Small Target Detection Method of Optical Remote Sensing Image Based on Multi-scale Information Fusion

Hongyan Li, Baoqing Xu, Ziyang Zhang, Weifeng Wang

Abstract—**This paper proposes a multi-scale information fusion based remote sensing small target detection method that aims to address the issue of low target detection accuracy in optical remote sensing images due to complex backgrounds, diversified scales, small targets, and different directions. Firstly, the architecture of the RepConv module significantly increases the detection accuracy of small targets without adding more inference time. Secondly, by introducing the ECA attention mechanism, a C3ECA module is constructed to effectively reduce the interference of complex background areas and achieve accurate positioning of the target area. The PANet structure in YOLOv5 is replaced by the BiFPN structure to balance the feature information of different scales and improve the detection performance of multi-scale objects. In addition, in order to solve the uncertainty of target direction and reduce the boundary discontinuity caused by angle regression, a circular smooth label method is used to provide an effective solution for target detection. The preprocessing method of image slices is employed to successfully achieve the target detection of high-resolution images. This approach greatly minimizes the issue of missed detection and erroneous detection of small objects in large images. The experimental findings indicate that the proposed approach markedly enhances the accuracy of remote sensing image detection and offers notable benefits in the recognition of small-scale targets.**

Index Terms—**Optical remote sensing image, Deep learning, Target detection, Circular smooth label**

I. Introduction

The era of high-quality optical remote sensing photographs has begun with the fast advancement of space remote sensing technology. These images are captured by satellites or drones from high altitudes and contain a wealth of surface information. The ground data in optical remote sensing photos has improved in quality and detail as resolution keeps rising. This trend provides rich and extensive possibilities for applications in multiple fields such as aerial reconnaissance [1], urban and rural construction [2], precision agriculture [3], and natural disaster monitoring [4].

In recent years, computer vision has advanced significantly, largely due to the growing applications and innovations in deep learning methods. Many excellent deep learning models for target detection have been proposed, including two-stage models like R-CNN [5-7], and one-stage models like SSD [8] and YOLO [9-12]. In the domain of detecting natural images, these models have exhibited remarkable performance. Researchers have initiated the incorporation of sophisticated detection algorithms into applications related to satellite remote sensing imaging, with the objective of addressing the challenges associated with this process.

In an effort to enhance item recognition in remote sensing photos, Yu et al [13] created the Stepwise Localization Bidirectional Pyramid Network (Sw-LBPN). Zhou et al [14] use of contextual transformation and data augmentation modules significantly increased the precision and effectiveness of remote sensing image identification. To improve the ability to express features, Zhao et al [15] used an embedded Reception Field Block (RFB) module with an RFBNet detector. Wan et al [16] successfully improved the object detection performance. Fu and colleagues [17] created a coordinate attention module that has an adaptable sensing field size, which improves the network's capacity to extract target characteristics that are multi-scale. Also, a CSandGlass module was used by Luo et al [18] to replace the residual module on the backbone feature extraction network, which greatly enhanced the performance of airplane target recognition in remote sensing photos. Conversely, Cao et al [19] improved the CSPDarknet53 architecture to retain adequate global contextual information by using Swin Transformer and introducing the Coordinate Attention (CA) module to enhance the accuracy of tiny object characteristics in remotely sensed pictures. Using the dual-branch architectural attention mechanism, Yi et al [20] improved the local module in the YOLOv8 feature extraction network. They also employed the visual converter block to maximize the feature map's representation, which improved the detection results' accuracy.

Although these methods have achieved certain results, they still have limitations for the detection of small targets under complex backgrounds. First of all, although the above methods improve the detection accuracy, they also increase unnecessary inference time and do not effectively solve the problem of small target detection in complex backgrounds.

Manuscript received January 3, 2024; revised April 25, 2024. This work is supported by the National Key R & D Program (2021YFE0105000), Natural Science Foundation of China (52074213).

Hongyan Li is a Senior Engineer of Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an 710054, China (e-mail: lihongyan@xust.edu.cn).

Baoqing Xu is a postgraduate student in Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an 710054, China (e-mail: 2586678038@qq.com).

Ziyang Zhang is a postgraduate student in Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an 710054, China (e-mail: yi_shang_0@163.com).

Weifeng Wang is a Professor of Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an 710054, China (e-mail: wangwf03@126.com).

Secondly, the above method still has room for improvement in target detection at different scales and in solving the uncertainty of target direction.

This work presents Ric-YOLOv5, a high-precision target identification technique, as a potential solution to the challenge of target recognition in satellite remote sensing images. The RepConv module is used to replace the traditional Conv module, which is integrated into the backbone network of YOLOv5; the ECA attention mechanism is introduced to construct the C3ECA module, which is used in place of the C3 module in YOLOv5; in YOLOv5, the BiFPN structure takes the role of the PANet structure; the image slicing pre-processing method is used, and the circular smooth labeling module is introduced. The enhanced model exhibits a notable increase in the assessment metrics of target detection and performs better in the DOTA dataset [21].

