基于跨模态融合与双曲图注意力机制的视频异常检测

姜迪[1,2,3]，赖惠成[1,2,3]，汪烈军[1,2,3]
（1.新疆大学计算机科学与技术学院，新疆 乌鲁木齐 830017；
2.新疆大学新疆维吾尔自治区信号检测与处理重点实验室，新疆 乌鲁木齐 830017；
3.丝路多语言认知计算国际合作联合实验室，新疆 乌鲁木齐 830017）

摘　要：针对视频异常检测中模态信息不平衡、视听噪声不平均以及模态异步等问题，提出了一个动态跨模态融合模块与双曲图注意力机制融合的多模态视频异常检测方法CM-HVAD，以准确检测异常行为。首先，提出了一种新的动态跨模态融合模块，动态压缩多模态数据特征，自主学习跨模态权重，动态平衡视觉特征和音视频特征并进行融合增强。然后，针对多模态数据中存在的模态异步问题，提出了模态一致性对齐模块，按时间帧序列对齐模态语义，确保多模态数据在时间和语义上的一致性。最后，引入了双曲图注意力机制，通过双曲空间的模式分离特性，有效捕捉正常和异常表示之间的层次关系，从而提高检测准确率。实验结果表明，所提方法在XD-Violence上AP达到了86.47%，在UCF-Crime上AUC达到了87.12%，性能优于基线方法。

关键词：视频异常检测；跨模态融合；双曲图注意力机制；多模态
中图分类号：TP391.41
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2025110

Video anomaly detection via cross-modal fusion and hyperbolic graph attention mechanism

JIANG Di[1,2,3], LAI Huicheng[1,2,3], WANG Liejun[1,2,3]
1. College of Computer Science and Technology, Xinjiang University, Urumqi 830017, China
2. Key Laboratory of Signal Detection and Processing, Xinjiang Uygur Autonomous Region, Xinjiang University, Urumqi 830017, China
3. Joint Laboratory of Silk Road Multilingual Cognitive Computing International Collaboration, Urumqi 830017, China

Abstract: To address the challenges of modality information imbalance, non-uniform audiovisual noise, and modality asynchrony in video anomaly detection, a multimodal video anomaly detection method called CM-HVAD was proposed for accurate anomaly detection. Firstly, a novel dynamic cross-modal fusion module was introduced to dynamically compress and reweight multimodal features through autonomous learning of cross-modal weights, thereby achieving balanced and enhanced fusion of visual and audio features. Secondly, to address the issue of modal asynchrony in multimodal data, a modal consistency alignment module was proposed, which aligned modal semantics along the temporal frame sequence to ensure both temporal and semantic consistency in multimodal data. Finally, a hyperbolic graph attention mechanism was incorporated to effectively capture the hierarchical relationships between normal and abnormal representations through the pattern separation property of hyperbolic space, thereby improving detection accuracy. The results show that the proposed method achieves 86.47% AP on XD-Violence and 87.12% AUC on UCF-Crime, outperforming baseline methods.

Keywords: video anomaly detection, cross-modal fusion, hyperbolic graph attention mechanism, multi-modal

收稿日期：2025–04–17；修回日期：2025–06–04
通信作者：赖惠成，lai@xju.edu.cn
基金项目：国家自然科学基金联合基金资助项目（No.U1903213）；新疆维吾尔自治区重点研发计划基金资助项目（No.2022B01008）
Foundation Items：The National Natural Science Foundation of China Joint Fund Project (No.U1903213), The Key Research and Development Program of Xinjiang Uygur Autonomous Region (No.2022B01008)

0　引言

近年来，视频异常检测[1]（VAD, video anomaly detection）已成为计算机视觉领域的研究热点之一，旨在从视频数据中自动识别非正常事件。随着城市化进程的加快和公共安全需求的增加，智能视频监控系统在机场、车站、学校等公共场所得到了广泛应用。然而，传统监控系统主要依赖人工观察，存在检测效率不足、易因人为疲劳导致漏检和误检等问题。因此，开发自动化视频异常检测技术具有重要的研究价值与现实意义[2]。

视频异常事件在安防等领域尤为重要，其典型表现包括突发性冲突、危险动作等行为，具有显著的突发性、多样性和复杂性等特点，这使其自动检测面临巨大挑战。早期方法[3-4]主要基于手工设计特征（如光流、纹理特征等），但这些方法在复杂场景下的泛化能力较差。此外，单一模态（如视频帧）信息往往难以全面描述异常事件，尤其是在光照变化、遮挡等复杂条件下[5]。因此，多模态学习[6-8]逐渐成为视频异常检测的主流方法之一。通过融合视频、音频、文本等多种模态信息，可以更全面地表征异常行为特征，从而提高检测的准确性和鲁棒性[9]。例如，音频信息可以帮助识别突发声响等异常声学线索，而文本信息（如字幕）可以提供上下文语义支持[10]。多模态融合技术的引入为视频异常检测提供了新的研究思路，同时也带来了跨模态对齐、信息异构性等新的挑战。

得益于深度学习的快速发展，长短期记忆[11]（LSTM, long short term memory）网络和注意力机制[12]等模型的应用，使视频异常检测性能得到显著提升。文献[13]使用 LSTM 网络学习时空特征，能够有效识别视频中的运动模式和外观变化，从而对异常行为进行精准定位。文献[14]提出了基于卷积神经网络（CNN, convolutional neural network）的 CNN-LSTM 框架，利用注意力机制提取时空特征并进行目标定位，以识别运动过程中的异常行为。此外，生成对抗网络和自监督学习等新兴技术[15-17]的引入，进一步推动了该领域发展。文献[18]提出了一种多流无监督框架，该框架由伪标记的视频序列和潜在的检测特征组成。上述方法不仅能够在公开数据集上取得优异的性能，还在实际监控场景中展现了广泛的应用潜力。尽管 VAD 技术取得了显著进展，但在复杂场景检测中仍面临诸多挑战[19]。文献[20]基于帧间一致性的自监督方法更新注释标签，但在低分辨率下易产生预测偏差。文献[21]提出的双流架构（2D-CNN 与 3D-CNN）通过特征增强机制处理复杂时空模式，但在高度动态场景中易出现信息丢失问题。文献[22]使用 ResNet 和简单循环单元（SRU, simple recurrent unit）提取时间特征，并行计算多类行为异常分类，但计算成本较高。上述研究在视频异常检测领域取得了显著进展，但在处理复杂交互行为时仍存在检测精度与鲁棒性方面的局限。

除单一模态视频帧外，声音等多维度输入同样有助于异常行为的定位[23]，仅使用视觉输入便可以准确地区分和识别异常行为，但同时也会导致无效的检测，而音频输入可以有效地辅助识别视觉上模糊和复杂的活动。因此，多模态比单模态更适用于视频异常检测。文献[24]表明，使用视听特征比仅使用视觉特征能获得更好的检测性能。在某些检测场景下，音频数据比复杂的视觉数据更有效，能够更好地区分异常和正常事件[25]。文献[26]采用多模态数据集，利用 3D-CNN 学习视听特征表示，用于动作识别和物体分类。然而，部分多模态数据不同模态的数据分布和特征表示差异较大且存在异步问题，这使多模态训练面临困难。文献[27]提出了对比语言−图像预训练（CLIP, contrastive language-image pre-training）增强的双模存储网络，通过设计分别存储视觉特征和文本描述特征的记忆模块来解决这一问题。总体而言，多模态学习方法通过结合视频、音频等多种信息，提高了视频异常检测性能，但在跨模态对齐和特征融合方面仍有改进空间。

图注意力机制[28]（GAT, graph attention network）是一种基于注意力机制的图神经网络[29]（GNN, graph neural network），旨在处理图结构数据。GAT 通过自适应地分配注意力权重来聚合邻居节点的信息，进而捕捉图中节点间的高阶关系。文献[30]结合空间注意图卷积（提取空间全局特征）与通道注意力图卷积（提取通道信息），构建包含归一化流的高斯模型检测框架，通过概率分布评估行为正常性。文献[31]设计了动态注意力增强图网络，集成全局上下文线索与时间注意力模块，利用潜在空间关系建模。文献[32]开发了基于动态图卷

积的异常图方法，通过时空嵌入特征提取和唯一性得分量化，实现时空异常检测。

然而，虽然上述多模态方法在VAD领域取得了显著进展，但仍存在一些亟待解决的问题。现有方法主要关注模型架构设计与正常/异常模式分离，对多模态数据固有特性（如模态信息不平衡、视听噪声不平均等）的优化仍有不足。因此，本文提出了一个动态跨模态融合模块与双曲图注意力机制融合的多模态视频异常检测（CM-HVAD, dynamic cross-modal fusion with hyperbolic attention for video anomaly detection）方法。在多模态融合部分，提出了动态跨模态融合（DCMF, dynamic cross-modality fusion）模块，动态压缩多模态数据特征，缓解信息冗余。设计了动态权重机制，依靠模态信息动态加权。此外，由于不同数据存在不同类型异构特征，本文通过构建场景感知控制单元，针对输入数据的场景异构特性（如运动强度、音频频谱等），自动生成融合策略。由于多模态数据还存在模态异步问题，即使输入时间帧同步，不同模态间仍可能时间帧序列不一致。因此，本文提出了模态一致性对齐（MCA, modality consistency alignment），按时间帧序列对齐模态语义。在采用图论区分正常和异常特征上，先前的研究方法多使用了图网络或图注意力机制，这些方法在准确区分特征方面仍然存在一定的进步空间。因此，本文引入了双曲图注意力（HGAtt, hyperbolic graph attention）机制，通过双曲空间的模式分离特性，学习正常和异常特征之间的时空关系。本文主要工作如下。

1) 提出了动态跨模态融合模块，动态压缩多模态数据特征，有效缓解信息梯度爆炸与冗余，并设计动态权重机制，实现对不同模态信息的自适应加权，提升多模态数据的利用率。

2) 针对多模态数据中存在的模态异步问题，提出了模态一致性对齐方法，按时间帧序列对齐模态语义，确保多模态数据在时间和语义上的一致性。

3) 引入了双曲图注意力机制，通过双曲空间的模式分离特性，更准确地学习正常和异常特征之间的时空关系，进一步增强模型对复杂场景中异常特征的捕捉能力。

4) 实验结果表明，在XD-Violence数据集上的检测指标平均精度（AP, average precision）达到了86.47%，在UCF-Crime数据集上的检测指标曲线下面积（AUC, area under the curve）达到了87.12%，证明了本文方法的实际效果，验证了本文方法的有效性。

1 相关工作

1.1 视频异常检测

视频异常检测旨在通过分析视频中的时空模式，自动识别偏离正常行为或视觉特征的事件。早期研究依赖手工设计[3-4]特征描述视频内容，通过概率模型（如高斯混合模型[33]）或稀疏编码[34]等构建正常模式基准，以统计偏差判定异常。但手工设计特征的表达能力有限，难以关联复杂时空特征，手动设定阈值导致泛化能力弱。因此，深度学习方法成为VAD主流特征学习模式。

