International Journal of Science and Engineering Applications
Volume 14-Issue 08, 33 –38, 2025, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1408.1008

# Research on Lizard Target Detection based on YOLOv5s-SSE

Xu Yang
School of Electric Information and Electrical Engineering
Yangtze University
Jingzhou, China

**Abstract**: In a complex natural environment, the traditional lizard detection method based on image processing is easily affected by the environment, resulting in high false detection and missed detection rates, which has a certain impact on the accurate monitoring and behavioral research of lizards. Therefore, designing an efficient lizard species detection model is convenient for more detailed detection of the distribution of local lizard populations and further study of lizard behavior patterns. To meet the above requirements, this paper proposes a lizard detection model based on YOLOv5s-SSE. This model integrates the Switchable Atrous Convolution (SAC) into the C3 module of the baseline model and combines it into the backbone feature extraction network. It can expand the receptive field and improve the multi-scale feature extraction capability without increasing the convolution kernel; Shuffle Attention mechanism (SA) is adopted to construct the channel attention mechanism and the spatial attention mechanism through grouping, and the information exchange between different groups is strengthened through channel shuffling, which not only better improves the expressive ability of the model, but also takes into account the advantage of lightweight; The EIoU loss function is used to measure the difference between the real box and the predicted box in multiple aspects to strengthen the focus on the real difference between the length and width of the bounding box and its confidence, thereby improving the positioning accuracy of the target bounding box and helping to capture small-scale lizard image information more accurately. This study takes eight species of lizards, including the Australian magic lizard and the red-eyed hawk lizard, as the main research objects, and constructs the LDD (Lizard Detection Dataset) dataset containing 8 types of lizards. Compared with the baseline model, the lizard detection model proposed in this study improves mAP@0.5:0.95, Precision, and Recall by 1.8%, 0.8%, and 2.2%, respectively. The lizard detection model proposed in this paper has good detection performance in complex environments, reduces the probability of false detection of the detection model, and realizes efficient detection and identification of lizard species.

**Keywords**: Object Detection; lizard; YOLOv5s; Switchable Atrous Convolution; Attention Mechanism

## 1. INTRODUCTION

With the rapid development of deep learning, object detection has been widely used in the field of wildlife identification. The early methods of target detection feature extraction mainly rely on manual feature extraction. With the continuous progress of machine learning, algorithms such as support vector machines[1] that learn the boundaries between targets and non-targets to achieve classification have been applied. In the field of deep learning, with the successful application of convolutional neural networks, object detection has also made significant progress. Object detection algorithms are divided into two-stage detection algorithms and single-stage detection algorithms. Although the two-stage algorithm has the advantage of high detection accuracy, the model complexity is relatively high and the number of parameters is also large. Representative two-stage algorithms include the R-CNN[2] algorithm, the FastR-CNN[3] algorithm, the SPP-Net[4] algorithm, etc. The single-stage algorithm model is simple, has fewer parameters, and is faster than the two-stage algorithm. Representative algorithms include the YOLO series[5-7] and SSD[8] algorithms.

Traditional monitoring methods have limitations such as low efficiency and difficulty in scalability, especially in complex natural scenes. Many researchers at home and abroad have conducted a series of studies on wildlife species detection and proposed a variety of methods to improve the detection performance of the model. Some studies have improved the architecture of existing target detection models. For example, Li[9] implemented the blood parrot object detection task by adding a two-layer routing attention and a visual universal transformer BiFormer to the YOLOv8n model, improving the recall rate and average precision by 1.4% and 1.0% respectively compared with the original version. To solve the problem of giant pandas being difficult to identify in complex environments, Lv[10] proposed a deep separable neck network based on yolov5n fused attention, and adopted the Alpha-IoU loss function. The mAP@0.5 value of the model was improved by 2.7% compared with the original version. In terms of the accuracy of target detection in complex environments, researchers are currently committed to improving the accuracy of detection models and enhancing their ability to identify targets in complex environments[11-13].

