EAI Endorsed Transactions
on Internet of Things
Research Article EAI.EU

Effective Object detection and Tracking using Attention-Driven YOLO v9 Model with Multi-Stage Cascaded Convolutional Model

Krishna Mohan A[1,*], P.V.N.Reddy[2], K. Satya Prasad[3]

[*1]Department of Electronics and Communication Engineering, JNTUK, Kakinada, Andhra Pradesh, India
[2]ECE Department, Ramireddy Subbarami Reddy Engineering College, Kadanuthala, Andhra Pradesh, India
[3]ECE Department, JNTUK, Kakinada, Andhra Pradesh, India

Abstract

INTRODUCTION: Object detection and tracking are essential for computer vision, particularly for vehicle monitoring within digital images and video streams. Traditional methods, such as background subtraction and template matching, rely on heuristic algorithm and handcrafted features, which often struggles with diverse vehicle appearance and complex backgrounds. These techniques, while foundational, exhibit limitations in flexibility and scalability, resulting in lower accuracy and high computational costs.
OBJECTIVES: In contrast, advanced Deep Learning (DL) approaches, particularly those utilizing Conventional Neural Network (CNNs), have revolutionized the field by enabling automatic feature extraction from large datasets. Despite their advantages, existing DL models like You Only Look Once (YOLO) face challenges in detecting small or closely packed vehicles and can be computationally intensive.
METHODS: This study proposed an Attention Driven YOLO v9 architecture that integrates with a proposed mechanism combining spatial and channel attention to detect the small size vehicle accurately.
RESULTS: Additionally the architecture incorporates multi stage cascaded convolution layers to enhance the feature extraction and robustness against occlusion and background noise. The model is trained using the UA-DETRAC dataset, providing a rich set of images for learning.
CONCLUSION: Performance evaluation metric such as Mean Average Precision (mAP), precision, recall, and tracking accuracy demonstrating significant improvement over traditional methods and existing state of the art models. This research contributes to the field by addressing the limitations of previous studies through technique to speed and accuracy in vehicle detection and tracking.

Keywords: Object Detection, Tracking, Vehicle, CNN, YOLO v9, Deep learning, Attention mechanism, Spatial Mechanism

Received on 18 December 2024, accepted on 24 April 2025, published on 11 June 2025

doi: 10.4108/eetiot.8231

1. Introduction

Object detection and tracking are critical components of computer vision, focusing on identifying and particularly in vehicle monitoring objects within digital images or video streams. Vehicle detection refers to the process of locating and classifying objects in a single framework, typically utilizing algorithms that draw bounding boxes around detected items [1]. This initial identification is important for subsequence tracking, where the aim is to follow the movement of these objects across multiple frames in real time [2]. The evolution of these technologies has been driven by advancements in Machine Learning (ML) and Deep Learning (DL), enabling more exact and efficient detection and tracking methods [3]. The integration of

*Corresponding author. Email: akrishnamohan286@gmail.com

these techniques has significant application in various fields, including surveillance, and autonomous vehicle and human computer interaction, highlights their importance in enhancing the safety and operational efficiency in numerous domains [4]. Vehicle detection and tracking employs a variety of methods that can be broadly categorized into traditional techniques and modern deep learning approaches [5]. Traditional methods include the background subtraction, frame differencing and template matching, which rely on analyzing changes between consecutive frames to identify moving vehicles. The field of object detection has evolved essentially over the past few years, transitioning from traditional methods reliant on handcrafted features to sophisticated DL techniques that leverage large datasets for improved exactness and effectiveness [5]. Traditional object detection methods, which emerged before the advent of DL, primarily used heuristic algorithm and manual feature extraction. These approaches, such as viola-jones detector and Histogram of Orientated Gradient (HOG), relief heavily on predefined criteria to identify vehicles within images [6]. For example, the Viola Jones detector employed Haar-like features to detect faces by analyzing intensity differences in the rectangular region of an image. Similarly, HOG calculated gradient orientation in localized portions of an image to detect pedestrians. While these methods laid the groundwork for vehicle detection, they were limited in their flexibility and scalability, often struggling with diverse vehicle appearances and complex backgrounds [7].

One of the important drawbacks of traditional methods is their reliance on handcrafted features, which are not universally applicable across various domains. This limitation often results in low exactness for common models and high computational complexity during region selection. Furthermore, traditional techniques typically involve a sliding window approach to scan images for potential vehicle locations, which can be computationally expensive and inefficient, particularly when dealing with high resolution images or real time application [8]. With the introduction of DL, particularly CNN (Conventional Neural Network), the landscape of vehicle detection began to shift dramatically. DL model can automatically learn relevant features from data, allowing them to generalize better across different tasks and datasets [9, 10]. This transition led to the development of two primary categories of object detection architectures: one stages and two stage detectors. One stage detectors, such as YOLO and Single Shot Multi Box Detector (SDD), perform object detection in a single pass through the network, offering faster processing speeds suitable for real time applications [11, 12]. Whereas, two stage detectors like R-CNN (Region based Conventional Neural Network) first generate region proposal and then classify these proposal in a second step. While two stage models tens to achieve higher accuracy due to their more complex processing, they are generally slower and less efficient than their one stage counterparts. Despite these advantages, the existing DL based models also face challenges [13, 14]. The one stage detector may struggle with detecting small vehicle or objects that are closely packed together due to their reliance on a single pass through the network. On the other hand tow stage model can be computationally intensive and may not be suitable for applications required real time performance. Additionally, the both types of models can suffer from issues related to occlusion and varying lighting conditional, which can importantly impact the detection exactness. In recent years, there has been a growing interest in enhancing vehicle detection capability through innovative architectures that incorporates attention mechanisms and multi stage processing [15, 16]. Attention driven models aim to improve feature extraction by focusing on relevant parts of an image while ignoring less important information. This approach allows models to better handle occlusions and variations in vehicle appearance, by dynamically adjusting their focus based on contextual information.

Moreover, this research addresses some limitation observed in previous studies by incorporating advanced techniques such as data augmentation which increase the diversity of training samples by applying various transformation to existing images, thereby improving model generalization capability. Overall, this innovative approach is expected to outperform existing state of art methods in term of both speed and accuracy. To overcome such limitation the proposed research focuses on developing an Attention Driven YOLO v9 model combine with a Multi-stage cascaded conventional Model for effective object detection and tracking. The approach lies in its ability to integrate attention mechanism directly into the YOLO v9 architecture while employing a multi stage cascade framework that enhanced feature extraction at various levels of abstraction. By leveraging attention mechanism, the model can selectively focus on critical features within an image, improving its robustness against occlusions and background noise. The model trained using the UA-ETRAC dataset, which provides a risk set of images for learning. This training phase is crucial for optimizing the models performance. Following the training, the predicted data will be loaded to test the models performance effectively. Performance estimation involves analyzing key metrics such as Mean Average Precision (mAP), Precision, Recall, and Tracking vehicle. Moreover, this research addresses some limitation observed in detection and tracking vehicles.

1.1 Research Contribution

The main contribution of the paper are as follows,

- To create an attention-driven Yolo v9 object detection and tracking capabilities. And to integrate attention mechanism into YOLO v9 architecture, allowing the model to focus on critical features within images, thereby improving robustness against occlusion and background noise.
- To employs a multi stage cascaded framework that enhanced feature extraction at various levels of abstraction, leading to improved detection accuracy and efficiency.
- To assess the models performance using key metrics such as MAP. Precision and recall

ensuring a comprehensive evaluation of its effectiveness in object detection and tracking.

