International Journal of Machine Learning and Cybernetics
https://doi.org/10.1007/s13042-025-02540-0

ORIGINAL ARTICLE

A multi-stream framework using spatial–temporal collaboration learning networks for violence and non-violence classification in complex video environments

Barun Pandey[1] · Upasana Sinha[1] · Kapil Kumar Nagwanshi[1]

Received: 1 July 2024 / Accepted: 2 January 2025

Abstract
Violence detection (VD) is a critical task in computer vision with applications in security, surveillance, and public safety. The proliferation of surveillance cameras and the increasing need for automated monitoring systems have underscored the importance of accurate and efficient violence detection algorithms. However, existing methods face several challenges, including limited performance in complex real-world scenarios, high false positive rates, and difficulties in capturing subtle violent behaviours. Addressing these challenges requires the development of advanced algorithms that can effectively differentiate between violent and non-violent activities while minimizing false positives. This paper proposes a Spatial–Temporal Context Collaboration Learning Multi-Stream Network (STCCLM-net) to address challenges in violence detection within video data, particularly in scenarios with crowded scenes, rapid motions, and occlusions. This framework leverages a dual-stream architecture comprising a spatial context extractor network (SCENet) and a temporal context extractor network (TCENet), incorporating a Spatial–Temporal Collaboration Unit (STCU) to optimize spatial and temporal features. The pre-processing stage involves frame difference and background suppression for motion capture and clutter removal. Feature extraction utilizes the VGG16 network for spatial and temporal context extraction, enhancing feature recognition. In this, fully connected VGG16 network (FC-VGG16 Net) efficiently handles sequential data with a one-dimensional structure, while convolutional VGG16 network (Conv-VGG16 Net) effectively models spatial–temporal sequences by leveraging convolution operations, enhancing the comprehensive modeling of spatial–temporal dynamics within video clip data. The STCU module explores the complementary nature of spatial and temporal features through an alternating co-attention mechanism, optimizing feature fusion. The classification stage labels video clips as violent or non-violent based on the extracted features. Together, these modules optimize the model's ability to discern complex patterns and salient features across spatial and temporal dimensions in video data. Experimental results showcase enhanced performance metrics, including increased precision and recall rates of overall 99.6%, validating its effectiveness in accurately classifying violent and non-violent actions on four video datasets.

Keywords Violence detection (VD) · Multi-stream · Video clips · VGG-Net · Spatial–temporal features · Feature fusion · Classification

1 Introduction

In the era of ever-expanding surveillance systems and the proliferation of digital content, the need for automated detection and classification of violent and non-violent behaviours in video data has become paramount [1]. Surveillance cameras, both in crowded urban settings and less populated environments, continuously capture an extensive array of human actions, making it imperative to develop robust and efficient methods for distinguishing

✉ Barun Pandey
barunpandey@gmail.com

[1] School of Studies in Engineering and Technology, Guru Ghasidas Vishwavidyalaya, Bilaspur, India

Published online: 16 January 2025

between violent and non-violent events [2, 3]. Violence detection (VD) in video streams poses a formidable challenge due to the inherent diversity of human actions. Actions, whether violent or benign, can exhibit substantial variation in their execution, ranging from interpersonal altercations in crowded spaces to isolated incidents in uncrowded environments. Compounding this complexity is the presence of overlapping actions, where benign interactions share visual characteristics with violent behaviors, such as hugging, handshaking, or enthusiastic discussions [4]. Moreover, the effectiveness of VD hinges on the ability to navigate the intricacies of spatial and temporal cues. Alterations in camera viewpoints, lighting conditions, and the presence of occlusions can drastically alter the visual representation of actions [5]. Therefore, a robust detection system must transcend these challenges, extracting discriminative features that encapsulate the essence of violent and non-violent behaviors, irrespective of external factors [6].

Existing methods for violent and non-violent classification have made substantial strides in leveraging advances in computer vision and deep learning. These methods typically rely on complex neural network architectures capable of extracting intricate spatial and temporal features from video streams. While these techniques have shown promise, they are not without their limitations. One of the primary challenges faced by existing methods is the inherent diversity and complexity of human actions [7–9]. Violent actions encompass a wide spectrum, ranging from physical altercations to the use of weapons, each with its unique visual signature. Furthermore, benign interactions, such as hugging or animated discussions, can share visual characteristics with violent behaviors, leading to classification ambiguities [10]. Another significant drawback is the sensitivity of existing methods to changes in camera viewpoints and environmental conditions. Actions captured from different angles or under varying lighting conditions can exhibit distinct visual appearances, which can confound existing classifiers. The ability to generalize across diverse scenarios remains a critical research challenge. Additionally, occlusion poses a formidable hurdle in violence classification, especially in crowded scenes. Individuals may partially or completely obscure each other, introducing occlusion-related visual artifacts [11]. Existing methods often struggle to distinguish between true violence and occlusion-induced anomalies, compromising detection accuracy. Moreover, many existing approaches require computationally intensive pre-processing steps before network training. While these steps are essential for feature extraction and model learning, they increase the overall processing time, making real-time VD a challenging endeavour [12].

Given these constraints, there arises a critical necessity for innovative strategies that mitigate the deficiencies of current methodologies. This study tackles the imminent requirement for the detection of violence and non-violence occurrences in various video contexts, encompassing both crowded and uncrowded scenarios. Utilizing the capabilities of deep learning, we propose a novel multi-scale spatio-temporal network designed to navigate the intricacies presented by diverse environments and activities. Our methodology comprises spatial feature extraction, designed to capture the distinctive shape and appearance of actions. Concurrently, we employ temporal feature extraction, emphasizing rapid and sudden movements characteristic of violent behaviors. The synergy between these spatial and temporal cues forms the backbone of our approach, providing a comprehensive understanding of actions in crowded and uncrowded settings. Furthermore, we tackle the challenges posed by occlusions, which are particularly pronounced in crowded scenes. Our proposed network not only excels in discerning violent and non-violent actions but also exhibits robustness against occlusion-related visual artifacts.

Motivated by these facts, we proposed a multi-stream framework namely, Spatial–Temporal Context Collaboration Learning Multi-Stream Network (STCCLM-net) to address the challenges posed by crowded scenes, rapid motions, and occlusions in video data for violence and non-violence detection. The framework aims to enhance the accuracy and efficiency of violence detection algorithms by integrating spatial and temporal features, optimizing feature recognition, and classification accuracy. By leveraging a dual-stream architecture with spatial and temporal context extractors, the framework focuses on extracting meaningful features indicative of violent actions while reducing noise from static or insignificant background elements. Ultimately, the objective is to provide a robust methodology for effectively identifying and classifying violent and non-violent actions in video data, thereby contributing to improved surveillance and security systems.

1.1 Main contributions

This paper aims is to enhance violence detection from video clip data by integrating spatial and temporal features, optimizing feature recognition, and improving classification accuracy in challenging scenarios such as crowded scenes, rapid motions, and occlusions. This aims to establish a robust detection framework named Spatial–Temporal Context Collaboration Learning Multi-Stream Network (STCCLM-net) that integrates spatial and temporal features through a dual-stream architecture to effectively classify violence and non-violence data across different

Bu proje Kocaeli Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon Birimi tarafından desteklenmektedir. Çalışmalar Koç Üniversitesi laboratuvarlarında gerçekleştirilmiştir. School of Studies in Engineering and Technology ve Indian Institute of Information Technology Allahabad çalışma ortaklarındandır.

video datasets and contributing to the development of more effective surveillance systems.

- A dual stream architecture named Spatial Context Extractor Network (SCENet) and Temporal Context Extractor Network (TCENet) is presented to extract spatial and temporal features from video frames, focusing on changes indicative of violence and enhancing discrimination between violent and non-violent actions. The incorporation of both FC-VGG16 Net and Conv-VGG16 Net enables the dual stream model to proficiently handle sequential features with varying dimensions and semantic levels, facilitating detailed analysis of spatial–temporal dynamics and enhancing the discrimination and understanding of temporal features within video clip data.
- A Spatial–Temporal Collaboration Unit (STCU) is introduced to explore the complementary nature of spatial and temporal features through a co-attention mechanism, optimizing feature recognition by alternating between spatial and temporal information. By exploring the complementary nature of spatial and temporal features, the STCU capture subtle patterns and significant features across both dimensions. This collaboration leads to improved feature recognition, which is crucial for accurately identifying violent actions in dynamic video contexts.
- A classification module is presented with the support of hidden feature map layer (HFmapL) that serves as a vital classifier within the STCCLM-net model architecture, designed to adaptively manage the dimensionality of feature representations. It integrates seamlessly with the multi-stream architecture of the model, including the SCENet and TCENet. This integration allows for a comprehensive analysis of both spatial and temporal features, improving the model's ability to discern violent actions from non-violent ones.
- Performance results demonstrate that the proposed model achieves high accuracy and AUC scores than existing methods, indicating its strong capability to correctly classify violent and non-violent actions across four datasets i.e., Hockey Fight dataset, Movies Fight dataset, BEHAVE dataset, and ViF dataset.

The remainder of the paper is structured as follows: Sect. 2 deals with the literature review focusing on the recent methods for the proposed framework, while Sect. 3 gives a detailed explanation of the proposed framework including a structural representation of the proposed model. Section 4 deals with the experimental set-up and the methodology that is required for the evaluation of the proposed framework using various performance metrics and lastly in Sect. 5, is the conclusion part where overall conclusion and scope of next further developments are mentioned.

2 Related work

Several recent works are presented in the domain of complex video scenes for violence/non-violence detection, and classification based on deep learning are analysed to highlight the limitations and emphasize the need for advancements in capturing complex spatial–temporal dynamics in video data about each method. For this analysis, the related work section has been categorized into two sub-sections: handcrafted feature-based methods and Spatial–temporal feature-based methods. These categorized sections are discussed in the following sub-sections.

2.1 Handcrafted feature-based methods

Handcrafted feature-based methods have played a significant role in VD in videos, primarily focusing on extracting specific features that can distinguish violent actions from non-violent ones. These methods typically rely on predefined algorithms to identify motion patterns, texture, and other visual cues within video frames. One prominent example of a handcrafted feature-based method is the use of the histogram of oriented gradients (HOG), where specific gradients or orientations in images are computed to capture the shape and appearance of objects. This method, when applied to video frames, allows for the extraction of features related to human motion and poses, which can then be analysed to identify violent behaviour. Another widely used approach is the extraction of motion boundary histograms (MBH), which captures motion information by calculating the gradient of optical flow fields. This technique enhances the robustness of motion-based features, particularly in complex scenarios where background noise and occlusions may pose challenges. Despite their utility, handcrafted feature-based methods have limitations, particularly in their sensitivity to noise, occlusions, and variations in lighting conditions. These methods often struggle in real-world surveillance scenarios, where video quality can be poor, and scenes are cluttered with multiple moving objects. Febin et al. [14] introduced a technique that combines MoBSIFT with a movement filtering algorithm, enhancing the MoSIFT descriptor with MBH. They emphasized the significance of motion information and proposed a movement filtering algorithm based on temporal derivatives to reduce computational complexity. However, their method suffered from computational expense and sensitivity to noise and occlusion, limiting its effectiveness in noisy surveillance environments. A semi-supervised hard attention network for the localization and

identification of violence was introduced by Mohammadi and Nazerfard [16]. This mechanism played a pivotal role in identifying essential regions within video data while effectively segregating them from non-informative segments. By implementing hard attention, the model aimed to enhance its accuracy by eliminating redundant data and emphasizing valuable visual information in higher resolutions. This approach was particularly notable for its ability to reduce the reliance on attention annotations in video violence datasets. A pre-trained I3D backbone architecture served as a foundation for the model, facilitating faster convergence and improved stability during training. The video violence recognition and localization, it also faced certain limitations that warrant consideration. These limitations encompassed aspects such as complexity, interpretability, and generalizability. Samuel et al. [20] proposed an approach that integrated big data analysis and BiLSTM to enhance the accuracy and efficiency of VD during live football events. In this framework, video frames underwent systematic segmentation, followed by individual feature extraction using the histogram of oriented gradients (HOG) function. The frames were then classified into three distinct categories: negative, violence, and human parts. These frames were subsequently employed to train the bidirectional long short-term memory (LSTM) network, enabling it to access information in both forward and reverse directions. However, it faced challenges related to dataset bias, cost of implementation, and technology dependence. Cobo and San Miguel [22] presented an efficient method for VD in surveillance videos, focusing on human skeletons and change detection techniques. They employed human pose extraction and change detection as the core components of their methodology. However, accurate human pose estimation proved challenging, especially in complex scenarios with multiple individuals. Additionally, change detection was susceptible to false positives, particularly in noisy video feeds. Limitations, including the challenges of accurate human pose estimation, susceptibility to false positives in change detection, and the system's limited scope of detectable violent actions.

2.2 Spatial–temporal feature-based methods

Currently, spatial–temporal feature-based methods have been proved to be a rather stable technique for VD in video content by utilizing spatial features describing the object of the frames and temporal features associated with motion and temporal shifts to describe violent actions crisply. As for the strengths of spatial–temporal feature-based methods, one may highlight the fact that the nature of violent actions is often dynamic and complex, implying interactions between the entities involved throughout the process. This dual focus on both space and time allows these models to discern subtle cues that may be indicative of violence, such as sudden movements, aggressive gestures, or unusual activity patterns. Segador et al. [13] presented CrimeNet, which integrates Vision Transformer (ViT) and Neural Structured Learning (NSL) methods while incorporating adversarial training to detect violence in video content. This approach utilized conventional convolutional neural networks (CNNs) for violence detection, which typically extracted temporal and spatial features from frames. Despite showing promise, these CNN-based methods often struggled with the intricate dynamics and nuances inherent in violent activities, resulting in suboptimal performance and a notable number of false positives. Wang et al. [15] introduced a pioneering approach that marked a significant milestone in this domain. This research addressed the growing security concerns in urban areas by leveraging video surveillance technology. Specifically, they concentrated on developing techniques for brute force detection and face recognition. This approach involved the use of CNNs to obtain temporal and spatial features from frames. These artificial features were designed to capture crucial spatiotemporal patterns within surveillance footage. Additionally, the role of trajectory features in behavior analysis and how the spatial pyramid pooling (SPP) layer and CNN model of multi-foot input handle face identification. Asad et al. [17] introduced VD model based on fusion of features from multiple frames. This approach used to combine both temporal and spatial features extracted from sequential frames. Utilizing a CNN, they extracted multi-level features from pairs of consecutive frames, incorporating both top and bottom layers to effectively capture motion information. Furthermore, they introduced the wide-dense residual module to bolster the learning of amalgamated spatial features from input frames. These methods were devised to mitigate the constraints related to generalizability and computational efficiency.

In a related study, Halder and Chatterjee [18] presented the CNN-based bidirectional long short-term memory (CNN-BiLSTM) model for violence detection in surveillance scenarios. It provided real-time VD and video data capture of motion and spatial characteristics. However, the model was computationally complex, posing challenges for real-time deployment, and its black box nature made it difficult to understand decision-making processes. Mohtavipour et al. [19] introduced a multi-stream CNN framework for deep VD. This framework utilized manually crafted features pertaining to visual attributes, motion speed, and representative imagery to enhance the precision of VD in video data. The process of crafting these features involved the selection and engineering of specific attributes to represent different aspects of violent actions. But this approach had drawbacks, as manually designing such

features was time-consuming and required domain expertise and these handcrafted features might not capture all relevant information accurately, leading to suboptimal performance. Wen et al. [21] presented a method for violence detection utilizing an automated mobile neural framework search (MNFS) network and convolutional LSTM techniques to extract spatiotemporal features from videos. Their approach integrated average and max pooling layers to capture detailed characteristics more effectively, reduced feature dimensionality to enhance classifier performance, and utilized a pre-trained MNFS network for feature extraction. However, the complexity, computational demands, and potential vulnerability to environmental changes or attacks were notable limitation. From the reviewed methods, several challenges of suboptimal performance and false positives in traditional CNN-based models, computational complexity and sensitivity to noise, and occlusion in certain techniques, efficiency, and accuracy issues, especially in detecting abnormal behavior, complexity and interpretability problems in models like CNN-BiLSTM, manual feature engineering challenges, dataset bias affecting generalizability, high implementation costs, and technology dependence. Environmental vulnerability in mobile-based approaches, accuracy issues in human pose estimation and change detection, and a limited scope of detectable violent actions in some methods also pose significant challenges in the field of VD in surveillance videos.