II. YOLOv5 Network Structure

YOLOv5, a single-stage algorithm, is the fifth generation of the YOLO family and a member of the regression family of target identification techniques. In contrast to conventional methodologies for identifying sliding windows and subsequent region delineation, YOLOv5 conceptualizes target detection as a regression problem. This approach enables optimization of end-to-end detection performance and brings significant innovation to the field of target detection. The YOLOv5 is available in five variants: YOLOv5l, YOLOv5x, YOLOv5m, YOLOv5n, and YOLOv5s. These variations have been designed to accommodate varying application situations and vary in terms of network depth and width. This series of versions offers a range of options, from lightweight to high performance, to suit various computing resources and performance requirements. The YOLOv5s model, which is extensively employed in the field of real-time target detection, is the smallest and quickest detection speed among them. As a result, YOLOv5s serve as the foundation for the research presented in this article.

III. Ric-YOLOv5 Network Structure

In this study, we improved the YOLOv5 target detection model and proposed a target detection model, Ric-YOLOv5, suitable for remote sensing images. The main improvements include (1) Without requiring more inference time, the architecture of the RepConv module can significantly increase the detection accuracy of small targets. (2) The ECA attention mechanism and the C3ECA module are designed to deal with the interference caused by complicated background areas in remote sensing image data sets. This helps the model to detect targets in regions more accurately. (3) The BiFPN structure replaces the PANet structure in YOLOv5 to achieve the balance of feature information at different sizes and thus improve the detection performance of multi-scale objects. (4) In order to solve the target orientation uncertainty problem and mitigate the boundary discontinuity caused by angular regression, the circular smoothing labeling method is adopted as an effective solution. (5) In order to skillfully achieve target detection on high-resolution images, a preprocessing method of image slicing is employed, which significantly reduces the problem of missing and misdetecting small targets in large images.

With these enhancements, the Ric-YOLOv5 model now exhibits better robustness and performance in challenges involving the identification of targets in remote sensing images. Figure 1 shows the architecture of the Ric-YOLOv5 neural network.

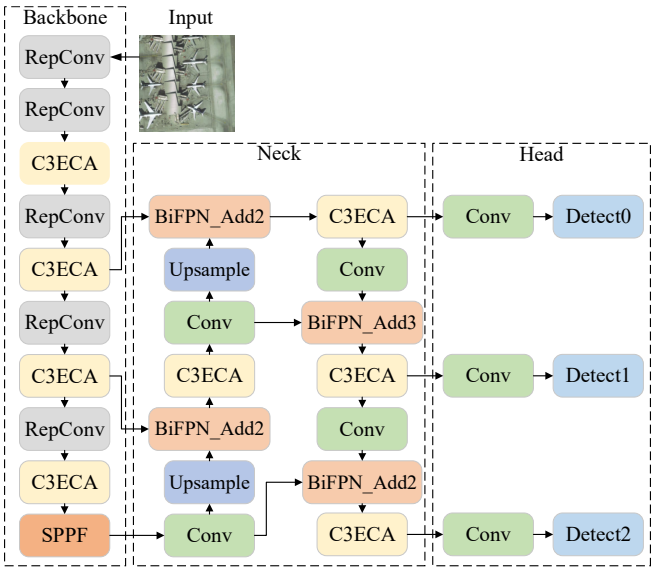
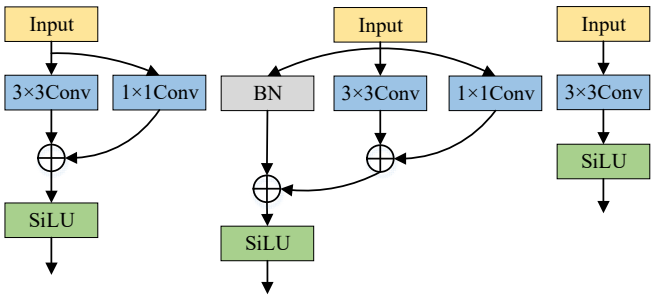


Fig. 1. Ric-YOLOv5 neural network structure

A. RepConv Module

The RepConv module has been seamlessly integrated into YOLOv5, utilizing the multi-branch structure of RepConv to effectively enhance the feature representation ability of small targets. As a result, the accuracy of tiny target detection improves. Furthermore, the reparameterization approach [22] is applied throughout the reasoning phase to merge the RepConv parallel branch into a single branch, preserving the YOLOv5 structure without adding to the reasoning time.



