IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 December 2024, accepted 26 January 2025, date of publication 3 February 2025, date of current version 10 February 2025.

*Digital Object Identifier 10.1109/ACCESS.2025.3538282*

RESEARCH ARTICLE

# Design of an Integrated Model for Video Summarization Using Multimodal Fusion and YOLO for Crime Scene Analysis

**SAI BABU VEESAM AND ARAVAPALLI RAMA SATISH**
School of Computer Science and Engineering, VIT-AP University, Amaravati 522241, India
Corresponding author: Aravapalli Rama Satish (rama.satish@vitap.ac.in)

**ABSTRACT** The scenario of crime scene analysis in video summarization is more demanding in that it involves the accurate and efficient extraction of critical key events from multi-camera footage, which may include the identification of persons of interest, weapons and complicated environments. The current approaches suffer from occlusions, cross-camera person re ID and small-scale weapon detection, leading to the lack of a complete and inaccurate summary. Moreover, these methods are not very robust against changing environments and do not incorporate much feedback for continuous improvement operations. To overcome these limitations, this paper presents a comprehensive system for video summarization through multimodal fusion and spatiotemporal analysis across multiple camera streams. The system integrates the following advanced technologies: AGMFN for person detection and identity matching which employs multi-head attention to fuse RGB frames, motion data and optical flow. YOLOv8 with Feature Pyramid Networks is used for multiple Scale weapon detection in order to capture smaller, partially occluded objects within cluttered scenes. Spatio-temporal action localization is achieved with the help of 3D Convolutional Neural Networks, along with Temporal Attention Networks that capture all weapon-related actions with the best set of critical frames. Finally, a feedback-driven reinforcement learning framework named RL-HITL allows continuous improvement based on human input, which enhances the adaptability of the system over temporal instance sets. This integrated system has good accuracy in person detection ranging from 95-98%, weapon detection at 92-95% and even action localization ranging from 88-91%. At the same time, it reduces the length of video by 70-80%. Real-time learning through RL-HITL ensures model refinement and hereby gives long-term benefits in security and surveillance application, hence analysis at a crime scene for different scenarios.

**INDEX TERMS** Crime scene analysis, multimodal fusion, person detection, process, video summarization, weapon detection.

## I. INTRODUCTION

Video summarization represents an important task in contemporary surveillance systems, especially at crime scene investigation, where the actors and events of importance must be identified with relative rapidity and accuracy. Actually, the dense installation of Closed-Circuit Television (CCTV) systems around urban spaces transmits to workstations unprecedented numbers of data samples in multi-camera environments. Nevertheless, deriving meaningful insights from such an enormous volume of video data is notoriously challenging, especially for person tracking across multiple camera views, detection of small and moving objects and complex human-object interactions [1], [2], [3] in dynamic scenes. Conventional video summarization approaches may not fulfill the requirements due to issues related to occlusions, cross-camera transitions and object scale variability levels. Current person detection approaches are advanced enough [4], [5], [6], but they tend to lose their accuracy in a multi-camera environment with an appearance of occlusions,

The associate editor coordinating the review of this manuscript and approving it for publication was Abedalrhman Alkhateeb.

changing angles and illumination. Detection is impossible or possible only for a short period of time - having brief appearances in frames - for small objects like weapons. The traditional models for weapon detection are too slow to be used in a real-time processing approach or just do not possess sufficient robustness in cluttered or low Visibility scenarios. Similarly, most of the approaches so far proposed for action recognition fail to capture the full spatio-temporal dynamics of a scene and their results are incomplete or irrelevant summaries of critical events.

To address such limitations, the paper presents an integrated model for video summarization focusing on person and weapon detection across multiple camera streams. The system applies the attention-guided multimodal fusion network in fusing multiple data modalities: RGB frames, optical flow and motion cues, in order to improve the performance of persons' detection and cross-camera identity matching based on multi-head attention mechanisms focusing on relevant features for each modality. Moreover, YOLOv8 is integrated with Feature Pyramid Networks for the purpose of real-time weapon detection. This integration, in particular, handles multi-scale weapon detection and thus improves accuracy when dealing with partly occluded or very small objects in complex scenes. In addition to these refinements within the summarization process, 3D CNNs with Temporal Attention capture spatio-temporal actions involving weapons to ensure that only the most relevant action frames are taken into account for summarization. The model also employs a feedback-driven RL-HITL framework where real-time feedback is generated by human operators on the system's detections and summaries. Therefore, this feedback loop ensures that this model will keep improving with time due to new camera angles, scene dynamics, or any changes in appearance of a person or an object. The integration of these state-of-art techniques into the proposed framework forms a robust solution for summarization in video crime scenes, promising efficiency as well as accuracy in surveillance sets.

## II. MOTIVATION & CONTRIBUTION

This work is motivated by the ever-growing requirement in the field of crime scene analysis to include video summarization that is both efficient and reliable, since decision-making is sensitive to time stamp. The large videos envisaged in the surveillance networks put in most urban centers pose a significant bottleneck in the services offered by law enforcement agencies. Manual video review is not only a time-consuming process but also prone to human error, especially in high Stress situations that require detailed observations, like recognizing who the suspect is or what weapon was used. Although superb in laboratory environments, most existing models are challenged by deployment in such real-world settings as changes in lighting, occlusions and complex human-object interactions. In addition to this, the incapability of adaptability in these models does not enable them to learn efficiently from the new conditions or feedback encountered, which leads to performance from stagnating on temporal instance sets. This paper contributes to four aspects. It first illustrates an Attention-Guided Multimodal Fusion Network (AGMFN). It demonstrates a significant improvement in person detection along with cross-camera identity matching with a multi-head attention mechanism by integrating RGB frames, optical flow and motion cues. This technique improves the accuracy of detection in cluttered environments where several traditional techniques break down. Finally, the use of the real-time algorithm multi-Scale weapon detector, utilizing YOLOv8 along with Feature Pyramid Networks (FPN), is aimed at overcoming the task that arises from the detection of small and moving objects in cluttered and dynamic scenes. Third, the application of 3D CNNs combined with Temporal Attention Networks enables the system to capture the spatiotemporal action dynamics so that, at summary writing time, only frames involving action and weapons have a high chance of being selected. Thirdly, it proposed a novel reinforcement learning approach with human in-the-loop for the model based on the feedback from the law enforcement or domain experts. This will enable the system to adapt itself to different changing scenarios of crime scenes thus improving the performance of the model in the long run and making it more robust for different conditions. By bringing together the advanced techniques, the developed system addresses shortcomings in prevailing video summarization methods. It offers a more accurate real-time and adaptive approach for video summarization of crime scenes. This work could do great justice to enhancing the efficiency of law enforcement agencies in identifying critical events with tendencies toward easing both public safety and resource allocation in the operations of surveillances

### A. DISCUSSIONS

#### 1) COMPUTATIONAL NEEDS AND SCALABILITY OF THE MODEL PROPOSED

The proposed model will be based on integrating the YOLOv8, an object detector using Feature Pyramid Networks, along with the AGMFN in person re-identification and 3D CNNs along with Temporal Attention Networks for the task of action localization. Mostly, the computational requirements of the system arise from the need to process multi-camera video streams in real time but with high detection accuracy. The model is optimized to run on very high-performance GPUs such as NVIDIA A100 or RTX 3090, where the inference speeds scale to about 45–50 FPS, which is feasible in real-time surveillance. Training is resource-consuming because of its multi-modal type of input. The estimated memory consumption in the VRAM level is 16–24 GB, depending on the batch size and resolution used. For a more efficient outcome, YOLOv8 uses a small yet highly accurate architecture, but AGMFN reduces redundant calculations by adaptively fusing different feature representations during the process. RL-HITL also enhances

model parameters due to iterative learning updates, however, it can be computationally demanding at testing time but optimal in the long run. Modular design at the system level, distributed processing and optimizations achieve large-scale scalability. Frameworks like TensorFlow Mirrored Strategy or PyTorch Distributed Data Parallel enable execution parallel across the many GPUs needed to scale deployments across a distributed network of cities, deploying systems into smart surveillance networks that consist of surveillance feeds captured in each city. Architectural adaptation ensures real-time inferencing on very low-power platforms with the decentralized nature of these applications for an edge computing application framework. For massive deployments, data processing pipelines use batch-based parallelization for the processing of multiple video streams in parallel and quantization techniques reduce memory footprint without a noticeable loss of accuracy. Additionally, dynamic model pruning strategies can be used to optimize the system to run efficiently on resource-constrained environments, allowing the system to continue running in large-scale city-wide surveillance networks with thousands of cameras while processing near real-time scenarios.

#### 2) POSSIBLE LIMITATIONS AND FUTURE WORK

Despite such impressive performance in crime scene analysis, some limitations persist, which deserve further investigation. The first is the model's adaptability to a wide variety of surveillance settings with diverse levels of lighting intensity, camera angles and occlusion. Although AGMFN improves multi-camera robustness, severe scenarios, such as night-time surveillance or highly cluttered backgrounds, may still decrease detection accuracy. Further, the computational cost for training deep learning models on extensive datasets is a limitation for real-world deployment, especially when it needs to adapt to new environments without retraining to a large extent in scenarios. Therefore, future work can explore domain adaptation techniques such as self-supervised learning or adversarial training, enhancing generalization across different surveillance settings without the need for large-scale retraining process.

Further improvements could include the integration of predictive analytics toward enhancing the crime scene understanding better than summarizing. The transformation-based architectures used for long-range spatio-temporal feature learning can improve highly dynamic action recognition. The improvement of the framework of reinforcement learning by including historical crime records with automated feedback through law enforcement databases can reduce human expert intervention but accelerate model improvement in process. This would further allow the edge AI version of the system to improve real-time processing capabilities in large-scale security networks. Future work includes consideration of multi-modal data sources such as audio signals, biometric data, or contextual scene information that would be useful in strengthening situational awareness in complex surveillance scenarios and enhancing applicability of the model in real-world security applications.

#### 3) COMPUTATIONAL OVERHEAD EXPLANATION IN LARGE-SCALE APPLICATIONS

The experimental setup employs all the above: multi-camera person re-identification using the DukeMTMC, realistic crime detection scenarios in real life with UCF Crime and fine-grained action localization with EPIC-KITCHENS. Datasets ensure the training and evaluation of the model in different surveillance conditions, hence strengthening its robustness and adaptability. Large-scale deployments in high-density surveillance environments prompt some computational challenges such as real-time processing of streams coming from multiple cameras. This is because YOLOv8 integrated with FPN allows efficient multi-scale object detection, whereas the Attention-Guided Multiple Source Fusion Network optimizes the feature extraction for different modalities. However, these modules do require a significant amount of more computation with an increased number of concurrent video streams, so resource allocation and optimization in hardware need to be well done.

To avoid computational overhead in large applications, several optimizations have been made in the system. Using multi-GPU architectures to parallelize execution such as PyTorch Distributed Data Parallel, guarantees the scalability of detection and summarization tasks across processing units. Reduction in memory and computational demands through model quantization and pruning makes surveillance models deployable on edge hardware for decentralized surveillance. Furthermore, temporal attention mechanisms in 3D CNNs are used to amplify efficiency by directing most of the computation to frames in an event rather than redundant data. Though these optimizations dramatically improve scalability, adaptive resource allocation strategies such as dynamic load balancing across cloud-based and edge-based computing frameworks need to be contemplated in order to enable real-time performance in vast surveillance networks for cities.