Lizard populations have different body sizes depending on their distribution areas, and the appearance of most lizards is similar to their surroundings, which leads to a decrease in the accuracy of lizard detection and an increase in the false detection rate. To this end, this study proposed a lizard target detection model based on YOLOv5s that integrates switchable hole convolution and Shuffle Attention mechanism. The model improves the feature extraction capability of multi-scale targets by controlling the size of the receptive field, thereby effectively reducing the probability of false detection of lizard targets.

## 2. METHOD

### 2.1 YOLOv5 Model

YOLOv5 has four different versions: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. Each model has an increasing number of parameters, computational complexity, and network depth and width. YOLOv5s is the smallest model in

International Journal of Science and Engineering Applications
Volume 14-Issue 08, 23 –26, 2025, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1408.1006

the YOLOv5 series. Therefore, it has a relatively low computational complexity and provides a relatively fast inference processing speed. YOLOv5s is a lightweight model, but it can still provide sufficient accuracy in many tasks. Therefore, YOLOv5s is selected as the experimental benchmark model. YOLOv5s is a single-stage target detection model with four parts: input, backbone feature extraction network, enhanced feature extraction network and output. The YOLOv5s model architecture is shown in Figure 1.

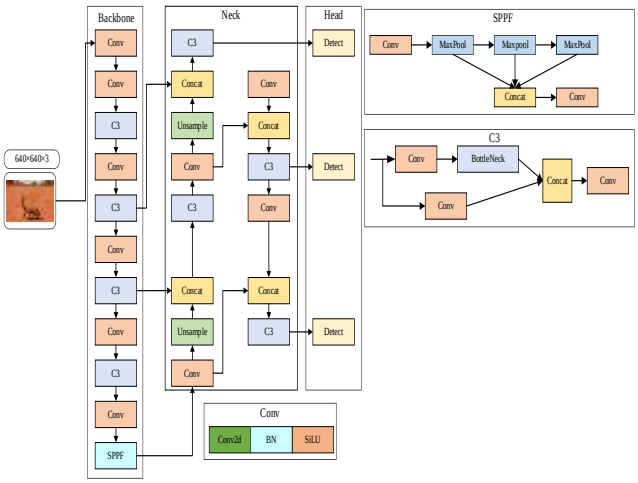



Figure 1. The structure of YOLOv5s model.

Input uses methods such as adaptive anchor box calculation to ensure that the input model data is correctly processed and recognized. Backbone, which consists of a Conv module, a C3 module, and an SPPF module, can extract local features such as texture and edges of the input image. It also enhances the network's feature extraction capabilities through residual connections and feature fusion. Neck mainly consists of two aspects: Feature Pyramid Network[14] and Path Aggregation Network[15]. Neck fuses and enhances the feature maps from Backbone at multiple scales, and the generated feature maps have better expressiveness. Head consists of a series of convolutional layers and fully connected layers, which are responsible for generating the prediction results required for target detection.

## 2.2 YOLOv5s-SSE Model

In order to solve the problem of false detection and missed detection of lizard species in complex and changeable natural environments, a single-stage target detector based on the attention mechanism and improved backbone network receptive field provides an intelligent technology for lizard species detection. This paper changes the C3 structure of the baseline model and uses switchable dilated convolution as the second convolution of Bottleneck, so that it can expand the receptive field area while keeping the convolution kernel size unchanged, enhance the extraction of multi-scale information in different natural environments, and improve the stability and performance of the model. The model uses Shuffle Attention mechanism to improve the model's perception of semantics and space, and suppress irrelevant information, and uses the EIoU loss function to make bounding box positioning more accurate, improve the model's ability to extract information about small lizards, reduce the probability of missed detections and false detections, and enable the model to complete lizard detection tasks more efficiently and accurately. YOLOv5s-SSE model is shown in Figure 2.

