中图法分类号： **文献标识码：** A **文章编号：** 1006-8961(XXXX)XX-0001-14

论文引用格式： Fu Qi, Zhu Tao, Liu Dexi, Cao Chonghua, Kuang Sirui, Huang Junting. TFC-Net: An Efficient Lightweight Time-Frequency Collaborative Method for Video Anomaly Detection[J/OL]. Journal of Image and Graphics, XXXX:1-14. DOI: 10.11834/jig.260488. (付淇, 朱涛, 刘德喜, 曹重华, 况思睿, 黄筠婷. 高效轻量化时频协同视频异常检测方法[J/OL]. 中国图象图形学报,XXXX:1-14. DOI: 10.11834/jig.260488.)
[DOI:10.11834/jig.260488]

高效轻量化时频协同视频异常检测方法

付淇[1,2*],朱涛[3],刘德喜[2,3],曹重华[3],况思睿[3],黄筠婷[3]

1. 豫章师范学院数学与计算机学院,南昌 330103; 2. 数据与知识工程江西省高校重点实验室(江西财经大学),南昌 330032; 3. 江西财经大学,南昌 330013

摘　要：目的　针对弱监督视频异常检测中深层时域聚合易平滑高频瞬态特征、削弱细粒度异常感知的问题,本文构建轻量化检测网络,并同时提高识别精度与推理效率。**方法**　提出轻量化时频协同网络(time-frequency collaborative network, TFC-Net)。TFC-Net采用多尺度时域分支建模宏观语义;频域补偿分支先将时序信号经快速傅里叶变换映射至频域,再通过复频谱交互抑制冗余噪声、增强微观动态表征,最后经逆傅里叶变换重构回时域,从而保留异常片段的时间定位能力;TFC-Net基于视频级多示例学习框架,无需复杂目标跟踪模块与高开销视觉语言模型。**结果**　TFC-Net在UCF-Crime、ShanghaiTech数据集上帧级曲线下面积(area under the curve, AUC)分别为88.91%、97.82%;检测头参数量为3.08 M,乘加运算量(multiply-accumulate operations, MACs)为5.90 G。**结论**　TFC-Net在维持较高检测精度的同时大幅降低计算开销,适用于低延迟、大规模视频异常筛查场景。

关键词： 视频异常检测;时频协同;频域补偿;多示例学习;轻量化网络

TFC-Net: An Efficient Lightweight Time-Frequency Collaborative Method for Video Anomaly Detection

Fu Qi[1,2*], Zhu Tao[3], Liu Dexi[2,3], Cao Chonghua[3], Kuang Sirui[3], Huang Junting[3]

1. School of Mathematics and Computer Science, Yuzhang Normal University, Nanchang 330103; 2. Key Laboratory of Data and Knowledge Engineering of Jiangxi Provincial Universities(Jiangxi University of Finance and Economics), Nanchang 330032; 3. Jiangxi University of Finance and Economics, Nanchang 330013

Abstract: Objective Weakly supervised video anomaly detection has attracted considerable attention because it can identify and localize abnormal events using only video-level labels, substantially reducing the cost of frame-level annotation. However, existing methods commonly rely on deep temporal aggregation to capture contextual dependencies and high-level semantic information. Although temporal aggregation is effective for modeling long-range video content, repeated feature integration and temporal smoothing can weaken high-frequency transient responses. As a result, subtle motion changes, short-duration abnormal events, and fine-grained dynamic patterns may be overlooked or inaccurately localized. This problem limits the ability of weakly supervised models to detect anomalies that differ from normal activities only in local temporal details. Therefore, this study develops a lightweight video anomaly detection network that enhances fine-grained anomaly perception while maintaining high detection accuracy and efficient inference. **Methods** We propose a lightweight

收稿日期：2026-08-05;**修回日期：**2026-11-16

*通信作者：付淇　419800621@qq.com

基金项目：国家自然科学基金项目(62676179);豫章师范学院江西省人才战略发展研究基地重点项目(YZRCZD-25-03)

Supported by: the National Natural Science Foundation of China (62676179), and the Key Project of the Jiangxi Provincial Research Base for Talent Strategy Development at Yuzhang Normal University(YZRCZD-25-03)

© 中国图象图形学报版权所有

time-frequency collaborative network, termed TFC-Net, for weakly supervised video anomaly detection. The proposed network establishes a complementary representation mechanism that jointly exploits temporal and frequency-domain information. First, TFC-Net introduces a multi-scale temporal branch to model the semantic evolution of video features at different temporal scales. The multi-scale design enables the network to capture both local temporal variations and broader contextual dependencies, thereby improving the representation of the overall activity structure. Through temporal modeling, the network can distinguish the semantic differences between normal and abnormal video segments and obtain contextual features that support video-level anomaly classification. Nevertheless, temporal aggregation alone may excessively smooth the feature sequence and suppress transient abnormal signals. To address this limitation, TFC-Net incorporates a frequency-domain compensation branch that provides additional information for fine-grained dynamic modeling. Specifically, fast Fourier transform is first applied to the temporal features to obtain their frequency-domain representations. Compared with direct temporal modeling, frequency-domain analysis provides a different perspective for characterizing variations within video sequences and allows the network to explicitly process temporal frequency components. The proposed branch performs complex-spectrum interaction in the frequency domain to jointly exploit the amplitude and phase information contained in the complex spectrum. This process enables the network to identify informative frequency responses associated with subtle and rapidly changing motion patterns. Through complex spectral interaction, redundant noise and less informative frequency components are suppressed, while frequency components that contribute to abnormal dynamic representation are enhanced. The resulting frequency-domain features are subsequently transformed back into the temporal domain through inverse Fourier transform. In this way, the frequency-domain compensation branch is integrated with the temporal representation and provides complementary dynamic information for anomaly detection. The reconstructed features preserve the temporal structure of video segments and retain the temporal localization capability required for identifying abnormal intervals. By combining the multi-scale temporal branch and the frequency-domain compensation branch, TFC-Net simultaneously represents global semantic context and local high-frequency dynamics. This time-frequency collaborative mechanism directly alleviates the feature-smoothing problem caused by deep temporal aggregation and strengthens the model's ability to perceive fine-grained and transient anomalies. TFC-Net is trained under a video-level multiple-instance learning framework. Each video is regarded as a bag of temporal instances, and the network learns to identify abnormal instances using only video-level supervision. This training strategy avoids the need for labor-intensive frame-level annotations while maintaining the ability to localize abnormal segments. In addition, TFC-Net does not rely on complex object-tracking modules or high-cost vision–language models. The simplified architecture reduces the dependence on additional computational components and makes the proposed method suitable for efficient deployment in large-scale video analysis systems. **Results** The proposed TFC-Net is evaluated on the UCF-Crime and ShanghaiTech datasets. Experimental results show that TFC-Net achieves frame-level AUC scores of 88.91% on UCF-Crime and 97.82% on ShanghaiTech, demonstrating its strong capability to detect abnormal events across different surveillance environments and anomaly categories. The results confirm that the time-frequency collaborative architecture effectively improves the representation of transient and fine-grained abnormal patterns while preserving the semantic information required for reliable anomaly detection. In terms of model complexity, the detection head contains only 3.08 million parameters and requires 5.90 G MACs. These results demonstrate that TFC-Net achieves high detection performance with a compact model structure and low computational cost. **Conclusion** This study proposes TFC-Net, a lightweight time-frequency collaborative network for weakly supervised video anomaly detection. The network combines multi-scale temporal modeling with frequency-domain compensation to overcome the loss of high-frequency transient information caused by deep temporal aggregation. Through fast Fourier transform, complex-spectrum interaction, and inverse Fourier reconstruction, TFC-Net suppresses redundant noise, enhances microscopic dynamic representation, and preserves the temporal localization of abnormal segments. The network further adopts a video-level multiple-instance learning framework, eliminating the need for complex tracking modules and computationally expensive vision–language models. On the UCF-Crime and ShanghaiTech datasets, TFC-Net achieves frame-level AUC scores of 88.91% and 97.82%, respectively, while maintaining only 3.08 million parameters and 5.90 G MACs in the detection head.

