

弱监督视频异常检测：不确定性引导加持双路径门控机制

张毅安¹⁾ 张 菁^{1),2)} 卓 力^{1),2)}

¹⁾(北京工业大学集成电路学院 北京 100124)

²⁾(北京工业大学计算智能与智能系统北京市重点实验室 北京 100124)

摘 要 弱监督视频异常检测仅依赖视频级别的标注即可实现异常事件的识别与定位,在真实常见的诸多应用中展现出更高的实用价值。尽管相关研究已经取得了一定进展,但在实际处理异常事件时,仍需面临语义多样性、模糊性及视频级标签噪声等带来的严峻挑战。当前主流方法大多基于多实例学习范式,但其性能常受限于以下关键瓶颈:首先,常规方法依赖标准动作识别特征,难以捕捉复杂、上下文相关的异常行为所需的高级语义信息;其次,经典的Top- k 实例选择策略容易过度关注最显著的异常片段,导致模型对持续时间短、视觉特征不突出的弱异常,或同一视频中多个离散异常事件的学习不够充分;最后,视频级标签存在固有噪声,异常视频中往往包含大量正常片段,极易误导模型的优化过程。上述因素共同作用,为模型学习鲁棒的异常判别特征带来了显著的不确定性。为增强模型对不确定性因素影响下异常事件的关注度,本文提出一种不确定性引导加持双路径门控机制的弱监督视频异常检测方法。首先,利用对比语言-图像预训练模型提取富含高级语义信息的视频片段级特征,并通过引入可学习的文本提示与视觉特征进行自适应融合,增强模型对特定异常事件的语义理解能力。在此基础上,进一步引入一种不确定性感知学习范式,利用专门设计的不确定性预测器量化模型对每一视频片段预测结果的置信度。同时,构建双路径门控选择器机制:软路径用于生成融合全局上下文的视频级表示,而硬路径能自适应地确定候选关键片段的数量,使模型动态聚焦于与异常事件最相关的视频片段,有效抑制无关信息或背景噪声的干扰。最后,为提升模型在噪声标签下的鲁棒性与泛化能力,引入不确定性加权的二元交叉熵损失,利用预测的置信度动态调整每个样本的贡献,通过有效削弱噪声样本的干扰,将训练重点引导至高置信度样本。其中,主损失与类鲁棒时间特征幅值损失、门控正则化损失协同优化,以增强特征空间的判别性和模型的泛化能力。在UCF-Crime、XD-Violence和ShanghaiTech三个公开基准数据集上的实验验证表明,所提方法的检测精度分别达到87.98%、83.4%和98.5%,验证了其在真实世界复杂场景下进行视频异常检测的有效性。

关键词 弱监督;视频异常检测;对比语言-图像预训练;不确定性;双路径门控机制

中图法分类号 TP391 DOI号 10.11897/SP.J.1016.2026.01716

Weakly-Supervised Video Anomaly Detection with Uncertainty-Guided Dual-Path Gating Mechanism

ZHANG Yi-An¹⁾ ZHANG Jing^{1),2)} ZHUO Li^{1),2)}

¹⁾(School of Integrated Circuits, Beijing University of Technology, Beijing 100124)

²⁾(Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, Beijing 100124)

Abstract Weakly supervised video anomaly detection (WS-VAD) enables the identification and localization of anomalous events using only video-level labels, demonstrating significant practical value in various real-world applications. Although progress has been made in related research, practical anomaly detection still faces substantial challenges arising from semantic diversity,

收稿日期:2025-07-14;在线发布日期:2026-03-26。本课题得到国家自然科学基金(62471013,61971016)资助。张毅安,硕士研究生,主要研究领域为视频异常检测。E-mail: zhangyian@emails.bjut.edu.cn。张菁(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为图像/视频处理、深度学习、计算机视觉。E-mail: zhj@bjut.edu.cn。卓力,博士,教授,主要研究领域为图像/视频处理、深度学习、计算机视觉。

ambiguity, and noise in video level labels. Current mainstream methods are predominantly based on the multiple instance learning (MIL) paradigm, yet their performance is often constrained by several critical bottlenecks. First, conventional methods rely on features derived from standard action recognition, which lack the high-level semantic information necessary for capturing complex, context-dependent anomalies. Second, the classic Top- k instance selection strategy tends to overemphasize the most salient anomalous segments, leading to insufficient learning of weak anomalies with short duration or inconspicuous visual features, as well as multiple discrete anomalous events within a single video. Finally, inherent noise in video-level labels where anomalous videos frequently contain numerous normal segments, can easily misguide the model optimization process. Collectively, these factors introduce significant uncertainty in learning robust anomaly discrimination features. To enhance model focus on anomalous events under such uncertainty, a weakly-supervised video anomaly detection method with uncertainty-guided dual-path gating mechanism is proposed in this study. Initially, a contrastive language-image pre-training (CLIP) model is employed to extract snippet-level features rich in high-level semantic information. Learnable text prompts are introduced to adaptively fuse with visual features, thereby enhancing the model's semantic understanding of specific anomalous events. Building on this, an uncertainty-aware learning paradigm is further incorporated, utilizing a dedicated uncertainty predictor to quantify the model's confidence in its predictions for each snippet. Concurrently, a dual-path gating selector mechanism is constructed: the soft path generates a video-level representation that integrates global contextual information, while the hard path adaptively determines the number of candidate key snippets. This enables the model to dynamically focus on the snippets most relevant to anomalous events, effectively suppressing irrelevant information or noisy interference in this process. Finally, to improve model robustness and generalization under noisy labels, an uncertainty weighted binary cross-entropy (BCE) loss is introduced, which dynamically adjusts each sample's contribution based on predicted confidence, effectively mitigating noise interference by prioritizing training on high-confidence samples. The primary loss operates synergistically with a class-robust temporal magnitude loss and a gating regularization loss to enhance feature space discriminability and model generalization. Experimental validation on three public benchmarks (UCF-Crime, XD-Violence, and Shanghai Tech) demonstrates detection accuracies of 87.98%, 83.4%, and 98.5%, respectively, confirming the effectiveness of the proposed method for video anomaly detection in complex real-world scenarios.

Keywords weakly-supervised; video anomaly detection; contrastive language-image pre-training; uncertainty; dual-path gating mechanism

1 引言

视频异常检测 (Video Anomaly Detection, VAD) 旨在自动识别和定位视频序列中偏离常规模式的事件。随着视频数据的爆炸式增长, 高效准确的 VAD 对于公共安全、应急响应及内容审查等应用领域具有至关重要的实际价值^[1, 2]。

基于深度学习 (deep learning) 的视频异常检测

是目前的主要研究技术^[3]。根据训练数据标注粒度的不同, 可分为全监督 (full-supervised)、弱监督 (weakly-supervised)、半监督 (semi-supervised) 和无监督 (unsupervised) 等机器学习范式。早期数据规模较小时, 普遍采用全监督学习范式, 需要精细到每一帧的“异常/正常”标签来训练模型, 如 Hinami 等^[4]联合检测并描述预先定义的异常行为, 但是在数据量规模逐步增长的背景下, 人工标注的高昂成本已难以实施。半监督范式下的视频异常检测方法

在训练阶段仅使用正常视频数据建模,测试阶段将偏离正常模式的样本判别为异常。例如,Ruff等^[5]通过神经网络学习一个紧凑的特征空间,使正常样本具有低熵表示,而异常样本则表现为高熵,从而实现有效区分。无监督范式不依赖任何标签信息,在训练阶段通过对视频样本进行特征聚类或表示学习来建模正常模式。例如,Liu等^[6]通过训练模型预测视频的下一帧,将误差较高的片段判定为异常。尽管无监督和半监督可以有效节省数据标注成本,但由于缺乏对异常事件本身的任何先验知识,加之真实世界中“正常”行为复杂多样,这类方法易产生较高的误报率,因而性能往往受限^[1]。相对而言,弱监督机器学习范式下的视频异常检测(Weakly Supervised Video Anomaly Detection, WS-VAD)近年来得到广大学者的关注和研究^[1, 7, 8],其特点在于仅需视频级别的标签进行训练,相对于前三种范式可以在精度和代价之间取得适中的性能。此外,WS-VAD在训练中引入粗粒度的异常视频样本,可能更有机会学习到与异常事件相关且更具判别性的特征,从而在检测性能和标注效率之间达到更为理想的平衡。

WS-VAD常采用多实例学习(Multiple Instance Learning, MIL)方式,首先将每个视频分为固定长度的片段,然后每个视频被抽象为一个“包(bag)”,视频中分割出的短片段(snippets)或帧则被视为包中的“实例(instances)”,其中正常视频被视为仅包含正常实例的负包,而异常视频被视为包含至少一个异常实例的正包。例如,Sultani等^[1]基于MIL排序框架,通过排序损失(ranking loss)训练模型,以最大化异常视频中得分最高的实例(最可能是异常的实例)与正常视频中得分最高的实例(最像异常的正常实例,即难负例)之间的得分间隔。考虑到正包中可能存在多个异常实例,RTFM^[7]采用Top- k 机制筛选特征幅度最大的 k 个实例,以增强正负包之间的特征区分能力。然而,该方法聚焦于固定数量的显著片段,难以充分捕捉视觉特征中那些微妙的弱异常模式。在视频异常检测任务中,一个比较常见棘手的问题是正常片段在数量上往往占绝对主导,导致严重的类别不平衡。此外,部分正常片段中存在一些边界模糊、难以明确归类的事件,容易引发模型的假阳性预测。这类问题被定义为不确定性(uncertainty),即模型对自身预测结果缺乏信心的量化表现。尽管Zhou等^[9]提出的不确定性调节的双存储单元(Dual Memory Units with Uncertainty