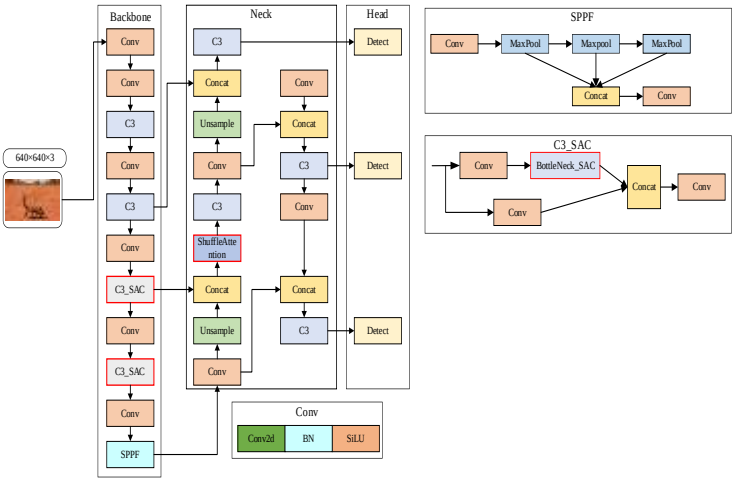



Figure 2. The structure of YOLOv5s-SSE model.

### 2.2.1 SAC module

The paper uses Switchable Atrous Convolution to improve the model's ability to recognize objects of different scales. Switchable Atrous Convolution[16] is based on Dilated Convolution[17]. Switchable Atrous Convolution consists of three main parts: the SAC module and two Global Context Modules located before and after the SAC module. the SAC is shown in Figure 3.

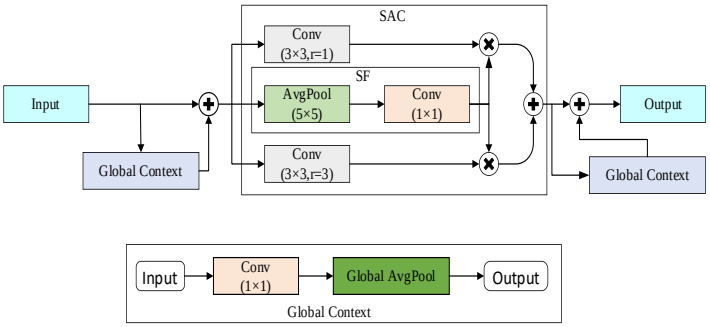



Figure 3. diagram of Switchable Atrous Convolution

The paper integrates the SAC convolution module into the C3 module of the Backbone network, reconstructing a 3×3 ordinary convolution in the Bottleneck structure of the C3 module into a switchable hole convolution of the same size, thereby constructing a new C3_SAC module. As shown in Figure 4. This module effectively suppresses the interference of irrelevant information by expanding the receptive field and enhancing the multi-scale feature extraction capability, further improving the detection capability of the model.

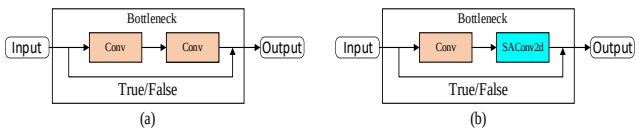



Figure 4. (a) Original module. (b)C3_SAC module

### 2.2.2 Shuffle Attention mechanism

Attention mechanisms have become an important component in improving the performance of neural networks. They help neural networks accurately focus on relevant elements of the

International Journal of Science and Engineering Applications
Volume 14-Issue 08, 23 –26, 2025, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1408.1006

