Based on YOLOv7-tiny Improved Model of Remote Sensing Image Detection

2023 16th International Conference on Advanced Computer Theory and Engineering (ICACTE) | 979-8-3503-1753-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICACTE59887.2023.10335430

Yue Wang
School of Computer and Information Engineering
Fuyang Normal University
Anhui, China
572363120@qq.com

Huaimeng Xiao
School of Computer and Information Engineering
Fuyang Normal University
Anhui, China
1656830222@qq.com

Ya Wang*
School of Computer and Information Engineering
Fuyang Normal University
Anhui, China
fync_wy80@163.com
*corresponding author

Jun Huang
School of Computer and Information Engineering
Fuyang Normal University
Anhui, China
jjun1110@163.com

***Abstract*— Object detection technology has become mature, but it is still challenging for remote sensing image detection. For example, the target size is small and difficult to separate from the surrounding background. The targets are distributed sparsely and unevenly, and the dense targets have occlusions, which makes it difficult to detect by the model. To solve this problem, this paper proposes an improved detection model based on YOLOv7-tiny. For the detection of small targets in data set, coordinate attention mechanism and loss function fusion method are used to improve the accuracy of detection of small targets. Combined with the idea of feature separation and fusion, the C5 module of YOLOv7-tiny model is improved to reduce the feature loss in the training process and improve the reasoning speed of the model. Compared with the original network model, the performance of the improved model is significantly improved, with mAP@0.5 reaching 0.935 and FPS reaching 141.13.**

Keywords—remote sensing image, real-time, attention mechanism, loss function

I. INTRODUCTION

In recent years, with the development of remote sensing technology, object detection of remote sensing image has attracted wide attention. Its study is of great significance for resource exploration, natural disaster assessment, military object detection and recognition[1].

Remote sensing image is different from ordinary image, and its particularity is mainly manifested in the following aspects [2,3].

Similar targets vary in size, such as the port of the ship as large as more than 300 meters, but only tens of meters. The image perspective is basically aerial overlooking. The detector trained well on conventional data sets may have poor effect on aerial remote sensing images. Most of the targets are small targets. Small targets have small amount of information and are difficult to detect. The image field of vision is relatively large and the background is complex, which will produce strong interference to the object detection.

At present, object detection algorithms based on deep learning are mainly divided into two categories: one-stage model and two-stage model. In the single-stage model, only one convolutional neural network is used to locate and classify all objects directly on the whole image, and the generation of candidate regions is ignored. The two-stage model first generates a series of potential target candidate regions on the image, and then carries out classification and boundary regression for each candidate region in turn. Common object detection algorithms include Fast R-CNN, Faster R-CNN and YOLO series.

Fast R-CNN[4] is a two-stage model, which integrates feature extraction, classification and regression into one step. In this way, it no longer needs to save feature vectors in the middle and solves the problem of storage space. And in the training can be the overall optimization, so can achieve higher accuracy. However, the selective search algorithm used is based on the underlying visual features of the image to directly generate candidate regions, which cannot be learned according to specific data sets and is very time-consuming.

On the basis of Fast R-CNN, Ren et al.[5] designed a Faster R-CNN network generated based on RPN candidate box. The input of Faster R-CNN is the feature map of the whole image extracted from the existing backbone network of Fast R-CNN. This sharing feature design not only makes full use of the feature extraction capability of CNN, but also saves calculation. Secondly, the concept of Anchor is proposed. RPN conducts classification (foreground or background) and regression based on anchor points with preset sizes, which not only ensures the generation of multi-scale candidate boxes, but also makes the model easier to converge. Compared with Fast R-CNN, Faster R-CNN has a great improvement in accuracy and speed, but it still cannot meet the requirement of real-time performance.

YOLO series algorithm is a typical single-stage model algorithm. In 2016, YOLOv1[6] used CNN convolutional neural network for feature extraction to identify target type and location. Although the detection speed is fast, the accuracy is significantly reduced. Subsequently, a series of algorithms about YOLO appeared one after another, and the detection speed and accuracy were further improved. In 2022, YOLOv7[7] introduced structural reparameterization. It effectively improves the detection efficiency of the algorithm by means of efficient long range aggregation network (ELAN), cascaded model scaling

strategy and training strategy combining auxiliary head and lead head, etc. The core of YOLO is fast detection, not high detection accuracy.