Key words: video anomaly detection; time-frequency collaboration; frequency-domain compensation; multiple instance

© 中国图象图形学报版权所有

learning; lightweight network

0 引 言

弱监督视频异常检测(weakly supervised video anomaly detection, WS-VAD)旨在仅利用视频级标签而非帧级精细标注,从长时间、复杂背景的视频序列中识别并定位异常事件,是智能监控、公共安全和自动化视频分析中的重要任务之一(Sultani 等, 2018; Ramachandra 等, 2022; 吉根林 等, 2024; 汪洋 等, 2025)。与全监督方法相比,WS-VAD 显著降低了人工标注成本,更符合真实监控场景中大规模视频数据的应用需求。

现有 WS-VAD 多采用多示例学习(multiple-instance learning, MIL),以视频级标签约束片段级异常分数。RTFM(Tian 等, 2021)通过特征幅度增强可分性,S3R(Wu 等, 2022)利用自监督稀疏表示建模正常上下文,MGFN(Chen 等, 2023a)结合全局浏览与局部聚焦。注意力与 Transformer 进一步增强长程依赖建模(Li 等, 2022; Cao 等, 2023; Hu 等, 2025)。视觉语言模型(vision-language model, VLM)则引入文本提示与跨模态对齐(Wu 等, 2024a, 2024b; Zhang 等, 2025),但附加语义建模也增加了计算与部署开销。

深层时间聚合有利于获得稳定语义,却也可能平滑短时、高频的动态变化。爆炸、斗殴等异常通常具有明显视觉线索,而盗窃、抢夺、快速接触和突然逃离等行为的判别信息可能集中在短时间窗口内;这些瞬态线索被平滑后,模型容易漏检或出现定位偏移。

从频域角度看,这一现象可以理解为现有时域模型中的低通滤波偏差。深度神经网络通常更容易学习低频模式,而对高频模式的拟合相对困难,这一现象也被称为谱偏差(Rahaman 等, 2019; Xu 等, 2019)。在 WS-VAD 场景中,谱偏差会进一步放大弱标签监督的不确定性,使模型更倾向于依赖稳定、显著和持续的低频语义线索,而忽视短暂、突发和局部的高频异常响应。因此,仅依赖单一时域表示的深度特征抽象,往往难以兼顾宏观语义理解与微观动态敏感性。

为缓解这一问题,频域分析提供了一种与时域建模正交且互补的视角。傅里叶变换能够将时间序列分解为不同频率分量,使模型显式感知周期模式、突发变化和高频动态。已有研究表明,频域信息对于捕捉细粒度结构、恢复高频细节和增强模型鲁棒性具有重要价值(Rahaman 等, 2019; Xu 等, 2019; Tan 等, 2024)。在视频异常检测中,也有研究尝试利用频率信息增强局部运动响应,或通过频域分解区分低频语义与高频异常扰动(Pi 等, 2024)。然而,现有方法多停留在频域特征增强或简单分解层面,对高频微观动态与宏观时序语义之间的联合建模仍不充分。尤其是在 WS-VAD 中,异常检测不仅要求判断异常是否存在,还需要在时间轴上准确定位异常片段,因此如何在利用频域敏感性的同时保留时间定位能力,仍是一个关键问题。

基于此,本文提出时频协同网络(time-frequency collaborative network, TFC-Net)。TFC-Net 构建了一个正交且互补的双流架构:一方面,时域分支通过多尺度时间聚合捕获宏观语义上下文;另一方面,频域分支引入“快速傅里叶变换(fast fourier transform, FFT) - 滤波 - 逆快速傅里叶变换(inverse fast fourier transform, iFFT)”的闭环机制,先将片段级特征映射至频域,再通过可学习的频域交互模块增强与异常相关的高频分量,最后利用 iFFT 将频域信息重构回时域,形成对微观动态敏感的增强时间表示。与单纯在频域中进行特征分解不同,TFC-Net 通过 iFFT 将频域补偿后的动态线索重新注入时域,使模型既能够捕捉短暂异常信号,又能够满足帧级异常定位需求。

在视频级 MIL 框架下,TFC-Net 结合双分支融合、一致性约束与时间平滑约束,兼顾语义鲁棒性和细微动态感知。在 UCF-Crime 和 ShanghaiTech 上的实验验证了其检测性能与计算效率。

本文的主要贡献如下:

1)本文提出时频协同建模新范式,针对现有时域建模过度平滑细微瞬时异常动态、高频特征表征能力不足的缺陷,打破单一时域特征学习的传统框架。通过构建时域宏观语义建模与频域微观动态增强相结合的时频协同学习机制,有效解决语义稳健性与动态特征敏感性难以兼顾的权衡问题,为视频时序特征精细化建模提供了新的研究思路。

© 中国图象图形学报版权所有

2)本文设计闭环频域补偿优化模块。该模块依托快速傅里叶变换、自适应频域滤波与逆傅里叶重构,建立从时域到频域再回到时域的闭环优化流程,可自适应滤除频域冗余噪声,并精细强化视频高频细微异常特征。该机制在有效弥补纯时域建模信息缺失的同时,大幅提升模型对短暂、细粒度异常行为的识别能力,克服传统时域聚合方法的低通滤波固有缺陷。

3)本文构建轻量化高精度检测架构。TFC-Net摒弃复杂时序跟踪模块与高开销视觉语言分支,基于视频级多示例学习搭建轻量化主干网络,在保障细粒度、高精度异常识别能力的同时,显著降低模型参数量与计算开销。在UCF-Crime和ShanghaiTech两大主流视频异常检测基准数据集上的实验结果表明,所提模型具备良好的检测性能、推理效率和泛化能力,可满足实际智能监控场景低延迟、规模化、实时化的异常筛查需求。

1 相关工作

1.1 WS-VAD的时间局限性

WS-VAD由基础MIL逐步发展为时序与多模态建模(朱新瑞 等, 2024; 苗仕林等, 2025)。Sultani等(2018)利用正常、异常视频包的排序损失定位异常;RTFM(Tian等, 2021)增强特征幅度的判别性,S3R(Wu等, 2022)引入自监督稀疏表示,MGFN(Chen等, 2023a)结合全局浏览与局部聚焦,以适应不同持续时间的事件。LN_Net(Shu等, 2026)进一步以高效时间建模和自适应聚焦减少冗余计算。

多模态线索和频率增强分别从不同角度改善异常表征(Majhi等, 2025; Pi等, 2024)。OrthoVAD(Zhu等, 2026)通过时间稀疏门控筛选片段,并以正常、异常原型的正交约束增强可分性。本文则针对时间聚合可能削弱瞬态线索的问题,结合谱偏差研究(Rahaman等, 2019; Xu等, 2019),引入频域补偿。

1.2 时间数据中的频域分析

频域分析为恢复聚合过程中减弱的细节提供了补充视角(Rao等, 2021; Pi等, 2024; Tan等, 2024; Yu等, 2025)。GFNet(Rao等, 2021)以可学习全局滤波器替代空间自注意力,展示了频域建模全局依赖的效率。

在工业视频中,INFP(Yu等, 2025)针对固定运动轨迹设计轨迹滤波与多特征融合,利用频率信息捕捉周期性及细微异常。本文将频域建模用于运动模式更复杂的通用监控场景。

与本文任务最接近的FE-VAD(Pi等, 2024)通过FFT显式分解并增强高低频信息,分别关注高级语义和局部异常。TFC-Net保留完整复频谱,侧重实部与虚部的联合交互。

具体而言,TFC-Net将RFFT所得实部与虚部沿通道拼接,经跨$2C$通道卷积交互后重构复频谱,再以iFFT和残差连接补偿片段特征。频率引导扩散模型(Tan等, 2024)则利用离散余弦变换区分局部高频与全局低频运动。