Regulation, UR-DMU)引入了不确定性学习机制,但其重点在于构建整体正常模式的空间表示,而非对每个片段预测置信度的细粒度评估。因此,在面对片段级别的模糊事件或噪声干扰时,现有模型仍倾向于做出过于自信的非黑即白判断,无法在训练过程中动态识别并处理这些具有不确定性的样本。由此可见,显式建模预测不确定性是一条更为合理的解决路径。通过引入不确定性引导的学习范式,可以让模型在面对模糊或含噪声的实例时表达“不确定”,并据此来动态调整不同样本在训练中的权重,从而将学习重点聚焦于信息量更高、判断更可靠的片段。这一策略有助于缓解标签噪声带来的干扰,进而学习到更鲁棒的异常判别特征^[7, 10]。

基于以上动机,本文将在WS-VAD框架下,首先通过不确定性引导弱监督视频异常检测。具体思路为,由于模型难以从较弱异常片段中学习到表征力强的特征,本质是无法有效处理预测过程中的不确定性。为此,本文将引入一个显式不确定性预测模块,让模型具备自我认知能力,即为每个片段的异常预测量化一个置信度分数。当模型面对一个模糊、难以判断的片段时,能够输出一个高不确定性值,而不是被迫做出一个自信但可能错误的分类。然而,仅仅识别出不确定性还不够,如何高效地利用所有片段信息来做出最终判断同样关键。前述Top- k 选择机制过于关注局部最显著的片段,简单的加权平均又会稀释关键异常信号。为此,本文提出了一种双路径门控机制(Dual-Path Gating Mechanism, DPGM):软路径加权聚合所有片段信息,以捕捉全局上下文,形成稳健的视频级判断;硬路径则聚焦于少数最关键的片段,以确保模型能学到最具判别力的局部特征。最后,将不确定性与选择性信息整合到一个不确定性加权的二值交叉熵(Binary Cross-Entropy, BCE)损失函数中,动态地指导模型的优化方向,将训练重点放在那些高置信度且信息丰富的样本上,从而全面提升模型的学习效率和泛化能力,帮助更有效地检测视频异常事件。

本文的主要贡献总结如下:

(1) 设计了一个专门的不确定性预测模块,以显式地量化模型对每个视频片段的预测置信度。与常规将不确定性用于构建正态潜空间不同,本方法将其作为动态校准因子,贯穿于模型训练与推理过程,直接指导实例选择与损失加权,从而赋予模型对模糊片段与噪声标签的内在鲁棒性。

(2) 提出了一种双路径门控机制,以克服现有

方法基于静态 Top- k 选择方法所固有的偏差。该机制通过软路径对全局片段进行加权平均, 聚合整体信息以形成稳健的视频级表示, 服务于分类决策; 同时, 通过硬路径执行 Top- k 选择, 聚焦于最具判别力的局部片段进行细粒度表示学习。两条路径相辅相成, 共同增强模型的判别能力。

(3) 引入一种基于不确定性加权的 BCE 损失, 降低高不确定性样本的影响, 使训练更专注于高置信度、信息丰富的样本, 有效缓解弱监督下的标签噪声问题, 提升模型泛化性与检测性能。

2 相关工作

2.1 弱监督视频异常检测

弱监督视频异常检测 (WS-VAD) 主流方法大多构建在 MIL 的框架之上^[1, 3, 11]。然而, 基于 MIL 的方法普遍面临一个核心挑战, 即实例选择偏差: 模型倾向于仅关注视频中最显著的异常片段, 这导致对持续时间短、视觉特征不明显 (即弱异常) 或多发的离散异常事件学习不足^[7]。针对该缺陷, 一些工作在实例选择和损失函数设计上进行了诸多改进。例如, 以鲁棒时间特征幅度 (Robust Temporal Feature Magnitude, RTFM)^[7] 为代表的方法采用基于固定 Top- k 的选择机制, 其核心思想是依据特征向量的模长选取最显著的片段, 并从理论角度分析了 k 值与类别分离度的关系。这类方法在一定程度上扩展了模型的关注范围, 但其本质仍是对少数最显著异常片段的持续优化。因此, 对于未被选入 Top- k 的、表现较为微妙的异常片段, 其特征表示往往难以得到充分学习, 从而限制了模型在复杂场景下的识别能力。相比之下, 本文将不确定性预测作为核心驱动力, 以引导门控自适应 Top- k 机制。该机制不依赖固定 k 值, 而是根据门控时序统计量与不确定性校准结果动态确定候选规模, 并在训练与推理阶段均采用不确定性信息进行片段加权与阈值调制, 从而在多段稀疏异常及边界模糊场景下实现更稳健的片段选择策略。

为了更直接地处理弱异常和多样化异常的学习问题, 研究者们探索了其他策略。例如, 何克磊等^[12]在 MIL 框架中引入原型学习层的思想, 通过基于原型度量的算法实现实例特征到包特征的映射。尽管这些方法在一定程度上缓解了对弱异常的忽视问题, 但如何定义和学习具有足够泛化能力的原型, 以及如何有效结合全局和局部信息仍是挑战。

基于上述问题, 本文通过显式建模和利用预测不确定性, 并结合双路径门控选择机制, 来更有效地提升对所有 (包括弱) 异常样本的关注度以及在噪声标签下的学习鲁棒性。

2.2 数据不确定性学习

近年来, 深度模型量化和利用预测不确定性受到了越来越多的关注。在人脸识别领域, Chang 等^[13]提出了一种数据不确定性学习方案, 展示了通过显式建模特征的不确定性, 可以提高模型在噪声数据下的识别鲁棒性, 其思想对于处理视觉数据中的不确定性具有借鉴意义。在时序动作定位任务中, Lee 等^[14]探索了不确定性建模的应用, 将视频中的背景帧视作相对于动作帧的分布外 (Out-Of-Distribution, OOD) 样本, 并提出通过一个不确定性模型来学习这种区分。一个奠基性的进展来自 Kendall 等^[15]在语义分割和深度估计等计算机视觉任务中, 开创性地展示了模型如何学习并预测偶然不确定性, 并将此不确定性整合到损失函数中。

尽管在视频异常检测领域对不确定性的显式建模和利用尚属较新方向, 但也已有一些代表性工作开始探索其潜力。前述 UR-DMU^[9]在其正常数据不确定性学习 (Normal Data Uncertainty Learning, NUL) 方案中, 核心思想是通过双记忆库存储原型, 并以高斯正态先验学习正常潜空间, 从而精细刻画正常数据的分布及其不确定性, 在测试阶段使用均值进行推理。然而, 该方法对不确定性的建模主要服务于构建鲁棒的正常模式表示, 并未直接用于校准片段级的预测置信度。与此不同, 本方法不引入记忆库, 而是以双门控和困难样本聚焦为主体框架, 并将不确定性作为贯穿训练与推理过程的统一校准量, 直接参与实例选择与阈值调制。这一设计不仅增强了对模糊与噪声样本的鲁棒性, 也避免了额外的记忆维护负担与潜在的模板漂移问题。

基于上述视觉任务中利用不确定性提升模型鲁棒性和判别能力的实践启发, 以及 UR-DMU^[9]在 VAD 领域对不确定性应用的具体探索, 本文专门设计了一个不确定性预测模块, 使其能够与主干网络一同训练, 并为每个实例的预测结果估计不确定性。与已有工作^[7]将不确定性主要用于选择性采样或伪标签的提纯不同, 本文选择了一条更直接的路径。此外, 本文引入 Kendall 等^[15]通过不确定性自动调整损失贡献的核心思想, 将预测的不确定性信息直接用于加权主要的分类损失函数。

2.3 选择性门控机制

为了动态地聚焦于信息量最丰富的片段,门控选择器作为一种显式的注意力机制被提出^[16]。它通过辅助网络多层感知机(MultiLayer Perceptron, MLP)为每个视频片段生成一个介于0到1之间的门控值,用于加权对应的片段特征,从而增强关键信息并抑制无关噪声。在Sultani等^[1]的经典工作中,对关键片段的选择是隐式的,即高异常分数的实例被自然地视为关键实例,选择过程是分类任务的副产品。显式门控的优势在于它将“选择”与“分类”解耦,使得模型可以直接学习“哪些片段是重要的”,这不仅增强了模型的可解释性,也提供了更灵活的控制。为了引导门控机制学到有意义的模式,研究者引入了正则化损失,Sultani等^[1]开创性地应用了两种约束:稀疏损失(sparsity loss)和平滑损失(smoothness loss),在视频异常检测任务上AUC从

74.44%提升至75.41%。

然而,现有的门控机制通常采用单一的注意力模式,这在处理复杂视频场景时可能受限。因此,更优的门控机制应该能够自适应地结合全局上下文信息与局部关键特征。基于此思考,本文设计了一种双路径门控选择器,通过软硬结合的协同组合:软路径通过加权聚合包含全局上下文信息的视频级表示,帮助做出更鲁棒的整体判别;硬路径基于门控值的Top-k选择策略,聚焦于少数最具判别力的局部信息,之后通过正则化损失来引导学习更优的模式。

3 方 法

本文通过施加不确定引导和双路径门控机制,提出弱监督视频异常检测方法,如图1所示。主要包括四个模块:

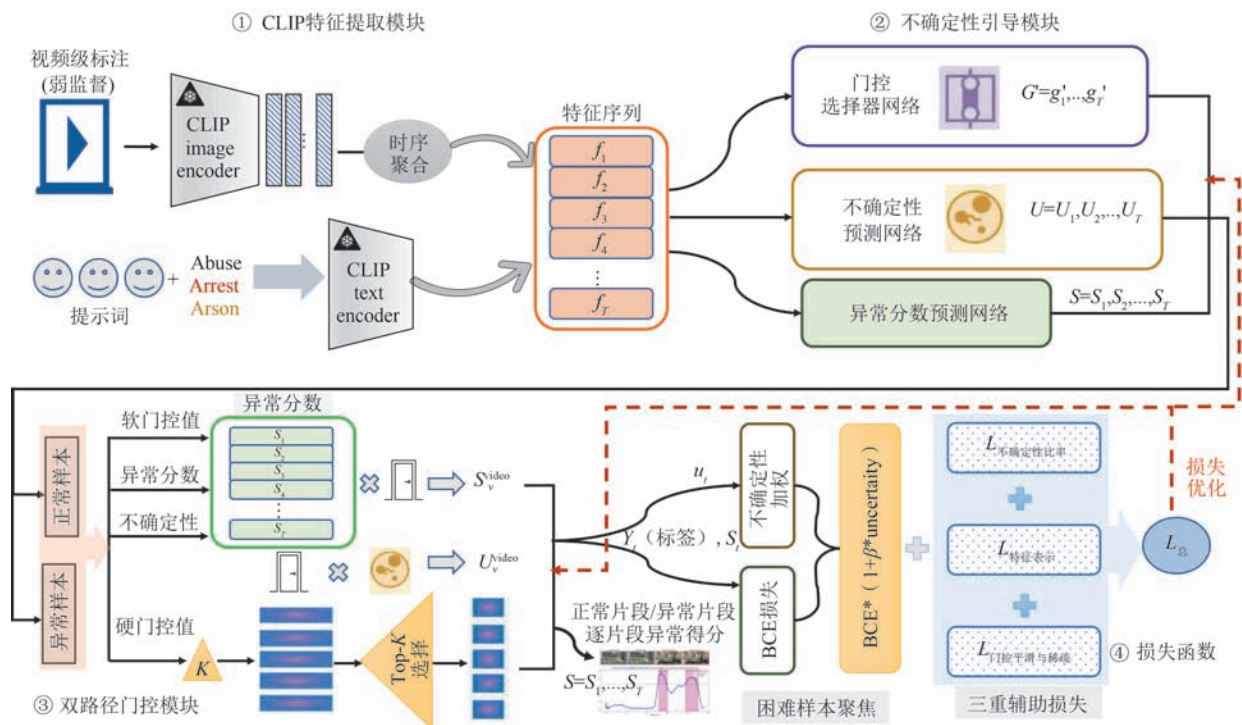


图1 所提WS-VAD总体框图

(1) CLIP 特征提取模块:首先将输入视频分割为固定长度的不重叠片段,并以每个片段的中心帧作为视觉代表;随后利用CLIP视觉编码器提取片段级特征,通过时序聚合模块捕捉片段间的时序依赖;同时,借助CLIP文本编码器生成文本提示特征,与视觉特征相融合,获得富含语义信息的融合特征序列。

(2) 不确定性引导模块:该模块以融合特征为

输入,通过三个并行子网络分别输出片段级门控值(反映片段重要性)、不确定性度量(反映预测置信度)以及初步的异常分数。

(3) 双路径门控模块:基于不确定性引导模块的输出,该模块设计软、硬两条路径。软路径通过加权聚合所有片段以捕捉全局上下文信息,硬路径则聚焦于最具判别力的局部片段,二者协同生成稳健的视频级表示。

(4) 损失函数：该模块将上述各模块的输出与视频级标签相结合，设计包含主损失与辅助损失的多项损失函数。其中，不确定性加权交叉熵损失作为主损失，结合特征幅值损失、不确定性比例损失及门控正则化损失，共同优化模型参数。

3.1 问题定义

在WS-VAD的设定下，本文给定一个包含 N 个视频的训练集：

$$D = \{(V_i, L_i)\} \quad (1)$$

其中， V_i 表第 i 个视频，而 $L_i \in \{0, 1\}$ 是其对应的视频级别标签。如果视频中包含任何异常事件，则 $L_i = 1$ ；反之，如果视频完全由正常事件构成，则 $L_i = 0$ 。每个视频可以被看作一个由 T_i 个时间步组成的序列。

在处理时，通常将其分割成一系列不重叠或轻微重叠的短片段。因此，视频 V_i 可以被表示为一个片段序列 V_i^T ，其中 T_i 是分割后的片段数量。

本文的目标是学习一个深度模型，该模型能够接收一个视频 V 作为输入，并为其所有片段输出一个逐片段的异常分数序列 S^T ，其中 S 表示片段 t 属于异常的可能性。在训练阶段，模型仅能利用视频级别的标签 L 进行优化。在测试阶段，模型需要利用学习到的能力，为给定的测试视频生成逐片段的异常分数，从而不仅能够判断整个视频是否异常，还能在时间上定位异常事件发生的位置。

3.2 CLIP 特征提取

本节从原始视频中提取具有时序上下文和语义信息的片段级特征，为异常检测提供高质量的输入。首先，利用CLIP视觉编码器为每个片段提取高级语义的初始视觉特征；随后，通过一个包含多尺度扩张卷积的模块来处理初始视觉特征序列，以捕捉视频中时序依赖关系；最后，将CLIP文本编码器生成的文本提示特征与时序聚合的视觉特征进行融合，增强特征的整体表征能力。

3.2.1 视觉特征提取

首先，将输入视频分割成 T 个等长且不重叠的片段(snippets)。对于每个片段，选取其中心帧 I_i 为该片段的关键视觉代表，并将其送入预训练的CLIP^[8]视觉编码器，提取得到片段级视觉特征 $f_i \in \mathbb{R}_{vis}$ ：

$$f_i = \text{ViT} - \text{CLIP}(I_i) \quad (2)$$

通过上述过程，原始视频被转换为一个初始的视觉特征序列：

$$F_{vis} = [f_1, f_2, \dots, f_T] \in \mathbb{R}^{T \times d_{vis}} \quad (3)$$

3.2.2 时序特征聚合

为捕捉视频中事件的动态演化和片段间的时序依赖关系，将初始的视觉特征序列 F_{vis} 送入一个时序特征聚合模块^[17]。该模块由多个并行的扩张卷积分支构成，每个分支采用不同的扩张率 $d_i \in \{1, 2, 4, 8\}$ ，以捕捉不同范围的时序依赖关系。随后，将聚合分支的输出特征进行融合，得到增强时序上下文信息的聚合特征序列 F_{agg} ：

$$F_{agg} = F_{vis} * d_i k = \sum_{s+d \cdot t} F_{vis}(s) k(t) \quad (4)$$

其中， k 为卷积核大小。

3.2.3 文本提示的视觉特征融合

为了增强特征的语义指向性并利用CLIP的多模态能力，本模型引入了可选的文本提示融合机制，设计思路借鉴了如Zhou等^[18]利用文本提示增强视频理解的工作。主要步骤如下：

步骤1：构造包含类别名称的提示模板：

$$P_c = [V_1][V_2] \dots [V_M][\text{CLASS}] \quad (5)$$

其中， V 为可学习的上下文向量，CLASS为类别名称占位符，在实际使用时由具体类别标签（如“violence”、“explosion”等）替换。该模板设计延续了Zhou等^[18]提出的连续提示学习机制，不再依赖如“a photo of a [CLASS]”等固定模板，而是通过训练自动优化上下文向量，以提升在不同数据分布下的语义对齐能力。

步骤2：生成文本提升特征

将每个类别对应的提示句 P_c 输入CLIP模型文本编码器(Text Encoder)，对应的文本特征表示 f_{text} ：

$$f_{\text{text}}^c = \text{CLIP}_{\text{text}}(P_c) \in \mathbb{R}^{d_{\text{text}}} \quad (6)$$

这里， f_{text}^c 表示CLIP语义空间中的类别原型向量。在实现过程中，固定文本编码器权重，仅优化上下文向量，以保持语义一致性并增强跨数据集的泛化能力。

步骤3：视觉与文本特征的融合

将时序聚合后的视觉特征与对应的文本提示特征沿特征维度进行拼接，并通过线性映射构建语义引导的融合特征：

$$f_{\text{fused}, t} = \text{Concat}(\alpha \cdot f_{\text{text}}^c, (1 - \alpha) F_{agg, t}) \cdot W \quad (7)$$

其中， α 为平衡视觉与文本模态影响的比例系数。本文主要关注视频片段的视觉异常模式，文本提示仅作为辅助语义引导。通过大量预实验验证，设定 $\alpha = 0.2$ ，可在充分利用语义信息的同时，有效避免

文本特征对模型的主导。

步骤4: 构建融合特征序列

得到的融合特征序列 F_{fuse} , 该序列同时具备动态与语义引导信息, 后续所有模块的训练与评估均基于该融合特征进行:

$$F_{\text{fuse}} = [f_{\text{fuse}, 1}, f_{\text{fuse}, 2}, \dots, f_{\text{fuse}, T}] \in \mathbb{R}^{T \times d_{\text{fuse}}} \quad (8)$$

3.3 不确定性引导机制

接下来, 对视频帧进行不确定预测, 以引导模型解决弱监督学习中的标签噪声和异常事件的模糊性问题。首先利用两层 1D 卷积提取时序敏感的高层特征 $f_{\text{conv}, t}$:

$$f_{\text{conv}, t} = \text{Conv 1D}_2(\text{ReLU}(\text{Conv 1D}_1(F_{\text{fuse}}))) \quad (9)$$

随后将其输入到三个紧密相关的子模块, 如图 2 所示: 门控选择器用于动态聚焦关键片段; 不确定性预测器用于量化模型预测的置信度; 异常分数预测器用于判断片段的异常程度。

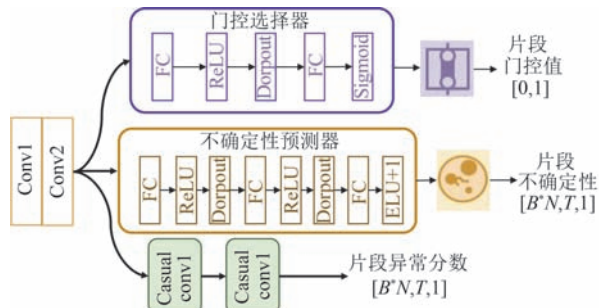


图2 不确定性引导模块

3.3.1 门控选择器

为动态识别并加权对异常判断最重要的视频片段, 本文设计了一个由 MLP 构成的门控选择器。以 $f_{\text{conv}, t}$ 作为输入, 输出一个标量门控 g_t , 其值域被限制在 $[0, 1]$ 内:

$$g_t = \text{Sigmoid}\left(\text{FC}_2\left(\text{Dropout}\left(\text{ReLU}\left(\text{FC}_1(f_{\text{conv}, t})\right)\right)\right)\right) \quad (10)$$

