软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(9):4362−4377 [doi: 10.13328/j.cnki.jos.006641]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

监控视频异常行为检测的概率记忆自编码网络*

肖进胜[1], 郭浩文[1], 谢红刚[2], 赵 陶[1], 申梦瑶[1], 王元方[1]

[1](武汉大学 电子信息学院, 湖北 武汉 430072)

[2](湖北工业大学 电气与电子工程学院, 湖北 武汉 430068)

通信作者: 谢红刚, E-mail: xiehg@hbut.edu.cn

摘 要: 异常行为检测是智能监控系统中重要的功能之一, 在保障社会治安等方面发挥着积极的作用. 为提高监控视频中异常行为的检测率, 从学习正常行为分布的角度出发, 设计基于概率记忆模型的半监督异常行为检测网络, 解决正常行为数据与异常行为数据极度不均衡的问题. 该网络以自编码网络为主干网络, 利用预测的未来帧与真实帧之间的差距来衡量异常程度. 在主干网络提取时空特征时, 使用因果三维卷积和时间维度共享全连接层来避免未来信息的泄露, 保证信息的时序性. 在辅助模块方面, 从概率熵和正常行为数据模式多样性的角度, 设计概率模型和记忆模块提高主干网络视频帧重建质量. 概率模型利用自回归过程拟合输入数据分布, 促使模型收敛于正常分布的低熵状态; 记忆模块存储历史数据中的正常行为的原型特征, 实现多模式数据的共存, 同时避免主干网络的过度参与而造成对异常帧的重建. 最后, 利用公开数据集进行消融实验和与经典算法的对比实验, 以验证所提算法的有效性.

关键词: 异常行为检测; 自编码网络; 概率模型; 记忆向量

中图法分类号: TP391

中文引用格式: 肖进胜, 郭浩文, 谢红刚, 赵陶, 申梦瑶, 王元方. 监控视频异常行为检测的概率记忆自编码网络. 软件学报, 2023, 34(9): 4362–4377. http://www.jos.org.cn/1000-9825/6641.htm

英文引用格式: Xiao JS, Guo HW, Xie HG, Zhao T, Shen MY, Wang YF. Probabilistic Memory Auto-encoding Network for Abnormal Behavior Detection in Surveillance Videos. Ruan Jian Xue Bao/Journal of Software, 2023, 34(9): 4362–4377 (in Chinese). http://www.jos.org.cn/1000-9825/6641.htm

Probabilistic Memory Auto-encoding Network for Abnormal Behavior Detection in Surveillance Videos

XIAO Jin-Sheng[1], GUO Hao-Wen[1], XIE Hong-Gang[2], ZHAO Tao[1], SHEN Meng-Yao[1], WANG Yuan-Fang[1]

[1](Electronic Information School, Wuhan University, Wuhan 430072, China)

[2](School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan 430068, China)

Abstract: Abnormal behavior detection is one of the important functions in the intelligent surveillance system, which plays an active role in ensuring public security. To improve the detection rate of abnormal behavior in surveillance videos, this study designs a semi-supervised abnormal behavior detection network based on a probabilistic memory model from the perspective of learning the distribution of normal behavior, in an attempt to deal with the great imbalance between normal behavior data and abnormal behavior data. The network takes an auto-encoding network as the backbone network and uses the gap between the predicted future frame and the real frame to measure the intensity of the anomaly. When extracting spatiotemporal features, the backbone network employs three-dimensional causal convolutional and temporally-shared full connection layers to avoid future information leakage and ensure the temporal sequence of information. In terms of auxiliary modules, a probabilistic model and a memory module are designed from the perspective of probability entropy and diverse patterns of normal behavior data to improve the quality of video frame reconstruction in the backbone network. Specifically, the

* 基金项目: 中国科学院光电信息处理重点实验室开放课题基金 (OEIP-O-202009); 国家自然科学基金 (61471272)
收稿时间: 2021-06-09; 修改时间: 2021-08-28, 2021-10-13, 2021-11-16; 采用时间: 2021-12-30; jos 在线出版时间: 2023-02-22
CNKI 网络首发时间: 2023-02-24

probabilistic model uses the autoregressive process to fit the input data distribution, which promotes the model to converge to the low-entropy state of the normal distribution; the memory module stores the prototypical features of normal behavior in the historical data to realize the coexistence of multi-modal data and avoid the reconstruction of abnormal video frames caused by excessive participation of the backbone network. Finally, ablation experiments and comparison experiments with classic algorithms are carried out on public datasets to examine the effectiveness of the proposed algorithm.

Key words: abnormal behavior detection; auto-encoding network; probabilistic model; memory vector

由于异常行为事件种类繁多、难以预测, 检测异常行为是智能视频监控系统中最有挑战的技术之一. 提高异常行为检测算法的效率和准确率, 将会极大地增强其在突发事件处置中的实用性. 监控视频中正常行为数据量庞大且易于获取, 对于数据驱动模型的训练十分有利, 使得利用正常行为建模来检测异常具有可行性. 针对异常事件定义模糊、分类界限不明确的问题, 本文利用其对立面正常行为数据来拟合正常行为分布, 以自编码网络为基础, 通过对视频帧的重构生成正常数据, 利用重建误差评判输入数据与正常数据之间的偏差程度, 以此来判断异常[1].

具体而言, 在数据量足够的前提下, 异常行为检测算法更加关注数据内部本质特征的提取. 为了充分利用视频中所包含的空间信息和时间信息, 本文在自编码器的设计上加入了三维卷积来实现时空信息的联合提取. 考虑到时间序列模型的先后顺序, 避免未来信息发生泄露, 使用了因果卷积和时间维度共享全连接层来保证时序. 为了提高网络重建正常帧的能力, 本文从概率熵和正常行为数据模式多样性的出发, 设计了概率模型和记忆模块辅助主干自编码网络进行视频帧重建. 概率模型利用自回归过程拟合输入数据的分布, 在训练数据全为正常行为数据的情况下, 使网络收敛于低熵, 减少对正常行为的意外程度[2]. 考虑到正常行为数据模式的多样性, 单一原型特征很难全面覆盖所有模式的正常行为数据[3], 因此用记忆模块存储模型历史视频中的正常原型特征, 实现正常行为的多模式数据共存. 记忆模块借助注意力机制的思想, 通过对记忆向量与概率模型输出的特征向量之间的加权计算, 实现记忆模块向主干自编码网络的信息注入. 记忆模块的加入避免了主干网络的过度参与而导致对异常视频帧的重建.

所提算法主要有以下几点贡献.

(1) 针对时序视频帧, 利用因果三维卷积和时间维度共享全连接层维护信息的时序性, 基于自编码网络及帧预测器实现对正常帧的预测.

(2) 利用自回归概率模型实现对正常帧的输入分布的拟合, 促使网络收敛于正常分布的低熵, 增强网络对正常行为的建模能力.

(3) 利用记忆模块存储多个正常行为的原型特征, 并对概率模型输出的向量进行更新, 融入自编码网络, 实现正常行为的多模式数据的共存.

本文第 1 节介绍异常行为检测方面的相关工作. 第 2 节主要介绍提出的异常检测算法, 描述算法的实现步骤, 给出算法中各模块、损失函数及异常分数的具体介绍. 第 3 节进行实验阐述, 分析算法的实验效果, 并与其他算法进行对比, 在多个指标上进行分析与评价. 第 4 节给出全文总结.

1 相关工作

异常行为检测[4]是计算机视觉领域重要的任务之一, 学者们提出了许多方法来提高检测的性能. 传统的人工特征方法着重于分析低层次视觉特征, 如导向梯度直方图[5]、光流直方图[6]、时空兴趣点[7]等. 这种方法严重依赖于手工设计的视频描述符表征能力, 具有局限性. 近年来, 利用神经网络提取特征[8]成为研究热点. 为了减少人力成本, 同时兼顾异常行为检测的准确率, 相比于无监督[9]、全监督[10]和弱监督[11]学习算法, 大多数学者倾向于使用半监督学习算法. 半监督学习算法是指在训练集中仅包含正常行为数据, 通过学习正常行为数据的分布模型, 将偏离正常分布的输入数据判为异常行为数据. 根据数据处理方式的不同, 半监督方法可以分为基于概率模型的方法、基于距离的方法和基于重构模型的方法. 参考这些方法, 本文提出了一种新的基于概率记忆模型的异常行为检测算法.

1.1 基于概率模型的方法

基于概率模型的方法是用数据属于正常行为分布的概率来表征数据的正常程度, 将概率值小于阈值的输入数

据视为异常数据. Feng 等人[12]假定概率密度函数符合高斯分布, 将多个高斯混合层进行堆叠, 形成了深度高斯混合模型, 增强了模型的表达能力; Amraee 等人[13]利用监控视频中的重点观测区域特征来构建多元高斯模型. Zhang 等人[14]利用统计直方图实现对运动和外观的联合建模. Colque 等人[15]设计了基于光流特征和能量熵的直方图来进行群体异常检测. 此外, Abati 等人[2]利用自回归过程来学习正常行为数据的概率分布, 学习到了更有效的数据分布. 基于概率模型的方法理论基础完整、易于实现; 但学习到准确的概率分布十分困难, 对于图像视频这种高维数据而言, 其概率密度函数往往非常复杂, 很难用单一分布来表征.

