

基于记忆增强和扩散模型的骨架视频异常检测*

张栖, 李元[†], 周彦钊, 焦建彬

(中国科学院大学电子电气与通信工程学院, 北京 100049)

(2024 年 11 月 14 日收稿; 2025 年 4 月 2 日收修改稿)

Zhang Q, Li Y, Zhou Y Z, et al. Skeleton-based video anomaly detection with memory enhancement and diffusion model[J]. Journal of University of Chinese Academy of Sciences, 2026, 43(5): 677-686. DOI: 10. 7523/j.ucas. 2025. 017.

摘要 传统的视频异常检测(VAD)方法主要依赖于从视频帧的整体或特定区域提取像素特征。为了降低非结构化噪声的影响,基于人体骨架的VAD方法备受关注。然而在开集场景下,此类方法面临2大挑战:对异常行为的错误重构及对正常行为拓展能力的不足。针对这个问题,提出一种基于记忆增强与扩散模型的骨架视频异常检测框架(MEDM-SVAD)。该框架通过引入记忆增强模块扩展模型对正常样本的记忆能力,避免异常行为因重构误差较小而被误判;此外,利用扩散模型显著提高对域外正常行为的泛化能力。在HR-STC、HR-Avenue和HR-UBnormal这3个公开数据集上的实验结果表明,相较于现有最先进的MoCoDAD算法, MEDM-SVAD在AUC指标上分别取得78.5%、90.1%和69.7%的结果,均有不同程度的性能提升,验证了其在多种场景下的有效性与优越性。

关键词 视频异常检测;骨架特征;扩散模型;记忆增强

中图分类号: TP181 文献标志码: A DOI: 10. 7523/j.ucas. 2025. 017

Skeleton-based video anomaly detection with memory enhancement and diffusion model

ZHANG Qi, LI Yuan, ZHOU Yanzhao, JIAO Jianbin

(School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences,
Beijing 100049, China)

Abstract Video anomaly detection (VAD) is a technique aimed at identifying abnormal events within videos and has significant applications in public safety and video content understanding. Traditional VAD methods primarily rely on extracting pixel features from either the entire video frame or specific regions. To reduce the impact of unstructured noise, VAD methods based on human skeletal data have garnered considerable attention. However, in open-set scenarios, these methods face two major challenges: erroneous reconstruction of abnormal behaviors and insufficient generalization to diverse normal behaviors. To address these issues, this paper proposes a skeleton-

* 中国科学院战略先导A项目(XDA27000000)资助

[†] 通信作者, E-mail: liyuan23@ucas.ac.cn

based video anomaly detection framework with memory enhancement and diffusion model (MEDM-SVAD). This framework incorporates a memory enhancement module to expand the model's memory capacity for normal samples, preventing abnormal behaviors from being misclassified due to minimal reconstruction error. Additionally, the diffusion model significantly improves the framework's ability to generalize to out-of-domain normal behaviors. Experimental results on three public datasets, HR-STC, HR-Avenue, and HR-UBnormal, show that the proposed method MEDM-SVAD achieves AUC scores of 78.5%, 90.1%, and 69.7%, respectively, demonstrating various levels of performance improvement over the current state-of-the-art MoCoDAD algorithm, thus verifying its effectiveness and superiority across multiple scenarios.

Keywords video anomaly detection; skeleton-based features; diffusion model; memory enhancement

视频异常检测(video anomaly detection, VAD)因其具有检测人类异常行为(如公共场所打架斗殴、道路交通事故、骚乱和踩踏事件^[1])的潜力而受到极大关注。传统的异常检测方法主要依赖从视频帧的整体^[2-6]或特定^[7-8]区域提取像素特征。然而,这种方法常常会弱化关键场景信息,并对非结构化噪声(如光照变化、监控角度差异及背景杂乱等)表现出较高的敏感性^[9],导致在精确提取目标对象特征时带来冗余信息^[10]。目前,基于人体骨架的VAD方法^[10-22]逐渐受到关注。这些方法通过姿态检测与跟踪算法提取目标的骨架特征,以减轻背景噪声的干扰,重点关注目标对象本身。此外,在VAD中发生异常事件的频率远低于正常事件,这导致了正负样本极度不平衡。因此利用监督学习,模型往往难以准确识别大量未知的异常行为。目前,主流方法是将异常检测视为一个开集问题^[23],即在训练阶段仅使用有具体标签的正常行为数据;在测试阶段,通过检测与已知正常行为的偏离来识别未知异常。闭集识别在固定的类集上训练和测试模型,与闭集问题不同,开集识别解决了模型需要区分已知类和未知类的更现实的情况。在开集场景,基于重构^[24-27]和预测^[28-29]的方法是当前VAD中2种主要的范式。基于重构的方法是利用自编码器对正常行为数据进行重构误差最小化训练,再在测试阶段通过重构误差来检测异常;基于预测的方法则依赖于行为数据的时间特性,训练一个网络模型,利用先前的行为数据预测未来的行为,并通过预测误差来识别异常。虽然这2种方法都能解决样本不平衡的问题,但目前仍存在2点局限。第一,当异常行为与正常行为具有相似特征时,在推理阶段可能仍然会被较好地重构或预测^[30],从而导致异常检测准确性的下降;第二,除已知类别外的正常行为为无

法拓展训练,导致模型学习能力有限。如何扩展模型的学习范围,使其不仅能识别训练样本中正常行为,还能识别训练样本外的正常行为^[11, 31],这也是开集场景中的关键研究问题。目前,基于人体骨架VAD方法重点在于改善人体骨架数据的建模方式,尚未充分考虑开集场景中这两方面的挑战。

