An Improved YOLOv8 Algorithm for Detecting Remote Sensing Target Images

Chao Chen, Bin Wu

***Abstract*—The features extracted by deep networks have strong semantic information of high-level features, but the features of small target regions cannot be correctly described after downsampling for many times, so that it is difficult to generate stable discriminant features. In remote sensing images, multiple small targets coexist and the scale changes greatly, which makes it easier to lose small targets' information. In order to describe the features of each small target with different scales in remote sensing images more accurately, adaptive deformations convolution and LSKNet dynamic adjustment field strategies were proposed to fuse multi-view feature information. At the same time, in order to locate the small target boundary more accurately, a more comprehensive loss function was proposed to guide the optimization direction. The experiment result on 2 common datasets (such as COCO) and 4 remote sensing data sets (such as HRRSD) shown: The detection accuracy of small targets have been improved by three to five percent point.**

***Index Terms*—Adaptive receptive field, Adaptive deformable conv, LSKNet, SWIOU loss function**

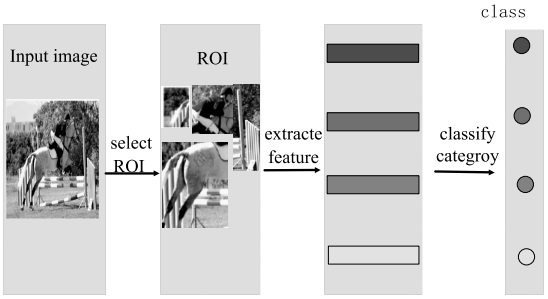


Fig. 1. Traditional common object classification's workflow

I. Introduction

In the field of computer vision, object detection is mainly involved in precise positioning and identification of specific objects in images or videos. Small target detection is one of the most challenging problems in the field of computer vision. Compared with large targets, small targets have smaller coverage area, lower spatial resolution and fewer usable features. The results are usually not satisfactory. In recent years, small target detection algorithms based on deep convolutional neural networks are also developed on common data sets. Here it is necessary to review the main algorithms in object detection.

A. Traditional universal target detection algorithm

Small target detection in high-resolution remote sensing images is a long-standing difficulty in detection tasks. Since most of the mainstream target detection algorithms are based on common data sets, it is necessary to review the traditional target detection algorithms. These methods have three main steps. First, the ROI of the region that may contain the target is screened and extracted; Then, the feature vector of each ROI is extracted; finally, the target category of the region is classified by the feature vector of ROI. The specific object classification's workflow is shown in Figure 1.

Manuscript received December 3, 2024; revised March 24, 2025. This work was supported in part by Object Detection Technology Based on Deep Learning (No.202307033), NSF of Sichuan Province (2023NSFSC0065) and Bridge Nondestructive testing and Engineering calculation Key Laboratory project of Sichuan University (No.2023QYY04).

Chao Chen is a PhD candidate of Southwest University of Science and Technology and an associate professor of Neijiang Normal University, Mianyang, Qinglong Avenue 59, P. R. China. (e-mail: ch10503@ 126.com).

Bin Wu is a professor of Southwest University of Science and Technology, Mianyang, Qinglong Avenue 59, P. R. China. (phone: 139- 0901-8585; e-mail: wubin@swust.edu.cn).

First, the whole image is scanned by sliding window[1], [2], [3], [4], [5], and then some fixed features are artificially set to describe the features of specific targets(For example: Haar[1], [2], [3], HOG[4], [5], SIFT[6], [7], SURF[8], etc.). In the identification phase, SVM(support vector machine)[9] can be used for object classification. Later scholars used bagging[10], cascade learning[11], adaboost[1], [2], [5] and other classification techniques to improve the accuracy and speed of target detection.

Dalal N et.al. used Hog[2], [12], [13], [14], [15] (essentially, it is the texture) to extract the target; Then, SVM classifiers were used to distinguish targets[4], [9], [10]; Drayer B et. al. proposed DPM, whose main principle is to construct a component detection method based on local features combined with textures[12], [13], [14]. A machine based on deformable parts[14] is a multipart model that learns and integrates with deformable losses. It shows strong robustness to deformed targets; In the detection phase, each characteristic component is excluded in turn, the target of interest is eventually detected. Later, scholars used the most advanced target detection system combined with other technologies to improve target detection's accuracy (For example: multi-scale detection, bounding box regression, context launching[8], etc.). The core of the mainstream algorithm is to extract multi-scale features by using feature pyramid; At the same time, the sliding window was compared with the key features on the image, which is also the basic framework of general target detection and pedestrian detection[10], [15]. However, the detection results are always much lower than those of deep learning-based methods[16]. Later, some scholars proposed semi-supervised learning and training methods[17], but, The efficiency of detection was not satisfactory. The attention mechanism[18], clustering[19] and selective search mechanism[20]were constructed in order

to extract significant features from images with a particular focus. However, the real-time detection of the target have not been well improved.

B. Object detection algorithm based on deep learning

At the 2012 Image Classification Challenge, the world's AlexNet algorithm based on GPU training was presented[21]. It opened a new era of deep learning(the detection accuracy exceeded the second place by several orders of magnitude). So the AlexNet[22] deep network was a milestone and laid a solid foundation for deep learning in the later stage, the efficiency of target identification was greatly improved.

Later, inspired by AlexNet network, some scholars designed a VGG model to increase the number of layers to 19[23] and used it for target detection. The feature maps extracted by VGG provided assurance for later detection, so it was an excellent early model in the field of target detection(The VGG model only expands in depth.). GoogLeNet proposed a network within a network structure[24], [25] to deepen the network. Later, some scholars added identical network layers in width[26], and the most influential one was ResNet which directly deepens the depth to 152 layers[26]. Later, a large number of target detection technologies were emerged, which were classified according to the detection principle. They are as shown in Figure 2.

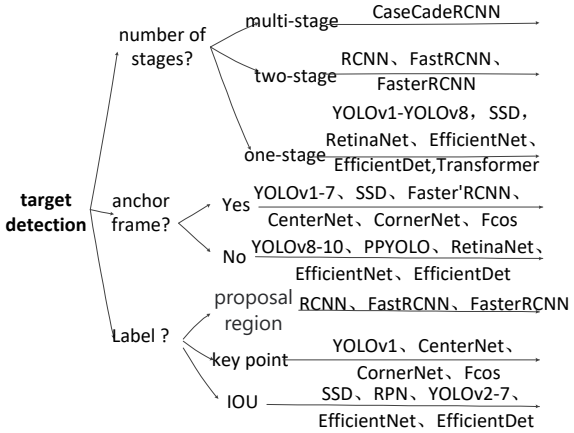


Fig. 2. Classification diagram of target detection technology based on different principle

(1)Two-stage target detection algorithm

Girshirk et. al.[27] proposed the RCNN target detection model, which had greatly improved the accuracy. However, the training model is too complex, which seriously affects the real-time target detection. Inspired by space pyramid matching, He et al.[28] proposed space pyramid network to adapt the size of input image. Faster RCNN model using RPN can effectively shorten the detection time[29], [30]. FastRCNN and Faster RCNN are often used as a baseline method to implement other improved algorithms. However, Faster RCNN has many problems such as large number of anchor frames, unbalanced positive and negative samples of anchor frames, high model complexity, long training time and low efficiency. A large number of proposal regions need to be calculated, which becomes the computational bottleneck of the Faster RCNN model. In order to solve the above problems, representative methods include RFCN[30] and Mask RCNN[31], [32], which eliminated these drawbacks to a certain extent. However, no special attention has been paid to the detection of small targets and multi-targets, especially for small targets with occluded, crowded images and less than 50 pixels, the detection effect was poor, and there was no experimental data for small targets in remote sensing images.

