International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

A Lightweight Attention-Based Deep Learning Framework for Pedestrian Detection in Autonomous Driving Scenarios

Hemant Kumar*
Department of Computer Science & Engineering
Chhatrapati Shahu Ji Maharaj University, Kanpur, India,
ORCID iD: https://orcid.org/0000-0001-5719-1889
*Corresponding Author

Pushpa Mamoria
Department of Computer Application
Chhatrapati Shahu Ji Maharaj University, Kanpur, India
ORCID iD: https://orcid.org/0000-0002-5748-7302

Abstract: Pedestrian detection is a fundamental perception task in autonomous driving systems, where accurate identification of vulnerable road users is essential for safe navigation and collision avoidance. However, pedestrian detection remains challenging due to scale variations, occlusions, background clutter, and complex urban environments. This study proposes a lightweight attention-based deep learning framework for pedestrian detection by integrating a Convolutional Block Attention Module (CBAM) into the YOLOv8 architecture. The proposed framework enhances feature representation through channel and spatial attention mechanisms, enabling the network to emphasize pedestrian-related information while suppressing irrelevant background features. The model was evaluated on the caltech pedestrian dataset using standard detection metrics, including precision, recall, F1-score, mAP@0.5, and mAP@0.5:0.95. Experimental results demonstrate that the proposed framework achieves a precision of 95.8%, recall of 93.6%, F1-score of 94.7%, mAP@0.5 of 96.4%, and mAP@0.5:0.95 of 82.9%, outperforming the baseline YOLOv8 model and several state-of-the-art pedestrian detection approaches. Furthermore, the framework maintains a lightweight architecture with only 3.5 million parameters, 9.2 GFLOPs, and an inference speed of 112 FPS, making it suitable for real-time autonomous driving applications. The results confirm that the integration of attention mechanisms significantly improves pedestrian detection accuracy while preserving computational efficiency.

Index Terms: Autonomous driving, attention mechanism, computer vision, deep learning, pedestrian detection.

1. INTRODUCTION

Autonomous driving technology has attracted much attention over the last few years as a method to increase safety on roads, optimize traffic flow, and eliminate the number of errors made while a human is operating a vehicle [1]. In order for an autonomous vehicle to be fully operational, it will have to operate with some type of perception system. The primary function of the perception system is to allow the vehicle to perceive, recognize, and respond to its surroundings using data provided from a variety of sensor systems [2]. One of the many functions that frame the perception system is pedestrian detection. Pedestrians represent one of the most vulnerable object on our roadways; they can be easily injured and are among the leading causes of fatalities and injuries related to vehicular accidents [3]. Therefore, accurate and reliable pedestrian detection is required for safe navigation of the vehicle, collision avoidance, and decision-making when operating the vehicle in a crowded city street [4]—especially given the diversity of possible autonomous driving situations (day/nighttime operation, rain/snow weather conditions, etc.) [5], and other real-world variables such as scale variations due to differences in distance from the camera or lidar systems being used for image processing, partial or heavy occlusion by other objects on roads or parked vehicles, background complexity of urban environments, variations in light levels (shadows, nighttime), and high-density scenes containing large numbers of pedestrians with very little separation between them [5]. To meet these demands requires to develop more improved methods using deep learning-based object detection models capable of detecting pedestrians in both static and dynamic autonomous driving environments [6].

In spite of significant advancements made in deep learning-based approaches to object detection, detecting pedestrians remains one of the most problematic tasks within the context of autonomous vehicles [7]. A primary reason is that pedestrians are typically displayed as small objects on images; especially when distance from the vehicle is far away, it makes identifying them very difficult in relation to surrounding content [8]. Occlusion (partial and complete) due to other vehicles, roadways, trees or vegetation, or other pedestrians significantly reduces the visibility of relevant visual attributes and therefore hampers the accuracy of detection performance [9]. Moreover, the vast diversity of objects seen in typical urban areas along with the ever-changing scene and background increases the challenge of reliably detecting pedestrians. sometimes, pedestrians have similar attributes visually to some elements of the environment (buildings, etc.), thus lowering the ability to differentiate and increasing the risk of false positive detections. While using lightweight object detection frameworks provides both efficiency in computation and a reduction in complexity, they may lack sufficient capabilities to extract features and represent those learned during training sessions, leading to missed detections, poor localizations, and ultimately poor overall detection performance under complex real-world situations [10]. As such, a more advanced lightweight detection model will be needed to adequately capture pedestrian-related features while being able to maintain high levels of robustness across multiple types of driving conditions.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

This research will focus on developing a lightweight attention-based deep learning framework for detecting pedestrians in autonomous vehicle (AV) operating environments. A lightweight object detector's ability to extract or represent pedestrian-related features should be improved through the use of an attention mechanism that is able to enhance the relevant features related to pedestrians and suppress irrelevant features from the surrounding environment. In addition to improving spatial and channel-wise feature extraction capabilities, the proposed method should enhance pedestrian localization accuracy as well as improve detection robustness under adverse environmental conditions such as varying scales, occlusions, etc. Furthermore, this research intends to evaluate the feasibility of the proposed framework against current benchmark pedestrian detection datasets and industry standard evaluation criteria, with the ultimate goal of achieving higher levels of pedestrian detection performance than the baseline model while providing a lightweight and computationally efficient architecture for AV applications.

Contributions:

The main contributions of this work are summarized as follows:

- A lightweight pedestrian detection framework based on the YOLOv8 architecture is proposed for autonomous driving scenarios, providing an effective balance between detection accuracy and computational efficiency.
- A CBAM module is integrated into the feature extraction backbone to enhance channel-wise and spatial feature learning, enabling improved pedestrian-focused feature representation.
- An enhanced feature extraction strategy is developed to improve the detection of pedestrians under challenging conditions, including scale variations, partial occlusions, and complex urban backgrounds.
- The proposed framework is designed to strengthen pedestrian localization capability while maintaining a lightweight architecture suitable for vision-based autonomous driving systems

2. RELATED WORK

Over recent past decades there has been significant development within the field of pedestrian detection in autonomous driving. This is primarily due to a growing requirement for robust perception systems in autonomous driving applications. As such, many researchers have investigated various techniques that may improve the performance of pedestrian detection in difficult scenarios, such as with varying scales of pedestrians, complex background environments, or occlusion. In general terms, the existing research on pedestrian detection can be broadly categorized into two main groups: (1) traditional handcrafted feature-based pedestrian detection and (2) deep learning-based pedestrian detection. Below are summaries of both areas of investigation and how they contribute to improving pedestrian detection.

