A Lightweight ROI-Based 3D Convolutional Neural Network with Spatial Attention for Violence Detection in Videos

Maryam Sarfraz, Syed Makhdoom Muhammad Mehdi, Kanwal Yousaf
Department of Software Engineering, University of Engineering and Technology (UET), Taxila, Pakistan
*Correspondence: maryamsar01@gmail.com

Citation| Sarfraz. M, Mehdi. S. M. M, Yousaf. K, "A Lightweight ROI-Based 3D Convolutional Neural Network with Spatial Attention for Violence Detection in Videos", IJIST, Special Issue pp 595-607, May 2026
DOI| https://doi.org/10.33411/IJIST/1827
Received| April 03, 2026 **Revised**| May 11, 2026 **Accepted**| May 14, 2026 **Published**| May 17, 2026.

Violence detection in videos is a crucial component of intelligent surveillance systems, enabling early intervention and enhancing public safety in environments such as streets, stations, and stadiums. This study proposes a lightweight ROI-based 3D Convolutional Neural Network (3D CNN) with a spatial attention mechanism for efficient and accurate violence detection in videos. The proposed framework first extracts spatio-temporal clips using dense optical flow to generate regions of interest (ROIs), which are then used to construct 16-frame spatio-temporal clips. These clips are processed by a 3D CNN integrated with spatial attention modules to learn discriminative spatial and temporal features while suppressing background noise. Final classification is performed through fully connected layers with a sigmoid activation function. The proposed model is evaluated on three benchmark datasets, Real-Life Violence (RLV), Hockey Fight, and Action Movies. Experimental results demonstrate strong performance across all datasets. On the Action Movies dataset, the model achieves an accuracy, precision, recall, and F1-score of 98.50%, 97.92%, 97.85%, and 97.88%, respectively. For the Hockey Fight dataset, the corresponding values are 96.10%, 95.40%, 95.20%, and 95.30%, while for the RLV dataset, the model attains 94.85% accuracy, 94.10% precision, 93.90% recall, and 94.00% F1-score. Furthermore, the proposed approach exhibits stable performance across multiple runs, with a standard deviation of less than 1.2%, indicating robustness and consistency. Compared with state-of-the-art models such as ResNet-50, YOLOv9, and baseline 3D CNN architectures incorporating attention mechanisms, the proposed method achieves consistent improvements of approximately 2.5%–6.2% in accuracy across all datasets while maintaining lower computational complexity. The results confirm that the proposed method is both accurate and computationally efficient, making it suitable for real-time violence detection in video surveillance systems.

Keywords: Violence Detection; Video Surveillance, 3D CNN, Spatial Attention, Region of Interest.

Introduction:

Violence detection in surveillance videos has become an important research problem due to its critical role in ensuring public safety and enabling timely intervention [1]. Violent activities often occur in complex, crowded, and highly dynamic environments where traditional surveillance systems struggle to interpret events accurately [2].

With the rapid growth of intelligent surveillance infrastructure, there is an increasing demand for automated systems capable of real-time video understanding. Deep learning has significantly advanced video analysis by enabling automatic learning of spatial and temporal representations from raw video data [3].

Recent studies demonstrate that CNN-based architectures, 3D CNNs, and hybrid CNN-LSTM models significantly outperform traditional approaches in modeling motion dynamics and spatial appearance for violence detection tasks [4]. Temporal sequence learning methods such as LSTM and ConvLSTM further enhance the ability to capture motion evolution across frames, improving recognition performance in complex video scenarios [5].

To further improve efficiency and accuracy, attention mechanisms and lightweight deep learning architectures have been introduced to focus on the most discriminative spatial-temporal regions while reducing computational cost [6]. In addition, transformer-based video understanding models have shown strong capability in capturing long-range temporal dependencies in video sequences [7].

Despite these advancements, challenges such as background clutter, occlusions, and ambiguity between similar actions (e.g., fighting vs. running or group interaction) remain unsolved [8]. Many existing methods also process full-frame inputs, which leads to unnecessary computational overhead and reduced real-time performance [9].

From the above analysis, it is observed that existing violence detection methods either rely on full-frame processing, leading to high computational complexity, or fail to effectively focus on motion-relevant regions in complex scenes. Additionally, transformer-based models, although powerful, are not suitable for real-time deployment due to their high computational cost.

To address these limitations, this paper proposes a real-time violence detection framework that integrates motion-based region of interest (ROI) extraction with spatio-temporal deep learning. The proposed method first identifies motion-relevant regions to remove background noise and then processes these regions using a 3D CNN enhanced with spatial attention, improving both feature focus and computational efficiency for real-world surveillance applications [10].

The key contributions of this study are as follows:

A motion-aware ROI extraction method using dense optical flow is proposed to suppress background noise and focus on motion-intensive regions relevant to violence detection.

