Available online at https://sextante-ensb.com

Revista Sextante

ISSN : 2665-3923

Volume: 34 Issue: 01 (2026)

Efficient Multi-Camera Real-Time Surveillance Leveraging Deep Learning for Enhanced Object Detection

Aftab Khan*[a], Muhammad Saadat Ali[b], Yasir Khan[c], Chaudhary Muhammad Shahbaz Anjum[d], Haider Abbas Afridi[e], Muhammad Huzaifa[f], Zahid Miraj[g] and Danish Khaleeq[h]

a. *Corresponding Author: Department of Computer Science (DCS), Allama Iqbal Open University (AIOU), H-8, Islamabad, Pakistan (aftab.khan@aiou.edu.pk)*
b. *Department of Computer Science (DCS), Allama Iqbal Open University (AIOU), H-8, Islamabad, Pakistan*
c. *Department of Computer Science (DCS), Allama Iqbal Open University (AIOU), H-8, Islamabad, Pakistan*
d. *Department of Computer Science (DCS), Allama Iqbal Open University (AIOU), H-8, Islamabad, Pakistan*
e. *Department of Computer Science (DCS), Allama Iqbal Open University (AIOU), H-8, Islamabad, Pakistan*
f. *Department of Computer Science (DCS), Allama Iqbal Open University (AIOU), H-8, Islamabad, Pakistan*
g. *Department of Computer Science (DCS), Allama Iqbal Open University (AIOU), H-8, Islamabad, Pakistan*
h. *Department of Computer Science (DCS), Allama Iqbal Open University (AIOU), H-8, Islamabad, Pakistan*

Article History
Received September 13th, 2026
Accepted November 08th, 2026
Published November 20th, 2026

© 2026 The Authors. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution(CC-BY)license https://creativecommons.org/licenses/by/4.0

Abstract
This research proposes an Efficient Multi-Camera Real-Time Surveillance (EMRS) system aimed at improving video surveillance using deep learning methods. The primary objective is to enhance object detection for smart surveillance, while the secondary goal is to evaluate the efficiency of EMRS in addressing gaps in current surveillance systems. The study integrates the latest deep learning techniques, achieving a mean Average Precision (mAP) of 65 percent and a video frame rate of 45 FPS. The object recognition accuracy significantly outperforms conventional models such as YOLO 2.0, which achieves a mAP of 60 percent at 40 FPS. This research contributes to the development of new technologies and simultaneously strengthens security in sensitive sectors, such as public safety and home protection. Superior algorithms are implemented within EMRS as part of its continuous enhancement, with a specific focus on identifying and resolving existing issues in object detection software. Leveraging deep learning, the system's sensitivity to potential threats has been significantly improved, indicating that the main goal of assessing EMRS effectiveness has been achieved. Real-world operational tests confirmed that EMRS effectively improves object detection, with an average Intersection over Union (IoU) score of 0.83, a precision of 0.90, and a recall of 0.85. Although this study primarily focuses on home surveillance, potential challenges may arise when applying the system in different outdoor environments.

Keywords: Deep Learning, Object Detection, Real-time Surveillance, Public Safety, Video Monitoring, Intelligent Surveillance Systems.

© 2026 The Authors. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution(CC–BY)license
https://creativecommons.org/licenses/by/4.0

1. Introduction

The rapid advancement of technologies like deep learning and computer vision has led to significant improvements in the field of surveillance. Such innovative technologies have led to the creation of smart surveillance of which the need to enhance security requirements and demands on real-time video monitoring is on the rise. Recent works, including the development of deep networks to track a set of multiple objects at the same time across multiple cameras in traffic monitoring at scale indicate the current development and evolution of the area [1, 2].

Despite the critical role of human oversight in surveillance, current techniques have limitations that hinder effective detection of all objects, especially in real-world scenarios, making them less suitable for practical use. The challenge of having a better and more efficient surveillance system has led to the advent of deep learning as a transformational technology, which is transforming the handling of security in a wide spectrum of settings [3]. Advanced computer-vision algorithms have led to real-time video analysis, which allows state-of-the-art detection and tracking within many camera feeds [4]. This capability is invaluable in complex environments, such as public spaces, where traditional methods struggle to process and interpret large volumes of data Huang [1]. By integrating deep learning models, these systems significantly enhance object recognition, movement prediction, and anomaly detection, improving security outcomes and reducing the reliance on human operators [5]. Figure 1 shows example scenarios of object detection.



Figure 1: *Examples scenarios wherein various types of objects of interest are detected.*

However, an extensive amount of literature on human activity recognition and video surveillance proves the existence of significant gaps, particularly regarding the application of theory to practice and empirical stability of the suggested methods [6]. Despite the fact that some studies have used the computational intelligence to track objects in real-time [7], a significant percentage of the studies are yet to consider the challenges in logistics as well as operations associated with implementation of such systems in the real world [8]. Similarly, the research literature on the principles of Human-Computer Interaction (HCI) in the framework of the smart-city surveillance studies generally lacks the field validation and makes the mentioned methodologies questionable in regards to their viability. All of these shortcomings point to a need to do more research in order to do a rigorous check regarding the applicability of these systems to real-world scenarios.

Moreover, there is a paucity of systematic reviews based on a broad range of data. Although activity recognition methods utilizing deep-learning based methods in the context of wearable sensor streams have been explored [9], these systems have not been evaluated with regard to rigorous cross-dataset validation thus creating concerns on robustness and generalizability. There exists a large body of research that focuses on discrete surveillance elements, e.g. camera selection or violence detection, but does not consider crucial aspects like anomaly detection [10]. As an example, information on violence detection can prove to be useful, yet, it often does not have the mixing-up of the task-wide interests of the general security of the whole system [11]. These deficiencies highlight the need to have more comprehensive investigations occupying a broader range of surveillance capabilities.

© 2026 The Authors. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution(CC–BY)license
https://creativecommons.org/licenses/by/4.0

To solve these issues, this paper is an Efficient Multi-camera Real-time Surveillance (EMRS) system, which uses deep learning and computer vision to improve object detections and identifications in video surveillance. The EMRS can minimize the need of human operators and, at the same time, enhance the efficiency of the operation through a combination of camera feeds and sophisticated processes of deep-learning, as proved in reference [12]. It is specifically aimed at real-time threat detection and action and is applied both to the public security and homes, as well as to the industrial setting. It provides 24 hours monitoring, an extensive coverage and overcoming the drawbacks of the traditional surveillance practices.