3 Proposed multi-stream violence and non-violence classification framework

In the proposed methodology, we introduced a STCCLM-net model to address the challenges posed by crowded scenes, rapid motions, and occlusions in video data for violence and non-VD. Initially, a frame difference thresholding key frame extraction (FD-KFE) method is employed to identify key frames with significant motion differences, utilizing metrics such as motion distance, edge magnitude, and Oriented FAST and Rotated BRIEF (ORB) points for efficient extraction of visual features indicative of violence. This applied frame differentiation to capture motion information by calculating the difference between consecutive frames, which is crucial for identifying rapid movements associated with violent actions. Additionally, background suppression is implemented to remove static elements from the scene, allowing the model to focus on dynamic foreground actions and reducing noise, thus enhancing robustness in complex environments. STCCLM-net leverages a dual-stream architecture consisting of a SCENet and a TCENet that incorporates a spatial–temporal collaboration unit. This module employs an alternating co-attention mechanism between the two streams to optimize spatial and temporal features. The pre-processing stage involves frame difference and background suppression for motion capture and clutter removal, respectively. The feature extraction stage utilizes VGG16 network for spatial and temporal context extraction. The STCU enhances feature recognition, and a fusion stage combines spatial and temporal information. A feature fusion module is employed to combine features originating from both the frames stream and the frame difference stream. Then different activation functions are employed to the outputs from the spatial and temporal streams, ensuring that both types of features are appropriately activated for further processing. The activated features undergo element-wise multiplication, effectively merging spatial and temporal information into a unified representation. This fusion process results in output feature maps that encapsulate the dynamics of the video to recognize significant events and actions. The HFmapL serves as a crucial classifier within the model architecture, processing extracted features to make informed predictions about violence detection in video data. It is designed to adaptively manage the dimensionality of feature representations, producing compressed, higher-dimensional, or equivalent-dimensional outputs based on the number of streams. This flexibility allows the model to retain essential information while optimizing performance. HFmapL employs activation functions such as hyperbolic tangent (hTan) and Sigmoid, introducing non-linearity that enables the capture of complex patterns within the data. It integrates seamlessly with other components, including the SCENet and TCENet, facilitating the fusion of spatial and temporal features for classifying labels as violent and non-violent. The proposed methodology framework is shown in the Fig. 1.

The proposed STCCLM-net model involves five modules: pre-processing module, feature extraction module, Spatial–Temporal Collaboration Unit module, feature fusion module and classification module.

3.1 Pre-processing module

Initially, the raw input videos are obtained from different sources. The next step is to convert videos into frames followed by preprocessing steps. To increase the transformation of visual data and reduce the model complexity, we applied several pre-processing steps for our dual stream network architecture. In one stream of our network, we utilize the difference between consecutive frames as input data. Calculating the frame difference allows to capture motion information explicitly. This is valuable for VD, as violent actions often involve rapid movements. This approach encourages the model to encode temporal variations between these frames, enhancing its ability to capture

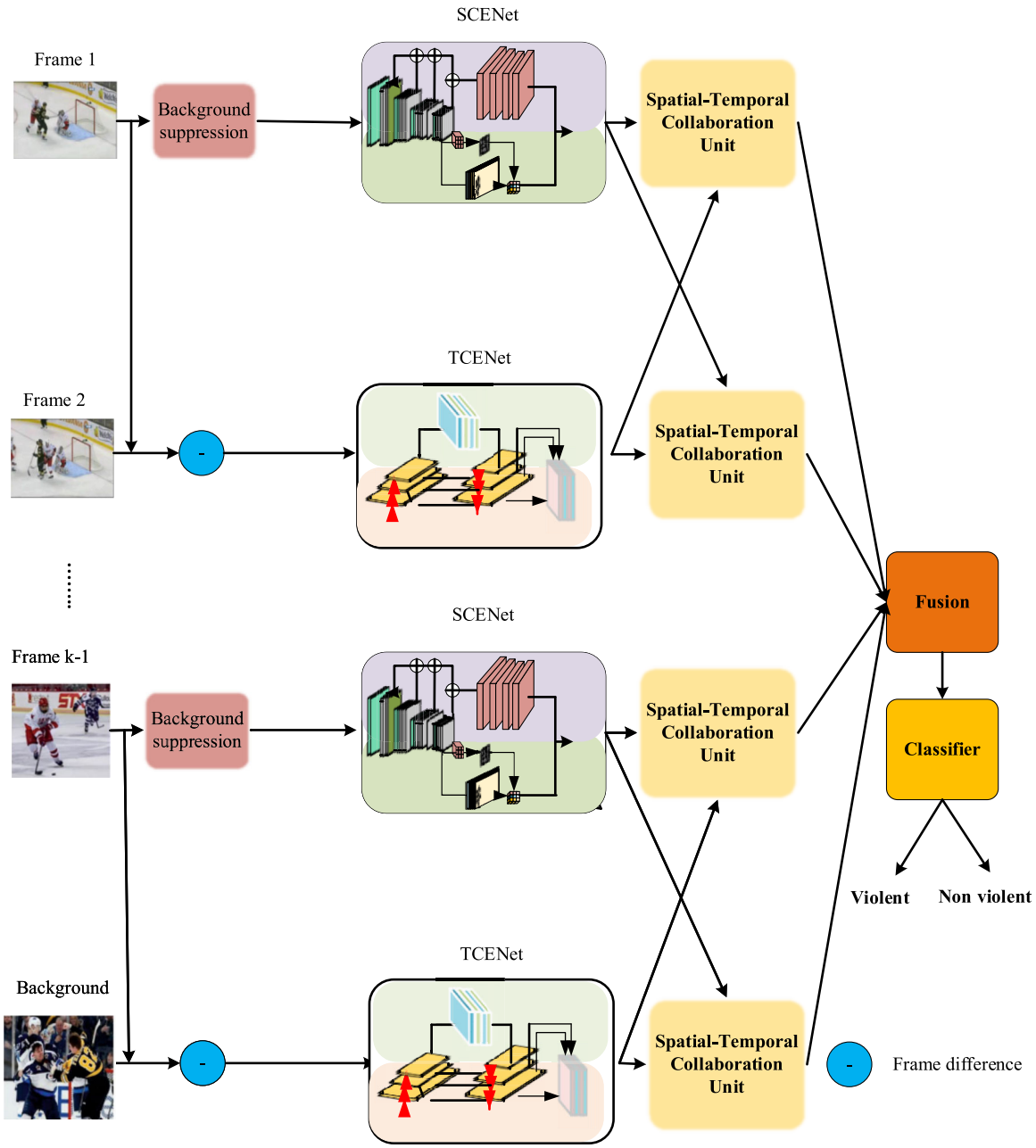


Fig. 1 Framework of the proposed methodology

motion-related information. It acts as an effective replacement for computationally demanding optical flow methods. In the other stream, rather than utilizing raw frames, we chose to employ background-suppressed frames. To achieve this, we implemented a straightforward method for background estimation, aiming to avoid introducing unnecessary computational complexity. Background suppression helps remove clutter and distractions in the scene, allowing the model to focus on foreground actions. This makes the model more robust to crowded and complex environments.

(i) Key frame extraction estimation

Key frame extraction is essential in pre-processing based on frame difference as it helps streamline the analysis of video data for violence detection. The key frames highlight relevant moments where violence or abrupt alterations in the scene occur, aiding in the extraction of motion important visual features indicative of violence. From these facts, we applied a frame difference thresholding key frame extraction (FD-KFE) method to identify frames with depth motion information differences compared to their neighbouring frames.

After frame conversion, we compute three metrics: motion distance (D), edge magnitude (E), and the number of ORB points (S) [23, 24]. To ensure consistency across different videos, we used min–max normalization method to normalize the metric scores, scaling them to a range of 0 to 1 inclusively. The motion distance D quantifies how much the metric has changed, indicating the level of motion between the frames. A larger value of D suggests significant movement, which is crucial for detecting dynamic events such as violence. The edge magnitude E reflects the strength of edges detected in the frames, which are indicative of significant changes in the scene. A higher value of E indicates more pronounced edges, suggesting that there may be important movements occurring in the video. The number of ORB points S serves as a measure of the richness of features available for analysis. A significant change in S indicates that new features have emerged or that existing features have been lost, which can be critical for identifying dynamic actions such as violence, which are expressed as:

$$\begin{aligned} D^1 &= \sum_{x=1}^{k} \left|D^x - D^{x\prime}\right|, E^1 = \sum_{x=1}^{k} \left|E^x - E^{x\prime}\right|, S^1 \\ &= \sum_{x=1}^{k} \left|S^x - S^{x\prime}\right| \end{aligned} \tag{1}$$

where D^1 denotes motion distance between consecutive frames, D^x signifies the value of the metric (e.g., pixel intensity, feature value) at frame k and $D^{x\prime}$ represented the value of the metric at the previous frame $k-1$. E^1 denotes edge magnitude between consecutive frames, E^x indicates Edge magnitude at frame k and $E^{x\prime}$ represents Edge magnitude at the previous frame $k-1$. S^1 is the number of ORB points detected between consecutive frames., S^x denotes the number of ORB points detected at frame k and $S^{x\prime}$ describes number of ORB points detected at the previous frame $k-1$.

Rather than simply summing the three scores, we assign dynamic weights to each score to capture changes effectively. The weight assigned to each score (W_s) is determined based on the variance of the score across consecutive frames. By computing the sum of the magnitude of differences between consecutive frame scores for each metric, the derived weights are then normalized to generate a fused score. This weighted approach allows for the nuanced incorporation of each metric's contribution to the overall assessment of key frame significance, expressed as:

$$\begin{aligned} D^W &= \frac{D^1}{D^1 + E^1 + S^1}, E^W = \frac{E^1}{D^1 + E^1 + S^1}, S^W \\ &= \frac{S^1}{D^1 + E^1 + S^1} \end{aligned} \tag{2}$$

where S^W is the weighted score based on the number of ORB points detected in the frame, E^W is the weighted score based on the edge magnitude detected in the frame and D^W signifies the weighted score based on the motion distance between consecutive frames.

The representation of fused score F_s at a frame is expressed as:

$$F_s = D^W D_1 + D^W E_1 + D^W S_1 \tag{3}$$

where these (D, E, S) metrics capture the dynamics and changes within the video frames. We then combine these metrics to derive a fused score (F_s), where each metric is dynamically weighted based on its contribution to frame significance. Frames with the highest fused scores, surpassing a predetermined threshold (set at 0.5), are selected for further analysis. This thresholding mechanism ensures that only frames with substantial changes or events are considered as key frames, enhancing the efficiency and accuracy of the VD process.

(ii) Background estimation

After key frame extraction, we applied background estimation in pre-processing step for violence detection as it helps discern changes in the scene and isolate foreground movements. This process helps in distinguishing between normal activities and potentially violent actions within a video sequence. Enhancing the accuracy and efficiency of violence detection algorithms involves focusing the analysis on regions where actions or movements occur by removing the background from the foreground. In this fact, we applied a straightforward method termed dynamic background modeling and suppression based Gaussian mixture model (DBM-GMM) to adaptively model complex background scenes in videos by dynamically updating Gaussian distributions, distinguishing between dynamic background and foreground changes. By continually refining the background model based on pixel intensity changes over time, DBM-GMM efficiently suppresses background elements, enabling accurate violence detection.

In DBM-GMM, each pixel (P) in the video frame (F_v) is analysed individually at time T. The pixels RGB value at coordinates (u, v) in frame F_v is denoted as $X^P(T)$. This pixel's values over time constitute a sample sequence, represented as:

$$\left[X^P(1), \ldots X^P(T)\right] = \{F_v(T), (u, v) : 1 \le T \le F\} \tag{4}$$

Let W_G represent the weight of the Gth Gaussian mode in the mixture. The priority levels are determined based on the weights, where higher weights indicate a stronger association with the background. described as:

$$W_G = \arg\max[\Pr ob_G] \tag{5}$$

At frame F, the GMM associated with pixel P consists of G weighted Gaussian functions in the RGB colour space, expressed as:

$$F(X) = \sum_{G=1}^{G} W_{G,T}^{P} \cdot F_G \left(X; \omega_{G,T}^{P} \sum\nolimits_{G,T}^{P} \right) \quad (6)$$

Hence, the priority levels help in distinguishing between background and foreground elements in the video. By associating modes with background based on their weights, the model can effectively suppress noise and focus on relevant actions, thereby improving the accuracy of violence detection.

Priority levels $W_G \sigma_G$ determine the association of modes as Background, with the first G_{Bm} modes labeled as such, determined by a threshold ($Thres_b$) in the range [0, 1], expressed as:

$$G_{Bm} = \arg Min \sum_{g=1}^{G} \omega_{G,T} > Thres_b \quad (7)$$

The range of threshold in Eq. (6) for background extraction is [0, 1] for a fixed threshold. This means that the threshold value is a constant between 0 and 1, and all pixels with a value greater than the threshold are considered foreground, while all pixels with a value less than or equal to the threshold are considered background.

The next step involves pixel classification which is assigned to the closest mode center under specific constraints of the pixel. If none of the modes meet these constraints, the priority mode is low with a replacement of new Gaussian centered on the current intensity $X^P(T)$ based on priori variance weight, given by:

$$\left\| X_T^P - \lambda_{G,T}^P \right\| \leq G_P \, \sigma_{G,T}^P \quad (8)$$

where F denotes the number of frames and G number of Gaussian modes in the mixture. The function $F_G \left(X; \omega_{G,T}^P \sum_{G,T}^P \right)$ represents the Gaussian density function associated with the Gth Gaussian mode of pixel p in frame T. In this function, $\sum_{G,T}^P$ denotes the center vector of the Gaussian mode G associated with the pixel p at frame F and $\omega_{G,T}^P$ corresponds to the weight of the mode G associated with the pixel p at frame F. In the given context, G_P serves as a constant coefficient that requires adaptation for each individual video.

Upon successful selection of a mode, the updated GMM parameters are used to reinforced based on the operation of Stauffer [25, 26], given by:

$$\sigma_{u,T+1}^2 = [1-\gamma]\, \sigma_{u,T}^2 + \gamma \left\| X_{T+1}^P - \lambda_{u,T+1} \right\|^2 \quad (9)$$

$$\omega_{v,T+1} = (1-\gamma)\, \omega_{v,T}, \ \forall v \neq u \quad (10)$$

Otherwise, a new Gaussian mode is distributed in the last replacement mode. This iterative process allows for continuous adaptation and refinement of the background model as new frames are processed.

3.2 Feature extraction module

The feature extraction stage consists of dual stream network architecture: Spatial Context Extractor Network (SCENet) and Temporal Context Extractor Network (TCENet). TCENet operates on a difference between consecutive frames to enhance the discrimination between violent and non-violent actions. SCENet operates on the background supressed frames. The model focusses on changes between consecutive frames that are indicative of violence. This frame difference helps to reduce noise from background elements, ability to identify relevant actions.

In both networks, VGG16 network is used first and then that output is passed to two parallel feature extraction branches for learning and performing inference. One of the branches is a feature fusion network with multi-layer, which supplies discriminative functionality to capture potential violent action candidates. In this way, the multi-scale spatial features can be extracted well for the model and localize the violent action in the frame well. It also enables the model to be scaled in terms of the size of violent actions, such that violence can easily be identified whether the actions are proximal to the camera or at a distance. The other branch makes use of deformable based RoI-pooling to produce invariant features, most useful in cases of occluded violent.