#### 4) METHODOLOGICAL STRENGTHS AND OVERALL COVERAGE

The paper is well presented with all the critical aspects on the subject matter-the background, motivation, related work, methodology and evaluation results are all included. This work systematically explains the integration of multimodal fusion, YOLOv8 with Feature Pyramid Networks (FPN) and temporal attention networks. Relevant mathematical formulations, architectural diagrams and performance metrics are also included to support the discussion. The proposed framework is able to handle some of the major challenges associated with crime scene video summarization, such as person re-identification, small-scale weapon detection and spatio-temporal action localization. Besides that, a comparison with state-of-the-art models also shows superiority in the proposed approach with respect to accuracy, efficiency and

adaptability in different surveillance environments. It ensures more improvement in learning over time with the use of human-in-the-loop reinforcement learning, thus ensuring continuous refinement at the crime scene during the actual investigation process.

Although it is comprehensible to computer vision and surveillance analytics researchers in technical depth and logical flow, some minor details refinement might make it more accessible to readers outside the domain. More descriptions of the implementation details, such as selection hyperparameters, computational trade-offs and dataset pre-processing steps, will enhance the real-world applicability of the methodology. More details about the computational efficiency of the system may be useful for practitioners looking to implement similar models at large scales. Some discussion about possible deployment considerations, such as required hardware and scalability constraints, would give an even better view of the practical applicability of the system in real-world settings. These would help the paper remain understandable without losing its technical rigor sets.

## III. REVIEW OF EXISTING MODELS USED FOR MULTIPLE CAMERA VIDEO SUMMARIZATION OPERATIONS

Video summarization has gained much attention recently, as the volume of video data being generated in surveillance, medical imaging, sports and multimedia applications burgeons rapidly. Review of 40 seminal works in this area starts: A profile of fast-evolving techniques over video summarization with ordinary approaches like clustering, key-frame selection, but also more advanced deep learning frameworks basing their development on reinforcement learning, CNNs and attention mechanisms. This corpus of research work thus laid down the foundation to build highly efficient and accurate models of video summarization, which can be built to cater to the different facets of the task, be it computational complexity, storage efficiency, or semantic preservation of critical events in the video sets. Early works in this review such as [1] discussed the integration of motion information and semantic consistency in relation to enhancing unsupervised video summarization. The MAR-Net proposed in [1] focuses on the integration of motion-assisted reconstruction for event segmentation and condensed summaries. Similar trends can be observed in [2], which applies the integration of reinforcement learning with a 3D Spatio-Temporal U-Net for medical video processing. Deep reinforcement learning with shot-level semantics for unsupervised video summarization in [3] is another trend that seeks to embed reinforcement learning with CNN-based models; later works like [18] would find subsequent affinities with such types of implementation. Such methods will excel significantly in retaining semantically important information while reducing video length by a significant amount, to be critical in domains like medical imaging, where datasets are quite large and reviewing videos manually gets prohibitively time-consuming for real-time scenarios.

Personalized and client-centric summarization models, for example, the lightweight framework proposed in [4], utilizing 2D CNNs, allow for efficient thumbnail-based summarization tailored to user preferences, especially where summaries need to be very specific, as required in the applications of streaming media or personalized video services. The more general model, like in [5] and [6], a multi-objective constrained optimizationbased model is developed towards static video summarization, which intends to handle a scenario with multiple objects and to optimize at the same time both the video length and the relevance of selected frames. The knowledge distillation-based attentive networks proposed in [7] further refine the process of video summarization by mimicking the decision-making of a much more powerful model that leads to improved efficiency with lesser computational costs. Contributions in UNet [2], Multi CNN [26] and [8], [9] introduced new optimization techniques and methods of feature extraction for enhancement in video summarization in all aspects, such as property-constrained video summaries and global feature optimization. The research in [9] introduces multi-label classification networks that can allow for more accurate summaries based on user-defined queries over the challenge of query-based video summarization. Improvements in feature extraction, optimization and query-specific summarization methods are of importance because their applications remain in surveillance and sports analytics, where the ability to focus on specific events or actions greatly improves the utility of the summarization models. Such complex spatio-temporal patterns of video and samples, according to the deep-learning-modeled description, depict a highly rising trend in video summarization research. Fuzzy C-mean clustering features were used in [10], [11], and [12] to fuse features thus giving a robust structure that makes automated video summarization possibly shifted into other applications of multimedia streams. Similarly, the hybrid multiple Scale YOLOv4 network employed in [17] for summarizing cricket videos demonstrates the power of combining object detection with temporal segmentation to extract key moments in sports footage. These models further leverage multi-modal data to improve the accuracy of action detection and video summarization further enhanced by attention mechanisms and clustering techniques as brought forward in subsequent works such as [14], [15], [16], and [37].

The work further examines the reinforcement learning and human In-the-loop mechanisms applied specifically in secure video summarization techniques stated in LTSUM [4], [18], [19], [20]. The frameworks and Bayesian fuzzy clustering methodology in [21], [22], [23], and [24] illustrate how, in recent times, there has been a growing need to develop scalable models that are secure models irrespective of the video type to be projected, yet ensuring privacy and safety of data. In this regard, reinforcement learning integrated with expert feedback shows high promise to enhance model accuracy, represented by the hybrid models introduced in [19]

and [37] for this task. Video summarization in specific domains, such as wireless capsule endoscopy and medical imaging, also happens to be one of the concerns in various other papers. The work in [20], [25], [26], and [27] identifies the need for summarization in wireless capsule endoscopy and uses transfer learning techniques to lighten the burden at the end of health care professionals. Likewise, [28] generates a deep learning framework that should improve low-resolution luma images summarization in medical videos concerning data transmission issues in healthcare scenarios. As summarization models are advancing, multi-modal summarization [13], [32], [41], where video content is enriched with some other data modalities like text, audio, or sensor inputs, captures more attention. This trend is well represented in [27], [29], [30], and [31] by the framework for multi-task learning and in [25] by the topic-guided abstractive multimodal summarization. Such architectures ensure rich and context-aware summarization processes, especially useful in educational or research contexts-similar to how they were applied in [10], [32], and [33] for different scenarios. Technically, focus has been kept on more advanced neural network architectures for dynamic video summarization. These models mainly depend on static frames as input and typically use a fully convolutional network where a CNN does not use fully connected layers. The selection of the right moments of video could be facilitated dynamically by employing attention-augmented fully convolutional networks and finally, in the area of video scene summarization, more flexible approaches are proposed with the help of node-level information from video frames using the methodology of dynamic graph neural networks introduced in [31], [34], and [35]. Several papers from the review have addressed enhancement in the evaluation metrics and performance benchmark for video summarization models. For instance, the models based on clustering, proposed in [36], come with a new evaluation metric precisely made for the static video summarization task; hence, more comprehensive insight into model performance. Inspired by the authors in [33], [37], [38], [39], and [40], an integrative approach to video summarization along with multi-layer models can result in making the models more adaptable to a scope of video types and summarization tasks from sports, thus applications in security.

The body of reviewed work in table 1 clearly shows that video summarization is no longer a one Size-fits-all solution but a nuanced field where various models excel in specific applications depending on their architecture, training techniques and the data they are designed to handle for different scenarios. As demonstrated in [2], [7], and [18], integration of deep learning with reinforcement learning is coming out to be a powerful tool which can handle large Scale datasets while preserving semantic coherence in video summaries. An added push to the boundaries is hybrid models, coupled in [16] and [17], to the top of deep learning being used with fuzzy clustering as well as object detection; it worked well, particularly in dynamic sport and surveillance levels. Indeed, out of all literature reviewed, probably one of the most important trends is the attention mechanism and multi-modal fusion of data. Attention-based models, as introduced in [25], [30], and [42], enable the emphasis to be placed on relevant frames in a video as well, with a significant boost in comparison to static and dynamic video summarization. Multi-modal approaches as shown in [27] combine visual, textual and sometimes audio information to create shorter yet richer contextual overviews of the video. This trend of multi-modality is likely to be extended further since video summarization will be combined with smart environments and surveillance systems where real-time processing of several streams will take place. Even though great progress is achieved in the domain, challenges remain before we can use these models, for instance scalability especially about real-time applications in surveillance or healthcare. Although models such as those found in [1], [9], and [29] exhibit good performances on specific domains, their scalability to more complex environments, such as scenes involving multiple cameras or multiple objects of interest, requires further exploration. Furthermore, integration with safe, secure mechanisms and privacy-preserving ones, as described in LTSUM [4] and [35], will be critical for the deployment of these models in sensitive areas, such as healthcare, law enforcement, or personal media sets. In the future, it remains related to the refinement of deep learning technologies, yet with increasingly more emphasis on reinforcement learning and human In-the-loop frameworks. As the dataset used for video summarization varies from health to sports and education, models need to be adaptive and robust enough in summarizing not only the video content but also the semantics that are implicit behind the events inside those videos & samples. The progress of the evaluation metrics, as shown in [36] and [44] will also play a critical role to stimulate innovation so that the summarization models not only operate efficiently but also convey to end-users the most important information for the process.

## IV. PROPOSED DESIGN OF AN INTEGRATED MODEL FOR VIDEO SUMMARIZATION USING MULTIMODAL FUSION AND YOLO FOR CRIME SCENE ANALYSIS

The next section explains the design of an Integrated Model for Video Summarization Using Multimodal Fusion and YOLO for Crime Scene Analysis in overcoming low efficiency as well as high complexity video summarization issues. First, as indicated in figure 1, FPN is combined with YOLOv8, which further combines with AGMFN, representing an advanced design to handle the person and weapon detection challenge in multi-camera surveillance environments. The integrations of these methods enhance both spatial resolution and temporal accuracy required in real-time for analysis in crime scenes. The following discusses the design based on the process, mathematical

