JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

JSCDM

Vol. 5 No. 2 (2024) 175-187
https://publisher.uthm.edu.my/ojs/index.php/jscdm

Attention-Enhanced YOLOv8 for Accurate Pedestrian Detection and Count Estimation

Mohammed Ahmed Jubair[1]*, Mohammed Mansoor Nafea[2]

[1] *Department of Computer Technical Engineering, College of Information Technology, Imam Ja'afar Al-Sadiq University, 66002 Al-Muthanna, IRAQ*

[2] *Department of Computer Technical Engineering, Al-Maarif University College, Al-Ramadi, 31001, IRAQ*

*Corresponding Author: mohamed.a.jubair@gmail.com
DOI: https://doi.org/10.30880/jscdm.2024.05.02.013

Article Info

Received: 17 September 2024
Accepted: 20 November 2024
Available online: 18 December 2024

Keywords

Autonomous vehicles (AVs), You Only Look Once (YOLO), Computer Vision(CV), Object Detection (OD)

Pedestrian Detection (PD), Deep Learning (DL)

Abstract

Autonomous vehicles (AVs) are crucial for improving the safety of highways by reducing man-made errors that account for the most dominant source of road accidents. They also offer significant potential for increased efficiency in transportation, lowering emissions, and enhancing mobility for the elderly and disabled. However, significant obstacles still need more study to be overcome. For example, in crowded metropolitan settings, AVs must correctly sense their surroundings to operate safely. Numerous research papers explore effective methods for precisely determining the surroundings of AVs. However, there are still challenges to detecting non-static objects, especially pedestrians. In this article, deep learning algorithms have been employed to realize the real-time detection of pedestrians, resulting in the advancement of AVS development. By enhancing feature extraction within the detection algorithm, This paper presents the Convolutional Block Attention Module (CBAM-YOLOv8) model, which integrates three CBAM modules into the backbone network of YOLOv8. Additionally, the DUA-YOLOv8 and ECA-YOLOv8 models will be introduced. The results obtained from the experiments on a combined dataset of 1520 images collected from the INRIA and ETH datasets indicate that CBAM-YOLOv8 slightly improves recall and mAP@0.5:0.95. The proposed model reached a mAP@0.5:0.95 of 0.57 and 42 FPS on an NVIDIA Tesla T4 GPU. The evaluation metrics for CBAM-YOLOv8 showed greater enhancements compared to SE-YOLOv8 and ECA-YOLOv8.

1. Introduction

Daily, the impact of autonomous vehicles (AVs) is experienced in the lives of humans and shows the potential to change how they are planned and piloted. One benefit that urban drivers can enjoy from the use of autonomous vehicles is the reduction in the time used in searching for a parking lot. Regardless of the benefits that can be derived from the use of AVs, making it a part of our daily lives will be impossible, except some of the challenges that accompany it are addressed. There is a huge problem of trust amongst different stakeholders including drivers, AV passengers, and pedestrians. All these people must demonstrate a certain level of trust for autonomous vehicles in order to make them a part of their daily lives. The functionality and uniqueness of AVs lie within the most current advancements in the areas of object tracking, detection, regulation, and protection [1]. This technology has been used in different areas of life including GPS and inertial measurement units for localization,

detecting and preventing obstacles through the use of computer vision, laser, LIDAR (light identification and ranging) [2], and sonar technology. All the aforementioned technology enable the prevention of road accidents, especially, those that occur as a result of human error. These human errors are responsible for a large amount of accidents that occur on the roads. In addition to these benefits, they benefit the user in terms of fuel economy, road safety, and eco-friendliness.

A key characteristic of AVs that makes them unique is their ability to detect objects around them to identify and categorize them. Through this process, a large number of specialized applications and disciplines like AV applications and pedestrian and face detection apps have emerged. Vision as a key concept in this area refers to the ability to evaluate and make inferences from a picture that is received. It goes beyond the mere perception of a picture in one's brain. This ability can also be stimulated in computers for the advancement of extant technologies, and this requirement can be achieved through object recognition technologies. Due to the advancements in existing technologies, the issues of accuracy and precision no longer pose a challenge as compared to the challenge of object identification. This implies that a key consideration in the advancement of technology is the ability of the technology to detect objects in rereal time It is critical for a vehicle to be able to sense and detect its environment in real-time.

AVs must be designed with capabilities that allow them to detect objects in a fast and effective manner [3]. However, the greatest challenge to achieving this lies in the selection of a model that supports real-time detection of objects in an accurate and precise way within the shortest possible time in the field of autonomous vehicles. To address this challenge, many scholars and professionals have employed the use of models like You Only Look Once series (YOLO~YOLOV6 series) [4], regional-based convolutional neural networks (R-CNN), faster regional-based convolutional neural networks (FR-CNN)[5], and YOLOLITE[6]. Despite the existence of several object detection models in autonomous vehicles, there are still huge challenges that have been observed including long computational time, and low detection accuracy, which impact the tracking abilities of the AVs in real time [7]. In recent times, it has become important to design a highly efficient object detection system that is characterized by short computational time, high accuracy level, and low cost of computation [8].

YOLOv8 is regarded as one of the best object detection models that has demonstrated capabilities in addressing the issue of accuracy associated with AVs. It has wide applications in the areas of 3D object detection, and detection of faces and images at one instance rather than in bits and pieces. The YOLO is one of the algorithms that is known for its ability to process images as a whole rather than using the region/window-based selection technique. Through this technique, the YOLO is able to achieve 45 FPS with higher speed as compared to other existing object detection systems and algorithms. Due to the wide acceptance and speed of YOLOv2, it has emerged as the best candidate as a foundation for YOLO-LITE. On the other hand, the development of YOLOv8 was achieved through the use of the YOLO series which made it possible to enhance the accuracy, and speed with which it detects objects while minimizing parameters. In comparison with state of the art model, YOLOv8 has demonstrated the highest level of accuracy. However, there is no evidence of thorough testing of this capability in autonomous vehicles.

The main contribution of this paper is proposing and assessing three modern YOLOv8-based models: CBAM-YOLOv8, DUA-YOLOv8, and ECA-YOLOv8 to improve pedestrian detection in AVs. The improvement in the features extracted from the image due to the integration of CBAM into the YOLOv8 enhances the model's performance in object detection. Our work is primarily motivated to overcome observed difficulties in detecting non-static objects such as pedestrians in complex contexts. Our suggested models demonstrate a significant increase in accuracy and recall based on trials conducted on the INRIA AND ETH datasets, respectively. As a result, they may be considered predictive of developments in AVs in the future.