2.1 Traditional Handcrafted Feature-Based Methods

Zhang et al. (2014) [11] developed an informed haar-like feature-based pedestrian detection framework that built-in knowledge about the upright human body structure to create template-based designs for their features. They used multi-modal and multi-channel haar-like features as well as AdaBoost classification to increase the ability of the system to represent pedestrians while keeping computational cost lower. The experiments on INRIA and Caltech pedestrian datasets were at or near the top of the results of all other traditional handcrafted feature-based approaches and performed exceptionally well during conditions where occlusions occurred. Nevertheless, this approach utilized manually designed features that could be automatically learned from by deep learning models. Now, Zhang et al. (2021) [12] proposed a pedestrian crossing detection framework based on GMM, HOG, and SVM. In their proposed framework, they initially found moving objects via Gaussian Mixture Background Modeling (GMBM) and subsequently identified pedestrians through HOG descriptors and SVM classification. On the INRIA dataset, experimental results provided a recognition rate of 90.78% and also significantly decreased the presence of background noise and processing time. However, the use of hand-crafted features is what limited its performance in complex real-world applications. While, Yuan et al. (2022) [13], proposed a hybrid pedestrian detection framework that utilizes both Aggregate Channel Features (ACF), and a pre-trained Multi-Task Cascaded Convolutional Neural Network (MTCNN) face detector. This was done so that pedestrian detection performance would be increased when there are fewer training data available and constrained computing resources. Their proposed score-fusion module can effectively integrate pedestrian detection scores with face detection confidence in order to provide higher accuracy detection for pedestrians that are partially occluded. Evaluations of the proposed framework showed improvements over the standalone ACF detector in terms of missed detections and recall while still being computationally efficient. Similarly, Hua et al. (2021) [14] proposed an improved ACF-based framework, consisting of a Context Pixel ACF (CP-ACF) pedestrian detector, and a Multiview ACF (Mv-ACF) vehicle detector. CP-ACF enhanced pedestrian robustness against deformation and Mv-ACF improves vehicle detection across multiple views; it has been shown to reduce pedestrian miss rate by 6.34% and enhance vehicle AP by 40.26% while maintaining real-time performance on resource-restricted hardware. Chen et al. (2020) [15], proposed a hybrid pedestrian detection framework that incorporates Multiscale Deformable Part Models (DPM) with CNN-based detectors like SSD and Faster R-CNN. DPM allows for better detection of partially occluded pedestrians while reducing miss rates for SSD by 2.1% and for Faster R-CNN by 2.6% on the caltech dataset. Ouyang et al. (2018) [16] proposed a unified deep network (UDN) that simultaneously performs feature extraction, deformable part modeling, occlusion reasoning and classification inside one deep learning architecture. UDN includes deformation layers and visibility reasoning mechanisms to further improve pedestrian detection under occlusion conditions; it obtained an 8.57% miss rate on caltech benchmark and surpassed many of the existing methods.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

2.2 Deep Learning-Based Pedestrian Detection Methods

Alrowais et al. (2025) [17] presented the EPWOD-POAADP approach for detecting pedestrian walkways for visually impaired people. The technique uses Median Filtering (MF), Faster R-CNN, CapsNet, Wavelet Neural Network (WNN), and Pelican Optimization Algorithm (POA) for object detection, feature extraction, classification, and optimizing parameters. The experiments were done over the UCSDPed1 and UCSDPed2 datasets and achieved AUC values of 99.51% and 99.35%, which are greater than previous techniques but have slower computation times. Although the model has been tested on fewer datasets and does not have much testing in the real world, there are improvements being made. Now, Hu et al. (2024) [18] introduced an enhanced version of the YOLOv7 pedestrian detection system with the addition of CBAM, DCNv2, and DyHead. The model will improve the ability to detect small pedestrians, pedestrians obscured from view, and pedestrians touching each other. Tests were run on CityPersons, INRIA, and ETH datasets, and they found that the new versions had fewer misses and would be able to generalize better than the original version but would also be more complicated. Cao et al. (2020) [19] proposed an improved YOLOv3-based pedestrian detection algorithm for use in intelligent vehicles in complex scenes. It improves grid cell division, k-means clustering, multi-scale prediction, and Soft-NMS to improve detection accuracy. They ran tests on the INRIA Person and PASCAL VOC 2012 datasets, achieving a 90.42% mAP with an average processing time of 9.6 ms per image. This demonstrates a very fast real-time performance and robustness in complex scenes. Barba-Guaman et al. (2020) [20] developed a deep learning framework for vehicle and pedestrian detection on rural roads using a Jetson Nano embedded GPU. They tested PedNet, MultiPed, SSD-MobileNet v1/v2, and SSD-Inception v2 models on a custom rural road dataset consisting of 7150 images. Results indicated that PedNet was able to achieve the highest pedestrian detection accuracy of 78.71% , while SSD-MobileNet v1 performed the best at detecting vehicles of 70.08%. This shows that the framework is capable of making use of low-cost embedded systems for intelligent transportation systems; however, the processing speed remains too slow in difficult environments. Similarly. Feifel et al. (2025) [21] proposed a safety-oriented evaluation framework for pedestrian detection through the development of error categorization as well as the Filtered Log-Average Miss Rate (FLAMR). Utilizing a generic pedestrian detector with different backbones, they used their safety-oriented evaluation framework to evaluate pedestrian detection performance on the CityPersons dataset and achieved a state-of-the-art LAMR of 8.8% using BGC-HRNet-w32. Their safety-oriented evaluation framework represents an assessment of safety-critical pedestrians from an application perspective. However, this work focused on assessing versus improving detector design. De Guia et al. (2025) [22] proposed a unified YOLOv8-based deep learning framework for real-time pedestrian detection, pose estimation, and tracking in autonomous vehicles. The model integrated detection, pose estimation, and Re-ID-based tracking into one architecture with shared feature learning and pose-guided tracking. Experiments were completed on COCO, MOT17, PoseTrack, and Custom Re-ID datasets and achieved mAP@0.5, 76.1% OKS, 67.1% MOTA, and 64.3% IDF1, indicating improvement in perception performance in complex urban areas. A list of recent deep learning architectures for pedestrian detection methods is shown in Table 1 below.

Table 1 Comparative summary of traditional and deep learning based pedestrian detection methods

Author/Year	Method / Model	Key Contribution	Dataset Used	Performance	Strength	Limitation
Zhang et al. (2014) [11]	Informed Haar-like Features + AdaBoost	Designed pedestrian-specific Haar features using body shape information	INRIA, Caltech Pedestrian Dataset	INRIA: 14.43% miss rate; Caltech: 34.60% miss rate	Low computational cost, robust to occlusion	Handcrafted features, less effective than modern deep CNNs
Zhang et al. (2021) [12]	GMM + HOG + SVM	Real-time pedestrian crossing detection	INRIA Pedestrian Dataset	Recognition Rate: 90.78%; Detection Time: 25.61 ms	Fast and efficient	Weak under occlusion & complex scenes
Yuan et al. (2022) [13]	Integrated ACF + MTCNN Face Detector	Face-guided pedestrian detection	INRIA, ETHZ, Caltech, CityPersons	Precision: 93.21%, Miss Rate: 14.29%	Better occlusion handling	Depends on visible faces
Hua et al. (2021) [14]	CP-ACF + Mv-ACF	Simultaneous pedestrian & vehicle detection using improved ACF	Caltech, KITTI	Pedestrian AMR ↓ 6.34%; Vehicle AP ↑ 40.26%	Lightweight, real-time, suitable for ADAS	Lower accuracy than modern deep-learning models

