Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 4, Issue 1, 2026

REAL-TIME AUTO START-STOP YOLO V11 SINGLE SHOT DETECTION APPROACH FOR TRACKING AND SURVEILLANCE SYSTEM

Attaullah Narejo*[1], Ali Nawaz Sanjrani*[2], Nouman Qadeer Soomro[3], Arshad Ali Matilo[4], Ali Bux Narejo[2]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, (UESTC) Chengdu, 611731, Sichuan, China.
[2]Department of Mechanical Engineering, Mehran University of Engineering and Technology, MUET, SZAB Campus, Khairpur Mir's, Sindh, Pakistan.
[3]Department of Software Engineering Mehran University of Engineering and Technology, SZAB Campus, Khairpur Mir's, Sindh, Pakistan
[4]Department of Informtica, Univeristy of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano (SA), Italy.
*[1]attaullahrehannarejo786@gmail.com, *[2]alinawaz.sanjrani@muetkhp.edu.pk

DOI: https://doi.org/10.5281/zenodo.18323194

Keywords

Automotive vehicle; Artificial Intelligence; YOLO v11; Advanced Driver Assistance Systems; Real-Time monitoring; Detection; Transformers

Article History

Received: 22 November 2025
Accepted: 03 January 2026
Published: 17 January 2026

Corresponding Author: *
Attaullah Narejo and Ali Nawaz Sanjrani
both contributed equally to this study.

Abstract

In this paper, the automotive vehicle systems auto stop and start have profoundly transformed the detection and tracking of human presence, enhancing safety and operational efficiency by using artificial intelligence (AI). Accurate vehicle detection advances intelligent transportation, autonomous driving, and traffic monitoring systems. This study explores the capabilities of YOLOv11, the latest innovation in the YOLO deep learning model series with integration of transformers for fast learning and decision making system, tailored explicitly for vehicle detections. Developed on the strengths of its predecessors, YOLOv11 incorporates significant architectural enhancements that improve detection speed, accuracy, and robustness in dynamic and complex traffic scenarios. This study analyzes accident prevention strategies by adopting auto start and stop methodology based on advanced object detection algorithms through YOLO v11. The proposed model examines the real-time application of model adaptability in conjunction with GPS technology, emphasizing their critical roles in bolstering vehicle safety and reducing accident occurrences. The findings demonstrate that YOLO v11 significantly surpasses traditional approaches in speed and accuracy with 99.75%. The results contribute to ongoing advancements in refining Advanced Driver Assistance Systems (ADAS) and enhancing the overall driving experience.

1. Introduction

In the current decade, the proliferation and rapid advancement of technologies have catalyzed significant innovations in an automotive safety systems and object tracking detection. However, the identification of vehicles is a fundamental and critical component of modern intelligent networks, where object detection and tracking play a crucial role. As the tracking relies on precise and instantaneous real-time data that improve traffic management, ensure safety, and enable surveillance systems for self-driving vehicle technologies (Sun et al., 2006). Machine learning supports artificial intelligence (AI) to brought the significant changes, fundamentally transforming the modern world's approach to vehicle interaction with its surroundings, and enhancing every possible human detection and real-time tracking capability in the latest technologies (Olugbade et al., 2022). Integration of AI in automotive applications made significant improvements in safety measures. It offers possible solutions to critical pressing challenges, such as theft prevention and accident reduction in real situations. The core of AI transformation is based on the capacity for detection models, which is functioning critically in different scenarios, including autonomous driving, personal vehicle security, and safety (Bhardwaj, 2023). Previous traditional human detecting models are often rely on sensor inputs and basic algorithms, which are somewhat limited in their efficacy (Alsheikh et al., 2014). Human and vehicle detection, with the rise of machine learning, is dependent on sensor data and rudimentary algorithms, which can be seen in limitations in effectiveness. The machine learning techniques suggests the researchers to develop an adaptive methods that can extract detailed features extraction during the real time data domain. Previously It has been seen that the shift gained observed significant momentum with the advancement of deep learning, primarily through the use of the Convolutional Neural Networks (CNNs) model (Aydin et al., 2023). These models have revolutionized the field by facilitating and enabling direct learning, thereby reducing dependence on manual feature extraction. In-depth learning-based approaches are addressed in the earlier limitations and establish protocols for limits using CNN architectures. Among these earlier creations, the YOLO family series models emerged as a pioneering and transformative solution, renowned for its high accuracy in real-time detection. The starting YOLO model redefined and refined detection and diagnosis by using a regression problem, that may allow the researchers for the simultaneous prediction of class probabilities and bounding boxes from raw image pixels and beyond raw data (Redmon, 2016). This single-stage framework offers and provides a momentous speed advantage over previous traditional models' two-stage detectors as the Region-CNN (R-CNN) (Alif & Hussain) and Faster R-CNN (Redmon & Farhadi), that require several networks to pass through to generate and refine model performance.

The latest YOLO evolution has achieved significant milestones, with each version after its predecessors to enhance the capabilities of the historical and previous ones. YOLOv1 was among the first to introduce the grid-based approach to predict bounding boxes and class probabilities for each cell data. (Alif & Hussain, 2024) The YOLOv2 and YOLOv3 are the initial models that improve accuracy for the small and complex object patterns by incorporating and adding features such as full batch normalization, anchor boxes in the raw pixels, And multi-scale detection. (Redmon, 2018; Redmon & Farhadi, 2017). YOLOv4 and YOLOv5 included better and enhanced features for the extraction and fusion techniques through advancements in Innovations such as both Cross-Stage Partial Networks (CSPNet) and Path Aggregation Networks (PANet) (Bochkovskiy et al., 2020; Jocher et al., 2022). While the focus of YOLOv6 and YOLOv7 shifted toward maximizing prioritized optimization based on speed of inference and computational efficiency, making them especially perfect for real-time applications (Li et al., 2022; Wang et al., 2023). YOLOv8 expanded its utility and enhances the performance by using the segmentation and tracking, while introducing hassel free detection mechanisms for greater generalizability across diverse datasets. (Jönsson Hyberg & Sjöberg, 2023). As YOLO, notably in its updated versions such as YOLO V11, demonstrates the ability to process images in real-time, making it invaluable for applications that require quick decision-making, such as collision avoidance systems. At the same time, the Vision Transformers (ViTs), has also pushed the boundaries of object detection and diagnosis (Dosovitskiy, 2020) and also leveraging self-attention mechanisms with attention scores, ViTs excel at capturing long-range dependencies in large-scale image recognition tasks. (Alif et al., 2024). However, due to real-time applications by viewing the latency constraints, such as autonomous driving and traffic monitoring suits the CNN-based models and also like YOLO models remain the preferred choice due to their efficiency and adaptability. YOLOv11's proficiency in vehicle detection offers significant potential for practical applications in systems demanding exceptional accuracy, efficiency, and real-time object identification. Its ability to manage various vehicle types, operate effectively under exciting ecofriendly conditions, and deliver high-speed inference makes it highly applicable in autonomous vehicles (Hussain et al., 2022), urban monitoring, and IoT sensor networks (Alsboui et al., 2022).

