Research on Vehicle and Pedestrian Detection Algorithms Based on an Improved YOLOv8

Yiming Sun
School of Integrated Circuits, Nanjing University of Information Science & Technology, Suzhou, Jiangsu, China

Abstract: Vehicle and pedestrian detection in complex road scenarios represents a critical challenge for autonomous driving environmental perception. Addressing issues such as reduced model reliability caused by interference from lighting variations, occlusions, and adverse weather conditions.
This paper addresses this critical issue by employing the YOLOv8 network as the foundational detection architecture, incorporating an attention mechanism to enhance the model's ability to extract, learn, and represent key object features.
Based on this dataset, the experimental section presents two comparison approaches: one directly employs the original YOLOv8 model, while the second integrates the SE (Squeeze-and-Excitation) attention mechanism into the YOLOv8 model.
Experimental results demonstrate that incorporating the SE attention mechanism achieves improvements in key metrics such as recall and mAP50-95, at the modest cost of increasing model parameters by 0.1 million and reducing inference speed by 3.4 FPS. Consequently, the proposed model exhibits superior overall performance and application potential in scenarios demanding high detection accuracy and real-time capabilities.

Keyword: YOLOV8; SE; Vehicle and Pedestrian Detection

1. Introduction

1.1 Research Background and Significance

Real-world road environments present numerous uncertainties, including abrupt changes in lighting conditions, adverse weather such as rain, snow, and fog, high-density traffic flows, and occlusion between objects. Such complex scenarios significantly increase the recognition difficulty for detection algorithms, severely constraining their accuracy and reliability. Consequently, effectively enhancing the robustness and adaptability of deep learning-based object detection algorithms in complex environments has become a core and pressing research issue in autonomous driving perception.

Among deep learning-based object detection algorithms, the YOLO series models exhibit notable advantages in detection speed and balanced accuracy due to their single-stage detection architecture. However, when confronted with the aforementioned complex road environments, these models still suffer from issues such as insufficient feature extraction and limited ability to distinguish between objects, resulting in overall robustness that falls short of practical application requirements. Recent studies indicate that incorporating attention mechanisms can effectively enhance a model's ability to learn and focus on key feature regions of objects, thereby improving detection performance in complex settings. This offers a crucial technical pathway for addressing the limitations faced by YOLO models.

Investigating how to effectively integrate attention mechanisms with YOLO series networks to enhance detection performance under complex outdoor conditions holds significant theoretical value and practical relevance. Accordingly, this paper adopts the YOLOv8 network as its foundational architecture. By embedding an attention mechanism to optimise its feature extraction process, we aim to enhance detection accuracy and efficiency for pedestrians and vehicles in complex scenarios. Concurrently, we will construct an image dataset incorporating multiple complex environmental factors-such as adverse weather, object occlusion, and lighting variations-to comprehensively validate the effectiveness of the proposed method.

1.2 Current State of Research

The continuous evolution of deep learning

technology has revolutionised the field of object detection. Deep learning-based object detection algorithms possess the capability to automatically learn and extract image features, enabling effective detection and recognition of diverse object categories. Currently, deep learning-based object detection algorithms are primarily categorised into two-stage approaches (e.g., the R-CNN series[1]) and single-stage approaches (e.g., YOLO[2] and SSD[3]). Among these, the YOLO series has emerged as a mainstream solution due to its favourable balance between speed and accuracy.

With continuous advancements in the field of deep learning, the network architectures underpinning object detection have undergone ongoing optimisation. From the early LeNet[4] and AlexNet[5] , to subsequent developments such as VGGNet[6] and ResNet[7] , deeper network architectures have significantly enhanced models' feature representation capabilities, thereby substantially improving detection accuracy. Feature extraction capability, as a core component of deep neural networks, exerts a decisive influence on detection performance. Concurrently, to address challenges such as target scale variations and morphological diversity, a series of technical solutions-including Feature Pyramid Networks (FPN) and Default Boxes mechanisms-have been extensively introduced and applied within contemporary detection frameworks.

1.3 Principal Research Contributions

The principal research endeavours of this paper are as follows:

(1) Construction of a complex scene evaluation subset based on the WiderPerson dataset;

(2) Integrating the SE attention mechanism module into the backbone and neck networks of the YOLOv8 architecture;

(3) Validating the proposed method's improvements in both accuracy and efficiency through comparative experiments.