其中, FC_1 和 FC_2 代表两个全连接层, ReLU 为修正线性单元激活函数, 被应用在 FC_1 后, Dropout 将 ReLU 层输出随机置为 0, Sigmoid 将 g_t 制在 $[0, 1]$ 内, 从而被应用为门控选择器软路径中的开关信号。由 g 组合为门控值序列:

$$G = [g_1, \dots, g_T] \in \mathbb{R}^T \quad (11)$$

以定量地表示每个片段的重要性, 并将用于后续的样本选择和损失正则化。

为了确保门控选择器学习到平滑且稀疏的注意

力模式, 避免过拟合或产生无意义的注意力分配, 本文引入了两种正则化损失^[1], 对门控值序列施加平滑约束, 以鼓励门控选择器在时间维度上做出平滑的决策。平滑损失为:

$$L_{\text{smooth}} = \sum_{t=1}^{T-1} (g_{t+1} - g_t)^2 \quad (12)$$

对门控值序列 G 施加稀疏性约束, 以鼓励模型只将少数几个真正关键的片段标记为高度相关, 而不是将注意力平均分配到所有片段上。稀疏损失为:

$$L_{\text{sparse}} = \sum_{t=1}^T |g_t| \quad (13)$$

3.3.2 不确定性预测器

为量化模型对预测的置信度, 本文设计了一个独立的 MLP 网络来预测每个片段的不确定性。以特征 $f_{\text{conv}, t}$ 作为输入, 输出一个非负的标量不确定性度量 U_t :

$$U_t = \text{ELU}\left(\text{FC}_4\left(\text{Dropout}\left(\text{ReLU}\left(\text{FC}_4(f_{\text{conv}, t})\right)\right)\right)\right) + 1.0 \quad (14)$$

值得注意的是, 本文采用 $\text{ELU}+1.0$ 作为输出层的激活函数以确保其值为正。不确定性的具体计算方式在模型训练和推理阶段有所不同, 以平衡计算效率和估计的鲁棒性。为了训练效率, 根据式(14), 不确定性 U_t 通过单次前向传播计算得到。

推理时为了获得更稳定和对模型参数扰动更不敏感的不确定性估计, 本文采用蒙特卡洛 Dropout (MC-Dropout)^[19]。首先, 对同一个输入片段特征 $f_{\text{conv}, t}$, 在保持 dropout 层激活的状态下进行 N 次随机前向传播, N 的取值直接影响不确定性估计的方差精度与计算成本, 随后得到一个不确定性预测集合:

$$\{U_t^{(1)}, U_t^{(2)}, \dots, U_t^{(N)}\}, \quad U_t^{(n)} = f_{\text{unc}}^{(n)}(f_{\text{conv}, t}) \quad (15)$$

其中, $f_{\text{unc}}^{(n)}$ 表示第 n 次随机 dropout 后的网络。

片段不确定性估计值 $U_{t, \text{mc}}$ 通过计算该集合的均值得到:

$$U_{t, \text{mc}} = \frac{1}{N} \sum_{n=1}^N U_t^{(n)} \quad (16)$$

该方法利用测试阶段 Dropout 的随机性来近似模型的后验分布, 从而为不确定性提供更为可靠的量化。尽管训练与推理阶段的计算方式存在差异, 但两者在参数共享与 Dropout 随机性机制上保持一致, 因此输出分布在统计意义上是等价的。多次采样仅用于降低估计方差, 而不会改变已学到的映射

关系。

3.3.3 异常分数预测器

异常评分网络负责为每个视频片段生成一个定量的异常分数, 以特征 $f_{\text{conv},t}$ 作为输入, 采用因果卷积 (Causal Convolution, CC) 作为核心组件:

$$f_{\text{cc},t} = \sum_{w=0}^{W-1} h(w) \cdot x(t-w) + b \quad (17)$$

其中, $f_{\text{cc},t}$ 表示第 t 个片段的卷积输出; W 表示卷积核大小; $h(w)$ 表示第 w 个滤波权重; $x(t-w)$ 表示输入的第 $t-w$ 个片段的特征; b 表示偏置项。该结构确保在时间 t 的输出 $f_{\text{cc},t}$ 仅依赖于当前及过去 $t-w$ 时刻的输入, 以严格遵守时序预测的因果关系。

本文使用由两层连续的因果卷积网络构成的深度建模模块, 并采用 Sigmoid 激活函数进行处理, 得到异常分数序列 $S \in [0, 1]$:

$$S_t = \text{Sigmoid}\left(\text{Conv}_{\text{causal},2}\left(\text{ReLU}\left(\text{Conv}_{\text{causal},1}\left(f_{\text{conv},t}\right)\right)\right)\right) \quad (18)$$

模块为视频中的每个片段输出一个异常分数序列 $S = [S_1, S_2, \dots, S_T]$ 。

3.4 双路径门控机制

为有效聚合各片段信息以生成稳健的视频级表示, 本节提出一种软硬结合的双路径选择机制, 以平衡全局上下文和局部关键信息的重要性。

3.4.1 正常/异常样本分离

根据视频级标签将批次数据在逻辑上分割为正常和异常样本子集。随后, 对每个视频的原始门控值序列 $G_v = [g_{v,1}, g_{v,2}, \dots, g_{v,T}]$ 进行归一化处理, 得到在时间维度和为 1 的权重分布 $g_{v,t}^{\text{norm}}$:

$$g_{v,t}^{\text{norm}} = \frac{g_{v,t}}{\sum_{i=1}^T g_{v,i} + \epsilon_{\text{norm}}} \quad (19)$$

其中, norm 是防零小常数。

3.4.2 软路径: 不确定性引导的门控加权聚合

利用得到的归一化的门控值 $g_{v,t}^{\text{norm}}$ 作为一种软注意力权重, 对视频中所有片段的异常分数 $s_{v,t}$ 和不确定性 $u_{v,t}$ 进行加权聚合。这种聚合方式的优势在于它全面地考虑了视频中所有时间点的信息, 但同时又根据门控选择器学习到的片段重要性以及模型对该片段预测的置信度 (间接通过不确定性加权损失影响门控和分数的学习), 赋予了那些被认为更关键且预测更可靠的片段更大的影响力。

具体而言, 对于每个样本 v , 其视频级别异常分数 S_v^{video} 和不确定性 U_v^{video} 计算如下:

$$S_v^{\text{video}} = \sum_{t=1}^T g_{v,t}^{\text{norm}} \cdot S_{v,t} \quad (20)$$

$$U_v^{\text{video}} = \sum_{t=1}^T g_{v,t}^{\text{norm}} \cdot U_{v,t} \quad (21)$$

这些视频级别的聚合值将用于后续的分类损失和不确定性相关损失的计算。

3.4.3 硬路径: 基于门控的自适应 Top- k 选择

传统的固定数量 Top- k 方法在选择代表性异常实例时, 难以适应不同视频中异常事件在时序长度、发生频率以及显著性程度上的动态变化。为解决此问题, 本文提出了一种基于门控值的自适应 Top- k 选择策略。

首先, 根据原始门控值序列 G_v 动态计算候选异常片段的数量 K_{abn} :

$$K_{\text{abn}} = \left\lceil \frac{\sum_{t=1}^T g_{v,t} \cdot \eta}{\epsilon_{\text{gate}}} \right\rceil \quad (22)$$

其中, η 是比例调节因子。较小的 η 值使 K_{abn} 更偏向高置信片段, 从而使输出更为稀疏和精确; 而较大的 η 值则有助于增强弱异常片段的覆盖能力, 提高召回率, 但可能引入一定噪声。因此, η 的取值直接影响模型在不同异常密度场景下的自适应性能。

随后, 通过一个结合了门控值和原始异常分数的综合得分 $S_{v,t}^{\text{comp}}$ 对片段进行排序:

$$S_{v,t}^{\text{comp}} = g_{v,t} \times S_{v,t} \quad (23)$$

最后, 根据此综合得分的排序, 为每个异常视频选择 Top- k_{abn} 个片段的特征作为代表性异常实例, 并根据相应策略选择代表性正常实例。整个动态门控选择与代表性实例识别的流程可以概括如算法 1:

算法 1. 基于门控的 Top- k 特征选择

输入:

- F_{conv} : 输入视频的特征序列
- G : 输入视频的原始门控值序列
- S : 输入视频的异常分数序列
- L : 输入视频的标签
- η : K_{abn} 的比例因子
- ϵ_{gate} : 用于计算 K_{abn} 的小常数

输出:

- $\text{FeatSelect}_{\text{abn}}$: 选择的代表性异常片段特征集
- $\text{FeatSelect}_{\text{nor}}$: 选择的代表性正常片段特征集

流程:

1. 初始化 $\text{FeatSelect}_{\text{abn}} \leftarrow \emptyset, \text{FeatSelect}_{\text{nor}} \leftarrow \emptyset$
2. 对于每个时间 t , 计算综合得分 $S_t^{\text{comp}[t]} = g_t \times S_t$
3. IF $L_{\text{video}} = \text{normal}$ THEN
4. $\text{FeatSelect}_{\text{norm}} = \{f_{\text{conv},i} | \forall t\}$ // 正常视频选择全部特