深度学习通过端到端特征学习，如重构网络（如自编码器[15]）、预测网络（如连续帧预测[35]）和分类网络（如多实例学习[36]），对视频数据进行检测。尽管深度学习模型提升了检测精度，但单一视频数据的输入仍存在挑战，如对多尺度异常感知度不均衡，过度依赖外观-运动模态的浅层融合特征等。因此，融合音频、文本等多模态信息的检测技术成为VAD高频研究之一。

1.2 双曲空间学习

双曲空间作为一种具有恒定负曲率的非欧几里得流形，近年来在表征层次化数据结构方面展现出独特优势[37]。相较于欧式空间，双曲空间的指数级扩展特性与树状结构、社交网络等现实数据的几何性质高度契合，为深度学习中的层次关系建模提供了新的数学工具。随着深度学习的发展，不同于早期主要集中于双曲嵌入的理论探索研究，双曲空间与神经网络的融合成为研究热点[30]。当前，双曲神经网络已在多个深度学习领域成功应用，如通过双曲门控机制建模层次化时序依赖[11]的递归网络、构建基于负曲率空间跨模态对齐模型的注意力机制[12]、利用双曲注意力捕捉层次化节点关系的图神经网络[29]等。

1.3 图注意力机制

图注意力机制通过动态学习节点关系权重，为VAD中的时空依赖建模提供了新范式。使用时空关系建模，可以通过注意力机制自动量化节点间关联度，克服传统固定邻接矩阵对复杂时空模式特征提

取不足的缺陷。进一步，通过多粒度语义感知，支持从像素级到对象级的多层次图构建，同时捕捉局部异常细节与全局行为模式偏离。文献[38]提出了一种基于时空图卷积网络的预测网络，将视频帧划分为网格或超像素区域，作为节点捕捉局部异常，构建一个法线图描述正常数据中关节的图连接，其中异常事件的关节是该图的离群点。文献[39]利用基于隐向量的注意力机制对包级特征加权，以固定时长视频片段为节点，建模长时序关联，针对视频中行走的异常，提出了一个端到端的异常行为检测网络。此外，GAT还能基于目标检测提取物体实例作为节点，对多对象进行场景交互。

1.4　跨模态视频异常特征融合

跨模态视频异常特征融合旨在整合来自不同模态（如视觉、音频、文本等）的信息，以提升模型对复杂场景的理解能力。在视频异常检测任务中，跨模态特征融合能够充分利用多模态数据的互补性，提高检测的准确性和鲁棒性。早期研究方法通常采用简单的操作来融合多模态特征。例如，文献[40]通过拼接视觉特征和音频特征来增强运动信息。文献[41]使用加权求和的方式融合视觉特征和音频特征。尽管这些方法实现简单，但未能充分考虑模态间的语义关联和差异性，导致融合效果受限。为了更有效地捕捉模态间的交互信息，提出了基于注意力机制的融合方法。例如，文献[42]设计了一种跨模态注意力模块，动态调整不同模态特征的权重。文献[43]引入多头注意力机制，分别从时间和空间维度对齐多模态特征。这些方法能够在一定程度上缓解模态间的噪声干扰，但对模态异步问题的处理能力有限。

图神经网络被广泛应用于跨模态特征融合任务中。通过将不同模态的特征表示为图结构，GNN能够显式建模模态间的复杂关系。例如，文献[44]提出了一种多模态图卷积网络，通过消息传递机制实现模态间的信息交互。文献[45]利用GAT动态调整模态间的连接权重。然而，现有方法在时间同步性问题上缺乏有效网络架构，不同模态间的时间或语义不一致会降低融合效果。此外，多模态数据中可能存在噪声，影响模型的鲁棒性，复杂的融合机制也会导致计算开销过大。为了解决上述问题，本文提出了一种动态跨模态融合模块，通过多模态数据压缩和动态权重机制缓解噪声干扰，提出模态一致性对齐方法解决模态异步问题，并结合HGAtt机制进行双曲面异常异构，从而提升视频异常检测的性能。

2　跨模态双曲图注意力视频异常检测

本文提出了一种全新的多模态融合与特征学习框架CM-HVAD，旨在通过动态跨模态融合、模态一致性对齐和双曲图注意力机制，全面提升视频异常检测的性能。网络框架如图1所示，整体架构由3个核心模块组成。首先，DCMF模块通过动态压缩多模态数据特征和动态权重机制，能够有效融合视听特征并缓解冗余特征。其次，MCA模块通过时间帧序列对齐模态语义，解决模态异步问题。最后，HGAtt模块利用双曲空间的层次化表示能力，进一步区分正常与异常特征。通过这3个模块的协同工作，本文方法能够显著提升视频异常检测的准确性和鲁棒性。

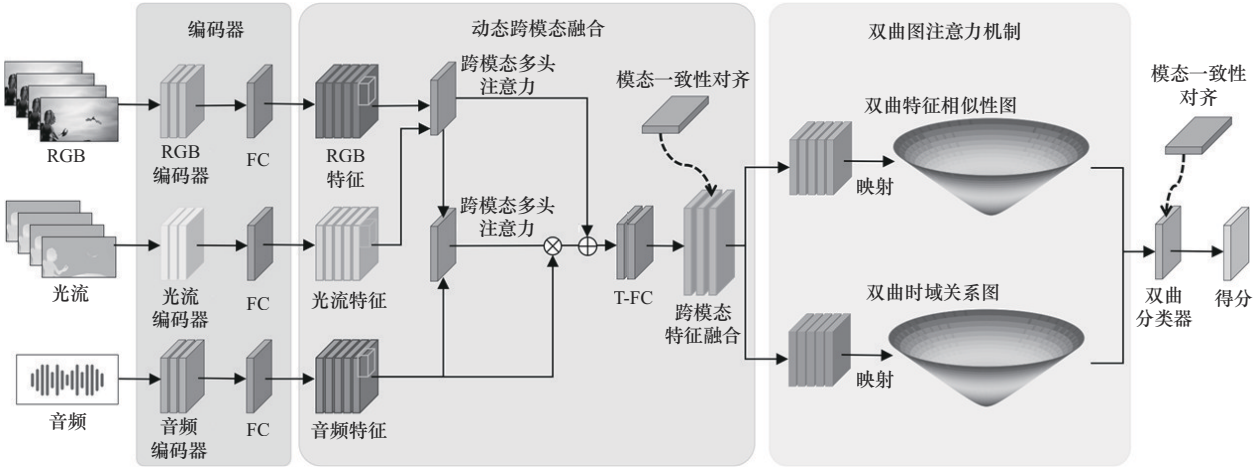

图1　基于跨模态融合与双曲图注意力机制的视频异常检测网络框架（CM-HVAD）

2.1　多模态特征提取

给定包含 K 个视频的数据集及其对应的视频级标签 $L=\{L_i\}\in\{0,1\}$，其中 $L_i=1$ 表示视频中存在异常事件，$L_i=0$ 表示视频中不存在异常事件。在进行视频异常检测时，每个视频被划分为多个包含 RGB、光流和音频的多模态实例，通过构造的 3 个编码器生成不同模态的特征。然后通过全连接（FC, fully connected）层聚合模态特征。多模态特征提取如图 2 所示，其中图 2(a)为 RGB 特征或光流特征提取架构，通过结合 3D 卷积与 ResNet 模块对视频流进行维度特征提取，输出维度为 1 024 的视频特征向量，为后续特征运算提供支持。图 2(b)为音频特征提取架构，通过 2D 卷积与全连接模式对音频流进行处理，获得输出维度为 128 的音频特征向量。

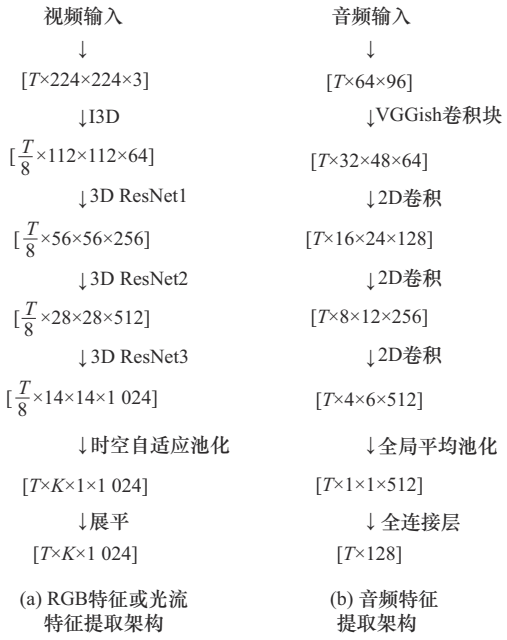

图2　多模态特征提取

2.2　动态跨模态融合

DCMF 模块由跨模态多头注意力（CMHA, cross-modal multi-head attention）模块和动态权重机制组成，如图 3 所示。CMHA 采用层次化融合的方式，通过分阶段的方式，逐步融合 3 个模态的信息，捕捉更复杂的交互关系，动态压缩多模态数据特征，减少数据冗余，并且可以根据任务需求调整每个步骤的模态融合方式，易于扩展到更多模态或更复杂的融合策略。调制机制通过调节因子动态地调节融合过程中各模态权重，有利于模态的可学习性。

1) 跨模态多头注意力模块（CMHA）。为了使模态间更好地进行交互，注意力机制使用双向模式，即模态 a 和模态 b 互为查询（Query）和键值对（Key），下面以单向模式为例介绍。

首先，将 RGB 特征 $\boldsymbol{F}_V$ 作为 Query ($\boldsymbol{Q}$)，将光流特征 $\boldsymbol{F}_L$ 作为 Key ($\boldsymbol{K}_L$)和 Value ($\boldsymbol{V}_L$)，从而通过跨模态多头注意力机制计算 $\boldsymbol{F}_V$ 和 $\boldsymbol{F}_L$ 的匹配分数。然后，对 $\boldsymbol{V}_L$ 进行加权求和，表示为模态 $\boldsymbol{F}_V$ 对模态 $\boldsymbol{F}_L$ 提取信息，生成特征中间表示 $\boldsymbol{F}_{VL}$。同理，将 $\boldsymbol{F}_{VL}$ 作为 Query ($\boldsymbol{Q}$)，将音频特征 $\boldsymbol{F}_A$ 作为 Key ($\boldsymbol{K}_A$)和 Value ($\boldsymbol{V}_A$)，通过跨模态多头注意力机制进一步融合，表示为模态 $\boldsymbol{F}_{VL}$ 对模态 $\boldsymbol{F}_A$ 提取信息，生成最终输出 $\boldsymbol{F}_{VLA}$。这种分阶段先融合前 2 个模态，再将结果与第 3 个模态融合的方式，由于在第一步中已经压缩了前 2 个模态的信息，从而在第二步减少了后续计算的开销。因此，不加动态权重机制的单向注意力机制的计算如式(1)所示。

$$\boldsymbol{F}_{VLA}=\mathrm{Attention}\left(\mathrm{Attention}\left(\boldsymbol{F}_V,\boldsymbol{K}_L,\boldsymbol{V}_L\right),\boldsymbol{K}_A,\boldsymbol{V}_A\right)=\mathrm{Softmax}\left[\frac{\mathrm{Softmax}\left(\dfrac{\boldsymbol{F}_V\boldsymbol{K}_L^{\mathrm{T}}}{\sqrt{d_{\boldsymbol{K}_L}}}\right)\boldsymbol{V}_L\boldsymbol{K}_A^{\mathrm{T}}}{\sqrt{d_{\boldsymbol{K}_A}}}\right]\boldsymbol{V}_A \tag{1}$$

