International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 6549-6563

https://doi.org/10.70517/ijhsa464555

Automatic Driving Target Detection Based on Transfer Learning and YOLOv5-BEs Algorithm

Jiahao Xue[1,*]
[1] School of Electronic Engineering, Heilongjiang University, Harbin, Heilongjiang, 150080, China
Corresponding authors: (e-mail: 20226171@s.hlju.edu.cn).

Abstract To address the limitations of current road target detection algorithms, including insufficient small target detection capability, slow speed, frequent misdetection and omission, and long training time, this paper proposes a high-precision target detection model integrating transfer learning and improved YOLOv5 algorithm to satisfy the high requirements of detection speed and accuracy in autonomous driving scenarios. The Efficient Channel Attention (ECA) attention mechanism is first added to the model in order to increase the accuracy and efficiency of the model by strengthening the attention to tiny target characteristics. Second, to increase the multi-scale fusion capacity and the underlying information of the feature map, the Weighted Bi-directional Feature Pyramid Network (BiFPN) is utilized in place of the Feature Pyramid Network+Path Aggregation Network (FPN+PAN). Meanwhile, Scalable Intersection over Union Loss (SIoU_Loss) is used instead of Complete Intersection over Union Loss (CIoU_Loss) to enhance the localization accuracy and further optimize the model training effect. This study also creates a framework for transfer learning that moves the YOLOv5-BEs' already-learned information from the source domain training dataset to the target domain dataset. This makes the model better at training on the small sample dataset. Empirical findings indicate that the suggested YOLOv5-BEs model performs better than current algorithms, improving the Mean Average Precision (mAP@0.5) by 1.3%~17.8%, and the Frames Per Second (FPS) by 4.46%~68.96%; through the transfer learning mechanism, the model's mAP @0.5 metric further reaches 68.2%, which is a 3% improvement from before transfer learning. The study's findings will offer an effective detection technique in the area of target identification for automatic driving, which has some potential uses.

Index Terms Transfer Learning, Target Detection, YOLOv5, Feature Extraction, Attention Mechanism

I. Introduction

Smart cars are progressively becoming more commonplace due to the economy's and technology's rapid development, but this also increases traffic accidents and congestion. Assisted driving systems developed based on artificial intelligence technology can effectively reduce the incidence of traffic accidents. In these systems, the target detection task is responsible for localizing and classifying various types of targets on the road and providing timely feedback of the detection results. However, due to the diversity of target detection types, the complexity of real-world traffic environments, and the high computational and memory demands of target detection algorithms, which results in insufficient detection accuracy, slow processing speed, and poor robustness, these problems make it difficult for existing algorithms to meet the needs of assisted driving systems on mobile devices in practical applications [1].

Traditional target detection methods predominantly rely on sensor-based or manual approaches. The topic of deep learning-based target identification has seen significant advancements and shown a wide range of possible applications in assisted driving systems thanks to the deep learning algorithms put forth by Bengio [2] et al. There are two types of target detection algorithms right now: one-stage detection algorithms [3] and two-stage detection algorithms [4]. The two-stage detection algorithm first generates candidate regions in the input image by region suggestion network and then performs bounding box regression and category prediction on these candidate regions. Kang [5] et al. used a region suggestion network with Faster R-CNN to generate candidate frames and optimized the detection results by identifying the small targets to improve the accuracy of the low categorization scoring frames. However, the method needs to be trained by repeatedly adjusting the hyperparameters to keep the balance of positive and negative samples. In contrast, the one-stage detection algorithm simplifies the detection process and offers notable speed and computational efficiency gains, making it more appropriate for embedded or mobile systems with constrained computational resources. Liu [6] et al. proposed a Single Shot MultiBox Detector (SSD) algorithm, which, by using preset bounding boxes at different feature layers for object prediction, avoids the step of candidate region generation in traditional detection methods, thus simplifying the detection process, but the detection effect is poor in small target scenes. Guo [7] et al. proposed a Foreign Object Debris (FOD) target detection algorithm based on improved YOLOv3, which solves the problems of insufficient localization accuracy and leakage detection by optimizing the feature extraction network and increasing the detection scale, but the overall detection accuracy decreases. Using the Receptive Field Block (RFB) module to increase the network's receptive field and the parameter-free Simple Parameter-Free Attention Module (SimAM) attention mechanism Bottleneck to enhance the network's feature extraction capabilities without the need for additional parameters, Wang Zhibin [8] et al. proposed an enhanced YOLOv5 model recognition algorithm. Despite the improvement in detection accuracy, this model's detection speed is nearly identical to that of the original model. To satisfy the demands of assisted driving systems for quick and precise target recognition, it is imperative to create more effective target detection algorithms.

The YOLOv5 algorithm has a complex network architecture and a large number of model parameters, which leads to its training process taking a long time to reach convergence. In addition, the final performance may be affected by random initialization parameters, which makes the training results unstable. Through the use of pre-trained model weights as initialization, the transfer learning algorithm can reduce training time and speed up the model's convergence, improving the model's robustness and generalization ability [9]. There exist five more typical pre-training weighting frameworks for the YOLOv5 model, which are YOLOv5l, YOLOv5s, YOLOv5n, YOLOv5x and YOLOv5m, where YOLOv5n, YOLOv5s, and YOLOv5m are lightweight weighting frameworks, and YOLOv5l and YOLOv5x are heavyweight weighting frameworks. Tang Lindong [10] et al. proposed an algorithm for complex road traffic target detection using YOLOv5s pre-training weighting framework, aiming at balancing the accuracy of the model with the training speed. To enhance the algorithm's immunity to interference in complex backgrounds, a multi-head self-attentive residual module and a CoordConv convolution are introduced, thus improving the feature extraction capability and robustness of the network. However, with the increase of model complexity, the detection speed shows a certain degree of degradation. Tu Chengfeng [11] et al. proposed a high-precision spam detection algorithm using the YOLOv5n pre-trained weight framework to reduce the number of parameters and computation, and combining the lightweight networks ShuffleNetv2 and GhostNet, which achieves the reduction of the model size while guaranteeing high accuracy. However, due to the small number of model parameters, its generalization ability has some deficiencies. Jia Weidi [12] et al. proposed a lightweight car face detection method based on the YOLOv5m pre-trained weight framework, which provides a lightweight structure and high detection performance. The method was able to reduce the model size and improve the performance by replacing the convolutional blocks in the original network and introducing upper and lower layer feature fusion modules. However, despite the lightweight improvement, the model still fails to meet the application requirements of embedded devices. The YOLOv5 series of lightweight pre-trained weight files face the problems of decreased detection rate, poor generalization, and excessive model size during the improvement process, which makes it difficult to meet the application requirements of autonomous driving target detection. Therefore, there is an urgent need to develop a pretraining weights framework based on automatic driving target detection devices with better applicability to improve the practicality and efficiency of the algorithm.

