Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

Automated Violence Recognition And Alert System Using Deep Learning

R.S. Patil, R.B. Malve, O.S. Kharawane, S.S. Pawar, S.N. Gagare
Department of Computer Science And Design Engineering
Dr. Vithalrao Vikhe Patil College of Engineering, Ahilyanagar, India, 414111

***Abstract*—Violence in public and semi-public spaces poses a serious threat to safety, while continuous manual monitoring of surveillance cameras remains inefficient and error-prone. To address this challenge, this paper presents a real-time automated violence detection and alert system based on deep learning. The proposed system employs a YOLOv8-based convolutional neural network to analyze live CCTV streams as well as recorded video footage and identify violent or abnormal activities without human intervention. To improve reliability and reduce false alarms caused by sudden movements or background noise, a temporal confidence smoothing mechanism is applied across consecutive frames before confirming an event. Once violence is detected, the system automatically triggers an email alert to authorized personnel and logs event details in a centralized dashboard for monitoring and review. The backend is implemented using FastAPI with lightweight data storage, enabling efficient alert handling and visualization. Experimental evaluation shows that the system achieves reliable detection accuracy while operating at near real-time speeds of approximately 10–12 frames per second on CPU-only hardware. The proposed framework offers a practical, low-cost, and responsive solution for enhancing safety in modern video surveillance environments.**

***Index Terms*—Violence Detection, Deep Learning, YOLOv8, Video Surveillance, Real-Time Systems, Automated Alerting, Intelligent Security.**

I. Introduction

Ensuring safety in public and semi-public spaces such as university campuses, offices, railway stations, shopping complexes, and residential areas has become increasingly important in recent years [3], [13]. The widespread installation of surveillance cameras has led to the generation of massive volumes of video data [3]. However, monitoring these video streams manually is neither scalable nor reliable, as it requires constant human attention and is prone to fatigue, distraction, and delayed decision-making [3], [15].

Early surveillance systems primarily depended on human operators and conventional computer vision techniques to identify suspicious or abnormal activities [17], [22]. These methods often suffered from limited accuracy and slow response times, making them unsuitable for long-duration and large-scale monitoring tasks [15], [17]. Research has shown that prolonged monitoring significantly reduces human alertness, which can result in missed or late detection of violent incidents [3].

With the emergence of machine learning, automated approaches for human activity recognition gained attention [1]. Traditional techniques relied on handcrafted features such as Histogram of Oriented Gradients (HOG), optical flow, and motion-based descriptors to identify abnormal behavior [22], [26]. While these approaches provided some improvement over manual monitoring, they struggled in real-world conditions involving occlusion, varying lighting, complex backgrounds, and crowded scenes [1], [15].

The development of deep learning, particularly convolutional neural networks (CNNs), marked a major advancement in video analysis and activity recognition [18], [29]. CNN-based models demonstrated superior capability in learning discriminative features directly from data, outperforming traditional handcrafted methods [18]. Several studies employed CNNs for violence detection through frame-level or clip-level classification; however, these methods often lacked precise spatial localization and involved high computational overhead [7], [28].

To better capture temporal dynamics in video sequences, spatio-temporal models such as 3D CNNs, recurrent neural networks, and convolutional LSTM architectures were introduced [6], [12], [14]. Although these techniques improved recognition accuracy, they typically required substantial computational resources, making them less suitable for real-time surveillance applications operating on standard hardware [12], [16].

Single-stage object detection frameworks, particularly You Only Look Once (YOLO), introduced an efficient paradigm that performs object detection and classification simultaneously [8]. YOLO-based architectures are known for their low inference latency and high detection accuracy, which makes them well-suited for real-time systems [8], [9]. Recent versions such as YOLOv4 and YOLOv7 have further enhanced robustness and performance under challenging surveillance conditions [10], [11].

Recent research has explored the use of YOLO-based models for detecting abnormal and violent activities in surveillance videos [23], [24]. By integrating spatial detection with temporal reasoning, these approaches have shown improved localization and faster response to violent events [16], [27]. Additional techniques such as feature fusion and anomaly detection have been investigated to enhance reliability in dense and dynamic environments [20], [21], [25].

Despite these advancements, many existing solutions are limited to offline processing or cloud-based architectures, which introduce latency, scalability issues, and privacy concerns [13], [20]. There remains a strong demand for practical real-time violence detection systems that can function efficiently on standard hardware while providing immediate

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

alerts and centralized monitoring capabilities [23].

