YOLO-FNC: An Improved Method for Small Object Detection in Remote Sensing Images Based on YOLOv7

Lanxue Dang, Gang Liu, Yan-e Hou and Hongyu Han

***Abstract*—The detection algorithms of small objects in remote sensing images is often challenging due to the complex background and limited pixels. This can lead to reduced accuracy in detection and an increased number of missed small objects. So this paper introduces YOLOFNC, an enhanced network based on YOLOv7. To improve the model's ability to capture features of small objects, an enhanced C3-Faster module based on the C3 module is designed and integrated into the YOLOv7 network. This module helps extract more features related to small objects. Additionally, we employ Normalized Wasserstein Distance (NWD) fusion GIoU as a novel loss function to refine the accuracy of network optimization weights and the small object regression framework. Furthermore, a coordinated attention (CA) mechanism is incorporated at strategic locations in the model to reduce redundant information in the feature layer and prevent the loss of important small object features. we conduct comparison experiments between YOLO-FNC and other commonly used object detection algorithms on DIOR, AI-TOD, and VisDrone datasets. The experimental results show that YOLO-FNC achieves 84.4% mAP on the DIOR dataset, 35.9% mAP on the AI-TOD dataset, and 52.6% mAP on the VisDrone dataset. Compared to YOLOv7 and other remote sensing object detection models, YOLO-FNC demonstrates better performance in object detection.**

***Index Terms*—small object detection, remote sensing images, deep learning, neural network, YOLOv7.**

I. Introduction

OBJECT detection is currently a popular area of research in computer vision and digital signal processing, with the advancements in remote sensing technology. The detection of objects in remote sensing images is being increasingly utilized in various technical fields such as drones [1], intelligent traffic monitoring [2], and aerospace [3]. Therefore, studying and addressing the challenges related to object detection in remote sensing images holds great significance for our future scientific and technological progress. With the increasing computing power, many algorithms, particularly convolutional neural networks, are widely used for various applications. The great success of AlexNet [4] in the 2012 image classification competition made object detection based on deep learning a new hot topic. On this basis, many scholars have developed more advanced object detection models. Object detection methods can be broadly categorized into two groups: region proposal-based algorithms and regression-based algorithms. The former uses a two-stage approach, first utilizing a region proposal network to identify potential regions containing objects, and then classifying and localizing each region for detection. Although these methods offer high accuracy, their models are complex, and their detection speed is relatively slow. Representative models in this category include the R-FCN [5] and Fast R-CNN [6] series. The latter is a regression-based algorithm that directly predicts the category and location of the object, completing the detection task in a single step. YOLO [7] is the most representative algorithm in this category. The YOLO algorithm has fast processing speed, small size, and good real-time performance. Yet, its detection precision typically lags behind two-stage methods.

Many scholars have made significant improvements to the detection performance of the YOLO series algorithms to improve their detection performance. Among the YOLO family of algorithms, YOLOv3 [8] is the most famous version. It employs multi-scale prediction, multi-level feature fusion, and a more robust network structure to improve detection accuracy and speed dramatically. YOLOv3 has achieved outstanding results in many vision tasks and competitions, becoming a significant milestone in object detection. It should be noted that YOLOv4 [9] also has a broad following and influence and surpasses YOLOv3 in some performance metrics, and thus also receives a high level of attention.TPH-YOLOv5 [10] applied a TPH prediction head to replace the original prediction head, which can detect objects at different scales, bringing some improvements in detecting small objects. The YOLOv5-Aircraft [11] includes scaling calibration into the normalization module to improve the effectiveness of features. They employ the Kullback-Leibler as a loss function and introduce CSandGlass module into their developed model to reduce information loss and reach higher detection precision and speed. The YOLO-extract [12] algorithm enhances feature extraction by removing layers and prediction heads with lower feature extraction ability, incorporating a new feature extractor, coordinate attention, and replacing the CIoU loss with the Focal-EIoU loss to speed up bounding box regression and reduce model loss. YOLOv7 [13] is a single-stage object detection model known for its good speed and accuracy. YOLOv7 introduces model re-parameters into the network architecture to re-parameterize the model. And adopt YOLOv5's cross-grid search label

Manuscript received January 24, 2024, revised July 25, 2024. This work was supported by the Science and Technology Development Plan Project of Henan Province, China (232102210013).

Lanxue Dang is a professor at Henan Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng, 475004, China (e-mail: danglx@vip.henu.edu.cn).

Gang Liu is a graduate student of the School of Computer and Information Engineering, Henan University, Kaifeng, 475004, China (e-mail:liugang@henu.edu.cn).

Yan-e Hou is a professor at Henan Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng, 475004, China (e-mail: houyane@henu.edu.cn).

Hongyu Han is a lecturer at the School of Computer and Information Engineering, Henan University, Kaifeng, 475004, China (corresponding author, e-mail:hanhongyu@henu.edu.cn).

allocation strategy. In addition, the E-ELAN efficient network architecture is used in the backbone. YOLOv7 is an excellent object detection algorithm. However, it still has some shortcomings in the task of detecting small objects. Because of its deeper network, E-ELAN structure does not fully extract features for small objects and is prone to losing some information. The GIoU loss function applied in YOLOv7 performs poorly in detecting small objects. Therefore, YOLOv7 still needs further research on small object detection.

We enhanced the existing model by using YOLOv7 as a base. Firstly, we enhanced the feature extraction capability of the YOLOv7 network by introducing the C3-Faster module, an improvement based on the C3 module and FasterNet[14] module. This enhancement reduces the number of model parameters and simplifies it. Secondly, we improved the network's ability to detect small objects by incorporating the Normalized Wasserstein distance (NWD) [15] and GIoU as the loss function in YOLOv7, leading to increased detection accuracy for smaller objects. Thirdly, we embedded the coordinate attention (CA) [16] mechanism in YOLOv7, effectively capturing both channel and spatial information, enhancing feature extraction for better target recognition, and enabling the model to focus on critical multiscale local and global information within large datasets.

II. Related Work

The use of remote sensing images is expanding rapidly, with object detection being a key focus for researchers. Various detection algorithms have been developed, including traditional methods and deep learning-based approaches. Traditional algorithms rely on manually designed features and are suited for detecting single objects but struggle with complex scenes [20]. For example, Chen et al. [21] used histogram oriented gradient description for feature extraction to detect vehicles. Based on YOLOv3, Zhang et al. [22] proposed a deep separation attention guidance network for detecting small vehicles in optical remote sensing images. Nevertheless, these methods are suitable for extracting shallow features and struggle to capture deep semantic information.