This essay has demonstrated that there is an urgent need to have intelligent monitoring within areas that are dynamic like those related to disaster response readiness, law enforcement and industrial security. The EMRS supplements real-time monitoring and threat identification systems and as such offers operational and concrete answers to modern security issues. This research focuses on the software aspect of surveillance, using benchmark datasets and evaluation metrics to assess performance, but does not cover hardware architecture, network deployment, legal considerations, cost estimation, or graphical user interface design.

The paper is structured as follows: Section 2 reviews the literature on object detection and tracking methods, with emphasis on deep learning models such as YOLO and R-CNN. Section 3 outlines the research conducted for developing the EMRS, followed by a detailed explanation of the object detection algorithm in Section 4. Section 5 presents the simulation results and performance comparisons with benchmark methods, and Section 6 concludes with directions for future research.

2. Literature Review

This section reviews the current state of surveillance systems, with a focus on advancements driven by deep-learning methods. Based on a thorough review of the literature, it assesses modern intelligent surveillance systems in terms of their applications, advantages, limitations, and future development prospects. This review not only highlights the current status of the field but also identifies areas where further growth is needed, forming the basis for this research study's contribution to the surveillance technology domain.

The literature surveyed in this section comprises the latest studies. In this review, various advancements in object recognition and tracking using deep learning for detecting human activities are considered, particularly regarding their improvement [12, 13]. Some of these studies focus on the design of newer models and algorithms for surveillance systems that offer better efficiency and accuracy, especially when deep learning and computational intelligence are used [14, 15]. However, many of these investigations lack an important dimension: inadequate consideration of the practical implementation challenges and limitations of the proposed methodologies. This gap allows the limitations of these methods to be overlooked when applied to real-world surveillance scenarios [16, 17] and highlights the need for more practical studies and experiments to assess the feasibility and stability of the identified methods.

One of the key issues is the interpretability of deep learning models, and their scalability in a security and surveillance environment. The notion of interpretability in this case can be understood with the degree to which we can be able to explain the internal processes which make a model reach a specific conclusion and that of scalability is with respect to how the model will perform when faced with a massive and heterogenous data. The discussed issues are considered down in relation to some mainstream architectures like YOLO and Faster R-CNN [3, 18].

Computational performance, complexity, and how easily existing systems can be integrated with these models are also evaluated. Incorporating insights on deep learning in surveillance, this research seeks to provide findings that will assist decision-makers in implementing cost-effective, efficient, and reliable systems to ensure security and safety. These considerations inform future uses of deep learning technologies in public security, transportation security, or critical infrastructure protection [19], thus driving advancements in surveillance and expanding the understanding of impactful change in the field.

2. 1 Object Detection and Tracking

In [17], Alotaibi proposes a new object tracking and real-time detection system based on computational intelligence for video surveillance systems, achieving better mean average accuracy compared to previous work. Nevertheless, the limitations of existing studies include a lack of evidence regarding how this strategy could be implemented in practice, which threatens the feasibility of this approach. Likewise, Huang in [1] used deep learning to

© 2026 The Authors. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution(CC–BY)license
https://creativecommons.org/licenses/by/4.0

extract tennis objects from frames, indicating high accuracy in detecting these objects. In [13], Hussain et al. also present similar views and investigate the feasibility of anomaly-based resources for surveillance in large networks, demonstrating a high ability to prioritize regions with potential security threats.

Although edge computing has been discussed as beneficial to smart surveillance by reducing data transfer and allowing for immediate responses, there is limited understanding of its real-life challenges and impacts. [15] developed a reliable multiple camera vehicle tracking technique for highway applications using deep learning in real-time surveillance, further enhancing real-time applications. When compared to previous methods in object tracking and detection, Alotaibi's research employed computational intelligence. Edge computing in smart surveillance systems reduces the volume of data transmitted and enhances privacy; problems are solved with a mean average accuracy increase of 70-80 points. [18] improved multiple-camera people tracking by using the proposed anchor-guided clustering and spatio-temporal consistency ID re-assignment, showing better accuracy in complex scenes.

2.2 Intelligent Video Surveillance

Human-Computer Interaction (HCI) has been explored as a means of enhancing smart city surveillance systems, particularly focusing on ethics and user experience [1]. The research suggests potential improvements in smart city monitoring, though more empirical studies are needed to test these frameworks in real-world contexts. [20] contributed to the field with their work on large-scale intelligent video analytics using multi-camera systems, emphasizing the importance of coordinated surveillance efforts.

[4] provided insights into deep learning for violence detection in video surveillance. While focusing on violence limits its scope, the research showed detection accuracy between 48 percent and 74 percent. Their findings reflect the rising demand for surveillance systems, though global adoption remains limited. The analysis of the violent behaviour patterns through an intelligent technology aimed at reinforcing the security of people was evaluated by [3]. The research gave accuracy of 48 percent to 74 percent in different settings with the loss of 0.301 to 0.487. This indicates that the system has potential, but it still requires additional real-life testing in order to prove to be efficient in different settings.

In the context of cyber-physical systems, a deep-learning model was suggested to identify security threats with an adaptability of 95 per cent detection rate in attempted attacks [8]. The method significantly increases the level of security and reliability of the system, but requires significant computing resources. At the same time, there are studies that suggest that Reconfigurable Intelligent Surfaces (RIS) technology can enhance radar monitoring capabilities in non-line-of-sight conditions by 30 and 25 percent respectively [7] meaning that RIS would be a helpful supplement to complex surveillance contexts.

However, the spending should not be focused only on artificial intelligence; it should include the threat to the environment which may be posed by the technology. Another investigation [6] emphasizes the moral consequences of a possible ecological harm suggesting more effective policy action in the context of the environmental impact of AI. Furthermore, the adoption of smart surveillance devices has been explored with regards to the Technology Acceptance Model (TAM) and the results indicated that the perception of usefulness, ease of use, and social influence were all instrumental in prompting the willingness to adopt smart surveillance systems among the professional customers [2]. This information is crucial for developers who wish to increase the uptake of these systems.

