计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.010144] [CSTR: 32024.14.csa.010144]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

骨架引导时空行为-场景对齐网络的视频异常检测①

张英俊, 陈志豪, 谢斌红, 张　睿

(太原科技大学 计算机科学与技术学院, 太原 030024)
通信作者: 陈志豪, E-mail: s202320211002@stu.tyust.edu.cn

摘　要: 针对复杂场景下视频异常检测中时空动态建模不充分与场景语义适配性不足的问题, 提出骨架引导的时空行为-场景对齐网络 (spatio-temporal behavior-scene alignment network, ST-BSAN). 该模型通过双核心模块协同优化实现鲁棒检测: 动态时空注意力模块 (dynamic spatio-temporal attention module, DSTAM) 集成空间与时间自注意力, 突破传统固定图结构与局部时序建模限制, 自适应捕捉关节动态关联与长程时序突变; 行为-场景对齐模块 (behavior-scene alignment module, BSAM) 构建动态记忆库, 通过余弦相似度度量行为-场景语义一致性, 抑制跨场景误检. 同时引入扩散概率模型生成多样化正常行为假设, 以 DSTAM 输出特征为条件约束生成过程, 解决单峰预测对正常行为多样性覆盖不足的问题. 在 HR-STC 和 UBnormal 数据集上的实验显示, ST-BSAN 的帧级 AUC 分别达 79.9% 和 70.1%, 较基线方法提升 2.3% 和 1.8%. 消融实验验证了 DSTAM 与 BSAM 的协同效应.

关键词: 视频异常检测; 行为-场景对齐; 空间自注意力; 时间自注意力; 扩散模型

引用格式: 张英俊,陈志豪,谢斌红,张睿.骨架引导时空行为-场景对齐网络的视频异常检测.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/10144.html

Skeleton-guided Spatio-temporal Behavior-scene Alignment Network for Video Anomaly Detection

ZHANG Ying-Jun, CHEN Zhi-Hao, XIE Bin-Hong, ZHANG Rui

(School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China)

Abstract: To address insufficient spatio-temporal dynamic modeling and inadequate scene semantic adaptability in video anomaly detection under complex scenarios, this study proposes a skeleton-guided spatio-temporal behavior-scene alignment network (ST-BSAN). Robust anomaly detection is achieved through the collaborative optimization of two core modules. The dynamic spatio-temporal attention module (DSTAM) integrates spatial self-attention and temporal self-attention, overcoming the limitations of traditional fixed graph structures and local temporal modeling, and adaptively capturing dynamic joint correlations as well as long-range temporal variations. The behavior-scene alignment module (BSAM) constructs a dynamic memory bank, measures behavior-scene semantic alignment via cosine similarity, and suppresses false detections across different scenes. In addition, a diffusion probabilistic model is introduced to generate diversified normal behavior hypotheses. The generation process is conditioned on features extracted by DSTAM, thereby addressing the limitation of unimodal prediction in covering the diversity of normal behaviors. Experiments on the HR-STC and UBnormal datasets show that ST-BSAN achieves frame-level AUCs of 79.9% and 70.1% respectively, outperforming baseline methods by 2.3% and 1.8%. Ablation studies further verify the synergistic effect of DSTAM and BSAM.

① 基金项目: 山西省科技成果转化引导专项项目 (202304021301020); 山西省产教融合研究生联合培养示范基地项目 (2022JD11); 吕梁市引进高层次科技人才重点研发项目 (2022RC08)
收稿时间: 2025-08-26; 修改时间: 2025-09-24, 2025-10-28; 采用时间: 2025-12-05; csa 在线出版时间: 2026-03-09

Key words: video anomaly detection; behavior-scene alignment; spatial self-attention; temporal self-attention; diffusion model

视频异常检测作为智能监控领域的核心技术, 旨在从复杂动态场景中精准识别违背常规行为模式的事件, 其研究成果对公共安全、工业质检、智慧城市等场景具有关键应用价值. 随着深度学习技术的发展, 基于骨架的行为分析因隐私保护性和语义抽象能力成为主流方向, 但该任务仍面临双重核心挑战.

一是动态时空依赖的精细化建模难题, 人体动作中关节间的临时协同关系与长程时间依赖难以通过固定拓扑结构有效捕捉[1]. 传统时空图卷积网络 (如 ST-GCN)[2]依赖预定义骨骼邻接矩阵, 仅能建模物理连接关节的静态关联, 却无法捕捉“拍手时双手非邻接交互”“投掷动作中躯干与手臂的动态耦合”等非固定依赖关系. 在时间建模方面, 卷积网络的局部感受野限制了对“行走→急停”等跨多帧以上时序突变的捕捉能力, 导致模型对快速异常事件的响应滞后.

二是场景语义的动态适应性瓶颈, 同一行为在不同场景下的合理性判定存在显著差异[3]. 现有方法因缺乏行为-场景的联合建模机制, 易导致跨场景误检. 例如, “快速奔跑”行为在操场场景中属于正常活动, 在医院走廊则因可能干扰医疗秩序而被判定为异常. 现有方法虽通过扩散模型生成多样化行为轨迹, 但未显式编码场景语义, 导致在“操场→医院”走廊场景切换时误检率显著提升.

传统基于像素的方法通过光流、帧间差分等底层特征建模正常模式, 虽能捕捉局部运动变化, 但受限于光照干扰、遮挡效应和隐私风险, 且难以刻画复杂行为的高层语义. 近年来, 基于人体骨架的异常检测方法通过关键点序列规避了外观隐私问题, 并借助图卷积网络 (GCN) 等工具实现关节运动建模. 然而, 主流方法如 ST-GCN 依赖静态骨骼拓扑, 无法动态捕获动作中临时关联的关节对; 时间建模方面, 卷积网络的局部感受野难以捕捉跨长时步的依赖关系, 导致模型对异常行为的时空敏感性不足.

在开放集场景下, 异常检测的核心矛盾进一步凸显: 模型需在仅接触正常样本的训练条件下, 泛化未知异常类型并适应多样化场景. 现有方法大多将行为与场景割裂处理, 忽略行为合理性对环境的强依赖性. 例如, MoCoDAD[4]等扩散模型虽通过生成增强正常行为多样性建模, 但其条件编码模块仍采用静态图结构, 且未引入场景语义约束, 导致在场景剧变时误检率显著上升.

针对上述挑战, 本文提出骨架引导的时空行为-场景对齐网络 (spatio-temporal behavior-scene alignment network, ST-BSAN), 通过双模块改进实现性能突破.

(1) 动态时空注意力模块 (dynamic spatio-temporal attention module, DSTAM): 引入空间自注意力 (spatial self-attention, SSA) 与时间自注意力 (temporal self-attention, TSA) 机制, 动态调整关节依赖权重并捕捉长程时序关联. 相较于传统 ST-GCN 的固定拓扑, SSA 可自适应强化关键关节对的交互, TSA 则通过全局时间注意力建模跨帧突变特征, 显著提升模型对异常行为的时空表征能力.

(2) 行为-场景对齐模块 (behavior-scene alignment module, BSAM): 构建动态更新的行为-场景联合记忆库, 通过余弦相似度度量行为特征与场景语义的一致性. 该模块可显式建模场景敏感逻辑, 抑制跨场景误判, 尤其适用于开放环境下的语义约束需求.

需强调的是, ST-BSAN 的改进并非局限于双模块, 而是构建“动态特征提取-场景语义对齐-模块协同优化”的系统性方案: DSTAM 以 SSA/TSA 突破传统静态时空建模瓶颈, 捕捉非邻接关节交互与长程时序突变; BSAM 通过动态行为-场景记忆库, 建立语义一致性判定准则; 二者经特征层间注入、异常分数加权融合深度协同, 将时空判别能力与场景适配性转化为检测性能提升. 这种全链路设计, 为复杂场景下的视频异常检测提供了更具鲁棒性的系统性解决方案.

