doi:10.3969/j.issn.1001-4616.2026.05.010

基于文本特征对齐的多模态视频异常检测

朱文佳[1],伊　君[2],侯思佳[3],张　婷[3],余　烨[3]

(1.安徽中安数科信息有限公司,安徽 合肥 230001)
(2.黄冈师范学院计算机与人工智能学院,湖北 黄冈 438000)
(3.合肥工业大学计算机与信息学院,安徽 合肥 230601)

[**摘要**] 弱监督多模态视频异常检测旨在仅利用视频级标签,通过融合视觉、音频等多模态数据,实现对视频中异常事件片段的精确识别与定位,是计算机视觉领域一项极具挑战性的任务. 当前方法面临两大瓶颈:一是跨模态时序相关性建模不足,限制了异常定位的精度;二是文本提示模块多依赖固定语义,缺乏动态上下文感知,导致语义引导能力有限. 为解决上述问题,提出一种基于文本特征对齐的多模态视频异常检测方法(multimodal video anomaly detection based on textual feature alignment,TFA-MVAD). 该方法首先设计了一个跨模态特征融合模块,通过动态注意力机制捕捉音视频片段间的时序依赖关系,实现多模态互补信息的深度整合,从而提升关键片段的定位能力. 其次,引入特征对齐模块,将预训练文本语义特征与融合后的音视频特征进行对齐,并借助对比学习策略增强跨模态语义一致性. 实验结果表明,该方法在 XD-Violence 数据集上表现优异,与现有最佳方法相比,其平均精度(Average Precision,AP)提升了 3.21%,验证了该方法的有效性.

[**关键词**] 多模态融合,文本信息,特征对齐,时序相关性
[**中图分类号**]TP391.41　[**文献标志码**]A　[**文章编号**]1001-4616(2026)05-0093-08

Multimodal Video Anomaly Detection Based on Textual Feature Alignment

Zhu Wenjia[1],Yi Jun[2],Hou Sijia[3],Zhang Ting[3],Yu Ye[3]

(1.Anhui Zhong'an Digital Technology Information Co.,Ltd.,Hefei 230001,China)
(2.School of Computer Science and Artificial Intelligence,Huanggang Normal University,Huanggang 438000,China)
(3.School of Computer and Information,Hefei University of Technology,Hefei 230601,China)

Abstract:Weakly supervised multi-modal video anomaly detection aims to accurately identify and localize anomalous event segments in videos by effectively integrating multi-modal data such as visual and audio information using only video-level annotations. It is a highly challenging task in the field of computer vision. On one hand,insufficient modeling of cross-modal temporal correlations limits the precision of anomaly event localization. On the other hand,the text prompt module relies on fixed semantic information and lacks dynamic contextual associations,which constrains the effectiveness of semantic guidance. To address these challenges,this paper proposes a multi-modal video anomaly detection method based on textual feature alignment(TFA-MVAD). The method first designs a cross-modal feature fusion module,which models the temporal correlations between audio-visual segments through a dynamic attention mechanism,achieving deep integration of complementary multi-modal information and significantly improving the localization capability of key segments. Secondly,a feature alignment module is introduced to align pre-trained textual semantic features with the fused audio-visual features,optimizing cross-modal semantic consistency through a contrastive learning strategy. Experimental results demonstrate that the proposed method achieves significant performance improvements on the XD-Violence dataset. Compared to the state-of-the-art methods,it increases the Average Precision(AP)by 3.21%,validating the effectiveness of the approach.

Key words:multimodal fusion,textual information,feature alignment,temporal correlation

视频异常检测[1](video anomaly detection,VAD)旨在通过提取和分析视频特征,实现对异常事件片段(如人员斗殴、儿童摔倒等)的精确识别与定位. 该技术依托对视频时空特征的解析,可实现实时预警,在

收稿日期:2025-06-09.
基金项目:国家自然科学基金面上资助项目(62372153)、安徽省自然科学基金面上资助项目(2308085 MF216).
通讯作者:伊君,博士,讲师,研究方向:计算机视觉、多媒体技术与三维重建,E-mail:junyi@ hgnu.edu.cn

智能安防与公共安全管理等领域具有重要的应用价值.

由于缺少帧级标注,现有 VAD 方法大多采用弱监督学习方式,即仅利用视频级别的标签信息完成异常检测. 根据所采用的数据模态,现有弱监督 VAD 方法可分为基于单模态与基于多模态两类. 基于单模态的方法[2-3]主要依赖视觉信息,使用时序建模技术(如 3D-CNN、LSTM 等)提取视频的时空特征. 然而,这类方法在复杂场景下存在一定局限性,例如对光照变化[4]、遮挡等环境干扰的鲁棒性较差,难以有效识别视觉特征不明显的异常行为. 基于多模态的方法[5-6]则通过引入音频等辅助模态信息,对视频模态形成有效补充. 音频信号可提供打斗时的尖叫声、玻璃破碎声等声学上下文特征,与视觉信息形成互补,不仅增强了异常事件的特征表征能力,也显著提升了模型在低光照、部分遮挡等挑战性场景下的检测性能.