The building on the achievements of previous versions and the successes of its earlier model as predecessor, while YOLOv11 represented the latest progression and advancement in this series of models. It introduces innovative architectural enhancements, incorporating advanced attention mechanisms, deeper feature extraction, and an improved detection strategy. These enhancements and modifications further address significant priority challenges in the field, such as detecting smaller, occluded, or rapidly moving vehicles within frames of pixel data, while ensuring that all preserve real-time inference capabilities. Furthermore, in detail, YOLOv11 is advanced and engineered for hardware acceleration and speed, making it ideal for edge devices in critical applications, including intelligent surveillance systems and object recognition.

This paper offers a significant advantage of YOLOv11 as a notable breakthrough in vehicle detection and recognition by examining previous settings and parameters. It is worth mentioning here that the said proposed model successfully overcomes the obstacles of detecting smaller, complex patterns, occluded, and moving vehicles in real-time settings. The proposed model's performance is analyzed using the performance and maintenance check and assessment benchmark, with performance metrics such as precision, mean average precision (mAP), accuracy, recall, and F1 score, as analyzed here in section 3. This study of the proposed model YOLOv11 offers a detailed comparison with its predecessors, YOLOv8 and YOLOv10. This research clearly emphasizes on accuracy, and robustness of YOLOv11, underscoring its potential for intelligent surveillance considerations, like an autonomous driving and traffic monitoring, where accuracy and reliability are paramount for detection and diagnosis.

2. Methods

2.1. Dataset and Pre-processing

The dataset used for evaluating YOLOv11 consists of 1,321 annotated images representing various vehicle types, such as cars, trucks, motorcycles, buses, bicycles, and persons. Captured under diverse real-world conditions such as day/night variations, weather effects, occlusions, and varying object distances, offers a realistic representation of intelligent transportation environments. Each (416×416) image is annotated with bounding boxes and class labels to ensure accurate localization and classification. The dataset was split into three sections and given to the model for 80:10:10 training, validation, and for testing, that ensures a stable consistent distribution for reliable assessment of the YOLOv11 model.

2.2. Proposed Approach YOLO11 Architecture with Data Augmentation

YOLOv11 initially builds upon the basis of its predecessors. First, we introduce key architectural enhancements that significantly boost detection accuracy and computational efficiency, particularly compared to YOLOv8. These improvements and advanced enhancement include step-by-step advanced layers, more optimized blocks, and better-refined feature extraction techniques, making YOLOv11 exceptionally well-suited for real-time vehicle detection tasks. The complete inclusive data augmentation strategy was implemented during training and testing of proposed model to enhance the robustness and generalization capabilities which enables the model capabilities across various real-world scenarios, including environment conditions, lighting variations, and object orientations.

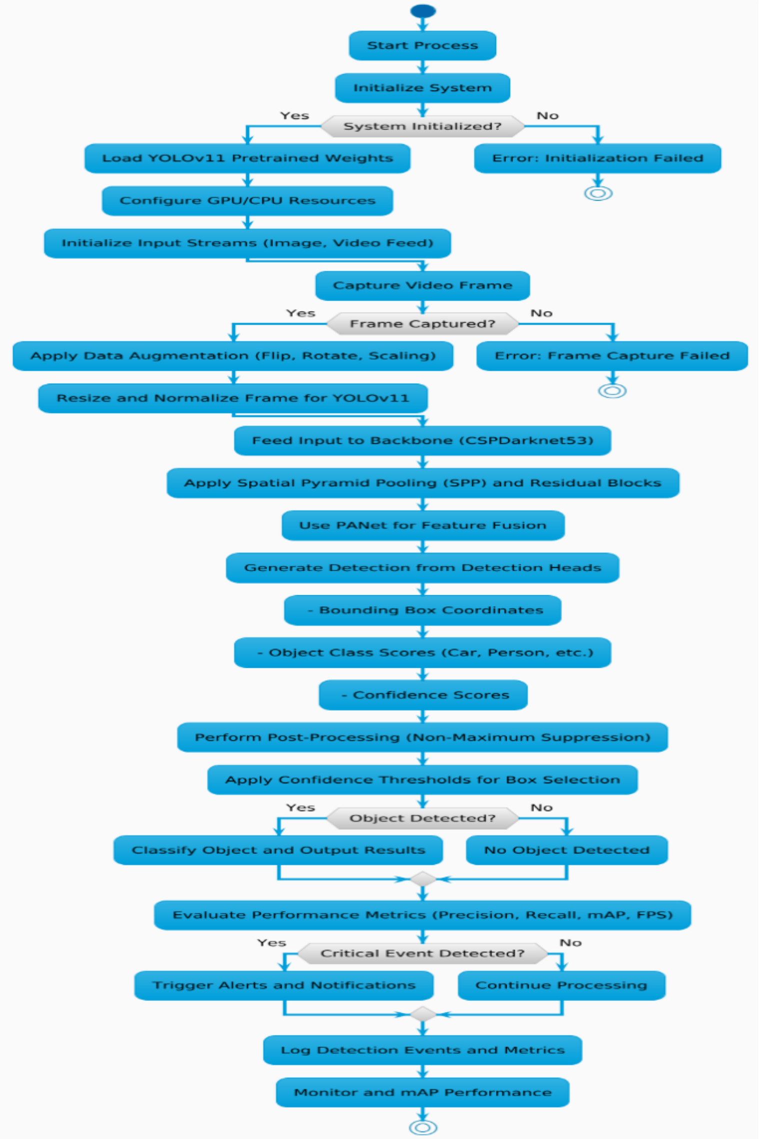

Start Process
Initialize System
Yes
System Initialized?
No
Load YOLOv11 Pretrained Weights
Error: Initialization Failed
Configure GPU/CPU Resources
Initialize Input Streams (Image, Video Feed)
Capture Video Frame
Yes
Frame Captured?
No
Apply Data Augmentation (Flip, Rotate, Scaling)
Error: Frame Capture Failed
Resize and Normalize Frame for YOLOv11
Feed Input to Backbone (CSPDarknet53)
Apply Spatial Pyramid Pooling (SPP) and Residual Blocks
Use PANet for Feature Fusion
Generate Detection from Detection Heads
- Bounding Box Coordinates
- Object Class Scores (Car, Person, etc.)
- Confidence Scores
Perform Post-Processing (Non-Maximum Suppression)
Apply Confidence Thresholds for Box Selection
Yes
Object Detected?
No
Classify Object and Output Results
No Object Detected
Evaluate Performance Metrics (Precision, Recall, mAP, FPS)
Yes
Critical Event Detected?
No
Trigger Alerts and Notifications
Continue Processing
Log Detection Events and Metrics
Monitor and mAP Performance