1 相关工作

视频异常检测的研究围绕数据表示与建模机制展开, 核心可分为像素级底层特征分析与姿态级语义特征建模两大技术路线. 像素级方法直接处理视觉原始信号, 姿态级方法则通过人体骨架规避隐私问题并聚焦行为语义. 近年来, 场景语义的动态适配成为提升开放集性能的关键突破方向.

1.1 基于像素的方法

基于像素的方法通过光流、帧间差分等底层特征建模正常行为, 核心假设为“异常是时空模式的显著偏离”. 早期如 Zhao 等[5]用稀疏编码重建正常帧检测异常, Liu 等[6]通过未来帧预测框架结合光流与 GAN 识别异常. Georgescu 等[7]引入自监督多任务学习增强特征判别力. 该类方法受限于光照干扰、隐私风险及语义鸿沟, 难以区分同行为在不同场景的合理性, 推动研究转向基于人体姿态的语义建模.

1.2 基于姿态的方法

基于人体姿态的异常检测方法通过提取骨架关键点序列 (如 OpenPose 的 18 关节坐标) 刻画行为模式, 既规避了像素级方法的隐私风险, 又能聚焦行为语义的高层建模. 近年来, 该方向围绕时空依赖建模的动态性、开放集泛化的生成能力、场景语义的适应性形成 3 条核心技术路线, 分别对应以下 3 个研究方向.

1.2.1 基于时空建模的方法

时空建模通过动态捕捉骨架的空间结构与时间依赖, 推动异常检测从静态拓扑向自适应结构演进. ST-GCN[2]首次引入图卷积建模人体骨架, 但固定邻接结构难以适配拍手等临时非物理连接行为. 本文 DSTAM 模块中的 SSA 机制, 正是在 ST-GCN 图卷积提取静态特征的基础上, 通过全局注意力动态优化关节依赖权重, 来弥补此局限; 后续 DG-STGCN[8]和 Dynamic GCN[9]以可学习图结构调整关节关系, 进一步验证动态优化关节关联的必要性. 刘禹含等[10]的 PAD-SGMA 方法虽借混合注意力提升建模能力, 但受限于局部区域注意力与传统时序建模的局部感受野, 而 DSTAM 的 TSA 机制摒弃局部时间窗口, 通过全序列时序注意力捕捉跨多帧突变, 突破局部时序建模瓶颈, 最终以空间与时间自注意力协同, 捕捉关节动态依赖与时序突变, 清晰呈现技术演进路径, 突破传统方法对动态时空交互的建模瓶颈.

1.2.2 基于生成模型的方法

生成模型通过学习正常行为分布解决开放集异常检测问题, 在姿态建模中从单个重建向多样化生成演进. 早期自编码器[11]以重建误差为异常判据, 受单峰分布限制, 难以覆盖行为多样性. 郑敬添[12]提出结合生成对抗网络与扩散模型的双向预测网络, 通过对抗训练提升重建质量以改进检测精度. MoCoDAD[4]利用扩散模型对骨架轨迹多阶段去噪重建, 显著增强对未知异常的建模能力. 然而, 现有方法多依赖静态图结构, 难捕捉动态关节依赖, 且未显式建模场景语义, 限制复杂场景泛化. 为此, 本文以 DSTAM 输出的动态特征向量为扩散模型生成条件, 引导生成过程更符合物理运动规律与语义逻辑, 提升异常建模精度与跨场景适应性.

1.2.3 行为与场景对齐方法

在开放环境中, 行为异常性因场景语义差异而变, 行为-场景对齐是提升检测精度的关键. 孙澈等[13]构建时空上下文图, 通过循环神经网络迭代更新图节点与边状态, 显式建模行为与场景交互. Chen 等[14]提出了 DecoAD, 通过解耦场景与动作并利用知识图谱探索两者关系, 有效提升异常检测性能, 但在复杂遮挡场景中易因特征提取不全导致对齐偏差. Yang 等[15]设计 Trinity 框架, 利用对比学习对齐多模态特征, 然而其泛化性受限于预训练场景分布. 现有方法多依赖静态建模或场景先验, 难以适应动态语义环境. 为此, 本文提出行为-场景对齐模块 (BSAM), 通过动态记忆更新与时空特征融合, 实现行为语义与场景语义的一致性对齐, 从而降低误检率, 提升多变环境下的检测稳定性.

2 方法

2.1 网络架构

本文提出的骨架引导的动态时空行为-场景对齐网络 (ST-BSAN) 针对复杂场景下的视频异常检测任务, 构建了一个融合时空动态特征提取、多样化生成与场景语义约束的端到端架构. 该框架通过 3 个核心模块协同工作: 动态时空注意力模块 (DSTAM) 捕捉骨架序列中的时空依赖关系, 扩散概率生成模块 (DPGM) 生成多样化正常行为假设, 以及行为-场景对齐模块 (BSAM) 评估生成行为与场景上下文的语义一致性. 这一设计有效解决了传统方法在时空建模灵活性、行为多样性覆盖以及场景适应性方面的不足. 其完整架构如图 1 所示.

整体网络采用 3 阶段架构: 条件编码阶段、多样化生成阶段和联合评估阶段. 在条件编码阶段, 动态时空注意力模块通过空间自注意力机制和时间自注意力机制, 分别建模关节间的动态空间关系和长程时序依赖. 如图 1 右上侧的条件编码阶段区域所示, 骨架序列 $X^{1:k}$ 输入该模块后, 空间自注意力机制通过计算关节特征相似性矩阵动态调整不同关节对的交互权重, 突破了传统固定拓扑结构的限制; 时间自注意力机制则通

过全局时序建模捕捉跨越多个时间步的行为变化模式, 特别关注异常发生时的时序突变特征. 两子模块输出经特征拼接后得到潜在表示h, 同时通过时间步编码器对扩散步骤t进行编码, 将其转换为一个潜在表示$\tau(t)$, 两者相加形成一个综合的条件信号$Z_{\text{cond}}=h+\tau(t)$, 作为后续生成过程的约束条件.

