计算机科学
COMPUTER SCIENCE

伪异常选择驱动学习的视频异常检测

赵松，傅豪，王洪星

引用本文

赵松，傅豪，王洪星.伪异常选择驱动学习的视频异常检测[J]．计算机科学, 2023, 50(5): 146-154.

ZHAO Song, FU Hao, WANG Hongxing. Pseudo-abnormal Sample Selection for Video Anomaly Detection [J]. Computer Science, 2023, 50(5): 146-154.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220400227

伪异常选择驱动学习的视频异常检测

赵 松 傅 豪 王洪星
重庆大学信息物理社会可信服务计算教育部重点实验室 重庆 400044
重庆大学大数据与软件学院 重庆 400044
(zhaosong@cqu.edu.cn)

摘 要 无监督视频异常检测方法通常使用正常的监控视频数据通过帧重构/帧预测方法来训练视频异常检测模型。然而,正常视频中往往包含大量的相似画面和背景帧,数据集冗余的问题尤为明显,因此不能高效地进行异常检测模型训练。针对该问题,提出了伪异常选择驱动学习的视频异常检测方法,从原始视频训练数据中迭代选取部分异常分数高的正常视频帧(伪异常帧)来构建新的训练池,用于学习和优化视频异常检测模型。在检测模型方面,设计了基于后继帧预测的双路 U-Net 骨干网络,以不同采样率的视频段分别作为两个支路的输入,从而从多个粒度上更好地提取和利用视频的时空特征。为了加强典型训练数据对帧预测任务和异常检测的影响,双路 U-Net 中设计了多层的记忆学习模块。在常用视频异常检测数据集上进行实验,验证了所提方法在检测精度和训练效率上的有效性。

关键词:视频监控;异常检测;样本选择;记忆模型;时空特征
中图法分类号 TP391.4

Pseudo-abnormal Sample Selection for Video Anomaly Detection

ZHAO Song,FU Hao and WANG Hongxing
Key Laboratory of Dependable Service Computing in Cyber Physical Society(Chongqing University),Ministry of Education,Chongqing 400044,China
School of Big Data & Software Engineering,Chongqing University,Chongqing 400044,China

Abstract Unsupervised video anomaly detection methods generally use normal video data to train an anomaly detection model through frame reconstruction or frame prediction. However,normal videos usually contain a large number of background frames and similar scenes,which are quite redundant,leading to inefficient modeling for video anomaly detection. To address this issue,this paper proposes a pseudo-abnormal sample selection method,which iteratively selects video frames with high abnormal scores from original videos to build a new concise training pool for video anomaly detection based on future frame prediction. As for the detection model,this paper designs a two-path U-Net architecture,where each path has a different sampling frequency on video frames so that spatial-temporal features of videos can be better extracted and utilized from multiple scales. In the two-path U-Net,each layer shares a memory module to strengthen the impact of typical training data for future frame prediction and video anomaly detection. Experimental evaluation on benchmark video datasets demonstrates the efficiency and effectiveness of the proposed method.

Keywords Video surveillance,Anomaly detection,Sample selection,Memory model,Spatial-Temporal feature

1 引言

视频异常检测的目标是在监控数据中发现不寻常或不符合预期的视频帧[1],这些视频帧即为异常帧。视频异常检测技术能够高效智能地监控异常事件,现已成为学术界[2]和工业界[3]的研究热点。然而,现实监控视频中的异常事件罕有发生,且异常的标准难以定义。收集真实场景下的所有异常画面是不现实的,这带来了异常样本缺乏的问题。因此,现有的无监督视频异常检测方法通常仅使用正常样本训练模型,根据视频帧的重构[4-5]或预测[6]误差来判别异常。这样的正常监控视频往往由大量相似画面和背景帧构成,数据冗余的问题尤为明显。以这样的全部样本来训练网络模型需要花费大量的时间,而且冗余的训练集可能导致样本空间不平衡,出现所得模型对某些样本的"过学习"现象[7]。

到稿日期:2022-04-22 返修日期:2022-09-12
基金项目:国家自然科学基金(61976029);重庆市技术创新与发展应用专项重点项目(cstc2021jscx-gksbX0033)
This work was supported by the National Natural Science Foundation of China(61976029) and Key Project of Chongqing Technology Innovation and Application Development(cstc2021jscx-gksbX0033).
通信作者:王洪星(ihxwang@cqu.edu.cn)

视频帧预测的方法关注目标帧与其期望（预测帧）之间的误差大小，更符合异常行为不符合预期发生的设定，被视为有效的解决方案。当前的基于帧重构和帧预测的方法也存在一定局限性，现有工作往往仅关注视频帧外观上的预测/重构任务或需要引入额外预训练模型来提取动作信息[4-6,8]，不能高效提取视频中的时空特征。受 SlowFast 模型[9]中快慢分支不同帧率输入的启发，本文设计了双路 U-Net 后继帧预测网络模型，由快、慢两个支路构成。其中慢支路输入稀疏帧，快支路输入连续帧，快、慢支路均能较好地捕获视频中的时空特征，且能通过跃层连接操作更好地从多个粒度上捕获视频的时空特征，在不需要额外信息输入的情况下，取得了较好的视频异常检测性能。

样本选择方法能够从原始样本集合中挑选部分典型样本形成新的训练集，从而缓解样本中的数据冗余问题。监控视频中，正确分类那些容易被误判的片段正是提高模型检测能力的关键。图 1(b)给出了常见视频异常检测流程，首先直接使用原始训练集进行训练，然后评估测试集。相比图 1 中的(b)，图 1 中的(a)为本文方法，引入了伪异常选择策略。每轮训练前，所提方法从原始训练集选取异常分数高的视频帧加入训练池，以此训练池训练模型，本文将被选视频帧称为伪异常帧。伪异常帧是正常视频中具有高异常分数的视频帧，伪异常选择策略能够过滤部分明显正常的视频帧，缓解训练数据冗余的问题。此外，所选伪异常帧构成的训练池中，多为人物重叠、人物进出监控画面等易被误判为异常的视频帧。基于此训练池优化模型，能提高模型的对类似画面的预测能力，减小误判概率。如图 1 所示，本文方法对于正常帧的异常分数普遍低于 0.5，评估的准确度更高，使正常、异常帧之间的区别更大，从而展示出了更好的检测效果。