The performance of algorithms or models in detecting small objects is still not satisfactory due to issues like the small volume of the objects and insufficient features in RSI. To address these problems, various local methods have been suggested. One category involves data strategy-based methods, such as multi-scale training [23] and data augmentation techniques like mosaic [9]. These methods can enhance small object detection but may also lead to increased computational costs. Akyon et al. [23] introduced a detection framework (SAHI) based on slicing-aided inference and fine-tuning, which not only improves the ability to detect small objects but also maintains higher memory utilization. Kisantal et al. [24] proposes oversampling of small object samples, followed by copying and pasting the small objects within the samples to provide sufficient small objects for matching with anchors, thereby improving the performance of small object detection. The second category is mainly based on feature enhancement learning methods. Pei et al. [25] proposed the feature extraction module LCB based on the improved YOLOv5s network and performed multi-scale feature fusion to extract features in remote sensing images. Dao et al. [26] proposed an asymmetric context modulation module to use the interaction of contextual information to highlight small objects better. In addition, Rabbi et al. [27] proposed an enhanced super-resolution generative adversarial network model, which converts images into super-resolution images and extracts feature information, improving the accuracy of detection performance of small object objects. The third category is methods based on sample allocation principles. The IoU metric is widely used in modern anchor-based detectors [8, 28, 29], for assigning positive and negative samples during training. However, small object detection requires careful manual setting of the size and aspect ratio of the anchor box to ensure sufficient positive samples. Zhang et al. [30] divided samples according to the statistical characteristics of objects, thereby avoiding additional hyperparameters. At the same time, changes in the position of small objects are very sensitive, and slight movements may cause significant interference to IoU, so Wang et al. [15] used NWD to build a detector for small objects.

The efforts made in the field of object detection have led to significant advances. However, there is still a need to improve small object detection in RSI. To address this, we have introduced the C3-Faster module into the YOLOv7 network, specifically designed to enhance feature extraction, improve inference speed, and reduce the number of model parameters. Additionally, we have incorporated the NWD combined with GIoU loss function to make the network more effective in detecting small objects. Furthermore, we have introduced the CA mechanism at appropriate positions of the model to capture channel and spatial information, thus enhancing the model's detection performance.

III. Method

This section describes the proposed improved detection model(YOLO-FNC). Next, the specific structure and methods of the model will be introduced.

A. Architecture

This section introduced the YOLO-FNC object detection algorithm proposed in this article. The network structure is shown in Fig. 1. Typically, high-level information can be well reflected and aggregated to a single point. Small objects originally contain fewer pixels. The aggregated features will be reduced if the number of network layers is increased. For example, a small object of 15*15 pixels may only have 1*1 features after convolution. The detection accuracy will drop significantly if there are multiple such small objects. This article believes that the E-ELAN module in YOLOv7 has too many convolutional layers, and the network depth is too high, which will lead to the loss of small object information. Therefore, our proposed method is as follows. Firstly, we introduced the C3-Faster module into the YOLOv7 model, replacing the E-ELAN module in the baseline model YOLOv7. This improvement enhances the feature extraction capability, reduces the model parameters, and improves the network's detection and inference speed. Secondly, in response to the features of small objects in

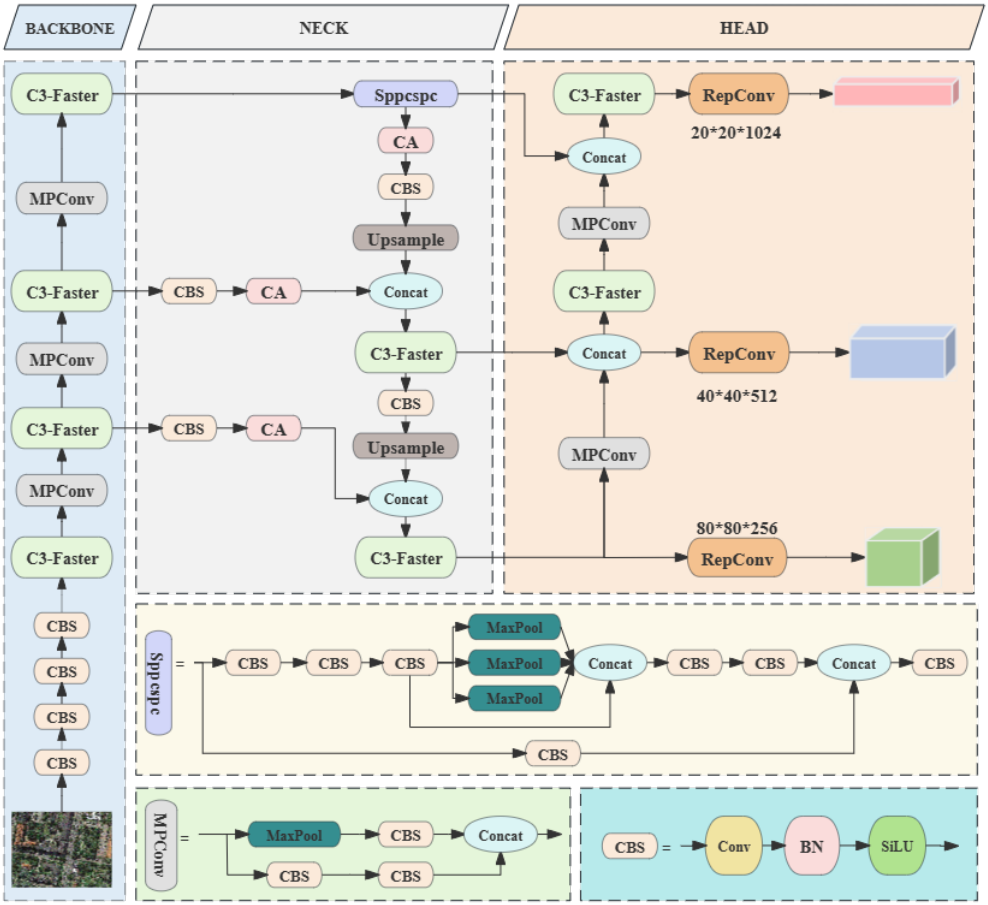

Fig. 1. Structure of the YOLO-FNC

remote sensing data sets, we introduced NWD and combined it with the GIoU loss function to enhance small object detection performance. Finally, we added a CA mechanism to the network at the location where different feature extraction layers interact and the location after the SPPCSPC module. This improvement enables the network to better learn useful small object information. Thus, the model emphasizes important object features, reduces interference from irrelevant information, and enhances the algorithm's ability to detect small objects.