征作为正常样本

5. RETURN $FeatSelect_{abn}, FeatSelect_{nor}$;
6. END IF
7. // 处理异常视频(正样本包)
8. 使用式(22)计算动态 K_{abn}
9. $K_{abn} = \max(1, \min(k_{abn}, T))$ // 确保 $K_{abn} \in [1, T]$
10. // 选择 Top- K_{abn} 个异常实例
11. for $i = 0$ TO $K_{abn} - do$
12. $idx = Indices_sorted_desc[i]$
13. $f_{conv, idx}$ 加入 $FeatSelect_{abn}$
14. END FOR
15. // 从异常包中选择正常实例
16. $K_{normbag} = DeterKForNorInAbnBag(K_{abn}, T)$
18. $FeatSelect_{nor} = SelectAndAddNorFeatures$
19. RETURN $FeatSelect_{abn}, FeatSelect_{nor}$;

3.5 弱监督视频异常检测

为实现更为准确的弱监督视频异常检测,本节采用端到端优化的复合损失函数,该函数主要由以下两部分构成:(1)困难样本聚焦:不确定性加权的二元交叉熵损失利用预测的置信度来动态调整样本权重,以有效缓解标签噪声的干扰;(2)三重辅助损失:包含用于增强特征判别性的特征幅值损失、用于维持期望不确定性比例的损失,以及用于规范门控选择器行为的平滑与稀疏损失。这些损失从不同维度对模型进行约束,以增强特征空间的判别性和模型的泛化能力。

3.5.1 困难样本聚焦

为有效训练模型以区分正常与异常视频片段,并特别应对在弱监督设定下由标签噪声和数据固有模糊性所带来的挑战,本文采用了不确定性加权的二元交叉熵(BCE)损失作为主要的分类目标。该损失利用由不确定性预测器输出的片段级不确定性度量 u_t 来动态调整每个样本的 BCE 损失权重,定义为:

$$L_{weightedBCE} = \frac{1}{T} \sum_{t=1}^T \left(\frac{1}{U_t + \epsilon_{unc}} \right) \cdot BCE(S_t, l_t), \quad (24)$$

$$BCE(S_t, l_t) = -[l_t \log(S_t) + (1 - l_t) \log(1 - S_t)] \quad (25)$$

其中, l_t 为二元标签 0 或 1, ϵ_{unc} 为防零小常数。

3.5.2 三重辅助损失

为使正常与异常片段特征在幅值上可分,根据 RTFM^[7] 的思想,本文引入异常损失 L_{abn} 和正常损失 L_{nor} :

$$L_{abn} = \sum_{f \in FeatSelect_{abn}} \text{Max} \left(0, m - \|f_{select, v}^{abn}\|_2 \right)^2, \quad (26)$$

$$L_{norm} = \sum_{f \in FeatSelect_{norm}} \|f_{select, v}^{norm}\|_2^2 \quad (27)$$

其中, m 是预设的边际值, $f_{select, v}^{abn}$ 和 $f_{select, v}^{norm}$ 分别为算法 1 选取的代表性异常与正常特征。该损失由超参数 α 控制。

为引导模型在预测正常与异常样本时的平均不确定性之间维持期望的比例关系,借鉴 UR-DMU^[9] 的思想,本文引入 $L_{uncratio}$:

$$L_{uncratio} = \left| \frac{\text{mean}(u_t | l_t = 0, t \in \text{sel})}{\text{mean}(u_t | l_t = 1, t \in \text{sel}) + \epsilon_{ratio}} - K \right| \quad (28)$$

其中, K 是目标比例常数, ϵ_{ratio} 是防零小常数。该损失由超参数 β 控制。

此外,为规范门控选择器的行为,引入平滑损失 L_{smooth} (式 12) 和稀疏损失 L_{sparse} (式 13)。总损失函数 L_{total} 定义为上述各项损失的加权和:

$$L_{total} = L_{weightedBCE} + \alpha(L_{abn} + L_{norm}) + \beta L_{uncratio} + \gamma(L_{smooth} + L_{sparse}) \quad (29)$$

其中, α, β, γ 为平衡各项损失的权重超参数。

4 实 验

4.1 实验设置

4.1.1 数据集

本文实验主要在以下三个被广泛应用于 WS-VAD 研究的公开数据集上进行:

(1) UCF-Crime^[1] 包含从真实世界监控摄像头和在线视频平台(如 YouTube)收集的 1900 个未剪辑长视频,总时长约 128 小时。在弱监督设定下,训练集包含 800 个正常视频和 810 个异常视频,测试集则包含 150 个正常视频和 140 个异常视频。

(2) XD-Violence^[20] 是一个大规模的暴力检测数据集,包含了从电影、网络视频、游戏和监控录像等多种来源收集的 4754 个视频片段,总时长超过 217 小时。它涵盖了 6 种主要的暴力或异常类别。训练集包含 3954 个视频(仅提供视频级标签),测试集包含 800 个视频(提供帧级标注)。

(3) ShanghaiTech Campus^[21] 由从 13 个不同背景场景中捕获的街道监控视频组成。ShanghaiTech 最初被提议作为半监督的 VAD 基准,后来被 Zhong 等人重组为 WS-VAD 基准^[22]。数据集共包含 437 个视频,其中 238 个视频用于训练,199 个视频用于测试。

4.1.2 评估指标

为了全面衡量所提出方法的性能,并与现有工作进行公平比较,本文采用了视频异常检测领域广泛使用的评估指标。

(1) 接受者操作特征 (Receiver Operating Characteristic, ROC) 曲线下面积 (Area Under Curve, AUC): 这是评估二分类模型性能最常用的指标之一。真阳性率 (True Positive Rate, TPR) 和假阳性率 (False Positive Rate, FPR) 为:

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{FP + TN} \quad (30)$$

其中, TP (True Positive) 是真阳例, FN (False Negative) 是假阴例, FP (False Positive) 是假阳例, TN (True Negative) 是真阴例。

ROC 曲线是 FPR 为横轴, TPR 为纵轴绘制的曲线, AUC 是该曲线下的面积:

$$AUC = \int_0^1 TPR(FPR^{-1}(x)) dx \quad (31)$$

该积分值为随机选择一个正样本得分高于一个负样本得分的概率。对于 UCF-Crime 和 ShanghaiTech 数据集, AUC-ROC 是主要评估标准。

(2) 平均精度 (Average Precision, AP): 对于像 XD-Violence 这样可能存在类别不平衡或者更关注正例 (异常事件) 检出情况的数据集, AP 是一个更为合适的评估指标^[21], 为精确率-召回率 (Precision-Recall Curve, PRC) 曲线下的面积。精确率 (Precision) 和召回率 (Recall):

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN} \quad (32)$$

在实践中, 通常通过对不同召回率水平上的精确率进行加权平均来计算:

$$AP = \sum_{k=1}^n (R_k - R_{k-1}) P_k \quad (33)$$

4.1.3 实施细节

对于所有数据集, 本文遵循了 CLIP-TSA^[23] 中关于视频片段划分的通用原则。具体而言, 每个视频首先被分割成一系列不重叠的、32 个片段, 每个片段包含 16 帧, 并从中提取代表性帧。使用预训练的 CLIP ViT-B/16 模型^[8] 提取 512 维视觉特征。数据集的标准训练集和测试集划分被严格遵守。本文采用 Adam 优化器^[24] 对模型参数进行优化, 权重衰减固定设置为 0.005, 学习率为 0.001, 批处理大小为 16。所有实验均在一台 3090Ti GPU 24G 上进行。为保证对比的公平性, 本文所选的对比方法均

在统一且可复现的实验配置下进行评测, 且其核心技术路线相似, 从而确保了比较结果的客观性与参考价值。

4.2 与先进方法的准确率对比

为了验证提出方法的有效性, 在三个广泛使用公开数据集上与当前先进的 WS-VAD 方法进行了全面的性能比较, 如表 1 所示。在 UCF-Crime 数据集上, 本文方法取得了 87.98% 的 AUC。这一结果均优于多个近期 SOTA 方法表现, 比如一些基于 CLIP 的方法如 IFS-VAD^[25]。相较于使用传统 I3D 特征的 SOTA 方法, 如 UMIL^[26] 和 MGFN^[27], 本文的方法也展现了竞争力。这初步表明, 本文的不确定性引导和门控选择机制能够在强大的 CLIP 特征基础上, 有效地处理 UCF-Crime 数据集中复杂的异常事件。

在极具挑战性的 XD-Violence 数据集上, 使用 AP 作为评估指标, 本文的方法达到了 83.40%, 与同样采用 CLIP 特征的 TPWNG^[28] 性能非常接近, 并优于一些早期的或基于传统特征的方法。这表明, 结合不确定性引导的门控选择机制与 CLIP 特征, 对于提升模型在多样化和嘈杂场景下的异常检测能力是有益的。

在 ShanghaiTech 数据集上, 本文方法取得了 98.50% 的 AUC, 这一性能超过了大多数对比方法, 包括同样基于 CLIP 特征的 CLIP-TSA^[23] 和 IFS-VAD^[25]。相较于使用 I3D 特征的方法, 如 UMIL^[26], 本文的优势更为明显。ShanghaiTech 数据集场景相对固定, 异常事件与正常背景的区别可能更为依赖于对微妙变化的捕捉和对正常模式的精确建模。本文不确定性引导机制有助于模型更好地从包含大量正常片段中学习“常态”, 并在测试时对偏离常态的事件给出更准确的判断, 同时门控机制帮助聚焦于真正的异常时刻。

4.3 主观结果展示

为了更直观地评估本文所提方法的有效性, 并深入理解其在实际场景中的行为模式, 本节对模型的检测结果进行主观定性分析并从以下三个角度展示: (1) 通过正例分析验证其成功识别并精准定位真实异常事件的能力; (2) 通过边界情况或检测结果不佳的负例分析模型在复杂或模糊场景下的潜在局限性; (3) 利用 t-SNE 技术^[37] 对模型处理前后的特征分布进行降维可视化, 以直观地证明本文方法在学习更具判别性的、可分离的特征表示方面的有效性。