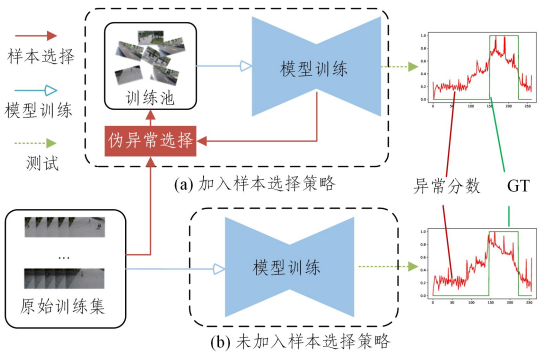


图 1　基于伪异常样本选择的视频异常检测

Fig. 1　Video anomaly detection with pseudo-abnormal sample selection

本文的主要贡献如下：

(1) 提出了伪异常选择驱动学习的视频异常检测方法，根据选择策略迭代地从原始训练集中选取训练数据加入训练池，并以此训练池来优化模型。

(2) 提出了双路 U-Net 后继帧预测网络，通过其两个支路不同帧采样率的输入和侧向连接操作，双路 U-Net 能够直接端到端地提取视频的时空特征，不需要输入光流等额外运动信息。

(3) 在所提双路 U-Net 中设计了多层的记忆机制。本文在每一层侧向连接中引入了一个共同的记忆模块，通过对典型时空特征的记忆，来加强训练数据中的典型行为模式产生的影响。

(4)实验结果表明，所提双路 U-Net 帧预测网络结合多层记忆机制，能够加强正常行为模式产生的影响，抑制异常画面的预测能力，从而提高正常画面与异常画面的区分度。本文的伪异常选择策略具有普适性，能够在保持较好检测性能的情况下，通过训练样本选择方法减少 80% 以上的训练耗时。

2　相关工作

2.1　视频异常检测

近年来涌现出了大量的视频异常检测方法[4,10-11]，它们往往通过自编码器的帧重构/帧预测误差来训练模型。Hasan 等[4]通过 3D 卷积建模视频中的时空特征，利用重构误差鉴别异常。Leng 等[10]通过多个视角的重构分数来判别异常。U-Net[12]作为优秀的自编码器，在视频异常检测领域被广泛使用。Liu 等[6]使用 U-Net 作为骨干网络，提出了基于帧预测的方法，通过历史帧序列预测目标帧，避免了直接重构目标帧的问题。该方法使用光流来加强运动特征，带来了额外的计算量消耗。Park 等[11]在 U-Net 中设计了记忆模块，强化了正常模式的影响。Fang 等[13]以 U-Net 为骨干网络，通过未来帧和历史帧的一致性来建模异常。然而，U-Net 源于图像任务，视频时空特征编码能力不足，在 ShanghaiTech 数据集上与未以 U-Net 作为骨干网络的方法存在较大差距。这些方法往往对视频时空特征进行了专门的设计。Xu 等[14]设计了双路的自编码器，包含一条单独的 RGB 支路和一条光流支路，将两个支路的特征融合后进行任务重构。Yu 等[15]将外观和运动特征作为互补的线索来优化模型。Chang 等[16]设计了双路的时空自编码器网络结构，以编码紧凑时空特征。Hao 等[17]在自编码器的基础上，设计了时空一致性的鉴别器，将时空一致性作为异常分数来鉴别异常。Georgescu 等[18]使用目标检测器来编码人物行为，通过多任务学习方法，设置了 4 个代理任务来应对 4 种类型的异常行为，在视频异常检测任务中取得了最好的效果。本文在 U-Net 结构的基础上，提出了双路 U-Net 后继帧预测网络模型，使用两个具有不同帧率输入的分支来编码视频特征，并通过侧向连接操作融合了不同粒度上的时空特征，展示出了更好的表现。

2.2　记忆模型

循环神经网络（Recurrent Neural Network，RNN）[19]和长短期记忆网络（Long Short-Term Memory，LSTM）[20]是用于捕获序列数据中长期依赖关系的常用方法，两者都通过隐藏状态来选择性地记录历史信息并将其作为记忆。但由于两者的结构限制，所含隐藏状态的大小十分有限，它们往往无法准确地记录顺序数据中的所有内容。为打破这一限制，最近的工作[21-22]关注于在端到端的网络中引入一个外部的记忆模块。这样专门的记忆库和记忆网络可以更好地执行记忆的读写操作[22-26]，表现出记忆功能。Gong 等[26]将记忆学习方法用于自编码器中，在特征空间引入了额外的记忆库来保存有价值的正常模式，以在测试阶段增强正常画面的重构能力。

本文扩展了记忆模块在异常检测方面的应用，在所提的双路U-Net后继帧预测网络中，本文将快、慢分支的每一层输出拼接后输入与之对应的记忆模块，以进行记忆学习。多层的记忆模块能够对正常模式进行多层次的记忆，增强正常画面的预测能力，抑制异常画面的预测效果，从而提高正常画面与异常画面的区分度。

2.3 训练样本的选择

训练样本在神经网络的学习中占有重要地位，训练样本的质量直接影响着模型的表现。训练样本集是否具有代表性将决定模型的学习效果。训练样本选择的目标是从原始训练集 D 中尽量多地删除冗余样本和噪声样本，然后得到一个具有代表性的压缩集 D_S，要求该压缩集能够保持或改善模型性能。根据数据选择策略和过程的不同，样本选择方法主要可以分为以下3种：基于抽样的方法、基于聚类的方法和基于近邻分类规则的方法[27]。文献[28]根据网络的训练情况动态地向训练集加入特定样本。Wang等[29]根据样本密度的大小选择更接近支持向量的数据训练SVM[30]。Tang等[31]提出了用于分类问题的根据样本相关性进行选择的算法。在视频异常检测领域，也有工作着力于解决训练数据冗余的问题。Cong等[32]通过字典选择方法，从训练数据中选择最优子样本集作为字典，通过字典对测试画面的重构代价来检测异常。借鉴这些思想，本文设计了伪异常样本选择框架，在每轮训练前，迭代地选取异常分数top-k的伪异常帧并将其加入训练池，并以该样本池训练模型。所选训练池中，多为人物重叠、人物进出监控画面等易被误判异常的视频帧。基于此训练池优化模型，能提高模型的对类似画面的预测能力，减小误判概率。