B. C3-Faster Module

The structure of C3-Faster is illustrated in Fig. 2. Its overall architecture is composed of a C3 module and a FasterNet module. The C3 module is primarily utilized to enhance the network's depth and receptive field to improve feature extraction capabilities. The C3 module is composed of three convolution (Conv) blocks and a BottleNeck. The first Conv block has a stride of 2, which can reduce the size of the feature map by half. The strides of the second and third Conv blocks are 1. This does not change the size and spatial resolution of the feature map, thereby better retaining the local information of the object. The Conv blocks in the C3 module all use 3x3 convolution kernels. Between each Conv block, a BN layer and LeakyReLU activation function are also added to improve the stability and generalization performance of the model. However, because it has a deep feature extraction layer, the C3 module is unsuitable for detecting small remote sensing objects. The FasterNet module has higher running speed and accuracy and is composed of a Partial convolution (PConv) layer 3*3 and two conventional convolution layers Conv1*1. Compared with conventional convolutional layers Conv, PConv reduces redundant calculations while optimizing memory access, making it a simple and efficient layer. So PConv can extract object feature information more efficiently. Together, they are shown as an inverse residual block, with increased channels in the middle layer and shortcut connections for reusing features. Therefore, this article combines the C3 and FasterNet modules to form the C3-Faster modules. The C3-Faster module provides enhanced performance in detecting small objects in remote sensing images.

C. NWD-GIoU Module

The loss function of YOLOv7 is Generalized Intersection over Union (GIoU), which is improved based on Intersection over Union (IoU). The expression for IoU is equation (1), and the expression for GIoU is equation (2).

$$IoU = \frac{|A \bigcap B|}{|A \bigcup B|} \tag{1}$$

$$GIoU = IoU - \frac{|C \backslash (A \bigcup B|)}{|C|} \tag{2}$$

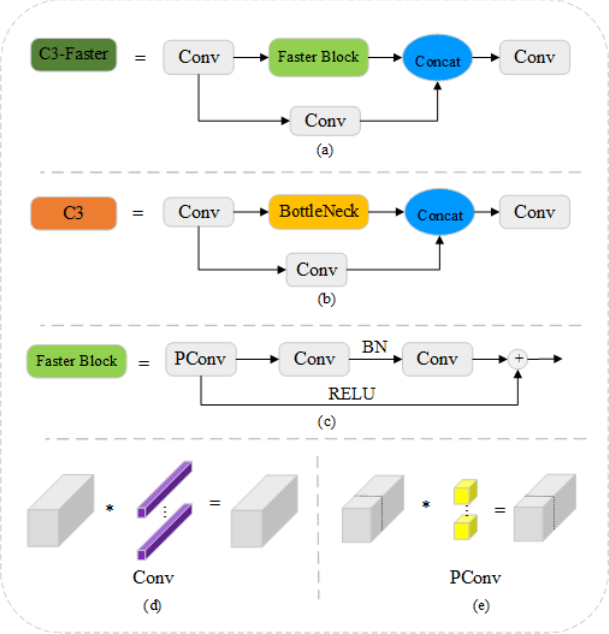

Fig. 2. Illustration of the C3-Faster module: (a) C3-Faster module; (b) C3 module; (c) Faster Block; (d) Conv module; (e) PConv.

Traditional object detection models for small objects are mainly developed and researched for relatively regular-sized objects. The IoU metric's sensitivity varies greatly for objects of different scales. In the YOLOv7 base model, the GIoU metric between boxes has such a problem when detecting small objects. Smaller position offsets lead to a sharp decrease in IoU values for small objects, but for large objects, the IoU values only change slightly for the same position offset. Therefore, the sensitivity of GIoU on small objects will cause the labels to easily become opposite, resulting in similarities between the features of positive and negative samples, making it difficult for the network to converge. Secondly, using GIoU as a metric, each Ground Truth (GT) The average number of positive samples distributed is less than 1. This is because the GIoU between some GTs and any anchor is less than the threshold. During training, there will be a lack of supervision information for small objects, thus reducing detection performance. Therefore, we choose NWD as the loss function for training and evaluating our model, which is not measured by intersection over union (IoU). NWD models bounding boxes using a two-dimensional Gaussian distribution and predict the similarity between object boxes and labeled object boxes through the corresponding Gaussian distribution. This makes it insensitive to the scale of the object. By modeling the predicted and the ground truth bounding boxes as Gaussian distributions, NWD provides an efficient and accurate method to measure the similarity between them. Its measurement method can be effectively applied to small objects, even when they do not overlap, which overcomes the limitations of using IoU as a loss function for detecting small objects. In addition, the scale invariance of NWD makes it more suitable for measuring similarities between small objects. Therefore, using it in small object detection in remote sensing images is more suitable. Therefore, this article integrates the NWD loss function with GIoU to improve the detection performance of the network. The NWD algorithm formula is defined as follows: This algorithm uses the Wasserstein distance in optimal transmission theory to calculate the distance between two distributions. For two 2D Gaussian distributions, $\mu_1 = N(m_1, \sum_1)$ and $\mu_2 = N(m_2, \sum_2)$, its second-order Wasserstein distance can be defined as :

$$W_2^2(\mu_1, \mu_2) = \|m_1 - m_2\|_2^2 + \left\|\sum_1^{1/2} - \sum_2^{1/2}\right\|_F^2 \quad (3)$$

Where $\|\|_F$ is Frobenius norm. Furthermore, for the Gaussian distributions N_a and N_b modeled according to the bounding boxes $A = (cx_a, cy_a, cw_a, ch_a)$ and $B = (cx_b, cy_b, cw_b, ch_b)$, the second-order Wasserstein distance can be further simplified to:

$$W_2^2(N_a, N_b) = \left\|([cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2}]^T, [cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2}]^T)\right\|_2^2 \quad (4)$$