Aiming at the above problems, this paper proposes an improved model based on yolov7-tiny. By improving the C5 module, adding attention mechanism, and introducing a normalized wasserstein distance (NWD)[8] loss function, it improves the real-time and accuracy of remote sensing image object detection.

II. Related Work

A. YOLOv7-tiny Network Model

In order to run codes in different environments, Wang et al. [6] designed three models, Yolov7-Tiny, YOLOv7 and Yolov7-W6, to run on edge GPU, ordinary GPU and cloud GPU respectively. Considering the existing hardware facilities, this paper finally choses YOLOv7-tiny as the benchmark model for this experiment.

The YOLOv7-tiny network model mainly includes three parts: Input, Backbone and Head. First, after a series of preprocessing such as data enhancement in the input part, the image is sent to the backbone network. The backbone network extracts features from the processed image and down-samples the image 32 times. Subsequently, the extracted features are processed by feature fusion to obtain features of three sizes: large, medium and small. Then, the final output result is tested. Its network architecture is shown in Fig. 1.

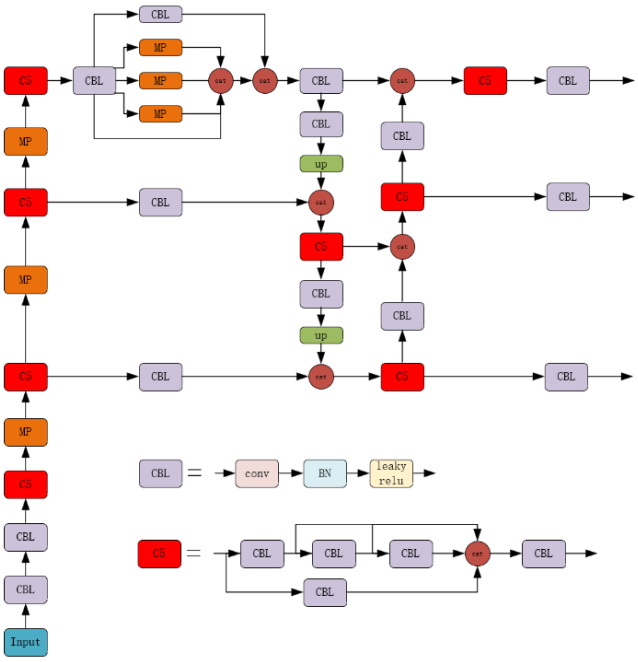


Fig. 1. YOLOv7-tiny network structure.

B. Attention Mechanism

The attention model[9] was originally used in machine translation and has now become an important concept in the field of neural networks. In the field of artificial intelligence, attention has become an important part of the structure of neural networks, and has a large number of applications in natural language processing, statistical learning, speech and computer.

The attention mechanism can be used to emphasize or select important information about the object and suppress some irrelevant details, which has a great effect on object detection.

This paper conducts experiments on various attention mechanisms, and finally found that the coordinate attention mechanism can extract image features better without increasing the model size.

C. Loss Function

The function of IoU[10] loss function is to make the prediction box close to the correct target to improve the positioning effect. The YOLOv7-tiny network uses CIoU[11] loss function, which takes into account geometric factors such as overlapping area, center distance and aspect ratio. As shown in equation (1).

$$LOSS_{CIoU} = 1 - CIoU = 1 - (IoU - \frac{d_o^2}{d_c^2} - \frac{v^2}{1-IoU+v}) \quad (1)$$

Where, d_0 is the Euclidean distance between the center points of the target frame and the prediction frame, d_c is the diagonal distance of the target frame, and v is the parameter measuring the consistency of the aspect ratio, as shown in equation (2).

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

Where, w^{gt}, h^{gt} are the width and height of the ground truth(gt), w^p, h^p are the width and height of the prediction frame.