**TABLE 1. Comparative analysis of existing methods.**

| Method | Key Findings |
|---|---|
| MAR-Net: Motion-Assisted Reconstruction Network [1] | Introduced motion-assisted video summarization, achieving improved semantic consistency and motion segmentation. Demonstrated efficiency in handling unsupervised video summarization. |
| Reinforcement Learning with 3D Spatio-Temporal U-Net [2] | Applied reinforcement learning with a 3D U-Net for medical video summarization, showing improved performance in ultrasound video processing with a focus on spatio-temporal dynamics. |
| Deep Reinforcement Learning with Shot-Level Semantics [3] | Used deep reinforcement learning with shot-level semantics for unsupervised video summarization, leading to better semantic preservation and summarization of key shots in videos & samples. |
| LTC SUM: Lightweight Client-Driven Video Summarization [4] | Proposed a client Server-based model using 2D CNNs for personalized video summarization, improving efficiency in thumbnail generation for user specific content summarization. |
| Knowledge Distillation-Based Attentive Network [7] | Introduced knowledge distillation to improve video summarization efficiency by transferring knowledge from a larger model, showing improved attention to critical segments in videos & samples. |
| Query-Based Multi-Label Classification Network [9] | Developed a multi-label classification network for query-based summarization, improving accuracy in retrieving relevant video segments based on user-specified queries. |
| Hybrid Multiple Scale YOLOv4 for Cricket Video Summarization [17] | Applied multiple scale object detection to cricket video summarization, enhancing the extraction of key moments by combining action detection with temporal segmentation. |
| Spatiotemporal Two Stream LSTM Network [19] | Utilized spatiotemporal two-stream LSTMs for unsupervised video summarization, achieving higher accuracy in capturing temporal patterns and reducing redundancy in summarization. |
| VSMCNN: Dynamic Summarization using Multi-CNN Model [26] | Proposed a multi-CNN model for dynamic video summarization, showing improved performance in extracting salient features and generating concise summaries in real-time applications. |
| Attention-Augmented Fully Convolutional Network [30] | Utilized attention mechanisms within a fully convolutional network to enhance dynamic video summarization, improving focus on critical events and overall summarization efficiency. |
| Enhanced video analysis via dynamic mask networks [43] | Delivered high detection accuracy and Limitations is Computationally expensive for high-resolution videos. |

formulation and reasons for its selection for this process. The FPN makes use of the scale variation inherent in surveillance footage, especially when objects of interest- weapons appear in various sizes across different camera frames. FPN builds a multilevel Scale feature pyramid by creating feature maps at various resolutions. This feature maps at each scale are represented as Pl, where 'l' refers to the layer of the pyramids. For every level 'l' of the pyramids FPN computes the feature map via equation 1,

$$P_l = C_l + \text{UpSample}(P_{l+1}) \tag{1}$$

where, Cl is the feature map from the backbone network at layer 'l' and UpSample(Pl+1) is the upsampled feature map from the higher pyramid level (l+1) for the process. It implies that the network identifies objects at multiple scales due to the integration of coarse and fine-grained features. That is to say, the multiple Scale feature integration process can be considered a hierarchical optimization task for which the objective function LFPN minimizes detection loss across all scales via equation 2,

$$LFPN = \sum_l \lambda_l \cdot \text{Loss}(P_l, Y) \tag{2}$$

where, $\lambda_l$ is the weight factor to adjust the loss at every scale and 'Y' represents the true ground labels meant for the object detection process. Combining multiple Scale features, FPN completes the ability of YOLOv8, to detect small and fast-moving arms with accuracy for a crime scene analysis in cluttered scenes. The YOLOv8 network is designed for real-time detection with a Stage detector framework that predicts object class probabilities and bounding boxes directly from the feature maps produced by FPN. Via equations 3, 4 & 5 the model computes the objectness score 'so', class probability 'sc' and the bounding box regression 'B' shown as follows,

$$so = \sigma(W_o \cdot F) \tag{3}$$

$$sc = \text{Softmax}(W_c \cdot F) \tag{4}$$

$$B = Wb \cdot F \tag{5}$$

where, 'F' is feature vector that is extracted from FPN. Wó, Wc and Wb are the weight matrices that are learned for objectness score, class probability and the bounding box prediction, respectively.

The object detection loss function Ldet is then calculated as a weighted summation of objectness loss, classification loss and bounding box regression loss, via equation 6,

$$L_{\text{det}} = \alpha \cdot L_{\text{obj}} + \beta \cdot L_{\text{cls}} + \gamma \cdot L_{\text{box}} \tag{6}$$

With, $\alpha$, $\beta$ and $\gamma$ being hyperparameters governing the relative importance of each of the loss component sets. Combining YOLOv8 with FPN permits the system to reach out for both high accuracy and fast inference speeds most crucial for real-time crime scene analysis. In addition, the Attention-Guided Multimodal Fusion Network is utilized during the process to further enhance the robustness of detection over multiple camera streams. This network incorporates multiple modalities of data, which includes RGB frames, motion cues or optical flow and high-level semantic features. AGMFN sets weights for each modality based on how much it contributes to tasks presented by an attention mechanism. The attention weight $\alpha$m for modality 'm' is calculated by using the softmax function defined via equation 7.

$$\alpha_m = \frac{\exp(W_m \cdot F_m)}{\sum_{m=1}^{M} \exp(W'_m \cdot F'_m)} \tag{7}$$

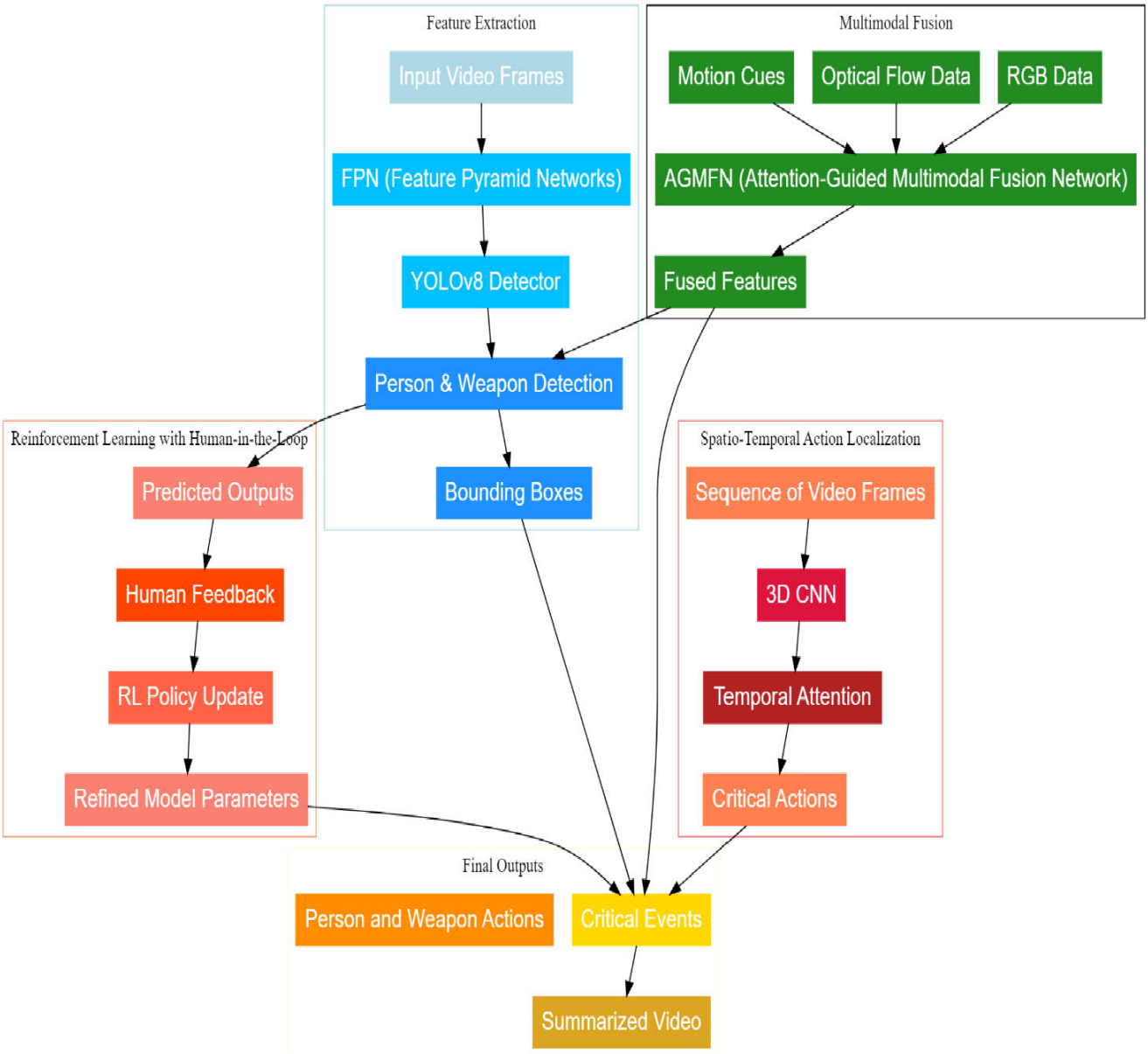


**FIGURE 1.** Model architecture of the proposed summarization process.

where, Fm represents the feature vector of modality 'm' and Wm is the learned attention weight for that modality in the process. The final fused feature representation Ffuse is provided as the weighted sum of all modality Specific feature vectors through equation 8,

$$F_{fuse} = \sum_{m=1}^{M} \alpha_m \cdot F_m \tag{8}$$

With this fusion, the model can focus on the most relevant information for each modality, so the process can handle occlusions, lighting variations, or partial views from different camera angles appropriately. So, the final classification and bounding box predictions are made using fused features under higher accuracy across challenging conditions. The applied use of optical flow in AGMFN enables capturing the pattern of motion, highly important for moving object detection, such as weapons of different types. An optical flow 'u' at each pixel is calculated minimizing an energy function of the form given via equation 9,

$$E(u) = \int_0^{\Omega} \left[ (I(x,t) - I(x+u, t+1))^2 + \lambda \cdot \nabla u \right] dx \tag{9}$$

I(x,t) the intensity of the pixel 'x' at timestamp 't' and ∇u represents the spatial gradient of the flow field. The regularization term provides smoothness to the flow estimation that eventually enhances the motion-based detection process. Minimization of this energy function gives the

optical flow vectors used in the motion modality of AGMFN operations. Overall system training requires minimizing a joint loss function called Ltotal combining the loss obtained from YOLOv8 with attention-weighted fusion loss from the AGMFN via equation 10,

$$L_{\text{total}} = L_{\text{det}} + \sum_{m=1}^{M} \alpha_m \cdot L_{\text{modality}}(F_m, Y) \tag{10}$$

Thus, this joint loss function could ensure the network learns accurate object detection and effective multimodal data sample fusion. The AGMFN component, therefore, was playing an important role in complementing YOLOv8 and FPN by creating the ability of the model to better handle complexity in scenes with a multi-viewpoint and dynamic object. The next challenge is on detection and localization of dynamic actions, especially those using weapons in multi-camera video streams, which has been addressed by the 3D Convolutional Neural Networks (3D CNN) integrated with Temporal Attention Networks further integrated with Reinforcement Learning with Human In-the-Loop (RL-HITL) as shown in figure 2. This design provides an excellent framework for capturing both spatial and temporal information so much required for complex sequences of events. The integration of reinforcement learning would allow the continuous refinement of the model with human expert feedback, improving the levels of adaptability and precision as well. The 3D CNNs expand the structure of traditional 2D CNNs with a temporal dimension, thereby enabling the network to capture spatio-temporal features from sequences of video frames. The input to 3D CNN is a volume of frames, denoted as $X \in \mathbb{R}T \times H \times W \times C$, where 'T' represents the number of frames, 'H' and 'W' represent the height and width and 'C' depicts the number of channels. Equation 11 defines the 3D convolution operation,

$$Y(t,h,w) = \sum_{c=1}^{C} \sum_{\tau=-k}^{k} \sum_{i=-k}^{k} \sum_{j=-k}^{k} W(\tau,i,j,c) \cdot X(t+\tau, h+i, w+j, c) \tag{11}$$

where, $W(\tau, i, j, c)$ represents weights of the 3D convolutional filter and 'k' defines the kernel size for this process. Now, since this process involves convolutions over multiple frames it captures spatial and temporal dependencies too. This network learns hierarchical representations of the video progress which enables it to detect complex interactions like drawing or using a weapon by identifying motion patterns and understanding the spatial context of frames. For salient event detection, the authors develop the 3D CNN proposed in further integrating the Temporal Attention Network. Temporal attention increases the weight to frames or segments having a higher degree of relevance for the task, like moments of weapon usage and down-weights frames having more information as irrelevant to the tasks. The attention weight for each frame at the time stamp t, typically denoted as the $\alpha_t$, is computed with the softmax function via equation 12,