However, $W_2^2(N_a, N_b)$ is a distance metric and cannot be used directly as a similarity metric. Therefore, we normalize using its exponential form and obtain a new metric called the Normalized Wasserstein Distance (NWD):

$$NWD(N_a, N_b) = exp(-\frac{\sqrt{W_2^2(N_a, N_b)}}{C}) \quad (5)$$

Where C is a constant closely related to the data set. The calculation expression of the loss function based on NWD is shown in equation (6):

$$L_{NWD} = 1 - NWD(N_g, N_t) \quad (6)$$

In the YOLOv7 model, the default choice for calculating the positioning loss is GIoU. In most cases, GIoU has the smallest regression error. However, considering that the data set used in this article is not entirely composed of small objects, L_{NWD} is not directly used to replace L_{GIoU}. By assigning appropriate fusion weights r to L_{NWD} and L_{GIoU}, the NWD-GIoU loss function is proposed as a measurement criterion. The calculation expression of NWD-GIoU is shown in equation (7):

$$L_{NWD-GIoU} = r \cdot L_{GIoU} + (1 - r) \cdot L_{NWD} \quad (7)$$

D. Coordinate Attention Module

The background of remote sensing images usually includes rich terrain and environment, as well as many multiple small objects. It isn't easy to extract effective features from these images. Therefore, it is crucial to introduce effective focus in important areas. Attention modules are widely used in deep learning to better focus on useful information and then enhance feature extraction capabilities. Traditional attention mechanisms calculate weights for each channel in order to enhance the model's performance. However, this introduces additional parameters to the model. The CA attention mechanism is simple and lightweight enough to fully utilize the extracted position information and effectively handle inter-channel relationships, improving the model's accuracy. The CA attention mechanism embeds position information into channel attention, which is compact and can be flexibly integrated into other classic mobile networks with almost no computational overhead. So we integrated Coordinate Attention into our modified YOLOv7 model to enhance object detection performance. Our experiments demonstrate that incorporating the CA mechanism into the YOLOv7

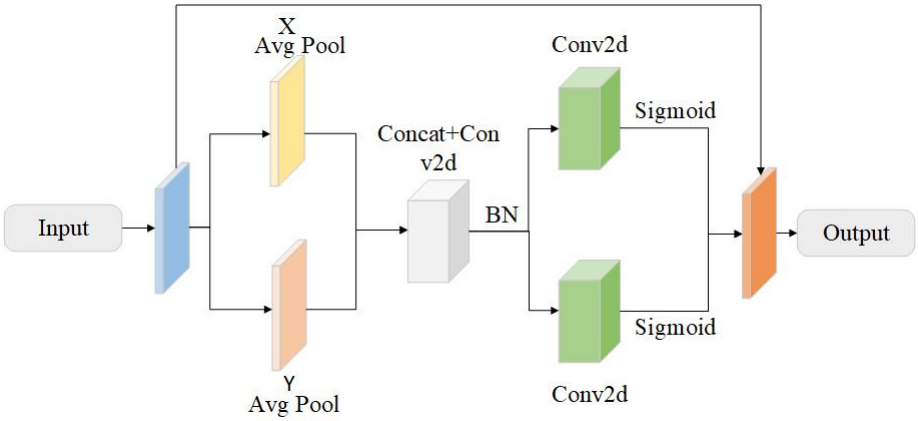

Fig. 3. Visual representation of the CA mechanism's structure.

structure enables the model to extract feature information of objects more effectively during the detection process, without increasing the number of parameters in the original network. Fig. 3 provides a visual representation of the CA mechanism's structure.

IV. Experiments Validation

To validate our proposed model, we performed comparative and ablation experiments on standard datasets. Next, we will introduce the data set, evaluation indicators, ablation experiments and comparative experiments in detail.

A. Datasets

We selected three data sets in the experiment: Dior[31], AI-TOD[32], and VisDrone[33] for small object detection in remote sensing images. The Dior dataset contains 20 categories and consists of 23,463 images and 190,288 instances. Dior-Vehicle is a separate category in the Dior data set. Most data sets are small objects, and car objects that consume less than 10 pixels are marked. AI-TOD is a very typical dataset of tiny objects in aerial images containing fewer tiny object pixels. AI-TOD provides 700,621 object instances for 8 categories in 28,036 aerial images, distributed among 8,605 In common categories, compared with existing aerial image object detection data sets, the average size of objects in AI-TOD is about 12.8 pixels, which is much smaller than other data sets and is easily confused with the background, making it very suitable for small object detection tasks. The VisDrone dataset was captured by the AISKYEYE team using various drones in 14 different cities across China. It comprises 288 videos, 261,908 video frames, and 10,209 static images. Over 2.6 million bounding boxes of interested objects were manually annotated, with a primary focus on pedestrians and vehicles, which were further classified into 10 categories.

B. Experiment Setup

The experimental environment of this work is based on the Windows 10. The environment is CUDA v11.0, the Pytorch version is v1.7, and the Python version is 3.6. The CPU is Intel(R) Xeon(R) Gold 5218R CPU @ 2.10 GHz, the GPU is Quadro GV100, the memory is 32GB, the input image size is 640 * 640 px, initial learning is 0.01, batch size 16, momentum is 0.937, weight Decay is 0.0005, The total epoch is 300.

TABLE I
CATEGORY NAME AND CORRESPONDING ABBREVIATIONS OF THE AI-TOD

AI	airplane	BR	bridge
ST	storage tank	SH	ship
SP	swimming pool	VE	vehicle
PE	person	WM	wind mill

C. Evaluation metrics

We use detection accuracy and detection speed as the evaluation indicators. Precision (P) mainly measures the degree of error detection by the model; recall rate (R) mainly measures the degree of failure to detect by the model; average precision (AP) is the area under the P-R curve; mAP is the average AP of all categories. The calculation methods are as follows:

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

$$AP = \int_0^1 P(R)d(R) \tag{10}$$

$$mAP = \frac{1}{n}\sum_{j=1}^{n} AP_{(j)} \tag{11}$$

Among them, TP is a true positive, TN is a true negative, FP is a false positive, and FN is a false negative. The n is the number of categories; $AP_{(j)}$ represents the precision of each category. The model complexity uses Params. The specific calculation equation is:

$$Params = C_o \times (K_w \times K_h \times C_i + 1) \tag{12}$$

where C_o represents the number of output channels, C_i represents the number of input channels, and K_w and K_h represent the width and height of the convolution kernel respectively.

