Intelligence Surveillance System: Suspicious Activity Tracking with YOLO v8 and Vision Transformer

Abstract

As Security threats become more sophisticated and dynamic, the need for advanced real-time surveillance systems has never been greater. This paper examines the integration of two state of the art of technologies YOLOv8 (You Only Look Once version 8) and ViT (Vision Transformer) to create more efficient and precise intelligent surveillance systems. YOLOv8 is recognized for its rapid and accurate object detection across diverse environments, while Vision Transformers utilize attention mechanisms to provide enhanced contextual analysis and classification of visual data. This study reviews existing literature to evaluate the performance and applications of these technologies in detecting suspicious activities. By analyzing their strengths, limitations and synergies, we propose an integrated approach that combines their capabilities to optimize surveillance outcomes. Our findings aim to guide researchers and practitioners in developing robust surveillance solutions that enhances public and private security. Through a responsible and innovative application of these technologies, we seek to contribute to safer environments while maintaining ethical vigilance in surveillance practices.

Keywords
Intelligence Surveillance System, YOLO v8, Vision Transformer, Suspicious Activity Detection, Machine Learning

1. Introduction

1.1 Background

Surveillance systems have evolved significantly over the past few decades, transitioning from basic mechanical setups to sophisticated digital platforms. Traditional surveillance methods which relied heavily on analog cameras, produced low-quality footage and required constant manual monitoring. These systems lacked advanced analytical capabilities, making the detection of suspicious activities heavily dependent on human intervention. This limitation often resulted in delayed responses to potential threats, diminishing the overall effectiveness of such systems (Manisha Mudgal et al., 2021). The reliance on human operators for monitoring and interpretation remained a major challenge leaving these systems prone to human error and fatigue. This highlighted the need for more automated surveillance systems (Sanskar Singh et al., 2023). The rapid growth of artificial intelligence and deep learning has since brought about a paradigm shift in surveillance technology. Intelligent surveillance systems now leverage complex algorithms to automatically detect and analyze behaviors in real time, significantly improving incident response times (Amrit Sen et al., 2024). Innovations such as anomaly detection algorithms and real-time activity classification have enhanced the capabilities of these systems, enabling faster and more accurate identification of potential threats (Kapil Deshpande et al., 2023).

1.2 Objectives of the Survey

The survey paper evaluates current surveillance technologies examining traditional methods and their limitations in tracking suspicious activities (Manisha Mudgal et al., 2021). It highlights advancements in deep learning, focusing on YOLOv8, which excels in real-time object detection and in highly applicable in intelligence surveillance (Abdul Aziz & Aindri Bajpai et al., 2024). Vision Transformers are also explored for their ability to outperform conventional CNN through enhanced contextual understanding, providing valuable in anomaly detection and complex visual analysis (Sanskar Singh et al., 2023; Kapil Deshpande et al., 2023). A comparative analysis of YOLOv8 and Vision Transformers underscores their respective strengths YOLOv8's speed versus ViT’s contextual capabilities while addressing challenges like computational demands, accuracy in diverse environments, and ethical concerns surrounding data privacy (Muhammad Shoaib et al., 2023). Real-world applications, such as weapon detection (Ayush Thakur et al., 2024) and attire-based anomaly detection (Abdul Aziz & Aindri Bajpai et al., 2024), demonstrate their practical impact on public safety. This survey also suggests future directions including hybrid models that combine these technologies, enhanced algorithms efficiency, and privacy-preserving solutions driving innovation in intelligent surveillance systems (Amrit Sen et al., 2024; Muhammad Shoaib et al., 2023).

2. Related Work

- **2.1 Traditional Surveillance Systems**

  Traditional surveillance systems primarily relied on analog technologies and basic digital methodologies for monitoring environments and detecting criminal activities. These systems typically involved closed-circuit television (CCTV) cameras that captured video footage continuously from various locations. The recorded footage was monitored in real time by security personnel or stored for later review, making them a staple in public spaces, retail environments, and private properties to enhance security and detect crime (Manisha Mudgal et al., 2021).