因此,为应对错误重构和预测的挑战,本文提出基于记忆增强与扩散模型的骨架VAD框架(memory-enhanced and diffusion model-based skeleton video anomaly detection framework, MEDM-SVAD),构建重构和预测模块统一进行异常行为推理学习。具体而言,在重构任务中,构建一个针对骨架数据的自编码器,并在解码器的各个层级引入记忆单元。这些记忆单元作为模型参数的一部分,在训练过程中能够不断更新,以强化对正常行为的记忆,从而提升异常检测的准确度。在预测任务中,利用骨架数据的结构化特性,通过聚类方法,提前获得能够代表正常行为的记忆模式。另一方面,为了扩展模型识别未知正样本的能力,利用扩散模型^[32]构造预测网络,将重构任务学习到的观察数据特征作为条件信息,用于预测任务中引导未来数据的生成。在推理时,通过调整随机采样次数,使模型能够生成多种可能的行为模式,以拓展其识别边界。在正常情况下,生成的运动是多样且合理的,但在异常情况下,生成的运动虽然也是多样的,但缺乏相关性^[11, 33]。

1 研究方法

根据人体骨架特征,提出MEDM-SVAD框架,通过构建记忆增强模块解决重构或预测时训练样本过拟合的问题,利用扩散模型扩展模型学习范围,总体框架如图1所示。将一个 N 帧视频段的

连续行为表示为 $X_a = \{X^1, \dots, X^N\}$, 其中每一帧行为 X^i 都可以表示成一个图 $X^i = (J, A)$, J 表示骨架节点组成的集合, A 表示节点连边关系, 节点的属性由 $\mathbb{R}^2$ 的空间坐标来表示。将 X_a 分为两部分: $X^{1:K}$ 为观察数据, $X^{K+1:N}$ 为未来数据。在此基础上, MEDM-SVAD 的训练分为 2 个阶段。第 1 阶段是对观察数据进行重构学习。这个阶段只将观察数据输入到重构网络, 而不输入未来数据, 其目的是确保该网络能专注于学习观察数据中骨架的原始特征表示, 而不受预测任务中可能的噪声和变化的干扰。第 1 阶段训练完成后, 固定重构网络的参数, 将重构任务学习到的特征, 作为第 2 阶段预测任务中扩散模型的条件用以指导未来数据从噪声中恢复。

1.1 记忆增强的重构网络

为对观察数据进行重构, 构造一个自编码器对数据进行条件编码, 以学习观察数据的高维嵌入。由于骨架数据在时间和空间维度上具有异质性, 且节点之间的影响具有空间依赖性和时间关联性, 传统的卷积神经网络^[34]使用共享的模块无法有效捕获时空网络中的异质性, 因此模型采用时空可分离图卷积网络(space time separable graph convolutional network, STS-GCN)^[35]作为自编码器的基网, 以捕获节点间复杂的局部时空相关性。

为了保证正常行为模式能够被记忆, 同时避免异常行为被很好地重构, 在自编码器中添加记忆单元。这些记忆单元分布在自编码器的不同层

级之间, 以保证在各个特征级别上都能有效地存储和提取正常行为模式。每个记忆单元表示为一个矩阵 $M \in \mathbb{R}^{L \times D}$, 矩阵的每一行 m_i 表示一个记忆槽。每一层的记忆单元都能从上一层的输出中接收信息, 并作为查询向量 $V_q \in \mathbb{R}^{1 \times D}$ 来检索相似的记忆槽。这些记忆槽随后组合成新的编码特征用于解码器重建。检索的过程采用注意力的方式, 表示如下

$$V_m = wM = \sum_{i=1}^L w_i m_i. \quad (1)$$

其中, 注意力系数 $w \in \mathbb{R}^{1 \times L}$ 是一个行向量, 由查询向量 V_q 和记忆单元 M 的转置相乘得到; w_i 则表示 w 的第 i 个分量。这种多层记忆结构不仅有助于保留更多有用的正常行为信息, 而且还能在测试阶段通过与存储的记忆比较来识别和放大异常行为的重建误差。

需要注意的是, 过多的记忆单元可能会导致过度的信息过滤, 降低网络记忆最具代表性的正常模式能力。因此使用跳跃连接^[36], 在编码器和解码器之间直接传递编码信息, 有助于改善自动编码器的重构质量, 如图 2 所示。