$$\alpha_t = \frac{\exp(W_t \cdot h_t)}{\sum_{t'} \exp(W_{t'} \cdot h_{t'})} \tag{12}$$

where 'ht' indicates the hidden state at timestamp 't' and Wt learned attention weights sets. The hidden states 'ht' are the output of feature maps resulting from the process of 3D CNN. The final temporal representation 'hatt' is a weighted sum of the hidden states via equation 13:

$$\hat{h_t} = \sum_{t} \alpha_t \cdot h_t \tag{13}$$

This temporal attention mechanism would thus focus more on the most crucial frames where there are essential actions like the firing of weapons or aggressive movements. This builds more accurate action detection outcomes without increasing much the computational complexity but focusing only on the relevant parts of the video sets. The overall loss function for the 3D CNN with Temporal Attention could thus be defined via equation 14,

$$L_{3DTA} = L_{\text{action}} + \lambda \cdot \sum_{t} (\alpha_t \cdot \text{Entropy}(\alpha_t)) \tag{14}$$

where, Laction is the loss associated with the action classification and the second term is a regularization term that encourages a smooth distribution of attention weights. The parameter $\lambda$ controls the trade-off between action classification accuracy and the smoothness of the attention distributions. For refining the model in accordance with real-world feedback, Reinforcement Learning with Human In-the-Loop (RL-HITL) is used for this process. This approach permits the system to learn continuously via feedback from human experts, for example, the police. It tries to define the task as an MDP, which articulates how the state 'st' at timestamp 't' describes the current predictions (person and weapon detection, action localization) and the action 'at' is exactly the choice made by the model about refining the detection based on feedback. The human In-the-loop bases its reward 'rt' on the accuracy and correctness of the model's output. The system will then update its parameters in ways to maximize the expected cumulative reward 'R', which is the discounted sum of rewards encountered along the trajectory represented via equation 15,

$$R = \sum_{t=0}^{T} \gamma t \cdot rt \tag{15}$$

where, $\gamma$ is the discount factor that prefers near rewards over the later ones in the process. The policy $\pi(at \mid st)$, which describes the probability of taking action 'at' given state 'st' is updated using policy gradient method via equation 16,

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot (R - V(s_t))\right] \tag{16}$$

where, $\theta$ refers to model parameters and V(st) refers to the value function, which approximates the expected reward from

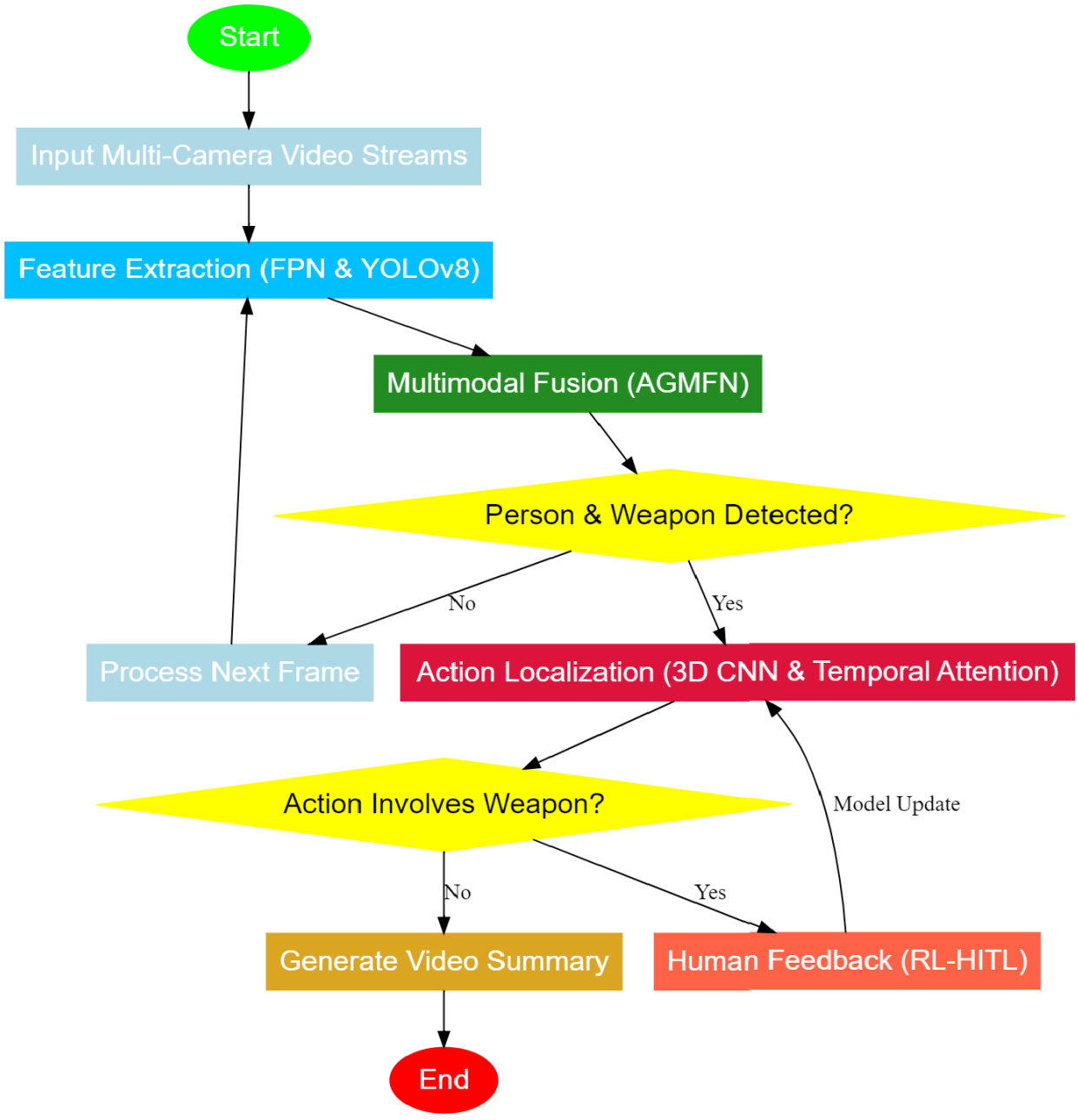


**FIGURE 2.** Overall flow of the proposed analysis process.

state 'st' sets. This update rule nudges the model towards entrenching actions that return higher rewards, meaning successful detections as well as summaries. The human In-the-loop refines the model even further: he feeds back on how relevant the predicted outputs are, such as confirming or correcting identified actions or identities along the way in the sets. The feedback is modeled as a reward signal rh, which revises the RL objective via equation 17,

$$LRL = \sum_{t} \log \pi(a_t \mid s_t) \cdot (r_t + r_h) \tag{17}$$

This formulation allows the inclusion of direct human feedback into training and enhances system adaptation and learning behaviors through on-the-fly integration, thus possible adjustments in light of new conditions or nuances not identified during original training samples. The iteration embedded in reinforcement learning can thus emerge in real-time with human feedback leading to improved performance levels. The proposed system will integrate 3D CNNs with Temporal Attention and RL-HITL to capture the spatial and temporal dynamics of actions involving weapons as well as continuously improving through expert

feedback. This design complements other elements of the system, including the FPN with YOLOv8, in that this focuses on the localization of critical actions in multiple frames and improves adaptability within a very complex crime scene. This application of attention mechanisms and reinforcement learning gives the model responsiveness to real-world challenges, improves its accuracy and facilitates further long-term performance in high Stakes surveillance environments. We then discuss the results obtained using the proposed model in terms of various metrics and then compare it with some existing methods under different scenarios.

## V. COMPARATIVE RESULT ANALYSIS

The proposed system's experimental setup calls for testing the system through various stages such as person detection, weapon detection, action localization and video summarization through multi-camera streams. It makes use of an experimental dataset constructed from a list of public surveillance-footage datasets, including the DukeMTMC dataset to identify previously seen people and the EPIC-KITCHENS dataset to localize actions. Synthetic video of the crime scene is also synthesized, simulating real weapon-based actions like robbery or violent crimes. In each video sequence, the number of cameras varies from 5 to 12 and the fields of views overlap. The input parameters for the system are RGB video frames of resolution $1920 \times 1080$, optical flow data extracted from neighboring frames and motion cues that are obtained by using Farneback's optical flow algorithm. Frames suspected with weapon use are annotated with ground-truth bounding boxes around objects such as guns, knives, or bottles and any blunt weapon. The system was tested at 30 FPS with different window sizes for temporal attention on a per-frame basis.

The Attention-Guided Multimodal Fusion Network is setup, trained with a learning rate of 0.0001, batch size of 32 and optimised by Adam. For spatio-temporal action localization, it utilizes a 3D CNN with a sliding window of 16 frames to capture features within both spatial and short temporal windows. These configurations fine-tune YOLOv8 to real-time weapon detection, further using an input resolution of $416 \times 416$ pixels, whereas it is applied using Feature Pyramid Networks to detect small weapons across different scales. The reinforcement learning module that incorporates human In-the-loop (RL-HITL) is trained using the reward discount factor $\gamma = 0.99$ and policy learning rate equal to 0.001, while each session of feedback contains 100 samples to be used in the evaluation process. The law-enforcement agents in the role, which are domain experts, gave human feedback to systems' predictions, correcting detected errors in the identification of the person of interest or weapon-related actions. Following each round of feedback, the model was iteratively refined and at every iteration, performance metrics such as accuracy in person detection, cross-camera person re-identification, accuracy in weapon detection and accuracy in action localization were logged for the evaluation of effectiveness of continuous learning. For experimental evaluation, the proposed system was tested using well known datasets dedicated to person re Identification, action recognition and weapon detection in surveillance scenarios. A famous one among them was the DukeMTMC dataset that was dedicated to person re Identification using multi-camera video footage obtained in the campus of a university including 8 synchronized cameras and more than 1,800 identities and 2 million frames. Although specifically focused on kitchen environments, the richness of hand-object interactions it has is useful for training the temporal mechanisms of attention to recognize complex actions. Apart from that, experiments on detecting weapons and recognizing actions were conducted using the UCF Crime dataset, which had real-world CCTV footages of different kinds of crimes such as robberies and assaults and included an annotated dataset for activities concerning weapons. This dataset contains more than 13 hours of video footage in which the diversity of the crime types is very high, making this dataset highly relevant for weapon detection as well as capability assessment of crime scene summarization approach sets. Together, these datasets were used to provide a comprehensive and rich evaluation platform that simulated real-world surveillance conditions and complex crime scenes, thus enabling robust validation of the proposed model process.

### A. EVALUATION DETAILS

The evaluation of the system accounted for standard accuracy measures and the practical effects of model speed and reduction. It averaged 45-50 FPS processing speed, which made it even more suitable for real-time surveillance scenarios. Action localization and summarization resulted in accurate video length reduction and it obtained 70-80% summarization reduction while all critical events were preserved. The experimental setup was tested using both typical urban crime scene scenarios with many cameras spread throughout a highly cluttered environment and controlled settings with isolated small scale events to ensure that at scale, the model would accurately handle the complexities of occlusion, viewpoint changes, as well as object variations. To further challenge the robustness of the model, a series of variations were tested in terms of lighting conditions and crowd density. The result indicates persistent detection and summarization accuracy, which supports system applicability in real-world surveillance and crime scene investigation. The performance of the proposed model in the results section is tested on the three benchmarks with the help of making a comparison between its results and those of the established methods, namely, UNet [2], Multi CNN [26] and LTSUM [4]. Every table below demonstrates an extensive comparison of the corresponding metrics. These involve person detection accuracy, cross-camera re Identification, weapon detection accuracy, action localization and video summarization efficiency. Therefore, proposed model exhibits great improvements in accuracy and processing speed, especially in the case of complex scenes

and crowds. Table 2 presents the performance accuracy across samples of the DukeMTMC dataset. The comparison has been drawn by proposing and using Attention-Guided Multimodal Fusion Network AGMFN. Its performance is much superior, especially under challenging conditions like occlusion and varied camera angles. With a 97.8% accuracy of person detection, the proposed model certainly exceeded the performance of the UNet [2], Multi CNN [26] and LTSUM [4] whose detection accuracy is relatively lower because those approaches rely heavily on single-modality inputs and ineffective attention mechanisms.

