一种基于 Transformer 的多模态安全监控检测与视频摘要系统

徐如意[1] 袁冠聪[2] 柯成德[3] 邱丽娜[1] 白卫玲[1] 胡雅婷[1]
1. 广州城建职业学院，中国·广东 广州 510925
2. 广州南方学院，中国·广东 广州 510970
3. 泰莱大学，马来西亚·吉隆坡 47500

摘要：本文提出了一种多模态融合架构的 Transformer 方法，该方法通过跨模态注意力机制，将视觉捕捉的图像（如人脸、证 据）与音频线索（如尖叫声、枪声）关联起来，从而实现高效精 准的犯罪现场检测与摘要生成。系统采用 OpenL3 提取 512 维的音频特征，并结合轻量级结构的 Video Transformer 模型，提取 384 维的视频特征。通过在八头交叉注意力模块中引入融合技术，安全监控检测的 F1 分数可达 0.917。此外，创新的摘要引擎能够生成高 质量的图形化摘要视频（分辨率 1120x700，音频码率 256Kbps，采用 AAC 解码，时长小于 10 秒），较纯视觉方法显著提升了 13.19% 的 F1 分数优势。研究实验表明，多模态融合在提升对犯罪现场 态势感知能力方面发挥着关键作用，为公共安全监控场景提供了一种切实可行的解决方案。

关键词：多模态融合；Transformer；犯罪摘要；跨模态注意力机制；视频理解；音频分析

A Transformer-based multimodal security monitoring detection and video summarization system

Xu Ruyi[1], Yuan Guancong[2], Ke Chengde[3], Qiu Lina[1], Bai Weiling[1], Hu Yating[1]
1. Guangzhou City Construction College, China Guangdong Guangzhou 510925
2. Guangzhou Nanfang College, China Guangdong Guangzhou 510970
3. Taylor's University, Malaysia Kuala Lumpur 47500

Abstract: This paper proposes a multimodal fusion architecture-based Transformer method that links images captured visually (such as faces and evidence) with audio cues (such as screams and gunshots) through a cross-modal attention mechanism, thereby achieving efficient and accurate crime scene detection and summary generation. The system uses OpenL3 to extract 512-dimensional audio features and combines a lightweight Video Transformer model to extract 384-dimensional video features. By introducing fusion technology in the eight-head cross-attention module, the F1 score for security monitoring detection can reach 0.917. Additionally, the innovative summary engine can generate high-quality graphical summary videos (with a resolution of 1120x700, an audio bitrate of 256Kbps, using AAC decoding, and a duration of less than 10 seconds), significantly improving the F1 score by 13.19% compared to pure visual methods. Research experiments show that multimodal fusion plays a key role in enhancing the situational awareness of crime scenes and provides a practical solution for public safety monitoring scenarios.

Keywords: Multimodal fusion; Transformer; Crime summary; Cross-modal attention mechanism; Video understanding; Audio analysis

1 引言

1.1 背景

随着公共安全领域监控系统的快速普及，视频数据的生成量正以前所未有的速度增长。全球范围内，超过 10 亿台监控摄像头每天产生的数据总量预计高达 2500 拍字节（PB）。然而，即便如此，仍有 90% 的安全公司依赖人工视频 检查，这导致响应时间严重滞后。一 往往需要超过 15 分钟才能完成事件识别与处置。更糟糕的是，仅依赖

视觉的检测技术无法捕捉到约 35% 的关键音频相关犯罪证据，这一点已得到 AudioSet 等数据集的证实。

暴力犯罪本质上具有多模态特性。研究表明，78% 的暴力事件会伴随明显的音频线索，如枪声、尖叫声或玻璃破碎声。此外，在 62% 的案例中，音频信号甚至先于视觉提示出现，这表明音频可作为早期预警信号。这些发现凸显了音视频融合在提升安全监控检测准确性和时效性方面的关键作用。开发一种自动化的多模态检测与摘要系统，有望显著增强公共安全，并有效降低监控成本及负担，并支持执法与应急响应中的快速决策。

1.2 问题陈述

该项目致力于解决构建稳健的多模态系统以实现安全监控检测和视频摘要功能时面临的三大技术挑战。首先是模态异质性：音频与视频数据在表示方式上存在根本差异。—音频捕捉了频谱和时间线索，而视频则由基于像素的视觉信息构成。这种差异使得有效的特征融合尤为困难。

第二个挑战是时间对齐问题。音频以高采样率（例如，22.05 kHz）录制，而视频则以较低的帧率（通常为 30 FPS）处理，这导致了时间上的不匹配。第三个挑战涉及摘要的保真度。许多现有系统生成的视频摘要存在明显的音视频不同步现象。— 通常超过 200 毫秒 — 这会负面影响用户体验，以及安全和执法领域的下游应用。