D. Evaluation Index

The performance metrics the paper used this time are model weight, mAP@0.5, and frames per second (FPS). The calculation of mAP is shown in equation (3).

$$mAP = \frac{\sum_{n=1}^{N}\int_0^1 P(R)dr}{N} \quad (3)$$

Among them: $P = \frac{TP}{TP+FP}$ $R = \frac{TP}{TP+FN}$.

mAP@0.5 indicates the mAP whose IoU threshold is 0.5. True Positive(TP) indicates that the positive category is predicted as positive, that is, the prediction is correct, the true value is 0, and the prediction is also 0. False Negative(FN) indicates that the positive class is predicted as negative, that is, the false prediction is 0 and the prediction is 1. False Positive (FP) indicates that the negative prediction to positive, that is, the false prediction, the true prediction is 1, and the prediction is 0. True Negative (TN) indicates that the negative class is predicted as negative, that is, the prediction is correct, and the true value is 1.

In the P-R curve, the area enclosed by the P-R curve and the coordinate axis is equal to the size of the AP value. The mAP can be obtained by averaging the AP values of all categories.

III. MODEL IMPROVEMENT

A. C2f Module

The YOLOv7-tiny network model is mainly stacked by C5 modules, in which C5 contains 5 convolution (CBL) and 3 shortcut connections. The structure of C5 module is shown in Fig. 1. According to ShuffleNetv2 proposed by Ma et al. [12], it can be known that excessive shortcut and convolution operations will increase the number of memory accesses and reduce the speed of model inference. Therefore, the paper uses C2f module instead of C5 module. C2f contains 4 convolution and 3 shortcut connections. Its structure is shown in Fig. 2.

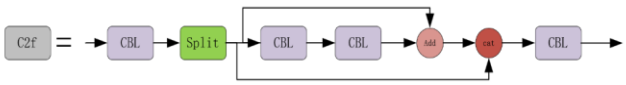


Fig. 2. C2f module structure.

First of all, the input features will undergo a convolution, and then the obtained feature maps will be divided into two parts. In one part, 3x3 convolution will be used to reduce dimension and eliminate redundant information. Then 3×3 convolution is used to raise dimension, and then it is added to the features before dimension reduction to enhance the feature performance. Then, it is spliced with another feature map, which can be regarded as feature reuse and reduce the loss of features. The above operation can reduce the training parameters and improve the model reasoning speed. Finally, the feature is extracted and output after another convolution.

Compared with C5 module, C2f can not only reduce the number of parameters in the model, but also improve the reasoning speed and accuracy of the model. There are three main reasons.

Reduce the degree of network fragmentation (such as parallel operation) and reduce the number of memory access. In each block, half of the feature channels (when c' = c/2) directly pass through the block and join the next block, which can be a feature reuse, reduce the loss of feature information, and reduce the number of training parameters. Reduce element-by-element operation to improve network reasoning speed.

In order to further explore how to use C2f module to maximize the detection effect, the paper replaces C5 module in YOLOv7-tiny network structure with C2f module, which can be divided into the following three cases. Case 1: Only the C5 module in Backbone is replaced by the C2f module. Case 2: Only replace the C5 module in the Head structure with the C2f module. Case 3: Replace all C5 modules in the network with C2f modules.

Experiments were carried out in view of the above situations, and the experimental results are shown in Table 1.

TABLE I. COMPARISON OF EXPERIMENTAL RESULTS UNDER DIFFERENT CONDITIONS

Model	Model Size (MB)	mAP@0.5	FPS
Case 1	11.0	0.903	120.31
Case 2	10.9	0.898	112.56
Case 3	**11.2**	**0.913**	**142.85**

It can be seen from Table 1 that when all C5 modules in the network are replaced by C2f modules, the model performance can be improved better, thus reflecting the effectiveness of C2f modules.

B. Selection of Attention Mechanism

In order to further obtain the information of feature map, this paper conducts experiments on several attention mechanisms, and the experimental results are shown in Table 2.

