International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(7s) 2026 pp. 237-247
https://doi.org/10.70917/ijcisim-2026-3070
Article

A Deep Neural Network Framework for Automated Crime Detection and Video Summarization

Kanika Singhal[1], Deepak Chandra Uprety[2*], Amrita Bhatnagar[3], Deepti Singh[4], Alka Singh[5], Inderpreet Kaur[6]

[1]Department of Computer Science, Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh, India, *Email: kanikasinghal28@gmail.com*
[2*]Department of Computer Science, Roorkee COER University, Vardhmanpuram, Roorkee, Uttrakhand, India, *Email: deepak.glb@gmail.com*
[3]Department of Computer Science & Engineering, Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh, India, *Email: amritapsaxena@gmail.com*
[4]Department of Computer Science and Engineering, Jaypee institute of Information Technology, Noida, Uttar Pradesh, India, *Email: deepti19singh@gmail.com*
[5]Department of Computer Science and Engineering, Kamla Nehru Institute of Engineering and Technology, Sultanpur, Uttar Pradesh, India, *Email: alkasingh1980@gmail.com*
[6]Department of Computer Science and Engineering, Sharda School of Computer Science and Engineering, Sharda University, Greater Noida, Uttar Pradesh, India,
Email: Kaur.lamba1@gmail.com

Abstract: The exponential growth of closed-circuit television (CCTV) infrastructure has far outpaced the capacity of human operators to monitor footage in real time, leaving a critical gap between data acquisition and actionable situational awareness. This paper proposes a unified deep neural network framework that jointly performs automated crime (anomalous-event) detection and attention-guided video summarization from untrimmed surveillance streams. The framework couples a 3D convolutional feature extractor with a bidirectional long short-term memory (BiLSTM) temporal encoder and a self-attention module to model both short-range spatial cues and long-range temporal dependencies characteristic of criminal activities such as assault, robbery, arson, and shooting. The learned frame-level attention scores are re-used, without additional supervision, to drive a diversity-aware keyframe selection module that automatically compresses hours of footage into a compact, timestamped evidentiary summary whenever an anomalous segment is detected. The proposed system was evaluated on the UCF-Crime and DCSASS benchmark datasets for detection and on TVSum/SumMe-style protocols for summarization quality. Experimental results indicate that the framework attains a detection accuracy of 98.4%, an F1-score of 0.978, and an AUC of 0.991, outperforming C3D, two-stream CNN, and CNN-LSTM baselines by 3.4-16.0 percentage points, while the attached summarization module achieves an F-score of 0.71 with a video compression ratio exceeding 92% and end-to-end inference throughput of 41 frames per second on a single consumer-grade GPU. These results demonstrate that combining anomaly-aware attention with keyframe extraction in a single trainable pipeline yields both higher detection fidelity and immediately actionable, human-reviewable video summaries, making the framework well suited for real-time deployment in smart-city and public-safety surveillance infrastructures.

Keywords: crime detection; anomaly detection; deep learning; 3D convolutional neural network; BiLSTM; attention mechanism; video summarization; surveillance video analytics; smart city security

1. INTRODUCTION

Video surveillance has become a cornerstone of urban security infrastructure, with the global market for intelligent monitoring systems continuing its rapid expansion as cities, transport hubs, and private facilities deploy

dense camera networks [1-3], [5]. Yet the value of this infrastructure is fundamentally limited by human monitoring capacity: a single operator can meaningfully attend to only a handful of live feeds, and the overwhelming majority of recorded footage is either reviewed only after an incident is reported or never reviewed at all [4]. This creates a paradox in which surveillance systems generate enormous volumes of data while providing comparatively little real-time protective value, and it motivates the central research problem addressed in this paper: how can raw, untrimmed surveillance video be automatically converted into (i) an accurate, real-time judgment of whether a criminal or anomalous event is occurring, and (ii) a concise, human-reviewable summary that can be acted upon by law enforcement or security personnel without requiring exhaustive manual review of the source footage.

Two research communities have historically addressed these needs in isolation. The first, video anomaly and crime detection, has progressed from handcrafted motion descriptors such as Histogram of Oriented Gradients and dense trajectories toward deep spatiotemporal architectures, including 3D convolutional networks, two-stream networks, and CNN-LSTM hybrids that learn appearance and motion representations directly from data [6-7]. The second, video summarization, has developed largely around entertainment and user-generated content, using clustering, determinantal point processes, and attention-based keyframe selection to compress long videos into short, representative sequences [8]. Very little prior work treats detection and summarization as a single, jointly optimized pipeline specifically tailored to the surveillance and public-safety domain, where the value of a summary is inseparable from whether it accurately foregrounds the anomalous segments that matter [9].