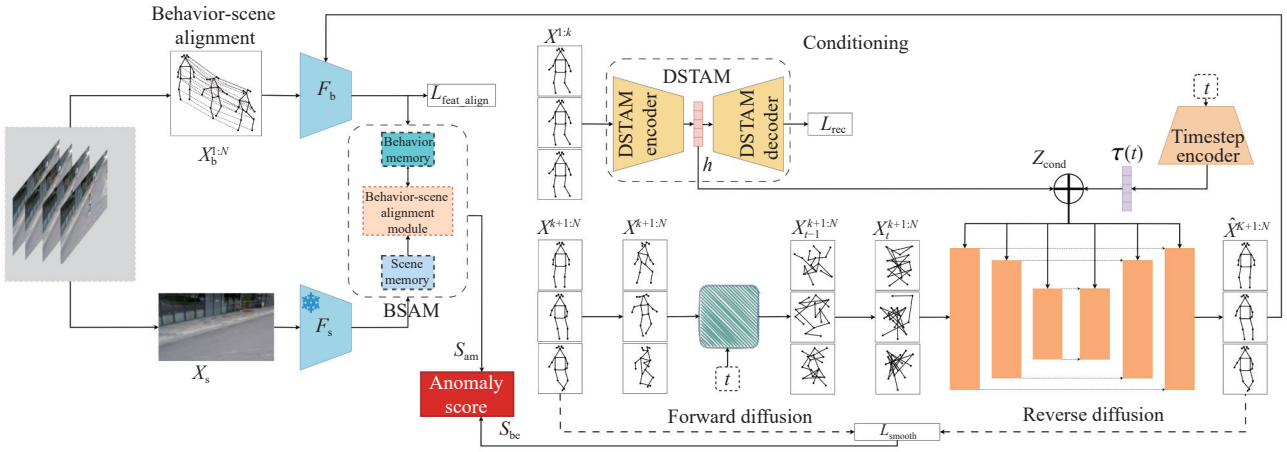

图 1 ST-BSAN 网络结构

在多样化生成阶段, 采用基于 DDPM 的生成框架, 将骨架序列$X^{k+1:N}$通过前向扩散和反向扩散过程生成多组未来骨架序列假设$\hat{X}^{1:k}$. 如图 1 右下侧多样化生成阶段区域所示, 扩散模型主干为基于 U-Net 的时空可分离图卷积 (STS-GCN) 架构, 分收缩 (3 层 STS-GCN 聚合关节与时序特征)、扩张 (3 层转置 STS-GCN 恢复特征维度, 设残差连接) 双阶段. 不同于传统的确定性预测方法, 扩散模型通过前向扩散对$X^{k+1:N}$采样时间步t加噪, 反向扩散以 DSTAM 输出的条件向量Z_{cond}为约束去噪, 显式建模正常行为的分布多样性, 从而更好地覆盖各种可能的正常行为模式. 确保生成的行为序列与历史观测具有时空连贯性和物理合理性.

在联合评估阶段, 模型通过融合行为异常分数与场景对齐度分数生成最终异常评分, 该过程如图 1 底部联合评估阶段区域所示: 首先, 扩散模型以条件向量为约束生成多组未来骨架假设, 通过平滑 $L1$ 损失计算每组生成骨架与真实骨架的差异, 取最小值作为行为异常分数S_{be}衡量当前行为与正常物理运动规律的偏差; 同时, BSAM 模块基于动态记忆库计算当前行为与场景的即时对齐度, 以及生成行为的平均对齐度, 融合得到场景对齐度分数S_{am}, 反映行为与场景的语义冲突程度; 最终总异常评分S_{anomaly}通过加权融合行为异常分数和场景对齐度分数, 实现对行为异常程度的综合评估, 平衡行为物理合理性与场景语义约束.

这种分层设计使 ST-BSAN 能够同时捕捉行为的时空动态特性、正常行为的分布多样性以及场景语义对行为合理性的约束, 从而在复杂场景下实现更准确、更鲁棒的异常检测.

2.2 动态时空注意力模块 (DSTAM)

为了在视频异常检测任务中更好地建模复杂的时空交互关系, 本文提出了动态时空注意力模块 (DSTAM). 该模块通过引入空间自注意力 (SSA) 和时间自注意力 (TSA) 机制, 旨在提升对人体动作中的关键时空依赖关系的捕捉能力, 从而显著提高异常行为的检测性能.

DSTAM 采用自编码器架构, 其核心思想是在空间维度上通过动态调整关节间的依赖关系, 在时间维度上通过关注长程时间依赖, 实现对复杂动作的精确建模. 编码器流程首先从输入的骨架数据中提取空间特征, 通过 SSA 增强关节间的动态关系, 再通过 TSA 捕捉时间步之间的依赖性, 最终将时空特征拼接为潜在表示h; 解码器则基于潜在表示h, 通过转置卷积网络重建骨架序列, 并引入重建损失L_{rec}约束生成过程, 迫使编码器学习到更具代表性的时空特征, 确保潜在表示h既能保留动作的关键动态信息, 又能过滤噪声干扰.

如图 2 所示, DSTAM 模块的编码器输入为骨架序列特征$X\in\mathbb{R}^{C_N\times T\times N}$ (C_N为输入通道数, T为时间步, N为关节数), 输出为增强时空特征$h\in\mathbb{R}^{(C_{\text{gcn}}+C_{\text{spa}}+2C_{\text{tem}})\times T\times N}$, 其流程分为空间特征提取与时间特征提取两阶段.

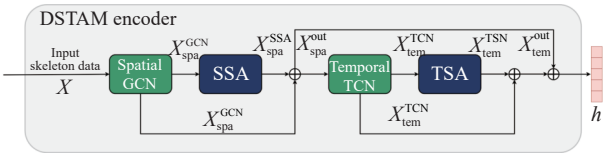

图 2 动态时空注意力模块

2.2.1 空间自注意力机制 (SSA)

空间自注意力机制 (SSA) 的引入是为了解决传统图卷积网络 (GCN) 在建模空间依赖关系时的局限性. GCN 通常依赖于固定的骨架拓扑结构, 这意味着它不能动态地调整不同关节之间的依赖关系. 然而, 在实际应用中, 不同的行为动作往往涉及关节间复杂的、动态变化的依赖. 例如, 在拍手动作中, 左右手虽然不直接相连, 但它们之间的时空关系却十分紧密. 为此, SSA 通过动态注意力机制替代固定拓扑, 自适应调整关节间权重. 具体实现如下.

首先, 通过 GCN 提取静态空间特征.

$$X_{\text{spa}}^{\text{GCN}} = \sigma(\Lambda^{-\frac{1}{2}}(A+I)\Lambda^{-\frac{1}{2}}XW_{\text{gcn}}) \tag{1}$$

其中, Λ为度矩阵, $W_{\text{gcn}} \in \mathbb{R}^{C_N \times C_{\text{gcn}}}$ 为可学习权重矩阵, σ为 ReLU 激活函数. 随后, SSA 基于多头注意力机制动态建模关节依赖.

$$Q_{\text{spa}} = W_q X_{\text{spa}}^{\text{GCN}},\ K_{\text{spa}} = W_k X_{\text{spa}}^{\text{GCN}},\ V_{\text{spa}} = W_v X_{\text{spa}}^{\text{GCN}} \tag{2}$$

其中, W_q, W_k, $W_v \in \mathbb{R}^{C_{\text{gcn}} \times C_{\text{gcn}}}$ 为线性变换参数. 通过计算归一化注意力权重:

$$A_{\text{spa}} = Softmax\left(\frac{Q_{\text{spa}}K_{\text{spa}}^{\text{T}}}{\sqrt{C_{\text{gcn}}}}\right) \tag{3}$$

得到动态关节依赖矩阵$A_{\text{spa}} \in \mathbb{R}^{N \times N}$, 并加权融合特征:

$$X_{\text{spa}}^{\text{SSA}} = A_{\text{spa}}V_{\text{spa}} \tag{4}$$

最终, 将 GCN 与 SSA 输出拼接为空间增强特征 $X_{\text{spa}}^{\text{out}} = Concat(X_{\text{spa}}^{\text{GCN}}, X_{\text{spa}}^{\text{SSA}})$, 显著提升对异常关节 (如摔倒时的手部动作) 的敏感性.

2.2.2 时间自注意力机制 (TSA)

时间自注意力机制 (TSA) 的加入是为了增强模型在处理跨时间步依赖时的能力. 视频中的动作通常跨越多个时间步, 而传统的时间卷积网络 (TCN) 只能处理局部的时间依赖关系, 难以捕捉到长时间跨度内的关键动态变化. 特别是在异常行为的检测中, 动作可能在短时间内发生剧烈变化 (如跌倒、急停等), 这些变化往往跨越多个时间步, 需要通过时间自注意力机制来敏感地捕捉. 其实现过程如下.