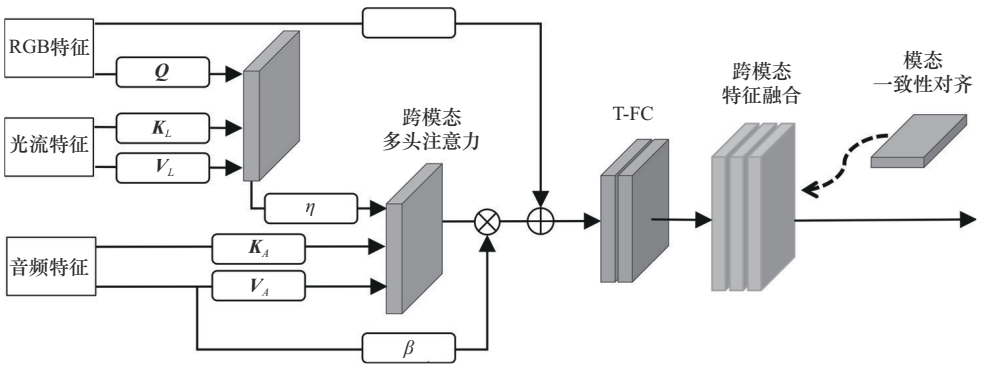

图3　动态跨模态融合模块

其中，$d_{\boldsymbol{K}_L}$表示$\boldsymbol{K}_L$的维度，$d_{\boldsymbol{K}_A}$表示$\boldsymbol{K}_A$的维度。由于在CMHA模块中采用多层结构，故在多层特征传递过程中，为了防止特征指数爆炸，通过模态压缩将特征数量减少一半。具体而言，记特征层为l，$\boldsymbol{F}_{VLA}^{l+1}=\dfrac{\boldsymbol{F}_{VLA}^{l}}{2}$，从而更快速与精炼地压缩存储模态信息。

同理可知，在双向注意力机制中，由于对$\boldsymbol{F}_V$、$\boldsymbol{F}_L$和$\boldsymbol{F}_A$输入顺序不同，最终结果不仅会得到融合特征$\boldsymbol{F}_{VLA}$，也会得到其他5种融合特征$\boldsymbol{F}_{VAL}$、$\boldsymbol{F}_{LVA}$、$\boldsymbol{F}_{LAV}$、$\boldsymbol{F}_{AVL}$和$\boldsymbol{F}_{ALV}$。这种双向模式可以捕捉模态内和模态间的依赖关系，获得不同方向模态信息流表示，其融合特征如式(2)所示。

$$\boldsymbol{F}_{\text{fusion}}=[\boldsymbol{F}_{VLA},\boldsymbol{F}_{VAL},\boldsymbol{F}_{LVA},\boldsymbol{F}_{LAV},\boldsymbol{F}_{AVL},\boldsymbol{F}_{ALV}] \tag{2}$$

其中，[⋯]表示拼接操作。随着CMHA模块不断更新，会自学习多层权重，穷举覆盖多模态关系中潜在的因果或依赖方向，根据任务需求选择最优信息流方向。因此最终模态如式(3)所示。

$$\begin{aligned}\boldsymbol{F}_{\text{fusion}}=[&\omega_1\boldsymbol{F}_{VLA},\omega_2\boldsymbol{F}_{VAL},\omega_3\boldsymbol{F}_{LVA},\\&\omega_4\boldsymbol{F}_{LAV},\omega_5\boldsymbol{F}_{AVL},\omega_6\boldsymbol{F}_{ALV}]\end{aligned} \tag{3}$$

其中，ω_1~ω_6是多个融合模态自学习特征总权重。

然后，在DCMF模块中，将CMHA模块生成的最终结果输入后续T-FC模块中，获得最后的跨模态融合特征。

2) T-FC模块。由一个Norm层、一个Transformer层、一个FC层和一个Norm层组成。其中Norm层对输入特征进行归一化，可以缓解梯度消失或爆炸问题，第一个Norm层确保输入Transformer层的特征分布更加均匀，第二个Norm层确保输出特征的分布更加稳定，同时，可以加速模型的收敛速度，减少训练时间。Transformer层可以捕捉输入特征的全局依赖关系，对于序列数据（视频和音频），Transformer层可以有效地建模长距离依赖，并同时进行并行计算，显著提高计算效率。FC层对特征进行非线性变换和维度映射，将特征的维度映射到目标维度，为后续双曲计算预设特征。具体而言，第一个Norm层和Transformer层负责提取全局特征，FC层负责对特征进行进一步的非线性变换和维度映射，第二个Norm层负责稳定输出特征的分布。具体不加动态权重机制的计算式如式(4)所示。

$$\text{T-FC}(\boldsymbol{F}_{\text{fusion}})=\text{Norm}(\text{FC}(\text{Trans}(\text{Norm}(\boldsymbol{F}_{\text{fusion}})))) \tag{4}$$

3) 动态权重机制。提出了一种动态权重调整机制，用于在多模态融合过程中灵活地调整每个模态的贡献。与传统的固定权重或单一调节因子方法不同，该机制引入了多个独立的调节因子，分别针对视频、音频和文本模态进行动态调整。具体来说，该机制通过以下步骤实现。对于每个模态，通过一个可学习的线性变换和非线性激活函数计算其调节因子。在视频模态中，其调节因子α的计算式如式(5)所示。

$$\alpha=\sigma(\omega_\alpha\boldsymbol{F}_V+b_\alpha) \tag{5}$$

其中，$\boldsymbol{F}_V$是视频模态特征，ω_α和b_α分别是可学习的权重矩阵和偏置向量，σ是Sigmoid激活函数，用于将调节因子的值限制在[0,1]范围内。与α类似，调节因子β和η的计算式分别如式(6)和式(7)所示。

$$\beta=\sigma(\omega_\beta\boldsymbol{F}_A+b_\beta) \tag{6}$$

$$\eta=\sigma(\omega_\eta\boldsymbol{H}+b_\eta) \tag{7}$$

其中，$\boldsymbol{F}_A$是音频模态特征，ω_β、b_β、ω_η和b_η分别是可学习的相应参数。在CMHA模块中，首先通过跨模态多头注意力机制生成$\boldsymbol{F}_V$与$\boldsymbol{F}_L$的中间特征表示$\boldsymbol{H}$。因此，$\boldsymbol{H}$具有模态交互信息，学习了第一个模态和第二个模态之间的交互关系，并且因为其基于$\boldsymbol{F}_V$和$\boldsymbol{K}_L$的匹配分数动态加权了$\boldsymbol{V}_L$，所以是上下文感知的。$\boldsymbol{H}$在第二步中用于与$\boldsymbol{F}_A$进一步融合。因此，$\boldsymbol{H}$虽然是视频和音频模态特征，但实际上，由于$\boldsymbol{F}_V$存在其独立的调节因子α，$\boldsymbol{H}$使用的调节因子η被视为音频模态调节因子。然后，对每个模态的特征进行动态调整其贡献。因此，增加动态权重机制后的特征表示式(1)更新为式(8)。

$$\begin{aligned}\boldsymbol{F}_{VLA}=&\text{Attention}\left(\eta\text{Attention}(\boldsymbol{F}_V,\boldsymbol{K}_L,\boldsymbol{V}_L),\boldsymbol{K}_A,\boldsymbol{V}_A\right)=\\&\text{Softmax}\left[\frac{\eta\text{Softmax}\left(\dfrac{\boldsymbol{F}_V\boldsymbol{K}_L^{\mathrm{T}}}{\sqrt{d_{\boldsymbol{K}_L}}}\right)\boldsymbol{V}_L\boldsymbol{K}_A^{\mathrm{T}}}{\sqrt{d_{\boldsymbol{K}_A}}}\right]\boldsymbol{V}_A\end{aligned} \tag{8}$$

由于多模态数据存在结构性差异，为动态适应不同场景的模态重要性变化，本文引入基于时序场景分析的二级权重调节，设计了可插拔模块场景感知门控单元（SAG, scene-aware gating unit）。

首先，SAG进行场景特征编码，设时间t的多模态特征$\boldsymbol{F}_V$、$\boldsymbol{F}_L$和$\boldsymbol{F}_A$为$\boldsymbol{F}_V^t$、$\boldsymbol{F}_L^t$和$\boldsymbol{F}_A^t$，通过双向LSTM（BiLSTM）编码构造联合特征空间建模场景上下文特征，如式(9)所示。其次，通过场景特

征生成模态权重系数如式(10)和式(11)所示。最后，将场景权重与动态权重调节因子结合，形成复合权重，如式(12)~式(14)所示。

$$\boldsymbol{h}_t = \mathrm{BiLSTM}([\boldsymbol{F}_V^t \| \boldsymbol{F}_L^t \| \boldsymbol{F}_A^t]; \boldsymbol{h}_{t-1}) \tag{9}$$

其中，$\|$表示拼接操作，$\boldsymbol{h}_t$捕获当前时刻的跨模态场景模式。具体而言，首先通过拼接操作保留了原始模态的独立信息量，而后双向LSTM通过门控机制实现模态间信息流动，例如，在异常倒地事件中，先听到声音（音频模态突显）后看到倒地动作（光流模态变化），双向LSTM的最终隐藏状态融合了这种因果-结果关联。$\boldsymbol{h}_t$本质是跨模态时序特征的压缩表示，实现了信息蒸馏。

$$\boldsymbol{g}_t = \mathrm{Softmax}(\omega_h \boldsymbol{h}_t + b_h) \tag{10}$$

$$[\alpha^*, \beta^*, \eta^*] = \boldsymbol{g}_t \tag{11}$$

其中，$\boldsymbol{g}_t$是调节因子，ω_h和b_h分别是可学习的权重矩阵和偏置向量，α^*、β^*和η^*分别是通过场景特征生成的模态权重系数。

$$\alpha' = \lambda_1 \alpha^* + (1 - \lambda_1)\alpha \tag{12}$$

$$\beta' = \lambda_2 \beta^* + (1 - \lambda_2)\beta \tag{13}$$

$$\eta' = \lambda_3 \eta^* + (1 - \lambda_3)\eta \tag{14}$$

其中，α'、β'和η'分别是场景权重与动态权重调节因子生成的复合权重，λ_1、λ_2和λ_3是可学习参数。因此，式(4)可更新为式(15)。

$$\text{T-FC}(\boldsymbol{F}_{\text{fusion}}) = \mathrm{Norm}(\mathrm{Trans}(\mathrm{FC}(\mathrm{Norm}(\alpha'\boldsymbol{F}_V + \boldsymbol{F}_{\text{fusion}}\beta'\boldsymbol{F}_A)))) \tag{15}$$