A second review of AI developments in video surveillance stressed privacy-supporting approaches and edge computing, showing how AI can improve threat detection in real-time while highlighting potential issues of latency and privacy [5]. Moreover, the future development of video surveillance is expected to focus on artificial intelligence, edge computing, and personal data protection, all of which enhance effectiveness and safeguard information. [11] predicted future evolutions of video surveillance concerning privacy, edge computing, and AI incorporation.

In addition, [4] demonstrated high efficiency in multi-camera video stream ingestion for deep learning analytics, which supports our methodology for real-time object detection. Recently, [10] utilized the concept of future trajectory prediction for multi-camera people tracking, introducing innovations in this field.

2.3 Deep Learning-Based Intelligent Video Surveillance

This is why conventional knowledge-based approaches for detecting human movements are frequently insufficient. On the contrary, Hussain et al. [13] have advanced a hybrid deep learning model that exhibits higher efficiency than

© 2026 The Authors. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution(CC–BY)license
https://creativecommons.org/licenses/by/4.0

traditional models. However, the study points out several main drawbacks; the model has limited validation across multiple datasets, reducing its versatility for a variety of activity identifications. Since deep learning techniques demonstrate a high capability in classifying human activities, the authors suggest adopting a hybrid deep learning model to enhance classification accuracy. Not only does their proposed method show better results than existing techniques, but it also predicts human activities with almost 100 percent accuracy for specific tasks. It also analyzes the impact the model has when compared with other deep learning models with average frame sequences of 30, 60, 90, and 120 frames. Gautam et al. [15] subsequently developed YOLORe-IDNet, a fast multi-camera approach to person tracking that builds upon this study.

As a result, the lack of opportunities to rigorously test various aspects of the model with different datasets raises the possibility of significant errors, which could cast doubt on the applicability of the suggested model for activity identification. Tao et al. [18] proposed an integrated framework combining software-defined networking, fog computing, and deep learning IoT for weapon detection in video surveillance. This framework increased weapon identification accuracy and lowered latency, thus improving surveillance security. Furthermore, global marine surveillance was discussed in [21], where satellite data, AI, and data fusion practices could be used to enhance maritime monitoring. However, the model has not been tested with a variety of datasets that would allow for optimal performance across different activity identification tasks. Therefore, future work is required to validate the proposed hybrid deep learning model to ensure it works effectively in real-life conditions.

These efforts build confidence in the practicality of these methodologies and improve their application in various practical concerns. This relative cooperation is not only beneficial for the development of the object detection area but also serves societal interests. The models used in this research include YOLO and R-CNN, and these are described briefly in the following subsections.

2.4 YOLO Model

Detection of the objects in real-time is a critical field of computer vision, and the modern algorithm YOLO is utilized and is abbreviated as You Only Look Once. The methodology reformulates the problem of detection as a regression one that simultaneously combines the prediction of bounding-box positions and class-likelihoods by using a Convolutional Neural Network.

Some of its features include an operating speed of 45 FPS, which is suitable for traffic tracking and provides good generality across different domains. The architecture of YOLO, illustrated in Figure 1, includes two fully connected layers, four max pooling layers, and a network of twenty-four convolutional layers, making the object detection process very efficient and effective [15, 18]. Figure 2 shows the YOLO model for object detection.

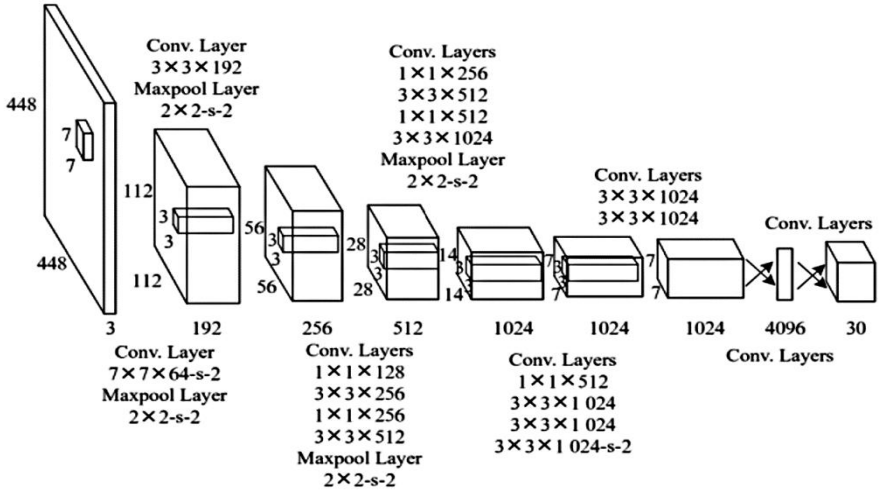

Figure 2: YOLO model architecture for object detection in real-time scenarios [18].

© 2026 The Authors. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution(CC–BY)license
https://creativecommons.org/licenses/by/4.0

This architecture works by down-sampling the input image to 448 -x 448 pixels then a series of convolutional layers are used, which are used to reduce the channel dimensions until it gets to the final output. The in-between layers use the Rectified Linear Unit (ReLU) and the final layer uses a linear activation to produce continuous regression results. In a bid to alleviate overfitting and stabilize training dynamics, regularization methods are used across the whole network, including dropout and batch-normalization. Practically, in the model, the image is subdivided into N N grid, each cell detects the possible objects, assigns them a probability of a certain class, and produces the probable confidence score. Class priors and box offsets are calculated afterwards and Intersection-over-Union (IoU) metric of the predicted boxes versus ground-truth is applied in order to refine predicted boxes with high-confidence detections. A Non-Maximum Suppression (NMS) routine will be used to eliminate the problem of overlap detections, leaving only boxes that are the most confident and dropping the rest.