2. Attention Mechanism and Enhanced YOLOv8 Model

2.1 SE Attention Mechanism

The SE (Squeeze-and-Excitation) module constitutes a lightweight channel attention mechanism[8] . It adaptively calibrates channel feature responses through a "squeeze-excitation-recalibration" operation. Specifically:

Squeeze: Compresses spatial features into channel descriptors via global average pooling.

Excitation: Learns weights for each channel through two fully connected layers, capturing nonlinear relationships between channels.

Scale: Multiplies the learned weights with the original feature map to complete feature re-scaling.

The SE module effectively enhances useful features while suppressing redundant ones, with minimal computational overhead, making it highly suitable for embedding into lightweight detection models.

2.2 Improved YOLOv8 Model

The structure of the improved YOLOv8 model presented herein is illustrated in Figure 1. To enhance the model's ability to extract key features of multi-scale objects, we embed SE modules after the terminal C2f module of the backbone and after key layers of the neck network (specific locations indicated in Figure 1). This design enables the network to prioritise the selection of more discriminative channel information before feature fusion, thereby enhancing the subsequent detection head's perception accuracy for objects, particularly small and occluded targets.

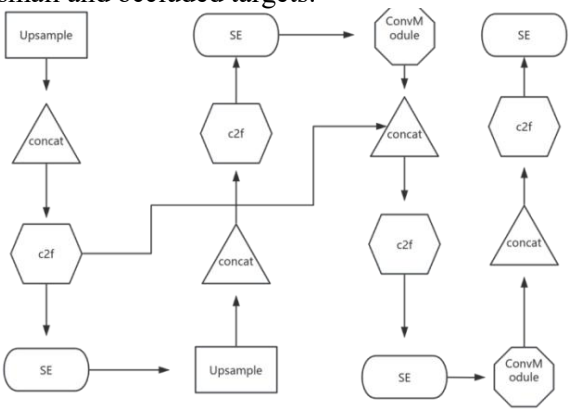


Figure 1. YOLOv8's Improved Model Architecture

3. YOLOv8 Network

3.1 YOLOv8 Network Analysis

The YOLOv8 network proposed herein [9] adopts a fully convolutional architecture with an encoder-decoder structure. As illustrated in Figure 2, it comprises three major modules: the Backbone feature extraction network, the Neck feature fusion network, and the Head detection prediction network.

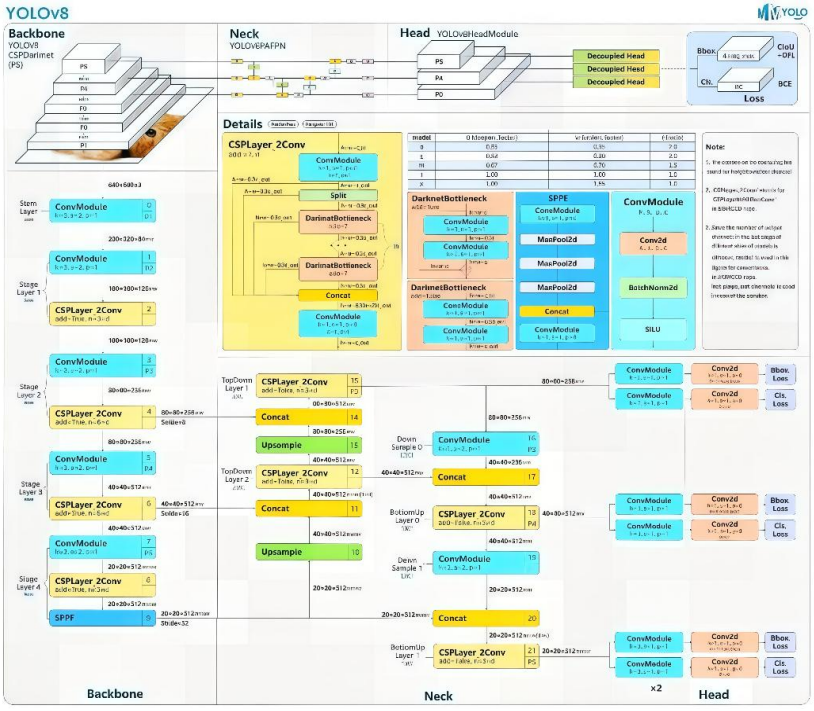


Figure 2. OLOv8 Network Fully Convolutional Architecture

3.1.1 Backbone: efficient feature extraction trunk network

Employing CSPDarknet as the foundational skeleton, the core is replaced with the proprietary C2f module. Combined with the SPPF (Spatial Pyramid Pooling-Fast) module, this constructs a lightweight yet expressive feature extraction chain.