input while suppressing irrelevant elements. Spatial attention mechanism and channel attention mechanism are commonly used in deep learning research. Both enhance the original features by aggregating the same features at all positions using different methods. Although studies such as CBAM[18] integrate the spatial attention mechanism and the channel attention mechanism into one module and achieve remarkable results, this design increases the difficulty of convergence and the amount of computation. The Shuffle Attention mechanism[19] adopted in this paper combines the spatial attention mechanism and the channel attention mechanism to construct a lightweight module that integrates the two attention mechanisms, thereby effectively solving the problems of high convergence difficulty and computational complexity. The Shuffle Attention mechanism first divides the features into multiple sub-features according to the channel dimension, then processes the sub-features through the shuffle unit, divides the sub-features into two groups, uses the channel attention mechanism and the spatial attention mechanism respectively, and finally performs feature aggregation. Among them, the sub-feature $X_a \in R^{C/G\times H\times W}$ is divided into $X_{a1}$ and $X_{a2} \in R^{C/2G\times H\times W}$ .In the channel attention mechanism, global information is first obtained through Global Averaging Pooling (GAP) to obtain the output $c \in R^{C/2G\times 1\times 1}$ . As shown in formula (1).

$$c = F_{gp}(X_{a1}) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{i=1}^{W} X_{a1}(i,j) \qquad (1)$$

Then use an adaptive scaling operation to learn the weight of each channel and adjust the importance of each channel. Finally, use the Sigmoid function for nonlinear transformation. As shown in formula (2).

$$X_{a1}^{'} = \sigma(F_C(c)) \bullet X_{a1} = \sigma(W_1 c + b_1) \bullet X_{a2} \qquad (2)$$

In the spatial attention mechanism, performing a Group Norm operation on $X_{a2}$ first can better capture and utilize the spatial information of the sub-feature map and obtain the spatial statistics of $X_{a2}$. Finally, $F_c(\square)$ is used to enhance the representation of the input, as shown in Equation (3).

$$X_{a2}^{'} = \sigma(W_2 \square GN(X_{a2}) + b_2) \square X_{a2} \qquad (3)$$

The results of the two branches of the spatial attention mechanism and the channel attention mechanism are spliced together to obtain $X_a^{'} = [X_{a1}^{'}, X_{a2}^{'}] \in R^{C/G\times H\times W}$ , and all sub-features are aggregated to form a whole with the same shape as the input. Then, channel shuffle is used to enhance the information exchange between different groups, which can better improve the expressiveness of the model.

### *2.2.3 EIoU*

This paper adopts Efficient Intersection over Union Loss[20] (EIoU), which is an enhanced loss function for bounding box regression. Although the CIoU loss function performs well by comprehensively considering geometric features such as overlapping area, center point distance, and aspect ratio, CIoU only focuses on the difference in aspect ratio and fails to reflect the true difference between length and width and their confidence levels. This limitation will affect the convergence efficiency of the algorithm. In order to improve the accuracy of the lizard species detection model, this study uses EIoU as the loss function. The EIoU loss value $L_{EIoU}$ consists of three parts: the intersection-over-union (IoU) loss $L_{IoU}$ , the distance loss $L_{dis}$ between the center point of the real box and the predicted box, and the aspect ratio loss $L_{asp}$ , as shown in formula (4).

$$L_{EIoU} = L_{IoU} + L_{dis} + L_{asp} = 1 - IoU + \frac{\rho^2(b,b^{gt})}{(w_c)^2+(h_c)^2} + \frac{\rho^2(w,w^{gt})}{(w_c)^2} + \frac{\rho^2(h,h^{gt})}{(h_c)^2} \qquad (4)$$

Among them, $b, b^{gt}$ is the coordinate of the center point of the predicted box and the true box, and $\rho$ calculates the Euclidean distance of the element center point. $w, w^{gt}$ is the predicted box width and the real box width. $h, h^{gt}$ is the predicted box height and the real box height. $w_c, h_c$ is the width and height of the smallest outer box that covers both boxes.

The EIoU loss function not only considers the overlapping area between the target detection box and the true target box, but also further introduces the center point distance and aspect ratio difference of the target box, making the target bounding box positioning more accurate. EIoU measures the difference between the true box and the predicted box in multiple aspects, making the loss function more robust to objects of different scales and shapes, making it easier to extract information about small-scale objects. It has better detection effects for objects of various scales, such as lizards living in different environments and of different sizes.