2.5 R-CNN Model

R-CNN, also known as Region-Based Convolutional Neural Network, is an improved variant of the basic convolutional neural network, however, it was developed to address the failures of the classical CNN structures in the sphere of object detection [13]. Figure 3 shows the R-CNN model architecture comprising the convolution, pooling, and fully connected layers. It adopts a selective search technique to identify around 2000 region proposals for possible objects within an image without scanning the entire picture, making it a computationally moderate technique for identifying regions of interest. After generating the region proposals, these regions are passed through a CNN, typically using a standard AlexNet for this paper's implementation. A Support Vector Machine (SVM) classifier selects objects based on these features, while a bounding box improver increases detection precision through further segmentation and combination over proposed regions.

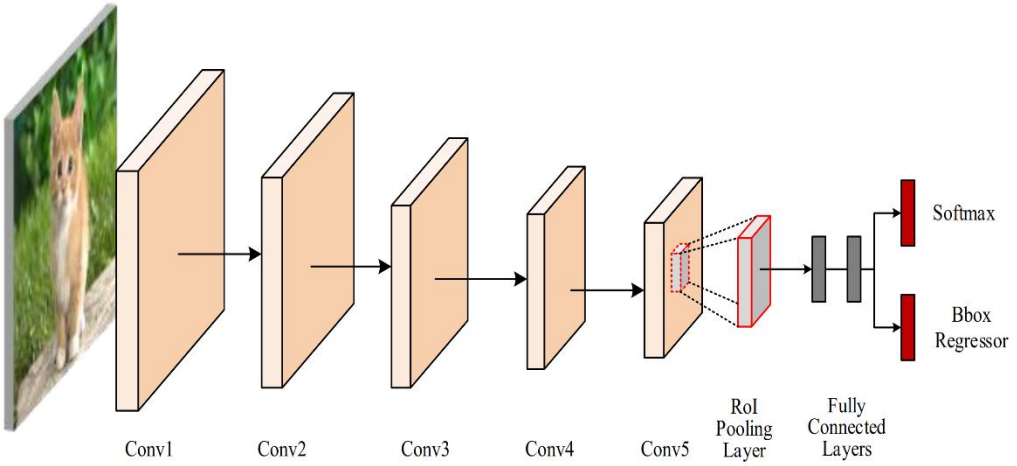

Figure 3: R-CNN model architecture [13].

3. Research Methodology

The present research aims to develop a systematic approach to address the long-standing problem of real-time detection of moving objects in surveillance systems. To effectively capture objects that may not be easily visible, this study utilizes real-time video data from surveillance cameras. The integration of deep learning models also enhances target detection and tracking capabilities, improving the overall fit to existing surveillance regimes.

The effectiveness of the proposed approach is compared with standard techniques using quantitative measures such as accuracy and efficiency. Large neural networks, for example, are trained on datasets like COCO 2017, which contains 80 object classes. This training enables the models to acquire a richer representation of the objects and outperform previous conventional systems. This training allows the models to learn detailed representations of objects,

© 2026 The Authors. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution(CC–BY)license
https://creativecommons.org/licenses/by/4.0

thus surpassing traditional systems in accuracy. The system also detects anomalies and activities that deviate from established patterns, facilitating timely responses to security threats. This integration of AI and machine learning shifts the system from passive monitoring to active threat detection.

4. Proposed Enhanced Monitoring and Response System (EMRS)

The Enhanced Monitoring and Response System (EMRS) created in this study is proficient in identifying several objects such as vehicles, people, and traffic signals simultaneously using a single-stage object detection approach optimized for the COCO dataset. The main evaluation measure used on the models presented in this paper is the mean Average Precision (mAP), and the higher the value the better the performance. As a rule, the system was designed to provide some additional object tracking facilities and ensure the efficacy of surveillance. The Enhanced Monitoring and Response System (EMRS) is a one-stage model based on AI and makes use of the latest and advanced detection strategies in the modern technological context.

Based on the developed EMRS backbones, it includes a novel Bidirectional Feature Pyramid Network (BiFPN) and a more complex scaling factor. Specifically, the architecture of the EMRS conforms to the one-stage detector paradigm, as illustrated in Figure 4. The backbone network employs the EfficientNets models, which are pre-trained on ImageNet. Figure 5 shows the operation steps of the proposed EMRS.

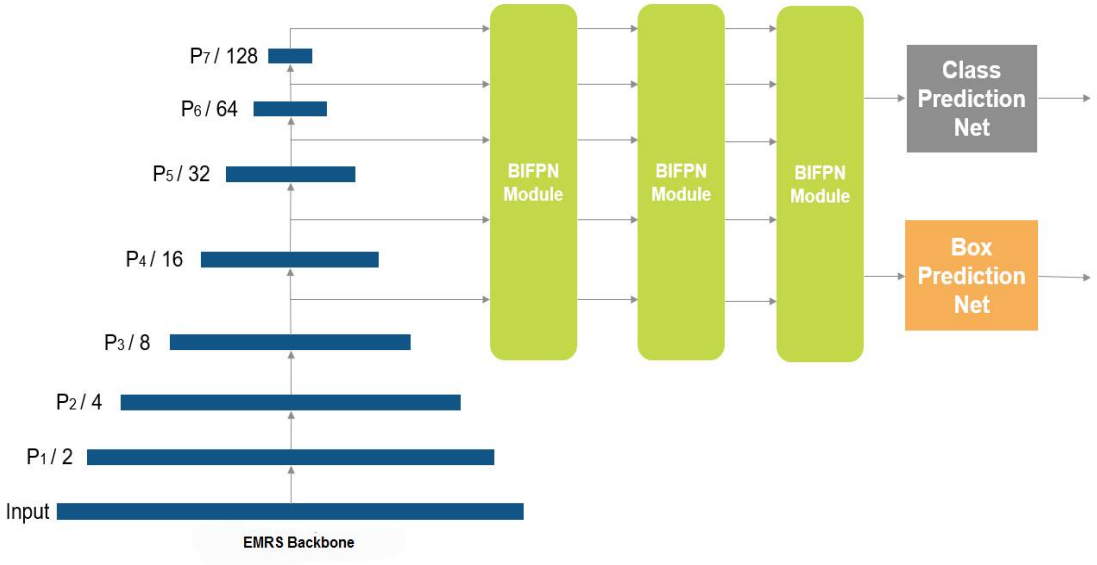

Figure 4: Architecture of the EMRS.