3 本文算法

伪异常选择驱动的视频异常检测训练框架如图2所示。为缓解视频异常检测训练集中的冗余问题，本文提出了伪异常选择学习方法。每次训练前，先从训练集选取异常分数top-k的帧加入训练池。为有效地提取视频中物体的时空特征，本文设计了双路U-Net后继帧预测网络，分别记为慢分支和快分支。其中快分支输入长为 T 的连续帧，慢分支输入长为 T/τ、间隔为 τ 采样的视频帧，所提模型解码器部分与U-Net解码器结构一致。两个分支由于输入帧的时间速率不同，所包含的时空特征也有差异。为了丰富特征中不同粒度的时空信息，双路U-Net每一层通过侧向连接将慢分支和快分支编码的特征进行融合。由于记忆模块能够强化训练数据产生的影响，本文在侧向连接中引入了记忆模块来强化正常视频中的典型特征。记忆模块中（图2中 MP_l，$l\in\{1,2,3,4\}$），将快、慢分支的输出进行拼接，得到两个分支融合后的特征 x_l，然后根据 x_l 读取并更新记忆库中的典型正常模式，得到更新后的特征 $\hat{x}_l$。

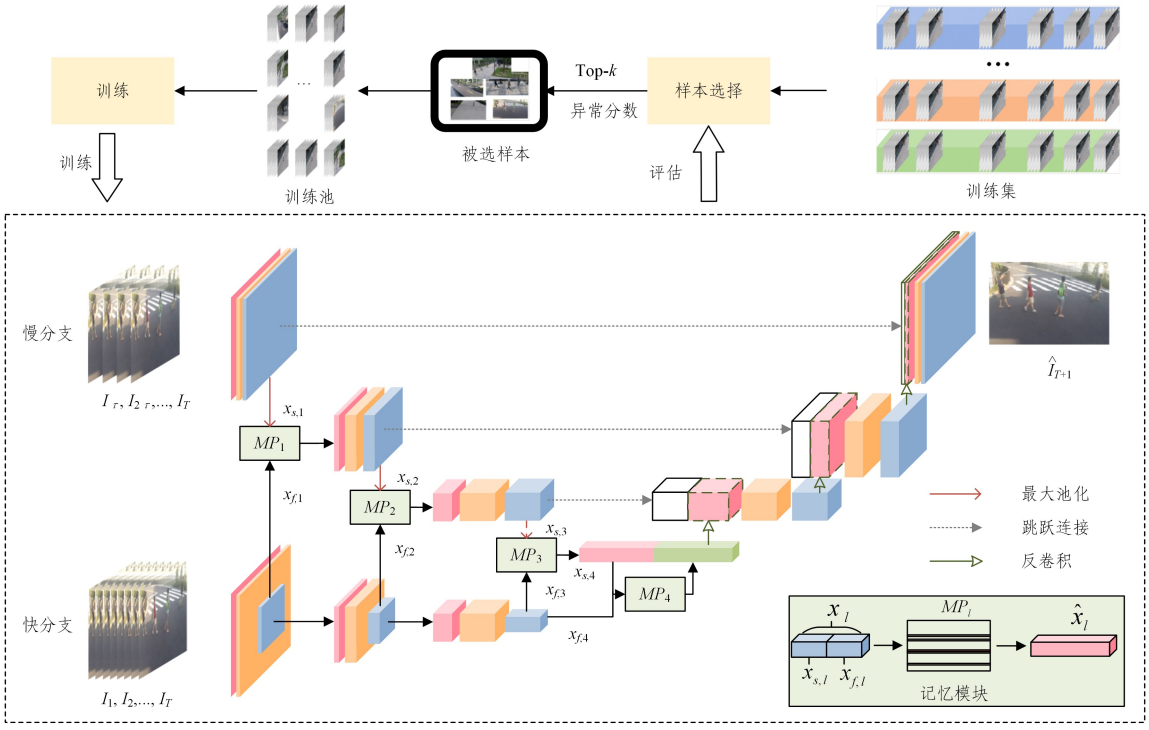


图2 伪异常选择驱动学习的视频异常检测训练框架

Fig. 2 Training framework of pseudo-abnormal sample selection for video anomaly detection

3.1 双路U-Net后继帧预测网络

由于U-Net[12]模型具有优秀的图片编码重构能力，其在视频异常检测领域也被广泛使用[6,11,25]。监控视频的相邻帧之间的空间轮廓信息（如背景和行人的外观）变化相对较小，而运动信息（如行走和跑步动作）的差异相对较大，因此仅仅提取视频中物体的空间特征远远不够。运动特征能够辅助判断一个行为是否异常，因此本文在U-Net的基础上设计了一个双路的编码器，旨在利用快、慢分支增强视频时空特征编码。其中慢分支的编码器的输入为间隔 τ 采样的视频帧，结构上与U-Net保持一致，网络参数为（Conv1 × 2^2-BN-ReLU +Conv1×2^2-BN-ReLU）。快分支的编码器输入为连续的 T 帧，与慢分支相比，其主要目的在于发掘更长输入

序列内的时空信息，以从多个粒度上捕获视频的时空特征。为使快分支与慢分支最大池化后的特征通道数相同，以便进行拼接操作，快分支的编码器设计了一个额外的 Conv1×3^2 卷积操作。

双路网络中，侧向连接常被用来融合两个分支中的信息[9]，该操作能在视频特征提取过程中合并不同粒度上的时空特征。如图 2 所示，本文在每一层对应的快分支和慢分支的编码器之间设计了一个横向连接。慢分支的输出保留了 U-Net 中的最大池化操作，使其与快分支输出的通道数相同，以便于拼接操作。双路后继帧预测网络中的层数共有 4 层（$L=4$），本文将第 $l\in\{1,\cdots,L\}$ 层拼接后的特征记为 x_l，由记忆模块更新后得到 $\hat{x}_l$，并将其作为下一层慢分支的输入。模型目标为对未来帧 I_{T+1} 进行预测，预测结果记为 $\hat{I}_{T+1}$。为保证 I_{T+1} 与 $\hat{I}_{T+1}$ 的一致性，本文通过最小化两者的 ℓ_2 距离来实现，损失 $\mathbb{R}_{pred}$ 如式(1)所示：

$$\mathscr{L}_{\text{pred}}=\| I_{T+1}-\hat{I}_{T+1}\|_2 \tag{1}$$

3.2　多记忆模块

在双路 U-Net 中，两个分支的输入帧率不同，且两个分支在结构上也有区别，编码后各自所提取的时空特征也有差异。本文在双路 U-Net 的每一层都使用了一个横向连接，将快分支和慢分支的特征进行融合，所得特征包含了两个分支中不同粒度的时空特征。每一层融合后的特征记为 $x_l\in\mathbb{R}^{H\times W\times C}$，将 $H\times W$ 记为 K。记忆模型可以通过读写其记忆库来强化特定样本产生的影响。因此，本文在双路 U-Net 结构的每一层都引入记忆模块，以强化正常样本中的时空特征。通过对不同层次特征的记忆学习，所提方法能够在预测时加强正常行为模式的影响，抑制异常画面的预测能力，提高正常/异常画面的区分度。本文所提多层记忆模块记为 $M_l\in\mathbb{R}^{m\times C}$，$l\in\{1,\cdots,L\}$ 代表对应层中正常视频帧的典型特征。记忆模块包含两个操作：记忆库读取和记忆库更新。