In summary, the present paper proposes a high-precision target detection model that integrates transfer learning with the enhanced YOLOv5 algorithm to address the limitations of contemporary road target detection algorithms. These limitations include small target detection, detection speed, vulnerability to misdetection and omission, and training time. To enhance the model's capacity for extracting features from small-scale objects and to improve detection accuracy and computational efficiency, the ECA mechanism is first implemented. Second, the traditional FPN+PAN architecture is replaced with the weighted BiFPN, which improves multi-scale feature fusion performance and low-level feature maps' capacity to extract information. To further enhance the model training effect and improve localization accuracy, SIoU_Loss is employed in place of CIoU_Loss. Finally, a transfer learning framework is proposed to utilize the pre-training knowledge of the YOLOv5-BEs model on the source domain dataset and migrate it to the target domain dataset. This approach is intended to accelerate the model convergence process and shorten the training time. The following are the study's main contributions:

(1) By adding the weighted BiFPN, the SIoU loss function, and the ECA attention mechanism, an enhanced YOLOv5 algorithm is suggested to overcome the algorithm's shortcomings in detecting small targets, detection speed, and false detection and omission. This improves the model's target recognition ability.

(2) A transfer learning framework based on YOLOv5-BEs is presented to enhance the generalization ability and robustness of the YOLOv5-BEs algorithm by transferring the pre-training knowledge learned during the model's training in the source domain to the target domain dataset.

(3) A target detection algorithm incorporating transfer learning and YOLOv5-BEs model is applied to a small-sample autopilot dataset, mAP@0.5 increased by 3%, and demonstrates the advantages of fast detection speed and high generalizability.

The architecture of this article: section 1 discusses the current needs and development context of autonomous driving and the shortcomings of target detection algorithms at the current stage, and presents the research contributions of this paper, section 2 specifies the structure of the proposed YOLOv5-BEs model, section 3 describes the mechanism and process of transfer learning, section 4 gives the model validation, ablation experiments, comparison experiments, and the results of transfer learning, and section 5 concludes the innovativeness of the YOLOv5-BEs model and the superiority of using the transfer learning mechanism are presented, and based on the current research results, potential directions for future research are proposed.

II. Method

Aiming at the YOLOv5 algorithm's problems of insufficient small-target detection capability, partial loss of original feature information, and prediction frame drift, which are difficult to meet the detection requirements of automatic driving target detection, this paper proposes an improved YOLOv5-BEs model. Specifically, the ECA mechanism is integrated into the backbone network to strengthen the model's ability to capture feature information of small targets, thereby enhancing both computational efficiency and recognition accuracy; BiFPN is used to replace the traditional FPN+PAN structure in order to strengthen the ability of the feature map's underlying information extraction and to improve the model's fusion of multi-scale features; in light of the issue of predicted bounding box drift associated with CIoU_Loss, this study proposes the adoption of SIoU_Loss as a replacement to further enhance localization accuracy and optimize model training efficacy. The framework of YOLOv5-BEs algorithm is shown in Figure 1.

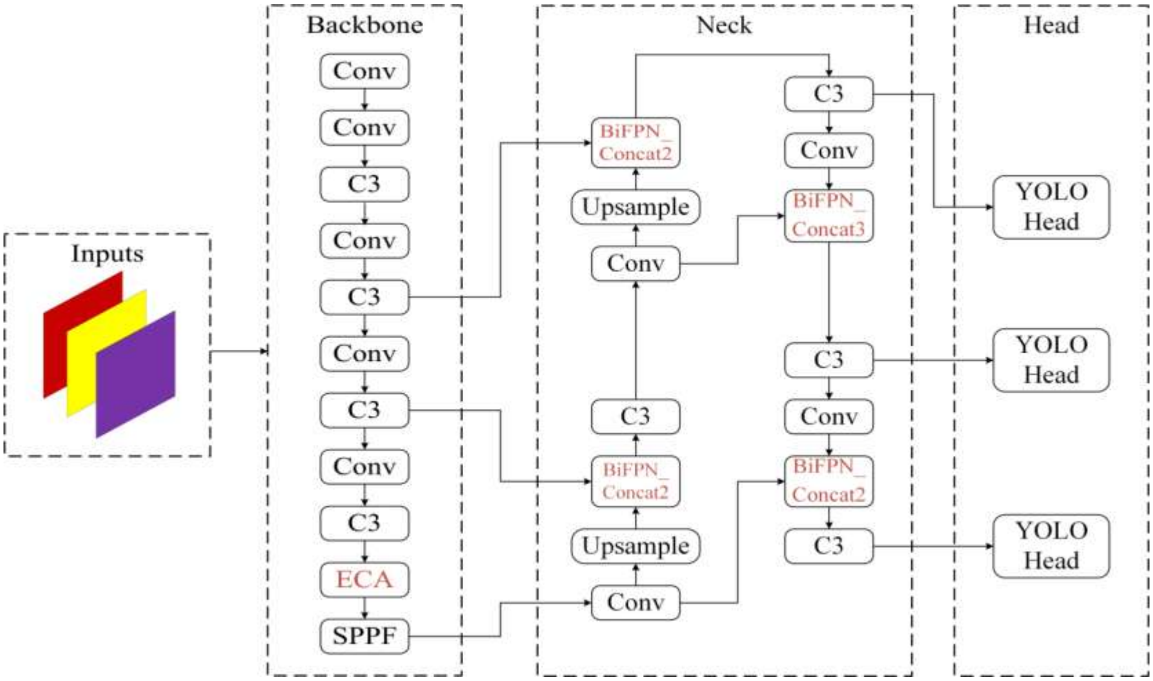


Figure 1: Network structure of YOLOv5-BEs

II. A. YOLOv5 model

The YOLO family [3] of algorithms is predicated on the concept of regression classification; it is a single-stage target detection technique. The algorithm generates multiple target candidate regions by dividing the grid and performs target classification and target location localization work in the same neural network. In this experiment, YOLOv5 model is chosen as the base model due to its lighter structure and wider application scenarios.

The YOLOv5 network architecture comprises four key components: the Inputs module cleans up the data, the Backbone network pulls out features, the Neck module combines features from different scales, and the Head network makes the final detection. First, the inputs perform data enhancement of the image and optimize the anchor frame generation and image processing strategy; second, the backbone network is in charge of extracting the features of the image, using the Focus structure to slice and duplicate the image, and optimizing the feature extraction procedure using the C3 and SPPF structures to minimize computational bottlenecks; subsequently, the Neck network employs the C3 structure and the FPN+ PAN structure to carry out the extracted feature information for the deeper extraction and fusion; finally, the head network performs the prediction and outputs the final result. The original YOLOv5 model demonstrates strong performance in large object detection, though it exhibits certain limitations when addressing the recognition of small distant targets. While the combined FPN+PAN architecture in YOLOv5's neck network effectively enhances feature fusion capabilities, this hierarchical aggregation approach inevitably causes partial loss of critical low-level feature information during the multi-scale integration process, ultimately compromising detection accuracy in complex scenarios. Even though the CIoU is utilized as the loss function and the frame aspect ratio's bounding scale information, the prediction frame's drift issue could still arise. The aforementioned issues make it challenging to adjust the model's parameters and base network scale alone to meet the scene requirements for automatic driving target detection. Therefore, we can further optimize the original YOLOv5 model by combining several enhanced modules.