D. Comparison Experiments

To display the results more clearly, we used the abbreviation of the category name to represent each category in AI-TOD, Dior and VisDrone datasets, as shown in TABLE I, Table II and Table III respectively.

TABLE II
CATEGORY NAME AND CORRESPONDING ABBREVIATIONS OF THE DIOR DATASET.

AE	Airplane	DA	Dam
AT	Airport	HA	Harbor
BF	Baseball field	OV	Overpass
BC	Basketball court	SH	Ship
BR	Bridge	SD	Stadium
CH	Chimney	ST	Storage tank
GTF	Ground track field	TC	Tennis court
ESA	Expressway service area	TS	Train station

TABLE III
CATEGORY NAME AND CORRESPONDING ABBREVIATIONS OF THE VISDRONE DATASET.

PD	Pedestrian	TK	Truck
PE	People	TC	Tricycle
BI	Bicycle	AT	Aw-tricycle
CA	Car	BU	Bus
VA	Van	MO	Motor

1) Results of the AI-TOD Dataset: As shown in TABLE IV, our model is tested on the AI-TOD dataset and compared with other methods, including anchor-based and anchor-free detectors. In addition to the previously mentioned YOLO series and Faster R-CNN, there are also SSD [34], RetinaNet [29], TridentNet [35], FoveaBox [36], RepPoints [37] and CornerNet [38]. Referring to the experimental results, it is obvious that when IoU is set to 0.5, our YOLO-FNC reaches a performance of 35.90% AP; compared with the baseline model YOLOv7, we exceed 14.20% AP. When IoU is greater than 0.5 and less than 0.95, YOLO-FNC achieves 15.48% AP performance. Compared with the baseline model YOLOv7, we exceed 6.45% AP. In both cases, our model also greatly outperforms other methods.

Comparing anchorless and anchor detectors, our model achieves the best detection performance in 5 categories, especially in some difficult tasks (e.g., vehicle, ship, person), which strongly proves the effectiveness of our method. The visual detection results are shown in Fig. 4. It is observed that YOLO-FNC not only generates accurate bounding boxes for small objects but also works robustly in various challenging scenarios - for example, densely arranged scenes, scenes with low background contrast, etc. Our methods can all successfully detect and locate objects with satisfactory accuracy.

2) Results of the Dior Dataset: As seen from TABLE V, in the test results of all models on the Dior data set, the accuracy in multiple small object categories such as cars, ships, and storage tanks has achieved relatively good results. The results marked in bold are the best results among all compared models. For larger category objects, such as baseball fields, basketball courts, etc., the detection accuracy of this model is also improved compared to other algorithms. This demonstrates the strong applicability of the algorithm proposed in this paper. Compared to the YOLOv7 baseline model, most categories in our tests are more accurate. This better proves that our algorithm is a relatively excellent object detection algorithm.

This is to demonstrate better the excellent performance of YOLO-FNC in small object detection in remote sensing images. This paper selects three types of scenes containing small objects from the Dior data set for detection and comparison. These three scenarios are ship detection under dense distribution, vehicle detection under partial shadow occlusion, and ground track field detection under complex background. We used YOLOv7 as the baseline comparison model. Fig. 5 shows the detection comparison between YOLOv7 and YOLO-FNC algorithms on the Dior data set. The detection results of YOLOv7 and YOLO-FNC are shown in the top line and bottom line respectively. The yellow thick line box represents the false detection object, and the red thick line box represents the missed detection object.

As shown in Fig.5, the red box in the first line of baseline represents missed detection, and the yellow box represents false detection. YOLO-FNC detects more small objects than YOLOv7. Moreover, missed detections and false detections are significantly reduced. As shown in Fig. 5(Left), YOLOv7 missed detecting objects with small pixels and dense distribution in the densely distributed ship detection scenario. However, YOLO-FNC detected more ships. Fig. 5(Middle) shows the detection of small objects obscured by shadows. Even when objects are occluded, YOLO-FNC can still detect smaller cars. On the contrary, YOLOv7 failed to detect the car in the shadow and misdetected the object. Fig. 5(Right) see that for object detection in complex backgrounds, YOLO-FNC can still detect track and field fields with smaller pixels in the image. However, YOLOv7 missed the detection of track and field fields with smaller pixels. In summary, in the task of small object detection in remote sensing images, YOLOv7 shows poor detection performance, high miss detection rate, and false detection rate. The YOLO-FNC model proposed in this article improves the detection performance in densely distributed and complex scenes.

3) Results of the VisDrone Dataset: The experimental results are shown in Table VI. Our model has achieved good results in multiple categories. And the mAP value of the YOLO-FNC algorithm is 52.6%, which is the best performance among the compared algorithms. This effectively proves the generality of YOLO-FNC.

E. Ablation Experiments

To illustrate the effectiveness of each part of our proposed work, we conducted ablation experiments on the three datasets. Among them, the DIOR data set contains a wide variety of objects, including a small number of medium and large-sized objects. In order to evaluate the performance on small objects, we selected the Vehicle class in the Dior data set to conduct ablation experiments. This category has a high percentage of all small objects in the dataset. The DIOR vehicle includes 6,421 images and over 32,000 objects. In order to further verify the effectiveness of the improved module, under the same experimental conditions, we conducted experiments on a larger AI-TOD data set with prominent small objects.