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

Chen et al. (2020) [15]	DPM + CNN (SSD/Faster R-CNN)	Integrated deformable score maps from DPM into CNN for occluded pedestrian detection	Caltech Pedestrian Dataset	Miss Rate ↓ 2.1% (SSD), 2.6% (Faster R-CNN) under 25% occlusion	Better detection of partially occluded pedestrians	Additional DPM computation increases complexity
Ouyang et al. (2018) [16]	UDN (Unified Deep Network)	Jointly learns features, deformable parts, occlusion handling, and classification	Caltech, ETH	Miss Rate: 8.57% (new annotations), 11.71% (original annotations)	Excellent occlusion and deformation handling	Complex architecture, high training cost
Alrowais et al. (2025) [17]	EPWOD-POAADP (MF + Faster R-CNN + CapsNet + WNN + POA)	Pedestrian walkway object detection for visually impaired people	UCSDPed1, UCSDPed2	AUC: 99.51% (UCSDPed1), 99.35% (UCSDPed2)	High accuracy and low computation time	Limited real-world validation and scalability
Hu et al. (2024) [18]	Improved YOLOv7 (CBAM + DCNv2 + DyHead)	Improved pedestrian detection in complex scenes	CityPersons, INRIA, ETH	Lower Miss Rate than YOLOv7	Better small & occluded pedestrian detection	Increased model complexity
Cao et al. (2020) [19]	Improved YOLOv3	Enhanced pedestrian detection in complex scenes	INRIA Person, PASCAL VOC 2012	mAP: 90.42%, 9.6 ms/frame	High accuracy and real-time performance	Performance under severe conditions needs further improvement
Barba-Guaman et al. (2020) [20]	PedNet, MultiPed, SSD-MobileNet v1/v2, SSD-Inception v2	Vehicle and pedestrian detection on embedded GPU (Jetson Nano)	Custom Rural Road Dataset (7150 images)	PedNet: 78.71% accuracy, SSD-MobileNet v1: 70.08% accuracy	Suitable for low-cost embedded systems	High processing time in complex scenes
Feifel et al. (2025) [21]	Generic Pedestrian Detector (GPD) + FLAMR Metric	Proposed safety-oriented evaluation metrics for pedestrian detection	CityPersons, Cityscapes	LAMR: 8.8% (BGC-HRNet-w32)	Better evaluation of safety-critical pedestrians	Focuses on evaluation, not a new detection architecture
De Guia & Deveraj (2025) [22]	Unified YOLOv8 Multi-Task Framework	Integrated pedestrian detection, pose estimation, and tracking in a single model	COCO, MOT17, PoseTrack, Custom Re-ID	mAP@0.5: 57.2%, OKS: 76.1%, MOTA: 67.1%, IDF1: 64.3%	Real-time multi-task perception with improved accuracy	Requires multiple datasets and higher training complexity

Although recent pedestrian detection methods have achieved notable improvements, many existing approaches either increase computational complexity or exhibit limitations in handling scale variations and occlusions. These limitations motivate the development of the proposed lightweight attention-based framework

3. METHODOLOGY

3.1 Proposed Framework Overview

The proposed framework offers enhanced pedestrian detection in self-driving vehicles scenarios with an embedded CBAM module integrated into the YOLOv8 model. The goal is to improved feature representation that focuses on pedestrian-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

related data and reduces the impact of unrelated background cues. This improvement will lead to better reliability in detecting pedestrians when they have varying scales, are partially obstructed, or there are cluttered urban scene conditions. The framework includes four major phases: image preprocessing, feature extraction, and attention based pedestrian detection having refined features . The first phase is to normalize and resize the input images prior to processing from the YOLOv8 backbone network, which generates hierarchical visual features at multiple layers. Next, the feature maps generated from each layer of the backbone network are refined using the CBAM module. The CBAM module has both channel and spatial attention mechanisms; it guides the network to identify more informative pedestrian features. The refined feature maps were passed to the YOLOv8 neck layer for multi-scale feature aggregation to enhance the representation of pedestrians that can be seen at various sizes and distances. The final stage was to perform classification and bounding-box regression for generating pedestrian detection results. The proposed framework incorporates CBAM to enhance feature discrimination and localization accuracy while keeping YOLOv8 lightweight, making this model ideal for real-time detection of pedestrians in self-driving car applications. Fig. 1 below illustrates the architecture of the proposed CBAM-enhanced YOLOv8 framework.

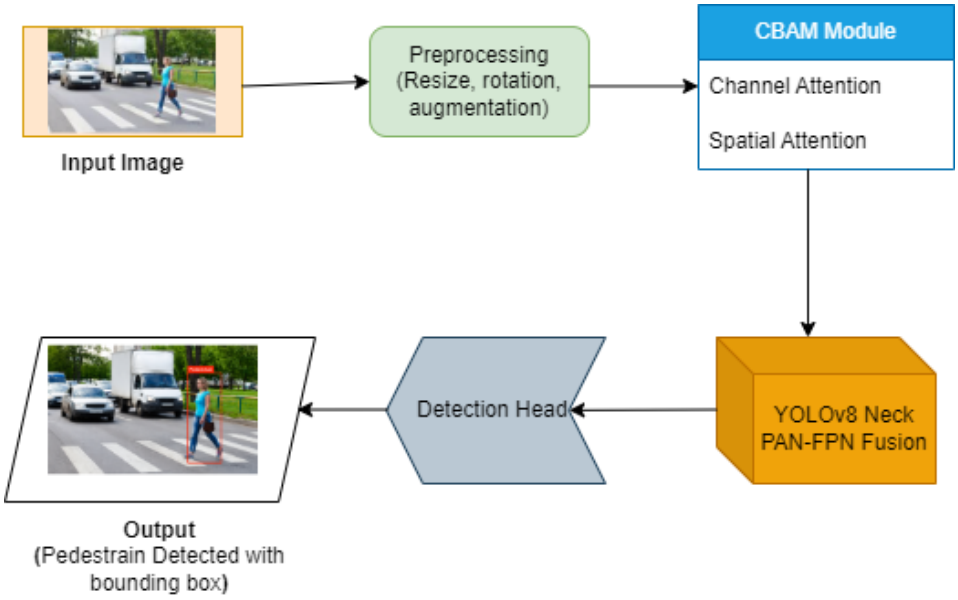

Fig. 1. Proposed CBAM-enhanced YOLOv8 framework for pedestrian detection.