TFC-Net通过iFFT将频域增强结果重构为时间表示,使微观动态与宏观语义在同一时间轴上融合,并保留片段定位能力。

1.3 多分支网络与特征融合策略

复合网络通过融合多层次表征支持行为识别(黄敏 等, 2022)。在WS-VAD中,MGFN与双记忆网络利用互补路径建模不同异常模式(Chen等, 2023a; Zhou等, 2023)。ProDisc-VAD(Zhu等, 2025b)利用少量可学习原型建模正常模式,并通过高置信度伪实例增强特征可分性。

多分支方法常通过特征拼接或分数融合整合互补信息(Cai等, 2021; Chen等, 2023a; Sun等, 2023; Zhou等, 2023)。DS-IB Net(Zhu等, 2025a)将双流建模与信息瓶颈结合,以紧凑表征兼顾正常动态与稀疏异常。

TFC-Net先在复频谱中交互建模,再将频率感知特征重构回时域,与多尺度时间特征融合,从而补充仅靠时域聚合难以保留的瞬态线索。

1.4 多模态与视觉语言模型方法

VLM与多模态学习通过文本语义和提示增强异常理解(任小龙 等, 2025; 桑农 等, 2026)。Vad-CLIP(Wu等, 2024a)学习视觉与文本提示的交互,STPrompt(Wu等, 2024b)利用时空提示辅助定位;Holmes-VAU(Zhang等, 2025)结合异常聚焦采样与InternVL2-2B,面向长视频异常理解和解释。

这些方法增强了语义表达,但预训练模型、跨模态匹配及附加推理模块也增加了计算开销。对于需要持续处理大规模视频流的监控系统,检测精度与低延迟需同时考虑。

© 中国图象图形学报版权所有

TFC-Net聚焦视频级MIL的视觉表征瓶颈，在预提取片段特征上进行时频协同建模；检测头不引入文本推理或目标跟踪，以较低开销增强对短暂异常的感知。

2 TFC-Net

2.1 方法动机

本节首先从频域视角分析现有时域建模范式的局限性，并据此引出TFC-Net的核心设计思路。在WS-VAD中，主流方法高度依赖深度时间聚合，如时序卷积或注意力机制，来提取宏观语义。然而，随着网络层数的加深，这种聚合操作不可避免地呈现出低通滤波器的特性。根据卷积定理，时域上的聚合对应于频域中的乘法，如公式(1)所示：

$$\mathcal{F}(y)(\omega)=\mathcal{F}(x)(\omega)\cdot\mathcal{F}(w)(\omega) \tag{1}$$

深度网络的权重分布$\boldsymbol{w}$在训练中倾向于平滑，其映射函数$\mathcal{F}(w)$往往呈现出低通响应倾向，导致：

$$\lim_{\omega\to\infty}\left|\mathcal{F}(w)(\omega)\right|\approx 0 \tag{2}$$

这种固有的过度平滑效应使得模型在捕获宏观语义的同时，无形中削弱了与快速盗窃及突然冲突等微小异常高度关联的高频瞬态脉冲$\delta(t)$。这也是纯时域模型频繁漏检细微异常的底层原因。

为了突破这一局限，本文将视角转向与时域正交的频域。离散傅里叶变换(discrete Fourier transform，DFT)具有固有的全局感受野：

$$X[k]=\sum_{n=0}^{N-1}x[n]\cdot e^{-i2\pi\frac{kn}{N}} \tag{3}$$

公式(3)表明，每个频率分量均由整个时间窗口内的所有信号点共同决定。在时域中容易被复杂背景噪声淹没的异常行为，在频域中往往表现为对正常周期性模式的显著破坏，从而展现出更强的线性可分性。然而，仅仅依赖频域分析会导致时间位置信息的丢失，无法满足异常检测必需的帧级定位要求。

因此，本文进一步引入iFFT来构建时频闭环机制。作为时序定位器，它能够恢复关键的时间位置信息：

$$x_{res}(t)=\mathcal{F}^{-1}(\mathcal{F}(x)(\omega)\odot H(\omega)) \tag{4}$$

公式(4)中，$H(\omega)$为频域滤波器，$\odot$表示哈达玛积(Hadamard product)。这一过程从原始信号中分离出特定的动态模式。通过iFFT重构的特征$x_{res}(t)$变成了一种频率感知的时间响应。它保留了时间维度t，使模型能够准确追踪高频瞬态发生的确切时刻。通过这种机制，TFC-Net在保留宏观语义的同时，精准补偿了丢失的微观动态，实现了两者在时域上的协同。

2.2 TFC-Net总体框架

TFC-Net在预提取的片段级视觉特征上构建时频互补的双分支检测头。如图1所示，频域分支利用复数—实数交互提取微观动态(第2.2.1节)，时域分支通过多尺度聚合建模语义上下文(第2.2.2节)。两分支特征融合后生成异常分数，无需额外的文本推理或目标跟踪模块。

2.2.1 频域分支与相位感知复频谱交互

频域分支通过提取全局周期模式，实现与时域特征的有效互补。如图1所示，该分支的处理流程包含三个连续的阶段：1)基于RFFT的频域变换；2)复数—实数交互建模；以及3)基于iFFT的逆频域变换。

首先，给定与时域分支平行的输入片段级特征$X_{feat}=\{X_t\}_{t=1}^{T}\in\mathbb{R}^{T\times C_{in}}$本文将其视为$C_{in}$个独立的一维时间信号，记为$X_{time}\in\mathbb{R}^{B\times C_{in}\times T}$。利用实数快速傅里叶变换(RFFT)$\mathcal{F}$，本文将信号从时域转换至频域：

$$F_{freq}=\mathcal{F}(X_{time})\in\mathbb{C}^{B\times C_{in}\times(T/2+1)} \tag{5}$$

其中T表示时间序列长度，C_{in}表示片段级特征的通道数(即特征维度)，B表示批大小。在该阶段，F_{freq}为复数张量，其振幅与相位分别编码了原始信号中各频率分量的强度与相移。由于RFFT沿时间维对片段级特征进行频率分解，不同时间变化速率会被映射到不同的频率位置。其中，持续、缓慢变化主要集中于低频区域，而短时间内发生的局部突变需要更多较高频率基函数参与表征，其频谱响应因而会扩展至高频区域。由此，原本混合在时间序列中的快速动态变化被转换为显式的高频复数系数，实现对短时异常动态线索的频域提取。

鉴于标准卷积神经网络难以直接处理复数运算，本文设计了相位感知复频谱交互模块，用于在保留幅度和相位信息的同时，以低成本方式建模复数频谱中的实部－虚部关系。如图2所示，通过通道拼接策略在实数域内实现复数特征的交互。给定频域特征，本文提取其实部$\Re(\cdot)$与虚部$\Im(\cdot)$，并沿通道