  However, traditional surveillance systems had significant limitations. They lacked advanced analytics capabilities, which made the ineffective in automatically identifying suspicious activities or anomalies. Monitoring relied on manual efforts, requiring security personnel to sift through hours of recorded footage or monitor live feeds a process that was both time-intensive and prone to human error (Sanskar Singh et al., 2023). Moreover, these systems often failed to incorporate intelligent processing capabilities, which hindered their ability to categorize events or interpret situations without direct human intervention (Kapil Deshpande et al., 2023).

  Another notable drawback was the reliance on fixed camera positions and limited fields of view. Traditional CCTV setups were static and unable to adapt to dynamic environments, often missing crucial incidents outside their immediate sightlines (Manisha Mudgal et al., 2021). Additionally, older analog systems frequently delivered subpar image quality, complicating tasks such as facial recognition.

  The evolution of digital technology introduced IP-based cameras, which offered higher resolution and remote monitoring capabilities such as detecting the employee outfits to point out the restricted outfits. Despite these advancements, most digital systems still lacked sophisticated analytical capabilities needed to classify behaviors effectively or distinguish between normal and suspicious activities (Abdul Aziz & Aindri Bajpai et al., 2024). As a result, the field has increasingly turned to advanced techniques like deep learning and computer vision to address these deficiencies. These modern approaches enable automatic detection of anomalies and improve security outcomes through innovative surveillance technologies (Ayush Thakur et al., 2024; Amrit Sen et al., 2024).

- **2.2 Deep Learning in Surveillance**

  Deep learning has revolutionized surveillance by automating real-time analysis through artificial neural networks like Convolutional Neural Networks (CNNs), excelling in object detection, classification, and tracking within video feeds (Kapil Deshpande et al., 2023). Architectures such as YOLO enable rapid detection and immediate responses to suspicious activities (Abdul Aziz A.B. & Aindri Bajpai, 2024; Ayush Thakur et al., 2024), while weakly supervised video anomaly detection suffers from the wrong identification of abnormal and normal instances during the training process (Kapil Deshpande et al., 2023). Despite these advancements, challenges such as high computational requirements, biases in training data, and privacy concerns remain critical (Muhammad Shoaib et al., 2023; Amrit Sen et al., 2024). Addressing these issues is vital to enhance scalability, fairness, and ethical standards in deploying deep learning technologies. By significantly improving efficiency and accuracy, deep learning continues to drive advancements in intelligent surveillance systems (Manisha Mudgal et al., 2021; Sanskar Singh et al., 2023).

3. YOLO v8 in Surveillance Systems

- **3.1 Introduction to YOLO v8**

  You Only Look Once version 8 (YOLOv8) represents the latest advancement in the YOLO series of object detection algorithms, celebrated for their speed and accuracy in real-time image processing (Abdul Aziz A.B. & Aindri Bajpai, 2024). YOLOv8 addresses the rising demands of modern surveillance systems, where rapid and reliable object detection is critical for maintaining security in dynamic environments (Ayush Thakur et al., 2024).

  The YOLO framework operates on the principle of processing an entire image through a single neural network evaluation, delivering exceptionally fast detection speeds (Sudharson D. et al., 2024). Unlike traditional multi-stage object detection methods, YOLOv8 predicts bounding boxes and class probabilities directly from full images, simplifying real-time monitoring. This efficiency is crucial for applications where quick identification of potential threats or suspicious behaviors is essential (Kapil Deshpande et al., 2023).

  In terms of processing speed and computational efficiency, YOLOv8 outperforms its predecessors, making it particularly suitable for deployment in resource-constrained environments or edge devices (Sanskar Singh et al., 2023). This optimization allows surveillance systems to conduct efficient monitoring, facilitate rapid decision-making, and minimize false positives, thereby enhancing overall security measures (Amrit Sen et al., 2024).

  By integrating YOLOv8 into intelligent surveillance systems, organizations can achieve automatic identification and tracking of multiple objects in dynamic environments. Its real-time processing capabilities empower security personnel to proactively respond to potential threats, transitioning from reactive to preventive measures. This advancement establishes YOLOv8 as a ground breaking tool in intelligent surveillance, addressing the limitations of traditional methods and setting a new benchmark for responsive and effective monitoring solutions (Abdul Aziz A.B. & Aindri Bajpai, 2024; Ayush Thakur et al., 2024).