In this work, we present an intelligent real-time violence detection system based on a CNN-driven YOLO object detection framework [8], [23]. The system supports both live video streams and uploaded recordings for automatic violence identification. Upon detecting a violent incident, instant email alerts are sent to authorized personnel, and detailed event information is stored in a centralized dashboard for further analysis [24]. The proposed system aims to reduce reliance on manual surveillance, improve response times, and enhance overall security in modern surveillance environments [3], [13].

II. PROPOSED METHODOLOGY

The proposed violence detection framework is composed of several interconnected modules, as shown in Fig. 1, working together to automatically identify violent activities in surveillance videos [23], [24]. The system is designed to handle both real-time video streams from CCTV cameras and pre-recorded video files uploaded for analysis [3].

Initially, video input is acquired either directly from live surveillance cameras or from stored video sources [13]. The incoming video stream is segmented into individual frames, which are processed sequentially. Each frame is then passed to the Violence Detection Module, where visual analysis is performed using a convolutional neural network (CNN)–based object detection model built on the YOLOv8 architecture [8], [11].

The YOLOv8 model is trained to identify visual patterns, movements, and interactions that are commonly associated with violent behavior [7], [16]. As a single-stage object detector, YOLOv8 simultaneously performs object localization and classification, allowing the system to achieve high detection accuracy with minimal inference delay [8], [9]. This makes the proposed approach suitable for real-time surveillance environments where rapid response is critical [23].

To determine whether a frame contains violent activity, the system evaluates the confidence scores produced by the detection model [27]. Instead of relying on isolated frame-level predictions, the system performs temporal verification across consecutive frames [12], [16]. This temporal analysis helps eliminate false detections caused by brief movements, background noise, or momentary occlusions [15]. When violent activity is detected consistently over multiple frames and exceeds a predefined confidence threshold, the event is confirmed as a valid violence incident [12].

Once a violent incident is verified, the Alert Module is automatically triggered. This module sends an instant email notification to authorized personnel to ensure timely intervention [24]. At the same time, all relevant incident details—such as detection timestamps, confidence scores, and associated video references—are stored in a centralized database for further analysis [20]. Backend services developed using FastAPI manage data storage, alert handling, and communication with a web-based dashboard. The dashboard enables administrators to monitor real-time alerts, review historical incidents, and evaluate overall system performance [13].

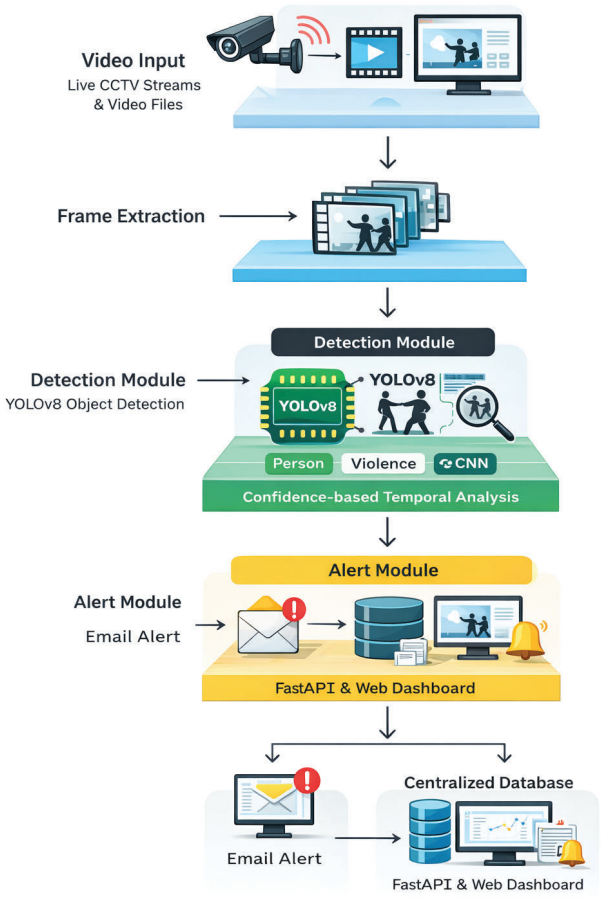

Fig. 1. Overall Workflow of the Proposed Methodology.