这种加权方式不仅保留了每个模态的特定信息，还通过调节因子动态和场景感知控制器调节各模态在融合过程中的重要性。

2.3 模态一致性对齐

在多模态学习中，RGB、光流和音频模态的原始数据可能存在时间异步性（如传感器时延或采样率差异）。尽管DCMF模块通过时间维度同步了输入特征，但不同模态的语义特征仍需在融合前对齐。为了解决异步数据的对齐问题，并充分利用多模态数据的同步特性，进一步增强特征的时间一致性，本节设计了MCA模块。

由于CMHA模块中输入特征存在6种模式，因此以时间流对比特征时，将同一时间的所有特征聚合在一起，并拉开与不同时间特征的距离，设计了模态一致性对齐损失（MCA Loss, modality consistency alignment loss）。MCA Loss如式(16)所示。

$$\mathcal{L}_{\text{MCA}} = -\frac{1}{N_F T}\sum_{t=1}^{T}\sum_{a,b,c\in\mathcal{M}}\sum_{d,e,f\in\mathcal{M}} \ln \frac{\exp\left(\frac{\phi\left(\text{T-FC}(\boldsymbol{F}_{\text{fusion}})_{abc}^t, \text{T-FC}(\boldsymbol{F}_{\text{fusion}})_{def}^t\right)}{\tau}\right)}{\sum_{k=1}^{T}\exp\left(\frac{\phi\left(\text{T-FC}(\boldsymbol{F}_{\text{fusion}})_{abc}^t, \text{T-FC}(\boldsymbol{F}_{\text{fusion}})_{def}^k\right)}{\tau}\right)} \tag{16}$$

其中，N_F是模态组合的数量，T是时间步长，$\mathcal{M} = \{\boldsymbol{F}_{VLA}, \boldsymbol{F}_{VAL}, \boldsymbol{F}_{LVA}, \boldsymbol{F}_{LAV}, \boldsymbol{F}_{AVL}, \boldsymbol{F}_{ALV}\}$，$a$、$b$、$c$、$d$、$e$和$f$是遍历$\mathcal{M}$的索引变量，$\phi()$是余弦相似度函数，$\text{T-FC}(\boldsymbol{F}_{\text{fusion}})_{abc}^t$是在时间点$t$处RGB、光流和音频模态按照abc的顺序融合特征，$\text{T-FC}(\boldsymbol{F}_{\text{fusion}})_{def}^t$是在时间点$t$处RGB、光流和音频模态按照def的顺序融合特征，τ是温度超参数。

MCA Loss通过对比学习，强制同一时间步的多模态组合特征在嵌入空间中相近，而不同时间步的多模态组合特征相互远离，对齐不同模态的特征，从而确保融合后的特征在语义上具有一致性，MCA Loss核心是最大化同一时间步内不同模态组合的余弦相似度，同时最小化不同时间步特征的相似度。余弦相似度是衡量特征向量的方向一致性，对幅度不敏感，适合多模态特征的对齐。温度超参数τ是调节相似度分布的相似程度。较小的τ会放大相似样本的差异，使模型更关注困难样本（如异步严重的模态对）。MCA Loss可以隐式时间对齐，即使原始数据存在微小异步，特征空间的对齐仍能保证融合后的语义一致性。同时提升鲁棒性，通过对比学习抑制噪声模态组合的影响。

虽然上述对RGB、光流和音频模态的原始数据实现了特征对齐，但异步数据和噪声干扰仍可能导致跨模态语义精度不够。为此，本节设计了噪声鲁棒性对齐与多尺度时序建模模块，通过噪声感知调节权重，添加下述3个可插拔组件，分别为噪声感知权重调节（NWM, noise-aware weight modulation）、抗噪温度超参数（NRTP, noise-resistant temperature parameter）和多尺度时序对齐（MTA, multi-scale temporal alignment）结构。

1) 噪声感知权重调节。首先对多模态特征$\boldsymbol{F}_V^t$、$\boldsymbol{F}_L^t$和$\boldsymbol{F}_A^t$进行L2归一化消除幅度影响，然后计算各模态特征$\boldsymbol{F}_i^t$的归一化熵值$\boldsymbol{H}_i^t$，最后，通过两层多

层感知机（MLP, multilayer perceptron）将熵值映射为置信度权重$\omega_{\mathrm{VLA}}^{t}$，具体如式(17)和式(18)所示。

$$\boldsymbol{H}_i^t = -\sum_{k=1}^{\frac{d}{3}} p_k \ln p_k,\quad p_k = \frac{\exp\left(\boldsymbol{F}_i^t[k]\right)}{\sum_{j=1}^{\frac{d}{3}} \exp\left(\boldsymbol{F}_i^t[j]\right)} \tag{17}$$

其中，$\boldsymbol{H}_i^t \in \left[0, \ln\left(\frac{d}{3}\right)\right]$为熵值，噪声越大$\boldsymbol{H}_i^t$越高。

$$\omega_{abc}^t = \sigma\left(\mathrm{NWM}\left([\boldsymbol{H}_a^t, \boldsymbol{H}_b^t, \boldsymbol{H}_c^t]\right)\right) \tag{18}$$

其中，ω_{abc}^t通过模态内特征熵估计噪声强度，低置信度组合（如受遮挡的光流）在损失计算中被自动抑制，其计算过程可微且与主干网络联合优化。

2) 抗噪温度超参数。为缓解噪声干扰，将固定温度超参数τ扩展为模态动态相关超参数τ_{abc}，通过低值放大相似样本间的差异，高值平滑相似度分布，增强噪声鲁棒性，根据噪声置信度权重ω_{abc}^t自适应调整，其计算如式(19)所示。

$$\tau_{abc} = \tau_0\left(1 + \varsigma \ln\left(1 + \frac{w_{abc}^t}{\zeta}\right)\right) \tag{19}$$

其中，τ_0为基准温度，ς和ζ分别为缩放系数和平滑系数，分别控制调整幅度与数值稳定性。该设计使模型在高噪声场景下自动增大温度（平滑相似度分布），且在清洁数据中保持敏锐的判别力。

3) 多尺度时序对齐。为捕捉多粒度模态异步，对每个模态特征并行施加窗口大小为{1,3,5}的平均池化，生成多尺度特征$\{\boldsymbol{F}^{t,s}\}_{s=1}^{3}$。扩展后的多尺度对比损失强制模型在局部运动（$s$=1）、短时序（$s$=3）和全局语义（$s$=5）3 个层级上对齐模态特征，其计算如式(20)所示。

$$\mathcal{L}_{\text{multi-scale}} = \sum_{s=1}^{3} \vartheta_s \mathcal{L}_{\mathrm{MCA}}^{s}\left(\boldsymbol{F}_{abc}^{t,s}, \boldsymbol{F}_{def}^{t,s}\right) \tag{20}$$

其中，ϑ_s为尺度权重，每个尺度的对比损失计算与原始MCA Loss相同。

2.4　双曲图注意力机制

本节提出了一种基于双曲图注意力机制的双曲图卷积网络。HGAtt通过引入双曲洛伦兹空间中的注意力机制，有效捕捉输入的图层次结构，从而增强正常与异常表示之间的区分能力。与现有的基于图或基于变换的方法相比，HGAtt在建模复杂层次关系方面具有显著优势。HGAtt主要由双曲空间转换、双曲特征相似性计算、洛伦兹线性变换和特征增强模块组成。

给定MCA最终输出融合特征$\boldsymbol{F}_{\text{T-fusion}}$=T-FC$(\boldsymbol{F}_{\text{fusion}}) \in \mathbb{R}^{T\times D}$，其中$D$表示特征维度，利用指数映射将融合的视听特征转换到双曲空间，学习到双曲特征$\boldsymbol{F}_H \in \mathbb{R}^{T\times D_H}$，其中$D_H$表示双曲空间的维度。然后，采用并行分支结构处理转换后的双曲特征图，以学习正常和异常特征的独特模式。每个分支通过独立的注意力机制捕捉特定特征，从而增强模型对正常和异常数据的区分能力。

首先，构建邻接矩阵$\boldsymbol{A} \in \mathbb{R}^{T\times T}$，用于表示双曲特征之间的相似性。邻接矩阵的每个元素$\boldsymbol{A}_{ij}$通过式(21)计算。

$$\boldsymbol{A}_{ij} = \mathrm{Softmax}\left(\exp\left(-d_L\left(\boldsymbol{F}_{H,i}, \boldsymbol{F}_{H,j}\right)\right)\right) \tag{21}$$

其中，$d_L(\cdot,\cdot)$表示双曲洛伦兹空间中的距离度量，根据计算$\boldsymbol{F}_H$片段i和片段j的洛伦兹距离d_L来评估其相似性，通过指数函数和Softmax激活函数，邻接矩阵的值被限制在(0,1)范围内。为了降低长距离冗余计算，加强双曲空间片段对的相似性，在Softmax归一化之前，进行了阈值操作，如式(22)所示。

$$\exp\left(-d_L\left(\boldsymbol{F}_{H,i}, \boldsymbol{F}_{H,j}\right)\right) = \max\left(\rho, \exp\left(-d_L\left(\boldsymbol{F}_{H,i}, \boldsymbol{F}_{H,j}\right)\right)\right) \tag{22}$$

其中，ρ是阈值。

当式(21)计算邻接矩阵$\boldsymbol{A} \in \mathbb{R}^{T\times T}$时，多数节点对（如相隔较远的视频片段）的关联性较低，计算存在冗余。因此，采用K最近邻（KNN, K-nearest neighbor）算法稀疏化，仅保留每个节点Top-K基于双曲距离的相似边构建双曲空间，将稠密矩阵转化为稀疏矩阵，降低计算和内存开销。