(a) Channels changed (b) Channels unchanged (c) Inference architecture

Fig. 2. RepConv module structure

The module expands the convolutional module by integrating a parallel 1×1 convolutional layer into the backbone's 3×3 convolutional layer. The 3 x 3 convolutional layer receives the outputs from the parallel branches during the inference stage. This update increases the detection accuracy of tiny items in the DOTA dataset without adding extra time to the inference process. The RepConv module makes use of a multi-branch structure during training. When there are more input feature channels than output feature channels, as in Fig. 2a, the module structure is altered. On the other hand, the module's structure is shown in Fig. 2b when

the number of input feature channels and output feature channels equals one. As illustrated in Fig. 2c, the outputs from the multi-branch structure are combined in the 3 × 3 convolution layer during the inference stage to create a single path structure.

B. C3ECA Module

DOTA datasets typically have a high concentration of complicated background, which significantly impairs object detection accuracy and causes interference. This is particularly detrimental to the identification of tiny objects. Therefore, this work uses the Efficient Channel Attention (ECA), an enhancement of the SE module, to reduce background interference and enhance the efficiency of tiny target identification. This innovation eliminates background interference, particularly when dealing with complex background information, which significantly limits feature extraction in DOTA datasets. This paper presents a unique attention mechanism, the C3ECA module, which is produced by integrating an efficient channel attention mechanism, ECA, inside the C3 module. The purpose of this mechanism is to improve the feature information within the target region and reduce the interference caused by the background. Following global average pooling (GAP), the ECA module eliminates the fully connected layer from the SE module and employs weight-shared one-dimensional convolution to learn features directly. A crucial factor in one-dimensional convolution is the convolution kernel size, or hyperparameter k, which establishes how well local interactions work across channels.

The ECA module aims to capture the interactions between channels by utilizing each channel and its adjacent k channels in the feature map. Fig. 3 vividly demonstrates the structure of the ECA module. In Fig. 3, it can be observed that the weights of the one-dimensional convolution are designed in a staggered arrangement, thus realizing the cross-channel interaction function.

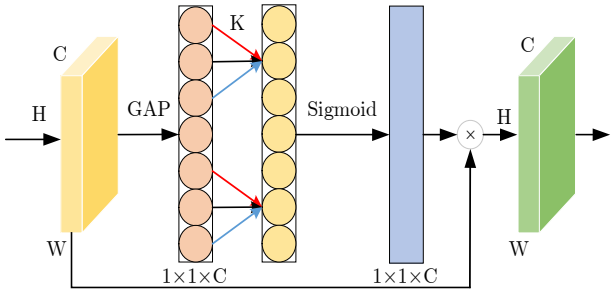


Fig. 3. ECA module structure

There are groups of weights, and the size of the convolutional kernel determines how many weights there are in each group. This greatly lowers the overall number of parameters because the weight values within each group are shared. While attention is calculated using two completely linked layers in the SE attention mechanism, it is computed using k closest neighbor channels in the ECA attention mechanism. It is evident that the ECA module's computational performance is directly impacted by the magnitude of k. In fact, there is a nonlinear mapping relationship between k and the number of channels C, which can be described by an exponential function as follows:

$$C = \phi(k) \approx \exp(\gamma * k - b) \tag{1}$$

Given that the number of channels is frequently an exponential multiple of two, the following formula may be employed to derive k:

$$k = \Psi(C) = \left| \frac{\log_2 C}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{2}$$

As illustrated in Fig. 4, a comparative analysis of the enhanced C3ECA module and the conventional C3 module is conducted within the context of the YOLOv5s network architecture. It can be clearly seen that the C3 module consists of multiple stacked bottleneck modules, while in this study, we introduced the ECA attention mechanism in these bottleneck modules, which effectively suppresses the background interference and improves the detection performance of small targets.

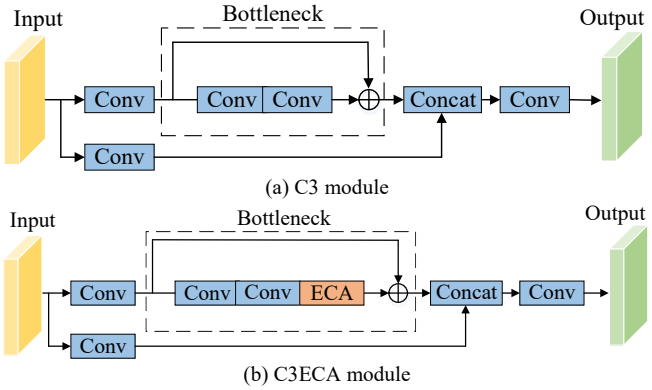


Fig. 4. Comparison of the architecture of the C3 module and the C3ECA module

C. Bidirectional Feature Pyramid Network

The YOLOv5 feature fusion network utilises PANet, which combines feature maps extracted at various levels to higher levels. In order to improve target identification efficiency, this allows the model to fully utilize multi-level and multi-scale feature information.