TABLE II. EXPERIMENTAL RESULTS OF DIFFERENT ATTENTION MECHANISMS

Attention Mechanism	Model Size (MB)	mAP@0.5	FPS
SE	11.5	0.815	105.63
ECA	11.3	0.707	74.50
CoordAtt	**11.2**	**0.925**	**140.32**
ShuffleAtt	11.7	0.656	113.81
Sim	13.4	0.573	64.30

Compared with other attention mechanisms,the coordinate attention mechanism (CoordAtt) [13] has a good performance of mAP@0.5 of 0.925 and FPS of 140.32. Therefore, the paper finally chooses CoordAtt. CoordAtt can capture not only cross-channel information, but also direction-sensing and position-sensitive information, which helps the model locate and identify the object of interest more accurately. Its structure is shown in Fig. 3.

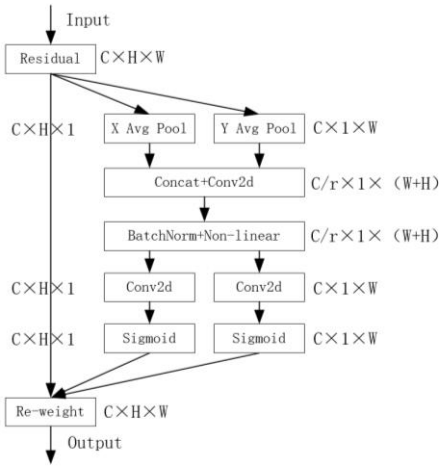


Fig. 3. Structure diagram of coordinate attention mechanism (CoordAtt).

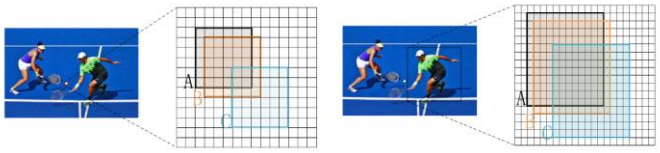
By embedding location information into the channel attention mechanism, CoordAtt enables mobile networks to focus on location over large areas without incurring significant computational overhead. In order to alleviate the loss of location information caused by global pooling, channel attention was decomposed into two parallel one-dimensional feature coding processes, which effectively integrated the spatial coordinate information and put it into the generated feature map. Specifically, CoordAtt uses two one-dimensional global pooling operations to aggregate input features into two independent direction-aware feature maps along the vertical and horizontal directions, respectively. Then, the two feature graphs embedded with specific directional information are encoded into two attention graphs respectively, each of which captures the long-term dependence relationship along a spatial direction in the

input feature graph. Location information can therefore be stored in the generated feature map. Then, the two attentions are mapped to the input feature graph to enhance the feature representation.

C. Improvement of Loss Function

IoU[9] is the most widely used metric for measuring the similarity between boundary frames. However, IoU can only take effect if there is overlap in bounding boxes. To solve this problem, Zhou et al. [14] proposed GIoU. However, when one bounding box contains another bounding box, GIoU is demoted to IoU. Therefore, YOLOv7 overcomes the limitations of IOU and GIOU by using DIoU[15] and CIoU[11], considering the geometric properties of overlapping area, center point distance and aspect ratio. GIoU, CIoU, and DIoU are mainly used to replace IOU in NMS and loss functions to improve the performance of object detection models.

However, as shown in Fig. 4, people can observe that IoU's sensitivity to objects of different scales varies greatly. Specifically, for tiny objects, a small positional bias will cause a significant drop in IoU (from 0.3 to 0.059), resulting in inaccurate label assignment.



$$IoU = \frac{|A \cap B|}{|A \cup B|} = 0.532 \rightarrow IoU = \frac{|A \cap C|}{|A \cup C|} = 0.059$$
$$IoU = \frac{|A \cap B|}{|A \cup B|} = 0.733 \rightarrow IoU = \frac{|A \cap C|}{|A \cup C|} = 0.3$$

Fig. 4. Sensitivity analysis of IoU to small and normal scale objects.

Where, each grid represents A pixel, box A represents the ground truth boundary box, box B and C represent the prediction boundary box with 1 pixel and 4 pixel diagonal deviation respectively.