其次，进行洛伦兹线性变换与特征聚合，在每一层l中，对双曲特征$\boldsymbol{F}_H^{l-1}$进行洛伦兹线性变换，并通过邻域聚合操作更新节点特征。

$$\boldsymbol{F}_{H,i}^{l} = \frac{\sum_{j=1}^{T} \boldsymbol{A}_{ij} \Lambda\left(\boldsymbol{F}_{H,i}^{l-1}\right)}{\sqrt{-\varepsilon}\left|\left\|\sum_{k=1}^{T} \boldsymbol{A}_{ik} \Lambda\left(\boldsymbol{F}_{H,i}^{l-1}\right)\right\|_{\mathcal{L}}\right|} \tag{23}$$

其中，$\Lambda()$是洛伦兹线性变换，ε是负曲率常数，用于调整双曲空间的几何特性，$\|\ \|_{\mathcal{L}}$是洛伦兹范数，用于定义双曲空间中的特征向量距离。

最后，在特征增强模块，为了进一步增强双曲特征的表达能力，设计了基于时间信息的特征增强模块。由于洛伦兹线性变换的特性，上述输入特征

向量 $\boldsymbol{F}_H[0]$ 表示双曲空间中的时间特征，因此对于时间维度的特征学习存在一定单一性。对此，为了学习视频中的时间特征关系，直接构建邻接矩阵 $\boldsymbol{B} \in \mathbb{R}^{T\times T}$ 表示其时间特征相似度，该邻接矩阵的每个元素 $\boldsymbol{B}_{ij}$ 通过式(24)计算。

$$\boldsymbol{B}_{ij} = \exp\left(-\frac{\left\| i-j \right\|}{\gamma}\right) \tag{24}$$

其中，γ 是控制时间距离的超参数。随后对其进行双曲度量，通过式(22)和式(23)计算出最终时间特征关系 $\boldsymbol{F}_T$。

2.5 双曲融合与分类

本节利用双曲几何的特性对来自 HGAtt 不同分支的特征进行融合，通过双曲分类器进行异常事件预测。

首先，将来自双曲特征相似性分支和时间关系分支的嵌入表示进行融合，通过双曲空间中的拼接操作将其融合为一个统一表示。

$$\boldsymbol{F}_{\text{sum}} = \boldsymbol{F}_H \oplus \boldsymbol{F}_T \tag{25}$$

这种融合方式不仅保留了双曲空间的几何结构，还能够有效捕捉视频片段之间的长程依赖和时间关系。但由于融合后的嵌入表示仍然保持双曲流形，难以直接进行线性分类，因此，引入双曲分类器来预测异常事件的置信度分数。其定义如式(26)所示。

$$\boldsymbol{F}_{\text{final}} = \sigma\left(\left(\epsilon + \epsilon < \boldsymbol{F}_{\text{sum}}, \omega > _{\mathcal{L}}\right) + b\right) \tag{26}$$

其中，ϵ 表示一个超参数，用于调节双曲分类器的输出范围，$< , > _{\mathcal{L}}$ 表示双曲空间中的洛伦兹内积，分别是双曲分类器的权重矩阵和偏置项。

为在弱监督条件下进行训练，将视频异常检测任务建模为一个多示例学习问题。其中，使用视频中 k-max 预测分数的均值作为视频的异常分数。对于正样本视频（包含异常事件），高分预测更可能对应异常片段。而对于负样本视频（仅包含正常事件），k-max 预测分数通常对应难分样本。基于此，目标函数定义为

$$\mathcal{L}_{\text{MIL}} = \frac{1}{N}\sum_{i=1}^{N} - Y_i \ln\left(\bar{S}\right) \tag{27}$$

其中，$\bar{S}$ 是视频中 k-max 预测分数的均值，Y_i 是视频级别的二值标签（1 表示异常，0 表示正常）。

通过优化该目标函数，能够有效区分正常与异常事件，同时充分利用双曲空间的几何特性来捕捉视频片段之间的复杂关系。

3 实验结果与分析

本节使用公开的视频异常检测数据集 XD-Violence[46] 测试本文方法，进行可行性分析，并对实验结果进行定性和定量分析，以验证本文方法的有效性。此外，为了保证实验结果的泛化性，还在 UCF-Crime 数据集[2] 上对本文方法进行评估。

3.1 数据集

采用 XD-Violence 数据集进行视频异常检测研究。该数据集是目前规模较大、场景多样和标注精细的异常行为检测数据集之一。XD-Violence 数据集包含 4 754 个视频片段，总时长超过 217 小时。视频来源包括电影、电视剧、监控视频和网络视频等，涵盖了室内外多种场景。数据集将视频片段分为正常视频和异常视频两类，其中异常视频包含多种异常行为。XD-Violence 数据集将视频片段划分为训练集和测试集，其中训练集包含 3 954 个视频片段，测试集包含 800 个视频片段。自发布以来，XD-Violence 数据集被广泛应用于视频异常检测领域，为异常检测算法的研究和评估提供了重要的数据基础。

3.2 实验设置

实验在 Ubuntu 22.04.4 操作系统下进行，使用 NVIDIA A40 GPU 进行模型训练和测试。本文方法是在上述数据集上使用基于多实例学习的损失函数，batch size 设置为 128。视觉特征以每秒 24 帧的采样率提取，使用窗口大小为 16 帧的滑动窗口方法。对于听觉数据，将每个音频记录分成 960 ms 的重叠片段，然后使用 96×64 的分辨率在训练过程中计算。在双曲空间中，利用 2 个双曲图卷积层来实现双分支机构，维度被设置为 32。负曲率常数 ε 设置为−1。在训练过程中，采用了随机梯度下降，初始学习率为 5×10^{-4}，使用余弦退火学习率调度更新，并训练了 300 个 Epoch，λ_1、λ_2 和 λ_3 初始设置为 0.5，γ 设置为 1，ϵ 设置为 2，τ_0 设置为 0.5，ς 设置为 0.5，ζ 设置为 0.1，ϑ_1 设置为 0.3，ϑ_3 设置为 0.5，ϑ_5 设置为 0.2，Dropout 设置为 0.1。

3.3 实验对比

CM-HVAD 使用多模态特征融合方式，通过有效整合时空特征与运动信息，增强了模型对异常事件的表征能力，从而显著提升检测精度。

表 1 展示了本文方法与其他视频异常检测方

法[2,27,43,47-54]在 XD-Violence 数据集 AP 指标上的性能对比。从表 1 可以看出，本文方法在 AP 指标上达到了 86.47%，显著优于其他方法。相较于文献[2]，提升了 10.79%；相比文献[47]和文献[48]，分别提升了 10.57% 和 8.66%；相比文献[53]，提升了 6.94%；相比文献[49]和文献[50]，分别提升了 5.04% 和 4.7%；相比文献[27]、文献[43]、文献[51]、文献[52]和文献[54]，分别提升了 4.08%、11.02%、24.46%、7.86% 和 4.37%。

表 1　不同检测方法在 XD-Violence 数据集的实验结果对比

年份	方法	AP
2018 年	文献[2]	75.68%
2021 年	文献[47]	75.90%
	文献[48]	77.81%
2022 年	文献[53]	79.53%
2023 年	文献[49]	81.43%
	文献[50]	81.77%
2024 年	文献[27]	82.39%
	文献[43]	75.45%
	文献[51]	62.01%
	文献[52]	78.61%
	文献[54]	82.10%
2025 年	本文方法	**86.47%**

在 XD-Violence 数据集的实验结果表明，与其他方法对比，本文方法在 AP 指标的检测上存在一定优势，能够对该数据集进行更有效的视频异常检测。

除 AP 指标外，在视频异常检测框架中，异常分数曲线的形态特征直接反映了模型对时序异常事件的敏感程度。图 4 展示了本文方法在 XD-Violence 数据集异常评分上的部分实例异常分数，其中灰色代表 Ground Truth 标记为异常。如图 4(b)所示，曲线在正常帧区间（t_1, t_2）呈现平稳态势，偶有振荡出现，其振幅范围较低，位于（0, 0.2）范围内，表明特征提取网络对常规模式具有稳定的表征能力。当发生异常事件时（视频开始与 t_3 时刻），其上升沿斜率 k 陡然增高，曲线出现显著峰值，长时间达到 1。在异常事件较少的视频中，如图 4(c)所示，在（t_1, t_2）、（t_3, t_4）和（t_5, t_6）时间轴内，曲线表明能够很好地识别异常状况，而在其他正常帧内，也保持良好的平稳曲线。在图 4(d)中，存在部分异常检测区间，曲线后期的振荡衰减（t_1, t_2）和（t_3, t_4）揭示了本文方法对场景恢复的适应过程。为更进一步表明正常与异常的区别，本文截取了部分正常与异常视频帧展示于图 4 中，从图中可以看出，当异常分数较高甚至接近 1 时，如图 4(b)所示，明显出现异常状况，而当异常分数较低甚至接近 0 时，如图 4(c)所示，事件依旧处于正常区间。图 4 异常分数展示出本文方法在 XD-Violence 数据

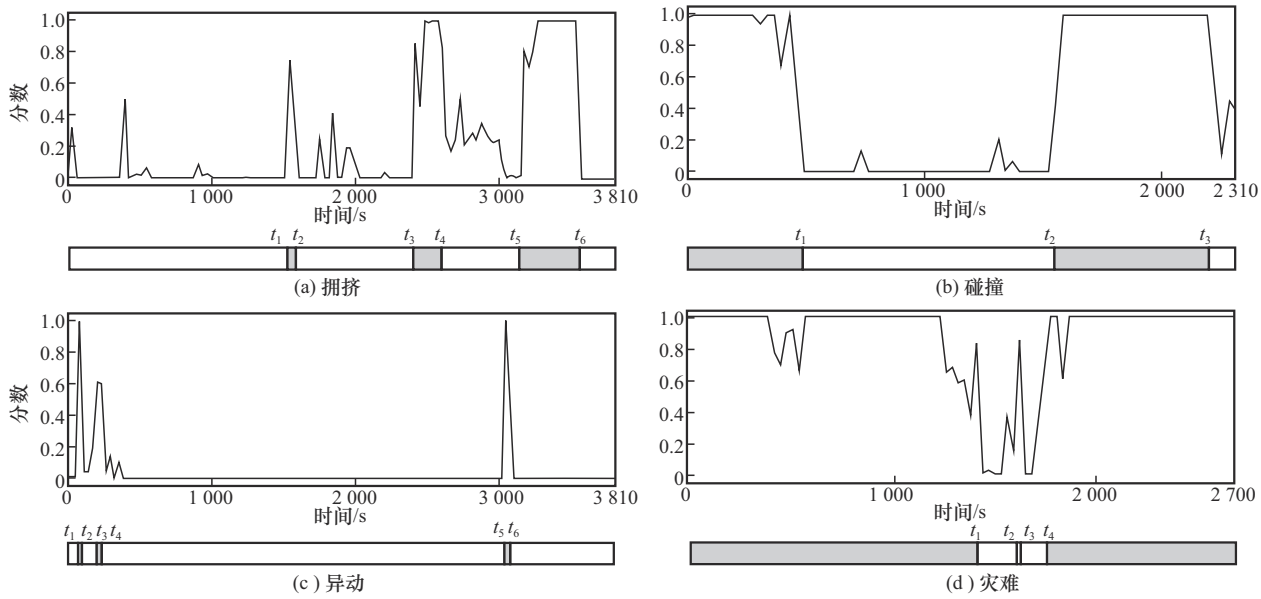

图 4　在 XD-Violence 数据集异常评分曲线的视觉比较

集上的异常曲线，进一步证明了视频异常检测的有效性和精确性。

此外，本文还绘制了AUC，如图5所示。当Epoch=1~5时，AUC从79.56%快速提升至90.37%，相对增幅达13.6%，表明模型在早期迭代中迅速学习到有效的特征表示。当Epoch=6~30时，AUC平均速率稳步增长，期间出现几次局部波动（如Epoch=16时下降），可能源于学习率未及时调整或训练数据中的噪声样本被重新加权。当Epoch=30~50时，AUC进入平稳，符合模型收敛的预期行为。第12个Epoch（AUC=94.72%）首次超过94%阈值，此时可考虑启用早停机制。在架构层面，最后10个Epoch速率提升较小，说明当前方法可能接近上限。