表 1 与现有方法的比较

方法	发表源	特征	UCF-Crime	XD-Violence	Shanghai Tech
			AUC(%)	AP(%)	AUC(%)
RTFM ^[7]	ICCV'21	I3D	84.30	77.81	/
Watanabe et al. ^[29]	ICIP'22	I3D	84.91	/	95.27
MSL ^[10]	AAAI'22	I3D	85.30	/	/
DAR ^[30]	TIFS'23	I3D	85.18	/	/
BE-WS-VAD ^[31]	TCSVT'23	I3D	84.05	78.08	95.00
NG-MIL ^[32]	WACV'23	I3D	85.63	78.51	85.63
CLAV ^[33]	CVPR'23	I3D	86.10	81.30	/
Zhang et al. ^[34]	CVPR'23	I3D	86.22	78.74	/
UMIL ^[26]	CVPR'23	I3D	86.75	/	96.78
MGFN ^[27]	AAAI'23	I3D	86.98	79.19	/
Fan et al. ^[35]	TCSVT'24	I3D	86.19	83.60	/
OE-CTST ^[36]	WACV'24	Video-Swin	86.99	81.78	/
CLIP-TSA ^[23]	ICIP'23	CLIP	87.58	82.19	98.32
TPWNG(大模型方法) ^[28]	CVPR'24	CLIP	87.79	83.68	/
IFS-VAD ^[25]	TCSVT'25	CLIP	86.57	83.14	97.95
所提方法		CLIP	87.98	83.40	98.50

注：以上对比方法数据结果来自原文，“/”表示代码未开源因而无法复现，粗体代表最优值。

4.3.1 正例分析

如图 3 所示,对于包含真实异常事件的视频,模型通常能够在其预测的异常分数曲线上,于真实异常发生的区间内产生显著的峰值或高分段。可以看到,图中绿框表示模型输出的异常分数(蓝色曲线)在描述异常发生的时间段内显著升高,并与真实的异常区间(粉色高亮区域)高度吻合。表明模型成功地检测并定位了异常事件。在上述示例图的非异常区域,模型的异常分数均能保持在相对较低的水平,低于设定的阈值。这表明模型对于这些视频中的正常片段具有较好的区分能力,能够有效地抑制将正常活动误判为异常的情况。

在 UCF-Crime 数据集上,图 3 右列第 1 幅图像显示了一个相对静止的室外场景,可能包含一些背景物体如卡车。“Abnormal Frame 2”和“Abnormal Frame 3”则显示了卡车启动或异常移动的情况,可能涉及盗窃或未经授权的车辆使用。该视频的异常分数在早期(约第 3 帧)迅速从低位(约 0.4)上升至 0.8 以上,并在此后持续缓慢爬升至接近 0.95 的水平,直到视频末尾才急剧下降。模型预测的异常区间(粉色高亮区域)同样覆盖了视频的绝大部分。以及在 XD-Violence 数据集上,例如图 3 右列第 2 幅图像,“Normal Frame 0”和“Normal Frame 1”似乎是办公室内相对正常的场景。而“Abnormal Frame 26”和“Abnormal Frame 27”则展示了一个会议室或办

公室内的激烈争执或冲突场面,人物情绪和动作都显得异常。模型同样在该视频的两个主要区域(约 25-30 帧和约 38-45 帧)给出了较高的异常分数,并标记为异常。

4.3.2 负例分析

本文方法也存在“检测结果不佳”或“边界情况”,如图 3 右列第 3 幅图像红框所示,在该视频的第 83-87 帧左右确实产生了一个异常分数的小峰值,并且根据阈值将其标记为一个预测的异常区间(粉色高亮区域)。这个峰值的绝对高度(约 0.25-0.28)远低于左侧第一个异常事件的峰值(约 0.9)。“Abnormal Frame 29”所代表的事件确实是一个真实的、需要被检测到的异常,其对应的异常分数峰值相对较低,这可能表明模型认为此事件的“异常程度”不如第一个事件那么显著。如果这个事件的真实重要性很高,但模型给出的分数仅略高于阈值,那么在阈值选择略微保守的情况下,这个事件就可能被漏检(成为假负例)。这提示模型需要关注不同强度和视觉显著性异常的响应一致性和敏感性。

4.3.3 热力图分析

为了更直观地评估本文提出的不确定性引导的弱监督视频异常检测方法在学习判别性特征方面的有效性,本文采用了 t -分布随机邻域嵌入(t -SNE)技术^[37]对特征分布进行降维可视化。本文对比了原始 CLIP 提取的片段特征与经过本文模型处理后得

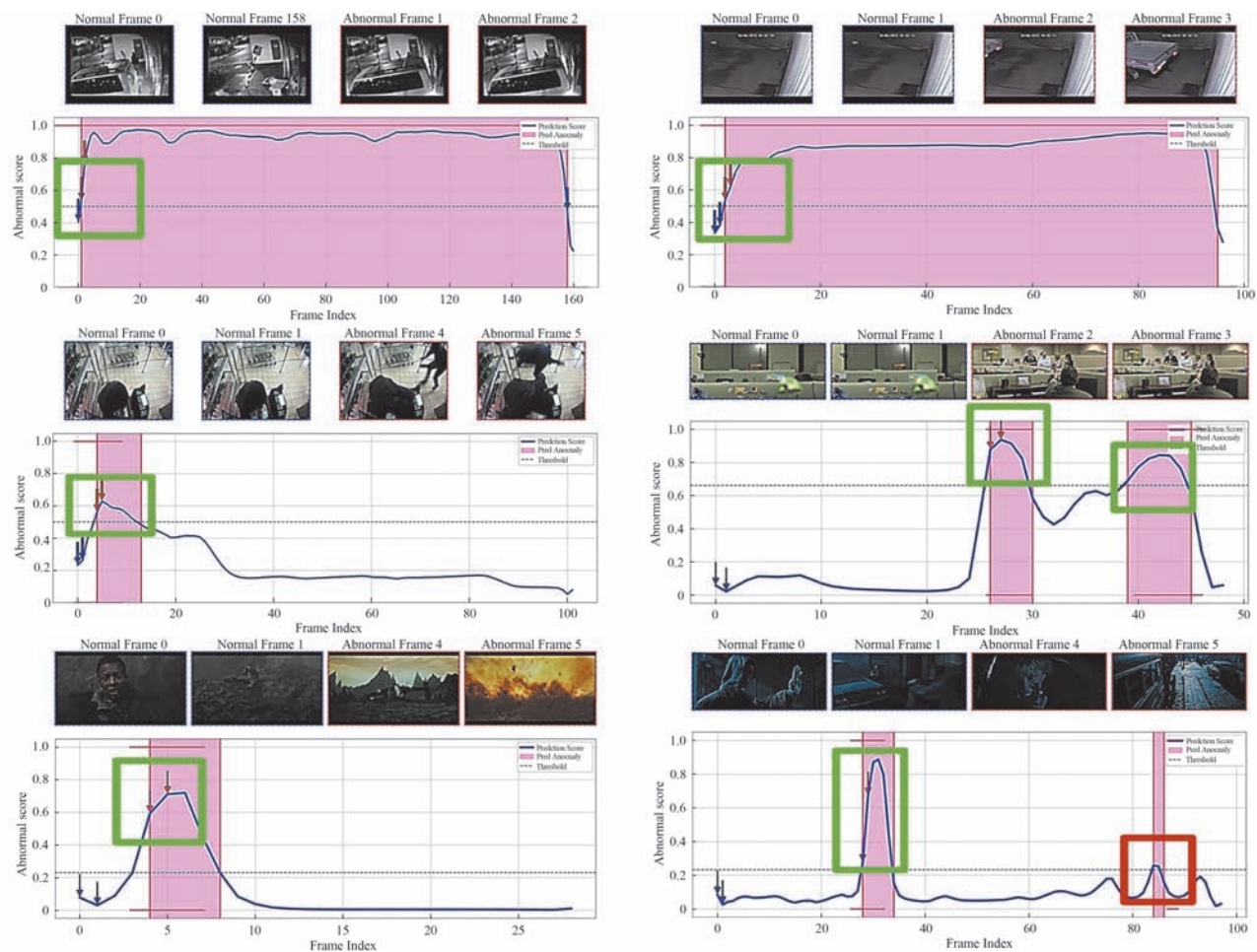


图3 主观结果展示

到的在UCF-Crime, XD-Violence测试集上的分布情况(ShanghaiTech数据集测试集中仅有异常一种类别,热力图无法呈现检测效果),如图4所示。

在原始特征图中正常(绿色点)和异常(橙色点)样本的分布呈现出显著的混合和重叠。尽管可以看出某些局部区域可能略有聚集的趋势,但整体上两类样本之间缺乏清晰的、全局性的可分界限。这表明原始的CLIP特征虽然包含了丰富的语义信息,但在直接应用于区分复杂的、多样化的异常视频片段时,其固有的特征空间结构可能不足以直接实现理想的分离。

在经过本文提出的模型框架进行特征聚合与优化后,正常与异常片段的特征分布展现出了显著改善的可分离性。可以看到,正常样本(绿色点)倾向于在特征空间的某些区域形成相对集中的簇,而异常样本(橙色点)则更多地分布在与之不同的区域,或者形成了自身独特的聚集模式。尽管两类样本之间仍然存在一定的交错区域(这反映了WS-VAD任务的固有

难度和异常的多样性),但相较于原始特征,模型处理后的特征在整体上呈现出更清晰的类间分离趋势和更紧凑的类内聚集特性。这直观地证明了本文的模型通过时序建模、不确定性引导的动态门控选择以及损失优化,能够有效地学习到更具判别性的特征表示,从而提升了区分正常与异常片段的能力。

4.4 消融研究

为了系统性地评估本文提出的不确定性引导模块、双门控选择机制以及困难样本聚焦策略的有效性,并量化各部分对模型最终性能的贡献,本节进行了一系列消融实验,如表2中所示。