记忆库读取模块如图 3 所示，输入的 x_l 经过记忆模块读取记忆项后，得到更新调节后的输出 $\hat{x}_l^k$。具体而言，每次输入 $x_l\in\mathbb{R}^{H\times W\times C}$ 被展开为 K 个大小为 $\mathbb{R}^{1\times 1\times C}$ 的查询，查询项表示为 x_l^k。查询时，先将 x_l^k 与记忆库中的每一个记忆项 M_l^j 计算余弦相似度，再由 Softmax 操作得到匹配概率 $w_l^{j,k}$，其计算式如式(2)所示：

$$w_l^{j,k}=\frac{\exp(x_l^k\,(M_l^j)^{\mathrm{T}})}{\sum\limits_{i\in m}\exp(x_l^k\,(M_l^i)^{\mathrm{T}})} \tag{2}$$

记忆读取时，查询项 x_l^k 与特定记忆项匹配概率越高，$w_l^{j,k}$ 就越大，在更新后的 $\hat{x}_l^k$ 中的占比也就越大。本文按照式(3)进行记忆聚合操作，得到更新后的特征 $\hat{x}_l^k$。通过 M_l 中全部的记忆项来更新 x_l^k，使得本文模型能够理解不同的正常模式，并抑制异常画面的预测效果。

$$\hat{x}_l^k=\sum_{j\in m}w_l^{j,k}M_l^j \tag{3}$$

记忆库需要不断训练更新，以尽可能多地保存更具代表性的样本，保证其影响力。记忆库中的每一个记忆项 M_l^j 都会参与查询项 x_l^k 的更新。为保证记忆项查询项 x_l^k 只与最相似的 M_l^j 保持较大的匹配概率 $w_l^{j,k}$，而与其余记忆项保持较低的匹配概率，本文将 M_l^j 作为 x_l^k 的正匹配对象，其余记忆项作为负匹配对象，通过对比学习优化记忆模块。对比学习方法保证了查询项只与某一记忆项最相关，而与其他记忆项相似度较低，从而保证记忆项是稀疏的。记忆模块对比学习损失函数如式(4)所示：

$$\mathscr{L}_{\text{mem}}=\sum_{l=1}^{L}\sum_{j=1}^{m}\left[-\log\frac{\exp(x_l^k\,(M_l^j)^{\mathrm{T}})}{\sum\limits_{i=1}^{m}\exp(x_l^k\,(M_l^i)^{\mathrm{T}})}\right] \tag{4}$$

本文通过预测误差和记忆模块损失来分别优化网络模型和记忆库参数，如式(5)所示：

$$\mathscr{L}_{\text{total}}=\mathscr{L}_{\text{pred}}+\mathscr{L}_{\text{mem}} \tag{5}$$

其中，$\mathscr{L}_{\text{pred}}$ 通过预测误差优化双路 U-Net 网络参数，根据 $\mathscr{L}_{\text{mem}}$ 优化记忆模块参数。

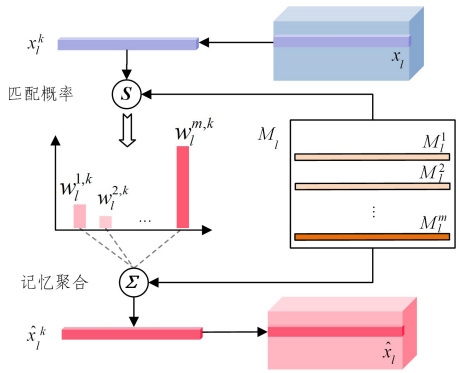


图 3　记忆读取模块

Fig. 3　Memory reading module

3.3　伪异常选择

预测帧的质量是衡量异常程度的标准[6]，式(6)中峰值信噪比（Peak Signal-to-Noise Ratio，PSNR）是一种被广泛用于衡量图像质量的方法。本文也使用 PSNR 来评估预测帧的质量，将 I_t 的 PSNR 值记为 $P(\hat{I}_t,I_t)$。对应地，式(6)中，$\hat{I}_t$ 表示预测帧，I_t 表示真实帧。$P(\hat{I}_t,I_t)$ 值越低，表示视频帧预测的质量越差，该视频帧为异常帧的可能性就越大。

$$P(\hat{I}_t,I_t)=10\log_{10}\frac{\max(\hat{I}_t)}{\|\hat{I}_t-I_t\|_2^2/N} \tag{6}$$

其中，N 表示视频帧中的像素点个数。本文按照式(7)将 $P(\hat{I}_t,I_t)$ 归一化到[0,1]，并以此表示各帧的异常值。其中具有更高的异常分数 $S(t)$ 的视频帧更有可能是异常帧。

$$S(t)=\frac{\max_t P(I_t,\hat{I}_t)-P(I_t,\hat{I}_t)}{\max_t P(I_t,\hat{I}_t)-\min_t P(I_t,\hat{I}_t)} \tag{7}$$

本文所提伪异常选择驱动学习的视频异常检测模型的训练流程如算法 1 所示。

算法 1　伪异常选择驱动学习

输入：Training set D；Selection budget D_{budget}，Initial model Θ；Selection size k

输出：Model Θ；Memory pool M

1. Randomly initialize training pool D_S from training set D by an initial

portion p;
2. $D=D\backslash D_S$;
3. While $D_S < D_{budget}$:
4. Calculate loss $\mathcal{L}_{total}$ on D_S by Eq. (5);
5. Update Θ and M by $\mathcal{L}_{total}$;
6. Calculate S(t) in D on Θ by Eq. (7);
7. $D_T=\{D_t, t\in \text{top-k}(S(t))\}$;
8. $D_S=D_S\cup D_T$, $D=D\backslash D_T$;
9. end

