Real-Time Video Violence Detection Using Shallow Convolutional Neural Networks

Nanfei Jiang

Investigation Department, Beijing Police College, Beijing, 102202, China

Corresponding author. E-mail: nanfei_jiang30@outlook.com

Received: Jan. 2, 2026; Accepted: Mar. 13, 2026

Detecting violent content in real-time videos has become vital when digital content management and safety for people are of the maximum priority. The research proposes a novel and effective hybrid framework, called Shallow Lightweight Convolutional Attention-based Unified Temporal Network (SLCAUT-Net), proposed to accurately identify violent actions in video streams while sustaining real-time processing competences. To extract spatial features, SLCAUT-Net uses a shallow convolutional neural network (SCNN) backbone, combined with motion-based features derived from optical flow and improved through a temporal attention mechanism. The proposed architecture uses a minimal number of convolutional layers to ensure fast inference while simultaneously capturing temporal dependencies and motion patterns that are critical to distinguishing violent from non-violent behavior. The dataset from Kaggle can be used to support training and valuation by using video clips of real violent and non-violent scenarios. Frame differencing, data augmentation, and lightweight attention modules are employed to increase robustness and reduce overfitting. Experimental assessments validate that SLCAUT-Net attains competitive accuracy (97%) while operating proficiently on low-resource devices, with atypical latency. The addition of temporal attention and motion cues with a shallow framework offers a novel solution to the challenges of violence detection in conditions that are dynamic and unrestricted. The research highlights the potential of hybrid shallow networks in real-time video surveillance and safety-critical applications.

Keywords: SLCAUT-Net, violence detection, shallow convolutional neural networks, real-time video analysis, computer vision, temporal attention.

http://dx.doi.org/10.6180/jase.202609_32.053

1. Introduction

Security cameras are increasingly deployed in public spaces to enhance safety, deter crime, and provide evidence in criminal trials. For law enforcement, rapid identification of significant events, like violent acts, is crucial for deterrence and crime reduction [1]. Public video surveillance systems offer valuable data for various security purposes, but the challenge of watching hours of video content impairs quick decision-making, hindering their effectiveness in preventing crime and violence [2].

Deep learning (DL), a key component of AI-driven computer vision, has led to advancements in surveillance systems, meeting the growing demand for safety in public areas, banks, and train stations [3]. The automatic identification of aggressive behavior in surveillance footage addresses issues like bullying, though performance challenges remain, as highperformance computers are typically required, making timely responses difficult [4].

An AIpowered industrial Internet of Things (IIoT) framework for violence detection can enhance security and notify law enforcement of violent incidents [5]. Vi-

sual stream monitoring, particularly in crowded areas or events like protests and sporting events, is vital for detecting violent acts and preventing potential dangers [6]. Action recognition systems are essential for combating assaults, harassment, altercations, and vandalism. Installed in public areas, prisons, and campuses, these systems can alert authorities to potential dangers, representing a significant technological advancement in today's world, where aggressive forces and extremists pose increasing threats [7].

Surveillance camera monitoring in public spaces is growing in importance due to advancing issues in public security and safety. Human oversight of security footage is crucial for identifying significant behaviors, extracting information, and triggering responses [8]. These systems are vital for monitoring high-security areas and tracking violent crimes such as robbery, theft, and public violence. Regular use of these systems ensures effective monitoring and control [9]. AI has demonstrated promise in violence detection, but ensuring operational efficiency and real-time crime detection is essential for public safety.

Timely communication of crime occurrences is also crucial [10]. CCTV cameras are increasingly used in both public and private sectors to monitor indoor and outdoor spaces, with traditional systems relying on stored footage to review incidents later [11]. In smart cities, surveillance cameras help with traffic control, violation detection, and weapon identification, playing a key role in crisis management in institutions like hospitals, schools, and retail stores [12].

SLCAUT-Net integrates motion-based features and temporal attention mechanisms to efficiently detect violent actions with high accuracy and low computational cost. By combining shallow convolutional networks (SCNN) with temporal attention and motion-based features, SLCAUT-Net is highly effective for real-time violence detection, even on low-resource platforms, making it stand out from other models with higher resource requirements. The system's ability to detect aggression in real-time is crucial for safety-critical applications, ensuring quick processing with minimal latency. Motion-based features and data augmentation enhance its robustness, ensuring reliable performance in dynamic and unstructured environments.