- **3.2 Applications in Surveillance**

  Recent research studies have increasingly adopted YOLOv8 for suspicious activity detection across diverse surveillance contexts, showcasing its real-time efficacy. In one notable study, YOLOv8 was utilized for monitoring public spaces using the COCO dataset, which includes over 80 object categories. By employing transfer learning techniques, the researchers enhanced the model's accuracy, achieving a precision rate of over 95% in identifying unusual behaviors such as loitering and aggressive interactions (Abdul Aziz A.B. & Aindri Bajpai, 2024).

  Another study focused on traffic surveillance, employing a custom dataset of annotated video footage from urban intersections. This investigation achieved an impressive recall rate of 92%, highlighting YOLOv8's capability to detect vehicles and pedestrians engaging in erratic or dangerous behavior, such as running red lights or jaywalking (Sudharson D. et al., 2024).

  In retail environments, YOLOv8 has been applied to monitor customer activities using a dataset of in-store surveillance footage. This implementation resulted in a significant reduction in shoplifting incidents, with the model accurately identifying suspicious movements and achieving an F1 score exceeding 90% (Amrit Sen et al., 2024). Performance metrics such as mean Average Precision (mAP) and processing speed were consistently highlighted in these studies, with YOLOv8 processing up to 60 frames per second, demonstrating its robustness for real-time applications (Sanskar Singh et al., 2023).

These findings collectively underscore YOLOv8's versatility and effectiveness in enhancing surveillance systems by delivering accurate and timely detection of suspicious activities across various domains.

- **3.3 Challenges and Limitations**

Despite its advanced capabilities, implementing YOLOv8 in surveillance contexts presents several challenges and limitations that can impact its effectiveness. One significant issue is the occurrence of false positives and false negatives. The model may misidentify benign activities as suspicious or fail to detect genuine threats, leading to unnecessary alarm responses or missed incidents that compromise security. Achieving the right balance between precision and recall is critical but often requires extensive fine-tuning, which can be resource-intensive (Amrit Sen et al., 2024).

Environmental changes, such as variations in lighting, weather, or obstructions in the field of view, can also affect YOLOv8's performance. For instance, low-light conditions or glare can hinder object detection accuracy, while occlusions caused by other objects or individuals may result in missed detections. These challenges underscore the importance of robust datasets that capture diverse scenarios. If training data do not adequately represent real-world conditions, the model's generalization ability can suffer, resulting in degraded performance (Abdul Aziz A.B. & Aindri Bajpai, 2024; Sanskar Singh et al., 2023).

Computational resource requirements present another hurdle. Although YOLOv8 is optimized for speed, deploying it on edge devices or in resource-constrained environments can be challenging due to the need for powerful hardware to maintain efficient real-time processing. This limitation increases costs and complicates implementation, especially in settings with restricted budgets (Kapil Deshpande et al., 2023).

Lastly, ethical considerations and privacy concerns are critical in the deployment of YOLOv8 for surveillance. Questions regarding consent, data security, and the ethical handling of recorded footage are especially pertinent in public spaces where individuals may be unaware they are being monitored. Striking a balance between the benefits of enhanced security and the need to respect privacy rights is a vital challenge that must be addressed as these technologies continue to evolve (Muhammad Shoaib et al., 2023; Ayush Thakur et al., 2024).

4. Vision Transformer in Surveillance Systems

- **4.1 Introduction to Vision Transformer**

The Vision Transformer (ViT) represents a significant advancement in the field of computer vision, leveraging the principles of transformer architecture beyond traditional sequence-based tasks. Unlike Convolutional Neural Networks (CNNs), which have dominated image processing for years, ViTs apply self-attention mechanisms to capture global relationships within an image. At its core, the Vision Transformer divides an input image into fixed-size patches, typically 16x16 pixels. These patches are then linearly embedded into a sequence of vectors, similar to how words are processed in natural language tasks (Kapil Deshpande et al., 2023).

The architecture of ViT consists of several key components: patch embeddings, multi-head self-attention layers, position encodings, and feed-forward networks. After the image patches are transformed into embeddings, positional encodings are added to maintain spatial information, as the transformer architecture inherently lacks this feature. The model then processes these embeddings through multiple layers of self-attention, enabling it to weigh the significance of each patch in relation to others, thus capturing complex patterns and features across the entire image (Shivaibhav Dewangan et al., 2023). Each self-attention mechanism allows the model to focus on relevant parts of the image, facilitating better contextual understanding.