首先, 对空间增强特征$X_{\text{spa}}^{\text{out}}$进行一维卷积操作 (卷积核大小为 3, 步长为 1).

$$X_{\text{tem}}^{\text{TCN}} = \sigma(X_{\text{spa}}^{\text{out}} * W_{\text{tem}}) \tag{5}$$

其中, *表示卷积操作, W_{tem}为可学习的一维卷积核, 输出局部时间特征$X_{\text{tem}}^{\text{TCN}} \in \mathbb{R}^{C_{\text{tem}} \times T \times N}$. 随后, TSA 计算时间注意力权重:

$$Q_{\text{tem}} = W_q X_{\text{tem}}^{\text{TCN}},\ K_{\text{tem}} = W_k X_{\text{tem}}^{\text{TCN}},\ V_{\text{tem}} = W_v X_{\text{tem}}^{\text{TCN}} \tag{6}$$

其中, W_q, W_k, $W_v \in \mathbb{R}^{C_{\text{tem}} \times C_{\text{tem}}}$ 为可学习参数. 归一化注意力权重为:

$$A_{\text{tem}} = Softmax\left(\frac{Q_{\text{tem}}K_{\text{tem}}^{\text{T}}}{\sqrt{C_{\text{tem}}}}\right) \tag{7}$$

得到动态时间依赖矩阵$A_{\text{tem}} \in \mathbb{R}^{T \times T}$, 并加权融合特征:

$$X_{\text{tem}}^{\text{TSA}} = A_{\text{tem}}V_{\text{tem}} \tag{8}$$

最终, 将 TCN 与 TSA 输出拼接为时间增强特征 $X_{\text{tem}}^{\text{out}} = Concat(X_{\text{tem}}^{\text{TCN}}, X_{\text{tem}}^{\text{TSA}})$, 增强对时序突变 (如突然静止) 的建模能力.

通过结合 SSA 和 TSA, DSTAM 编码器最终输出的潜在向量$h = Concat(X_{\text{spa}}^{\text{out}}, X_{\text{tem}}^{\text{out}})$, DSTAM 模块能够在空间和时间维度上同时动态调整特征的权重, 捕捉复杂的时空依赖关系. SSA 关注空间中关节之间的动态依赖, 而 TSA 则关注时间步之间的长程依赖, 二者相互补充, 使得模型能够更加准确地表征动作中的关键时空特征, 为后续扩散模型生成多样化未来姿态提供高区分度的条件编码, 从而显著提升异常检测性能.

2.3 行为-场景对齐模块 (BSAM)

传统的异常检测方法主要依赖行为特征进行判断, 而忽略了行为所发生的场景信息. 本文提出了行为-场景对齐模块 (BSAM), 旨在通过引入场景信息, 增强行为异常检测的准确性.

如图 3 所示, BSAM 模块的核心思想是将行为特征和场景特征结合, 通过计算它们之间的对齐度, 识别出那些不符合环境背景的异常行为. 这一模块的架构包括行为记忆、场景记忆、行为-场景对齐模块和对齐记忆这 4 个主要组成部分. 首先, 行为特征和场景特征分别被输入到行为记忆和场景记忆模块中, 经过处理后得到行为和场景的对齐权重. 然后, 这些权重被组合成一个联合表示, 并送入行为-场景对齐模块, 计算行为与

场景之间的对齐度. 最终, 对齐记忆模块通过计算对齐分数, 生成异常分数S_{am}, 用来判断当前行为是否异常.

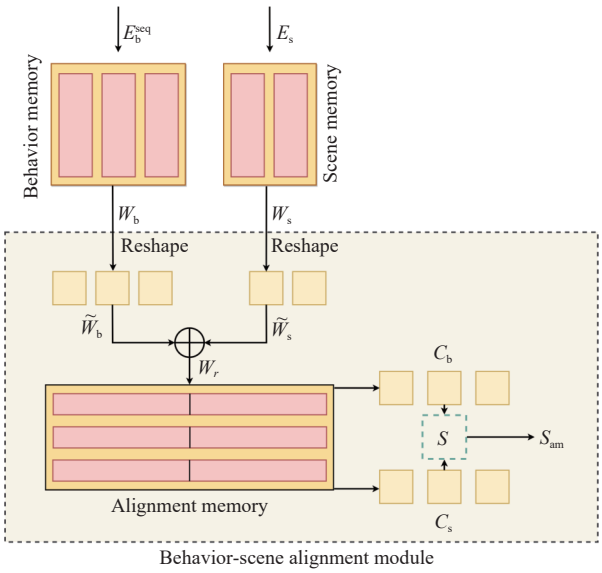

图 3 行为-场景对齐模块

BSAM 模块的输入包括两类特征: 行为特征和场景特征. 行为特征是从骨架数据中提取的, 表示人体的运动模式或姿态信息, 如站立、奔跑等, 经行为特征提取器F_{b} (GCN) 输出为 128 维向量, 选择 GCN 是因 BSAM 核心任务为评估行为与场景语义一致性, 需稳定的行为类别特征, GCN 基于骨骼拓扑可高效聚合关节类别相关特征, 不受行为动态变异干扰; 场景特征来自 RGB 图像, 表示环境的背景信息, 如图书馆、操场等, 经预训练 CNN 提取为 256 维向量. 为确保维度一致性, 场景特征通过线性投影层降维至 128 维, 与行为特征维度对齐. 行为特征和场景特征分别通过行为记忆和场景记忆模块进行处理. 每个记忆模块存储并更新对应的特征模式, 通过计算输入行为特征与行为记忆模块中存储特征的相似度, 得到行为特征的对齐权重W_{b}; 同理, 得到场景特征的对齐权重W_{s}. 具体地, 行为特征和场景特征的对齐度通过余弦相似度来计算, 公式如下:

$$W_{\text{b}}^i = \frac{\exp(sim(E_{\text{b}}^{\text{seq}}, m_{\text{b}}^i))}{\sum_{j=1}^{N_{\text{b}}} \exp(sim(E_{\text{b}}^{\text{seq}}, m_{\text{b}}^j))} \tag{9}$$

$$W_{\text{s}}^i = \frac{\exp(sim(E_{\text{s}}, m_{\text{s}}^i))}{\sum_{j=1}^{N_{\text{s}}} \exp(sim(E_{\text{s}}, m_{\text{s}}^j))} \tag{10}$$

其中, m_{b}^i和m_{s}^i分别是存储在记忆模块中的行为和场景的第i个特征, N_{b}和N_{s}分别是行为和场景特征样本的数量, $sim(\cdot,\cdot)$是计算相似度的函数.

在得到对齐权重后, BSAM 将行为特征和场景特征的对齐权重拼接在一起, 形成联合表示W_r. 该表示将行为和场景的对齐信息结合, 为后续的异常分数计算提供基础. 联合表示通过以下公式构造:

$$W_r = [W_{\text{b}}, W_{\text{s}}] \tag{11}$$

其中, W_{b}和W_{s}分别是行为和场景的对齐权重向量. 通过将这两种信息结合, BSAM 能够在更高层次上捕捉行为与场景之间的复杂交互关系. 联合表示W_r随后被输入到对齐记忆 (alignment memory) 中, 用于更新行为与场景之间的对齐关系. 每次新的行为和场景输入时, 系统会根据当前的对齐度计算更新对齐内存. 更新过程通过加权平均方式进行, 公式如下:

$$m_r^i \leftarrow f\left(m_r^i + \sum_{v \in U_i} v_{p,i} W_r^v\right) \tag{12}$$

其中, U_i是与第i个记忆元素相关的查询索引集合, $v_{p,i}$是对齐度的权重, $f(\cdot)$为 $L2$ 正则化函数, 用于约束记忆槽更新幅度, 避免参数震荡.