2. Review of Related Work

This section presents an elaborate review of several empirical and theoretical research on autonomous vehicles, focusing on models and techniques that can be deployed in deep learning to improve AVs. The subsequent paragraphs present related works that have focused on pedestrian detection.

In the study carried out by Shao et al., (2024) [9], a lightweight UAV image target detection technique was proposed based on YOLOv8 and called Aero-YOLO. The proposed method was aimed at solving the problems of low computational power, low level of detection accuracy from small target sizes in images, as well as missed detections resulting from densely arranged targets. This proposed technique achieves this through the replacement of the initial Conv module with BSConv, and also replacing the C2f module with C3 to reduce the parameters of the model, extend the receptive field, and improving the computational efficiency of the model. Results of the experiment conducted on the VisDrone 2019 dataset show that the proposed algorithm is capable of enhancing the speed and accuracy of pedestrian and vehicle detection with a mAP of 43% and 41.61 M parameters.

The research carried out by Gong et al., (2024) [10] suggests using several dilated convolutions to sample feature images to increase the algorithm's performance for target detection and feature extraction while avoiding

the loss of information due to recurring sampling. Furthermore, a lightweight shuffle-based efficient channel attention (SECA) technique is presented to carry out parallel processing for every sub-feature map channel and grouping in the channel dimension. To improve the channel feature information for multiscale feature representation, a new branch is presented. The approach is thoroughly tested on many difficult pedestrian recognition datasets, yielding mean average precision (mAP) values of 87.73%, 34.7%, 93.96%, and 95.23% on PASCAL VOC 2012, MS COCO, Caltech Pedestrian, and INRIA Person, in that order. The outcomes of the experiment show how successful the strategy is.

In the study conducted by Chen and Wan [11], a new transformer-fusion-based YOLO (TF-YOLO) detector was introduced to detect pedestrians in a wide range of illumination settings like heavy rain, smog, and nighttime. The detector is intended to address the problem whereby there is no correspondence between the application environment's illumination conditions and the experimental data's lighting conditions, which can potentially cause significant degradation in the model's ability to detect effectively. To effectively integrate latent interactions between multimodal pictures (visible and infrared images), the author has developed a unique transformer-fusion module that seamlessly integrates into a two-stream backbone network. On the difficult multi-scenario multi-modality dataset, the suggested TF-YOLO significantly decreases the miss rate of the state-of-the-art strategy by around 6% and significantly increases its average precision by 3.3%.

Luo et al. [12] To address the issues raised by the present infrared pedestrian-vehicle detection systems, poor identification accuracy, and significant computing load, an infrared pedestrian-vehicle detection method A proposition is put forth, predicated on an improved iteration of YOLOv5. First, to improve the accuracy of recognizing tiny objects, a head made expressly for that purpose has been incorporated into the model to use shallow feature information fully. Second, to solve difficulties with target overlap and category imbalance, the Focal Generalized Intersection over Union (GIoU) is used in place of the initial loss function. Third, the model's computational load is reduced without noticeably lowering detection accuracy by utilizing the distribution shift convolution optimization feature extraction operator. According to test findings, the enhanced algorithm's average accuracy (mAP) is 90.1%. In particular, the modified algorithm's Giga Floating Point Operations Per Second (GFLOPs) is only 9.1. On comparable GFLOPs, however, the enhanced algorithms performed better than the others, including YOLOv6n (11.9), YOLOv8n (8.7), and YOLOv7t (13.2).

Nan et al. [13] suggest that an enhanced pedestrian recognition method be used to address the current YOLO's high missed detection rates and poor accuracy for occlusion and multi-scale pedestrian targets. To improve cross-scale feature extraction capabilities, the YOLO backbone is changed. In front of the YOLO layers, a spatial pyramid pooling module and two attention methods are implemented at distinct points to improve the pedestrian feature fusion capabilities of different sizes. The real scenario prunes the network structure to optimize the model training efficiency and detect performance deterioration caused by the extremely complicated network module. Results of the experiments show that the YOLO-SSC model can significantly minimize missed detection rates in the event of occlusion and enhance the accuracy and speed of detecting medium and small pedestrian targets when compared to YOLOv3 and other models.

In conclusion, this work has focused on extending the past works conducted on the problem of pedestrian detection by proposing a new model based on YOLOv8 to be used in AVs. Some similar work that has been done in the past are [9] and [10], where some different lightweight architectures and attention mechanisms have been introduced for enhancing the detection in UAV images and complex scenarios, while our contribution focused on using pyramid spatial pooling and attention mechanisms for dealing with occlusions and pedestrian targets in different scales. These enhancements are incorporated into our model to deliver enhanced detection rate and speed in an efficient platform for pedestrian recognition in real-time AVs.

3. Materials and Methods

3.1. YOLOv8

This paper introduces a computer vision-based recognition and detection algorithm for object detection and pedestrian population statistics. With this algorithm, AVs can acquire real-time information on pedestrian population and behavior dynamics, facilitating rapid management and strategy development for AVs. This approach aims to optimize pedestrian safety and enhance financial benefits.

Due to the low density of the pedestrian population in the INRIA dataset and the real-time requirements for population statistics, this research utilized the latest version of the YOLO (You Only Look Once)v8 model. YOLOv8 is a one-stage object detection algorithm. Figure 1 shows the network architecture diagram of YOLOv8, which consists of four main modules: input station, backbone, head, and predictor, along with four core components: CBS, MP, C2F, and SPPF. The preprocessing method of the YOLOv8 model is an advancement in this series, incorporating improvements and optimizations from previous versions such as YOLOv4 and YOLOv5. Additionally, the use of Mosaic data augmentation is particularly effective for detecting small objects [15], [16] Regarding its architecture, YOLOv8 incorporates the C2F Network in its backbone. Also, the YOLOv8 constantly

improves the learning capability of the network with no interruption of the gradient route through the use of expand, shuffle, and merge cardinality methods. More so, the application of convolution is made to improve the cardinality and channel of the computational blocks in the architecture. Through this process, various sets of computational blocks are enabled to learn a wider range of features, which in turn improves the overall performance and feature diversity of the model [17].