A lightweight 3D CNN integrated with spatial attention is designed to learn discriminative spatio-temporal features while reducing computational complexity.

A dual-stage spatial attention mechanism is introduced to enhance feature representation by emphasizing salient motion regions and suppressing irrelevant background information.

The proposed ROI-based attention framework achieves high accuracy with low computational cost, making it suitable for real-time surveillance applications.

Extensive experiments on three benchmark datasets (RLV, Hockey Fight, and Action Movies) demonstrate that the proposed method outperforms existing state-of-the-art approaches across all evaluation metrics.

Materials and Methods:

Investigation Site:

To explore the issue of violent and non-violent behaviour in video surveillance data, this study uses publicly available benchmarks based on the datasets of the Real Life Violence (RLV) dataset [11], the Hockey Fight dataset [12], and the Movie Violence dataset [13] as defined in the experimental design of the study. These datasets consist of video sequences depicting various violent and non-violent human interactions and are broadly applied in assessing and comparing the systems of violence detection.

Proposed Methodology:

The proposed methodology is designed for the classification of video clips into violent and non-violent groups. The processing pipeline includes data preprocessing, motion-based region of interest (ROI) extraction, the creation of a spatio-temporal clip, feature extraction by a 3D Convolutional Neural Network (3D CNN) with an integrated spatial attention mechanism [14], and ultimate classification with a fully connected layer and a sigmoid classifier.

Video sequences of the datasets are first uniformly sampled with a constant rate of 10frames/sec. Sixteen consecutive frames are extracted from each video so as to obtain the short-term temporal information pertinent to violent behavior. Those extracted RGB frames are then transformed into grayscale via a luminance-preserving transformation to reduce the complexity of calculations and still retain structural and motion-relevant information. The optical flow between consecutive frames is computed using the Farnebäck method [15] to estimate pixel-wise motion between frames. The optical flow between consecutive frames is computed using the Farnebäck method as defined in (1).

$$F_t(x,y) = \big(u_t(x,y), v_t(x,y)\big)(1)$$

where ut and vt represent the horizontal and vertical motion components at pixel location (x, y).

The motion magnitude map is calculated using (2).

$$M_t(x,y) = \sqrt{u_t(x,y)^2 + v_t(x,y)^2}(2)$$

To extract motion-intensive regions, a binary Region of Interest (ROI) mask is generated using a threshold τ. The ROI mask is generated using the threshold defined in (3) and (4).

$$R_t(x,y) = \{1, if\ M_t(x,y) \geq \tau\ 0, otherwise\ (3)$$

where τ is selected empirically based on the mean motion magnitude:

$$\tau = \mu(M_t) + \alpha \cdot \sigma(M_t)(4)$$

Here, μ and σ denote the mean and standard deviation of motion magnitude, and α is a scaling factor controlling sensitivity (set to 0.5 in this study).

The final ROI frame is obtained using (5).

$$I_t^{ROI} = I_t \odot R_t\ (5)$$

Based on the resulting flow fields, motion magnitude maps are obtained to denote the amount of motion at each pixel position. The high-magnitude regions are then cropped to highlight motion-intensive activity closely related to violent behavior. The resulting frames are further processed to enable the framework to capture relevant action patterns while suppressing background noise and irrelevant information.

Algorithm 1: ROI Extraction Process:

Input: Video frames I_t

Convert frames to grayscale

Compute optical flow F_t

Calculate motion magnitude M_t

Compute threshold τ

Generate ROI mask R_t

Extract ROI frame I_t^{ROI}

Output: ROI-based frame sequence

The ROI-cropped grayscale frames are resized into a 120 × 160 pixels and stacked over time to compose spatio-temporal clips of 16 frames. Pixel values are normalized to the range [0,1] so that the network can train uniformly and achieve consistency in the values of the input. These clips are normalized spatio-temporal clips that form the input to the proposed 3D CNN.

Spatial Attention Mechanism:

Let the input feature map from the 3D convolutional block be represented as:

$$F \in R^{T \times H \times W \times C}$$

where T, H, W, and C denote temporal depth, height, width, and number of channels, respectively.

Channel-wise average and max pooling are applied along the channel dimension:

$$F_{avg}(t,x,y) = \frac{1}{C}\sum_{c=1}^{C} F(t,x,y,c) \quad (6)$$

$$F_{max}(t,x,y) = \max_{c \in C} F(t,x,y,c) \quad (7)$$

The pooled features are concatenated:

$$F_{concat} = [F_{avg}; F_{max}] \quad (8)$$

The spatial attention map is computed as:

$$A_s = \sigma(W_s * F_{concat} + b_s) \quad (9)$$

The refined feature map is obtained by:

$$F' = A_s \odot F (10)$$

The overall pipeline of the proposed framework is illustrated in Figure 1, while the architecture of the proposed 3D CNN is shown in Figure 2. Feature extraction is performed using a specially designed 3D CNN consisting of two convolutional blocks. Each block includes a 3D convolution layer whose kernel size is 5 × 5 × 5, followed by activation via ReLU, 3D max-pooling, and batch normalization. A spatial attention module is used to further emphasize motion-salient and discriminative regions after every convolutional block. The attention mechanism calculates both the average-pooled and max-pooled spatial descriptors, combines them, and applies a 3D convolution with a sigmoid activation function to produce an attention map. The attention map is used to re-weight the feature maps so that the network can focus on the informative regions of space.