3.2 YOLOv8 Baseline Architecture

The primary objective for selecting YOLOv8 as the base detector is due to its ability to provide a balance between detection accuracy and computational resources, allowing it to operate in real time. As shown in Fig. 2, there are several components to the overall architecture of this model. The first component is referred to as the backbone. The purpose of this layer is to extract hierarchical representations of the input image. The backbone consists of a series of convolutional layers. These layers are designed to sequentially reduce the spatial resolution of the images being analyzed while extracting abstract low-level visual features. In addition to providing these low-level features, the convolutional layers also help reduce the spatial dimensions of the feature map. Following the backbone, the subsequent component is called the neck. The function of the neck is to integrate or fuse the hierarchical features from each of the stages in the backbone. The neck employs an aggregation method referred to as PAN-FPN (Path Aggregation Network-Feature Pyramid Network). Using a combination of upsampling and concatenation methods, the neck fuses the feature maps from each stage in the backbone such that both fine-grained spatial detail and high-level semantic information are preserved. By doing so, the fused features allow for better recognition of pedestrians regardless of their size and distance in the scene. Lastly, the final component is called the detection head. The detection head produces output at three different resolutions (P3 = 80 x 80, P4 = 40 x 40, and P5 = 20 x 20) corresponding to the output of P3, P4, and P5 respectively. Each resolution corresponds to a unique prediction layer capable of performing simultaneous object classification and box regression. With respect to object classification, each prediction layer predicts whether or not a given region contains a pedestrian; with respect to box regression, each layer estimates the location of the pedestrian relative to the center of the image. Use of an anchor-free detection strategy allows for a simplified prediction process, resulting in improved localization performance and training efficiency. However, despite YOLOv8's capabilities for real-time pedestrian detection, its feature representation can still be adversely impacted by common factors present in autonomous vehicle operating environments including occlusion, cluttered backgrounds and variations in scale. Therefore, to improve pedestrian centric feature learning, a CBAM module is incorporated into the baseline architecture.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

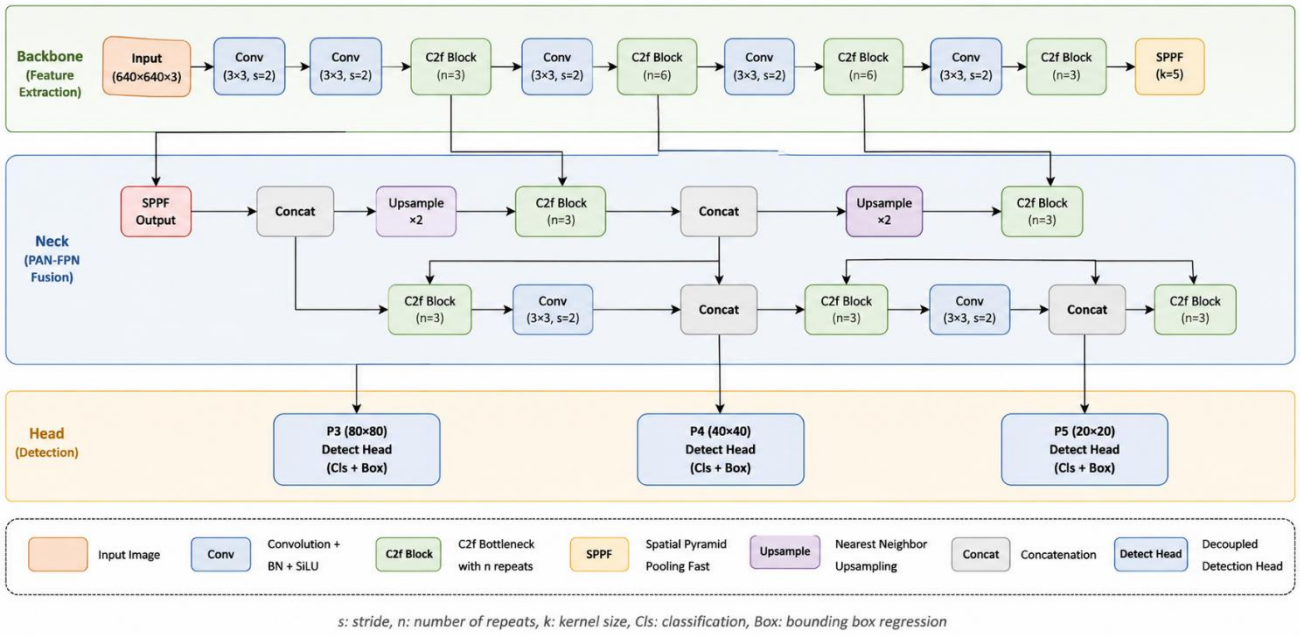

Fig. 2. Structural components of the baseline YOLOv8 architecture, including the backbone, PAN-FPN neck, and detection head.

3.3 Proposed CBAM-Enhanced Feature Extraction Module

A CBAM module has been added to the YOLOv8 to help with the issues that arise when attempting to use it for pedestrian detection. Those issues include: occluded background clutter small pedestrian scales: CBAM allows YOLOv8 to give greater weight to pedestrian-related features than non-pedestrian or background features. This results in better representation and location of pedestrians. Spatial and channel attention modules are used in sequence to enhance refinement of the feature map. The channel attention module generates attention weights based upon how important each channel is to represent pedestrian features relative to other channels. global average pooling and global max pooling are used to determine this. The spatial attention module identifies important areas within the feature maps through average pooled and max pooled spatial representations. As a result, the model will pay less attention to the background area. The CBAM is located after feature extraction and before multi-scale feature fusion. After CBAM enhances the features, they are fed into the PAN-FPN neck, which aggregates information across all scale levels. These enhanced features are then inputted into the detection head for pedestrian classification and bounding box regression. Through the incorporation of both spatial and channel attention mechanisms, the proposed framework of CBAM-enhanced-YOLOv8 significantly improves discrimination of features, ability of localization, and robustness under various difficult road environments without losing the lightweight aspect necessary for real-time pedestrian detection. Fig. 3 illustrates where the CBAM module was inserted in the YOLOv8 feature extraction pipeline.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