C2f Module: Employing a "Convolution Splitting-Parallel Convolution-Feature Fusion" design, it maintains lightweight network parameters while increasing gradient flow branching. This captures richer image detail features and high-level semantic features, effectively mitigating convergence degradation issues during deep network training.

SPPF Module: Aggregates feature maps through multi-scale pooling operations, expanding the feature receptive field while reducing computational load. This enhances feature extraction capabilities for objects of varying sizes.

3.1.2 Neck: multi-scale feature fusion neck network

Employing a hybrid PAN-FPN (Path Aggregation Network-Feature Pyramid Network) architecture, this component prioritises optimising feature propagation and fusion efficiency:

Removes redundant 1×1 convolutional layers prior to upsampling, feeding features from different backbone layers (high-resolution shallow features, low-resolution deep features) directly into the upsampling process to minimise feature loss;

Unified adoption of the C2f module enhances cross-scale feature fusion capabilities, particularly improving feature capture for small and distant targets.

3.1.3 Head: decoupled detection-prediction head network

The innovative Anchor-Free + decoupled head architecture substantially improves detection efficiency and accuracy:

Decoupled Detection Head: Separates "object classification" and "bounding box regression" into two independent convolutional branches. The classification branch focuses on learning semantic features, while the regression branch optimises bounding box coordinates. Parallel computation accelerates model convergence and reduces task interference.

Anchor-Free Design: Eliminates reliance on predefined anchor boxes by directly predicting target centre coordinates, aspect ratio, and confidence scores. This approach reduces hyperparameter dependency and computational redundancy associated with anchor box design while improving adaptability to irregularly shaped objects.

4. Experiments and Analysis

4.1 Computational Methods and

Experimental Setup

4.1.1 Metric calculation methods

Object detection model performance is evaluated using the following metrics:

(1) Precision: Measures the accuracy of positive samples in detection results.

$$\text{Precision} = TP / (TP + FP) \quad (1)$$

(2) Recall: Measures the model's ability to cover true positive samples.

$$\text{Recall} = TP / (TP + FN) \quad (2)$$

(3) Mean Average Precision (mAP)

$$mAp=\frac{1}{k}\sum_{i=1}^{k} AP\_i \quad (3)$$

(4) Loss function

box_loss: Localisation error between predicted bounding boxes and ground truth boxes, calculated using CIoU.

cls_loss: Target classification error.

dfl_loss: Bounding box distribution focus loss, employed to enhance regression accuracy.

(5) The F1 curve illustrates the model's prediction performance across different confidence thresholds.

$$F1=2\cdot\frac{Precision\cdot Recall}{Precision+Recall} \quad (4)$$

By comprehensively evaluating these five metrics, we can assess the specific impact of the attention mechanism on YOLOv8's object detection performance across multiple dimensions. Precise rate and recall respectively reflect model performance from the core dimensions of "prediction accuracy" and "object completeness"-precise rate indicates the reliability of the model's prediction of positive samples, while recall measures the model's coverage capability for true targets; Meanwhile, mAP@.5 and mAP@.5: 95 provide quantitative metrics for overall model performance from the perspective of "overlap standard adaptability": mAP@.5 focuses on comprehensive performance under relaxed overlap requirements, while mAP@.5 95 rigorously tests model robustness through evaluation across multiple gradient IoU thresholds.

4.1.2 Experimental configuration

The experimental environment parameters for the YOLOv8 model are shown in Table 1:

Table 1. Experimental Environment Parameters

Experimental Environment	Parameters
PyTorch Version	Windows 11
Python version	3.8.10
CUDA Version	11.1
GPU	Nvidia GeForce RTX 3070 Laptop GPU
CPU	AMD Ryzen 7 5800H with Graphics
Memory	16GB
Graphics Memory	16GB

4.1.3 Dataset introduction

The Widerperson dataset was employed, with a subset comprising 9,000 samples extracted from the original dataset. This subset not only encompasses an exceptionally diverse range of scene types but also features a notably high density of human targets within the samples.