本文通过结合不确定性引导模块、双门控选择机制和困难样本聚焦策略,使得模型性能达到最佳(表2第1行),在UCF-Crime、XD-Violence和ShanghaiTech数据集上分别达到87.98%AUC、83.40%AP和98.50%AUC。表2第2行去除了困难样本聚焦策略,在三个数据集上性能分别下降了0.55%、0.50%和0.07%,这是因为不确定性可以

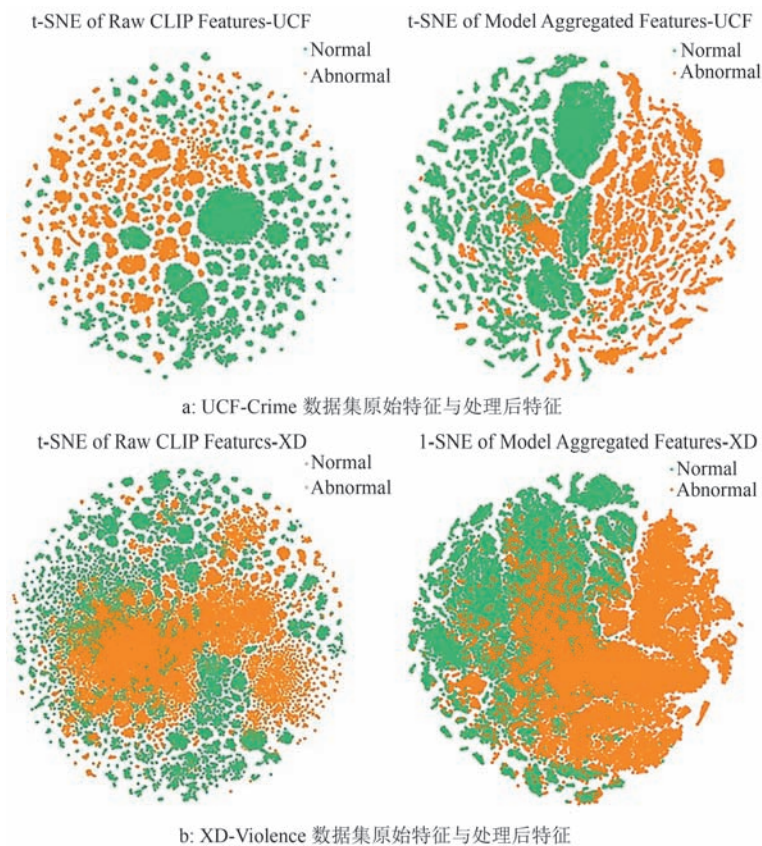


图4 热力图对比

表2 消融结果

不确定性引导	双路径门控	困难样本聚焦	UCF数据集	XD数据集	SH数据集
			AUC(%)	AP(%)	AUC(%)
✓	✓	✓	87.98	83.40	98.50
✓	✓		87.43 ↓ 0.55	82.90 ↓ 0.50	98.43 ↓ 0.07
✓		✓	86.90 ↓ 1.08	83.02 ↓ 0.38	98.37 ↓ 0.13
	✓	✓	86.78 ↓ 1.20	82.88 ↓ 0.52	98.36 ↓ 0.14
✓			86.10 ↓ 1.88	82.27 ↓ 1.13	98.33 ↓ 0.17

调节损失权重,有效抑制噪声样本干扰,对提升模型鲁棒性至关重要。第3行去除了双路径门控模块,在三个数据集上性能分别下降1.08%、0.38%和0.13%,这是因为选择机制在提炼判别性信息方面的关键作用。第4行去除了不确定性引导模块,性能分别下降了1.20%、0.52%和0.14%,体现了不确定性引导模块对模型性能的稳定增益。最后,在本文的基础模型上加入不确定性引导模块(包含不确定性预测器),性能则下降了1.88%、1.13%和0.17%,综上,消融实验结果清晰地表明,不确定性引导模块、双门控选择机制以及困难样本聚焦策略三者相辅相成,共同为模型性能提升做出了不可或缺贡献。

4.5 参数分析

4.5.1 随机前向传播次数 N

为在推理效率与估计精度之间取得平衡,视频异常检测领域通常将 N 设置为10或20,如表3所示,不同 N 取值的实验结果表明,当 $N=5$ 时,模型在UCF-Crime、XD-Violence和ShanghaiTech数据集上的性能分别为87.65%AUC、82.96%AP和98.42%AUC,此时方差尚不稳定,估计精度较低;当 $N=10$ 时,三项性能分别提升0.28%、0.32%和0.06%至87.93%、83.28%和98.48%,结果趋于稳定;而当 $N=20$ 时,模型达到最佳精度,进一步增加0.05%、0.12%和0.02%,达到87.98%、83.40%和98.50%,获得了最佳精度。

表3 随机前向传播次数 N			
N	UCF数据集	XD数据集	SH数据集
	AUC(%)	AP(%)	AUC(%)
5	87.65	82.96	98.42
10	87.93	83.28	98.48
20	87.98	83.40	98.50

4.5.2 比例调节因子 η 参数分析

为探究比例调节因子 η 在不同数据集上的最优取值，实验针对UCF-Crime、XD-Violence和ShanghaiTech三个数据集的不同异常分布特征，分别设置了不同的 η 搜索区间，并以0.05为步长进行网格搜索，以确定各数据集对应的最优 η 值，结果如表4所示。实验结果表明， η 的最优取值与数据集的异常分布特征高度相关。具体而言，UCF-Crime数据集在 $\eta=0.45$ 时达到最优性能，AUC最高为87.98%。该数据集的异常片段通常较长且分布稳定，较小的 η 有助于维持较高的选择稀疏性。XD-Violence数据集在 $\eta=0.55$ 时表现最佳，AP值为83.40%。该数据集包含多段弱异常事件，较大的 η 有助于扩展K值范围，从而提升对多段异常的召回能力。ShanghaiTech数据集在 $\eta=0.35$ 时取得最优结果，AUC为98.50%。由于该数据集视频较短且异常集中，较小的 η 能有效抑制噪声干扰。

表4 比例调节因子 η 参数分析			
数据集	搜索范围	最佳 η	对应指标
UCF-Crime	0.30-0.70	0.45	87.98AUC
XD-Violence	0.40-0.80	0.55	83.40AP
ShanghaiTech	0.25-0.6	0.35	98.50AUC

4.5.3 不确定性加权系数参数分析

为了探究不确定性加权损失在总损失函数中的最佳权重，并验证模型性能对该关键超参数的敏感性，本节对不确定性加权系数进行了一系列参数分析实验。从图5可以看出，模型性能对不确定性加权系数的选择表现出一定的敏感性。当系数值过小（如0.01或0.05）时，不确定性引导机制对模型训练的正则化或优化作用可能不够充分，导致性能未达到最佳。在这组实验中，当不确定性加权系数为0.1时，模型在三个数据集上均取得了最佳的性能。这表明在该系数值下，不确定性相关损失项与主要分类损失和其他辅助损失之间达到了较好的平衡，使得模型能够最有效地利用不确定性信息来提升鲁棒性和判别能力。

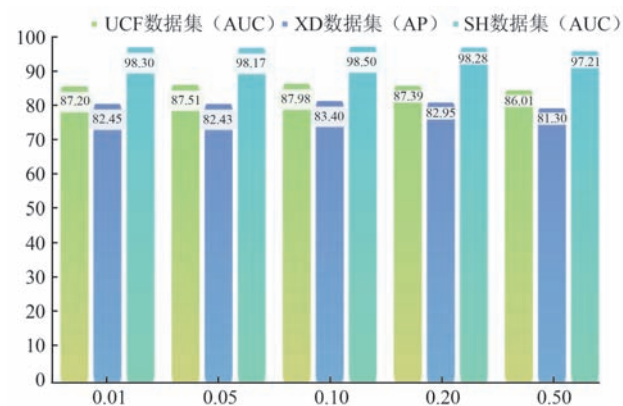


图5 不确定性加权系数对实验结果的影响

4.6 计算复杂度分析

为了评估本文所提模型在实际应用中的部署潜力与效率，本节对模型复杂度进行了分析。如图6所示，测试了在三个基准数据集上参数量 (Parameters) 和推理时间 (Inference Time)，并与两种代表性的SOTA方法RTFM^[7]和UR-DMU^[9]进行了比较。本文提出的模型在性能、模型大小和推理效率之间取得了良好的平衡。相较于一些SOTA方法，它在显著减少参数量或大幅提升推理速度的同时，保持了良好的异常检测能力，这使其在实际应用部署中更具潜力。

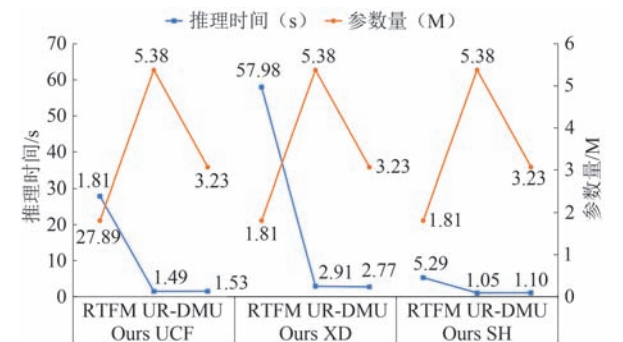


图6 不同方法的性能和效率

5 结 论

本文针对弱监督视频异常检测中存在的标签噪声、实例模糊性以及微妙异常关注不足等核心挑战，提出了一种不确定性引导的双路径门控机制的弱监督视频异常检测方法。首先设计了一个专门的不确定性预测模块，用于量化模型对每个视频片段预测的置信度，使模型能够主动识别并应对弱监督设定下的模糊性与噪声。其次提出了一种双路径门

控选择机制;通过软路径的加权聚合来捕捉全局上下文信息,形成稳健的视频级表示;同时通过硬路径的自适应 Top- k 选择来聚焦最具判别力的局部片段特征。最后引入了一种不确定性加权的二元交叉熵损失,利用预测的置信度来动态调整每个样本在训练中的贡献,有效缓解了标签噪声问题。