- To train the proposed model using the UA-DETRAC dataset, providing a diverse set of images for effective learning and evaluation.

1.2 Paper Organization

The paper is organized in the basis of research on the attention driven YOLO v9 model combined with a multi stage cascade Conventional model for object detection and tracking. Section 1 includes an introduction that outlines the research background, contribution of the paper. As well as section 2 reviews the related work, analyzing conventional methods and identifying existing problems. In section 3, the proposed methodology details data pre-processing and the implementation of object detection, including the channel prior attention mechanism and spatial attention mechanism. Section 4 presents the result and discussion, covering dataset description, exploratory data analysis, performance metrics, and a comparative analysis. Finally section 5 concludes the paper by summarizing finding and discussing potential future work, highlights the contribution of the proposed methods to advancements in object detection and tracking technologies.

2. Related Work

This section deliberates the analysis of the conventional research in the vehicle detection and tracking such as DL and ML also computer vision techniques. By exploring various methodologies the study address the challenges such as occlusion, varying traffic conditions, and the need from real time processing.

In the conventional system, the logistic vehicle speed detection method using YOLO (LV-YOLO), which enhances traffic management by segmentation vehicle with U-Net and detecting their speed based on the Boxy Vehicle dataset. The YOLO achieves a mAP in better percentage, outperforming existing systems by up to 5.42% in vehicle detection and 4.81% in speed Prediction [17]. Similarly the existing study [18], the use of DL, particularly CNN in vehicle detection and tracking highlighting framework like faster R-CNN and YOLO for optimized performance. It also addresses tracking challenges with algorithm such as Deep SORT and Tractor, evaluating effectiveness through metrics like IoU, precision, and recall while emphasizing the importance of balancing real time processing and accuracy for traffic surveillance. As well as, the prevailing study [19], focused on estimating traffic entity to improve intelligent transportation systems, emphasizing vehicle recognition an counting as critical steps in the process. It leverage deep learning technologies, particularly CNN utilizing data from open source libraries such as MB7500, KITTI, an FLIR, with image annotation and augmentation techniques applied to enhanced dataset size and quality. A hybrid model combining Faster R-CNN and YOLO with a majority voting classifier is trained on the processed data, achieving a detection accuracy of up to 98%, surpassing YOLO's 95.8% and faster R-CNNs 97.5%. The proposed approach demonstrates superior performance in estimating traffic density, indicating its potential to enhance road traffic management effectively.

The study [20], introduced DETR-SPP, a one stage vehicle detection network that enhances real time detection speed and accuracy by modifying the Detection Transformer (DETR) architecture with spatial pyramid pooling, focusing on vehicle classes from the MS COCO 2017 dataset. The model achieves a mAP of 51.31%, surpassing the DETR baseline by 5.19%, with a Wilcoxon signed-rank test p-value of 0.03, confirming its effectiveness in vehicle detection. Similarly, the enhanced Histogram of Oriented Gradients (HOG) [21], approach for night time vehicle detection, utilizing background illumination removal and saliency models to extract vehicle lights, followed by SVM classification and non-maximum suppression (NMS) for improved accuracy. Experimental result demonstrate significant enhancements in vehicle recognition accuracy in low light conditions, although specific numerical improvements were not detailed. Generally, the study [22] in YOLO-GNS, a novel algorithm for detection of special vehicles from UAVs, which enhanced feature extraction through a single Stage Headless (SSH) context structure and reduced computational costs using Ghost Net's linear transformations. Experimental result demonstrate a 4.4% increase in average detection accuracy and a 1.6 improvement in detection frame rate, highlighting its effectiveness for monitoring illegal activities in various scenario. Correspondingly, the edge intelligence-based improved YOLO v4 vehicle detection algorithm that enhances detection capabilities using an efficient channel attention (ECA) mechanism and a high resolution network (HRNet), achieving an average precision increase from 82.03% to 86.22%. Additionally, an improved DeepLabv3+ segmentation algorithm utilizing MobileNetv2 and soft pooling boosts Mean Intersection over Union (mIoU) from 73.63%, importantly advancing traffic information processing.

The paper [23], presented a vehicle detection and classification method using the YOLO v5 architecture, leverage transfer learning to fine tune the pre trained model on extensive dataset that capture various traffic condition, including occlusions and different weather scenario. The improved YOLO v5 model outperforms traditional detection method in accuracy and execution time, demonstrated effectiveness on publicly available dataset like PKU, COCO, and DAWN. Similarly, the prevailing study [24], used automatic multiple vehicle detection and tracking framework that combines computer vision with Partial Differential Equation (PDE) based model using a Haar Cascade Classifier and Active Contour based segmentation for vehicle detection. The tracking method employs DL for multi scale analysis and vehicle matching, with simulation results demonstrating the effectiveness of the proposed approach in various traffic scenario. The study [25], present the modified cascade R-CNN that enhance vehicle detection by integrating contextual information, improving features extraction for small and occulted objects through an improved features pyramid and

predictive optimization module. Experimental results show that this method outperforms state of art vehicle detection in accurately detecting small and shielded vehicle. The existing study [26], showed the automatic vehicle classification has become crucial for intelligent transportation systems, particularly during mobility restriction like those imposed during COVID-19, where controlling vehicle classification methods, which often prioritize prediction exactness at the expense of real time performance and resource efficiency, by proposing a new techniques that utilizes adaptive histogram equalization and Gaussian mixture model to enhance vehicle image quality and employs an ensemble DL approach for classification.

Similarly, to reduce the false detection rate of vehicle target caused by occlusion, an improved vehicle detection method based on an enhanced YOLO v5 network has been used in an existing study [27], utilizing the Flip-Mosaic algorithm to enhance the network's ability to observe small targets. Investigational results indicate that this data enhancement technique significantly improves detection accuracy and reduces false detection rates across a multi-type vehicle dataset collected in various traffic states. The novel multi stage CNN [28], for vehicle detection that operates proficiently on a Central Processing Unit (CPU), eliminating the need for GPUs commonly required by traditional methods. The MSCNN framework, which includes stages for boundary detection and vehicle classification, achieves an average precision of 72.1% on the KITTI dataset, demonstrating it's effective for practical application in intelligent transportation systems. Likewise, the study [29], focused to develop a mobile application using augmented reality to assist elderly users in identifying traffic signals and signboards in real time through deep learning techniques. By comparing the single shot multi box detector (SSD) model with two stage faster R-CNN, the study finds, that the SSD model with mobile Net is faster and comparably accurate, while also addressing occlusion challenges through image segmentation techniques for robust object detection suitable for mobile deployments.

The paper [30] used, a Normalization-based Attention Module (YOLOv5-NAM) integrated into the YOLOv5 model for vehicle detection and tracking method for small target vehicles, and a real-time tracking approach (JDE-YN) that embeds feature extraction in the prediction head. Experimental results on the UA-DETRAC dataset show that YOLOv5-NAM improves mAP by 1.6%, while the JDE-YN method enhances the MOTA value by 0.9% compared to their respective original models. Similarly, the prevailing study [31], improved YOLOX_S detection model addressed the challenges of misdetection and omission of small target in vehicle detection by implementing several enhancement. Key modification include clipping redundant parts of the original network to boost inference speed, integrating a coordinate attention module within the residual structure to preserve feature information, and adding an adaptive features fusion module to enrich small target features, ultimately achieving an average detection accuracy of 77.19% on the experimental dataset, despite a decrease in detection speed to 29.73 fps.