To ensure scientific, standardised, and credible model evaluation, the subset is further divided into training, validation, and test sets.

Training set: Used to train the model, constituting the majority of the data. (Number of annotated boxes: 220,394; Number of images: 7,200)

Validation set: Used to test model performance during training, assist parameter tuning, and prevent overfitting. (Annotation boxes: 26,890; Images included: 900)

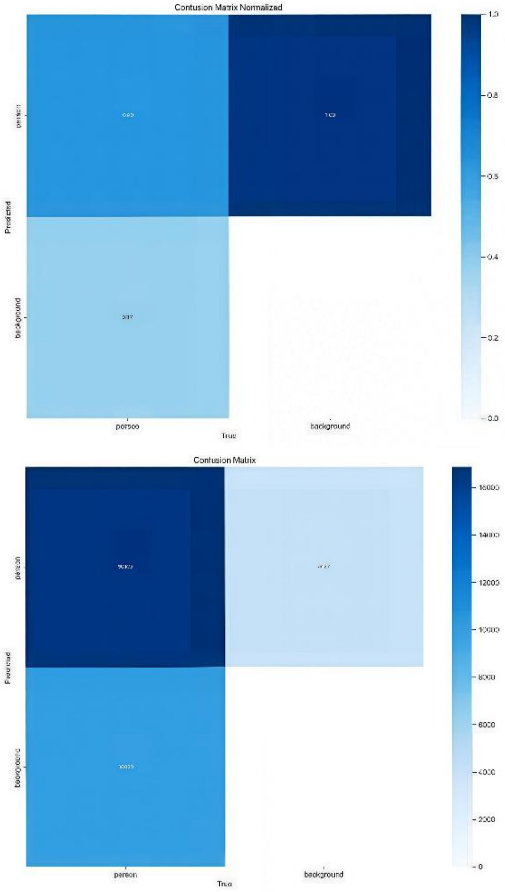
Test set: Used for the final evaluation of model performance. It must not be used for training and serves to test the model's generalisation ability. (Number of annotated boxes: 26,191; Number of images included: 900)

As a classic open-source dataset in the field of pedestrian detection, Widerperson's design limitations serve both the standards of academic research and impose constraints on practical applications, focusing on three main aspects. Firstly, there are usage permissions and scenario restrictions. The dataset explicitly states that it is only for non-commercial academic research and prohibits use in commercial product development or profit-oriented projects. For example, if a company wishes to train a commercial security detection model based on its annotated complex scene pedestrian data (such as crowded streets and occluded pedestrian samples), separate authorisation is required. While this limitation protects the academic nature of the dataset, it also raises the threshold for commercial implementation. Secondly, there are limitations in data scope and annotation. Widerperson samples mainly come from publicly available scene images and do not cover specialized domain data (such as pedestrians in military areas or medical scenarios), and the annotations only include pedestrian bounding boxes, lacking in-depth information such as

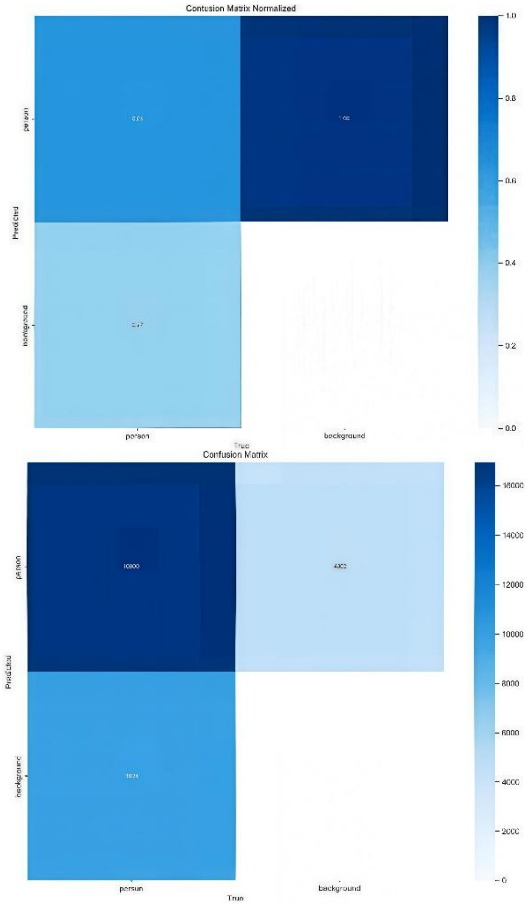
pedestrian behaviours and postures. This means that researchers who need to conduct specialised studies like "abnormal pedestrian behaviour recognition" must supplement additional data, increasing research costs. Moreover, the restriction on the update frequency of this dataset is also contentious. Its latest version is several years old and has not promptly included new scenario samples such as "interactions between pedestrians and autonomous vehicles" following the widespread adoption of new energy vehicles, reducing the adaptability of models trained on this dataset to emerging real-world scenarios. However, this limitation also compels researchers to explore data augmentation techniques, which to some extent promotes the development of related technologies.