In the spatial feature extraction process, the architecture draws inspiration from the complementary nature of feature representation across different layers from CNN [27]. Each video frame undergoes processing in SCENet to extract two types of features: high-level fully-connected features and middle-level convolutional features. These features are respectively extracted from the fully-connected layer and pooling layer of SCENet. The high-level fully-connected features capture abstract and semantic information from the frame, while the middle-level convolutional features retain spatial relationships and local patterns. To facilitate seamless integration into subsequent analysis stages, the extracted features serve as inputs for the temporal-sequenced model. Leveraging the pre-trained VGG16 Net from ImageNet for spatial feature extraction ensures a robust foundation for extracting meaningful spatial features from video frames. Then the operation process of feature extraction via SCENet involves passing each frame through the network architecture, where it undergoes hierarchical processing. As the frame traverses through SCENet, it undergoes transformations and feature

extraction operations across various layers. The high-level fully-connected features capture global context and semantic information, while the middle-level convolutional features retain spatial details and local patterns.

Initially, we first establish the sets used for training expressed as:

$$Train_{set} = [G_{Bm} H_s]_{s \in [1,F]}, F = \left\{ H_s^{(1)}, H_s^{(2)}, \ldots, H_s^{(b)} \right\} \tag{11}$$

where G in Gbm typically denotes the index of the Gaussian mode, while bm stands for background model. b signifies the batch size, $[H_s]$ represents the frame at the s step for the sample in the batch. Subsequently, the $Train_{set}$ is introduced into the VGG network, known as VGGNet, with the input handling process executed by the pre-trained VGG16 Net, referred to as VGG. The outputs derived from the pooling layer (PL) and fully connected (F_C) layer are then defined as follows:

$$Conv_f = \left[F^s_{channel\ z} \right]_{z \in [1, Z \times Z],\, S \in [1,F]}, \tag{12}$$

$$F_C = [f_c(s)]_{s \in [1.F]} \tag{13}$$

where $Conv_f$ denotes the convolution layer features with a feature cube characterized by dimensions $F \times Z \times Z \times F_{chaneel}$. Here, $Z \times Z$ denotes the spatial dimensions of the cube, while $F_{channel}$ represents the number of feature channels. This depiction enables the extraction of detailed spatial action information, allowing the model to capture complex spatial patterns and relationships within the video frames. On the other hand, the D-dimensional FC-features of the sth frame in the video sequence. These features encapsulate high-level global characteristics, complementing the local details captured by $Conv_f$. By combining both local and global features, the model gains a comprehensive understanding of the spatial dynamics within the video content, allowing for effective analysis and interpretation of spatial information. This dual representation approach enhances the ability to discern meaningful spatial patterns.

The paper uses a TCENet which combines VGG16-Net and deform-based RoI pooling to model the temporal dynamics of sequentially ordered features extracted from VGGNet. Taking advantage of memory and forgetting mechanism of VGG16 Net, the model is capable of adjusting relevance of current and past observations and adapts the model to continuously decide how much information of current and previous observations it should retain or let go at any given time step. This adaptability enables TCENet to effectively acquire discriminative information and learn dependencies within sequential features. By incorporating this mechanism, TCENet facilitates the modeling of temporal dynamics, enabling the extraction of meaningful patterns and dependencies critical for accurate VD analysis. This integration enhances the model's capacity to discern subtle changes over time, leading to more precise discrimination and understanding of temporal features within the video clip data.

The features extracted from various layers of the network exhibit diverse dimensions, presenting a challenge for sequential modeling. The TCENet, mentioned previously is fully connected, which is ideal for sequential data with a 1-d structure like fully connected activations from a deep CNN [28]. At the same time, when working with convolutional features that are themselves characterized by the presence of spatial layouts, the FC-VGG16 Net has numerous limitations and can lose valuable spatial context information. To tackle this challenge and fully leverage spatial information across time, the Conv-VGG16 Net is introduced to simultaneously model spatial–temporal sequential data. Unlike the standard VGGNet, which relies on multiplication operations, the Conv-VGG16 Net employs convolution operations, leveraging its inherent convolutional structure to extend the capabilities of the FC-VGG16 Net and effectively adapt to spatial–temporal sequences. By incorporating both FC-VGG16 Net and Conv-VGG16 Net, the model proficiently handles sequential features with varying dimensions and semantic levels, enabling comprehensive and detailed modeling of spatial–temporal dynamics within the video clip data.

3.3 Spatial–Temporal Collaboration Unit module

Inspired by the mechanism of human visual attention, we introduce a Spatial–Temporal Collaboration Unit (STCU) module to leverage the captured temporal and spatial features for enhanced recognition. Its primary purpose is to explore the complementary nature of spatial and temporal features. Specifically, the co-attention layer in this unit alternates between optimizing spatial and temporal information and creates temporal and spatial attention features. In this process, the STCU module applied attention mechanisms to optimize both alternates between spatial and temporal features.

Recognizing the limited temporal attention capabilities of the VGG16 Net, a strategically devised co-attention layer is introduced to dynamically assign distinct weights to sequential features at each processing step. Illustrated in Fig. 2, the proposed STCU aims to enhance the model's spatial–temporal attention mechanisms. Initially, the sequential features undergo a transformation by passing through a deep multi-layer perceptron (MLP) followed by a hyperbolic tangent activation function (Tan h), resulting in the generation of unnormalized attention weights (A_w) denoted as:

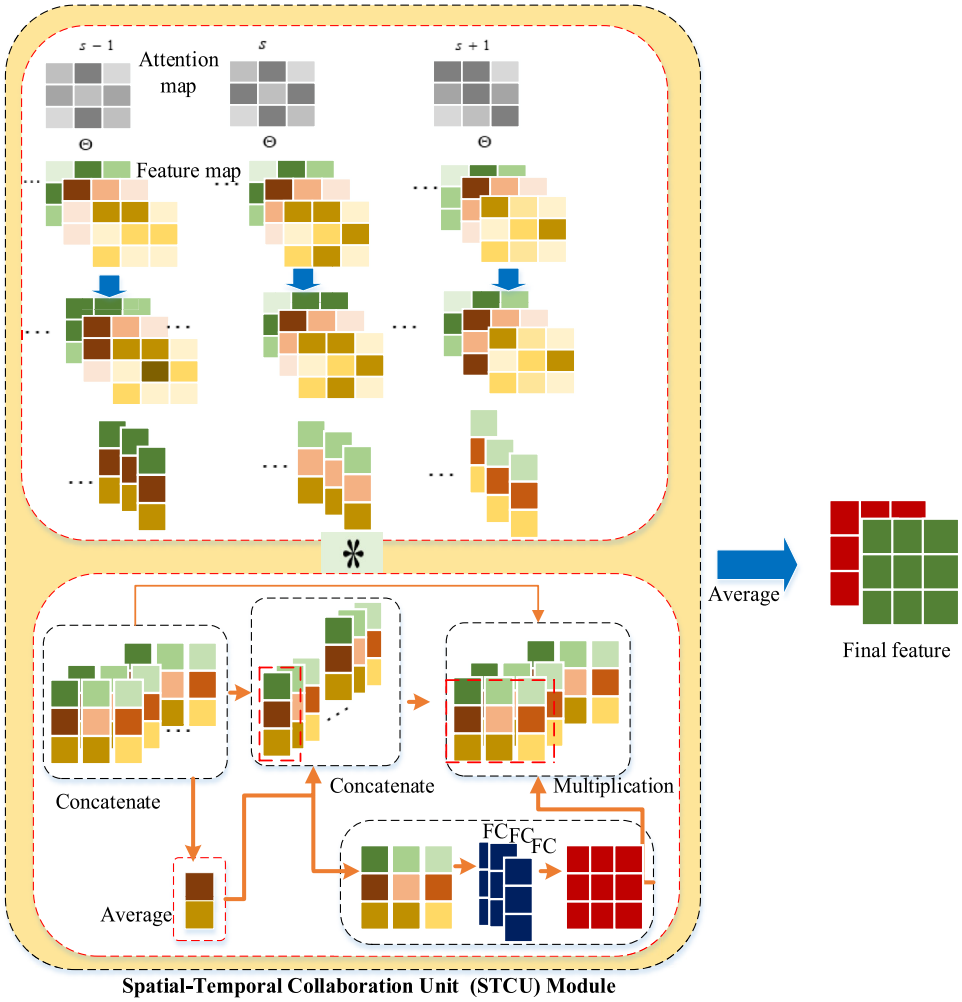


Fig. 2 Schematic structure of STCU

$$A_w = \mathrm{Tan}\,h\,(m_d v_s + B_A) \tag{14}$$

Here, m_d stands for mean distance, v_s represents the output of the VGG16-Net at the sth step, while A_w and B_A correspond to the parameters of the deep MLP and bias, respectively.

Following the initial processing steps, the importance weights for each frame are computed using a normalization technique called softmax:

$$a_w = \rho\left(T_{att}^{s}/\, v_s\right) = \frac{Exp[A_s]}{\sum_{s=1}^{F} Exp[A_s]} \tag{15}$$

In this equation, a_w denotes the normalized attention weights, indicating the frames to prioritize based on their significance. T_{att}^{s} represents the attention value generated for each frame, while F signifies the total number of frames in the video sequence. The softmax operation transforms the raw attention scores into a probability distribution, ensuring that the weights assigned to each frame sum up to 1. This normalization enables the model to effectively allocate attention across the frames, highlighting those that contribute most substantially to the analysis and recognition of temporal patterns within the video data.

Following the determination of the temporal attention distribution, the final video frame representation is computed by taking the expectation of the feature vectors across all time steps, expressed as:

$$F_S = \sum_{s=1}^{F} a_w\, v_s \tag{16}$$

where F_S denotes the global features generated by the temporal attention distribution. Taking advantage of the unique temporal and spatial traits present in the feature sets obtained from both the FC-VGG16 Net and Conv-VGG16 Net, this study introduces STCU module, as depicted in Fig. 2. Here, the effectiveness is delved in producing high-quality spatial–temporal features and its capability to enhance the performance of spatial–temporal saliency. By synergistically merging spatial and temporal features, the STCU module enables the extraction of comprehensive and informative representations, thereby enhancing the model's capacity to detect and capture subtle patterns and

significant features across both spatial and temporal dimensions within the video dataset.

In response to the output of the Conv-VGG16 Net, a STCU module is specifically engineered to discern the locations within the convolutional cubes that hold significant relevance to the current prediction. Rather than employing a deep MLP to generate spatial attention weights, the STCU module leverages convolutional operations to compute spatial attention maps.

By applying softmax, the sth important scores at each location $[1, Y \times Y]$ are subsequently derived as:

$$\widetilde{A}_s = m_h \mathrm{Re}Lu\left[\widetilde{kc}_o * \widetilde{v}_s + B_s\right] \tag{17}$$

$$\widetilde{a}_s(Y) = \rho\left(T^s_{att}(Y) / \widetilde{v}_s(Y)\right) = \frac{Exp\left\{\widetilde{A}_s(Y)\right\}}{\sum_{i=1}^{Y\times Y} Exp\left\{\widetilde{A}_s(i)\right\}} \tag{18}$$

where $[Y \times Y]$ represents the spatial size. In this process, the ReLU activation function $\mathrm{Re}Lu(\cdot)$ is applied, B_s accompanied by a bias, while $\underset{s}{\tilde{v}}$ denotes a 3-dimensional tensor extracted from the Conv-VGG16 Net. $\tilde{m}^o_d$ represents the 2-dimensional convolution kernels, while m^c_d assists in channel integration.

Notably, the spatial attention map is decomposed into (Y), delineating the significance of the Yth position within the spatial attention map at the sth step. $[Y \times Y]$ represents the spatial size of the attention map, facilitating detailed attentional focus across spatial dimensions. Through this mechanism, the STCU module effectively directs attention to spatial features crucial for accurate predictions.

In order to intensify the focus on critical informative areas within the feature map, a novel feature vector is derived as follows:

$$\tilde{Y}_s = \sum_{i=1}^{Y\times Y}\left[Y_s(i) = \tilde{a}_s(i)\ \Theta\ \tilde{v}_s(i)\right] \tag{19}$$

Here, $\tilde{Y}_s$ is derived to each channel of the feature map through attention map. The symbol Θ represents the element-wise product operation, and $\tilde{Y}_s$ serves as the spatial feature vector at the s step.

After a series of the aforementioned operations, including the utilization of a deep MLP m^s_d, activation $\sigma(\cdot)$, normalization, and discriminative spatial–temporal feature $\beta_{spatial-temporal}$ is generated by consolidating the temporal contribution of each spatial feature vector $\tilde{Y}_s$ across various time steps. Through this iterative process, the model iteratively refines its understanding of temporal dynamics while aggregating spatial information, ultimately enhancing its ability to discern and interpret intricate patterns, represented as:

$$A_S(s) = \rho\left(T^s_{att}/\underset{s}{\tilde{Y}}\right) = \frac{Exp\left\{\sigma\left(m_d\ \underset{s}{\tilde{Y}}\right)\right\}}{\sum_{s=1}^{F}\exp\left[\sigma\left(m_d\ \underset{s}{\tilde{Y}}\right)\right]} \tag{20}$$

$$\beta_{spatial-temporal} = \sum_{s=1}^{F} A_S(s)\ \underset{s}{\tilde{Y}} \tag{21}$$

Therefore, $\beta_{spatial-temporal}$ represents the video-level feature based on spatial–temporal perspective of the video content. Moreover, by dynamically learning to identify crucial spatial–temporal cues within the context and autonomously filtering out redundant information within an observed frame.

3.4 Feature fusion module

The feature fusion module combines features originating from both the frames stream and the frame difference stream. Firstly, the multi-scale spatial features extracted from the frames stream are processed using a Leaky ReLU activation layer. Similarly, the feature maps obtained from the frame difference stream, representing temporal features, are passed through a Sigmoid activation layer. Following activation, the features from both streams undergo element-wise multiplication, considering both spatial and temporal aspects from STCU. This operation enables the fusion of multi-scale spatial features with temporal cues, culminating in the generation of final output feature maps that encapsulate both spatial and temporal information.

In order to comprehensively capture and integrate temporal, spatial, and semantic information present in videos, this study introduces a weighted fusion methodology to combine outputs derived from the STCU. At the feature level, the fusion process is delineated as follows:

$$f_w = \eta\, f_{spatial-temporal} + (1-\eta)\, f_S \tag{22}$$

Here, f_S and $f_{spatial-temporal}$ represent features condensed using principal component analysis (PCA) [29], enabling effective information retention while reducing the dimensionality of F_S and $\beta_{spatial-temporal}$. Afterwards, the features from both streams are merged into a dimension corresponding to the number of action types using an FC layer. The weight parameter η represents the significance of the output from the STCU network, while the resulting fused vector, denoted by f_w, is generated through this weighted fusion process. To prevent overfitting, a dropout operation is introduced between the FC layer and the weighted fusion module. This fusion technique facilitates the creation of holistic video representations encompassing varied temporal, spatial, and semantic attributes, thereby augmenting the model's capability to discern and comprehend intricate patterns.

To enable the comprehensive training of the STCU network in an end-to-end fashion, the fused vector f_w undergoes a softmax operation:

$$\tilde{P_{v,t}} = H_{soft\max}(f_w) \tag{23}$$

In this context, $\tilde{P_{v,t}}$ represents the prediction vector, which indicates the probability distribution across action categories predicted for the tth video. This softmax operation ensures that the predicted probabilities are well-calibrated and collectively sum up to 1, providing a reliable representation of the model's confidence in different action categories. By incorporating this softmax activation, the STCU network is seamlessly trained to optimize its parameters across both spatial and temporal dimensions.

3.5 Classification module

In this module, the STCCLM-net model based VGG16-Net serves as the backbone for multi-stream classification, distinguishing between violent and non-violent occurrences. The HFmapL plays a crucial role as the classifier. The extraction of fused features is characterized by three distinctive scenarios. Firstly, when the number of streams in HFmapL is less than that in both the input and output layers, the extraction results in compressed dimension representations of the trained data. Secondly, if the HFmapL streams count surpasses that of the input and output layers, the extraction yields higher-dimensional depictions of the trained records. Finally, when the HFmapL streams count equals that of the input streams, the feature represents an equivalent-dimensional manifestation of the trained data. This process is expressed as follows:

$$\begin{cases} H_{mapL} = \sigma\left[W_m^1 I + f_w h_v^1\right] \\ \hat{O} = \sigma\left[W_m^2 H_{mapL} + f_w h_v^1\right] \end{cases} \tag{24}$$

where the activation function σ stands for the neuron activation function, with roles primarily utilizing Tan h and *Sigmoid*. Matrices $W_m^1 \in Z^{x\times y}$ and $h_v^1 \in Z^x$ denote the weighted matrix and offset vector of the HFmapL. Matrices $W_m^2 \in Z^{x\times y}$ and $h_v^2 \in Z^x$ indicate the weighted matrix and bias vector of the observed layer, respectively. The HFmapL of the networks is denoted by $H = \{h_1, h_2, \ldots h_x\} \in Z^x$. $\hat{O} = \{\hat{O}_1, \hat{O}_2, \ldots \hat{O}_y\}^T \in Z^x$ and $I = \{\hat{O}_1, \hat{O}_2, \ldots \hat{O}_y\} \in Z^y$ imply the input and output layers of the STCCLM-net correspondingly.