The study addressed the limitation of existing vehicle detection and tracking methods, particularly the challenges of misdetection and omission of small targets, especially in complex traffic scenario with occlusion. Traditional approaches often struggle with high, small or occluded vehicle, leading to issues such as false positive and missed detections. To overcome these issues, the proposed method for object detection and tracking incorporates an attention mechanism into the YOLO framework.

2.1 Problem Identification

Several conventional research has been limited in vehicle detection particularly with YOLO framework,

- Traditional YOLO struggles with detecting vehicles that are occluded by other object or vehicle. This often leads to a significant loss of key features, making it difficult for the model to accurately identify and classify vehicle in complex traffic environment [32].
- Previous YOLO version lack refined attention mechanism that can prioritize relevant feature in particular visible vehicles. This limitation hinders models ability to focus on unobstructed parts of vehicle when occlusion occurs [33].

Many existing dataset do not adequately represented real-world scenario involving occlusion, which can lead to over fitting, poor generalization of the model in practical application, and a declined in both accuracy and efficiency [34, 35]

3. Proposed Methodology

The Attention-Driven YOLO v9 model with a Multi-Stage Cascaded Convolutional model represent a significant advancement in vehicle detection and tracking technology. This innovative approaches leverages the latest enhancement in the YOLO architecture, integrating advanced attention mechanism to improve feature extraction and retention and using UA-DETRAC dataset. This combination of cutting edge techniques positions the YOLO v9 model as a robust solution for application in intelligent transportation systems, detection of object and tracking the vehicles.

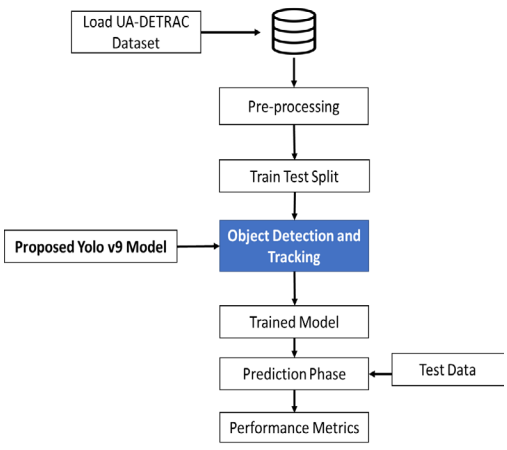

Figure 1. Proposed Flow

Figure 1 illustrates the end to end pipeline for developing and evaluating an object detection and tracking system using a proposed YOLO v9 model. By loading the UA-DETRAC datasets, a well-known dataset designed for vehicle detection and tracking. Next the data undergoes the pre-processing to prepare it for modelling, which may include the step such as resizing images, normalization, or annotation refinement. Following this, the dataset is divided into training and testing splits, ensuring separate data for model training and testing. The core object detection and tracking step, powered by the proposed YOLO v9 model. This model is trained using the training dataset to learn the characteristics and patterns necessary for detecting and tracking objects, such as vehicle, effectively. After the model is trained, it transition into the prediction phase, where it processes the test data to make predictions about object location and movement. Finally, the proposed model performance is measuring the performance metrics, which likely including the indicators such as mAP, value of precision, probability of detection, and F measures and tracking efficiency. These metric provides the insights into effectiveness in real world applications, enabling an evaluation of how well the proposed YOLO v9 model performs in the context of the object detection and tracking.

3.1 Data Pre-processing

Data Pre-processing is a crucial step in preparing the UA-DETRAC dataset for training the attention Driven YOLO v9 model for vehicle detection and tracking. The quality of the input data directly impacts the models performance, making effective preprocessing essential for achieving accurate and reliable results. This process involves several key techniques tailored to enhance the dataset suitable for ML application.

3.2 Object Detection and Tracking using proposed Attention-Driven YOLO v9 Model with Multi-Stage Cascaded Convolutional Model

Traditional methods for vehicle detection and tracking often struggles with accuracy identifying vehicle lighting conditions. These limitations can lead to significant degradation in detection performance, as conventional algorithm may fail to recognize partially obscured vehicle or misclassify them due to overlapping features. To overcome these challenges, the proposed attention driven YOLO v9 model with a multi stage cascaded conventional model is introduce as a robust solution that enhance both detection and tracking efficiency. The proposed attention driven YOLO v9 model with a multi stage cascade convolution model marks a significant advancement in vehicle lighting conditions are prevalent. This model introduce a novel attention mechanism that enhances feature extraction by allowing the network to focus on the most relevant parts of the input data, with channel mechanism. The proposed novel YOLO v9 model is illustrated.

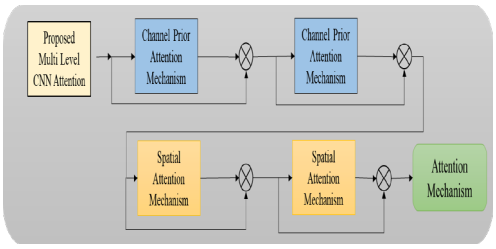

Figure 2. Proposed Attention Driven YOLO v9 model

The Figure 2 illustrate the proposed multi-level CNN attention framework employs both channel and spatial attention mechanism to enhance feature learning in CNN addressing the limitations of traditional object detection methods. This framework integrates a channel prior attention mechanism and a spatial attention mechanism, which works in parallel pathways to refine feature representation extracted from the CNN model. By processing input feature through these distinct attention pathways, the framework identifies and amplifies the most relevant channels while simultaneously highlight critical spatial regions within the feature maps. The iterative application of these attention mechanisms allows for a more precise enhancement of features, ultimately leading to improved detection accuracy. The final output from both attention pathway are combined into a unified representation, leveraging the strengths of each mechanism. This innovative approach not only capture complementary aspects of the feature but also ensure robust performance across various tasks, such as image recognition and classification, making it a significant advancement in the field. Similarly, the architecture of the proposed model in figure 3,

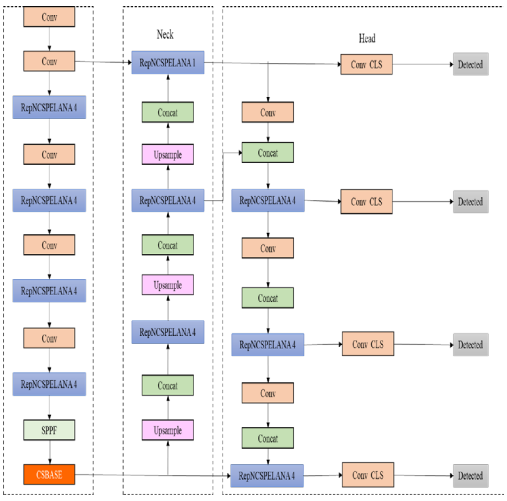

Figure 3. Architecture of the proposed model

Figure 3, depicts a deep learning model architecture designed for object detection processing input images of size 64.*640*3. The conventional layer that reduces spatial dimension from 640*640 to 320*320. And then to 160*160, utilizing RepNCSP ELANA blocks for efficient features representation. Feature are down sampled using

Adown layers to produce higher levels features maps at scale of 80*40*40, and 20*20. A spatial pyramid pooling layer enhanced (SPPLEN A) extracts multi scale contextual information, which is concatenated with other feature map for robust fusion. Finally, CB-Fuse and CB-Linear modules refine the feature maps for detection at three scales to 80*80, 40*40, and 20*20 ensuring the effectiveness detection of object of varying.

