文章编号：1672-3961(2025)05-0110-10 DOI：10.6040/j.issn.1672-3961.0.2025.031

基于视频描述增强和双流特征融合的视频异常检测方法

郑晓[1]，陈鹤[2]，周东傲[3*]，宫永顺[1]
（1.山东大学软件学院，山东 济南 250101；2.山东交控科技有限公司，山东 济南 250022；3.中国人民解放军军事科学院，北京 100091）

摘要：针对现有异常检测方法在语义上下文利用和时空特征建模方面的不足，提出一种基于视频描述增强和双流特征融合的视频异常检测方法。自动化提取视频描述，利用对比语言-图像预训练（constrastive language-image pre-training，CLIP）模型进行编码，作为视频上下文语义特征辅助视频异常检测；引入一种时空自适应嵌入模块，分别捕捉视频中细微的时序变化和复杂的空间结构，并进行有效的时空融合；利用精心设计的跨模态对齐模块将上下文语义特征与时空视觉特征进行深度融合，更准确地捕捉异常事件的时空-语义联合特征。试验结果显示，该方法在 ShanghaiTech 和 CUHK Avenue 数据集上的检测指标曲线下面积 A_{UC} 分别达到 97.54%和 90.54%，证明该方法在公开视频异常检测数据集上表现优异，具有强大的鲁棒性，为视频异常检测提供一种有效的解决方案。

关键词：视频异常检测；视频描述；时空自适应嵌入；时序 Transformer；空间 Transformer
中图分类号：TP391 **文献标志码**：A
引用格式：郑晓，陈鹤，周东傲，等. 基于视频描述增强和双流特征融合的视频异常检测方法[J]. 山东大学学报（工学版），2025，55(5)：110-119.
ZHENG Xiao，CHEN He，ZHOU Dongao，et al. Video anomaly detection method based on video caption augmentation and dual-stream feature fusion[J]. Journal of Shandong University（Engineering Science），2025，55(5)：110-119.

Video anomaly detection method based on video caption augmentation and dual-stream feature fusion

ZHENG Xiao[1]，CHEN He[2]，ZHOU Dongao[3*]，GONG Yongshun[1]
（1. School of Software，Shandong University，Jinan 250101，Shandong，China；2. Shandong Jiaokong Technology Co.，Ltd.，Jinan 250022，Shandong，China；3. PLA Academy of Military Science，Beijing 100091，China）

Abstract：To address the limitations in semantic context utilization and spatio-temporal feature modeling in existing anomaly detection methods，a video anomaly detection method based on video caption augmentation and dual-stream feature fusion was proposed. Video captions were automatically extracted and encoded using the contrastive language-image pre-training（CLIP）model to serve as auxiliary semantic context information for anomaly detection. A spatio-temporal adaptive embedding module was introduced to capture subtle temporal variations and complex spatial structures within videos，enabling effective spatio-temporal feature fusion. A cross-modal alignment module was further designed to deeply integrate contextual semantic features with spatio-temporal visual features，allowing more accurate capture of joint spatio-temporal-semantic representations of anomalous events. Experimental results showed that the method achieved area under the curve A_{UC} scores of 97.54% on the ShanghaiTech dataset and 90.54% on the CUHK Avenue dataset. The results confirmed the performance and robustness of the method across multiple public video anomaly detection datasets，providing an effective solution for this critical task.

Keywords：video anomaly detection；video caption；spatio-temporal adaptive embedding；temporal Transformer；spatial Transformer

收稿日期：2025-03-12
基金项目：山东省优秀青年基金（海外）资助项目（2022HWYQ-044）；山东交控科技有限公司科技资助项目（1480024005）
第一作者简介：郑晓（2000— ），男，山东德州人，硕士研究生，主要研究方向为计算机视觉、异常检测。E-mail：xzheng@ mail.sdu.edu.cn
*通信作者简介：周东傲（1990— ），男，湖南新邵人，助理研究员，博士，主要研究方向为人工智能与信号检测。
E-mail：zhoudongao08@ nudt.edu.cn

0　引言

随着城市化进程加速和公共安全需求不断增加,视频监控系统的应用越来越广泛。在大型商场、机场、学校等公共场所广泛部署监控摄像头已成为维护社会秩序和保障公共安全的核心手段。视频监控技术的普及使安全监管成本不断下降,但当前识别暴力、偷窃、交通事故等异常事件仍然主要依赖人工监控。人工监控不仅消耗大量人力物力,还容易出现人为疏漏和误检。随着计算机视觉技术的迅猛发展,视频异常检测的研究受到广泛关注。该技术能够在视频流中自动化识别异常或可疑行为,几乎不需要人工干预,在维护公共安全中的重要性日益凸显。

目前,监控视频异常检测方法大致分为无监督方法[1-3]和弱监督方法[4-6]。无监督方法主要基于正常标签的视频数据进行训练,训练过程中没有明确的异常数据标签,通过训练,模型能够在测试过程中识别偏离正常数据模式的异常事件。一些学者对无监督视频异常检测方法进行研究,例如:文献[2]提出一种卷积自编码器,通过提取外观和运动信息进行异常检测;文献[3]提出一种融合时序多尺度建模的自编码器网络架构,通过构建未来帧预测模型实现异常检测。弱监督方法引入少量异常数据进行训练,通常比无监督方法表现更优。因此,弱监督视频异常检测逐渐成为当前的研究热点。一些学者对弱监督视频异常检测方法进行研究,例如文献[6]通过引入多实例学习进行异常检测。然而,在面对复杂多样的场景及异常事件时,现有检测方法往往难以提供足够的性能支持。

1　相关工作

近年来,一些学者尝试探索时空建模方法提升检测效果,例如:文献[7]通过使用压缩与激励网络(squeeze-and-excitation networks, SENet)将 RGB(red, green, blue)特征与光流特征进行融合,使用卷积长短期记忆(convolutional long short-term memory, ConvLSTM)网络实现全局时空感知,以实现异常检测;文献[8]提出一种结合卷积神经网络(convolutional neural network, CNN)与 Transformer[9]的端到端架构(hybrid CNN and Transformer, TransCNN),利用 CNN 提取空间特征,再利用 Transformer 学习异常事件的依赖关系;文献[10]提出 Transformer 增强双流网络(Transformer enhanced dual-stream network, TDS-Net),同时提取 RGB 特征和光流特征,使用 Transformer 网络进行序列模式学习,从而进行异常检测。然而,当视频中的时空特征发生显著变化时,现有方法往往难以有效捕捉细微差异,导致异常检测精度降低[7-8,10-11]。