尽管基于多模态的视频异常检测方法已取得显著进展,但仍面临以下挑战:(1)现有研究大多未能充分考虑不同模态片段之间的时序相关性,导致模型难以准确判定异常事件发生的具体时空位置;(2)尽管已有尝试引入文本信息构建提示模块以提供额外语义约束,但目前多依赖固定文本作为提示,缺乏与上下文相关的动态短语或句子,因语义信息不足而限制了检测性能的进一步提升.

针对上述问题,本文提出了一种基于文本特征对齐的多模态视频异常检测方法(multimodal video anomaly detection based on textual feature alignment,TFA-MVAD). 该方法的主要创新点包括:(1)设计了一个跨模态特征融合模块,用于捕捉视频及音频序列中的关键异常片段,提取并融合多模态互补信息,有效学习不同帧之间的依赖关系;(2)设计了一个跨模态特征对齐模块,将文本特征与融合后的音视频特征进行对齐,不仅增强了特征的判别能力,也保证了多模态特征在语义层面的一致性,从而提升了异常分类的准确性.

1 相关工作

Wu 等[7]于 2020 年发布了首个多模态视频异常检测数据集 XD-Violence,该数据集同时提供了视频与音频模态信息,推动了多模态视频异常检测研究的发展. 文献[7]还提出了一种并行的三分支网络 HL-Net(holistic and localized network),通过捕捉视频片段的全局与局部关系来整合多模态特征. 实验结果表明,多模态数据能显著提升检测性能,但其所采用的早期特征融合策略在模态交互深度方面存在明显不足.

为克服上述局限,后续研究主要从以下几个方向展开深入探索:

(1)跨模态融合优化

Pang 等[8]引入注意力机制构建动态交互网络,实现了模态间的自适应特征融合,但在噪声抑制和特征对齐方面表现不足. 针对该问题,Pu 等[9]提出了跨模态感知局部唤醒模块,通过音频引导的全局注意力机制来增强视觉特征表示,并结合自适应局部先验校准视频片段权重,有效降低了时间维度上的信息冗余. Sun 等[10]则提出多尺度瓶颈变换器融合模块,通过渐进式瓶颈令牌压缩策略实现高效的多模态特征融合,并采用加权机制突出关键特征,但在处理模态异构性方面仍存在改进空间.

(2)噪声鲁棒性与上下文建模

Yu 等[11]开发的 MACIL-SD 框架通过对比学习和自蒸馏技术的协同作用,显著增强了模型对模态噪声的鲁棒性. Pu 等[12]在上下文建模方面取得进展,其提出的时间上下文聚合模块和提示增强学习机制有效强化了模型的时序建模能力与语义判别性能.

(3)特征表示优化

在特征表示方面,Zhou 等[13]将特征学习引入双曲空间,利用双曲几何的层次特性突破欧氏空间的表示瓶颈,但面临双曲运算带来的计算复杂度挑战. Ghadiya 等[14]进一步提出跨模态融合适配器与双曲洛伦兹图注意力机制,通过几何空间建模实现模态贡献的动态平衡和特征分离增强,为模态不平衡问题提供了新颖的解决方案,但其计算效率仍需优化.

2 模型总体框架

如图 1 所示,TFA-MVAD 模型主要由跨模态特征融合模块(cross-modal feature fusion module,CM-FFM)和特征对齐模块(feature alignment module,FAM)组成. 首先将输入视频序列划分为若干长度为 1 s 的非重叠片段,每个片段包含 16 帧,将其输入至预训练的 I3D 模型中提取视觉特征,表示为 $\boldsymbol{V}_i \in \mathbf{R}^{T\times d_v}$,其中 T 为片段

总数,d_v 为视觉特征维度. 音频处理方面,将音频切分为长度 960 ms 的重叠片段,每个片段与视频片段的结尾对齐,并利用预训练的 VGG 网络提取音频特征 $\boldsymbol{A}_i \in \mathbf{R}^{T\times d_a}$,$d_a$ 为音频特征维度. 随后,将视觉特征 $\boldsymbol{V}_i$ 与音频特征 $\boldsymbol{A}_i$ 输入到跨模态特征融合模块,输出音视频融合特征 X. 该融合特征再经由全连接层(fully connected layer,FC)进一步处理,得到最终的音视频特征表示 $\boldsymbol{X}^c$.