Correspondingly, the mathematical formalize the operations within the proposed attention driven YOLO v9 model is defined the key components and their interaction through a series of equations that capture the essence of the models architecture and attention mechanism,

$$H = \delta\left(M_2\varepsilon(M_1 y)\right)$$
(1)

In equation 1 $y = \frac{1}{H\times W}\sum\frac{H}{i=1}\sum\frac{W}{j=1}F_{ij}$ represent the Global Average Pooling (GAP) of the feature map. The parameters $M1 \in R^{c\times C_r}$ and $M2 \in R^{C} *^{C^r}$ are learned weights in the attention module, while δdenotes the sigmoid activation function and ϵ represent the Rectified Linear Unit (ReLU) activation function.

$$H = H_2 \otimes H_1 \otimes H,$$
(2)

$$H_1 = \delta\left(M_2(M_1 y) + M_2(M_1 z)\right)$$
(3)

$$H_2 = \delta\left(A2B_7\left(y(H_1), z(H_1)\right)\right)$$
(4)

Here, equation 3, captures channel attention by combining the GAP y and Global Max Pooling (GMP)z. The shared weight parameters ensure the both pooling operations contributing to refine the channel importance. Similarly, equation 4, applying a 7*7 convolution kernel to the concatenated outputs.

$$H = H \times \delta\left(M_2\varepsilon(M_1 y_c^h)\right) \times \delta\left(M_2\varepsilon(M_1 y_c^w)\right)$$
(5)

Where, $y_c^h = \frac{1}{W}\frac{W}{j=1}F_{hj}$ and $Y_c^w = \frac{1}{H}\sum\frac{H}{i=1}F_{iw}$ represent the channel-wise global average pooling for heights and width respectively. This mechanism allows for accurate localization of objects by considering both horizontal and vertical spatial information.

$$H = \delta\left(A1B_k(y)\right)$$
(6)

Whereas, equation 6, introduce 1D conventional with kernel size$k$, represented as$A1Bk(y)$, demonstrating that efficient channel attention can be achieved without relying on complex parameter designs.

$$H^1 = f_c[F^{in}] \otimes F^{in}$$
(7)

$$H^2 = f_h[H^1] \otimes H^1$$
(8)

$$H^3 = f_w[H^2] \otimes H^2$$
(9)

$$F^{out} = f_c[H^3] \otimes H^3$$
(10)

The equation 7, 8,9,10 represent the channel, height, width, while considering both channel and height information, and final feature map. The final output feature map F^{out} is generated by reapplying the channel attention mechanism.

$$f_c[H] = \delta\left(A1B_{k_1}\left(avgpool(H)\right)\right) = \delta\left(A1B_{k_1}, (y)\right)$$
(11)

The equation defines the channel attention function, where $A1B_{k1}$ represent a convolution operation with kernel size determined by the function defined in equation 11.

$$k = \Phi(C) = \left[\frac{log_2(C)}{2} + \frac{1}{2} + \frac{1}{10}\right]_{odd}$$
(12)The kernel size K is derived based on the number of channel$C$. The function ensures that the result is rounded to the nearest odd number to maintain compatibility with conventional operations.

$$f_h[H] = \delta\left(A1B_{k_2}\left(avgpool(H)\right)\right) = \delta\left(A1B_{k_2}(y)\right)$$
(13)

$$f_w[H] = \delta\left(A1B_{k_2}\left(avgpool(H)\right)\right) = \delta\left(A1B_{k_2}(y)\right)$$
(14)

The equation 12, 13 outlined the calculation of height attention and width attention, employing a similar approach as height attention to effectively capture width specific information.

3.2.1 Channel Prior Attention Mechanism

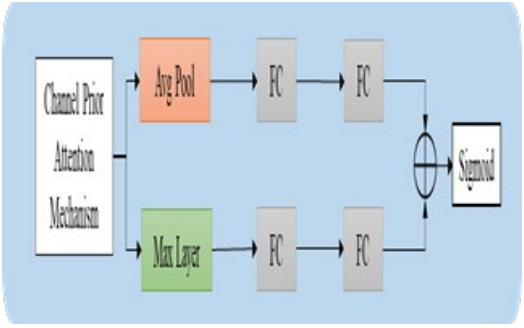

Figure 4. Channel Attention Mechanism

The figure 4 outlines the channel prior attention mechanism, which selectively emphasizes the important features channel in a feature map. Initially, the input feature pass through the average pooling and the max pooling operations along the spatial dimension, producing two separate representations. These outputs are then fed into two shared Fully Connected (FC) layers, which act as learnable mechanism to capture channel wise dependencies. The result from the two pathways are merged using an element wise addition operation. Finally, a sigmoid activation function is applied, generating a channel wise attention map that highlights essential channels while suppressing less relevant ones.

3.2.2 Spatial Attention Mechanism

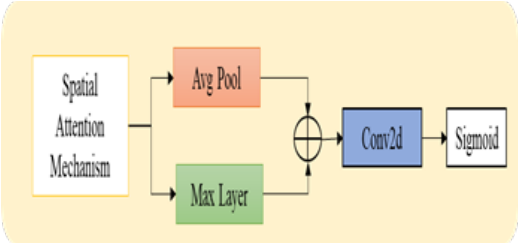

Figure 5. Spatial Attention Mechanism

The Figure 5, illustrate a Spatial Attention Mechanism, which focuses on spatial regions within a feature map that are more significant for the task. The input feature map undergoes average pooling and ma pooling, but this time across the channel dimension, resulting in two 2D spatial maps. These maps are combined using an elements-wise addition operation and processed by a 2D conventional layer (Conv2D) to capture spatial relationships. The output is passed through a sigmoid activation function, creating a spatial attention map that assigned higher weights to relevant spatial regions.

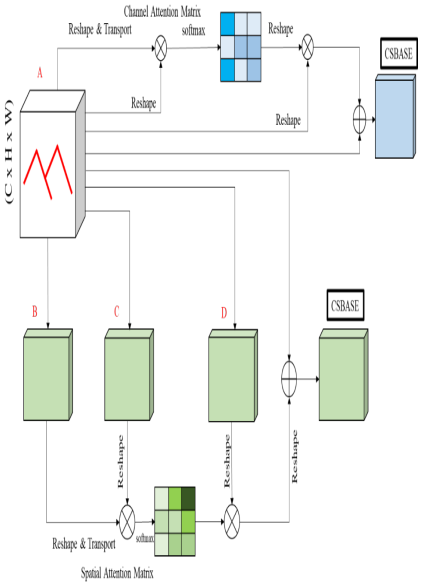

Figure 6. Proposed Channel attention mechanism

The channel attention mechanism is a clear approach in DL aimed at enhancing the performance of CNNs by selectively emphasizing or suppressing information across feature map channels. The process begins with an input feature map, represented as a 3D tensor (Height* width* channels) which is visualized as a cube as labelled "A". This feature map undergoes global spatial average pooling, collapsing its spatial dimensional into a single scalar values for each channel, thereby generating a channel descriptor that captures the global importance of each channel subsequently, a channel wise weight generation process employs a compact multi-layer structure, including full connected layer connected layers and nonlinear activities like ReLU and sigmoid, to refine these channel weights into a normalized vectors that indicated the significance of each channel. The computer weights are then applied to the input feature map through an element-wise multiplication, scaling each channels based on its importance and resulting in an output features map where relevant channels are emphasized while less significant ones are suppressed.