Section 1 introduces the concept of video violence detection, Section 2 reviews existing methods, Section 3 presents SLCAUT-Net, focusing on temporal attention, motion properties, and SCNN, while Section 4 evaluates the model's performance. The contributions and future research directions are outlined in Section 5.

2. Materials and methods

2.1. Related Works

AI for semi-automated violence identification aims to improve public safety by automatically detecting violent situations. Among four trained models, the Vision Transformer achieved the highest accuracy (99%) in identifying violent situations, demonstrating the potential of AI in public safety and continuous monitoring [13]. The Large-scale Anomaly Detection (LAD) database is a new standard for detecting anomalies in video sequences, identifying fourteen different abnormalities in 2000 video sequences [14]. Multi-task deep neural networks (DNNs) were introduced for anomaly detection, with a 3D neural network handling local spatial data and a recurrent CNN managing global temporal features. "Video violence recognition" involves detecting violent actions, like fights, in video content using automated systems, while "Large-scale Anomaly Detection" focuses on identifying unusual patterns in large video datasets to detect rare or abnormal events. The proposed deep learning architecture for detecting unsuitable content in videos, like those on YouTube, used EfficientNet-B7 CNN trained on ImageNet [15].

SLCAUT-Net integrates a shallow CNN for rapid spatial feature extraction and temporal attention to prioritize significant frames, improving the accuracy of real-time violence detection with motion-based features using optical flow. Additionally, BiLSTM and attention mechanisms were employed to retain more contextual information compared to traditional machine learning classifiers. A weakly supervised anomaly detection system, using Temporal Context Aggregation (TCA) and Prompt-Enhanced Learning (PEL), showed improved detection performance with fewer parameters and lower computational cost [16]. Video anomaly detection (VAD) is gaining traction for practical applications but still struggles with the classification of complex events [17].

The Video Assistant Referee (VAR) task, utilizing cross-modalities for anomaly detection, showed improved semantic linkages and video analysis [18]. The AI-based VD-Net architecture for identifying violent activity in public and private spaces uses lightweight CNN blocks for accurate detection and timely alarms. The proposed model outperforms current approaches in crowd anomaly detection, improving surveillance in crowded areas [19]. A DL framework using CNN and support vector machines (SVM) was also suggested for identifying irregularities in video surveillance, including applications like monitoring crowd activities during the Hajj [20].

2.2. Area of Improvements

Video anomaly and violence detection techniques have advanced, but challenges remain, especially with current models that use complex topologies, struggling with scalability and real-time processing on large video datasets. Models like vision transformers and multi-task DNNs require improvements in contextual awareness and accuracy for dynamic, congested environments. SLCAUT-Net addresses these issues by using a shallow convolutional network for fast inference, combining temporal attention and motion-based features to improve detection precision in complex settings.

3. Results and discussion

3.1. Methodology

SLCAUT-Net combines motion data from optical flow with an SCNN for efficient spatial feature extraction, capturing dynamic patterns through temporal attention. It enhances resilience with frame differencing and data augmentation, and its shallow convolutional layers enable quick inference. Motion-based features help focus on rapidly changing regions, while temporal attention prioritizes significant frames, improving detection in dynamic environments like crowded spaces. The integration of SCNN for spatial extraction and temporal neural networks enables SLCAUT-Net to capture both immediate and long-term action dynamics, making it ideal for real-time video analysis in surveillance. The complete structure of the suggested approach for video violence detection is represented in Fig. 1.

3.2. Dataset acquisition

The open-source Kaggle structure was used to gather the Smart-City CCTV Violence Detection Dataset (SCVD): https://www.kaggle.com/datasets/toluwaniaremu/smartcity-cctv-violence-detection-dataset-scvd/data. The SCVD standard is a new specification that incorporates phone camera recordings into existing violence detection datasets, like the NTU CCTV-fights database and the real-time violence scenarios dataset (RLVS). It is the first to include a class for weapons detection in videos, allowing the dataset to consider any portable gadget that could pose a threat to people or property as a weapon.

- Data Exploration

  The most important attribute, is flexibility, which is followed by other factors like weight and balance. A similar feature importance analysis might be used in video violence detection research to determine which characteristics are most important for precise violence identification in CCTV footage. The model can concentrate on crucial elements for better and more effective violence identification by giving priority to the most significant traits.