© 中国图象图形学报版权所有

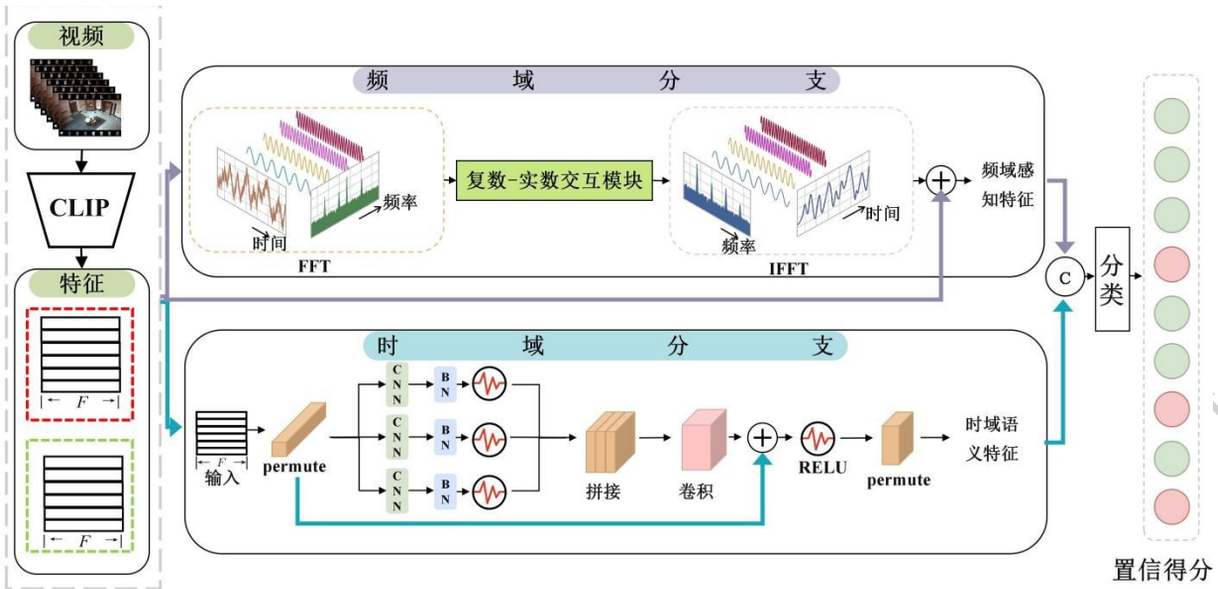


图1　TFC-Net 总体框架

Fig. 1　Overall framework of TFC-Net

维度进行拼接，将其映射为实值张量 F_{stack}。

$$F_{stack} = \text{Concat}(\Re(F_{freq}), \Im(F_{freq})) \in \mathbb{R}^{(T/2+1)\times 2C} \tag{6}$$

随后，本文采用参数为 $\boldsymbol{W}$ 的卷积层 f_{conv} 对 F_{stack} 进行映射。通过跨 $2C$ 个通道的联合卷积操作，该模块构建了实部与虚部之间的非线性依赖关系：

$$H_{inter} = f_{conv}\left(F_{stack}; W\right) \in \mathbb{R}^{(T/2+1)\times 2C} \tag{7}$$

公式(7)中，H_{inter} 表示交互后的混合特征。为明确卷积映射对不同频率成分的增强作用，将 H_{inter} 沿通道维拆分为 H_{real} 和 H_{imag}。对于第 k 个频率位置，交互后的频谱幅值可表示为：

$$\hat{A}_k = \sqrt{H_{real,k}^2 + H_{imag,k}^2} \tag{8}$$

公式(8)中，k 表示频率位置，$H_{real,k}$ 和 $H_{imag,k}$ 分别表示卷积交互后第 k 个频率位置对应的实部和虚部，$\hat{A}_k$ 表示该位置交互后的频谱幅值。

在此基础上，用交互前后的幅值变化比例衡量该频率成分的响应增强程度：

$$g_k = \frac{\hat{A}_k}{|F_{\text{freq},k}| + \varepsilon} \tag{9}$$

公式(9)中，$|F_{freq,k}|$ 表示交互前第 k 个频率位置的频谱幅值，g_k 表示该频率成分经卷积交互后的幅值变化比例，ε 为避免分母为零而设置的极小正数。当 $g_k>1$ 时，该频率成分的响应得到增强；当 $g_k<1$ 时，其响应受到抑制。在异常检测目标的约束下，跨 $2C$ 通道的卷积能够根据不同频率成分对异常判别的作用调整其响应强度，使与短时异常相关的高频成分获得更明显的响应，同时削弱部分冗余频率信息。为了将调制后的特征恢复至复数形态以进行后续处理，本文将输出特征沿通道维重新拆分为 H_{real} 和 H_{imag}，并重构为复数形式：

$$\hat{F}_{freq} = H_{real} + i \cdot H_{imag} \in \mathbb{C}^{(T/2+1)\times C} \tag{10}$$

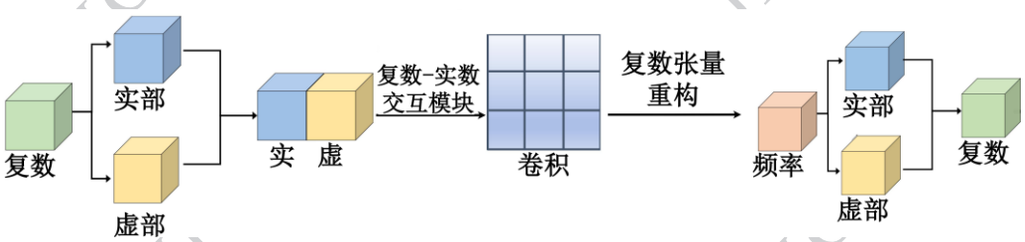


图2　复数-实数交互模块

Fig. 2　Complex-real interaction module

© 中国图象图形学报版权所有

最后，通过iFFT将频域特征还原为时域表示。在此处引入残差连接，将其与原始输入相加，得到最终的频域感知特征。这种机制使该分支只需聚焦于学习相对于原始特征的频域残差补偿，从而进一步提升了模型训练的稳定性。

2.2.2 时域分支与多尺度聚合

时域分支（如图1下半部分所示）侧重于捕获跨越不同时间跨度的语义依赖关系。为了满足时间卷积的输入维度要求并提取多尺度上下文，本文设计了以下处理流程：

对于输入的时间特征 $X_{\text{time}} \in \mathbb{R}^{B\times C\times T}$，本文首先执行维度置换（permute）操作，将输入重塑为 $C\times T$ 格式，以适应一维卷积的通道优先要求。这些特征被同时输入到三个具有卷积核大小 $k\in\{3,5,7\}$ 的独立一维卷积分支中。

$$H_k=\sigma(BN(Conv_k(X_{perm})))\quad k\in\{3,5,7\} \tag{11}$$

其中 C 表示特征通道数，T 表示时间步数 $Conv_k$ 表示卷积核大小分别为3、5和7的一维卷积层，它们定义了不同的时间感受野；BN 表示用于标准化特征分布和加速收敛的批量归一化；σ 是非线性激活函数。

随后，在不同尺度上提取的这些特征图沿着通道维度进行拼接，形成一个聚合的多尺度特征表示：

$$h_{cat}=\text{Concat}[h_3,h_5,h_7] \tag{12}$$

为了智能地融合这些并行的特征流，本文采用一个1×1卷积层来处理 h_{cat}，实现多尺度特征图的自适应加权融合。

最后，在经过ReLU激活后，本文再次应用置换操作将特征维度还原至 $T\times C$。为了增强梯度传播并保留底层特征，本文引入了残差连接，将融合后的特征与经1×1卷积 W_r 对齐维度后的原始输入 X_{time} 进行逐元素相加：

$$X_{time_out}=\text{ReLU}(Conv_{1\times1}(h_{cat})+W_r(X_{time})) \tag{13}$$

最终输出是一个鲁棒的时域特征表示，它编码了短期和长期的上下文信息，为后续的异常判别提供了丰富的语义基础。

2.2.3 优化目标

如图1所示，TFC-Net的最后阶段是特征融合。频率感知特征和时域语义特征通过通道拼接操作合并为一个增强的特征向量。

为了训练TFC-Net并实现双分支之间的有效协同，本文设计了一个包含三部分的复合损失函数。首先，鉴于弱监督设置下缺乏帧级标签，本文采用了Top-kMIL损失。与仅关注单一最大异常分数的传统方法不同，Top-k策略在每个视频包中选择 k 个最高分数的集合 Ω_k，以覆盖异常事件的完整持续时间。本文使用二元交叉熵（BCE）作为基础分类损失：

$$\mathcal{L}_{mil}=-\frac{1}{K}\sum_{j\in\Omega_k}[y\log(s_j)+(1-y)\log(1-s_j)] \tag{14}$$