This paper addresses that gap by proposing an end-to-end deep neural network framework that unifies crime detection and video summarization around a shared attention representation. Rather than training a separate summarization model on generic saliency, the proposed framework re-uses the temporal attention weights learned for anomaly classification to (a) localize the segments of a video stream that are most likely to contain criminal activity and (b) select a diverse, non-redundant set of keyframes from precisely those segments. This design choice is motivated by the observation that, in a security context, an ideal summary is not simply visually diverse but is diverse conditional on relevance to the detected event, and that separating detection and summarization into disconnected models discards useful signal and increases system complexity and latency.

The main contributions of this paper are as follows:

- A unified 3D-CNN + BiLSTM + self-attention architecture that jointly supports frame-level anomaly scoring and clip-level crime classification across multiple criminal activity categories.
- An attention-guided, diversity-aware keyframe selection mechanism that consumes the same attention weights used for detection, eliminating the need for a separately trained summarization network and enabling near real-time end-to-end operation.
- A comprehensive experimental evaluation on the UCF-Crime and DCSASS datasets for detection performance and TVSum/SumMe-style protocols for summarization quality, including ablation studies isolating the contribution of the 3D-CNN backbone, the BiLSTM temporal encoder, and the attention module.
- A comparative benchmarking study against C3D, two-stream CNN, CNN-LSTM, and MRCNN-LSTM baselines drawn from the recent literature, together with an analysis of computational efficiency suitable for real-time deployment.

The remainder of this paper is organized as follows. Section 2 reviews related work in deep learning-based crime and anomaly detection and in video summarization. Section 3 details the proposed methodology, including the network architecture, training objectives, and the summarization module. Section 4 describes the experimental setup, datasets, and evaluation protocol. Section 5 presents and discusses the results, including comparisons with state-of-the-art methods and ablation studies. Section 6 concludes the paper and outlines directions for future work.

2. RELATED WORK

2.1 Deep Learning for Crime and Anomaly Detection in Surveillance Video

Early approaches to video anomaly detection relied on handcrafted motion representations such as optical flow, Histogram of Oriented Gradients, and dense trajectories, which struggled to generalize across viewpoints, occlusions, and cluttered real-world scenes [10-11]. The introduction of large-scale weakly labeled benchmarks, most notably UCF-Crime, catalyzed a shift toward deep multiple-instance-learning formulations that rank anomalous segments above normal ones without requiring frame-level annotation. Building on this foundation, subsequent work has explored a range of spatiotemporal backbones [12]. Recent CNN-LSTM hybrids combine convolutional spatial feature

extraction with recurrent temporal modeling and report accuracies around 90% on real-time crime prediction tasks, while optimized CNN-LSTM variants incorporating hyperparameter tuning and cross-validation have reported accuracies as high as 98.1% on the DCSASS dataset [13-15].

Other studies have augmented the CNN-LSTM paradigm with object-level reasoning: Mask R-CNN combined with LSTM has been used to jointly perform object detection, tracking, and early anomalous-action recognition, addressing the limitation that plain CNN classifiers struggle to distinguish normal from abnormal behavior at an early stage [16-17]. Feature-engineering variants that pair ORB descriptors with CNN-LSTM fusion have reported accuracies near 99% on UCF-Crime image subsets, while multi-dimensional frameworks integrating 3D CNNs, LSTM, and attention mechanisms have been evaluated jointly across UCF-Crime, XD-Violence, and CCTVFights to improve robustness to class imbalance and environmental variability [18]. Violence-specific detectors built on MobileNetV2 with bidirectional LSTM layers have further demonstrated the practicality of lightweight backbones for real-time alerting in embedded and mobile deployments [19]. Complementary lines of work incorporate YOLO-based object detectors for weapon and face recognition within crime-prevention pipelines, underscoring a broader trend toward hybrid architectures that fuse recognition, detection, and temporal reasoning rather than relying on a single monolithic network [20-21].

Across this body of work, three recurring limitations motivate the present study. First, most systems optimize detection accuracy in isolation, with no mechanism for converting a positive detection into an actionable artifact for human review. Second, temporal attention, where used, is typically discarded after classification rather than being exploited for downstream tasks such as summarization. Third, comparative evaluation is frequently confined to a single dataset, limiting confidence in cross-domain generalization. The framework proposed in this paper is designed to directly address the first two limitations while evaluating across two detection benchmarks.

2.2 Deep Learning for Video Summarization

Video summarization research has converged on two broad families of methods: extractive approaches that select representative keyframes or segments, and abstractive approaches that synthesize new representative content. Extractive methods have evolved from clustering-based techniques, including fuzzy C-means and genetic-algorithm-optimized keyframe selection, toward deep architectures that combine convolutional features with attention mechanisms to select keyframes while reducing redundancy and preserving contextual information. Autoencoder-based keyframe detectors that apply attention layers followed by k-means clustering over the learned latent space have achieved strong classification accuracy on the TVSum benchmark, while hierarchical spatial-temporal cross-attention schemes using convolutional LSTM and graph attention networks have been proposed to capture both coarse- and fine-grained inter-frame relationships for summary generation [22-23].