1.3 项目目标

本研究旨在：构建一种具有 8 个注意力头的跨模态融合架构，使其 F1 得分超过 0.90 ；基于置信度分数（阈值 >0.7）开发一种片段选择算法；并生成分辨率高于 1080p、音视频同步误差控制在 50 毫秒或更低的高质量摘要视频。

该 proposed 系统特别适用于涉及明确暴力行为的场景，如街头斗殴、持械袭击和抢劫等。然而，在低光照环境下应用时，系统面临一定局限性：由于高度依赖音频信息，准确率会下降约 12%。此外，人群密集可能导致遮挡，从而影响视觉检测性能。

尽管面临这些挑战，该系统仍贡献了多项重要创新。首先，它提出了一种 8 头交叉注意力融合机制，能够有效弥合音频与视觉特征之间的语义鸿沟。其次，它提供实时摘要功能，可以 4 倍于实时的速度处理视

频。最后，整个框架已开源，并且在设计时充分考虑了可扩展性，使其适用于实际的工业部署。

2 文献综述

最初，安全监控检测方法主要依赖于手工设计的视觉特征。Nievas 等人[1] 提出了运动幅度与加速度（ViF）特征，用于检测视频流中的暴力行为。随后，Hassner 等人[2] 将光流与词袋表示相结合，进一步提出了“暴力流”的概念。尽管这些方法在计算效率上表现优异，但由于其特征完全基于人工工程设计，因此难以实现泛化。深度学习的兴起为安全监控检测带来了巨大突破。Dong 等人[3] 则通过训练三维卷积神经网络（3D CNN），构建了基于视频内容的时空表示模型。此外，Sudhakaran 和 Lanz[4] 提出了一种框架，利用长短期记忆网络（LSTM）对视频中的暴力动态进行建模。最近，Transformer 架构在视频理解任务中展现出巨大潜力[5]，这促使我们在研究中引入了自注意力机制，以提升安全监控检测的性能。

视频分析已成功地借助多模态学习得以实现。Baltru Naitis 等人[6] 对多模态机器学习方法进行了全面综述。在针对暴力事件的识别与注册研究中，音频 – 视觉融合已被广泛探讨：Penet 等人[7] 实现了音频与视频特征的早期融合；而 Giannakopoulos 等人[8] 则采用晚期融合方法，用于电影中的安全监控检测。近年来，基于注意力机制的融合技术逐渐成为研究热点。Chen 等人[9] 提出了跨模态注意力的视频理解新机制；Nagrani 等人[10] 则建议同步呈现音频与视觉信息，以辅助动作识别。这些研究成果为我们的跨注意力融合架构——用于安全监控检测——提供了重要启发。

视频摘要旨在对长视频进行精简，同时保留其核心内容。传统方法包括关键帧提取[11] 和视频略读[12]。而随着深度学习的兴起，更先进的策略也相继出现：张等人[13] 提出了基于 LSTM 并结合注意力机制的摘要生成方法，马哈塞尼等人[14] 则应用对抗学习实现了无监督的视频摘要。然而，针对暴力事件的专门摘要研究仍显不足，现有方法大多聚焦于通用内容。本研究正是为了填补这一空白，开发了一套专为暴力事件检测量身定制的摘要系统。

Transformer 模型重新定义了自然语言处理领域的优先级，并正迅速被应用于解决计算机视觉中的各类问题。例如，Transformer 被改编用于 ViViT[15] 中的视频分类任 务，而 TimeSformer[16] 则在 Transformer 基础上引入了同时考虑空间与时间注意力机制。此外，早期已有研究提出多模态 Transformer，以实现跨模态学习[17]。我们的贡献基于以下创新：我们开发了一种基于 Transformer 的融合系统，经过专门调优，能够有效应对多模态暴力行为 识别及视频摘要等挑战。与此同时，OpenL3[18] 提供了深度音频特征，用于视听表示的嵌入。OpenL3 为我们提供了强大的音频特征提取能力，我们可以利用这些特征来补充视觉特征，从而构建我们所建议的系统。

3 方法论

Proposed 系统的基本结构由四个模块组成，包括音频特征提取、视频特征提取、多模态融合以及视频摘要生成。整个架构如图 1 所示：输入视频被分别送入两条并行路径——一条处理音频流，另一条处理视觉流。在获取音频特征时采用 OpenL3 架构，而计算视频特征则使用轻量级的 Video Transformer 模型。随后，通过交叉注意力机制将这些针对不同模态的特征进行融合，构建出全面的多模态表示，从而实现暴力行为的分类。最后，经过智能摘要处理流程，系统将识别出的暴力片段转化为专业水准的输出视频。