公式（14）中，$y\in\{0,1\}$ 是视频级标签，s_j 是由融合特征预测的异常分数。k 的大小设置为 $k=\lceil T/p\rceil$，根据视频长度 T 进行自适应调整，其中 p 为控制Top-k采样比例的缩放因子，本文取 p=16，即每个视频包中选取约1/16的片段参与损失计算。该损失函数迫使模型识别异常视频中的高置信度异常片段，同时抑制正常视频中的虚假警报。

其次，为了确保时域分支和频域分支在决策层面达成共识，本文引入了分支一致性损失。尽管两个分支侧重于正交的特征维度，但它们对同一时间步 t 的异常判断本质上应该是一致的。本文使用平滑L1损失来约束时域分支（s_t^{time}）和频域分支（s_t^{freq}）预测的异常分数之间的差异：

$$\mathcal{L}_{con}=\frac{1}{T}\sum_{t=1}^{T}\text{SmoothL1}(s_t^{time},s_t^{freq}) \tag{15}$$

$\mathcal{L}_{con}$ 损失作为一种相互监督机制，迫使模型学习特定模态的特征，同时防止单一分支出现过度自信的错误预测，从而实现真正的协同检测。

此外，无论是纯正常视频还是异常视频中的正常片段，通常都表现出连续且稳定的时序演变规律，这意味着预测分数的剧烈波动往往源于噪声干扰。为了利用这一时间先验来约束模型行为，本文引入了时间平滑度损失。通过最小化相邻时间步之间预测得分的差异，该损失有效增强了输出结果的时间平滑性：

$$\mathcal{L}_{sm}=\frac{1}{T-1}\sum_{t=1}^{T-1}(s_t-s_{t+1})^2 \tag{16}$$

为了避免抑制真正突发异常的响应信号，此损失函数仅应用于正常视频样本。最后，通过最小化以下加权总损失函数，对TFC-Net进行端到端优化：

$$L_{total}=\lambda_{mil}L_{mil}+\lambda_{con}L_{con}+\lambda_{sm}L_{sm} \tag{17}$$

公式（17）中，λ_{mil}、λ_{con} 和 λ_{sm} 表示用于平衡各项损失的超参数。

© 中国图象图形学报版权所有

3 实验及结果分析

3.1 实验环境

为了验证TFC-Net在弱监督视频异常检测任务中的有效性和效率，本文在UCF-Crime和ShanghaiTech两个主流公开视频异常检测数据集上进行定量实验，并与当前主流方法进行对比分析。UCF-Crime包含1900个真实世界监控视频，涵盖13类异常事件，其中1610个视频用于训练，290个视频用于测试。ShanghaiTech包含采集自13个不同校园场景的437个视频片段，本文采用该数据集的标准弱监督划分设置。

基于PyTorch实现TFC-Net，使用预训练CLIP ViT-B/16视觉编码器离线提取512维片段级特征，并在其上训练检测头。实验硬件为NVIDIA GeForce RTX 3060显卡及Intel Core i5-13490F处理器。检测头采用Adam优化器训练50个epoch，权重衰减为5×10^{-4}，初始学习率为1×10^{-4}，采用余弦退火。批量大小为32，包括16个正常视频和16个异常视频。损失权重设为$\lambda_{mil}=1.0$，$\lambda_{con}=0.5$，$\lambda_{sm}=0.1$。

3.2 评价指标

采用帧级接收者操作特征曲线下面积（area under the curve，AUC）评价正常与异常片段的判别能力。

3.3 对比实验

表1比较ShanghaiTech上的帧级AUC及对应特征，包括膨胀三维卷积网络（inflated 3D ConvNet，I3D）、VideoSwin、CLIP及结合文本的方案。

表2在更具挑战性的UCF-Crime数据集上，TFC-Net取得了88.91%的AUC，高于RTFM、MGFN、FE-VAD、VadCLIP、STPrompt、explainable fine- grained video anomaly detection（Ex-VAD）、distilling aggregated knowledge with disentangled attention（DAKD）和discriminative score suppression（DSS）等纯视觉或轻量视觉语言基线。

Holmes-VAU结合InternVL2-2B、层级指令数据与异常聚焦采样，取得88.96%的AUC，其监督形式、模型规模和推理流程与视频级MIL不同。TFC-Net仅处理片段特征，无需密集跟踪或多模态大语言模型（multimodal large language model，MLLM）辅助，即取得接近的检测性能，体现了轻量时频建模的效果。

表1 ShanghaiTech数据集AUC性能对比

Table 1 ShanghaiTech AUC comparison

Method	Feature	AUC(%)
Sultani等(2018)	I3D	85.33
MIST(Feng等, 2021)	I3D	94.83
RTFM(Tian等, 2021)	I3D	97.21
S3R(Wu等, 2022)	I3D	97.48
MSL(Li等, 2022)	VideoSwin	97.20
FE-VAD(Pi等, 2024)	I3D	98.12
TEVAD(Chen等, 2023b)	I3D/Text	97.80
ProDisc-VAD(Zhu等, 2025b)	CLIP	97.98
Skynet-V1(Ma等, 2025)	CLIP	97.85
DSIB-Net(Zhu等, 2025a)	CLIP	**98.13**
TPWNG(Yang等, 2024)	CLIP	98.00
VadCLIP(Wu等, 2024a)	CLIP	97.49
DAKD(Dalvi等, 2025)	CLIP	98.10
TFC-Net(Ours)	CLIP	97.82

表2 UCF-Crime数据集AUC性能对比

Table 2 UCF-Crime AUC comparison

Method	Feature	AUC(%)
Sultani等(2018)	I3D	77.92
RTFM(Tian等, 2021)	I3D	84.30
MGFN(Chen等, 2023b)	I3D	86.98
DSS(Xu等, 2025)	I3D	86.37
FE-VAD(Pi等, 2024)	I3D	87.13
ProDisc-VAD(Zhu等, 2025b)	CLIP	87.12
Skynet-V1(Ma等, 2025)	CLIP	87.60
DSIB-Net(Zhu等, 2025a)	CLIP	87.65
TPWNG(Yang等, 2024)	CLIP	87.79
VadCLIP(Wu等, 2024b)	CLIP	88.02
STPrompt(Wu等, 2024a)	CLIP	88.08
Ex-VAD(Huang等, 2025)	VLM	88.29
DAKD(Dalvi等, 2025)	CLIP	88.34
Holmes-VAU(Zhang等, 2025)	VLM/MLLM	**88.96**
TFC-Net(Ours)	CLIP	88.91

© 中国图象图形学报版权所有

3.4 计算效率分析

本文在检测阶段统一统计主要乘加运算和完整注册参数。THOP依赖算子hook，可能漏计多头注意力、矩阵乘法、einsum、自定义图卷积或未覆盖的冻结模块。因此，MACs按Conv、Linear、Attention、Mat-Mul和einsum等运算逐模块解析计算，参数量按模型全部注册参数累加。

检测头统计不含离线视觉骨干，但包含在线参与检测的文本编码器、可学习提示及跨模态模块。本文以公开checkpoint核对参数规模：纯fp32权重的理论大小约为参数量×4字节，实际文件还可能包含元数据或优化器状态。STPrompt无官方代码，其复杂度沿用原论文报告并单独标注。

表3显示，RTFM、MGFN和TFC-Net的THOP参数量与checkpoint规模基本对应；Holmes-VAU-ATS的估计略低。VadCLIP的THOP结果为35.17M、理论权重134.16MB，而完整注册参数为162.26M、checkpoint为619.10MB，差异主要来自未完整计入的文本侧组件，而非离线视觉CLIP。因此，表3用于展示工具统计与权重文件的差异；后续效率比较统一采用完整注册参数量与解析式MACs，以保证各检测模型的统计边界一致。