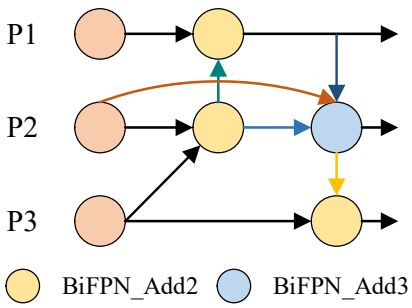


Fig. 5. BiFPN module structure

The advent of FPN has led to the development of numerous cross-scale feature fusion networks, including PANet and NAS-FPN, which have gained popularity due to the increasing prevalence of multi-scale feature fusion. However, the incorporation of learnable weights, which assist in the assessment of the significance of certain input traits, and the more even distribution of data across several scales, collectively enhance the performance of FPN. Therefore, instead of using the neck portion of the feature fusion approach, BiFPN is employed in this research. The

BiFPN model's schematic structure is seen in Fig. 5, where P1, P2, and P3 stand for the various scale characteristics that the Backbone component produces. The weighted aggregation approach is employed by the feature fusion modules BiFPN_Add2 and BiFPN_Add3, which fuse the features of the current layer with those of the preceding layer.

D. Circular Smooth Label

At present, target detection bounding boxes are mainly divided into three types: horizontal bounding boxes, rotational bounding boxes and customized bounding boxes. It is important to keep in mind that while recognizing objects in remote sensing photos, the target's orientation may vary or be random. As a result, the bounding box labeling technique should be flexible enough to change depending on the target object's true shape. Given the features of changeable object orientation in remote sensing picture target recognition, the rotating bounding box is undoubtedly a better option.

This paper performs rotated bounding box target detection for the DOTA dataset and introduces angular loss. Four major components comprise the total loss: angle loss, classification loss, regression loss, and confidence loss. The angular regression problem is converted into a classification problem in the computation of angular loss, and Circular Smooth Label (CSL) is adopted to solve it. The approach of angular segmentation restricts the range of predicted results and effectively mitigates potential boundary issues. The circular smooth labels are given smooth label values with a tolerance and are encoded cyclically and periodically.

Equation (3) displays the CSL expression.

$$CSL(x)=\begin{cases} g(x), & \theta-r<x<\theta+r \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The window function's radius θ is connected to the angle of the current bounding box in the case when $g(x)$ is the window function. The window function has four main characteristics, including periodicity, monotonicity, maximum value and symmetry. In practical applications, window functions, which include impulse, rectangle, trigonometric, and Gaussian functions, are frequently employed. The network model can effectively calculate the angular separation between the expected and the real label by expertly setting the window function. The model performs better because of this architecture, which causes the loss value to gradually drop as the forecast gets better. In addition, the periodic nature of the window function effectively solves the problem of angular periodicity, as shown by CSL in Fig. 6.

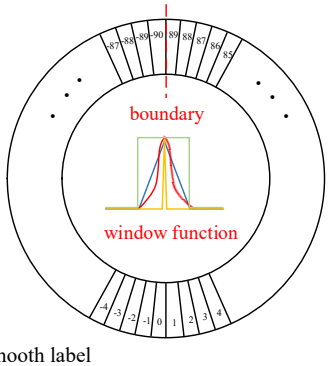


Fig. 6. Circular smooth label

IV. Experimental Design

The experimental data set used in this work was taken from the DOTA public remote sensing data set, which is sourced from a variety of platforms and sensors, including the JL-1, GF-2, and Google Earth satellites. There are 2806 high-resolution photos in the collection, which span 188,282 distinct instances. These examples relate to 15 different categories, including Small Vehicle (SV), Basketball Court (BC), Tennis Court (TC), Harbor (HA), Soccer Ball Field (SBF), Plane (PL), Swimming Pool (SP), Roundabout (RA), Large Vehicle (LV), Storage Tank (ST), Ship (SH), Bridge (BR), and Helicopter (HC), Ground Track Field (GTF), Baseball Diamond (BD). These photos' resolutions span from 800×800 to 4000×4000, providing a broad range of realistic application scenarios.

The picture in the dataset is too big, therefore direct model training is not appropriate. Therefore, the image cutting method is used for preprocessing. In order to ensure that no target information is lost during the cutting process, the conventional approach is to ensure that there are overlapping areas in the image after cutting. In this study, the original image is cut by the overlap distance of each 200 pixels, and a sub-image with a size of 1024×1024 pixels is generated. This process generated multiple sub-images with a resolution of 1024×1024, yielding a total of 21046 such images. Of these, 2105 photos made up the test set and 16837 images were chosen at random to form the training set.

The AMD Ryzen 76800HCPU, 3.20 GHz, 16 GB RAM, NVIDIA GeForce RTX3060 graphics card, Windows 11 operating system, and CUDA version 11.6 were the specifications of the machine used for the tests, which were carried out with the PyTorch deep learning framework. With a momentum term of 0.9, an initial learning efficiency of 0.001 for the weights, and a decay coefficient of 0.0005, the network was trained via asynchronous stochastic gradient descent. Batch-size = 16 and 200 iteration rounds.