在训练阶段, 重构网络采用编码器 E 将观察数据嵌入到高维空间, 随后由解码器 D 进行重构, 并通过最小化重构误差来优化网络参数

$$L_{\text{rec}} = \|D(E(X^{1:K})) - X^{1:K}\|_2^2. \quad (2)$$

其中, $\|\cdot\|_2^2$ 表示 L_2 范数平方。另外借鉴先前工作^[24], 针对注意力系数 w 添加交叉熵损失, 如下所示

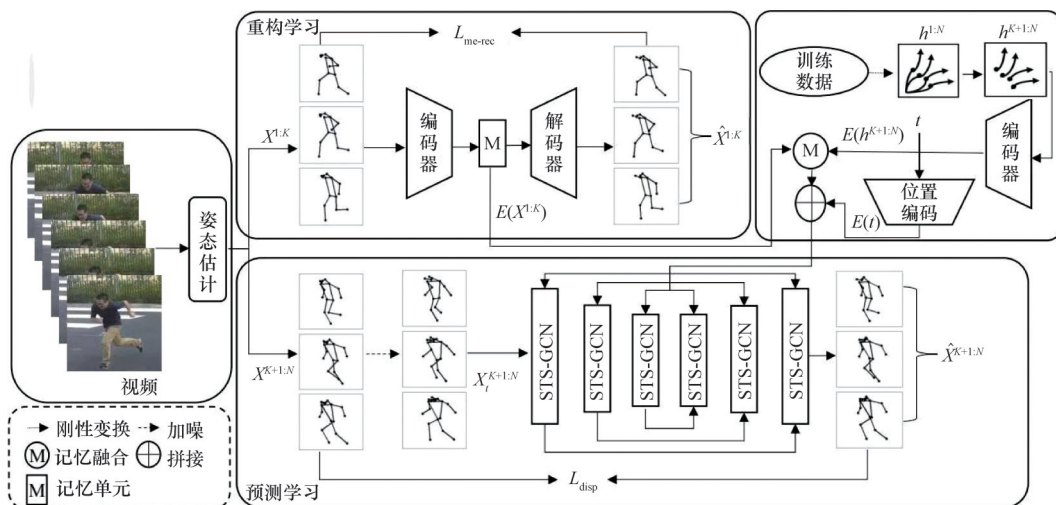


图1 MEDM-SVAD 总体框架图

Fig. 1 Overall framework of MEDM-SVAD

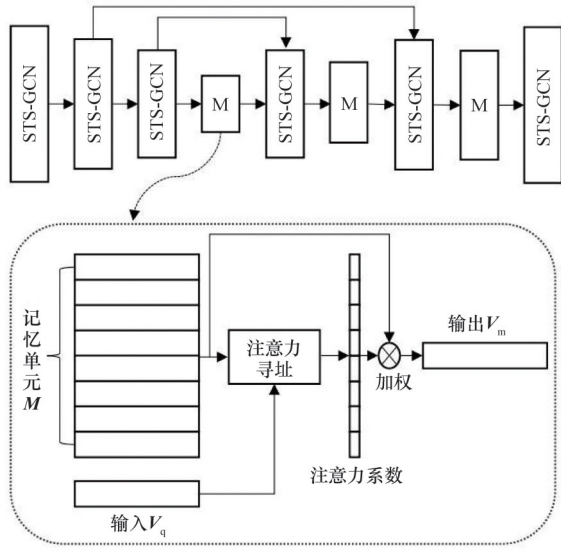


图2 重构网络结构图

Fig. 2 Architecture of reconstructed network

$$L_{\text{ent}} = \sum_{i=1}^O \sum_{j=1}^L -w_{i,j} \log(w_{i,j}). \quad (3)$$

其中, O 表示记忆单元数目, $w_{i,j}$ 表示第 i 个单元的第 j 个记忆槽的系数。最终得到总体损失函数

$$L_{\text{me-rec}} = \lambda_1 L_{\text{rec}} + \lambda_2 L_{\text{ent}}. \quad (4)$$

其中, λ_1 和 λ_2 分别表示重构损失和交叉熵损失的平衡权重。为确保重构质量, 重构损失需要在总体损失中占主导地位, 因此 λ_1 设置为 1.0, λ_2 为 0.0002。

1.2 记忆增强的预测网络

基于扩散模型的预测网络负责还原噪声数据。利用 U-Net^[37] 作为扩散模型的网络部分, 其输入包括 4 部分: 噪声数据 $\mathbf{X}_t^{K+1:N}$ 、观察数据的高维嵌入 $E(\mathbf{X}^{1:K})$ 、时间嵌入 $E(t)$ 以及代表正常行为的记忆模式的嵌入 $E(\mathbf{h}^{K+1:N})$ 。

1.2.1 输入数据融合

与重构网络的单一输入不同, 预测网络的输入数据更加多样且具有噪声, 这使得构造一种可以随着训练而不断优化的记忆单元变得困难, 因此需要通过预先设定的方式来获取。

首先, 对用于训练的骨架数据做刚性变换(平移和旋转)^[38], 将这些数据限制在一个有限的空间区域, 但不破坏其本身所包含的运动特征, 如图 3 所示。具体而言, 对于一条骨架数据 $\mathbf{X}_a$, 以未来数据的第 1 帧 $\mathbf{X}^{K+1}$ 为中心点, 将整个骨架平移, 直到中心点位于坐标原点为止。这一步骤是为了消除骨架序列在空间中的绝对位置信息, 从

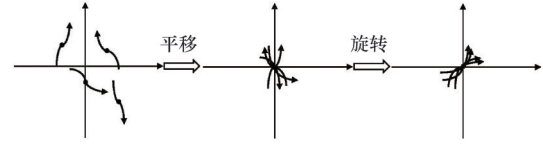


图3 刚性变换示意图

Fig. 3 Illustration of rigid transformation

而降低对初始位置差异的敏感性。接着保持中心点固定, 将整个骨架旋转, 直到最后一帧 $\mathbf{X}^N$ 位于正轴。此操作旨在调整骨架序列的方向, 进一步规范不同序列之间的位置差异。通过刚性变换, 骨架序列在空间坐标的差异被标准化, 使得具有相似运动模式的序列在后续的分析中能够显示出更小的距离。

然后, 使用 K 均值算法将标准化后的数据进行统计聚类, 得到代表正常行为的记忆模式。其中每个记忆模式在数据维度上与骨架序列一致, 记为 $\mathbf{H}_c = \{\mathbf{h}^1, \dots, \mathbf{h}^N\}$ 。在进行预测任务时, 选取记忆模式的未来数据部分 $\mathbf{h}^{K+1:N}$, 并使用重构网络的编码器 E 将其嵌入到高维空间, 得到记忆模式的嵌入 $E(\mathbf{h}^{K+1:N})$ 。因为记忆模式和重构网络的记忆单元实际上是使用同一批数据得到的, 所以用重构网络的自编码器去嵌入记忆模式, 并不会破坏记忆模式的特征信息。最后, 将记忆模式的嵌入 $E(\mathbf{h}^{K+1:N})$ 和观察数据的嵌入 $E(\mathbf{X}^{1:K})$ 进行记忆融合, 具体操作如下: 将骨架数据的观察嵌入 $E(\mathbf{X}^{1:K})$ 与记忆模式所对应的嵌入 $E(\mathbf{h}^{K+1:N})$ 拼接。再通过线性降维和池化, 获得信息更加紧凑密集的融合嵌入。

1.2.2 训练和推理

扩散模型的基本思想是通过一系列逐步的随机过程将数据从其原始状态逐渐转换为纯噪声状态, 然后再通过一个逆过程将噪声数据恢复到原始数据。正向扩散过程是一个马尔可夫链, 逐步将数据 $\mathbf{x}_0$ 通过添加高斯噪声转化为噪声数据 $\mathbf{x}_T$ 。每一步的噪声添加过程可以表示为