表3 THOP统计结果比较

Table 3 Comparison based on THOP profiling

Method	MACs	Params	FP32	Ckpt
RTFM(Tian等, 2021)	63.29	24.71	94.26	94.30
MGFN(Chen等, 2023b)	73.33	28.65	109.29	109.30
VadCLIP(Wu等, 2024b)	47.56	35.17	134.16	619.10
Holmes-VAU-ATS(Zhang等, 2025)	16.48	6.17	23.53	24.78
TFC-Net(Ours)	**5.90**	**3.08**	**11.74**	**11.76**

注：FP32表示THOP参数量×4bytes的理论权重大小；Ckpt表示公开checkpoint或等价权重文件大小，均按1MB=10242bytes统计。

表4中，TFC-Net检测头为5.90GMACs、3.08M参数；fp32理论权重11.74MB与实际checkpoint 11.76MB相符。

表4 计算效率比较

Table 4 Computational efficiency comparison

Method	MACs	Params	FP32	Ckpt
RTFM(Tian等, 2021)	63.29	24.71	94.26	94.30
MGFN(Chen等, 2023b)	73.33	28.65	109.29	109.30
VadCLIP(Wu等, 2024b)	70.89	162.26	618.97	619.10
STPrompt(Wu等, 2024a)†	44.80	31.50	120.16	-
Holmes-VAU-ATS(Zhang等, 2025)	18.77	6.49	24.76	24.78
TFC-Net(Ours)	**5.90**	**3.08**	**11.74**	**11.76**
vs RTFM	90.7%↓	87.5%↓	-	87.5%↓
vs MGFN	92.0%↓	89.2%↓	-	89.2%↓
vs. VADCLIP	91.7%↓	98.1%↓	-	98.1%↓
vs STPrompt	86.8%↓	90.2%↓	-	-
vs Holmes-VAU-ATS	68.6%↓	52.5%↓	-	52.5%↓

注：†STPrompt未公开官方代码，表中数值来自其原论文复杂度报告；FP32表示完整注册参数量×4bytes的理论权重大小；Ckpt表示公开check-point或等价权重文件大小，均按1MB=10242bytes统计。

© 中国图象图形学报版权所有

表4中各方法的完整参数量与权重规模基本吻合。相较RTFM和MGFN，TFC-Net分别减少90.7%和92.0%的MACs、87.5%和89.2%的参数及checkpoint大小；相较VadCLIP，分别减少91.7%、98.1%和98.1%；相较STPrompt的论文报告值，MACs减少86.8%。

Holmes-VAU在表4中仅计入异常聚焦时间采样器（Holmes-VAU-ATS）的特征级打分模块，不含InternVL2-2B的生成流程。相较该模块，TFC-Net仍减少68.6%的MACs、52.5%的参数与checkpoint大小。

表5进一步计入视觉骨干，比较TFC-Net与完整Holmes-VAU：前者包含CLIP ViT-B/16视觉编码器，后者包含InternVL2-2B视觉语言模型。VadCLIP和STPrompt同样依赖CLIP特征，端到端计算也主要由共享视觉编码器决定。

结果表明，TFC-Net在端到端口径下仍仅需4.497×10^4 GMACs和89.27M参数，相比Holmes-VAU分别减少95.6%和96.0%。图3进一步以线性坐标展示了这种差异，能够更直观地体现本文方法相对大规模MLLM系统的轻量化优势。

表5　端到端计算效率比较

Table 5　End-to-end computational efficiency comparison

Method	Backbone	MACs (G)	参数量(M)
TFC-Net(Ours)	CLIP ViT-B/16	**4.497　104**	**89.27**
Holmes-VAU(Zhang等, 2025)	InternVL2-2B	1.03×106	2212.25
vs Holmes-VAU	-	95.6% ↓	96.0% ↓

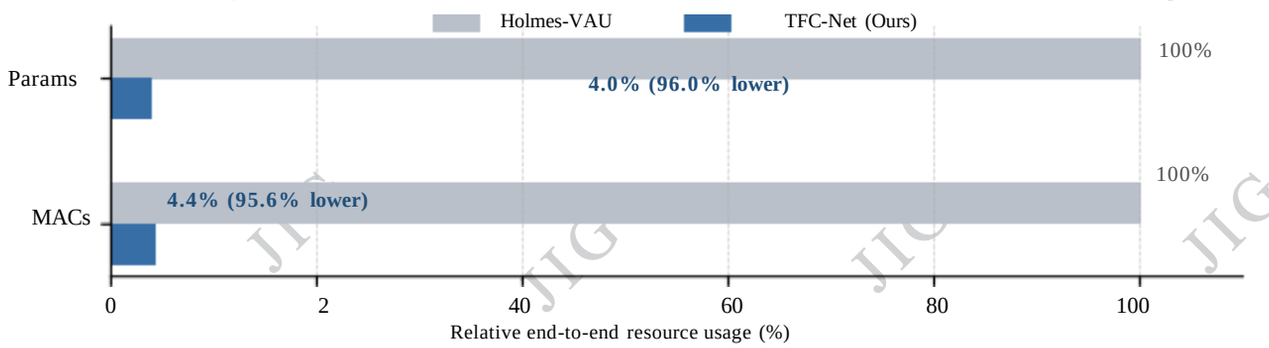


图3　相对Holmes-VAU的端到端资源占用

Fig. 3　End-to-end resource usage relative to Holmes-VAU

Holmes-VAU的完整系统包含视觉编码器、跨模态投影与语言模型，计算量随输入帧数、图像patch数、prompt长度及生成token数变化。表5依据公开实现与模型规模估计，用于比较轻量筛查与MLLM理解系统的资源量级；表4则比较预提取特征后的检测开销。

3.5　消融实验

为验证TFC-Net各组成部分的有效性，本文围绕双流架构、频域交互方式、复频谱分量以及损失函数设计进行了消融实验。所有实验均在相同训练设置下完成，并以UCF-Crime数据集上的帧级AUC作为评价指标。

表6中，单独时域、频域分支的AUC分别为86.47%和87.93%，融合后达到88.91%。时域分支建模连续上下文，频域分支补充动态线索，联合结果支持二者的互补性。

表7比较两种频域建模方式：朴素幅度谱仅保留能量信息，再经一维卷积（one-dimensional convolution，Conv1D）变换；复频谱交互则保留实部与虚部，并以实数域卷积建模其耦合关系。后者获得更高AUC，支持在幅度信息之外保留相位所含时间结构的设计。

为解释复频谱交互的收益，本文分别保留幅度、相位及完整复数表示，在相同训练设置下比较帧级

© 中国图象图形学报版权所有

表 6　双流架构有效性

Table 6　Effectiveness of the dual-stream architecture

时域分支	频域分支	AUC(%)
✓	–	86.47
–	✓	87.93
✓	✓	**88.91**

表 7　相位感知复频谱交互模块有效性

Table 7　Phase-aware interaction module

频域建模方式	AUC(%)
幅度谱建模 w/Conv1D	87.66
复频谱交互建模 w/Real-Imag Conv	**88.91**

AUC。幅度刻画频率能量，相位反映时间结构，该对照用于检验两类信息的互补性，结果如表8所示。

表 8　幅度与相位分量贡献分析

Table 8　Magnitude and phase components

保留分量	AUC(%)
仅幅度(相位=0)	86.95
仅相位(幅度=1)	86.94
完整复数(幅+相)	**88.91**

表8中，完整复数建模优于单独保留幅度或相位，支持两类信息联合用于异常判别。

最后，本文验证了所提出约束的必要性。如表9，移除一致性约束(L_{con})会导致0.76%的下降，表明其确保了稳健的决策。移除平滑度约束(L_{sm})会导致0.53%的下降，表明时间连续性有助于滤除噪声。

表 9　不同损失函数组件的影响

Table 9　Effects of loss components

Lmil	*L con*	*Lsm*	AUC(%)
✓	–	–	87.35
✓	–	✓	88.15
✓	✓	–	88.38
✓	✓	✓	**88.91**

3.6　超参数敏感性

如表10所示，本文在UCF-Crime数据集上对大范围的λ_{sm}值进行了细粒度分析。结果表明，模型性能在不同超参数设置下相对稳定。具体而言，当λ_{sm}=0.1时达到最佳性能(88.91%)。这表明，虽然所提出的TFC-Net能够从参数调整中获益，但它在合理的数值范围内保持了极具竞争力的性能。