(2) Single-stage target detection algorithm

For example, for 224*224 images, the speed of detection based on the two-stage algorithm is less than 22 frames per second. Later, scholars put forward the representative method of YOLO series(YOLOv1[33],YOLOv2[34],YOLOv3[35], [36], [37], [38], [39]). The detection speed of YOLOv3 model can reach 20 FPS(Only slightly less accurate). Later, scholars took CSPDarknet-53[40], [41], [42] as the skeleton network and PANet as the fusion network to carry out the end-to-end target detection algorithm of multi-scale feature fusion. The CSP module were proposed to solve the problem of a large amount of forward computation in the network[40], whose general idea is to divide the input features into two ways at the time of input. Then, carry out cross-layer, multi-scale and multi-feature fusion at the end. The YOLOv4 algorithm proposed that the PAN structure can transfer features from bottom to top, and then carry out feature fusion for different network layers. YOLOv4 is superior to the one-stage algorithm in real-time detection speed[41], [42]. However, after analyzing the COCO data set, it was found that there are still many small targets that are not detected. And, the targets in the multi-target crowded state still have false detection or missing detection. Later, many improved YOLOv5[43], [44], [45], [46] introduced multiple CSP small modules embedded in the backbone network; In YOLOv5, the more superior GIOU_Loss is used as the loss of regression frame, the accuracy of detection has been improved. However, there were still some small targets that were not detected. And there were still some false detection or missing detection for multi-target occlusion or crowded targets.

SSD algorithm is divided into six scales for target detection[47], and the accuracy of small target detection was improved to a certain extent. Later, some scholars proposed that DSSD[48] realized small target detection, but it was not applied to small target detection in remote sensing images. The transformer model has quickly gained traction in the computer vision space, especially in the field of object recognition and detection. After investigating the results of the most advanced target detection methods, the authors noted that transformer model was superior to mature CNN-based detectors in almost every video or image dataset[49], [50], [51], [52], but it was not used for small target detection in remote sensing images because the original spatial information of small targets was disrupted when image subblocks were split.

These models fix structure and can not change the receptive field size adaptively. This does not meet the need of high level networks to encode semantic features in spatial locations. An image may have multiple objects of different scales, and different positions of the image correspond to objects of different scales. The convolution filter encodes semantic features in spatial positions, so it is expected to have an object detection model of the adaptive convolu-

tion filter's receptor field. Inspired by literature[53], [54], [55], we proposed to extract shallow features by adaptive adjustment of receptive field to fuse multi-scale features. In the later stage, adaptive deformable convolution and LSKNet that dynamically adjusted receptive field strategies were introduced to enhance the interaction of multi-scale feature information. Finally, in order to locate small targets faster and more accurately, a comprehensive SWIOU loss function [56], [57], [58], [59], [60], [61] combined with the above improved backbone network was proposed to achieve accurate detection of small targets in remote sensing images.

II. Model Construction

A. Dynamic receptive field based on adaptive three-way deformable convolution and LSKNet

The convolution kernel of the convolution filter in the traditional CNN layer is independent of the scale of the input image (generally fixed to 3*3), and the receptive field of the convolution kernel is also fixed. It was found that CNN wihch is the closer the positive sample scaleis to the optimal input image scale of the convolutional filter. the larger the output of the convolutional filter, the higher the detection accuracy [53], [54].

In order to more accurately describe the characteristics of small targets with different scales in the same scene, an adaptive receptive field was proposed[53]. The semantic information of high-level features is strong, but the description ability of low-level features is insufficient, which makes it impossible to generate stable discriminative features by using deep or shallow features alone. In respect of the above issues, adaptive deformable convolution and LSKNet dynamically adjusted receptive field strategies were proposed to fuse the discriminative features of small targets.

B. Adaptive three-way deformable convolution

Adaptive convolution was added different dilatation values to the standard convolution filter, calculate the convolution separately, and then select the maximum activation value as the output. It was as shown in Figure 3.

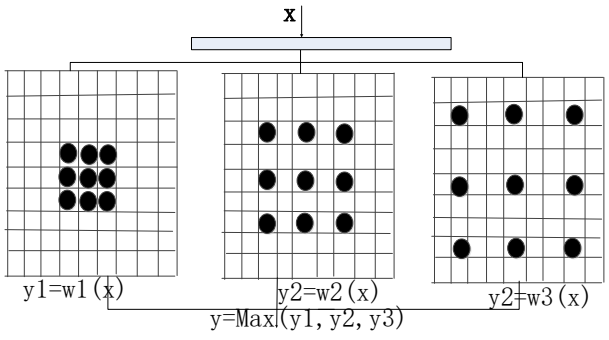


Fig. 3. Adaptive convolution structure with step sizes of 1, 2 and 3 respectively

The selected dilatation value is determined autonomously by the input feature map, and the process is carried out without adding any other parameters. The information includes the size of the receptive field corresponding to the current feature point (r), the distance between two adjacent feature points(j) and the step size of the current convolutional filter(S). Formula (1) calculates the feature distance in the output feature map, which is equal to the feature distance of the input feature map multiplied by the step size of the convolutional filter, k is the convolution kernel scale, formula (2) calculates the receptive field.

$$\mathrm{j_{out}} = \mathrm{j_{in}} \times S \tag{1}$$

$$\mathrm{r_{out}} = \mathrm{r_{in}} + \mathrm{j_{in}} \times (k-1) \tag{2}$$

The difference between standard 2D convolution and 2D deformable convolution are shown in Figure 4.

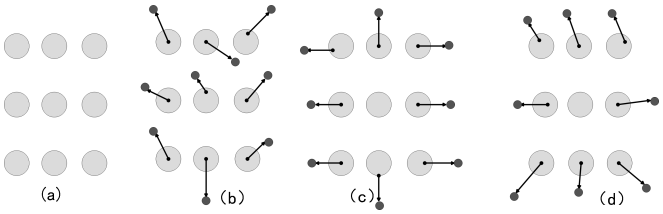


Fig. 4. Standard convolution and deformable convolution (Note: (a) is standard convolution, (a) -(c) is deformable convolution, and c was chosen as the deformable convolution kernel in this paper.)

In order to extract more information in the small target domain, two-dimensional deformable convolution was introduced for feature extraction. This is shown in Figure 5.