BSAM 模块通过计算行为特征与对齐记忆槽之间的对齐权重C_{b}和场景特征与对齐记忆槽之间的对齐权重C_{s}, 来评估当前行为是否与场景相符, 进而实现异常行为检测. 行为特征对齐权重C_{b}和场景特征对齐权重C_{s}的计算公式如下:

$$C_{\text{b}}^i = \frac{\exp(d(W_{\text{b}}, m_r^{i,;L_{\text{b}}}))}{\sum_{j=1}^{N} \exp(d(W_{\text{b}}, m_r^{j,;L_{\text{b}}}))} \tag{13}$$

$$C_{\text{s}}^i = \frac{\exp(d(W_{\text{s}}, m_r^{i,;L_{\text{s}}}))}{\sum_{j=1}^{N} \exp(d(W_{\text{s}}, m_r^{j,;L_{\text{s}}}))} \tag{14}$$

最终, BSAM 计算行为与场景之间的异常分数S_{am}, 该分数用于评估当前行为是否异常. 异常分数的计算公式为:

$$S_{\text{am}} = \|C_{\text{b}} - C_{\text{s}}\| \tag{15}$$

该公式通过计算行为对齐权重C_{b}和场景对齐权重C_{s}之间的差异, 若两者差异显著 (如行为对齐集中于“奔跑”, 场景对齐集中于“图书馆”) 表明行为与场景语

义冲突, 可能为异常.

综上所述, 行为-场景对齐模块 (BSAM) 通过结合行为特征与场景特征, 计算它们之间的对齐度和异常分数, 从而实现异常行为的检测. 该方法不仅提高了异常检测的精度, 还增强了系统对复杂场景的适应能力. 通过记忆模块的动态更新, BSAM 能够不断学习新的行为-场景组合, 提高对新环境的识别能力.

2.4 损失函数

模型训练通过分阶段优化策略实现物理运动规律与场景语义约束的协同学习, 核心损失函数分为预训练阶段损失与联合微调阶段损失, 分别对应扩散模型的物理合理性建模与行为-场景对齐的语义约束.

预训练阶段仅启用扩散损失 ($L_{\text{diffusion}}$), 专注于训练扩散模型生成符合人体运动学的正常行为轨迹, 公式为:

$$L_{\text{diffusion}} = \frac{1}{M}\sum_{m=1}^{M} Smooth\ L1\left(X^{k+1:N}, \hat{X}^{k+1:N}\right) \tag{16}$$

该阶段以历史骨架序列$X^{1:k}$经条件模块编码的条件向量Z_{cond}为输入, 通过平滑 $L1$ 损失最小化生成骨架与真实骨架的关节坐标差异, 抑制“关节穿透”等物理不合理生成, 确保扩散模型习得正常行为的基础运动模式, 仅更新扩散模型与 DSTAM 参数, 行为特征提取器F_{b}与行为-场景对齐模块 (BSAM) 保持冻结.

联合微调阶段引入特征对齐损失$L_{\text{feat_align}}$与行为-场景对齐损失L_{am}, 总损失公式为:

$$L_{\text{total}} = L_{\text{diffusion}} + \lambda \cdot L_{\text{feat\_align}} + \mu \cdot L_{\text{am}} + v \cdot \|\theta\|_2^2 \tag{17}$$

其中, 特征对齐损失$L_{\text{feat_align}} = \|E_{\text{b}}^{\text{real}} - E_{\text{b}}^{\text{gen}}\|_2^2$强制生成骨架特征与真实骨架特征在语义空间一致, 优化F_{b}模块参数, 行为-场景对齐损失:

$$L_{\text{am}} = 1 - \cos\left(Concat(W_{\text{b}}^{\text{gen}}, W_{\text{s}}), Concat(W_{\text{b}}^{\text{real}}, W_{\text{s}})\right) \tag{18}$$

通过软地址权重衡量生成行为与真实行为的场景对齐一致性, 驱动 BSAM 动态更新行为-场景联合记忆库; $L2$ 正则化项$\|\theta\|_2^2$避免过拟合.

损失权重通过五折交叉验证确定, 默认$\lambda = 0.3$, $\mu = 0.2$, $v = 1\times10^{-4}$. 预训练阶段 (前 20 epoch) 学习率为1×10^{-4}, 联合微调阶段 (后 80 epoch) F_{b}学习率降至5×10^{-5}, 并每批次更新 BSAM 记忆库中 10% 的记忆槽.

2.5 异常分数

在推理阶段, 模型通过融合行为异常分数S_{be}与场景对齐度分数S_{am}生成最终异常评分S_{anomaly}, 实现对行为异常程度的量化评估. 该设计通过分层建模行为的物理规律性与场景语义约束, 有效解决动态多元场景下的跨模态检测难题.

行为异常分数基于扩散模型的多样化生成能力, 旨在衡量生成的未来行为与真实行为的运动连续性偏差. 具体流程为: 首先, 扩散模型以历史骨架序列$X^{1:k}$和动态时空注意力模块 (DSTAM) 输出的条件向量以及时间步的编码信息为输入, 通过反向扩散过程生成K组未来骨架假设$\hat{X}^{k+1:N}$. 随后, 对每组生成骨架计算其与真实未来骨架$X^{k+1:N}$的平滑 $L1$ 损失 (Huber 损失).

$$s_i = \begin{cases} 0.5 \cdot \left\|X^{k+1:N} - \hat{X}^{k+1:N}\right\|_2^2, & \text{if } \left\|X^{k+1:N} - \hat{X}^{k+1:N}\right\|_2 < 1 \\ \left\|X^{k+1:N} - \hat{X}^{k+1:N}\right\|_2^2 - 0.5, & \text{otherwise} \end{cases} \tag{19}$$

最终取所有生成样本的最小损失作为行为异常分数:

$$S_{\text{be}} = \min_{1 \leqslant i \leqslant M} s_i \tag{20}$$

场景对齐度分数通过行为-场景对齐模块 (BSAM) 动态记忆库机制实现, 综合评估行为与场景的语义一致性, 计算公式为:

$$S_{\text{am}} = \gamma \cdot S_{\text{am\_initial}} + (1-\gamma) \cdot \overline{S}_{\text{am\_generated}} \tag{21}$$

其中, $S_{\text{am_initial}}$为初始骨架序列$X^{1:N}$经F_{b}提取特征后输入给 BSAM 模块得到, 反映即时行为与场景的对齐度; $\overline{S}_{\text{am\_generated}}$. 为扩散模型生成的$K$组骨架同样输入给 BSAM 模块得到的平均对齐度.

最终异常评分通过加权融合行为异常与场景对齐度分数, 公式为:

$$S_{\text{anomaly}} = \alpha \cdot S_{\text{be}} + (1-\alpha) \cdot (1 - S_{\text{am}}) \tag{22}$$

其中, $\alpha + (1-\alpha) = 1$, 默认$\alpha = 0.6$, 优先关注行为物理合理性; $1 - S_{\text{am}}$将场景对齐度转换为异常倾向, 值越大表示场景冲突越显著. 该设计通过 DSTAM 与 BSAM 的协同作用, 实现物理规律与语义逻辑的分层建模: DSTAM 输出的h主要约束扩散生成的物理合理性, BSAM 的动态记忆库则通过实时更新行为-场景对齐模式 (如图书馆-静立、操场-奔跑写法), 提升对跨场景异常的敏感性.

3 实验

3.1 数据集与评价指标

在本文的实验部分, 本文采用了 3 个具有代表性

的公开数据集来验证所提出方法的有效性, 这些数据集分别是 UBnormal、HR-ShanghaiTech Campus (HR-STC) 和 HR-Avenue. 这些数据集涵盖了丰富的人类行为场景, 并包含了多种典型的正常与异常行为模式, 为评估基于骨架的视频异常检测方法提供了理想的测试平台.

开放集数据集 UBnormal 含 29 个场景共 551 个合成视频片段, 异常类型包括跌倒、打斗等, 严格遵循开放集策略划分数据集, 训练集仅含正常行为, 侧重评估模型泛化能力; HR-STC 含 13 个场景, 训练集包含 330 个视频共 274515 帧, 测试集包含 107 个视频共 42883 帧, 涵盖 130 个异常事件, 主要考验模型检测精度与鲁棒性; HR-Avenue 含 37 个视频, 训练集 15 328 帧、测试集 15 324 帧, 视频分辨率为 640×360, 涵盖 47 种异常行为, 聚焦走廊场景下突发异常的检测能力.