## 3. EXPERIMENTS

### 3.1 Experimental Environment and Parameters

After image preprocessing, the image size is 640+640, the number of iterations is 300, the initial learning rate is 0.01, the batch size is 16, and the weight decay is 0.0005. The experimental environment of this paper is shown in Table 1.

These hyperparameter settings (learning rate, decay rate, batch size, and number of iterations) work together to optimize the

International Journal of Science and Engineering Applications
Volume 14-Issue 08, 23 –26, 2025, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1408.1006

model's learning efficiency of the input image features and enable it to achieve good performance within a reasonable time. The learning rate primarily regulates the trade-off between convergence speed and training stability, while the batch size is more limited by computing power and task complexity.

Table.1 Experimental environment.

| | |
|---|---|
| Operating system | Windows11 |
| CPU | i5-12490F |
| GPU | NVIDIA GeForce RTX 4060 |
| CUDA | 11.7 |
| framework | Pytorch1.13.0 |
| Python | 3.9.18 |
| Pycharm | 2023.1.3（community） |

## 3.2 DataSet

The lizard data used in this research experiment are mainly lizard images collected from different open source data platforms. The collected lizard images were manually annotated and data augmented using the online annotation tool MakeSense to construct the LDD dataset (Lizard Detection Dataset). The LDD dataset contains images of eight types of lizards, including the Australian devil lizard, red-eyed hawk lizard, Komodo dragon, peacock needle lizard, green iguana, double-crested crested lizard, pleated parasitic lizard, and bearded dragon. It is divided into training set and validation set in a ratio of 7:3, with 1,446 images in the training set and 620 images in the validation set. In order to enhance the detection model's ability to detect targets of different scales, lizard images of the same species but of different scales are added to the dataset. The lizard images are labeled and screened from multiple angles to improve the diversity of the dataset. Some images of the LDD dataset are shown in Figure 5.

Figure 5. Schematic diagram of lizard detection dataset

Table.2 Composition of lizard detection dataset

| Lizard Species | Number of Images |
|---|---|
| Komodo Dragon | 369 |
| Green Iguana | 219 |
| Frill-necked Lizard | 209 |
| Thorny Devil | 259 |
| Emerald Swift | 233 |
| Plumed Basilisk | 320 |
| Red-eyed Crocodile Skink | 203 |
| Bearded Dragon | 254 |
| Total | 2066 |

## 3.3 Experimental Results

### *3.3.1 Ablation experiments*

In order to verify that the optimized detection model has better detection effect in lizard species detection, this study conducted an ablation experiment on the detection model. The Bottleneck convolutional layer in the C3 module is replaced by the SAC module to form the C3_SAC module, which replaces the two C3 modules in the backbone network, improving the model's ability to extract multi-scale features and model performance; the Shuffle Attention mechanism module is introduced and compared with the ParNetAttention mechanism and the S2Attention mechanism; the loss function is replaced by the EIoU function and compared with the CIoU function and the WIoU function. The ablation experiments are shown in Table 3.

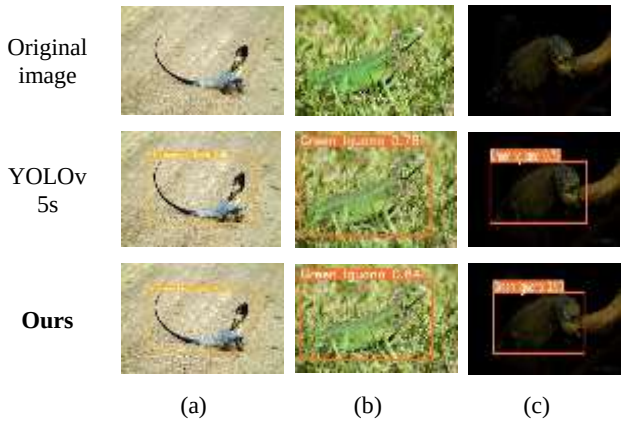
Table.3 The ablation experiments

| YOLOv5s baseline model | | | Precision | Recall | mAP@0.5:0.95 |
|---|---|---|---|---|---|
| C3_SAC | SA | EIoU | | | |
| — | — | — | 0.888 | 0.831 | 0.543 |
| ✓ | — | — | 0.885 | 0.841 | 0.549 |
| ✓ | ✓ | — | 0.892 | 0.820 | 0.556 |
| ✓ | ✓ | ✓ | 0.896 | 0.853 | 0.561 |