For small targets, it is difficult to match a high-quality anchor to ground truth. A simple way to do this is to lower the threshold of IoU, which increases the number of positive samples for small targets, but also decreases the overall quality. At present, some works want to make tag allocation more adaptive, such as adaptive training sample selection (ATSS) [16], The probabilistic anchor assignment (PAA) [17] and the optimal transport assignment (OTA) [18]. However, these methods are still based on IoU.

Wang et al. [8] designed a new evaluation index that can better measure the similarity between tiny objects. The principle is as follows.

For small objects, there tend to be a few background pixels in their bounding box, as most real objects are not strictly rectangular. In these bounding boxes, foreground pixels and background pixels are concentrated at the center and boundary of the bounding box. In order to better describe the weights of different pixels in bounding box, it can be modeled as two-dimensional Gaussian distribution, where the center pixel has the highest weight and the importance of pixels gradually decreases from center to boundary. Concretely, for horizontal bounding box $R = (cx, cy, w, h)$, where (cx, cy), w and h represent center coordinates, width and height respectively. The equation of its intrinsic ellipse is shown in equation (4).

$$\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} = 1 \tag{4}$$

Where the central coordinate of the (μ_x, μ_y) ellipse, σ_x, σ_y is the half-axis length along the x and y axes. So $\mu_x = cx$, $\mu_y = cy$, $\sigma_x = \frac{w}{2}$, $\sigma_y = \frac{h}{2}$.

The probability density function of two-dimensional Gaussian distribution is shown in equation (5).

$$f(x|\mu, \boldsymbol{\Sigma}) = \frac{exp\left(-\frac{1}{2}(x-\mu)^T \boldsymbol{\Sigma}^{-1}(x-\mu)\right)}{2\pi|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \tag{5}$$

Where x, μ, and Σ represent coordinates (x, y), mean vectors, and covariance matrices of Gaussian distributions. When $(x-\mu)^T \boldsymbol{\Sigma}^{-1}(x-\mu) = 1$.

The ellipse in equation (5) will be the density contour of two-dimensional Gaussian distribution. Therefore, the horizontal boundary frame $R = (cx, cy, w, h)$ can be modeled as a two-dimensional Gaussian distribution $N(\mu, \varepsilon)$, where $\mu = \begin{pmatrix} cx \\ cy \end{pmatrix}$, $\boldsymbol{\Sigma} = \begin{pmatrix} \frac{w^2}{4} & 0 \\ 0 & \frac{h^2}{4} \end{pmatrix}$.

In addition, the similarity between boundary frames A and B can be translated into the distribution distance between the two Gaussian distributions.

The distribution distance is calculated according to the Wasserstein distance of optimal transport theory. Second-order Wasserstein distance is shown in equation (6).

$$W_2^2(N_a, N_b) = \left\| \left(\left[cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2} \right]^T, \left[cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2} \right]^T \right) \right\|_2^2 \tag{6}$$

Where, $W_2^2(N_a, N_b)$ is expressed as a distance metric. (cx, cy), w and h represent center coordinates, width and height respectively, $\| \cdot \|$ represents Frobenius norm.

Finally, normalization of its exponential form is used to obtain a new metric, called Normalized Wasserstein Distance (NWD), as shown in equation (7).

$$NWD(N_a, N_b) = exp\left(\frac{\sqrt{W_2^2(N_a, N_b)}}{C} \right) \tag{7}$$

The IoU loss can reduce the gap between training and testing, but the IOU-based loss function does not provide the gradient to optimize the network in the following cases, such as no overlap and gt completely includes prediction boxes or vice versa.

When it comes to small goals, both of these things happen. While DIoU and CIoU can handle both cases, they are too sensitive to changes in position to detect small targets. Therefore, the introduce a loss function based on NWD, as shown in equation (8).

$$Loss_{NWD} = 1 - NWD(N_p,\ N_g) \tag{8}$$

Where, N_p is the Gaussian distribution of the prediction box, and N_g is the Gaussian distribution of the gt box.