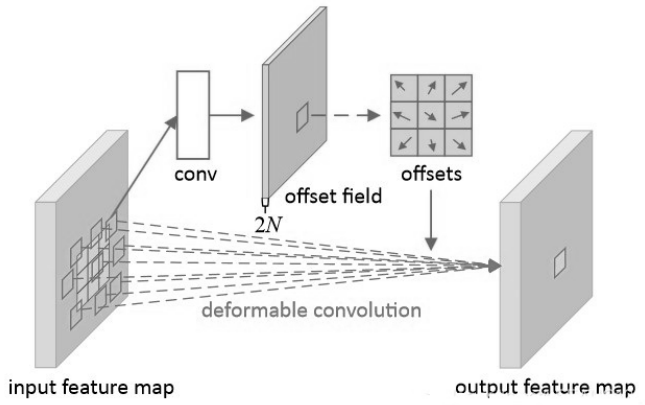
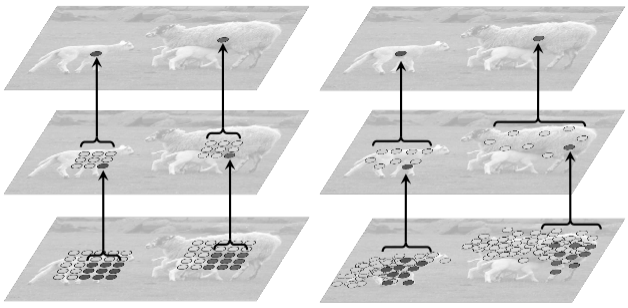


Fig. 5. Schematic diagram of deformable convolution

The operation of two-dimensional deformable convolution is shown in Figure 6.



(a) standard convolution (b) deformable convolution

Fig. 6. Schematic diagram of standard convolution and deformable convolution that applied to image operation

DCN adds 2D offset to sample the position of the rectangular grid in the standard convolution. Combining the advantages of deformable convolution and adaptive deformable convolution, adaptive deformable convolution was designed to incorporate multi-scale information. Convolution with fixed step and adaptive deformable convolution are shown in Figure 7.

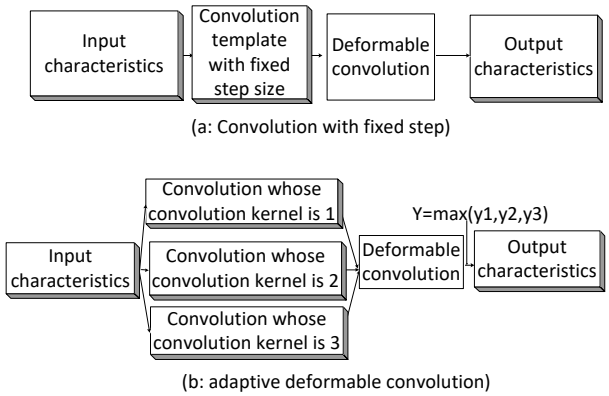


Fig. 7. Schematic diagram of deformable convolution with fixed step size and adaptive step size

C. Adjust spatial receptive field based on improved LSKNet

It is easy to misdetect small targets in remote sensing image because of the large scale change and the existence of many different kinds of small targets in the same scene during downsampling. We embed the CABM module into LSKNet. The LSKNet module scales its spatial receptive field in real time. And, the CABM module can pay more attention to the characteristics of the small target itself. Therefore, various small objects in remote sensing scenarios can be better modeled, where: CABM represents the spatial channel attention mechanism. LSKNet(Large K) means large convolution kernel. In this paper, the convolution kernel sizes are 5*5 and 7*7 respectively; $F_1^{1^*1}$ and $F_2^{1^*1}$ respectively indicate the different-channel convolution whose convolution kernel is 1. LSKNet is shown in Figure 8. Where: CBAM (

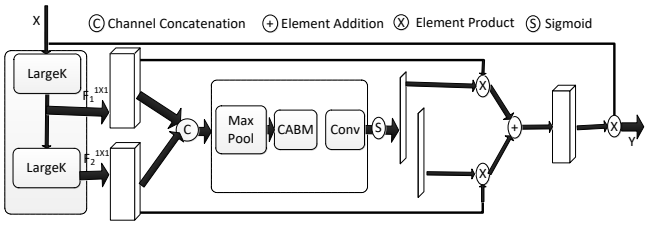


Fig. 8. Schematic diagram of LSKNet module

Convolutional Block Attention Module) is a hybrid attention mechanism that can simultaneously consider the channel and direction information of an image. The CBAM module can effectively improve the efficiency and performance of target detection.

Finally, the receptive field is dynamically adjusted based on adaptive three-way deformable convolution and LSKNet module. The module is a plug and play lightweight component that can be placed in any network for use. This is shown in Figure 9. The adaptive three-way deformable convolution

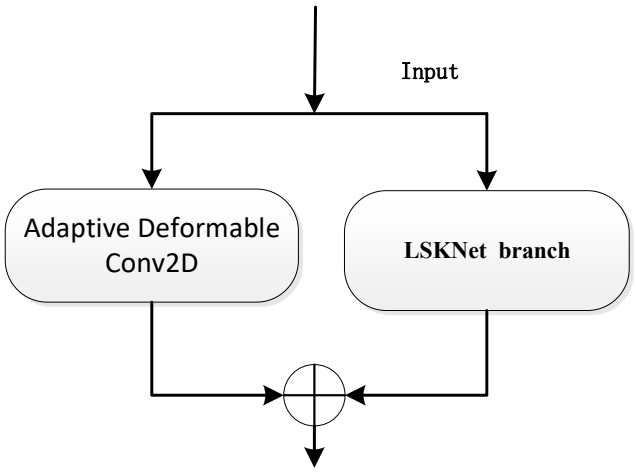


Fig. 9. Adaptive deformable convolution and adaptive LSKNet module

and LSKNet modules can extract more significant features and boundary information, so they can well deal with the multi-scale small target detection problem of remote sensing image.

D. Improved SWIOU loss function

At the same time, in order to locate the small target boundary faster and more accurately, a comprehensive loss function combined in our improved deep neural network was proposed to achieve accurate detection of small targets in remote sensing images.

1) Loss function of classical target detection: The loss function of target detection is shown in Equation 3 [56], [57], [58], [59], [60], [61].

$$
\begin{aligned}
&\text{Loss} = \\
&\lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{obj} \left[\left(x_i - \hat{x}_i^j \right)^2 + \left(y_i - \hat{y}_i^j \right)^2 \right] + \\
&\lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{obj} \left[\left(\sqrt{w_i^j} - \sqrt{\hat{w}_i^j} \right)^2 + \left(\sqrt{h_i^j} - \sqrt{\hat{h}_i^j} \right)^2 \right] - \\
&\sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{obj} \left[\left(\hat{C}_i^j \log \left(C_i^j \right) + \left(1 - \hat{C}_i^j \right) \log \left(1 - \hat{C}_i^j \right) \right] - \\
&\lambda_{\text{noobj}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{\text{noobj}} \left[\left(\hat{C}_i^j \log \left(C_i^j \right) + \left(1 - \hat{C}_i^j \right) \log \left(1 - \hat{C}_i^j \right) \right] - \\
&\sum_{i=0}^{s^2} I_{ij}^{obj} \sum_{j \in class}^{B} \left[\left(\hat{P}_i^j \log \left(P_i^j \right) + \left(1 - \hat{P}_i^j \right) \log \left(1 - \hat{P}_i^j \right) \right] \right.
\end{aligned}
\tag{3}
$$

2) SIoU Loss Function: SIoU further considers the multiple differences between real box and predicted box. For example, angle cost, distance cost, shape cost, IoU cost, etc. We redefined loss function which is related to the bounding box's coordinate, it is as shown in Figure 10.