表 10　权重系数λ_{sm}敏感性分析

Table 10　Sensitivity analysis of λ_{sm} on UCF-Crime

取值	0	0.01	0.05	0.1	0.5	1	2	5
λ_{sm}	86.36	87.95	88.09	**88.91**	88.41	88.46	88.68	88.45

3.7　可视化分析

图4展示了TFC-Net在UCF-Crime数据集中若干典型案例的可视化结果。观察可知，预测轨迹在异常发生的时刻精准地产生了响应峰值，且高分区间在时间轴上与地面真值展现出高度的一致性。针对枪击、爆炸等瞬态异常，预测曲线表现出尖锐且集中的脉冲式响应；而对于逮捕、交通事故等持续性异常，异常分数则在相应时段内保持稳健的高位波动。这有力地验证了模型对短期突变与长期演化模式的同步捕捉能力，完美契合了时频协同建模的设计初衷：即利用频域分支捕捉瞬态细微特征，同时依托时域分支建模宏观上下文。

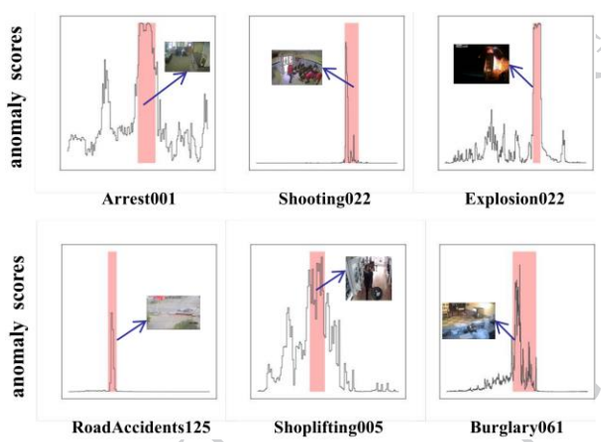


图4　UCF-Crime数据集定性结果

Fig. 4　UCF-Crime qualitative results

尽管提出的TFC-Net已取得了显著进展，但分析表明模型在极端模型下仍存在短板：在伴随剧烈相机抖动的场景中，模型偶尔会产生误报。这是由于相机的全局运动引入了高频噪声，在频谱特征上模拟了瞬态异常的动态过程。如何在频域中实现相机自运动与目标运动的分解，依然是未来进一步提升模型鲁棒性的开放课题。

© 中国图象图形学报版权所有

4 结束语

本文提出了时频协同网络TFC-Net，旨在缓解弱监督视频异常检测中固有的谱偏差与特征过度平滑问题。通过将快速傅里叶变换、滤波与逆快速傅里叶变换构成的闭环机制与多尺度时间聚合器相结合，该框架实现了高频微观动态与宏观时序语义的协同建模。TFC-Net在UCF-Crime和ShanghaiTech等主流基准数据集上展现出具有竞争力的性能，但其核心定位仍是一个高效的下游异常筛查模块，而非独立的端到端边缘系统。当前设计仍存在一定局限性：对离线预提取特征的依赖限制了底层视觉表征的动态微调，而大型视觉骨干网络的计算开销也构成了端到端部署的性能瓶颈，使系统难以直接在资源严重受限的边缘设备上独立运行。未来工作将探索自适应时频窗口与多分辨率频率分析技术。此外，当前框架主要针对视觉谱偏差设计，而在高度依赖尖叫或爆炸声等视听线索的复杂场景中，纯视觉方案仍存在模态感知瓶颈。因此，将时频协同机制拓展至多模态视听联合架构，将是未来推进复杂真实环境落地部署的重要方向。

参考文献(References)

Sultani W, Chen C and Shah M. 2018. Real-world anomaly detection in surveillance videos//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6479-6488 [DOI:10.1109/CVPR.2018.00678]

Ramachandra B, Jones M J and Vatsavai R R. 2022. A survey of single-scene video anomaly detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(5): 2293-2312 [DOI: 10.1109/TPAMI.2020.3040591]

Wang Y, Zhou J G, Yan J and Guan J H. 2025. Survey of anomaly detection methods in surveillance videos based on deep learning. Journal of Image and Graphics, 30(3): 615-640 (汪洋, 周脚根, 严俊, 关佶红. 2025. 基于深度学习的监控视频异常检测方法综述. 中国图象图形学报, 30(3): 615-640) [DOI: 10.11834/jig.240329]

Tian Y, Pang G S, Chen Y H, Singh R, Verjans J W and Carneiro G. 2021. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 4975-4986 [DOI: 10.1109/ICCV48922.2021.00493]

Wu J C, Hsieh H Y, Chen D J, Fuh C S and Liu T L. 2022. Self-supervised sparse representation for video anomaly detection//European Conference on Computer Vision. Tel Aviv, Israel: Springer: 729-745 [DOI:10.1007/978-3-031-19778-9_42]

Chen Y X, Liu Z Z, Zhang B H, Fok W, Qi X J and Wu Y C. 2023a. MGFN: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection//Proceedings of the AAAI Conference on Artificial Intelligence. Washington DC, USA: AAAI Press: 387-395 [DOI: 10.1609/aaai.v37i1.25112]

Li S, Liu F and Jiao L. 2022. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press: 1395-1403 [DOI: 10.1609/aaai.v36i2.20028]

Cao C Q, Zhang X, Zhang S Z, Wang P and Zhang Y N. 2023. Weakly supervised video anomaly detection based on cross-batch clustering guidance//2023 IEEE International Conference on Multimedia and Expo (ICME). Brisbane, Australia: IEEE: 2723-2728 [DOI: 10.1109/ICME55011.2023.00463]

Hu Q, Zhang J, Mi J S, Yuan Z and Li M Z. 2025. TIEOD: Three-way concept-based information entropy for outlier detection. Applied Soft Computing, 170: 112642 [DOI: 10.1016/j.asoc.2024.112642]

Wu P, Zhou X R, Pang G S, Zhou L R, Yan Q S, Wang P, et al. 2024a. VadCLIP: Adapting vision-language models for weakly supervised video anomaly detection//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press: 6074-6082 [DOI: 10.1609/aaai.v38i6.28423]

Wu P, Zhou X R, Pang G S, Yang Z W, Yan Q S, Wang P, et al. 2024b. Weakly supervised video anomaly detection and localization with spatio-temporal prompts//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM: 9301-9310 [DOI:10.1145/3664647.3681442]

Zhang H X, Xu X H, Wang X, Zuo J L, Huang X N, Gao C X, et al. 2025. Holmes-VAU: Towards long-term video anomaly understanding at any granularity//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 13843-13853 [DOI: 10.1109/CVPR52734.2025.01292]

Rahaman N, Baratin A, Arpit D, Draxler F, Lin M, Hamprecht F A, et al. 2019. On the spectral bias of neural networks//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 5301-5310. https://proceedings.mlr.press/v97/rahaman19a.html

Xu Z Q J, Zhang Y, Luo T, Xiao Y and Ma Z. 2019. Frequency principle: Fourier analysis sheds light on deep neural networks. arXiv preprint, 2019[2026-08-03]. https://arxiv.org/abs/1901.06523

Tan X, Wang H, Geng X and Wang L. 2024. Frequency-guided diffusion model with perturbation training for skeleton-based video anomaly detection. arXiv preprint, 2024 [2026-08-03]. https://

© 中国图象图形学报版权所有

arxiv.org/abs/2412.03044

Pi R, Xu J and Peng Y. 2024. FE-VAD: High-low frequency enhanced weakly supervised video anomaly detection//2024 IEEE International Conference on Multimedia and Expo (ICME). Niagara Falls, Canada: IEEE: 1-6 [DOI:10.1109/ICME57554.2024.10688326]