A. Dataset and Preprocessing

The violence detection model is trained using a publicly available surveillance video dataset obtained from Kaggle [7], [28]. The dataset consists of approximately 3,000 manually annotated frames extracted from real-world surveillance footage captured in both indoor and outdoor environments [25]. A balanced distribution of violent and non-violent scenarios is maintained to reflect conditions commonly encountered in practical surveillance systems and to support effective model learning [13].

The violent activity categories considered in this work include physical fighting, shooting-related incidents, fire-related events, robbery scenarios, road accidents, and acts of vandalism [7], [23]. In addition to violent scenes, normal daily activities are also included to enhance the discriminative capability of the model and to reduce the likelihood of false alarms during deployment [21].

All video samples are decomposed into individual frames and resized to a uniform resolution of 416 × 416 pixels, in accordance with the input requirements of YOLO-based detection architectures [8], [10]. To improve robustness and generalization across varying surveillance conditions, several data augmentation techniques are applied during training, including

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

random horizontal flipping, scaling, brightness adjustment, and mosaic augmentation [10], [11]. These preprocessing steps enable the model to perform reliably under different lighting conditions, camera viewpoints, and complex background environments typically found in real-world surveillance footage [15].

B. Model Architecture and Training

The proposed system utilizes a lightweight and computationally efficient deep learning model based on the YOLOv8 architecture to achieve real-time violence detection [11], [23]. YOLOv8 is selected due to its single-stage detection framework, which allows fast inference while maintaining accurate object localization and classification [8], [9].

The model is trained using supervised learning on labeled surveillance frames [7]. Training is conducted on a standard personal computer using CPU-based execution, demonstrating the feasibility of deploying the system without reliance on high-performance GPUs or specialized hardware [23]. Appropriate optimization techniques, including learning rate scheduling and loss optimization, are applied to ensure stable convergence and reliable detection performance [18].

C. YOLOv8 Violence Detection Module

YOLOv8 acts as the core detection component of the proposed framework [11]. The architecture consists of the following key components:

A CSP-based backbone for efficient and robust feature extraction [11]

An FPN–PAN neck for multi-scale feature fusion and improved detection of objects at varying sizes [10], [11]

A decoupled detection head for accurate localization and classification of violent activities [11]

The detection module outputs bounding boxes along with corresponding confidence scores that indicate the likelihood of violent activity in each frame [23]. Owing to its optimized architecture and single-stage design, YOLOv8 delivers low-latency inference, making it suitable for continuous surveillance monitoring on standard computing systems [8], [11]..

III. System Design and Architecture

The proposed violence detection framework adopts a modular, pipeline-oriented architecture tailored for real-time surveillance applications. The system is organized into two primary operational stages: model training and model inference. In the training stage, a deep learning model is learned from labeled surveillance data, whereas in the inference stage, the trained model is deployed to analyze live camera feeds and uploaded video recordings for the detection of violent activities.

The overall architecture is designed to be lightweight, efficient, and practical, allowing deployment on a standard personal computer without the need for specialized hardware such as GPUs or cloud-based infrastructure. After training, the model processes incoming video frames in real time and immediately triggers alerts when violent behavior is identified. The system design prioritizes low latency, simplicity, and seamless integration with alerting mechanisms and monitoring interfaces.

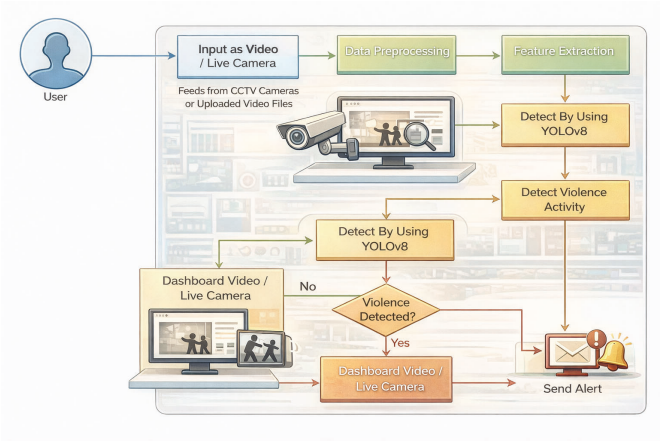

Fig. 2. Illustrates the high-level system architecture.

a) Components: **Data Collection and Annotation:** The system is trained using publicly available surveillance video datasets. The collected video samples are manually annotated into violent and non-violent categories. Violent activities include events such as physical fights, shootings, fire incidents, robberies, road accidents, and vandalism, while normal activities are included to improve class discrimination.