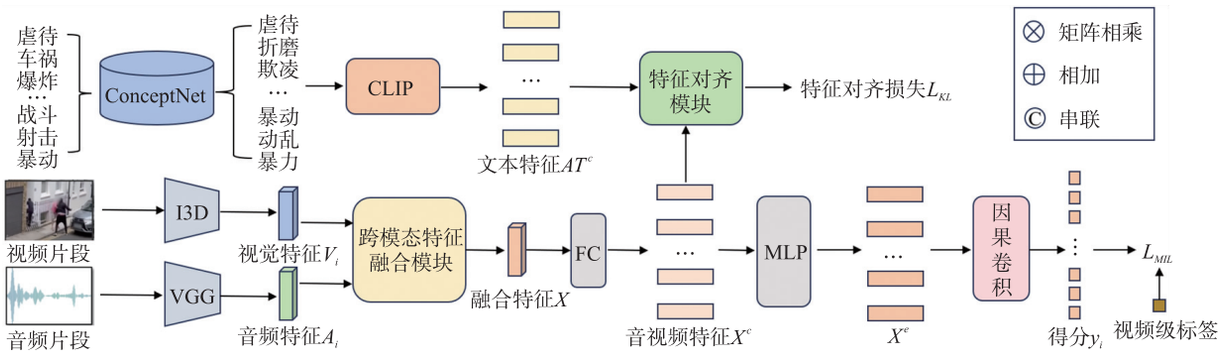


图 1　TFA-MVAD 模型结构图

Fig. 1　Structure of TFA-MVAD

文本特征的提取首先基于字典构建方法[15],根据训练集中所有可能的异常事件类型,构建包含 N 个异常事件类型和 1 个正常事件的字典,总项数为 N+1.

以 XD-Violence 数据集为例,其包含 6 种异常事件类型. 针对某一事件类型,如“打架”,将其输入至 ConceptNet[16]中,获取语义相近的文本描述(如“斗殴”、“厮打”等)及其相关性分数. 选取其中相关分数最高的 M 个文本作为该事件类型的键,并以相应分数作为值,构建该类型的字典项. 字典构建完成后,使用预训练的 CLIP 模型提取文本特征. 对于某一异常事件类型 c,从其字典中选择一组键 $\{\boldsymbol{k}_i^c\}_{i=1}^{M}$ 作为上下文提示符,并将 c 与所有键文本一同输入 CLIP 模型,得到一组文本特征向量 $\{\boldsymbol{K}_i^c\}_{i=1}^{M+1}$. 最终,通过对所有特征向量取平均,得到基于文本信息的提示特征,计算公式如下:

$$\boldsymbol{T}^c=\frac{1}{M+1}\sum_{i=1}^{M+1}\boldsymbol{K}_i^c. \tag{1}$$

文本提示信息融合了多种事件类型之间的语义关联,有助于增强音视频特征的语义表示. 将文本特征 $\boldsymbol{T}^c$ 与音视频特征 $\boldsymbol{X}^c$ 输入至特征对齐模块,在该模块中,两类特征通过一组共享的语义中心进行自适应聚合,并计算同一中心内音视频特征与文本特征之间的局部跨模态相似性,进而通过特征对齐损失函数施加约束. 为获得高层语义表示,同时将音视频特征 $\boldsymbol{X}^c$ 输入到 MLP 中进行特征降维,得到输出特征 $\boldsymbol{X}^e$,将其输入到因果卷积层,以预测每个片段的异常得分 y_i,计算公式如下:

$$y_i=\sigma(f_t(\boldsymbol{X}^e)), \tag{2}$$

其中,$f_t(\cdot)$为核大小为 Δt 的因果卷积层,$\sigma(\cdot)$为 sigmoid 函数,y_i 表示第 i 个片段的异常得分.

获取所有片段的异常得分后,采用 top-k 策略筛选出得分最高的 k 个关键片段,作为异常候选片段. 这些片段的异常得分与视频级标签共同用于构建多实例学习(multi-instance learning,MIL)损失函数.

2.1　跨模态特征融合模块

首先,将原始视频特征输入两个卷积核大小为 3 的二维卷积层(Conv2D-C1 和 Conv2D-C2)进行特征变换,分别生成特征 $\boldsymbol{V}_{c1}$ 和 $\boldsymbol{V}_{c2}$. 其中,$\boldsymbol{V}_{c1}$ 主要用于捕获视频内部的时空关联性,为后续视频相似性矩阵构建提供基础;$\boldsymbol{V}_{c2}$ 则侧重于挖掘与音频信号的相关性,以支持跨模态交互. 同时将 $\boldsymbol{V}_i$ 进行全局平均池化(global average pooling,GAP),提取视频的全局特征 $\boldsymbol{V}_g$.