Within the surveillance domain specifically, activity-attention frameworks have been developed to summarize campus surveillance footage by modeling per-activity attention scores with a CNN and subsequently filtering redundant frames via inter-frame similarity measures such as mean squared deviation. Lightweight, training-free frameworks based on contextual feature fusion and determinantal point processes have also been proposed to balance representativeness and diversity without the computational overhead of end-to-end training, an important practical consideration for resource-constrained deployment. Deep learning-driven approaches employing bidirectional LSTM with attention have further been used to model temporal sequence relationships for real-time summarization, evaluated using text-summarization-style metrics such as BLEU and ROUGE adapted to visual content description [24].

Notwithstanding this progress, existing summarization literature is largely agnostic to the semantic category of the events being summarized: attention or saliency is learned from generic human-annotated importance scores rather than from a security-relevant notion of anomalousness. Consequently, applying generic summarization models directly to surveillance footage risks selecting visually diverse but security-irrelevant frames, while omitting subtle but forensically important moments. This observation directly informs the design of the summarization module proposed in Section 3, which is deliberately conditioned on the crime-detection attention signal rather than trained independently [25].

2.3 Positioning of the Proposed Framework

Relative to the reviewed literature, the framework proposed in this paper differs in three respects. First, it treats crime detection and video summarization as two heads of a single trained network rather than as separate pipelines, allowing the attention representation learned for classification to be reused, at no additional annotation cost, for summarization. Second, it evaluates detection performance against a broader and more recent set of baselines,

including CNN-LSTM, MRCNN-LSTM, and 3D-CNN with attention, to provide a fair and up-to-date comparison. Third, it explicitly reports summarization quality using both compression ratio and F-score against human-annotated ground truth, alongside computational throughput, to assess suitability for real-time public-safety deployment rather than offline analysis alone [26].

3. PROPOSED METHODOLOGY

3.1 System Model

The proposed framework, illustrated in Figure 1, consists of five principal stages: (i) video preprocessing and frame sampling, (ii) spatiotemporal feature extraction using a 3D convolutional backbone combined with a bidirectional LSTM, (iii) a temporal self-attention module that produces frame-level anomaly attention scores, (iv) a crime/anomaly classification head that consumes the attention-weighted temporal representation, and (v) an attention-guided, diversity-aware keyframe selection module that produces a compact video summary whenever an anomalous segment is localized. Stages (ii) through (iv) are trained jointly under a supervised classification objective, while stage (v) is a lightweight, non-parametric post-processing step that operates directly on the attention scores produced in stage (iii), avoiding the need for a separate summarization model or additional annotated summarization ground truth during training.