3.3. Preprocessing

Preprocessing for video violence detection involves making data more useful through enhancements like frame differencing, motion transformations, and synthetic frame creation. These advancements emphasize motion dynamics (action) and scene transitions, enhancing the resilience of the model and improving violence detection accuracy.

3.3.1. Data Augmentation

Data augmentation techniques are essential for improving model performance in video violence detection, especially when sparsely labeled data is used. These techniques generate realistic examples to maximize model generalization and reduce overfitting. Evolutionary-based generative methods are used for audio-based applications, and learning reinforcement frameworks automate model performance improvement. Data augmentation strategies can be extended to video violence detection by manipulating video frames and audio data with videos. The research aims to enhance model robustness in dynamic settings and acquire underrepresented datasets of in-car violence.

3.3.2. Frame differences

The frame difference can serve as a valuable method for displaying motion and movement between consecutive frames and identifying violent actions. More specifically, the difference between frame $_s$ and the next frame, frame $_{s+1}$, is obtained in this Eq. (1):

$$\text{frame\_dif}_s = \text{frame}_{s+1} - \text{frame}_s \tag{1}$$

This process collects motion and frame changes that are critical for identifying sudden or rapid behaviors in video footage. The result of omitting the consecutive frames accentuates the motion data and enables the model to detect large changes in the scene, indicating possible violent actions. When applied to in-context or real-time detection of violent actions using the program, the output conveys the motion dynamics while keeping the geospatial context.

3.4. Spatial Feature Extraction and detection of Real-time video violence using SLCAUT-Net

The SCNN in SLCAUT-Net emphasizes aggressive behaviours and important body interactions by extracting crucial low-level spatial characteristics from video frames. To emphasize salient points in time, at which relevant events, such violent events, take place, the T-Net performs with