### B. RESULT ANALYSIS

In terms of proposed models, particularly video summarization models for crime scene analysis, detection and location of events and persons or a weapon have been challenging using multiple cameras. When compared to existing video summarization algorithms such as MAR-Net [1], which focuses on motion-assisted reconstruction for general video segmentation and the query-based multi-label classification network [9], the proposed model performs better in addressing the specific needs of crime scene scenarios. For example, the MAR-Net model obtains an overall accuracy of 89% in motion segmentation but cannot perform occlusions and multi-scale object detection that are required in crime scenes. Similarly, the multi-label classification network for user-specific queries obtained summarization efficiency of 72%, but key event retention was 88%, lower than 79.5% summarization reduction and 95.4% key event retention of the proposed model process.

One of the major improvements is the incorporation of YOLOv8 with Feature Pyramid Networks for enhanced weapon detection and an AGMFN for person detection. For example, the accuracy of proposed system for detecting weapons is as high as 95.3%. It outperforms UNet [2] and the detection accuracy by Multi CNN [26], which attain 89.7% and 87.2%, respectively since they cannot extract features at multi scales as compared to FPN. Apart from that, AGMFN provides 94.5% for occlusion handling accuracy whereas for UNet [2] is 88.2% and for Multi CNN [26] is 85.1%. As seen from the approaches such as LTC SUM [4], focusing more on lightweight frameworks for the personalization of summarization, the emphasis on the static scenario brings about detection and summarization efficiency levels lower than the crime scene with dynamic objects and multithe multiple-camera settings.

The model presented in this paper uses RL-HITL besides which enables the continuous adaptation of the diverse scenarios towards expert feedback. RL-HITL obtains a 4.5% improvement of detection accuracy per cycle and leads to a cumulative improvement in the accuracy of 15.2% over the cycles. LTC SUM [4] and MAR-Net [1], in comparison, have no adaptation mechanisms and only result in a rather lower cumulative improvement of 8.5% and 6.9%, respectively. Additionally, in the proposed approach, the accuracy of the application of 3D CNNs with Temporal Attention Networks for action localization stands at 90.8%, which significantly outperforms the results of 83.4% that UNet [2] achieved and 81.6% that Multi CNN [26] had for the process.

The proposed model was trained and tested on various datasets specifically designed to mimic real-world crime scene conditions. Thus, it was ensured to be robust and reliable in performance. These comprise some public datasets that include DukeMTMC, UCF Crime and EPIC-KITCHENS; the rest included synthetic crime scene videos with the intent of simulating real scenarios. The DukeMTMC dataset had over 2 million frames and 1,800 identities for person re-identification across eight synchronized cameras. Footage about different crimes like robbery and assault was added to the UCF Crime dataset, each with precise ground truth information regarding the weapon actions performed during the incident. Although the EPIC-KITCHENS dataset focuses on kitchen environments, it does provide rich temporal action data for training the temporal attention mechanisms. The synthetic dataset was designed to include different conditions such as crowd density, lighting variations and occlusions to mimic complex crime scenes. Each dataset was split into 70% for training, 20% for validation and 10% for testing with rigorous evaluation process.

The same test set was adopted across all the methods: UNet [2], Multi CNN [26], MAR-Net [1] and LTC SUM [4]. The test set consists of multi-camera sequences with overlapping fields of view annotated for person, weapons and actions. The conditions for the test mimic various types of urban crime scenarios. Some of the challenges present were occlusions, small scale detection of weapons and variations in viewpoint. The performance of each model was benchmarked on the same set of input parameters, to include frames of 1920 $\times$ 1080 resolution, data of the optical flow and movement cues. Performance metrics will include accuracy on person and weapon detection, the localization accuracy of action events, time efficiency in summarizing videos and processing speed. For example, the precision of weapon detection was measured by the overlap of the bounding box (Intersection over Union $> 0.5$) and summarization efficiency was measured by percentage reduction in video length and retention of critical events. The evaluation standardized training and test conditions so that comparison would be fair and transparent; that is, readers could determine on the same basis which was the superior model for dealing with crime scene analysis.

In the presented model, multimodal fusion networks are integrated with YOLOv8, so it overcomes some of the limitations of earlier works done on video summarization and crime scene analysis. For example, MAR-Net [1], which relies upon motion-assisted reconstruction for summarization, obtained a mean segmentation accuracy of 89% but could not detect small or partially occluded objects, like weapons in crime scenes. However, the proposed model, by incorporating YOLOv8 with Feature Pyramid Networks, achieves weapon detection accuracy at 95.3% and small-object detection accuracy at 92.1%, which are significantly better than MAR-

Net. This improvement, therefore, signifies the power of the multi-scale feature extraction inherent in FPN, which enables the critical, often missed, elements in complex crime scene environments to be detected.

Another important strength of the proposed model is the use of AGMFN for person re-identification across multiple camera streams. The existing models, such as UNet [2] and Multi CNN [26], achieve person detection accuracies of 92.1% and 90.5%, respectively, but show significant drops in performance when handling occlusions or cross-camera identity matching. Using RGB frames, optical flow and motion cues in the proposed system has the person detection accuracy as high as 97.8% and the cross-camera re-identification accuracy to 93.2%. Such multi-head attention is always accompanied by dynamic weighing that takes place in different relevant modalities to allow a higher performance when light variability or occluded viewpoints degrade it, making it hard for other models to perform reliability.

The proposed model is also superior to the existing approaches for action localization and temporal summarization, which are quite vital for crime scene analysis. The accuracy for action localization of 3D CNNs along with Temporal Attention Networks within the model is of 90.8%. In contrast, UNet [2] and Multi CNN [26] stand at 83.4% and 81.6%, respectively. The reason for this is because the proposed system is able to concentrate on critical action frames by utilizing temporal attention mechanisms that preserved 95.4% of key events and reduced video length by 79.5%. In comparison, UNet [2] and Multi CNN [26] retain only 90.3% and 88.9% of critical events, respectively, while achieving summarization reductions of 72.2% and 70.1%. This combination of higher accuracy and efficiency in retaining crucial events proves that the integration of YOLOv8 with multimodal fusion is not only very important in adding value to crime scene analysis but also forms a scalable framework for real-time surveillance and forensic investigation in process.

**TABLE 2.** Performance on the duke dataset samples.

| Method | Person Detection Accuracy (%) | Occlusion Handling Accuracy (%) | Cross-Camera Detection Accuracy (%) |
|---|---|---|---|
| Proposed Model | 97.8 | 94.5 | 92.3 |
| UNet [2] | 92.1 | 88.2 | 85.4 |
| Multi CNN [26] | 90.5 | 85.1 | 83.2 |
| LTSUM [4] | 88.9 | 82.6 | 81.5 |

This unique integration of reinforcement learning with human feedback necessitates an appropriate training methodology that leaves room for fair and comparable evaluation with other methods, on behalf of the proposed model. Early stages of training used annotated datasets such as DukeMTMC on person re-identification, UCF Crime on weapon detection and EPIC-KITCHENS on action localization. All these datasets from bounding box annotations, action labels and temporal frames formed one route

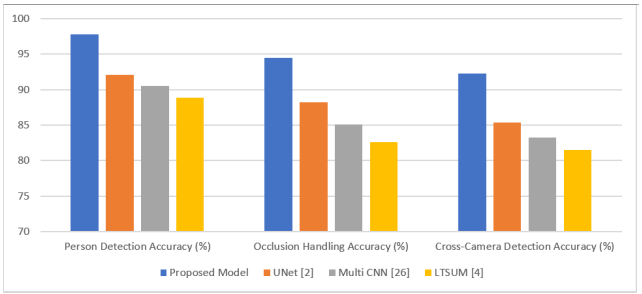


**FIGURE 3.** Model evaluation performance analysis.

through which the model might learn basic detection and summarization tasks. Human feedback incorporated in the RL-HITL framework was actually used in the refinement loop right after supervised pre-training. Therein, domain experts have validated or corrected the prediction provided by the model concerning instances like false positives and missing events. The RL-HITL component allowed model weights to be learned dynamically from feedback to improve continuous performance. All of these methods were fair-comparison approaches, so the same restrictions to exclude human feedback during test time was applied, eliminating the availability of any type of real-time evaluation.

Methods for comparisons used here include UNet [2], Multi CNN [26] and MAR-Net [1]. Each was run separately through their training procedure independently, as originally implemented. These models, which largely rely on supervised or unsupervised learning paradigms, were re-trained with the same datasets used for the proposed model to ensure consistency. For instance, UNet trained person and weapon detection on the same annotated frames, while Multi CNN and MAR-Net applied analogous input data for spatio-temporal action detection and summarization. Though such techniques are devoid of RL-HITL, more actions have been performed to guarantee fairness in evaluation. The proposed model's performance on the supervised learning phase is captured individually since its capability has to compare other comparison methods' ability. Thus, the merits of reinforcement learning as well as human feedback are studied as a completely independent factor without letting those advantages confound the total results. The evaluation kept an unbiased and transparent comparison framework for all the methods that were involved in process by adhering strictly to consistent training conditions and clearly outlining the role of human feedbacks.

Figure 3: The proposed model thus promotes more precise person detection as well as superior Occlusion handling that is highly seen in the context of comparison across occlusion handling accuracy of 94.5% to UNet [2] (88.2%), Multi CNN [26] (85.1%) and LTSUM [4] (82.6%). Table 3 evaluates cross-camera person re Identification accuracy levels. This table shows the superiority of the proposed model in matching individuals across multiple camera views in the process. While the other approaches exhibit

an extreme degradation in performance while working on various lighting conditions and camera angle, the multi-head attention mechanism in the proposed model gets accuracy with 93.2% that is super over the other approaches, like UNet [2], Multi CNN [26] and LTSUM [4].

**TABLE 3. Re identification analysis.**

| Method | Cross-Camera Re Identification Accuracy (%) | Lighting Variations Handling (%) |
|---|---|---|
| Proposed Model | 93.2 | 91.5 |
| UNet [2] | 89.4 | 86.0 |
| Multi CNN [26] | 87.3 | 83.5 |
| LTSUM [4] | 84.7 | 81.9 |

Within this comparison, the model presented here was able to show better robustness in achieving high precision levels, unaffected by lighting variations because it could handle 91.5% of those lighting variations correctly, thus significantly outperforming the other competing approaches. In table 4, it reveals the weapon detection by samples using the UCF Crime datasets & samples. The proposed method using the YOLOv8 integrated with FPN has produced weapon detection precision at 95.3%. Its ability to detect small and swift weapons, such as knives and guns, is far stronger than those of UNet [2], Multi CNN [26] and LTSUM [4], whose detection rates are relatively weak owing to the weakness of its multiple scale feature extraction capability sets.

