络的读写操作来对特征进行动态更新。为了能够动态适应不同的监控场景,模型在训练时自适应调整记忆项的数目,并采用模态分离损失增加各个记忆项之间的区分度。在测试时,不需要融合未来帧的特征从而实现在线检测。在两个公开数据集上与现有方法的对比结果验证了所提方法的有效性。在未来工作中,将考虑更精细化的样本重采样策略,使得在训练时能够利用视频中更多的样本,从而达到更好的检测效果。

## 参 考 文 献


[1] HUANG T, WU K J, WANG D C, et al. Video Anomaly Detection Based on Improved Time Segmentation Network[J]. Computer Engineering, 2022, 48(11): 137-144.

[2] FENG L J, ZHAO C H. Transfer increment for generalized zero-shot learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(6): 2506-2520.

[3] FENG L J, ZHAO C H, LI X. Bias-Eliminated Semantic Refinement for Any-Shot Learning[J]. IEEE Transactions on Image Processing, 2022, 31: 2229-2244.

[4] 储岳中,乔雨楠. 多注意力结合光流的视频超分辨方法[J]. 重庆工商大学学报(自然科学版), 2022, 39(4): 1-8.

[5] 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. 2017: 1933-1941.

[6] WANG X Z, CHE Z P, JIANG B, et al. Robust unsupervised video anomaly detection by multipath frame prediction [J]. arXiv: 2011.02763, 2021.

[7] ZHOU W H, LI Y X, ZHAO C H. Object-Guided and Motion-Refined Attention Network for Video Anomaly Detection[C]// 2022 IEEE International Conference on Multimedia and Expo (ICME). 2022: 1-6.

[8] LIU W, LUO W X, LIAN D Z, 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.

[9] YE M C, PENG X J, GAN W H, et al. Anopcn: Video anomaly detection via deep predictive coding network[C]// Proceedings of the 27th ACM International Conference on Multimedia. 2019: 1805-1813.

[10] CHAI Z, ZHAO C H, HUANG B. Multisource-refined transfer network for industrial fault diagnosis under domain and category inconsistencies[J]. IEEE Transactions on Cybernetics, 2021, 52(9): 9784-9796.

[11] SONG P Y, ZHAO C H. Slow down to go better: A survey on slow feature analysis[J]. IEEE Transactions on Neural Networks and Learning Systems, Early Access.

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

[13] LIU W, LUO W X, LI Z X, et al. Margin Learning Embedded Prediction for Video Anomaly Detection with A Few Anomalies [C]// IJCAI. 2019: 3023-3030.

[14] ZHONG J X, LI N N, KONG W J, et al. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1237-1246.

[15] PURWANTO D, CHEN Y T, FANG W H. Dance with self-attention: A new look of conditional random fields on anomaly detection in videos[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 173-183.

[16] TIAN Y, PANG G S, CHEN Y H, 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.

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

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

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

[20] GOODFELLOW I, POUGET-ABADIE J , MIRZA M, et al. Generative adversarial nets[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. 2014: 2672-2680.

[21] AKCAY S，ATAPOUR-ABARGHOUEI A，BRECKON T P. Ganomaly：Semi-supervised anomaly detection via adversarial training[C]// Asian Conference on Computer Vision. 2018：622-637.

[22] HU H，GU J Y，ZHANG Z，et al. Relation networks for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：3588-3597.

[23] 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. 2020：322-339.

[24] ZHOU B，ANDONIAN A，OLIVA A，et al. Temporal relational reasoning in videos[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018：803-818.

[25] WESTON J，CHOPRA S，BORDES A. Memory networks[J]. arXiv：1410. 3916，2014.

[26] SUKHBAATAR S，WESTON J，FERGUS R. End-to-end memory networks[J]. Advances in Neural Information Processing Systems，2015，28：1-9.

[27] CARREIRA J，ZISSERMAN A，QUO V. action recognition? a new model and the kinetics dataset[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：6299-6308.

[28] GIRDHAR R，CARREIRA J，DOERSCH C，et al. Video action transformer network[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：244-253.

[29] LUO W X，LIU W，GAO S H. 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.

[30] WAN B，FANG Y M，XIA X，et al. Weakly supervised video anomaly detection via center-guided discriminative learning [C]// 2020 IEEE International Conference on Multimedia and Expo(ICME). 2020：1-6.

[31] FENG J C，HONG F T，ZHENG W S. Mist：Multiple instance self-training framework for video anomaly detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：14009-14018.

[32] ZHANG J G，QING L Y，MIAO J. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection[C]// 2019 IEEE International Conference on Image Processing(ICIP). 2019：4030-4034.

[33] ZHU Y，NEWSAM S. Motion-aware feature for improved video anomaly detection[J]. arXiv：1907. 10211，2019.

[34] VAN DER MAATEN L，HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research，2008，9(11)：2579-2605.

**ZHOU Wenhao**，born in 1998，master. His main research interest is video and image anomaly detection.

**ZHAO Chunhui**，born in 1979，Ph. D，professor. Her main research interests include statistical machine learning and data mining for industrial application.

（责任编辑：何杨）