According to the ablation test performance comparison in Table 3, the detection performance indicators of the proposed model (YOLOv5s+C3_SAC+SA+EIoU) have been significantly improved. Compared with the baseline model, the precision has increased by 0.8%, the recall rate has increased by 2.2%, and the mAP@0.5:0.95 value has increased by 1.8%. In the improved experiment, the model that changed the C3 module to the C3_SAC module (YOLOv5s+C3_SAC) improved the recall rate of the baseline model by 1% and the mAP@0.5:0.95 value by 0.6%, indicating that the switchable dilated convolution effectively expanded the receptive field, improved the ability to extract multi-scale features, and enhanced the model's ability to identify lizards of different environments and different body shapes. The introduction of the stochastic attention mechanism (SA) significantly improved all model metrics, with accuracy increasing by 0.4% and mAPA@0.5:0.95 by 1.3% compared to the baseline model. By combining channel-

wise and spatial-wise attention mechanisms, the SA attention mechanism not only enhances the model's expressiveness but also offers the advantage of being lightweight. The EIoU loss function incorporates differences in the center distance and aspect ratio of the object bounding box, making the model more robust to objects of varying scales and shapes. Ablation experiments show that the optimized lizard species detection model significantly improves detection accuracy compared to the baseline model.

The number of existing public lizard datasets is small and they mainly come from documentary materials, online images and manual photography. Although deep learning methods have been introduced into wildlife target detection models, most datasets lack images of small objects. Therefore, this paper adds lizard images of different scales to the dataset. This paper enhances the detection capabilities of different image scales by expanding the receptive field, among other methods. This model not only improves accuracy compared to the baseline model, but also achieves more accurate recognition in natural environments with colors similar to those of lizards, mitigating the baseline model's false positives and missed detections. This is shown in Figure 6.




(a) (b) (c)

As shown in Figure 6(a), under standard lighting conditions, our model improves the lizard detection capability by 5% compared to the baseline model. As shown in Figure 6(b), in the complex situation where the environment and the lizard's appearance color are similar, the recognition ability of the optimized model reaches 84% compared with the baseline model. As shown in Figure 6(c), the model proposed in this study can more accurately identify targets in low-light environments than the baseline model, with an accuracy of 90%, which is 18% higher than the baseline model.

### *3.3.2 Comparative experiment*

To verify the effectiveness of the proposed model, the optimized model was compared with common object detection models such as Faster-RCNN and YOLOv4 on the same dataset. mAP@0.5 and mAP@0.5:0.95 were used as evaluation indicators. The performance comparison of each model is shown in Table 4. Experimental comparisons show that the proposed model achieves a mAP@0.5:0.95 value of 0.561, improving the mAP@0.5:0.95 values of other models by 0.4%, 0.8%, and 1.8%, respectively. Its mAP@0.5 value also improves by 1.3%, 3.1%, and 1.2%, respectively, compared to other models. The improved model achieved excellent results in lizard detection.

