程清楠, 杨晓昶. 基于改进 SlowFast 的暴力行为识别方法[J]. 智能计算机与应用,2026, 16(8):206-211. DOI:10.20169/j.issn.2095-2163.25012101

基于改进 SlowFast 的暴力行为识别方法

程清楠, 杨晓昶
(上海第二工业大学 计算机与信息工程学院, 上海 201209)

摘 要: 人体动作识别在视频监控、人机交互和医疗保健等领域具有广泛应用,其中暴力行为识别尤为重要,暴力事件因其突发性和偶发性,对社会稳定构成巨大威胁。本文旨在提高暴力行为识别的准确性,解决传统人工监控方式难以高效覆盖所有场景的问题。针对暴力事件的突发性和偶发性,本文在 SlowFast 模型基础上提出了一种改进的方法。首先,引入 3D 时空通道注意力模块,通过结合慢网络中的时序、空间、通道信息,提升暴力行为发生瞬间的判断精度,减少误判和漏判。其次,增加时空增强步骤,对双流特征进行优化,突出暴力行为的关键时空区域,提升了模型对细微变化的敏感性。实验结果表明,改进后的模型在暴力行为识别任务中的准确性和效率得到了明显提升,为自动化视频监控和安全预警系统提供了技术支持。

关键词: 视频动作识别; SlowFast; 暴力行为识别; CBAM; 智能监控
中图分类号: TP391.41 **文献标志码**: A **文章编号**: 2095-2163(2026)08-0206-06

Violence identification method based on improved SlowFast

CHENG Qingnan, YANG Xiaochang
(School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai 201209, China)

Abstract: Human action recognition has broad applications in fields such as video surveillance, human-computer interaction, and healthcare, with violent behavior recognition being particularly important due to the sudden and occasional nature of violent incidents, which pose a significant threat to social stability. This paper aims to improve the accuracy of violent behavior recognition and address the issue of traditional manual monitoring being unable to efficiently cover all scenarios. To tackle the suddenness and occasional nature of violent events, an improved method based on the SlowFast model is proposed. First, a 3D spatiotemporal channel attention module is introduced, which combines temporal information from the slow network with spatial and channel information, significantly enhancing the accuracy of identifying violent behaviors at critical moments, thereby reducing false positives and false negatives. Second, a spatiotemporal enhancement step is added to optimize the dual-stream features, emphasizing the key spatiotemporal regions of violent behaviors and improving the model's sensitivity to subtle changes. Experimental results show that the improved model significantly enhances the accuracy and efficiency of violent behavior recognition tasks, providing technical support for automated video surveillance and security warning systems.

Key words: video action recognition; SlowFast; violent behavior recognition; CBAM; intelligent surveillance

0 引 言

人体动作识别在视频监控、人机交互、医疗保健、智能人机界面等领域具有广泛的应用场景[1-4]。暴力行为识别是人体行为识别及视频理解的重要子领域,由于暴力事件具有偶发性的特点,部署大量人力进行监控不仅耗费资源还容易产生漏检和误判问题。同时,暴力事件的发生往往会对社会秩序和稳定性构成较大的威胁,因此通过自动化暴力行为检测系统能够有效提升检测效率,精准识别暴力行为,避免因资源有限导致的监控盲区,节省人力与物力成本[5]。

暴力行为识别领域的深度学习方法主要可分为 3 类:双流卷积网络、3D 卷积网络和长短期记忆网络(LSTM)。双流卷积网络通过分别处理空间流和时间流,从静态视频帧和密集光流中提取特征,有效整合了时空信息[6]。然而,其对光流的依赖限制了方法的应用场景与性能上限。Tran 等[7] 提出的

作者简介: 程清楠(2001—),女,硕士研究生。主要研究方向:图像处理与机器视觉; 杨晓昶 (1997—),男,硕士研究生。主要研究方向:图像处理与机器视觉。Email:1478857785@qq.com。

收稿日期: 2025-01-21

C3D 网络通过三维卷积从视频帧中直接提取时空特征，避免了对光流等预处理步骤的需求，但仍未充分挖掘时空特征的交互性。LSTM 能够解决卷积神经网络（CNN）在建模长时间依赖时的不足，但由于其依赖 CNN 全连接层的特征，缺乏对时空细节的精确捕捉能力[8]。