4.2 Experimental Analysis and Detection Results

As illustrated in Figures 3 (a) and (b), the introduction of the SE mechanism yields more pronounced values along the main diagonal of the model's normalised confusion matrix, indicating enhanced classification accuracy.



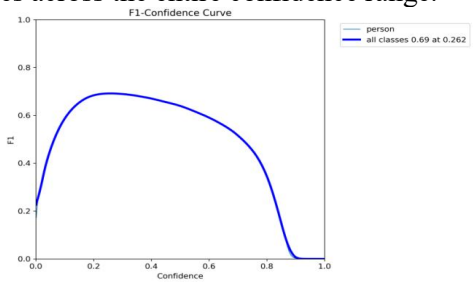
(a) YoloV8 Model Confusion Matrix



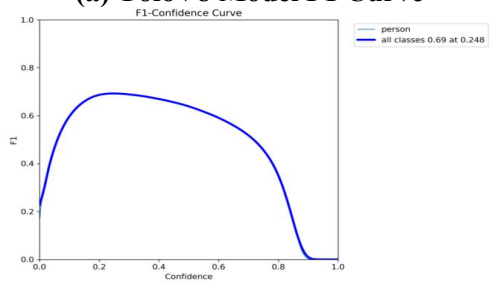
(b) YOLOV8 + Self-Attention Mechanism Model Confusion Matrix

Figures 3. (a) and (b) Model Confusion Matrix

Comparing the F1 curves in Figures 4 (a) and (b), the improved model demonstrates enhanced F1 scores across the entire confidence range.



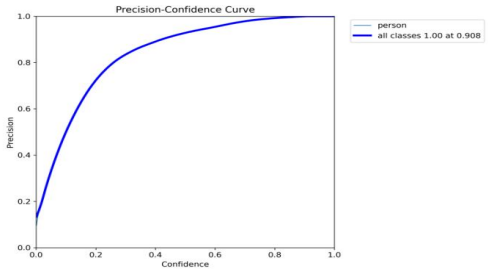
(a) YoloV8 Model F1 Curve



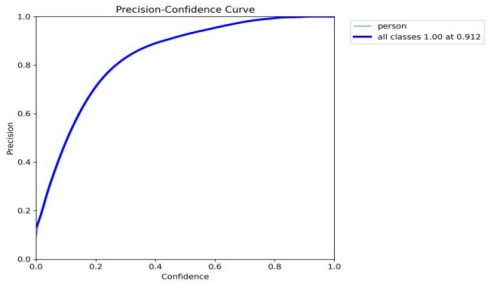
(b)YoloV8 + Self-Attention Mechanism Model F1 curve

Figure 4. (a) and (b) Model F1 Curve

Comparing the accuracy-confidence curves in Figures 5 (a) and (b), the improved model achieves full-category precise detection at higher confidence thresholds, demonstrating superior performance.



(a)YoloV8 Model Precision-Confidence Curve

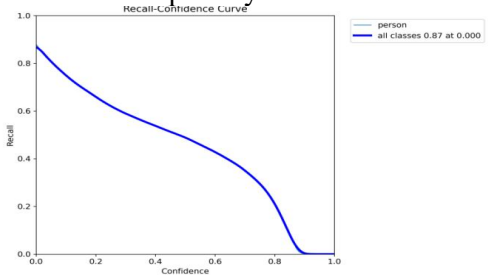


(b) YoloV8 + Self-Attention Mechanism Model Precision-Confidence Curve

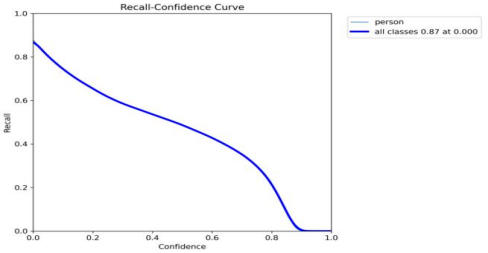
Figures 5. (a) and (b) *Model* Precision-Confidence Curve