视觉-语言模型通过融合上下文语义信息,学习更通用的视觉表示,在各种任务中展示出强大的性能[12-13]。基于此,研究者们开始探索将多模态学习范式应用于异常检测领域,通过引入类别文本提示提升异常行为的识别准确率。文献[14]提出提示增强学习(prompt-enhanced learning, PEL)方法,引入外部知识库 ConceptNet,根据类别文本创建文本提示,与视觉特征进行对齐,以增强视觉特征的语义判别能力;文献[15]提出适配视觉-语言模型的弱监督视频异常检测(adapting vision-language models for weakly supervised video anomaly detection, VadCLIP)方法,利用冻结的对比语言-图像预训练(constrastive language-image pre-training, CLIP)[12]图像编码器提取视觉特征,根据视觉特征进行二分类,同时利用文本编码器编码异常类别的文本信息,与视觉特征进行融合,实现细粒度的视频异常检测。然而,上述方法中简单的类别文本可能不足以描述复杂的现实场景,例如“奔跑”在体育场中可能视为正常行为,但在银行或商场中则可能视为异常行为。

有些研究尝试利用视频描述提升视频异常检测性能,例如:文献[16]提出文本增强的视频异常检测(text empowered video anomaly detection, TEVAD)模型,通过构建视频-文本对比学习框架,强制正常样本的视觉特征与描述文本在嵌入空间中对齐;文献[17]提出基于记忆库的文本引导方法,通过计算测试视频与存储的正常文本-视频原型之间的相似度识别异常。但以上方法仅关注视频与文本的全局相似度,未能建模异常行为的时序演化特性。

为应对上述挑战,本研究提出一种基于视频描述增强和双流特征融合的视频异常检测方法。针对语义表征不足的问题,自动化提取视频描述,借助 CLIP 对视频描述进行编码,以获取视频的上下文语义特征;针对现有方法时空建模不足的问题,提出一种时空自适应嵌入模块,利用时序 Transformer 和空间 Transformer 分别捕捉视频中的时序动态和空间结构,显著提升视频异常检测精度。本研究提出一种跨模态对齐模块,将视频的时空特征与上下文语义特征进行深度融合,帮助模型同时学习视觉信息和语义信息,从而更全面地理解视频内容,提升异常检测的鲁棒性和准确性。本研

究为视频异常检测任务的实际应用提供一种高效且可靠的解决方案。

2 方法设计与实现

本研究所提模型框架图如图1所示。将每个输入视频划分为多个帧数为 n 的视频片段。这些片段通过预训练的膨胀三维卷积网络（inflated 3D convnet，I3D）提取RGB和光流特征，并进行融合处理。将融合后的特征输入本研究设计的时空自适应嵌入模块中，以学习复杂的时空特征变化，有效捕捉视频中的细微时序动态和复杂空间结构。同时，通过基于视频理解模型CogVLM2-Video的视频描述生成模型（video captioning model based on the video understanding model CogVLM2-Video，CogVLM2-Caption）[18]自动提取每个输入视频的视频描述，利用CLIP对视频描述进行编码，得到上下文语义特征。将得到的时空特征和上下文语义特征通过跨模态对齐模块进行特征融合，送入多任务联合学习模块，进行视频异常检测。

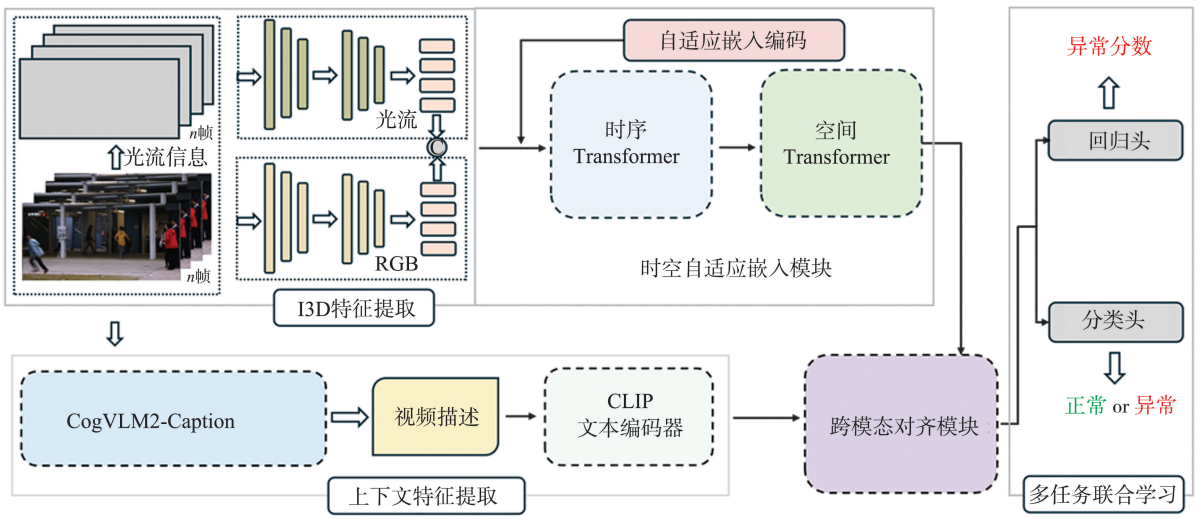


图1 基于视频描述增强和双流特征融合的异常检测框架图

Fig.1 The anomaly detection framework based on video caption augmentation and dual-stream feature fusion

2.1 I3D特征提取

对于每个帧数为 N 的视频输入 $\boldsymbol{O}$，本研究将其划分为 m 个等长视频序列 $\boldsymbol{o}_i$，帧数为 n，$m=N/n$，序列之间互不重叠。视频输入 $\boldsymbol{O}$ 具体表示为

$$\boldsymbol{O}=\{\boldsymbol{o}_i\}_{i=1}^{m}。\tag{1}$$

给每个 $\boldsymbol{o}_i$ 分配一个标签，标签“0”表示该序列为正常视频片段，标签“1”表示该序列为异常视频片段。

I3D模型在Kinetics数据集[19]上进行预训练，学习大量的时空特征表示。这些先验知识能够迁移并应用于各种视频分析任务中。因此，本研究使用I3D模型进行视频时空特征提取，其模型架构如图2所示。