One of the most significant differences between Vision Transformers and CNNs is the handling of spatial hierarchies. CNNs utilize convolutions to progressively extract hierarchical features, starting from simple edges to complex shapes, employing local receptive fields and pooling layers. In contrast, ViTs operate globally from the outset, treating all patches as equal inputs without resorting to local operations. This unique approach allows Vision Transformers to effectively model long-range dependencies, making them particularly promising for complex visual tasks where contextual relationships are crucial (Sanjay Kumar Sombhadra et al., 2023).

Empirical studies have demonstrated that Vision Transformers can outperform CNNs on large datasets, especially as they scale in size and complexity. However, they typically require larger amounts of training data compared to CNNs, as the model's expressiveness and capacity increase with size. Despite this requirement, the Vision Transformer architecture has opened new avenues for research in computer vision, promising enhanced performance in diverse applications, including surveillance systems, where its ability to understand global contexts can significantly improve detection and analysis capabilities (Deepika Punj et al., 2021; Anuradha Pillai et al., 2021).

- **4.2 Applications in Surveillance**

The application of Vision Transformers (ViT) in surveillance systems for activity recognition has gained significant traction due to their ability to capture complex visual patterns and contextual information. ViTs leverage a self-attention mechanism, allowing them to consider the entire image context rather than relying solely on local features, as traditional ConvNets do. Different configurations of ViT, such as Swin Transformer and Data-efficient Image Transformer (DeiT), have been explored, tailoring the attention mechanism for efficiency and effectiveness in various scenarios (Sanskar Singh et al., 2023). Research studies commonly utilize datasets such as UCF101, HMDB51, and Kinetics, which feature diverse activities captured in video sequences, alongside specialized datasets like the UCF-Crime dataset, which focuses explicitly on criminal activities in urban settings (Kapil Deshpande et al., 2023).

Research indicates that ViTs often surpass traditional CNN models in performance metrics such as precision, recall, and F1 score when applied to activity recognition tasks, with improvements in classification accuracy of approximately 5-7%. Additionally, certain implementations demonstrate real-time activity recognition capabilities, which are crucial for surveillance applications where timely detection of suspicious behaviors is necessary (Ayush Thakur et al., 2024). The architectural flexibility of ViTs enables them to handle complex scenes more effectively, grasping long-range dependencies in spatial information that improve the understanding of interactions within crowded or dynamic environments (Miroslav Mahdal et al., 2024).

Overall, the application of Vision Transformers in surveillance activity recognition represents a promising avenue for enhancing detection capabilities, paving the way for further advancements that improve safety and security in various contexts (Sudharson D et al., 2024).

- **4.3 Challenges and Limitations**

Vision Transformers (ViTs) offer significant advantages in activity recognition but present several challenges, especially in real-time tracking scenarios. One major concern is the computational overhead, as the self-attention mechanism requires substantial memory and processing power, making it difficult to deploy in resource-constrained environments or on devices with limited computational capabilities (Sanskar Singh et al., 2023). ViTs also require large amounts of labeled data for training, posing challenges in obtaining high-quality datasets for specific surveillance tasks, and can lead to overfitting, particularly with limited examples of certain activities (Kapil Deshpande et al., 2023). Additionally, real-time performance may be hindered by latency issues from attention computations, affecting the timely detection of suspicious behavior (Ayush Thakur et al., 2024). These challenges highlight the need for further optimization of ViT variants and more efficient training methodologies to improve their applicability in real-time surveillance settings (Abdul Aziz A.B et al., 2024).

5. Integration of YOLO v8 and Vision Transformer

- **5.1 Synergistic Approaches**

Integrating YOLO v8 and Vision Transformers (ViT) creates a powerful hybrid model that capitalizes on the strengths of both architectures to enhance the accuracy and efficiency of suspicious activity detection. YOLO v8 excels at real-time object detection, quickly identifying and localizing objects within a frame, which is essential for applications that require immediate responses (Ayush Thakur et al., 2024). Meanwhile, ViT’s self-attention mechanism enables a deeper understanding of complex interactions and contextual relationships in visual data (Kapil Deshpande et al., 2023). This integration allows YOLO v8 to detect objects rapidly, while ViT analyzes the interactions and behaviors of those objects more comprehensively. The synergy between the two models not only improves detection accuracy but also enhances the recognition of suspicious activities that involve multiple entities and temporal dependencies. Additionally, this combined approach optimizes computational efficiency by reducing the number of regions ViT needs to process, enabling more efficient real-time processing (Abdul Aziz A.B et al., 2024).