II. B.ECA attention mechanism

The ECA attention mechanism [13] is a neural network architecture for image processing tasks. The ECA module may efficiently concentrate on the relationship between image channels while retaining the high efficiency to improve feature performance, avoiding the potential drawbacks of the dimensionality reduction operation when compared to the SE attention mechanism [14]. The detailed architecture of the ECA module is illustrated in Figure 2. Its core principle involves integrating a channel attention mechanism into the convolutional operation process. The input feature with dimensions (H, W, C) first undergoes global average pooling to compress the two-dimensional features (H, W) of each channel, resulting in a feature map of size (1, 1, C). Subsequently, by defining the coverage range of local cross-channel interactions as k, a 1D convolutional kernel with size k is applied to process the feature map, thereby capturing inter-channel dependencies. An output with dimensions (H, W, C) is then produced by multiplying the acquired weights element-wise with the appropriate channels of the input feature map after normalizing them using a sigmoid activation function. Cross-validation is avoided because the k-value can be found using Eq. (1) because it is proportional to the number of channels.

$$k = \left| \frac{\log_2(C) + b}{\gamma} \right|_{odd} \tag{1}$$

In the formula: b, γ are usually set to 1 and 2 to adjust the ratio between the convolution kernel size and the number of channels; C denotes the number of channels of the input feature map; and $| \ |_{odd}$ denotes the closest odd number of k to the value of this function.

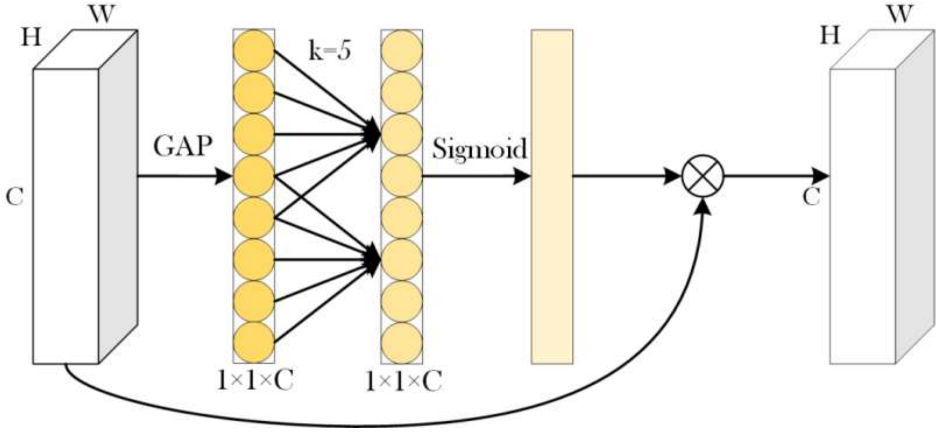


Figure 2: Structure of ECA module

II. C.Weighted bidirectional feature pyramid

BiFPN is a new neural network design that is frequently applied to computer vision problems, particularly semantic segmentation and target identification [15]. The BiFPN exhibits a marked distinction from the conventional FPN by establishing bidirectional cross-layer relationships between the feature pyramid's neighboring levels and implementing an adaptive weighted feature fusion mechanism. This improvement makes BiFPN more efficient in dealing with the information interaction between features at different scales, and it can effectively integrate features at different levels, which helps to capture multi-scale features and ensure that the feature integration achieves the optimal effect, thus improving the efficiency and performance of the model.

Cross-scale connectivity is realized through the construction of bidirectional pathways that integrate both top-down and bottom-up architectures. By fusing features extracted from the backbone network with appropriately scaled counterparts in the bottom-up pathway, this design ensures enhanced retention of shallow semantic information while concurrently mitigating the degradation of deep semantic features. In this process, BiFPN dynamically assigns weights to input features based on their importance and scales the weights to within the interval [0, 1] by Fast Normalized Fusion (FNF). In addition, each bi-directional path is treated as a feature network layer, and this structural layer is reused to achieve more advanced feature fusion.The modular design of BiFPN enables it to be easily embedded into various neural network architectures, which effectively improves the accuracy of the target detection and semantic segmentation tasks, and at the same time enhances the neural network's capability of understanding and interpreting the multiscale information, thus significantly improving the model performance and adaptability. adaptability. The network architecture of BiFPN is shown in Figure 3.

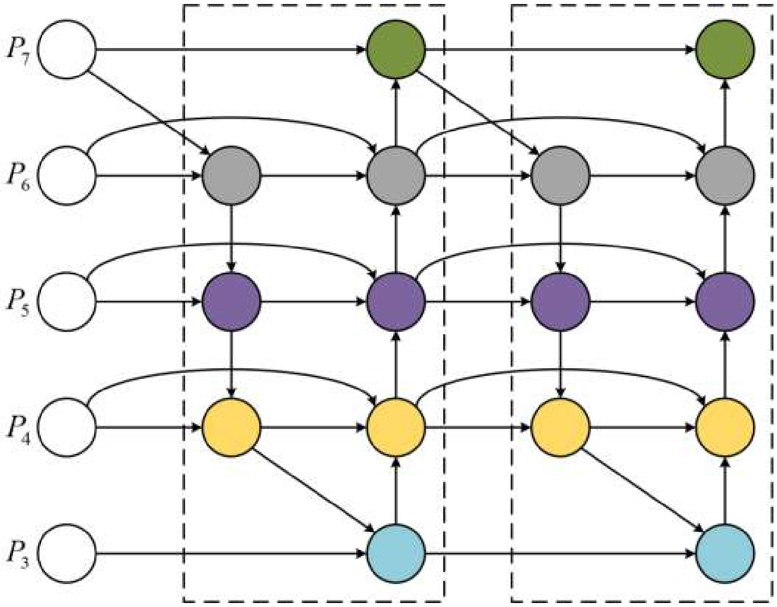


Figure 3: Structure of BiFPN module

II. D.SIoU loss function

The loss function in the YOLOv5 algorithm is CIoU. Factors such as the distance between the centroids, the aspect ratio, and the non-overlapping and overlapping regions of the real and predicted frames are all considered by the CIoU loss function, which is calculated as follows

$$Loss_{CIoU} = 1 - IoU + \frac{\rho^2}{c^2} + \alpha v \tag{2}$$

$$v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \tag{3}$$

$$\varepsilon = \frac{v}{(1 - IoU) + v} \tag{4}$$

In the formula: IoU represents the ratio of intersection between the actual and anticipated frames; the euclidean distance between the centers of the actual and predicted frames is denoted by ρ ; the diagonal length of the smallest closure zone that can hold both the real frame and the predicted frame is denoted by c ; v is used to measure the similarity of the width-to-height ratio; ε is the weighting function; (w^{gt} , h^{gt}) and (w , h) are the width and height of the real and predicted frames, respectively. When IoU is high, the weight of aspect ratio increases; when IoU is low, the weight of intersection ratio increases.

Given that CIoU has the problem of prediction frame drift during the training process, which leads to slow convergence speed and inefficiency, SIoU introduces the vector angle required for the regression between the real frame and the prediction frame on the basis of CIoU, and redefines the penalty term, which effectively prevents the prediction frame from drifting around and makes it fit the real frame better, which then improves the convergence speed of the network and the training quality. The angle loss is shown in Figure 4.