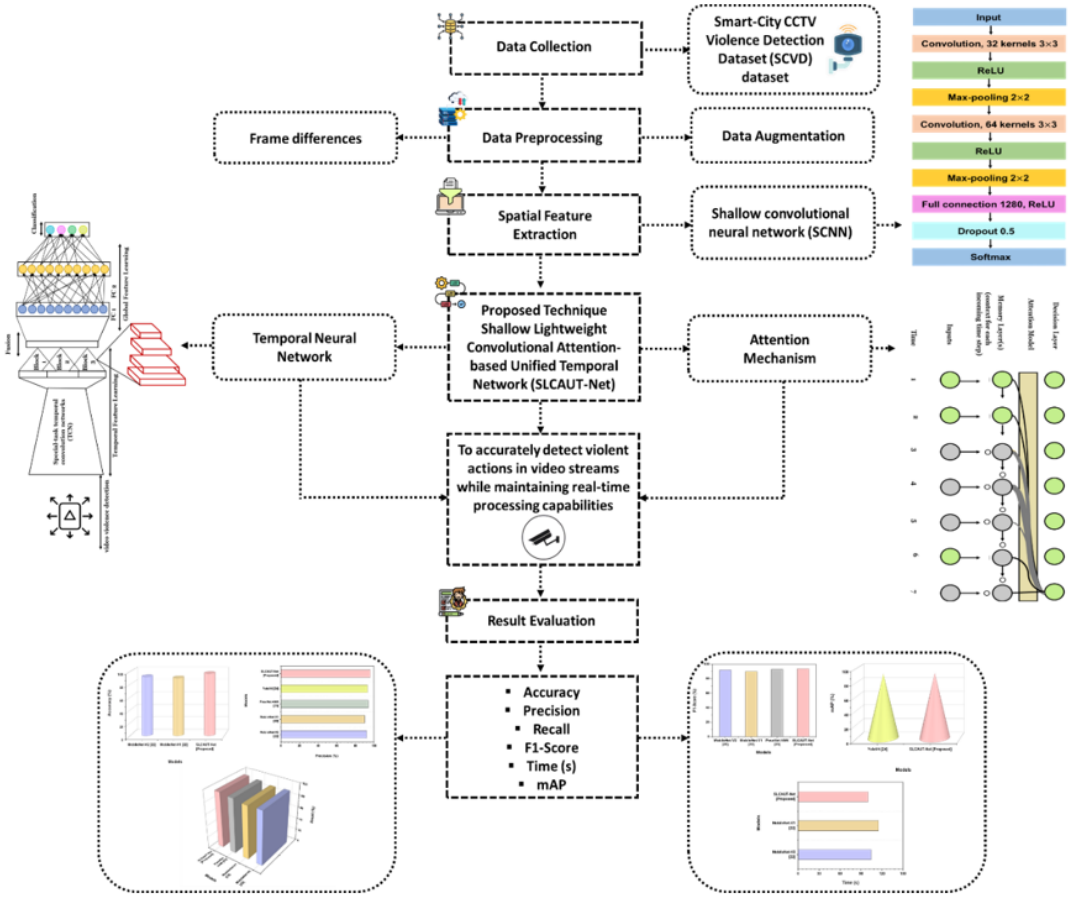

Fig. 1. The suggested method's detailed sequence for detecting video violence.

the Attention mechanism. This approach ensures real-time performance also optimizes the processing performance by prioritizing the relevant frames. By reducing the inclusion of non-essential frames, the T-Net increases detection performance and reduces computational cost.

3.4.1. Shallow convolutional neural network (SCNN)

The research describes a SCNN model, an end-to-end DL architecture that is designed for real-time identification of violent video. This model is suitable for surveillance or safety critical applications because of its efficiency with respect to limited hardware resources, high dependability and accuracy with better processing latency. The model accepts video frames from time at regular intervals, which were then pre-processed into a grayscale image by the model to capture the specific features of violent acts, such as human pose motions and body part interactions. The softmaxFunction is defined as the following Eq. (2):

$$\text{Softmax}(z)_j = \frac{f^{z_j}}{\sum_{j=1}^{m} f^{z_j}} \tag{2}$$

Where z_j is the yield of the j^{th} node of the output layer, and m indicates how many nodes there are in the output layer, and likewise how many categories are present in the specific classification task. The softmax function converts the model's output to an expected probability. Through the Rectified Linear Unit (ReLU) activation function, nonlinearity is incorporated into the model's structure. In Eq. (3), ReLU permits the training of intricate feature representations that are capable of identifying minute differences between violent and non-violent behavior.

$$e(w) = \max(0, w) \tag{3}$$

The SCNN model in SLCAUT-Net uses a convolutional block to capture high-dimensional spatial features of violent behaviors, with ReLU activation and dropout regularization to prevent feature dependency. The global feature aggregator layer has 1280 neurons, and the output layer uses softmax to normalize the final feature vector. This shallow model provides acceptable performance with low latency, making it efficient for real-time video analysis. However, varying video resolutions, frame rates, lighting, and camera angles can affect its performance, requiring

adaptation for robust detection. The shallow architecture enables faster inference but sacrifices some depth in feature extraction, balancing computational efficiency with the ability to capture complex behaviors.

3.5. Attention Mechanism

The framework for detecting real-time video violence uses an AI-based approach that utilizes attention processes to maximize the video mechanisms. The algorithm differentiates between violent and non-violent actions by highlighting the most important elements of an action.

The attention process is characterized by a weighted sum of video frame attributes, with a softmax function used to ensure all attention weights sum up to one. The output is produced by multiplying the importance weights by the appropriate frame attributes and adding the results. The method is described in Eq. (4):

$$d_q^j = \frac{\exp\left(h_q^j\right)}{\sum_{l=1}^{o} \exp\left(h_q^l\right)} \tag{4}$$

The critical scores in this Eq. (5) are scaled using the exponential function, where d_q^j is the weight of attention assigned to the j^{th} input at time q. The denominator normalizes the weights, ensuring that the total of all inputs equals 1.

$$h_q^j = w_h^Q \sigma\left(V_h\left[e_{q-1}, b_{q-1}\right] + X_h e_q + c_h\right) \tag{5}$$