V. Experimental Results And Analysis

A. Evaluation Index

The model's performance is assessed in this research using the mean average precision (mAP), recall (R), precision (P), and average precision (AP) of AP values across all categories. Below are the formulae for these metrics:

$$P=\frac{TP}{TP+FP} \tag{4}$$

$$R=\frac{TP}{TP+FN} \tag{5}$$

$$AP=\frac{\sum_n P}{n} \tag{6}$$

$$mAP=\frac{\sum_{i=1}^{N} AP}{N} \tag{7}$$

where N is the number of categories, n is the total number of data, TP stands for true instances, FP for false true cases, and FN for the number of missed detections by the model. An area under a curve, represented by the AP value, may be visualised by using the accuracy rate as the vertical

axis and the recall rate as the horizontal axis. In this experiment, the evaluation index is the mean average precision (mAP) at the point where the intersection and concurrency ratio thresholds between the detection frame and the real frame are equal to 0.5.

B. Ablation Experiment

This study conducts a series of module ablation experiments on the DOTA dataset to fully validate the effectiveness of the proposed algorithm. The experiments aim to investigate the effects of RepConv, C3ECA, bidirectional feature pyramid network, and circular smoothing labelling module on the algorithm's performance. This study included five sets of ablation trials, the results of which are displayed in Table I.

TABLE I
ABLATION EXPERIMENT

Group	RepConv	C3ECA	BiFPN	CSL	mAP
1					75.7
2	√				77.1
3	√	√			80.2
4	√	√	√		81.5
5	√	√	√	√	**82.2**

It is evident from the data in Table I that the method suggested in this work yielded a number of important experimental findings. The benchmark model used in the first set of experiments was the YOLOv5 algorithm, which was compared in detail with the subsequent experiments. The mAP of the benchmark model is 75.7%. In the second set of tests, the original algorithm structure is maintained by substituting the RepConv module for the Conv module in the backbone network. This simple yet effective improvement significantly improves the detection accuracy of small-scale objects and increases mAP by 1.4%. The third group of experiments further introduces the RepConv module and the C3ECA module. The combined impact of these two modules efficiently aids in the model's correct target area location and enhances detection accuracy by 3.1%. Based on the findings of the third set of trials, the fourth group presents a bidirectional feature pyramid network module. By increasing mAP by 1.3%, this module successfully improves multi-scale object detection performance. Finally, the fifth set of experiments further optimized the boundary regression, especially after increasing the rotation angle of the target, using a circular smooth label to improve the accuracy of target positioning. This improvement improves the overall accuracy of the algorithm mAP to 82.2%. When the improved Ric-YOLOv5 network is trained for 200 rounds, the loss, recall rate and mAP have basically converged. The comparison charts for the loss, recall rate, and mAP of the Ric-YOLOv5 network and the original network are shown in Figures 7, 8, and 9.

After a thorough analysis of these data, it can be said that the algorithms presented in this research have significantly advanced the field of remote sensing picture target recognition, and that each of the modular techniques offered has improved performance to a large degree. These experimental findings not only confirm the algorithms' efficacy but also offer compelling evidence in favor of further study and use of remote sensing image processing.