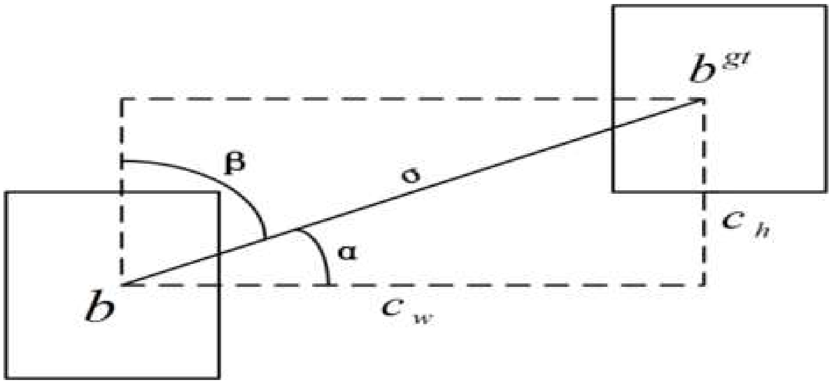


Figure 4: Angular loss diagram

In order to ensure that the prediction frame moves quickly towards the X or Y axis during the regression process and approximates the true frame along the axis of interest, the convergence process will preferentially work to minimize α when $\alpha \le \frac{\pi}{4}$; and vice versa for β. To reach this goal, SIoU introduces the angular cost function, which is calculated as follows

$$\wedge = 1 - 2\sin^2(\arcsin x - \frac{\pi}{4}) \tag{5}$$

Therefore, the SIoU_Loss regression loss function is formulated as

$$Loss_{SIoU} = 1 - IoU + \frac{\Delta + \Omega}{2} \tag{6}$$

$$\Delta = \sum_{t=x,y}(1 - e^{-\lambda\rho_t}) \tag{7}$$

$$\rho_x = (\frac{b_{c_x}^{gt} - b_{c_x}}{c_w})^2, \rho_y = (\frac{b_{c_y}^{gt} - b_{c_y}}{c_h})^2 \tag{8}$$

$$\Omega = \sum_{t=w,h}(1 - e^{-w_t})^\theta \tag{9}$$

$$\omega_w = \frac{|w - w^{gt}|}{\max(w, w^{gt})}, \omega_h = \frac{|h - h^{gt}|}{\max(h, h^{gt})} \tag{10}$$

$$x = \frac{c_h}{\sigma} \tag{11}$$

$$\sigma = \sqrt{(b_{c_x}^{gt} - b_{c_x})^2 + (b_{c_y}^{gt} - b_{c_y})^2} \tag{12}$$

$$c_h = \max(b_{c_y}^{gt} - b_{c_y}) - \min(b_{c_y}^{gt} - b_{c_y}) \tag{13}$$

In the formula: Δ denotes the distance loss considering the angular loss; Ω denotes the shape loss; λ denotes the weight of the distance value under the time priority; b is the prediction frame; b^{gt} is the real frame; the smallest outside rectangle of the prediction and real frames' width and height are denoted by c_w、c_h、b_{c_y}、$b_{c_y}^{gt}$ is the y-coordinate of the prediction and real frames; θ denotes the weight on the shape loss, which takes the range of [2, 6]; the distance between the centers of the real and predicted frames is shown by σ. SIoU incorporates the angle information into the regression process, which effectively reduces the total degrees of freedom of the loss function and thus improves the model accuracy and robustness.

III. The Proposed Method

III. A. Transfer learning mechanisms

Transfer learning [16] is a machine learning methodology that centers on the idea of applying knowledge gained by a model in one task or environment to another task or environment as an initial training condition. The key to transfer learning is to identify and utilize the knowledge similarity between the source task and the target task. By utilizing existing knowledge and pre-trained models, transfer learning can effectively alleviate the overfitting problem caused by too small a training dataset, reduce the dependence on large-scale datasets, and shorten the time for model training. This greatly increases the learning efficiency in addition to improving the model's performance.

Five more typical pre-training weights frameworks exist for the YOLOv5 model, as shown in Table 1. Their differences are mainly reflected in the different numbers of feature extraction modules and convolutional kernels in the algorithm structure, and the different widths and depths of the networks obtained.The two frameworks, YOLOv5l and YOLOv5x, are larger in size and have superior performance in terms of size and performance, but have higher demands on computational resources and require longer training times.YOLOv5n, as the smallest in size in the YOLOv5 family, fastest weights, the width of its input feature map and the depth of the network are smaller, so the detection performance is poorer, and the practical application ability is not strong.The YOLOv5s and YOLOv5m frameworks are widely used in routine target detection tasks, with a more balanced size and performance than the other, but due to the higher demand on the training time and detection speed in this experiment, the frameworks are not applicable to the automatic driving target detection . Aiming at the limitations of the traditional YOLOv5 series of pre-training weights frameworks, this experiment proposes a YOLOv5-BEs pre-training weights framework incorporating transfer learning.

Table 1: Parameters of YOLOv5 weighting frameworks

Model	Params (M)	mAP@ 0.5:0.95	FLOPS @640 (B)	mAP@ 0.5	Size (pixels)
YOLOv5n	1.9	28.0	4.5	45.7	640
YOLOv5s	7.2	37.4	16.5	56.8	640
YOLOv5m	21.2	45.4	49.0	64.1	640
YOLOv51	46.5	49.0	109.1	67.3	640
YOLOv5x	86.7	50.7	205.7	68.9	640

In transfer learning, data features and their distributions are commonly referred to as domains, where the domain from which knowledge is acquired is termed the source domain, while the domain requiring further learning is designated as the target domain. In this experiment, 1000 images data were collected and labeled, but due to the relatively small amount of data and insufficient features, directly used to train the YOLOv5-BEs model with a large number of parameters will lead to overfitting, long training time and lower accuracy, therefore, this experiment adopts the Transfer learning mechanism in order to improve the performance and training efficiency of the model and to enhance its generalization, and through the YOLOv5-BEs algorithm on the source domain dataset to obtain the YOLOv5-BEs pre-training knowledge, and then the YOLOv5-BEs pre-training knowledge is put into the small sample dataset for further training.

III. B. Migratory learning process

The specific process of transfer learning based on the YOLOv5-BEs model is shown in Figure 5. First, the VOC2007 dataset is regarded as the source domain, and the YOLOv5-BEs model is pre-trained using this dataset, so as to obtain the pre-trained model containing the weights of YOLOv5-BEs. Then, the pre-trained model is applied to the small sample dataset of this experiment for training and is trained in comparison with the unpretrained YOLOv5-BEs model on the same dataset. By comparing the performance of the two on the small sample dataset, the effect of transfer learning can be effectively evaluated and the whole process of transfer learning can be finalized.