The paper combine the loss function based on NWD and the CIoU loss function. On the one hand, the method can reduce the sensitivity of IoU to the position deviation of small objects and improve the accuracy of the model to detect small objects. On the other hand, the method can maintain the accuracy of the original model to detect normal-size objects.

In order to further explore the effect of the improved loss function on the improvement of object detection, this paper conducts experiments on loss functions of different proportions, and the experimental results are shown in Table 3.

TABLE III. EXPERIMENTAL RESULTS OF LOSS FUNCTIONS OF DIFFERENT PROPORTIONS

NWD：CIoU	mAP@0.5
0.9：0.1	0.931
0.8：0.2	0.934
0.7：0.3	**0.935**
0.6：0.4	0.933
0.5：0.5	0.926
0.4：0.6	0.928
0.3：0.7	0.923

It can be seen from Table 3 that when NWD: CIoU is 0.7:0.3, the object detection of the model achieves the best effect, mAP@0.5 is 0.935.

IV. EXPERIMENTAL ANALYSIS

A. Data Set

The experimental data set of this paper was GRS-HRRSD-Dataset[19]. HRRSD data set is a data set released by the University of Chinese Academy of Sciences in 2019. HRRSD contains 21761 images obtained from Google Earth and Baidu Map, with spatial resolution ranging from 0.15m to 1.2m. There are 55,740 instances of targets in the HRRSDS, with about 4,000 samples in each category. The HRRSDS contain 13 types of targets. The 13 categories are: airplanes, baseball fields, basketball courts, Bridges, intersections, track fields, harbours, parking lots, boats, storage tanks, T-junctions, tennis courts and cars. In this paper, 15,667 pictures were selected as the training set, 1,741 pictures as the verification set, and 4,353 pictures as the test set.

B. Experimental Parameters

Experiments run under the of windows 10 operating system, python 3.10.6, Pytorch 1.13.0 and cuda 11.6. The GPU model used in this experiment is NVIDIA Quadro RTX 4000 with a capacity of 8GB. The superparameters of the experiment were set as follows: the training image size was 640x640, the batch size was 4, all models trained 80 epochs and did not use the pre-training weight, the initial learning rate was set as 0.01, ADAM was used as the optimizer, the momentum was 0.937, and there were 3 rounds of preheating training.

C. Comparison Between YOLOv7-tiny Network Model and Improved Network Model

As shown in Fig. 5 and Fig. 6, people can see that compared with the original YOLOv7-tiny model, the improved network model has a great improvement, mAP@0.5 from 0.905 to 0.935. Among them, the detection accuracy of cars, parking lots, T-intersections and baseball stadiums has the greatest improvement, which increases by 4.9%, 7.6% and 5.9% respectively. 4.0%. Other categories of testing have also seen different improvements.

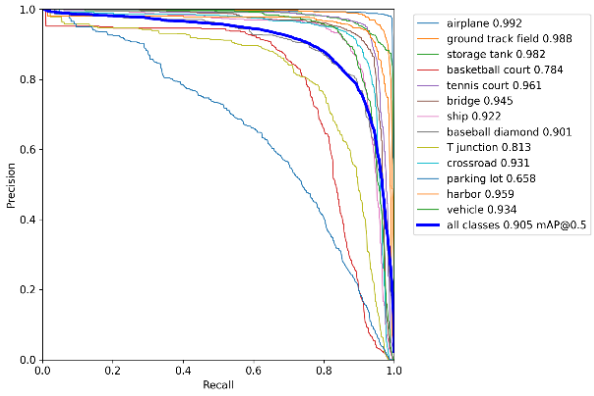


Fig. 5. P-R curves of YOLOv7-tiny network model.