The video frame's significance at time r is represented in this instance by h_q^j. To determine the importance, the context vector b_{q-1}, the current hidden state e_q, as well as the earlier concealed state e_{q-1} are utilized. The parameters V_h, X_h, and c_h are taught by the training method. Eq. (6) uses the function of non-linear activation σ, which provides the model non-linearity and assists in identifying complex patterns in the video data, to assist the system better differentiate between violent and non-violent behaviors:

$$\bar{z}_q = \sum_{q}^{Q} d_q^Q e_q \tag{6}$$

3.5.1. *Temporal Neural Network (TNN)*

The video clip can be viewed as a temporal network in which every frame or series of frames is regarded as a state $w(s)$, in the system. To identify changes in the network state, a set of control nodes is associated with each action or behavior. These control nodes could be generated using a variety of violent action-signaling video properties, such as motion vectors and appearance features. Eq. (7) maintains their genetic form:

$$w(s+1) = B(s)w(s) + A(s)v(s) \tag{7}$$

Where, the system state at time s is represented by $w(s) \in \mathbb{R}^M$, which in the context of video detection can represent states or characteristics extracted from video frames, $B(s) \in \mathbb{R}^{M\times M}$, the adjacency depicts the dynamics of transitions between various frames or activities, the control nodes at time s are represented by $v(s) \in \mathbb{R}^{M_C(s)}$, which could be specific features or areas of the video where notable behavioral changes are anticipated, the input points, or features or areas of the frame affecting the detecting process, are represented by $A(s) \in M_C(s)^{M\times M_C(s)}$. It can be accomplished by using the proper control nodes to reach the target state $w\left(s_e\right)$, which represents a violent incident in the video. The state evolution of the temporal network can be expanded as follows in Eqs. (8) to (12):

$$\begin{aligned} w\left(s_e\right) &= B\left(s_e-1\right) B\left(s_e-2\right) \hat{B} \cdot \hat{B} \cdot \hat{B} \cdot B\left(s_e+1\right) A\left(s_e\right) v\left(s_e\right) \\ &+ \cdots + B\left(s_e-1\right) A\left(s_e-\right) v\left(s_e-1\right) v\left(s_e-1\right) \end{aligned} \tag{8}$$

The matrix of controllability is slated for (9):

$$\begin{aligned} D\left(s_t, s_e\right) &= \left[B\left(s_e-1\right) B\left(s_e-2\right) \ldots B\left(s_e+1\right) A\left(s_t\right); \ldots\right. \\ &\left.; B\left(s_e-1\right) A\left(s_e-2\right); A\left(s_e-1\right)\right] \end{aligned} \tag{9}$$

When two matrices are concatenated to form $[B; A]$, and $D\left(s_t, s_e\right) \in \mathbb{R}^{M\times M_C}$. Consequently:

$$w\left(s_e\right) = D\left(s_t, s_e\right) V \tag{10}$$

Here,

$$V = \left[V\left(s_e\right)^S; V\left(s_e+1\right)^S; \ldots; V\left(s_e+1\right)^S\right]^S \in \mathbb{R}^{M_C} \tag{11}$$

In this case, the driver nodes C(s) have complete control over the temporal network if:

$$\operatorname{rank}\left(D\left(s_t, s_e\right)\right) = M \tag{12}$$

The temporal network dynamics in detecting video violence are influenced by changes in camera angles or video content, which is crucial for comprehending how violent acts appear or change in video sequences, similar to how temporal networks develop over time in control theory. The challenge with this is the need to find the rank of the controllability matrix and the number of control nodes $(M_C(s))$ at each time step, which complicates this procedure. In video violence detection, processing large numbers of video frames continuously leads to high computational costs, which grow exponentially over time. SLCAUT-Net

uses optical flow to capture motion dynamics by analyzing pixel movement between frames, calculating object velocity, and focusing on dynamic regions with rapid movement. This enhances the model's ability to detect violent actions in real-time, optimizing detection performance.

3.5.2. *Combining Temporal and Spatial Features to accurately and Instantaneously Identify Video Violence*

The attention mechanism integrates motion-based characteristics from optical flow and spatial features from the shallow convolutional backbone with temporal dependencies to create the fusion of spatial and temporal characteristics. This hybrid approach enables us to accurately detect aggression by using both spatial and temporal signals, ensuring high accuracy and real-time performance with no computational overhead. Algorithm 1 shows the SLCAUT-Net algorithm for video violence detection.

SLCAUT-Net is a real-time model that uses spatial and temporal data, along with an attention mechanism, to classify violent content in videos. It requires powerful hardware for real-time processing, improving detection accuracy and enabling timely interventions. While it excels in most scenarios, SLCAUT-Net struggles with detecting subtle or complex violent behaviors, especially in dynamic environments with cluttered backgrounds or fast-moving subjects. It adapts to lighting variations and camera angle changes, but occlusions can still affect detection accuracy by hiding subjects.