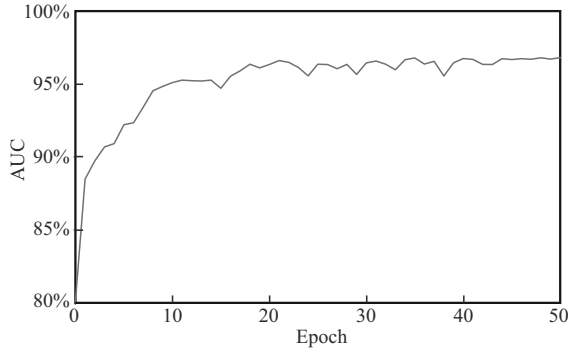

图5　在XD-Violence数据集上AUC指标验证

综上所述，本文方法在XD-Violence数据集上具有显著优势。

3.4　消融研究

本节首先对DCMF、MCA和HGAtt模块进行了消融研究，如表2所示。由表2可知，DCMF作为基础特征融合模块，仅使用表明单纯动态融合虽能平衡模态权重，但缺乏时序对齐和层次建模能力。单独使用MCA略低于单独使用DCMF，证明时序对齐对模态异步问题存在部分改善效果。单独使用HGAtt显著高于前两者，验证了双曲空间在表征异常层级关系上的优势。融合DCMF与HGAtt模块，较单模块HGAtt提升了2.22%，说明动态融合能优化双曲空间的输入特征质量（如场景中异常碰撞声与视觉动作的匹配度提升）。融合DCMF与MCA模块，表明时序对齐与动态融合的协同效应存在优势。DCMF、MCA和HGAtt三模块联合使用较最优双模块组合（DCMF+MCA）再提升1.75%，因此证明，HGAtt补充了前两者缺失的层次推理能力。此外，DCMF+MCA性能远超DCMF+HGAtt，说明时序对齐是多模态融合的基础前提。增加HGAtt带来的AP提升从单模块（79.94%→82.16%）到全模块（84.72%→86.47%）逐渐减小，符合模型性能收敛特性，证明了DCMF、MCA和HGAtt组合的有效性。

表2　对模块DCMF、MCA和HGAtt的消融研究

DCMF	MCA	HGAtt	AP
√	×	×	77.80%
√	×	√	82.16%
√	√	×	84.72%
×	√	×	77.31%
×	×	√	79.94%
√	√	√	**86.47%**

对模块DCMF、MCA和HGAtt的消融研究表明了各模块的不可或缺，其AP指标展示了相应模块对本文方法的不同贡献与组合提升，为验证各模块的有效性，本文采用增量贡献法计算各模块对AP的边际贡献，具体结果如图6所示。

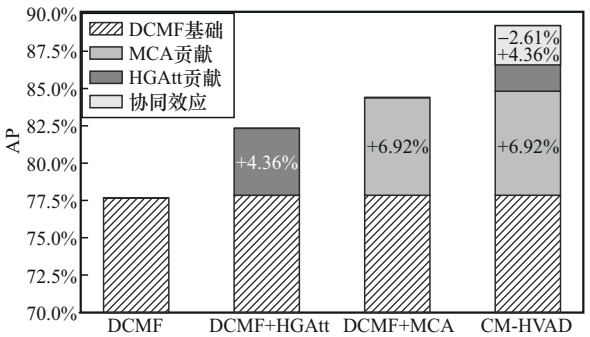

图6　DCMF、MCA和HGAtt模块贡献度研究

计算方式如下，已知DCMF模块的基础值为77.80%，HGAtt模块增量为82.16%，MCA模块的增量为84.72%。计算可知，DCMF与HGAtt模块的协同效应为（DCMF+HGAtt）−77.80%=+4.36%，DCMF与MCA模块的协同效应为（DCMF+MCA）−77.80%=+6.92%，而三部分模块总协同效应为86.47%，三模块间的负协同效应为CM-HVAD−（77.80%+6.92%+4.36%）=−2.61%。

从图6中计算可知，在以DCMF为基线时，相比于HGAtt的4.36%，MCA模块的贡献显然更大（+6.92%），证明了时序对齐对多模态融合的关键

作用。而 HGAtt 与 DCMF 模块之间的增幅相对较小。此外，负协同效应表明，三模块联合时存在轻微性能抵消问题。指标结果表明，DCMF 高于单独 MCA 或 HGAtt，说明动态融合是本文方法的核心，时序对齐在增幅动态融合上格外重要，而双曲图注意力机制对于提升性能不可或缺。

此外，为验证多模态对本文方法有效性，分别对 Video+Audio、Flow+Audio、Video+Flow 和 Video+Flow+Audio 这 4 种类型多模态分别在 Epoch=10、30 和 50 的情况下进行消融研究，如表 3 所示。当 Epoch=50 时，Video+Audio 模态 AP 达到了 83.48%，仅次于 Video+Flow+Audio 的 86.47%，比 Flow+Audio 和 Video+Flow 分别高 7.85% 和 3.19%。模态 Flow+Audio 性能相对最低，50 轮 AP 为 75.63%，表明单独依赖光流难以捕捉静态异常特征。模态 Video+Flow 相对 Flow+Audio 带来 5.66% 的 AP 提升，证明时空信息对视频异常检测场景具有关键作用。Video+Audio 的组合对于本文方法实验结果提升最大，证明视频信息与空间信息的组合对于补全场景异常检测有相当提升空间。

表 3　在 XD-Violence 数据集的多模态研究 AP

Epoch	Video+Audio	Flow+Audio	Video+Flow	Video+Flow+Audio
10	76.53%	70.27%	74.95%	78.32%
30	82.11%	75.46%	79.61%	85.30%
50	83.48%	75.63%	80.29%	**86.47%**

其中，CMHA 对三模态和两模态的提取机制不同是导致模态 Video+Audio、Flow+Audio 和 Video+Flow 低于 Video+Flow+Audio 的重要因素。结果表明，三模态融合是本文方法的最优组合。

3.5　多模态自适应优化研究

在 DCMF 模块中，本文设计了四级调节机制实现动态优化。首先，多模态动态权重调节因子（式(5)~式(7)）通过可学习的 $\{\omega,b\}$ 参数组，针对 RGB($\boldsymbol{F}_V$)、光流($\boldsymbol{F}_L$)和音频($\boldsymbol{F}_A$)分别生成[0,1]范围的权重；其次，SAG 通过双向 LSTM 生成可学习的场景权重，并与动态权重调节因子结合生成复合权重（式(9)~式(14)）；然后，跨模态多头注意力机制（式(8)）通过 Softmax 激活函数动态计算模态间交互强度；最后，T-FC 模块（式(15)）的 Transformer 层自动学习场景相关的特征变换模式。

为验证 DCMF 参数自动调整能力，本节设计了场景−权重关联实验。α'、η'和β'分别是 DCMF 生成的多模态复合权重，为直观展示不同场景类型间的参数差异度，将三者之和进行归一化。本节在 XD-Violence 数据集进行场景−权重关联实验。动态权重机制导致不同帧测试的权重占比结果均不同，因此，本文在不同场景中分别选取单帧视频进行测试，从而获得该图片的动态权重因子的权重占比实验结果。

表 4 展示了在不同场景类型下，多模态复合权重 α'、η'和β'的权重贡献度变化。在视觉显著场景下（如异常燃烧），RGB 模态的融合权重因子 α' 占比为 0.61，在特征融合中占主导地位。在运动主导场景下（如异常碰撞），光流模态的融合权重因子 η' 占比为 0.58，而在声音敏感场景下（如气体急剧膨胀异常音频），音频模态的融合权重因子 β' 会获得更高权重占比结果。

表 4　在 XD-Violence 数据集场景−权重占比关联实验

场景类型	α'	η'	β'
视觉显著	**0.61**	0.23	0.16
运动主导	0.30	**0.58**	0.12
声音敏感	0.28	0.19	**0.53**

为有效证明 DCMF 模块中四级调节机制的参数动态调整能力，本节对动态调节因子进行了消融实验。在消融研究中，表 3 展示了多模态复合权重 α'、η'和β'分别为 0 时的 AP 值，验证了不同模态对实验结果的不同影响，进一步，为研究不同比重的静态权重占比对实验结果的影响，本节设定了针对不同模态的静态权重占比消融实验，通过设置其中特定模态的权重比例为定值（0.3、0.5、0.8），从而对 XD-Violence 数据集进行实验，验证四级调节机制的参数动态调整能力，具体结果如表 5 所示。

表 5 展示了动态调节因子中静态权重占比对模型性能影响的系统性消融实验结果。通过控制不同模态的静态权重占比（0.3、0.5、0.8），本文定量评估了四级调节机制中参数动态调整的有效性。实验结果表明，各模态对最终性能的贡献度存在显著差异：当光流模态静态权重设为 0.3 时，模型取得最优 AP 值 84.61%；当音频模态权重提升至 0.8 时，性能急剧下降至 55.48%，降幅达 29.13%。值得注

意的是，静态权重的增加普遍导致性能衰减，但当视频模态权重设为0.5时，模型仍能维持83.15%的较高精度。这些发现验证了动态调节机制的必要性，并揭示不同模态特征对最终预测结果的差异化敏感性，为多模态融合中的参数优化提供了重要实证依据。

表5　动态调节因子静态权重占比消融实验

静态权重占比	α'	η'	β'	AP
0.3	√			78.52%
		√		**84.61%**
			√	81.97%
0.5	√			**83.15%**
		√		77.81%
			√	69.14%
0.8	√			**81.09%**
		√		70.36%
			√	55.48%