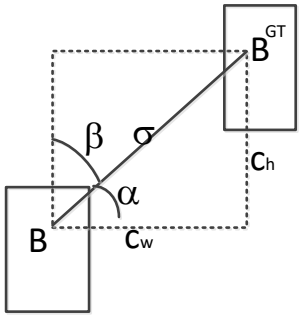


Fig. 10. Vector angle between real box and prediction box

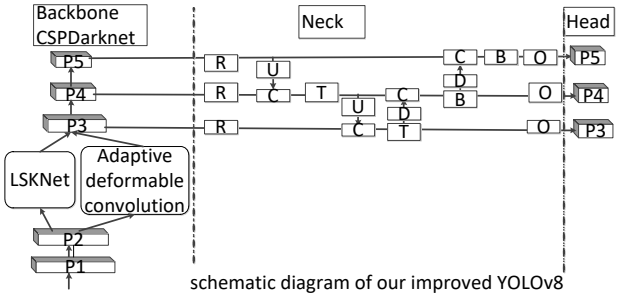


Fig. 11. Schematic diagram of our improved YOLOv8

The relevant parameters in SIoU are shown in formula (4)-(13).

$$\begin{aligned}\Lambda &= 1-2^{*}\sin^{2}\left(\arcsin\left(\frac{c_h}{\sigma}\right)-\frac{\pi}{4}\right)\\ &= \cos\left(2^{*}\left(\arcsin\left(\frac{c_h}{\sigma}\right)-\frac{\pi}{4}\right)\right)\end{aligned} \tag{4}$$

Where

$$\frac{c_h}{\sigma}=\sin(\alpha) \tag{5}$$

$$\sigma=\sqrt{\left(b_{c_x}^{gt}-b_{c_x}\right)^{2}+\left(b_{c_y}^{gt}-b_{c_y}\right)^{2}} \tag{6}$$

$$c_h=\max\left(b_{c_y}^{gt},b_{c_y}\right)-\min\left(b_{c_y}^{gt},b_{c_y}\right) \tag{7}$$

$$\Delta=\sum_{t=x,y}\left(1-e^{-\gamma\rho_t}\right)=2-e^{-\gamma\rho_x}-e^{-\gamma\rho_y} \tag{8}$$

$$\rho_x=\left(\frac{b_{c_x}^{gt}-b_{c_x}}{c_w}\right)^{2} \tag{9}$$

$$\rho_y=\left(\frac{b_{c_y}^{gt}-b_{c_y}}{c_w}\right)^{2} \tag{10}$$

$$r=2-\Lambda \tag{11}$$

$$\Omega=\sum_{t=w,h}\left(1-e^{-w_t}\right)^{\theta}=\left(1-e^{-w_t}\right)^{\theta}+\left(1-e^{-w_h}\right)^{\theta} \tag{12}$$

$$w_W=\frac{w-w^{gt}}{\max\left(w,w^{gt}\right)},w_h=\frac{h-h^{gt}}{\max\left(h,h^{gt}\right)}. \tag{13}$$

It includes four parts: angle cost, distance cost, shape cost and IoU cost. The final definition of SIoU loss function is shown in Equation (14) :

$$\text{Loss}_{\text{SIoU}}=1-IoU+\frac{\Delta+\Omega}{2} \tag{14}$$

3) WIOU loss function: Considering the area between the predicted frame and the actual frame, the loss function is weighted to solve the possible deviation of the evaluation result of the traditional loss function. WIOU can evaluate test results more accurately, and the final definition of WIoU loss function is shown in equation (15) [61], [62], [63] :

$$\text{Loss}_{\text{WToUv3}}=r\,\text{Loss}_{\text{WToUv1}}\left(r=\frac{\beta}{\delta\alpha^{\beta-\delta}}\right) \tag{15}$$

Where:

$$L_{\text{WIOUv1}}=\text{R}_{\text{WIOU}}L\text{IoU} \tag{16}$$

$$\text{R}_{\text{WIou}}=\exp\left(\frac{(x-x_{gt})^{2}+(y-y_{gt})^{2}}{\left(w_g^{2}+h_g^{2}\right)}\right) \tag{17}$$

4) SWIoU loss function: Inspired by SIoU and WIoU [60], [61], [62], [63], this paper proposed an improved loss function SWIoU to optimize the object detection model. SWIoU loss function can establishe the relation between various losses by introducing the information of angle, distance, aspect ratio, loss function weighting, etc., and effectively increased the message interaction between each other. The final SWIoU loss function is shown in equation (18).

$$\text{Loss}_{\text{SWIoU}}=1-IoU+\frac{\Delta+\Omega}{2}\frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{18}$$

III. Algorithm flow

In order to verify the effectiveness of our algorithm, a comparative experiment was conducted using mainstream YOLOv8 target detection algorithms as the baseline. (Note: The improved SWIOU loss function was used in the latest YOLOv8 algorithms.). Inspired by references[64], [65], [66], [67], the improved module which is as shown in Figure 9 was embedded between CSP1_1 and CSP1_2 of YOLOv8; The improved YOLOv8 network is shown in Figure 11.

IV. Experiment

The proposed method was introduced for comparative experiments. The experimental hardware and software configurations, remote sensing image data sets, parameter Settings and performance evaluation indicators adopted are as follows.

A. Software and hardware configuration

For the neural network models in the experiment, SGD with Momentum was used as the optimization algorithm for model training (Momentum=0.9). The learning rate was 0.01; The learning rate decay rate was 0.1. All experiments were completed under the Win10 operating system, the hardware environment was a dual-core 3.4 GHz CPU, 256 GB RAM, two NVIDIA TESLA 100 GPU(32G). And, deep learning framework was Pytorch, programming tools was Pycharm2023. CUDA versions was 10.1. The performance evaluation indexes in this experiment were mAP, loss, FLOPs, etc.

B. Experimental data set

Comparative experiments were conducted on 6 data sets(COCO, VOC, HRRSD, DIOR, Dota, NMPUCHR, ect.)

C. Parameter setting and evaluation index

1) Parameter setting: In the experiment, the adaptive deformable convolution and adaptive LSKNet module(Figure 9) proposed in this paper were embedded into YOLOv8 target detection algorithms. Comparative experiments were conducted on 6 data sets(COCO,VOC, HRRSD, DIOR, Dota, NMPUCHR, ect.).

2) Detailed evaluation index: There are five detailed evaluation indexes, each of which is described below.

(1)mAP

The mAP50 and MAP50-95(average precision) were used in this paper. Accuracy refers to the proportion of records correctly detected using the test set to the total number of classified records, calculated as shown in equation(19) :

$$acc = \frac{TP}{TP + FP} \tag{19}$$

Where TP represents the number of correctly classified records and FP represents the number of incorrectly classified test data. For example, ImageNet has about 1000 categories, and when the model predicts an image, it gives a ranking of 1000 categories from highest probability to lowest probability. Here, the pre-trained model is loaded on the ImageNet dataset. The mAP50 represents the mAP value at 50% loU threshold; The mAP50-95 is a more rigorous evaluation metric that calculates mAP values within the 50%-95% loU threshold range and then takes the average; The mAP50-95 can more accurately evaluate the performance of the model under different LOU thresholds.