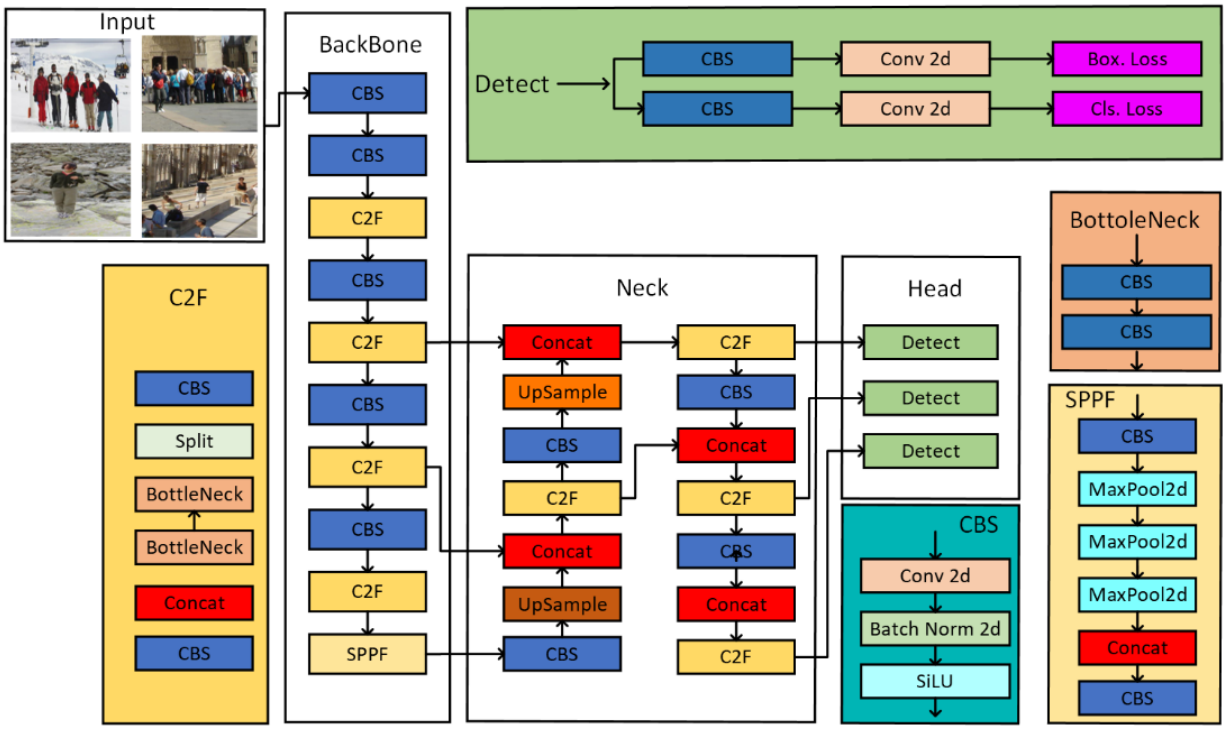


Fig. 1 *The network structure of YOLOv8 [14]*

YOLOv8 introduces several optimizations and enhancements over its predecessors, including YOLOv5. These improvements encompass architectural changes, training methodologies, and post-processing techniques. Here are the key optimizations in YOLOv8:

1. New Neural Network Architecture:
   - C2F Network: YOLOv8 incorporates the C2F (Cross-Stage Partial and Focus) Network in its backbone. This network utilizes expand, shuffle, and merge cardinality techniques to enhance the model's learning ability without disrupting the gradient path, allowing the model to learn more diverse features.
   - Feature Pyramid Network (FPN) and Path Aggregation Network (PAN): The combination of FPN and PAN produces feature maps at different resolutions and scales, which enhances the accuracy of objects of different shapes and sizes.
2. Anchor-Free Detection: Unlike YOLOv5, which uses anchor boxes for bounding box prediction, the technique adopted by YOLOv8 is an anchor-free technique that carries out direct prediction of object centers and minimizes the number of box predictions, thereby increasing the speed of post-processing and training enhancements.
   - Larger and More Diverse Dataset: YOLOv8 was trained on a mixture of the COCO dataset and several other datasets, providing a broader range of images and improving performance across diverse image types compared to YOLOv5, which was primarily trained on the COCO dataset.
   - Mosaic Augmentation: Both YOLOv5 and YOLOv8 through the use of mosaic augmentation, which merges four random images into one, enhancing the model's robustness by providing more varied training samples.
3. Advanced Labeling Tool: RoboFlow Annotate: YOLOv8 introduces RoboFlow Annotate, a new labeling instrument that simplifies the procedure required for image annotation. Features like auto labeling, labeling shortcuts, and customizable hotkeys characterize this instrument. This tool improves the efficiency of preparing training data compared to the Labeling used in YOLOv5.
4. Improved Post-Processing Techniques: Soft-NMS: YOLOv8 employs Soft-NMS (Non-Maximum Suppression), which uses a soft threshold in the overlapping of bounding boxes rather than eliminating them. This helps

retain useful detections and improves overall prediction accuracy compared to the standard NMS used in YOLOv5.

3.2. Acquisition of Materials

The benefits of autonomous driving are clear: improved road safety, reduced driver fatigue, lower fuel consumption, and decreased traffic congestion, all of which are precious areas of research. This work used a combined dataset from the INRIA and ETH datasets, which are widely used images for training and evaluating object detection algorithms, particularly those focused on pedestrian detection. It features diverse, high-quality images from various urban settings, aiding in the development of robust computer vision models. During the dataset preparation, initially sourced the data from the Roboflow Universe. Ultimately, our dataset comprised 1520 images, divided into 150 for the training set, 300 for the test set, and 170 for the validation set. To enhance training and testing accuracy, re-labeled the dataset. An example of the process of dataset labeling is presented below.

(a) (b) (c) (d)

Fig. 2 *Sample steps for explaining data for the entire human being*

3.3. Pre-Processing of Data

3.3.1 Mixup Data Augmentation

Mixup refers to the method of augmenting data innovatively, which is dependent on a straightforward [18], data-independent principle. It creates new training samples and labels through linear interpolation. The following is the formula used in data processing:

$$\widetilde{x} = \lambda x_i + (1-\lambda)x_j \tag{1}$$

$$\widetilde{y} = \lambda y_i + (1-\lambda)y_j \tag{2}$$

In this context, the two data pairs (xi, yi) and (xj, yj) are training sample pairs from the original dataset (each consisting of a training sample and its corresponding label). The parameter λ follows the distribution of β. The resulting $\tilde{x}$ is the training sample created through the Mixup data augmentation, and $\tilde{y}$ is its corresponding label. Figure 3 illustrates the data results for Pedestrians images after applying the Mixup data augmentation with various fusion proportions, where lamα and lamβ are the image fusion ratios and lamα+lamβ=1.