4. Result and Discussion

4.1 Dataset Description

The UA-DETRAC dataset as a critical resource for training and evolution the attention driven YOLO v9 model in vehicle detection and tracking applications. This benchmark dataset comprise 100 challenging video sequence captured from real world traffic environment, totaling over 140000 frames, each annotated with vehicle types, occlusion levels, illumination conditions, and truncation ratios. It encompasses diverse traffic scenarios such as urban highways and intersections, allowing the model to learn vehicle detection under various conditions. The dataset contains annotation for over 8250 vehicle and approximately 1.21 million bounding boxes, providing a solid foundation for robust vehicle detection algorithm. The image serve as input for the model, which is divided into an 80% of training set and 20% test set to facilitate evaluation and prediction. During training the proposed attention driven YOLO v9 model with a multi stage cascade convolution model is utilized. The challenging attributes, including occlusion and varying lighting conditions, enhance the model's ability to recognize vehicle that may be particularly hidden.

4.2 Exploratory Data Analysis

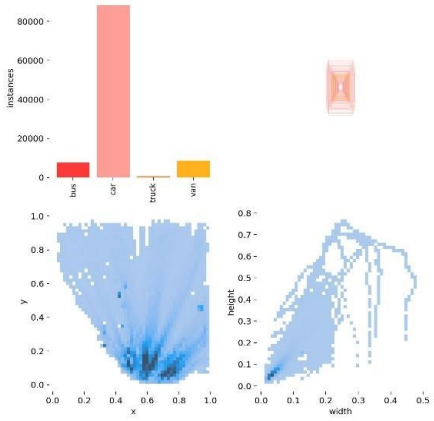

Figure 7. Labelling of object in Dataset

From Figure 7, it shows the analysis of dataset and it's surmised. The dataset comprised of a huge number of vehicle that can be existed in uneven and dense distribution. A first illustration shows the number of vehicle in each type and their instance. Where the second subfigure shows the vehicle locating points which is bounded in the dataset. The below subfigures demonstrate

both x-axis and y-axis, similarly height and width of the radiation. Moreover the figure 7, presents the correlogram labels.

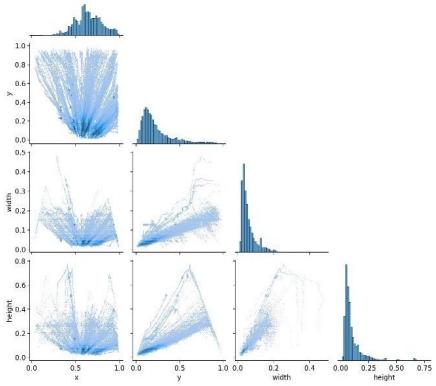

Figure 8. Correlogram Label

The figure 8, deliberates the Correlogram labels. The Correlogram is known as the graphical representation which shows the correlation co-efficient among variables in the dataset. Moreover the labels in correlation denotes the feature or variables that are being examined. The YOLO v9 model ensures training on UA-DETRAC dataset for its challenging detecting vehicle along with limited data.

Figure 9. Sample Dataset Images

Figure 10. Object Detection

From the figure 9 and 10 the sample images have identified various objects, including cars, buses, vans and truck. The model not only detect the types of vehicle present but also detect their movement, providing valuable insights into the dynamic of the scene.

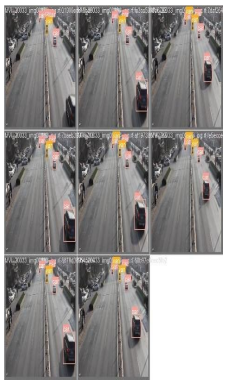

Figure 11. Object Detection and Tracking

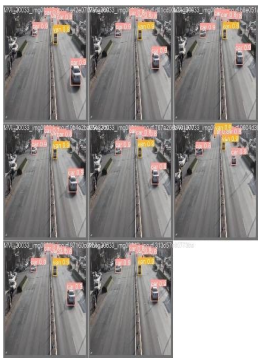

Figure 12. Object Tracking

In figure 11 the model has accurately detected the type of vehicle present, identifying it as a specific category, such as car, bus, van and truck. Alongside this classification, the model also provide the name of the vehicle and quantifies its movement values, offering detailed insights into both the type and dynamics of the vehicle within the image. This comprehensive analysis enhances understanding of vehicle behavior and contribution to effective monitoring in various applications.

4.3 Experimental Research

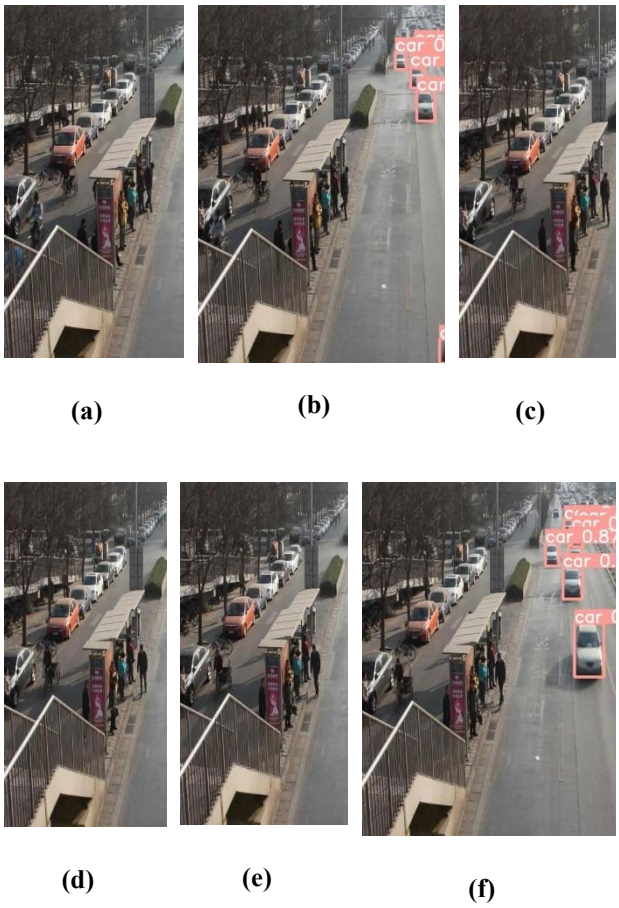

Figure 13. Experimental Results of Vehicle Detection

Figure 13 shows the experimental result of vehicle detection on UA-DETRAC. The proposed framework detects different vehicle of different sizes with their class labels by using YOLO v9 with multi stage cascaded conventional model.