1.2 基于距离的方法

基于距离的方法是利用距离度量函数或模型来评判输入数据与正常行为数据之间的距离. 最典型的方法为利用支持向量机来拟合正常数据的判别边界, 此类方法易受噪声、离群点等的干扰. Ma 等人[16]利用多种低层特征如直方图、密集轨迹等的量化数据来训练支持向量机的包络面, 将决策空间之外的数据判为异常数据. Ionescu 等人[17]学习目标的外观和运动信息并进行加权融合, 利用 K 均值聚类将正常数据划分为多个正常行为类别簇, 对每一个聚类中心, 其余簇均为伪异常, 并以此为基础训练多个支持向量机; 若测试数据不在任一簇中, 则为异常数据. Ramachandra 等人[18]利用孪生网络训练距离度量函数, 通过测试数据与参照数据之间的距离来得到最终的异常得分, 该方法需要预先存储大量的参照数据.

1.3 基于重构模型的方法

基于重构模型的方法是利用重构数据与原始数据之间的误差来表示原始数据的异常程度. 重构模型在正常行为数据约束下进行训练以学习正常行为数据的内在特征, 若测试数据属于正常行为数据, 则重构模型可以以较小的重建误差来重构输入数据; 而对于异常行为数据, 重构模型不能很好地对数据进行重建, 因而重建误差大, 异常得分高. 此类模型不能学习正常行为数据和异常行为数据之间的边界, 受训练数据分布影响较大. 目前使用最广泛的重构生成模型当属生成对抗网络和自编码网络. 生成对抗网络通过生成器和对抗器的联合训练实现对输入数据的重构或预测. Liu 等人[19]利用生成对抗网络来预测未来帧, 根据预测的未来帧与真实帧之间的差异来检测异常. 从网络输入角度来看, Vu 等人[20]针对 RGB 图、光流图、时空特征分别训练生成对抗网络, 用多个网络的生成误差加权和来衡量异常程度. 以“编码-解码-编码”为思路, Akcay 等人[21]针对测试数据在生成器编码空间和隐空间之间的差距来判断异常. 此外, 对于另一种重构方法, 自编码网络利用编解码器对输入数据进行重构, 用网络输入与输出之间的距离来衡量输入偏离正常分布的程度. Hasan 等人[22]使用了堆叠的卷积层来实现对视频帧特征的编码, 利用编码器的镜像结构组成的解码器来进行图像重建. 由于卷积操作在二维平面上实施, 会导致时间信息的丢失, 袁静等人[23]对稀疏去噪自编码网络添加梯度差约束, 提高网络性能. 为了更好地保留时间信息, Chong 等人[24]设计了时空自编码器, 利用基于卷积的长短期记忆 (convolutional long-short term memory, Conv-LSTM)[25]结构进行编码, 以捕捉视频中包含的时空信息; Luo 等人[26]基于时域一致稀疏编码在堆叠的循环神经网络上实现了对群体异常行为的检测. Park 等人[3]设计了记忆模块辅助基于卷积的自编码网络进行图像重构.

相较于上述方法, 本文的工作基于重构模型, 更着力于增强正常帧的重建质量. 在重建时考虑到正常行为分布的概率熵和正常行为数据模式的多样性, 分别设计了概率模型和记忆模块. 概率模型用于约束网络提取出更有效的特征, 记忆模块用于多模式数据的共存. 两个模块相互贴合, 进一步提高了网络对于正常帧的重建质量, 对于异常帧则不能正常重建. 最终网络对于异常帧有良好的区分能力, 实现异常行为检测率的提高.

2 本文算法

基于概率记忆模型的异常行为检测网络整体框架如图 1 所示. 自编码网络通过编码-解码过程学习数据本身的性质, 从而实现对数据的重建. 考虑到真实世界中隶属于正常情况的模式的多样性, 单一原型特征的图像重构模型很难对多样的正常模式进行准确建模, 因此本文设计了隐向量概率分布的记忆模型. 通过概率模型约束隐空间分布, 同时利用记忆模块存储不同正常行为的原型特征, 提高对监控视频异常行为的检测率.

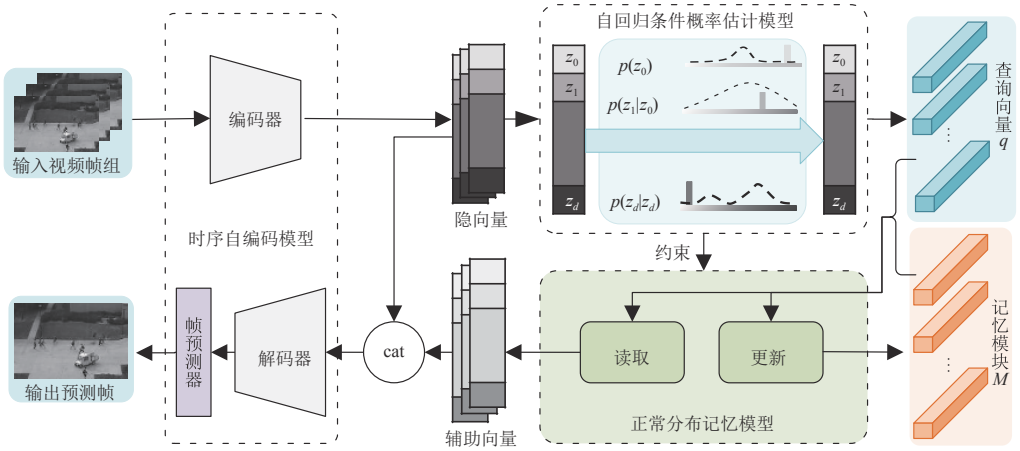


图 1　基于概率记忆模型的异常行为检测网络结构示意图

2.1　时序自编码网络

首先, 网络输入数据为多帧的视频帧组, 所设计的时序自编码网络架构由编码器、解码器和帧预测器 3 部分组成. 编码器包括 5 层下采样层和 2 层时间维度共享全连接层, 实现输入 x 到其编码 z 的映射; 解码器结构与编码器结构对称, 实现隐向量 z 到多维时间特征图 $\tilde{x}$ 的映射; 帧预测器包含一个时空特征融合结构, 将 $\tilde{x}$ 转换为最终的预测帧 y. 图 2 详细给出了时序自编码网络参数及特征图尺寸[2].

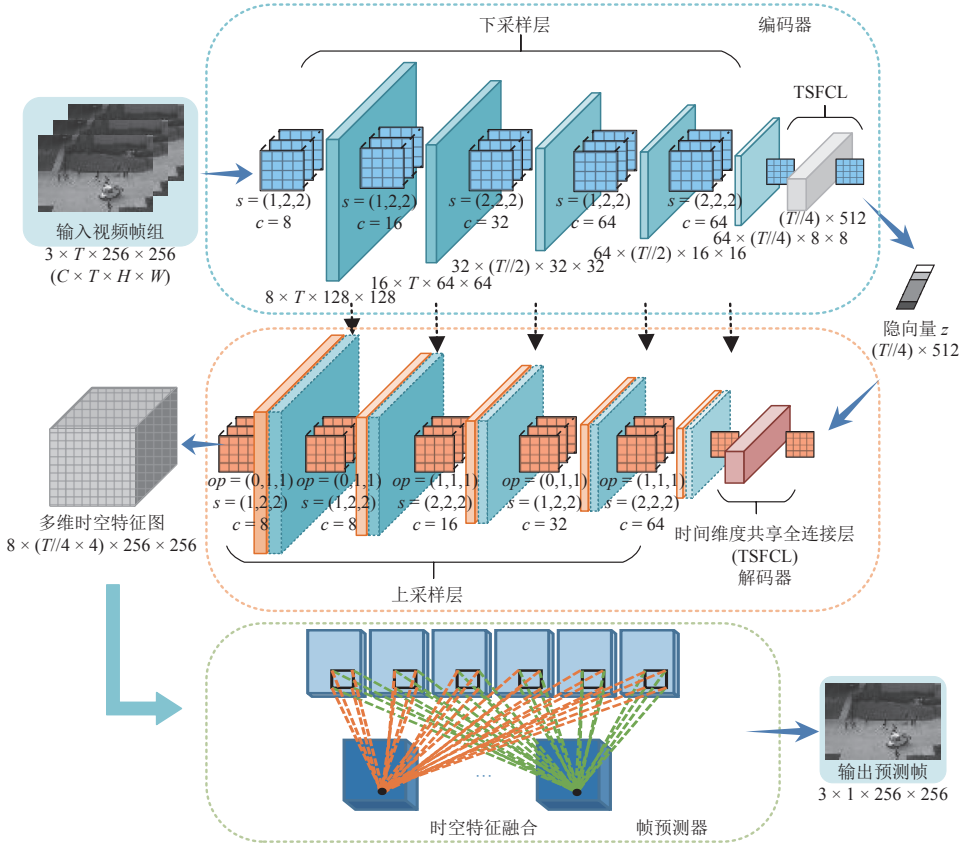


图 2　时序自编码网络结构图

图 2 中 T 表示时间轴长度. 编码器接收一个 3×T×256×256 的张量, 经过下采样层的卷积操作、解码器上采样层的反卷积操作和时空信息融合结构的卷积操作, 最终得到 3×1×256×256 维度的输出预测帧. 为了充分利用多层次卷积特征, 将输入端信息通过跳层连接辅助输出端进行图像重建, 使得上采样层的每一层的输入维度扩充为之前的两倍. 跳层连接的加入有利于每层参数更新分布均匀, 提高泛化性, 可以极大改善网络的性能. 在保证信息的时序性上, 自编码网络在采样层使用了因果 3D 卷积, 以及编码器和解码器的连接使用了时间维度共享全连接层.