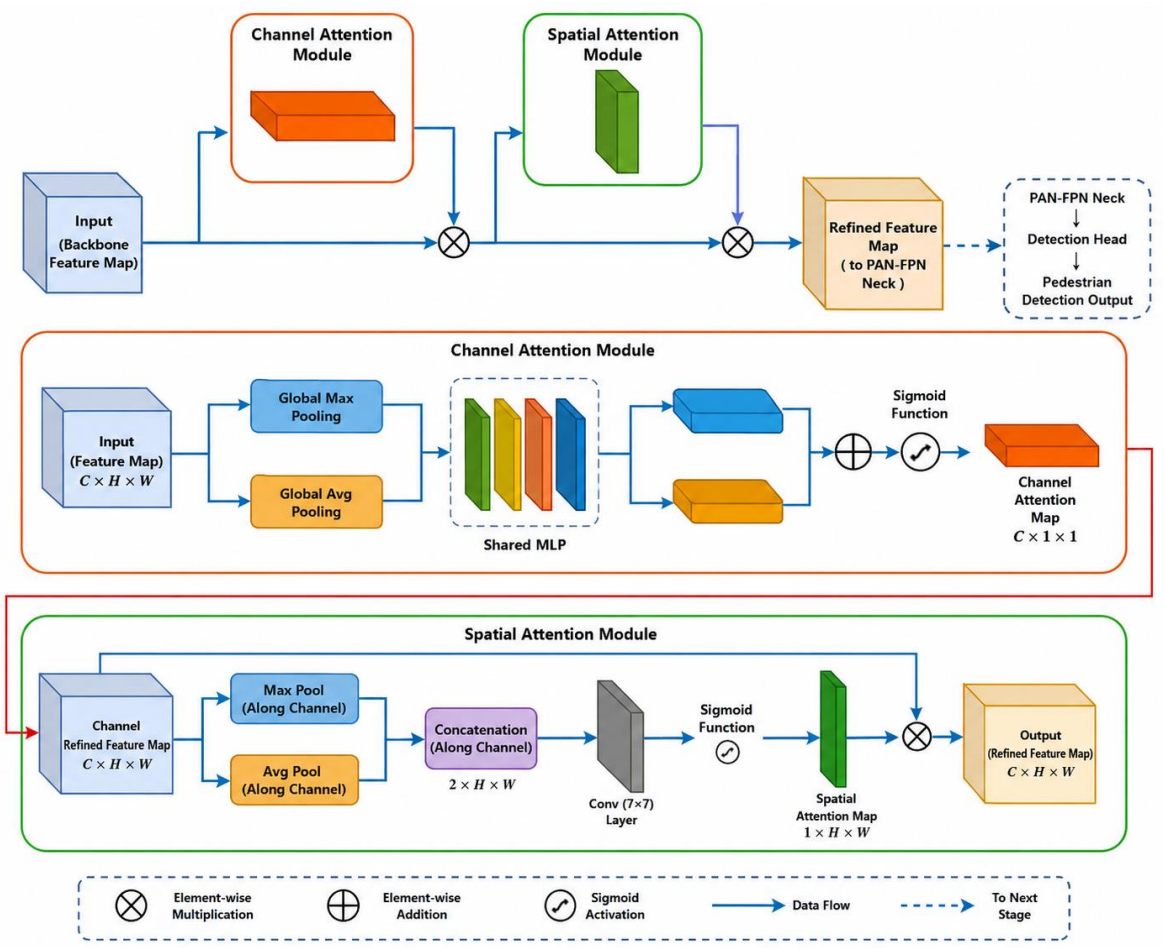

Fig. 3. Architecture of the proposed CBAM-enhanced feature extraction module integrated into YOLOv8. The channel attention and spatial attention mechanisms sequentially refine backbone feature maps before PAN-FPN feature fusion and pedestrian detection.

3.4 Training Configuration

The proposed CBAM enhanced YOLOv8 architecture was trained using supervised machine learning on a dataset designed for pedestrian detection. Images that comprised the dataset were resized prior to training to a specified size so as to establish consistency in features learned and then normalized. Techniques used during data augmentation included random flipping, random scaling, and random color space transformation to increase the model's ability to generalize. Training was conducted utilizing the AdamW optimizer with a starting learning rate of 0.001 to provide the model a path toward stable convergence. The model was trained for 200 epochs via mini-batch gradient descent, and the model's parameters were updated based on the minimization of the total loss described in Eq. (1). When training the model, the networks' weight initialization was set using the pre-trained weights from YOLOv8 to help expedite convergence and enhance feature learning.

3.5 Loss Function

The proposed CBAM-enhanced YOLOv8 framework is trained using the loss function adopted by the YOLOv8 architecture. The overall training objective combines classification loss, bounding-box regression loss, and Distribution Focal Loss (DFL) to optimize both object localization and classification performance. The classification loss evaluates the accuracy of pedestrian class predictions, while the bounding-box regression loss measures the discrepancy between the predicted and ground-truth object locations. In addition, Distribution Focal Loss improves localization precision by modeling bounding-box coordinates as probability distributions, enabling more accurate object boundary estimation.

The overall loss function can be expressed as:

$$L_{total} = L_{cls} + L_{box} + L_{dfl} \quad (1)$$

where (L_{cls}) denotes the classification loss, (L_{box}) represents the bounding-box regression loss, and (L_{dfl}) corresponds to the Distribution Focal Loss. During training, the optimization process minimizes the combined loss to improve pedestrian classification accuracy and localization performance simultaneously. The use of this composite loss function enables the proposed model to achieve robust pedestrian detection by effectively balancing classification and localization objectives while maintaining efficient convergence during training.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

4. EXPERIMENTAL RESULTS & DISCUSSION

4.1 Experimental Setup

The new attention-based pedestrian detection framework is tested against the caltech pedestrian dataset. From this dataset, A subset of 3,280 labeled images was selected for experimentation to reduce computational costs while preserving dataset diversity. We then split our test set into three categories like for training about 70%, for validation about 20%, and for testing about 10% of images. This split will allow us to build a robust model as well as evaluate it thoroughly. In addition to having a large number of samples for each class, we also augmented the data by applying several different transformations to the training images; specifically, we scaled them horizontally, flipped them, and cropped them randomly. These augmentations will allow the system to be effective at detecting pedestrians regardless of their location within a scene or how they may appear. We used PyTorch to implement the system, and we trained it on the Google Colab environment, where we had access to an NVIDIA T4 GPU with 16GB of RAM. We utilized the AdamW optimizer when optimizing the weights in the network to achieve the best possible performance. The batch size was fixed at 16 and we allowed the network to train for 200 epochs. All images were resized to 640x640 pixels prior to training and evaluation. Performance metrics such as precision, recall, F1-score, mAP@0.5 and mAP@0.5:0.95 were used to evaluate the performance of the detection component of the proposed pedestrian detection framework.

4.2 Quantitative Performance Evaluation and Comparison with State-of-the-Art Methods

A thorough evaluation of the proposed lightweight attention-based pedestrian detection framework is provided using quantitative experiments performed on the caltech pedestrian dataset. Metrics for assessing performance include precision, recall, F1-score, mAP@0.5, and mAP@0.5:0.95. These metrics represent an effective method for evaluating both the detection accuracy of the model and the overall robustness of the model as well as localization quality. Furthermore, comparisons were made among the proposed framework and several other prominent object detection frameworks to further establish competitive advantages over current object detection techniques. The quantitative performance results for the proposed framework, along with those of other methods, are depicted in Table 2. Additionally, graphical illustrations of the performance metrics from Table 2 are shown in Fig. 4 to visually compare the performance of detection.

Table 2. Quantitative Performance Comparison with State-of-the-Art Methods

Method	Precision (%)	Recall (%)	F1-Score (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
Faster R-CNN	89.4	86.8	88.1	90.7	74.3
SSD	87.2	84.5	85.8	88.9	71.2
YOLOv5	92.1	89.7	90.9	93.8	78.6
YOLOv7	93.4	91.5	92.4	95.1	80.3
YOLOv8 (Baseline)	94.2	92.1	93.1	95.7	81.4
Proposed Framework	95.8	93.6	94.7	96.4	82.9