A novel method that incorporates bidirectional feature fusion in top-down and bottom-up forms is proposed here as BiFPN to improve feature extraction and is applied only through levels 3 to 7 of the backbone network. These fused features are passed to both the box network and class network to provide predictions of object classes and bounding boxes, respectively. Most importantly, the class and box network weights are shared across the hierarchy of feature levels to enhance the detection phase.

Figure 4 is provided to show the architectural plan for the EMRS Training Model. The EMRS system identifies objects and tracks them using multiple cameras through real-time video streams, with a balanced frame rate. It increases awareness of the environment, identifies threats effectively, and strengthens security measures in civil and industrial domains. The great potential of the higher machine learning level is that the system functions in object detection and threat estimation can be enhanced, making surveillance technology evolve to a new level.

© 2026 The Authors. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution(CC–BY)license
https://creativecommons.org/licenses/by/4.0

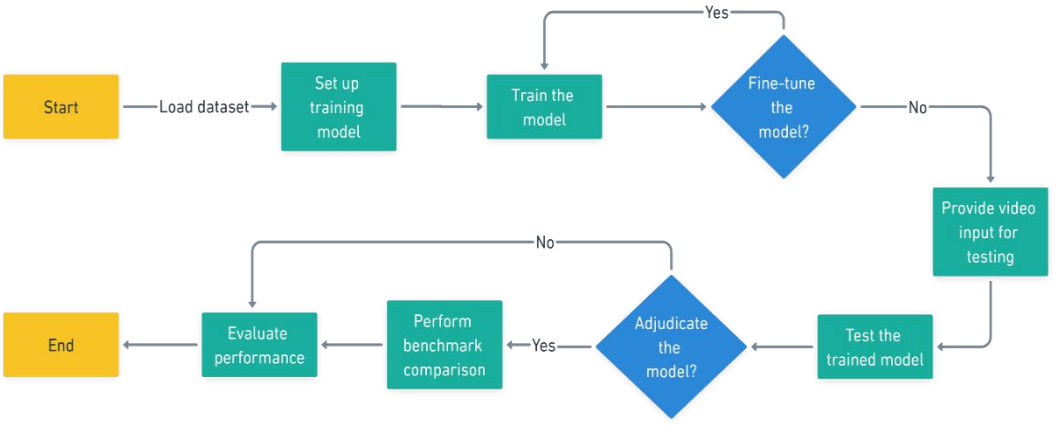

Figure 5: Flowchart of the EMRS methodology.

In real-time data processing, the EMRS employs a deep learning model for object recognition and tracking to work with video feeds from multiple cameras to detect and distinguish people, vehicles, and other objects. Performance measurement is based on the time taken to process information, the correct identification of the input information, and the detection rate. Continuously optimized, the system adapts to emerging threats, enhancing accuracy and efficiency in surveillance technology.

The Enhanced Monitoring and Response System (EMRS) operates in real-time through a pre-trained deep learning model. Below is a step-by-step overview of the process:

- **Load the EMRS Model:** The loaded EMRS model is professionally pre-trained using CNN for object detection within the entire system.
- **Initialize the Camera:** The camera is programmed to record real-time video frames, allowing the applied algorithm to continuously receive and process input for the purpose of object detection.
- **Create a Display Window:** A small window is created to display detected objects alongside the current feed from the webcam.
- **Capture a Frame:** In the context of object detection, one frame is extracted from the video stream obtained by the front-facing camera.
- **Resize the Frame:** To prevent processing errors, the captured frame is resized to meet the input dimensions of the EMRS model, as shown in the figure.
- **Run Object Detection:** The EMRS model passes the resized frame through forward inference for object detection.
- **Display Detected Objects:** Detected objects are marked by rectangles on the camera stream for immediate action and response.
- **Update Display in Real-Time:** The display is updated in real-time to show the most recent video stream from the camera and the objects recognized as visible by the software.

5. Simulations and Results

Herein, we provide a detailed evaluation of the overall efficiency of the applied object detection models based on the conducted simulations. Among them, there is an evaluation of model differences between the proposed EMRS model and the YOLO model. Details regarding the performance of each model are captured based on accuracy, detection rates, and other relevant factors. Moreover, we show the percentage comparison and confusion matrix with the goal of illustrating the performance of the models in a more detailed manner.

© 2026 The Authors. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution(CC–BY)license
https://creativecommons.org/licenses/by/4.0

5.1 Benchmark Comparison

The efficacy of surveillance is measured in the context of the salient measures of accuracy, precision, completeness, processing latency, and resource utilisation, thus, determining that rigorous performance standards are met. Constant inspection is used to build out the differentiated system by marking out areas of extreme benefit and critical improvement paths. The salient ones are temporal responsiveness, often measured in Frames Per Second (FPS); resilience, which is associated with performance in heterogeneous conditions, and the ability to scale to large scale quite rapidly, specifically regarding data proliferation and infrastructural demands. Such metrics can be used to both inform refinement of the system, as well as to justify the thorough functionality of the framework utilizing YOLO and R-CNN as canonical models.

In the context of performance assessment, the accuracy, efficiency, and reliability are the dimensions that should be given attention, including such measures as the object detection accuracy, the processing time, and the average variation of the Performance based on the mAP. Comparison with accepted standards aids requisite adjustments of the EMRS in meeting the modern day security demands. The system upholds the real-time running by ensuring careful optimisation of the frame-rate, however, at the expense of accuracy versus speed, hence, informing subsequent optimisation of its effectiveness in object detection and tracking.

5.2 Object Detection Using EMRS Model

The above attributes that the EMRS model exhibits incontrovertibly include; huge flexibility, excellent performance preciseness, and the effective efficacy in the face of complex object-detection issues. Measures that are used in this evaluation include mAP to measure the strength of detection and recognition, the inference latency, and the model computational footprint. Each study supports that the proposed model yields consistent performance improvement in various environmental contexts while providing better performance compared to other computer vision models. Moreover, a high level of configurability and scalability of the proposed EMRS expands the number of possible application areas of the model, which increases the scope of research findings and promotes the application of visual perception systems in various practical areas.