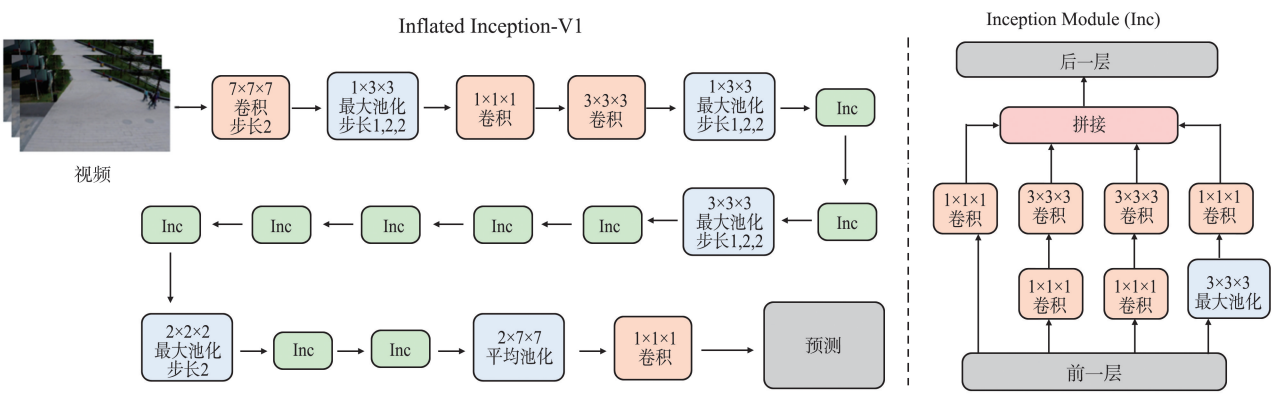


图2 I3D网络结构图

Fig.2 I3D network architecture

该网络基于Inception-V1[20]架构，通过扩展2D卷积和池化操作到3D时空卷积捕捉视频中的时空特征，输入视频分别通过RGB流（提取静态外观特征）和光流（捕捉动态运动信息）双路径处理，最终

输出时空融合特征。本研究基于 I3D 模型进行特征提取,提出视频描述增强方法,自动化提取视频描述,用 CLIP 编码生成语义上下文特征辅助检测,结合时空自适应嵌入模块,独立建模细微时序变化与复杂空间结构,实现高效时空融合;深度融合语义特征与时空视觉特征,精准捕捉时空-语义联合异常特征。

2.2　时空自适应嵌入模块

时空自适应嵌入模块通过结合时空自适应嵌入编码和时空 Transformer,更有效地捕捉视频数据中的时空特征,适用于视频异常行为检测。网络输入为 I3D 模型提取的视频帧特征。这些特征表示视频帧中的高维时空信息。在将特征输入网络前,本研究将每个视频帧的特征划分为相同大小的图像特征块 $\boldsymbol{P}_{(f,j)}$,表示第 f 个视频帧的第 j 个图像块。这种划分方式能够保留每个图像块中的局部时空信息,使模型更精确地捕捉异常行为的空间细节和时间演变信息。

2.2.1　时空自适应嵌入

本研究创新性地引入一种时空自适应嵌入编码。该编码旨在通过动态调整特征表示,更好地识别视频中的复杂时空语义。设计一个可学习的嵌入编码 $\boldsymbol{E}^{\mathrm{a}}$, $\boldsymbol{E}^{\mathrm{a}} \in \mathbf{R}^{T\times M\times d}$,其中 T 为帧数,M 为图像特征块数量,d 为特征嵌入维度,通过优化一组可学习参数动态调整特征表示。该自适应嵌入编码能够为每个时间点和空间位置生成特定的特征表示,从而识别视频中的复杂时空特征语义。在处理过程中,本研究将 $\boldsymbol{P}_{(f,j)}$ 与其对应的时空自适应嵌入编码 $\boldsymbol{E}^{\mathrm{a}}_{(f,j)}$ 进行拼接,将拼接后的特征输入一个多层感知器(multilayer perceptron, MLP)中进行特征变换,获得一个新的时空特征表示 $\boldsymbol{F}^{\mathrm{v}}_{(f,j)}$。$\boldsymbol{F}^{\mathrm{v}}_{(f,j)}$ 结合图像块初始特征和时空自适应嵌入增强信息。上述过程可表示为

$$\boldsymbol{F}^{\mathrm{v}}_{(f,j)} = \mathrm{MLP}(\mathrm{Concat}(\boldsymbol{P}_{(f,j)} + \boldsymbol{E}^{\mathrm{a}}_{(f,j)})), \tag{2}$$

式中 Concat(·)为将 $\boldsymbol{P}_{(f,j)}$ 和 $\boldsymbol{E}^{\mathrm{a}}_{(f,j)}$ 进行拼接。

自适应嵌入的引入使模型能够根据不同输入内容动态调整特征表示,更好地适应视频中的复杂时空特征语义。通过这种动态特征调整,模型能够处理各种场景变化和运动模式,有效提升模型的泛化能力和鲁棒性。

2.2.2　时空 Transformer

为了更好地捕捉视频中的细微时序变化和复杂空间结构,本研究分别引入时序 Transformer 和空间 Transformer 两个子模块,以实现对输入视频进行更有效的时空特征融合。

时序 Transformer 专门用于捕捉视频帧之间的时序依赖关系。异常事件通常与时间序列中的某些模式或变化相关,因此视频数据的时序依赖性是异常检测的一个关键因素。时序 Transformer 的自注意力机制计算每个视频帧之间的相似性,通过加权方式融合帧级特征,获得时间上下文特征表示。给定时空特征表示 $\boldsymbol{F}^{\mathrm{v}}$,分别计算每个时序自注意力层的查询矩阵 $\boldsymbol{Q}$、键矩阵 $\boldsymbol{K}$ 和值矩阵 $\boldsymbol{V}$,计算式分别为

$$\boldsymbol{Q} = \boldsymbol{F}^{\mathrm{v}}\boldsymbol{W}_{Q}, \tag{3}$$

$$\boldsymbol{K} = \boldsymbol{F}^{\mathrm{v}}\boldsymbol{W}_{K}, \tag{4}$$

$$\boldsymbol{V} = \boldsymbol{F}^{\mathrm{v}}\boldsymbol{W}_{V}, \tag{5}$$

式中,$\boldsymbol{W}_{Q}$、$\boldsymbol{W}_{K}$、$\boldsymbol{W}_{V}$ 为可学习参数矩阵。时序自注意力的核心是通过 softmax 函数计算每一帧之间的自注意力分数 $\boldsymbol{A}$,建模时间依赖关系,具体计算式为

$$\boldsymbol{A} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{h}}}}\right)\boldsymbol{V}, \tag{6}$$