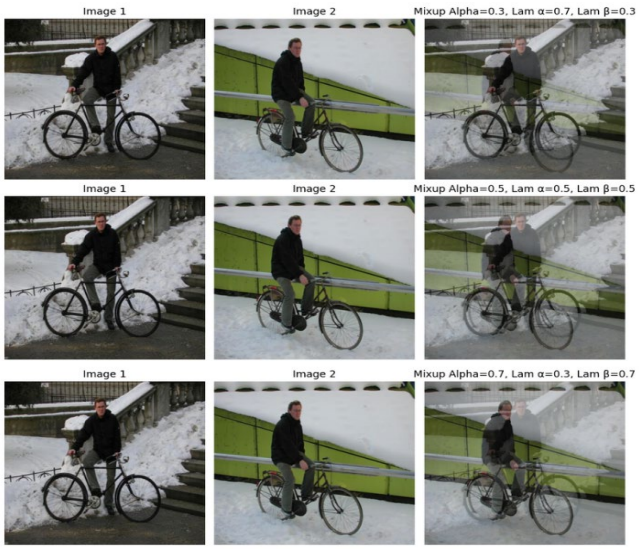


Fig. 3 *Results of data enhanced using the Mixup process with a variety of fusion proportions*

3.3.2 Augmentation of Mosaic Dataset

The Mosaic dataset has been augmented via the YOLOv4 network, which randomly involves cutting and merging four images into a single new image to create training data. This process significantly enriches the detection dataset, enhances the network's robustness, and lowers GPU memory usage [16]. Figure **4** illustrates the Mosaic data augmentation workflow.

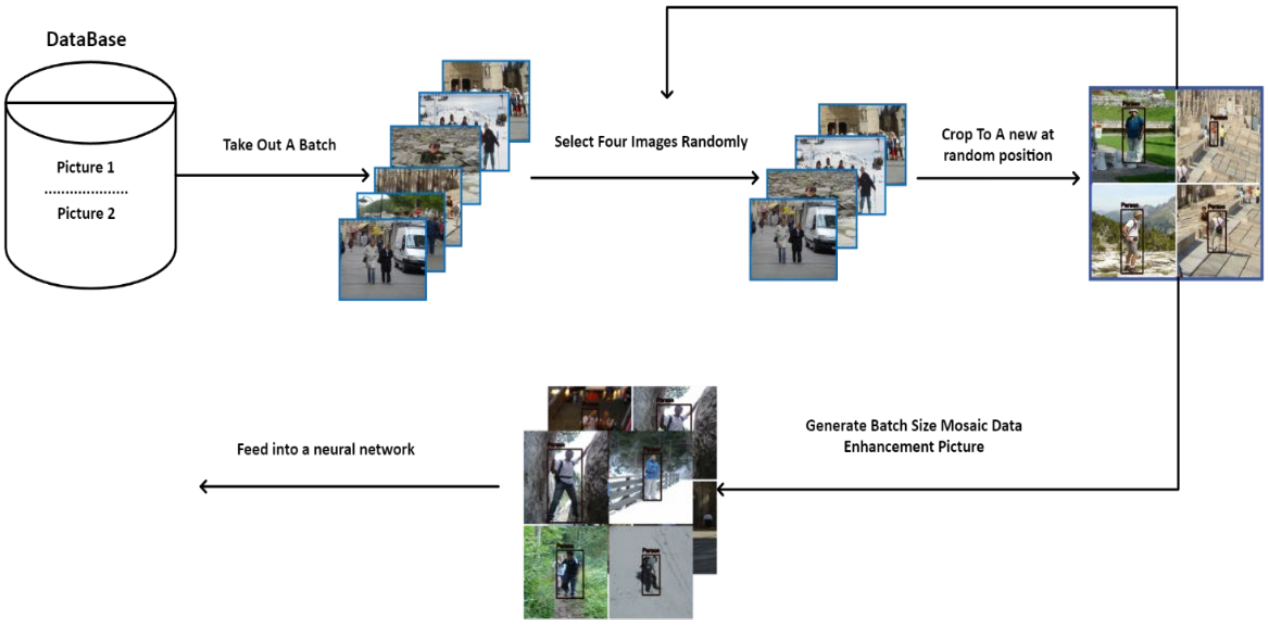


Fig. 4 *Workflow of Mosaic data augmentation*

3.4. Experimental Environment

The project computer featured an NVIDIA Tesla T4 GPU, equipped with 2560 CUDA cores and 16GB of video memory. CPU was an 8-core (Intel Xeon Gold 6330) running at (2.20 GHz), complemented by 30 GB of RAM. The system operated on Linux and utilized PyTorch version 2.0.1, Python version 3.10, and CUDA version 11.7

3.5. Training Parameters

The experiment displays the training parameters utilized in its training process, as detailed in Table 1.

J. of Soft Computing and Data Mining Vol. 5 No. 2 (2024) p. 175-187

Table 1 *Training parameters*

Parameters	Value
Learning Rate	0.001
Batch Size	16
Image Size	640 × 640
Weight Decay	0.004
Momentum	0.9
Epochs	80

3.6. Training Parameters

To assess the algorithm's performance, this study used several evaluation metrics: mean average precision (mAP), recall (R), precision (P), F1 score, and frames per second (FPS), where precision denotes the proportion of true positive samples among all samples predicted to be positive. The precision formula is:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

The recall represents the proportion of actual positive samples correctly identified among all positive samples in the data set. The calculation formula is:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

The F1 score is the weighted average of precision and recall. The F1 is calculated as follows:

$$\text{F1} = \left(\frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}} \right) = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