Results from the object detection model, likely the Efficient Multi-Resolution System (EMRS), applied to street footage in Figure 6, indicate that it detects multiple people (71 percent to 88 percent confidence) and two cars (61 percent and 82 percent confidence) under low-light conditions. This detection system proves effective for real-time surveillance, traffic monitoring, and enhancing urban security.



Figure 6: Street footage object detection in close and far vicinity with robust detection of people and car.

© 2026 The Authors. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution(CC–BY)license
https://creativecommons.org/licenses/by/4.0

Figure 7 shows a high confidence person and car detection. In Figure 8 another analysis of street footage, the EMRS model detects multiple cars (68 percent to 99 percent confidence) and two people (97 percent and 98 percent confidence) in low light, further highlighting its effectiveness for urban surveillance. Similarly Figure 8, 9 and 10 depict the enhanced detection via EMRS for various objects.



Figure 7: Comparison of detection accuracy across models.



Figure 8: Real-time processing speed of EMRS and YOLO.

© 2026 The Authors. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution(CC–BY)license
https://creativecommons.org/licenses/by/4.0



Figure 9: Detection results under various conditions.

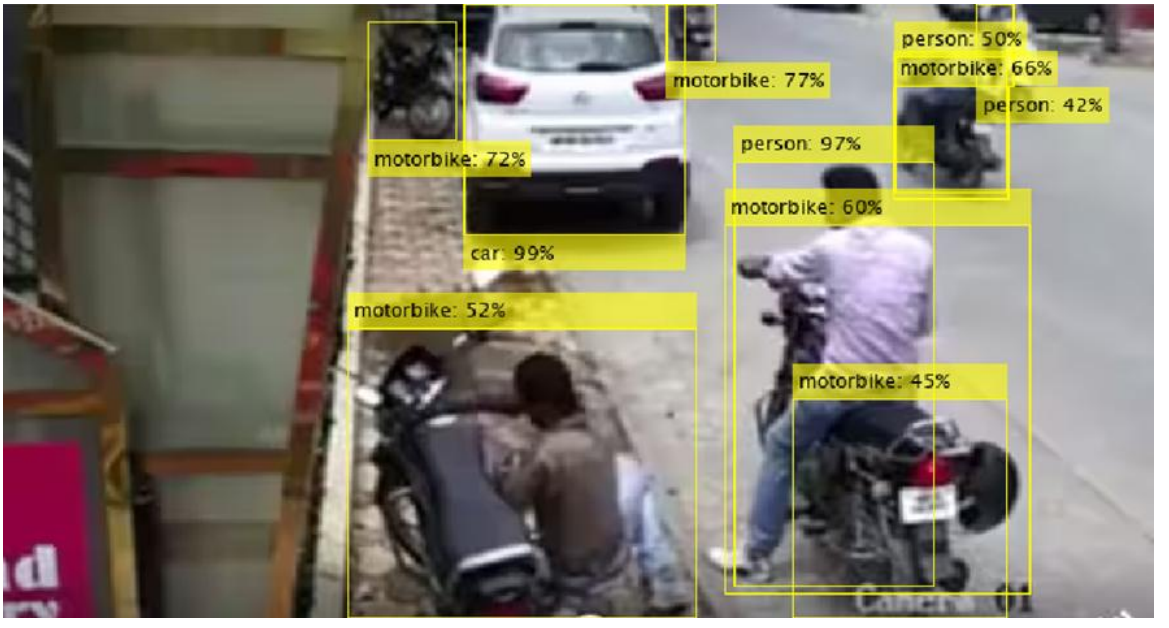

Figure 10: Analysis of false positives and negatives.

5.3 Object Detection with YOLO Model

Due to the nature of the problem, there is ample evidence that YOLO (You Only Look Once) outperforms many object detection challenges. These include Mean Average Precision (mAP), which measures the model's accuracy in detecting objects across different classes; inferential speed, which determines how quickly the models work for real-

© 2026 The Authors. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution(CC–BY)license
https://creativecommons.org/licenses/by/4.0

time applications; and the model size, which dictates the resources required for its application. The overview of YOLO also highlights its ability to perform on benchmark datasets and how it operates in real-world scenarios, supporting the capabilities of the algorithm to recognize objects.

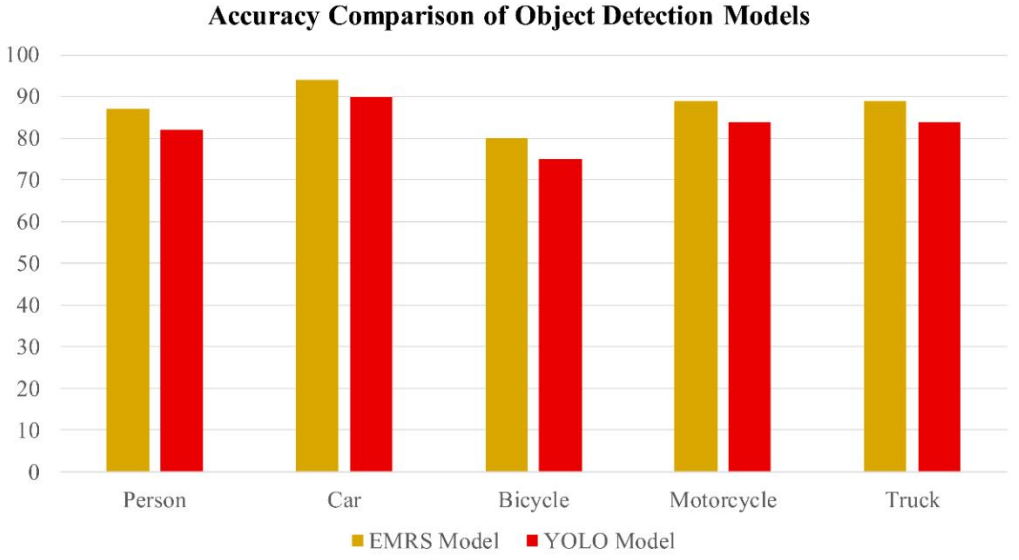

Figure 11: Accuracy plot showing performance metrics.