为使异常检测模型的训练数据更有意义，降低训练数据的冗余性，本文在每轮模型迭代前，尽可能地从训练集中选择画面、人物复杂的伪异常视频帧加入训练池。通过对训练池内这些伪异常视频片段的学习，模型在评估时能够提高对类似正常画面的预测能力，使这些画面保持较低的异常分数。所提方法首先用 2% 的训练集随机初始化训练池 D_S，并以此训练池初始化模型。每轮训练完成后，进入推理阶段，对待选择样本进行评估。按照式(7)的评估结果，将训练集中异常分数 top-k（具体为训练集大小的 1%）的伪异常样本集合 D_T 加入训练池 D_S，并在 D_S 上继续优化所提双路 U-Net 网络模型，如此迭代，直到 D_S 与训练集 D 的占比达到预设样本选择比例 D_{budget}。正如前文所述，所选取的异常分数 top-k 的伪异常样本往往是画面、人物复杂的视频帧，通过对这些视频帧的学习，提高了正常帧和异常帧的区分度，便于设定阈值来判别异常。

4 实验

4.1 数据集与实验设置

为评估所提方法(SelAD)在视频异常检测上的实验表现，本文在主流的异常检测数据集 UCSD Ped2[33]，Avenue[34] 和 ShanghaiTech[6] 上进行了验证。数据集示例如图 4 所示，其中第一行为正常示例，第二行为异常示例。

(a) Ped2 (b) Avenue (c) ShanghaiTech

图 4 数据集示例

Fig. 4 Sample of experimental datasets

(1) UCSD Ped2。Ped2 包含 16 个训练视频和 12 个测试视频，每帧分辨率为 240×360 像素。训练集包含 2550 帧，测试集包含 2010 帧，其中包含各种拥挤的场景，异常情况包括自行车、车辆、滑板、轮椅穿越行人区。

(2) Avenue。Avenue 数据集是固定相机视角捕获的走廊数据集，包含 16 个训练和 21 个测试视频片段，包含 15328 个训练帧，15324 个测试帧。训练视频包含正常情况下的视频，测试视频包含标准和异常事件视频，其中检测的主要是行人的各种异常，如异常动作(跑步)、错误移动方向和异常物体(如自行车)等。

(3) ShanghaiTech。ShanghaiTech(SHTech)是异常检测现有基准中最大的数据集，包含 130 个异常事件和逾 270000 个训练帧。该数据集包含 13 个场景，这些场景具有复杂的光照条件和摄像机角度。此外，ShanghaiTech 数据集中引入了由突然运动引起的异常，如追逐和争吵，这些异常与实际场景更相符。

本文以接受者操作特性曲线(Receiver Operating Characteristic curve, ROC)为评测标准，曲线下面积(Area Under Curve, AUC)表示 ROC 曲线面积。AUC 越高表明模型性能越好，该算法根据阈值来检测异常是否发生。实验参数设置如下：记忆库的维度 m 设置为 10，batch_size 设置为 12，学习率设置为 2×10^{-4}。本文在 SHTech 数据集上将每一轮的 epochs 设置为 1，Avenue 和 Ped2 上的 epochs 设置为 5，训练的最后一轮 epochs 设为原来的 2 倍，D_{budget} 设置为占原始训练集的 20%。训练时，使用训练集原始大小 2% 的样本初始化训练池，然后在此训练池中进行一轮训练。每轮训练完成后，随机从训练集中选取 30% 的样本进行评估，选择其中异常分数 top-k(训练集原始大小的 1%)的伪异常样本加入训练池。如此迭代至训练池占原始训练集大小的 20% 时，训练完成。

4.2 实验结果比较

4.2.1 伪异常选择

本文从运行效率(训练耗时)和准确率(AUC)两个角度出发，评估了所提伪异常样本选择方法的实际表现。表 1 列出了在 3 个视频异常检测数据集上，本文模型设置不同选择样本比例时的结果(本文所注样本选择百分比皆表示总选择量占原始训练集的比例，100% 表示在全训练集上训练)。

表 1 不同样本选择策略下的结果比较(AUC)

Table 1 Results comparison of different sample selection strategies(AUC)

(单位：%)

样本选择策略	Ped2	Avenue	SHTech
random-20%	93.64	85.31	70.58
top-10%	95.81	87.42	70.68
top-20%	**97.03**	**87.45**	**72.09**
100%	**97.28**	**89.16**	**72.54**

表 1 中，random-20% 选择策略表示每轮迭代从训练集随机选取 1% 的样本加入训练池，总选择量达 20% 时停止迭代。top-10% 表示随机选取 1% 的样本初始化模型，每轮迭代从训练集中选取异常分数前 1% 的伪异常样本加入训练池，并在此训练池中训练，直至选择量达 10%。top-20% 表示随机选取 2% 的样本初始化模型，迭代训练方式与 top-10% 相同，直至选择量达 20%。综合比较，使用伪异常样本选择方法能够在 3 个数据集中均取得比随机选择更好的效果，这验证了伪异常样本选择方法能够在减小训练集大小的情况下，相比随机方法更能提升检测性能。其中 top-20% 选择策略能够较好地保持在全训练集上的检测性能，这表明 top-20% 伪异常选择策略更有效。

为验证伪异常选择方法对训练耗时和性能产生的影响，

本文对 MNAD 方法[11]和所提 SelAD 方法在 Avenue 和 ShanghaiTech 数据集上进行了实验。表 2 列出了在使用/不使用伪异常选择驱动学习的情况下，两种方法的耗时与 AUC 值。样本选择比例 100%对应的 epochs 为 20，其余样本选择策略下的训练轮数为 20，且在 SHTech 上每轮训练 1 个 epoch，在 Avenue 上每轮训练 5 个 epochs，即本文样本选择策略下的总 epochs 比全训练集更多。但由于训练池比全训练集小，因此仍能够明显减少时间消耗。

表 2　伪异常样本选择策略对运行效率和检测精度的影响

Table 2　Impact of pseudo-abnormal sample selection method on training time and accuracy

方法	数据集	样本选择	耗时/h	AUC/%
MNAD[11]	Avenue	100%	22.2	88.45
	Avenue	top-20%	2.5	86.81
	SHTech	100%	35.8	70.51
	SHTech	top-20%	6.3	70.16
SelAD	Avenue	100%	17.8	89.16
	Avenue	top-20%	1.8	87.45
	SHTech	100%	34.0	72.54
	SHTech	top-20%	5.6	72.09

使用伪异常样本选择方法后，训练池缩减为原始训练集的 20%，在 Avenue 和 ShanghaiTech 数据集上的耗时分别仅为传统方法的 1/10 和 1/5，并保持了较好的检测性能。MNAD 方法[11]和本文方法的实验表明，伪异常样本选择方法具有一定的普适性，能够提高训练样本的质量，缩短训练耗时，节约计算资源。