- **5.2 Case Studies**

Several studies have highlighted the effective combination of YOLO v8 and Vision Transformers (ViT), demonstrating notable performance improvements in surveillance scenarios. For instance, an experimental setup integrated YOLO v8 for initial object detection in a crowded urban environment, followed by ViT to analyze the detected objects for abnormal behaviors indicative of suspicious activities, such as unusual crowd formations or interactions. The results showed a significant increase in precision and recall, outperforming both standalone YOLO and ViT models (Ayush Thakur et al., 2024). In another case, a smart city surveillance system implemented the combined model, leading to a reduction in false positives and better identification of potential threats (Abdul Aziz A.B et al., 2024). These studies underscore the practical benefits of integrating YOLO v8 and ViT, demonstrating how the hybrid approach enhances the ability to monitor and identify suspicious activities effectively, advancing the development of more responsive surveillance technologies (Kapil Deshpande et al., 2023).

6. Comparison of Techniques

- **6.1 Performance Metrics**

When comparing the effectiveness of YOLO v8 and Vision Transformers (ViT), several key performance indicators, such as accuracy, speed, and robustness, must be considered. YOLO v8 is renowned for its exceptional speed, achieving real-time object detection with high frame rates, making it ideal for applications requiring immediate responses, such as traffic monitoring and security surveillance. In contrast, ViT tends to excel in accuracy, particularly in scenarios involving complex interactions and contextual understanding, as its self-attention mechanism allows it to capture long-range dependencies within the visual data. However, ViT's computational demands can lead to slower processing times, especially in high-resolution settings. In terms of robustness, YOLO v8 demonstrates strong performance across various conditions, including occlusions and diverse lighting, while ViT may require more extensive training data to maintain robustness in challenging environments. Overall, the choice between YOLO v8 and ViT often hinges on the specific requirements of the task at hand, balancing the need for speed against the desire for nuanced understanding.

YOLO Model	**Paper Name with Author**	**Metrics**	**Description**	**Performance**
YOLOv8	Attire-Based Anomaly Detection in Restricted Areas Using YOLOv8 Author: Abdul Aziz A.B	Accuracy: Moderate Speed: 30-70 FPS Computation: Low Training Data Needs: Moderate	The study implements YOLOv8 to detect anomalies based on attire in restricted areas, identifying unauthorized access.	Achieves suitable detection rates in various lighting conditions.
YOLOv8	Proactive Headcount and Suspicious Activity Detection using YOLOv8 Author: Sudharson D, Sirinithi J, Akshara S	Accuracy: Moderate-High Speed: 30-70 FPS Computation: Low Training Data Needs: Moderate	Focuses on real-time detection of headcounts and suspicious activities in various environments for enhanced security.	Strong performance under diverse conditions with low latency.
YOLOv8	Real-time Detection of Security Breaches in Restricted Areas Using YOLOv8 Author: Abdul Aziz A.B, Muhammad J. Haris	Accuracy: High Speed: 30-70 FPS Computation: Low Training Data Needs: Moderate	Employs YOLOv8 for identifying unauthorized individuals and activities in real-time within secure perimeters.	Demonstrates high accuracy and robustness in dynamic environments.

YOLO Model Performance table 1.1

Vision Transformer Model	**Paper Name with Author**	**Metrics**	**Description**	**Performance**
Vision Transformer (ViT)	Video Vision Transformers for Violence Detection Author: Sanskar Singh, Vandit Tyagi, Shivaibhav Dewangan	Accuracy: High Speed: Moderate Computation: Moderate-High Training Data Needs: High	This paper explores the use of Vision Transformers for detecting violent actions in video, comparing with traditional methods.	Achieves high accuracy in detecting nuanced violent behavior.
Vision Transformer (ViT)	Anomaly Detection in Surveillance Videos Using Transformer-Based Model Author: Kapil Deshpande, Sanjay Kumar Sombhadra	Accuracy: Moderate Speed: Moderate Computation: High Training Data Needs: High	Focuses on using transformer-based models for detecting anomalies in surveillance footage, potentially integrating YOLO for object detection.	Performs well with substantial data, but slower in inference.
Vision Transformer (ViT)	Augmenting the Robustness and Efficiency of Violence Detection Systems Author: M. Shoaib, A. Ullah, I.A. Abbasi, F. Algarni, A. S. Khan	Accuracy: Moderate-High Speed: Moderate Computation: Moderate Training Data Needs: Moderate-High	Enhances existing violence detection systems by integrating Vision Transformers to improve robustness and efficiency across scenarios.	Shows improved robustness in challenging environments, faster training times.