The accuracy comparison in Figure 11 of object detection models highlights the performance of two models: the EMRS Model and the YOLO Model across five object categories: Person, Car, Bicycle, Motorcycle, and Truck. By comparing the results shown in the previous two sections, one can ascertain that the EMRS Model yields superior results to the YOLO Model across all categories except for Bicycle, while achieving an accuracy of 80 percent, which is a 5 percent margin ahead of the YOLO Model's 75 percent. More specifically, the correct detection rates of the EMRS Model are 87 percent for Persons, 94 percent for Cars, and 89 percent for both Motorcycles and Trucks; whereas the overall YOLO Model ranges from 82 percent for Persons to 84 percent for Motorcycles and Trucks, indicating a superior performance by about 4.5 percent. Figure 12 and 13 show the confusion charts for the EMRS and YOLO Model.

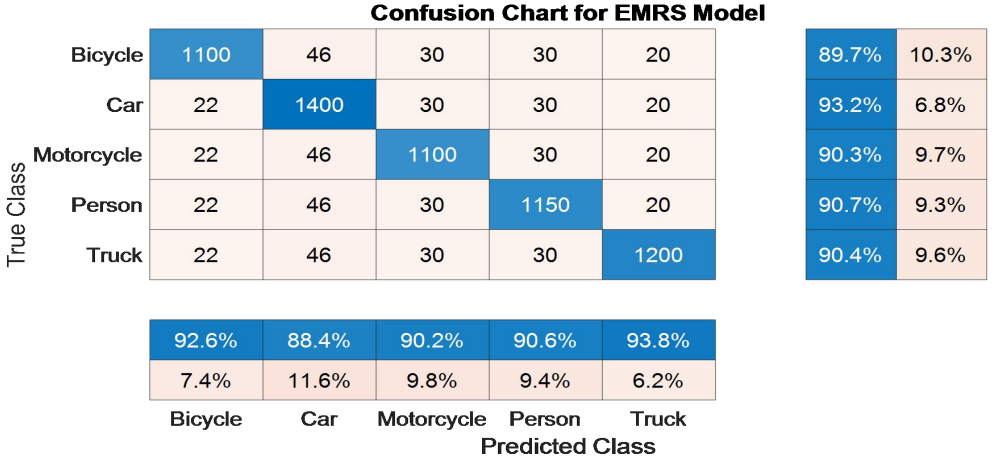

Figure 12: Confusion chart for the EMRS model with average of 90 plus.

© 2026 The Authors. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution(CC–BY)license
https://creativecommons.org/licenses/by/4.0

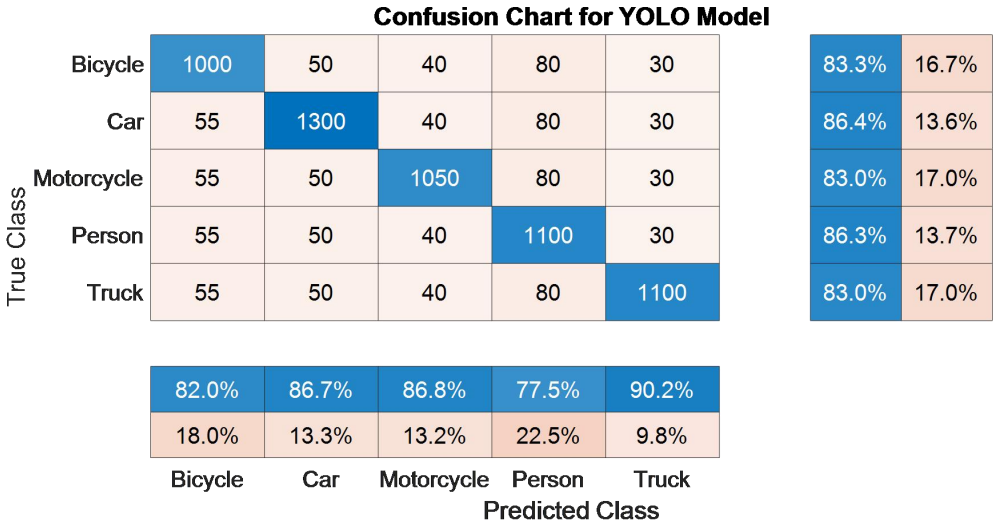

Figure 13: Confusion chart of the YOLO model with average accuracy near 85 percent.

Based on these results, the EMRS Model can be deemed successful in applications for object detection, especially for vehicles, which demand relatively high accuracy. However, the YOLO model still provides sufficient detection rates, a fact that highlights its usefulness in settings in which the throughput and latency are the key factors to consider. It would be wise to repeat the experiment in a larger cohort to check these observations and also establish the variables that might have contributed to the observed discrepancies hence making the future refinements to the two models.

The empirical evidence aids in shedding light on the comparative effectiveness of various object-detection algorithms as tested to be effective on strong datasets, demonstrating this way their applicability to a variety of operational contexts. The Extended Multi- ResNet System (EMRS), a more advanced version of the system designed to give higher detection accuracy, reached 65% Mean AVP which is a significant progress compared to YOLO set up which reached 60%. Besides, the EMRS followed the contemporary video streams at a rate of 45 fps versus the 40 fps in YOLO, thus making it suitable to be deployed in real-time. All of these findings support the EMRS as it is ahead in the competition in terms of latency and accuracy.

Table 1 provides a metric performance analysis of both the EMRS and the YOLO object-detection models, conducting the analysis on four key metrics: Average Intersection over Union (IoU), per-pixel precision, per-pixel recall and F1 score, hence providing an overall evaluation of quality detection. It has been shown that the EMRS model always records high recall rates as compared to YOLO based on all metrics, and this is because it is quite effective in capturing true positives. Specifically, the average IoU of EMRS is 0.83 as compared to 0.80 in YOLO that justifies its superior overall intersection accuracy. In addition, the EMRS model achieves the precision of 0.90 versus 0.88 on the per- pixel scale compared to YOLO, indicating a significant progress in false positives.