(1) 因果残差采样层

一般卷积网络中信息流传递较弱, 为了减少信息的丢失, 在每个采样层设计了残差结构, 采样层结构如图 3 所示. 下采样层主干用两个因果 3D 卷积来进行特征的提取, 在原有的 3D 卷积核上加上时间维度的掩膜, 从而保证了信息的时序性. 采样层的分支结构将输入添加到该层输出中, 经过了一个步长为 1 的 3D 卷积来保证输入输出维度一致. 这种残差结构使得网络能够向更深的设计发展, 时空特征图拥有更大的感受野, 更好地捕获较高层次的语义信息. 同理, 上采样层也采用残差结构. 主干采用 3D 反卷积和 3D 卷积进行特征重建, 分支采用 3D 反卷积使输入数据维度与输出数据维度匹配.

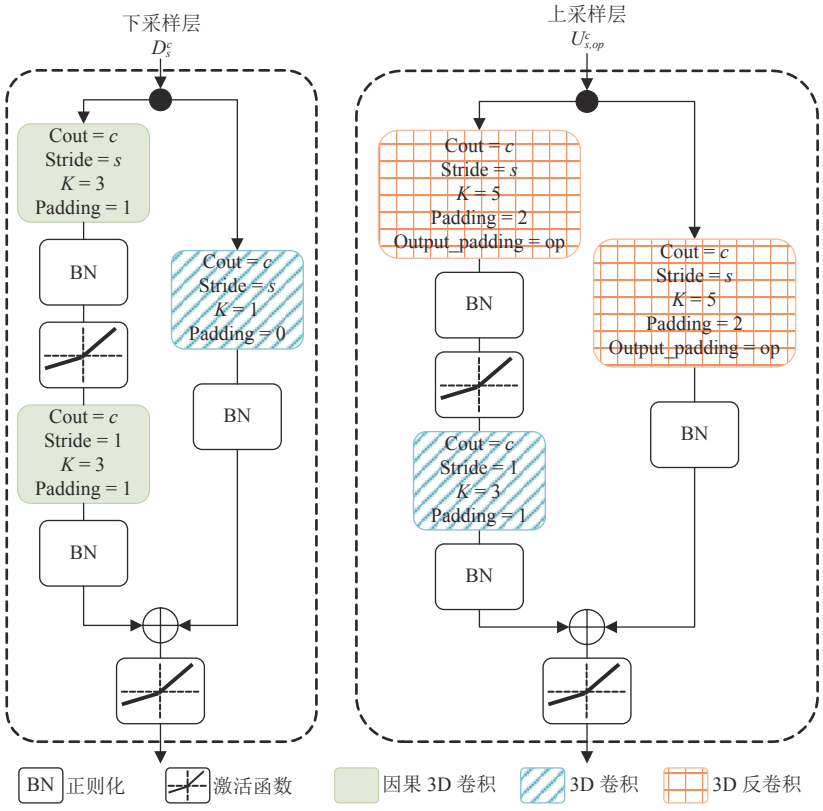


图 3 采样层结构示意图

(2) 时间维度共享全连接层

经过编码器输出的特征图在时间尺度上不具有置换不变性, 因此在编码器末尾用时间维度共享全连接层 (TSFCL) 替换普通的全连接层. 对于特征图, 按照时间维度分成单个矩阵向量, 分别输入全连接层进行特征空间转换. 对于不同时间维度的矩阵向量, 共用同一全连接层参数. 输出经过全连接层变换后, 按照时间顺序进行堆叠, 从而获得在时间上彼此独立的特征图. 时间维度共享全连接层在进行特征提取和空间转换时, 保证时间顺序一致性.

2.2 自回归条件概率估计模型

自编码网络通过正常样本学习正常特性, 在遇到异常样本时根据学习到的正常特性将异常样本辨别出来. 根

据认知心理学, 对异常行为的判断可以根据对正常行为的记忆能力来表示, 也可以由该行为引起的意外程度来表示. 前者我们用自编码网络对数据的重建来实现, 后者可以根据在预期模型下发生的概率较低来进行建模. 针对训练样本 x 的概率密度函数 $p(x)$, 我们通过隐向量 z 的真实分布 $p(z)$ 来辅助建模, 形成描述所有观察结果的因果关系集. 对 $p(x)$ 进行分解, 如公式 (1). 其中 $p(x|z)$ 是针对观察结果的条件似然函数, 由重建误差实现近似.

$$p(x)=\int p(x|z)p(z)dz \tag{1}$$

针对 $p(z)$ 的估计, 构建了概率模型 $h(z;\theta_h)$, 并增加约束条件为最小化隐向量 z 的微分熵. 隐向量 z 的微分熵表示如公式 (2) 所示. 其中, $p^*\left(z;\theta_f\right)$ 表示由编码器 $f(\cdot)$ 产生的隐向量 z 的真实分布, $D_{KL}(\cdot\|\cdot)$ 表示前一分布与后一分布的 Kullback-Leibler 散度, 简称 KL 散度, H 表示信息熵, E 表示数学期望. 最小化隐向量 z 的微分熵, 可以确保概率模型 $h(z;\theta_h)$ 模拟的参数分布与真实分布 p^* 之间的信息差距很小, 而且使隐含在编码器 $f(\cdot)$ 产生的编码向量所属的分布的微分熵最小. 这种关系在学习正常行为模式时是至关重要的约束点. 如果我们将编码器视为发出潜在分布的源, 面对正常行为数据时, 其期望的行为应当收敛于以低熵作为固有特征的“平凡”过程. 因为异常行为在训练过程中不会出现, 所以不会产生较大的“意外”; 降低信息熵意味着模型所包含的意外程度减少, 从而可以更加准确地对正常行为模式进行建模. 因此, 在编码器的训练过程中, 我们的目标是希望编码器表现出较低的微分熵. 在这一约束条件下, 编码器会被训练得更加注重提取在训练集中经常出现的、易于预测的正常行为的特征, 这种特征对于判别新样本是否是正常行为样本最为有用.

$$\begin{aligned}E_{z:p^*(z;\theta_f)}[-\log h(z;\theta_h)]&=E_{z\sim p^*(z;\theta_f)}\left[-\log h(z;\theta_h)+\log p^*\left(z;\theta_f\right)-\log p^*\left(z;\theta_f\right)\right]\\&=D_{KL}(p^*\left(z;\theta_f\right)\|h(z;\theta_h))+H\left[p^*\left(z;\theta_f\right)\right]\end{aligned} \tag{2}$$

具体实现上, 为了避免模型采用特定的分布族, 通过自回归过程来估计 z 的分布 $p(z)$, 如公式 (3) 所示. 对 $p(z)$ 的估计转换为对每个 $p(z_i|z_{<i})$ 的估计, 其中 $<$ 表示随机变量的顺序, d 为隐向量 z 的维度. 针对 $p(z_i|z_{<i})$ 的估计, 这里引用了有序堆叠全连接层[2], 确保每个 $p(z_i|z_{<i})$ 都是仅根据 $\{z_0,z_1,\ldots,z_{i-1}\}$ 计算得到的, 并以多项式的形式对概率密度函数进行逼近. 首先, 利用 Sigmoid 函数对隐向量 z 进行激活, 将其投射到 [0, 1] 范围内, 并在 B 个量箱上进行量化. B 为超参数, 实验中将其设为 100. 即对于第 i 个隐向量元素 z_i, 以 B 维分类函数 $\phi(z_i)$ 的形式表明其属于哪一个量箱类别. 通过对每一个隐向量元素以连续分布 $p(z_i|z_{<i})$ 在 B 维分布函数上的建模, 成就了概率模型 $h(z;\theta_h)$ 的基础.

$$p(z)=\prod_{i=1}^{d}p(z_i|z_{<i}) \tag{3}$$

另外, 在有序堆叠全连接层中, 沿编码维度设计了 d 个不同的卷积核. 每个卷积核添加不同的掩膜使得整个卷积过程仅可以观察到上一个时间步中的整个特征向量和当前时间步中的部分特征向量, 卷积核沿时间轴移动进而捕获蕴含在特征图中的时间信息. 不同类型的有序堆叠全连接层的串连形成隐向量概率分布估计器, 如图 4 所示. 类型 A 严格依赖于先前的特征元素, 仅用作模型的起始层. 类型 B 仅掩盖后续特征元素, 使得当前输出与先前输入和当前输入有关. 隐向量概率分布估计器通过设计的掩膜卷积核实现自回归估计和不同类型网络层的堆叠, 实现对输入分布的建模.