Figure 1. YOLOv11 Frame Work

This combination of architectural advance innovation and all inclusive data augmentation that establishes YOLOv11 as a robust solution for intelligent surveillance and autonomous driving. A spectrum of color fraction was altered to acquaint with variability in image hue, improving and enhancing the model's generalization to various lighting conditions and object hues.

$$P' = \text{adjust\_hue}(P, \alpha h) \qquad (1)$$

Where hue adjustment, αh is set to 0.015, and P denotes the original image. A fraction of the saturation is modified to vary color intensity, simulating different environmental conditions. It enhances the model's robustness in detecting objects across images with varying saturation levels.

$$P' = \text{adjust\_saturation}(P, \alpha s) \qquad (2)$$

where αs = 0.75 specifies the degree of saturation applied to the image.

Brightness was varied proportionally to the original image value, enhancing the model's detection capability to detect objects under different lighting conditions and situations, including daylight, nighttime, and shadowed settings in the images.

$$P' = \mathrm{adjust\_brightness}(P, \alpha v) \qquad (3)$$

where αv = 0.5 controls the brightness adjustment in the image.

The images were randomly interchanged within a fixed range of angles that increases model's fitness to recognize objects across varied orientations.

$$P' = \mathrm{rotate}(P, \theta) \qquad (4)$$

where $\theta \in [-180^\circ, 180^\circ]$ is the randomly chosen rotation angle. The image was translated horizontally and vertically by a fraction of its size, simulating partial object occlusion to enhance the model's capability in detecting objects that are not fully visible in real-world scenarios.

$$P'(a', b') = P(a + ta \cdot c, b + tb \cdot h) \qquad (5)$$

where t_a, t_b $\in [-0.1, 0.1]$, and c represent the width, and h represents the height of the image. A gain factor scaled the image to simulate varying object distances, improving detection across different object sizes.

$$P'(a', b') = \mathrm{scale}(P, \alpha s) \qquad (6)$$

Here, αs = 0.5 controls the scaling factor in the image. Shearing was utilized to mimic the experience of observing objects and check imbalance from various perspectives, therefore improving the model's capability to comprehend object distortions in the images.

$$P'(a', b') = \mathrm{shear}(P, \theta s) \qquad (7)$$

Here, $\theta s \in [-180^\circ, 180^\circ]$ represents the shearing angle. A random perspective transformation was applied to mimic 3D effects, enhancing the model's robustness to depth and orientation variations.

$$P' = \mathrm{perspective}(P, \alpha p) \qquad (8)$$

Here, αp = 0.001 controls the perspective transformation and transform. The image was inverted vertically with a fixed probability to enhance dataset changeability while preserving the objects' inherent features from the pictures.

$$P' = \mathrm{flipud}(P) \qquad (9)$$

with a probability of p = 0.0. image flipped horizontally with a confident likelihood to improve and stabilizes model learning of symmetrical objects and to enhance dataset diverseness.

$$P' = \mathrm{fliplr}(P) \qquad (10)$$

with a probability of p = 0.5. Mosaic augmentation was applied by combining four images, simulating complex scene compositions, and improving the model's understanding of diverse object interactions.

$$P' = \mathrm{mosaic}(P_1, P_2, P_3, P_4) \qquad (11)$$

P_1, P_2, P_3, P_4 are four images combined into one mosaic image.

The YOLO11 framework marks a notable and essential improvement compared to previous versions in the historical data, especially the last one, YOLOv8. It incorporates novel layers, modules, and refinements that together enhance computational efficiency and detection accuracy, making it highly effective and suitable for real-time applications such as vehicle recognition.

2.2.1. Backbone is improved

The backbone of YOLO11 is responsible for extracting meaningful and essential features from the input image at multiple scales. This involves a series of convolutional layers and custom blocks that generate feature maps at varying resolutions. YOLO11 integrates the C3k2 block, preserves the Spatial Pyramid Pooling Fast (SPPF) block from earlier versions, and incorporates additional enhancements through the C2PSA block. Further the conventional layers as Conv1 (1, 64, 3, 2) and Conv2 (conv1, 128, 3, 2), the input image or feature map, are processed with 64 filters, resulting in 64 feature maps of reduced spatial size due to the stride. The output of (64 feature maps) is processed with 128 filters, generating 128 feature maps with further reduced spatial dimensions. These operations are used in YOLO11 to progressively extract hierarchical features while lowering the input's spatial dimensions. This helps focus on high-level features and reduces the computational complexity of downstream tasks.