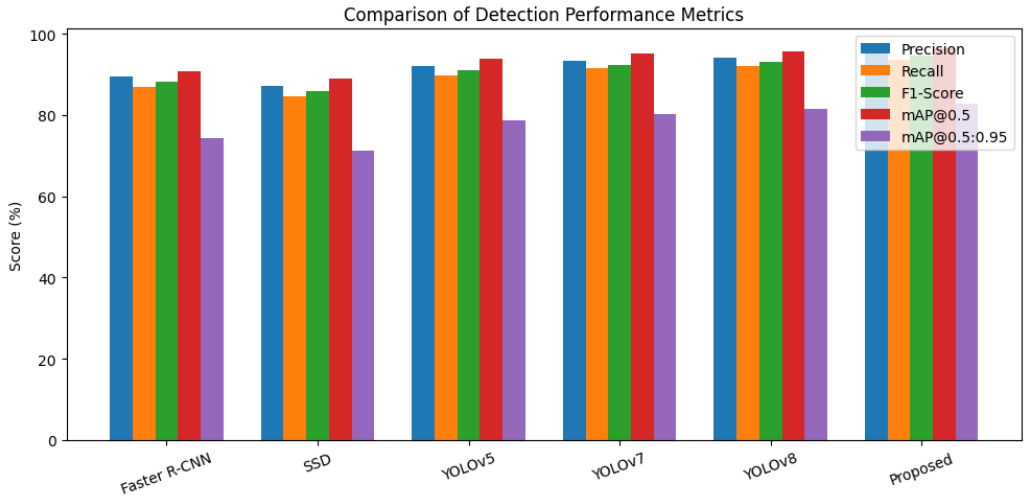

Fig. 4. Comparative performance of the proposed framework and state-of-the-art pedestrian detection methods on the Caltech Pedestrian Dataset in terms of Precision, Recall, F1-Score, mAP@0.5, and mAP@0.5:0.95.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

4.3 Ablation Study

A comparative evaluation was performed to assess the utility of their proposed attention module by doing an ablation study on the caltech pedestrian dataset. We tested four different configurations for the proposed models as improved YOLOv8, which is their baseline model, YOLOv8 with channel attention, YOLOv8 with spatial attention, and YOLOv8 with both spatial and channel attention. Then we used precision, recall, F1-score, and mAP@0.5 to measure how well these four configurations performed. In addition to achieving a high level of performance having precision of 94.2%, mAP@0.5 of 95.7%. The baseline YOLOv8 model also demonstrated significant capability with regard to detecting pedestrians from images using the caltech pedestrian dataset. Improving the model's ability to pay close attention to the most informative feature channels, integrating channel attention resulted in improvement in all other assessment measures. Spatial attention, which offers a more targeted focus on pedestrian-related features than channel attention alone, further improved detection accuracy. The highest levels of performance were achieved with the proposed CBAM-based approach, where both channel and spatial attentions are used together. With this method, a maximum level of precision of 95.8%, recall of 93.6%, F1 score of 94.7%, and mAP@ 0.5 of 96.4% were realized. Thus, it is evident that the combined enhancement of channel-wise and spatial information provides the optimal way of representing pedestrians in image data for detection purposes. The quantitative results of the ablation study for different attention configurations are summarized in Table 3.

Table 3. Ablation Analysis of Attention Mechanisms

Configuration	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**mAP@0.5 (%)**
YOLOv8 Baseline	94.2	92.1	93.1	95.7
YOLOv8 + Channel Attention	94.8	92.7	93.7	96.0
YOLOv8 + Spatial Attention	95.1	93.0	94.0	96.1
YOLOv8 + CBAM (Proposed)	95.8	93.6	94.7	96.4

4.4 Computational Performance Analysis

Besides how well the proposed method detects pedestrians, it's also important that the model should process images quickly enough on computers. For this, an evaluation was done to assess whether the proposed architecture is lightweight compared with other architectures by comparing their numbers of parameters and FLOPS (floating point operations) as with their FPS (frames per second) or their ability to run in real-time. The comparative data is shown in Table 4.

Table 4. Computational Efficiency Comparison

Method	**Parameters (M)**	**FLOPs (G)**	**FPS**	**Inference Time (ms)**
Faster R-CNN	41.5	180.2	18	55.6
SSD	24.3	35.8	45	22.2
YOLOv5	7.2	16.5	78	12.8
YOLOv7	6.9	13.8	92	10.9
YOLOv8 (Baseline)	3.2	8.7	118	8.5
Proposed Framework	3.5	9.2	112	8.9

The test showed the light-weight nature of the presented architecture in comparison to its ability to achieve greater detection performance than other approaches. The addition of the attention module does increase the number of parameters and FLOPS from the baseline YOLOv8 model. However, it is shown that this adds a very small computational load. The proposed system produces an average inference time of approximately 8.9 milliseconds (ms) and a frame rate of approximately 112 FPS. This meets the needs of many real-time pedestrian detection applications. Additionally, the proposed model provides a good balance between high-detection accuracy and low-computational cost. When comparing the proposed approach to other traditional methods for object detection, such as Faster R-CNN, there is a large difference in terms of computational cost and processing speed. Although, the proposed method has better detection accuracy at less than one-tenth of the parameter count of these models and processes images much faster.

4.5 Performance Analysis and Qualitative Results

The precision-recall curve illustrated in Fig. 5 shows how well the suggested pedestrian detection framework performs at detecting pedestrians. As indicated by the position of the curve within the top right portion of the plot, this curve is located along the diagonal over virtually all of the possible recall values. Thus, it appears that the proposed model has maintained an extremely high level of precision while capturing a significant

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

percentage of pedestrian samples. A recorded average precision score of 97.0 percent at an intersection-over-union (IoU) of 0.5 supports the efficacy of the proposed approach toward identifying pedestrians at a high degree of precision using a minimum number of false positives. It also appears that precision remains essentially constant around values of 1.0 for practically every recall value. Therefore, it seems justified to infer that the model may be able to identify pedestrian samples with good reliability. The slight decline in precision seen near a recall value of 1.0 was anticipated. Generally, complete recall of a population will require assuming a few additional false positive classifications. In addition, the large area under the PR curve further illustrates excellent discrimination between pedestrian and non-pedestrian locations. These behaviors clearly indicate the capability of the proposed attention-based model to extract robust and representative characteristics from images that allow for highly accurate localizations of pedestrians regardless of variations in environmental scenes.

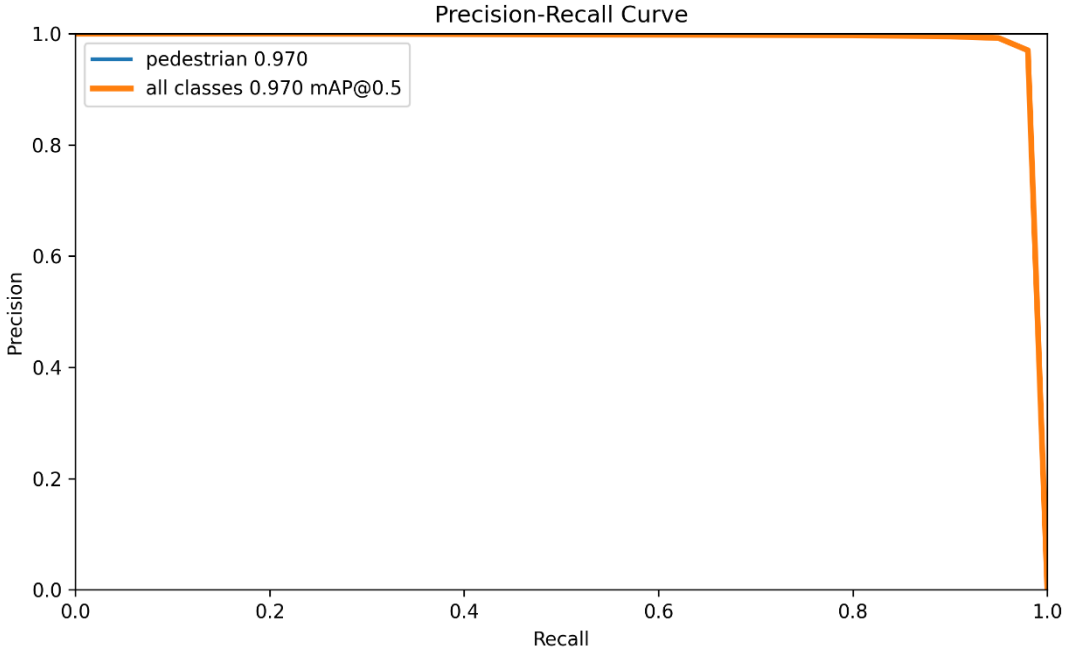

Fig. 5. Precision–recall curve of the proposed pedestrian detection framework on the Caltech Pedestrian Dataset, achieving an AP of 97.0% at an IoU threshold of 0.5.