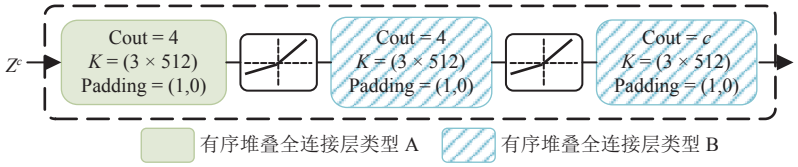


图 4 隐向量概率分布估计器结构示意图

2.3 正常分布记忆模型

网络在训练过程中, 不会接触到异常行为样本, 只在测试时利用重建误差来量化视频帧的异常程度. 而大多数真实监控视频帧中所包含的异常行为涉及的范围并不大或者有些异常行为表现不甚明显. 深度神经网络强大的重

建能力甚至可以重建这部分异常场景, 会产生对异常行为的漏检. 同时, 根据前人工作[2,3]所述, 对单一模式正常行为数据进行建模, 没有考虑到现实场景中正常行为数据模式的多样性, 也会对检测结果造成一定的影响. 在正常行为数据经过自编码网络和概率模型输出学习到的特征后, 这些特征是杂乱没有规律的. 但同一场景或相似场景的正常数据特征在空间中的距离会特别相近, 如同一室内建筑, 公路和街道两个相似场景等. 若将这些相似特征进行聚类, 那么每一类特征就是正常数据的一种模式. 通常数据集里的摄像头是固定的, 所包含的场景也很有限. 每种模式象征着正常数据的同一场景或相似场景, 对应一个原型特征表示, 位于每一类聚类特征的中心. 通过获得正常行为数据的各种模式, 向自编码网络注入更多与输入特征最接近模式的信息, 可以更针对性地重建出该场景下的正常帧. 基于上述讨论, 我们提出了一种正常分布记忆模块. 该模块包含了多个模拟正常行为的原型特征向量, 即“记忆向量”, 可以使得模型针对每个模式的正常视频获得其独有的属性. 记忆模块的设计一方面可以存储多项正常行为原型特征向量, 通过加权求和对视频数据进行重建, 弥补单个原型特征不足以表示多种正常行为数据的缺陷; 另一方面, 记忆模块辅助神经网络进行正常视频帧的重建, 减少神经网络的过度参与和表示能力, 降低异常行为的漏检率. 在具体实现上, 用概率模型输出的特征向量来生成记忆向量. 由于概率模型输出向量表现为多维, 将这些特征按照时间维度进行独立划分形成“查询向量”. 查询向量通过与记忆向量的距离形成记忆向量的权重来提取并融合记忆模块蕴含的信息, 一方面可以减少记忆向量的数量, 另一方面将新数据的模式保留在记忆模块中. 因此, 记忆模块拥有两个最基本的功能, 即“读取”和“更新”[3].

查询向量通过计算与记忆模块中各记忆向量的距离来形成各记忆向量的融合权重, 通过对记忆向量进行加权融合, 形成新的查询向量. 该向量包含了原有查询向量的信息和记忆模块中的信息, 对正常行为特征表述更为全面. 读取操作与注意力机制操作类似, 如图 5.

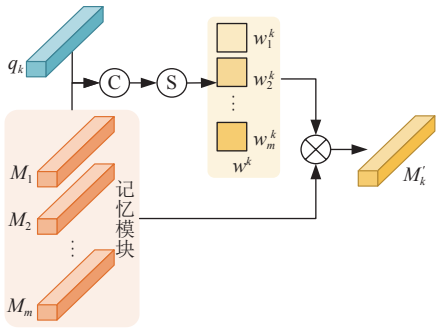


图 5 记忆模块读取操作示意图

在图 5 中, 第 i 个记忆向量 M_i 对于第 k 个查询向量 q_k 的权重系数 w_i^k 是其余弦距离的相对正则化, 如公式 (4) 所示. 其中 T 表示转置操作. 然后计算所有记忆向量的加权平均值, 获得更新后的特征向量 M'_k, 如公式 (5) 所示. 在读取记忆模块获得更新的特征向量时, 采用了所有的记忆向量而非与查询向量最接近的记忆向量, 在于重建帧时可以考虑各种正常行为模式. 然后将其与隐向量沿通道维度连接起来, 一并输入解码器, 这样可以使用记忆模块中的正常行为模式来辅助重构输入帧, 从而减少了自编码网络的表示能力, 减轻自编码器的负担, 同时可以覆盖所有已记录的正常行为模式.

$$w_i^k = \frac{\exp(q_k M_i^{\mathrm{T}})}{\sum_{j=1}^{m} \exp(q_k M_j^{\mathrm{T}})} \tag{4}$$

$$M'_k = \sum_{i=1}^{m} w_i^k M_i \tag{5}$$

记忆模块的更新是根据原有记忆向量和查询向量的关系, 利用当前的输入信息对记忆模块的正常行为模式进行更新. 因为记忆向量需要根据正常分布数据不断地输入, 不断更新调整, 才会提高记忆向量的泛化性, 对所有训练数据起到良好的记忆效果. 受前人工作[3]的启发, 设计的具体更新操作流程如图 6 所示.

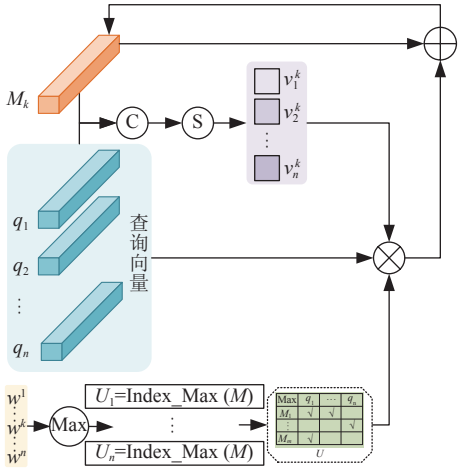


图 6　记忆模块更新操作示意图

首先, 根据公式 (4) 从匹配概率组 $w^k=\left[w_1^k,w_2^k,\ldots,w_n^k\right]$ 中寻找与 q_k 匹配概率最大的记忆向量 M_i. 由于可能存在多个记忆向量与查询向量最接近, 所以将最大索引值排列形成矩阵 U_k, 用来表示与 q_k 最接近的记忆向量的索引值矩阵. 与此同时, 对于第 k 个记忆向量 M_k, 按照公式 (6) 获取其与每个查询向量之间的匹配概率 v_i^k, 并对其归一化. 最后利用归一化后的匹配概率对相应的查询向量进行加权平均, 并将其累加到原始记忆向量中实现对索引值矩阵里记忆向量的更新, 如公式 (7) 所示. 其中 $f(\cdot)$ 表示 L2 正则化. 通过对查询向量的加权求和, 可以在更新过程中将更多的注意力集中在与记忆模块最近的查询向量上. 记忆模块在模型训练和测试时都拥有更新操作. 但在测试时, 模型会接触到异常行为视频. 所以还使用了正则化分数[3]来控制更新操作使用的时机. 当正则化分数高于阈值时, 将不用于记忆模块的更新.

$$v_i^k=\frac{\exp(q_iM_k^{\mathrm{T}})}{\sum_{j=1}^{n}\exp(q_jM_k^{\mathrm{T}})} \tag{6}$$

$$M_k\leftarrow f\left(M_k+\sum_{j\in U^k}\frac{v_j^kq_j}{\max_{i\in U^k}v_i^k}\right) \tag{7}$$

2.4　目标函数及异常分数

根据图 1 网络的 3 个部分, 网络训练的目标函数和判断视频异常程度的异常分数的计算从重建误差、概率熵、记忆特征 3 个方面展开.

重建误差定义为预测帧与真实帧的均方误差 *MSE* (mean square error), 如公式 (8) 所示. 其中 M 和 N 分别是 t 时刻图像 I_t 的长和宽. 通过最小化预测帧与真实帧之间的重建误差, 从而迫使网络学习正常视频帧的结构特征. 在测试时, 使用 *PSNR* (peak signal to noise ratio) 生成异常分数, 如公式 (9) 所示.

$$\ell_{\mathrm{rec}}=\frac{\left\|\hat{I}_t-I_t\right\|_2^2}{M\times N}=\frac{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left[\hat{I}_t(i,j)-I_t(i,j)\right]^2}{M\times N}=MSE \tag{8}$$

$$S_{\mathrm{rec}}=10\times\log_{10}\left(\frac{MAX_I^2}{MSE}\right)=PSNR \tag{9}$$

概率模型在正常行为数据的作用下使网络收敛于低熵, 以编码器生成的隐向量为边界函数, 概率模型的目标函数最终可以表示为每一个概率密度分布和其分类函数的交叉熵损失, 得到的自回归概率损失函数也即异常分数如公式 (10).

$$\ell_{\mathrm{atr}}\left(\theta_f,\theta_h\right)=E\left[-\sum_{i=1}^{d}\sum_{k=1}^{B}\phi(z_i)_k\log(p(z_i|z_{<i})_k)\right]=S_{\mathrm{atr}} \tag{10}$$

记忆模块存储了不同模式的原型特征, 我们使用特征紧密损失来减少类内差异, 使用特征分离损失来增大类间差距[3]. 这样可以保证记忆向量的辨别能力和多样性. 特征紧密损失鼓励受概率模型约束的查询向量接近记忆模块中的存储项, 从而减少类内差距. 因此特征紧密损失定义为查询向量与其最接近的记忆向量的 *MSE*, 同时也表征了异常程度, 如公式 (11) 所示. 其中 n 和 d 分别表示查询向量的行和列维度, p_p 定义为与 q 最接近的记忆向量. 如果一味强调减少类内差距, 对于无监督的训练而言, 很容易使所有的特征向量收敛于同一分布, 从而造成模型退化, 记忆模块丧失记录各种正常行为模式的能力. 因此, 使用特征分离损失来增大类间差距, 使特征向量在类内接近而在类间疏远. 特征分离损失具体如公式 (12). p_n 为与 q 次接近的记忆向量, α 为超参数, 表示正负样本之间的最小间隔, 实验中设置为 1.