3.6. Comparative Analysis

To improve real-time video violence identification, an AI-based strategy was set into place. Basic performance measurements were recall, time (s), accuracy, F1-score, precision, and mAP. The preferred approach, SLCAUT-Net, was contrasted with baseline models such as MobileNet-V1 [20], MobileNet-V2 [20], PoseNetArtificial Neural Network (PoseNetANN) [21], and you only look once version 4 (YoloV4). This evaluation demonstrates that SLCAUT-Net can provide efficient, accurate real-time violence detection while outperforming competing models in terms of speed and classification accuracy.

- Accuracy

  Accuracy is described as the percentage of appropriately recognized frames among all processed frames. For real-time violence detection in dynamic circumstances, more accuracy means better decisions and fewer misclassifications. It evaluates how well the technique distinguishes between violent and non-violent performances in a video clip overall. Table 1 present the findings of the accuracy comparison.

Table 1. Comparison of video violence detection accuracy results.

Models	Accuracy (%)
MobileNet-V2 [20]	92%
MobileNet-V1 [20]	90%
SLCAUT-Net [Proposed]	97%

- Precision

  The proportion of violent frames that the model correctly detects out of all the frames it labels as violent is known as precision. Violent actions are identified based on the model's positive prediction accuracy. Real-time violence detection in dynamic video streams with high accuracy results in fewer false positives. Table 2 display the findings of the precision comparison.

Table 2. Comparison outcomes of precision for video violence detection.

Models	Precision (%)
MobileNet-V2 [20]	92%
MobileNet-V1 [20]	90%
PoseNet ANN [21]	94%
YoloV4 [22]	93%
SLCAUT-Net [Proposed]	96%

- Recall

  The model's recall gauges how well it can recognize violent frames in videos. It is the percentage of real violent frames (true positives + false negatives) to true positive detections. High recall shows that the model is effective at identifying the majority of violent incidents, reducing the number of missed detections in surveillance situations. Table 3 shows the findings of the recall and F1-Score comparison.

Table 3. Comparison outcomes of recall and F1-score for video violence detection.

Models	Recall (%)	F1-Score (%)
MobileNet-V2 [20]	91%	92%
MobileNet-V1 [21]	90%	90%
PoseNet ANN [21]	93%	93%
SLCAUT-Net [Proposed]	95%	94%

- F1-Score

  By balancing recall and accuracy, the F1-score is a performance statistic that offers a single assessment of the model's capacity to recognize aggressive behavior. It

Algorithm 1. SLCAUT-Net algorithm for video violence detection

```
import tensorflow as tf
import from tensorflow.Keras.layers Attention,LSTM,Dense,Flatten,Softmax,ReLU,
Dropout,Conv2D,and MaxPooling2D
import Sequential from tensorflow.Keras.models
timeDistributed from tensorflow.Keras.layers import
def SLCAUT_Net(input_shape=(28,28,1),num_classes=2,timesteps=10):
model = Sequential()
input_shape=(timesteps,) +
input_shape)) model.add(TimeDistributed(Conv2D(32,kernel_size=(3,3),strides=
(1,1),padding='same')
TimeDistributed(ReLU()) is added to the model.
time distributed(MaxPooling2D(pool_size=(2,2)))) model.add(
TimeDistributed(Conv2D(64,kernel_size=(3,3),padding='same')) as model.add(
TimeDistributed(ReLU()) is added to the model
time distributed(MaxPooling2D(pool_size=(2,2)))) model.add(
(TimeDistributed(Flatten()))) model.add
LSTM(128,return_sequences=False) model.add(
(Dense(1280))model.add
(ReLU()) model.add
(Dropout(0.5)) model.add
(Attention()) model.add
Density(num_classes) + model.add
Softmax(model.add)
return model
model = SLCAUT_Net()
model. compile(metrics = [ 'accuracy' ], loss ='categorical '_crossentropy , optimizer ='adam'
model. summary()
```

is highly helpful when there is an imbalance between the aggressive and non-violent courses. The F1-score, which calculates the harmonic mean of accuracy and recall, offers a thorough assessment.

- Mean Average Precision (mAP)

  The efficacy of recognition models is evaluated using a metric known as mAP. The model's ability to identify violent behaviors is evaluated using mAP, which computes the average accuracy across several categories and times. A higher mAP indicates better detection accuracy, which balances false positives and false negatives over a complete dataset. Table 4 present the findings of the mAP comparison.