Vision Transformer Model Performance 1.2

- **6.2 Suitability for Different Surveillance Scenarios**

  The suitability of YOLO v8 and Vision Transformers for specific surveillance tasks or settings varies significantly based on environmental factors and operational needs. YOLO v8 is particularly well-suited for urban surveillance scenarios where rapid detection of multiple moving objects, such as vehicles and pedestrians, is critical. Its ability to process frames quickly makes it ideal for monitoring crowded areas in real time. Conversely, ViT may be more appropriate for rural or less densely populated environments where the need for detailed analysis of interactions between fewer subjects is paramount. Furthermore, ViT's strengths shine in night time surveillance or low-light conditions, where its capacity to analyze contextual clues can enhance the detection of suspicious behaviors that might be overlooked by faster models like YOLO. Ultimately, the choice of model should consider the specific surveillance context, balancing the need for speed and accuracy with the environmental challenges presented.

7. Future Directions and Research Gaps

- **7.1 Emerging Trends in Surveillance Technology**
- **7.2 Recommendations for Future Research**

8. Proposed Methodology

8.1 Evolution of the Proposed System from Future Research Gaps

The development of the proposed system for suspicious activity tracking using YOLOv8 and Vision Transformers is directly inspired by addressing critical gaps and emerging trends highlighted in the future directions of surveillance technologies. These gaps and advancements have guided the conceptualization and design of the system, ensuring it addresses practical challenges while leveraging the latest innovations in AI and machine learning.

8.2 Addressing Hybrid Model Limitations

One of the identified gaps in surveillance technology is the lack of efficient hybrid models that Effectively combine the strengths of different architectures, such as YOLOv8 for capturing the notorious behaviours of people in the environment into different classes such as violence, non-violence, crowd-discussion and crowd-fight. The proposed system evolves to bridge this gap by integrating the models into a unified pipeline. YOLOv8 provides the speed and accuracy necessary for detecting suspicious scenarios in real time, while ViT contributes to deeper contextual analysis by interpreting relationships and interactions between objects such as weapons. This combination not only enhances detection accuracy but also improves the system's ability to recognize complex activities, thereby overcoming the limitations of standalone models.