为系统评估SAG模块在多模态场景中的动态调节能力，本文实验在XD-Violence数据集的场景子集（视觉显著、运动主导和声音敏感）及全数据集上进行了对比分析，如表6所示。由表6可知，相较于基线方法CM-HVAD，引入SAG模块后系统表现出显著的场景适应性优化。

表6　SAG模块在XD-Violence数据集的不同场景对比实验

方法	视觉显著	运动主导	声音敏感	全数据集
CM-HVAD	**92.15%**	88.27%	86.52%	86.47%
+SAG	91.83%	89.04%	88.31%	86.69%

SAG模块在运动主导和声音敏感场景中分别取得89.04%和88.31%的AP值，较基线方法提升0.77%和1.79%，表明其对非视觉模态的特征动态融合具有增强作用。值得注意的是，声音敏感场景的改善幅度最大（+1.79%），验证了SAG模块在音频特征权重分配上的有效性。在视觉显著场景中，SAG模块的AP值（91.83%）与基线方法（92.15%）保持相近（仅降低0.32%），说明模块在优先依赖视觉模态时未引入显著干扰，体现了权重调节的鲁棒性。在全数据集场景下，SAG以86.69%的AP值超越基线方法（86.47%），综合提升0.22%，证实其在不同场景混合的实际应用中仍保持性能增益。

SAG模块通过动态权重调节，在运动/声音主导场景中表现优于静态融合策略，且不损害视觉模态的原始性能。这一特性使其更适应多模态数据分布的异质性，为复杂场景下的视频异常检测提供了更优的融合方案。

3.6　语义对齐优化研究

为验证MCA模块面向噪声的准确度与鲁棒性，在XD-Violence数据集分别加视觉噪声、听觉噪声或多模态噪声，评估了噪声感知权重调节（NWM）、抗噪温度参数（NRTP）和多尺度时序对齐结构（MTA）3个模块在噪声条件下的性能表现。其中，视觉噪声为空间噪声（模拟局部遮挡）和时序噪声（模拟帧率不稳定），听觉噪声为频带噪声（模拟环境声干扰），多模态噪声为高斯噪声（空间噪声+时序噪声+频带噪声），实验结果如表7~表9所示。

表7　对NWM、NRTP和MTA的视觉噪声消融实验

噪声类型	NWM	NRTP	MTA	AP
空间噪声	×	×	×	71.37%
	√	×	×	77.64%
	√	×	√	78.09%
	√	√	×	79.82%
	×	×	√	71.75%
	√	√	√	**80.44%**
时序噪声	×	×	×	74.61%
	√	×	×	76.39%
	√	×	√	81.25%
	√	√	×	77.94%
	×	×	√	80.83%
	√	√	√	**81.67%**

在表7中，空间噪声为正态分布的空间独立噪声。该噪声作用于视频模态，随机施加于视频帧上半部分，噪声标准差设为0.5，模拟真实场景中的局部遮挡现象。时序噪声采用随机丢弃帧策略，丢弃率设为0.3，模拟现实场景中传感器的间歇性失效现象。实验结果表明，在仅施加空间噪声时，模型AP从71.37%提升至77.64%，当同时启用NWM和MTA模块时，AP进一步提升至78.09%。值得注意的是，当3种模块共同作用时，模型在空间噪声

条件下达到最优性能80.44%。在时序噪声实验中，完整框架（NWM+NRTP+MTA）展现出最强的鲁棒性，取得81.67%的AP值，显著优于74.61%。

表8　对NWM、NRTP和MTA的听觉噪声消融实验

噪声类型	NWM	NRTP	MTA	AP
频带噪声	×	×	×	76.34%
	√	×	×	81.30%
	√	×	√	81.79%
	√	√	×	83.01%
	×	×	√	77.26%
	√	√	√	**83.32%**

表9　对NWM、NRTP和MTA的多模态噪声消融实验

噪声类型	NWM	NRTP	MTA	AP
空间噪声 + 时序噪声+ 频带噪声	×	×	×	67.18%
	√	×	×	74.93%
	√	×	√	78.42%
	√	√	×	75.29%
	×	×	√	73.71%
	√	√	√	**79.26%**

在表8中，频带噪声为频谱受限的基于标准正态分布的加性白高斯噪声，通过频段掩码施加于低频（模拟背景环境噪声）、中频（模拟人声/音乐干扰）和高频（模拟电子设备噪声），噪声强度由信噪比（SNR, signal-to-noise ratio）（SNR=10 dB）控制。实验结果表明，单独引入NWM模块后AP由76.34%显著提升至81.30%，而NWM与MTA模块的协同作用时AP可进一步提升至81.79%。值得注意的是，当NWM与NRTP模块共同作用时，模型表现出更强的噪声鲁棒性，AP达到83.01%。特别地，3种模块联合使用时取得最优性能83.32%。

在表9中，实验采用多模态复合噪声干扰模式，包括空间噪声、时序噪声和频带噪声，以模拟真实场景中的复杂干扰环境。实验结果表明，在未启用任何抗噪模块时，模型在复合噪声条件下的基线性能仅为67.18% AP，单独引入NWM模块后，性能显著提升至74.93% AP，而NWM与MTA模块的协同作用可进一步提升至78.42% AP。特别值得注意的是，当3种模块共同作用时，模型展现出最强的抗干扰能力，达到79.26% AP的最高性能，相比基线提升幅度达12.08%。这一系列实验结果不仅验证了各模块在应对多模态复合噪声时的有效性，更揭示了模块间的协同增强效应，为复杂噪声环境下的多模态学习提供了重要的方法参考和理论依据。

3.7　特征可视化分析

为验证本文方法的特征表征能力，本节进行特征可视化分析，如图7所示。其中，图7(a)是特征输入时，提取部分正常与异常特征进行可视化，图7(b)是输入在双曲网络中的部分正常和异常特征，图7(c)是在双曲网络中针对输入的部分正常和异常特征进行双曲空间变化后的特征模式。在特征输入时，如图7(a)所示，正常与异常特征混合在一起，其特征难以相互区别，无法区分正常或异常事件。当经过DCMF与CMA，以及双曲空间变化后，如图7(b)所示，特征已经明显出现特征聚集现象，正常和异常事件已经可以进行区分，但其部分边界依旧存在模糊现象。在进行HGAtt后，如图7(c)所示，正常事件与异常事件已经能简单区分，其边界趋于明显。特征可视化分析表明，双曲空间的几何特性拉大了异常特征与正常特征的类间距离，证明本文方法使用双曲空间更适合建模层次化表征，有效提升检测精度。

3.8　泛化性分析

本节采用UCF-Crime数据集对本文方法进行泛化性分析，该数据集是视频异常检测领域的基准数据集之一，由University of Central Florida于2018年

(a) 输入特征可视化　　(b) 中间特征可视化　　(c) 输出特征可视化

图7　特征可视化分析

构建，包含1 900个未裁剪的长时监控视频，涵盖真实场景下的13类异常事件，包括抢劫、打架、交通事故、纵火等，其中正常视频采集自监控公共场所，异常视频包含人工标注的时空边界，覆盖室内外多种光照条件及摄像机视角，包含真实监控中的自然噪声（如运动模糊、低分辨率），增强模型鲁棒性，有效衡量模型在类别不平衡下的性能。

为了保证本文方法的泛化性，本节在UCF-Crime数据集进行评估。

表10展示了本文方法与其他视频异常检测方法在UCF-Crime数据集上AUC指标的性能对比。从表10可以看出，本文方法的AUC指标达到了87.12%，显著优于其他方法。相较于文献[2]，提升了9.20%；相比文献[47]和文献[48]，分别提升了2.23%和2.82%；相比文献[53]，提升了4.68%；相比文献[49]和文献[50]，分别提升了0.90%和0.15%；相比文献[43]、文献[51]、文献[52]和文献[54]，分别提升了1.65%、6.84%、3.75%和0.41%。实验结果表明，与其他先进方法对比，本文方法在UCF-Crime数据集上的AUC指标具有一定的优势，能够对该数据集进行有效的视频异常检测，证明了本文方法的泛化性。

表10 不同检测算法在UCF-Crime数据集的实验结果对比

年份	方法	AUC
2018年	文献[2]	77.92%
2021年	文献[47]	84.89%
	文献[48]	84.30%
2022年	文献[53]	82.44%
2023年	文献[49]	86.22%
	文献[50]	86.97%
2024年	文献[43]	85.47%
	文献[51]	80.28%
	文献[52]	83.37%
	文献[54]	86.71%
2025年	本文方法	**87.12%**

4 结束语

本文针对视频异常检测中模态信息不平衡、视听噪声不平均以及模态异步等问题，提出了一种基于动态跨模态融合模块与双曲图注意力机制融合的多模态视频异常检测方法CM-HVAD。通过动态跨模态融合模块DCMF自主学习跨模态权重，动态平衡视觉特征和音视频特征并进行融合增强，有效解决了模态信息不平衡问题。同时，模态一致性对齐模块MCA按时间帧序列对齐模态语义，缓解了模态异步问题。此外，引入双曲图注意力机制HGAtt，利用双曲空间的模式分离特性，有效捕捉了正常和异常表示之间的层次关系，进一步提高了检测准确率。在公开数据集XD-Violence上的实验结果表明，本文方法的检测精度AP值达到了86.47%，其性能明显优于对比方法，验证了本文方法的有效性。此外，还在UCF-Crime数据集上AUC指标达到了87.12%，验证了本文方法的泛化性。

未来工作将重点关注以下几个方面：1) 探索更高效的跨模态融合策略，进一步提升模型对多模态信息的利用效率；2) 研究更鲁棒的模态对齐方法，以应对更复杂的实际应用场景；3) 将本文方法应用于更广泛的视频异常检测任务，如群体异常检测和跨场景异常检测等。

参考文献：

[1] NAYAK R, PATI U C, DAS S K. A comprehensive review on deep learning-based methods for video anomaly detection[J]. Image and Vision Computing, 2021, 106: 104078.

[2] SULTANI W, CHEN C, SHAH M. Real-world anomaly detection in surveillance videos[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6479-6488.

[3] 何平, 李刚, 李慧斌. 基于深度学习的视频异常检测方法综述[J]. 计算机工程与科学, 2022, 44(9): 1620-1629.
HE P, LI G, LI H B. A survey on deep learning based video anomaly detection[J]. Computer Engineering & Science, 2022, 44(9): 1620-1629.

[4] NAKAHATA M T, THOMAZ L A, SILVA A F D, et al. Anomaly detection with a moving camera using spatio-temporal codebooks[J]. Multidimensional Systems and Signal Processing, 2018, 29(3): 1025-1054.