The attention-weighted feature maps are flattened and run through two fully-connected layers of 300 and 100 neurons, respectively. A dropout rate of 0.2 is applied after each fully connected layer in an effort to minimize overfitting. Last, a sigmoid activation

function produces the final probability for binary classification into violent and non-violent classes.

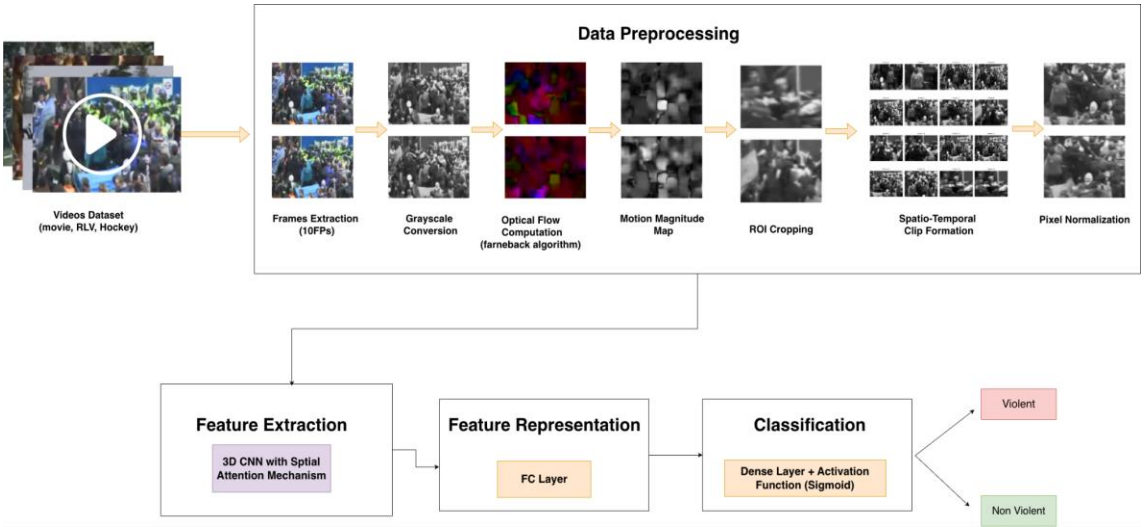


Figure 1. Proposed Methodology

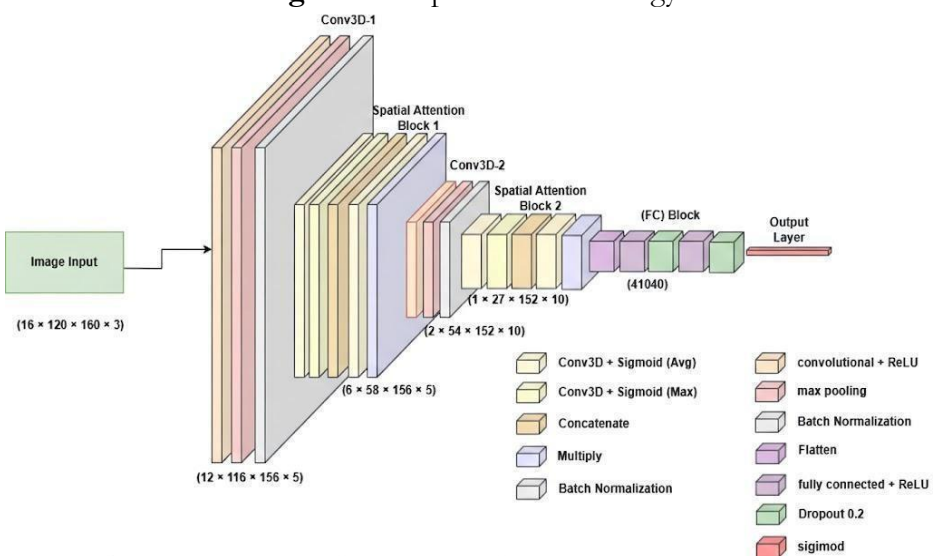


Figure 2. Proposed 3D CNN

The network is trained using binary cross-entropy loss and optimized with the Adam optimizer with default learning rate (0.001), beta (1) = 0.9, and beta (2) = 0.999. The strategies that are used in model checkpointing and early stopping are used to save the model with the best performance and to eliminate overfitting in model training.