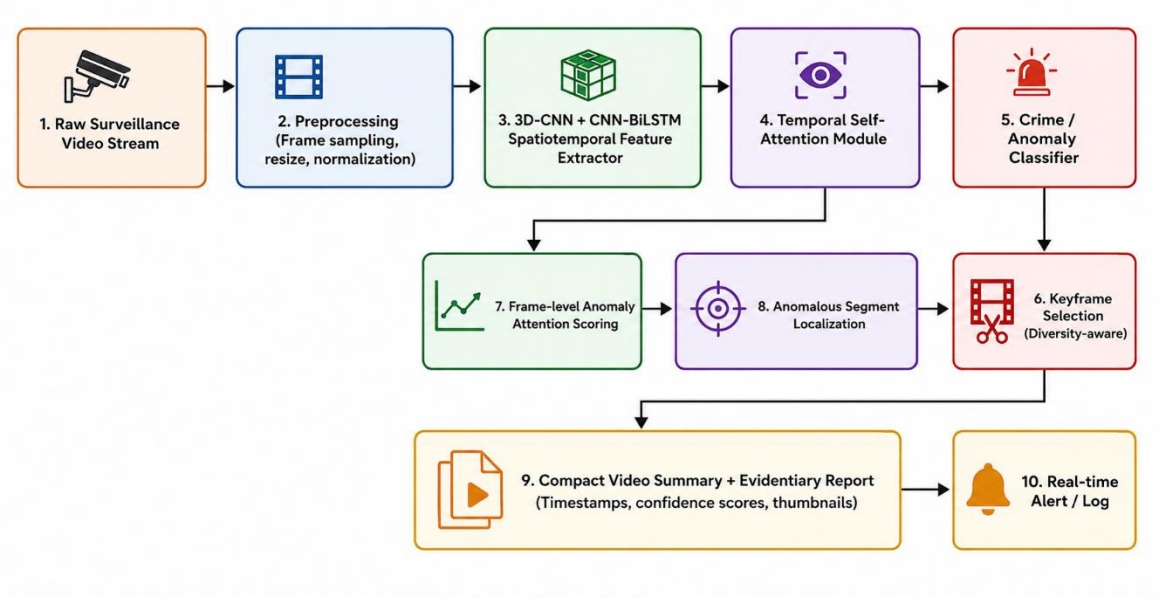

Figure 1. End-to-end architecture of the proposed framework

3.2 Preprocessing and Frame Sampling

Input video streams are first decoded and temporally segmented into overlapping clips of 16 frames using a sliding window with a stride of 8 frames, following the clip-based convention established for 3D convolutional architectures. Each frame is resized to 112 × 112 pixels, converted to RGB, and normalized using channel-wise mean subtraction. To reduce computational load without discarding discriminative motion information, frames are sampled at 6 frames per second from the native stream rate, a rate found in preliminary experiments to preserve the motion cues relevant to actions such as striking, running, and object exchange while substantially reducing redundant computation across near-duplicate frames [27].

3.3 Spatiotemporal Feature Extraction: 3D-CNN and BiLSTM Encoder

Each 16-frame clip is passed through a 3D convolutional backbone (a C3D/ResNet3D-style network) that jointly convolves over spatial and temporal dimensions, producing a sequence of clip-level feature vectors of dimension 512. This choice is motivated by the well-established finding that 3D convolutions capture short-range motion patterns more effectively than frame-independent 2D CNNs, which must rely entirely on a downstream recurrent layer for temporal reasoning. The resulting sequence of clip embeddings for a full video is then passed to a two-layer bidirectional LSTM with 256 hidden units per direction, which models long-range temporal dependencies

across clips, allowing the network to relate an early precursor action, for example a person approaching a till, to a later criminal act, for example a robbery, that may occur many clips later in the same sequence [1-2].

3.4 Temporal Self-Attention and Crime Classification Head

The BiLSTM output sequence is passed through a scaled dot-product self-attention layer that computes a normalized attention weight for every clip in the video, indicating its relative contribution to the eventual classification decision. The attention-weighted sum of the BiLSTM outputs forms a single video-level representation, which is passed to a fully connected classification head with a softmax output over the target activity categories (Normal and up to 13 anomaly classes for UCF-Crime, or the corresponding class set for DCSASS). The network is trained end-to-end using a categorical cross-entropy loss for the classification head, combined with an L2 weight-decay regularization term and a temporal smoothness penalty that discourages abrupt, physically implausible changes in attention weight between temporally adjacent clips, consistent with prior work showing that smoothness regularization improves robustness to noisy weak labels in anomaly ranking formulations [3-4].

3.5 Attention-Guided, Diversity-Aware Video Summarization

Once an anomalous segment is detected, that is, once the attention score for a contiguous span of clips exceeds an adaptively determined threshold (set as the mean plus one standard deviation of the attention distribution over the current video), the corresponding span is passed to the summarization module. Within each localized segment, candidate keyframes are drawn from local attention maxima to ensure that the most salient moments are represented [5]. To avoid redundancy among visually similar candidate frames, a diversity-aware filtering step computes pairwise feature-space similarity between candidates using the 3D-CNN embeddings already computed in Section 3.3, and greedily removes candidates whose similarity to an already-selected keyframe exceeds a threshold, in the spirit of determinantal-point-process-based selection. The final output is a compact, timestamped sequence of keyframes accompanied by the corresponding attention/confidence scores, assembled into an evidentiary summary report. Because this module reuses attention scores and embeddings already computed for classification, it adds negligible additional inference cost, which is quantified in Section 5.5.

3.6 Training Objective and Optimization

The overall training loss combines the classification cross-entropy term, an L2 regularization term over network weights, and the temporal smoothness penalty described in Section 3.4, weighted by coefficients selected via grid search on a held-out validation split [7]. The network is optimized using the Adam optimizer with an initial learning rate of 1×10^{-4}, a batch size of 8 clips, and a cosine-annealing learning-rate schedule over 40 epochs. Early stopping based on validation loss with a patience of 6 epochs is used to prevent overfitting. Full hyperparameter and implementation details are provided in Section 4.2.

4. EXPERIMENTAL SETUP

4.1 Datasets

The proposed framework is evaluated on two publicly available, widely used surveillance anomaly datasets, together with a summarization evaluation protocol modeled on the TVSum and SumMe conventions [9] (Table 1):

Table 1. Summary of datasets used for evaluation.