The hyperparameter tuning process of the STCCLM-net method employs the ADAM optimizer [30]. ADAM, a variant of stochastic gradient descent (SGD), is widely utilized to update network weights during training sessions. Renowned for its optimization prowess, ADAM operates based on the principles of adagrad while introducing additional flexibility. By combining adagrad with momentum, ADAM offers enhanced optimization capabilities. Parameters W_{itetn} and L_{itetn} representing the currently trained iteration (*itetn*), undergo parameter enhancement within ADAM, as expressed below:

$$\hat{Y}_W = \frac{y_w^{(itetn+1)}}{1-(\delta_1)^{(itetn+1)}} \tag{25}$$

$$X_W = \frac{x_w^{(itetn+1)}}{1-(\delta_2)^{(itetn+1)}} \tag{26}$$

$$W^{itetn+1} \leftarrow W^{itetn} - \eta \frac{\hat{y}}{\sqrt{\hat{x}_w + \in}} \tag{27}$$

where the inclusion of $\in$ as a safeguard against division by zero is fundamental in preventing computational errors, especially in scenarios where gradients may approach zero during training iterations (*itetn*). Moreover, the parameters δ_1 and δ_2 play pivotal roles in controlling the momentum and scaling of the gradients during optimization. By effectively managing the second moments of gradients and gradient forgetting features, ADAM optimizes the learning process, allowing for efficient and stable updates of network parameters.

To facilitate the training process of our network, we leverage the cross-entropy (CE) loss function for evaluating classification tasks, defined as:

$$\psi_{loss} = \frac{1}{N(L_h \times L_w)} \sum_{h=1}^{H} \sum_{w=1}^{W} q^{hw} \times \log r^{hw} + \left(1-q^{hw}\right) \times \log\left(1-r^{hw}\right) \tag{28}$$

Here, N represents the total number of samples in our dataset, h and w denote the height and width of the labels, respectively, q^{hw} represents the label at location (h, w). The value of q^{hw} lies within the range of [0, 1], where $q^{hw} = 0$ indicates the absence of violence, while 1 signifies the predicted value at location (h, w). By minimizing the cross-entropy loss during training, our network learns to better predict the likelihood of violence occurrences at various spatial locations within the input frames.

4 Experiments and results

This section presents the evaluation of the proposed model based on experimental setup, evaluation datasets, evaluation metrics and simulation results in terms of different metrics.

4.1 Datasets

In our work, we considered four datasets i.e., Hockey Fight dataset, Movies Fight dataset, BEHAVE dataset and violent flow (ViF) dataset.

(a) Dataset 1: The Hockey Fight dataset collected from NHL match videos, where the prime concern of the videos is fighting amongst players [32]. It includes a collection of 500 fighting video clips and 500 non-fighting clips considering the demography and od factors of ice fighting carefully. At an original frame rate of 25, together with average clip lengths of approximately two seconds, consisting of 50 frames, the videos provide a good insight of the aggressive incidents that occur in professional hockey. Each frames has a fixed 720 × 560 pixels resolution to eliminate inter-frame difference in view quality in the dataset. And by extracting the actual real-world clips from high-energy sporting events, Hockey Fights filled the gap that is necessary to research violent behavior in a context of competitive sport environment.

(b) Dataset 2: The Movies Fight dataset [33] presents a diverse array of video excerpts extracted from a wide range of action films, encompassing cinematic depictions of physical combat and non-combat sequences. The original frame rate is commonly around 24 to 30 frames per second (fps), with 24 fps being the most typical for film-based content. Comprising 100 carefully selected videos showcasing fight scenes and an equal number of videos featuring non-fight actions, this dataset offers a rich tapestry of visual content drawn from popular movie genres. The fight sequences, meticulously curated from action-packed movie sequences, showcase choreographed battles and intense confrontations, while the non-fight clips, sourced from datasets like UCF-101, provide contrasting scenarios of everyday activities. With varying lighting conditions, backgrounds, and narrative contexts, the Movies Fight dataset encapsulates the cinematic portrayal of violence and action, making it a valuable resource for exploring the nuances of fight choreography and visual storytelling in film.

(c) Dataset 3: The BEHAVE dataset [34] is a comprehensive collection of video recordings capturing a diverse range of group activities and interactions, including running, chasing, following, and physical altercations. Comprising four extensive videos filmed from different camera positions and locations, this dataset offers a detailed glimpse into human behaviour in various social contexts. The original videos in the dataset are typically recorded at 25 frames per second (fps). With a total frame count of approximately 200,000 frames and a resolution of 640 × 480 pixels per frame, the BEHAVE dataset provides a rich source of visual data for studying group dynamics and behavioural patterns. To facilitate experimental analysis, a subset of 280 clips was extracted from the original videos, featuring 80 instances of fighting actions and 200 non-fight scenes representing a spectrum of everyday activities. By capturing real-world interactions and activities, the BEHAVE dataset serves as a valuable resource for exploring social dynamics, conflict resolution, and group behaviour in diverse settings.

(d) Dataset 4: The ViF dataset [35] is a curated collection of 246 video clips sourced from various sources, including online platforms like YouTube. These videos capture a wide range of scenarios depicting instances of violent behaviour, with approximately half of the videos (123) showcasing physical altercations or aggressive interactions classified as "violent," while the remaining videos (123) portray non-violent activities. The videos in the dataset usually have a original frame rate of 30 frames per second (fps). The resolution of the videos in this dataset is standardized at 320 × 240 pixels, ensuring consistency in visual quality across the dataset. One of the distinctive characteristics of the Violent Flow dataset is the diversity of contexts in which violent actions are observed. The videos encompass a variety of settings, such as sports events, public gatherings, and street scenes, reflecting the complexity and variability of real-world scenarios where violence may occur. This diversity presents a challenge for violence detection algorithms, as they must be robust and adaptable to different environments and behaviours. Furthermore, the dataset includes videos of varying lengths, ranging from 50 to 150 frames, capturing the temporal evolution of events leading up to and during violent incidents. This temporal information is crucial for analyzing the dynamics of violent interactions and extracting meaningful features for violence detection algorithms. Some sample images are shown in the Fig. 3. The data description is summarized in the Table 1.

4.1.1 Experimental settings

The experiments were performed using Python 3.6 as well as Tensorflow (GPU) 1.14 toolbox. The computational environment was Intel(R) Core (TM) i7-9750H CPU (9th Gen.,) having a clock speed of 2.60 GHz and RAM of 16 GB. Moreover, a 6 GB NVIDIA GeForce RTX 2060 were used to provide computational support for the experiment. All experiments were performed under 64-bit Windows 10 Home running on the experiment machine.

The preprocessing of the sentinel 1 data uses pre-trained weights from venerable VGG-16 model to extract spatial and temporal feature. Only the added Conv and FC units are enabled and trained on the input data while keeping the pre-trained model the same. All of the modules within the architecture have three convolutional layers in them where

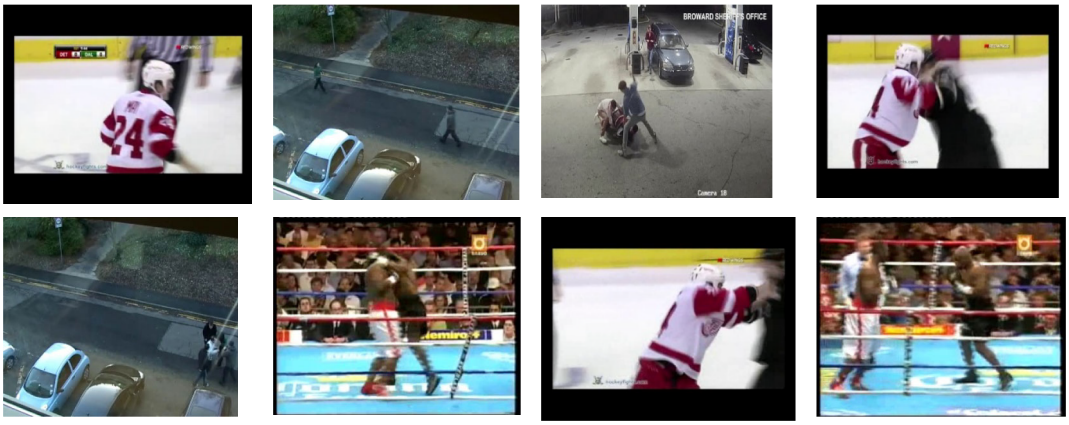


Fig. 3 Sample video frames from four datasets

Table 1 Data description

Dataset	Samples	Video size (pixels)	Violence video clips	Non-violence video clips
Hockey Fights	1000	360 × 280	500	500
Movies Fight	200	360 × 250	100	100
BEHAVE	280	640 × 480	80	200
ViF	246	320 × 240	123	123

the filter sizes and the number of filters is chosen appropriately. In the same regard, to overcome con councils inherent in deep learning networks such as the exploding gradient problem arising from sequences, a gradient clipping technique is applied. However, dropout regularization is applied between the layers of VGG16 net to avoid eventual overfitting the network has a dropout rate of 0.2. The model works in batches of videos, opting for a batch size of 5 videos, with each video having up to 30 frames. Training the model employs the ADAM optimization function, initializing with a learning rate of 10^{-3} and spanning 100 epochs for each training set to minimize the cross-entropy loss. Towards considering the multi-stream classification problem, groups of 10 consecutive frames of size 100 × 100 are fed into the model, as well as the spatial and temporal features are extracted. The data division was set up at 20:80 for training and testing, with class labels shown as "0 or 1" instead of one-hot, and a batch size of 5 samples per iterate. With SGD optimization, and fixing the learning rate to 0.001, the decay rate of 1e−6, the model works with the cross-entropy loss function.

4.2 Results

This section presents two evaluation results: evaluation of the proposed method on each dataset and comparative results between the proposed and the existing methods in terms of different metrics on four datasets.

4.2.1 Evaluation of the proposed method on each dataset

As shown in Fig. 4, the visualization results of the proposed model on preprocessing components of input frame, key frame estimation, and background estimation across the four datasets (Hockey Fight, Movies, BEHAVE, and ViF) reveal key insights of the model processes video data for violence detection. Starting with the input frame (a), the raw video data from each dataset is captured, reflecting the various environments, lighting conditions, and activities. The key frame estimation (b) isolates specific frames within each video that contain significant motion or abrupt scene changes, which are critical indicators of violent actions. For instance, in the Hockey Fight dataset, key frames might capture the exact moments of physical altercation, while in the Movies dataset, they could highlight intense combat scenes. The background estimation (c) further refines this process by suppressing static elements within the scene, effectively filtering out noise and distractions that do not contribute to the detection of violent behaviour. This step is particularly crucial in datasets like BEHAVE, where crowded scenes with overlapping activities can obscure the focus on violence. By emphasizing dynamic foreground activities, the model improves its sensitivity to violent actions, even in complex settings with rapid motions and occlusions. Overall, these preprocessing steps work in tandem to optimize the model's performance, ensuring that it accurately captures and analyses the most relevant features across different video

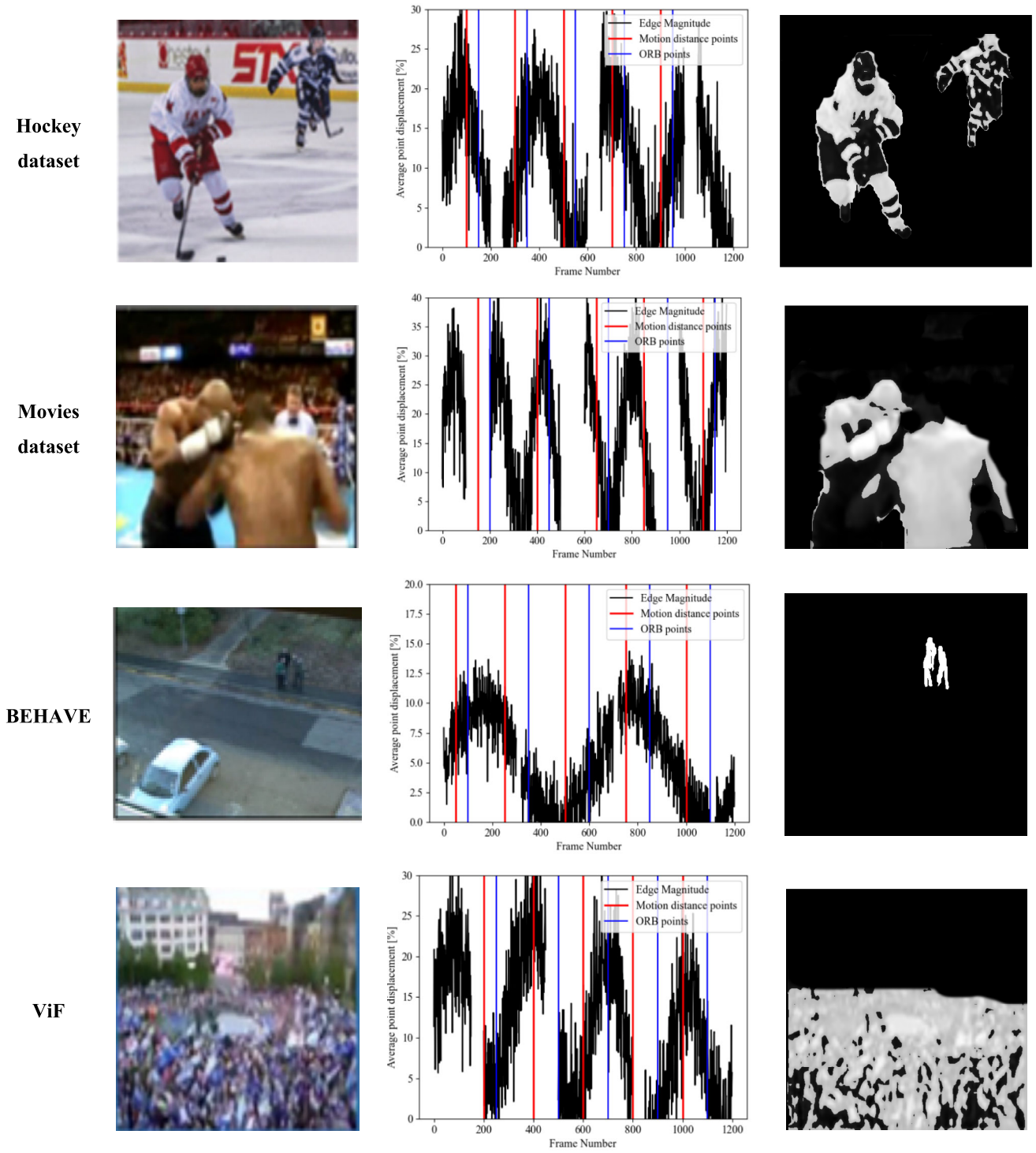


Fig. 4 Visualization results of the proposed model on preprocessing components **a** input frame, **b** key frame estimation, **c** background estimation for each dataset, respectively

contexts, leading to more reliable violence detection across all datasets.