本文通过在 UCF-Crime, XD-Violence 和 ShanghaiTech 数据集上进行实验,验证了所提出方法的有效性。实验结果表明,与当前主流 WS-VAD 方法相比,本文在上述三个数据集上分别达到 87.98%、83.40%、98.50% 的精度,在各项评估指标上均取得了具有竞争力的性能。特别地,消融研究,也清晰地表明不确定性引导机制、动态门控选择以及困难样本聚焦策略,即不确定性加权 BCE 损失,对于提升模型整体性能的关键作用。主观结果分析和特征可分性可视化进一步揭示了本文模型在准确识别异常事件、有效利用不确定性信息以及学习更判别性特征表示方面的能力。

但是本文仍存在一些值得未来进一步探索和改进的方向。例如,如何更精细地对不同类型的不确定性(偶然不确定性与认知不确定性)进行区分和建模,并将其更深入地整合到模型的各个组件中,可能会带来性能的进一步提升。

参 考 文 献

- [1] Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6479-6488
- [2] Sun C, Wu Y W, Jia Y D. Video anomaly detection with context modeling and reasoning. Chinese Journal of Computers, 2024, 47(10): 2368-2386 (in Chinese)
(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)
- [3] Pang G S, Shen C, Cao L, Van Den Hengel A. Deep learning for anomaly detection: a review. ACM Computing Surveys, 2021, 54(2): 38
- [4] Hinami R, Mei T, Satoh S. Joint detection and recounting of abnormal events by learning deep generic knowledge//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 3648-3657
- [5] Ruff L, Vandermeulen R A, Görnitz N, Binder A, Müller E, Müller K R, Kloft M. Deep semi-supervised anomaly detection.//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020
- [6] Liu W, Luo W, Lian D, Gao S. Future frame prediction for anomaly detection—A new baseline//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 3421-3430
- [7] Tian Y, Pang G, Chen Y, Singh R, Verjans J W, Carneiro G. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 4975-4986
- [8] Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 8748-8763
- [9] Zhou H, Yu J, Yang W. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023, 37(3): 3769-3777
- [10] Li S, Liu F, Jiao L. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2022: 1395-1403
- [11] Wang F, Jiang Y, Zhang R, Wei A, Xie J, Pang X. A survey of deep anomaly detection in multivariate time series: taxonomy, applications, and directions. Sensors, 2025, 25(1): 190
- [12] He K L, Shi Y H, Gao Y, et al. A multiple instance convolutional neural network based on prototype learning. Chinese Journal of Computers, 2017, 40(6): 1182-1193 (in Chinese)
(何克磊, 史颖欢, 高阳, 等. 一种基于原型学习的多示例卷积神经网络. 计算机学报, 2017, 40(6): 1182-1193)
- [13] Chang J, Lan Z, Cheng C, Wei Y. Data uncertainty learning in face recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 5710-5719
- [14] Lee P, Wang J, Lu Y, Byun H. Weakly-supervised temporal action localization by uncertainty modeling//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2021, 35(3): 1854-1862
- [15] Kendall A, Badrinarayanan V, Cipolla R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 7482-7491
- [16] Zhang J, Shi X, Xie J, Ma H, King I, Yeung D Y. GAAN: Gated attention networks for learning on large and spatiotemporal graphs//Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence. Monterey, USA, 2018: 339-349
- [17] Zheng W J, Fu S C, Chen J Z, et al. Few-shot graph anomaly detection in extremely weakly supervised scenarios. Chinese Journal of Computers, 2025, 48(4) (in Chinese)
(郑文捷, 傅司超, 陈嘉真, 等. 极其弱监督场景下的小样本图异常检测. 计算机学报, 2025, 48(4): 928-948)
- [18] ZhouKY, YangJK, Loy C C, LiuZW. Learning to prompt for

- vision-language models. *International Journal of Computer Vision*, 2022, 130(9): 2337-2348
- [19] Gal Y, Ghahramani Z. Dropout as a bayesian approximation: representing model uncertainty in deep learning//*Proceedings of the 33rd International Conference on Machine Learning*. New York, USA, 2016: 1050-1059
- [20] Wu P, Liu J, Shi Y, Sun Y, Shao F, Wu Z, Yang Z. Not only look, but also listen: learning multimodal violence detection under weak supervision//*Proceedings of the European Conference on Computer Vision*. Glasgow, UK, 2020: 322-339
- [21] Liu W, Luo W, Lian D, Gao S. Future frame prediction for anomaly detection - a new baseline//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA, 2018: 6536-6545
- [22] Zhong J X, Li N, Kong W, Liu S, Li T H, Li G. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, USA, 2019: 1237-1246
- [23] Joo H K, et al. Clip-tsa: clip-assisted temporal self-attention for weakly-supervised video anomaly detection//*Proceedings of the 2023 IEEE International Conference on Image Processing*. Kuala Lumpur, Malaysia, 2023: 2005-2009
- [24] Kingma D P, Ba J. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014
- [25] Zhong Y, Zhu R, Yan G, Gan P, Shen X, Zhu D. Inter-clip feature similarity based weakly supervised video anomaly detection via multi-scale temporal MLP. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025, 35(2): 1961-1970
- [26] Lv H, Yue Z, Sun Q, Luo B, Cui Z, Zhang H. Unbiased multiple instance learning for weakly supervised video anomaly detection//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, Canada, 2023: 8022-8031
- [27] Chen Y, Liu Z, Zhang B, Fok W, Qi X, Wu Y C. MGFN: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection//*Proceedings of the AAAI Conference on Artificial Intelligence*. Washington, USA, 2023, 37(1): 387-395
- [28] Yang Z, Liu J, Wu P. Text prompt with normality guidance for weakly supervised video anomaly detection//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, USA, 2024: 18899-18908
- [29] Watanabe Y, Okabe M, Harada Y, Kashima N. Real-world video anomaly detection by extracting salient features in videos. *IEEE Access*, 2022, 10: 125052-125060
- [30] Liu T, Zhang C, Lam K M, Kong J. Decouple and resolve: Transformer-based models for online anomaly detection from weakly labeled videos. *IEEE Transactions on Information Forensics and Security*, 2023, 18: 15-28
- [31] Yang Z, Guo J, Wang D, Huang X, Bao Y, Wang Y. Towards video anomaly detection in the real world: A binarization embedded weakly-supervised network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, 34(5): 4135-4140
- [32] Park S, Kim H, Kim M, Kim D, Sohn K. Normality-guided multiple instance learning for weakly supervised video anomaly detection//*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. Waikoloa, USA, 2023: 2664-2673
- [33] Cho M, Kim M, Hwang S, Park C, Lee K, Lee S. Look around for anomalies: weakly-supervised anomaly detection via context-motion relational learning//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, Canada, 2023: 12137-12146
- [34] Zhang C, et al. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, Canada, 2023: 16271-16280
- [35] Fan Y, Yu Y, Lu W, Han Y. Weakly-supervised video anomaly detection with snippet anomalous attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, 34(7): 5480-5492
- [36] Majhi S, Dai R, Kong Q, Garattoni L, Francesca G, Brémond F. OE-CTST: Outlier-embedded cross temporal scale transformer for weakly-supervised video anomaly detection//*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. Waikoloa, USA, 2024: 8559-8568
- [37] Maaten L V D, Hinton G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008, 9(11): 2579-2605



ZHANG Yi-An, M. S. candidate.

Her research interest is video anomaly detection.

ZHANG Jing, Ph. D., professor. Her current research interests include image/video processing, computer vision and deep learning.

ZHUO Li, Ph. D., professor. Her current research interests include image/video processing, computer vision and deep learning.

Background

As a fundamental task in computer vision, Video Anomaly Detection (VAD) aims to automatically identify and localize events that deviate from normal patterns in video sequences. With the explosive growth of video data, VAD has become increasingly valuable in application such as public safety, emergency response, and content moderation. Early approaches to VAD predominantly relied on fully supervised learning, which, despite its effectiveness, demands costly fine-grained frame-level annotations. Unsupervised and semi-supervised paradigms reduce this annotation burden but often suffer from high false-positive rates due to the lack of prior knowledge about anomalous events and the inherent complexity and diversity of what constitutes “normal” behavior in real-world scenarios. In contrast, Weakly-Supervised Video Anomaly Detection (WS-VAD), which requires only video-level labels during training, achieves a more favorable trade-off between detection performance and annotation cost, and has thus emerged as a prominent direction in recent years.

Most existing WS-VAD methods are built upon the Multiple Instance Learning (MIL) framework, where videos are treated as “bags” and video snippets as “instances”. For example, a seminal work introduced a ranking loss to maximize the score margin between the highest-scoring instances in positive (anomalous) and negative (normal) bags. Subsequent studies have adopted Top- k selection strategies to jointly optimize multiple highly

salient anomalous instances. While these methods have advanced the field, they still face several critical challenges. First, conventional instance selection strategies tend to focus predominantly on the most salient anomalous segments, often resulting in insufficient learning of weak anomalies, those with short durations or subtle features, or failing to capture multiple discrete anomalous events within a single video. Moreover, when faced with ambiguous or difficult-to-define events, existing models frequently produce overconfident but erroneous predictions, lacking the ability to quantify predictive uncertainty, a limitation that hinders both reliability and interpretability. Additionally, since anomalous videos inherently contain numerous normal segments, such label noise can significantly bias the optimization process and degrade model performance.

To address these issues, this paper proposes a weakly-supervised video anomaly detection method guided by an uncertainty-aware dual-path gating mechanism. The proposed approach explicitly estimates the prediction uncertainty for each video segment and leverages this information to steer the dual-path gating module and refine the loss function. Extensive experiments demonstrate the effectiveness of the proposed method, highlighting its potential to improve both detection accuracy and robustness in real-world applications.

This research was supported by the National Natural Science Foundation of China under Grants 62471013, 61971016.