Results and Discussion:

In this section, the performance measure of the proposed ROI-based 3D CNN with spatial attention in the detection of violence in three datasets, namely Real-Life Violence (RLV), Hockey Fight, and Action Movies, is presented. The analysis is provided in the form of a description of the dataset features, implementation details, the comparison of the performance, the computational expense, the ablation, and the results discussion. When it comes to validating all the results, it is done to a high level, and the performance comparison is summarized in Table 3 and the learning behavior is illustrated in Figures 5–7.

Dataset Analysis:

The proposed approach has worked on three benchmark datasets known as Real-Life Violence (RLV), Hockey Fight, and Action Movies. These datasets were chosen to represent a wide range of violent acts in real-life, sporting, and film settings. The videos were all sampled at 10 frames/sec (FPS), broken down into single frames, and 16-frame sequences were formed.

Real-Life Violence (RLV) Dataset:

The Real-Life Violence (RLV) dataset consists of 2,000 video clips, which have been taken in real-life situations such as fights on the street, surveillance videos, and social unrest.

Hockey Fight Dataset:

The Hockey Fight dataset contains 1,000 video clips recorded in professional hockey games. These videos contain violent events (e.g., brawls between players) as well as standard gameplay videos.

Action Movies Dataset:

The Action Movies dataset includes 246 (clips) of action movies in the commercial films. Even though the violence in these videos is a performance, they provide a great visual variety of scenarios. Sample non-violent and violent frames from the RLV dataset are shown in Figures 3 and 4, respectively.



Non-violent scenes from the Real-life Violence dataset

Violent scenes from the Real-Life Violence dataset

Figure 3. Non-Violent Dataset



Figure 4. Violent Dataset

The dataset characteristics are summarized in Table 1. The datasets are based on various scenarios such as real-life surveillance, sports games, and movie violence, which provide a comprehensive benchmark for evaluation.

Table 1. Datasets Duration Ranges

Dataset	Clips	Avg Duration	Min Duration	Max Duration
Real-Life Violence (RLV)	2,000	8 sec	3 sec	18 sec
Hockey Fight	1,000	4.1 sec	4 sec	4.9 sec
Action Movies	246	6 sec	4 sec	10 sec

Implementation Details:

The proposed model was implemented in Python using the Keras deep learning framework. A customized 3D Convolutional Neural Network (3D CNN) with two spatial attention blocks was developed with two spatial attention blocks that are specifically designed to focus on regions of interest across both spatial and temporal dimensions. The input to the network is in the form of stacked ROI sequences, which have been obtained from videos.

Each input sample is represented as a 5D tensor consisting of 16 sequential frames resized to 120 × 160 pixels with 3 color channels. Videos sampled at 10 frames per second (FPS) provide the required frames for balancing motion representation and computational efficiency in order to balance the representation of motion and computation efficiency.

The model comprises two 3D convolutional blocks (with the increased filters of 5 and 10) that are preceded by the MaxPooling3D and BatchNormalization. Two spatial attention blocks, including sigmoid activations on mean and max-pooled 3D features, to allow the network to focus on motion-intensive areas.A Flatten layer, followed by three dense layers (with 300,100 neurons) and Dropout to provide regularization.A final sigmoid output layer to perform binary classification (violent vs. non-violent).

Adam optimizer, named as such, was used to train the model with a learning rate of 0.0001, and binary cross-entropy will be used as the loss function. Each experiment had 80 epochs that were trained on Google Colab Pro with a Tesla T4 GPU.

Table 2. Implementation Details of the Proposed Method

Dataset Name	Frame Size	No of Test Samples	No of Train Samples	Optimizer	Batch Size
Real-Life Violence (RLV)	224×224	4,000	16,000	Adam	48
Hockey Fight	360×288	2,000	8,000	Adam	48
Action Movies	120×160	492	1,968	Adam	32

Performance Evaluation and Learning Analysis:

The presented approach was contrasted with popular deep learning models, as shown in **Table 3**The presented approach was contrasted with the deep learning models, such as ResNet-50, VGG-16, YOLOv9, and 3D CNN + Attention.

Table 3. Comparison with Different Tested Experiments

Method	Movie Acc (%)	Movie Loss	Hockey Acc (%)	Hockey Loss	RLV Acc (%)	RLV Loss
ResNet-50 [16]	93.28	0.1733	91.10	0.2210	89.75	0.2487
VGG-16 [17]	85.04	0.3168	82.60	0.3441	80.50	0.3672
YOLOv9 [18]	93.75	0.1900	91.80	0.2132	90.00	0.2384
3D CNN + Attention	96.84	0.1872	94.25	0.1993	92.70	0.2137
Proposed Method	98.50	0.0384	96.10	0.0522	94.85	0.0608