As shown in the Table 2 the proposed violence detection model achieves higher results on higher threshold values across four distinct datasets: Hockey Fight, Movies, BEHAVE, and ViF because at higher threshold, one is more sensitive to violence. When moving to higher values of the threshold, the model is more selective in labeling

behavior as violent, which improves both precision and recall. This is clearly seen in the Table 2, in which precision, recall and F1-score increases steadily with the increase in the threshold value from 0.1 to 0.9. Higher thresholds allow the model to bring into the foreground more details and various shades of violent leading to increase effectiveness. However, the higher threshold values also make it able to minimize the false positives when other gentle motion is taken as a violent one, which in turn, increases precision. The positive difference in performance for all the four datasets established that the proposed model is Iiable and flexible enough to work with any form of violence. Moreover, the experiment shows that the use of a threshold equal to 0.9 maximizes the average of the precision/recall trade-off measurements and provides the highest F1-score for all the datasets under investigation.

Performance of model on violence and non-violence detection in videos as presented in Fig. 5a–h. For the Hockey Fight Detection dataset, the visualization thereby demonstrates the spatial and temporal features learned by the model sufficient for violent behaviour detection during hockey games. Feature map results show where in the frame the model finds most relevant cues in terms of fights or aggressive involving instances, and the detection results denote the model's capacity to classify these instances as violent. Likewise, on the Movies dataset, the visualization shows how the model differentiates between violent clips and those that are not violent from a host of movie clips. The feature map results offer the detailed information of which of the visual inputs the model focuses on at the time of violence detection and the detection results indicate how effectively the model is able to classify violent segment and non-violent segment. In the same dataset, BEHAVE, the results of the detection are shown in a way that illuminates the model's functionality for identifying violent behaviour in real-life surveillance videos. Here, the feature map results may point at particular actions, contacts, or objects connected with violence, for example, fights, gestures. The detection results give researchers an insight into model's performance in certain environments and lighting conditions, enhancing feedback on model improvement. Last but not least, on the ViF dataset, the visualization demonstrates how the STCCLM-net model works detecting violence in the scenarios depending on video clips, sport incidents, streets, or staged performances. By exploring the feature map activation and the results of the detections. In general, the presence of the visualization of detection results allows for the desired representation of the STCCLM-net model on various datasets.

Table 2 Performance of threshold sensitivity analysis for background subtraction in terms of metrics

Datasets	Threshold value	Precision	Recall	F1-score
Hockey Fight	0.1	0.55	0.75	0.63
	0.3	0.65	0.85	0.73
	0.5	0.75	0.90	0.81
	0.7	0.85	0.95	0.89
	0.9	0.95	0.98	0.96
Movies	0.1	0.50	0.70	0.58
	0.3	0.60	0.80	0.68
	0.5	0.70	0.90	0.78
	0.7	0.90	0.95	0.96
	0.9	0.90	0.98	0.93
BEHAVE	0.1	0.45	0.65	0.53
	0.3	0.55	0.75	0.63
	0.5	0.65	0.85	0.73
	0.7	0.75	0.90	0.81
	0.9	0.95	0.95	0.95
ViF	0.1	0.40	0.60	0.48
	0.3	0.50	0.70	0.58
	0.5	0.60	0.80	0.68
	0.7	0.70	0.90	0.78
	0.9	0.90	0.95	0.96

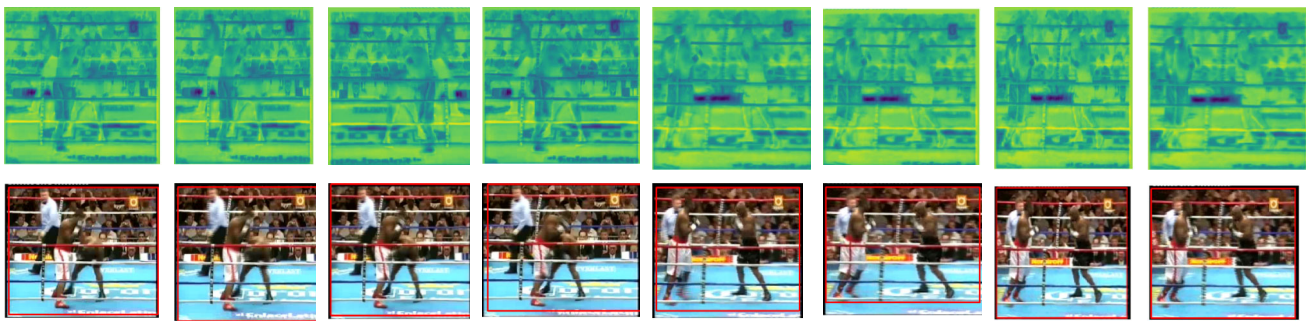
Figure 6 shows the analysis outcomes of the detection results on violence detection scenarios such as crowded scenes, rapid motions, and occlusion, in the four datasets of Hockey Fight, Movies, BEHAVE, and ViF, thereby disclosing some of the most valuable findings regarding the capability of the proposed mode in handling complicated video settings. The first row of input frames from each dataset demonstrates the difficulty of these scenes with crowded scenarios, high movement speed, and cases of target occlusion, whether partially or fully. The second row, heatmap results demonstrate that the proposed model highlights the regions of interest in each frame, where it expects violent motion. The signal strength and distribution of the heatmaps show how well the model is able to identify important features and motion in dense or obscured situations. Our last row, shown in the third row, provides the detection results and validates that indeed these scenes are correctly categorized as either violent or non-violent by the model. The analysis shows that the model is insensitive to the most critical aspects of the task—to recognize violent behaviour in cases, when the activities occur in sports (Hockey Fight dataset) or when several subjects interact in the frame and it is difficult to say which of them is violent (BEHAVE dataset). The overlying visualization in all four datasets clearly demonstrates that the model the work was done on is not only muddle-headed but also generalized and suitable for different settings of violence detection in videos.

(a) Hockey dataset result: Violence

(b) Hockey dataset result: Non-Violence

(c) Movies dataset result: Violence

(d) Movies dataset result: Non-Violence



(e) BEHAVE dataset result: Violence

(f) BEHAVE dataset result: Non-Violence

(g) ViF dataset result: Violence

(h) ViF dataset result: Non-Violence

Fig. 5 **a**–**h** Detection results of continuous frames on each dataset: **a** Hockey Fight Detection dataset, **b** Movies dataset, **c** BEHAVE dataset, **d** ViF dataset. First row represents the feature map result output using STCU and second row represents the classification result output for each dataset, respectively

From the confusion matrix analysis of four datasets including Hockey Fight Detection dataset, Movies dataset, BEHAVE dataset, and ViF dataset as depicted in Fig. 7, the proposed STCCLM-net, valuable information about the

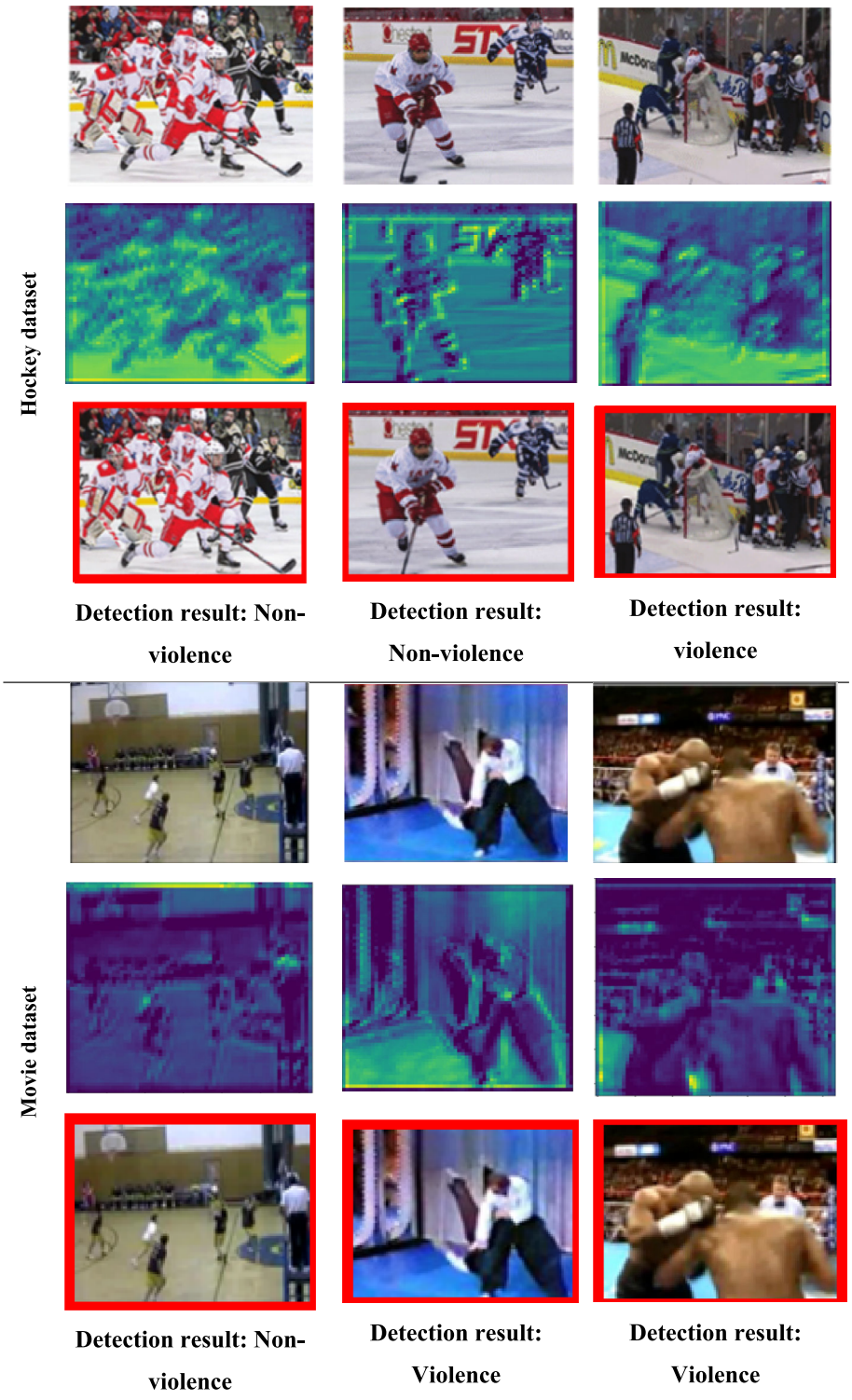
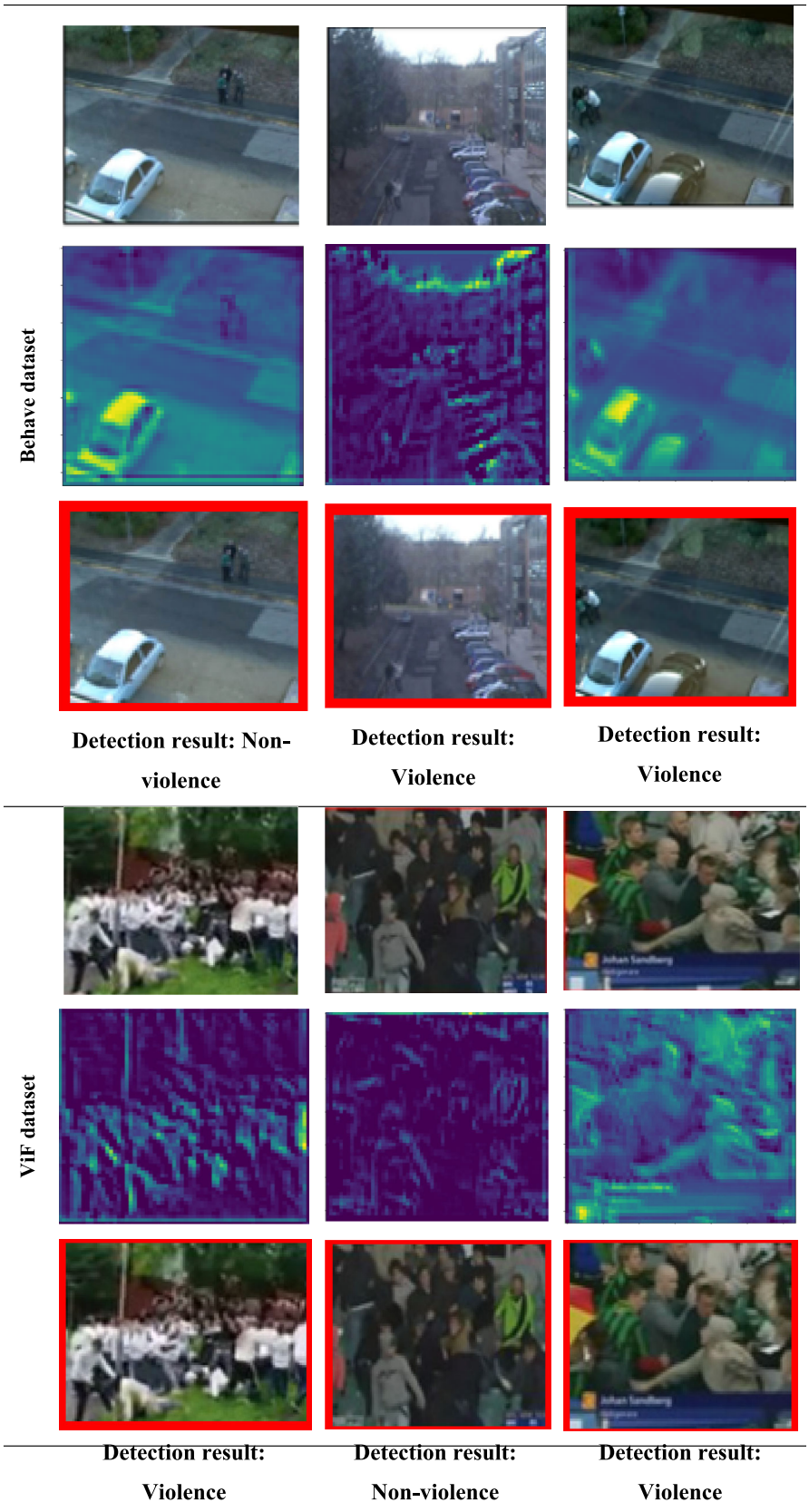


Fig. 6 Visualization of detection results on violence detection scenarios, like crowded scenes, rapid motions, and occlusion for four datasets, respectively. The first row shows that the input frames from four datasets. Second row shows the heatmap results and the third row shows the detection results for crowded, rapid motions, and occlusion scenes on four datasets, respectively

