Wuhan University
Journal of Natural Sciences

2024, Vol.29 No.5, 439-452

Article ID 1007-1202(2024)05-0439-14 DOI https://doi.org/10.1051/wujns/2024295439

Cite this article: LIU Beibei, DENG Yansong, LYU He, *et al*. Improved YOLOX Remote Sensing Image Object Detection Algorithm[J]. *Wuhan Univ J of Nat Sci*, 2024, 29(5): 439-452.

Improved YOLOX Remote Sensing Image Object Detection Algorithm

LIU Beibei, DENG Yansong†, LYU He, ZHOU Chenchen, TANG Xuezhi, XIANG Wei

College of Electronic and Information, Southwest Minzu University, Chengdu 610041, Sichuan, China

Abstract: Remote sensing image object detection is one of the core tasks of remote sensing image processing. In recent years, with the development of deep learning, great progress has been made in object detection in remote sensing. However, the problems of dense small targets, complex backgrounds and poor target positioning accuracy in remote sensing images make the detection of remote sensing targets still difficult. In order to solve these problems, this research proposes a remote sensing image object detection algorithm based on improved YOLOX-S. Firstly, the Efficient Channel Attention (ECA) module is introduced to improve the network's ability to extract features in the image and suppress useless information such as background; Secondly, the loss function is optimized to improve the regression accuracy of the target bounding box. We evaluate the effectiveness of our algorithm on the NWPU VHR-10 remote sensing image dataset, the experimental results show that the detection accuracy of the algorithm can reach 95.5%, without increasing the amount of parameters. It is significantly improved compared with that of the original YOLOX-S network, and the detection performance is much better than that of some other mainstream remote sensing image detection methods. Besides, our method also shows good generalization detection performance in experiments on aircraft images in the RSOD dataset.

Key words: remote sensing images; object detection; YOLOX-S; attention module; loss function

CLC number: TP751

0 Introduction

In recent years, with the development of space remote sensing technology, the acquisition of high-resolution remote sensing images[1] has become more convenient. Object detection in remote sensing images has become one of the current research hotspots. Research on object detection technology in remote sensing images can provide important information for disaster prediction, national defense security and smart city construction. In past researches, many excellent methods have been proposed to deal with this task, especially deep learning-based methods and traditional methods. These studies have greatly promoted the development of remote sensing object detection technology and brought considerable economic value in many fields[2,3].

Traditional remote sensing image object detection methods are usually image-based processing methods. In brief, threshold segmentation is firstly used for image preprocessing, then the Scale Invariant Feature Transform (SIFT) method is used to extract image features. Finally, template matching and methods based on shallow learning are used to judge the target. Over the past decades, researchers have developed many traditional

Received date: 2023-09-01

Foundation item: Supported by the National Natural Science Foundation of China (72174172, 71774134) and the Fundamental Research Funds for Central University, Southwest Minzu University (2022NYXXS094)

Biography: LIU Beibei, male, Master candidate, research direction: object detection, deep learning, E-mail: liubeibei199809@outlook.com

† Corresponding author. E-mail: 21500059@swun.edu.cn

methods to deal with this problem, including template matching sliding window, Histogram of Oriented Gradients (HOG) and Bag of Words (BOW)[4], etc. However, these traditional methods require manual design of feature priors, leading to extensive time spent on feature engineering for specific tasks. Consequently, this results in long processing times, low detection accuracy and poor generalization ability. In contrast, deep learning-based object detection algorithms automatically extract object features and directly predict object category and location information on the feature map, significantly enhancing the detection performance and efficiency. Deep learning-based object detection algorithms can be divided into two categories according to whether they include a region proposal generation stage. One category is two-stage object detection algorithms[5], such as Region-based Convolutional Neural Network (R-CNN), Spatial Pyramid Pooling Network (SPP-Net) [6], Fast R-CNN[7] and Faster R-CNN[8]. The other category is single-stage object detection algorithms[9], such as Single Shot MultiBox Detector (SSD)[10], RetinaNet[11], and You Only Look Once (YOLO) algorithm[12]. Although deep learning-based object detection algorithms exhibit good performance, remote sensing images are captured from a bird's-eye view at different altitudes. The backgrounds in these images are complex and the sizes of the objects are generally small, which introduces interference to the detection and recognition of the object.

To solve the above issues, scholars have conducted a series of researches. Cheng *et al*[13] proposed the Remote Sensing Image Convolutional Neural Network (RICNN) model, which effectively dealt with the problem of target rotation changes in remote sensing images. Yang *et al*[14] aimed at small targets densely arranged in any direction in remote sensing images, and improved the detection ability of the model for targets in complex backgrounds by introducing a supervised multi-dimensional attention network. Wang *et al*[15] used the Adaptive Spatial Feature Pyramid (ASFP) module to fuse multi-scale feature information to enhance the information interaction of different scale feature layers. Nayan *et al*[16] utilized upsampling and skip connections to extract multi-scale features at different depths during the training process to improve the accuracy of small object detection. Xi *et al*[17] proposed a dual-stream representation learning generative adversarial network by recovering the missing information of the high-frequency and low-frequency components in high-resolution images, which in turn improved the detection accuracy of small objects in low-resolution aerial images. Nie *et al*[18] proposed an improved ship detection method based on Mask R-CNN, which enhanced the accurate detection of equipment in remote sensing images.

Although the above methods exhibit satisfactory detection performance on remote sensing images, the detection efficiency of small objects in complex backgrounds remains unsatisfactory. Thus, to address this issue, this research proposes a lightweight remote sensing image object detection network model which utilizes YOLOX-S as the basic network model. And it is optimized for different subproblems. The main contributions of this research are summarized as follows:

1) We employ the YOLOX model, which has demonstrated commendable results in various scene detection tasks, for target object detection in remote sensing images. On the basis of YOLOX model, we integrate the Efficient Channel Attention (ECA) mechanism to improve the network's ability in extracting important features, reducing the influence of redundant information such as background, and ultimately improving the detection accuracy of small objects in remote sensing images.

2) We optimize the loss function of the YOLOX-S model. Specifically, for the model's boundary regression loss, we employ the Alpha Intersection over Union (α-IoU) loss function as a replacement for the IoU loss function to improve robustness to small sample datasets and noise.

3) Experimental evaluations conducted on the NWPU VHR-10 remote sensing image dataset demonstrate that our method can achieve better detection performance than most mainstream methods. Additionally, we conduct generalization experiments on aircraft images in the RSOD dataset, and achieve a good detection effect. These results show that our proposed model has a good generalization ability.