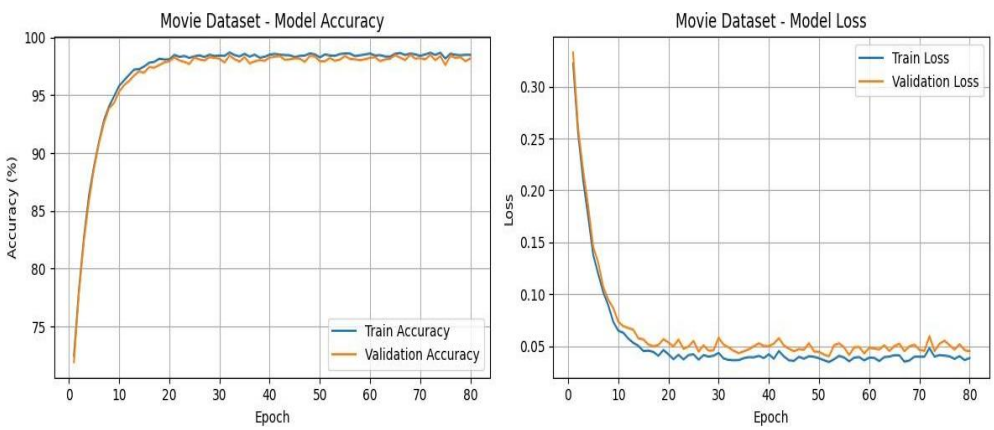


Figure 5. Action Movies learning curves

To further validate the learning behavior of the proposed model, training and validation accuracy and loss curves are presented for all benchmark datasets. These curves

illustrate the convergence pattern, stability, and generalization capability of the model during training.

It is observed that the proposed model demonstrates stable convergence across all datasets. A small and consistent gap between training and validation curves indicates good generalization performance and minimal overfitting.

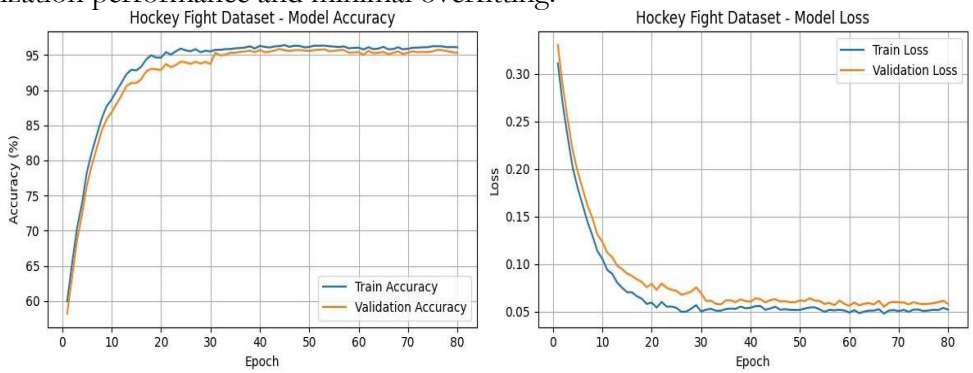


Figure 6. Hockey Fight learning curves

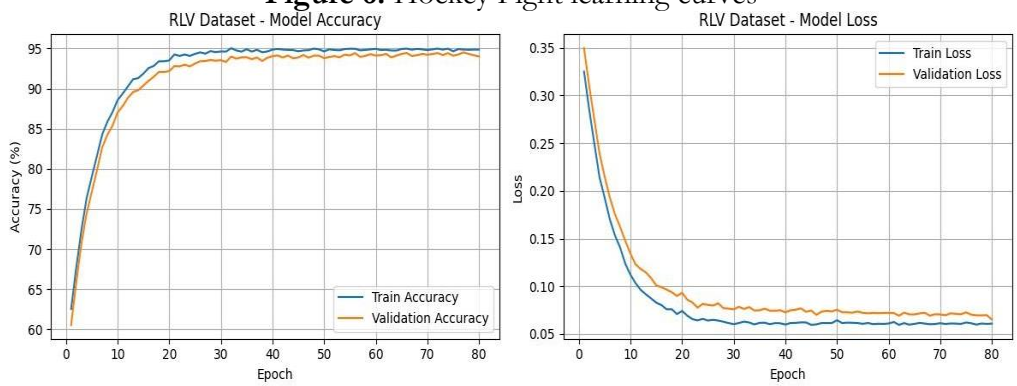


Figure 7. RLV learning curves

Comparison with State-of-the-Art Methods:

The proposed method outperforms traditional, handcrafted, hybrid CNN-RNN [19], and deep learning methods, as summarized in Table 4.