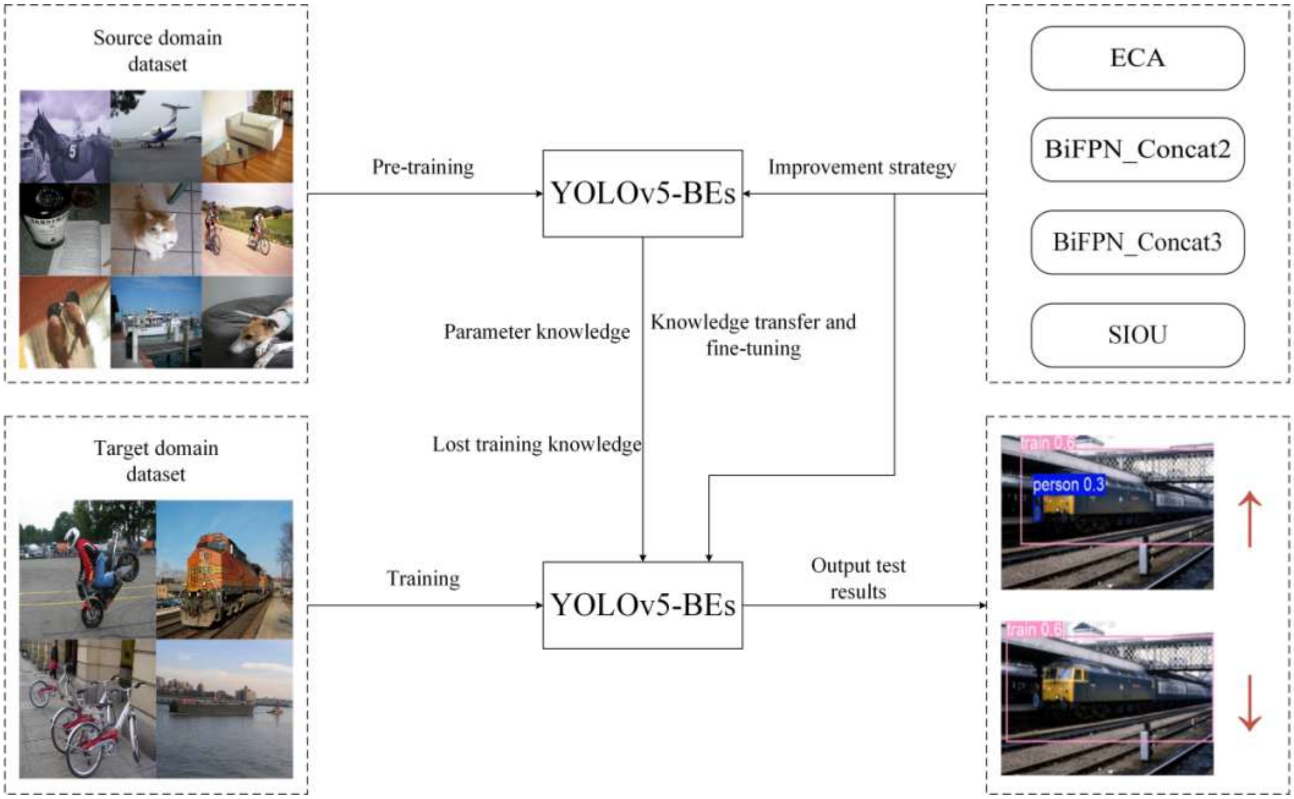


Figure 5: The process of migratory learning

IV. Experiments and Analyses

IV. A. Experimental environment

(1) Simulation environment

The hardware configuration of this experiment is illustrated in the following diagram: The CPU employed is an Intel Core i5-12500H@2.50 GHz processor, operating under a Windows 11 x64 system architecture. The system is equipped with 16.0 GB of RAM and an NVIDIA GeForce RTX 3050 GPU with 12.0 GB of dedicated video memory. The experimental environment utilizes PyTorch 2.4.0 as the deep learning framework, implemented through Python 3.9.19 programming language. GPU acceleration is enabled via CUDA 12.4 parallel computing architecture.

(2) The data set

The source domain dataset is the publicly available VOC2007 dataset from the Pascal VOC challenge, as shown in Figure 6. This dataset randomly divides the 9963 samples into training, validation, and test sets in a 7:2:1 ratio.

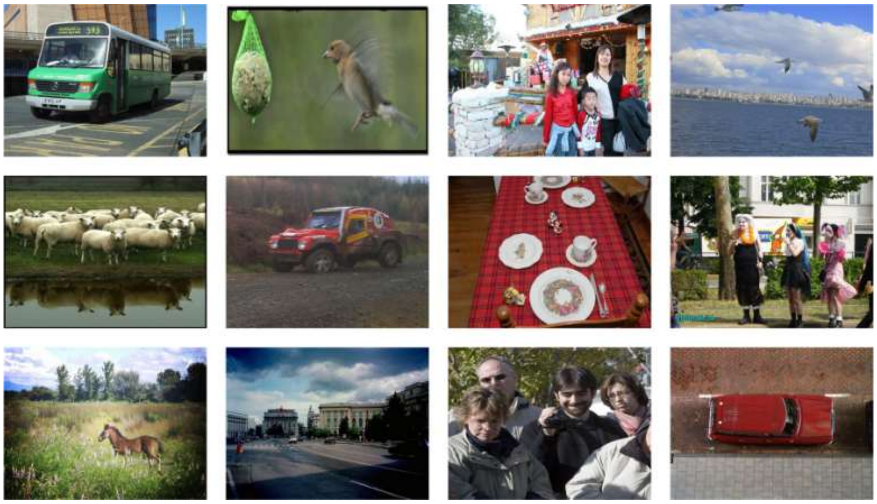


Figure 6: Source domain dataset

The target domain is a small sample dataset, as shown in Figure 7. This dataset contains 20 object categories, each image is labeled, and the object categories covered include animals, transportation, furniture, and so on. This experiment only focuses on the human and transportation category labels, i.e., person, bicycle, boat, bus, car, motorcycle, and train. The dataset comprises a total of 1,000 samples, which were randomly partitioned into training and test sets following an 8:2 ratio.

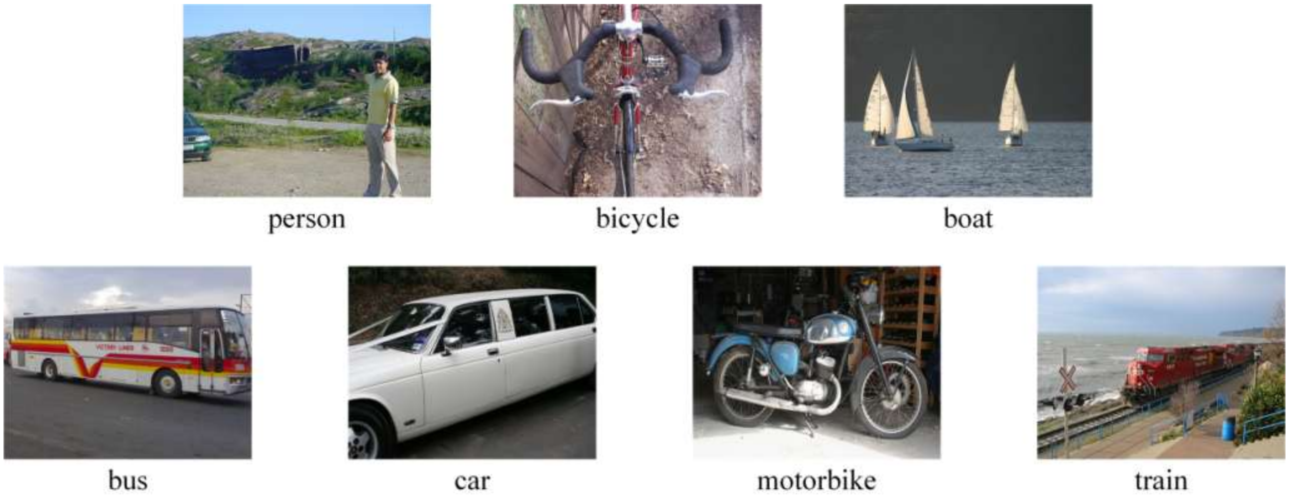


Figure 7: Target domain dataset

(1) Parameter setting

The algorithm setup for this experiment is shown in Table 2.