Table.4 Comparative experiment

| Model | mAP@0.5 | mAP@0.5:0.95 |
|---|---|---|
| Faster-RCNN | 0.879 | 0.557 |
| YOLOv4 | 0.861 | 0.553 |
| YOLOv5s | 0.880 | 0.543 |
| **ours** | 0.892 | 0.561 |

## 4. CONCLUSION

The YOLOv5s-SAC detection model proposed in this study Switchable Atrous Convolution to improve the model's ability to recognize objects at multiple scales. Shuffle Attention mechanism is used to enhance the model's spatial and channel perception, allowing it to focus on relevant information and suppress irrelevant information. The EIoU loss function is used to improve the model's sensitivity to small objects. On the self-made LDD lizard detection dataset, the detection model proposed in this paper improved mAP@0.5:0.95 by 1.8%, Precision by 0.8%, and Recall by 2.2% compared with the baseline model. The lizard detection model used in this paper achieves excellent recognition results, reducing the baseline model's probability of missed or false detections for lizards of varying sizes. This detection model also meets the accuracy requirements for lizard species detection in complex environments, providing model support for lizard population monitoring and research on lizard behavior patterns. However, while this model improves detection performance, it inevitably increases the model's parameters and computational complexity. Future work will focus on reducing the number of model parameters and improving detection speed while maintaining accuracy.

## 5. REFERENCES

[1] Balasubramaniam V. Artificial intelligence algorithm with SVM classification using dermascopic images for melanoma diagnosis[J]. Journal of Artificial Intelligence and Capsule Networks, 2021, 3(1): 34-42.

[2] Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition.2024:580-587.

[3] Girshick R .Fast R-CNN[C]//International Conference on Computer Vision.IEEE Computer Society, 2015.

[4] He, Kaiming, Xiangyu,et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2015,37(9):1904-1916.

[5] Redmon J. Yolov3: An incremental improvement[R]. arXiv preprint arXiv:1804.02767, 2018.

[6] Wang C Y, Bochkovskiy A, Liao H Y M. Scaled-YOLOv4: Scaling cross stage partial network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.2021: 13029-13038.

[7] Wang C Y , Bochkovskiy A , Liao H Y M .YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[J]. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,2023: 7464-7475.

[8] Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector[C]//Leibe B, Matas J, Sebe N, et al. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I. Cham: Springer International Publishing, 2016: 21-37.

[9] Li P L,Zhang S M,Shen L,*et al*. YOLOv8 blood parrot target detection and tracking method with double layer routing attention mechanism[J].Journal of Dalian Ocean University,2024,39(02):318-3266.

[10] Lv H T, Jia X L. Lightweight giant panda object detection model integrating attention mechanism[j].Laser Journal,2024.45(08):61-68.

[11] Yang W H, Liu T Y, Zhou J C, *et al*. CNN Swin Transformer forest wildlife image object detection algorithm based on improved YOLOv5s[J].Scientia Silvae Sincae,2024,60(03):121-130.

[12] Mingyu Z ,Fei G ,Wuping Y , *et al*.Real-Time Target Detection System for Animals Based on Self-Attention Improvement and Feature Extraction Optimization[J].Applied Sciences,2023,13(6):3987-3987.

[13] Tong Z M, Chen X H, Wang B F, *et al*. A wheat ear detection and counting method based on improved YOLOv5s[J]. Journal of Nanjing Agricultural University,2024,47(06):1202-1211.

[14] Ghiasi G, Lin T Y, Le Q V. NAS-FPN: Learning scalable feature pyramid architecture for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7036-7045.

[15] Zhou L, Rao X H, Li Y H, *et al*. A lightweight object detection method in aerial images based on dense feature fusion path aggregation network[J]. ISPRS International Journal of Geo-Information, 2022, 11(3): 189.

[16] Qiao S, Chen L C, Yuille A. Detectors: Detecting objects with recursive feature pyramidand switchable atrous convolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.2021:10213-10224.

[17] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[J]. arXiv preprint arXiv:1511.07122, 2015.

[18] Woo S, Park J, Lee J Y, *et al*. CBAM: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 3-19.

[19] Zhang Q L, Yang Y B. SA-Net: Shuffle attention for deep convolutional neural networks[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 2235-2239.

[20] Zhang Y F, Ren W Q, Zhang Z, *et al*. Focal and efficient IOU loss for accurate bounding box regression[J]. Neurocomputing, 2022, 506: 146-157.