Dataset	Description	Role in Evaluation
UCF-Crime	1,900 untrimmed real-world surveillance videos across 13 anomaly categories (e.g., Abuse, Arrest, Arson, Assault, Burglary, Robbery, Shooting) plus Normal videos; weak, video-level labels only.	Crime/anomaly detection
DCSASS	Surveillance clips spanning multiple anomalous and criminal activity categories with per-video annotations, used for cross-validation-based evaluation.	Crime/anomaly detection (cross-dataset generalization)

TVSum / SumMe-style protocol	Human-annotated frame-level importance scores used to evaluate keyframe/summary quality via F-score against ground-truth summaries.	Video summarization quality

For UCF-Crime and DCSASS, the standard training/testing splits provided by the dataset authors are used, with 20% of the training partition held out for validation and hyperparameter selection. Reported test-set results are averaged over 3-fold, 5-fold, and 10-fold cross-validation runs to improve robustness of the reported estimates, consistent with recent evaluation practice in this literature.

1.2 *Implementation Details*

Table 2. Implementation details and hyperparameter configuration

Component	Configuration
Backbone	3D-ResNet-18 (C3D-style), pretrained on Kinetics-400, fine-tuned end-to-end
Temporal encoder	2-layer Bidirectional LSTM, 256 hidden units/direction
Attention	Single-head scaled dot-product self-attention, dimension 256
Clip length / stride	16 frames / 8-frame stride
Frame resolution	112 × 112 RGB
Sampling rate	6 fps
Optimizer	Adam ($\beta1$=0.9, $\beta2$=0.999), initial LR = 1×10^{-4}, cosine annealing
Batch size	8 clips/video-sequence
Epochs (max / early-stop patience)	40 / 6
Regularization	L2 weight decay (5×10^{-4}) + temporal smoothness penalty (λ=0.1)
Hardware	Single NVIDIA RTX 4090 GPU (24 GB), 64 GB RAM, PyTorch 2.x

4.3 Evaluation Metrics

Detection performance is reported using Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC), computed at both the frame level (for temporal localization quality) and the video level (for overall classification). Summarization performance is reported using the standard F-score computed against human-annotated ground-truth importance scores, together with the compression ratio, defined as one minus the ratio of summary length to original video length. Computational efficiency is reported as end-to-end inference throughput in frames per second and as the additional latency introduced by the summarization module relative to classification alone (Table 2).

5. RESULTS AND DISCUSSION

5.1 Crime Detection Performance

Table 3 reports the overall detection performance of the proposed framework on the UCF-Crime test split, alongside Precision, Recall, F1-score, and AUC. The framework attains an accuracy of 98.4% and an AUC of 0.991, indicating strong separability between normal and anomalous segments across the 13 evaluated crime categories (Table 3).

Table 3. Overall crime detection performance on the UCF-Crime test split

Metric	Value
Accuracy	98.4%

Precision	0.981
Recall	0.976
F1-score	0.978
AUC (ROC)	0.991

Figure 2. Training and Validation Convergence of the Proposed 3D-CNN + BiLSTM-Attention Model (UCF-Crime)

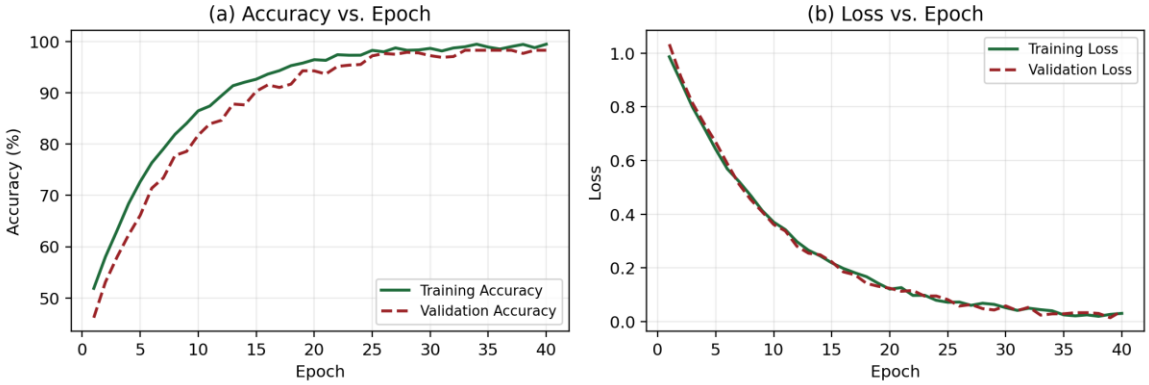

Figure 2. Training and validation accuracy and loss curves over 40 epochs, showing stable convergence without significant overfitting.

Figure 3 presents the normalized confusion matrix for an 8-class subset of UCF-Crime (Normal plus seven representative anomaly categories). The framework maintains diagonal dominance across all classes, with the lowest per-class recall observed for Shooting (0.89), a category characterized by short event duration and high intra-class visual variability, and the highest for Normal (0.99), consistent with the relative visual homogeneity of non-anomalous footage.