SlowFast 模型结合了双流网络的优势，采用快慢网络结构：慢网络（Slow Path）以低帧率提取语义信息，而快网络（Fast Path）则以高帧率捕捉快速变化的动态信息，聚焦短时间尺度的动作变化[9]。与传统方法相比，SlowFast 模型通过优化时空特征的捕捉与融合，展现出更强的表现力与灵活性，为暴力行为识别提供了一个理想的基础框架。其结构如图 1 所示。

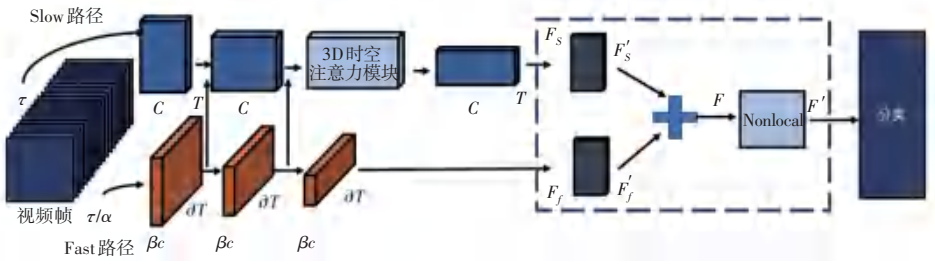


图 1　改进的 SlowFast 模型结构图

Fig. 1　Structure diagram of the improved SlowFast model

1　改进方法

在暴力行为识别任务中，如何有效地从视频中提取并区分暴力行为与非暴力行为的细节至关重要。虽然 SlowFast 模型通过其快慢网络结构能够捕捉到不同时间尺度的特征，但在暴力行为的细微识别上，时空信息的表达与交互仍存在一定的局限性[10]。暴力行为通常具有快速、短暂且突发的特征，这使得时空信息的精细提取尤为关键。针对这一挑战，本文提出以下改进方案：

（1）3D 时空通道注意力模块的引入：本文提出一种基于 CBAM 的时空通道注意力模块，将 SlowFast 模型中慢网络的时序信息与空间信息相结合，以时序信息辅助空间信息与通道信息，提高空间信息与通道信息提取的准确度。该模块不仅有助于提高暴力行为的识别准确度，还能增强模型对暴力行为边界的判断力，减少误判和漏判。

（2）时空信息增强步骤的引入：本文利用 Non-Local 模块对 SlowFast 模型的双流信息进行时空特征增强，通过对双流特征进行精细的时空增强，有效突出暴力行为中的关键时空区域。模型可以更准确地聚焦于暴力行为发生的瞬间，提高对暴力行为的特征敏感性。

1.1　3D 时空通道注意力

SlowFast 模型的慢网络专注于低帧率、长时间跨度的视频特征，适合捕捉视频的全局空间和时间信息[11]，因此，在 SlowFast 网络的慢路径退化块（Degraded Bottleneck Block，DBB）中嵌套注意力模块，能够平衡全局信息与局部细节的提取，尤其是在暴力行为的复杂动作场景下，可显著提高识别精度。

在 SlowFast 模型的慢网络的退化块中包含两个（1 * 1 * 1）卷积和一个（1 * 3 * 3）卷积，前者主要用于特征维度的变换，后者则通过空间域卷积操作捕捉图像的结构信息与空间关系，是慢网络理解和表征图像的重要层级[12-15]。因此本文在（1 * 3 * 3）的卷积块后引入 3D 时空通道注意力模块，过滤冗余信息，增强关键特征，提升该卷积层提取信息的准确性，使网络更关注关键的空间区域。而且，提取后的信号通过与残差路径连接，能够将此时的注意力权重影响持续到后续的信息流动，提高网络的整体表达能力。具体嵌套结构如图 2 所示。

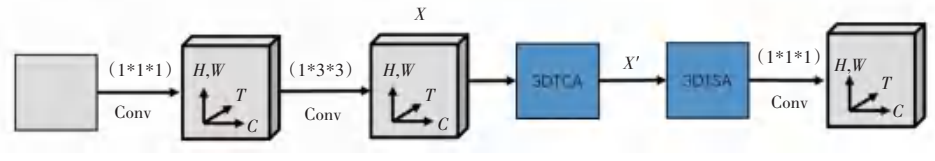


图 2　注意力模块嵌套结构

Fig. 2　Nested structure of attention modules

本文提出的注意力模块基于 CBAM[16] 构建,使用平均池化与最大池化分别提取不同特征,并在此基础上采用 3D 卷积进行通道信息之间的相互注意,保证局部特征不被遗漏的同时也将时序信息对通道信息的影响考虑在内,以时序信息辅助空间信息与通道信息,提升二者提取的准确度。其中,时间通道注意力(3D Temporal-Channel Attention, 3DTCA)具体结构如图 3 所示,时间空间注意力(3D Temporal-Spatial Attention, 3DTSA)具体结构如图 4 所示。