$$\ell_{\text{cpt}} = \frac{1}{n\times d}\sum_{i=0}^{n-1}\sum_{j=0}^{d-1}[q(i,j)-p_p(i,j)]^2 = S_{\text{cpt}} \tag{11}$$

$$\ell_{\text{spt}} = \frac{\left[\left\|q-p_p\right\|_2^2 - \left\|q-p_n\right\|_2^2 + \alpha\right]_+}{n\times d} \tag{12}$$

综上所述, 结合各模块的特性, 整体网络训练的目标函数 ℓ 如公式 (13) 所示. 其中各项损失在计算时均有正则化操作, 故其范围均在 $[0,1]$ 之间, 各项系数取相同取为 1. 最终的异常分数 S 如公式 (14). 由于整个异常行为检测网络各个模块的侧重点不同, 其反映在各模块异常分数上的诉求可以通过加权平均的方式都得到关注.

$$\ell = \ell_{\text{rec}} + \alpha_{\text{atr}}\ell_{\text{atr}} + \alpha_{\text{cpt}}\ell_{\text{cpt}} + \alpha_{\text{spt}}\ell_{\text{spt}} \tag{13}$$

$$S = \lambda_{\text{rec}}Norm(S_{\text{rec}}) + \lambda_{\text{atr}}Norm(S_{\text{atr}}) + \lambda_{\text{cpt}}Norm\left(S_{\text{cpt}}\right) \tag{14}$$

$$\lambda_{\text{rec}} + \lambda_{\text{atr}} + \lambda_{\text{cpt}} = 1 \tag{15}$$

3 实　验

3.1 实验配置

本文网络使用 Python 3.6 实现, 利用 PyTorch 1.1.0 框架搭建, 训练和测试借助武汉大学超算完成. 代码运行在 x86_64 架构的 CentOS 7.5 系统, 计算资源为 Nvidia Tesla V100 16 GB, 依赖 CUDA10 和 CUDNN7 支持. 视频帧在输入到网络之前被缩放到 256×256 大小, 3 通道的像素值归一化到 $[-1,1]$. 网络训练共有 60 个 epoch, 采用初始学习率为 2e–4 的 Adam 优化器, 学习率按照 CosineAnnealingLR[27]的方式呈下降趋势进行调整, 批处理块大小 batch size 设置为 4. 自编码网络隐向量设为 512 维; 概率模型使用 100 阶多项式来拟合概率密度函数; 记忆模块含有 10 个 512 维的记忆向量. 本文在 UCSD Ped2 和 ShanghaiTech 数据集上进行性能测试.

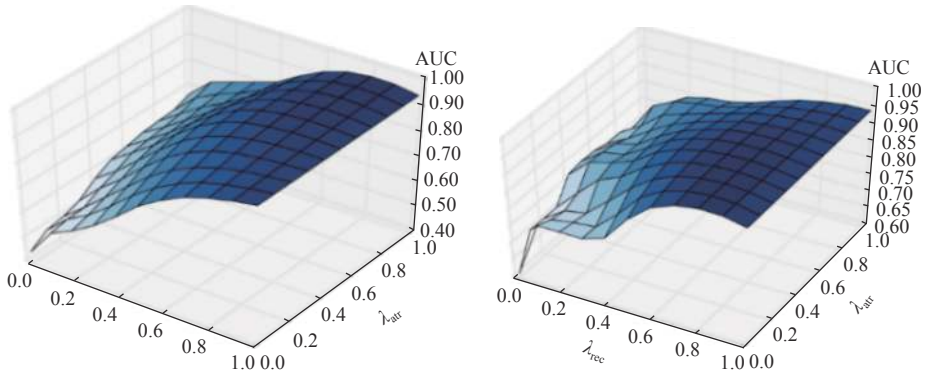
为了综合评估所提算法的性能, 异常行为检测研究中, 经常使用受试者工作特征 (receiver operating characteristic, ROC) 曲线及其所对应的面积 (area under the curve, AUC) 来作为模型特性的衡量指标[28]. 我们同样借助于 ROC 和 AUC 两个指标. 它们会针对每一个阈值计算假正类率 (false positive rate, FPR) 和真正类率 (true positive rate, TPR), 并以此为横纵轴绘制 ROC 曲线, 进而可计算曲线下面积即 AUC. ROC 可以在正负样本不均衡的情况下使用. 当 TPR 越大、FPR 越小, ROC 就越接近 (0, 1) 点, 此时 AUC 越接近 1, 说明模型的性能越好. 相较于 ROC 不直观的曲线表达方式, AUC 可以定量地描述出模型的性能, 表示预测为正的概率值比预测为负的概率值大的可能性, 即算法根据得分值将随机挑选的正样本排在负样本前面的概率[29]. 此外, 异常检测属于二分类问题, 本文还使用了 F_1 值即 $F1$-score 指标[30]测试性能, 该值为精确率 (precision, P) 与召回率 (recall, R) 的调和平均值. $F_1 = (2\times P\times R)/(P+R)$. F_1 值越大, 则算法分类的效果越好.

3.2 模型超参数设置

视频帧组的不同长度会对结果产生不同的影响: 视频帧组小, 模型特征提取简单但蕴含的时序变化较少; 视频帧组大, 蕴含的时间特征充分但模型处理数据量大而复杂. 考虑到硬件条件的限制及前人的经验值, 最终以 T=5 和 T=9 对模型进行训练, 分别代表利用 4 帧预测第 5 帧和利用前 8 帧预测第 9 帧.

根据公式 (14), 为了探究每个模块对最终异常分数的影响, 寻找最佳评分项系数组合, 以 0.1 为步长, 通过网

格化搜索得到不同 T 下的最优模型在不同系数下的性能如图 7 所示. 底平面轴分别为 λ_{rec} 和 λ_{atr}, z 轴为对应参数所计算出的 AUC 值, 图中颜色越深代表 AUC 值越大. 从图中可以看出, 两个模型都很大程度上依赖于重建误差, 最优参数组合集中在 λ_{rec} 轴的后半部. 在辅助模块方面, 对输入数据的概率估计占 T=5 模型的绝大比重, 最优参数组合集中在 λ_{atr} 轴的前半部; T=9 模型恰恰相反, 记忆模块的决定权更大. 当输入视频帧组较短时, 模型学习输入正常数据分布相对简单; 由于视频帧之间较为接近, 记忆向量发挥空间不大, 对于网络的贡献有限. 而当数据维度大时, 有足够的信息支撑记忆向量空间, 但高维的数据使得概率分布拟合较为困难, 因此记忆模块提供信息的权重大. 从模型指标性能上来看, T=5 模型的最佳 AUC 为 0.958 050, 略低于 T=9 模型的 0.958 954, 证明本文模型对长度大的帧视频组有更大的性能优势.



(a) T=5 系数组合网格搜索 (b) T=9 系数组合网格搜索

图 7 不同系数组合对结果的影响

3.3 消融实验

为了探究各模块的作用, 在 UCSD Ped2 数据集上对模型开展消融实验, 比较模型中不同模块的性能. 具体有: 对于主干自编码网络, 探究跳层连接对结果的影响; 对于整体网络, 比较主干自编码网络、概率模型、记忆模块两两组合对结果的影响. 训练的模型各自取第 3.2 节系数的最优性能, 消融实验 AUC 指标结果如表 1 所示.

表 1 消融实验结果

主干自编码网络		概率模型	记忆模块	T=5	T=9
有跳层	无跳层				
—	√	—	—	0.850	0.860
√	—	—	—	0.948	0.944
√	—	√	—	0.953	0.951
√	—	—	√	0.946	0.954
√	—	√	√	0.958	0.959

从表 1 可知, 跳层连接使用输入端特征信息辅助引导输出端进行特征重建, 可以极大地提高网络的重建能力. 对于辅助模块, 增加概率模型和记忆模块后, 网络的性能都得到了有效的提升. 概率模型的加入使训练网络收敛于低熵, 约束网络隐向量特征空间, 学习到更有效的正常数据特征, 于是指导重建的效果变好. 记忆模块的加入在概率模型输出的特征向量基础上, 保存了多项正常行为的原型特征, 避免了单一原型特征不足以表示多种正常行为数据的缺陷, 所以提高了重建数据的多样性; 并且记忆模块类似注意力机制的设计, 通过加权求和增加离特征最接近的正常模式比重对视频数据进行重建, 也同样提高了重建质量. 两者相互结合时, 网络性能提升更多. 因为主干网络注入了经过选择后的辅助重建信息后, 可以降低主干网络的表示能力, 防止主干网络过度参与而重建出异常帧.