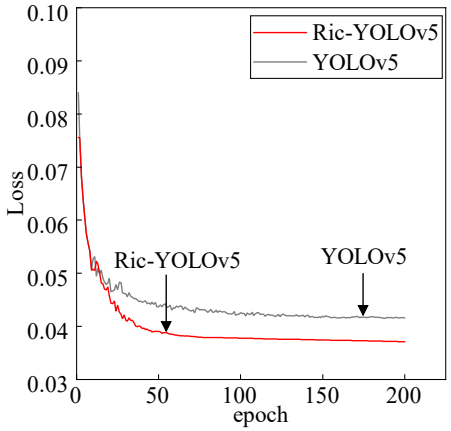


Fig. 7. Loss comparison

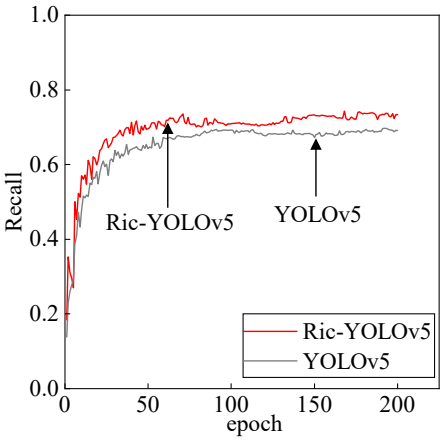


Fig. 8. Recall rate comparison

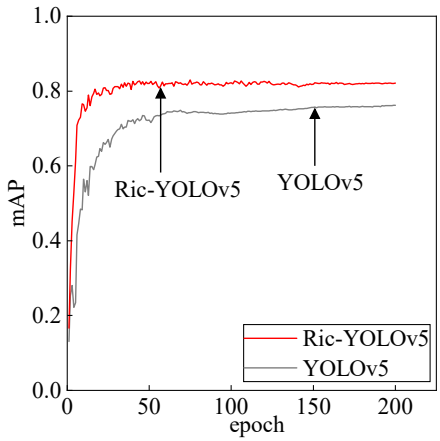


Fig. 9. mAP comparison

C. Comparison Experiment

To evaluate the performance of the updated algorithm against well-known target identification methods as R^2CNN, YOLOv3, YOLOv4, MaskOBB, YOLOv6s, and YOLOv8, a number of comparison experiments were conducted using the DOTA dataset. Table II displays the experiment results. These comparative trials offer a better knowledge of how well the enhanced algorithms perform in contrast to other algorithms.

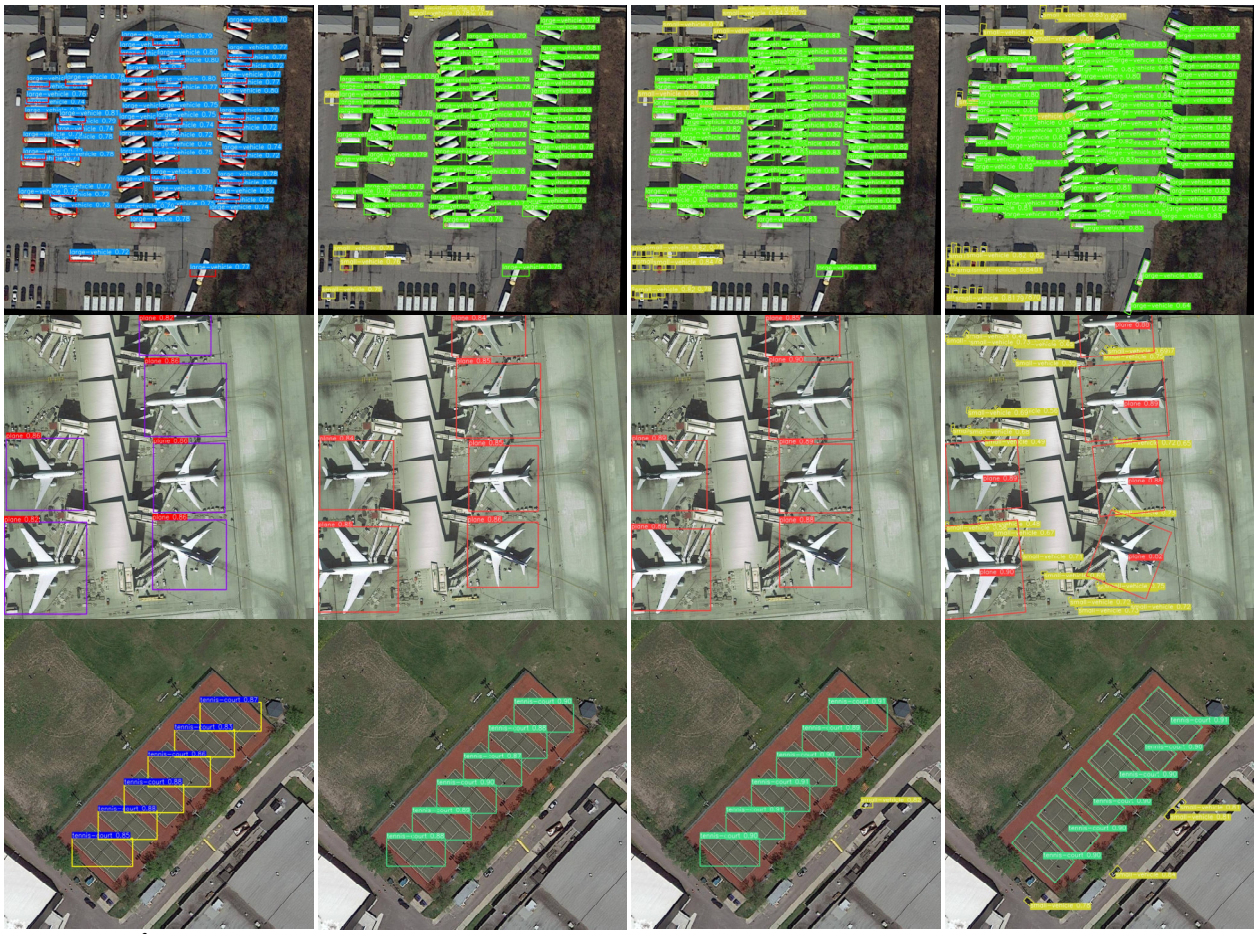
The data shown in Table II provides a clear understanding of how various target recognition algorithms perform on remote sensing photos.