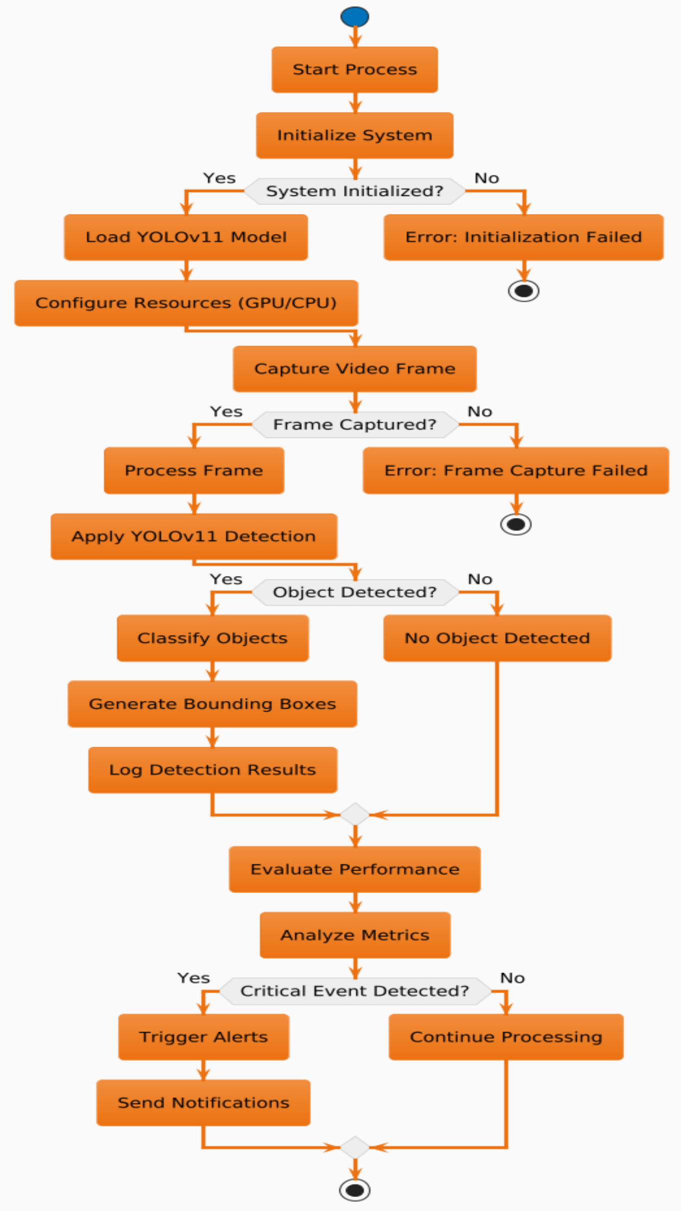


Figure 2 Process Flow Diagram of YOLO11

In contrast to the C2f block used in YOLOv8, YOLO11 adopts the more efficient C3k2 block, which is derived from the Cross-Stage Partial (CSP) network. The block of C3k2 features like convolutions with kernel size of 2 to lower complexity while maintaining performance. Block of C3K2 is $C3k2(X) = Conv(Split(A))Conv(Merge(Split(A)))$split (A) partitions the feature map into two streams, one propagated through the bottleneck layer and the other integrated with the resulting outputs.

In YOLO11, the SPPF block performs multi-scale spatial pooling by applying max pooling with kernel sizes of 5, 3, and 1, followed by the concatenation of the resulting feature maps. Formally, it is defined as:

SPPF(A)=Concat({MaxPool(A,k)∣k∈{5,3,1}}) (12)

YOLOv11 represent the C2PSX block which enhance the attention mechanism and the same is applied in spatial features map.

$$C2PSX(A) = Attention(Concat(A_{path1}, A_{path2})) \quad (13)$$

As it permits the model to focus over the specific region in the image, thereby enhancing its performance in detecting small and stop-up objects. YOLOv11 is enhanced to aggregate multi-scale feature representations and forward them to the detection head, as it incorporates the block of C3k2, which enhances the efficiency and effectiveness of feature aggregation. The features aggregation, where the model utilizes up-sampling and concatenation layers to merge feature maps across multiple scales in the environment.

$$F_{upsample(a)} = Upsample(F_{previous}) \quad (14)$$

$$F_{concat(a)} = Concat\left(F_{upsample}, F_{lower}\right) \quad (15)$$

$$C3k2_{neck(a)} = Conv_{small}(Concat(F_{concat})) \quad (16)$$

In routine the application of C3k2 block; afterward chain confirms an approach for well-organized feature aggregation where the YOLOv11, plays a crucial role in creating the model's final predictions, class probabilities, outputting bounding boxes, and confidence scores, much like its previous predecessors. The proposed architecture utilizes and uses detection layers at three various scales where everyone has its stage, at low (P3), intermediate (P4), and high (P5), that effectively recognize objects of multiple extents. By processing various feature maps at each level scale, YOLO11 ensures its capability and ability to detect small and large vehicles efficiently and accurately.

$$Detect(P3, P4, P5) = BoundingBoxes + Lclass \quad (17)$$

2.2. Hyperparameters Settings and Loss function

The training and validation Process is very critical and challenging to analyze for YOLOv11 following the same structure as earlier version as YOLOv8 and YOLOv10 model structure, analyses ensures consistency of framework, that allowing for straight performance comparison among both models. Key hyperparameters were carefully chosen to optimize and enhance YOLOv11's performance while maintaining computational efficiency and accuracy. The training configured with learning rate of η= 0.01, which gradually decayed across epochs using a cosine annealing schedule, represented by the following equation:

$$a_t = a_0 x 0.5(1 + cos(\frac{e}{E}\pi)) \quad (18)$$

Where with e denoting the running epoch, E the maximum epochs, and a batch size of 64 chosen to ensure optimal computational efficiency. Training employed a momentum of 0.972 and weight decay of 0.0005, with 100 epochs to ensure stable optimization and convergence. Table 1 highlights the key hyperparameters utilized as:

Table 1. Setting of Hyperparameters for YOLO11

Description of hyperparameters	Setting Value
Initial learning rate	0.01
Batch Size	64 samples
Momentum coefficient	0.972
Weight decay parameter	0.0005
Training duration	100 Epoch
Optimizer	SGD

2.3. Implementation and Visualization

In this study, the model was trained using the stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.01, batch size of 64, momentum coefficient of 0.972, and weight decay of 0.0005. Training was performed for 100 epochs. The multi-scale training strategy implemented on the dataset, and the selected size of input is randomly varied between 320 × 320 and 640 × 640 pixels, thereby enhancing the image and the model’s robustness across different image resolutions and object scales. Therefore, the primary evaluation metric was Mean Average Precision (mAP), calculated over multiple intersection over union thresholds from 0.5 to 0.95 to assess recognition performance.

$$\text{mAP} = \frac{1}{|\,\text{IOU Threshold}\,|} \Sigma_{\text{IOU Threshold}} \text{AP} \quad (19)$$

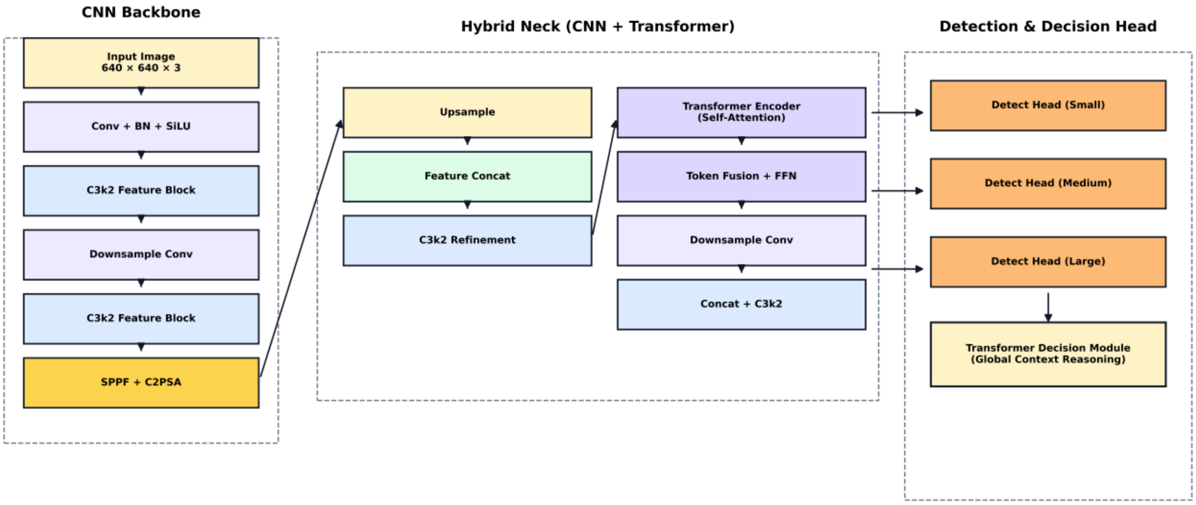


Figure 3 YOLOv11 Hybrid Architecture with Intelligent Decision-Making

Finally, the performance matrices are also calculated such as precision, recall, and attention scores are used to evaluate detection and diagnosis accuracy. Early stopping was applied to prevent overfitting and to retain the best model. This study ensured a consistent evaluation of YOLO11 compared to YOLOv8 and YOLOv10.