式中,d_{h} 为键矩阵的维度,通常用于缩放,以避免梯度消失或梯度爆炸。通过时序自注意力机制,模型能够有效捕捉视频帧之间的细微时序变化,对于识别时序上不连续的异常事件尤为关键。经过时序自注意力处理后,更新后的时空特征 $\boldsymbol{F}^{\mathrm{v}}_{\mathrm{update}}$ 通过一个前馈神经网络(feedforward neural network, FFN)进一步处理:

$$\mathrm{FFN}(\boldsymbol{F}^{\mathrm{v}}_{\mathrm{update}}) = \sigma(\boldsymbol{F}^{\mathrm{v}}_{\mathrm{update}}\boldsymbol{W}_{\mathrm{F}} + \boldsymbol{b}_{\mathrm{F}})\boldsymbol{W}_{\mathrm{N}} + \boldsymbol{b}_{\mathrm{N}}, \tag{7}$$

式中,$\boldsymbol{W}_{\mathrm{F}}$、$\boldsymbol{W}_{\mathrm{N}}$ 为可学习参数矩阵,$\boldsymbol{b}_{\mathrm{F}}$、$\boldsymbol{b}_{\mathrm{N}}$ 为偏置矩阵,σ 为 ReLU 激活函数。这一过程增强了视频特征的非线性变换能力,帮助模型更好地学习复杂的特征表示,提高异常检测的鲁棒性和准确性。

空间 Transformer 用于捕捉每个视频帧内部的空间依赖关系。在视频异常检测中,帧内的空间信息同样重要。空间 Transformer 的架构设计与时序 Transformer 类似,但自注意力机制计算的是帧内各图像块之间的关系。通过这一机制,模型能够有效识别帧内对象与背景之间的复杂关系,在对空间结构的复杂性和多样性建模时表现出色。这种方法不仅有助于捕捉帧内对象的运动特征,还能增强对异常行为的空间识别能力,使模型在识别复杂场景中的异常事件时更加精确、鲁棒。

2.3　上下文特征提取

在现实场景中,视频数据通常蕴含丰富的语义信息,例如场景信息、物体空间关系及其与环境的交互等。这些信息不仅能够帮助模型理解数据的高层特征,还能为异常检测任务提供强有力的先验知识。视频的语义信息通常来源于文本描述、视频标注或视频字幕等,能够提供对场景或对象的详细

解释。针对现有研究方法语义表征不足这一局限性,本研究考虑如何获取丰富的语义信息作为先验知识,增强模型的异常检测能力。直接做法显然是依靠人工对视频进行标注,然而人工标注成本太高,难以满足实际需求。随着多模态大模型的快速发展,自动生成视频描述或视频字幕的技术逐渐成熟[18,21]。这些模型通过联合学习视频和文本模态,自动生成与视频内容相匹配的文本描述,提供丰富的视频语义信息。

本研究提出利用 CogVLM2-Caption 自动生成视频描述,将生成的描述作为语义先验知识,以增强异常检测模型的语义表征能力。CogVLM2-Caption 的生成过程如图 3 所示。从输入视频中提取视频帧,通过 ViT(Vision Transformer)[22]进行编码。这些视觉特征经过 Adapter 网络转换映射到文本特征空间,同时拼接时间戳信息。转换后的视觉特征与经过分词器处理的提示文本特征共同输入 Llama3 模型,以生成视频描述信息。通过引入视频描述,模型能够更精准地捕捉场景中的上下文信息,显著提升异常事件检测精度与鲁棒性。

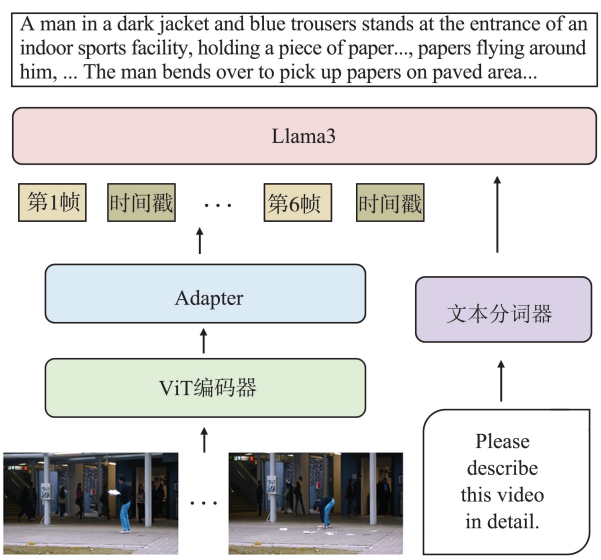


图 3 CogVLM2-Caption 网络结构图

Fig.3 CogVLM2-Caption network architecture

在获取视频描述后,本研究使用 CLIP 文本编码器对视频描述进行编码,获得对应的语义嵌入表示。使用 CLIP 编码的主要原因在于,通过在大规模文本-图像对数据上进行联合训练, CLIP 能够有效捕捉文本与视觉内容之间的语义关联,具备强大的多模态对齐能力。具体编码过程为:视频描述经过分词转换为离散的文本表示,通过嵌入层映射到高维向量空间,通过 Transformer 结构对输入序列进行自注意力计算,捕捉不同令牌间的关联性,最终输出一个全局语义嵌入向量,充分反映视频描述的语义信息。这一上下文语义特征向量将与视频的视觉特征进行融合,最终用于异常检测任务。

2.4 跨模态对齐模块

本研究分别提取视频的时空特征和上下文语义特征。为了充分融合这两种模态信息,本研究提出一种跨模态对齐模块,将上下文语义特征 $\boldsymbol{F}^{\mathrm{s}}$ 作为指导,对时空特征 $\boldsymbol{F}^{\mathrm{v}}_{\mathrm{update}}$ 进行动态调控,实现跨模态信息的有效融合,获得最终特征

$$\boldsymbol{F}_{\mathrm{final}}=\boldsymbol{F}^{\mathrm{v}}_{\mathrm{update}}+\tau(\boldsymbol{F}^{\mathrm{s}}\boldsymbol{W}_{\mathrm{S}}+\boldsymbol{b}_{\mathrm{S}})\odot(\boldsymbol{F}^{\mathrm{v}}_{\mathrm{update}}\boldsymbol{W}_{\mathrm{A}}+\boldsymbol{b}_{\mathrm{A}}),\tag{8}$$

式中,$\boldsymbol{W}_{\mathrm{S}}$、$\boldsymbol{W}_{\mathrm{A}}$ 为可学习参数矩阵,$\boldsymbol{b}_{\mathrm{S}}$、$\boldsymbol{b}_{\mathrm{A}}$ 为偏置矩阵,τ 为 sigmoid 激活函数。$\boldsymbol{F}^{\mathrm{v}}_{\mathrm{update}}$ 经过全连接层进行线性变换;$\boldsymbol{F}^{\mathrm{s}}$ 经过一层全连接层进行映射,通过 sigmoid 激活函数归一化,与时空特征相乘,以调整时空特征的表达强度。为了保留原始时空特征信息,本研究引入残差连接模块,将调整后的时空特征与原始特征相加,以增强模型的稳定性和特征完整性。通过上述双流融合机制,本研究能够在不同模态之间建立有效的特征映射,提升模型对复杂视频场景的理解能力,为后续异常事件检测提供更丰富的时空语义联合表征。