(2)Evaluation loss

Evaluation loss is as follow: evaluation of the loss of the target boundary frame (va1/box_loss), evaluate the target object's loss (val/obj_loss), evaluation target object classification loss (val/cls_loss).

(3)Floating-point computation(FLOPs)

The FLOPs parameter directly affects the usage scenario of the target detection model, because the storage of handheld devices is generally not very large(directly affects the computing speed of GPU). Parameter quantification(PARAMs: conv_param = (kernel_size * in_channel + bias)* out_channel; Where, the kernel_size represents the size of the convolution kernel, in_channel is the input channel number, out_ channel is the output channel number, bias is the biased number.). FLOPs is the floating point operation's abbreviations. Therefore, this quantity can be used to measure the computational complexity of an algorithm or model.

(4)training time

training time is the cost of the training target detection model. The unit of measurement for training time is the hour.

(5)Model Memory (M)

It is the memory space that is occupied by the target detection model.

V. Experimental results and analysis

The following two technical points were integrated into YOLOv8 algorithms in our experiment: (1: the adaptive deformable convolution; 2: adaptive LSKNet module; 3: loss function.). Then, a data set was selected as the experimental object for training, evaluation, testing and so on.

A. Experimental results and analysis of YOLOv8 algorithm on COCO dataset

There are more targets in the COCO dataset that occupies a smaller percentage. The feature of image data in COCO data set is very close to the real scene, but there are many kinds of targets and complex background, so it is used to verify the effectiveness of the algorithm. The improved YOLOv8 algorithm had an increase of 2.5 percentage points on mAP50. There is a 0.05 percentage point increase in mAP50:95. The remaining parameters hardly changed, which shown that there was a certain improvement effect on the COCO dataset. The specific comparative data is shown in Table I. The comparisonon mAP50 data on the COCO dataset is shown in Figure 12. Figure 12 shows the mAP50 changes of all targets

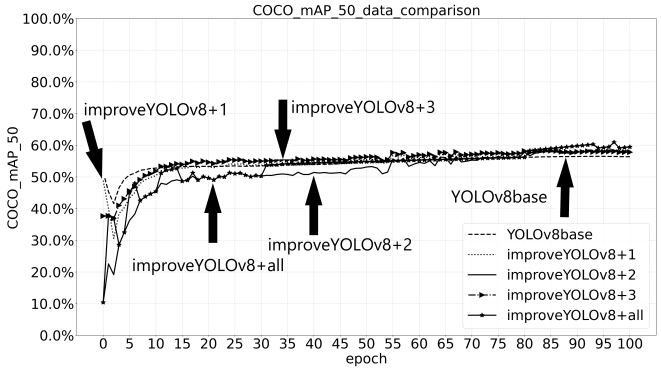


Fig. 12. Comparison mAP50 of COCO dataset

in the COCO dataset before and after the improvement of the three algorithms. The target detection rate is improved and converges stably. The improved algorithm converged more smoothly to the stable loss value. And, the loss value was still high; That is to say, our algorithm will improve the accuracy of target detection if we continuou training.

B. Experimental results and analysis of YOLOv8 algorithm on VOC dataset

There are also more targets in the VOC dataset that occupies a smaller proportion. The image data in VOC data set is very close to the real scene and is the preferred general data set for target detection algorithms. The improved YOLOv8 had increased 2.5 percentage points on mAP50. The remaining parameters hardly changed, which shown a good effect on the VOC data set. The specific comparative data is shown in Table II. The improved YOLOv8 algorithm has an increase of 0.1-3.1 percentage points on mAP50. The remaining parameters (FlOPs, etc.) hardly change, which shows that the detection of small targets on the VOC dataset has a certain effect. The comparisonon mAP50 data on the VOC dataset is shown in Figure 13. As can be seen from the comparison in Figure 13, the improved algorithm can achieve rapid convergence more quickly and stably. The corresponding mAP50 is higher than the original algorithm. the improved algorithm can converge more smoothly to the stable loss value. The overall mAP50 corresponding to most improved algorithms has a certain improvement. The improved algorithm can converge more smoothly to the stable loss value. The whole mAP50 corresponding to the improved algorithm was improved to some extent.

TABLE I
COMPARISON TABLE OF EXPERIMENTAL DATA OF COCO DATASET

algorithm index	mAP50	mAP50-95	val/box_loss	val/obj_loss	val/cls_loss	FLOPs (G)	Time (h)	Model Memory (M)
YOLOv8	0.563830	0.361721	0.042777	0.062211	0.015859	22.44	23.53	14.8
YOLOv8-1	0.584250	0.374402	0.042613	0.0598789	0.0151098	22.44	23.53	14.8
YOLOv8-2	0.577997	0.364875	0.021232	0.057420	0.0130731	22.44	23.53	14.8
improved YOLOv8	0.588325	0.366820	0.058551	0.015367	0.689772	22.44	23.53	14.8

TABLE II
COMPARISON TABLE OF EXPERIMENTAL DATA ON VOC DATASET

algorithm index	mAP50	mAP50-95	val/box_loss	val/obj_loss	val/cls_loss	FLOPs (G)	Time (h)	Model Memory (M)
YOLOv8	0.823151	0.580720	0.024751	0.022938	0.005039	22.4	23.53	13.7
YOLOv8-1	0.857564	0.613959	0.01199	0.032211	0.003621	22.4	23.53	13.7
YOLOv8-2	0.831642	0.560251	0.026748	0.023090	0.004817	22.4	23.53	13.7
improved YOLOv8	0.848762	0.598536	0.021345	0.004227	0.811197	22.4	23.53	13.7

TABLE III
EXPERIMENTAL DATA COMPARISON TABLE OF HRRSD DATASET

algorithm index	mAP50	mAP50-95	val/box_loss	val/obj_loss	val/cls_loss	FLOPs (G)	Time (h)	Model Memory (M)
YOLOv8	0.928590	0.652640	0.017919	0.001505	0.922260	22.40	23.53	19.6
YOLOv8-1	0.956731	0.650155	0.023736	0.018217	0.001482	22.40	23.53	19.6
YOLOv8-2	0.959693	0.672338	0.009827	0.020796	0.001088	22.40	23.53	19.6
improved YOLOv8	0.959414	0.674457	0.016870	0.001338	0.955186	22.40	23.78	19.8