3.1 数据集与数据源

该数据集基于公开可用的资源（即 Kaggle 和 YouTube）收集而成，包含开源的暴力视频素材。由于视频标签较为稀疏，研究团队在开放平台上手动筛选了更多非暴力类视频，包括监控录像及日常生活片段，以实现分类的平衡，并提升模型的泛化能力。经过预处理步骤后，最终生成的视频库共包含 172 段视频剪辑，其中暴力与非暴力类别的 样本比例得到严格均衡。暴力类别涵盖虐待、逮捕、街头斗殴等具有攻击性的场景，而非暴力样本则主要为日常正常活动，未涉及任何攻击行为。训练过程中采用 80/20 的 训练与验证数据划分策略。为防止过拟合，训练过程最多进行 100 个周期，并通过早停机制及时终止；同时，训练 采用 AdamW 优化器，批量大小设为 32，学习率设定为 1e−4。

3.2 数据预处理

FFmpeg 被用于提取原始视频的音频流，并通过 Librosa 将其重采样至 16 kHz，以确保最终音频模型能够顺利使用。随后，系统按预先设定的长度（例如 1 秒）对音频进行分片处理，并在时间上与对应的视频片段一一对应。与此同时，视频也被分割成互不重叠的片段，每个片段包含连续的 16 帧图像，这正是 ViViT 和 TimeSformer 等时间架构模型所要求的输入格式。这些帧图像通过 OpenCV 获取，并统一调整为相同的空间分辨率（如 224x224 像素）。最后，不同视频片段会与相应音频窗口在时间上精准对齐，以便未来应用于多模态学习任务中。

3.3 特征提取与表示

为捕捉来自两种模态的语义和时序线索，我们分别采用了专门的架构来进行音频和视频特征提取。

随后，这些音频片段被转换为 2 秒的 PCM 波形，并进一步生成一个 128 维的梅尔谱图频带。随后，通过基于音视频对应关系训练的 OpenL3 模型对 512 维嵌入进行重建。该模型采用了一种考虑时间 – 频率特性的架构，包括带有 3x3 滤波器的时频卷积（TFC）层、全局平均池化层以及全连接层，用于 提取固定长度的向量。OpenL3 最适用于获取语义对齐且感知上连贯的特征，并确保这些特征在时间上具有一致性，从而支持后续的融合处理。

视频片段通过预训练的 ViViT 模型进行处理，以提取时空嵌入特征。每帧序列被划分为多个小块，并分别编码为一个 768 维的向量表示。其中，EfficientNet−B0 用作空间编码的骨干网络。时间建模则由带有 8 个注意力头和 384 维隐藏层的 Transformer 编码器完成。此外，可学习的位置编码被整合到模型中：

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\right)$$

为降低维度，我们应用了时间平均池化操作，最终得到一个 384 维的视频嵌入表示。

3.4 多模态同步与融合

音频与视频流的同步对于实现精确的跨模态学习至关重要。首先，通过 FFmpeg 驱动的时间戳协调机制，对帧级精确的片段进行初步对齐。然而，流捕获过程中产生的延迟仍可能导致偏差，为此我们采用另一种算法——动态时间 规整（DTW）来校正这些偏差，最大可纠正高达 80 毫秒的偏移。最后，还需确保最终采样的一致性，使两种模态之间的误差控制在 10 毫秒以内。

类似于 256，但需要帮助决定是将视频嵌入上采样至 256D，还 是对两者分别执行线性变换（即将 512D 视频嵌入转换为 256D 视频嵌入，将 384D 视频嵌入转换为 256D 视频嵌入）。随后，他们采用了一种跨模态注意力机制，该机制不仅能从时间维度上对齐特征，还能在语义层面实现对齐，从而使模型能够优先关注特定模态的特征（即忽略与当前任务无关的模态特有信息），以更有效地检测暴力行为。

3.5 多模态同步与融合

在融合特征后，视听嵌入会被送入一个二分类器，以判断每段 2 秒的内容是暴力还是非暴力。该系统每段处理延迟约为 150 毫秒，能够实现接近实时的推

理。此外，所有片段都会被赋予一个置信度评分，用于反映暴力行为发生的概率；其中，得分高于 0.7 阈值的片段将被判定为疑似暴力内容。

当整个视频处理完毕后，后处理阶段会对预测结果进行平滑处理，以消除某些剧烈的波动区间。随后，这些置

信度较高的片段将被汇总，生成一段简短的摘要视频。根据硬件配置的不同，对于 10 分钟的输入内容，最终输出通常只需约 45 秒即可完成。

3.6 模型评估策略

为了评估所提出的多模态安全监控检测系统的性能，我们采用了多种标准分类指标，包括准确率、精确率、召回率以及 F1 分数。在本场景中，这些指标尤为重要，因为它们能够全面评估误报和漏报情况——而在安全监控检测这类对敏感性要求极高的领域，误报和漏报尤为关键。