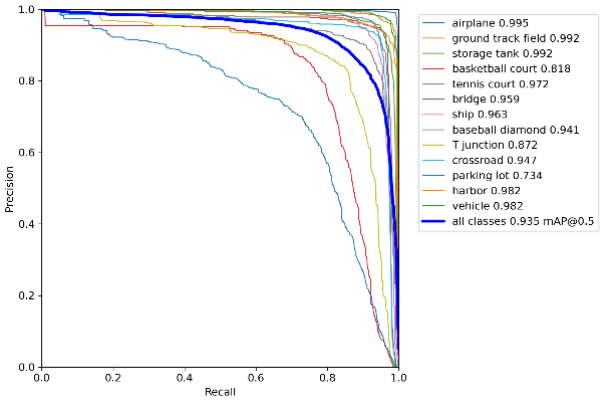


Fig. 6. P-R curves of Improved network model(ours).

For the detection of small targets in the data set, the paper selects two groups of typical images from the test set and tested them respectively with the original model and the improved model. The results are shown in Fig. 7 and Fig. 8.

From Fig. 7 and Fig. 8, people can see that the improved model has a great improvement compared with the original model. For example, the accuracy of the improved model in the first picture of each group has been increased by 0.1 ~ 0.3.

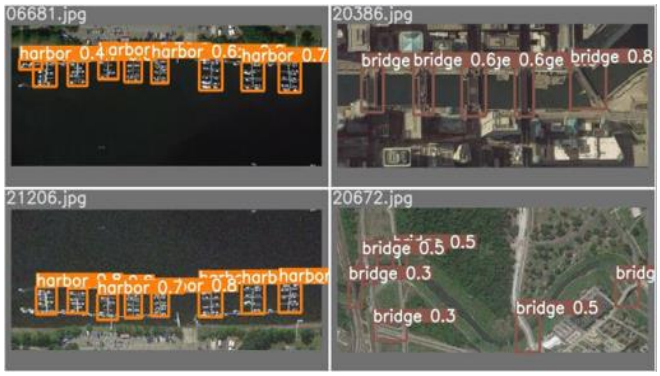


Fig. 7. The test results of YOLOv7-tiny network model.

Fig. 8. The test results of improved network model.

To further verify the effectiveness of the improved model, the paper conducted experiments on other models, such as Fast R-CNN, Fast R-CNN+GACL-Net, Faster R-CNN, Faster R-CNN+GACL-Net, YOLOv7-tiny (GACL-Net represents the method of lifting target location) and the results are shown in Table 4.

TABLE IV. COMPARISON OF EXPERIMENTAL RESULTS OF DIFFERENT NETWORK MODELS

Network model	Input Size	mAP@0.5	FPS
Fast R-CNN	640×640	0.665	0.65
Fast R-CNN+GACL-Net	640×640	0.68	0.56
Faster R-CNN	640×640	0.815	7.0
Faster R-CNN+GACL-Net	640×640	0.821	6.2
YOLOv7-tiny	640×640	0.905	89.29
Ours	640×640	**0.935**	**141.23**

As can be seen from Table 4, when inputting pictures of the same size, compared with the model based on Fast R-CNN, the model based on Faster R-CNN has great improvement in accuracy and speed. In addition, experimental results show that compared with the two-stage model, the speed of the first-stage model (YOLOv7-tiny) is about 10 times that of the two-stage model, and its accuracy is also greatly improved. By improving the YOLOv7-tiny, the researchers further improved the accuracy and speed of the model, with mAP@0.5 and FPS reaching 0.935 and 141.13 respectively.

V. CONCLUSIONS

Aiming at the particularity of remote sensing images, the paper propose an improved YOLOv7-tiny network model. The C2f module is designed by separating and merging and reducing the element by element operation, which improves the reasoning speed of the model while extracting the detail information of the picture. Secondly, the coordinate attention mechanism is added to further enhance the extraction of image features. Finally, the loss function based on NWD is introduced and combined with the CIoU loss function to improve the accuracy of the model for small object detection. Experiments show that the improved YOLOv7-tiny model is superior to the original network model and the traditional network model.

The proposed method can effectively improve the model's detection of remote sensing images, not only improve the accuracy of object detection, but also improve the reasoning speed of the model, and meet the real-time requirements. Next, the paper amplified the data set, increased the detection categories in the data set, and expanded the detection range, so as to improve the detection performance of the model in practical applications.