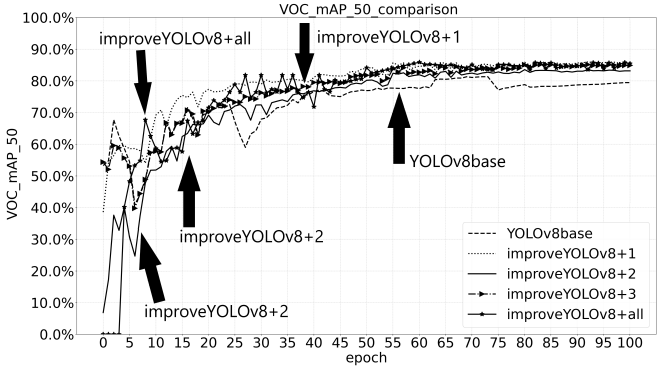


Fig. 13. Comparison mAP50 of VOC dataset

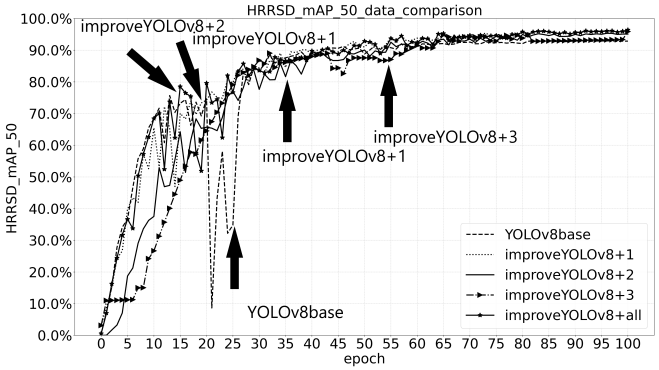


Fig. 14. Comparison mAP50 of HRRSD datasetn

C. Experimental results and analysis of YOLOv8 algorithm on HRRSD dataset

HRRSD is a large-scale dataset with moderate quantity, balanced distribution among classes, more small targets, but the background is relatively simple. The experimental data on the HRRSD dataset is shown in Table III.

The comparisonon mAP50 data on the HRRSD dataset is shown in Figure 14.

The improved YOLOv8 algorithm has an increase of 3.1 percentage points on mAP50. The remaining parameters (FlOPs, etc.) hardly change, which shows that the detection of small targets on the HRRSD dataset has a certain effect. mAP50-95 also has an increase of 0.22 percentage points. That is to say, the improved algorithm can meet higher detection requirements.

D. Experimental results and analysis of YOLOv8 algorithm on DIOR dataset

The DIOR dataset is a recently released dataset of large-scale and high-resolution remote sensing targets. The ground objects in this data set have rich differences in scale, color, texture and other features, which makes the performance of existing detection models in the data set poor. In order to verify the effectiveness of the improved algorithm, we conducted experiments on DIOR remote sensing data set, and the relevant data corresponding to most models is shown in Table IV.

The comparisonon mAP50 data on the DIOR dataset is shown in Figure 15.

Based on the DIOR experimental data, our two technical points were integrated into the YOLOv8 algorithm. Compared to the basic detection algorithm, the mAP50 improved by 2.9% respectively. The corresponding three loss

TABLE IV
COMPARISON TABLE OF EXPERIMENTAL DATA ON DIOR DATASET

algorithm index	mAP50	mAP50-95	val/box_loss	val/obj_loss	val/cls_loss	FLOPs (G)	Time (h)	Model Memory (M)
YOLOv8	0.788320	0.572010	0.019148	0.001264	0.866980	22.40	23.53	14.8
YOLOv8-1	0.813759	0.583668	0.014548	0.019297	0.001110	22.40	23.53	14.8
YOLOv8-2	0.814296	0.589214	0.025268	0.018767	0.001164	22.40	23.53	14.8
improved YOLOv8	0.817666	0.589539	0.018979	0.001227	0.891142	22.40	23.58	15.5

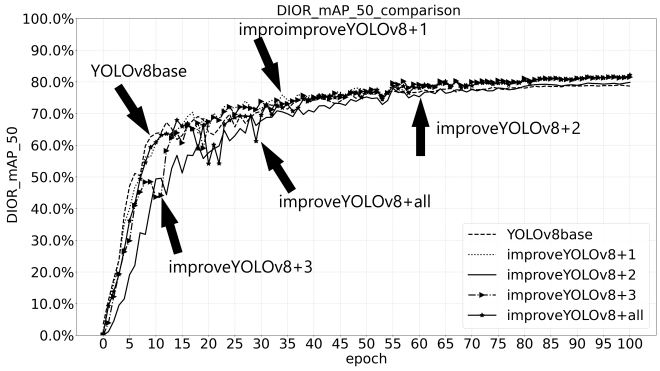


Fig. 15. Comparison table of experimental data on DIOR dataset

values still have a certain downward trend. The remaining parameters (FLOPs, training time, model memory) barely increased (i.e. no increase in detection costs). Neither the training time nor the model increased.

E. Experimental results and analysis of YOLOv8 algorithm on DOTA dataset

DOTA is a large-scale data set with both vertical bounding box and rotating bounding box annotations. It contains 16 object categories and 400,000 object instances. Some annotation information has been added to significantly increase the number of target instances. And, a new target class(container cranes) has been added. The experimental-related indicators of DOTA data set is shown in Table V. Based on DOTA experimental data, our three technical points were integrated into the YOLOv8 algorithm. Compared with the basic detection algorithm, mAP50 improved by 2.7

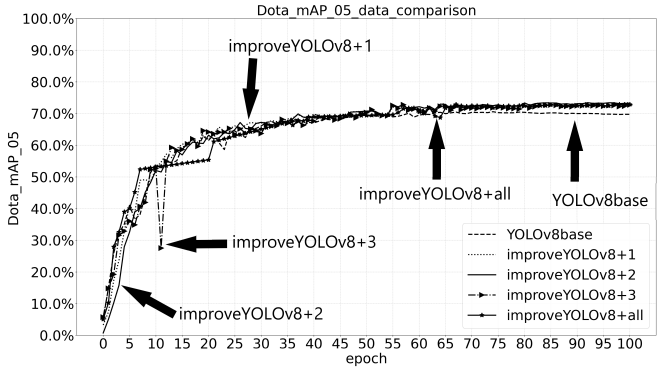


Fig. 16. Comparison table of experimental data on DOTA dataset

According to DOTA experimental data, our 3 technical points were integrated into the YOLOv8 algorithm.

Compared with the basic detection algorithm, the mAP50 improved by 2.7% respectively. At the same time, the mAP50:95 increased by 0.03% respectively. The corresponding three loss values still have a certain downward trend. The remaining parameters (FLOPs, training time, model memory) barely increased (i.e. no increase in detection costs), and neither the training time nor the model increased.

F. Experimental results and analysis of YOLOv8 algorithm on NWPUCHR dataset

The NWPUCHR dataset contains 10 types of objects, and the ground objects in this dataset have rich differences in scale, color, texture and other features. And, fully considered the imaging conditions of real scenes, clouds and other factors, which makes the performance of these mainstream models poor. In order to more fairly reflect the effectiveness of the proposed algorithm, we conducted experiments on the NWPUCHR dataset, and the main performance indicators is shown in Table VI. It is again verified that the proposed multi-scale fusion module pays more attention to the boundary information of the target. The comparisonon mAP50 data on the NWPUCHR dataset is shown in Figure 17.