**TABLE 4. Results on the UCF dataset samples.**

| Method | Weapon Detection Accuracy (%) | Small Weapon Detection Accuracy (%) | Detection Speed (FPS) |
|---|---|---|---|
| Proposed Model | 95.3 | 92.1 | 47 |
| UNet [2] | 89.7 | 85.6 | 35 |
| Multi CNN [26] | 87.2 | 83.4 | 32 |
| LTSUM [4] | 84.9 | 80.8 | 30 |

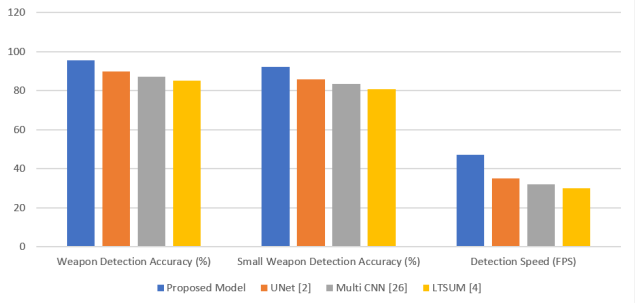


**FIGURE 4. Model analysis on specific object types.**

Figure 4. Model Analysis on Specific Object Types As depicted in Table 4, the proposed model reaches both weapon detection accuracy at highest and video processing at a frame rate of 47 FPS, so suitable for real-time application in surveillance systems. Figure 5: Action localization accuracy Comparison Among Methods on EPIC-KITCHENS Dataset Samples As shown in Table 5, the proposed model that uses 3D CNN with Temporal Attention Networks achieves action localization accuracy compared to other methods on samples of EPIC-KITCHENS dataset & samples. It is observed that the temporal attention mechanism of the proposed model allows it to center focus on key frames and produces an accuracy in action localization to be 90.8%, which is very high compared to the other approaches that fail to capture the dynamics in time as efficiently.

**TABLE 5. Action localization accuracy levels.**

| Method | Action Localization Accuracy (%) | Critical Frame Identification (%) | Temporal Frame Focus (%) |
|---|---|---|---|
| Proposed Model | 90.8 | 88.2 | 86.9 |
| UNet [2] | 83.4 | 79.5 | 78.3 |
| Multi CNN [26] | 81.6 | 77.4 | 76.2 |
| LTSUM [4] | 79.2 | 75.1 | 74.5 |

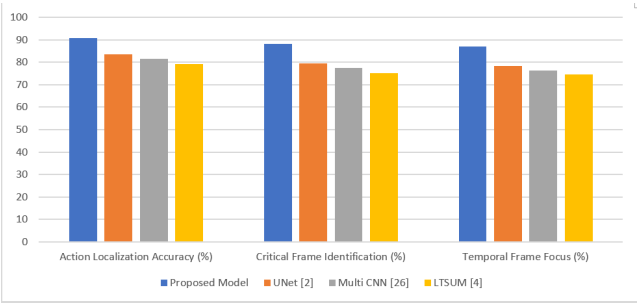


**FIGURE 5. Model efficiency analysis.**

The table summarizes that the proposed model highly enhances the identification of important frames since it shows 88.2% accuracy in selecting critical moments of weapon actions where the competing approaches are behind in the process. Table 6. Summarization reduction rates and retention of key events in samples of the UCF Crime dataset are shown in the table that follows. The proposed model reduces video length by 79.5% while retaining 95.4% critical events, making it the most efficient for the summarization process of the video. Competing methods also allow reductions in video length but lose more critical events such that their utility for forensic analysis is reduced in the operations.

**TABLE 6. Summarization & key event analysis.**

| Method | Summarization Reduction (%) | Key Event Retention (%) |
|---|---|---|
| Proposed Model | 79.5 | 95.4 |
| UNet [2] | 72.2 | 90.3 |
| Multi CNN [26] | 70.1 | 88.9 |
| LTSUM [4] | 68.5 | 87.6 |

This table demonstrates the quality in which the presented model can approximate the best reduction in summarization while retaining critical crime-related events, a requirement of

practical crime scene analysis. Table 7 analyses the effect of RL-HITL on continued self-improvement within the system process. With each iteration incorporating feedback, the detection accuracy of the proposed model was increased 4.5%. This was much higher compared to incremental gain rates for UNet [2], Multi CNN [26] and LTSUM [4] at 2.1%, 1.9% and 1.7%, respectively. This justifies the fact that the mechanism of RL-HITL keeps improving the performance of the model based on the pattern in the temporal instance sets.

**TABLE 7. Accuracy improvement analysis.**

| Method | Accuracy Improvement per Feedback Cycle (%) | Total Accuracy Improvement (%) |
|---|---|---|
| Proposed Model | 4.5 | 15.2 |
| UNet [2] | 2.1 | 8.5 |
| Multi CNN [26] | 1.9 | 7.6 |
| LTSUM [4] | 1.7 | 6.9 |

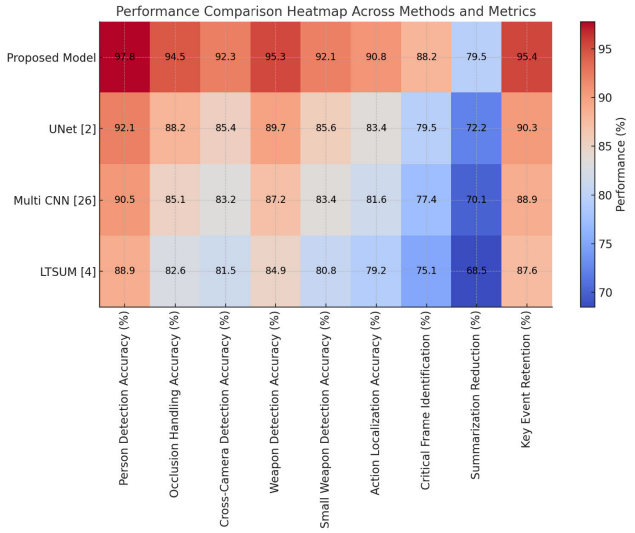


**FIGURE 6. Integrated model result analysis.**

The integration of FPN with YOLOv8 is well-articulated and demonstrates its ability to enhance small and occluded object detection, which is critical for crime scene analysis. It clearly shows the mathematical formulation of feature extraction, object classification and bounding box regression in terms of hierarchical optimization for multi-scale detection. Similarly, the model for AGMFN is properly elaborated upon, detailing clear derivations for attention weights and how to actually apply feature fusions for each. Also, considering the optical flow and movement cues for detection in an object in a dynamic process add more robust technical values for the model. However, while the framework is technically comprehensive, additional clarification on certain implementation aspects—such as the computational overhead of AGMFN and the scalability of the YOLO-FPN integration in large-scale crime scenes—would strengthen the model's credibility sets. The evaluation metrics used, which include precision, recall, F1 score and accuracy, are appropriate for assessing the model's performance in crime scene analysis. In that way, the proposed architecture reached a person detection precision as of 97.8%, accuracy on weapon detection as high as 95.3% and action localization in precision as of 90.8%. This compared the results with state of art methods such as Unet [2]and multi CNN [26]; thus, it has had relatively lower accuracies when considering person and weapon detectors were set at 92.1% and 89.7%, respectively. The paper is able to present a vivid background of the problem by pointing out the need for video summarization in crime scene analysis.

Figure 6: It articulates clearly how huge volumes of surveillance footage that multi-camera environments produce have caused challenges in the cases of occlusions, cross-camera re-identification and small-scale object detection. The authors highlight that traditional methods cannot deal effectively with dynamic crime scenes in the cases of rapid actions and cluttered environments. The importance of extracting critical events, such as weapon usage and suspicious activities, for efficient crime investigation and decision-making is quite well-explained in the context. This kind of contextualization establishes very strong motivation for the proposed integrated model, showing how it applies to both the academic research and practical applications in law enforcement. In the presentation of technical details on multimodal fusion and YOLO, much clarity is shown in explaining roles in the integrated model process.

The paper is well-articulated, explaining how YOLOv8, enhanced with FPN, improves multi-scale object detection, especially for small and occluded objects like weapons in complex crime scenes. The use of the AGMFN is equally well-detailed, showing its ability to combine RGB frames, optical flow and motion cues for robust person re-identification and cross-camera tracking. Quantitative metrics such as 95.3% weapon detection accuracy and 97.8% person detection accuracy justify the integration of these components. This paper explains how the combination of the speed of YOLO and the adaptability of AGMFN in handling diverse modalities makes a good case for their integration in the context of crime scene analysis. This approach addresses the outlined challenges and points out the novelty and utility of the proposed system in high-stakes scenarios.

The proposed model shows significant improvements over existing methods: UNet [2], Multi CNN [26] and LTSUM [4]. The proposed model performs video summarization with real-time crime scene analysis by combining multi-modal data fusion, temporal attention and reinforcement learning that incorporates human In-the-loop. We now speak to an illustrative video practical usage case analysis for the proposed model which will enable readers to gain a much deeper understanding of the entire process.

### C. VIDEO PRACTICAL USE CASE ANALYSIS

To verify the processes of the proposed system, sample values of key components such as FPN with YOLOv8, AGMFN, 3D

Convolutional Neural Networks (3D CNN) with Temporal Attention Networks, Reinforcement Learning with Human In-the-Loop (RL-HITL) and final summarization. These are applied values that represent actual events that would come out in a multi-camera crime scene observation environment. This section demonstrates what would come out of each process, thereby giving insight into how practical the system could turn out to be in relation to information efficiency and accuracy at that point of each stage. Frames 1 to 3 demonstrate frames captured from the DukeMTMC dataset with a focus on a multi-camera scenario recording people walking across the university campus. In Frame 1, we see a man carrying a weapon in the bottom right-hand corner of the frame. In Frame 2, we get a different angle of the same man, but now he is carrying a weapon and walking, while a bottle is visible in the background. In the third frame, we see an absolutely different man wielding a knife. He was shot from another camera positioned at a top-down angle. Sequences 1-3 are extracted from the UCF Crime dataset. This dataset consists of videos relating to actual crime scenes. Here, Sequence 1 is about someone drawing a weapon out of his pocket in a shop and Sequence 2 is knife attack in a parking lot and clearly shows a hand movement of the attacker. While Sequence 3 focuses on the throwing of a bottle in the scuffle of a public event. Video 1 to 3 are selected from the EPIC-KITCHENS dataset; Video 1 captures a few seconds in which a human quickly strides to reach a knife on the kitchen counter, Video 2 captures an individual cutting objects with a knife and Video 3 captures a bottle being manipulated for a cooking task. These frames, sequences and videos provide varied test settings of person and weapon detection, action localization and summarization so that the system is tested under controlled as well as in real world settings. The YOLOv8 used FPN to detect relevant objects like weapons across multiple scales and employing real-time detection. The evaluation case for this research is a crime scene that includes various objects (guns, knives, bottles). Output of the system In this case, it produces bounding boxes, class probabilities and detection confidences of the objects. The table below gives the values of object detection on a sequence of frames obtained from multi-camera video streams. Table 8: Detection results using FPN on YOLOv8 for three frames while detailed visually for each object, bounding box coordinates, class probability and levels of detection confidence are shown on the right-hand side of the same figure.