2.5 多任务联合学习

受文献[23]启发,本研究设计一个多任务联合学习网络,能够同时学习异常检测任务和异常分类任务,提高模型对视频数据中异常行为的识别和分类能力。这种联合学习策略不仅可以帮助网络识别异常事件,还可以捕捉异常事件的时间序列模式,使网络在处理复杂异常检测任务时表现得更加鲁棒和高效。

异常分数预测子任务的目标是对异常分数进行回归。将时空自适应嵌入 Transformer 输出的高维时空特征表示 $\boldsymbol{F}_{\mathrm{final}}$ 通过一个全连接层进行线性变换;加入非线性激活函数 ReLU,以增强模型的非线性表达能力;最终的输出层是一个单层全连接网络,输出一个标量,通过 sigmoid 激活函数转换为分数,以表示异常程度。异常分数回归网络的具体过程可表示为

$$\hat{S}=\tau((\max(0,\boldsymbol{F}_{\mathrm{final}}\boldsymbol{W}_1^{\mathrm{reg}}))\boldsymbol{W}_2^{\mathrm{reg}}),\tag{9}$$

式中,$\hat{S}$ 为异常分数,$\boldsymbol{W}_1^{\mathrm{reg}}$、$\boldsymbol{W}_2^{\mathrm{reg}}$ 为可学习参数矩阵。

对于异常分数预测的子任务,使用均方误差损失 L_{mse} 进行监督,计算式为

$$L_{\mathrm{mse}}=\frac{1}{n}\sum_{i=1}^{n}(S_i-\hat{S}_i)^2,\tag{10}$$

式中,S_i 为真实标签,$\hat{S}_i$ 为预测的异常分数。

异常类别分类子任务的网络和异常分数回归网络相似,区别在于分类任务中不使用 sigmoid 函数进行激活,网络的输出层直接输出视频片段的类别标签。这种设计允许模型根据视频的时空特征直接预测异常事件的具体类别,实现异常行为的精确分类。异常分类网络的具体过程可以表示为

$$\hat{Y}=(\max(0,\boldsymbol{F}_{\text{final}}\boldsymbol{W}_1^{\text{cls}}))\boldsymbol{W}_2^{\text{cls}}, \tag{11}$$

式中,$\hat{Y}$ 为预测类别,$\boldsymbol{W}_1^{\text{cls}}$、$\boldsymbol{W}_2^{\text{cls}}$ 为可学习参数矩阵。

对于异常类别分类任务,本研究采用交叉熵损失函数 L_{cls} 优化模型,计算式为

$$L_{\text{cls}}=-\sum_{i=1}^{C}\hat{Y}_i\ln Y_i+\alpha\|\boldsymbol{W}\|_2^2, \tag{12}$$

式中:$\|\boldsymbol{W}\|_2^2$ 为 L2 正则项,防止模型过拟合;α 为超参数,用于平衡损失和正则化项。

最终的多任务联合学习目标函数为

$$L_{\text{final}}=L_{\text{cls}}+\beta L_{\text{mse}}, \tag{13}$$

式中,β 为超参数,用于调整不同损失在联合学习中的相对重要性。通过联合学习异常分数回归和异常类别分类,模型能够在视频异常检测任务上取得更好的性能。

3　试验验证及分析

3.1　试验细节

所有试验均在 NVIDIA Tesla A100 40G 上进行,试验环境为 CUDA 11.6,Pytorch 版本为 1.12.1。试验中设置 $n=16$,即将视频序列划分为包含 16 个不重叠帧的视频片段,并将连续 5 个视频片段(即 80 帧)输入 I3D 网络以提取视频特征。本研究从 I3D 网络的最后一个池化层提取维度为 1 024 的特征,并将 RGB 和光流的输出特征拼接,作为最终视频特征。可学习的嵌入编码维度设置为 10。时序 Transformer 和空间 Transformer 均采用单层网络结构,注意力头数设定为 4。训练过程中,使用 Adam 优化器[24]更新网络参数,学习率设置为0.000 3,训练轮数为 300。

3.2　数据集和验证指标

为了更好地验证本研究方法的有效性和鲁棒性,本研究使用帧级检测指标曲线下面积 A_{UC} 作为性能评估指标,测试模型在各数据集上的表现。本研究在以下 3 个数据集上进行试验验证。

(1) CUHK Avenue 数据集[25]。该数据集包含 16 个训练视频和 21 个测试视频,视频拍摄地点为香港科技大学校园,视频分辨率为 640 像素×360 像素,帧率为 25 帧/s,每个视频片段时长约为 2 min。异常行为包括逆行、异常跑步及在非正常区域停留等。CUHK Avenue 数据集中的视频场景背景变化较大,且行人的行为多样化,使异常检测更具复杂性。

(2) ShanghaiTech 数据集[26]。该数据集涵盖 13 个城市公共场所场景,如商场、地铁站、校园等,场景复杂且人流密集,光照条件和摄像机角度多变。包含 330 个训练视频和 107 个测试视频,帧率为 15 帧/s,涵盖打架、偷窃、摔倒、突然奔跑等 130 个异常事件。由于场景背景复杂,正常行为与异常行为的边界模糊,人群密度高,异常检测任务具有很大的挑战性。

(3)LAD2000 数据集[23]。该数据集包含 1 440 个训练视频和 560 个测试视频,视频分辨率为 226 像素×400 像素,帧率为 25 帧/s。视频主要来自公共网站和监控摄像头等,涉及停车场、车站、街道等 1 895 个场景,包含车祸、拥挤、坠落、打架等 14 类异常事件。LAD2000 数据集涵盖多种场景和异常事件,因此在异常检测任务中具有极强的实用性和挑战性。