Table 2: Model parameter settings

Epochs	100
Batch size	16
Initial learning rate (lr0)	0.01
Loop learning rate (lrf)	0.01
optimizer	SGD
learning rate momentum	0.937
Workers	8
Input Image Size	640× 640

(2) Evaluation indicators

The models in this experiment are mainly modeled by Precision (P), Recall (R), The average mean average precision at IOU threshold of 0.5, mAP@0.5, and frames per second, FPS, are used to reflect these four indicators.

The accuracy rate can be defined as the ratio of successfully correctly predicted positive samples of the items detected by the model to all positive samples.

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

In the formula: The number of accurately identified positive samples is called TP, whereas the number of negative samples that were identified as positive is called FP.

Recall is the proportion of positive samples identified by the model to the number of correct positive samples.

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

In the formula: The number of positive samples that are mistakenly identified as negative samples is denoted by FN.

The average accuracy mean at an IOU threshold of 0.5 is the mean value of the accuracy of the model in predicting multiple categories at an intersection and integration ratio (IOU) threshold of 0.5.

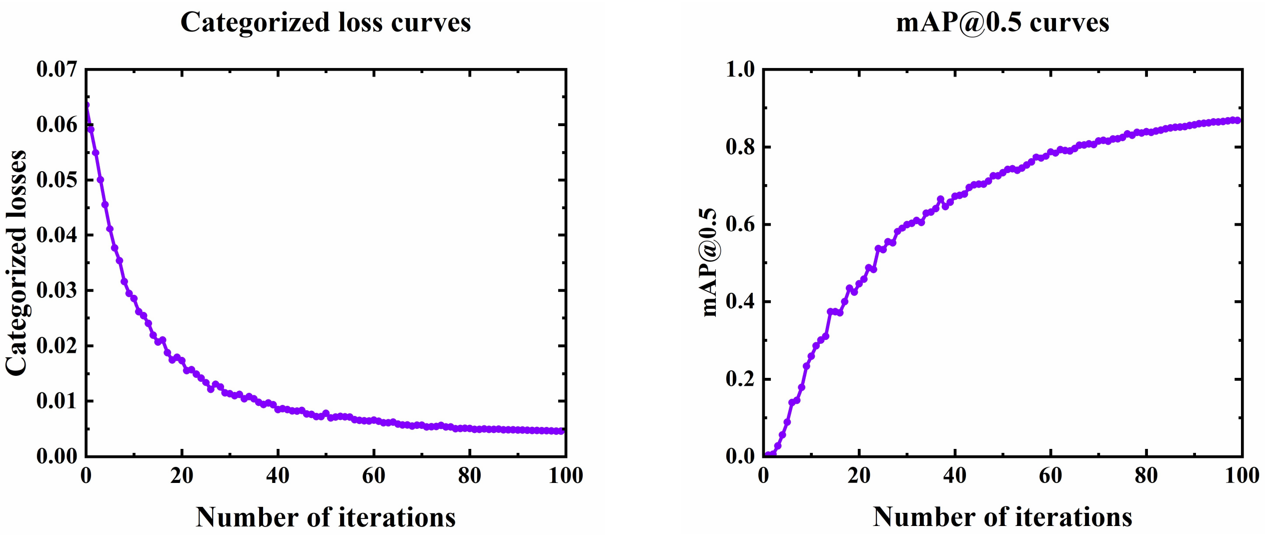
$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{16}$$

$$AP = \int_0^1 P(r)dr \tag{17}$$

In the formula: The region that the coordinate axis and PR curve enclose is known as AP. All the above evaluation indexes are used for the performance test of this experimental model, the higher the value, the better the model's performance is demonstrated to be.

IV. B. Model validation

The classification loss curves and mAP@0.5 curves for all categories of the YOLOv5-BEs model are shown in Figure 8. Observing Figure 8 (a), it can be seen that the classification loss shows a decreasing trend with the increasing number of iterations. When training first started, the loss value decreases significantly to 0.0072 at the 50th iteration, after which the loss value tends to stabilise and approaches 0. This indicates that the training process is effective and the model has strong classification ability. In Figure 8 (b), the mAP@0.5 indicator curve shows a gradual upward trend. When the iteration reaches 90 times, the mAP@0.5 value has reached 0.855, and subsequently the curve tends to be stable and finally reaches 0.868, which represents that the model has higher classification performance.



(a)

(b)

Figure 8: Model performance curve

Table 3 shows the results of the mean average precision mean (mAP@0.5) of the YOLOv5-BEs model on seven different categories. From Table 4, it can be observed that there is variability in the performance of the model on each category. Specifically, the categories "car" and "bus" contribute significantly to the sum of mAP@0.5, with values of 94.9 and 94.8, respectively, while the category "boat" has a relatively high mAP@0.5. The reason for the relatively low mAP@0.5 is that the boundary and background of this category are blurred, and the introduction of the ECA module provides a larger sensory field, which makes the model pay more attention to the global information and high-level features, which results in the recognition effect of this category being affected. Overall, the YOLOv5-BEs model maintains a high level of the mAP@0.5 metric.

Table 3: Target detection results

Evaluation indicators	bike	boat	bus	car	person	motorcycle	train	all
mAP@0.5/ (%)	89.3	76.0	94.8	94.9	92.0	90.6	90.6	89.5
P/ (%)	87.3	72.8	87.9	90.5	87.5	85.6	86.5	85.4
R/ (%)	81.7	68.1	88.2	90.5	83.1	85.7	77.6	82.1

IV. C. Ablation experiments

Ablation experiments are conducted for the YOLOv5-BEs model proposed in this experiment with the aim of exploring the enhancement effect of the improved module, and the experimental results are shown in Table 4. In Table 4, " √ " indicates that the module is adopted.

Table 4: Results of ablation experiments

Serial number	YOLOv5	SIoU	ECA	BiFPN	P/%	R/%	mAP@0.5/%	FPS
1	√	-	-	-	85.1	81.1	87.2	83.81
2	√	√	-	-	87.7	78.7	87.4	85.64
3	√	√	√	-	86.5	81.5	87.9	86.72
4	√	√	√	√	85.4	82.1	89.5	88.52

Analyzing Table 4, it can be seen that: after replacing the base model loss function with SIoU, the model recall is reduced, but the precision rate and FPS are substantially improved by 2.6%, 1.83, and the mAP is improved by 0.2%, respectively, which improves the model performance; subsequently, with the introduction of the ECA attention mechanism, the model recall and the mAP are improved by 0.4% and 0.7%, respectively, and the FPS is improved by 2.91; Finally, with the introduction of the BiFPN module, the model performance is further enhanced, and the precision rate, recall rate, mAP, and FPS are improved by 0.3%, 1.0%, 2.3%, and 4.71, respectively, compared with the original model. The above results clearly show that the proposed method in this experiment significantly improves the performance metrics, and the overall performance of the model is all enhanced.

IV. D. Comparative experiments

In order to further reflect the superiority of the YOLOv5-BEs model, the training results of this paper's model are compared with YOLOv3, YOLOv5, YOLOv7, Li [17], and Zhou [18] target detection algorithms on the VOC2007 dataset, using mAP@0.5 and FPS as the evaluation metrics, and the comparison results are shown in Table 5.