音频信息作为一维时序信号,通常表示为时间上的振幅或频谱序列,因此采用一维卷积处理其特征. 将得到的音频特征 $\boldsymbol{A}_i$ 输入至两个卷积核大小为 3 的一维卷积层(Conv1D-C1 和 Conv1D-C2),得到处理后的特征 $\boldsymbol{A}_{c1}$ 和 $\boldsymbol{A}_{c2}$,并对 $\boldsymbol{A}_i$ 进行 GAP 操作,得到音频全局特征 $\boldsymbol{A}_g$.

进一步,分别利用音频特征 $\boldsymbol{A}_i$ 引导视频时间相关性的建模,同时利用视觉特征 $\boldsymbol{V}_i$ 引导音频时间相关性的建模. 视频相似性矩阵 $\boldsymbol{S}_{v\leftarrow a}(\boldsymbol{V}_{c1},\boldsymbol{A}_{c2})\in\mathbf{R}^{T\times T}$ 和音频相似性矩阵 $\boldsymbol{S}_{a\leftarrow v}(\boldsymbol{A}_{c1},\boldsymbol{V}_{c2})\in\mathbf{R}^{T\times T}$ 计算公式如下:

$$\boldsymbol{S}_{v\leftarrow a}(\boldsymbol{V}_{c1},\boldsymbol{A}_{c2})=\mathrm{softmax}\left(\frac{\boldsymbol{V}_{c1}\boldsymbol{A}_{c2}^{\mathrm{T}}}{\sqrt{d_1}}\right),\tag{3}$$

$$\boldsymbol{S}_{a\leftarrow v}(\boldsymbol{A}_{c1},\boldsymbol{V}_{c2})=\mathrm{softmax}\left(\frac{\boldsymbol{A}_{c1}\boldsymbol{V}_{c2}^{\mathrm{T}}}{\sqrt{d_2}}\right),\tag{4}$$

其中,d_1 和 d_2 分别表示 $\boldsymbol{A}_{c2}$和 $\boldsymbol{V}_{c2}$的维度.

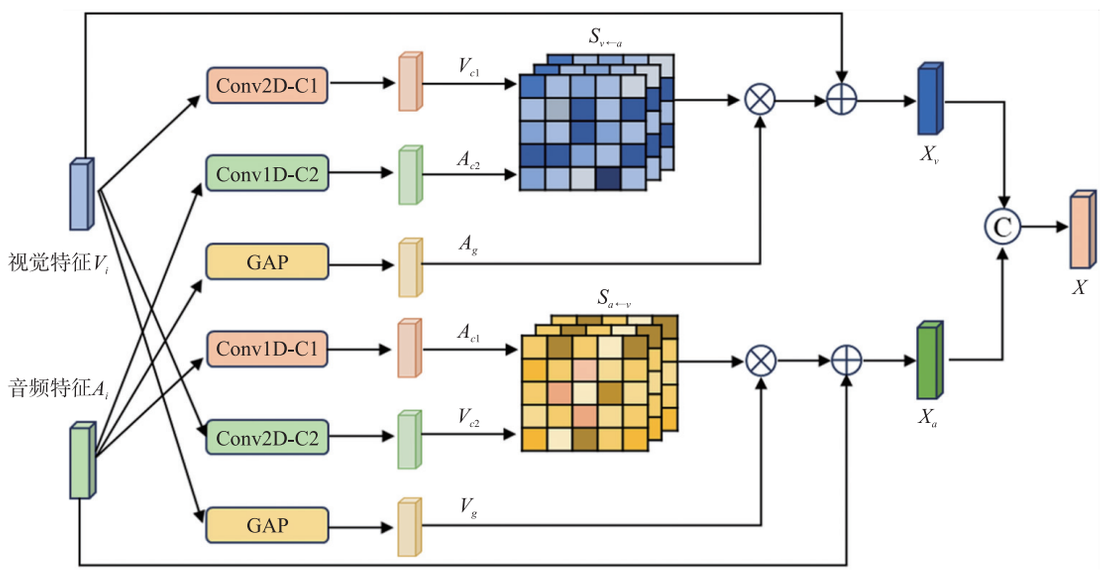


图 2　CM-FFM 模块结构图

Fig. 2　Structure of CM-FFM

增强以后的特征 $\boldsymbol{X}_v$ 和 $\boldsymbol{X}_a$ 可表示为:

$$\boldsymbol{X}_v=\boldsymbol{V}_i+\boldsymbol{W}_v(\boldsymbol{S}_{v\leftarrow a}(\boldsymbol{V}_{c1},\boldsymbol{A}_{c2})\boldsymbol{A}_g),\boldsymbol{X}_a=\boldsymbol{A}_i+\boldsymbol{W}_a(\boldsymbol{S}_{a\leftarrow v}(\boldsymbol{A}_{c1},\boldsymbol{V}_{c2})\boldsymbol{V}_g),\tag{5}$$

其中,W_a 和 W_v 为可学习参数.