1 Related Work

1.1 Remote Sensing Image Detector Based on YOLO Series Algorithm

The YOLO object detection algorithm was firstly proposed by Joseph Redmon. It constitutes an end-to-end network model that directly predicts target object bounding boxes and categories. YOLO redefines object detection as a regression problem, employing a single Convolutional Neural Network (CNN) applied to the

whole image. This network divides the image into grids and predicts the class probability and bounding box coordinates for each grid. YOLOv3[20] was proposed in 2018. It uses the global region of the image for training, which can better distinguish the target from the background. Based on YOLOv3, Cao *et al*[21] integrated a 104×104 detection scale into the improved model to enhance the sensitivity and detection ability of the network towards small objects such as planes and ships in remote sensing images. In April 2020, Bochkovskiy *et al*[22] proposed YOLOv4 based on YOLOv3, incorporating Cross Stage Partial Network (CSP) and Path Aggregation Network (PAN) structures, and adopting some practical techniques to achieve a balance between detection speed and accuracy. Yu *et al*[23] introduced feature layer scaling to YOLOv4 to improve the detection accuracy of bridge objects in remote sensing images. In June 2020, Ultralytics introduced a new object detection network framework, YOLOv5, in which some improvements in details have been implemented based on YOLOv4. In order to solve the problem of category label imbalance caused by sparse distribution of objects in remote sensing images, Zhao *et al*[24] integrated the Aggregated-Mosaic method into the YOLOv5 model, thereby enhancing the stability of training and inference processes.

1.2 Attention Mechanism

In computer vision tasks, such as image recognition, semantic segmentation and object detection, attention mechanism can enhance essential features of images while suppressing some irrelevant features to improve the model accuracy. In 2018, Woo *et al*[25] proposed inferring attention maps along two independent dimensions of channel and space to strengthen target features and improve object detection accuracy. Wang *et al*[26] designed an encoder-decoder module and constructed a residual attention network based on it. Experimental results show that better outputs can be achieved by continuously refining the feature maps. Inspired by these findings, we incorporated attention mechanism into our network. The experimental results indicate that these attention mechanisms can significantly enhance detection accuracy.

2 Method

2.1 Analysis of YOLOX Network

YOLOX[27] inherits the developmental ideas of the previous YOLO series network, effectively achieving a balance between detection accuracy and processing speed, thus rendering it highly adaptable for remote sensing image object detection. YOLOX network structure has four models: YOLOX-S, YOLOX-M, YOLOX-L and YOLOX-X. Considering the model reasoning speed and lightweight, we choose YOLOX-S as the basic model.

The structure of YOLOX network is divided into three key components: Backbone, Neck and Output. The Backbone network is Cross Stage Partial Darknet (CSP-Darknet). Through multiple convolutional and pooling processing within the Backbone network, feature maps of different sizes are extracted and transferred to the Neck part of the modular structure. The Neck component consists of Path Aggregation Feature Pyramid Network (PAFPN) [28], which initially utilizes the Feature Pyramid Network (FPN) architecture to fuse low-level and high-level features. This fusion enhances both position and semantic information across feature images of different scales. Subsequently, a top-down Path Aggregation Network (PANet) is utilized to establish connections between high-level features and those of the lower layers. These enrich the feature information obtained and also enhance feature fusion capabilities. For the Output, a decoupled head is used, incorporating a generalized crossover joint loss function and the innovative Simulated Optimal Transport Assignment (SimOTA) dynamic positive sample matching method. An anchor-free method is adopted to alleviate the constraints imposed by prior bounding boxes, utilizing the CSPDarknet architecture which is known for its excellent feature extraction performance. In addition, we incorporate the Focus layer from YOLOv5 for data augmentation, along with Mosaic augmentation, enabling more comprehensive end-to-end predictions. Although YOLOX exhibits commendable detection performance, this research identifies several areas for improvement to address the following challenges:

1) The objects in remote sensing images are dense and the background is complex, which brings great difficulties to the object detection task. YOLOX employs the CNN-based CSPDarknet Backbone network to capture local feature information through convolution kernels. However, this approach often ignores the relationship between global feature information, resulting in indistinguishable target information and background information. Consequently, this approach affects the effectiveness of detecting targets in remote sensing images.

2) Remote sensing images[29] often exhibit a large aspect ratio, with objects appearing relatively small in size, thus complicating accurate object localization. YOLOX utilizes the IoU loss function for bounding box regression. However, this choice may cause gradient disappearance during the calculation process, resulting in slow convergence speed and low bounding box regression accuracy. Consequently, this approach affects the precision of object localization in remote sensing images.

2.2 Improved YOLOX Network

Aiming to address the shortcomings of YOLOX in remote sensing image object detection, this research proposes the YOLOX-NR model (shown in Fig. 1).

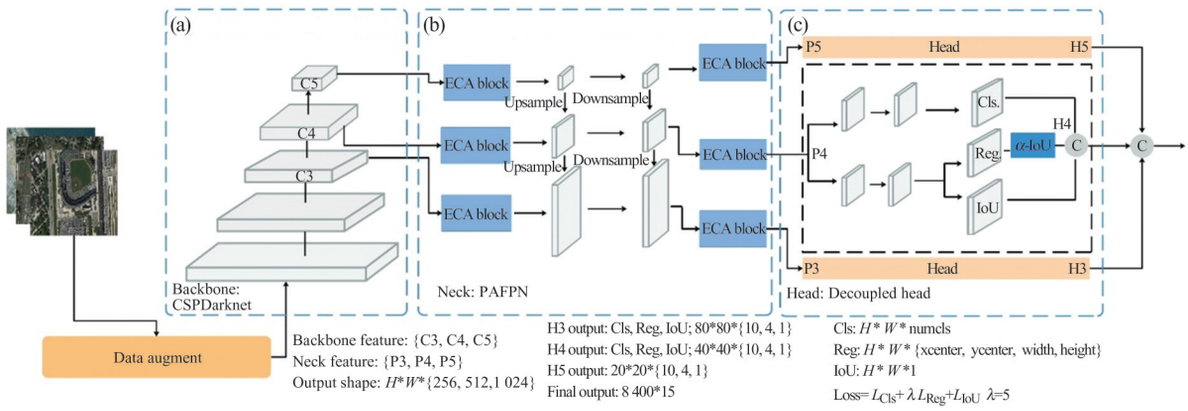

Fig. 1 YOLOX-NR model structure diagram

(a) The Backbone of YOLOX-NR is based on CSPDarknet, with three added ECA attention block. C3, C4, and C5 are feature maps of different scales extracted by the Backbone network. (b)The Neck uses the structure of the PANet and is optimized by using ECA attention block. (c) Decoupled head is the detection of YOLOX-NR, with added α-IoU loss function. H3, H4, H5 are three output branches of different scales

YOLOX-NR is an improved model based on YOLOX. In our experiments, the detection results of YOLOX-NR are better than the original YOLOX. We introduced three ECA modules following the Backbone CSPDarknets output to improve the efficiency of feature extraction. Moreover, ECA blocks are also incorporated to the back of PANet in the Neck section. This strategic placement can avoid dimensionality reduction and enable the model to focus on useful information for detection while suppressing irrelevant information such as background noise. This refinement facilitates enhanced detection accuracy of small objects in remote sensing images. Additionally, we refined the loss function by replacing the IoU loss function with the α-IoU loss function. This modification provides more flexible bounding box regression accuracy by modulating the parameter α, thereby improving the model's resilience to small datasets sizes and noise.