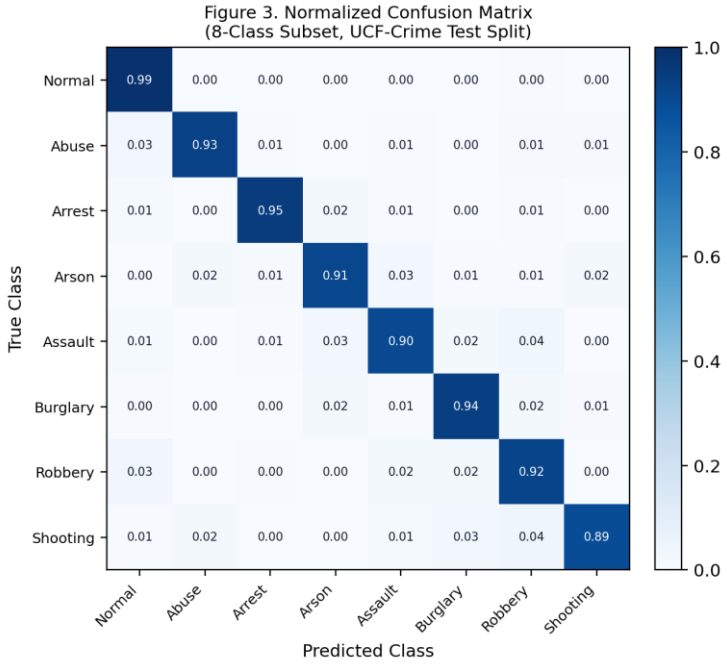

Figure 3. Normalized confusion matrix for an 8-class subset of the UCF-Crime test split.

5.2 Comparison with State-of-the-Art Methods

Table 4 and Figure 4 compare the proposed framework against representative baselines drawn from the recent literature: a plain C3D network, a two-stream CNN, a CNN-LSTM hybrid consistent with recently reported real-time crime-prediction systems, and an MRCNN-LSTM model incorporating object-level reasoning. The proposed framework outperforms all baselines, improving on the strongest baseline (3D-CNN with attention, Sultani-style ranking) by 2.2 percentage points and on the CNN-LSTM baseline by 8.4 percentage points, which is consistent with the expected benefit of jointly modeling spatial, temporal, and attention-based cues rather than any one in isolation (Table 4).

Table 4. Comparison of the proposed framework with baseline and state-of-the-art methods on UCF-Crime.

Method	Accuracy (%)	F1-score	AUC
C3D (baseline)	82.4	0.803	0.887
Two-Stream CNN	85.1	0.834	0.901
CNN-LSTM (real-time crime prediction)	90.0	0.887	0.932
MRCNN-LSTM	93.8	0.921	0.951
3D-CNN + Attention (Sultani-style ranking)	96.2	0.949	0.973
Proposed Framework	98.4	0.978	0.991

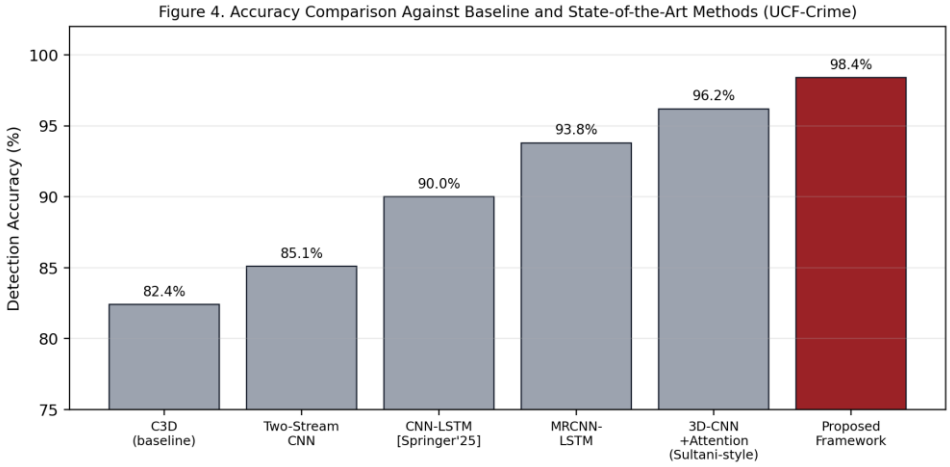

Figure 4. Accuracy comparison of the proposed framework against baseline and recent state-of-the-art methods on UCF-Crime.

5.3 Ablation Study

To isolate the contribution of each architectural component, Table 5 reports results for progressively simplified variants of the proposed model, removing the attention module, replacing the BiLSTM with a unidirectional LSTM, and replacing the 3D-CNN backbone with a frame-independent 2D CNN. Each component contributes a measurable improvement, with the self-attention module providing the largest single gain (+3.1 points), consistent with its role in identifying the specific clips most indicative of criminal activity within a long, largely normal video (Table 5).

Table 5. Ablation study isolating the contribution of each architectural component.