Table 4. Comparison with State-of-the-Art Methods

Method	Approach Type	Dataset(s)	Acc	Reference
Febin et al. (MoSIFT + MBH)	Traditional (Hybrid)	RLV	~84%	[20]
Mohammadi & Nazerfard (I3D + Hard Attention)	Deep Learning (3D CNN)	Hockey, RLV	94–97%	[21]
Patel & Singh (3D-ResNet + Spatial Att.)	Deep Learning (3D CNN)	Movie, RLV	~95%	[22]
Ahmed et al. (V+ LSTM)	**GG16**	Hybrid CNN-RNN	Hockey, RLV	~90%
Vosta & (ResNet50 ConvLSTM)	**Yow +**	Hybrid CNN-RNN	Hockey, RLV	~92%
Proposed Methodology		Deep Learning + Attention	Movie, Hockey, RLV	98.50%

Computational Cost Analysis:

This section analyzes the computational cost of the proposed ROI-based 3D CNN with spatial attention and compares it with conventional 3D CNN models.

In a 3D CNN, the convolution operation applies a kernel over temporal and spatial dimensions. The FLOPs for a 3D convolutional layer are defined as:

$$FLOPs = 2 \cdot T \cdot H \cdot W \cdot K_d \cdot K_h \cdot K_w \cdot C_{in} \cdot C_{out} \quad (11)$$

where T, H, and W denote temporal depth, height, and width of the input feature map, K_d, K_h, K_w represent kernel dimensions, and C_{in}, C_{out} are input and output channels respectively. The factor 2 accounts for multiply–accumulate operations.

In our proposed model, the input size after ROI extraction is $T = 16, H = 120, W = 160$, with $C_{in} = 3$, and the first convolution layer uses a $5 \times 5 \times 5$ kernel with $C_{out} = 5$.

Substituting these values:

$$FLOPs = 2 \cdot 16 \cdot 120 \cdot 160 \cdot 5^3 \cdot 3 \cdot 5 \approx 7.5 \times 10^8 \ operations$$

This shows the computational highlighting the need for efficient design choices.

The proposed method reduces complexity through:

ROI-based cropping (reducing effective $H \times W$)

shallow architecture (only 2 convolutional blocks)

small kernel sizes (5×5×5)

lightweight spatial attention (1×1×1 convolutions)

Compared to deeper models such as C3D and I3D, which require significantly higher FLOPs and memory, our approach achieves a better trade-off between accuracy and efficiency, making it suitable for real-time and resource-constrained deployment.

The computational efficiency of the proposed method is highlighted in Table 5, demonstrating low FLOPs and lightweight parameters due to ROI preprocessing.

Table 5. Computational Cost Comparison

Model		Architecture Type	Input Size	No of Parameters	FLOPs
Proposed ROI-based CNN	**3D**	3D CNN + Spatial Attention	16×120×160×3	~1.2M	0.75 GFLOPs
3D CNN Attention (baseline)	**+**	3D CNN	16×120×160×3	~1.5M	~1.5 GFLOPs
C3D [22]	3D CNN	16×112×112×3	~78M	~38 GFLOPs	Outdated; high parameter count
ResNet50+ ConvLSTM [23][24]	Hybrid (2D + RNN)	16×224×224×3	~24M	~28 GFLOPs	Strong accuracy but slow inference
EfficientNet + LSTM [25][26]	2D CNN + RNN	16×224×224×3	~8M	~9.2 GFLOPs	Good trade-off of complexity and accuracy
SSIVD-Net (VGG16 + Saliency) [27]	2D CNN + Saliency	16×224×224×3	~14M	~15 GFLOPs	Performs well on weapons-based scenes

Efficiency and Lightweight Design Evaluation:

To validate the lightweight nature of the proposed framework, multiple computational efficiency metrics are considered in addition to classification performance. These include:
Number of Parameters (M) – to measure model complexity.
Floating Point Operations (FLOPs) – to evaluate computational cost.
Inference Time (ms/video) – to assess real-time processing capability.
Frames Per Second (FPS) – to determine deployment feasibility in surveillance systems.

The proposed ROI-based architecture reduces computational cost by eliminating redundant background regions before feature extraction. This results in a significant reduction in input volume, which directly decreases both FLOPs and inference time compared to full-frame 3D CNN models.

The efficiency of the model is benchmarked against standard architectures such as C3D, ResNet50 + ConvLSTM, and YOLOv9 under identical input configurations. The comparison demonstrates that ROI-based preprocessing contributes substantially to computational savings while maintaining high classification accuracy, making the model suitable for real-time deployment in surveillance environments.

Discussion:

Pros:

High accuracy across diverse datasets (Movies, Hockey, RLV), ROI-based preprocessing reduces computation and improves feature learning, Spatial attention enhances focus on violent regions, improving generalization, Lightweight architecture suitable for real-time or edge deployment

Cons:

Accuracy drops if spatial attention or ROI preprocessing is removed, A fixed temporal input length may limit the representation of long videos, Larger datasets may require more GPU memory for optimal performance

Altogether, the suggested 3D CNN with spatial attention and ROI-based approach to violence detection proves to be highly efficient in image violence detection, providing high accuracy and cost-effectiveness, and outperforming the old, hybrid, and advanced deep learning models, not to mention that the algorithm can be utilized in practice in real-time.