The results of ablation experiments on the DIOR-Vehicle data set are shown in TABLE VII. We used YOLOv7 as a baseline and expand upon it. By adding only the C3-Faster module based on the YOLOv7 algorithm, AP has been significantly improved, and the model parameters have been reduced from 37.19M to 33.28M. Subsequently, only the NWD module was introduced in YOLOv7, the precision and recall rate were slightly improved, and other indicators

TABLE IV
COMPARISON RESULTS OF ALGORITHMS ON AI-TOD DATASET

Method	AI	BR	ST	SH	SP	VE	PE	WM	AP@.5:.95	AP@.5
Faster R-CNN	22.71	3.87	20.18	19.02	8.90	11.88	4.49	0.32	11.40	27.00
YOLOv3	7.14	2.60	3.66	10.69	0.61	8.50	2.13	0.40	4.50	14.20
YOLOv5	9.29	6.37	20.28	32.86	0.86	19.28	5.28	0.70	11.90	28.50
YOLOv7	4.52	0	30.40	15.30	0	19.90	2.14	0	9.03	21.70
RetinaNet	0.01	6.62	1.84	20.87	0.06	5.67	1.75	0.53	4.70	13.60
SSD	14.52	3.13	10.89	13.05	1.92	7.84	3.12	1.48	7.00	21.70
TridentNet	9.67	0.77	12.28	17.11	3.20	11.87	3.98	0.94	7.50	20.90
FoveaBox	13.75	0.00	18.51	17.70	0.03	11.42	3.38	0.00	8.10	19.80
RepPoints	2.92	2.34	21.37	26.40	0.00	15.16	5.39	0.00	9.20	23.60
CornerNet	10.63	11.81	14.05	16.94	4.41	6.96	4.82	2.67	9.00	25.40
Ours	16.90	4.07	38.00	27.20	0.56	27.50	6.59	3.03	15.48	35.90

Fig. 4. Detection results were achieved by our YOLO-FNC on the AI-TOD set. One color stands for one object class.

were slightly improved. We also embedded the CA attention mechanism at an appropriate location in YOLOv7, which significantly increased the accuracy by 3.4% compared with the baseline model. When the three modules are added to the model in pairs, it can be seen that multiple indicators have improved. Finally, three modules introduced YOLOv7

TABLE V
COMPARISON RESULTS OF ALGORITHMS ON DIOR DATASE

Category	SSD	PANet	RetainNet	YOLOv3	YOLOv4	YOLOv5	YOLOv7	Ours
AE	59.5	61.9	53.3	72.2	96.0	94.3	95.7	95.2
AT	72.7	70.4	77.0	29.2	87.9	84.0	82.9	86.0
BF	72.4	71.0	69.3	74.0	94.7	94.1	96.3	97.1
BC	75.7	80.4	85.0	78.6	91.9	86.8	87.4	87.8
BR	29.7	38.9	44.1	31.2	60.0	57.6	58.8	59.3
CH	65.8	72.5	73.2	69.7	90.8	91.3	91.4	93.1
DA	56.6	56.6	62.4	26.9	69.9	73.7	68.1	72.4
ESA	63.5	68.4	78.6	48.6	92.1	74.2	86.1	88.7
ETS	53.1	60.0	62.8	54.4	87.5	73.0	75.8	80.0
GC	65.3	69.0	78.6	31.1	87.6	80.8	82.8	83.6
GTF	68.6	74.6	96.6	61.1	83.7	83.6	88.8	88.3
HA	49.4	41.6	49.9	44.9	55.6	70.6	73.4	75.5
OV	48.1	55.8	59.6	49.7	68.7	68.1	70.9	70.8
SH	59.2	71.7	71.1	87.4	94.6	94.5	95.3	95.4
SD	61.0	72.9	68.4	70.6	83.8	92.3	96.7	97.2
ST	46.6	62.3	45.8	68.7	88.4	85.1	86.2	86.2
TC	76.3	81.2	81.3	87.3	95.7	93.0	95.1	95.4
TS	55.1	54.6	54.2	29.4	44.4	66.3	61.6	66.3
VE	27.4	48.2	45.1	48.3	62.1	80.6	82.4	83.6
WI	65.7	86.7	83.4	78.7	90.4	83.2	85.2	85.4
mAP(%)	58.6	63.8	65.7	57.1	81.3	81.4	83.0	84.4

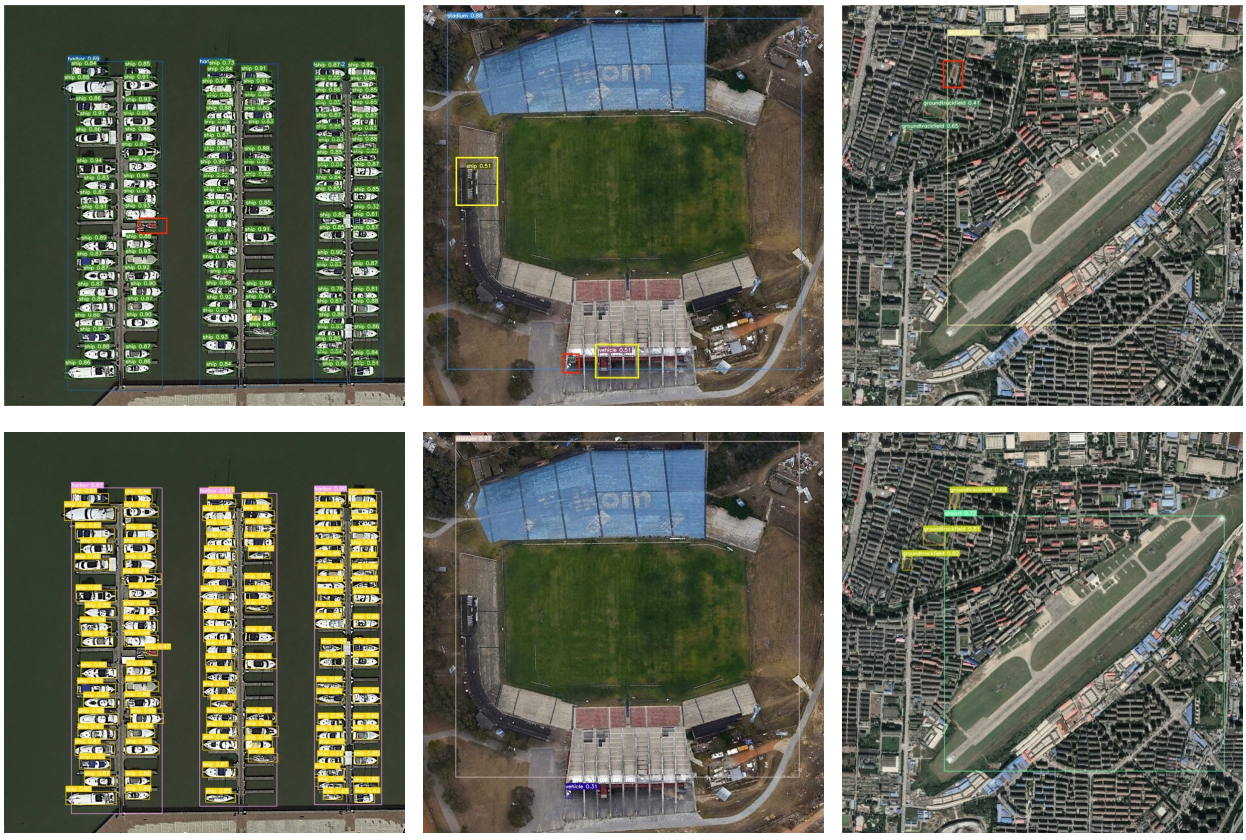

Fig. 5. Comparison of small object detection results on DIOR dataset. Left: ship detection under dense distributed environments. Middle: vehicle detection under partial shadow occlusion. Right: ground track field detection under complex background.