Fig. 6 continued



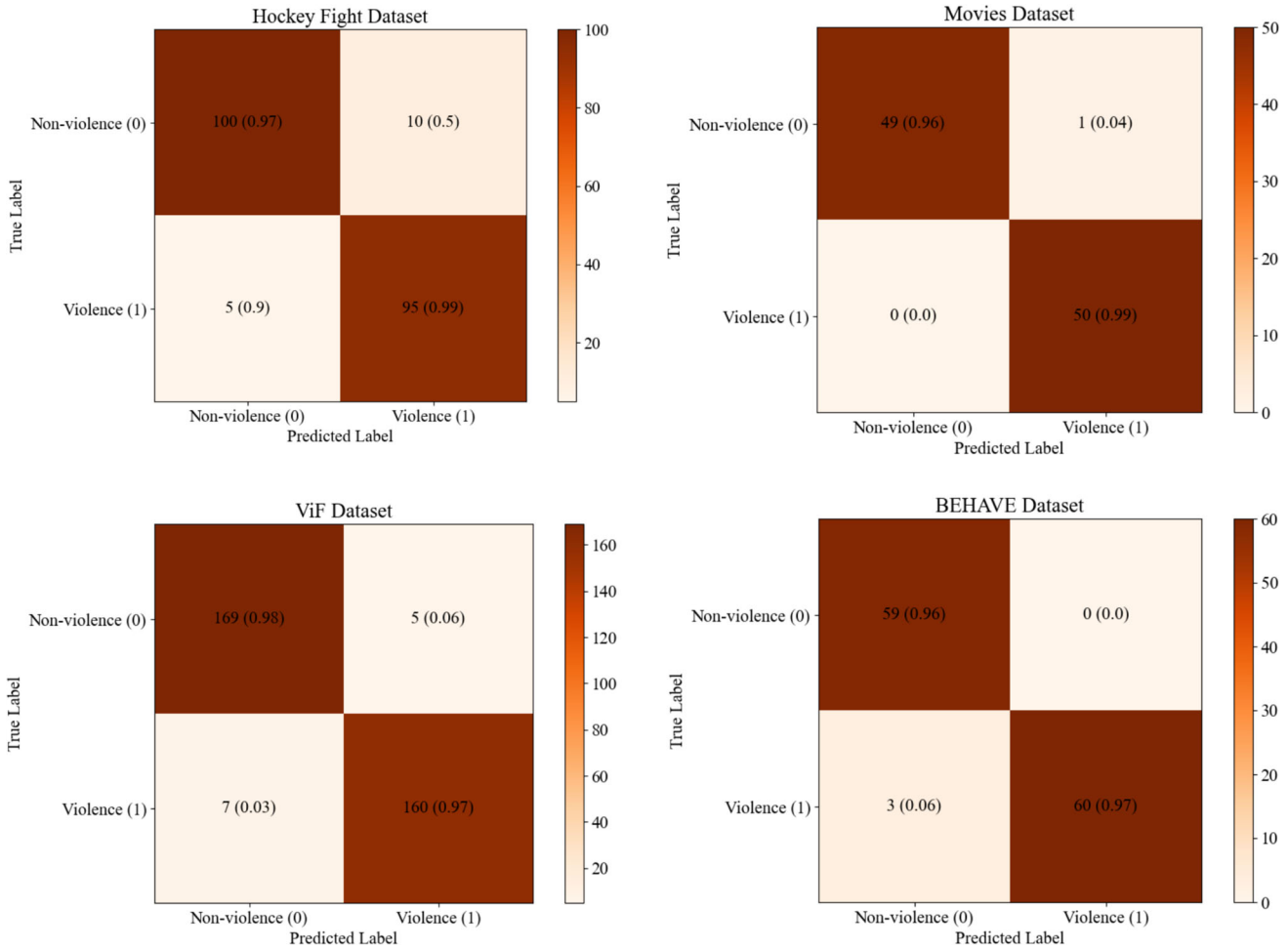


Fig. 7 Confusion matrix on each dataset: **a** Hockey Fight dataset, **b** Movies dataset, **c** BEHAVE dataset, **d** ViF dataset

violence detection performance across different scenario is obtained. Since it is possible to know the true labels and compare it with the predicted labels plotted against each other on two axes, namely as x and y axes, one is in a position to evaluate the ability of the model in classification of violent and non-violent events. High values of both TPs, i.e., the number of cases where the model correctly predicts that the video contains violence, and TNs, i.e., the cases where the model correctly predicts that the video does not, prove the efficiency of the model. On one hand, false positive, which are cases where non-violence snippets were categorized as violence, may suggest areas of improvement of the model or particular difficulties in violence classification in the particular datasets On the other hand, false negatives, which include violence snippets that were classified as non-violence, may indicate specific challenges in violence detection within each of the datasets. Finally, based on the observed results, on four datasets, the STCCLM-net stresses the labels.

On analysing the model accuracy and loss curve results using the STCCLM-net architecture yields valuable insights into the performance dynamics of the implemented models across the Hockey Fight dataset, Movies dataset, BEHAVE dataset, and ViF dataset, shown in Fig. 8a, b. Across all datasets, the accuracy curves exhibit a consistent trend of increasing accuracy with epochs, indicating the models' progressive learning and improved ability to correctly classify violence and non-violence events over time. However, variations in the rate of accuracy improvement are observed among datasets, reflecting differences in dataset characteristics such as scene complexity, variability, and the prevalence of violence. Notably, the loss curves for each dataset display a consistent downward trend, signifying effective error reduction and enhanced model performance throughout the training process. While the rate of loss reduction varies across datasets, likely influenced by dataset-specific challenges and nuances, the overall trend underscores the models' capacity to learn and capture violence-related patterns effectively. The STCCLM-net model converges quickly due to its well-designed architecture which utilizes multiple convolutional layers and incorporates dropout regularization (with a 0.2 probability)

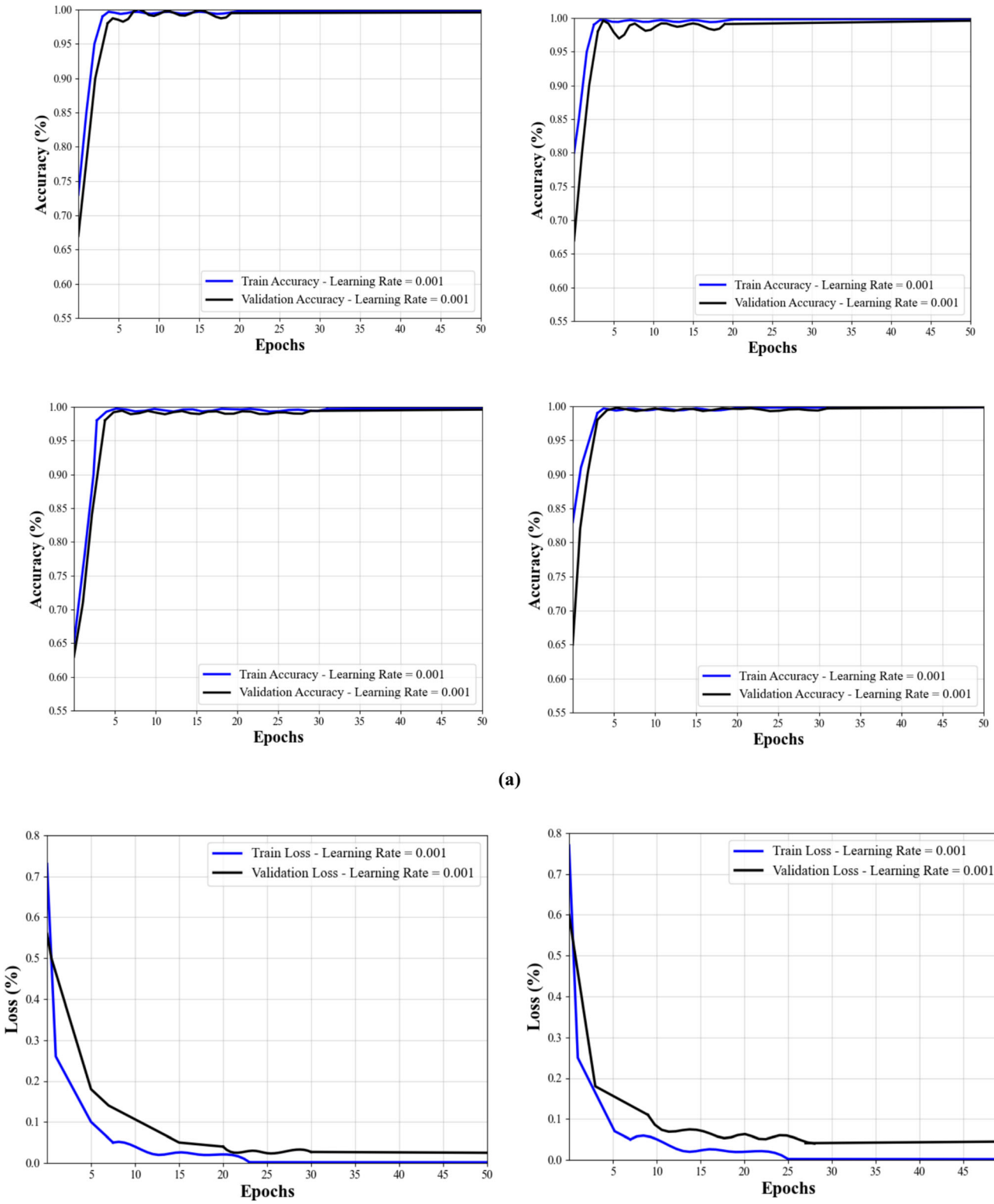
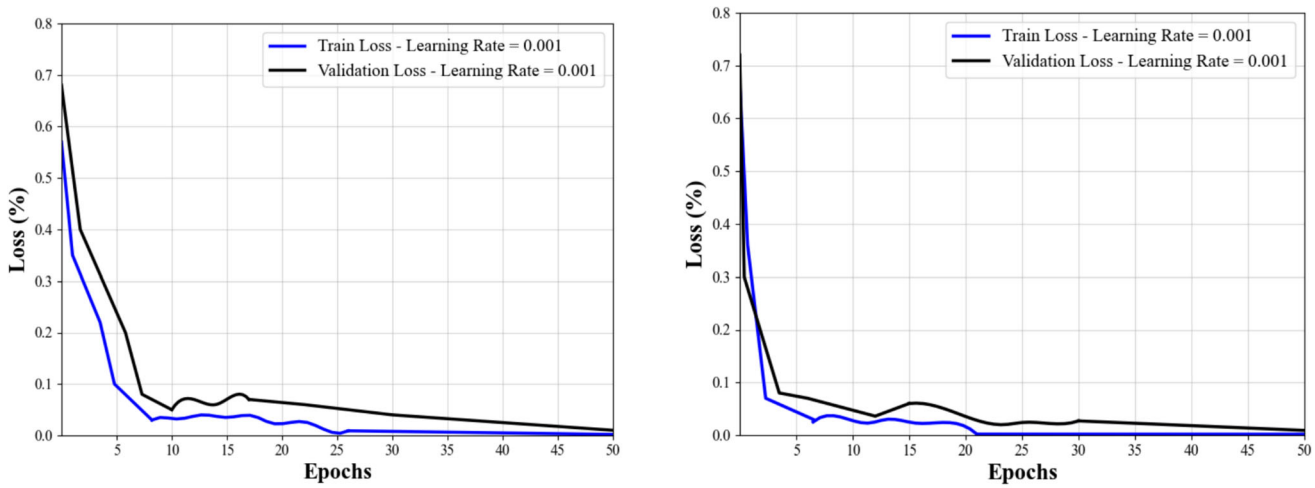


Fig. 8 **a**, **b** Model accuracy and loss curve results on each dataset: **a** Hockey Fight dataset, **b** Movies dataset, **c** BEHAVE dataset, **d** ViF dataset



(b)

Fig. 8 continued

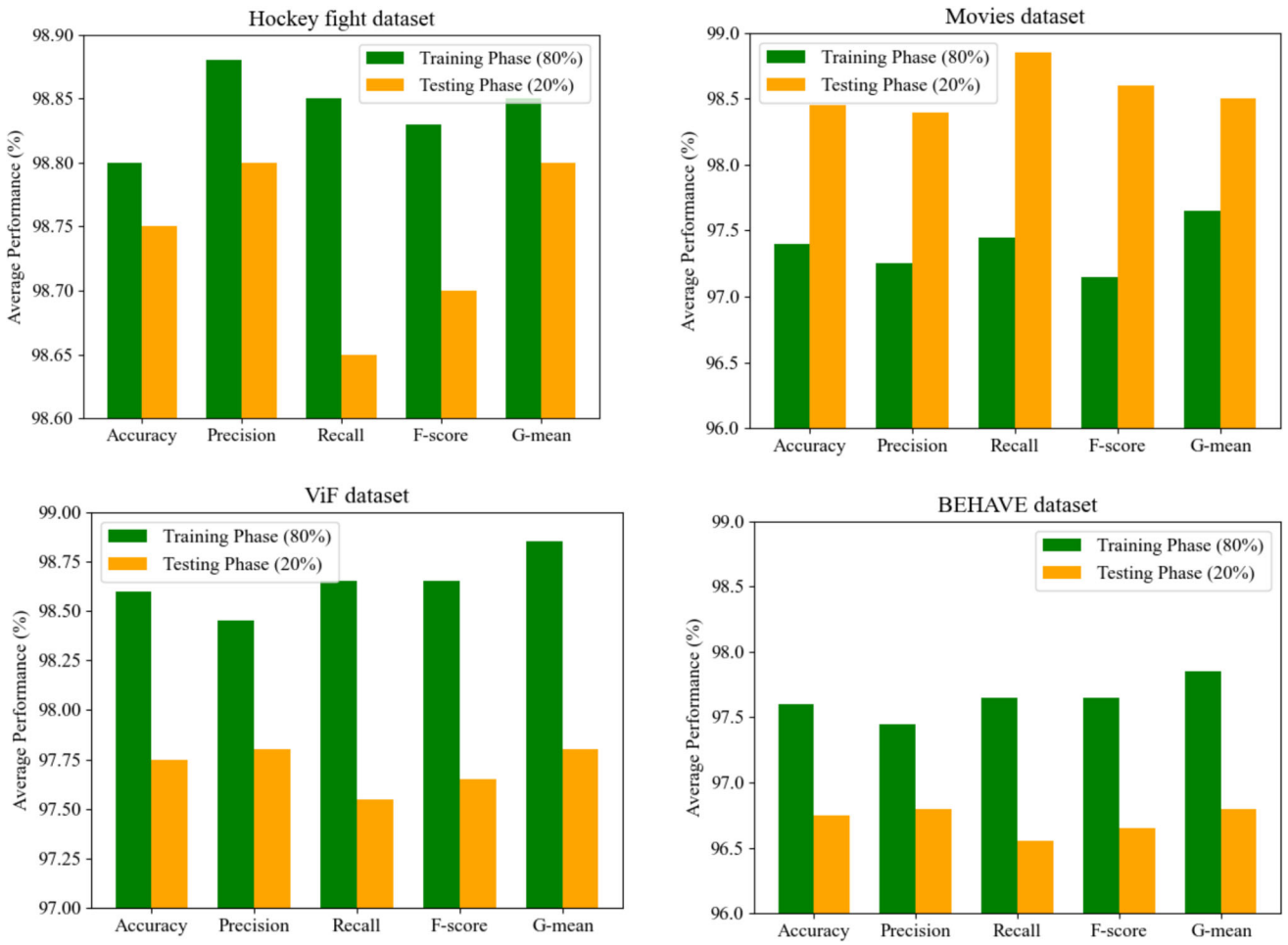


Fig. 9 Average performance results of the proposed model under various metrics (accuracy, precision, recall, F-score and G-mean) on each dataset

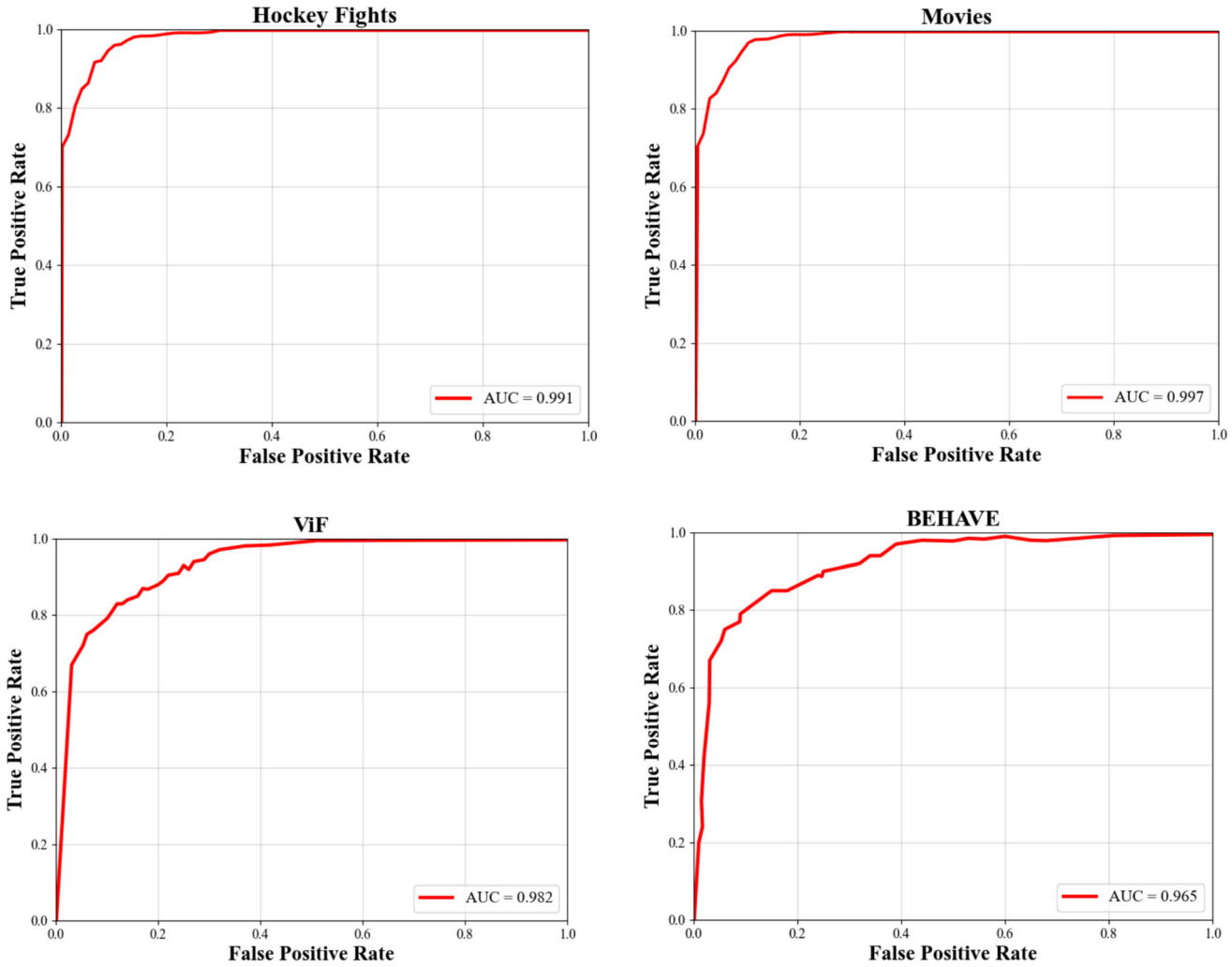


Fig. 10 Performance results of ROC curve on each dataset: **a** Hockey Fight Detection dataset, **b** Movies dataset, **c** BEHAVE dataset, **d** ViF dataset