TABLE II
COMPARATIVE EXPERIMENTS OF DIFFERENT TARGET DETECTION ALGORITHMS

Model	SV	LV	PL	ST	SH	HA	GTF	SBF	TC	SP	BD	RA	BC	BR	HC	mAP
R^2CNN	59.6	51.2	79.9	73.1	55.4	54.9	65.2	55.3	89.7	54.1	66.1	53.2	67.1	36.1	49.3	60.7
YOLOv3	68.3	70.2	90.1	61.3	83.6	78.9	48.8	51.1	95.3	78.3	54.8	33.3	53.6	28.9	**81.4**	65.2
YOLOv4	65.7	77.7	89.7	81.3	86.7	80.3	62.7	60.9	90.6	66.8	73.1	63.1	70.6	50.6	57.8	71.8
MaskOBB	68.8	60.8	89.8	86.6	73.3	66.8	68.7	**66.1**	90.5	68.4	80.9	67.1	**88.1**	52.1	65.5	72.9
YOLOv6s	76.6	75.2	89.5	86.7	86.6	74.1	72.8	54.6	90.1	**69.1**	**85.8**	70.1	83.9	54.6	63.5	75.5
YOLOv8	87.2	91.4	95.8	85.3	95.6	86.1	**74.6**	60.3	94.6	61.8	82.4	71.1	80.4	60.2	69.9	79.8
Ric-YOLOv5	**90.1**	**94.6**	**97.7**	**87.5**	**97.8**	**88.2**	74.1	63.5	**97.3**	63.2	85.3	**74.2**	82.6	**63.1**	73.5	**82.2**

The suggested R^2CNN algorithm, which is based on Faster R-CNN, has a mAP value of 60.7%; however, it has some issues processing multi-category remote sensing images. These issues are primarily caused by the requirement to generate a lot of horizontal frames using the RPN, which causes the horizontal frames to overlap, which lowers the detection accuracy overall. With a detection accuracy of 65.2%, YOLOv3 does particularly well when it comes to identifying major target categories like helicopters (HC). However, the algorithm is constrained by poor localization accuracy as well as weak detection of small objects, and the overall detection effect needs to be improved. As opposed to the first two algorithms, YOLOv4 obtains a mAP value of 71.8%, indicating a notable improvement in performance. However, the algorithm has low detection efficiency when dealing with complex scenes and dense targets. MaskOBB has a mAP value of 72.9%, and while it works well in the majority of detection circumstances, its accuracy in detecting dense objects, such large vehicles (LV), may still be improved. With a mAP score of 75.5% and a quite good accuracy, YOLOv6s does well when it comes to detecting tiny targets in complicated backgrounds. Nevertheless, the technique is vulnerable to the missed detection issue, necessitating more study and development. The YOLOv8 algorithm achieved an average accuracy of 79.8%. In particular, it outperforms most of the compared algorithms in the detection of the Ground Track Field (GTF) category. The algorithm's shortcomings, which show up as a lack of recognition efficacy, persist when it comes to the identification of tiny spinning objects.

After extensive experimental validation, the method proposed in this study shows a significant improvement in accuracy compared with existing techniques. Specifically, the algorithm achieves an average precision mean of 82.2 %, which is in the leading position among the compared algorithms. In addition, the algorithm shows significant performance improvement in dense target detection such as Small Vehicle (SV), Plane (PL), ship (SH) and other targets. This finding provides additional evidence of the algorithm's efficacy in processing remote sensing images.



(a) R^2CNN (b) YOLOv5s (c) YOLOv6s (d) proposed algorithm

Fig. 10. Comparison of detection effect of some algorithms

Using R^{2}CNN, YOLOv5s, YOLOv6s, and the enhanced algorithms suggested in this study, among other methods, some of the photos chosen for this study were recognized in order to assess how well various target detection algorithms performed in processing the images in the DOTA test set. Figure 10 displays a comparison of the detection findings. The graphic clearly shows that the classic R^{2}CNN, YOLOv5s, and YOLOv6s algorithms frequently have missed detection for tiny objects, such small automobiles. Large differences in target scales in remote sensing images are a challenge that these algorithms struggle to handle, and they are unable to fully extract the features of small targets in complex backgrounds or focus enough on targets that are hard to classify. However, the enhanced method shown in this work improves the capability of multi-scale target localization and feature extraction in complicated backdrops by the incorporation of modules including bidirectional feature pyramid network, C3ECA, and RepConv. As a result, it can generate more accurate detection frames, reducing both false positives and missed detections to some extent. Together, these improved modules make the algorithm in this study perform well in dealing with small targets in complex backgrounds.