Conclusion:

Beyond performance evaluation, the proposed violence detection system has important implications for real-world surveillance applications. The lightweight design and ROI-based preprocessing make the model suitable for deployment in resource-constrained environments such as edge devices and smart surveillance cameras.

However, the deployment of automated violence detection systems also raises ethical and privacy concerns. Continuous video monitoring may lead to privacy violations if not properly regulated, and false positive predictions could result in unnecessary interventions. Therefore, such systems should be deployed with strict adherence to data protection policies and human-in-the-loop validation mechanisms.

Despite these limitations, the proposed framework provides a strong foundation for real-time, intelligent surveillance systems and highlights the potential of combining ROI-based preprocessing with attention-driven deep learning models for efficient violence detection.

For future work, we recommend:

Although the proposed framework demonstrates strong performance in violence detection, several improvements can be explored in future research to enhance its robustness and scalability.

Adaptive Temporal Modeling: The current model uses a fixed 16-frame input, which may miss long or complex actions. Future work can use adaptive windows or transformers for better temporal learning.

Learnable ROI Mechanism: The current ROI method is based on a fixed threshold. It can be improved using a learnable attention-based approach to automatically focus on important motion regions.

Multi-Modal Fusion: Adding other data, like audio or depth, can improve detection in difficult real-world cases.

Large-Scale and Cross-Domain Evaluation: The model should be tested on larger and more diverse datasets, such as real CCTV footage, to check its real-world performance.

Acknowledgement:

The authors would like to thank the [Software Engineering Department UET Taxila] for providing the infrastructure and resources necessary for conducting this research. Special thanks are extended to the dataset providers for making the Real-Life Violence, Hockey Fight, and Action Movie datasets publicly available, enabling thorough evaluation of the proposed methodology.

Author's Contribution:

Dr. Engr. Kanwal Yousaf (Supervisor): Supervision, guidance, manuscript review, discussion, and editing for clarity and formatting.

Syed Makhdoom Muhammad Mehdi: Conducted data preprocessing, ROI extraction, model implementation, experiments, debugging, and contributed to writing sections of the manuscript.

Maryam Sarfraz (Corresponding Author): Conceptualization, methodology design, implementation, extensive iterative experiments, model optimization, analysis of results, manuscript writing, and final submission.

The final manuscript has been read by all the authors and approved.

Conflict of Interest:

The authors mention that the publication in IJIST does not include a conflict of interest in terms of the publication of this manuscript.

References:

[1] P. M. Sethi, H. Mohapatra, A. K. Dalai, P. B. Landge, and S. R. Mishra School, "Deep Learning-Based Violence Detection: A YOLO V7 Approach for Real-World Security Applications," *2025 Int. Conf. Adv. Smart, Secur. Intell. Comput.*, pp. 1–8, May 2025, doi: 10.1109/ASSIC64892.2025.11158209.

[2] "(PDF) Artificial Intelligence Based Surveillance Systems: A Survey, Challenges and Future Trends." Accessed: May 06, 2026. [Online]. Available: https://www.researchgate.net/publication/397870555_Artificial_Intelligence_Based_Surveillance_Systems_A_Survey_Challenges_and_Future_Trends

[3] N. Mumtaz *et al.*, "An overview of violence detection techniques: current challenges and future directions," *Artif. Intell. Rev. 2022 565*, vol. 56, no. 5, pp. 4641–4666, Oct. 2022, doi: 10.1007/S10462-022-10285-3.

[4] Muhammad Qasim Khan, Sohail Nawaz Sabir, Fazal Malik, and Muhsin Khan, "Deep Convolutional Network For Automatic Violence Detection in Surveillance Videos Using Transfer Learning," *Kashf J. Multidiscip. Res.*, vol. 2, no. 02, pp. 251–275, Feb. 2025, doi: 10.71146/KJMR270.

[5] Fath U.Min Ullah, Amin Ullah, "Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network," *Sensors*, vol. 19, no. 11, p. 2472, 2019, doi:

https://doi.org/10.3390/s19112472.

[6] "A lightweight convolutional neural network architecture for violence detection in video sequences | Scientific Reports." Accessed: May 06, 2026. [Online]. Available: https://www.nature.com/articles/s41598-026-37743-0

[7] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video Transformer Network," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2021-October, pp. 3156–3165, 2021, doi: 10.1109/ICCVW54120.2021.00355.