Comparing the recall-confidence curves, PR curves, and labels_correlogram.jpg between the YOLOv8 model and the YOLOv8+SE attention mechanism model in Figure 6 (a) and (b), Figure 7 (a) and (b), Figure 8(a) and (b), and Figure 9 (a) and (b), it is evident that the improved model exhibits no significant changes in these curves, maintaining its original performance. labels_correlogram.jpg reveal that the improved model exhibits no significant deviation in these curves, maintaining its original performance.

Comparing the loss function curves in Figures 10(a) and (b) reveals that the final values of evaluation metrics (precision, recall, mAP, etc.) for the improved model are higher, and the rate at which they plateau remains virtually unchanged. This indicates superior performance in both "classification/detection accuracy" and "generalisation capability".

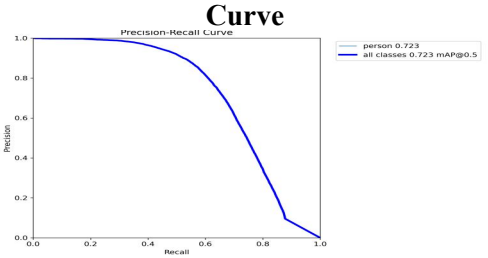


(a) YoloV8 Model Recall-Confidence Curve

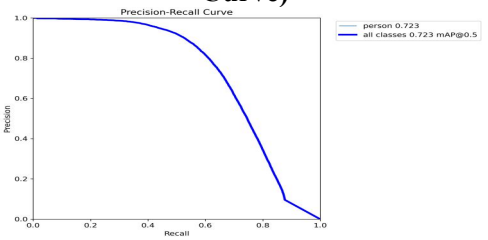


(b) YoloV8 + Self-Attention Mechanism Model Recall-Confidence Curve

Figure 6. (a) and (b) Model Recall-Confidence Curve

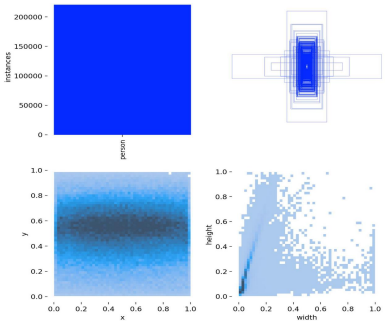


(a) YoloV8 Model PR Curve (Precision-Recall Curve)