在评价指标方面, 本文采用视频异常检测领域中广泛使用的接收者操作特征曲线 (receiver operating characteristic, ROC) 下的面积 (area under the curve, AUC) 来衡量模型的检测性能. AUC 值通过逐渐调整正常分数的阈值来计算, 反映了模型在不同阈值下区分正常与异常行为的能力. AUC 值越高, 表明模型的异常检测性能越出色. 这一指标能够综合反映模型的真正率 (true positive rate, TPR) 和假正率 (false positive rate, FPR), 是评估视频异常检测任务中模型性能的权威指标.

3.2 实验设置

本实验基于 PyTorch 框架, 在 NVIDIA A10 GPU 上采用混合精度训练 (FP16) 与分布式策略 (DDP). 模型在多样化行为生成子网络中, 以 U-Net 为骨干网络, 集成动态时空注意力模块 (DSTAM, 4 头注意力机制) 增强条件约束; 同时, 独立部署行为-场景对齐模块 (BSAM), 通过构建动态记忆库 (行为特征维度 128, 场景特征维度 256) 实现跨模态语义对齐, 记忆库采用每批次更新 10% 的在线策略维持动态性. DSTAM 条件自编码器联合扩散模型生成多样化动作假设, 训练阶段采用 AdamW 优化器 (学习率 0.001), 扩散步数 10 步, 批量大小 32, 窗口长度 6 帧. 数据预处理包含关节点中心化与鲁棒归一化, 测试阶段通过 30 帧滑动平均滤波优化异常评分曲线, BSAM 的行为-场景对齐权重采用余弦相似度计算. 所有实验在数据集上以 100 训练周期完成, 关键参数通过五折交叉验证确定.

3.3 实验结果与分析

3.3.1 对比实验

为了验证 ST-BSAN 模型在视频异常检测任务中的有效性, 本研究将其与当前主流的检测方法进行了对比, 涵盖了基于重建和基于预测的技术. 表 1 展示了各方法在 HR-STC、HR-Avenue 和 UBnormal 数据集上的 AUC (%) 表现, 其中加粗数字表示最佳结果.

表 1 在 HR-STC、HR-Avenue 和 UBnormal 数据集上的帧级 AUC 异常检测结果比较 (%)

Method	HR-STC	HR-Avenue	UBnormal
Conv-AE[11]	69.8	84.8	—
Pred[6]	72.7	86.2	—
MPED-RNN[16]	75.4	86.3	60.6
GEPC[17]	74.8	58.1	53.4
Multi-timescale prediction[18]	77.0	88.3	—
Normal graph[19]	76.5	87.3	—
BiPOCO[20]	74.9	87.0	50.7
STGCAE-LSTM[21]	77.2	86.3	—
SSMTL++[22]	—	—	62.1
COSKAD[23]	77.1	87.8	65.0
MoCoDAD[4]	77.6	89.0	68.3
TrajREC[24]	77.9	89.4	68.0
TSGAD[25]	**81.77**	—	—
FG-Diff[26]	78.6	**90.7**	68.9
ST-BSAN (Ours)	79.9	89.1	**70.1**

在 HR-STC 数据集上, ST-BSAN 取得了 79.9% 的帧级 AUC, 优于基于扩散模型的 MoCoDAD (77.6%) 和基于图卷积的 STGCAE-LSTM (77.2%), 分别提升了 2.3% 和 2.7%. 这一优势源于 ST-BSAN 对时空动态关系的细粒度建模能力, 尤其是通过动态时空注意力模块 (DSTAM) 中的时间自注意力机制 (TSA), 实现了对跨帧时序突变特征的全局建模, 增强了对短时异常行为的时空敏感性. 与传统重构类方法如 Conv-AE (69.8%) 和 MPED-RNN (75.4%) 相比, ST-BSAN 分别提升了 10.1% 和 4.5%, 这主要是因为传统重构类方法采用的静态编码器在动态时空依赖建模方面存在局限性.

在 HR-Avenue 数据集上, ST-BSAN 的帧级 AUC 为 89.1%, 低于 FG-Diff 的 90.7%. 这可能是因为 FG-Diff 借助 2D-DCT 分离运动频域特征, 优先重建全局低频信息, 还通过扰动训练生成类正常样本, 增强对未见过正常变体的泛化, 适配该数据集长时序全局运动建模需求.

在 UBnormal 上, ST-BSAN 的帧级 AUC 达到了 70.1%, 显著优于 SSMTL++ (62.1%) 和 BiPOCO (50.7%), 分别提升了 8.0% 和 19.4%. 这一显著差异可能归因于 ST-BSAN 对场景语义与行为关联性的显式建模策略, 通过分离场景无关的共性特征与场景敏感的特异性特征, 有效抑制了开放集中的跨场景误检, 提升了模型的鲁棒性.

综上所述, ST-BSAN 通过时空动态建模与场景语义对齐的协同优化, 在 3 个数据集上均取得了优异的性能, 尤其在 HR-STC 和 UBnormal 数据集上优势明显. 这表明 ST-BSAN 不仅能够有效捕捉视频中的时空动态信息, 还能够适应场景中的复杂变化, 为异常检测任务提供了精度与鲁棒性兼具的解决方案, 具有重要的实用价值.

3.3.2 消融实验

为验证 ST-BSAN 框架中关键模块的有效性, 本文在 HR-STC 和 UBnormal 数据集上设计了系统性消融实验, 重点评估动态时空注意力 (DSTAM) 与行为-场景对齐 (BSAM) 对视频异常检测性能的贡献度. 实验结果如表 2 所示, 其中“√”表示启用对应模块, “×”表示禁用该模块. SSA 表示空间自注意力机制, TSA 代表时间自注意力机制, BSAM-Dynamic 指支持动态更新的场景记忆库, BSAM-Static 则为固定场景记忆库. 实验重点考察各模块组合对时空建模能力和场景适应能力的影响, 其中 AUC 指标反映综合检测性能.

表 2 消融实验结果 (%)

SSA	TSA	BSAM-Dynamic	BSAM-Static	HR-STC	UBnormal
×	×	×	×	77.6	68.3
√	×	×	×	78.1	68.8
×	√	×	×	77.9	68.6
√	√	×	×	78.4	69.3
√	√	×	√	79.5	69.8
√	√	√	×	**79.9**	**70.1**

从表 2 实验结果可以看出, ST-BSAN 框架通过模块化组合实现了渐进式性能提升: 仅引入动态时空注意力模块 (DSTAM) 即可在基线模型 MoCoDAD 基础上带来 0.8%–1.0% 的 AUC 增益, 验证了时空自注意力机制对长程依赖建模的有效性; 当进一步加入动态行为-场景对齐模块 (BSAM-Dynamic) 后, 模型在 HR-STC 和 UBnormal 数据集上的 AUC 分别跃升至 79.9% 和 70.1%, 较静态场景对齐版本 (BSAM-Static) 提升 0.4%–0.5%, 表明动态记忆更新机制对复杂场景语义突变具有显著适应能力. 这些模块的有序集成与功能互补, 共同构建了更适配复杂场景的异常检测能力, 验证了 ST-BSAN 框架模块化设计的科学性与有效性.