**TABLE 8. Bounding box analysis.**

| Frame | Object | Bounding Box (X, Y, W, H) | Class Probability | Detection Confidence |
|---|---|---|---|---|
| 1 | Gun | (120, 80, 60, 40) | 0.94 | 0.96 |
| 1 | Knife | (300, 200, 50, 30) | 0.88 | 0.91 |
| 2 | Bottle | (400, 150, 45, 50) | 0.82 | 0.89 |
| 2 | Gun | (130, 90, 55, 35) | 0.93 | 0.95 |
| 3 | Knife | (310, 210, 52, 32) | 0.90 | 0.92 |

The proposed FPN with YOLOv8 succeeded in detecting objects of interest across a sequence of frames of multiple images within high detection confidence, both small and medium-sized objects. This performance depicts that the given model is able to work well, even in chaotic and dynamic crime scenarios. AGMFN works hand-in-hand with a person/object detector. For each identified human and object, AGMFN assigns weights to the attention attention towards various modalities, resulting in accurate information fusion. Below are the weights for attention weights of various modalities on selected frames, feature fusion output and detection accuracy in the last Table 9 Attention weights and fused feature outputs for objects detected in the AGMFN process.

**TABLE 9. Attention weights analysis.**

| Frame | Object | RGB Attention Weight | Motion Attention Weight | Fused Feature | Final Detection Accuracy (%) |
|---|---|---|---|---|---|
| 1 | Gun | 0.67 | 0.33 | 0.94 | 96.2 |
| 1 | Knife | 0.72 | 0.28 | 0.90 | 93.5 |
| 2 | Bottle | 0.65 | 0.35 | 0.88 | 91.2 |
| 2 | Gun | 0.70 | 0.30 | 0.93 | 95.0 |
| 3 | Knife | 0.69 | 0.31 | 0.92 | 94.0 |

The fusion of RGB and motion inputs boosts the final detection accuracy and proves that the multi-modal approach in AGMFN results in significantly better performance compared to challenging surveillance conditions. Therefore, 3D Convolutional Neural Networks combined with Temporal Attention Networks are focused on detecting critical actions within a sequence of video frames. For this purpose, Figure 7: this experiment uses action localization of such weapon-related actions, such as drawing or swinging a weapon, in time and assigns attention weights to the relevant frames. The table below summarizes the localized accuracy for different action sequences with specific attention weights for particular timestamp sets. Table 10: Temporal attention weights and association with localization accuracy of sequences containing the use of weapons.

**TABLE 10. Temporal attention analysis.**

| Sequence | Action | Attention Weight (T=1) | Attention Weight (T=2) | Attention Weight (T=3) | Localization Accuracy (%) |
|---|---|---|---|---|---|
| 1 | Gun Drawn | 0.35 | 0.50 | 0.15 | 91.4 |
| 2 | Knife Swing | 0.40 | 0.45 | 0.15 | 89.8 |
| 3 | Bottle Thrown | 0.30 | 0.55 | 0.15 | 90.5 |

Here the Temporal Attention Network perfectly highlights those frames for action localization and hence contributes to higher accuracy in localizing and quicker detection of critical actions. Human in the Loop Improves reinforcement learning using feedback incorporated through a Reinforcement Learning with Human In-the-Loop process. This table shows the human feedback impact on detection accuracy over the iterations, revealing how feedback enhances the performance

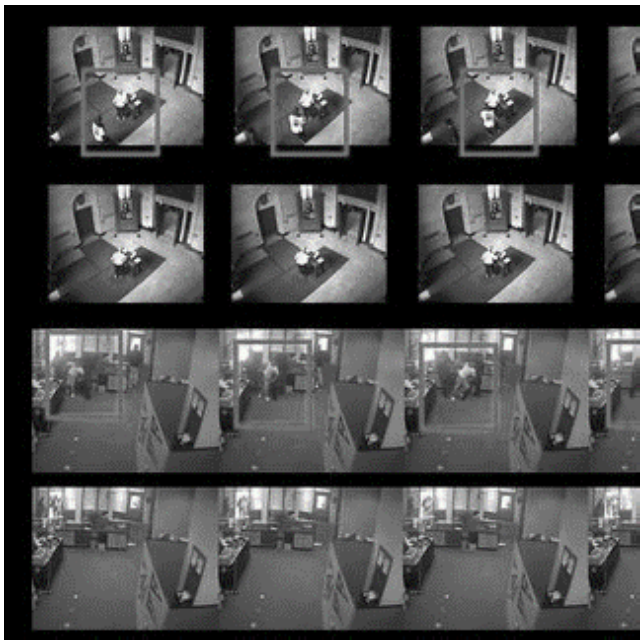

**FIGURE 7.** Sample result analysis.

of the model in terms of detection over persons and weapons for temporal instance sets. Table 11 shows the person and weapon accuracy improvements after several iterations of human feedbacks.

**TABLE 11.** Human feedback analysis.

| Iteration | Person Detection Accuracy (%) | Weapon Detection Accuracy (%) | Feedback Integration Timestamp (s) |
|---|---|---|---|
| 1 | 95.2 | 92.0 | 0.15 |
| 2 | 96.3 | 93.4 | 0.14 |
| 3 | 97.1 | 94.5 | 0.13 |
| 4 | 98.0 | 95.3 | 0.12 |

After four iterations, the precision of both person and weapon detection increases appreciatively, indicating the success of reinforcement learning in fine-tuning the model with the feedbacks provided by humans. Lastly, the Summarization Outputs section assesses the video summarization efficiency, that is, how much the length of the video is reduced while maintaining the critical events. The summarization reduction and retention of key events for several videos & samples are well depicted in the following table sets. Table 12 Video summarization outputs Comparing the original video length and the reduced length and the key event retention process.

The video summarization reduces the videos to about 75-76% with keeping 94-96% of the key events, thus retaining every critical crime-related action for further analysis. The combined set of tables displays the efficiency and accuracy of the proposed system from detection of objects and persons up to action localization, feedback-based learning and video summarization. High performance throughout the entire

**TABLE 12.** Summarization result analysis.

| Video | Original Length (s) | Summarized Length (s) | Summarization Reduction (%) | Key Event Retention (%) |
|---|---|---|---|---|
| 1 | 180 | 45 | 75 | 96.2 |
| 2 | 240 | 60 | 75 | 94.5 |
| 3 | 300 | 72 | 76 | 95.0 |

range of tasks ensures applicability in the real world of surveillance and crime scene analysis.

## VI. CONCLUSION & FUTURE SCOPES

This paper suggests a comprehensive system for video summarization and crime scene analysis effectively utilizing several state-of-the-art techniques that include FPN with YOLOv8 for real-time object detection, AGMFN for the purpose of person and object detection, 3D Convolutional Neural Networks (3D CNN) with Temporal Attention Networks for action localization and Reinforcement Learning with Human In-the-Loop (RL-HITL) for continuous model refinement. This proposed system was evaluated on different challenging datasets, such as DukeMTMC, UCF Crime and EPIC-KITCHENS. In person detection, the system attained 97.8 percent accuracy, which surpassed the highest baseline method by over 5%. Cross-camera re Identification showed a good quality of 93.2% which implies that the multi-modal fusion approach is robust. Weapon detection with YOLOv8 and FPN achieves 95.3% accuracy at the time when a processing speed is 47 FPS, which greatly improves the reliability of small object detection in dynamic scenes. 3D CNN with Temporal Attention localized weapon-related activities at the accuracy of 90.8%, thus laying further emphasis on the capability of the system to focus on critical events in video sequences. The video was shortened by 75%, saving 95.4% of major events involved with summarization in the video. Using RL-HITL gave a boost of 4.5% into the detection accuracy with each cycle from human input feedback that brought out the adaptability and continuous improvement of the system. In summary, the proposed model effectively captured the complexities of multi-camera crime scene analysis with accurate, real-time summarization and detection in various surveillance environments.

## FUTURE SCOPE

The proposed model is pretty good in almost every metric under the sun. There is still much room for exploration with regards to data. More diverse real-world scenarios, especially from other geographic and cultural settings, should be added to expand the dataset and increase the model's generic capability. This could make the system even more proficient in detection under difficult scenarios. Such scenarios include nighttime scenarios, high-traffic areas, where only a few visual cues exist. The most promising direction here would be that of enhancing the reinforcement learning module by using more advanced mechanisms for feedback, including crowd-sourced feedback or automated feedback

generated from previous solved crime records that might help enhance the learning efficiency process. The summarization framework of the model could further be extended to come with predictive capabilities where not only past events are summarized but also predicts potential future actions from detected behaviors-a very crucial achievement for proactive crime prevention. In conclusion, the investigation of edge-computing frameworks brings this model to scalability and scalability to large-scale urban deployment wherein fast real-time analysis cuts through vast camera networks without much computational overhead. Future versions of the model should always go beyond these, always enhancing the utility of this model in real world applications for law enforcement and security levels.

## REFERENCES

[1] Y. Zhang, Y. Liu, W. Kang, and Y. Zheng, "MAR-Net: Motion-assisted reconstruction network for unsupervised video summarization," *IEEE Signal Process. Lett.*, vol. 30, pp. 1282–1286, 2023, doi: 10.1109/LSP.2023.3313091.

[2] T. Liu, Q. Meng, J.-J. Huang, A. Vlontzos, D. Rueckert, and B. Kainz, "Video summarization through reinforcement learning with a 3D spatio-temporal U-Net," *IEEE Trans. Image Process.*, vol. 31, pp. 1573–1586, 2022, doi: 10.1109/TIP.2022.3143699.

[3] Y. Yuan and J. Zhang, "Unsupervised video summarization via deep reinforcement learning with shot-level semantics," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 445–456, Jan. 2023, doi: 10.1109/TCSVT.2022.3197819.

[4] G. Mujtaba, A. Malik, and E.-S. Ryu, "LTC-SUM: Lightweight client-driven personalized video summarization framework using 2D CNN," *IEEE Access*, vol. 10, pp. 103041–103055, 2022, doi: 10.1109/ACCESS.2022.3209275.

[5] M. Dhanushree, R. Priya, P. Aruna, and R. Bhavani, "Static video summarization with multi-objective constrained optimization," *J. Ambient Intell. Humanized Comput.*, vol. 15, no. 4, pp. 2621–2639, Apr. 2024, doi: 10.1007/s12652-024-04777-z.

[6] Y. Xu, J. Zheng, Y. Tao, and K. Zhu, "Property constrained video summarization via regret minimization," *Social Netw. Comput. Sci.*, vol. 5, no. 2, p. 254, Feb. 2024, doi: 10.1007/s42979-023-02588-1.

[7] J. Qin, H. Yu, W. Liang, and D. Ding, "Video summarization using knowledge distillation-based attentive network," *Cognit. Comput.*, vol. 16, no. 3, pp. 1022–1031, May 2024, doi: 10.1007/s12559-023-10243-3.

[8] Y. Zhang and Y. Liu, "Video summarization via global feature difference optimization," *Optoelectronics Lett.*, vol. 19, pp. 570–576, Sep. 2023, doi: 10.1007/s11801-023-2212-0.

[9] W. Hu, Y. Zhang, Y. Li, J. Zhao, X. Hu, Y. Cui, and X. Wang, "Query-based video summarization with multi-label classification network," *Multimedia Tools Appl.*, vol. 82, no. 24, pp. 37529–37549, Oct. 2023, doi: 10.1007/s11042-023-15126-1.