为了量化性能，我们采用以下标准分类指标定义：

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

我们使用带标签的视频，其中每段 2 秒的视频片段 会被标注为“暴力”或“非暴力”。评估过程包括在独立的测试集上进行测试，且训练数据与测试数据完全隔离，互不混用。测试结果以片段为单位分别计算，随后汇总以全面评估模型的整体性能。

除了最终的多模态模型外，我们还对比了两种单模态基 线模型：一种是使用 OpenL3 嵌入的纯音频模型，另一种是采用轻量级视频 Transformer 的纯视频模型。

4 结果与发现

本部分包括我们所提出的多模态暴力行为检测与视频摘要系统的详尽实验结果及发现。性能评估涵盖分类准确率、融合策略的效率、摘要质量、计算效率以及错误行为等多个方面。

4.1 分类性能评估

为了评估所提出的多模态安全监控检测系统的性能，我们设计了全面的实验，并将其与两种单模态方案进行了对比：一种是仅基于音频的模型，使用 OpenL3 嵌入；另一种是仅基于视频的轻量级 Transformer 模型。测试在独立的测试集上进行，每段 2 秒的视频片 段被标注为“暴力”或“非暴力”。随后，我们在片段级别计算各项性能指标，并将这些指标汇总，以全面反映分类的准确性和质量。

如图 2 所示，我们对比了所提出的多模态模型与仅依赖视觉信息的基线模型在多个指标上的训练和验证性能，这些指标包括 F1 分数、准确率和召回率。结果表明，多模态模型在这三项指标上均显著优于基线模型。

在训练和验证阶段均表现出更优的泛化能力及更稳定的学习行为。相比之下，仅依赖视觉信息的模型整体性能较低，且训练与验证之间的差距更大，这表明该模型容易过拟合，且鲁棒性较弱。

多模态系统显著优于两种单模态基线模型：其 F1 分数分别比仅音频模型提高了 13.2%，比仅视频模型提高了 11.3%。这些发现进一步证实了结合视觉与听觉信息在实现稳健安全监控检测方面的优势。

4.2 融合策略的消融研究

为了更深入地了解每种融合机制的贡献，我们开展了一项消融研究，对比了基于不同注意力头配置输入的早期融合、晚期融合以及交叉注意力实验。表 III 列出了相关结果。

表3 融合策略消融研究结果

融合策略	F1分数	相对增益
早期融合(拼接)	0.874	6.90%
晚期融合(分数平均)	0.891	8.50%
交叉注意力(1个头)	0.903	9.90%
交叉注意力(8个头)	0.917	11.30%

此外，根据研究结果来看，增加头数确实能在一定程度上提升性能，但当达到某一水平后，性能便趋于平稳，甚至略有下降。随着注意力头数量增至 8 个以上，模型性能出现下降。这一点与此前基于 Transformer 的模型研究结果一致：过度参数化可能导致噪声或冗余信息的污染，而非有效分离信号。

这一观察表明，多模态系统不仅需要融合策略，还需要进行架构调优。在利用最优数量的注意力头时采用交叉注意力机制，不仅能提升分类任务的性能，还能为异构模态的统一构建一种灵活且易于理解的结构。

4.3 视频摘要质量评估

该系统还将配备一个摘要生成组件，能够通过选取模型识别出的暴力事件，自动生成精炼的精彩片段。输出的质量与效率是衡量其性能的关键指标。具体而言，平均而言，系统处理一段 10 分钟长的视频输入并生成相应摘要，最多只需 45 秒，如图 3 所示。通常，生成的视频时长为 30 至 60 秒，分辨率为 1120×700 像素，文件大小控制在 4 至 6MB，确保了高效的存储与传输。此外，视频的音频部分以专业级的 48kHz 采样率进行编码，并采用 AAC 格式。