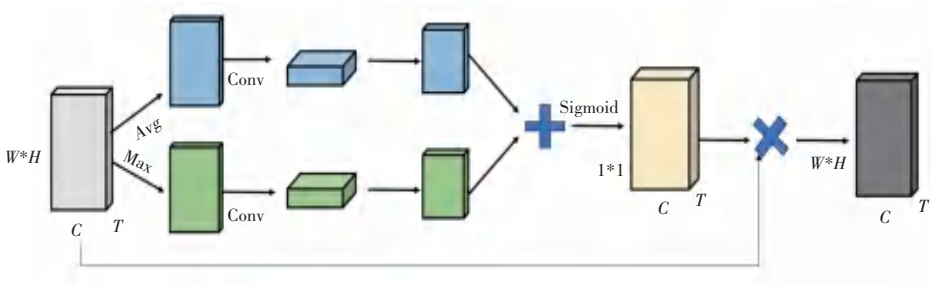


图 3 时间通道注意力

Fig. 3 3D Temporal-channel attention

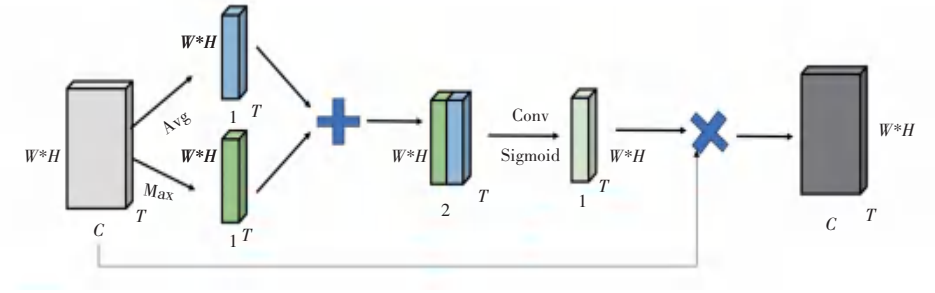


图 4 时间空间注意力

Fig. 4 3D Temporal-spatial attention

注意力模块总体设计采用串联设计,即先通过时间通道注意力模块,再通过时间空间注意力模块。在 3DTCA 模块中,输入特征图为 $X \in R^{C\times T\times H\times W}$,其中 C 为通道数,T 为输入帧数,H、W 为输入图像的长和宽,经过池化,对时间维度分别进行全局平均池化和全局最大池化得到:$\mathrm{AvgPool}_{\mathrm{time}}(X) \in R^{C\times 1\times H\times W}$ 和 $\mathrm{MaxPool}_{\mathrm{time}}(X) \in R^{C\times 1\times H\times W}$。将池化后的 2 个特征图特征进行拼接,送入共享的多层感知机(MLP),计算公式如下:

$$M_{\mathrm{channel}} = \sigma(\mathrm{MLP}(\mathrm{Concat}(\mathrm{AvgPool}_{\mathrm{time}}(X), \mathrm{MaxPool}_{\mathrm{time}}(X)))) \tag{1}$$

激活函数为 ReLU 和 Sigmoid,得到通道注意力权重为 $M_{\mathrm{channel}} \in R^{C\times 1\times H\times W}$。最后将通道注意力的输出权重 M_{channel} 与输入特征映射 X 按元素相乘,最终得到通道特征映射,最终输出记为 $X' \in R^{C\times T\times H\times W}$,计算公式如下:

$$X' = X \cdot M_{\mathrm{channel}} \tag{2}$$

将 3DTCA 模块输出的特征图送入 3DTSA 模块作为输入。首先,对特征映射 X' 进行三维全局最大池化和全局平均池化,由此产生两个特征:$\mathrm{AvgPool}_{\mathrm{channe}}(X') \in R^{1\times T\times H\times W}$ 和 $\mathrm{MaxPool}_{\mathrm{channe}}(X') \in R^{1\times T\times H\times W}$。沿着通道维度将池化后的特征拼接,并通过(1 * 1 * 1)3D 卷积运算对通道维度进行线性组合,实现跨通道的信息交互和融合,再通过 Sigmoid 激活函数得到空间时间注意模块的输出特征映射:$M_{\mathrm{spatial}} \in R^{1\times T\times H\times W}$,其计算公式如下:

$$M_{\mathrm{spatial}} = \sigma(\mathrm{Conv3D}(\mathrm{Concat}(\mathrm{AvgPool}_{\mathrm{channel}}(X'), \mathrm{MaxPool}_{\mathrm{channel}}(X')))) \tag{3}$$

最后,通道注意力的输出特征 M_{spatial} 与输入特征映射 X' 按元素相乘,生成最终的通道特征映射,记为 $X'' \in R^{C\times T\times H\times W}$,公式如下:

$$X'' = X' \cdot M_{\mathrm{spatial}} \tag{4}$$

1.2 时空信息增强

SlowFast 模型中快慢网络的信息经过卷积块以及时空通道注意力模块的提取之后,包含了大量的时空与通道信息[17],本文引入 Non-Local 模块实现时空信息增强,对快慢网络双流信息融合后的信息进行时空信息的增强处理,实现对全局时空依赖关系的建模,突出双流特征中的关键时空区域,进一步增强 SlowFast 模型在暴力行为检测方面的表达能

力，特别是在复杂场景中有助于提升分类准确率，实现快流和慢流特征的优势互补。

Non-Local 模块的核心思想是计算任意两个位置特征的关系，并通过加权汇聚生成增强特征[18]。其基本操作公式如下：

$$y_i = \frac{1}{C(x)} \sum_j f(x_i, x_j) g(x_j) \tag{5}$$

其中，x_i、x_j 分别表示输入特征图中第 i 和 j 位置的特征；$g(x_j)$ 是对特征 x_i、x_j 的变换函数，本文使用 1 * 1 卷积的线性变换；$f(x_i, x_j)$ 是相似性函数，其具体计算公式如下：

$$f(x_i, x_j) = \text{Softmax}(x_i^{\mathrm{T}} x_j) = \frac{e^{x_i^{\mathrm{T}} x_j}}{\sum_k e^{x_i^{\mathrm{T}} x_j}} \tag{6}$$

具体来说，输入为快慢网络拼接后的融合特征：$F_{\text{fusion}} \in R^{C \times T \times H \times W}$，其中 C 为通道数，T、H、W 分别为时间步、空间高度和宽度。Non-Local 模块首先对其进行特征变换，本文使用两个独立的 1 * 1 卷积将输入特征映射为查询特征 Q 和键特征 K，计算公式如下：

$$Q = W_Q * F_{\text{fusion}}, K = W_K * F_{\text{fusion}} \tag{7}$$

之后即计算查询特征和键特征之间的相似性并使用相似性权重对值特征进行加权汇聚，其计算公式如下：

$$F_{\text{enhanced}}(i) = \sum_j S(i,j) V(j) = \sum_j \text{Softmax}(Q_i^{\mathrm{T}} K_j) V(j) \tag{8}$$

最终将 F_{enhanced} 与 F_{fusion} 相加得到输出特征。通过 Non-Local 模块对任意位置的全局时空关系建模，在 SlowFast 模型快慢网络融合的基础上实现进一步优化，更突出关键时空区域的信息，不仅提升模型鲁棒性与分类准确率，也让模型更适用于暴力行为检测领域。

1.3　损失函数

暴力行为识别是一个二分类任务，因此本文采用了二分类交叉熵损失函数（Binary Cross-Entropy Loss），也称为对数损失函数，这是二分类问题中广泛应用的标准损失函数[19]。该损失函数的核心作用是衡量模型预测的概率分布与实际标签的概率分布之间的差异。其直接优化概率输出，使模型能够更精确地预测类别概率。

暴力行为识别的目标是区分暴力与非暴力行为，该损失函数能推动模型在决策边界附近做出更精确的判断，从而提高分类性能，特别是在面对复杂的、细微差异的暴力行为时，能够保证更好的泛化能力和鲁棒性[20]。定义公式如下：

$$L(y, y') = -\frac{1}{N} \sum_N [y_i \log(y'_i) + (1 - y_i) \log(1 - y'_i)] \tag{9}$$