2.4. YOLOv11 Hybrid Architecture with Transformer-Based Decision Making

The proposed **YOLOv11 Hybrid Architecture with Transformer-Based Decision Making** integrates convolutional neural networks (CNNs) and Transformer modules to jointly enhance multi-scale feature representation, long-range dependency modeling, and global reasoning for object detection. The CNN backbone is designed to efficiently extract hierarchical spatial features from an input image of size $640 \times 640 \times 3$. It begins with a **Conv + Batch Normalization + SiLU** block to capture low-level visual patterns, followed by **C3k2 feature blocks** that improve feature reuse and gradient flow while maintaining computational efficiency. Downsampling convolution layers progressively reduce spatial resolution while enriching semantic representations. At the deepest stage of the backbone, **Spatial Pyramid Pooling–Fast (SPPF)** combined with **Channel-wise Partial Self-Attention (C2PSA)** aggregates multi-scale contextual information and enhances salient features, enabling robust and efficient feature extraction with minimal computational overhead as shown in Figure 3.

Building upon the backbone outputs, the **hybrid neck** combines CNN-based feature fusion with Transformer-based global modeling to refine multi-scale representations. Backbone features are upsampled and concatenated with higher-resolution feature maps, followed by **C3k2 refinement blocks** to ensure spatial and semantic consistency across scales. The fused features are then processed by a **Transformer encoder**, where self-attention mechanisms capture long-range spatial dependencies that are difficult for conventional CNNs to model. A **token fusion module and feed-forward network (FFN)** further refine the contextual representations. Subsequently, downsampling and **Concat + C3k2** operations align the features for the detection stage.

Finally, the **detection and decision head** employs three multi-scale detection branches (small, medium, and large) together with a **Transformer-**

based decision module that performs global reasoning across detection outputs. This module improves context-aware confidence estimation, reduces false positives, and ensures consistent classification in complex scenes. Compared to traditional CNN-only YOLO architectures, the proposed YOLOv11 hybrid design introduces Transformer encoders in both the neck and decision stages, achieving superior detection robustness while preserving real-time efficiency and scalability for edge and real-world deployment.

2.5. Experimental Setup of Proposed Methodology

The proposed model was trained using the SGD optimizer with an initial learning rate of 0.01, batch size of 64, momentum coefficient of 0.972, and weight decay of 0.0005, for a total of 100 epochs. A multi-scale training strategy was applied, where input images were randomly resized between 320 × 320 and 640 × 640 pixels to improve robustness across different scales and views. Data augmentation, including random brightness and contrast adjustments along with normalization, was employed to increase dataset diversity and reduce overfitting. Training and validation performance were monitored using early stopping, checkpointing, and TensorBoard visualization. All experiments were conducted on a laptop equipped with a 13th Gen Intel® Core™ i9-13900HX CPU (2.20 GHz) and an NVIDIA GeForce RTX 4060 GPU.

3. Experimental Results

The performance of YOLOv11 was evaluated on a vehicle detection dataset using different standard object detection metrics, including precision, recall, F1 score, mean average precision (mAP), inference time, and a confusion matrix.

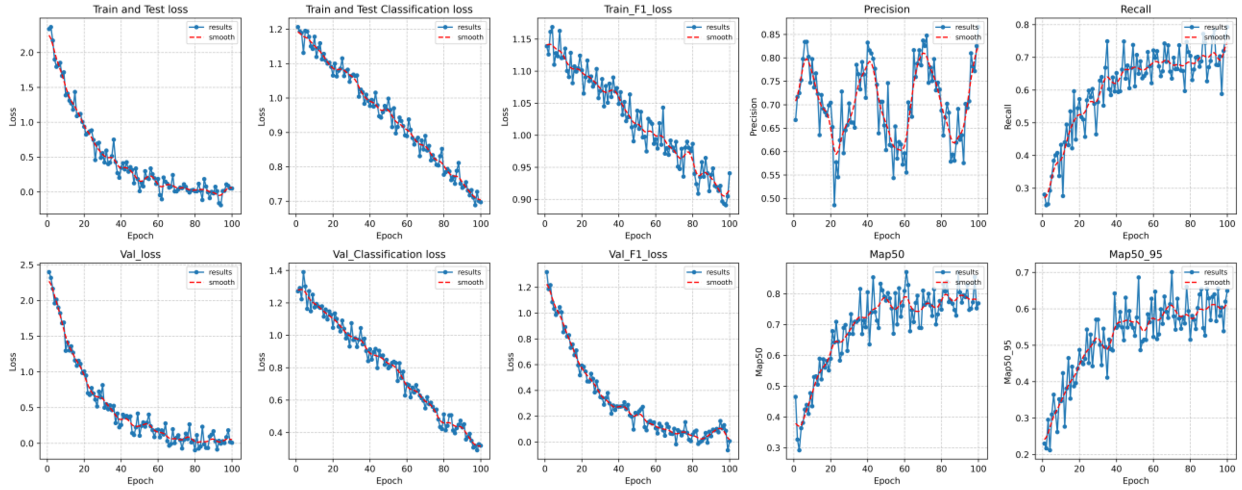


Figure 4. Training, validation, and testing performance metrics (F1 loss, precision, recall, mAP@50, and mAP@50–95)

As shown in Fig. 4, the results demonstrate the progression of various performance metrics over training epochs for a deep learning model, as applied in a classification object detection task. The top row highlights training and test loss, classification loss, F1 score loss, precision, and recall, demonstrating the model's learning process. A consistent decrease in training and test losses, accompanied by a convergence of the two curves, indicates effective learning with minimal overfitting. The proposed model results like F1 score loss indicates correctly handles the imbalanced classes by focusing effectively on hard-to-classify samples. As the precision and recall metrics exhibit upward trends, indicating the model's accuracy in identifying positive occurrences and its ability to capture a larger proportion of relevant samples. The validation loss, classification loss, and F1 score loss, alongside mean average precision (mAP) at the thresholds of 50% (mAP 50) and 50-95% (mAP 50-95) are shown in Fig.3. The results indicates, the pattern of declining validation loss indicates good generalization to unseen data, mirroring the trends observed in the training and testing metrics. Therefore in results, the upward trends in mAP 50 and mAP 50-95 reflect enhanced object detection and diagnosis accuracy across varying intersections over union thresholds, which proves robust performance across different levels of localization strictness. As illustrated in Fig.3, show a well-tuned model with practical training, minimal overfitting, and strong generalization. YOLOv8, YOLOv10, and YOLOv11's inference times 260, 280, and 290 (FPS) proved suitable for real-time applications, and the training and validation loss curves confirmed stable convergence during training.