Table 4. Comparison outcomes of mAP for video violence detection.

Models	mAP (%)
YoloV4 [22]	91.73
SLCAUT-Net [Proposed]	93

- Time (s)

  Time is the amount of time needed for the model to analyze and categorize every video frame. It is a crucial performance indicator that shows how quickly inference occurs in real time. Reducing this time allows for efficient processing, which guarantees quick identification of violent moments while preserving accuracy in dynamic video settings. Table 5 display the outcomes of the time (s) comparison.

Table 5. Comparison outcomes of time (s) for video violence detection.

Models	Time (s)
MobileNet-V2 [20]	104.37
MobileNet-V1 [20]	114.92
SLCAUT-Net [Proposed]	100.2

In real-time video violence detection, existing techniques such as MobileNet-V2, MobileNet-V1, PoseNet ANN, and YoloV4 have several drawbacks. Though these models are effective in detecting and classifying objects, are highly latency-prone and fail to effectively process motion-dependent features and temporal correlation, which are critical for accurate violence detection. Even though MobileNet [20] models are light, detection accuracy is poor, especially for complex movements in dynamic scenarios. YoloV4 [21] is a wonderful technology, but it can be computationally expensive for real-time usage, and PoseNet

ANN can struggle with pose-specific scenarios.

4. Conclusion

Real-time video violence detection is crucial for moderating digital content and enhancing public safety. SLCAUT-Net, a hybrid architecture, effectively detects violent events from video streams by using a shallow convolutional neural network (SCNN) for rapid spatial feature extraction, motion-based data, and a temporal attention mechanism to improve detection accuracy. The model, trained on the SCVD dataset, achieves high performance with accuracy (97%), recall (95%), precision (96%), mAP (93%), F1 (94%), and a latency of 100.2s. SLCAUT-Net enhances public safety by enabling quick, real-time detection of violent incidents and triggering immediate alerts to law enforcement, significantly reducing response times. It integrates easily into existing surveillance systems, making it ideal for real-time public safety and security. However, the shallow architecture and limited feature extraction depth may hinder its ability to handle complex aggressive behaviors, suggesting the need for future research with deeper architectures or hybrid models.

Declarations

Funding

Authors did not receive any funding.

Conflicts of interests

Authors do not have any conflicts.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Nanfei Jiang is responsible for designing the framework, analyzing the performance, validating the results, and writing the article.

References

[1] V. D. Huszar, V. K. Adhikarla, I. Négyesi, and C. Krasznay, (2023) *"Toward fast and accurate violence detection for automated video surveillance applications"* **IEEE Access *11***: 18772–18793. DOI: 10.1109/ACCESS.2023.3245521.

[2] A. N. Sai and K. S. Prasad, (2023) *"Machine learning software for the detection of violence from CCTV live footage"* **Journal of Image Processing and Artificial Intelligence *9***(3): DOI: 10.46610/JOIPAI.2023.v09i03.002.

[3] F. U. M. Ullah, K. Muhammad, I. U. Haq, N. Khan, A. A. Heidari, S. W. Baik, and V. H. C. de Albuquerque, (2021) *"AI-assisted edge vision for violence detection in IoT-based industrial surveillance networks"* **IEEE Transactions on Industrial Informatics *18***(8): 5359–5370. DOI: 10.1109/TII.2021.3116377.

[4] D. Freire-Obregón, P. Barra, M. Castrillón-Santana, and M. De Marsico, (2022) *"Inflated 3D ConvNet context analysis for violence detection"* **Machine Vision and Applications *33***: 15. DOI: 10.1007/s00138-021-01264-9.

[5] A. J. Naik and M. T. Gopalakrishna, (2021) *"Deep-violence: Person violent activity detection in video"* **Multimedia Tools and Applications *80***(12): 18365–18380. DOI: 10.1007/s11042-021-10682-w.

[6] S. Vosta and K. C. Yow, (2022) *"A CNN-RNN combined structure for real-world violence detection in surveillance cameras"* **Applied Sciences *12***(3): 1021. DOI: 10.3390/app12031021.

[7] M. Asad, J. Yang, J. He, P. Shamsolmoali, and X. He, (2021) *"Multi-frame feature-fusion-based model for violence detection"* **The Visual Computer *37***(6): 1415–1431. DOI: 10.1007/s00371-020-01878-6.