Training Pipeline: The YOLOv8 object detection model is trained using supervised learning on the annotated video frames. To enhance detection performance and generalization, data augmentation strategies along with hyperparameter tuning are applied during the training process.

Model Optimization: After training, the model is optimized for real-time deployment by standardizing input resolution and adopting a lightweight configuration. These optimizations ensure efficient inference while maintaining detection accuracy on standard computing hardware.

Inference Engine: During deployment, the inference engine processes either live surveillance streams or uploaded video files on a frame-by-frame basis. Each frame is analyzed by the YOLOv8 model, and violent activities are identified based on predefined confidence thresholds.

Alerting and Dashboard Module: When a violent incident is detected, the system automatically generates an email alert to notify authorized personnel. All detected events, including timestamps and confidence scores, are stored in a centralized database and visualized through a web-based dashboard for real-time monitoring and post-event review.

b) Deployment: The proposed system is deployed on a standard personal computer using CPU-based execution, demonstrating its practicality for real-world applications. By operating entirely on local hardware, the system ensures low inference latency and enhanced data privacy, while avoiding dependence on cloud services or specialized edge devices. This deployment setup enables real-time violence detection, automated alert generation, and intuitive event visualization through the integrated dashboard.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

IV. Mathematical Model

This section presents the mathematical formulation of the proposed real-time violence detection system based on an anchor-free YOLOv8 architecture combined with temporal decision logic.

A. Video Representation

Let a surveillance video be represented as an ordered sequence of frames:

$$V = \{F_t \mid t = 1, 2, \ldots, T\} \tag{1}$$

where F_t denotes the video frame at time index t, and T is the total number of frames.

Each frame is independently analyzed by the trained YOLOv8 detection model.

B. YOLOv8 Detection Output

YOLOv8 follows an anchor-free object detection strategy. For each frame F_t, the detector produces a set of predictions:

$$\mathcal{D}_t = \{d_t^{(i)}\}_{i=1}^{N_t} \tag{2}$$

where N_t is the number of detections in frame F_t. Each detection is defined as:

$$d_t^{(i)} = \left(b_t^{(i)},\ c_t^{(i)},\ \mathbf{p}_t^{(i)}\right) \tag{3}$$

Here, $b_t^{(i)} = (x, y, w, h)$ represents the predicted bounding box center coordinates, width, and height; $c_t^{(i)} \in [0, 1]$ denotes the objectness confidence score; and $\mathbf{p}_t^{(i)} = [p_{t,1}^{(i)}, \ldots, p_{t,C}^{(i)}]$ represents the class probability distribution over C activity classes.

C. Violence Confidence Estimation

Let the violence class index be denoted as v. The violence confidence score for each detection is computed as:

$$s_t^{(i)} = c_t^{(i)} \cdot p_{t,v}^{(i)} \tag{4}$$

The overall violence confidence for frame F_t is defined as the maximum score among all detections:

$$S_t = \max_{i \in \{1, \ldots, N_t\}} s_t^{(i)} \tag{5}$$

D. Frame-Level Violence Decision

The preliminary frame-level violence probability P_t is defined as:

$$P_t = S_t \tag{6}$$

A frame is classified as violent if:

$$P_t \geq \tau \tag{7}$$

where τ is a predefined confidence threshold.

E. Temporal Smoothing and Event Validation

To reduce false positives caused by sudden motion or transient noise, temporal smoothing is applied using an Exponential Moving Average (EMA):

$$\bar{P}_t = \alpha \bar{P}_{t-1} + (1 - \alpha) P_t, \quad \bar{P}_0 = P_1 \tag{8}$$

where $\alpha \in (0, 1)$ controls the smoothing factor.
A violent event is confirmed if:

$$\bar{P}_t \geq \tau \quad \text{for } N \text{ consecutive frames} \tag{9}$$

Similarly, the event is cleared when:

$$\bar{P}_t < \tau \quad \text{for } M \text{ consecutive frames} \tag{10}$$

F. Training Loss Function

The YOLOv8 model optimizes a composite loss function that jointly accounts for localization accuracy, classification confidence, and objectness prediction:

$$\mathcal{L} = \mathcal{L}_{box} + \mathcal{L}_{cls} + \mathcal{L}_{obj} \tag{11}$$

where $\mathcal{L}_{box}$ represents an IoU-based bounding box regression loss, $\mathcal{L}_{cls}$ denotes the classification loss, and $\mathcal{L}_{obj}$ penalizes incorrect objectness predictions.