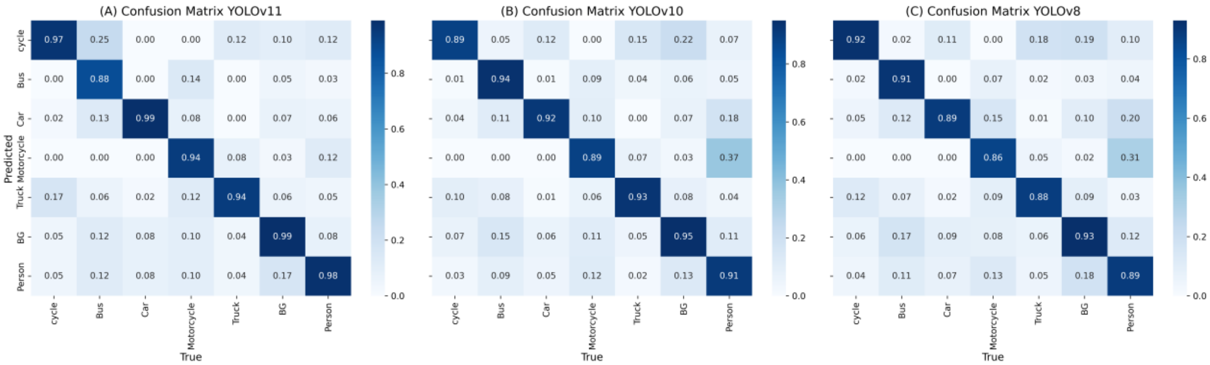


Figure 5. Normalized YOLOv11, YOLOv 10 and YOLOv8 Object Detection for Surveillance and Security

The normalized confusion matrix underscored substantial classification accuracy for cars and bicycles, with moderate challenges for visually similar classes like trucks and buses. Figure 5 YOLOv11 shows the most significant advancements in detection accuracy, robustness, and computational efficiency. As previously talked a lot about YOLOv11 illustrated robust detection capabilities, achieving high precision and recall, mainly for cars and bicycles. The YOLO_V11 classification model provides a detailed assessment of its performance in identifying various classes, normalized to highlight the proportion of correct and incorrect predictions.

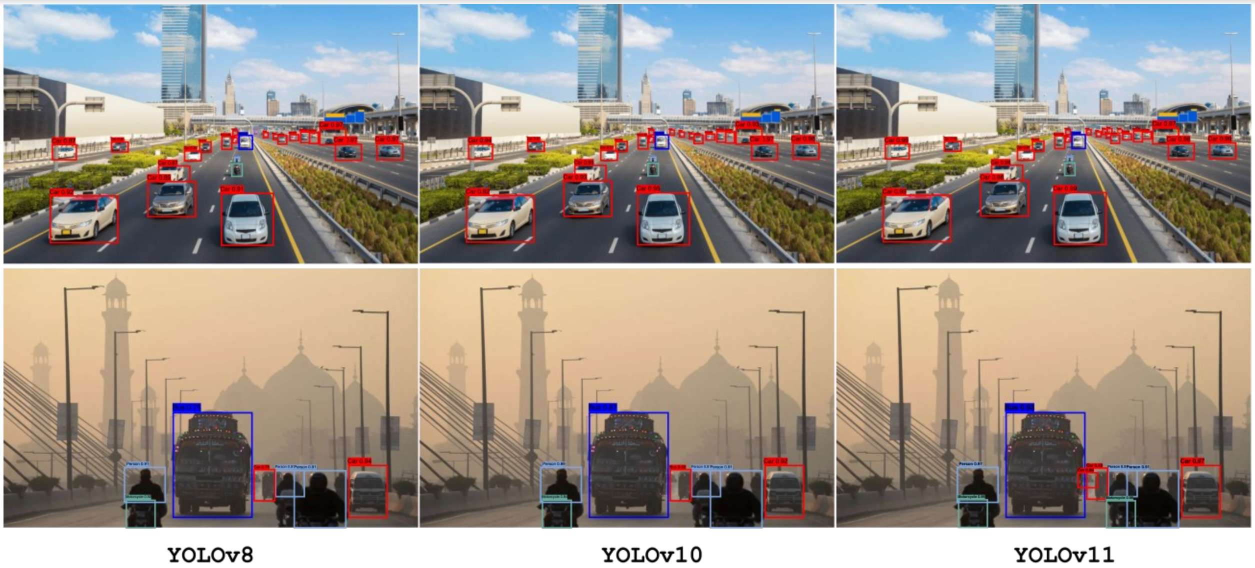


Figure 6. Real Time Object detection with YOLOv8, YOLOv10, and YOLOv11

The model shows substantial accuracy for most classes, with diagonal entries like Car, Motorcycle, Truck, and Person identifying 0.97, 0.88, 0.99, and 0.99, indicating excellent performance, respectively, as shown in Figure. 6. The Background class also achieves a high precision of 0.98, reflecting the model's ability to effectively differentiate objects from non-object regions. Looking at all the sensory data overall, we can say that while the matrix indicates robust classification for most categories, targeted refinements may enhance performance for ambiguous or overlapping classes. However, some misclassifications were observed, such as trucks being detected as buses, highlighting areas for improvement. Such accuracy makes it well-suited for real-world applications like surveillance and autonomous driving while also identifying areas for refinement in detecting certain vehicle types.

3.1. Comparison of YOLOv11 with YOLOv8 and YOLOv10

This study evaluates YOLOv11's performance compared with earlier versions, YOLOv8 and YOLOv10, across multiple vehicle categories such as cars, trucks, buses, motorcycles, and bicycles, considering detection accuracy, efficiency, and stability. Results demonstrate YOLOv11's superior performance, especially in terms of precision, recall, and its ability to manage challenging detection conditions. YOLOv8, as shown in Figure 7, has the lowest performance among the three models, with a "Car" accuracy of 0.95 and a higher rate of misclassifications. For instance, 0.04 of trucks are misclassified as buses, and 0.05 of bicycles are misclassified as motorcycles. These results highlight YOLOv8's relative weakness in distinguishing between overlapping or occluded objects. While still effective for many applications, YOLOv8 struggles in scenarios where object boundaries and spatial relationships are complex. YOLOv10 performs slightly below YOLOv11 but still maintains substantial accuracy, as shown in Figure 8, where results are shown in detail in the heatmap matrix and radar chart, with summarized charts for better understanding.