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{a_t} \mathbf{x}_{t-1}, (1 - a_t) \mathbf{I}). \quad (5)$$

其中, $t \sim U(1, T)$ 是当前扩散的时间步数, $a_t \in (0, 1)$ 表示第 t 步的噪声方差, $\mathbf{I}$ 为单位矩阵, T 为扩散的总步数。另外, 通过重参数化技巧, $\mathbf{x}_t$ 可以由 $\mathbf{x}_0$ 和噪声 $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I})$ 直接计算得到, 过程如下

$$\mathbf{x}_t = \sqrt{\bar{a}_t} \mathbf{x}_0 + \sqrt{1 - \bar{a}_t} \boldsymbol{\varepsilon}. \quad (6)$$

其中, $\bar{a}_t = \prod_{i=1}^t a_i$ 。反向扩散过程是从噪声 $\mathbf{x}_T$ 逐

步恢复数据 $\mathbf{x}_0$ 。反向扩散过程也是一个马尔可夫链,通过学习一个参数化的高斯分布来近似真实地反向分布,表示如下

$$p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t) = N(\mathbf{x}_{t-1}; u_{\theta}(\mathbf{x}_t, t), \sum_{\theta}(\mathbf{x}_t, t)). \quad (7)$$

其中, $u_{\theta}(\mathbf{x}_t, t)$ 表示模型预测的均值; $\sum_{\theta}(\mathbf{x}_t, t)$ 是模型预测的方差,通常可以固定为 $(1 - a_t)\mathbf{I}$ 。

本文中,扩散模型用于构造预测网络。在正向扩散过程中,从高斯分布 $N(0, \mathbf{I})$ 中采样随机位移 $\boldsymbol{\varepsilon}^{K+1:N} \in \mathbb{R}^{2 \times |U|}$,并将其加到 $\mathbf{X}^{K+1:N}$,如下式所示

$$\mathbf{X}_t^{K+1:N} = \sqrt{a_t} \mathbf{X}^{K+1:N} + \sqrt{1 - a_t} \boldsymbol{\varepsilon}^{K+1:N}. \quad (8)$$

其中,输入数据 $\mathbf{X}^{K+1:N}$ 对应式(6)中的 $\mathbf{x}_0, \boldsymbol{\varepsilon}^{K+1:N}$ 对应 $\boldsymbol{\varepsilon}$ 。

在反向扩散过程中,通过预测网络 ε_{θ} 去估计随机位移,从而去除噪声。为了能够让估计更加近似 $\boldsymbol{\varepsilon}^{K+1:N}$,将观察数据的高维嵌入 $E(\mathbf{X}^{1:K})$ 作为条件,用于引导网络的预测。预测网络的优化在形式上表示为最小化预测位移和真实位移之间的误差

$$L_{\text{disp}} = \left\| \boldsymbol{\varepsilon}^{K+1:N} - \varepsilon_{\theta}(\mathbf{X}_t^{K+1:N}, E(t), E(\mathbf{X}^{1:K}), E(\mathbf{h}^{K+1:N})) \right\|_2^2. \quad (9)$$

其中, $\mathbf{X}_t^{K+1:N}$ 直接输入到预测网络中, $E(t)$ 、 $E(\mathbf{X}^{1:K})$ 和 $E(\mathbf{h}^{K+1:N})$ 则与网络隐藏层的输出进行相加融合。

在推理阶段,首先获取观察数据的高维嵌入 $E(\mathbf{X}^{1:K})$;然后从高斯分布中采样随机噪声 $\mathbf{z}^{K+1:N} \sim N(0, \mathbf{I})$,作为降噪的初始状态;再通过预测网络逐步合成未来数据。根据式(7),可得到每一步的降噪过程,如下所示

$$\mathbf{X}_{t-1}^{K+1:N} = u_{\theta}(\mathbf{X}_t^{K+1:N}, t) + \mathbf{z}^{K+1:N} \sqrt{1 - a_t}. \quad (10)$$

其中均值 $u_{\theta}(\mathbf{X}_t^{K+1:N}, t)$ 的设计是为了最大化恢复原始数据的可能性,在形式上可以表示为从 $\mathbf{X}_t^{K+1:N}$ 中减去预测的噪声

$$u_{\theta}(\mathbf{X}_t^{K+1:N}, t) = \frac{1}{\sqrt{a_t}} \left(\mathbf{X}_t^{K+1:N} - \frac{1 - a_t}{\sqrt{1 - a_t}} \varepsilon_{\theta}(\mathbf{X}_t^{K+1:N}, E(t), E(\mathbf{X}^{1:K}), E(\mathbf{h}^{K+1:N})) \right). \quad (11)$$

为了能够生成更多合理的未来数据,进行 P 次上述采样过程,得到的结果记为 $\mathbf{Z}_1^{K+1:N}, \dots, \mathbf{Z}_P^{K+1:N}$ 。最后计算所有生成数据和真实数据的误差

$$s_i = \left\| \mathbf{X}^{K+1:N} - \mathbf{Z}_i^{K+1:N} \right\|_2^2. \quad (12)$$

并将所有误差的平均值作为个体的异常分数。这里需要指出,在同一帧中可能会出现多个行人,这种情况下取所有行人中异常分数的最大值作为整个帧的异常分数。

2 实验与分析

2.1 数据集

采用 VAD 领域常用的 3 个公开数据集 HR-STC^[10]、HR-Avenue^[10] 和 HR-UBnormal^[16] 进行评估实验,所使用的骨架数据通过姿态估计和跟踪算法获取^[39]。

3 个数据集在原数据集上都过滤了与人无关的异常事件。其中 HR-STC 是目前最大的数据集之一,涵盖 13 个不同的摄像头拍摄的 300 000 帧,经过过滤后的测试集包含 130 个异常事件。HR-Avenue 是在香港中文大学的步行街道采集,包含 16 个训练视频和 21 个测试视频,经过过滤后包含 47 个异常事件。HR-UBnormal 是通过 Cinema4D 软件从 2D 自然图像合成的 29 个场景,每个场景有 19 个视频片段,包含正常和异常事件。