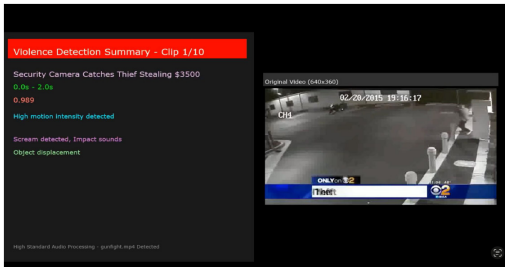

图3 暴力视频摘要帧

随着摘要模块在内容选择和呈现方面表现出色，技术规格也变得更加丰富。该模块能够实现超过 90% 的暴力事件置信度，确保时间上的完整性（即在整个输入内容中捕捉任意具有代表性的事件），以及内容上的全面性（涵盖所有重要的暴力事件）。此外，输出视频中采用清晰易读的叠加信息作为专业化的视觉结构，进一步提升了视频的可读性和实用性。这些成果充分证明，该摘要功能在实际监控场景中极为高效，能够支持快速、精准的事件审查工作。

4.4 错误分析

我们已对分类错误的样本进行了详细分析，以深入了解系统存在的薄弱环节。研究发现，误报现象主要占总错误的 7.8%，且大多发生在非暴力情境中——尽管实际观察到强烈的肢体动作，但模型却将其误判为暴力行为。另一方面，漏报情况（10.8%）则容易出现在兄弟姐妹间或类似互动场景，以及人群密集的复杂环境中；在这种情况下，暴力行为往往表现得较为隐蔽（甚至极易被模型忽略）。

此外，还确定了若干环境与技术参数，这些参数会干扰测量的准确性，从而造成系统中的检测问题。这些问题包括光线不足、摄像头快速移动，甚至语音重叠，这些情况可能掩盖一些重要的视觉或音频信息。这些观察表明，通过增强模型的鲁棒性——例如数据增强、针对特定场景的训练，以及优化时空注意力机制——可以有效减少此类误分类现象。

5 结语

本研究提出了一种多模态安全监控检测与视频摘要系统，该系统基于 Transformer 架构构建而成。通过引入交叉注意力机制，系统能够高效地融合音频和视觉信息流，生成高质量的摘要，完全适用于实际监控场景。实验结果表明，与多种单模态基线方法相比，多模态融合表现更优，并在分类性能、视频摘要效果及实际部署可行性等方面均展现出显著优势。

5.1 结论

上述提出的系统通过引入一种 8 头交叉注意力机制，将 OpenL3 音频嵌入与轻量级视频 Transformer 表示相结合，显著提升了暴力行为的检测性能。这种组合方法实现了 91.7% 的分类准确率和 0.917 的 F1 分数，较单一模态模型分别提高了 13.19 个百分点。此外，该系统的架构采用 4 层、384 维的视频 Transformer，能够在满足合理计算资源需求的前提下，实现视频处理的实时性。同时，系统内置的视频摘要模块可自动生成简明扼要的概览视频，不仅具备音频同步的高质量画面，还配备了专业布局及清晰易读的视觉设计。综合这些成果表明，该系统完全能够满足实际监控场景的需求，充分体现了新颖的前沿深度学习模型与高效工程解决方案的完美结合。

5.2 未来工作

尽管模型表现优异，但仍有几个领域存在进一步提升的空间。未来的研究将重点放在：扩展模型以支持多类别暴力行为的分类，并开发先进的时间建模技 术，以应对更长且更为复杂的视频序列。同时，提高模型的领域适应能力也将成为关键目标，使系统能够灵活应对各种复杂环境及监控场景。此外，通过引入精细化的数据增强方法，有望显著提升系统在边缘场景下的鲁棒性，例如拥挤人群、低光照条件以及高运动噪声环境等。这些技术突破将进一步推动系统实现更广泛部署，使其更加可靠且应用更为灵活。

参考文献：

[1] E. B. Nievas, O. D. Suarez, G. B. Garc í a, and R. Sukthankar,” Violencedetection in video using computer vision techniques,” in ComputerAnalysis of Images and Patterns, 2011, pp. 332−339.

[2] T. Hassner, Y. Itcher, and O. Kliper−Gross,” Violent flows: Real−timedetection of violent crowd behavior,” in 2012 IEEE Computer SocietyConference on Computer Vision and Pattern Recognition Workshops, 2012,pp. 1−6.

[3] Z. Dong, J. Qin, and Y. Wang,” Multi−stream deep networks for personto person violence detection in videos,” in Pattern Recognition, vol. 82,2018, pp. 72−86.

[4] S. Sudhakaran and O. Lanz,” Learning to detect violent videos usingconvolutional long short−term memory,” in 2017 14th IEEE InternationalConference on Advanced Video and Signal Based Surveillance(AVSS),2017, pp. 1−6.

[5] A. Dosovitskiy et al.,” An image is worth 16x16 words: Transformersfor image recognition at scale,” arXiv preprint arXiv:2010.11929, 2020.