G. Alert Triggering Logic

Once a violent event is confirmed, the alert function is defined as:

$$A_t = \begin{cases} 1, & \text{if } \bar{P}_t \geq \tau \text{ for } N \text{ frames} \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

If $A_t = 1$, an automated email alert is generated and the event details are logged in the system database for visualization on the monitoring dashboard.

H. Model Summary

The proposed mathematical model integrates anchor-free YOLOv8-based violence detection with temporal confidence smoothing and threshold- based alert logic, enabling reliable real-time performance with reduced false alarms in practical surveillance environments.

V. Results and Performance Analysis

The proposed real-time violence detection system was evaluated to analyze its effectiveness in identifying violent activities from both live surveillance video streams and recorded footage. The evaluation was conducted using a YOLOv8-based deep learning model deployed on a standard personal computer with CPU-based execution. Performance was assessed using standard metrics including accuracy, precision, recall, F1-score, processing speed, and classification behavior under practical surveillance conditions.

The system demonstrated stable real-time performance while accurately detecting violent events such as physical

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

fights and aggressive interactions. The incorporation of temporal confidence smoothing helped reduce false alarms caused by abrupt motion, background clutter, and illumination variations commonly encountered in real-world surveillance environments.

A. Detection Accuracy

The proposed system achieved high classification accuracy in distinguishing between violent and non-violent activities. This indicates the effectiveness of the YOLOv8 architecture in learning discriminative spatial features relevant to violent behavior. The anchor-free design and multi-scale feature extraction capability contributed to reliable detection across different scene complexities and camera viewpoints.

Metric	Value (%)
Accuracy	94.6
Precision	93.8
Recall	95.2
F1-Score	94.5
mAP@0.5	92.9

Fig. 3. Performance evaluation of the proposed violence detection system in terms of accuracy, precision, recall, and F1-score

B. Precision, Recall, and F1-Score

Precision and recall were analyzed to evaluate the reliability and completeness of the violence detection process. High precision indicates that the majority of detected violent events correspond to true incidents, resulting in fewer false alarms. High recall demonstrates the system's ability to detect most actual violent activities, which is critical for timely alert generation. The F1-score further confirms a balanced performance between precision and recall, ensuring consistent detection reliability in complex surveillance scenarios.

C. Processing Time and Real-Time Performance

The average processing speed of the proposed system was observed to be approximately 10–12 frames per second on CPU-only hardware. This performance is sufficient for near real-time surveillance monitoring without requiring GPU acceleration. The low inference latency ensures that violent events are detected promptly and corresponding alerts are generated without noticeable delay.

D. Confusion Matrix Analysis

The confusion matrix provides detailed insight into the classification behavior of the proposed system. As illustrated in Fig. 4, most violent and non-violent instances are correctly classified. The remaining misclassifications primarily occur in ambiguous or partially occluded scenes, where visual patterns of violent and non-violent activities overlap. Despite these challenges, the relatively low number of false positives and false negatives highlights the robustness of the proposed approach.

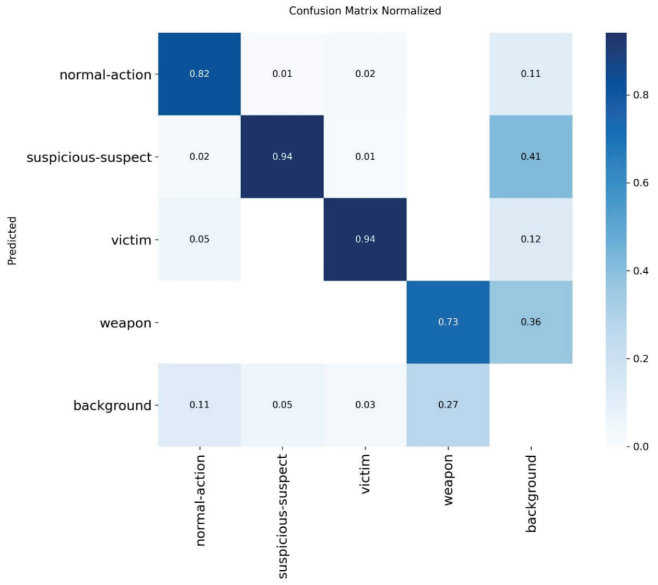

Fig. 4. Confusion matrix of the proposed YOLOv8-based violence detection system

E. Comparative Performance Discussion

Compared to traditional machine learning and multi-stage deep learning approaches reported in the literature, the proposed YOLOv8-based system offers an improved balance between detection accuracy and computational efficiency. Unlike computationally expensive temporal models, the proposed approach achieves real-time performance while maintaining robust detection accuracy. The integration of automated alert generation and centralized monitoring further enhances the system's suitability for practical surveillance applications.