Table 5: Comparative experimental results

Model	mAP@0.5/%	FPS
YOLOv3	79.8	73.67
YOLOv5	87.2	78.70
YOLOv7-tiny	88.2	84.74
Li [17]	71.7	60.18
Zhou [18]	80.4	52.39
YOLOv5-BEs	89.5	88.52

From the comparison results in Table 5, it can be seen that under the same experimental environment and the same dataset, the YOLOv5-BEs model proposed in this experiment improves the mAP by 9.7%, 2.3%, 1.3%, 17.8%, and 9.1%; and the FPS improves the FPS by 20.16%, respectively, compared to YOLOv3, YOLOv5, YOLOv7, Li [17], and Zhou [18], 12.48%, 4.46%, 47.09%, and 68.96%, respectively. The comparison shows that the YOLOv5-BEs model outperforms the other models and has better detection effect on self-driving target recognition with strong practical application capability.

IV. E. Results of Transfer experiments

In order to verify the impact of the proposed transfer learning method on the model training efficiency and performance, the training results of the YOLOv5-BEs model without transfer learning are compared with the training results with transfer learning. Table 6 shows the results of model precision, recall, and mAP@0.5 after five times of training for each of the two approaches, and Figure 8. shows the time-contrast histograms for five times of training for each of the two approaches.

The results of all five training sessions that used transfer learning are noticeably superior to those that did not, as indicated in Table 6. Specifically, the maximum improvement in precision, recall and mAP@0.5 are 17%, 7.3% and 11%, respectively; and the minimum improvement is 11.4%, 2.7% and 7.8%, respectively. The comparison of training time demonstrated in Figure 9. shows that the training time of the model after five times of using transfer learning is less than the training without transfer learning, showing faster convergence efficiency. Figure 10. shows the comparison of the detection effect on the concern categories before and after transfer, the model's detection accuracy is significantly enhanced, reducing omissions and false detections.

Table 6: Experimental results before and after transfer

Before	P/%	R/%	mAP@0.5/%	After	P/%	R/%	mAP@0.5/%
1	58.1	57.4	56.2	1	75.1	60.1	67.2
2	58.9	55.2	56.5	2	70.3	60.9	66.2
3	61.6	54.3	58.3	3	76.8	61.6	68.2
4	63.0	54.6	57.2	4	73.4	59.8	66.9
5	64.0	55.1	58.9	5	75.5	59.9	66.7

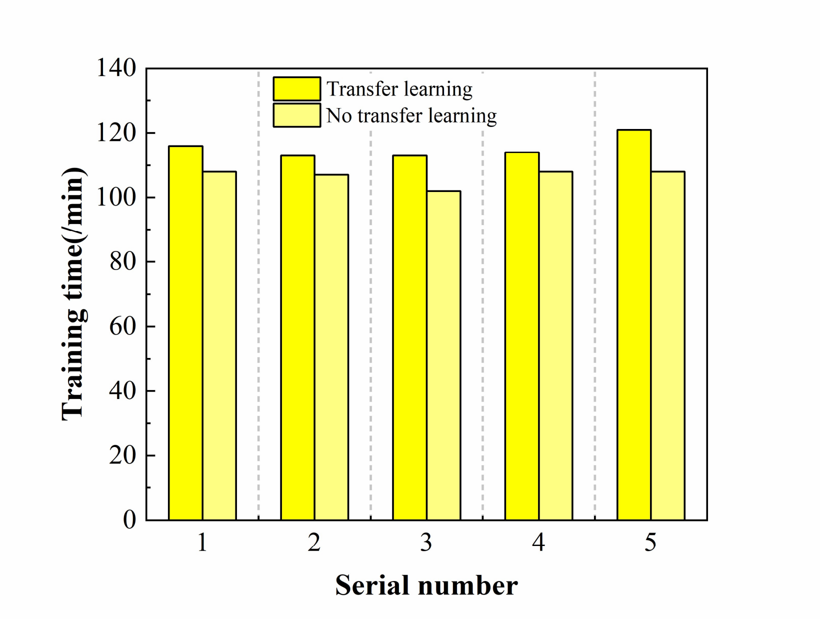
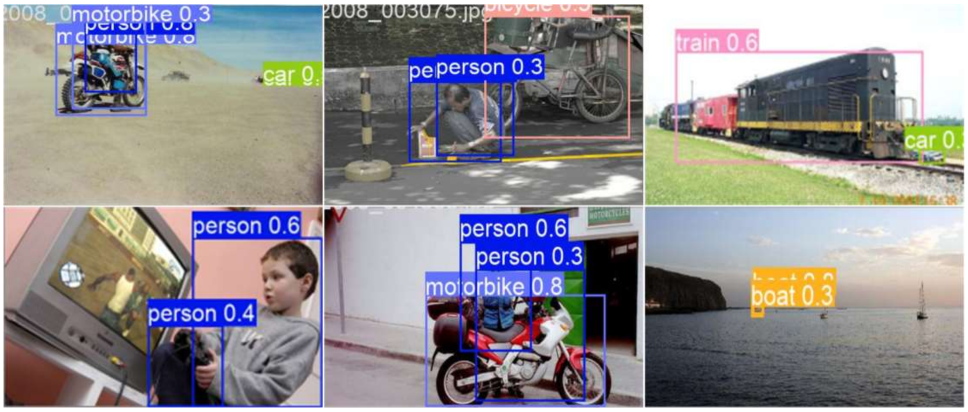
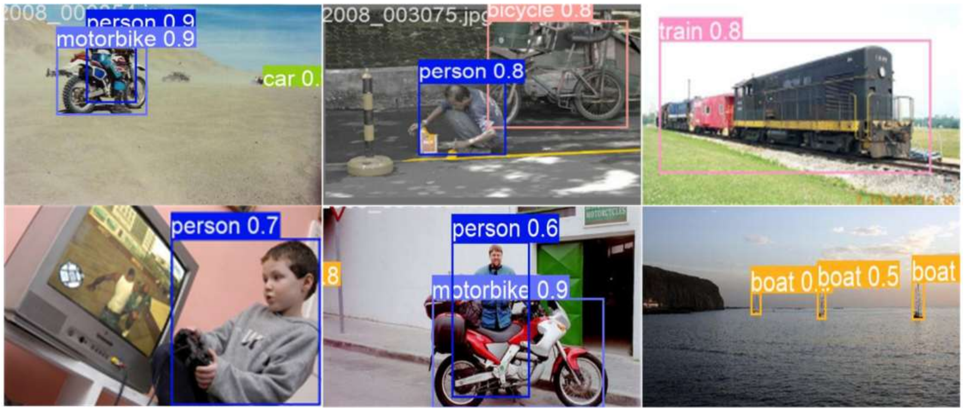


Figure 9: Histogram of training time



(a) YOLOv5-BEs unmigrated detection effect



(b) YOLOv5-BEs transfer detection effect

Figure 10: Comparison of detection results before and after YOLOv5-BEs transfer

As shown in Figure 10, the detection accuracy of YOLOv5-BEs model is effectively improved and compensates for some of the missed and misdetected cases.