Table 3 Classification results on each dataset in terms of various metrics (TP = true positive; TN = true negative; FP = false positive; FN = false negative)

Datasets	TP	TN	FP	FN	Precision	Recall	F1-score	Accuracy
Hockey Fight dataset	95	10	5	100	0.99	0.5	0.9	0.97
Movies dataset	49	1	0	50	0.96	0.04	0.0	0.99
BEHAVE dataset	169	5	7	160	0.98	0.06	0.03	0.97
ViF dataset	59	0	3	60	0.96	0.0	0.06	0.97

to enhance feature extraction while reducing overfitting risks. By leveraging pre-trained networks like VGG16 for spatial–temporal feature extraction, the model starts with a robust foundation. Gradient clipping is applied to ensure stable and faster convergence during training, particularly in deep learning networks, preventing issues like the exploding gradient problem. The model processes batches of videos, improving learning dynamics and accelerating convergence. The accuracy and loss curves demonstrate consistent improvement, indicating effective learning without overfitting. These findings underscore the robustness of the STCCLM-net architecture in learning from diverse datasets and highlight its potential for accurate violence detection in real-world scenarios.

ROC curve and AUC, as presented in the results in Figs. 9 and 10, present an informative picture of the ability of STCCLM-net model to produce the necessary discrimination for the Hockey Fight Detection dataset, Movies dataset, BEHAVE dataset, and ViF dataset. The ROC graphs represent the true positive rate (sensitivity) against

Table 4 Ablation study results of key components on four datasets in terms of AUC, accuracy, ROC and AP

Datasets	Key components	Accuracy (%)	AUC (%)	ROC (%)	AP (%)
Hockey Fight dataset					
	VGG16-net	92.50	91.0	88.7	89.5
	Without pre-processing	83.1	81.0	82.8	84.2
	With pre-processing	90.2	93.1	92.6	94.7
	VGG16-net + SCENet and TCENet	93.75	93.1	91.2	90.8
	VGG16-net + STCU	93.84	93.21	91.8	91.2
	VGG16-net + HFmapL	93.75	91.87	89.5	89.0
	Full STCCLM-net	98.38	96.76	97.5	97.2
Movies dataset					
	VGG16-net	91.72	90.4	87.5	88.0
	Without pre-processing	82.1	78.0	79.2	76.4
	With pre-processing	89.2	93.3	94.7	97.6
	VGG16-net + SCENet and TCENet	93.84	93.3	91.4	90.5
	VGG16-net + STCU	93.14	92.7	90.2	89.8
	VGG16-net + HFmapL	92.67	91.2	88.0	87.5
	Full STCCLM-net	98.20	97.68	99.1	98.8
BEHAVE dataset					
	VGG16-net	90.50	88.5	85.7	86.2
	Without pre-processing	76.4	76.3	78.5	79.9
	With pre-processing	87.9	88.9	92.1	93.5
	VGG16-net + SCENet and TCENet	91.80	90.2	87.4	87.9
	VGG16-net + STCU	92.25	91.1	88.6	88.5
	VGG16-net + HFmapL	91.80	89.8	86.2	85.5
	Full STCCLM-net	97.10	97.65	98.7	98.4
ViF dataset					
	VGG16-net	92.00	90.9	88.0	89.3
	Without pre-processing	83.2	81.1	76.8	78.5
	With pre-processing	86.8	91.2	89.9	85.8
	VGG16-net + SCENet and TCENet	93.45	92.6	90.3	91.0
	VGG16-net + STCU	93.12	91.9	89.7	90.5
	VGG16-net + HFmapL	92.75	91.0	88.8	89.2
	Full STCCLM-net	98.75	98.12	99.8	98.0

the false positive probability (1—specificity), across different discrimination points. The ROC curves of all the datasets show the upward trend, which proves that the STCCLM-net has high true positive rate and low false positive rate, which can prove its effectiveness in the classification of violence and non-violence events. Further, the AUC value remains consistently high at 0.99 for all the datasets thus confirming the generalizability of the method in detecting high discriminant patterns of the proposed STCCLM-net at different datasets. Such high AUC value suggests that the model can rank positive instances over negative ones significantly well for violence detections tasks and thus affirming our findings. These results demonstrate and confirm the efficiency and robustness of the STCCLM-net architecture regarding high-level results. The plotted results are given in Table 3 below.

4.2.2 Ablation experiment: impact of key components in STCCLM-net

In this section, the functions of the specific elements of the proposed STCCLM-net architecture are analyzed by performing an extensive ablation study. The objective is to examine the effect of each component, in terms of the accuracy of violence detection, across different datasets systematically. The following configurations are utilized for the ablation experiments are:

Baseline model: VGG16-net without any modifications.

With and without pre-processing: Employs key frame estimation and background subtraction estimation.

Feature extraction mechanism: Extraction models of SCENet and TCENet are utilized with the support of VGG16-net model.

Table 5 Comparison results of violent/non-violent classification models for the proposed model and the existing methods on each dataset

Datasets	Schemes	Accuracy (%)
Hockey Fight dataset	U-net + LSTM	96.1
	Conv LSTM	94.5
	CNN-LSTM	98.7
	CNN-BiLSTM	99.2
	MNAS network + ConvLSTM	99.0
	SSHA	97.0
	MoBSIFT	90.1
	Proposed method	**99.6**
Movies dataset	U-net + LSTM	99.1
	Conv LSTM	98.0
	CNN-LSTM	99.0
	CNN-BiLSTM	99.3
	MNAS network + ConvLSTM	99.0
	SSHA	99.0
	MoBSIFT	98.8
	Proposed method	**99.7**
Behave dataset	ViF + DL	82.02
	RVD + CNN	85.25
	MoWLD + SMGAO	87.16
	CNN-LSTM	95.3
	Cascade DNN	95.43
	Proposed method	**99.8**
ViF dataset	VGG-F + SVM	93.0
	CNN + Bi-ConvLSTM	90.5
	CNN + ConvLSTM	96.0
	3D-CNN-STF	98.0
	CNN + LSTM	98.21
	CNN-BiLSTM	98.5
	MNAS network + ConvLSTM	96.0
	Proposed method	**99.8**

The best results are indicated as bold font

STCU module: VGG16-net with the STCU integrated.

HFmapL classifier: VGG16-net with the HFmapL classifier mechanism but without the STCU.

Full STCCLM-net: The complete model incorporating SCENet, TCENet, STCU and the spatio-temporal attention mechanism.

Each configuration is evaluated on the following datasets: Hockey Fight dataset, Movies dataset, BEHAVE dataset, ViF dataset. Measures of accuracy, area under the curve, precision, recall, and F1 are measured. Table 4 presents the overall performance of all the model configurations as captured by the four datasets in the ablation study. We record performance metrics which include accuracy, AUC, precision, recall and F1-score. A summary for the contribution of the ablation study is presented in Table 4, showcasing the performance of each model architecture in the datasets.

The analysis results underscore the transformative impact of integrating multiple components into the STCCLM-net architecture for violence detection in videos. The performance gains across all datasets demonstrate that each component—SCENet, TCENet, STCU, and the spatio-temporal attention mechanism—contributes significantly to the model's efficacy. The baseline VGG16-net provides a solid foundation, but the incorporation of SCENet and TCENet reveals their essential role in capturing spatial and temporal nuances, respectively, which are critical for accurate violence detection. The introduction of the STCU module marks a pivotal improvement, particularly in precision and recall, which are crucial for reducing the risk of false negatives—a key concern in violence detection. This enhancement ensures that the model not only identifies violent incidents more reliably but also minimizes the likelihood of overlooking them,

Table 6 Number of parameters and FLOPs of each major component used in the STCCLM-net model

Component	Parameters (millions)	FLOPs (billions)
VGG16 Backbone	138	15.5
SCENet (Spatial Context)	5	2.3
TCENet (Temporal Context)	7	3.8
STCU (Spatial–Temporal Unit)	2	1.2
HFmapL	1.5	0.9

Table 7 Computational and parametric quantities of the STCCLM-net model

Component	Description	Computational complexity	Number of parameters
VGG16 Backbone	Convolutional layers and fully connected layers	$O(H \cdot W \cdot C \cdot K^2 \cdot F)$	$\sim$ 138,000,000
SCENet	Additional convolutional layers for spatial context	$O(H \cdot W \cdot F_{SC} \cdot K_{SC}^2 \cdot L_{SC})$	153,856
TCENet	Convolutions and attention for temporal context	$O(T \cdot H \cdot W \cdot F_{TC} \cdot K_{TC}^3 \cdot L_{TC})$	331,776
STCU	Fusion and context modelling unit	$O(F_s \cdot F_t \cdot F_{STCU})$	134,217,728
Classifier (HFmapL)	Intermediate feature map refinement	$O(F_{HFmapL} \cdot H_{HFmapL} \cdot W_{HFmapL})$	25,000,000

H and W = height and weight of the input image; C = number of channels in the feature map; K = kernel size for convolutional layers; F = number of filters in the convolutional layers; T = number of frames in the temporal sequence; L_{SC} = number of additional convolutional layers in SCENet; K_{SC} = kernel size for SCENet layers; F_{SC} = number of filters in SCENet; L_{TC} = number of 3D convolutional layers in TCENet; K_{TC} = kernel size for TCENet layers; F_{TC} = number of filters in TCENet; F_s = feature dimension from SCENet; F_t = feature dimension from TCENet; F_{STCU} = feature dimension in STCU; F_{HFmapL} = number of features in HFmapL; H_{HFmapL} and W_{HFmapL} height and width of the feature maps in HFmapL

which is vital in real-world scenarios where missing such events could have serious consequences. The full STCCLM-net configuration, which synergistically combines all these components, demonstrates the most significant performance improvements, with accuracy reaching up to 98.75% and AUC values up to 98.12%. These results indicate that the model is highly effective in distinguishing violent from non-violent scenes, achieving near-perfect ROC and AP scores. The consistently high values further confirm the model's robustness in maintaining a delicate balance between precision and recall, ensuring that it is not only accurate but also reliable in diverse video contexts. Overall, the analysis highlights the importance of a comprehensive approach that leverages the strengths of each component, resulting in a highly effective violence detection system that is well-suited for deployment in real-world applications.

4.2.3 Comparative results between the proposed and the existing methods

The performance of the proposed model is compared with existing methods such as U-net + LSTM [31], Conv LSTM [22], CNN-LSTM [44], CNN-BiLSTM [17], MNAS network + ConvLstm [21], SSHA [16], and MoBSIFT [14], Violent Flow dataset with Deep Learning (ViF-DL) [36], Recurrent Violent Detector-CNN (RVD-CNN) [37], Motion Weber Local Descriptor-Stochastic Multi-Goal Agent Optimization (MoWLD-SMGAO) [38], Cascade Deep Neural Network (Cascade DNN) [39], Visual Geometry Group—Fast Model—Support Vector Machine (VGG-F-SVM) [40], CNN + Bidirectional Convolutional LSTM (CNN-Bi-ConvLSTM) [41], CNN-Convolutional LSTM (CNN-ConvLSTM) [42] and 3D CNN-Spatiotemporal Feature (3D-CNN-STF) [43].

Table 5 however shows the comparative analysis of violent and non-violent event classification using different approaches including the present STCCLM-VGG16 net model and state of the art works. Table 5 shows the accuracy achieved on each of these models when tested on these datasets. The numerical analysis of the results across the four datasets: Relying on four datasets which are Hockey Fight, Movies, BEHAVE, and ViF—four datasets mentioned in the earlier sections—, the results show that the proposed method significantly outperforms other existing methods in the domain of violent/non-violent classification. It is to be pointed that applying the proposed method in the context of the Hockey Fight dataset provides the accuracy of 99.6% which is even higher than the recent achievements of CNN-BiLSTM—99.2% and MNAS network with ConvLSTM—99.0%. The best competitor, CNN-BiLSTM, drops by 0.4% in accuracy, demonstrating the proposed method's potential to learn and differentiate violent/non-violent motion in HI sports. The higher accuracy compared with other models as Conv LSTM (94.5%) and MoBSIFT (90.1%) reinforces the efficiency of the proposed technique mainly concerning abundant noise and

Table 8 Impact of frame rate variation performance in terms of metrics on four datasets

Datasets	Frame rate (FPS)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Hockey Fight	24	85.1	83.2	87.1	85.1
	48	87.3	85.5	89.2	87.3
	60	96.5	94.1	95.9	95.5
	120	92.1	91.3	92.9	92.1
	240	93.2	92.5	94.0	93.2
Movies	24	82.5	80.9	84.2	82.5
	48	85.1	83.5	86.8	85.1
	60	88.3	86.9	89.8	88.3
	96	96.2	96.3	97.2	96.2
	120	91.5	90.7	92.4	91.5
BEHAVE	25	78.1	75.9	80.4	78.1
	50	81.1	79.3	83.0	81.1
	60	95.2	92.5	96.1	94.2
	100	86.3	85.1	87.6	86.3
	200	87.9	86.7	89.3	87.9
ViF	30	80.3	78.5	82.2	80.3
	60	83.2	81.4	85.1	83.2
	90	86.1	84.5	87.8	86.1
	120	98.5	97.1	96.1	98.5
	180	90.1	89.2	91.2	90.1

complicated movement patterns of players in dynamic and crowded scenes of Hockey Fight dataset.

Likewise, in the Movies dataset, the proposed method even surpassed all the previously used methods with the maximum accuracy of 99.7% and just slightly being higher than CNN-BiLSTM, which reached only 99.3% accuracy. The high accuracy of the proposed method indicates its proficiency in processing and interpreting the varied and often subtle cues in movie scenes that differentiate violent from non-violent actions. When compared to other methods like U-net + LSTM (99.1%) and SSHA (99.0%), the proposed method's enhanced accuracy underscores its effectiveness in leveraging spatial and temporal features for action recognition. This improvement is also evident in the BEHAVE and ViF datasets, where the proposed method achieved a remarkable 99.8% accuracy, substantially outperforming traditional and deep learning approaches like ViF + DL (82.02%) and VGG-F + SVM (93.0%), respectively. These results not only establish the proposed method as the top-performing model across diverse datasets but also highlight its potential for practical applications in real-world scenarios requiring precise and reliable violent/non-violent classification.

4.2.4 Analysis of model parameters and FLOPs of the STCCLM-net model

The model's performance and efficiency can be evaluated through its parameters and floating-point operations per second (FLOPs). The STCCLM-net comprises several key components, including the VGG16 Backbone, SCENet (Spatial Context), TCENet (Temporal Context), STCU (Spatial–Temporal Unit), and hidden feature map layer (HFmapL).