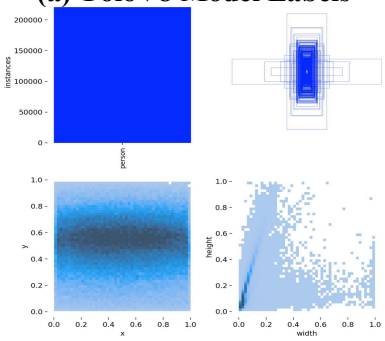


(b) YoloV8 + Self-Attention Mechanism Model Precision-Recall Curve

Figure 7. (a) and (b) Model Precision-Recall Curve

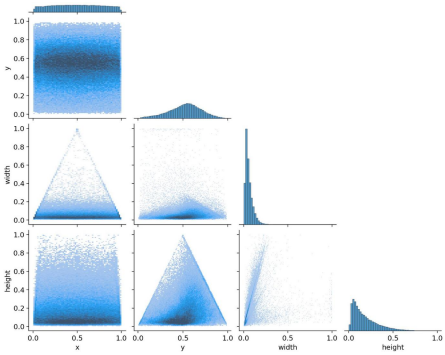


(a) YoloV8 Model Labels

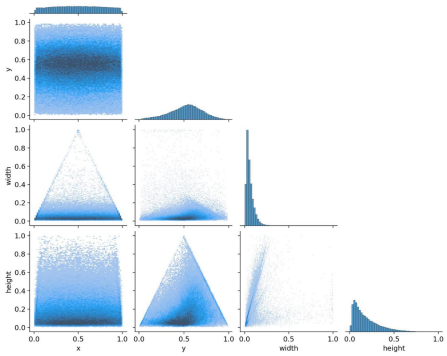


(b) YoloV8 + Self-Attention Mechanism Model Labels

Figure 8. (a) and (b) Model Labels

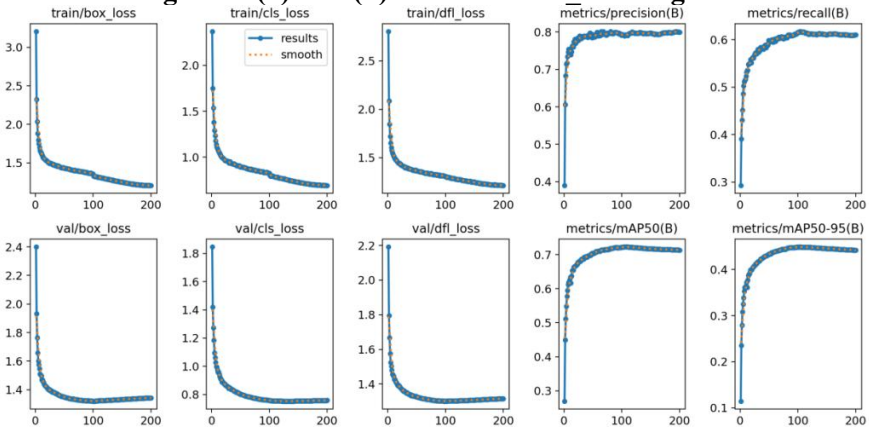


(a) YoloV8 Model labels_correlogram -- Illustrating the relationship between the centre point's x and y coordinates and the height and width of the bounding box

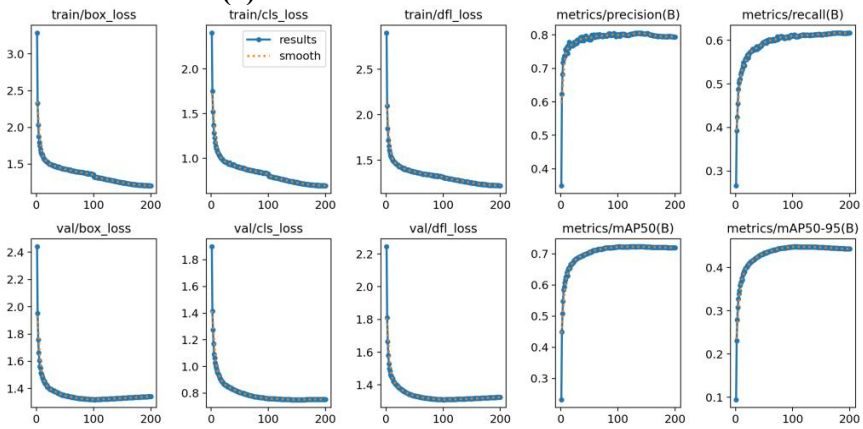


(b) YoloV8 + Self-Attention Mechanism Model labels_correlogram-Illustrating the relationship between the centre point's x and y coordinates and the height and width of the bounding box

Figure 9. (a) and (b) Model Labels_correlogram



(a) YoloV8 Model Loss Functions



(b)YoloV8 + Self-Attention Mechanism Model Loss Functions

Figures 10(a) and (b) Model Loss Functions

Table 2. Comparison of Evaluation Results from Multiple Models

Model Name	Params (M)	Size (MB)	Inference Time(ms)	FPS	Precision	Recall	mAP50	mAP50-95
runs/ train/ yolov8s4/ weights/ best.pt	11.1	21.48	8.27	120.86	0.785	0.626	0.731	0.448
runs/ train/ yolov8s-attention-SE2/ weights/ best.pt	11.2	21.59	8.52	117.44	0.786	0.629	0.731	0.449

4.3 Model Evaluation: and Experimental Results Analysis

In this experiment targeting object detection tasks, we compared the performance of the YOLOv8S model against the YOLOv8S-Attention-SE2 model incorporating an attention mechanism (SE2). Analysis was

conducted across multiple dimensions including model parameter count, size, inference speed, and accuracy, as shown in Table 2.

4.4 Practical Detection Performance

(a) YOLOV8 Model

(b) Model with SE Mechanism Injected

Figure 11. (a) and (b) Practical Detection Performance

Comparing Figures 11 (a) and (b), the model incorporating the SE mechanism demonstrates superior background-person separation, achieving higher recall (red box in Figure (b)) and improved precision (red arrow in Figure (b)).