Combined with Table 6, Figure 9. and Figure 10, it can be seen that although the YOLOv5-BEs model itself has a good detection effect, but without transfer learning, the actual detection still exists in the case of miss-detection and false-detection, and the detection accuracy is not high, which limits the actual detection effect of the model. After transfer learning, the model not only improves the detection effect but also increases the detection accuracy and speed, which fully proves the effectiveness of transfer learning in improving the model accuracy and enhancing the model performance.

In order to better reflect the detection performance of the YOLOv5-BEs weights in this experiment, the experimental weights are compared with the other weights of the YOLOv5 series, namely YOLOv5n, YOLOv5s, and YOLOv5m, in a small-sample dataset, and the evaluation criteria used are the checking rate and mAP@0.5, as well as the training time, and the results of the comparison are shown in Table 7.

Table 7: Results of weighting comparison

Weighting	P/ (%)	mAP@0.5/ (%)	Training time/min
YOLOv5n	68.9	65.4	70.6
YOLOv5s	70.8	70.3	135.1
YOLOv5m	68.2	62.9	916.5
YOLOv5-BEs	76.8	68.2	107.6

Based on the experimental data presented in Table 7, under the same experimental environment and dataset framework, analysed from the perspective of a single evaluation metric, YOLOv5-BEs achieves the highest value in the checking accuracy metrics, YOLOv5n performs optimally in terms of training time consumed, YOLOv5s achieves the highest value in the mAP@0.5 metrics, and YOLOv5m performs mediocrely in all evaluation metrics.Taken together, although YOLOv5n has a significant advantage in training time, it has a 7.9% and 2.8% gap with YOLOv5-BEs in the two evaluation metrics of checking rate and mAP0.5, respectively; YOLOv5s, despite a slightly higher mAP@0.5, has a 6.0% gap with YOLOv5-BEs in checking rate, and has a muchIt takes 27.5 min. based on the above comparison, YOLOv5-BEs achieves a better balance in all evaluation indexes, and its comprehensive performance exceeds that of other weighting models, which fully proves the superiority of YOLOv5-BEs in improving the model accuracy and enhancing the overall performance of the model.

IV. F. Discussion

In this research, we offer a high-accuracy target identification model that combines the enhanced YOLOv5 algorithm with transfer learning. By integrating three improvement strategies, the model achieves high detection accuracy, and subsequently accelerates the convergence process and shortens the training time by fusing transfer learning. Although the introduction of multiple improvement mechanisms increases the computational burden and cost of the model, the need for accuracy is higher than the need for lightweighting in autonomous driving target detection scenarios. Therefore, although the computational cost of YOLOv5-BEs model training in the source domain is high, the process can be completed in advance in practical applications, and after transfer learning, the model will improve the training efficiency in the application process, thus reducing the computational cost. The target detection model based on source-domain training proposed in this paper can not only be migrated to other datasets, but also has good application prospects in the fields of road crack detection and steel crack detection.

V. Conclusion

In response to the limitations of existing road object detection algorithms regarding insufficient small object detection capability, frequent occurrences of false positives and missed detections, as well as excessive training duration, this paper proposes an enhanced YOLOv5-BEs model through systematic improvements. The ECA attention mechanism is added to the backbone network to improve the model's attention to small target features. To improve the model's feature extraction capacity, multi-scale feature fusion weighted BiFPN is used instead of FPN+PAN in order to optimize the fusion impact of feature information at different scales. To fix the problem of prediction frame drift in the regression process, this paper further optimizes the loss function and uses SIoU_Loss instead of CIoU_Loss. This makes the model better at detecting things. Finally, the transfer learning mechanism is applied to improve the model convergence speed and reduce the training time. The experimental results show that YOLOv5-BEs improves the detection accuracy of road targets by 2.3% and the detection frame rate by 12.48% compared with YOLOv5s, and reduces the training time by 27.5 min after transfer learning, obtaining a model that can be more suitable for road target detection. In this paper, the advantages of the improved algorithm in terms of detection speed and accuracy are verified through ablation experiments and comparison with other target detection algorithms. Future work will focus on integrating the proposed model into assisted driving systems.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, author-ship, and/or publication of this article.

Data Sharing Agreement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

[1] Deng Yaping, Li Yingjiang. Review of YOLO algorithm and its applications to object detection in autonomous driving scenes[J]. Journal of Computer Applications, 2024, 44(06): 1949-1958.

[2] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444

[3] Mi Zeng, Lian Zhe. Review of YOLO methods for universal object detection[J]. Computer Engineering and Applications, 2024, 60(21): 38-54.

[4] Shan Xianying, Zhang Lin, Li Zehu. Review of research progress in object detection driven by deep learning[J]. Computer Engineering and Applications, 2024, 1: 1-18.

[5] Kang M, Leng X G, Lin Z, et al. A modified faster R-CNN based on CFAR algorithm for SAR ship detection[C]//2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP). IEEE, 2017: 1-4

[6] Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector[C]//European Conference on Computer Vision. Cham: Springer, 2016: 21-37

[7] Guo Xiaojing, Sui Haod. Application of Improved YOLOv3 in foreign object debris target detection on airfield pavement[J]. Computer Engineering and Applications, 2021, 57(8): 249-255.

[8] Wang Zhibin, Feng Lei, Zhang Shaobo, et al. Vehicle type recognition based on improved YOLOv5 and video images[J]. Science Technology and Engineering, 2022, 22(23): 10295-10300.

[9] Li Xinyao, Li Jingjing, Zhu Lei, et al. Efficient transfer learning of large models with limited resources: a aurvey[J]. Chinese Journal of Computers, 2024, 47(11): 2491-2521.

[10] Tang Lindong, Yun Lijun, Luo Ruilin, et al. Complex road traffic target detection algorithm based on improved YOLOv5s[J]. Journal of Zhengzhou University(Engineering Science), 2024, 45(03): 64-71.

[11] Tu Chengfeng, Yi Anlin, Yao Tao, et al. High-precision garbage detection algorithm of lightweight YOLOv5n[J]. Computer Engineering and Applications, 2023, 59(10): 187-195.

[12] Jia Weidi, Yu Pengfei, Yu Guohao, et al. Lightweight car front detection method based on improved YOLOv5m[J]. Electronic Measurement Technology, 2023, 46(12): 125-133.

[13] Liu H, Zhang Y, Chen Y. A symmetric efficient spatial and channel attention (ESCA) module based on convolutional neural networks[J]. Symmetry, 2024, 16(8): 952.

[14] Lai Qinbo, Ma Zhenghua, Zhu Rong. Uav image object detection based on attention mechanism and dilated convolution[J]. Computer Applications and Software, 2025, 42(02): 227-235.

[15] Tan M X, Pang R M, Le Q V. EfficientDet: scalable and efficient object detection[C] // 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 10778-10787

[16] Zhao Z, Alzubaidi L, Zhang J, et al. A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations[J]. Expert Systems With Applications, 2024, 242: 1-6.

[17] Li Tian, Lin Guimin, Yu Yekai. Research on vehicle target detection for 3improved YOLOv5s[J]. Auto Time, 2024, (01): 16-18.

[18] Zhou Qing, Tan Gongquan, Yin Songlin, et al. Road object detection algorithm based on improved YOLOv5s[J]. Chinese Journal of Liquid Crystals and Displays, 2023, 38(05): 680-690.