4.4 Performance Metrics

Performance Metrics are primarily used for observing the efficiency of thee projected research by utilizing various metrics such as mAP, Precision, Probability of Detection, F1 measures.

Accuracy: The primary measure used to assess the model is accuracy, calculated as the proportion of accurate prediction to total prediction according to the equation (15).

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Number\ of\ all\ predictions} \quad (15)$$

Precision: It is defined as the number of accurate positive prediction made and can be represented by an equation (16).

$$Precision = \frac{No.of\ correctly\ predicted\ objects}{Total\ no.of\ \ predictions} \quad (16)$$

Recall: Probability of detection is the proportion of positive instance correctly predicted by the model among all positive instance in the data, as defined by an equation (17).

F1-Measure: also known as F1 score is represented by an equation (17) and combines the harmonic mean of probability of detection and value of precision.

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (17)$$

Mean Average Precision: The weight for each threshold in the measurement is determined by t increase in recall from the previous threshold, as shown in equation 18,

$$mAp = \frac{1}{N}\sum_{i=1}^{N} \quad AP_i \quad (18).$$

4.5 Performance Analysis

The performance of proposed algorithm is examine utilizing metrics like precision, probability of detection, F1 measures and mAP. Likewise, the Confusion Matrix (CM) is utilized for identifying the performance of proposed research. It summarizes and anticipates the function of the classification algorithm. Therefore, the confusion matrix shows the number of positive and negative predictions for each class. Figure 13, represent the confusion matrix of the model.

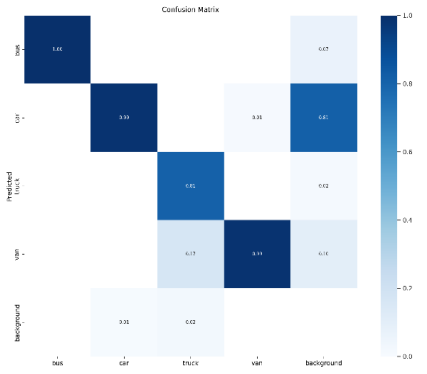

Figure 14. Confusion Matrix

From figure 14, the confusion matrix is demonstrated the performance of a classification model distinguishing between five classes namely background, bus, car, truck and van. The diagonal cells indicate high accuracy, with values close to 1.0 for most classes, suggesting effective classification. Specifically, the model achieves perfect classification for bus (1.00) and near perfect accuracy for cars and van (99%). However, it struggle with truck, correctly identifying only 81% with 17% misclassification

as van and background for 2%. The classification of background is accurately predicted 97% of the time, with minimal confusion from other classes. Key insights highlights the strong performance of car, van, and bus. Whereas the misclassification is between van and truck.

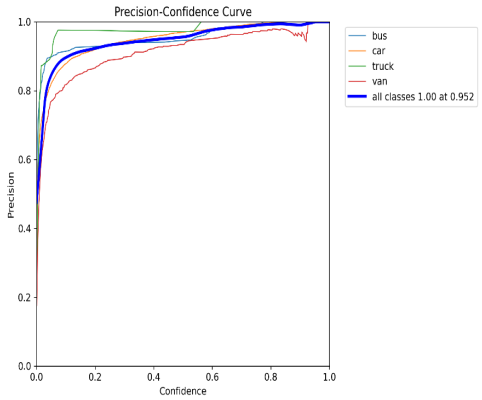

Figure 15. Precision Confidence Curve

From figure 15, the precision confidence curve assesses the relationship between the models prediction confidence as x-axis and precision as y-axis. The thick blue line represent average performance across all classes, attains high precision for >0.95 at most of the confident levels, peaking near 1.0 as confidence rises. The Bus in blue and car in orange categories consistently exhibits the highest precision, closely aligning with the overall curve. In contrast thee truck in green shows the initial performance and the van in red class has low precision overall.

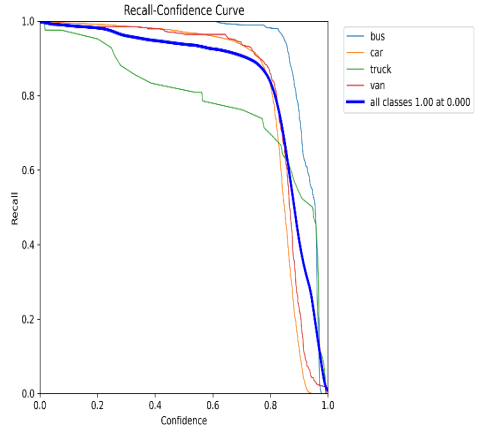

Figure 16. Recall Confidence Curve

The figure 16, illustrates recall confidence curve of the relationship between prediction confidence threshold and recall, measuring the proportion of true positive identified. Recall is highest at lower threshold but declines sharply as confidence increase. The van and truck classes experience a rapid drop in recall indicating reduced detection ability, while the car class maintains more stable performance. The aggregated curve for all classes shows balanced overall performance. This suggests a need to enhance detection for van and truck classes an consider adjusting confidence thresholds to balance high recall with low false positive.

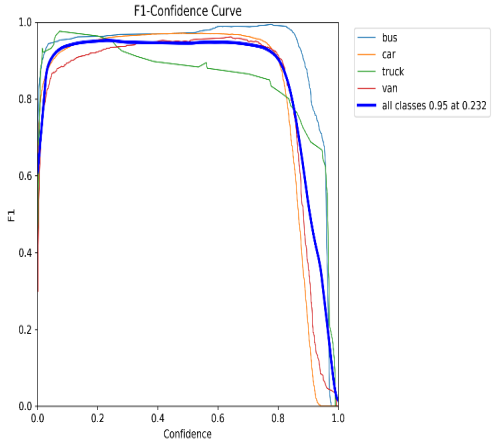

Figure 17. F1 Measure Confidence Curve

Similar to figure 15, 16 the F1 confidence curve evaluates the F1 score, which is illustrated in figure 17, a harmonic mean of precision and recall as a function of confidence thresholds. The thick blue lines shows the average F1 score peaks at the 0.95 but drops significantly at higher thresholds due to the decreasing recall. The bus category maintains the highest F1 measures across the range, while the car orange drops sharply after a threshold of approximately 0.8. The truck which is green are categorized to initial F1 measure. And the van in green class consistently exhibits lower F1 measures, indicating weaker performance.

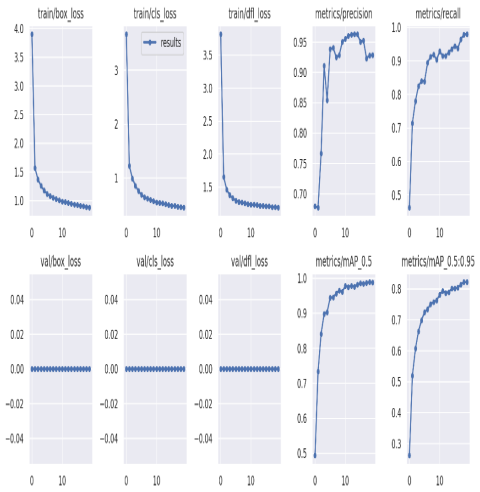

Figure 18. Training and Validation Metrics

The figure 18, illustrate the progression of training and validation metrics, indicating successful model optimization. The training loss curves for box loss, classification loss, and distribution focal loss show the downwards trends, with an initial rapid decrease flowed by a plateau, significantly convergence and improved predictions. Validations loss values remains the close to

zero, suggesting effective generalization without over fitting. Precision increase and plateaus near 0.95, while recall stabilized above 0.9 reflection strong detection of true positive, mAP shows robust performance, with mAP at 0.5 stabilizing around 0.95 and mAP at 0.5:0.95 reaching approximately 0.8 across various intersection over union threshold.