Zhu X R, Qian X Y, Shi Y Z, Tao X D and Li Z Y. 2024. Video anomaly detection with long-and-short-term time series correlations. Journal of Image and Graphics, 29(7): 1998-2010 (朱新瑞, 钱小燕, 施俞洲, 陶旭东, 李智昱. 2024. 长短期时间序列关联的视频异常事件检测. 中国图象图形学报, 29(7): 1998-2010) [DOI: 10.11834/jig.230406]

Miao S L, Tang Y P, Zhang C J, Liu C K, Zhang J T and Cao J F. 2025. Intelligent temporal behavior detection for lunar space station. Journal of Image and Graphics, 30(9): 3039-3049 (苗仕林, 唐业鹏, 张淳杰, 刘传凯, 张济韬, 曹建峰. 2025. 面向月球空间站的智能时序行为检测. 中国图象图形学报, 30(9): 3039-3049) [DOI: 10.11834/jig.250135]

Shu L, Zhu T, Jiang J L, Yu Q, Li S Y and Peng Y. 2026. LN_Net: Lightweight non-interventionist network for weakly-supervised video anomaly detection//PRICAI 2025: Trends in Artificial Intelligence. Singapore: Springer: 412-428 [DOI: 10.1007/978-981-95-7084-3_26]

Majhi S, D'Amicantonio G, Dantcheva A, Kong Q, Garattoni L, Francesca G, et al. 2025. Just dance with pi! A poly-modal inductor for weakly-supervised video anomaly detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 24265-24274 [DOI: 10.1109/CVPR52734.2025.02260]

Zhu T, Liu Y, Luo K W, Huang L J, Tu X Y, Cheng Y H, et al. 2026. OrthoVAD: Weakly supervised video anomaly detection via prototype orthogonality learning//2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE: 11242-11246 [DOI: 10.1109/ICASSP55912.2026.11462316]

Rao Y, Zhao W, Zhu Z, Lu J and Zhou J. 2021. Global filter networks for image classification. Advances in Neural Information Processing Systems, 2021, 34: 980-993.

Yu Q, Zhu K, Cao Y and Kang Y. 2025. INFP: Industrial video anomaly detection via frequency prioritization//Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. Santiago de Compostela, Spain: IJCAI: 2296-2304 [DOI: 10.24963/ijcai.2025/256]

Huang M, Shang R X and Qian H M. 2022. Composite deep neural network for human activities recognition in video. Pattern Recognition and Artificial Intelligence, 35(6): 562-570 (黄敏, 尚瑞欣, 钱惠敏. 2022. 面向视频中人体行为识别的复合型深度神经网络。模式识别与人工智能, 35(6): 562-570) [DOI: 10.16451/j.cnki.issn1003-6059.202206008]

Zhou H, Yu J and Yang W. 2023. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection//Proceedings of the AAAI Conference on Artificial Intelligence. Washington DC, USA: AAAI Press: 3769-3777 [DOI: 10.1609/aaai.v37i3.25489]

Zhu T, Yu Q, Dong X R, Li S Y, Liu Y, Jiang J L, et al. 2025b. ProDisc-VAD: An efficient system for weakly-supervised anomaly detection in video surveillance applications//Proceedings of the 2025 IEEE International Conference on Systems, Man, and Cybernetics. Vienna, Austria: IEEE: 2532-2537 [DOI: 10.1109/SMC58881.2025.11343488]

Cai R, Zhang H, Liu W, Gao S and Hao Z. 2021. Appearance-motion memory consistency network for video anomaly detection//Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Press: 938-946 [DOI:10.1609/aaai.v35i2.16177]

Sun S and Gong X. 2023. Long-short temporal co-teaching for weakly supervised video anomaly detection//2023 IEEE International Conference on Multimedia and Expo (ICME). Brisbane, Australia: IEEE: 2711-2716.

Zhu T, Yu Q, Song H, Cheng Y, Li S, Liu Y, et al. 2025a. DS-IB Net: Ultra-lightweight weakly-supervised video anomaly detection through synergistic dual streams and information bottleneck//Proceedings of the 7th ACM International Conference on Multimedia in Asia. Kuala Lumpur, Malaysia: ACM: 1-7 [DOI: 10.1145/3743093.3771006]

Ren X L, Zhang F F, Zhou W T and Zhou L. 2025. Complementary multi-type prompts for weakly-supervised temporal action location. Journal of Image and Graphics, 30(3): 842-854 (任小龙, 张飞飞, 周琬婷, 周玲. 2025. 多类型提示互补的弱监督时序动作定位. 中国图象图形学报, 30(3): 842-854) [DOI: 10.11834/jig.240354]

Sang N, Huang K Q, Zhao Y, Gao C X, Kao Y Y, Tan C C, Wang X, Wu M Q and Yin W T. 2026. Advances in video and image security research in the era of large models. Journal of Image and Graphics, 31(6): 2222-2259 (桑农, 黄凯奇, 赵耀, 高常鑫, 考月英, 谭创创, 王翔, 武美奇, 尹文体. 2026. 大模型时代的视频与图像安全研究进展. 中国图象图形学报, 31(6): 2222-2259) [DOI: 10.11834/jig.250656]

Feng J C, Hong F T and Zheng W S. 2021. MIST: Multiple instance self-training framework for video anomaly detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14009-14018 [DOI: 10.1109/CVPR46437.2021.01379]

Chen W, Ma K T, Yew Z J, Hur M and Khoo D A A. 2023b. TEVAD: Improved video anomaly detection with captions//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Vancouver, Canada: IEEE: 5549-5559 [DOI: 10.1109/CVPRW59228.2023.00587]

Ma J, Wang J, Zhang M and Zhou G. 2025. Skynet-V1: Towards early warning of video abnormal events via a spatial-temporal causal-

© 中国图象图形学报版权所有

enhanced MoE framework//Proceedings of the 33rd ACM International Conference on Multimedia. Dublin, Ireland: ACM: 6586-6595 [DOI: 10.1145/3746027.3755805]

Yang Z W, Liu J and Wu P. 2024. 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: IEEE: 18899-18908 [DOI: 10.1109/CVPR52733.2024.01788]

Dalvi J, Dabouei A, Dhanuka G and Xu M. 2025. Distilling aggregated knowledge for weakly-supervised video anomaly detection//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Tucson, AZ, USA: IEEE: 5439-5448 [DOI:10.1109/WACV61041.2025.00531]

Xu C, Li C G and Xing H J. 2025. Discriminative score suppression for weakly supervised video anomaly detection//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Tucson, AZ, USA: IEEE: 9569-9578 [DOI: 10.1109/WACV61041.2025.00928]

Huang C, Shi Y, Wen J, Wang W, Xu Y and Cao X. 2025. Ex-VAD: Explainable fine-grained video anomaly detection based on visual-language models//Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: PMLR.

作者简介

付淇(通信作者),博士,副教授,硕士生导师,主要研究方向为信息检索、多模态情感分析、视频异常检测等。E-mail: 419800621@qq.com

朱涛,硕士研究生,主要研究方向为计算机视觉、三维图像重建、视频异常检测。E-mail:taozhustu@foxmail.com

刘德喜,博士,教授,博士生导师。主要研究方向为信息检索、多模态情感分析、视频异常检测等。E-mail: liudexi@jxufe.edu.cn.

曹重华,博士,讲师,主要研究方向为人工智能、计算机视觉等。E-mail:caozhonghua@jxufe.edu.cn.

况思睿,硕士研究生,主要研究方向为信息检索、财会多模态智能分析、视频异常检测等。E-mail:kuangsirui@jxufe.edu.cn.

黄筠婷,本科生,主要研究方向为深度学习、计算机视觉、弱监督视频异常检测。E-mail:1254660507@qq.com

© 中国图象图形学报版权所有