5. Comparison of Fundamental Model Attributes

Parameter Count and Model Size: The YOLOv8S model has 11.1 million parameters and a size of 21.48 MB; the YOLOv8S-Attention-SE2 model exhibits a slight increase in parameters to 11.2 million, with its size also rising marginally to 21.59 MB. This indicates that the introduction of the SE2 attention mechanism results in a slight increase in model complexity. However, the models remain relatively lightweight overall and should not impose significant storage pressure on deployment environments.

Inference speed: Measured in FPS (frames per second), YOLOv8s achieves 120.86 FPS, while YOLOv8s-Attention-SE2 achieves 117.44 FPS. Evidently, incorporating the attention mechanism results in a certain degree of reduction in inference speed. This is because the attention mechanism increases the computational load of the model, causing the single-frame inference time to rise from 8.27ms to 8.52ms. However, from a practical application perspective, the inference speeds of both models can meet the basic requirements for real-time detection, and the speed difference is acceptable in many scenarios.

6. Detection Accuracy Comparison

Precision and Recall: YOLOv8s achieved a precision of 0.785 and a recall of 0.626; YOLOv8s-Attention-SE2 improved precision to 0.786 and recall to 0.629. These modest gains in precision and recall indicate that incorporating the attention mechanism improves the model's ability to reduce false positives (enhancing precision) and false negatives (enhancing recall), thereby strengthening object recognition accuracy.

mAP metrics: mAP50 (mean average precision at IoU threshold 0.5) remained consistent at 0.731 for both YOLOv8s and YOLOv8s-Attention-SE2. mAP50–95 (mean average precision across IoU thresholds from 0.5 to 0.95): YOLOv8s achieved 0.448, while YOLOv8s-Attention-SE2 improved to 0.449. mAP50–95 more comprehensively reflects a model's overall detection performance across varying IoU thresholds. This improvement indicates that incorporating the attention mechanism optimises the model's object localisation and classification capabilities under more stringent evaluation criteria.

7. Conclusions

The YOLOv8s-Attention-SE2 model, incorporating the SE2 attention mechanism, achieves modest improvements across multiple key detection metrics (precision, recall, mAP50–95) despite incurring a slight increase in model parameters, size, and inference speed. The improvement in mAP50-95 indicates that the SE mechanism enhances the discriminative power of learned features by amplifying useful channels and suppressing redundant ones, thereby achieving more precise localisation under stricter IoU thresholds. In object detection scenarios demanding high detection accuracy while maintaining real-time inference speeds, the yolov8s-attention-SE2 model offers greater application advantages over the original yolov8s model.

References

[1] Tian Xuan, Wang Liang, Ding Qi. A Review of Deep Learning-Based Image Semantic Segmentation Methods [J]. Journal of

Software, 2019, 30(02): 440-468.
[2] Shao, Y. H., Zhang, D., Chu, H. Y., et al. Review of Deep Learning-Based YOLO Object Detection [J]. Transactions of Electronics and Information Technology, 2022, 44(10): 3697-3708.
[3] Zhao Yongqiang, Rao Yuan, Dong Shipeng, et al. Review of Deep Learning Object Detection Methods [J]. Chinese Journal of Image and Graphics, 2020, 25(04): 629-654.
[4] Chen Xianchang. Research on Deep Learning Algorithms and Applications Based on Convolutional Neural Networks [D]. Zhejiang Gongshang University, 2014.
[5] Lu Hongtao, Zhang Qinchuan. A Review of Deep Convolutional Neural Networks in Computer Vision Applications [J]. Data Acquisition and Processing, 2016, 31(01): 1-17.
[6] Li Bingzhen, Liu Ke, Gu Jiaojiao, et al. Research Review on Convolutional Neural Networks [J]. Computer Era, 2021, (04): 8-12+17.
[7] Zhang Ke, Feng Xiaohan, Guo Yurong, et al. Review of Deep Convolutional Neural Network Models for Image Classification [J]. Chinese Journal of Image and Graphics, 2021, 26(10): 2305-2325.
[8]Ren Huan, Wang Xuguang. Review of Attention Mechanisms [J]. Computer Applications, 2021, 41(S1): 1-6.
[9] Han Qiang. Research on an Improved YOLOv8 Algorithm for Small Object Detection [D]. Jilin University, 2023.