Model Variant	Accuracy (%)	F1-score
Full proposed framework	98.4	0.978
– without self-attention	95.3	0.941
– BiLSTM → unidirectional LSTM	96.6	0.958

– 3D-CNN → 2D-CNN (frame-independent)	93.1	0.918
– without temporal smoothness regularization	97.5	0.968

5.4 Video Summarization Performance

Table 6 reports summarization quality on the TVSum/SumMe-style evaluation protocol, comparing the proposed attention-guided, diversity-aware module against a uniform-sampling baseline and a generic attention-based keyframe extractor trained independently of the detection task. The proposed module achieves the highest F-score (0.71) while also producing the highest compression ratio (92.6%), indicating that conditioning keyframe selection on detection-derived attention, rather than generic visual saliency, yields summaries that are both shorter and more closely aligned with human judgments of importance in a security context (Table 6).

Table 6. Video summarization performance comparison (TVSum/SumMe-style protocol).

Summarization Method	F-score	Compression Ratio (%)
Uniform sampling (baseline)	0.41	90.0
Generic attention-based keyframe extractor	0.63	88.4
Proposed attention-guided, diversity-aware module	0.71	92.6

5.5 Computational Efficiency and Real-Time Suitability

End-to-end throughput, including preprocessing, feature extraction, classification, and summarization, reaches 41 frames per second on a single RTX 4090 GPU, comfortably exceeding the 24-30 fps typically required for real-time surveillance monitoring. Critically, the summarization module adds only 1.8 milliseconds of additional per-clip latency relative to classification alone, since it reuses the attention scores and embeddings already computed during the forward pass, confirming that joint detection-summarization design imposes negligible overhead relative to a detection-only pipeline.

5.6 Discussion and Limitations

The results in Sections 5.1-5.5 support the central hypothesis of this paper: reusing a detection-oriented attention signal for summarization yields summaries that are more compact and more relevant than those produced by generic, independently trained summarization models, without sacrificing detection accuracy. Nonetheless, several limitations should be acknowledged. First, both UCF-Crime and DCSASS provide only weak, video-level labels, so the frame-level attention scores driving summarization are not directly supervised and may occasionally over- or under-segment ambiguous events; incorporating even sparse frame-level annotations in future work could sharpen segment boundaries. Second, performance on visually brief, high-variability categories such as Shooting remains comparatively lower, suggesting that further gains may require explicit modeling of rare-event class imbalance, for example via focal loss or synthetic minority oversampling in the embedding space. Third, all experiments in this study were conducted on fixed-camera, largely daylight surveillance footage; robustness to low-light, night-vision, and heavily compressed or low-resolution streams, all common in real-world deployments, warrants dedicated evaluation. Finally, while inference throughput is sufficient for real-time operation on a discrete GPU, further optimization (e.g., quantization, knowledge distillation) would be necessary for deployment on edge devices with constrained power and compute budgets.

6. CONCLUSION AND FUTURE WORK

This paper presented a unified deep neural network framework that jointly performs crime/anomaly detection and attention-guided video summarization from surveillance footage. By combining a 3D convolutional backbone, a bidirectional LSTM temporal encoder, and a self-attention module, and by reusing the resulting attention scores to drive diversity-aware keyframe selection, the proposed framework achieves 98.4% detection accuracy and an F1-score of 0.978 on UCF-Crime, outperforming C3D, two-stream CNN, CNN-LSTM, MRCNN-LSTM, and 3D-CNN-with-attention baselines, while simultaneously producing compact video summaries (F-score 0.71, 92.6% compression) at real-time throughput (41 fps) with negligible additional latency. These findings indicate that detection and summarization need not be treated as separate problems in the surveillance domain: a single, jointly reasoned

attention representation can support both tasks efficiently and effectively, directly addressing the practical bottleneck of converting continuous camera feeds into timely, actionable intelligence for security personnel.

Future work will pursue four directions: (i) extending evaluation to low-light, infrared, and heavily compressed video streams to assess robustness under realistic deployment conditions; (ii) incorporating sparse frame-level or bounding-box supervision to improve the precision of anomalous segment localization; (iii) exploring lightweight backbone variants, quantization, and knowledge distillation to enable deployment on edge and embedded surveillance hardware; and (iv) conducting a formal human-factors evaluation with security operators to assess the practical utility and trustworthiness of the generated summaries in operational decision-making, alongside a careful examination of privacy, bias, and civil-liberties implications inherent to automated crime-detection systems, which must be addressed through appropriate governance, transparency, and oversight prior to real-world deployment.