Table 1. Proposed Model Result

Class	Images	Instance	Precision	Recall	mAP 50	mAP 50-95
Bus	500	288	0.93	1	0.995	0.929
Car	500	3554	0.927	0.991	0.992	0.766
Truck	500	42	0.975	0.941	0.988	0.819
Van	500	338	0.88	0.985	0.973	0.773
All	500	4222	0.928	0.979	0.987	0.822

Table 1, exhibits the proposed model performance in vehicle detection across various classes, processing 500 images and detecting 4222 instances with an overall precision of 0.928 and Probability of detection of 0.979. The mAP at 50% IoU is 0.987, which the mAP across IoU thresholds from 50% to 35% is 0.822, indicating strong detection capabilities. For bus the model attains a value of precision of 0.93 and the perfect recall of 1.0, with mAP50 at 0.995 and mAP 50-95 at 0.929. In Car detection, it maintains a precision of 0.927 and recall of 0.991, with mAP50 at 0.992 and mAP 50-95 at 0.766. Truck detection shows high precision 0.975 and solid recall 0.941, with mAP 50 at 0.988 and mAP 50-95 at 0.819. For Vans, the model records a precision of 0.88 and recall of 0.985, with mAP50 at 0.973 and mAP 50-95 at 0.773. Overall, the model demonstrates robust detection performance, particularly surpassing in Bus and Truck categories while maintaining strong results for Cars and Vans, representing its reliability for practical applications in vehicle detection tasks.

4.6 Comparison Analysis

This section illustrate the comparative analysis of the proposed mechanism, with the existing approaches depending on the performance metrics. The table 1 deliberates the comparative analysis of YOLO v9 with YOLO v8 model at mAP values.

Table 2. Comparative Analysis of Proposed Model with Existing Methods [36]

Existing Study	Faster-RCNN	SSD 300	SSD 512	YOLO-V2 Baseline	YOLOv5	Proposed model
Bus	85.49	81.56	84.56	80.86	84.61	99.5
Cars	84.4	84.05	84.46	82.63	86.03	99.2
Van	70.49	71.85	76.64	72.22	77.25	97.3
Other Vehicle	50.29	50.29	50.29	59.57	63.15	98.8
mAP	72.67	73.08	75.99	73.82	77.6	98.7

The table 2 compares the performance of various object detection model in Faster- RCNN, SSD300, YOLO V2 baseline, YOLO v5 and the proposed model across different vehicle categories and their overall mAP. The proposed model significantly outperformance all other in bus detection with an accuracy of 99.5%, surpassing Faster RCNN attains 85.49%, SDD300 for 81.56%, SSD512 for 84.56%, and YOLO-v2 for 80.86%. It also attains 99.2% accuracy for car detection, with YOLO v5 being the closest competitor at 86.03%. For van detection other vehicles, it excels at 98.8%, well above the 50.29% to 63.15% range of other models. Overall, the proposed model attains a mAP of 98.7% significantly higher than YOLO v5. This demonstrates the proposed model greatly enhances vehicle detection performance across all categories compared to existing state of art models, making it a strong candidate for higher accuracy vehicle recognition application.

Table 3. Comparative Analysis of Proposed Model [30]

Model	Cars	Buses	Vans	Other	mAP
YOLOv5s	69.7	43.3	74.3	13.9	50.3
YOLOv5-NAM	0.7	4.60	73.1	17.5	51.7
YOLOv5s	7.12	42.4	73.5	14.3	50.4
YOLOv5-NAM	72	44.6	72.6	18.6	51.9
Proposed model	99.2	99.5	97.3	98.8	98.7

Similarly, table 3 compares the performance of various models including two configurations of YOLO v5 and YOLO v5-NAM and a proposed model across four vehicle categories, like Car, Bus, Van and other, along with that mAP. The YOLO v5s model achieves mAP values of 0.503 and 0.504, with its highest accuracy for van at 0.743 and it's lower for buses at 0.433. The YOLO v5-NAM model shows slight improvement with mAP values of 0.517 and 0.519, performing best with buses 0.46 and other vehicle for 0.175, but still lacks high accuracy. In contrast, the proposed excels across all categories, achieving a mAP of 98.7, with scores of 99.2 for cars, 99.5 for buses, and 97.3 for van and the other vehicle is 98.7, demonstrating a significant improvements in detection accuracy compared to the YOLO v5 models.

The proposed method demonstrate high performance in vehicle detection, achieving an impressive of 98.7, significantly surpassing the moderate accuracy levels of existing YOLO v5 models.

5. Conclusion

The proposed attention driven YOLO v9 model with Multi stage cascaded convolutional model demonstrated exceptional performance in object detection and tracking tasks. The model attained a remarkable mAP of 98.7, with precision values of 99.2 for cars, 99.5 for buses, 97.3 for vans, and 98.9 for other vehicle, showcasing its high accuracy across various classes. Additionally, the model maintains excellent recall rate, ensuring effective identification of objects in diverse scenarios. By integrating advanced techniques such as attention mechanism a multi stage processing, the proposed model addressed the limitation of previous approaches, enhancing both accuracy and real time application. The result indicates that this innovative approach not only outperforms existing state of art model but also sets a new benchmark for future research in vehicle detection and tracking systems. Future work may explore further enhancement by applying the model to more complex dataset and varying environmental conditions to improve its robustness and adaptability.

6. Declaration

Conflict of Interest

There is no conflict of interest.

Funding Support

There is no funding support for this study.

Data Availability Statement

Not Applicable.

References

[1] M. T. Hosain, A. Zaman, M. R. Abir, S. Akter, S. Mursalin, and S. S. Khan, "Synchronizing Object Detection: Applications, Advancements and Existing Challenges," *IEEE Access,* 2024.

[2] A. S. Patel, R. Vyas, O. Vyas, M. Ojha, and V. Tiwari, "Motion-compensated online object tracking for activity detection and crowd behavior analysis," *The Visual Computer,* vol. 39, no. 5, pp. 2127-2147, 2023.

[3] H. Alqahtani and G. Kumar, "Machine learning for enhancing transportation security: A comprehensive analysis of electric and flying vehicle systems," *Engineering Applications of Artificial Intelligence,* vol. 129, p. 107667, 2024.

[4] S. Kurian, M. Faiz, M. Zubair, and S. Ahamed, "Advancing Safety and Efficiency in Human-Robot Interaction," in *2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT)*, 2024: IEEE, pp. 1-8.

[5] S. Kanagamalliga, P. Kovalan, K. Kiran, and S. Rajalingam, "Traffic Management through Cutting-Edge Vehicle Detection, Recognition, and Tracking Innovations," *Procedia Computer Science,* vol. 233, pp. 793-800, 2024.

[6] L. Zhang, W. Xu, C. Shen, and Y. Huang, "Vision-based on-road nighttime vehicle detection and tracking using improved HOG features," *Sensors,* vol. 24, no. 5, p. 1590, 2024.