2.2.1 ECA attention module

Squeeze-and-Excitation Network (SENet) is the most common channel domain attention mechanism, which mainly includes three steps: Squeeze, Excitation and Attention[30]. As shown in Fig. 2(a), firstly, global average pooling is carried out for each channel separately. Then two nonlinear Fully Connected (FC) layers are applied, followed by a Sigmoid function to generate channel weight values. The two FC layers are designed to capture nonlinear channel interactions while finishing dimensionality reduction. The dimensionality reduction operation maps channel features from a high-dimensional space to a low-dimensional space and then maps them back. This operation can reduce the complexity of the model and decrease the model parameters, but it obstructs the direct correspondence between weights and channels.

ECA Network (ECANet) is an extension and improvement of SENet, which avoids dimension reduction and effectively realizes cross-channel communication. Moreover, introducing fewer parameters and computations can bring significant gains. After global average pooling of channels without dimensionality reduction, ECANet captures cross-channel interaction information by considering each channel and its K neighbors.

Considering the requirements of model performance, inserting the efficient channel attention mechanism ECANet[31] into the network can effectively improve the model performance. Additionally, due to its lightweight design, the number of model parameters will not greatly increase. As shown in Fig. 2(b), ECA obtains a $1\times1\times C$ feature map after global average pooling and

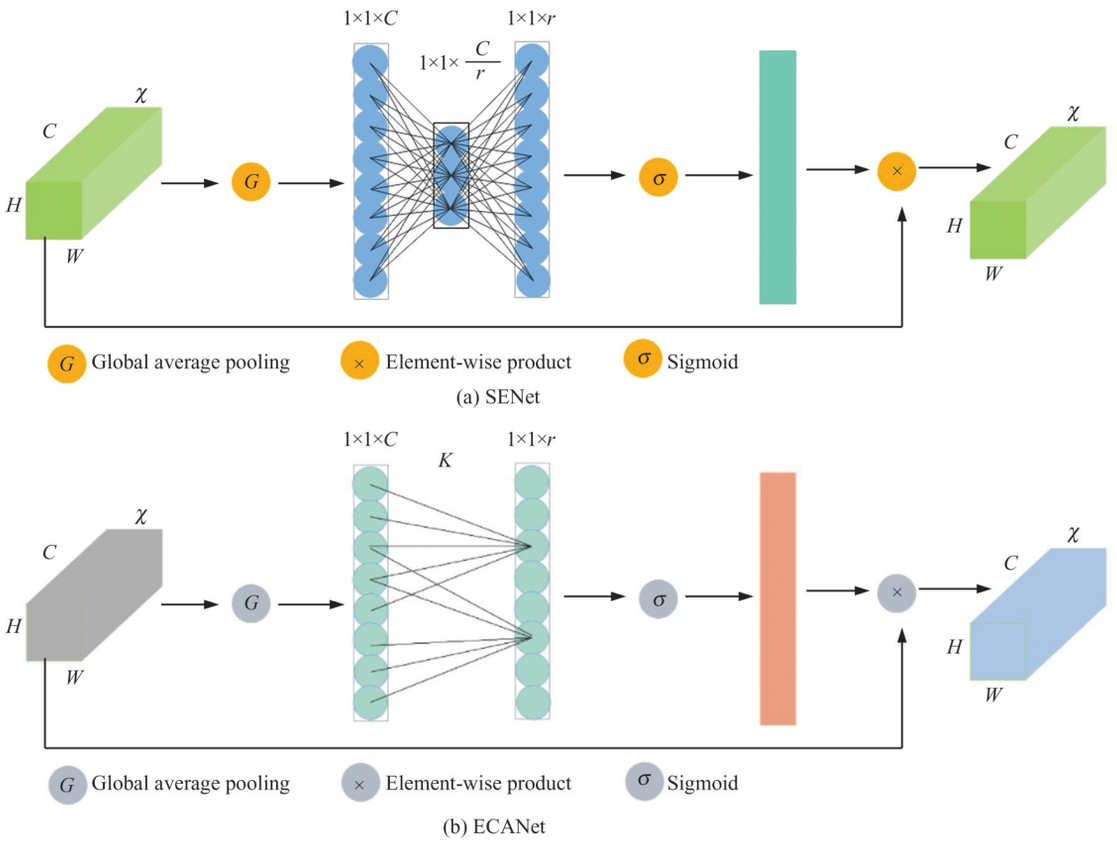

Fig. 2 Structure comparison diagram of SE module (a) and ECA module (b)

calculates the adaptive convolution kernel size K, where K refers to the coverage of local channel communication, that is how many neighbors will participate in the attention prediction of a single channel. The interaction coverage (the size of the convolution kernel) is proportional to the channel dimension. The weights for each channel are obtained by fast 1D convolutions of size K, as shown in Eq. (1).

$$W=\sigma(C1\mathrm{D}_K(y)) \tag{1}$$

where σ represents the activation function, 1D indicates 1D convolution, y indicates aggregated features without dimensionality reduction, C indicates the number of channels, then the normalized weights and the original input feature maps are multiplied channel by channel to generate the weighted feature maps.

For the determination of K value, the optimal coverage can be manually adjusted in different network structures and different numbers of convolution modules, but manual adjustment will lead to a lot of waste of computing power resources. Since the convolution kernel size K is proportional to the channel dimension C, it can be inferred that there is a mapping relationship between K and C. The mapping relationship is shown in Eq. (2).

$$C=\phi(K)=2^{\gamma\times K-b} \tag{2}$$

In order to enable layers with a larger number of channels to carry out more cross-channel interactions, the convolution kernel size of one-dimensional convolution is self-adaptive according to a function. The formula for calculating the convolution kernel size K is:

$$K=\varphi(C)=\left|\frac{\log_2(C)}{\gamma}+\frac{b}{\gamma}\right|_{\mathrm{odd}} \tag{3}$$

where C represents the number of channels, $|\bullet|_{\mathrm{odd}}$ indicates the nearest odd number, b and γ are hyperparameters, and set to be 1 and 2, respectively.

In this research, the ECA module is added, which can avoid dimension reduction and enable the model to focus on the object in the image while ignoring the influence and interference of the background. This improves the detection accuracy of the model.