此外, 从对不同视频组的长度 T, 基础网络增加某一模块的效果可知, 两个模块对于不同视频组的长度具有

互补作用, 保证在不同视频组长度下都能发挥出较好的综合性能. 当视频帧组长度为 5 时, 如上所述, 记忆模块提升的作用不如概率模型, 甚至于出现记忆模块的辅助会使网络收敛于次局部极大值, 不能完全发挥网络的优势; 当视频帧组长度增加至 9 时, 由于视频帧组时空特征信息量增多, 记忆模块便发挥作用, 存储一部分历史时空信息, 辅助网络进行视频帧重建, 从而提升了网络的性能. 由于概率模型利用自回归对输入数据分布进行拟合, 其输出的特征向量更能表示正常数据分布, 以该向量为查询向量更贴合记忆模块中的记忆向量的特性, 因此网络能展现更好的性能. 综合来看, 本文算法模块对最终网络模型的性能会产生有利影响, 印证了算法的有效性.

3.4 与经典算法对比

为了横向比较本文所提算法的性能, 在公开数据集 UCSD Ped2 和 ShanghaiTech 上与经典算法比较 AUC 指标. 这两个数据集是异常行为检测领域针对半监督学习使用最多的数据集, 在此仅列出部分算法对比. UCSD Ped2 数据集实验对比结果如表 2 所示.

表 2 中大多数算法遵循对正常分布进行建模来检测异常行为的思想, 算法性能逐年提升. 而 PMAE 算法效果处于佼佼者地位. 其中 ENC、MNAD 使用的是重新训练的模型; LSAND 是作者提供的训练好的模型; 其余对比算法指标使用参考文献中的已有数据. 从对比结果可以看出, PMAE 算法与使用了自回归概率方法的 LSAND 和其他时序神经网络算法相比, 性能又得到了提升. 而且 TSC、Stacked RNN 和 MNAD 这些使用了时序信息建模的方法, 性能都表现较好, 说明 UCSD Ped2 这种简单的数据集上充分利用时序特征信息非常重要. PMAE 算法的记忆模块记忆了历史正常数据的多类时空信息特征, 就发挥了关键作用.

对于规模大、场景多的 ShanghaiTech 数据集, 在此数据集上验证的算法较少, 对比结果如表 3 所示. 从表 3 可以看出, 所提 PMAE 算法在 ShanghaiTech 数据集上的性能十分具有挑战性, 超过并接近了许多经典实验算法. 尤其是针对同样使用了记忆模型的 MNAD 算法相比, PMAE 算法表现更好, 说明额外增加概率模型后, 对 ShanghaiTech 这种复杂的数据集提取了概率分布信息, 并约束了网络减少对复杂正常数据的意外程度. 最终学习到更有效更普遍的正常数据特征信息, 会对性能提升很大. 使用了自回归概率方法的 LSAND 算法和用对抗生成网络学习数据分布的 FFP+MC 算法实验效果较好, 而其他时序神经网络算法效果表现不突出也体现了这一点, 验证了本文算法概率模型的有效性.

表 2 UCSD Ped2 数据集性能对比

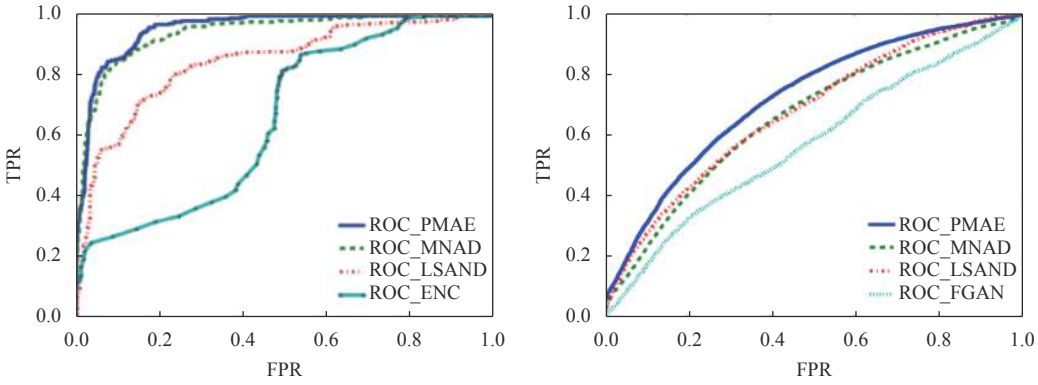
算法	AUC
ConvAE (2016)[22]	0.850
ENC (2017)[24]	0.656
TSC (2017)[26]	0.910
Stacked RNN (2017)[26]	0.922
FFP+MC (2018)[19]	0.954
LSAND (2019)[2]	0.844
MNAD (2020)[3]	0.950
PMAE_T5 (本文算法)	0.958
PMAE_T9 (本文算法)	0.959

表 3 ShanghaiTech 数据集性能对比

算法	AUC
ConvAE (2016)[22]	0.609
TSC (2017)[26]	0.679
Stacked RNN (2017)[26]	0.680
FFP+MC (2018)[19]	0.728
LSAND (2019)[2]	0.708
FGAN (2022)[10]	0.570
MNAD (2020)[3]	0.678
PMAE_T5 (本文算法)	0.722
PMAE_T9 (本文算法)	0.729

根据 AUC 指标对比结果, 所提算法较其他对比算法有了一定提升. 在 ROC 指标上, 现挑选了近几年的部分算法绘制 ROC 曲线如图 8 所示, 各算法使用的视频帧组长度按照最短长度截取, 同时对数据进行对齐计算出指标. 根据图 8 曲线高度可以观察到, 所提 PMAE 算法的曲线位置高于其他 3 种对比方法, 且更接近于 (0, 1) 点, 因此体现出较好的性能, 验证了所提算法的有效性. 由于 ShanghaiTech 数据集涵盖了多种场景, 异常行为种类也较 UCSD Ped2 数据集多, 因此十分具有挑战性. 从图 8 中算法在两数据集上的表现也可以看出, UCSD Ped2 数据集的 ROC 曲线明显比 ShanghaiTech 数据集上的 ROC 曲线更加接近 (0, 1) 点, 所以对于场景复杂的数据集而言还存

在很大的改进空间. 综合来看, 无论是在 UCSD Ped2 数据集还是在 ShanghaiTech 数据集上, 所提算法性能均有可观之处, 十分具有竞争力.