2.2 评估指标

为了对 MEDM-SVAD 性能进行全面评估,采用受试者工作特征曲线下面积(receiver operating characteristic area under the curve, ROC-AUC)、召回率(recall, R)、精确度(precision, P)和 F 度量(F-measure, F)4 个评价指标进行实验验证。其中,正类定义为异常行为,负类定义为正常行为。ROC-AUC 用于衡量模型在不同阈值下对异常行为的区分能力; R 反映模型对异常行为的检测覆盖率; P 表示模型预测为异常的样本中真正异常样本所占的比例;而 F 作为 P 和 R 的调和平均,综合评价模型的检测性能。

2.3 数据预处理与实验参数

将每个人的运动划分成视频片段,每个片段固定分为 6 帧,其中前 3 帧作为观察数据,后 3 帧为未来数据。模型训练使用 Adam 优化器,初始学习率为 $1e^{-4}$,第 1 阶段的重构网络训练 50 个周期后衰减 0.8,然后加载并冻结重构网络的参数,在第 2 阶段对整个框架训练的学习率指数衰减 20 个训练周期。扩散模型的扩散次数 T 为 10,噪声控制因子 α_1 和 α_T 分别设为 0.999 9 和 0.02。异常分数的判断阈值为 0.5。

2.4 基于骨架的 VAD 算法的性能验证

使用6个基于骨架的异常检测算法和本文所提出的 MEDM-SVAD 进行性能验证。其中,MPED-RNN^[10]从人体运动的全局和局部视角出发提取骨架数据的特征表示,并使用循环神经网络^[40]构造相互独立的重构和预测双分支网络。MT-Prediction^[14]从不同时间尺度上预测过去和未来数据,再使用全连接层将不同时间尺度的数据结合。BiPOCO^[17]使用条件变分自编码器(conditioned variational-autoencoder, CVAE)学习正常行为的后验分布,并使用编码的观察数据来预测未来序列。Normal Graph^[19]首次使用 STS-GCN 对骨骼序列进行编码,与直接将骨架坐标输入到网络的方式相比,能够更明确地分析骨架的空间模式。COSKAD^[16]基于 STS-GCN 编码,并将正常行为的嵌入映射到隐空间的狭窄区域,区域之外的行为则被视为异常。MoCoDAD^[11]使用 STS-GCN 将观察数据编码为条件嵌入,利用扩散模型预测未来数据,并通过误差损失来发现异常。

在3个公开数据集上进行比较后,实验结果如表1所示(表中R-A表示ROC-AUC,下同)。由于MT-Prediction和Normal Graph的相关代码未开源,因此本文只引用其论文中报告的ROC-AUC结果。对于其他未公开的评估指标(如*R*、*P*等),在表格中用“—”表示缺失数据。可以看出,STS-GCN 编码在骨架数据的特征提取方面展现了显著的优势。这种编码方式能够更有效地捕捉人体骨架在时空维度的复杂动态特征,从而提升对行为模式的识别能力。此外,MoCoDAD作为当前最先进的方法,也证明了扩散模型能够更好地捕捉数据分布,从而预测更加真实的骨架数据。

本文提出的 MEDM-SVAD 相比于 MoCoDAD,

增加了记忆增强模块以及重复采样,在 HR-STC 数据集上*P*和*R*分别提高5.5和3.3个百分点;在 HR-Avenue 数据集上*R*和*F*分别提高7.1和2.2个百分点;在3个数据集上的 ROC-AUC 分数分别提高0.9、1.1和1.3个百分点。这些结果表明,该方法在推理阶段对易错的异常行为和未见过的正常行为保持了较高的鲁棒性。

2.5 消融学习

为深入探索 MEDM-SVAD 中记忆增强模块和扩散模型的重要性,在 HR-STC 数据集上分别进行了消融研究。

2.5.1 记忆增强模块的验证

针对记忆增强模块,设置6种不同的方案,结果如表2所示,其中*w/o*分别表示加入和删除跳跃链接。

首先验证了加入记忆模块(add memory, AM)或删除记忆模块(delete memory, DM)对模型性能的影响。当不用记忆模块时,模型的 ROC-AUC 为 77.1%;而当记忆模块在预测网络中启用时,模型的 ROC-AUC 上升为 78.1%;而仅在重构网络中启用时,ROC-AUC 为 77.9%/77.5%。这表明,无论是在重构网络还是预测网络中加入记忆模块,都能提升模型性能,说明记忆模块能够有效增强模型对已知正常行为的记忆,从而减少异常行为的误检。此外,当记忆模块同时应用于重构和预测网络时,模型的 ROC-AUC 提升至 78.5%/78.4%。这一提升可归因于重构网络和预测网络中记忆的互补性:重构网络中的记忆模块通过学习捕捉隐式记忆,而预测网络中的记忆模块则通过统计聚类获取显式记忆。这2种记忆的结合共同增强了模型的整体性能。

其次,实验验证了跳跃链接对模型的影响。方案3和方案5相比,ROC-AUC、*R*、*P*、*F*分别提高

表1 HR-STC、HR-Avenue、HR-UBnormal 数据集上的性能比较
Table 1 Performance comparison on the HR-STC, HR-Avenue, and HR-UBnormal datasets %

模型	HR-STC				HR-Avenue				HR-UBnormal			
	R-A	<i>R</i>	<i>P</i>	<i>F</i>	R-A	<i>R</i>	<i>P</i>	<i>F</i>	R-A	<i>R</i>	<i>P</i>	<i>F</i>
MPED-RNN	72.7	60.7	74.7	69.2	86.3	73.3	89.4	81.9	61.2	49.3	64.1	55.3
MT-Prediction	77.0	—	—	—	88.3	—	—	—	—	—	—	—
Normal Graph	76.5	—	—	—	87.3	—	—	—	—	—	—	—
BiPOCO	74.9	68.3	76.8	70.9	87.0	76.3	88.3	84.1	52.3	43.7	56.1	50.3
COSKAD	77.1	71.3	79.3	73.2	87.8	79.2	90.5	86.3	65.5	51.4	67.1	62.9
MoCoDAD	77.6	70.8	78.2	76.4	89.0	78.1	91.6	85.1	68.4	59.3	70.2	64.1
MEDM-SVAD	78.5	74.1	83.7	77.2	90.1	85.2	92.8	87.3	69.7	63.7	72.3	67.1