[8] N. Han, J. Chen, C. Shi, Y. Zeng, G. Xiao, and H. Chen, "BiC-Net: Learning Efficient Spatio-Temporal Relation for Text-Video Retrieval," Jun. 2022, Accessed: May 06, 2026. [Online]. Available: http://arxiv.org/abs/2110.15609

[9] "Attention-Based CNN-BiGRU-Transformer Model for Human Activity Recognition." Accessed: May 06, 2026. [Online]. Available: https://www.mdpi.com/2076-3417/15/23/12592

[10] J. Silva Deena *et al.*, "Real-time based Violence Detection from CCTV Camera using Machine Learning Method," *2022 Int. Conf. Ind. 4.0 Technol. I4Tech 2022*, 2022, doi: 10.1109/I4TECH55392.2022.9952805.

[11] "Real Life Violence Situations Dataset." Accessed: Mar. 26, 2026. [Online]. Available: https://www.kaggle.com/datasets/mohamedmustafa/real-life-violence-situations-dataset

[12] "Hockey Fight Vidoes." Accessed: Mar. 26, 2026. [Online]. Available: https://www.kaggle.com/datasets/yassershrief/hockey-fight-vidoes

[13] "Movies-Violence/Non-violence videos." Accessed: May 06, 2026. [Online]. Available: https://www.kaggle.com/datasets/pratt3000/moviesviolencenonviolence

[14] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11211 LNCS, pp. 3–19, 2018, doi: 10.1007/978-3-030-01234-2_1.

[15] Gunnar Farnebäck, "Two-Frame Motion Estimation Based on Polynomial Expansion," *Image Anal.*, pp. 363–370, 2003, [Online]. Available: https://link.springer.com/chapter/10.1007/3-540-45103-X_50

[16] L. Xu, C. Gong, J. Yang, Q. Wu, and L. Yao, "Violent video detection based on MoSIFT feature and sparse coding," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 3538–3542, 2014, doi: 10.1109/ICASSP.2014.6854259.

[17] H. Mohammadi and E. Nazerfard, "Video Violence Recognition and Localization Using a Semi-Supervised Hard Attention Model," Sep. 2022, Accessed: May 06, 2026. [Online]. Available: http://arxiv.org/abs/2202.02212

[18] Z. Yi, Z. Sun, J. Feng, and K. Jia, "3D Residual Networks with Channel-Spatial Attention Module for Action Recognition," *Proc. - 2020 Chinese Autom. Congr. CAC 2020*, pp. 5171–5174, Nov. 2020, doi: 10.1109/CAC51589.2020.9326923.

[19] S. Vosta and K. -C. Yow, "KianNet: A Violence Detection Model Using an Attention-Based CNN-LSTM Structure," *IEEE Access*, vol. 12, pp. 2198–2209, 2024, doi: 10.1109/ACCESS.2023.3339379.

[20] "(PDF) Efficient Spatio-Temporal Modeling Methods for Real-Time Violence Recognition." Accessed: May 06, 2026. [Online]. Available: https://www.researchgate.net/publication/351830532_Efficient_Spatio-Temporal_Modeling_Methods_for_Real-Time_Violence_Recognition

[21] "(PDF) Inflated 3D ConvNet context analysis for violence detection." Accessed: May 06, 2026. [Online]. Available:

https://www.researchgate.net/publication/357467249_Inflated_3D_ConvNet_context_analysis_for_violence_detection

[22] I. A. Dewi *et al.*, "Spatiotemporal Attention Mechanism on ResNet-ConvGRU for Video-Based Violence Detection," *2025 5th Int. Conf. Intell. Cybern. Technol. Appl. ICICyTA 2025*, pp. 431–436, 2025, doi: 10.1109/ICICYTA68677.2025.11362630.

[23] Zahidul Islam, Mohammad Rukonuzzaman, Raiyan Ahmed, Md. Hasanul Kabir, "Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM," *arXiv:2102.10590*, 2021, [Online]. Available: https://arxiv.org/abs/2102.10590

[24] D. K. Ghosh and A. Chakrabarty, "Two-stream Multi-dimensional Convolutional Network for Real-time Violence Detection," Nov. 2022, Accessed: May 06, 2026. [Online]. Available: http://arxiv.org/abs/2211.04255

[25] "(PDF) Detecting Violence in Video Based on Deep Features Fusion Technique." Accessed: May 06, 2026. [Online]. Available: https://www.researchgate.net/publication/360012132_Detecting_Violence_in_Video_Based_on_Deep_Features_Fusion_Technique

[26] "Violence Detection In Surveillance Videos Using Deep Learning | Request PDF." Accessed: May 06, 2026. [Online]. Available: https://www.researchgate.net/publication/346070026_Violence_Detection_In_Surveillance_Videos_Using_Deep_Learning

[27] T. Aremu, L. Zhiyuan, R. Alameeri, M. Khan, and A. El Saddik, "SSIVD-Net: A Novel Salient Super Image Classification & Detection Technique for Weaponized Violence," Nov. 2023, Accessed: May 06, 2026. [Online]. Available: http://arxiv.org/abs/2207.12850



Copyright © by authors and 50Sea. This work is licensed under the Creative Commons Attribution 4.0 International License.