其中，y_i 是第 i 个样本的实际标签，y'_i 是第 i 个样本的预测概率。

2　实验结果与分析

2.1　实验设置与数据集

本文模型训练使用 Ubuntu 20. 04 系统、显卡为 NVIDIA 3080、深度学习框架采用 PyTorch 1. 8 搭配 CUDA 10. 1，编程环境为 Python 3. 8。

本文实验面向社区环境下的暴力行为检测，视频数据主要来源于监控画面，采用 Real Life Violence Situations[14] 数据集进行训练和评估。该数据集旨在弥补现实生活中人际之间暴力相关数据不足而构建的大型公开数据集，该数据集包含了 1 000 个暴力视频和 1 000 个非暴力视频，涵盖了不同环境、多种形式的暴力行为。数据主要来源于 YouTube 平台，其中，暴力视频包括多种现实街头打斗场景，涉及多种环境与条件，非暴力视频则包括体育运动、日常行走等人类活动。实验中将数据集按 7∶2∶1 的比例划分为训练集、测试集与验证集。

2.2　评价指标

本文采用精确度（Precision，P）、召回率（Recall，R）、$F1$ 分数以及模型参数量作为算法的评价指标。精确度表示模型预测的正样本中有真实正样本的比例；召回率表示预测结果中正确预测样本的比例；$F1$ 分数则作为精确度和召回率的调和平均数，能够综合考虑精确度和召回率的平衡，而模型参数量则直接影响模型的计算复杂度、存储需求和推理速度等。具体计算公式如下：

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{12}$$

其中，TP 表示被正确预测的正样本；FP 表示被误判为正样本的负样本；FN 表示被误判为负样本的正样本。

2.3　实验分析

2.3.1　暴力行为识别结果

为验证本文提出的改进 SlowFast 模型的暴力行

为识别方法的识别效果与优越性,将改进后的模型与原模型进行对比实验。对比结果见表 1。

表 1 模型改进前后的性能对比

Table 1 Comparison of model performance before and after improvements %

方法	Precision	Recall	*F*1 Score
SlowFast	87. 6	74. 9	80. 8
本文方法	90. 4	88. 7	89. 5

由表 1 可以看出,改进后的方法相比原始的 SlowFast 模型在识别精确度与召回率上均有不同幅度的提升,图 5 及图 6 分别为改进后的模型以及原始模型的识别结果。对比可见,优化后的模型在人群密集的场所、存在部分遮挡或光线较暗的场景下均能实现精准的识别。而原始模型则无法识别出马赛克模糊的部分,以及在昏暗拥挤场景下均未能完成准确检测。



图 5 优化后模型识别结果

Fig. 5 Optimized model recognition result



图 6 原始模型识别结果

Fig. 6 Original model recognition results

为进一步验证本文算法的性能,本文与其他现有模型进行了对比,结果见表 2。

由表 2 可见,本文方法在精确度、召回率、*F*1 分数方面均优于其他对比模型,表明本文方法在检测准确性和稳定性上具有优势。同时,本文模型参数量和大小也相对适中,适合在计算资源较为充裕的环境中应用。为后续开展自动化视频监控和安全预警系统提供了有力的技术支持。

表 2 与现有模型的性能对比

Table 2 Performance comparison with existing models

方法	Precision/ %	Recall/ %	*F*1 Score/ %	模型参数量/ 百万
SlowFast	87. 6	74. 9	80. 8	22
I3D	89. 0	76. 0	82. 2	27
ResNet+2D 卷积	85. 5	72. 5	78. 6	10
本文方法	90. 4	88. 7	89. 5	20

2.3.2　消融实验

为进一步验证本文改进方案的有效性，本文采用了不同的改进策略进行消融实验，得到实验结果对比见表 3。

表 3　消融实验

Table 3　Ablation experiment

算法	3DTSA	3DTCA	Non-local block	Precision/%	Recall/%
SlowFast	×	×	×	87.6	74.9
模型 1	√	×	×	87.5	86.1
模型 2	×	√	×	90.9	85.4
模型 3	×	×	√	89.7	87.3
模型 4	√	√	√	90.4	88.7