4.2.2　模型性能

为验证本文异常检测模型的有效性，本文方法与主流的视频异常检测方法在全训练集(100%训练集)的情况下进行了实验比较。由于 Liu 等[6]，Park 等[11]，Gong 等[26]和 Fang 等[13]提出的方法均以 U-Net 为骨干网络(U-Net based)，与本文模型最为相近，是主要的比较目标。此外，本文也给出了视频异常检测领域最近的研究成果，以作为参考，这些方法未使用 U-Net 为骨干网络(Non U-net based)，它们通常会引入额外的操作(光流、特征预处理)等。表 3 列出了在全训练集训练的情况下各模型的检测精度。

表 3　全训练集下的视频异常检测结果

Table 3　Results of video anomaly detection on whole datasets

(单位：%)

模型类别	方法	Ped2	Avenue	SHTech
Non U-Net based	Hasan 等[4]	84.98	80.05	60.94
	Yu 等[15]	97.32	89.61	74.80
	Chang 等[16]	96.52	86.38	73.32
	Wang 等[35]	—	87.01	79.30
	Georgsu 等[18]	**99.82**	**92.82**	**90.18**
	Li 等[36]	95.10	88.80	73.90
	Hao 等[17]	96.91	86.62	73.79
	Zhong 等[37]	97.70	88.90	70.70
	Chang 等[38]	96.71	87.13	73.70
U-Net based	Liu 等[6]	95.43	85.11	**72.80**
	Park 等[11]	96.91	88.45	70.51
	Gong 等[26]	94.13	83.32	71.12
	Fang 等[13]	96.20	86.80	—
	SelAD(Ours)	**97.28**	**89.16**	72.54

表 3 所列的未以 U-Net 作为骨干网络(Non U-Net based)的方法中，Georgesu 等[18]使用了预训练的目标检测器，通过更精细的输入，将异常检测的目标聚焦于人的行为模式，以多任务学习方法训练模型，在现有方法中取得了最好的表现。Wang 等[35]也使用了预训练的目标检测器，使得异常检测更加关注人的行为，在 ShanghaiTech 数据集上取得了较好的表现。Chang 等[38]设计了分别关注于外观和运动特征的双路自编码器模型，通过聚类等操作强制外观编码器和运动编码器关注于主要正常模式，在 ShanghaiTech 数据集取上取得了较好的表现，但由于其 3D 卷积主干网络编码能力的局限性，在 Ped2 和 Avenue 数据集上表现稍逊。

以 U-Net 为骨干网络(Non U-Net based)的方法中，本文的双路 U-Net 两个分支能够发掘细粒度的时空特征，在 Ped2 和 Avenue 数据集上都取得了有竞争力的表现，分别为 97.28%和 89.16%。这是因为数据集 Ped2 和 Avenue 的分辨率较低，画面中的人物体积小且密集，细粒度的特征更有利于判别异常。本文方法在 U-Net 的基础上，加强了不同粒度的时空特征的影响，因此表现更佳。由于 U-Net 能通过跳跃连接减少降维带来的信息损失，能较好地反映出重构/预测误差，因此 Liu 等[6]，Park 等[11]和 Gong 等[26]提出的方法在 Ped2 和 Avenue 上的总体表现也较好。数据集 ShanghaiTech 存在突发的异常(如追逐打闹)，需要更加关注人物运动特征，因此更具挑战性。Park 等[11]仅使用记忆模块强化正常模式产生的影响，缺乏对运动特征的优化，在 ShanghaiTech 数据集上表现稍逊。Liu 等[6]引入了额外的光流信息，以额外的算力消耗为代价提升了性能。可以看到，本文所提的双路 U-Net 后继帧预测网络具有强大的时空特征编码能力，结合多层记忆学习方法强化正常模式产生的影响，在 3 个数据集上均展示出了优秀的综合表现。

4.3　消融实验

4.3.1　双路 U-Net

(1)快慢分支设计。为验证本文快、慢分支的必要性，本文分别对单分支、双分支和侧向连接下的网络设计进行了实验比较，结果如表 4 所列。

表 4　不同双路 U-Net 设计的影响

Table 4　Impact of different two-path U-Net designs

(单位：%)

Method	Ped2	Avenue	SHTech
Slow	95.68	86.72	70.65
Fast	95.81	85.80	69.23
Slow→Fast	96.87	87.46	71.62
Slow←Fast	97.28	**89.16**	**72.54**
Slow+Fast	**97.37**	89.15	72.31

Slow 表示仅使用慢分支，并在每一层引入记忆模块的结果，由于缺乏更丰富的时空特征，其表现较差。表 4 中第二行的 Fast 表示仅使用快分支，并在每一层引入记忆模块的结果，由于快分支更多地压缩了特征信息，相比单独的慢分支设计并没有效果提升。Slow→Fast 表示以快分支作为主分支，将慢分支为辅助分支，通过跃层连接融合信息，并在每一个跃层连接中加入记忆学习操作，快慢分支的融合带来了较大的

效果提升(1%左右)。类似地,Slow+Fast 表示双向的跃层连接,Slow←Fast 表示以慢分支作为主分支,将快分支作为辅助分支。实验结果表明,跃层连接操作能够从多个粒度上捕获视频的时空特征,提升模型的表现。两个分支中的特征是互补的,在 U-Net 的基础上,以慢分支为主分支、以快分支为辅助分支时的表现最佳。

(2)输入帧率设计。为确定双路 U-Net 后继帧预测网络中,两个分支输入帧率的差异对效果的影响,本文在 ShanghaiTech 数据集上设置了不同的帧率组合进行实验。图 5 给出了快、慢分支上不同帧率输入对模型的 AUC 和耗时影响。

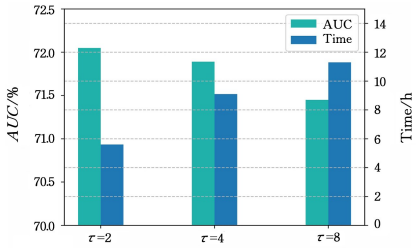


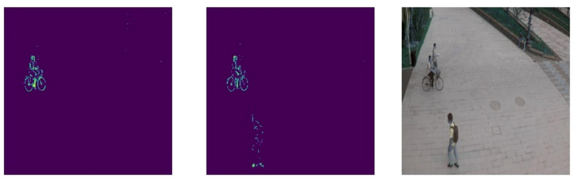
图 5 不同帧采样率的结果比较(电子版为彩图)