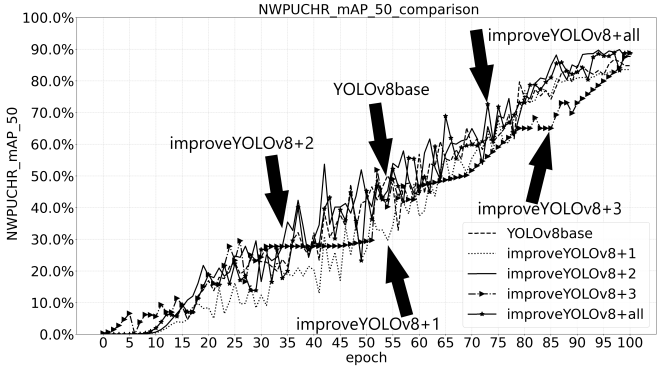


Fig. 17. Comparison table of experimental data on NWPUCHR dataset

On the NWPUCHR dataset, these algorithms were applied for experiments, and, the mAP50 increased by 4.2%. The remaining parameters (FLOPs, training time, model memory) barely increased (i.e. no increase in detection costs). Neither the training time nor the model increased.

VI. CONCLUSION

In order to detect small objects on remote sensing images more accurately, adaptive deforming convolution and LSKNet with automatic adjusted receptive field strategies were proposed to integrate multi-scale features. The improved SWIOU loss function was used to optimize the optimization direction in the small target detection phase. The experimental results shown that the mAP50 of detecting small targets on 4 remote sensing data sets(such as HRRSD)

TABLE V
EXPERIMENTAL DATA COMPARISON TABLE ON DOTA DATASET

algorithm index	mAP50	mAP50-95	val/box_loss	val/obj_loss	val/cls_loss	FLOPs (G)	Time (h)	Model Memory (M)
YOLOv8	0.715793	0.546474	0.020446	0.001562	0.878559	22.4	23.53	14.8
YOLOv8-1	0.725582	0.452423	0.039581	0.034781	0.002051	22.4	23.53	14.8
YOLOv8-2	0.723549	0.458144	0.038489	0.034586	0.001737	22.4	23.53	14.8
improved YOLOv8	0.742296	0.549214	0.018767	0.001164	0.895166	22.4	23.53	14.8

TABLE VI
EXPERIMENTAL DATA COMPARISON TABLE ON NWPUCHR DATASET

algorithm index	mAP50	mAP50-95	val/box_loss	val/obj_loss	val/cls_loss	FLOPs (G)	Time (h)	Model Memory (M)
YOLOv8	0.848060	0.443712	0.042503	0.024823	0.006742	22.40	23.53	19.6
YOLOv8-1	0.876221	0.446064	0.046343	0.039504	0.006114	22.40	23.53	19.6
YOLOv8-2	0.887562	0.490955	0.041986	0.036767	0.007458	22.40	23.54	19.6
improved YOLOv8	0.890110	0.507170	0.038009	0.007719	0.892850	22.40	23.53	19.8

were improved by 3 to 5 percentage points. The mAP50:95 were improved by 1 to 3 percentage points. The remaining parameters hardly changed, which shown that there was a certain improvement effect on 4 remote sensing data sets.

REFERENCES

[1] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," *2009 IEEE 12th International Conference on Computer Vision*, 2019.

[2] J. M. J. Viola P A, "Rapid object detection using a boosted cascade of simple features," *Computer Vision and Pattern Recognition*, 2001.

[3] J. F. S. C. Harzallah, H., "Combining ecient object localization & image classication," *2009 IEEE 12th International Conference on Computer Vision*, p. 237244, 2009.

[4] T. B. Dalal N, "Histograms of oriented gradients for human detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 886–893.

[5] J. M. J. Viola P, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[6] D. Lowe, "Object recognition from local scale-invariant features (sift)," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 1, no. 2, 2004, p. 11501157.

[7] M. Lienhart, R., "Speeded-up robust features (surf) original publication," in *Computer vision and image understanding*, 2002.

[8] T. Bay H, Ess A, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 110, no. 3, 2008, pp. 346–359.

[9] O. E. Hearst M A, Dumais S T, "Support vector machines," in *IEEE Intelligent Systems and Their Applications*, vol. 13, no. 4. IEEE, 1998, pp. 18–28.

[10] M. R. Opitz D, "Popular ensemble methods: An empirical study," *Journal of Artificial Intelligence Research*, vol. 11, no. 1, pp. 169–198, 1999.

[11] F. Y, "Experiment with a new boosting algorithm," in *Morgan Kaufmann*.

[12] M. R. Opitz D, "Popular ensemble methods: An empirical study," *Journal of Artificial Intelligence Research*, vol. 11, no. 1, pp. 169–198, 2019.

[13] T. B. Dalal N, "Histograms of oriented gradients for human detection," *IEEE Computer Society Conference on Computer Vision & Pattern Recognition*, pp. 886–893, 2005.

[14] B. W. Hazgui M, Ghazouani H, "Data augmentation for genetic programming-driven late merging of hog and uniform lbp features for texture classification," *Vietnam Journal of Computer Science*, vol. 11, no. 02, pp. 211–239, 2024.

[15] P. Ren, W. Fang, and S. Djahel, "Distinctive image features from scale-invariant key points," in *International Journal of Computer Vision*, vol. 20, 2003, pp. 91–110.

[16] H. S. C. H. Wu X, Sahoo D, "Recent advances in deep learning for object detection," in *Neurocomputing*, 2020, pp. 396–405.

[17] S. H. Rosenberg C, Hebert M, "Semi- supervised self- training of object detection models," *IEEE Workshop on Applications of Computer Vision*, 2005.

[18] K. C. Itti L, "Training deformable object models for human detection based on alignment and clustering," *European Conference on Computer Vision*.

[19] B. T. Drayer B, "Training deformable object models for human detection based on alignment and clustering," *Computer Vision ECCV 2014: 13th European Conference*, pp. 406–420, 2014.

[20] G. T. Uijlings J R R, Van De Sande K E A, "Selective search for object recognition," in *International Journal of Computer Vision*, vol. 104, no. 2, 2013, pp. 154–171.

[21] H. G. E. Krizhevsky A, Sutskever I, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[22] Y. X. Z. Zhang, "The kfiou loss for rotated object detection international conference on learning representations," in *Computer Vision and Pattern Recognition*, 2022, pp. 54–62.

[23] J. Y. Szegedy C, Liu W, "Going deeper with convolutions," in *IEEE Computer Society*, 2015, pp. 1–9.

[24] C. J. M. Senthilkumar S, Brindha K, "An optimized handwritten polynomial equations solver using an enhanced inception v4 model," *Multimedia Tools and Applications*, vol. 83, no. 17, pp. 50 691–50 710, 2024.

[25] P. F. Sara M, "Complexity loss in physiological time series of patients in a vegetative state," in *Nonlinear Dynamics Psychol Life*, 2010, pp. 1–13.

[26] G. R. Ren S, He K, "Faster rcnn: Towards realtime object detection with region proposal networks," in *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, 2017, pp. 1137–1149.

[27] D. T. Girshick R, Donahue J, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[28] R. S. He K, Zhang X, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, 2015, pp. 1904–1916.

[29] G. R, "Fast rcnn," in *Computer Science*, 2015, pp. 1440–1448.

[30] W. Y. Liu Y, "Synthetic aperture radar image target recognition based on improved fusion of r-fcn and src," *Proceedings of the 4th International Conference on Computer Science and Software Engineering Pageg*, pp. 53–60, 2021.

[31] W. Y. Wen W, Wu C, "Learning structured sparsity in deep neural networks," *Cornell University*, 2016.

[32] D. P. He K, Gkioxari G, "Mask rcnn. ieee international conference on computer vision," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017.