3.3　试验验证

3.3.1　对比试验

本研究将所提方法与多种视频异常检测方法进行对比,不同方法的 A_{UC} 指标比较结果如表 1 所示,其中最优结果加粗表示。在 CUHK Avenue 数据集上,本研究方法的 A_{UC} 为 90.54%,优于其他对比方法,表明本研究方法在固定场景的异常检测任务中表现出显著优势,其强大的时空特征提取能力及额外引入的上下文视频描述信息使模型能够更加精确地识别各种异常行为。在 ShanghaiTech 数据集上,本研究方法表现更加突出,A_{UC} 达到 97.54%,显著优于其他对比方法,表明本研究方法在复杂多样化的场景下具有更强的鲁棒性,尤其是在处理多样的场景和光照条件变化时表现更为优越。LAD2000 数据集包含低分辨率、长时间的视频序列,在该数据集中,本研究方法同样表现出色,A_{UC} 达到 86.93%,明显高于其他对比方法,进一步验证本研究方法不仅适用于高分辨率、短时视频的异常检测任务,在低分辨率、长时视频场景中同样展现出强大的泛化能力和出色的异常检测性能。上述试验结果表明,通过引入视频描述,本研究方法能够进一步提高对复杂异常行为的理解能力,有效提升视频异常检测性能。

表1 不同方法对比结果
Table 1 Comparison results of different methods

方法	网络	A_{UC}/%		
		CUHK Avenue	ShanghaiTech	LAD2000
文献[23]方法	I3D	89.33	92.97	86.28
STD[27]	Spatio temporal dissociation	87.10	73.70	—
TransCNN[8]	Hybrid CNN	89.60	94.60	—
文献[7]方法	I3D	87.47	—	86.49
TDS-Net[10]	I3D	89.02	95.82	86.07
PEL[14]	I3D(类别文本)	—	97.32	—
VadCLIP[15]	CLIP(类别文本)	—	97.49	—
文献[28]方法	Auto-encoder	83.10	83.10	—
文献[29]方法	Context-aware	88.50	74.10	—
本研究方法	I3D(视频描述)	**90.54**	**97.54**	**86.93**

注:时空解耦(spatio-temporal dissociation, STD)方法。“—”表示该方法在该数据集上没有进行测试或未报告结果。

为评估方法的计算效率,本研究在 CUHK Avenue 数据集上对不同方法的推理时间进行比较,结果如表2所示。试验测量了从 I3D 特征加载到模型输出检测结果的完整推理时间。结果表明,相比文献[28]提出的方法,本研究方法的推理时间更短。这主要得益于本研究使用的视频描述提取网络和上下文编码器均为固定预训练模型,不参与训练或微调,在提高检测精度的同时,并未引入显著的计算负担。虽然本研究方法的推理时间相比文献[7]和 TDS-Net[10]方法更长,但已能够满足实时检测要求。试验结果充分验证本研究在计算效率方面的有效性。

表2 不同方法推理时间对比
Table 2 Comparison of inference time among different methods

方法	推理时间/s
文献[7]方法	0.18
文献[28]方法	0.26
TDS-Net[10]	0.13
时空自适应 Transformer	0.15
本研究方法	0.21

注:时空自适应 Transformer 为本研究方法未引入上下文特征提取模块和跨模态对齐模块。

3.3.2 消融试验

本研究通过消融试验验证不同模块和特征对视频异常检测性能的影响。

在 CUHK Avenue 数据集上分析不同模态特征及不同模块对视频异常检测性能的影响,分别验证 RGB 特征、光流特征及二者融合特征的效果(均使用 I3D 骨干网络提取特征),对比循环神经网络(recurrent neural network, RNN)、门控循环单元(gated recurrent unit, GRU)和长短期记忆(long short-term memory, LSTM)等时序模型在不同模态特征下的表现,结果如表3所示,其中最优结果加粗表示。对于 RGB 和光流的单模态特征,LSTM 的 A_{UC}分别为 64.80% 和 66.60%,明显优于 RNN 和 GRU;对于 RGB 和光流的融合特征,RNN 表现最佳,A_{UC}达到 72.20%,高于 GRU 和 LSTM。上述结果表明,融合 RGB 和光流特征对提升视频异常检测性能至关重要。在引入 Transformer 后,不同模态特征的检测性能均显著提升。在引入时空 Transformer(无自适应嵌入编码)后,相比于仅使用 Transformer 模块,性能获得提升,表明本研究提出的时空 Transformer 可以更有效捕捉细微时序动态和复杂空间结构。当引入本研究提出的时空自适应 Transformer 模型后,异常检测性能进一步提升,有效验证了所提时空自适应 Transformer 的有效性。在引入视频描述后,本研究的模型在各个模态下均达到最佳检测性能,尤其在 RGB 和光流融合特征下,模型的 A_{UC}达到 90.54%,进一步验证上下文特征提取模块的作用。消融试验结果证明各模块能够显著提升模型性能。

表3 不同时序模型在不同模态特征下的性能比较
Table 3 Performance comparison of different temporal models with various modal features

模型	A_{UC}/%		
	RGB 特征	光流特征	RGB 和光流特征融合
RNN	52.60	51.60	72.20
GRU	61.40	62.40	63.20
LSTM	64.80	66.60	68.20
Transformer	86.74	85.09	89.02
时空 Transformer	87.33	85.68	89.50
时空自适应 Transformer	87.82	85.93	89.76
本研究模型	**88.60**	**86.64**	**90.54**

本研究针对多任务联合学习模块对模型性能的影响进行试验分析,结果如表4所示。当仅使用异常分数预测子任务学习时,A_{UC}为 88.37%;当使用多任务联合学习模块时,性能取得显著提升。

在 CUHK Avenue 数据集上,超参数 β 对模型性能影响的试验结果如图 4 所示。当 $\beta=10$(即分类损失与回归损失的权重比为 1∶10)时,本研究方法获得最佳性能,A_{UC} 达到 90.54%。上述结果表明,在多任务学习中,合理设置 β 可以有效平衡分类损失和回归损失,提升模型整体性能。本研究的所有试验均设置 $\beta=10$,以确保在不同任务之间取得最佳权衡,实现最优的异常检测效果。