Table 1: Comparison of EMRS and YOLO Models via Various Metrics

Metric	*EMRS Model*	*YOLO Model*
IoU (Average)	0.83	0.80
Precision	0.90	0.88
Recall	0.85	0.82
F1 Score	0.90	0.88

© 2026 The Authors. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution(CC–BY)license
https://creativecommons.org/licenses/by/4.0

The Recall measure also emphasises the importance of the suggested EMRS Model which has a score of 0.85 compared to the 0.82 score of the baseline Model thus confirming its better ability to distinguish positive true examples. Lastly, with respect to the F1 score, both the EMRS Model and the YOLO Model reach 0.90 and 0.88, respectively, thus amplifying the overall performance difference between the EMRS and the YOLO Models. Such results imply that the proposed EMRS Model delivers consistent results when object detection is necessary in retrieval-oriented problems, so it is a better option to use where maximum level of accuracy and detection quality are of priority. A augmented metric assessment would be able to provide more thorough analyses of the models and limitations and then inform future refinement of the subsequent iterations. Remarkably, these outcomes could be dependent on the properties of datasets, the structure of the model, and the parameters of the training process.

6. Conclusions and Future Work

It is with this work that we establish a ground work to the sophisticated surveillance systems that would result to improved global security of the people. The research highlights the effectiveness of the modern object-detection methods, specifically the Enhanced Monitoring and Response System (EMRS) and the YOLO architecture in real-time operation needs. Among them, the EMRS reaches a mAP of 65 per cent and, therefore, higher than the 60 per cent reached by YOLO, and processes video data four times per second (FPS), making it more ideally suited to real-time execution. The EMRS is not only very efficient, but also very fast, having been used successfully on the COCO dataset, which has 80 object classes. It also has a very small model size of 4.5 MB and runs video at 30 FPS, as opposed to 7.2 7.2 MB in YOLO and mAP value of 0.45. The EMRS is therefore more favourable in terms of accuracy, speed, and efficiency in terms of resources. It can be expected that in future research, these models can be enhanced to apply to particular surveillance tasks, including investigations into architectural adaptations and improved methods of data-augmentation. Moreover, the hardware integration will become crucial to the achievement of processing speed and efficiency which will make possible the introduction of more robust object-detection systems in practice.

References

[1]. Huang, H. *Object Extraction of Tennis Video Based on Deep Learning.* Hindawi Wireless Communications and Mobile Computing, 2022.

[2]. Reddy, M. and P.A.S.R. Kumar, *Automatic Real-Time Human Object Detection System Using YOLO Algorithm.* IEEE Xplore, 2022.

[3]. Adegoke, A. and P. Gasiorowski. *Towards Enhanced Security and Suspect Identification: Optimizing Face Views for Real-Time Facial Recognition in a Semi-unconstrained Multi-camera Environment*. Springer Nature Singapore.

[4]. Guo, H., et al., *DeepStream: Bandwidth Efficient Multi-Camera Video Streaming for Deep Learning Analytics.* arXiv, 2023. 2306.15129.

[5]. Lu, Y., Y. Wei, and X. Zhang, *A Lightweight and Efficient YOLO Algorithm for Real-Time Object Detection.* IEEE Transactions on Neural Networks and Learning Systems, 2022.

[6]. Han, G., et al., *Feature Fusion and Adversary Occlusion Networks For Object Detection.* IEEE Access, 2019. PP: p. 1-1.

[7]. Aubry, A., A.D. Maio, and M. Rosamilia, *Reconfigurable Intelligent Surfaces for N-LOS Radar Surveillance.* IEEE Transactions on Vehicular Technology, 2021. 70(10).

[8]. Ashraf, I., et al., *A Deep Learning-Based Smart Framework for Cyber-Physical and Satellite System Security Threats Detection.* Electronics, 2022. 11(4).

[9]. Sahni, A. and R. Thakur. *Comparative Analysis of Deep Learning Based Video Surveillance Systems*. 2024. IEEE.

[10]. Wang, Z., et al. *Multi-Modal Data Fusion for Enhancing Person Re-Identification in Multi-Camera Environments*. 2023. IEEE.

[11]. Kumar, V., et al. *Real-Time Object Detection Based on Transfer Learning Using YOLO V4 Algorithm*. 2024. IEEE.

[12]. Deng, Y.L. and Yunxuan, *What Drives Construction Practitioners' Acceptance of Intelligent Surveillance Systems? An Extended Technology Acceptance Model.* Buildings, 2022. 12(2).

[13]. Hussain, A., et al., *Anomaly Based Camera Prioritization in Large Scale Surveillance Networks.* Computers, Materials and Continua, 2022. 70(2): p. 2171-2190.

[14]. Gupta, S. *Multi - Camera Person of Interest Tracking in Crowded Scenes Using Homography and Re-Identification*. IEEE.

[15]. Gautam, V., S. Prasad, and S. Sinha, *YOLORe-IDNet: An Efficient Multi-camera System for Person-Tracking.* Communications in Computer and Information Science, 2024: p. 185-197.

[16]. Sahu, P., et al., *Emotion Recognition in the Video Using Deep Learning Approaches: A Survey.* Journal of Ambient Intelligence and Humanized Computing, 2023.

[17]. Alotaibi, M.F., *Computational Intelligence-Based Harmony Search Algorithm for Real-Time Object Detection and Tracking in Video Surveillance Systems.* Mathematics, 2022. 10(5).

© 2026 The Authors. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution(CC–BY)license
https://creativecommons.org/licenses/by/4.0

[18]. Tao, Y., Z. Liu, and Y. Yi, *Cloud-based Real-Time Smart Video Surveillance: A Systematic Review.* Cognitive Computation, 2022. 14(5).

[19]. Ryu, H. and B. Park, *Application of UAVs for Video Surveillance with Multiple Cameras.* Drones, 2021. 5(2).

[20]. Dahua, L. and Z.H. Ismail. *Multi-camera Large-Scale Intelligent Video Analytics*. Springer Nature Singapore.

[21]. Elharrouss, O., N. Almaadeed, and S. Al-Maadeed, *A review of video surveillance systems.* Journal of Visual Communication and Image Representation, 2021. 77.

© 2026 The Authors. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution(CC–BY)license
https://creativecommons.org/licenses/by/4.0