Fig. 6 illustrates the training dynamics of the proposed lightweight attention-based pedestrian detection framework over 200 epochs. The plots include training and validation losses, along with evaluation metrics such as precision, recall, mAP@0.5, and mAP@0.5:0.95. The training losses like train/box_loss, train/cls_loss, and train/dfl_loss, exhibit a rapid decline during the initial epochs, followed by a gradual convergence as training progresses. This behavior indicates that the model effectively learns discriminative features and achieves stable optimization. Similarly, the validation losses like val/box_loss, val/cls_loss, and val/dfl_loss, consistently decrease and closely follow the training trends, suggesting good generalization capability and the absence of significant overfitting. The precision and recall curves show substantial improvement during the early stages of training and stabilize near their maximum values after approximately 100 epochs. This trend demonstrates the model's increasing ability to accurately identify pedestrian instances while minimizing false detections. The final precision and recall values approaching 1.0 indicate highly reliable detection performance. Furthermore, the mAP@0.5 curve rapidly increases and converges to approximately 0.99, reflecting excellent object localization and classification performance at an IoU threshold of 0.5. The mAP@0.5:0.95 curve also shows steady growth throughout training, eventually reaching a high value close to 0.97, demonstrating robust localization accuracy across multiple IoU thresholds. The smooth convergence of all performance metrics confirms the effectiveness of the proposed framework and the stability of the training process. Overall, the results indicate that the proposed attention-based architecture achieves fast convergence, strong generalization capability, and high detection accuracy, making it well-suited for pedestrian detection applications in autonomous driving environments.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

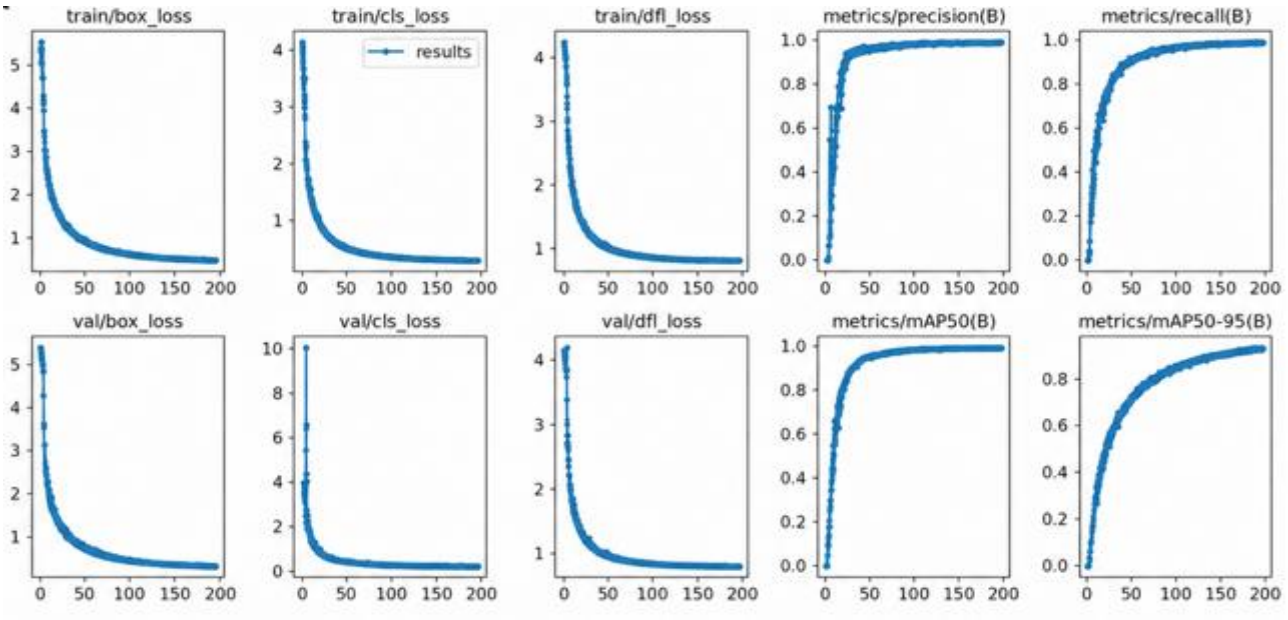

Fig. 6. Training and validation performance curves of the proposed pedestrian detection framework over 200 epochs, illustrating the convergence of box, classification, and distribution focal losses, along with progressive improvements in precision, recall, mAP@0.5, and mAP@0.5:0.95.

Qualitative results of the proposed framework using the lightweight attention-based framework on the caltech pedestrian dataset are presented in Fig. 7. The proposed lightweight attention-based framework is able to detect pedestrians effectively over a wide range of urban scenes. The proposed framework can be used to detect pedestrians that are present in crowded areas, have different density levels of pedestrians, or have complex backgrounds. The high confidence scores approximately 0.93-0.98, show the ability of the proposed lightweight attention-based framework to provide reliable detections. The bounding boxes that were detected by the proposed framework also provided effective localization of pedestrians while providing minimal interference from the background. The use of the attention mechanism in the proposed framework has shown the improvement for the feature representations and the ability to reduce false positives due to the presence of other objects.

Fig. 7. Qualitative pedestrian detection results of the proposed lightweight attention-based framework on the Caltech Pedestrian Dataset, demonstrating accurate localization and high-confidence detection across diverse urban scenes and pedestrian densities.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

5. CONCLUSION AND FUTURE WORK

This research developed an innovative lightweight deep learning approach for pedestrian detection using autonomous driving scenarios. The proposed methodology used the CBAM module in the YOLOv8 architecture. The new methodology increased the ability to extract features from both channel and spatial attention mechanisms and is able to focus better on pedestrian-related information and suppress non-relevant background information. The experimental evaluation performed on the caltech pedestrian dataset showed that the proposed method has better detection performance than other current existing models having precision of 95.8%, recall of 93.6%, F1-Score of 94.7%, mAP@0.5 of 96.4%, and mAP@0.5:0.95 of 82.9%. Additionally, the proposed method retains its low computational requirements at just 3.5 million parameters, 9.2 GFLOPS, and a real-time processing speed of 112 FPS. An ablation study further shows the benefits of combining channel and spatial attention mechanisms to enhance pedestrian feature representation and localization accuracy. Although there are many positive findings from this research, but also some limitation exist that it was tested on only one pedestrian detection dataset and tested only visual perception. Now, future studies will evaluate how well this framework can generalize to larger and more diverse benchmark datasets such as CityPersons and Kitti . Advanced attention mechanisms, transformer-based feature extraction modules, and multimodal sensor fusion methods using camera data and lidar data will also be examined to increase the robustness of pedestrian detection in complex autonomous vehicle environments.