表 4　不同损失函数对模型性能影响

Table 4　The impact of different loss functions on model performance

L_{cls}	L_{mse}	A_{UC}/%
×	√	88.37
√	√	90.54

注:√表示使用该损失函数,×表示不使用该损失函数。

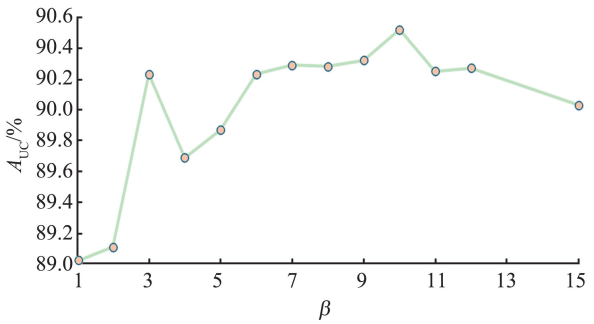
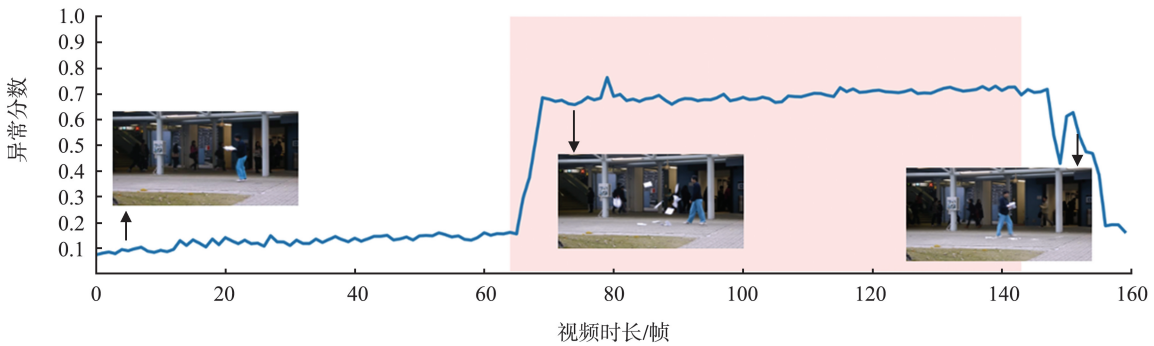


图 4　参数 β 对模型性能的影响

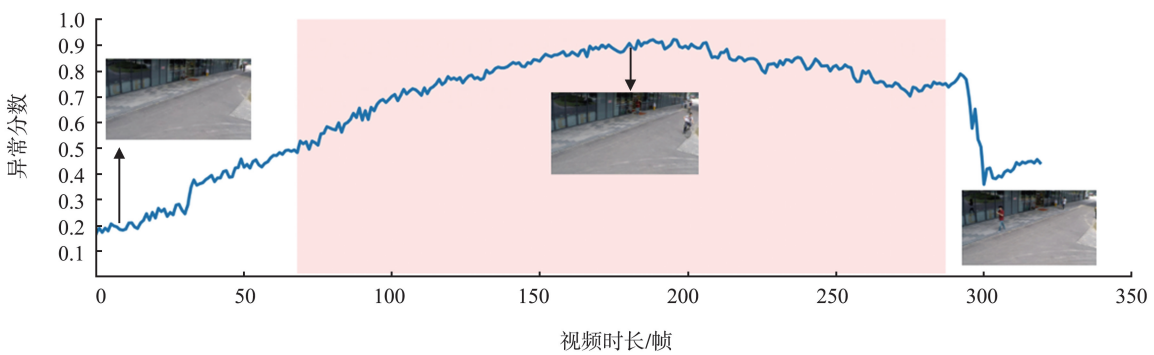
Fig.4　The impact of parameter β on model performance

3.4　可视化结果分析

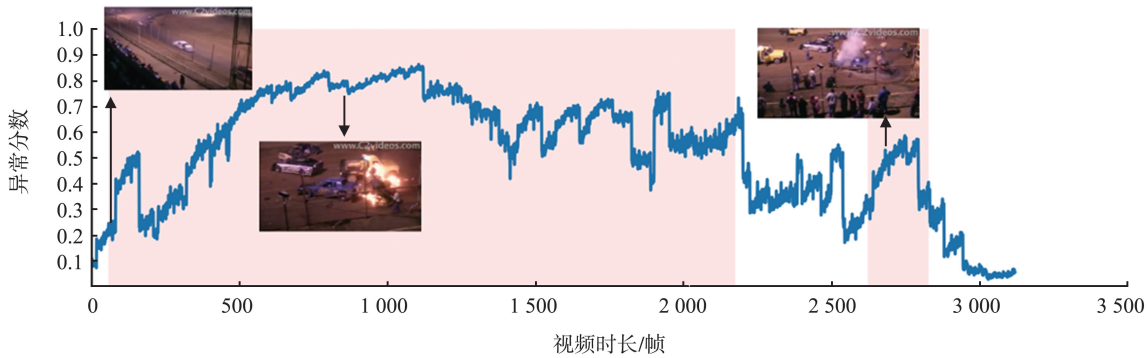
本研究方法在 3 个数据集上的异常检测效果如图 5 所示,其中蓝线表示模型预测的异常分数,粉色背景表示真实的异常标签。选取 CUHK Avenue 数据集中 Avenue_a_20 测试视频进行检测,结果如图 5(a)所示。该视频中,一个男孩在某个时间段突然将手中的文件抛出,触发异常行为。随着这一异常事件的发生,模型预测的异常分数显著上升。当男孩捡回文件并恢复正常行为后,异常分数逐渐下降,回归到正常范围。选取 ShanghaiTech 数据集中 02_0161 测试视频进行检测,结果如图 5(b)所示。视频中出现一名骑自行车的人,在该场景中为异常行为,导致模型异常分数升高。随着自行车骑行者离开视频视线范围,场景恢复正常,异常分数显著下降。选取 LAD2000 数据集中 v_Fire_a_s005_c001测试视频进行检测,结果如图 5(c)所示。视频描述一起赛车相撞后引发火灾的异常事件。在车辆起火的时刻,异常分数迅速上升;由于视频中途拍摄角度发生变化,起火车辆短暂未被捕捉,导致异常分数暂时降低。结果表明,本研究方法能够有效区分大多数视频片段中的正常帧和异常帧,展示出较强的鲁棒性和精度。本研究在 CUHK Avenue 数据集上的一个失败案例如图 5(d)所示。视频最左侧区域存在行人异常行为,但该行人身体部分被遮挡,导致视觉信息不完整,CogVLM2-Caption 模型提取的视频描述未能包含异常行为描述。受此影响,本研究提出的方法未能准确识别该异常事件。