3.3.3 其他实验

为验证模型中扩散主干引入 SSA/TSA 的必要性, 设计表 3 开展专项实验: 控制条件模块均启用 DSTAM (含 SSA+TSA)、BSAM 为动态更新版本, 仅改变扩散主干的注意力机制配置. 结果显示, 当在扩散主干单独引入 SSA 或 TSA 时, 模型在 HR-STC 和 UBnormal 数据集上的帧级 AUC 分别下降 0.8%–1.0%; 同时引入 SSA+TSA 时, AUC 进一步下降 1.3%–1.5%. 这表明 SSA/TSA 仅适配条件模块的“特征提取”任务, 在扩散主干的“噪声估计与运动生成”环节中属于冗余机制, 其引入会干扰扩散过程的平滑性, 导致性能下降, 因此模型设计在扩散主干中排除了这些模块.

表 3 扩散主干引入 SSA/TSA 的有效性验证 (%)

模型配置	HR-STC	UBnormal
ST-BSAN (Ours)	**79.9**	**70.1**
ST-BSAN (扩散主干+SSA)	79.1	69.5
ST-BSAN (扩散主干+TSA)	78.9	69.3
ST-BSAN (扩散主干+SSA+TSA)	78.4	68.8

为优化记忆库容量配置, 开展行为记忆库容量调优实验. 从表 4 结果显示, 行为记忆库容量对模型性能呈非线性影响: 容量过小时, 难以充分覆盖多样正常行为, 易误判合理动作; 容量过大, 会因纳入边缘特征、噪声引入干扰并增加计算开销. 128 槽中等容量能在正常行为表征完整性与对齐计算高效性间达最优平衡, 支撑复杂场景判别鲁棒性. 利用同样方法, 经实验验证, 场景记忆库、行为-场景对齐记忆库容量分别设为 32 槽、64 槽.

表 4 行为记忆库容量对模型 AUC 的影响 (%)

记忆库容量 (槽)	HR-STC	UBnormal
64	79.1	67.5
128	**79.9**	**70.1**
256	79.5	69.8
512	79.0	68.3

为验证分层训练策略对性能的增益, 开展不同训练方法对比实验. 从表 5 结果可见, 两阶段训练 (预训练+微调) 在 HR-STC、UBnormal 数据集优势显著, 较端到端训练分别提升 0.8%、1.1%, 验证“预训练+联合微调”分层优化的有效性. 预训练学习正常行为物理规律, 为特征提取、模态生成奠基; 联合微调引入场景语

义约束, 增强复杂场景行为判别力. 仅预训练因缺语义约束, 跨场景检测表现不足, 凸显场景语义对齐模块对泛化能力的关键作用.

表 5 不同训练方法对模型 AUC 的影响 (%)

训练方法	HR-STC	UBnormal
端到端训练	79.1	69.0
仅预训练	79.4	69.3
两阶段训练 (预训练+微调)	**79.9**	**70.1**

因特征对齐损失 (关联 λ) 与场景对齐损失 (关联 μ) 对异常检测效果影响关键, 结合多任务学习经验及损失项量纲分析, 先将 λ 初始范围设为{0.1, 0.2, 0.3, 0.4, 0.5}、μ 设为{0.1, 0.2, 0.3}开展五折交叉验证. 从表 6 的实验结果表明, 不同 λ 与 μ 组合对模型在 HR-STC 和 UBnormal 数据集上的 AUC 表现影响显著, 当 λ 取 0.3、μ 取 0.2 时, 模型在两个数据集上均达最优, 此参数组合能有效平衡两种损失, 助力模型在不同场景异常检测中发挥更佳性能; 而随着 λ、μ 偏离该组合, 模型 AUC 呈下降趋势, 该实验结果验证了多损失联合优化中参数协同的重要性.

表 6 五折交叉验证确定 λ 与 μ 参数的平均 AUC 结果表 (%)

λ	μ	HR-STC	UBnormal
0.1	0.1	78.2	68.8
0.1	0.2	78.5	69.2
0.1	0.3	77.9	68.5
0.2	0.1	79.0	69.5
0.2	0.2	79.3	69.8
0.2	0.3	78.7	69.1
0.3	0.1	79.5	69.9
0.3	0.2	**79.9**	**70.1**
0.3	0.3	79.4	69.7
0.4	0.1	79.1	69.3
0.4	0.2	79.0	69.5
0.4	0.3	78.5	68.9
0.5	0.1	78.5	68.7
0.5	0.2	78.1	68.5
0.5	0.3	77.7	68.2

表 7 数据显示, 参数优化遵循“分层协同”逻辑有序开展. 在固定$\alpha = 0.5$ (中性值) 探索γ的影响时, γ本质是场景对齐度中实时行为与生成行为的融合权重, 当γ取 0.5 时, HR-STC 与 UBnormal 数据集的平均 AUC 分别达 79.0%、69.5%, 实现场景对齐度模块内实时与生成行为约束的最优平衡. 在此基础上, 固定$\gamma =$ 0.5 (对齐度最优值) 进一步搜索α (行为异常与场景冲突的融合权重), 发现$\alpha = 0.6$对行为物理合理性与场景语义冲突的协同约束效果最佳. 最终, 表 7 的结果也验证, 经分层优化的最优参数组合 ($\gamma = 0.5$, $\alpha = 0.6$) 可使行为物理规律约束与场景语义逻辑约束深度耦合, 在不同场景异常检测任务中均展现最优性能.

表 7 五折交叉验证确定 γ 和 α 参数的平均 AUC 结果表 (%)

γ	α	HR-STC	UBnormal
0.3		78.2	67.8
0.4		78.8	68.6
0.5	0.5	79.0	69.5
0.6		78.7	69.2
0.7		78.1	68.9
	0.4	77.9	68.9
	0.5	79.0	69.5
0.5	0.6	**79.9**	**70.1**
	0.7	79.3	69.7
	0.8	79.1	69.3

3.4 可视化分析

为验证 ST-BSAN 框架的检测性能, 图 4 在 HR-STC、UBnormal 和 HR-Avenue 数据集场景下, 对核心模块 DSTAM 与 BSAM 开展可视化验证. 对于 DSTAM 模块, 通过正常帧与异常帧的对比呈现, 可清晰观测到其对异常行为时空动态关联的捕捉能力. 当行人出现摔倒、扔书包等异常时, 模型自动聚焦人体骨骼非相邻节点的相关性, 以亮色关联突出显示, 与正常行为的局部相邻关节关联形成鲜明区别, 精准识别跨肢体的运动传导及协同模式. 针对 BSAM 模块, 同一行为在不同场景展现出差异化判定结果, 如打电话踱步行为, 在斑马线场景下, 模型依据场景原型判定其异常, 以边界框精准圈定; 在小巷场景中则判定为正常. 全方位体现 ST-BSAN 框架“时空动态特征捕捉+场景语义理解”的协同优势, 为复杂场景下的视频异常检测提供有力的支撑.

为直观呈现模型的异常检测性能, 图 5 展示了视频流中的异常评分时序变化与关键帧可视化结果. 异常评分曲线在正常行为段 (如行走) 保持较低水平, 当异常事件发生时, 评分显著上升并持续高于通过五折交叉验证确定的判定阈值 (HR-STC 数据集最优阈值为 0.55); 对应视频帧中, 边界框精准锁定异常目标 (奔跑者), 背景场景信息 (街道环境) 为行为判定提供语义支撑. 异常区域内出现的分数冒低现象, 是由于扩散模型生成了与“快速行走”相似的轨迹样本, 导致行为异常分数暂时降低. 异常事件结束后, 评分快速回落至基线水平, 这种变化直观地反映了异常事件的发生时刻, 证明了模型在异常检测中的有效性.