表 2 记忆模块的消融学习

Table 2 Ablation study of the memory module

%

方法	重构网络			预测网络		R-A	R	P	F
	AM(<i>w</i>)	AM(<i>o</i>)	DM	AM	DM				
方案 1	—	—	✓	—	✓	77.1	63.1	78.9	69.2
方案 2	—	—	✓	✓	—	78.1	67.3	80.3	71.3
方案 3	✓	—	—	—	✓	77.9	71.9	81.1	73.8
方案 4	✓	—	—	✓	—	78.5	74.1	83.7	77.2
方案 5	—	✓	—	—	✓	77.5	70.3	79.7	70.6
方案 6	—	✓	—	✓	—	78.4	71.9	82.2	75.3

0.4、1.6、1.4 和 3.2 个百分点;方案 4 和方案 6 相比,ROC-AUC、*R*、*P*、*F* 分别提高 0.1、2.2、1.5 和 1.9 个百分点。这表明跳跃链接可以增加记忆模块对最具代表性的正常行为模式的记忆能力,从而更好地检测出观察数据中的异常行为。

2.5.2 扩散模型的验证

对于扩散模型,本文在预测网络中将其与 CVAE、掩码自编码器(masked autoencoder, MAE)以及生成对抗网络(generative adversarial network, GAN)3 种可替代方案进行了比较,比较结果见表 3。

结果表明,CVAE 在 4 种模型架构中表现最为欠佳,其 ROC-AUC 分数较 MAE 低 7.2 个百分点,这一现象凸显出自编码器结构在预测任务中泛化能力存在明显局限。相比之下,MAE 虽然同属自编码器变体,但其引入的随机掩码机制显著增强了模型对缺失数据的鲁棒性。该机制与扩散模型中添加高斯噪声的处理方式具有相似性,但 MAE 的核心在于从部分观测数据中重构完整数据,这导致其在训练数据分布外的样本预测能力受限。值得注意的是,GAN 作为生成模型,在 4 个指标上较 MAE 提升均超过 7.5 个百分点,这一显著差异证实了生成模型在预测任务中的性能优势。特别地,作为另一种生成式方法,扩散模型拥有 GAN 所不具备的多样性,这种多样性不仅能够保证训

练数据范围内的预测质量,同时还能够预测超出范围的数据。

2.6 参数化分析

为了研究预测网络中的记忆模块聚类中心数以及扩散模型重复采样数对所提出方法在 3 个数据集上性能的影响,还进行了敏感性分析,实验结果如图 4 所示。图 4(a)展示了不同聚类中心数的实验结果。可以观察到,随着聚类中心数的逐渐增加,ROC-AUC 值也随之提升,聚类中心数为 1 000 时 ROC-AUC 值达到峰值,随后开始下降。这表明,聚类中心数过少会导致所得到的记忆模式覆盖范围较窄,而过多则会引入冗余。同样,从图 4(b)可以看出,生成样本数在达到峰值后继续增加会导致性能下降,这可能是由于生成了一些不合理的样本。基于此,聚类中心数和生成样本数的最优超参数分别选定为 1 000 和 20。

2.7 实例分析

为了解 MEDM-SVAD 的局限性,针对跑步和骑车两种异常行为展开分析,并通过可视化对比成功与失败的检测实例,深入挖掘模型产生误判的原因。图 5(a)和 5(b)所示的可视化结果中,绿色骨架表示正确的实例,红色骨架则表示错误实例。由图可见,成功检测的异常行为的骨架通常展现出较为清晰且易于区分的特征。然而,受距离与角度等客观因素影响,跑步与骑车的骨架数据在某些情况下与正常行走的骨架特征高度相似,导致行为表征出现模糊性,从而引发模型的误判。该现象揭示了模型在应对高相似度骨架特征时的局限性。为提升检测性能,后续研究可考虑在骨架信息之外融合视觉特征,以进一步增强模型的判别能力并减少误检。

表 3 扩散模型的消融学习

Table 3 Ablation study of the diffusion model

%

方法	R-A	R	P	F
CVAE	59.1	49.6	63.3	55.8
MAE	66.3	57.7	70.2	60.9
GAN	73.8	67.2	77.8	70.0
Diffusion Model	78.5	74.1	83.7	77.2