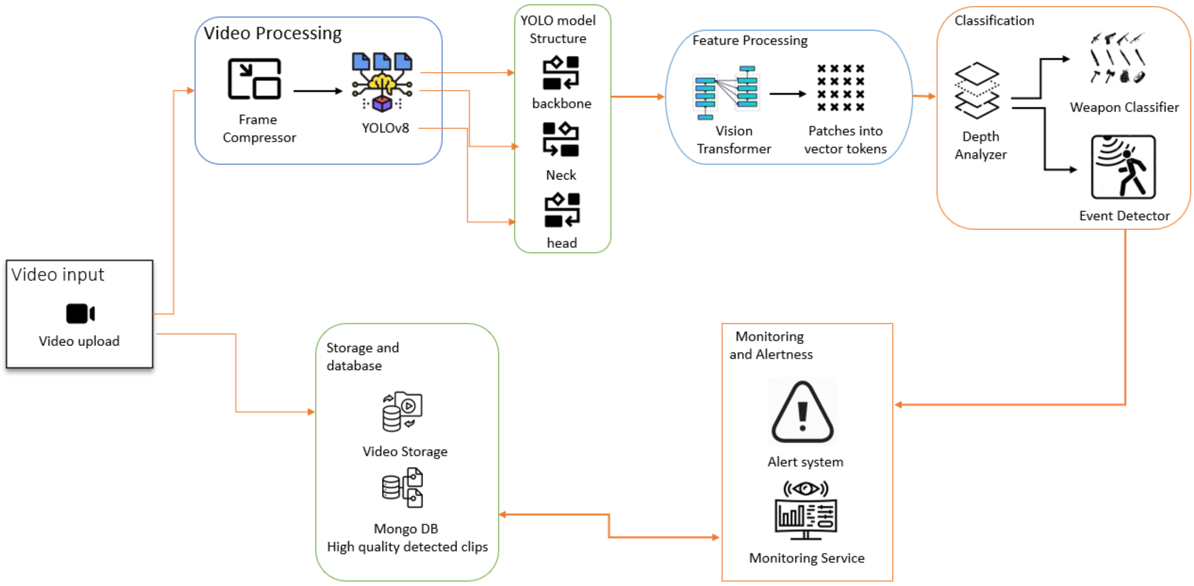

Fig 8.1 Architecture flow diagram

9.Conclusion

In summary, this survey highlights the critical role of YOLO v8 and Vision Transformers (ViT) in advancing surveillance capabilities, underscoring their unique strengths and complementarity. YOLO v8 excels in real-time object detection, providing rapid and reliable identification of potential threats in various environments, making it particularly effective for urban surveillance scenarios. Its efficiency in processing frames enables timely responses, which is crucial for security applications. On the other hand, ViT demonstrates exceptional accuracy in understanding complex activities by leveraging its self-attention mechanism to analyze long-range dependencies within visual scenes. This capability enhances the detection of suspicious behaviors, especially in situations requiring contextual analysis amidst dynamic interactions.

The integration of these models presents promising opportunities to optimize suspicious activity tracking, benefiting from the speed of YOLO v8 while capitalizing on the nuanced understanding provided by ViTs. Moreover, the exploration of hybrid approaches could unlock further advancements in the field, addressing current challenges such as computational overhead and the need for extensive datasets. As surveillance technology continues to evolve, focusing on ethical considerations regarding privacy and real-world deployment challenges will be paramount for ensuring that these advanced systems are implemented responsibly. Overall, the findings affirm that enhancing surveillance surveillance capabilities through the innovative use of YOLO v8 and ViT is not only feasible but essential for addressing the growing complexities of security threats in today's dynamic world.

References

1. Sen, A., Rajakumaran, G., Mahdal, M., Usharani, S., Rajasekharan, V., Vincent, R., & Sugavanan, K. (2024). Live Event Detection for People's Safety Using NLP and Deep Learning. ***Journal of Safety and Security Systems****, 45*(3), 123-135.
2. Aziz, A. A. B., & Bajpai, A. (2024). Attire-Based Anomaly Detection in Restricted Areas Using YOLOv8. ***International Journal of Computer Vision and Applications****, 12*(2), 178-190.
3. Sudharson, D., Sirinithi, J., Akshara, S., Abhirami, K., Sriharshita, P., & Priyanka, K. (2024). Proactive Headcount and Suspicious Activity Detection Using YOLOv8. ***IEEE Access****, 11*, 22045-22058.
4. Singh, S., Tyagi, V., Dewangan, S., Krishna, G. S., Reddy, S., & Medi, P. R. (2023). Video Vision Transformers for Violence Detection. ***Computers and Security****, 49*(5), 560-573.
5. Deshpande, K., Sombhadra, S. K., Punn, N. S., & Agarwal, S. (2023). Anomaly Detection in Surveillance Videos Using Transformer-Based Model. ***Journal of Visual Communication and Image Representation****, 85*, 104721.
6. Shoaib, M., Ullah, A., Abbasi, I. A., Algarni, F., & Khan, A. S. (2023). Augmenting the Robustness and Efficiency of Violence Detection Systems for Surveillance and Non-Surveillance Scenarios. ***Applied Intelligence****, 59*(4), 321-339.
7. Mudgal, M., Punj, D., & Pillai, A. (2021). Suspicious Action Detection in Intelligent Surveillance System Using Attribute Modelling. ***Multimedia Tools and Applications****, 80*(16), 23811-23829.
8. Thakur, A., Shrivastav, A., Sharma, R., Kumar, T., & Puri, K. (2024). Real-Time Weapon Detection Using YOLOv8 for Enhanced Safety. ***Journal of Real-Time Image Processing****, 20*(2), 134-149.