Precision is the parameter used to measure a model's capability to identify negative samples correctly. The ability of the model to achieve a higher precision rate is a reflection of the model's excellent performance in terms of differentiating negative samples. On the other hand, the model's ability to correctly identify positive samples is referred to as recall, and a higher recall rate achieved by the model reflects the model's high performance in that regard. Jointly the recall and precision can be measured together using the F1 parameter to provide a balanced assessment of the model's performance. A higher F1 score indicates that the model has achieved strong performance across both precision and recall, reflecting its robustness. The average precision (AP) is the mean of the highest precision values across various recall levels (typically computed separately for each category). The equation below is used to compute this parameter:

$$AP = \frac{1}{11} \sum_{0,0.1\ldots1.0} P_{smooth}(i) \tag{6}$$

In Pascal VOC 2008[19], the IOU threshold is fixed at 0.5. In cases of multiple detections of objects, the one who has the highest confidence is considered positive., while the others are deemed negative. The smoothed precision-recall (PR) curve calculates precision at 10 evenly spaced thresholds (including 11 points) along the horizontal axis from 0 to 1 and averages these values to obtain the final average precision (AP) value. The mean average precision (mAP) is the average of the average precision values for all classes. Below is the formula for the computation of this parameter:

$$\text{m}AP = \frac{\sum_{j=1}^{S} AP(j)}{S} \tag{7}$$

Where S represents the total number of categories, and the Numerator is the sum of the APs of all categories. This study involved the detection of only one kind of object. Therefore, AP equals mAP.

4. Improved YOLOv8

4.1. Improved YOLOv8 via Attention Mechanism

The attention mechanism is described as a popularly used method of data processing deployed in different machine-learning tasks in a wide variety of fields [20]. The term attention mechanism in the field of computer vision refers to the mechanism through which relationships within the original data can be recognized,

highlighting critical characteristics. The different kinds of attention mechanisms include multi-order attention, channel attention, and pixel attention amongst several other kinds of attention mechanisms. The CBAM is made up of both a spatial attention module and a channel attention module [21]. Figure 5 below shows the architecture of the module 5.

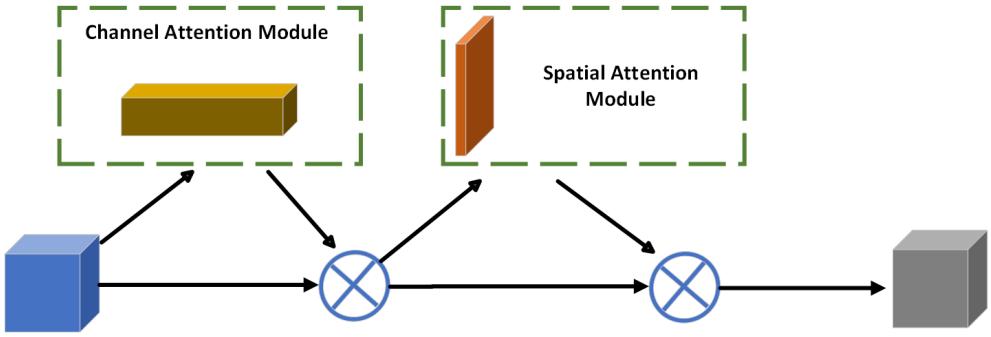


Fig. 5 *Diagram of the CBAM module structure*

CBAM [22] can be described as a lightweight attention mechanism capable of carrying out attention operations within the spatial and channel dimensions. It consists of a spatial attention module (SAM) and a channel attention module (CAM). While the latter allows the network to focus more on the foreground of the image as well as relevant areas with useful details, the former, which is SAM allows the network's attention to be focused on areas that possess rich contextual information even without navigating the whole image [23], [24].

4.2. YOLOv8-based CBAM Attention Mechanism

The CBAM was integrated into the YOLOv8 architecture [17], [22], illustrated in Figure 6. This module aims to improve the network's ability to extract features. In the current study, the incorporation of the attention mechanism into the main network altered a number of the original weights, causing the network to produce erroneous prediction results. This problem is addressed by adopting a selective attention mechanism to improve the feature extraction process without compromising the original network features.

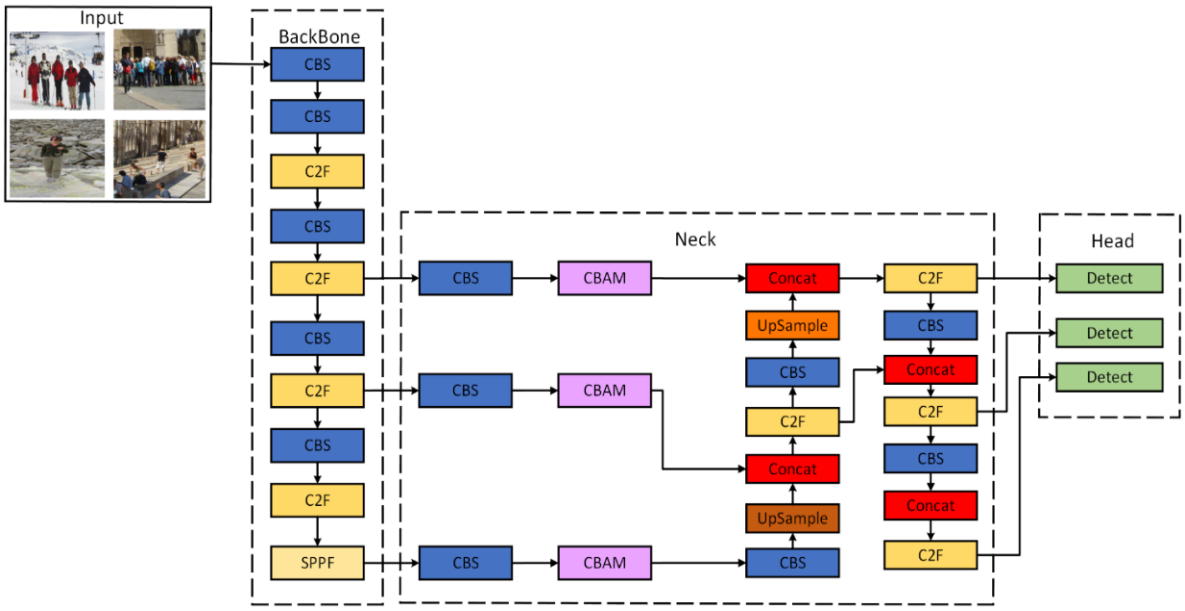


Fig. 6 *The YOLOv8 network structure with the CBAM attention mechanism*

The channel attention unit produces two feature maps of the size of 1 × 1 × C t through the application of the global average pooling (GAP) and global max pooling (GMP) to the input feature map of H × W × C. A two-layer, multilayer perceptron receives the two feature maps. ReLU serves as the activation function, while C/r, or the reduction rate, demonstrates the number of neurons present within the MLP's first layer. The weights of two layers of the neural network are shared, and the second layer has C neurons. Subsequently, element-wise computation is employed to combine the output features, and the final channel binding property is produced by