at the same time. Compared with the YOLOv7 baseline model, precision increased by 1.64%, recall increased by 4.1%, mAP@0.5 increased by 3.47%, mAP@.5:.95 increased by 3.79%, and parameters decreased by 4.69 M.

The results of ablation experiments on the AI-TOD data set are shown in TABLE VIII. The experimental conditions are the same as above. The introduction of the C3-Faster module has improved precision, recall, and mAP, and reduced the amount of parameters. With the introduction of NWD, mAP increased most significantly, mAP@0.5 increased from 21.75 to 33.9, and mAP@.5:.95 increased from 9.03 to 14.02. The embedding of the CA attention mechanism has also significantly improved various indicators. The experimental results are significantly improved when each module is embedded in the baseline model in pairs. Finally, the three modules were embedded in YOLOv7 at the same time. Compared with the YOLOv7 baseline model, precision increased by 3.32%, recall increased by 10.81%, mAP@0.5 increased by 14.15%, mAP@.5:.95 increased by 6.45%, and parameters decreased by 3.89 M.

TABLE VI
COMPARISON RESULTS OF ALGORITHMS ON VISDRONE DATASE

Method	PD	PE	BI	CA	VA	TK	TC	AT	BU	MO	mAP(%)
Faster R-CNN	20.9	14.8	7.3	51.0	29.7	19.5	14.0	8.8	30.5	21.2	21.8
RetinaNet	27.0	13.0	14.0	59.0	50.0	54.0	25.0	30.0	59.0	35.0	35.6
YOLOv3	22.3	20.6	6.7	59.7	21.2	21.3	10.1	6.9	37.9	23.7	23.0
YOLOv4	24.8	12.6	8.6	64.3	22.4	22.7	11.4	7.6	44.3	21.7	30.7
YOLOv5-TPH	29.0	16.7	15.6	68.9	49.7	45.1	27.0	24.7	61.8	30.9	37.3
YOLOv7	54.1	36.8	23.4	82.0	57.8	59.7	34.5	30.2	72.8	51.6	50.3
Ours	57.3	39.2	24.9	83.3	59.7	62.2	35.8	33.8	74.6	55.0	52.6

TABLE VII
ABLATION EXPERIMENTAL ON THE DIOR-VEHICLE DATASET

Baseline	C3-Faster	NWD-GIoU	CA	Precise(%)	Recall(%)	AP@0.5(%)	AP@.5:.95(%)	Parameters(M)
✓				80.76	67.10	76.67	42.52	37.19
✓	✓			81.01	68.35	78.16	43.76	33.28
✓		✓		79.79	68.01	76.69	42.48	37.19
✓			✓	84.21	64.05	76.23	42.22	37.23
✓		✓	✓	82.78	67.58	78.13	45.13	36.51
✓	✓		✓	82.35	70.15	79.54	45.97	32.60
✓	✓	✓		81.69	69.50	79.79	46.07	32.56
✓	✓	✓	✓	82.40	71.20	80.14	46.31	32.50

TABLE VIII
ABLATION EXPERIMENTAL ON THE AI-TOD DATASET

Baseline	C3-Faster	NWD	CA	Precise(%)	Recall(%)	AP@0.5(%)	AP@.5:.95(%)	Params(M)
✓				70.36	23.46	21.75	9.03	37.23
✓	✓			75.48	25.48	25.82	10.74	33.32
✓		✓		70.04	32.67	33.9	14.02	37.23
✓			✓	78.74	26.73	27.95	11.51	37.26
✓		✓	✓	70.29	35.64	34.9	14.83	36.49
✓	✓		✓	77.87	30.38	30.95	12.89	33.35
✓	✓	✓		76.25	33.92	34.42	14.47	33.32
✓	✓	✓	✓	73.68	34.27	35.9	15.48	33.34

TABLE IX
ABLATION EXPERIMENTAL ON THE VISDRONE DATASET

Baseline	C3-Faster	NWD	CA	Precise(%)	Recall(%)	AP@0.5(%)	AP@.5:.95(%)	Params(M)
✓				59.20	50.60	50.29	27.89	37.24
✓	✓			60.02	50.95	51.53	28.46	33.32
✓		✓		59.52	53.23	51.62	29.46	37.23
✓			✓	58.55	50.85	50.34	27.84	37.26
✓		✓	✓	61.13	50.91	51.99	28.76	36.49
✓	✓		✓	59.09	52.17	51.70	28.56	33.36
✓	✓	✓		62.22	50.09	51.59	28.40	33.33
✓	✓	✓	✓	60.76	52.30	52.57	29.42	33.34

The results of ablation experiments on the VisDrone dataset are shown in TABLE IX. The ablation experimental steps on this data set are the same as above. By embedding the three modules individually and in pairs into the model, multiple detection indicators have been effectively improved. When the three modules are embedded in the model at the same time, the precision and recall rate are improved, mAP@0.5 and mAP@.5:95 reach 52.57% and 29.42% respectively, and the number of parameters also decreases.

In summary, the experimental results above show the effectiveness of each part of our work. Our model has shown good detection performance in three datasets.

V. CONCLUSION

To solve the problem of low detection accuracy and serious missed detection of small objects in remote sensing images, this article proposes an improved object detection network based on YOLOv7. Firstly, introducing the C3-Faster module into the YOLOv7 network can simply and effectively extract spatial features and enhance the ability to extract object features. Secondly, we introduced NWD combined with GIoU as the position regression loss function in the network to improve the detection effect of small objects in remote sensing images. Finally, we embed CA into the YOLOv7 model, which cannot only focus on channel information and position information at the same time, reducing redundant feature information, but is also flexible and lightweight enough to improve the detection performance of the network effectively.