In the STCCLM-net model, the number of parameters is influenced by the complexity of the individual components, such as the VGG16 network, SCENet, TCENet and STCU. FLOP measures the computational workload required to perform a forward pass through the network. This metric is particularly important for understanding the efficiency of the model during inference. FLOPs provide insight into how many operations (such as multiplications and additions) are needed to process an input, which directly impacts the speed and resource requirements of the model. FLOPs can be calculated based on the operations performed in each layer of the model. This often involves counting the number of multiplications and additions for convolutional layers, fully connected layers and HFmapL. The VGG16 Backbone provides the spatial feature extraction capability of the network. It contains approximately 138 million parameters and has around 15.5 billion FLOPs. SCENet (Spatial Context) module enhances the spatial context by focusing on local features within the spatial domain. SCENet has about 5 million parameters and approximately 2.3 billion FLOPs. TCENet (Temporal Context) designed to capture temporal dynamics, TCENet processes sequences of frames to extract temporal features. It includes roughly 7 million parameters and 3.8 billion

FLOPs. The STCU integrates both spatial and temporal features to provide a comprehensive context for classification. This unit has around 2 million parameters and 1.2 billion FLOPs. The HFmapL aggregates and refines the features extracted by the preceding modules. It consists of approximately 1.5 million parameters and 0.9 billion FLOPs. The number of parameters and FLOPs of each major component is shown in the Table 6.

4.2.5 Analysis of computational and parametric quantities of the STCCLM-net model

The analysis of computational and parametric quantities for the STCCLM-net model provides insight into its efficiency and performance capabilities. By examining the computational complexity and parameter counts of key components—VGG16 Backbone, SCENet, TCENet, STCU, and HFmapL—this analysis highlights the model integrates spatial and temporal features for effective video violence classification. Understanding these quantities is crucial for evaluating the model's balance between accuracy and resource demands. Table 7 provides a comprehensive overview of the resource requirements and complexity associated with the STCCLM-net. Firstly, the use of the VGG16 Backbone for spatial feature extraction, despite its substantial computational complexity and large parameter count, provides robust and high-quality feature representations crucial for accurate violence classification. Its deep architecture, while computationally intensive, ensures that the spatial features are captured with high granularity and detail, which is essential for distinguishing between violent and non-violent events in videos. The addition of SCENet and TCENet enhances the model's capability by introducing specialized mechanisms for spatial and temporal context enrichment. SCENet's relatively modest parameter count (153,856) compared to its contribution highlights its efficiency in augmenting spatial features without excessively increasing computational load. Similarly, TCENet's parameter count (331,776) is balanced by its ability to handle complex temporal dependencies, which is crucial for analyzing video sequences. The STCU's role in fusing spatial and temporal features adds significant value, with a high parameter count (134 million) reflecting its comprehensive approach to integrating features from both streams. The HFmapL further refines the feature representations, contributing to the model's overall performance with a moderate increase in parameters. Collectively, the model's design achieves a favorable trade-off between computational complexity and parameter efficiency, enabling high-performance classification while maintaining manageable computational resource requirements. This balance ensures that STCCLM-net is both powerful and practical for real-world applications in video analysis.

4.2.6 Analysis of frame rate influence across four datasets

The analysis explores the impact of frame rate variation on the accuracy and performance of proposed violence detection model across four distinct datasets: Hockey Fight, Movies, BEHAVE, and ViF. By evaluating the proposed model responses to different frame rates, the study aims to identify the optimal conditions for accurate violence classification. The Hockey Fight Dataset originally has a frame rate of 24 FPS, but can be increased to 48 FPS (doubling the original frame rate), 60 FPS (standard for sports videos), 120 FPS (high-speed recording for detailed analysis), and 240 FPS (extremely high frame rate for specialized applications). Similarly, the Movies Fight Dataset, which has a cinematic standard of 24 FPS, can be increased to 48 FPS, 60 FPS, 96 FPS (used in some high-frame-rate cinematic productions), and 120 FPS. The BEHAVE Dataset, with an original frame rate of 25 FPS (PAL standard), can be increased to 50 FPS, 60 FPS, 100 FPS, and 200 FPS. Lastly, the Violent Flow (ViF) Dataset, which has a standard frame rate of 30 FPS for web videos, can be increased to 60 FPS, 90 FPS, 120 FPS, and 180 FPS.

Table 8 reveals that specific frame rates achieve higher percentages due to the increased temporal resolution, enabling the models to capture more detailed and nuanced information about violent behavior. For instance, in the Hockey Fight dataset, the 60 FPS frame rate achieves the highest accuracy (96.5%) and F1-score (95.5%), indicating that this frame rate strikes an optimal balance between capturing sufficient detail and avoiding excessive data redundancy. As a result, shortening the time between frames, the model can better capture the essential moments of a fight, including fast punches and falls. The proposed method's architecture is designed to extract both spatial and temporal features, which gives the method the ability to effectively differentiate between fast movements that are characteristic of violent behavior. In the same case in the Movies dataset, 96 FPS frame rate yielded the highest accuracy (96.2%) and F1-score of (96.2%), making it most appropriate for detecting violence in movies. Therefore, using the BEHAVE dataset, the maximum accuracy is obtained at 60 FPS frame rate at 95.2% and F1-score as 94.2% which makes 60 FPS frame rate most appropriate when detecting actual violence. Currently, the multi-scale spatio-temporal network can focus on the analysis of the fight choreography that contributes to the increased classification accuracy. However, for ViF dataset, obtained the highest overall accuracy (98.5%) and F1-score (98.5%) at 120 FPS of web videos; thus, it can be concluded that this frame rate is the most appropriate for detection of violence in web videos. The improved performance at specific frame rates can be attributed to the increased ability of the proposed model to: the advantages are as follows: (i) motion

tracking that is able to capture fine details about behaviour and movement. (ii) Event detection that is able to capture sequences of events in faster speeds. (iii) Pattern and anomaly detection of violent behaviour and (iv) minimized false positives and negatives.

5 Conclusion

In this paper, we presented the multi-stream framework based on STCCLM-net for the delineation of violent and non-violent actions in video data for tackling some major issues such as complex scene and frequent movement. Subsequently, by combining a SCENet and TCENet on this basic model, the spatial and temporal features required for the action recognition are distinguished and learnt. At the preprocessing stage of this framework, which includes frame differencing and background suppression, the extraction of meaningful features, which are fundamental to violence detection, is improved. The STCCLM-net integrates spatial and temporal information by using spatial temporal collaboration unit and obtains superior performance in violent actions detection in video frames. This is because the model uses VGG16 Net for spatial feature extraction with accuracy and density that are important in the portrayal of the results of the visual data. In general, the STCCLM-net makes it possible to carry out the violence classification tasks at a high level of performance compared to the traditional approaches. These make it a potential candidate for practical uses on the variety of spaces and time granularity levels and for providing accurate and reliable localization of violent actions in dynamic video streams. However, the performance of the STCCLM model greatly depends on the selection of a greater number of hyper-parameters and therefore it is difficult to determine suitable hyper–parameters for various datasets. The work done in this paper could be extended in the future as more detailed investigations probing the incorporation of new contextual information like audio and text data into the proposed STCCLM model for realistic scenarios.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Barun Pandey, Upasana Sinha, Kapil Kumar Nagwanshi. The first draft of the manuscript was written by Barun Pandey and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. Conceptualization: Barun Pandey; Methodology: Barun Pandey, Upasana Sinha; Formal analysis and investigation: Barun Pandey, Upasana Sinha; Writing—original draft preparation: Barun Pandey, Kapil Kumar Nagwanshi; Writing—review and editing: Upasana Sinha, Kapil Kumar Nagwanshi; Supervision: Kapil Kumar Nagwanshi.

Funding There is no funding for this study.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest Authors declares that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants and/or animals performed by any of the authors.

Informed consent There is no informed consent for this study.

References

1. Singh S, Dewangan S, Krishna GS, Tyagi V, Reddy S, Medi PR (2022) Video vision transformers for violence detection. arXiv preprint http://arxiv.org/abs/2209.03561
2. Liu J, Dai P, Han G, Sun N (2023) Combined CNN/RNN video privacy protection evaluation method for monitoring home scene violence. Comput Electr Eng 106:108614
3. Marcondes FS, Durães D, Gonçalves F, Fonseca J, Machado J, Novais P (2021) In-vehicle violence detection in carpooling: a brief survey towards a general surveillance system. In: Distributed computing and artificial intelligence, 17th international conference. Springer International Publishing, pp 211–220
4. Mehmood A (2021) Abnormal behavior detection in uncrowded videos with two-stream 3D convolutional neural networks. Appl Sci 11(8):3523
5. Appavu N (2023) Violence detection based on multisource deep CNN with handcraft features. In: 2023 IEEE international conference on advanced systems and emergent technologies (IC_ASET). IEEE, pp 1–6
6. Tian Y, Pang G, Chen Y, Singh R, Verjans JW, Carneiro G (2021) Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4975–4986
7. Chelali M, Kurtz C, Vincent N (2021) Violence detection from video under 2D spatio-temporal representations. In: 2021 IEEE international conference on image processing (ICIP). IEEE, pp 2593–2597
8. Hashmi TS, Haq NU, Fraz MM, Shahzad M (2021) Application of deep learning for weapons detection in surveillance videos. In: 2021 international conference on digital futures and transformative technologies (ICoDT2). IEEE, pp 1–6
9. Traoré A, Akhloufi MA (2020) Violence detection in videos using deep recurrent and convolutional neural networks. In: 2020 IEEE international conference on systems, man, and cybernetics (SMC). IEEE, pp 154–159
10. Jianjie S, Weijun Z (2020) Violence detection based on three-dimensional convolutional neural network with inception-ResNet. In: 2020 IEEE conference on telecommunications, optics and computer science (TOCS). IEEE, pp 145–150
11. Liang Q, Cheng C, Li Y, Yang K, Chen B (2021) Fusion and visualization design of violence detection and geographic video. In: Theoretical Computer Science: 39th National Conference of Theoretical Computer Science, NCTCS 2021, Yinchuan, China, July 23–25, 2021, Revised Selected Papers 39. Springer Singapore, pp 33–46

12. Ahmad W, Munsif M, Ullah H, Ullah M, Alsuwailem AA, Saudagar AK, Muhammad K, Sajjad M (2023) Optimized deep learning-based cricket activity focused network and medium scale benchmark. Alex Eng J 73:771–779
13. Rendón-Segador FJ, Álvarez-García JA, Salazar-González JL, Tommasi T (2023) CrimeNet: neural structured learning using vision transformer for violence detection. Neural Netw 161:318–329
14. Febin IP, Jayasree K, Joy PT (2020) Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm. Pattern Anal Appl 23(2):611–623
15. Wang P, Wang P, Fan E (2021) Violence detection and face recognition based on deep learning. Pattern Recognit Lett 142:20–24
16. Mohammadi H, Nazerfard E (2023) Video violence recognition and localization using a semi-supervised hard attention model. Expert Syst Appl 212:118791
17. Asad M, Yang J, He J, Shamsolmoali P, He X (2021) Multi-frame feature-fusion-based model for violence detection. Vis Comput 37:1415–1431
18. Halder R, Chatterjee R (2020) CNN-BiLSTM model for violence detection in smart surveillance. SN Comput Sci 1(4):201
19. Mohtavipour SM, Saeidi M, Arabsorkhi A (2022) A multi-stream CNN for deep violence detection in video sequences using handcrafted features. Vis Comput 38(6):2057–2072
20. Fenil E, Manogaran G, Vivekananda GN, Thanjaivadivel T, Jeeva S, Ahilan AJ (2019) Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM. Comput Netw 151:191–200
21. Wen X, Lai H, Gao G, Xiao Y, Wang T, Jia Z, Wang L (2023) Video anomaly detection based on cross-frame prediction mechanism and spatio-temporal memory-enhanced pseudo-3D encoder. Eng Appl Artif Intell 126:107057
22. Garcia-Cobo G, SanMiguel JC (2023) Human skeletons and change detection for efficient violence detection in surveillance videos. Comput Vis Image Underst 233:103739
23. Qin Y, Xu H, Chen H (2014) Image feature points matching via improved ORB. In: 2014 IEEE international conference on progress in informatics and computing. IEEE, pp 204–208
24. Duan FF, Meng F (2020) Video shot boundary detection based on feature fusion and clustering technique. IEEE Access 8:214633–214645
25. Xia H, Song S, He L (2016) A modified Gaussian mixture background model via spatiotemporal distribution with shadow detection. SIViP 10:343–350
26. Dong E, Han B, Jian H, Tong J, Wang Z (2020) Moving target detection based on improved Gaussian mixture model considering camera motion. Multimed Tools Appl 79(11):7005–7020
27. He T, Liu Y, Yu Y, Zhao Q, Hu Z (2020) Application of deep convolutional neural network on feature extraction and detection of wood defects. Measurement 152:107357
28. Khan MA, Sharif M, Akram T, Raza M, Saba T, Rehman A (2020) Hand-crafted and deep convolutional neural network features fusion and selection strategy: an application to intelligent human action recognition. Appl Soft Comput 87:105986
29. Zhuang D, Jiang M, Kong J, Liu T (2021) Spatiotemporal attention enhanced features fusion network for action recognition. Int J Mach Learn Cybern 12(3):823–841
30. Bock S, Goppold J, Weiß M (2018) An improvement of the convergence proof of the ADAM-Optimizer. arXiv preprint http://arxiv.org/abs/1804.10587
31. Ye L, Yan S, Zhen J, Han T, Ferdinando H, Seppänen T, Alasaarela E (2022) Physical violence detection based on distributed surveillance cameras. Mob Netw Appl 27(4):1688–1699
32. Bianculli M, Falcionelli N, Sernani P, Tomassini S, Contardo P, Lombardi M, Dragoni AF (2020) A dataset for automatic violence detection in videos. Data Brief 33:106587
33. Serrano I, Deniz O, Espinosa-Aranda JL, Bueno G (2018) Fight recognition in video using hough forests and 2D convolutional neural network. IEEE Trans Image Process 27(10):4787–4797
34. Lopez DJ, Lien CC (2023) Two-stage complex action recognition framework for real-time surveillance automatic violence detection. J Ambient Intell Humaniz Comput 14(12):15983–15996
35. Gao Y, Liu H, Sun X, Wang C, Liu Y (2016) Violence detection using oriented violent flows. Image Vis Comput 48:37–41
36. Sharath Kumar YH, Naveena C (2023) A deep learning based system to estimate crowd and detect violence in videos. Artificial intelligence for societal issues. Springer International Publishing, Cham, pp 45–57
37. Mugunga I, Dong J, Rigall E, Guo S, Madessa AH, Nawaz HS (2021) A frame-based feature model for violence detection from surveillance cameras using ConvLSTM network. In: 2021 6th international conference on image, vision and computing (ICIVC). IEEE, pp 55–60
38. Naik AJ, Gopalakrishna MT (2022) Automated violence detection in video crowd using spider monkey-grasshopper optimization oriented optimal feature selection and deep neural network. J Control Autom Electr Syst 33(3):858–880
39. Baba M, Gui V, Cernazanu C, Pescaru D (2019) A sensor network approach for violence detection in smart cities using deep learning. Sensors 19(7):1676
40. Xia Q, Zhang P, Wang J, Tian M, Fei C (2018) Real time violence detection based on deep spatio-temporal features. Biometr Recognit 10996:157–165
41. Hanson A, PNVR K, Krishnagopal S, Davis L (2019) Bidirectional convolutional LSTM for the detection of violence in videos. In: Lecture notes in computer science 2019, pp 280–295
42. Jahlan HM, Elrefaei LA (2022) Detecting violence in video based on deep features fusion technique. arXiv preprint http://arxiv.org/abs/2204.07443
43. Ullah FUM, Ullah A, Muhammad K, Haq IU, Baik SW (2019) Violence detection using spatiotemporal features with 3D convolutional neural network. Sensors 19(11):2472
44. Ullah FU, Obaidat MS, Muhammad K, Ullah A, Baik SW, Cuzzolin F, Rodrigues JJ, de Albuquerque VH (2022) An intelligent system for complex violence pattern analysis and detection. Int J Intell Syst 37(12):10400–10422