其中，模型 1 至模型 3 分别为加入时间通道注意力模块、时间空间注意力模块、时空增强模块的模型，模型 4 为本文提出的完整模型，可以看出，时空通道注意力模块的引入与时空信息增强步骤的增加，均可提升 SlowFast 模型对暴力行为的识别精确度。而同时使用三者，不仅可以利用时空通道注意力模块极大地提升时空通道信息的提取能力，而且通过时空信息增强可以极大地提升这些信息的交互与利用效率，最终使识别精确度提升 2.8%。

3　结束语

针对暴力行为识别实际需求，本文提出一种改进的 SlowFast 暴力行为识别模型。在 SlowFast 模型中慢网络的退化块中引入时空通道注意力模块，结合慢网络的时序信息与空间特征，增强暴力行为细节的提取能力；增加时空信息增强步骤，通过 Non-Local 模块有效地捕捉全局时空依赖关系，强化模型对关键时空区域的识别能力。本文提出的改进后的模型在 Real Life Violence Situations 数据集上的识别准确率提升了 2.71%，证实了该改进方法的有效性，为暴力行为识别领域提供了新的研究视角和方法。

参考文献

[1] 朱相华，智敏，殷雁君. 基于 2D CNN 和 Transformer 的人体动作识别[J]. 电子测量技术，2022，45(15)：123-129.

[2] Sajjad M，Ullah F U M，Ullah M，et al. A comprehensive survey on deep facial expression recognition：Challenges，applications，and future guidelines[J]. Alexandria Engineering Journal，2023，68(4)：817-840.

[3] Huang D Y，Lin C L，Chen Y. Securable networked scheme with face authentication[J]. IET Biometrics，2022，11(2)：97-108.

[4] Jayaraman U，Gupta P，Gupta S，et al. Recent development in face recognition[J]. Neurocomputing，2020，408：231245.

[5] Vijeikis R，Raudonis V，Dervinis G. Efficient violence detection in surveillance[J]. Sensors，2022，22(6)：2216.

[6] Simonyan K，Zisserman A. Two-stream convolutional networks for action recognition in videos[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. NeurIPS，2014：568-576.

[7] Tran D，Bourdev L，Fergus R，et al. Learning spatio-temporal features with 3D convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway，NJ：IEEE，2015：4489-4497.

[8] Hochreiter S，Schmidhuber J. Long Short-Term Memory[J]. Neural Computation，1997，9(8)：1735-1780.

[9] Jayaraman U，Gupta P，Gupta S，et al. Recent development in face recognition[J]. Neurocomputing，2020，408：231-245.

[10] 周璇，易剑平. 基于时间上下文模块的人体动作识别方法[J]. 国外电子测量技术，2022，41(10)：72-79.

[11] Feichtenhofer C，Fan H，Malik J，et al. SlowFast networks for video recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2019：6202-6211.

[12] 袁野，黄丽清，叶锋，等. 基于集成学习双流神经网络的实时面部篡改视频检测模型[J]. 计算机工程与科学，2023，45(3)：470-477.

[13] Li S B，Wang Z Y，Liu Y X，et al. FSformer：Fast-slow transformer for video action recognition[J]. Image and Vision Computing，2023，137：104740.

[14] Zhang Y，Ibrayim M，Hamdulla A. Research on cow behavior recognition based on improved SlowFast with 3D CBAM[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway，NJ：IEEE，2023：470-475.

[15] Zhang S，Liu H，Sun C，et al. MSTA-SlowFast：A student behavior detector for classroom environments[J]. Sensors，2023，23(11)：5205.

[16] Woo S，Park J，Lee J Y，et al. CBAM：Convolutional block attention module[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway，NJ：IEEE，2018：3-19.

[17] Diao X，Xu Y. A SlowFast-based violence recognition method[C]//Proceedings of the 6th Asian Conference on Artificial Intelligence Technology. Piscataway，NJ：IEEE，2022：1-6.

[18] Wang X L，Girshick R，Gupta A，et al. Non-local neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition . Piscataway，NJ：IEEE，2018：7794-7803.

[19] Soliman M，Kamal M H，El-Massih Nashed M，et al. Violence recognition from videos using deep learning techniques[C]//Proceedings of the 9th International Conference on Intelligent Computing and Information Systems. Piscataway，NJ：IEEE，2019：80-85.

[20] 张立国，马子荐，金梅，等. 基于非局部感知网络的运动目标跟踪方法[J]. 激光与光电子学进展，2023，60(4)：391-394.