[7] V. K. Patil, P. Nawade, R. Nagarkar, and P. Kadale, "Object Detection and Tracking Face Detection and Recognition," *Integrating Metaheuristics in Computer Vision for Real-World Optimization Problems,* pp. 25-54, 2024.

[8] D.-y. Ge, X.-f. Yao, W.-j. Xiang, and Y.-p. Chen, "Vehicle detection and tracking based on video image processing in intelligent transportation system," *Neural Computing and Applications,* vol. 35, no. 3, pp. 2197-2209, 2023.

[9] M. Zohaib, M. Asim, and M. ELAffendi, "Enhancing Emergency Vehicle Detection: A Deep Learning Approach with Multimodal Fusion," *Mathematics,* vol. 12, no. 10, p. 1514, 2024.

[10] M. Rasheed, M. Kaleem, M. A. Mushtaq, N. Ahmed, S. Rasheed, and S. Batool, "Deep Learning Approaches for Classification of Vehicles in Intelligent Transportation Systems," 2024.

[11] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using YOLO: Challenges, architectural successors, datasets and applications," *multimedia Tools and Applications,* vol. 82, no. 6, pp. 9243-9275, 2023.

[12] A. A. Mustapha and M. S. Yoosuf, "Exploring the efficacy and comparative analysis of one-stage object detectors for computer vision: a review," *Multimedia Tools and Applications,* vol. 83, no. 20, pp. 59143-59168, 2024.

[13] A. Thakur and S. K. Mishra, "An in-depth evaluation of deep learning-enabled adaptive approaches for detecting obstacles using sensor-fused data in autonomous vehicles," *Engineering Applications of Artificial Intelligence,* vol. 133, p. 108550, 2024.

[14] J. Zhai, B. Li, S. Lv, and Q. Zhou, "FPGA-based vehicle detection and tracking accelerator," *Sensors,* vol. 23, no. 4, p. 2208, 2023.

[15] M. Carranza-García, J. Torres-Mateo, P. Lara-Benítez, and J. García-Gutiérrez, "On the performance of one-stage and two-stage object detectors in autonomous vehicles using camera data," *Remote Sensing,* vol. 13, no. 1, p. 89, 2020.

[16] B. Karbouj, G. A. Topalian-Rivas, and J. Krüger, "Comparative Performance Evaluation of One-Stage and Two-Stage Object Detectors for Screw Head Detection and Classification in Disassembly Processes," *Procedia CIRP,* vol. 122, pp. 527-532, 2024.

[17] N. G. Rani, N. H. Priya, A. Ahilan, and N. Muthukumaran, "LV-YOLO: Logistic vehicle speed detection and counting using deep learning based YOLO network," *Signal, Image and Video Processing,* vol. 18, no. 10, pp. 7419-7429, 2024.
[18] J. Wilson and A. York, "Enhancing Road Traffic Surveillance: Deep Learning Techniques for Vehicle Detection and Tracking," *Journal of Computer Technology and Software,* vol. 1, no. 1, pp. 5-9, 2024.
[19] U. Mittal, P. Chawla, and R. Tiwari, "EnsembleNet: A hybrid approach for vehicle detection and estimation of traffic density based on faster R-CNN and YOLO models," *Neural Computing and Applications,* vol. 35, no. 6, pp. 4755-4774, 2023.
[20] K. SP and P. Mohandas, "DETR-SPP: a fine-tuned vehicle detection with transformer," *Multimedia Tools and Applications,* vol. 83, no. 9, pp. 25573-25594, 2024.
[21] Y. Sakagawa, K. Nakajima, and G. Ohashi, "Vision based nighttime vehicle detection using adaptive threshold and multi-class classification," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences,* vol. 102, no. 9, pp. 1235-1245, 2019.
[22] Z. Qiu, H. Bai, and T. Chen, "Special vehicle detection from UAV perspective via YOLO-GNS based deep learning network," *Drones,* vol. 7, no. 2, p. 117, 2023.
[23] A. Farid, F. Hussain, K. Khan, M. Shahzad, U. Khan, and Z. Mahmood, "A fast and accurate real-time vehicle detection method using deep learning for unconstrained environments," *Applied Sciences,* vol. 13, no. 5, p. 3059, 2023.
[24] T. Barbu, S.-I. Bejinariu, and R. Luca, "Automatic Vehicle Detection and Tracking using Cascade Classifiers and CNN-based Multi-scale Feature Extraction."
[25] X. Han, "Modified cascade RCNN based on contextual information for vehicle detection," *Sensing and Imaging,* vol. 22, no. 1, p. 19, 2021.
[26] P. Jagannathan, S. Rajkumar, J. Frnda, P. B. Divakarachari, and P. Subramani, "Moving vehicle detection and classification using gaussian mixture model and ensemble deep learning technique," *Wireless Communications and Mobile Computing,* vol. 2021, no. 1, p. 5590894, 2021.
[27] Y. Zhang, Z. Guo, J. Wu, Y. Tian, H. Tang, and X. Guo, "Real-time vehicle detection based on improved yolo v5," *Sustainability,* vol. 14, no. 19, p. 12274, 2022.
[28] J. Kim, S. Hong, and E. Kim, "Novel on-road vehicle detection system using multi-stage convolutional neural network," *IEEE Access,* vol. 9, pp. 94371-94385, 2021.
[29] D. M. Chilukuri, S. Yi, and Y. Seong, "A robust object detection system with occlusion handling for mobile devices," *Computational Intelligence,* vol. 38, no. 4, pp. 1338-1364, 2022.
[30] J. Wang, Y. Dong, S. Zhao, and Z. Zhang, "A high-precision vehicle detection and tracking method based on the attention mechanism," *Sensors,* vol. 23, no. 2, p. 724, 2023.
[31] Z. Liu, W. Han, H. Xu, K. Gong, Q. Zeng, and X. Zhao, "Research on vehicle detection based on improved YOLOX_S," *Scientific reports,* vol. 13, no. 1, p. 23081, 2023.
[32] S. Sutikno, A. Sugiharto, R. Kusumaningrum, and H. A. Wibawa, "Improved car detection performance on highways based on YOLOv8," *Bulletin of Electrical Engineering and Informatics,* vol. 13, no. 5, pp. 3526-3533, 2024.
[33] M. M. Rana, M. S. Hossain, M. M. Hossain, and M. D. Haque, "Improved vehicle detection: unveiling the potential of modified YOLOv5," *Discover Applied Sciences,* vol. 6, no. 7, p. 332, 2024.
[34] J. He, H. Chen, B. Liu, S. Luo, and J. Liu, "Enhancing YOLO for occluded vehicle detection with grouped orthogonal attention and dense object repulsion," *Scientific Reports,* vol. 14, no. 1, p. 19650, 2024.
[35] T. Deng, X. Liu, and L. Wang, "Occluded vehicle detection via multi-scale hybrid attention mechanism in the road scene," *Electronics,* vol. 11, no. 17, p. 2709, 2022.
[36] M. A. Z. Al Bayati and M. Çakmak, "Real-Time Vehicle Detection for Surveillance of River Dredging Areas Using Convolutional Neural Networks," *Int. J. Image Graph. Signal Process,* vol. 15, pp. 17-28, 2023.