Fig. 5 Results comparison of different frame rates

由于现有方法[6, 11, 25]通常将帧预测输入长度设置为 4,本文也将慢分支输入固定为 4 帧。本文设置 τ 为 2,4,8 这 3 种情况,对应的快分支输入分别为连续 8,16,32 帧。图 5 中左侧绿色柱体对应模型 AUC,右侧蓝色柱体对应模型训练耗时。可以看到,所提模型的训练耗时与 τ 的取值呈正相关。这是因为将更长的视频序列输入快分支时,会使模型参数增加,带来更大的计算量。当 τ 取值较大时,会带来输入快、慢分支的视频帧差异过大的问题,不利于融合不同粒度的时空特征,因此相应的异常检测精度并没有得到提高。本文中,在慢分支输入间隔 $\tau=2$ 采样 4 帧、快分支输入连续 8 帧的情况下,双路 U-Net 的综合表现最好。

4.3.2 多记忆模块

现有方法仅对 U-Net 编码器最后一层输出进行记忆学习(对应本文 MP_4),强化正常模式的影响。这样的单记忆模块对浅层的信息记忆力不够充分。本节在 ShanghaiTech 数据集上对比了所提的多记忆模块($MP_l, l\in\{1,\cdots,L\}$)与单记忆模块的预测误差,以验证多记忆模块的有效性。为了便于展示,本文将误差图的像素值映射到[0,255]区间,并过滤了其中低于 100 的像素值,结果如图 6 所示。



(a)多记忆模块 (b)单记忆模型 (c)目标帧

图 6 多记忆模块与单一记忆模块的误差图

Fig. 6 Error maps of the proposed method using multi-memory and a single memory

图 6(a)给出了本文所提多记忆模块对其的预测误差图,图 6(b)给出了单个记忆模块对其的预测误差图,图 6(c)给出了预测目标帧。可以看到,图 6(a)中的预测误差高度集中于画面中的骑行者和自行车,图 6(b)中关于骑者和自行车的误差则较小(更暗),且图 6(b)中进入画面的正常行为也出现了部分预测错误。具体而言,在 ShanghaiTech 数据集上,两种方案的精度表现如表 5 所列。所提多记忆模块通过对多层次正常特征的记忆学习,提高了对正常行为的预测准确度,而单记忆模块对正常行为中的典型特征记忆学习的能力不足。与单记忆模块相比,所提方法能够准确地反映出画面中的正常/异常行为,降低画面中正常行为对异常分数的影响,呈现出更好的异常检测效果。

表 5 多/单记忆模块对精度的影响

Table 5 Impact of multi-memory and single memory on accuracy

(单位:%)

多/单记忆模块	SHTech
多记忆模块	72.54
单记忆模块	71.08

4.3.3 训练样本选择策略

本文方法的训练流程为:首先用训练池优化模型,每一轮训练完成后,评估待选训练集的异常分数 $S(t)$,然后选取其中异常分数最高的 k 个样本加入训练池,继续按照此流程进行模型优化和样本选择,直到样本池大小达到预设样本选择比例。由此可以看出,样本选择策略是决定模型表现的关键。为了验证伪异常选择策略的优越性,本文设置了 random-k,top-k 和 bottom-k(每次迭代选择异常分数最低的 k 个样本加入训练池)3 种选择策略。为了验证本文 top-k 训练样本选择策略对提升精度的有效性,本文在 ShanghaiTech 数据集上根据这 3 种样本选择策略进行实验,对应的 ROC 曲线如图 7 所示。

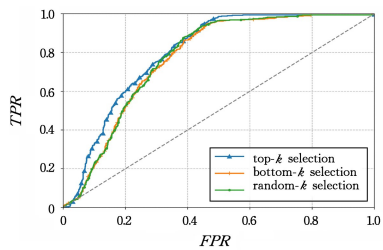


图 7 3 种选择策略下的 ROC 曲线图

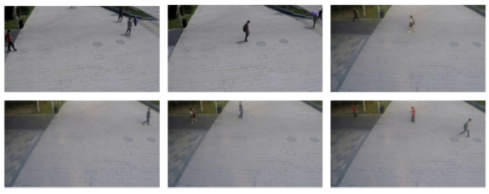
Fig. 7 ROC curves of three selection strategies

这 3 种样本选择策略中,top-k 选择策略展示出了比 random-k 和 bottom-k 更好的效果。实验结果表明,random-k 选择策略意味着随机地选择训练数据,这样可能会丢失一些对模型训练有价值的视频帧。bottom-k 选择异常分数 $S(t)$ 最低的视频帧,所得训练池由大量相似简单的、明显正常的视频帧构成。基于 random-k 和 bottom-k 的训练池缺少关键的时空特征信息,所训练的模型特征编码能力变差,异常检测效果较差。在使用同样大小的样本量进行训练的情况下,基于伪异常样本选择方法(top-k)的表现最佳,可提高异常检测的准确率。

为了更直观地展示 top-k 和 bottom-k 选择策略在训练时的选择结果,本文在 ShanghaiTech 数据集上对两者进行了对比实验。本文在第二轮训练时,将原始训练集中帧序号为

2000至3000之间的被选视频帧进行了可视化展示，结果如图8所示。

(a) top-k



(b) bottom-k

图8　不同选择策略的帧选择结果(电子版为彩图)

Fig. 8　Selected frames of different selection strategies

图8(a)给出了top-k伪异常样本选择的结果，可以看出，被选样本包含了更多人物重叠和人物出入监控镜头的视频帧(见图中黄色方框)。在测试过程中，正确判别与图8(a)类似的视频帧是提高模型表现的关键。而图8(b)中的场景则重复率高，且大多由简单画面构成，以此类样本训练模型不利于识别复杂的正常视频帧。这表明与random-k和bottom-k方法相比，所提伪异常样本选择能够提高模型对正常复杂场景的判别能力，从而更高效地训练视频异常检测模型。

4.3.4　样本选择比例

选择策略中样本选择比例的设置是迭代停止的关键，样本选择比例的大小决定了从原始训练集往训练池中增加数据的次数，从而决定了训练池的大小。样本选择比例过大则不能达到去除冗余训练数据的效果，太小则正常样本不够丰富。为验证样本选择比例对结果的影响，本文在ShanghaiTech数据集上分别设置了不同样本选择比例来进行实验。观察图9中的数据可知，在伪异常选择驱动学习的框架下，模型总体表现会随样本选择比例的增加而提升。综合表2中的时间消耗，当样本选择比例取20%时，所得表现与100%数据集相当，在效率和准确率方面的性能最佳。

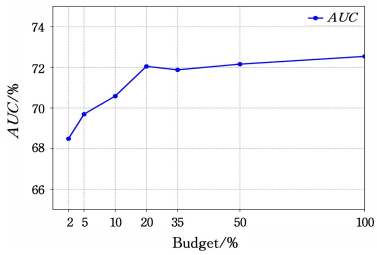


图9　不同样本选择比例下的表现

Fig. 9　Results under different selection budgets

结束语　本文提出了伪异常选择驱动学习的视频异常检测方法，可以在使用更少训练数据的情况下高效地完成视频异常检测模型的训练。本文设计了双路U-Net后继帧预测网络来编码视频中不同粒度的时空特征，并且通过分层的记忆力机制，强化了正常场景中典型时空特征产生的影响。所提伪异常样本选择方法具有普适性，能够减轻异常检测训练集中的数据冗余特性。本文将异常分数top-k的视频帧用于模型训练，在大幅缩短训练耗时的情况下，达到了与使用全部视频帧直接训练模型相近的效果。后续工作可考虑在引入部分异常数据的情况下，更加深入地研究样本选择策略。

参 考 文 献

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

[2] CHALAPATHY R, CHAWLA S. Deep learning for anomaly detection: A survey[J]. arXiv: 1901.03407, 2019.