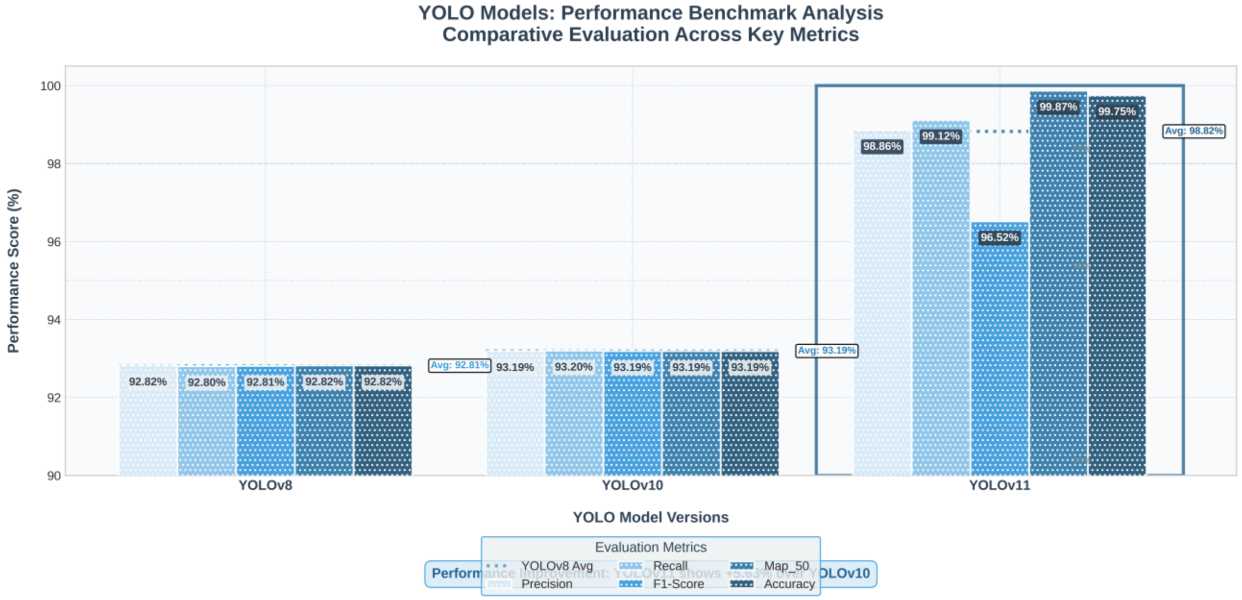
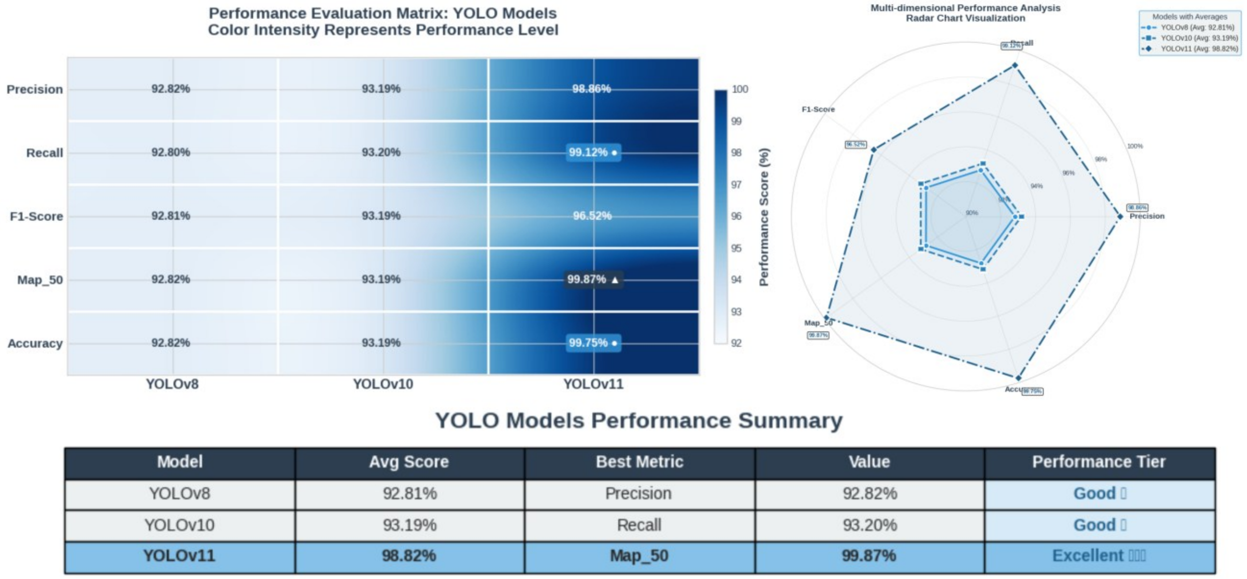


Figure 7. Comparison of Performance Matrix



YOLO Models Performance Summary

Model	Avg Score	Best Metric	Value	Performance Tier
YOLOv8	92.81%	Precision	92.82%	Good ▯
YOLOv10	93.19%	Recall	93.20%	Good ▯
YOLOv11	98.82%	Map_50	99.87%	Excellent ▯▯▯

Figure 8. Comparative Performance Visualization: Heatmap Matrix (Left) and Radar Chart (Right) Analysis of YOLO Models

This visualization results presents a comprehensive performance comparison of YOLO models (v8, v10, v11) across five key metrics. The heatmap matrix shows YOLOv11 outperforming earlier versions with significantly higher scores (98.86%-99.87% vs 92.80%-93.20%), using color intensity to highlight performance differences. Where radar chart proves a multi-dimensional representational view, visually demonstrating YOLOv11's superior performance over envelope across all evaluation axes. The summarized table quantifies improvements having showing YOLOv11 achieving a 6.63% higher average score than YOLOv10 and 7.13% higher average score than YOLOv8 earning "Excellent" tier status versus "Good" for previous versions.