ACKNOWLEDGMENT

This research was funded by Anhui Provicial Overseas Study Visit Project (gxgwfx200050), Fuyang Normal University of Natural Science Research Project (2018kyqd0028), Fuyang Normal University of Science Research and Innovation Team (kytd202004), Anhui Graduate Graduate Innovation and Entrepreneurship Practice Project （2022cxcysj189).

REFERENCES

[1] Ying Liu, Luyao Geng, Weidong Zhang, Yanchao Gong, and Zhijie Xu, "Survey of Video Based Small Target Detection," Journal of Image and Graphics, Vol. 9, No. 4, pp. 122-134, December 2021. doi: 10.18178/joig.9.4.122-134

[2] G.T. Nie, and H. Huang, "A Survey of Object Detection in Optical Remote Sensing Images," Acta Automatica Sinica, vol.47, pp.1749-1768,2021.（in Chinese）

[3] Y.R. Liao,H.N. Wang, C.B.Lin, Y.Li,Y.Q.Fang, and S.Y.Ni, "Research progress of deep learning-based object detection of optical remote sensing image," Journal on Communications ,vol.43, pp.190-203, 2022.（in Chinese）

[4] R. Girshick," Fast R-CNN,"2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, *USA,* pp. 1440-1448, December 2015.

[5] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, pp. 1137-1149, June 2017.

[6] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 779-788,June 2016.

[7] C.Y.Wang, A. Bochkovskiy, and H. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," arXiv prints arXiv:2207.02696, 2022.

[8] J.W.Wang, C.Xu, W.Yang and L.Yu," A Normalized Gaussian Wasserstein Distance for Tiny Object Detection," arXiv preprint arXiv:2110.13389 ,2021.

[9] Z.Y.Niu, G.Q.Zhong, and H.Yu, "A Review on the Attention Mechanism of Deep Learning," Neurocomputing ,vol. 452, pp. 48-62, 2021.

[10] G. Brauwers and F. Frasincar, "A General Survey on Attention Mechanisms in Deep Learning," IEEE Transactions on Knowledge and

Data Engineering, vol. 35, pp. 3279-3298, April 2023.
[11] Z. Zheng ,P.Wang,D.W.Ren,W.Liu,R.G,Ye,Q.H.Hu and W.M.Zuo, "Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation," IEEE Transactions on Cybernetics, vol. 52, pp. 8574-8586, Aug. 2022.
[12] N.N.Ma, X.Y.Zhang, H.T.Zheng and J.Sun, "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design," arXiv preprint arXiv:1807.11164, 2018.
[13] Q. Hou, D. Zhou and J. Feng, "Coordinate Attention for Efficient Mobile Network Design," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 13708-13717, 2021,
[14] D.F.Zhou, J.Fang, X.B.Song, C.Y.Guan, J.B.Yin, Y.C.Dai and R.G.Yang, "IoU Loss for 2D/3D Object Detection," 2019 International Conference on 3D Vision (3DV), Quebec City, QC, Canada, pp. 85-94, 2019.
[15] Z.H.Zheng, P.Wang,W.Liu,J.Z.Li, R.Ye and D.W.Ren, "Distance-IoU Loss: Faster and Better Learning for bounding box Regression ," arXiv preprint arXiv:1911.08287, 2019.
[16] S. Zhang, C. Chi, Y. Yao, Z. Lei and S. Z. Li, "Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 9756-9765, 2020.
[17] K., Kang and H. S.Lee," Probabilistic Anchor Assignment with IoU Prediction for Object Detection ",arXiv preprint arXiv:2007.08103, 2020.
[18] Z. Ge, S. Liu, Z. Li, O. Yoshie and J. Sun, "OTA: Optimal Transport Assignment for Object Detection," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 303-312, 2021.
[19] Y. Zhang, Y. Yuan, Y. Feng and X. Lu, "Hierarchical and Robust Convolutional Neural Network for Very High-Resolution Remote Sensing Object Detection," IEEE Transactions on Geoscience and Remote Sensing, vol. 57, pp. 5535-5548, Aug. 2019.