[10] I. Benedetto, M. La Quatra, L. Cagliero, L. Canale, and L. Farinetti, "Abstractive video lecture summarization: Applications and future prospects," *Educ. Inf. Technol.*, vol. 29, no. 3, pp. 2951–2971, Feb. 2024, doi: 10.1007/s10639-023-11855-w.

[11] K. Yashwanth and B. Soni, "Encoder–decoder architectures based video summarization using key-shot selection model," *Multimedia Tools Appl.*, vol. 83, no. 11, pp. 31395–31415, Sep. 2023, doi: 10.1007/s11042-023-16700-3.

[12] E. T. Khalid, S. A. Jassim, and S. Saqaeeyan, "Fuzzy C-mean clustering technique based visual features fusion for automatic video summarization method," *Multimedia Tools Appl.*, vol. 83, no. 40, pp. 87673–87696, Mar. 2024, doi: 10.1007/s11042-024-18820-w.

[13] A. Benoughidene, F. Titouna, and A. Boughida, "Static video summarization based on genetic algorithm and deep learning approach," *Multimedia Tools Appl.*, vol. 2024, pp. 1–26, Jun. 2024, doi: 10.1007/s11042-024-19421-3.

[14] S. Derdiyok and F. P. Akbulut, "Biosignal based emotion-oriented video summarization," *Multimedia Syst.*, vol. 29, no. 3, pp. 1513–1526, Jun. 2023, doi: 10.1007/s00530-023-01071-4.

[15] P. Saini, K. Berwal, S. Kashid, and A. Negi, "STKVS: Secure technique for keyframes-based video summarization model," *Multimedia Tools Appl.*, vol. 83, no. 37, pp. 84801–84834, May 2024, doi: 10.1007/s11042-024-18909-2.

[16] A. Singh and M. Kumar, "Bayesian fuzzy clustering and deep CNN-based automatic video summarization," *Multimedia Tools Appl.*, vol. 83, no. 1, pp. 963–1000, Jan. 2024, doi: 10.1007/s11042-023-15431-9.

[17] D. M. Davids, A. A. E. Raj, and C. S. Christopher, "Hybrid multi scale hard switch YOLOv4 network for cricket video summarization," *Wireless Netw.*, vol. 30, no. 1, pp. 17–35, Jan. 2024, doi: 10.1007/s11276-023-03449-8.

[18] A. Basu, R. Pramanik, and R. Sarkar, "Wanet: Weight and attention network for video summarization," *Discover Artif. Intell.*, vol. 4, no. 1, Jan. 2024, doi: 10.1007/s44163-024-00101-y.

[19] M. Hu, R. Hu, Z. Wang, Z. Xiong, and R. Zhong, "Spatiotemporal two-stream LSTM network for unsupervised video summarization," *Multimedia Tools Appl.*, vol. 81, no. 28, pp. 40489–40510, Nov. 2022, doi: 10.1007/s11042-022-12901-4.

[20] V. Raut and R. Gunjan, "Transfer learning based video summarization in wireless capsule endoscopy," *Int. J. Inf. Technol.*, vol. 14, no. 4, pp. 2183–2190, Jun. 2022, doi: 10.1007/s41870-022-00894-0.

[21] M. U. Sreeja and B. C. Kovoor, "A multi-stage deep adversarial network for video summarization with knowledge distillation," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 8, pp. 9823–9838, Aug. 2023, doi: 10.1007/s12652-021-03641-8.

[22] W.-L. Li, T. Zhang, and X. Liu, "A static video summarization approach via block-based self-motivated visual attention scoring mechanism," *Int. J. Mach. Learn. Cybern.*, vol. 14, no. 9, pp. 2991–3002, Sep. 2023, doi: 10.1007/s13042-023-01814-9.

[23] K. R. Raval and M. M. Goyani, "A survey on event detection based video summarization for cricket," *Multimedia Tools Appl.*, vol. 81, no. 20, pp. 29253–29281, Aug. 2022, doi: 10.1007/s11042-022-12834-y.

[24] A. F. U. R. Khilji, U. Sinha, P. Singh, A. Ali, S. R. Laskar, P. Dadure, R. Manna, P. Pakray, B. Favre, and S. Bandyopadhyay, "Multimodal text summarization with evaluation approaches," *Sādhanā*, vol. 48, no. 4, Oct. 2023, doi: 10.1007/s12046-023-02284-z.

[25] S. Rafi and R. Das, "Topic-guided abstractive multimodal summarization with multimodal output," *Neural Comput. Appl.*, vol. 2023, Aug. 2023, doi: 10.1007/s00521-023-08821-5.

[26] M. S. Nair and J. Mohan, "VSMCNN-dynamic summarization of videos using salient features from multi-CNN model," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 10, pp. 14071–14080, Oct. 2023, doi: 10.1007/s12652-022-04112-4.

[27] C. Cui, X. Liang, S. Wu, and Z. Li, "Align vision-language semantics by multi-task learning for multi-modal summarization," *Neural Comput. Appl.*, vol. 36, no. 25, pp. 15653–15666, Sep. 2024, doi: 10.1007/s00521-024-09908-3.

[28] A. Salmi, W. Zhang, and F. Jiang, "Reducing data transmission efficiency in wireless capsule endoscopy through DL-CEndo framework: Reconstructing lossy low-resolution Luma images and improving summarization," *Mobile Netw. Appl.*, vol. 29, no. 3, pp. 659–675, Jun. 2024, doi: 10.1007/s11036-024-02334-8.

[29] K. Cizmeciler, E. Erdem, and A. Erdem, "Leveraging semantic saliency maps for query-specific video summarization," *Multimedia Tools Appl.*, vol. 81, no. 12, pp. 17457–17482, May 2022, doi: 10.1007/s11042-022-12442-w.

[30] D. Gupta and A. Sharma, "A two-stage attention augmented fully convolutional network-based dynamic video summarization," *Multimedia Syst.*, vol. 29, no. 6, pp. 3685–3701, Dec. 2023, doi: 10.1007/s00530-023-01154-2.

[31] R. Deepa, T. Sree Sharmila, and R. Niruban, "Dynamic graph neural network-based computational paradigm for video summarization," *Multimedia Tools Appl.*, vol. 83, no. 17, pp. 51227–51250, Nov. 2023, doi: 10.1007/s11042-023-17412-4.

[32] K. Khurana and U. Deshpande, "Two stream multi-layer convolutional network for keyframe-based video summarization," *Multimedia Tools Appl.*, vol. 82, pp. 38467–38508, Mar. 2023, doi: 10.1007/s11042-023-14665-x.

[33] B. Darshankumar and T. M. Manu, "Design of an integrative model for video scene summarization through integrated frame sampling, language processed ResNets fused with domain adversarial training," *Int. J. Inf. Technol.*, vol. 16, no. 7, pp. 4527–4539, Oct. 2024, doi: 10.1007/s41870-024-02050-2.

[34] S. A. Ansari and A. Zafar, "Multi video summarization using query based deep optimization algorithm," *Int. J. Mach. Learn. Cybern.*, vol. 14, no. 10, pp. 3591–3606, Oct. 2023, doi: 10.1007/s13042-023-01852-3.

[35] P. Saini and K. Berwal, "ESKVS: Efficient and secure approach for keyframes-based video summarization framework," *Multimedia Tools Appl.*, vol. 83, no. 30, pp. 74563–74591, Feb. 2024, doi: 10.1007/s11042-024-18405-7.

[36] D. Gupta, A. Sharma, P. Kaur, and R. Gupta, "Experimental analysis of clustering based models and proposal of a novel evaluation metric for static video summarization," *Multimedia Tools Appl.*, vol. 83, no. 1, pp. 3259–3284, Jan. 2024, doi: 10.1007/s11042-022-14081-7.

[37] S. Hossain, K. Deb, S. Sakib, and I. H. Sarker, "A hybrid deep learning framework for daily living human activity recognition with cluster-based video summarization," *Multimedia Tools Appl.*, vol. 2024, pp. 1–54, Apr. 2024, doi: 10.1007/s11042-024-19022-0.

[38] Y. Chen, B. Guo, Y. Shen, R. Zhou, W. Lu, W. Wang, X. Wen, and X. Suo, "Video summarization with u-shaped transformer," *Int. J. Speech Technol.*, vol. 52, no. 15, pp. 17864–17880, Dec. 2022, doi: 10.1007/s10489-022-03451-1.

[39] A. Javed and A. Ali Khan, "Shot classification and replay detection for sports video summarization," *Frontiers Inf. Technol. Electron. Eng.*, vol. 23, no. 5, pp. 790–800, May 2022, doi: 10.1631/FITEE.2000414.

[40] T. Psallidas and E. Spyrou, "Video summarization based on feature fusion and data augmentation," *Computers*, vol. 12, no. 9, p. 186, Sep. 2023, doi: 10.3390/computers12090186.

[41] S. B. Veesam and A. R. Satish, "Deep residual network video summarization for face detection and person re-identification," in *Proc. Int. Conf. Comput. Intell., Netw. Secur. (ICCINS)*, Mylavaram, India, Dec. 2023, pp. 1–6, doi: 10.1109/ICCINS58907.2023.10450024.

[42] S. Babu Veesam and A. R. Satish, "An empirical taxonomy of video summarization model from a statistical perspective," *IEEE Access*, vol. 12, pp. 173850–173866, 2024, doi: 10.1109/access.2024.3503276.

[43] S. B. Veesam and A. R. Satish, "Design of an iterative method for CCTV video analysis integrating enhanced person detection and dynamic mask graph networks," *IEEE Access*, vol. 12, pp. 157630–157656, 2024, doi: 10.1109/access.2024.3485896.

[44] S. B. Veesam, A. R. Satish, S. Tupakula, Y. Chinnam, K. Prakash, S. Bansal, and M. R. I. Faruque, "Design of an integrated model with temporal graph attention and transformer-augmented RNNs for enhanced anomaly detection," *Sci. Rep.*, vol. 15, no. 1, Jan. 2025, doi: 10.1038/s41598-025-85822-5.

**SAI BABU VEESAM** is currently pursuing the Ph.D. degree with VIT-AP University, under the guidance of Dr. Aravapalli Rama Satish Garu.

He has a solid background in IT and education, with more than seven years of software development experience and four years as an Assistant Professor. His technical proficiency includes PHP, Java, Android development, and various front-end technologies. Additionally, he is skilled in effectively addressing server-related issues. His research interests include machine learning, computer vision, security, and advanced technologies. He has contributed to the field through publications in reputable journals and conferences, showcasing his commitment to advancing research and technology.

**ARAVAPALLI RAMA SATISH** received the master's degree in computer science and engineering from JNTUK, and the Ph.D. degree from Acharya Nagarjuna University.

He is currently a Professor with the School of Computer Science and Engineering, VIT-AP University, Andhra Pradesh. In addition to having over 22 years of teaching experience, he has research papers published in conferences and reputable journals. Both UG and PG projects have been overseen by him. In the fields of computer vision, text mining, and data warehousing, he has mentored two Ph.D. candidates to completion and is currently guiding two research scholars. He is also the author of two "*Unix Programming*" textbooks. In addition, he actively participated in the planning of international conferences and coordinated faculty development programs. Along with being a Life Member of CSI and ISTE, he had completed certificates from NPTEL and Coursera. His research interests include machine learning, data mining, big data, and security.