[3] FU Z, HU W, TAN T. Similarity based vehicle trajectory clustering and anomaly detection[C]// Proceedings of the IEEE International Conference on Image Processing. 2005: II-602.

[4] HASANM, CHOI J, NEUMANN J, et al. Davis. Learning temporal regularity in video sequences[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 733-742.

[5] ABATI D, PORRELLO A, CALDERARA S, et al. Latent space autoregression for novelty detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 481-490.

[6] 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.

[7] HAO H W, JIANG R R. Training sample selection method for neural networks based on nearest neighbor rule[J]. Acta Automatica Sinica, 2007, 33(12): 1247-1251.

[8] MARKOVITZ A, SHARIR G, FRIEDMAN I, et al. Graph embedded pose clustering for anomaly detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2020: 10536-10544.

[9] FEICHTENHOFER C, FAN H, MALIK J, et al. Slowfast networks for video recognition[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6202-6211.

[10] LENG J X, TAN M P, HU B, et al. Video Anomaly Detection Based on Implicit View Transformation[J]. Computer Science, 2022, 49(2): 142-148.

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

[12] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmen-tation[C]// Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. 2015: 234-241.

[13] FANG Z, LIANG J, ZHOU J, et al. Anomaly Detection with Bi-directional Consistency in Videos[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(3): 1079-1092.

[14] XU D, YAN Y, RICCI E, et al. Detecting anomalous events in videos by learning deep representations of appearance and mo-

tion[J]. Computer Vision and Image Understanding, 2017, 156: 117-127.
[15] YU G, WANG D, CAI X, et al. Cloze test helps: Effective video anomaly detection via learning to complete video events[C]// Proceedings of the ACM International Conference on Multimedia. 2020: 583-591.
[16] CHANG Y, TU Z, XIE W, et al. Clustering driven deep autoencoder for video anomaly detection[C]// Proceedings of the European Conference on Computer Vision. 2020: 329-345.
[17] HAO Y, LI J, WANG N, et al. Spatiotemporal consistency-enhanced network for video anomaly detection[J]. Pattern Recognition, 2022, 121: 108232.
[18] GEORGESCU M, BARBALAU A, IONESCU R, 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. 2021: 12742-12752.
[19] GRAVES A, MOHAMED A, HINTON G. Speech recognition with deep recurrent neural networks[C]// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. 2013: 6645-6649.
[20] SHI X, CHEN Z, WANG H, et al. Convolutional LSTM Network: A machine learning approach for precipitation nowcasting [C]// Advances in Neural Information Processing Systems. 2015: 802-810.
[21] WESTON J, CHOPRA S, BORDES A. Memory networks[J]. arXiv: 1410.3916, 2014.
[22] WANG X, DU Y, LIN S, et al. adVAE: a self-adversarial variational autoencoder with Gaussian anomaly prior knowledge for anomaly detection[J]. Knowledge-Based Systems, 2019, 190: 105187.
[23] HAN T, XIE W A. Zisserman. Memory augmented dense predictive coding for video representation learning[C]// Proceedings of the European Conference on Computer Vision. 2020: 312-329.
[24] SUKHBAATAR S, SZLAM A, WESTON J, et al. End-to-end memory networks[J]. arXiv: 1503.08895, 2015.
[25] MILLER A, FISCH A, DODGE J, et al. Key-value memory networks for directly reading documents[J]. arXiv: 1606.03126, 2016.
[26] 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 International Conference on Computer Vision. 2019: 1705-1714.
[27] ZHOU Y, REN Q C, NIU H B. Research on training sample data selection methods[J]. Computer Science, 2020, 47(11A): 402-408.
[28] PLUTOWSKI M, WHITE H. Selecting concise training sets from clean data[J]. IEEE Transactions on Neural Networks, 1993, 4(2): 305-318
[29] WANG J, NESKOVIC P, COOPER L N. Training data selection for support vector machines[C]// International Conference on Natural Computation, 2005: 554-564.
[30] CORTES C, VAPNIK V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297.
[31] TANG K, LIN M, MINKU F, et al. Selective negative correlation learning approach to incremental learning[J]. Neurocomputing, 2009, 72(13/14/15): 2796-2805.
[32] CONG Y, YUAN J, LIU J. Sparse reconstruction cost for abnormal event detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2011: 3449-3456.
[33] MAHADEVAN V, LI W, BHALODIA V, et al. Anomaly detection in crowded scenes[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2010: 1975-1981.
[34] LU C, SHI J, AND JIA J. Abnormal event detection at 150 fps in matlab[C]// Proceedings of the IEEE International Conference on Computer Vision. 2013: 2720-2727.
[35] WANG Z, ZOU Y, ZHANG Z. Cluster Attention Contrast for Video Anomaly Detection[C]// Proceedings of the International Conference on Multimedia. 2020: 2463-2471.
[36] LI B, LEROUX S, SIMOENS P. Decoupled appearance and motion learning for efficient anomaly detection in surveillance video [J]. Computer Vision and Image Understanding, 2021, 210: 103249.
[37] ZHONG Y, CHEN X, JIANG J, et al. A cascade reconstruction model with generalization ability evaluation for anomaly detection in videos[J]. Pattern Recognition, 2022, 122: 108336.
[38] CHANG Y, TU Z, XIE W, et al. Video anomaly detection with spatio-temporal dissociation[J]. Pattern Recognition, 2022, 122: 108213.

ZHAO Song, born in 1997, postgraduate, is a student member of China Computer Federation. His main research interests include computer vision and video anomaly detection.

WANG Hongxing, born in 1985, Ph.D, associate professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include computer vision and pattern recognition.

(责任编辑:杨雪敏)