[8] G. Garcia-Cobo and J. C. SanMiguel, (2023) *"Human skeletons and change detection for efficient violence detection in surveillance videos"* **Computer Vision and Image Understanding *233***: 103739. DOI: 10.1016/j.cviu.2023.103739.

[9] S. M. Mohtavipour, M. Saeidi, and A. Arabsorkhi, (2022) *"A multi-stream CNN for deep violence detection in video sequences using handcrafted features"* **The Visual Computer *38***(6): 2057–2072. DOI: 10.1007/s00371-021-02266-4.

[10] F. J. Rendón-Segador, J. A. Álvarez-García, J. L. Salazar-González, and T. Tommasi, (2023) *"CrimeNet: Neural structured learning using vision transformer for violence detection"* **Neural Networks *161***: 318–329. DOI: 10.1016/j.neunet.2023.01.048.

[11] A. Alshalawi, W. Abdul, and G. Muhammad, (2025) *"Advanced detection of violence from video: Performance evaluation of transformer and state-of-the-art convolution neural network transformer"* **IEEE Access**: DOI: 10.1109/ACCESS.2025.3564435.

[12] B. Wan, W. Jiang, Y. Fang, Z. Luo, and G. Ding, (2021) *"Anomaly detection in video sequences: A benchmark and computational model"* **IET Image Processing *15***(14): 3454–3465. DOI: 10.1049/ipr2.12258.

[13] K. Yousaf and T. Nawaz, (2022) *"A deep learning-based approach for inappropriate content detection and classification of YouTube videos"* **IEEE Access *10***: 16283–16298. DOI: 10.1109/ACCESS.2022.3147519.

[14] Y. Pu, X. Wu, L. Yang, and S. Wang, (2024) *"Learning prompt-enhanced context features for weakly supervised video anomaly detection"* **IEEE Transactions on Image Processing**: DOI: 10.1109/TIP.2024.3451935.

[15] P. Wu, J. Liu, X. He, Y. Peng, P. Wang, and Y. Zhang, (2024) *"Toward video anomaly retrieval from video anomaly detection: New benchmarks and model"* **IEEE Transactions on Image Processing *33***: 2213–2225. DOI: 10.1109/TIP.2024.3374070.

[16] M. Khan, A. El Saddik, W. Gueaieb, G. De Masi, and F. Karray, (2024) *"VD-Net: An edge vision-based surveillance system for violence detection"* **IEEE Access *12***: 43796–43808. DOI: 10.1109/ACCESS.2024.3380192.

[17] A. Mehmood, (2021) *"Efficient anomaly detection in crowd videos using pre-trained 2D convolutional neural networks"* **IEEE Access *9***: 138283–138295. DOI: 10.1109/ACCESS.2021.3118009.

[18] R. Sharma and A. Sungheetha, (2021) *"An efficient dimension reduction-based fusion of CNN and SVM model for detection of abnormal incidents in video surveillance"* **Journal of Soft Computing Paradigm *3***(2): 55–69. DOI: 10.36548/jscp.2021.2.001.

[19] S. Habib, A. Hussain, W. Albattah, M. Islam, S. Khan, R. U. Khan, and K. Khan, (2021) *"Abnormal activity recognition from surveillance videos using convolutional neural network"* **Sensors *21***(24): 8291. DOI: 10.3390/s21248291.

[20] J. C. Vieira, A. Sartori, S. F. Stefenon, F. L. Perez, G. S. De Jesus, and V. R. Q. Leithardt, (2022) *"Low-cost CNN for automatic violence recognition on an embedded system"* **IEEE Access *10***: 25190–25202. DOI: 10.1109/ACCESS.2022.3155123.

[21] B. Omarov, S. Narynov, Z. Zhumanov, A. Gumar, and M. Khassanova, (2022) *"A skeleton-based approach for campus violence detection"* **Computers, Materials & Continua *72***(1): 315–331. DOI: 10.32604/cmc.2022.024566.

[22] M. T. Bhatti, M. G. Khan, M. Aslam, and M. J. Fiaz, (2021) *"Weapon detection in real-time CCTV videos using deep learning"* **IEEE Access *9***: 34366–34382. DOI: 10.1109/ACCESS.2021.3059170.