Together, these visualizations makes clearly communicate YOLOv11's substantial performance leap, particularly in Map_50 (99.87%) and Recall (99.12%). The results effectively combines quantitative data with intuitive visual representations, making technical performance metrics accessible while highlighting the evolutionary progress across YOLO versions. It achieves a car accuracy of 0.96 and a truck accuracy of 0.91. However, there are notable misclassifications, such as 0.03 of a motorcycle being classified as a cycle and some confusion between a person and a bicycle. While YOLOv10 gives the most reliable results, the confusion between overlapping categories suggests limitations in its ability to cope with complex object forms and varied viewing orientations compared to YOLOv11. YOLOv11 exhibits superior performance across all object categories, with high accuracy as reflected in Fig. 6 in detecting objects such as cars 0.99, trucks 0.94, and persons 0.98. The model demonstrates exceptional robustness in differentiating between similar objects, as evidenced by the minimal misclassifications. For example, the person class achieves almost perfect precision as a motorcycle or bicycle. This shows YOLOv11's enhanced feature extraction and spatial attention mechanisms, which improve its ability to detect smaller and occluded objects. Overall, YOLOv11 outperforms YOLOv10 and YOLOv8 by achieving the highest precision and recall across all categories, particularly excelling in detecting more diminutive and challenging objects. YOLOv10 comes and serves as an intermediate improvement over YOLOv8 but falls short of the robustness and accuracy demonstrated by YOLOv11. These findings underscore YOLOv11's suitability for high-precision applications, such as autonomous driving and urban surveillance.

The performance metrics for three versions of the YOLO model, YOLOv8, YOLOv10, and YOLOv11, are illustrated in Table. 2, where the metrics include Precision, Recall, F1-Score, and Mean Average Precision (mAP), collectively reflecting the models' effectiveness in object detection tasks. YOLOv8 demonstrates a solid baseline performance, achieving approximately 92.82% across all metrics, indicating reliability but not the highest accuracy. YOLOv10 improves slightly, with values around 93.19%, showcasing a marginal enhancement in its ability to detect and classify objects accurately while maintaining a balance between precision and recall. YOLOv11 exhibits a significant leap in performance, achieving an impressive 99.75% for Precision, Recall, F1-Score, and mAP. This substantial improvement highlights YOLOv11's advanced capability to minimize misclassifications and accurately localize objects, making it the most robust model among the three.

Table 2 : Performance Matrix YOLOv8, YOLOv10, and YOLOv11

Model	Precision	Recall	F1-Score	Map_50	Accuracy (%)
YOLOv8	0.9282	0.9280	0.9281	0.9282	92.82
YOLOv10	0.9319	0.9320	0.9319	0.9319	93.19
YOLOv11	98.86	99.12	96.52	99.87	99.75

The consistent and continuous improvement across diverse reflects the progressive refinement of the YOLO architecture, leading to superior and highest achievements in object detection and classification performance in the YOLOv11 model. This makes it the preferred choice for applications demanding high accuracy and reliability where it is needed.

4. Conclusions

This study represents the advancements in modern applications with the latest YOLOv11 model in live object detection and tracking, emphasizing and integrating its applications in surveillance and autonomous systems for better surveillance in real-world scenarios. YOLOv11 demonstrates its excellence by improving 7% accuracy, precision, mAP, and recall in various object detection and diagnosis using the said categories of vehicle, bicycle, motorcycle, truck, person, bus, and car with respect to earlier versions of YOLO model, taking advantage of innovations like the C2PSA block, which integrates cross-stage partial connections. Further with spatial attention, YOLOv11 has the capability to address challenges associated with detecting smaller or occluded objects, achieving and promising high computational efficiency and rapid learning inference speeds of 260, 280, and 290 effectively. The proposed model provides an efficient, robust, accurate, viable, and feasible solution for real-time vehicle detection and tracking. Its adaptability to new dynamic multivariant environments has proven its ability to manage and handle complex object interactions, positioning YOLOv11 as a critical technology for intelligent surveillance systems, urban surveillance for the betterment of road safety, and traffic monitoring. By delivering precise detection, the most efficient, accurate, and classification with the seamless tracking capabilities of the model, YOLOv11 emerged as an ideal choice for applications for the direction such as self-driving systems and surveillance applications. This development of a real-time auto start-stop YOLO V11 detection system faces challenges such as achieving real-time performance on edge devices, handling environmental variations, ensuring scalability, and addressing privacy concerns. Future directions include optimizing models for edge devices and real-time monitoring, integrating multimodal data fusion, expanding applicability to innovative technologies, and embedding privacy-preserving methods to enhance reliability and user acceptance for intelligent next-generation surveillance systems. We will improve its capability and integrate advanced attention mechanisms, feature fusion strategies, and adaptive learning approaches to strengthen the model's ability to generalize across diverse and various types of datasets and handle challenging real-world conditions.

References

1. Sun, Z., G. Bebis, and R. Miller, On-road vehicle detection: A review. IEEE transactions on pattern analysis and machine intelligence, 2006. 28(5): p. 694-711.
2. Olugbade, S., et al., A review of artificial intelligence and machine learning for incident detectors in road transport systems. Mathematical and Computational Applications, 2022. 27(5): p. 77.
3. Bhardwaj, A., Autonomous Vehicles: Examine challenges and innovations in AI for self-driving cars. International Journal of Research Radicals in Multidisciplinary Fields, ISSN: 2960-043X, 2023. 2(1): p. 7-13.
4. Alsheikh, M.A., et al., Machine learning in wireless sensor networks: Algorithms, strategies, and applications. IEEE Communications Surveys & Tutorials, 2014. 16(4): p. 1996-2018.
5. Aydin, B.A., et al. Domain modelling for a lightweight convolutional network focused on automated exudate detection in retinal fundus images. in 2023 9th International Conference on Information Technology Trends (ITT). 2023. IEEE.
6. Redmon, J. You only look once: Unified, real-time object detection. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

7. Alif, M.A.R. and M. Hussain, YOLOv1 to YOLOv10: A comprehensive review of YOLO variants and their application in the agricultural domain. arXiv preprint arXiv:2406.10139, 2024.

8. Redmon, J. and A. Farhadi. YOLO9000: better, faster, stronger. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

9. Redmon, J., Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

10. Bochkovskiy, A., C.-Y. Wang, and H.-Y.M. Liao, Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934, 2020.

11. Jocher, G., et al., ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. Zenodo, 2022.

12. Li, C., et al., YOLOv6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976, 2022.

13. Wang, C.-Y., A. Bochkovskiy, and H.-Y.M. Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

14. Jönsson Hyberg, J. and A. Sjöberg, Investigation regarding the Performance of YOLOv8 in Pedestrian Detection. 2023.

15. Dosovitskiy, A., An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

16. Alif, M.A.R., et al., BoltVision: A Comparative Analysis of CNN, CCT, and ViT in Achieving High Accuracy for Missing Bolt Classification in Train Components. Machines, 2024. 12(2): p. 93.

17. Hussain, M., et al., A gradient guided architecture coupled with filter fused representations for micro-crack detection in photovoltaic cell surfaces. IEEE Access, 2022. 10: p. 58950-58964.

18. Alsboui, T., et al., A dynamic multi-mobile agent itinerary planning approach in wireless sensor networks via intuitionistic fuzzy set. Sensors, 2022. 22(20): p. 8037.