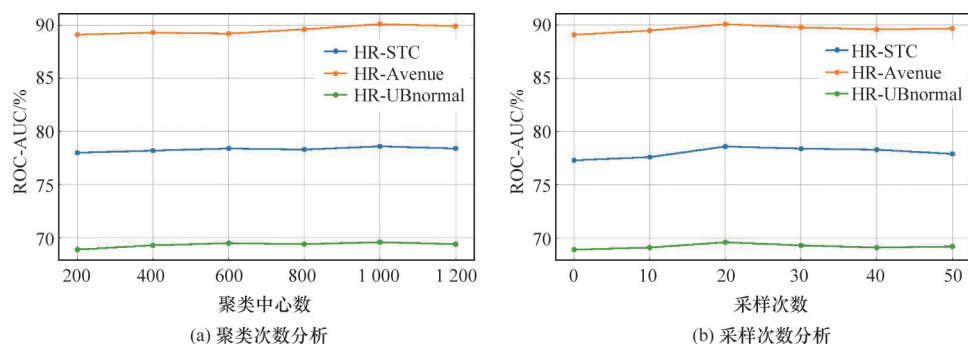


图 4 超参数分析图

Fig. 4 Hyperparameter analysis

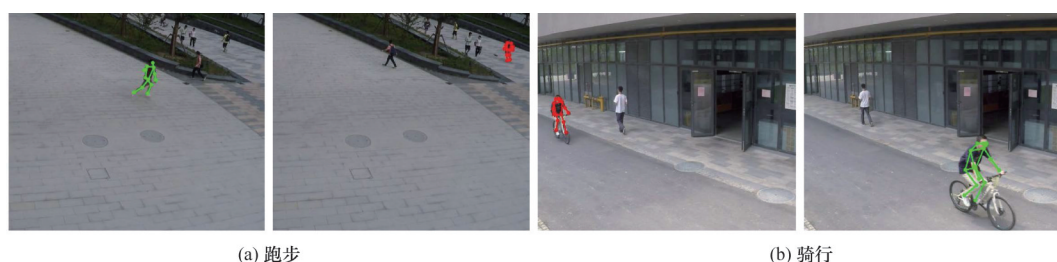


图 5 实例分析的可视化结果

Fig. 5 Visualization results of case analysis

3 总结与未来工作

本文针对骨架 VAD 中正常与异常行为难以区分的问题,提出 MEDM-SVAD 框架。该框架包括两大部分:记忆增强的重构网络和记忆增强的预测网络。在重构网络中,通过引入可学习的记忆单元,增强对正常观察数据的记忆能力;在预测网络中,利用扩散模型更强的多样化生成能力,扩展模型对正常行为的识别范围。此外,针对预测网络的特殊结构,通过统计聚类生成一组固定的未来数据记忆模式,以增强对正常未来数据的记忆能力。实验结果表明,该框架在基于骨架的视频异常行为检测方面优于现有其他方法,表明所提方法具有更好的鲁棒性。在未来的研究中,将进一步优化该框架,以提升其实时性。

参考文献

- [1] Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 18-23, 2018, Salt Lake City, UT, USA. IEEE, 2018: 6479-6488. DOI: 10.1109/CVPR.2018.00678.
- [2] Hasan M, Choi J, Neumann J, et al. Learning temporal regularity in video sequences [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 27-30, 2016, Las Vegas, NV, USA. IEEE, 2016: 733-742. DOI: 10.1109/CVPR.2016.86.
- [3] Liu W, Luo W X, Lian D Z, et al. Future frame prediction for anomaly detection: a new baseline [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 18-23, 2018, Salt Lake City, UT, USA. IEEE, 2018: 6536-6545. DOI: 10.1109/CVPR.2018.00684.
- [4] Luo W X, Liu W, Gao S H. A revisit of sparse coding based anomaly detection in stacked RNN framework [C]//2017 IEEE International Conference on Computer Vision (ICCV). October 22-29, 2017, Venice, Italy. IEEE, 2017: 341-349. DOI: 10.1109/ICCV.2017.45.
- [5] Chong Y S, Tay Y H. Abnormal event detection in videos using spatiotemporal autoencoder [C]//Advances in Neural Networks-ISNN 2017. Cham: Springer International Publishing, 2017: 189-196. DOI: 10.1007/978-3-319-59081-3_23.
- [6] Xu D, Yan Y, Ricci E, et al. Detecting anomalous events in videos by learning deep representations of appearance and motion [J]. Computer Vision and Image Understanding, 2017, 156: 117-127. DOI: 10.1016/j.cviu.2016.10.010.
- [7] Hinami R, Mei T, Satoh S. Joint detection and recounting of abnormal events by learning deep generic knowledge [C]//2017 IEEE International Conference on Computer Vision (ICCV). October 22-29, 2017, Venice, Italy. IEEE, 2017: 3639-3647. DOI: 10.1109/ICCV.2017.391.
- [8] Coşar S, Donatiello G, Bogorny V, et al. Toward abnormal trajectory and event detection in video surveillance [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27 (3): 683-695. DOI: 10.1109/