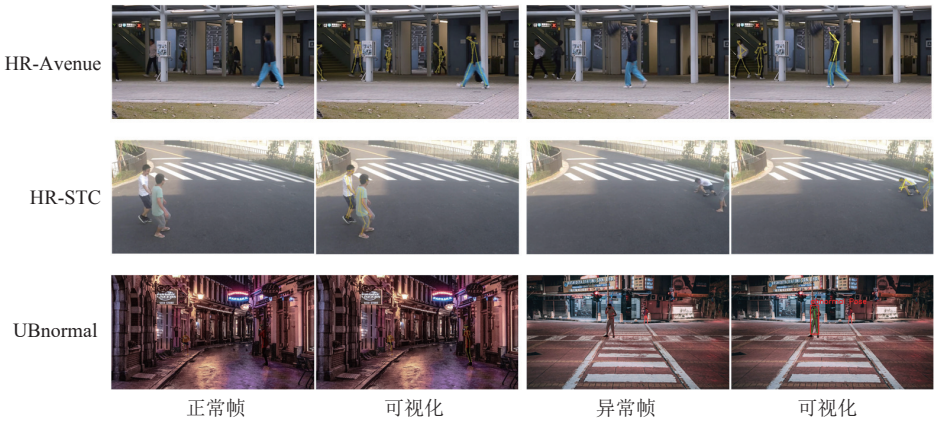

图 4　多数据集下 DSTAM 时空动态关联与 BSAM 场景语义的可视化验证

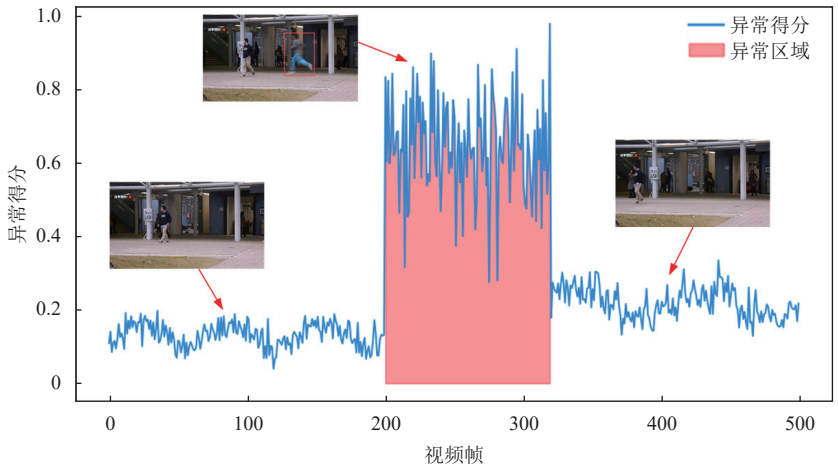

图 5　数据集 HR-Avenue #04 异常分数可视化图

4　结语

针对视频异常检测中时空动态突变难捕捉、行为-场景语义协同性弱的问题, 本文构建融合动态时空注意力模块 (DSTAM) 与行为-场景对齐模块 (BSAM) 的检测框架: DSTAM 突破静态建模局限, 借时空注意力动态交互精准捕获人体动作突变; BSAM 以动态记忆库机制, 实时适配场景语义漂移与正常行为模式演进, 强化跨模态语义约束鲁棒性. 实验验证, 该框架有效提升异常检测精度, 未来可进一步拓展模型对多样行为的泛化能力, 突破现有扩散模型仅基于已有行为扩充的局限, 实现对更丰富多样正常行为的覆盖.

参考文献

1 付荣华, 刘成明, 刘合星, 等. 骨架引导的多模态视频异常行为检测方法. 郑州大学学报 (理学版), 2024, 56(1): 16–24. [doi: 10.13705/j.issn.1671-6841.2022284]

2 Yan SJ, Xiong YJ, Lin DH. Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the 32th AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018.

3 Yang GQ, Luo ZM, Gao JZ, *et al*. A multilevel guidance-exploration network and behavior-scene matching method for human behavior anomaly detection. Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne: ACM, 2024. 5865–5873.

4 Flaborea A, Collorone L, Di Melendugno GMD, *et al*. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 10284–10295.

5 Zhao B, Li FF, Xing EP. Online detection of unusual events in videos via dynamic sparse coding. Proceedings of the 2011 Conference on Computer Vision and Pattern Recognition. Colorado Springs: IEEE, 2011. 3313–3320.

6 Liu W, Luo WX, Lian DZ, *et al*. Future frame prediction for anomaly detection—A new baseline. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6536–6545.

7 Georgescu MI, Bărbălău A, Ionescu RT, *et al*. Anomaly detection in video via self-supervised and multi-task learning. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 12737–12747.

8 Duan HD, Wang JQ, Chen K, *et al*. DG-STGCN: Dynamic spatial-temporal modeling for skeleton-based action recognition. arXiv:2210.05895, 2022.

9 Chen YX, Zhang ZQ, Yuan CF, *et al*. Channel-wise topology refinement graph convolution for skeleton-based action recognition. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 13339–13348.

10 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. 计算机应用, 2024, 44(8): 2551–2557. [doi: 10.11772/j.issn.1001-9081.2023081157]

11 Hasan M, Choi J, Neumann J, *et al*. Learning temporal regularity in video sequences. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 733–742.

12 郑敬添. 基于生成模型的视频人体异常行为检测方法研究[硕士学位论文]. 重庆: 重庆理工大学, 2024. [doi: 10.27753/d.cnki.gcqgx.2024.000794]

13 孙澈, 武玉伟, 贾云得. 上下文建模与推理的视频异常事件检测. 计算机学报, 2024, 47(10): 2368–2386.

14 Chen CLZ, Liu XY, Song MK, *et al*. Unveiling context-related anomalies: Knowledge graph empowered decoupling of scene and action for human-related video anomaly detection. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(8): 8071–8085. [doi: 10.1109/TCSVT.2025.3546107]

15 Yang ZY, Radke RJ. Context-aware video anomaly detection in long-term datasets. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 4002–4011.

16 Morais R, Le V, Tran T, *et al*. Learning regularity in skeleton trajectories for anomaly detection in videos. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 11988–11996.

17 Markovitz A, Sharir G, Friedman I, *et al*. Graph embedded pose clustering for anomaly detection. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10536–10544.

18 Rodrigues R, Bhargava N, Velmurugan R, *et al*. Multi-timescale trajectory prediction for abnormal human activity detection. Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass: IEEE, 2020. 2615–2623.

19 Luo WX, Liu W, Gao SH. Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection. Neurocomputing, 2021, 444: 332–337. [doi: 10.1016/j.neucom.2019.12.148]

20 Kanu-asiegbu AM, Vasudevan R, Du XX. BiPOCO: Bi-directional trajectory prediction with pose constraints for pedestrian anomaly detection. arXiv:2207.02281, 2022.

21 Li NJ, Chang FL, Liu CS. Human-related anomalous event detection via spatial-temporal graph convolutional autoencoder with embedded long short-term memory network. Neurocomputing, 2022, 490: 482–494. [doi: 10.1016/j.neucom.2021.12.023]

22 Barbalau A, Ionescu RT, Georgescu MI, *et al*. SSMTL++: Revisiting self-supervised multi-task learning for video anomaly detection. Computer Vision and Image Understanding, 2023, 229: 103656. [doi: 10.1016/j.cviu.2023.103656]

23 Flaborea A, di Melendugno GMD, D'Arrigo S, *et al*. Contracting skeletal kinematics for human-related video anomaly detection. Pattern Recognition, 2024, 156: 110817. [doi: 10.1016/j.patcog.2024.110817]

24 Stergiou A, De Weerdt B, Deligiannis N. Holistic representation learning for multitask trajectory anomaly detection. Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2024. 6715–6725.

25 Noghre GA, Pazho AD, Tabkhi H. An exploratory study on human-centric video anomaly detection through variational autoencoders and trajectory prediction. Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2024. 995–1004.

26 Tan XF, Wang HS, Geng X, *et al*. Frequency-guided diffusion model with perturbation training for skeleton-based video anomaly detection. arXiv:2412.03044, 2024.

(校对责编: 张重毅)