获得增强后的音视频特征后,将其进行串联拼接,得到融合后的特征. 该融合模块的核心在于利用双向跨模态注意力机制建立时序相关性,通过音视频特征间的交互引导实现特征的跨模态融合.

2.2　特征对齐模块

将提取到的文本特征 $\boldsymbol{T}^c$ 与音视频融合特征 $\boldsymbol{X}^c$ 输入至 FAM 模块进行聚类操作. 首先,将音视频特征 $\boldsymbol{X}^c$ 映射到聚类空间,采用 K-Means 聚类方法得到 K 个聚类中心$\{c_1,c_2\cdots c_K\}$. 随后,计算各聚类中心对应的聚合特征,公式如下:

$$\boldsymbol{g}_j=\mathrm{nor}\left(\sum_{i=1}^{T}\frac{\exp(\boldsymbol{x}_i^c\boldsymbol{c}_j^{\mathrm{T}}+b_j)}{\sum_{k=1}^{K}\exp(\boldsymbol{x}_i^c\boldsymbol{c}_k^{\mathrm{T}}+b_k)}(\boldsymbol{x}_i^c-\boldsymbol{c}_j)\right),\tag{6}$$

其中,x_i^c 表示任意音视频特征向量,c_j 为第 j 个聚类中心,nor 表示 L_2 归一化,b_j 为可学习权重. 聚合后的音视频特征集合表示为 $G=\{\boldsymbol{g}_1,\boldsymbol{g}_2,\cdots\boldsymbol{g}_K\}$.

将文本特征同样映射到该聚类空间,并利用已获得的 K 个聚类中心进行聚类,以相同方式计算聚合文本特征:

$$\boldsymbol{g}_j^{\mathrm{text}}=\mathrm{nor}\left(\sum_{i=1}^{T}\frac{\exp(\boldsymbol{t}_i^c\boldsymbol{c}_j^{\mathrm{T}}+b_j)}{\sum_{k=1}^{K}\exp(\boldsymbol{t}_i^c\boldsymbol{c}_k^{\mathrm{T}}+b_k)}(\boldsymbol{t}_i^c-\boldsymbol{c}_j)\right),\tag{7}$$

其中,$\boldsymbol{t}_i^c$ 表示任意文本特征向量. 得到聚合文本特征 $G^{\mathrm{text}}=\{\boldsymbol{g}_1^{\mathrm{text}},\boldsymbol{g}_2^{\mathrm{text}},\cdots\boldsymbol{g}_K^{\mathrm{text}}\}$.

由于音视频特征和文本特征共享同一组聚类中心,因此可实现有效的跨模态特征对齐. 采用欧氏距离计算两类聚合特征之间的局部相似度:

$$S(G,G^{\mathrm{text}})=\sqrt{\sum_{i=1}^{K}(\boldsymbol{g}_i-\boldsymbol{g}_i^{\mathrm{text}})^2},\tag{8}$$

其中,$S(G,G^{\text{text}})$表示聚合音视频特征与聚合文本特征之间的相似度. 若该值大于设定阈值(本文取 0.7),则认为两类特征已实现对齐.

2.3 损失函数

模型的损失函数由两部分组成:特征对齐损失和 MIL 损失. 两者共同优化模型,分别从跨模态语义对齐和异常片段识别两个角度提升检测性能.

特征对齐损失用于确保文本模态与音视频模态在语义空间中的一致性. 通过 KL(Kullback-Leibler)散度进行度量. KL 散度可衡量两个概率分布之间的差异,其计算公式如下:

$$L_{\text{KL}}=E[\log S-\log p], \tag{9}$$

其中,L_{KL}表示 KL 散度损失,E 表示期望,p 为一致性标签. 若聚合音视频特征 G 与聚合文本特征 G^{text}已对齐,则 $p=1$,否则,$p=0$. 训练过程中通过最小化 L_{KL}来缩小预测分布 S_i 和真实标签 p 之间的差异.

MIL 损失基于交叉熵构建,用于根据片段异常得分生成视频级异常检测结果. 对于异常视频,选取 top-k 片段的得分均值作为视频级得分,其中 $k=\lfloor T/16+1\rfloor$,T 为片段个数,其计算公式如下:

$$L_{ce}=\sum_{i=1}^{N}-[p_i\log(y_i)-(1-p_i)\log(1-y_i)], \tag{10}$$

其中,p_i 表示视频级标签(异常为 1,正常为 0),y_i 为预测得分,N 为样本数量.