2.2.2 Optimization of loss function

The loss function of the YOLOX-S model is mainly composed of three parts: bounding box regression loss ($\mathrm{Loss_{IoU}}$) [32]; class loss ($\mathrm{Loss_{cls}}$) and confidence loss ($\mathrm{Loss_{conf}}$). Among them, the bounding box regression loss of YOLOX-S model adopts the IoU loss function[33].

IoU refers to the interaction over union, which is an important index for evaluating detector performance in the field of object detection. It reflects the detection effect between the predicted box and the ground truth box. The calculation formula is shown in Eq. (4).

$$\mathrm{IoU}=\frac{|A\cap C|}{|A\cup C|} \tag{4}$$

where A represents the predicted box and C indicates the ground truth box. The greater the overlap between the prediction box and the ground truth box, the higher the IoU value. The calculation formula of IoU loss function is shown in Eq. (5).

$$\mathrm{Loss}_{\mathrm{IoU}}=1-\mathrm{IoU} \tag{5}$$

However, during the calculation process using the IoU loss function, there might be scenarios where the prediction box and the ground truth box do not intersect. This can result in the gradient disappearance problem, which will reduce both the convergence speed and detection accuracy.

To address the existing problems of IoU loss function, this paper adopts α-IoU loss function[34] to solve these deficiencies. The calculation formula is shown in Eq. (6).

$$\mathrm{Loss}_{\alpha-\mathrm{IoU}}=1-\mathrm{IoU}^{\alpha}+\rho^{\alpha_2}(B,B^{\mathrm{gt}}) \tag{6}$$

That is, the Box-Cox transformation is firstly applied to the IoU loss, and it is generalized as Power IoU loss, which incorporates an additional power regularization term along the power parameter α. B represents the predicted bounding box, and B^{gt} represents the ground truth bounding box. $\rho^{\alpha_2}\left(B,B^{\mathrm{gt}}\right)$ indicates penalty term computed based on B and B^{gt}, which can be used to prevent the gradient disappearance problem when calculating the IoU loss. Besides, the α-IoU loss can significantly exceed the existing IoU-based loss. By adjusting α, the detector gains more flexibility in achieving different levels of bounding box regression accuracy. When $\alpha > 1$, it increases the loss and gradient for high IoU objects, thus improving the bounding box regression accuracy. In most situations, the detector performs well when $\alpha = 3$. The final improved objective loss function formula is shown in Eq. (7).

$$\mathrm{Loss}=\mathrm{Loss}_{\alpha-\mathrm{IoU}}+\mathrm{Loss}_{\mathrm{conf}}+\mathrm{Loss}_{\mathrm{cls}} \tag{7}$$

When performing object detection in remote sensing images, using the IoU loss function to locate the loss may lead to a slower convergence rate, especially for small objects. This is because the small size of the object can cause the predicted box and the ground truth box to overlap which makes it difficult to accurately determine the moving direction for the bounding box. Consequently, this results in poor localization performance in remote sensing images. In this research, the α-IoU loss function is adopted to address this issue. This loss function aims to improve the convergence speed of localization loss by increasing the loss and gradient of high IoU objects, and thereby improving the regression accuracy of bounding box.

3 Experiment

3.1 Datasets

The datasets used in this research are the NWPU VHR-10 remote sensing image dataset[35] and RSOD dataset[36]. Firstly, the NWPU VHR-10 dataset is used to evaluate and train the entire model, and then the aircraft images of the RSOD dataset are used to evaluate the generalization ability and robustness of the improved model[37].

The NWPU VHR-10 dataset, annotated by the Northwestern Polytechnical University team, is used for training and testing in this experiment. The images in the dataset are cropped from Google Earth and Vaihingen datasets. The NWPU VHR-10 small target dataset consists of 800 images, which are divided into two parts: positive and negative image sets. The positive image set consists of 650 images, each containing at least one target, while the negative image set comprises 150 images with no targets. Negative image sets are only used in semi-supervised learning or weakly supervised learning. However, this research uses supervised learning, so only positive sample sets are used. The NWPU VHR-10 positive sample set covers 10 classes of objects, which are aircraft, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge and vehicle. Before the experiment, the positive sample set of NWPU VHR-10 dataset was randomly divided into a training set and a test set in a ratio of 7:3, in which the training set had 455 images and the testing set had 195 images. The numbers of different targets included in the training set and test set are shown in Table 1.

The RSOD dataset was annotated by Wuhan University and used to evaluate the generalization ability and robustness of the improved model in this experiment. The aircraft images from the RSOD dataset are selected for this generalization evaluation. The RSOD dataset[38] contains 446 aircraft remote sensing images and 4 993 aircraft targets. The brightness and contrast in the

Table 1 Numbers of targets contained in the dataset

Category	Train set	Test set	Total
Airplane	516	241	757
Ship	237	65	302
Storage tank	541	114	655
Baseball diamond	273	117	390
Tennis court	379	145	524
Basketball court	119	40	159
Ground track field	114	49	163
Harbor	159	65	224
Bridge	85	39	124
Vehicle	361	237	598

images are diverse, and there are interferences such as occlusion, shadow and distortion. During the experiment, all 446 aircraft remote sensing images were used as the test set to verify the generalization ability of the model.

3.2 Experimental Details

3.2.1 Experimental platform

The experiments in this paper have certain requirements on hardware devices, and GPU need to be used for accelerated operations. The environment required for the experiment is built on the server. The operating system is Ubuntu 18.04, the GPU is GTX1080TI, the memory is 11 GB, and the PyTorch deep learning framework is selected. The specific environment required for the experiment is shown in Table 2.

3.2.2 Training details

During the training process, the official pre-trained weight yolox-s.pt of YOLOX-S is downloaded for transfer learning. The experiment sets the total training period epochs to 300, enables Mosaic data augmentation, with a range of Mosaic augmentation from 0.1 to 2, and turns off data augmentation for the last 15 epochs.

Table 2 Experimental environment configuration

Configuration	Parameter
GPU	GTX1080TI
System	Ubuntu 18.04
Python version	3.7
Deep learning framework	PyTorch 1.8.0
CUDA	11.1

The optimizers used for this training are Stochastic Gradient Descent (SGD) and Adamax, with a weight decay coefficient of 0.000 5, an initial learning rate of 0.01, a minimum learning rate is 5% of the set current learning rate, and a learning rate decay strategy using the cosine annealing algorithm. Additionally, the momentum and Nesterov parameters of the SGD optimizer are set to 0.9 and True, respectively. The batch size is set to 8, and the image input size is 640×640. Each group of experiments is trained for 5 times.