In this article, we performed ablation experiments on the DIOR-Vehicle and AI-TOD, VisDrone datasets and compared them with YOLOv7. After a series of ablation experiments, we proved that each part of our proposed work is correct and feasible. In comparative experiments, we

performed small object detection on the Dior dataset and AI-TOD dataset, as well as the VisDrone dataset. Experimental results show that YOLO-FNC has obvious advantages. The mAP of YOLO-FNC on the Dior data set reached 84.4%, the mAP on the AI-TOD data set reached 35.9%, and the mAP on the VisDrone reached 52.6%. The results are better than those of other compared algorithms. It effectively proves the feasibility and superiority of the algorithm proposed in this article. In future work, we will further explore more lightweight and effective network models to improve the performance of small object detection on remote sensing images.

REFERENCES

[1] L. Tan, X. Lv, X. Lian and G. Wang, "YOLOv4_Drone: UAV image target detection based on an improved YOLOv4 algorithm," *Computers & Electrical Engineering*, vol. 93, 107261, 2021.

[2] J. Zhang, R. Wang, R. Liu, D. Guo, B. Li, and S. Chen, "DSP-Based Traffic Target Detection for Intelligent Transportation," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1-12, 2022.

[3] Y. Wang, X. Wang, Z. Wang, J. Liu, and S. Xu, "Aerospace target detection based on complex background," in *2020 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pp. 505-510, 2020.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[5] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks," in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

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

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi,"You Only Look Once: Unified, Real-Time Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779-788, 2016.

[8] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement,"*arXiv:1804.02767*, 2018.

[9] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "Yolov4: Optimal Speed and Accuracy of Object Detection," *arXiv preprint arXiv:2004.10934*, 2020.

[10] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 2778-2788, 2021.

[11] S. Luo, J. Yu, Y. Xi, and X. Liao, "Aircraft Target Detection in Remote Sensing Images Based on Improved YOLOv5," *IEEE Access*, vol. 10, pp. 5184-5192, 2022.

[12] Z. Liu, Y. Gao, Q. Du, M. Chen, and W. Lv, "YOLO-Extract: Improved YOLOv5 for Aircraft Object Detection in Remote Sensing Images," *IEEE Access*, vol. 11, pp. 1742-1751, 2023.

[13] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7464-7475, 2023.

[14] J. Chen, S. H. Kao, H. He, W. Zhuo, S. Wen, C. H. Lee and S. H.G. Chan, "Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12021-12031, 2023.

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

[16] Q. Hou, D. Zhou, and J. Feng, "Coordinate Attention for Efficient Mobile Network Design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13713-13722, 2021.

[17] J. Han, J. Ding, J. Li, and G. S. Xia, "Align Deep Features for Oriented Object Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-11, 2022.

[18] X. Yang, J. Yan, Z. Feng, and T. He, "R3det: Refined Single-stage Detector with Feature Refinement for Rotating Object," in *Proceedings of the AAAI conference on Artificial Intelligence*, vol. 35, no. 4, pp. 3163-3171, 2021.

[19] J. Pang, C. Li, J. Shi, Z. Xu, and H. Feng, "R2-CNN: Fast Tiny Object Detection in Large-scale Remote Sensing Images,"*IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5512-5524, 2019.

[20] H. Zhou, L. Wei, C. P. Lim, S. Nahavandi, and R. Sensing, "Robust Vehicle Detection in Aerial Images using Bag-of-words and Orientation Aware Scanning,"*IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 12, pp. 7074-7085, 2018.

[21] Z. Chen, C. Wang, C. L Wen, X. H. Teng, Y. P. Chen, H. Y Guan, H. Luo, L. J. Cao and J. Li, "Vehicle Detection in High-resolution Aerial Images via Sparse Representation and Superpixels,"*IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 1, pp. 103-116, 2015.

[22] Z. Zhang, Y. Liu, T. Liu, Z. Lin, and S. Wang, "DAGN: A Real-Time UAV Remote Sensing Image Vehicle Detection Framework," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 11, pp. 1884-1888, 2020.

[23] F. C. Akyon, S. O. Altinuc, and A. Temizel, "Slicing Aided Hyper Inference and Fine-tuning for Small Object Detection," in *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 966-970, 2022.

[24] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for Small Object Detection,"*arXiv preprint arXiv:1902.07296*, 2019.

[25] W. Pei, Z. Shi, and K. Gong, "Small Target Detection with Remote Sensing Images based on an Improved YOLOv5 Algorithm," *Frontiers in Neurorobotics*, vol. 16, 2022.

[26] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric Contextual Modulation for Infrared Small Target Detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 950-959, 2021.

[27] J. Rabbi, N. Ray, M. Schubert, S. Chowdhury, and D. Chao, "Small-object Detection in Remote Sensing Images with End-to-end Edge-enhanced GAN and Object Detector Network," *Remote Sensing*, vol. 12, no. 9, 2020.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[29] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980-2988, 2017.

[30] 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," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9759-9768, 2020.

[31] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object Detection in Optical Remote Sensing Images: A Survey and A New Benchmark," *ISPRS journal of photogrammetry and remote sensing*, vol. 159, pp. 296-307, 2020.

[32] J. Wang, W. Yang, H. Guo, R. Zhang, and G. S. Xia, "Tiny Object Detection in Aerial Images," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 3791-3798, 2021.

[33] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and Tracking Meet Drones Challenge," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 44, no. 11, pp. 7380-7399, 2021.

[34] A. C. Berg, C. Y. Fu, C. Szegedy, D. Anguelov, D. Erhan, S. Reed and W. Liu, "SSD: Single Shot Multibox Detector," in *Computer Vision-ECCV 2016: 14th European Conference*, pp. 21-37, 2016.

[35] Y. Li, Y. Chen, N. Wang, and Z. Zhang, "Scale-aware Trident Networks for Object Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6054-6063, 2019.

[36] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "FoveaBox: Beyound Anchor-Based Object Detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 7389-7398, 2020.

[37] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point Set Representation for Object Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9657-9666, 2019.

[38] H. Law and J. Deng, "Cornernet: Detecting Objects as Paired Keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 734-750, 2018.