- TCSVT. 2016. 2589859.
- [9] Zimek A, Schubert E, Kriegel H P. A survey on unsupervised outlier detection in high-dimensional numerical data[J]. *Statistical Analysis and Data Mining: the ASA Data Science Journal*, 2012, 5(5): 363-387. DOI: 10.1002/sam.11161.
- [10] Morais R, Le V, Tran T, et al. Learning regularity in skeleton trajectories for anomaly detection in videos [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 15-20, 2019, Long Beach, CA, USA. IEEE, 2019: 11988-11996. DOI: 10.1109/CVPR.2019.01227.
- [11] Flaborea A, Collorone L, D' Amely Di Melendugno G M, et al. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection [C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). October 1-6, 2023, Paris, France. IEEE, 2023: 10284-10295. DOI: 10.1109/ICCV51070.2023.00947.
- [12] Jain Y, Sharma A K, Velmurugan R, et al. PoseCVAE: anomalous human activity detection [C]//2020 25th International Conference on Pattern Recognition (ICPR). January 10-15, 2021, Milan, Italy. IEEE, 2021: 2927-2934. DOI: 10.1109/ICPR48806.2021.9412132.
- [13] Markovitz A, Sharir G, Friedman I, et al. Graph embedded pose clustering for anomaly detection [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 13-19, 2020, Seattle, WA, USA. IEEE, 2020: 10536-10544. DOI: 10.1109/CVPR42600.2020.01055.
- [14] Rodrigues R, Bhargava N, Velmurugan R, et al. Multi-timescale trajectory prediction for abnormal human activity detection [C]//2020 IEEE Winter Conference on Applications of Computer Vision (WACV). March 1-5, 2020, Snowmass Village, CO, USA. IEEE, 2020: 2626-2634. DOI: 10.1109/WACV45572.2020.9093633.
- [15] Barbalau A, Ionescu R T, Georgescu M I, et al. SSMTL++: revisiting self-supervised multi-task learning for video anomaly detection [J]. *Computer Vision and Image Understanding*, 2023, 229: 103656. DOI: 10.1016/j.cviu.2023.103656.
- [16] Flaborea A, D' Amely di Melendugno G M, D' Arrigo S, et al. Contracting skeletal kinematics for human-related video anomaly detection [J]. *Pattern Recognition*, 2024, 156: 110817. DOI: 10.1016/j.patcog.2024.110817.
- [17] Kanu-Asiegbu A M, Vasudevan R, Du X X. BIPOCO: Bi-directional trajectory prediction with pose constraints for pedestrian anomaly detection [EB/OL]. arXiv 2022: 2207.02281v1. (2022-07-05) [2025-03-18]. <https://arxiv.org/abs/2207.02281v1>.
- [18] Li N J, Chang F L, Liu C S. Human-related anomalous event detection via spatial-temporal graph convolutional autoencoder with embedded long short-term memory network [J]. *Neurocomputing*, 2022, 490: 482-494. DOI: 10.1016/j.neucom.2021.12.023.
- [19] Luo W X, Liu W, Gao S H. Normal graph: spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection [J]. *Neurocomputing*, 2021, 444: 332-337. DOI: 10.1016/j.neucom.2019.12.148.
- [20] 刘禹含, 吉根林, 张红苹. 基于骨架图与混合注意力的视频行人异常检测方法[J]. *计算机应用*, 2024, 44(8): 2551-2557. DOI: 10.11772/j.issn.1001-9081.2023081157.
- [21] 杨学存, 李杰华, 陈丽媛, 等. 基于人体骨架的扶梯乘客异常行为识别方法[J]. *安全与环境学报*, 2024, 24(2): 636-643. DOI: 10.13637/j.issn.1009-6094.2022.2404.
- [22] 朱高伟. 基于骨架数据的列车司机异常行为检测[J]. *计算机科学与技术*, 2024, 14(7): 42-50. DOI: 10.12677/csa.2024.147162.
- [23] Acsintoae A, Florescu A, Georgescu M I, et al. UBnormal: new benchmark for supervised open-set video anomaly detection [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 18-24, 2022, New Orleans, LA, USA. IEEE, 2022: 20111-20121. DOI: 10.1109/CVPR52688.2022.01951.
- [24] Gong D, Liu L Q, Le V, et al. Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). October 27-November 2, 2019, Seoul, Korea. IEEE, 2019: 1705-1714. DOI: 10.1109/ICCV.2019.00179.
- [25] Nguyen T N, Meunier J. Anomaly detection in video sequence with appearance-motion correspondence [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). October 27-November 2, 2019, Seoul, Korea. IEEE, 2019: 1273-1283. DOI: 10.1109/ICCV.2019.00136.
- [26] Maas A L, Hannun A Y, Ng A Y. Rectifier nonlinearities improve neural network acoustic models [C]//2013 International Conference on Machine Learning (ICML). June 16-21, 2013, Atlanta, Georgia, USA. PMLR, 2013, 30(1): 3.
- [27] Fan Y X, Wen G J, Li D R, et al. Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder [J]. *Computer Vision and Image Understanding*, 2020, 195: 102920. DOI: 10.1016/j.cviu.2020.102920.
- [28] Lu Y W, Kumar K M, Nabavi S S, et al. Future frame prediction using convolutional VRNN for anomaly detection [C]//2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). September 18-21, 2019, Taipei, China. IEEE, 2019: 1-8. DOI: 10.1109/avss.2019.8909850.
- [29] Yu G, Wang S Q, Cai Z P, et al. Cloze test helps: effective video anomaly detection via learning to complete video events [C]//Proceedings of the 28th ACM International Conference on Multimedia. October 12-16, 2020, Seattle, WA, USA. ACM, 2020: 583-591. DOI: 10.1145/3394171.3413973.
- [30] 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]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). October 10-17, 2021, Montreal, QC, Canada. IEEE, 2021: 13568-13577. DOI: 10.1109/ICCV48922.2021.01333.
- [31] Zhang X Y, Li N Q, Li J W, et al. Unsupervised surface anomaly detection with diffusion probabilistic model [C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). October 1-6, 2023, Paris, France. IEEE, 2023: 6759-6768. DOI: 10.1109/ICCV51070.2023.00624.
- [32] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[EB/OL]. arXiv 2020: 2006.11239v2. (2020-12-16) [2025-03-18]. <https://arxiv.org/abs/2006.11239>.
- [33] Xiao Z S, Kreis K, Vahdat A. Tackling the generative learning trilemma with denoising diffusion GANs[EB/OL]. arXiv 2022: 2112.07804v2. (2022-04-04) [2025-03-18]. <https://arxiv.org/abs/2112.07804v2>.
- [34] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90. DOI: 10.1145/3065386.
- [35] Sofianos T, Sampieri A, Franco L, et al. Space-time-separable graph convolutional network for pose forecasting [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). October 10-17, 2021, Montreal, QC, Canada. IEEE, 2021: 11189-11198. DOI: 10.1109/ICCV48922.2021.01102.
- [36] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 27-30, 2016, Las Vegas, NV, USA. IEEE, 2016: 770-778. DOI: 10.1109/CVPR.2016.90.
- [37] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation [C]// Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. October 5-9, 2015, Munich, Germany. Cham: Springer International Publishing, 2015: 234-241. DOI: 10.1007/978-3-319-24574-4_28.
- [38] Shi L S, Wang L, Zhou S P, et al. Trajectory unified transformer for pedestrian trajectory prediction [C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). October 1-6, 2023, Paris, France. IEEE, 2023: 9641-9650. DOI: 10.1109/ICCV51070.2023.00887.
- [39] 余保玲, 虞松坤, 孙耀然, 等. 基于 DeepPose 和 Faster RCNN 的多目标人体骨骼节点检测算法[J]. 中国科学院大学学报, 2020, 37(6): 828-834. DOI: 10.7523/j.issn.2095-6134.2020.06.015.
- [40] Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization[EB/OL]. arXiv 2015: 1409.2329v5. (2015-02-19) [2025-03-18]. <https://arxiv.org/abs/1409.2329>.