(a) UCSD Ped2 数据集 ROC 曲线　　(b) ShanghaiTech 数据集 ROC 曲线

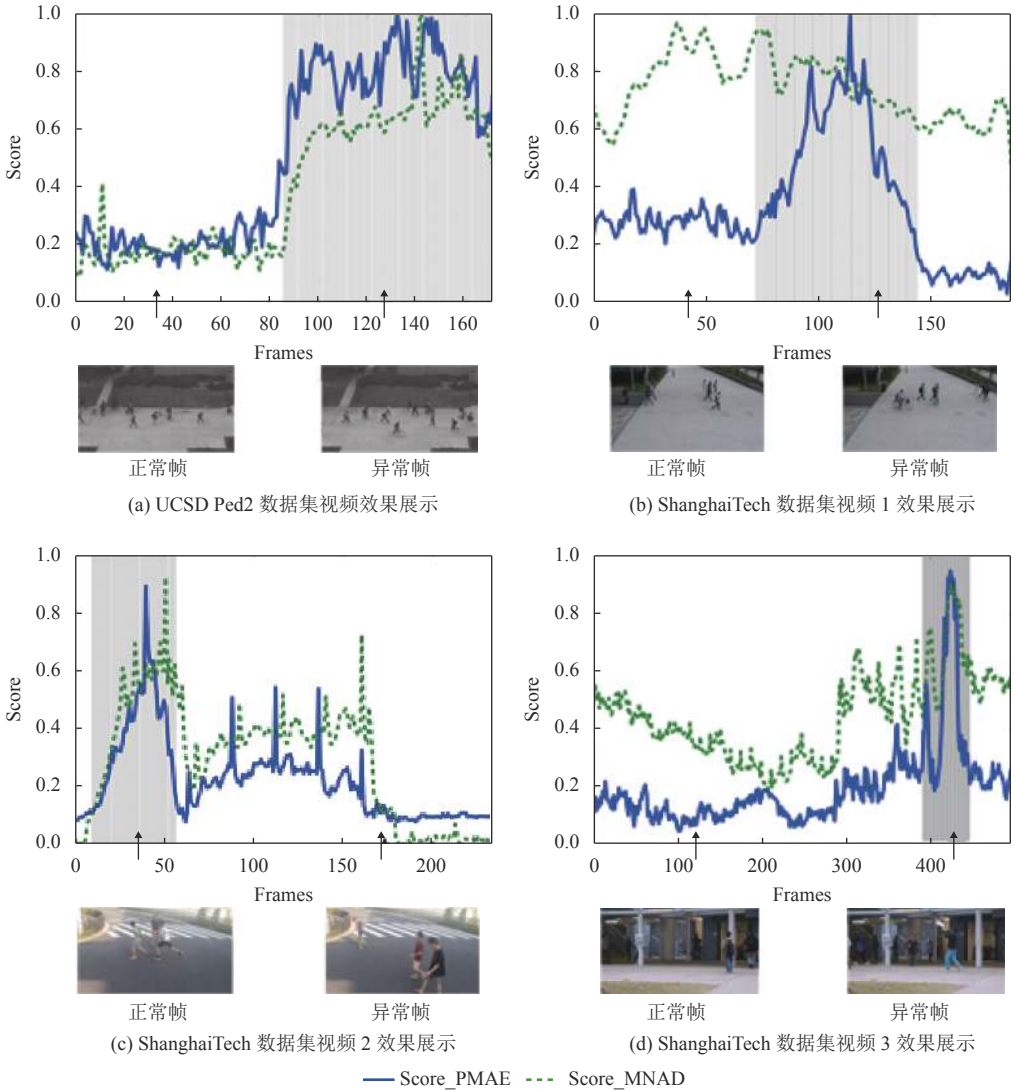
图 8　公开数据集不同算法 ROC 曲线

3.5　实验效果展示

算法训练完成后, 对于输入的视频帧组, 每一帧都会输出相应的一个异常得分, 以此判断是否产生异常. 图 9 列举了 UCSD Ped2 数据集和 ShanghaiTech 数据集中部分视频的异常得分图, 可以更直观地展示算法效果. 对比算法采用 MNAD[3]. 有色矩形代表真实帧在此处为异常帧. 从图 9 可以看出本文算法相较于对比算法的优越性. 若算法在异常片段打分越高, 则性能越好; 在正常片段打分越低, 性能也越好. 图 9(a) 取自 UCSD Ped2 数据集, 异常事件为骑自行车; 本文算法相较于对比算法在异常片段异常得分更高. 图 9(b) 取自 ShanghaiTech 数据集, 异常事件为骑自行车; 在对比算法虚警的情况下, 本文算法能很好地定位异常事件发生的时间. 图 9(c) 取自 ShanghaiTech 数据集, 异常事件为奔跑打闹; 本文算法异常分数曲线在正常片段出现尖刺, 但整体效果优于对比算法. 图 9(d) 取自 ShanghaiTech 数据集, 异常事件为高空抛物; PMAE 算法在部分异常片段的异常得分没有对比算法高, 但在正常数据片段的异常得分相较于对比算法更低, 且得分变化趋势更加明显. 通过在不同场景下视频的效果对比, 展示了本文所提算法的优越性和实用性.

图 9 的曲线对比显示两种算法在打分结果上均有各自效果好的地方. 为了定量描述本文算法对于正常行为和异常行为分类的效果, 将 PMAE 和 MNAD 两种算法在 UCSD Ped2 和 ShanghaiTech 数据集上分别计算了 F_1 值如表 4 所示. 从检测精度和虚警率的角度来看, PMAE 算法在性能上要优于 MNAD, 表明本文算法对于正常行为和异常行为的分类性能也有所提升, 对于多场景和多类别的异常行为检测有较强的泛化性和鲁棒性.

PMAE 算法基于重构网络进行未来帧的预测, 为了更直观地展现网络对于正异常帧所表现出来的不同的重构能力. 图 10 以 4 种场景为代表, 将网络重构的未来帧与真实帧进行可视化, 并比较他们之间的差值. 图 10 最上两行为 ShanghaiTech 数据集中的场景, 第 3 行和第 4 行为 UCSD Ped2 数据集中的场景; 图中左边 3 列分别为正常帧的真实帧、预测帧、差分帧, 右边 3 列分别为异常帧的真实帧、预测帧及差分帧. 从图 10 可以看出, 所提算法网络对正常帧可以以减小的重建误差进行重建, 而对于异常帧中的异常目标则重建误差较大, 即图 10 右边异常帧的差分帧白色区域更多且差别更加明显. 对于正常帧中的缓慢运动目标或背景, 网络在重建时会有极小的误差, 但误差值远小于异常目标的重建误差值; 而对于异常目标, 并非整体目标都不能重建, 只是误差的连通域相较于正常目标的误差连通域要大, 这也侧面反映出神经网络强大的重构能力, 是很多其他算法漏检率高的根本原因. 总体来说, 本文所提算法对于正常帧和异常帧的重建效果不同, 是算法能很好地区分正异常行为的前提和保证.



(a) UCSD Ped2 数据集视频效果展示

(b) ShanghaiTech 数据集视频 1 效果展示

(c) ShanghaiTech 数据集视频 2 效果展示

(d) ShanghaiTech 数据集视频 3 效果展示

图 9　实验效果展示

表 4　数据集 F_1 值对比

算法	UCSD Ped2	ShanghaiTech
MNAD (2020)[3]	0.924	0.600
PMAE_T9 (本文算法)	0.941	0.633

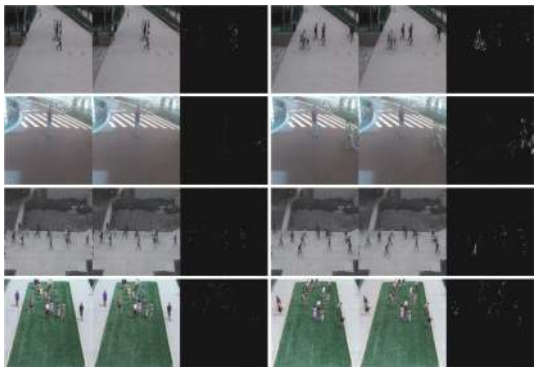


图 10　预测帧与真实帧对比

为了测试算法在现实场景中的应用效果, 将 FGAN[10]和 PMAE 两种算法在收集的 WHU 真实监控视频上进行离线分析, 得到的对比得分曲线如图 11 所示. 图 11 中第 1 行为 WHU 某地监控室内场景, 异常行为为行人奔跑, 第 2 行为室外场景, 异常行为为交通工具; 每行内左起第 1 幅图为真实视频帧, 第 2 幅图为 PMAE 算法的重建帧, 第 3 幅图为 FGAN 算法针对对应视频的异常分数变化曲线, 第 4 幅图为 PMAE 算法的异常分数变化曲线. 在室内场景异常行为发生时, FGAN 算法产生了剧烈的震荡, 这是网络过拟合的表现, 而 PMAE 算法则整体趋势较好, 重构的预测帧在异常行为区域出现模糊, 相较于真实帧差别明显, 表明检测出了异常; 在室外场景异常行为发生时, 两种算法都能检测出异常, 并且当小车运动速度低时的异常分数要略微低于小车运动速度高时的异常分数. 总体来说, 两种算法在实际监控视频中性能表现良好, 且 PMAE 算法的效果要优于 FGAN 算法的效果.

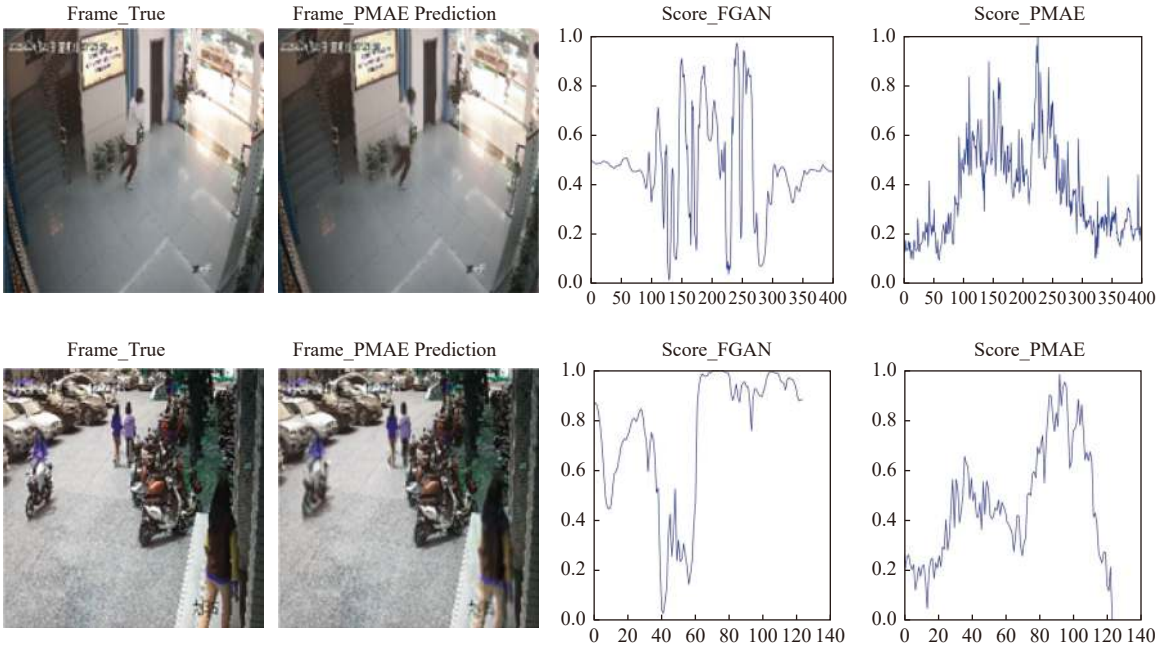


图 11 监控视频分析结果

4 结 论

针对正常行为数据和异常行为数据极度不均衡的问题, 本文从数据量大且易于获取的正常行为数据出发, 设计了基于概率记忆模型的半监督异常行为检测算法. 通过学习正常行为分布, 将偏离正常分布的数据判为异常行为数据. 算法的训练集全部为正常视频数据, 因此称之为半监督学习. 算法主干采用自编码网络进行视频帧重建, 在设计时使用三维因果卷积和时间维度共享全连接层来保证数据的时序性. 为了更好地拟合正常行为分布, 从概率熵和正常行为数据模式多样性的两个角度设计了辅助模块, 辅助主干网络进行视频帧重建. 概率模型利用自回归拟合输入数据分布, 促使模型收敛于正常分布的低熵状态. 记忆模块存储了历史数据中的正常行为原型特征, 实现正常行为多模式数据的共存, 同时避免主干网络的过度参与导致对异常视频帧的重建. 在公开数据集上进行了客观指标的实验验证和结果分析, 并与经典算法进行比较. 结果显示, 所提算法的辅助模块在训练时提高了正常帧的重建效果, 降低了正常视频帧误检率, 实现了更好的性能. 此外, 在测试时, 算法对数据集中的正常帧和异常帧实现了不同的重构效果, 有利于区分异常帧, 从而提高了异常行为的检测率. 最后, 算法在实际监控视频数据中进行了测试, 对于检测出异常行为也有一定的效果.