3.3 Evaluation Indicator

In terms of the accuracy of the network model, this experiment uses Average Precision (AP), Mean Average Precision (mAP) and Precision-Recall (P-R) curve, along with other common object detection evaluation indicators to evaluate the performance of the algorithm in this research[39].

Precision is used to measure the correctness of model testing, and recall is used to evaluate the comprehensiveness of model testing. The calculation formulas are as follows:

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \end{aligned} \tag{8}$$

The meanings of TP, FP and FN in the formula are shown in Table 3.

Table 3 Description of evaluation index formula

Name	Definition
FN	Number of positive samples incorrectly determined as negative samples
FP	Number of negative samples erroneously determined as positive samples
TP	Number of positive samples correctly classified
TP+FN	Actual number of positive samples
TP+FP	Total number of positive samples predicted

The P-R curve can be drawn according to the calculated precision and recall rate, and the area of the graph formed by the P-R curve is the AP of a single class, as shown in Eq. (9).

$$\mathrm{AP}=\int_0^1 P(r)\mathrm{d}r \tag{9}$$

where P represents Precision and r represents Recall. For mAP, the PASCAL VOC evaluation metric is adopted in this paper, which calculates the average AP for each class at the IoU threshold of 0.5. The calculation formula is shown in Eq. (10).

$$\mathrm{mAP}=\frac{\sum_{i=1}^{C}\mathrm{AP}_i}{C} \tag{10}$$

where C represents the number of classes and AP_i represents the AP value of each class.

In terms of model complexity, the evaluation metrics used in this experiment consist of the number of parameters (weights of the model) and Floating-point Operations (FLOPs):

$$\begin{aligned}\mathrm{Parameters}&=[i\times(f\times f)\times o]+o\\ \mathrm{FLOPs}&=H\times W\times\mathrm{Parameters}\end{aligned} \tag{11}$$

where i represents the number of input channels, f represents the size of the convolution kernel, o represents the number of output channels, and $H\times W$ represents the size of the output feature map.

3.4 Analysis of Experimental Results

3.4.1 Experiments on attention mechanisms

In order to verify the role of the ECA module, this research uses three different attention mechanisms, namely Convolutional Block Attention Module (CBAM)[25], SE and ECA for comparative experiments. Each experiment uses the same set of parameters and employs the official pretrained weights of YOLOX-S for transfer learning. Additionally, to unify the control variables, the three attention mechanisms are only added between the Backbone and Neck of YOLOX-S. The loss function and optimizer of the experiment are both set to IoU and SGD, and the training rounds are unified to 300 epochs. The final results are shown in Table 4.

As can be seen from Table 4, when ECA attention module is added to YOLOX-S model, its mAP value is higher than that when the CBAM module and SE module are added. Additionally, the number of parameters and FLOPs after adding ECA module are also less than the latter two, which reduces the complexity of the model and the requirements of hardware equipment to a certain extent. Therefore, the attention mechanism selected in this study is ECA module.

Table 4 Comparative experiment of different attention mechanisms

Models	mAP/%	Parameters/10^6	FLOPs/G
YOLOX-S+CBAM	93.2	9.03	26.69
YOLOX-S+SE	93.7	8.98	26.67
YOLOX-S+ECA	**94.4**	8.94	26.66

The bold font in the table is the maximum value

3.4.2 Results on the NWPU VHR-10 dataset

Figure 3(a) and 3(b) show the changes of mAP curve and P-R curve of the two algorithms during the training process, respectively. It can be seen from the figure that the mAP value of our method is higher than that of the YOLOX-S model, and the area formed by its P-R curve is also larger than that of YOLOX-S. This indicates that our method is better than the YOLOX-S model in the accuracy of remote sensing image detection. Figure 3(c) and 3(d) show the changes of the P-R curves of the YOLOX-S model and our method at the IoU thresholds of 0.5, 0.6, 0.7, 0.8 and 0.9, respectively. When IoU = 0.5, the area formed by P-R curve is the largest, indicating that the image object detection at this time is the most accurate. The sum of the area of five P-R curves of our method is larger than that of the YOLOX-S model, which further verifies that the detection performance of our method is better than that of the YOLOX-S model.

Figure 4 shows the test results of the proposed algorithm on part of the data in the NWPU VHR-10 dataset. As shown in the Fig. 4, although the detection accuracy of harbor object is slightly low, the detection accuracy of other objects is high, so the overall detection effect is excellent, which also demonstrates the reliability of the proposed algorithm for remote sensing image object detection.

3.4.3 Model generalization verification experiment

In order to verify the generalization of the improved method based on the YOLOX-S model proposed in this research, we tested it on the aircraft images of the RSOD dataset. Briefly, the weights trained by the previous improved model on the NWPU VHR-10 dataset are directly used for inference testing on the aircraft images of the RSOD dataset, without training process. Its mAP and mAP50: 95 can reach 86.70% and 48.47% respectively. This indicates that the YOLOX-NR model has