模型的总损失函数为:

$$L=L_{ce}+\lambda L_{\text{KL}}, \tag{11}$$

其中,λ 为平衡两项损失的权重系数.

3 实验数据及结果分析

3.1 数据集及评价指标

为验证本文方法在视频异常检测任务中的有效性,在公开的多模态异常检测数据集 XD-Violence 上开展实验. XD-Violence 是目前较新且规模较大的多模态暴力检测数据集,包含 4 754 个未经剪辑的视频片段,涵盖监控录像、电影、车载摄像头和游戏等多种场景. 数据集共包含六类异常事件,分别为虐待、车祸、爆炸、斗殴、暴乱和枪击,视频总时长约为 217 h. 其中训练集包含 3 954 个视频,测试集包含 800 个视频. 评价指标[17]采用帧级平均精度(average precision,AP),该指标广泛用于视频异常检测任务,能够综合评价模型在异常/正常片段分类及其置信度排序方面的性能.

3.2 实验细节

实验在 Windows 操作系统下完成,基于 PyTorch 框架和 CUDA 10.0 环境实现. 硬件配置为 Intel Core i7-8700K CPU、Nvidia GeForce RTX 2080 Ti GPU 及 32GB 内存.

具体实验设置如下:视频以每秒 24 帧采样,使用 16 帧的非重叠滑动窗口提取视觉特征. 音频被切分为 960 ms 的重叠片段,并转换为 96×64 维的 log-Mel 频谱图作为特征表示. 模型共训练 50 个 epoch,批大小设为 128,优化器选用 Adam,初始学习率为 0.001. 全连接层的隐藏层维度设置为 256. 损失函数中权重系数 λ 设为 0.5,MIL 损失中 top-k 的 k 值取 9(λ 和 k 均为经实验验证取得最佳性能的经验值).

3.3 实验结果分析

在 XD-Violence 数据集上,将本文所提方法 TFA-MVAD 与其他先进的异常检测方法进行了对比实验,实验结果如表 1 所示. 可以看出,TFA-MVAD 在该数据集上取得了 84.32%的最佳 AP,优于其他对比方法,表现出最佳检测性能.

表 1 在 XD-Violence 数据集上实验结果对比

Table 1 Comparison of experimental results on the XD-Violence dataset

算法	模态	AP/%	算法	模态	AP/%	算法	模态	AP/%
CUPL[18]	A+V	81.43	AlMarri et al[21]	A+V	71.40	ALAN[23]	A+V+T	82.3
UR-DMU[19]	A+V	81.77	MSBT[8]	A+V	82.54	TFA-MVAD(Ours)	A+V+T	84.32
RM Module[20]	A+V	78.64	AVCL[22]	A+V	81.11			

具体分析如下:与仅使用音频和视频的方法相比,TFA-MVAD 的性能显著优化. 相较于同样采用音视

频跨模态融合机制的 AVCL 方法,本文方法在 AP 上提升了 3.21%,验证了所提出的 FA 模块的有效性. 与多尺度融合策略的 MSBT 方法相比,TFA-MVAD 提升了 1.78%,说明本文的多尺度特征融合策略能有效地提取视频中的局部和全局信息,与通过建模片段间关系融合音视频特征的 RM Module 方法相比,本文方法提升幅度达 5.68%,进一步体现出跨模态特征融合模块的优越性. 与致力于增大异常-正常特征差异的 UR-DMU 方法相比,TFA-MVAD 仍领先 2.55%,表明本文在特征表示学习方面的策略能更清晰地区分正常与异常样本. 此外,与同样使用了 3 种模态的 ALAN 方法相比,本文方法仍取得了 2.02%的性能提升.

3.4　消融实验

为进一步验证本文方法中各组件的重要性,在 XD-Violence 数据集上开展消融实验. 验证内容包括 CM-FFM 和 FAM 模块的有效性验证. 实验结果如表 2 所示.

表 2　在 XD-Violence 数据集上的各模块消融实验结果

Table 2　Ablation study results of modular components on the XD-Violence dataset

序号	Baseline	CM-FFM	FAM	AP/%
1	√	×	×	74.2
2	√	√	×	78.46
3	√	×	√	79.69
4	√	√	√	84.32

在基准实验(Baseline)中,使用 I3D 和 VGG 模型分别作为视频和音频的特征编码器. 在提取音视频特征后,未进行任何增强,仅通过简单的特征拼接进行融合,并以 MIL 损失作为约束. 实验结果表明,基准模型在 XD-Violence 数据集上的 AP 为 74.2%,该结果为后续模块的改进提供了基线参考. 首先,在引入 CM-FFM 模块后,模型的 AP 提升了 4.26%,表明该模块能够有效促进跨模态交互,实现音频和视频特征间互补. 随后,引入 FAM 模块,与 Baseline 相比,AP 增加了 5.49%,说明该模块能有效利用文本信息对音视频融合过程进行提示,增强模型性能. 从表中第 2 和第 3 行可以看出,单独添加 FAM 模块比单独添加 CM-FFM 模块在 AP 上进一步提高了 1.23%,进一步证明了文本信息在融合中的重要性. 最终,同时使用 CM-FFM 和 FAM 模块时,模型达到最佳性能,AP 为 84.32%.