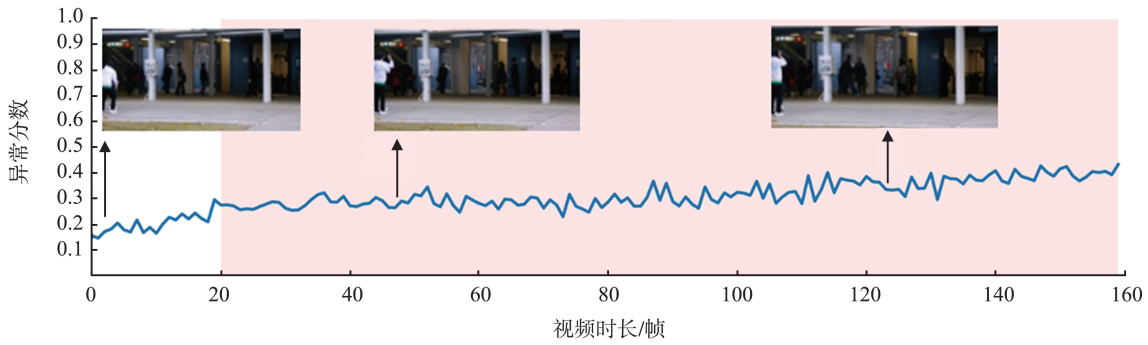
(a) CUHK Avenue数据集中Avenue_a_20测试视频检测结果



(b) ShanghaiTech数据集中02_0161测试视频检测结果



(c) LAD2000数据集中v_Fire_a_s005_c001测试视频检测结果



(d) CUHK Avenue数据集中Avenue_a_18测试视频检测结果

图 5 本研究方法在各数据集上的检测效果

Fig.5 The detection performance of the proposed method across various datasets

4 结论

本研究提出一种全新的基于视频描述增强上下文语义特征和双流特征融合的视频异常检测方法,以提升现有方法在复杂场景下的检测性能。通过自动化提取视频描述并结合 CLIP 进行编码,有效提升模型对视频内容的语义理解;设计的时空自适应嵌入模块显著加强复杂时空特征的感知能力,时序 Transformer 与空间 Transformer 的协同建模进一步提升模型对细微时序变化与空间结构的捕捉能力;提出的跨模态对齐模块实现视觉与语义特征的深度融合,从而更精准地检测视频异常。试验结果表明,本研究方法在 3 个公开数据集上均取得卓越的性能,验证了模型的鲁棒性和有效性,为视频异常检测提供一种创新且实用的解决方案。但当前模型在遮挡推理和细粒度语义区分上仍存在局限。这些限制主要源于空间 Transformer 对遮挡区域的特征补全能力不足,文本描述与视觉异常的对齐粒度不够精细。未来可以进一步优化模型结构和训练策略,以应对更复杂的场景和多样化的异常事件。

参考文献:

[1] DENG H Q, ZHANG Z X, ZOU S H, et al. Bi-directional frame interpolation for unsupervised video anomaly detection[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2023: 2633-2642.

[2] CHANG Y P, TU Z G, XIE W, et al. Video anomaly detection with spatio-temporal dissociation[J]. Pattern Recognition, 2022, 122: 108213.

[3] 吕浩, 易鹏飞, 刘瑞, 等. 用于视频异常检测的时序多尺度自编码器[J]. 图学学报, 2022, 43(2): 223-229.
LYU Hao, YI Pengfei, LIU Rui, et al. Sequential multi-scale autoencoder for video anomaly detection[J]. Journal of Graphics, 2022, 43(2): 223-229.

[4] DI MAURO M, GALATRO G, FORTINO G, et al. Supervised feature selection techniques in network intrusion detection: a critical review[J]. Engineering Applications of Artificial Intelligence, 2021, 101: 104216.

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

[6] SULTANI W, CHEN C, SHAH M. Real-world anomaly detection in surveillance videos[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 6479-6488.

[7] XU D, WU P, YUAN L. Video anomalous behaviour detection based on compressed-inflated attention module

[J]. Journal of Computer Science and Electrical Engineering, 2024, 6(2): 2663-1946.

[8] ULLAH W, HUSSAIN T, ULLAH F U M, et al. TransCNN: hybrid CNN and Transformer mechanism for surveillance anomaly detection[J]. Engineering Applications of Artificial Intelligence, 2023, 123: 106173.

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: ACM, 2017: 6000-6010.

[10] HUSSAIN A, ULLAH W, KHAN N, et al. TDS-Net: Transformer enhanced dual-stream network for video anomaly detection[J]. Expert Systems with Applications, 2024, 256: 124846.

[11] 黄少年, 文沛然, 全琪, 等. 基于多支路聚合的帧预测轻量化视频异常检测[J]. 图学学报, 2023, 44(6): 1173-1182.
HUANG Shaonian, WEN Peiran, QUAN Qi, et al. Future frame prediction based on multi-branch aggregation for lightweight video anomaly detection[J]. Journal of Graphics, 2023, 44(6): 1173-1182.

[12] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. [S.l.]: PMLR, 2021: 8748-8763.

[13] LI J N, LI D X, XIONG C M, et al. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//International Conference on Machine Learning. Baltimore, USA: PMLR, 2022: 12888-12900.

[14] PU Y J, WU X Y, YANG L L, et al. Learning prompt-enhanced context features for weakly-supervised video anomaly detection[J]. IEEE Transactions on Image Processing, 2024, 33: 4923-4936..

[15] WU P, ZHOU X, PANG G, et al. VadCLIP: adapting vision-language models for weakly supervised video anomaly detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2024: 6074-6082.

[16] CHEN W L, MA K T, YEW Z J, et al. TEVAD: improved video anomaly detection with captions[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Vancouver, Canada: IEEE, 2023: 5549-5559.

[17] SHI Y Z, YAMASHITA T, HIRAKAWA T, et al. Caption-guided interpretable video anomaly detection based on memory similarity[J]. IEEE Access, 2024, 12: 63995-64005.

[18] HONG W Y, WANG W H, DING M, et al. CogVLM2: visual language models for image and video understanding[EB/OL]. (2024-08-29)[2025-03-01]. https://arxiv.org/abs/2408.16500v1

[19] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017: 4724-4773.

[20] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015: 1-9.

[21] CHEN T S, SIAROHIN A, MENAPACE W, et al. Panda-70M: captioning 70M videos with multiple cross-modality teachers[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024: 13320-13331.

[22] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. (2021-06-03)[2025-03-01]. https://arxiv.org/abs/2010.11929v2

[23] WAN B Y, JIANG W H, FANG Y M, et al. Anomaly detection in video sequences: a benchmark and computational model[J]. IET Image Processing, 2021, 15(14): 3454-3465.

[24] KINGMA D P, BA J. Adam: a method for stochastic optimization[EB/OL]. (2017-01-30)[2025-03-01]. https://arxiv.org/abs/1412.6980v9

[25] LU C W, SHI J P, JIA J Y. Abnormal event detection at 150 FPS in MATLAB[C]//2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE, 2013: 2720-2727.

[26] 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). Venice, Italy: IEEE, 2017: 341-349.

[27] CHANG Y P, TU Z G, XIE W, et al. Video anomaly detection with spatio-temporal dissociation[J]. Pattern Recognition, 2022, 122: 108213.

[28] QIU S M, YE J F, ZHAO J C, et al. Video anomaly detection guided by clustering learning[J]. Pattern Recognition, 2024, 153: 110550.

[29] YANG Z Y, RADKE R J. Context-aware video anomaly detection in long-term datasets[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, USA: IEEE, 2024: 4002-4011.

(编辑:孙亚彤)