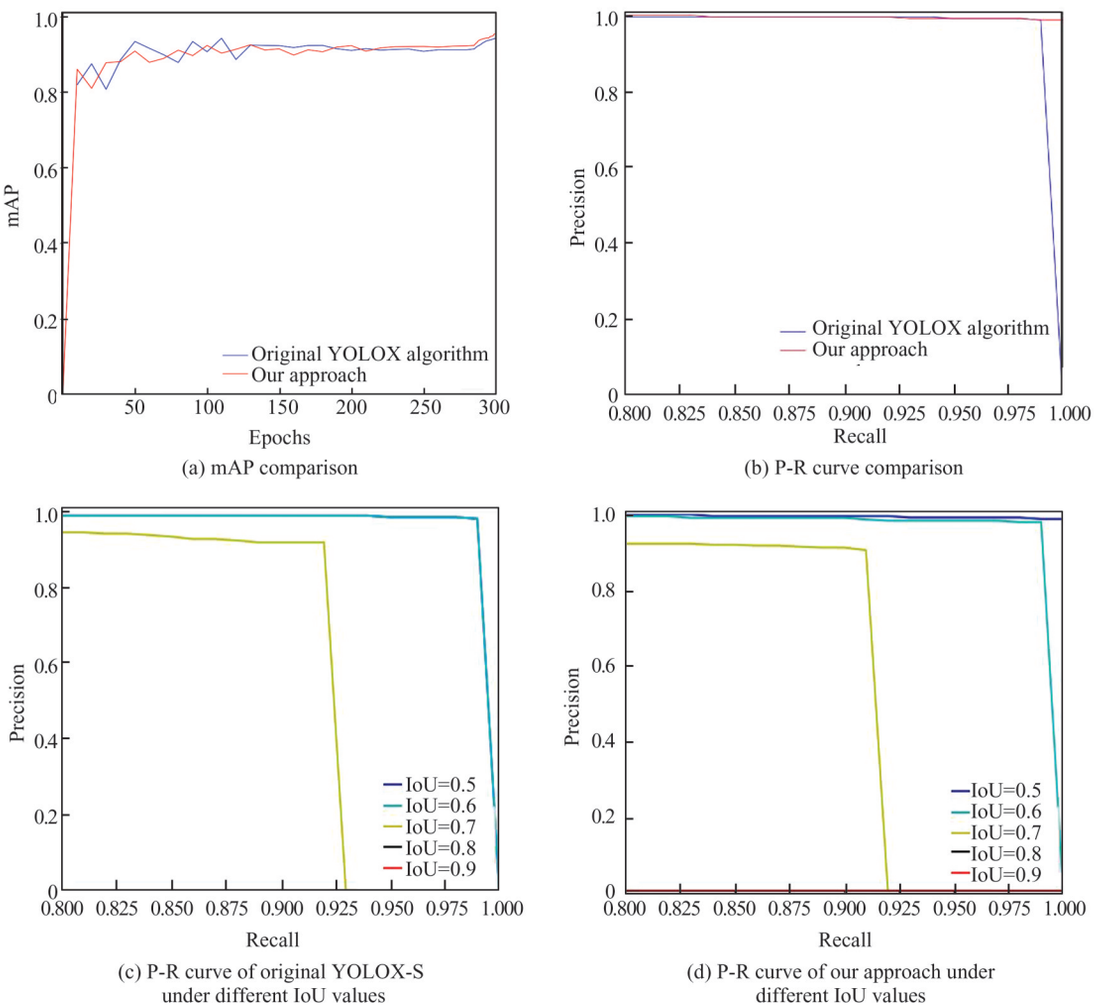

Fig. 3 Performance comparison between our method and the original YOLOX-S model

relatively good generalization performance and can be extended to the detection of other remote sensing image objects.

Figure 5 shows the performance of the YOLOX-NR model on aircraft images from the RSOD dataset. As observed in the figure, the YOLOX-NR model consistently achieves good performance and accurate object detection under varying conditions of brightness and contrast, occlusion, shadow, and small object interference. This indicates that the YOLOX-NR model exhibits relatively good generalization properties, suggesting its potential applicability to the detection of other objects in remote sensing images.

3.4.4 Ablation experiment

In order to further verify the effectiveness of each module in the YOLOX-NR model, we conducted an ablation experiment on each module using the NWPU VHR-10 dataset in this research. The results of the ablation experiments are shown in Table 5.

As observed from Table 5, the first group employed YOLOX-S as the benchmark model without incorporating any improved modules, and its detection accuracy reached 94.2%. In the second group, the loss function optimized and the optimizer was modified based on the first group. Specifically, the IoU loss function was replaced with the α-IoU loss function, and the SGD optimizer was substituted with the Adamax optimizer, resulting in an mAP of 94.5% without introducing additional parameters. In the third group, the channel attention module ECA was added between the Backbone and Neck of the YOLOX-S model based on the second group, leading to an accuracy of 95.0% with a slight increase in computational overhead. On the basis of the third group, the fourth experiment also incorporated the ECA module between the Neck and Head of YOLOX-S

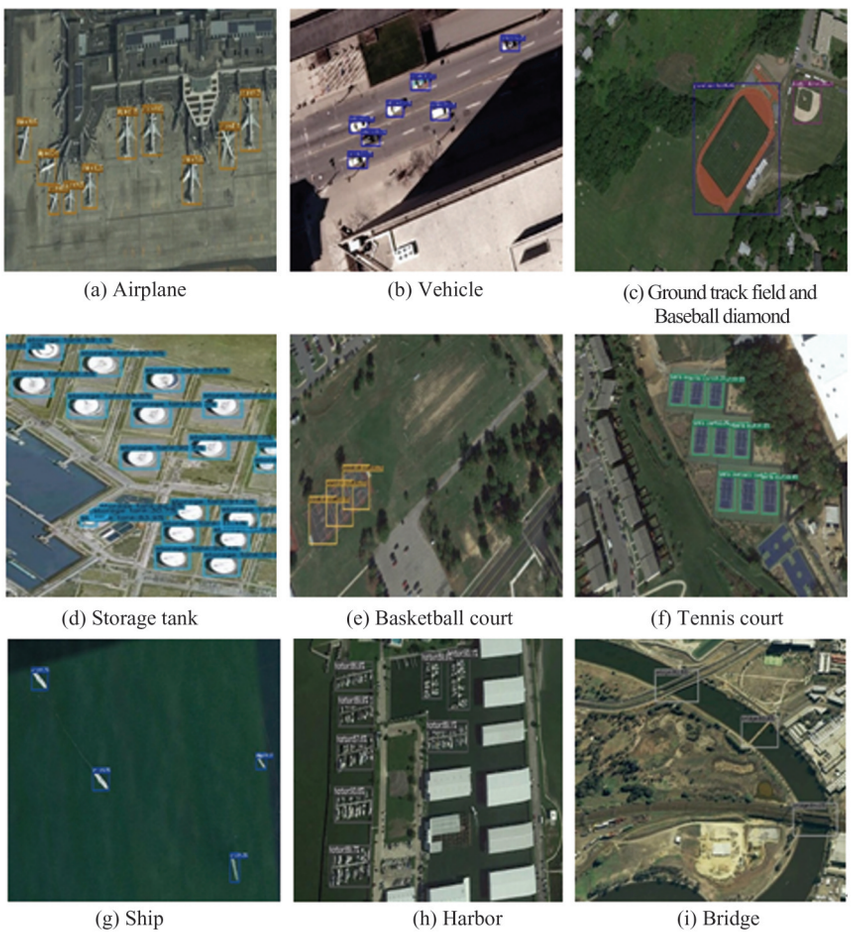

Fig. 4 The detection results of different targets in NWPU VHR-10 dataset using our method

Fig. 5 Detection performance of YOLOX-NR in RSOD dataset aircraft images

Table 5 Comparison of ablation experiments of each module of the improved algorithm

YOLOX-S	α-IoU+Adamx	ECA (between Backbone and Neck)	ECA (between Neck and Head)	mAP/%	mAP50:95/%
√	×	×	×	94.2	61.85
√	√	×	×	94.5	62.70
√	√	√	×	95.0	63.05
√	√	√	√	**95.5**	**62.86**

"√" indicates the structure is included in the model, and "×" indicates that the structure is not included in the model. The bold font in the table is the maximum value

model. The number of parameters and floating-point operations remained unchanged, and the mAP reached 95.5%, which further improved the detection accuracy.