3.5　可视化分析

使用 t-SNE 方法对原始特征和中间特征的处理效果进行可视化. 如图 3(a)和(b)所示,原始样本特征呈现出混杂无序的分布状态. 具体表现为:不同异常事件的特征相互重叠,未形成清晰的聚类边界;正常

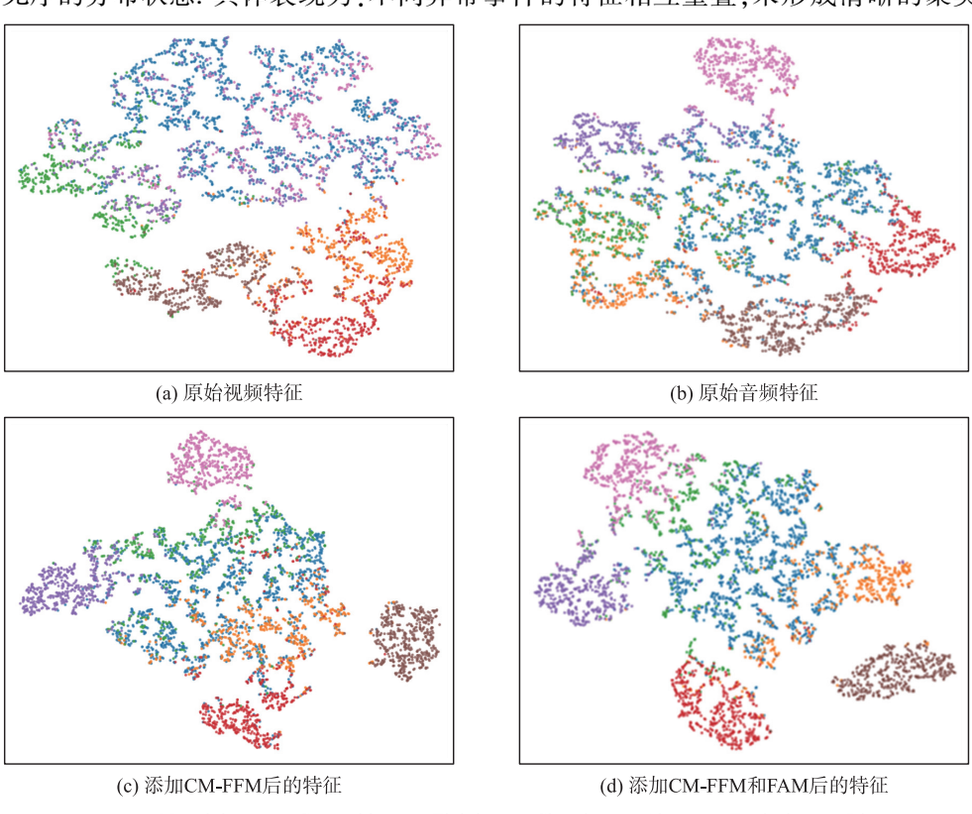


图 3　基于 t-SNE 方法的视频特征可视化图

Fig. 3　Visualization of video features based on t-SNE method

事件与异常事件之间也存在大量特征重叠. 这一现象说明,原始特征缺乏足够的判别性,难以有效区分不同类型的异常事件.

相比之下,图 3(c)展示了经 CM-FFM 模块处理后的特征分布,可以看出,不同类型的异常事件开始呈现聚类趋势,例如 shooting 和 fighting 事件的特征分别聚集在不同区域;正常事件与异常事件之间的分离程度也明显提高. 结果表明,通过融合音视频互补信息,CM-FFM 模块有效增强了特征的判别能力,促进了不同类别之间的分离. 图 3(d)为进一步引入 FAM 模块后的特征可视化结果. 图中可见,不同异常事件之间的边界更加清晰,类间距离进一步增大,聚类结构更为明显. 这说明通过引入文本提示信息并进行特征对齐,FAM 模块显著提升了特征的区分度,能够更好地区分不同类型的异常事件.