致谢 本论文的数值计算得到了武汉大学超级计算中心的计算支持和帮助.

References:

[1] Lee S, Kim HG, Ro YM. BMAN: Bidirectional multi-scale aggregation networks for abnormal event detection. IEEE Trans. on Image Processing, 2019, 29: 2395–2408. [doi: 10.1109/TIP.2019.2948286]

[2] Abati D, Porrello A, Calderara S, Cucchiara R. Latent space autoregression for novelty detection. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 481–490. [doi: 10.1109/CVPR.2019.00057]

[3] Park H, Noh J, Ham B. Learning memory-guided normality for anomaly detection. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 14360–14369. [doi: 10.1109/CVPR42600.2020.01438]

[4] Wang ZG, Zhang YJ. Anomaly detection in surveillance videos: A survey. Journal of Tsinghua University (Science and Technology), 2020, 60(6): 518–529 (in Chinese with English abstract). [doi: 10.16511/j.cnki.qhdxxb.2020.22.008]

[5] Xiao T, Zhang C, Zha HB, Wei FY. Anomaly detection via local coordinate factorization and spatio-temporal pyramid. In: Proc. of the 12th Asian Conf. on Computer Vision. Singapore: Springer, 2014. 66–82. [doi: 10.1007/978-3-319-16814-2_5]

[6] Reddy V, Sanderson C, Lovell BC. Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. In: Proc. of the 2011 CVPR Workshops. Colorado: IEEE, 2011. 55–61. [doi: 10.1109/CVPRW.2011.5981799]

[7] Dollar P, Rabaud V, Cottrell GW, Belongie S. Behavior recognition via sparse spatio-temporal features. In: Proc. of the 2005 IEEE Int'l Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. Beijing: IEEE, 2005. 65–72. [doi: 10.1109/VSPETS.2005.1570899]

[8] Xiao JS, Shen MY, Lei JF, Zhou JL, Klette R, Sui HG. Single image dehazing based on learning of haze layers. Neurocomputing, 2020, 389: 108–122. [doi: 10.1016/j.neucom.2020.01.007]

[9] Wang SQ, Zeng YJ, Liu Q, Zhu CZ, Zhu E, Yin JP. Detecting abnormality without knowing normality: A two-stage approach for unsupervised video abnormal event detection. In: Proc. of the 26th ACM Int'l Conf. on Multimedia. Seoul: ACM, 2018. 636–644. [doi: 10.1145/3240508.3240615]

[10] Xiao JS, Shen MY, Jiang MJ, Lei JF, Bao ZY. Abnormal behavior detection algorithm with video-bag attention mechanism in surveillance video. Acta Automatica Sinica, 2022, 48(12): 2951–2959 (in Chinese with English abstract). [doi: 10.16383/j.aas.c190805]

[11] Zhong JX, Li NN, Kong WJ, Liu S, Li TH, Li G. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 1237–1246. [doi: 10.1109/CVPR.2019.00133]

[12] Feng YC, Yuan Y, Lu XQ. Learning deep event models for crowd anomaly detection. Neurocomputing, 2017, 219: 548–556. [doi: 10.1016/j.neucom.2016.09.063]

[13] Amraee S, Vafaei A, Jamshidi K, Adibi P. Anomaly detection and localization in crowded scenes using connected component analysis. Multimedia Tools and Applications, 2018, 77(12): 14767–14782. [doi: 10.1007/s11042-017-5061-7]

[14] Zhang Y, Lu HC, Zhang LH, Ruan X. Combining motion and appearance cues for anomaly detection. Pattern Recognition, 2016, 51: 443–452. [doi: 10.1016/j.patcog.2015.09.005]

[15] Colque RVHM, Caetano C, de Andrade MTL, Schwartz WR. Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. IEEE Trans. on Circuits and Systems for Video Technology, 2017, 27(3): 673–682. [doi: 10.1109/TCSVT.2016.2637778]

[16] Ma K, Doescher M, Bodden C. Anomaly detection in crowded scenes using dense trajectories. University of Wisconsin-Madison, 2015. https://pages.cs.wisc.edu/~kma/downloads/anomaly-detection.pdf

[17] Ionescu RT, Khan FS, Georgescu MI, Shao L. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7834–7843. [doi: 10.1109/CVPR.2019.00803]

[18] Ramachandra B, Jones MJ, Vatsavai RR. Learning a distance function with a Siamese network to localize anomalies in videos. In: Proc. of the 2020 IEEE Winter Conf. on Applications of Computer Vision. Snowmass: IEEE, 2020. 2587–2596. [doi: 10.1109/WACV45572.2020.9093417]

[19] Liu W, Luo WX, Lian DZ, Gao SH. Future frame prediction for anomaly detection-a new baseline. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6536–6545. [doi: 10.1109/CVPR.2018.00684]

[20] Vu H, Nguyen TD, Le T, Luo W, Phung D. Robust anomaly detection in videos using multilevel representations. In: Proc. of the 2019 AAAI Conf. on Artificial Intelligence. Honolulu: Association for the Advancement of Artificial Intelligence (AAAI), 2019. 5216–5223. [doi: 10.1609/aaai.v33i01.33015216]

[21] Akcay S, Atapour-Abarghouei A, Breckon TP. GANomaly: Semi-supervised anomaly detection via adversarial training. In: Proc. of the 14th Asian Conf. on Computer Vision. Perth: Springer, 2018. 622–637. [doi: 10.1007/978-3-030-20893-6_39]

[22] Hasan M, Choi J, Neumann J, Roy-Chowdhury AK, Davis LS. Learning temporal regularity in video sequences. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 733–742. [doi: 10.1109/CVPR.2016.86]

[23] Yuan J, Zhang YJ. App1ication of sparse denoising auto encoder network with gradient difference information for abnormal action detection. Acta Automatica Sinica, 2017, 43(4): 604–610 (in Chinese with English abstract). [doi: 10.16383/j.aas.2017.c150667]

[24] Chong YS, Tay YH. Abnormal event detection in videos using spatiotemporal autoencoder. In: Proc. of the 14th Int'l Symp. on Neural Networks. Hokkaido: Springer, 2017. 189–196. [doi: 10.1007/978-3-319-59081-3_23]

[25] Shi XJ, Chen ZR, Wang H, Yeung DY, Wong WK, Woo WC. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In: Proc. of the 28th Int 'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2015. 802–810. [doi: 10.5555/2969239.2969329]

[26] Luo WX, Liu W, Gao SH. A revisit of sparse coding based anomaly detection in stacked RNN framework. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 341–349. [doi: 10.1109/ICCV.2017.45]

[27] Loshchilov I, Hutter F. SGDR: Stochastic gradient descent with warm restarts. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[28] Sabokrou M, Fayyaz M, Fathy M, Klette R. Deep-cascade: Cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Trans. on Image Processing, 2017, 26(4): 1992–2004. [doi: 10.1109/TIP.2017.2670780]

[29] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters, 2006, 27(8): 861–874. [doi: 10.1016/j.patrec.2005.10.010]

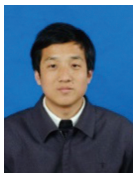
[30] Borji A, Cheng MM, Jiang HZ, Li J. Salient object detection: A benchmark. IEEE Trans. on Image Processing, 2015, 24(12): 5706–5722. [doi: 10.1109/TIP.2015.2487833]

附中文参考文献:

[4] 王志国, 章毓晋. 监控视频异常检测: 综述. 清华大学学报(自然科学版), 2020, 60(6): 518–529. [doi: 10.16511/j.cnki.qhdxxb.2020.22.008]

[10] 肖进胜, 申梦瑶, 江明俊, 雷俊峰, 包振宇. 融合包注意力机制的监控视频异常行为检测. 自动化学报, 2022, 48(12): 2951–2959. [doi: 10.16383/j.aas.c190805]

[23] 袁静, 章毓晋. 融合梯度差信息的稀疏去噪自编码网络在异常行为检测中的应用. 自动化学报, 2017, 43(4): 604–610. [doi: 10.16383/j.aas.2017.c150667]



肖进胜(1975－), 男, 博士, 副教授, CCF 高级会员, 主要研究领域为计算机视觉, 图像处理与分析.

赵陶(1996－), 男, 硕士, 主要研究领域为图像处理与分析.

郭浩文(1999－), 男, 硕士生, 主要研究领域为图像处理与分析.

申梦瑶(1994－), 女, 硕士, 主要研究领域为图像处理.

谢红刚(1975－), 男, 博士, 副教授, 主要研究领域为人工智能, 图像处理与分析, 深度学习.

王元方(1997－), 男, 硕士, 主要研究领域为计算机视觉, 图像处理与分析.