J. of Soft Computing and Data Mining Vol. 5 No. 2 (2024) p. 175-187

activating the sigmoid function. Most importantly, it is worth noting that obtaining the feature of spatial attention involves the multiplication of the channel attention feature by the original input feature map [21]. The spatial attention module's input is the previous stage's feature map. Two distinctive feature maps with a size of H × W × 1 are obtained after GMP and GAP. Next, the concatenation function is executed. Following the feature map's dimensionality reduction, sigmoid activation produces the spatial attention feature. To obtain the final feature map, the spatial attention is multiplied by the input feature map [22].

5. Experiment Results

To evaluate the effect of CBAM-YOLOv8, the method of this work borrowed the Dual Attention Network (DANet) and the Effective Channel Attention (ECA) modules to replace the CBAM modules in the ablation experiments. The DANet network has two main components: the channel attention module and the position attention module [25]. Figure 7 shows the module structure.

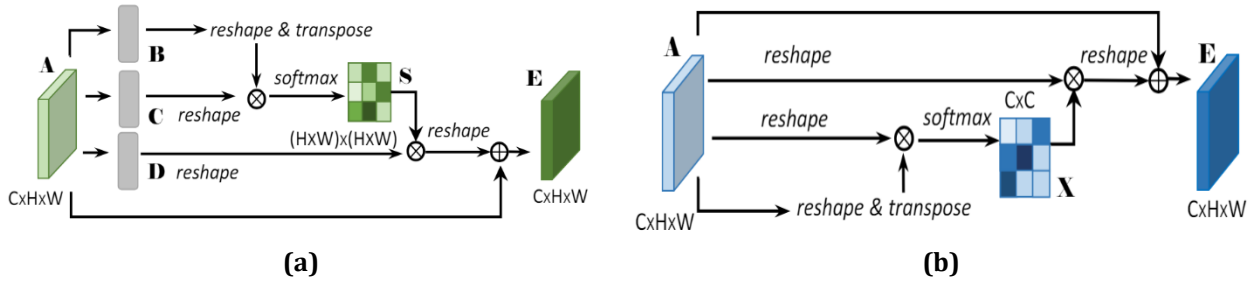


Fig. 7 *Structure of the DANet. (a) Position attention module; (b) Channel attention module*

DANet leverages dual attention mechanisms—position attention and channel attention—to capture rich contextual dependencies. By integrating local features with global spatial and channel information, DANet significantly improves feature representation. This enhanced feature extraction can improve performance in various tasks, including scene segmentation and object detection. Incorporating DANet into your model can help selectively aggregate essential features, thus improving overall accuracy and robustness. Through the ECA module, a cross-channel local interaction method is introduced to help avoid dimensionality reduction in a way that effectively prevents the negative effects on the learning performance of channel attention. The ECA module includes a one-dimensional convolution guided by nonlinear adaptation, which captures the information generated by cross-channel localities by considering each channel and its neighboring channels. It is a highly lightweight, plug-and-play block with minimal parameters that delivers substantial performance improvements [26]. Figure 8 below illustrates the architecture of the ECA.

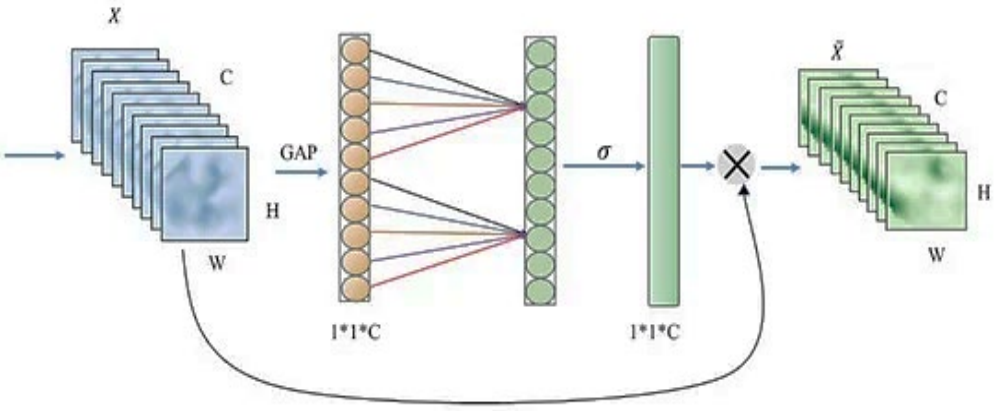


Fig. 8 *ECA Structure*

5.1 Experimental Results of Comparison of Object Detection Network

The selection of the model for object detection for pedestrian detection involved applying numerous widely known models to both the training and testing datasets. The comparison used performance parameters like F1 Score, Recall, Precision, and mAP@0.5. Ultimately, YOLOv8 was chosen as the object detection model for the subsequent experiments of our study. The comparison of the performance evaluation metrics of each model on

the previously mentioned aggregated dataset is shown in Table 1, with the experimental results shown in Table 2.

Table 2 *Comparison of object detection algorithms.*

Methodology	Recall	Precision	F1	FPS	mAP@ 0.5	mAP@ 0.5:0.95
EfficientNetB7	97.12%	94.16%	0.95	45	93.41%	81.50%
MobileNetV2	80.40%	88.03%	0.84	42	87.03%	71.60%
EfficientDet	90.98%	85.66%	0.89	36	93.91%	82.20%
Faster R-CNN	87.17%	90.00%	0.87	27	96.04%	78.90%
YOLOv5s	80.04%	90.26%	0.85	42	91.82%	86.60%
YOLOv7	88.70%	92.50%	0.90	62	96.90%	85.10%
YOLOv8	91.64%	97.80%	0.94	60	98.21%	87.80%

Table 2 shows that YOLOv8 outperformed the other detection algorithms examined overall. It came in first place for recall, mAP@0.5:0:0.95, and detection speed and second for precision, F1 score, and mAP@0.5. For instance, YOLOv8 had a recall rate that was 3.3% greater than YOLOv7's. They are essentially better when comparing the remaining signs to the other target detection techniques. Ultimately, YOLOv8 was employed as the experiment's target identification technique.