VI. Conclusion

Building upon the YOLOv5 architecture, this paper suggests an enhanced method for remote sensing image target recognition called Ric-YOLOv5. This study first designed the RepConv module and used it to replace the conventional Conv module in the backbone network, thereby significantly improving the detection accuracy of small-scale targets. This was done in response to the challenges posed by complex backgrounds, diverse scales, the prevalence of small targets, and the diversity of target directions in remote sensing images. Furthermore, the ECA attention mechanism is used in the construction of the C3ECA module. This enhancement lessens the impact of background noise while simultaneously improving the accuracy of recognizing tiny objects. We substituted the BiFPN structure for the PANet structure in order to better match the detection requirements of multi-scale targets. Furthermore, we successfully addressed the problem of border discontinuity and target direction uncertainty by employing the circular smooth label approach. Simultaneously, the high-resolution picture pre-processing approach of image slicing is used, which successfully mitigates the issue of tiny target false and missed detection. The results of the experiments confirm the benefits of the Ric-YOLOv5 algorithm for small-scale target detection that this research proposes. The quality of target recognition tasks in remote sensing images is greatly enhanced by this advancement. Subsequent investigations endeavor to enhance the model's equilibrium between inference velocity and detection precision, rendering it more applicable to realistic situations.

References

[1] P. Stodola, J. Drozd, K. Šilinger, J. Hodický, and D. Procházka, "Collective Perception Using UAVs: Autonomous Aerial Reconnaiss -ance in a Complex Urban Environment," *Sensors*, vol. 20, no. 10, pp. 2926–2953, 2020.

[2] A. H. Aljaddani, X.-P. Song, and Z. Zhu, "Characterizing the Patterns and Trends of Urban Growth in Saudi Arabia's 13 Capital Cities Using a Landsat Time Series," *Remote Sensing*, vol. 14, no. 10, pp. 2382–2407, 2022.

[3] H. Opedes, S. Mücher, J. E. M. Baartman, S. Nedala, and F. Mugagga, "Land Cover Change Detection and Subsistence Farming Dynamics in the Fringes of Mount Elgon National Park, Uganda from 1978–2020," *Remote Sensing*, vol. 14, no. 10, pp. 2423–2443, 2022.

[4] P. Ye, "Remote Sensing Approaches for Meteorological Disaster Monitoring: Recent Achievements and New Challenges," *International Journal of Environmental Research and Public Health*, vol. 19, no. 6, pp. 3701–3729, 2022.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

[6] R. Girshick, "Fast R-CNN," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, 2015.

[7] 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*, pp. 1137–1149, 2017.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," Computer Vision -ECCV 2016, Lecture Notes in Computer Science, pp. 21–37, 2016.

[9] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

[10] Redmon, Joseph, and Ali Farhadi. "YOLO9000: better, faster, stronger," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263–7271, 2017.

[11] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement," *arXiv preprint arXiv: 1804.02767*, 2018.

[12] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv: 2004.10934*, 2020.

[13] N. Yu, H. Ren, T. Deng, and X. Fan, "Stepwise Locating Bidirectional Pyramid Network for Object Detection in Remote Sensing Imagery," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2023.

[14] Q. Zhou, W. Zhang, R. Li, J. Wang, S. Zhen, and F. Niu, "Improved YOLOv5-S object detection method for optical remote sensing images based on contextual transformer," *Journal of Electronic Imaging*, vol. 31, no. 4, pp. 43049, 2022.

[15] Y. Zhao, J. Zhao, C. Zhao, W. Xiong, Q. Li, and J. Yang, "Robust Real-Time Object Detection Based on Deep Learning for Very High Resolution Remote Sensing Images," *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1314-1317, 2019.

[16] H. Rizeei et al., "YOLO-HR: Improved YOLOv5 for Object Detection in High-Resolution Optical Remote Sensing Images," *Remote Sensing*, vol. 15, no. 3, pp. 614–621, 2023.

[17] Hongjian Fu, Hongyang Bai, Hongwei Guo, and Weiwei Qin, "Object Detection Method of Optical Remote Sensing Image with Multi-attention Mechanism," *Acta Photonica Sinica*, vol. 51, no. 12, pp. 312–320, 2022.

[18] S. Luo, J. Yu, Y. Xi, and X. Liao, "Aircraft Target Detection in Remote Sensing Images Based on Improved YOLOv5," *IEEE Access*, pp. 5184–5192, 2022.

[19] X. Cao, Y. Zhang, S. Lang, and Y. Gong, "Swin-Transformer-Based YOLOv5 for Small-Object Detection in Remote Sensing Images," *Sensors*, vol. 23, no. 7, pp. 3634–3650, 2023.

[20] H. Yi, B. Liu, B. Zhao and E. Liu, "Small Object Detection Algorithm Based on Improved YOLOv8 for Remote Sensing," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 1734–1747, 2024.

[21] G.-S. Xia et al., "DOTA: A Large-scale Dataset for Object Detection in Aerial Images," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3974–3983, 2018.

[22] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "RepVGG: Making VGG-style ConvNets Great Again," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13733–13742, 2021.