[5] WEI D L, LIU Y, ZHU X G, et al. MSAF: multimodal supervise-attention enhanced fusion for video anomaly detection[J]. IEEE Signal Processing Letters, 2022, 29: 2178-2182.

[6] FLABOREA A, COLLORONE L, MELENDUGNO G M D A D, et al. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2023: 10284-10295.

[7] PAULRAJ S, VAIRAVASUNDARAM S. M^{2}VAD: multiview multimodality transformer-based weakly supervised video anomaly detection[J].

Image and Vision Computing, 2024, 149: 105139.

[8] WU P, LIU J, HE X T, et al. Toward video anomaly retrieval from video anomaly detection: new benchmarks and model[J]. IEEE Transactions on Image Processing, 2024, 33: 2213-2225.

[9] FENG C, CHEN Z Y, OWENS A. Self-supervised video forensics by audio-visual anomaly detection[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 10491-10503.

[10] DEV P P, HAZARI R, DAS P. MCANet: multimodal caption aware training-free video anomaly detection via large language model[C]//International Conference on Pattern Recognition. Berlin: Springer, 2024: 362-379.

[11] 仇媛, 常相茂, 仇倩, 等. 基于长短期记忆网络和滑动窗口的流数据异常检测方法[J]. 计算机应用, 2020, 40(5): 1335-1339.
QIU Y, CHANG X M, QIU Q, et al. Stream data anomaly detection method based on long short-term memory network and sliding window[J]. Journal of Computer Applications, 2020, 40(5): 1335-1339.

[12] 朱张莉, 饶元, 吴渊, 等. 注意力机制在深度学习中的研究进展[J]. 中文信息学报, 2019, 33(6): 1-11.
ZHU Z L, RAO Y, WU Y, et al. Research progress of attention mechanism in deep learning[J]. Journal of Chinese Information Processing, 2019, 33(6): 1-11.

[13] ASAD M, YANG J, HE J, et al. Multi-frame feature-fusion-based model for violence detection[J]. The Visual Computer, 2021, 37(6): 1415-1431.

[14] KUPPUSAMY P, HARIKA C. Human action recognition using CNN and LSTM-RNN with attention model[J]. International Journal of Innovative Technology and Exploring Engineering, 2019, 8: 1639-1643.

[15] 陈佛计, 朱枫, 吴清潇, 等. 生成对抗网络及其在图像生成中的应用研究综述[J]. 计算机学报, 2021, 44(2): 347-369.
CHEN F J, ZHU F, WU Q X, et al. A survey about image generation with generative adversarial nets[J]. Chinese Journal of Computers, 2021, 44(2): 347-369.

[16] 张文林, 刘雪鹏, 牛铜, 等. 基于正样本对比与掩蔽重建的自监督语音表示学习[J]. 通信学报, 2022, 43(7): 163-171.
ZHANG W L, LIU X P, NIU T, et al. Self-supervised speech representation learning based on positive sample comparison and masking reconstruction[J]. Journal on Communications, 2022, 43(7): 163-171.

[17] WU J C, HSIEH H Y, CHEN D J, et al. Self-supervised sparse representation for video anomaly detection[C]//European Conference on Computer Vision. Berlin: Springer, 2022: 729-745.

[18] LIU R K, LIU W M, DUAN M F, et al. MemFormer: a memory based unified model for anomaly detection on metro railway tracks[J]. Expert Systems with Applications, 2024, 237: 121509.

[19] 杨静, 吴成茂, 周流平. 基于全局-局部自注意力网络的视频异常检测方法[J]. 通信学报, 2023, 44(8): 241-250.
YANG J, WU C M, ZHOU L P. Novel video anomaly detection method based on global-local self-attention network[J]. Journal on Communications, 2023, 44(8): 241-250.

[20] PEIXOTO B M, LAVI B, DIAS Z, et al. Harnessing high-level concepts, visual, and auditory features for violence detection in videos[J]. Journal of Visual Communication and Image Representation, 2021, 78: 103174.

[21] ULLAH W, HUSSAIN T, KHAN Z A, et al. Intelligent dual stream CNN and echo state network for anomaly detection[J]. Knowledge-Based Systems, 2022, 253: 109456.

[22] QASIM M, VERDU E. Video anomaly detection system using deep convolutional and recurrent models[J]. Results in Engineering, 2023, 18: 101026.

[23] KUMARI P, BEDI A K, SAINI M. Multimedia datasets for anomaly detection: a review[J]. Multimedia Tools and Applications, 2024, 83(19): 56785-56835.

[24] PHAM L, NGUYEN T, LAM P, et al. Toolchain for comprehensive audio/video analysis using deep learning based multimodal approach: use case of riot or violent context detection[C]//Proceedings of the 2024 International Conference on Content-Based Multimedia Indexing (CBMI). Piscataway: IEEE Press, 2024: 1-4.

[25] PU Y J, WU X Y, WANG S J, et al. Semantic multimodal violence detection based on local-to-global embedding[J]. Neurocomputing, 2022, 514: 148-161.

[26] JAAFAR N, LACHIRI Z. Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance[J]. Expert Systems with Applications, 2023, 211: 118523.

[27] WU Y L, MAO Z Y, YU C Y, et al. Enhancing weakly supervised anomaly detection in surveillance videos: the CLIP-augmented bimodal memory enhanced network[C]//Proceedings of the 2024 18th International Conference on Control, Automation, Robotics and Vision (ICARCV). Piscataway: IEEE Press, 2024: 756-762.

[28] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[J]. arXiv Preprint, arXiv: 1710.10903, 2017.

[29] CORSO G, STARK H, JEGELKA S, et al. Graph neural networks[J]. Nature Reviews Methods Primers, 2024, 4: 17.

[30] ZHU H L, QIAO K X, XU Z G. Video anomaly behavior detection method based on attention-enhanced graph convolution and normalizing flows[J]. Signal, Image and Video Processing, 2025, 19(5): 352.

[31] SONG W F, LI S, CHANG T, et al. Dynamic attention augmented graph network for video accident anticipation[J]. Pattern Recognition, 2024, 147: 110071.

[32] CHIRANJEEVI V R, MALATHI D. Anomaly graph: leveraging dynamic graph convolutional networks for enhanced video anomaly detection in surveillance and security applications[J]. Neural Computing and Applications, 2024, 36(20): 12011-12028.

[33] REYNOLDS D A. Gaussian mixture models[J]. Encyclopedia of Biometrics, 2009, 741(3): 659-663.

[34] 赵仲秋, 季海峰, 高隽, 等. 基于稀疏编码多尺度空间潜在语义分析的图像分类[J]. 计算机学报, 2014, 37(6): 1251-1260.
ZHAO Z Q, JI H F, GAO J, et al. Sparse coding based multi-scale spatial latent semantic analysis for image classification[J]. Chinese Journal of Computers, 2014, 37(6): 1251-1260.

[35] LIU W, LUO W X, LIAN D Z, et al. Future frame prediction for anomaly detection-a new baseline[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6536-6545.

[36] LI X D, LANG Y N, CHEN Y F, et al. Sharp multiple instance learning for DeepFake video detection[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 1864-1872.

[37] 贾香恩, 董一鸿, 朱锋, 等. 异构图卷积网络研究进展[J]. 计算机工程

与应用, 2021, 57(9): 36-49.
JIA X E, DONG Y H, ZHU F, et al. Research progress of heterogeneous graph convolutional networks[J]. Computer Engineering and Applications, 2021, 57(9): 36-49.

[38] LUO W X, LIU W, GAO S H. Normal graph: spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection[J]. Neurocomputing, 2021, 444: 332-337.

[39] 肖进胜, 申梦瑶, 江明俊, 等. 融合包注意力机制的监控视频异常行为检测[J]. 自动化学报, 2022, 48(12): 2951-2959.
XIAO J S, SHEN M Y, JIANG M J, et al. Abnormal behavior detection algorithm with video-bag attention mechanism in surveillance video[J]. Acta Automatica Sinica, 2022, 48(12): 2951-2959.

[40] REHMAN A U, ULLAH H S, FAROOQ H, et al. Multi-modal anomaly detection by using audio and visual cues[J]. IEEE Access, 2021, 9: 30587-30603.

[41] KUMARI P, SAINI M. An adaptive framework for anomaly detection in time-series audio-visual data[J]. IEEE Access, 2022, 10: 36188-36199.

[42] GHADIYA A, KAR P, CHUDASAMA V, et al. Cross-modal fusion and attention mechanism for weakly supervised video anomaly detection[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2024: 1965-1974.

[43] ALMARRI S, ZAHEER M Z, NANDAKUMAR K. A multi-head approach with shuffled segments for weakly-supervised video anomaly detection[C]//Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). Piscataway: IEEE Press, 2024: 132-142.

[44] ZHOU H L, HE L F, CHEN B Y, et al. Multi-modal diagnosis of Alzheimer's disease using interpretable graph convolutional networks[J]. IEEE Transactions on Medical Imaging, 2025, 44(1): 142-153.

[45] JIA X G, JIANG M, DONG Y H, et al. Multimodal heterogeneous graph attention network[J]. Neural Computing and Applications, 2023, 35(4): 3357-3372.

[46] WU P, LIU J, SHI Y J, et al. Not only look, but also listen: learning multimodal violence detection under weak supervision[C]//European Conference on Computer Vision. Berlin: Springer, 2020: 322-339.

[47] WU P, LIU J. Learning causal temporal relation and feature discrimination for anomaly detection[J]. IEEE Transactions on Image Processing, 2021, 30: 3513-3527.

[48] TIAN Y, PANG G S, CHEN Y H, et al. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 4955-4966.

[49] ZHANG C, LI G R, QI Y K, et al. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 16271-16280.

[50] ZHOU H, YU J Q, YANG W. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(3): 3769-3777.

[51] ZANELLA L, MENAPACE W, MANCINI M, et al. Harnessing large language models for training-free video anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2024: 18527-18536.

[52] JAIN Y, DABOUEI A, XU M. Cross-domain learning for video anomaly detection with limited supervision[C]//European Conference on Computer Vision. Berlin: Springer, 2024: 468-484.

[53] WU P, LIU X T, LIU J. Weakly supervised audio-visual violence detection[J]. IEEE Transactions on Multimedia, 2022, 25: 1674-1685.

[54] TAN W J, YAO Q, LIU J F. Overlooked video classification in weakly supervised video anomaly detection[C]//Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). Piscataway: IEEE Press, 2024: 212-220.

[作者简介]

姜迪（1997– ），男，山东济宁人，新疆大学博士生，主要研究方向为视频异常检测、深度学习、目标检测等。

赖惠成（1963– ），男，四川德阳人，新疆大学教授、博士生导师，主要研究方向为视频/图像信息处理、图像理解与识别等。

汪烈军（1975– ），男，四川眉山人，博士，新疆大学教授、博士生导师，主要研究方向为视频通信处理、图像识别与处理等。