3.4.5 Comparison with other models

In order to verify the effectiveness of method, we compared it with several mainstream algorithms on NWPU VHR-10. Among these, BOW[40] represents a machine learning method, while the others are based on deep learning technology. These deep learning-based methods include R-CNN-based RICNN, single-stage algorithm SSD, deformable Region-based Fully Convolutional Network (R-FCN), traditional Faster R-CNN, and Multi-Scale Convolutional Neural Network (MSCNN), etc. The mAP values of each class are shown in Table 6. From the total average precision of each class, the method proposed in this paper (95.5%) is higher than other methods in accuracy. The traditional BOW object detection method has the lowest mAP value (24.57%), and the detection performance of deep learning methods is significantly better than that of traditional object detection algorithms. From the detection accuracy rate of individual classes, the method in this research demonstrates superior performance in identifying airplanes, ships, storage tanks, baseball diamonds, basketball courts, ground track fields, harbors and vehicles. In particular, the detection accuracy of small objects such as ships and storage tanks, as well as complex objects like baseball diamonds, ground track fields, and basketball courts, which are difficult to distinguish from the background and surrounding objects, has been significantly improved. Furthermore, while the YOLOX-S model exhibits substantial improvements in detection effectiveness compared to other models, it encounters issues such as false detections and missed detections in habors, which lead to lower mAP values. The method proposed in this research effectively addresses these problems, improving the mAP of harbor by 7.18% and enhancing the overall detection performance.

Table 6 Performance comparison between the proposed model and other methods on NWPU VHR-10 dataset

%

Category	BOW[4]	RICNN[13]	SSD[10]	Deformable R-FCN[41]	Faster R-CNN[8]	Multi-Scale CNN[42]	YOLOX-S	YOLOX-NR
Airplane	24.96	88.35	95.70	87.30	94.60	99.30	98.87	**99.89**
Ship	58.49	77.34	82.90	81.40	82.30	92.00	94.34	**95.59**
Storage tank	63.18	85.27	85.60	63.60	65.30	83.20	93.04	**96.36**
Baseball diamond	9.03	88.12	96.60	90.40	95.50	97.20	96.35	**95.32**
Tennis court	4.72	40.83	82.10	81.60	81.90	90.80	**99.28**	98.61
Basketball court	3.22	58.45	86.00	74.10	89.70	92.60	99.82	**100.00**
Ground track field	7.77	85.73	58.20	90.30	92.40	98.10	98.62	**99.15**
Harbor	52.98	68.60	54.80	75.30	72.40	85.10	88.05	**95.23**
Bridge	12.16	61.51	41.90	71.40	57.50	71.90	**82.79**	81.64
Vehicle	9.14	71.10	75.60	75.50	77.80	85.90	90.87	**93.01**
mAP	24.57	72.63	75.90	79.10	80.90	89.60	94.20	**95.50**

The bold font in the table is the maximum value

4 Discussion

In order to solve the problems of complex background and dense small target detection in remote sensing images, the YOLOX-NR proposed in this research integrates the efficient channel attention and optimizes the boundary regression loss function. This integration significantly enhances the detection accuracy of small and medium-sized objects in remote sensing images.

In this research, we proved the effectiveness of ECA attention module in detecting small objects in remote sensing images. As shown in Table 4 and Table 5, when the YOLOX-S model is added to ECA attention module, its mAP value is higher than the original model, as well as models augmented with the CBAM or SE modules. This superiority is attributed to the stronger feature extraction capabilities of the ECA channel attention module compared to CBAM or SE modules. Addi-

tionally, the ECA channel attention can effectively focus on useful information for detection while suppressing irrelevant background information, thereby improving the detection accuracy of small objects in remote sensing images. Furthermore, we demonstrated the feasibility of the α-IoU loss function. As shown in Table 5, adding α-IoU loss function can improve the object detection accuracy. This is because the α-IoU loss function can accelerate the convergence of positioning loss by modulating α, increase the loss and gradient of high IoU targets, and thereby improving the regression accuracy of bounding boxes.

In summary, the method proposed in this research effectively solves the problem of dense small object detection in remote sensing images and exhibits excellent detection performance in actual remote sensing image object detection tasks. Nevertheless, since this study employs horizontal bounding boxes for object detection, it may ignore the directional information such as the rotation angle of objects. Therefore, in the future work, we will take into account factors such as the orientation information of objects and explore optimization methods to further research in the domain of remote sensing image tasks.

5 Conclusion

To solve the problems of complex background and poor detection performance of dense small objects in remote sensing images, this study proposes a lightweight remote sensing image object detection algorithm which is based on an improved YOLOX-S model. This algorithm integrates the ECA mechanism to improve the network's ability on extract important features from images. Additionally, we use the α-IoU loss function as a replacement for the IoU loss function in bounding box regression, thereby improving the localization accuracy of objects. Experimental results show that the mAP in the improved method exceeds that in the original YOLOX model by 1.3% on the NWPU VHR-10 dataset. Moreover, in the generalization experiment using RSOD remote sensing dataset for aircraft images, our method also exhibits commendable object detection performance, indicating that our method exhibits good generalization ability.

References

[1] Dai L J, Liu C. Research on remote sensing image of land cover classification based on multiple classifier combination [J]. *Wuhan University Journal of Natural Sciences*, 2011, **16** (4): 363-368.

[2] Yu D W, Ji S P. A new spatial-oriented object detection framework for remote sensing images[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, **60**: 1-16.

[3] Shivappriya S N, Priyadarsini M J P, Stateczny A, *et al*. Cascade object detection and remote sensing object detection method based on trainable activation function[J]. *Remote Sensing*, 2021, **13**(2): 200.

[4] Zhang Y, Jin R, Zhou Z H. Understanding bag-of-words model: A statistical framework[J]. *International Journal of Machine Learning and Cybernetics*, 2010, **1**(1): 43-52.

[5] Sandoval C, Pirogova E, Lech M. Two-stage deep learning approach to the classification of fine-art paintings[J]. *IEEE Access*, 2019, **7**: 41770-41781.

[6] He K M, Zhang X Y, Ren S Q, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, **37**(9): 1904-1916.

[7] Girshick R. Fast R-CNN[C]//2015 *IEEE International Conference on Computer Vision* (*ICCV*). New York: IEEE, 2015: 1440-1448.

[8] Ren S Q, He K M, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks[C]//*Proceedings of the* 28*th International Conference on Neural Information Processing Systems*. New York: ACM, 2015: 91-99.

[9] de Vos B D, Berendsen F F, Viergever M A, *et al*. A deep learning framework for unsupervised affine and deformable image registration[J]. *Medical Image Analysis*, 2019, **52**: 128-143.

[10] Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector[C]//*Computer Vision – ECCV* 2016. Cham: Springer, 2016: 21-37.