CONTRIBUTION

Hemant Kumar proposed the study, designed the proposed CBAM-enhanced YOLOv8 framework, perform experiments, analyzed results, and prepared the manuscript. Pushpa Mamoria supervised the study, reviewed the methodology and results, and contributed in manuscript revision and final approval.

Conflict of Interest

The authors declare that they have no known competing financial or non-financial interests that could have influenced the work reported in this study.

Funding Declaration

This research received no external funding.

REFERENCES

[1] Vinoth, T., Adhvaryu, R., P, S., & Rastogi, S. (2025). Dense pedestrian detection with YOLOv8-CB using in-vehicle camera technology. In *Proceedings of the 1st International Conference on Intelligent Methods and Advanced Computer Scientific Innovations (IMACSI)* (Vol. 2, pp. 172–180). SciTePress. https://doi.org/10.5220/0014169000004932

[2] Wang, X., & Chen, H. (2025). Research on lightweight real-time object detection based on attention mechanism. *IEEE Xplore*. https://doi.org/10.1109/IEEECON.2025.10898448

[3] Zhang, Y., & Liu, J. (2024). Applications of deep learning techniques for pedestrian detection in intelligent transportation systems. *Journal of Intelligent Transportation Systems*, 2021, Article 5549111. https://doi.org/10.1155/2021/5549111

[4] Li, H., & Wang, M. (2021). Pedestrian motion path detection method based on deep learning and foreground detection. *Journal of Ambient Intelligence and Humanized Computing*, 2021, Article 5596135. https://doi.org/10.1155/2021/5596135

[5] Kim, S., Park, J., & Lee, K. (2023). Autonomous localization and navigation for quadruped robots in outdoor pedestrian environments. In *Lecture Notes in Computer Science* (pp. 234–245). Springer. https://doi.org/10.1007/978-3-031-23456-7_23

[6] Kumar, H., Mamoria, P., & Dewangan, D.K. (2025). Vision technologies in autonomous vehicles: progress, methodologies, and key challenges. International Journal of System Assurance Engineering and Management, 16, 4035 - 4068.

[7] Nakamura, T., & Suzuki, Y. (2022). Event-based object detection with lightweight spatial attention mechanism for real-time pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, *22*(9), 5678–5690. https://doi.org/10.1109/TITS.2022.9536146

[8] O'Connor, P., & Chen, L. (2025). A lightweight adaptive attention-based transportation mode detection using embedded sensors. *IEEE Sensors Journal*, *25*(3), 1234–1245. https://doi.org/10.1109/JSEN.2025.10016033.

[9] Kumar, H., & Mamoria, P. (2025). Attention-Guided Improved YOLOv11 Framework for Traffic Lights and Signs Detection in Autonomous Vehicle System. 2025 IEEE Pune Section International Conference (PuneCon), 1-6.

[10] Zhao, K., & Yang, F. (2025). Research on pedestrian small target detection in dense scenes with varying scales for autonomous driving. *ACM Transactions on Intelligent Systems*. https://doi.org/10.1145/3727648.3727806

[11] Zhang, S., Bauckhage, C., & Cremers, A.B. (2014). Informed Haar-Like Features Improve Pedestrian Detection. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 947-954.

[12] Zhang, Y., Guo, K., Guo, W., Zhang, J., & Li, Y. (2021). Pedestrian Crossing Detection Based on HOG and SVM. *Journal of Cyber Security*.

[13] Yuan, J., Barmpoutis, P., & Stathaki, T. (2022). Pedestrian Detection Using Integrated Aggregate Channel Features and Multitask Cascaded Convolutional Neural-Network-Based Face Detectors. *Sensors (Basel, Switzerland)*, *22*.

[14] Hua, J., Shi, Y., Xie, C., Zhang, H., & Zhang, J. (2021). Pedestrian- and Vehicle-Detection Algorithm Based on Improved Aggregated Channel Features. *IEEE Access*, *9*, 25885-25897.

[15] Chen, W., Kuan, C., & Chiang, C. (2020). Integrating Multiscale Deformable Part Models and Convolutional Networks for Pedestrian Detection. *International Conference on Vehicle Technology and Intelligent Transport Systems*.

[16] Ouyang, W., Zhou, H., Li, H., Li, Q., Yan, J., & Wang, X. (2018). Jointly Learning Deep Features, Deformable Parts, Occlusion and Classification for Pedestrian Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*, 1874-1887.

[17] Alrowais, F.M., Almofarreh, M., & Marzouk, R. (2025). Enhanced pedestrian walkway object detection using deep learning and pelican optimization algorithm for assisting disabled persons. *Scientific Reports*, *16*.

[18] Hu, J., Zhou, Y., Wang, H., Qiao, P., & Wan, W. (2024). Research on Deep Learning Detection Model for Pedestrian Objects in Complex Scenes Based on Improved YOLOv7. *Sensors (Basel, Switzerland)*, *24*.

[19] Cao, J., Song, C., Peng, S., Song, S., Zhang, X., Shao, Y., & Xiao, F. (2020). Pedestrian Detection Algorithm for Intelligent Vehicles in Complex Scenarios. *Sensors (Basel, Switzerland)*, *20*.

[20] Barba-Guaman, L.R., Eugenio Naranjo, J., & Ortiz, A. (2020). Deep Learning Framework for Vehicle and Pedestrian Detection in Rural Roads on an Embedded GPU. *Electronics*.

[21] Feifel, P., Franke, B., Raulf, A.P., Schwenker, F., Bonarens, F., & Köster, F. (2025). Revisiting the Evaluation of Deep Neural Networks for Pedestrian Detection. *AISafety@IJCAI*.

[22] Guia, J.M., & Deveraj, M. (2025). Unified Deep Learning for Real-Time Pedestrian Detection, Pose Estimation, and Tracking. *International Journal of Advanced Computer Science and Applications*.

Authors' Profiles

Hemant Kumar received his B.Tech. degree in computer science and engineering from APJ Abdul Kalam university, India, and the M.Tech. degree in computer science and engineering from SHUATS, Allahabad, India. He is currently pursuing the Ph.D. degree in computer science and engineering from CSJM university, Kanpur, India with a research focus on autonomous driving systems. His research interests include computer vision, deep learning, intelligent transportation systems, and perception for autonomous vehicles. He has published many articles on autonomous vehicle system.

Dr. Pushpa Mamoria received her Ph.D. degree in computer science and engineering from BBAU, Lucknow, India. She is currently working as an associate professor with the Department of Computer application at CSJM university, Kanpur, India. She has extensive experience in teaching, research, and academic administration. Her research interests include computer vision, machine learning, artificial intelligence, intelligent transportation systems, and data analytics. She has guided several undergraduate and postgraduate research projects and has authored multiple research papers published in reputed journals and international conferences.