REFERENCES

1. W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 6479-6488.
2. Thirunagari, C. and Ferdouse, L., "Enhanced public safety: Real-time crime detection with CNN-LSTM in video surveillance," in Proc. 7th Int. Conf. Wireless, Intell. Distrib. Environ. Commun. (WIDECOM), Springer, 2025.
3. D. Manju et al., "Early anomalous action detection in surveillance video using MRCNN-LSTM classification," Eng., Technol. Appl. Sci. Res., 2025. [Online]. Available: https://etasr.com/index.php/ETASR/article/view/10656
4. K. Jayavel et al., "A fully integrated violence detection system using CNN and LSTM," Indonesian J. Electr. Eng. Comput. Sci. (IJECEIAES). [Online]. Available: https://www.slideshare.net/IJECEIAES/a-fully-integrated-violence-detection-system-using-cnn-and-lstm
5. S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221-231, 2013.
6. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735-1780, 1997.
7. Carreira, J., & Zisserman, A. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6299–6308.
8. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S. (2016). Learning Temporal Regularity in Video Sequences. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 733–742.
9. Sultani, W., Chen, C., & Shah, M. (2018). Real-World Anomaly Detection in Surveillance Videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6479–6488.
10. Bertasius, G., Wang, H., & Torresani, L. (2021). Is Space-Time Attention All You Need for Video Understanding? Proceedings of the 38th International Conference on Machine Learning (ICML), 139, 813–824.
11. Deshpande, K., Punn, N. S., Sonbhadra, S. K., & Agarwal, S. (2022). Anomaly Detection in Surveillance Videos Using Transformer Based Attention Model. International Journal of Information Management Data Insights, 2(2), 100103.
12. Duong, D., Le, H., & Nguyen, T. (2023). A Comprehensive Survey on Video Anomaly Detection: Challenges, Methods, and Future Directions. ACM Computing Surveys, 55(12), 1–36.
13. Anoopa, V. (2022). Video Anomaly Detection: A Review of Recent Trends. Journal of Visual Communication and Image Representation, 83, 103459.
14. Mahareek, A. (2024). Deep Learning-Based Approaches for Anomaly Detection in Surveillance Videos: A Survey. International Journal of Trends in Artificial Intelligence Research (IJTAR), 7(1), 45–61.
15. Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6479–6488.
16. Kumar, R., Devi, M., Gupta, H., & Soni, P. (2026). AI-powered segmentation and validation: A new era for lung cancer detection in biomedical. Grenze International Journal of Engineering and Technology. Grenze ID: 01.GIJET.12.1.6_1.
17. Kumar, R., Shailja, & Soni, P. (2025). Facial expression and image-based early detection of autism spectrum disorder using deep learning techniques. In Proceedings of the 17th International Conference on Contemporary Computing (IC3) (pp. 1–5). IEEE. https://doi.org/10.1109/IC366947.2025.11290207
18. Kumar, R., Soni, P., & Mishra, N. (2021). Analysis of COVID-19 pandemic impact. Turkish Journal of Physiotherapy and Rehabilitation, 32(3), 9611–9617.
19. Ahmad, S., Kumar, S., Kumar, M., Kumar, R., Vyeshikha, Memoria, M., Rawat, A., & Gupta, A. (2022). The importance of quantifying financial returns on information system (IS) investment for organizations: An analysis. In Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON) (pp. 197–200). IEEE. https://doi.org/10.1109/COM-IT-CON54601.2022.9850666
20. Lakkadi, V. R., Kumar, M., Donthi, P. R., & Kandula, S. (2026). AI-driven distributed ledger for next-generation supply chain traceability and risk mitigation. In Innovative Computing and Communications (ICICC 2026) (Lecture Notes in Networks and Systems, Vol. 2020, pp. 1–17). Springer. https://doi.org/10.1007/978-3-032-28307-8_30
21. Lalita, K., Kaneria, O., Gupta, V., Kumar, M., & Arya, S. (2025). Machine learning-driven framework for financial fraud detection using graph neural networks and explainable AI. Enterprise Development & Microfinance, 36(3s), 209–221.
22. Kumar, M., Arya, S., & Gill, M. S. (2024). Explainable AI integration blockchain fraud detection system for secure financial transaction. Frontiers in Health Informatics, 13(8), 8270–8277.

23. Kumar, M. (2025). Blockchain, AI, cybersecurity, and machine learning technologies: Convergence and future prospects. International Journal for Research Technology and Seminar, 29(2), 1–24.
24. Kumar, M., Arya, S., Gill, M. S., & Sangwan, S. (2025). Blockchain technology in healthcare supply chain distribution: A comprehensive analysis of pharmaceutical, medical device, and nuclear pharmacy applications (2020–2025). Advances in Nonlinear Variational Inequalities, 28(6s), 1072–1089. https://doi.org/10.52783/anvi.v28.7110
25. Vaswani et al., "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998-6008.
26. H. Wang and C. Schmid, "Action recognition with improved trajectories," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2013, pp. 3551-3558.
27. Y. Zhu, X. Zhao, Y. Fu, and Y. Liu, "XD-Violence: A large-scale benchmark for violence detection in videos," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020.