5.2 Contrast Experiment Results of Introducing Attention Mechanism

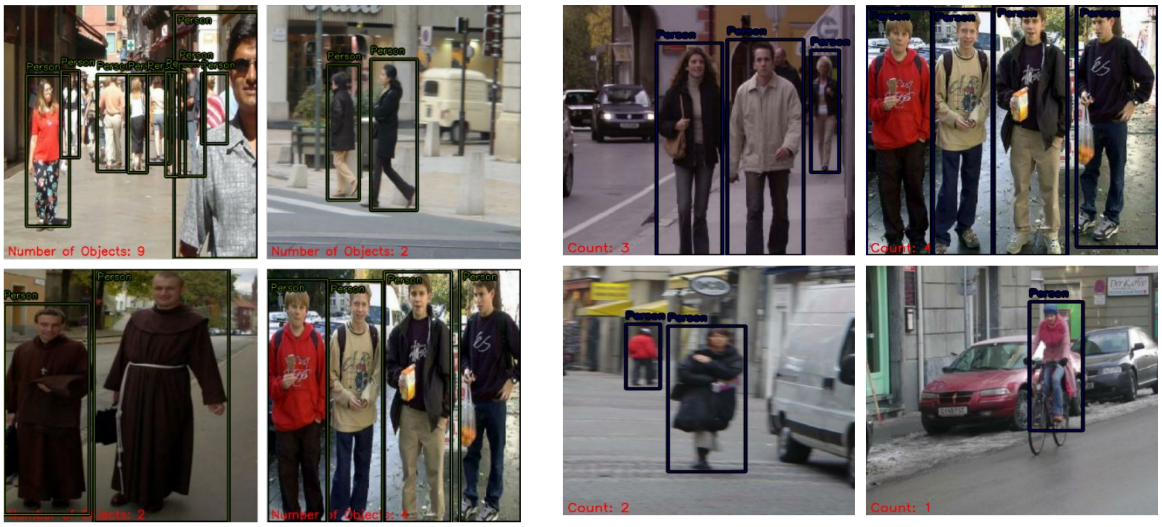
The efficiency and effectiveness of the enhanced algorithm, CBAM, was deployed as the attention mechanism, the YOLOv8 object detection algorithm was included, and experiments were conducted on the combined dataset. Table 3 below shows the results of the experiments, with recall rates mAP@0.5 and mAP@0.5:0.95 serving as indicators.

Table 3 *Comparative experiments for attention mechanism*

DUANet	CBAM	ECA	Precision	Recall	F1	mAP@0.5	mAP@0.5:0.95
×	×	×	97.80%	91.64%	0.94	98.21%	87.80%
✓	×	×	91.31%	89.43%	0,90	93.48%	85.30%
×	✓	×	98,24%	95.97%	0.97	98.09%	89.10%
×	×	✓	96.35%	90.33%	0.93	98.41%	86.30%

Table 3 shows that the DUA-YOLOV8 algorithm's accuracy rate dropped by 4.73%, the recall rate dropped by 2.21%, the mAP dropped by 2.50%, and compared to the original YOLOv8 method. The ECA-YOLOv7 algorithm's accuracy rate fell by 0.80%, the recall rate rose by 0.69%, and mAP also dropped. Table 3 displays the findings, which indicate that the DUA-YOLOv8 and ECA-YOLOv8 algorithms produced less of an impact than the original YOLOv8 and raised the computational burden and model parameters. The CBAM-YOLOV7 algorithm improved by 0.20% in accuracy, 1.30% in the recall, and 1.30% in mAP@0.5:0.95 compared to the original YOLOv8 algorithm. Based on the experimental results obtained in this work, it can be concluded that the algorithm proposed in this work outperformed the original method, and the algorithm used the DUA and ECA modules by comparing and evaluating the experimental results. The CBAM module outperformed the SE-YOLOv8 and ECA-YOLOv8 by introducing a spatial attention module, enhancing the channel attention module, and assessing the data in two dimensions to ascertain the order from the channel to the area. The CBAM-YOLOv8 algorithm's detection impact on the INRIA Dataset is displayed in Figure 9.

J. of Soft Computing and Data Mining Vol. 5 No. 2 (2024) p. 175-187



(a) The CBAM-YOLOv8 algorithm's **(b) The YOLOv8 algorithm (no tricks)**

Fig. 9 *YOLOv8 prediction result graph*

4.2 Contrast Experiment Results of Introducing Attention Mechanism

Figure 9 (b) shows the prediction of the YOLOv8 algorithm without additional training skills. This research has experimented with several training strategies in the ablation experiment, utilizing various methods to process the model, all based on the original YOLOv8 algorithm, including Mixup and mosaic processing. The results confirmed the experimental findings acquired with the previously mentioned treatments through the studies. Table 4 displays the outcomes of the experiment.

Table 4 *Set of training and assessment capabilities*

Group	Mosaic	MixUp	Precision	Recall	F1	mAP@0.5	mAP@0.5:0.95
1	×	×	97.80%	91.64%	0.94	98.21%	87.80%
2	✓	×	96.55%	94.64%	0.95	97.26%	88.40%
3	×	✓	95.45%	93.75%	0.94	97.65%	86.90%
4	✓	✓	96.80%	93.64%	0,95	97.57%	89.50%

In the YOLOv8 ablation experiment, each group of experiments corresponds to a set of training and assessment capabilities. In Table 4, '✓' indicates the use of the training technique, while '×' denotes its inapplicability. Overall, combining Mosaic and Mixup simultaneously proved more effective than using either method alone or not using either.