[33] N. H. Couturier R, Gregori P, "A deep learning object detection method to improve cluster analysis of two-dimensional data," *Multimedia Tools and Applications*, vol. 83, no. 28, pp. 71 171–71 187, 2024.

[34] G. R. Redmon J, Divvala S, "You only look once: unified, real time object detection," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

[35] Z. H. Chen C, Wu B, "An image recognition technology based on deformable and cbam convolution resnet50," *IAENG International Journal of Computer Science*, vol. 50, no. 1, pp. 274–281, 2023.

[36] T. S. Madanan M, Muthukumaran N, "Rsa based improved yolov3 network for segmentation and detection of weed species," *Satya*

Nadella Multimedia Tools and Applications, vol. 83, no. 12, 34913-34942.

[37] L. M. C. H. Adiono T, Ramadhan R M, "Fast and scalable multicore yolov3-tiny accelerator using input stationary systolic architecture," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, pp. 1774–1787, 2023.

[38] R. P. K. M. Adarsh, P., "Yolo v3-tiny: Object detection and recognition using one stage improved model," *In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems, Coimbatore, India*, pp. 687–694, 2020.

[39] W. Y. Xu D, "Improved yolo-v3 with densenet for multi-scale remote sensing target detection," *Sensors*, vol. 20, pp. 4276–4289, 2020.

[40] L. J. Yang T, "Remote sensing image object detection based on improved yolov3 in deep learning environment," *Journal of circuits Systems and Computers*, vol. 32, no. 15, pp. 1–13, 2023.

[41] X. L. Li J, Xu Z, "Vehicle and pedestrian detection method based on improved yolov4-tiny," *Optoelectronics Letters*, vol. 19, no. 10, pp. 623–628, 2023.

[42] K. D. Arulalan V, "Efficient object detection and classification approach using htyolov4," *Computer Systems Science and Engineering (English)*, vol. 44, no. 2, pp. 1703–1717, 2023.

[43] M. S. N. Patil S M, Pawar S D, "Yolov4-based hybrid feature enhancement network with robust object detection under adverse weather conditions," *Signal, Image and Video Processing:*, vol. 18, no. 5, pp. 4243–4258, 2024.

[44] K. J. K. Clayton S, "Active serial port: A component for jcspnet embedded systems," *Communicating Process Architectures, 2010:*, pp. 85–98, 2010.

[45] C. Y. Wang J, "Improved yolov5 network for real-time multi-scale traffic sign detection," *Neural Computing and Applications*, pp. 1–13, 2022.

[46] Q. W. Z. Shanshan W, Weiwei T, "High-voltage transmission line foreign object and power component defect detection based on improved yolov5," *Journal of Electrical Engineering & Technology:*, vol. 19, no. 1, pp. 851–866, 2024.

[47] E. D. Liu W, Anguelov D, "Ssd: Single shot multibox detector," *European Conference on Computer Vision, 2016,*, pp. 21–37, 2016.

[48] N. K. Takashima T, Mitani T, "Double-side silicon strip detector (dssd) with va32ta applied for medium energy particle detector in high-count rate environment," *IEEE Transactions on Nuclear Science*, vol. 51, no. 5, pp. 2004–2007, 2004.

[49] S. A. Plisiecki H, "Extrapolation of affective norms using transformer-based neural networks and its application to experimental stimuli selection," *Behavior Research Methods*, vol. 56, no. 5, pp. 4716–4731, 2024.

[50] W. X. X. Sun E, Zhou D, "Transformer-based few-shot object detection in traffic scenarios," *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, vol. 54, no. 1, pp. 947–958, 2024.

[51] S. A. Plisiecki H, "Extrapolation of affective norms using transformer-based neural networks and its application to experimental stimuli selection," *Behavior Research Methods*, vol. 56, no. 5, pp. 4716–4731, 2024.

[52] L. Y. Xin W, Liu R, "Transformer for skeleton-based action recognition: A review of recent advances," *Neurocomputing*, no. 537, pp. 164–186, 2023.

[53] L. Qishuo, "Research on object detection methods based on deep learning," *Beijing University Of Posts and Telecommunications*, 2020.

[54] W. X. Ouyang W, Zeng X, "Deepid-net: Deformable deep convolutional neural networks for object detection," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 1–10, 2016.

[55] Z. Z. Li Y, Hou Q, "Large selective kernel network for remote sensing object detection," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16 794–16 805, 2023.

[56] G. J. Y. Rezatofighi H, Tsoi N, "Generalized intersection over union: A metric and a loss for bounding box regression," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 658–666, 2019.

[57] L. W. Zheng Z, Wang P, "Ssd: Single shot multibox detectordistance-iou loss: Faster and better learning for bounding box regression," *Computer Vision ECCV 2016: 14th European Conference*, pp. 21–37, 2016.

[58] M. B. L. C. Li H, Zhou Q, "Alpha-sganet: A multi-attention-scale feature pyramid network combined with lightweight network based on alpha-iou loss," *Plos One*, vol. 17, no. 10, 2022.

[59] M. X. He J, Erfani S, "Alpha-iou: A family of power intersection over union losses for bounding box regression," *Cornell University*, 2021.

[60] Z. Gevorgyan, "Siou loss: More powerful learning for bounding box regression," *Cornell University*, 2022.

[61] K. R. Hussain M, "In-depth review of yolov1 to yolov10 variants for enhanced photovoltaic defect detection," *Agronomy*, vol. 14, no. 8, pp. 351–386, 2024.

[62] S. Y. Feng W, Liu M, "The use of a blueberry ripeness detection model in dense occlusion scenarios based on the improved yolov9," *Agronomy*, vol. 14, no. 8, pp. 1860–1872, 2024.

[63] Cho.Y.J, "Weighted intersection over union (wiou) for evaluating image segmentation," *Pattern Recognition Letters*, pp. 101–107, 2024.

[64] A. D. Sapkota R, Meng Z, "Comprehensive performance evaluation of yolov10, yolov9 and yolov8 on detecting and counting fruitlet in complex orchard environments," *Cornell University*, 2022.

[65] H. M. Sundaresan Geetha A, Alif M A R, "Comparative analysis of yolov8 and yolov10 in vehicle detection: Performance metrics and model efficacy," *Vehicles*, vol. 6, no. 3, pp. 1364–1382, 2024.

[66] L. J, "Ld-yolov10: A lightweight target detection algorithm for drone scenarios based on yolov10," *Electronics*, vol. 13, no. 16, pp. 3269–3275, 2024.

[67] S. Y. Chen C, Wu B, "An improved yolov5 algorithm for detecting target," *IAENG International Journal of Computer Science*, vol. 51, no. 10, pp. 1454–1461, 2024.

Chao Chen is an associate professor from Key Laboratory of Numerical Simulation in Sichuan University, Neijiang Normal University, Neijiang, Hongqiao Street 1, P. R. China. He has published 12 core papers, 5 invention patents, and has applied for 25 computer software copyrights.

Bin Wu is a professor and doctoral supervisor from Southwest University of Science and Technology; He is also a part-time doctoral supervisor of China Academy of Engineering Physics, an outstanding expert with outstanding contributions in Sichuan province, and an academic and technical leader in Sichuan province.