[11] Lin T Y, Goyal P, Girshick R, *et al*. Focal loss for dense object detection[C]//2017 *IEEE International Conference on Computer Vision* (*ICCV*). New York: IEEE, 2017: 2999-3007.

[12] Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection[C]//2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). New York: IEEE, 2016: 779-788.

[13] Cheng G, Zhou P C, Han J W. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2016, **54**(12): 7405-7415.

[14] Yang X, Yang J R, Yan J C, *et al*. SCRDet: Towards more

robust detection for small, cluttered and rotated objects[C]// 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). New York: IEEE, 2019: 8231-8240.

[15] Wang P J, Sun X, Diao W H, *et al*. FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2020, **58**(5): 3377-3390.

[16] Nayan A A, Saha J, Nokib Mozumder A, *et al*. Real time multi-class object detection and recognition using vision augmentation algorithm[J]. *International Journal of Advanced Science and Technology*, 2020, **29**(5): 14070-14083.

[17] Xi Y, Jia W, Zheng J, *et al*. DRL-GAN: Dual-stream representation learning GAN for low-resolution image classification in UAV applications[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021, **14**: 1705-1716.

[18] Nie X, Duan M Y, Ding H X, *et al*. Attention mask R-CNN for ship detection and segmentation from remote sensing images[J]. *IEEE Access*, 2020, **8**: 9325-9334.

[19] Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-Time object detection[C]// 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). New York: IEEE, 2016: 779-788.

[20] Redmon J, Farhadi A. YOLOv3: An incremental improvement[EB/OL]. [2018-04-08]. http://arxiv.org/abs/1804.02767.

[21] Cao C Q, Wu J, Zeng X D, *et al*. Research on airplane and ship detection of aerial remote sensing images based on convolutional neural network[J]. *Sensors*, 2020, **20**(17): 4696.

[22] Bochkovskiy A, Wang C Y, Liao H M. YOLOv4: Optimal speed and accuracy of object detection[EB/OL]. [2020-04-23]. https://arxiv.org/abs/2004.10934.

[23] Yu P D, Wang X, Liu J H, *et al*. Bridge target detection in remote sensing image based on improved YOLOv4 algorithm [C]//2020 4*th International Conference on Computer Science and Artificial Intelligence*. New York: ACM, 2020: 139-145.

[24] Zhao B Y, Wu Y F, Guan X R, *et al*. An improved aggregated-mosaic method for the sparse object detection of remote sensing imagery[J]. *Remote Sensing*, 2021, **13**(13): 2602.

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

[26] Wang F, Jiang M Q, Qian C, *et al*. Residual attention network for image classification[C]//2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). New York: IEEE, 2017: 6450-6458.

[27] Ge Z, Liu S T, Wang F, *et al*. YOLOX: Exceeding YOLO series in 2021[EB/OL]. [2021-08-06]. http://arxiv. org/abs/2107.08430.

[28] Wang W H, Xie E Z, Song X G, *et al*. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network[C]//2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). New York: IEEE, 2019: 8439-8448.

[29] Shen Y C, Jin H, Du B. An improved method to detect remote sensing image targets captured by sensor network[J]. *Wuhan University Journal of Natural Sciences*, 2011, **16**(4): 301-307.

[30] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]// 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). New York: IEEE, 2018: 7132-7141.

[31] Wang Q L, Wu B G, Zhu P F, *et al*. ECA-net: Efficient channel attention for deep convolutional neural networks[C]// 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). New York: IEEE, 2020: 11531-11539.

[32] He Y H, Zhu C C, Wang J R, *et al*. Bounding box regression with uncertainty for accurate object detection[C]//2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). New York: IEEE, 2019: 2883-2892.

[33] Yu J H, Jiang Y N, Wang Z Y, *et al*. UnitBox: An advanced object detection network[C]//*Proceedings of the* 24*th ACM international conference on Multimedia*. New York: ACM, 2016: 516-520.

[34] He J B, Erfani S, Ma X J, *et al*. Alpha-IoU: A family of power intersection over union losses for bounding box regression[EB/OL]. [2022-01-22]. http://arxiv. org/abs/2110.13675.

[35] Cheng G, Han J W. A survey on object detection in optical remote sensing images[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2016, **117**: 11-28.

[36] Xiao Z F, Liu Q, Tang G F, *et al*. Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images[J]. *International Journal of Remote Sensing*, 2015, **36** (2): 618-644.

[37] Chen X, Liu J H, Xu F, *et al*. A novel method of aircraft detection under complex background based on circular intensity filter and rotation invariant feature[J]. *Sensors*, 2022, **22** (1): 319.

[38] Long Y, Gong Y P, Xiao Z F, *et al*. Accurate object localization in remote sensing images based on convolutional neural networks[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2017, **55**(5): 2486-2498.

[39] Yang C C, Ma J Y, Zhang M F, *et al*. Multiscale facet model for infrared small target detection[J]. *Infrared Physics and Technology*, 2014, **67**: 202-209.

[40] Xu S, Fang T, Li D R, *et al*. Object classification of aerial images with bag-of-visual words[J]. *IEEE Geoscience and Remote Sensing Letters*, 2010, **7**(2): 366-370.

[41] Xu Z Z, Xu X, Wang L, *et al*. Deformable ConvNet with aspect ratio constrained NMS for object detection in remote sensing imagery[J]. *Remote Sensing*, 2017, **9**(12): 1312.

[42] Guo W, Yang W, Zhang H J, *et al*. Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network[J]. *Remote Sensing*, 2018, **10** (1): 131.

改进YOLOX的遥感图像目标检测算法

刘蓓蓓, 邓彦松†, 吕赫, 周晨晨, 唐学智, 向伟

西南民族大学 电子信息学院, 四川 成都 610041

摘要: 遥感图像目标检测是遥感图像处理的核心任务之一。近年来, 随着深度学习技术的发展, 遥感图像中的目标检测技术取得了很大的进步。然而, 遥感图像中存在的小目标密集、背景复杂以及目标定位精度差等问题, 使得遥感目标的检测仍然比较困难。为了解决这些问题, 本文提出了一种基于改进YOLOX-S的遥感图像目标检测算法。首先, 引入高效通道注意力模块ECA, 提升网络对图像中重要特征的提取能力, 抑制背景等无用信息; 其次, 优化损失函数,提升目标边界框的回归精度。我们在公开的遥感图像数据集NWPU VHR-10上评估了本文算法的有效性, 实验结果表明该算法的检测精度能达到95.5%, 在不增加参数量的情况下, 较原有YOLOX-S网络有明显提升, 且检测性能大幅优于其他一些主流的遥感图像检测方法。除此之外, 在RSOD数据集中的飞机图像的泛化性实验中, 我们的方法也表现出不错的检测性能。

关键词: 遥感图像; 目标检测; YOLOX-S; 注意力模块; 损失函数

□