6. Conclusion

This study integrated three CBAM modules into the YOLOv8 algorithm to enhance structural optimization. An enhanced YOLOv8 algorithm integrating an attention mechanism has been proposed. In addition, this study introduced DUA-YOLOv7 and ECA-YOLOv7 for comparative experiments. Additionally, a large-scale dataset for pedestrian detection was collected comprising 1520 images from the INRIA and ETH Person datasets. These datasets are available for researchers, offering valuable data support for visual research in AVs. The accuracy rate, memory rate, and mAP all improved; FLOPS increased only by 0.02 G without increasing the computational pressure. The algorithm presented in this paper has shown effective results in detecting and accurately counting the number of people. The CBAM-YOLOv8 algorithm was introduced to improve the accuracy of object detection. Furthermore, the study investigated the advantages of incorporating Mosaic and Mixup data augmentation techniques. Future research will focus on optimizing the algorithm's network structure and integrating it into the hardware environment for field applications. Pedestrian detection is a critical component in the perception system of AVs. Accurate and reliable pedestrian detection can significantly enhance AVs' safety and operational effectiveness. This research introduces an enhanced object detection model based on YOLOv8, integrated with the Convolutional Block Attention Module (CBAM). It is optimized explicitly for pedestrian detection using a combined dataset from the INRIA and ETH datasets. This research contributes to the advancement of AV technology by

addressing one of the most critical challenges in autonomous driving—reliable and accurate pedestrian detection. The improved model sets a new benchmark for future research and development in this domain. In addition, enhanced pedestrian detection directly translates to improved public safety. By reducing the risk of collisions with pedestrians, the model promotes wider acceptance and trust in AV technology among the public. It's important to acknowledge that this study has limitations that should not be overlooked. Specifically, external environmental disruptions were not considered in our analysis. Furthermore, the dataset predominantly includes photos taken in favorable lighting conditions and typical weather, which may not fully reflect real-world scenarios, especially regarding issues like overlapping objects. To address these shortcomings, additional pedestrian datasets that capture a broader spectrum of environmental conditions in future research endeavors. Second, the detection findings contain both missing and incorrect detection data. To improve the confidence level of bug detection, this work used a strategy of increasing the number of negative training samples. Additionally, to address missed detection, we used a strategy that involved filtering out samples with high loss values from each training iteration and adding them to the training set for the next iteration. This allowed the detection model to focus more on frequently missed samples.

Acknowledgment

This research work received support and funding from Al-Maarif University College.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the paper's publication.

Author Contribution

The authors confirm their contribution to the paper as follows: ***study conception, design, and draft manuscript preparation; reviewed the results and approved the final version of the manuscript:*** *Mohammed Ahmed Jubair, Mohammed Mansoor Nafea;* ***data collection, analysis, and interpretation of results, and draft manuscript preparation****.*

References

[1] Faisal, A., Kamruzzaman, M., Yigitcanlar, T., & Currie, G. (2019). Understanding autonomous vehicles. Journal of transport and land use, 12(1), 45-72.
[2] Wei, P., Cagle, L., Reza, T., Ball, J., & Gafford, J. (2018). LiDAR and camera detection fusion in a real-time industrial multi-sensor collision avoidance system. Electronics, 7(6), 84.
[3] Galvao, L. G., Abbod, M., Kalganova, T., Palade, V., & Huda, M. N. (2021). Pedestrian and vehicle detection in autonomous vehicle perception systems—A review. Sensors, 21(21), 7267.
[4] Mostafa, S. A., Ravi, S., Zebari, D. A., Zebari, N. A., Mohammed, M. A., Nedoma, J., ... & Ding, W. (2024). A YOLO-based deep learning model for Real-Time face mask detection via drone surveillance in public spaces. *Information Sciences*, 120865.
[5] Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence, 39(6), 1137-1149.
[6] Huang, R., Pedoeem, J., & Chen, C. (2018, December). YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers. In 2018 IEEE international conference on big data (big data) (pp. 2503-2510). IEEE.
[7] Ren, H., Jing, F., & Li, S. (2024). DCW-YOLO: Road Object Detection Algorithms for Autonomous Driving. IEEE Access.
[8] Nafea, M. M., Tan, S. Y., Jubair, M. A., & Abd Mustafa, T. (2022). A Review of Lightweight Object Detection Algorithms for Mobile Augmented Reality. International Journal of Advanced Computer Science and Applications, 13(11).
[9] Galdran, A., Carneiro, G., & González Ballester, M. A. (2021). Balanced mixup for highly imbalanced medical image classification. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24 (pp. 323-333). Springer International Publishing.
[10] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.
[11] Hoiem, D., Divvala, S. K., & Hays, J. H. (2009). Pascal VOC 2008 challenge. World Literature Today, 24(1), 1-4.
[12] Shao, Y., Yang, Z., Li, Z., & Li, J. (2024). Aero-YOLO: An Efficient Vehicle and Pedestrian Detection Algorithm Based on Unmanned Aerial Imagery. Electronics, 13(7), 1190.
[13] Gong, L., Wang, Y., Huang, X., Liang, J., & Fan, Y. (2024). An improved YOLO algorithm with multi-sensing for pedestrian detection. Signal, Image and Video Processing, 1-14.

[14] Chen, Y., Ye, J., & Wan, X. (2023). TF-YOLO: A Transformer–Fusion-Based YOLO Detector for Multimodal Pedestrian Detection in Autonomous Driving Scenes. World Electric Vehicle Journal, 14(12), 352.
[15] Luo, X., Zhu, H., & Zhang, Z. (2024). IR-YOLO: Real-Time Infrared Vehicle and Pedestrian Detection. Computers, Materials & Continua, 78(2).
[16] Nan, X., Lu, W., Chongliu, J., Yuemou, J., & Xiaoxia, M. (2023). Simulation of Occluded Pedestrian Detection Based on Improved YOLO. Journal of System Simulation, 35(2), 286.
[17] Reis, D., Kupec, J., Hong, J., & Daoudi, A. (2023). Real-time flying object detection with YOLOv8. arXiv preprint arXiv:2305.09972.
[18] Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7464-7475).
[19] Li, Y., Fan, Q., Huang, H., Han, Z., & Gu, Q. (2023). A modified YOLOv8 detection network for UAV aerial image recognition. Drones, 7(5), 304.
[20] Niu, Z., Zhong, G., & Yu, H. (2021). A review of the attention mechanism of deep learning. Neurocomputing, 452, 48-62.
[21] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 37(9), 1904-1916.
[22] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 3-19).
[23] Muhammad, M. B., & Yeasin, M. (2020, July). Eigen-cam: Class activation map using principal components. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.
[24] Ying, X., Wang, Y., Wang, L., Sheng, W., An, W., & Guo, Y. (2020). A stereo attention module for stereo image super-resolution. IEEE Signal Processing Letters, 27, 496-500.
[25] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3146-3154).
[26] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11534-11542).