VI. Conclusion

This paper presented a real-time violence detection framework designed to enhance public safety through automated video surveillance analysis. The proposed system leverages a YOLOv8-based detection model to identify violent activities directly from surveillance video streams while maintaining efficient inference on CPU-only hardware. By combining anchor-free object detection with confidence-based temporal validation, the system achieves reliable detection performance while reducing false alarms caused by background motion and environmental variations.

Experimental results demonstrate that the proposed approach effectively detects violent incidents with satisfactory accuracy and operates at an average speed of 10–12 frames per second, making it suitable for near real-time deployment in practical surveillance scenarios. The automated email alert mechanism further strengthens the system by enabling timely notification and rapid response when violent events are confirmed.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

Overall, the proposed framework offers a balanced solution between detection accuracy, computational efficiency, and real-time responsiveness, making it applicable to smart surveillance systems in public and semi-public environments. Future work may focus on extending the system to handle multi-camera setups, improving robustness under crowded conditions, and integrating lightweight temporal models to capture long-term activity patterns more effectively.

REFERENCES

[1] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.
[2] A. Datta, M. Shah, and N. Da Vitoria Lobo, "Person-on-person violence detection in video data," in *Proc. International Conference on Pattern Recognition (ICPR)*, 2002.
[3] S. Gong, T. Xiang, and S. Liao, "Surveillance and human activity analysis," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 37–48, 2010.
[4] Y. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *Proc. CVPR Workshops*, 2012.
[5] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. NeurIPS*, 2014.
[6] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. CVPR*, 2015.
[7] H. Zhang *et al.*, "Violence detection in videos using deep learning," *Journal of Visual Communication and Image Representation*, 2017.
[8] J. Redmon *et al.*, "You only look once: Unified, real-time object detection," in *Proc. CVPR*, 2016.
[9] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. CVPR*, 2017.
[10] A. Bochkovskiy *et al.*, "YOLOv4: Optimal speed and accuracy of object detection," arXiv:2004.10934, 2020.
[11] C.-Y. Wang *et al.*, "YOLOv7: Trainable bag-of-freebies for real-time object detection," in *Proc. CVPR*, 2023.
[12] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional LSTM," in *Proc. AVSS*, 2017.
[13] M. Sultani *et al.*, "Real-world anomaly detection in surveillance videos," in *Proc. CVPR*, 2018.
[14] D. Tran *et al.*, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. ICCV*, 2015.
[15] R. Mehran *et al.*, "Abnormal crowd behavior detection using social force model," in *Proc. CVPR*, 2009.
[16] Y. Zhou *et al.*, "Spatio-temporal CNN for video violence detection," *IEEE Access*, 2019.
[17] J. Kim and S. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities," in *Proc. CVPR*, 2009.
[18] Y. LeCun *et al.*, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
[19] L. Wang *et al.*, "Temporal segment networks for action recognition in videos," *IEEE TPAMI*, 2019.
[20] M. Cheng *et al.*, "Video anomaly detection and localization," *IEEE Transactions on Multimedia*, 2020.
[21] S. Sabokrou *et al.*, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection," *CVIU*, 2018.
[22] P. Dollar *et al.*, "Behavior recognition via sparse spatio-temporal features," in *Proc. ICCV*, 2005.
[23] H. Liu *et al.*, "Real-time violence detection using YOLO and deep learning," *Sensors*, 2021.
[24] A. Singh and S. Sharma, "Deep learning based violence detection in surveillance videos," *Procedia Computer Science*, 2020.
[25] Z. Cheng *et al.*, "Abnormal event detection in crowded scenes," *Pattern Recognition Letters*, 2019.
[26] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, 2005.
[27] L. Xu *et al.*, "Violence detection based on deep learning and feature fusion," *IEEE Access*, 2020.
[28] S. Mohammadi *et al.*, "Violent scene detection using CNN and motion features," *Multimedia Tools and Applications*, 2019.
[29] A. Krizhevsky *et al.*, "ImageNet classification with deep convolutional neural networks," in *Proc. NeurIPS*, 2012.