4 结论

在复杂场景下的视频异常检测任务中,现有方法往往未能充分考虑不同模态之间的时序相关性,导致难以准确判定异常事件发生的具体时空位置. 本文提出了一种基于文本特征对齐的视频异常检测方法 TFA-MVAD. 该方法通过跨模态模块提取并融合音视频模态中的互补信息,建模帧间依赖关系;同时构建可学习的字典生成文本提示特征,用于对齐多模态融合特征,并通过匹配异常上下文与其对应提示特征以抑制非异常信息,从而增强多模态特征在语义层面的一致性和连贯性. 在 XD-Violence 数据集上的实验及对比分析结果表明,所提出的 TFA-MVAD 方法能够有效实现异常片段的精准检测.

[参考文献]

[1] 汪洋,周脚根,严俊,等. 基于深度学习的监控视频异常检测方法综述[J]. 中国图像图形学报,2025,30(3):615-640.

[2] Cheng B,Wenjia Z,Sun X,et al. A novel spatio-temporal information knowledge distillation strategy for weak-supervised video anomaly detection[C]//The 7th International Conference on Pattern Recognition and Artificial Intelligence,Hangzhou,China,2024:398-403.

[3] 李歌,肖洪兵,闫善武,等. 基于多尺度时空特征融合的视频异常事件检测[J]. 燕山大学学报,2025,49(1):74-82.

[4] 余烨,陈维笑,陈凤欣. 面向车型识别的夜间车辆图像增强网络 RIC-NVNet[J]. 中国图像图形学报,2023,28(7):2054-2067.

[5] Zhou X,Peng X,Luo Y,et al. Learning weakly supervised audio-visual violence detection in hyperbolic space[J]. Image and Vision Computing,2024,151:105286.

[6] Chen J,Li L,Su L,et al. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle,WA,USA,2024:18319-18329.

[7] Wu P,Liu J,Shi Y,et al. Not only look,but also listen:Learning multimodal violence detection under weak supervision[C]//16th European Conference on Computer Vision. Glasgow,UK,2020:322-339.

[8] Pang W F,He Q H,Hu Y,et al. Violence detection in videos based on fusing visual and audio information[C]//ICASSP 2021-2021 IEEE international conference on acoustics,speech and signal processing(ICASSP). Toronto,Ontario,Canada,2021:2260-2264.

[9] Pu Y,Wu X. Audio-guided attention network for weakly supervised violence detection[C]//2022 2nd International Conference on Consumer Electronics and Computer Engineering(ICCECE). Guangzhou,China,2022:219-223.

[10] Sun S,Gong X. Multi-scale bottleneck transformer for weakly supervised multimodal violence detection[C]//IEEE International Conference on Multimedia and Expo(ICME). Niagara Falls,Canada,2024:1-6.

[11] Yu J,Liu J,Cheng Y,et al. Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection[C]//Proceedings of the 30th ACM International Conference on Multimedia. Lisbon,Portugal,2022:6278-6287.

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

[13] Zhou X,Peng X,Wen H,et al. Learning weakly supervised audio-visual violence detection in hyperbolic space[J]. Image and Vision Computing,2024,151:105286.

[14] Ghadiya A,Kar P,Chudasama V,et al. Cross-modal fusion and attention mechanism for weakly supervised video anomaly

detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA, 2024:1965-1974.

[15] Sharma M K, Sheet D, Biswas P K. Local instance and context dictionary-based detection and localization of abnormalities[J]. Machine Vision and Applications, 2021, 32(3):69.

[16] Speer R, Havasi C. Representing general relational knowledge in ConceptNet 5[J]. European Language Resources Association (ELRA), 2012.

[17] 朱文佳,张婷,程茹秋,等. 考虑背景失真的无参考视频质量评价方法[J]. 南京师大学报(自然科学版),2025,48(3):102-111.

[18] Zhang C, Li G, Qi Y, et al. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canada, 2023:16271-16280.

[19] Zhou H, Yu J, Yang W. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Washington, DC, USA, 2023, 37(3):3769-3777.

[20] Wu P, Liu X, Liu J. Weakly supervised audio-visual violence detection[J]. IEEE Transactions on Multimedia, 2023, 25:1674-1685.

[21] AlMarri S, Zaheer M Z, Nandakumar K. A multi-head approach with shuffled segments for weakly-supervised video anomaly detection[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, HI, USA, 2024:132-142.

[22] Meng J, Tian H, Lin G, et al. Audio-visual collaborative learning for weakly supervised video anomaly detection[J]. IEEE Transactions on Multimedia, 2025, 27:1-12.

[23] Wu P, Liu J, He X, et al. Toward video anomaly retrieval from video anomaly detection:new benchmarks and model[J]. IEEE Transactions on Image Processing, 2024, 33:2213-2225.

[责任编辑:陆炳新]