Improved remote sensing image target detection based on YOLOv7[*]

XU Shuanglong, CHEN Zhihong[**], ZHANG Haiwei, XUE Lifang, and SU Huijun

School of Integrated Circuit Science and Engineering, Tianjin University of Technology, Tianjin 300384, China

(Received 8 April 2023; Revised 21 September 2023)

Remote sensing images are taken at high altitude from above, with complex spatial scenes of images and a large number of target types. The detection of image targets on large scale remote sensing images suffers from the problem of small target size and target density. This paper proposes an improved model for remote sensing image detection based on you only look once version 7 (YOLOv7). First, the small-scale detection layer is added to reacquire tracking frames to improve the network's recognition ability of small-scale targets, and then Bottleneck Transformers are fused in the backbone to make full use of the convolutional neural network (CNN)+Transformer architecture to enhance the feature extraction ability of the network. After that, the convolutional block attention module (CBAM) mechanism is added in the head to improve the model's ability of small-scale target. Finally, the non-maximum suppressed (NMS) of YOLOv7 algorithm is changed to distance intersection over union-non maximum suppression (DIOU-NMS) to improve the detection ability of overlapping targets in the network. The results show that the method in this paper can improve the detection rate of small-scale targets in remote sensing images and effectively solve the problem of high overlap and is tested on the NWPU-VHR10 and DOTA1.0 datasets, and the accuracy of the improved model is improved by 6.3% and 4.2%, respectively, compared with the standard YOLOv7 algorithm.

Document code: A **Article ID:** 1673-1905(2024)04-0234-9

DOI https://doi.org/10.1007/s11801-024-3063-z

With the rapid development of the satellite sensor technology, high spatial resolution remote sensing data have attracted extensive attention in military and civilian applications[1]. Satellite images taken on the earth's surface are analyzed to identify the spatial and temporal changes that have occurred naturally or man-made. Real-time prediction of change provides an understanding related to the land cover[2], environmental changes, habitat fragmentation, coastal alteration, urban sprawl, etc. Therefore, remote sensing image target detection and recognition has important practical value.

With the rapid development of deep learning[3], more powerful tools capable of semantic learning, advanced and deeper features have been introduced to solve the problems in traditional architecture. The target detection algorithm based on deep learning[4], with its advantages of flexible structure, automatic feature extraction and powerful data processing capability, has many advantages, such as high performance, wide practical application scenarios, convenient and simple use. There are two main types of deep learning based target detection algorithms, one is the two-stage algorithm represented by fast region-based convolutional neural network (fast R-CNN)[5], faster R-CNN[6], etc. This type of method possesses high detection accuracy, but the detection speed is slow. Single-stage methods are regression-based detection algorithms that employ end-to-end target detection methods, such as single shot multibox detector (SSD)[7], you only look once (YOLO)[8-10], etc. These algorithms do target detection directly in the feature extraction layer without generating candidate regions, which greatly saves detection time and has no significant disadvantage in detection accuracy.

With the continuous pursuit of small target detection accuracy and building a deeper network, researchers have done a lot of research. Ref.[11] proposed an adaptive learning method to select the best data enhancement strategy to obtain a certain performance improvement in small target detection. Ref.[12] proposed a deconvolutional object detection network (DODN) model by optimizing the detection frame filtering mechanism, which replaces the anchored frame mechanism by building a two-level deconvolutional network, and then generates the region of interest by region proposal network (RPN). The detection accuracy of the model is improved. Ref.[13] proposed a learning rotation-invariant convolutional neural network (RICNN), which introduces and learns new rotation-invariant structures to improve the detection performance based on the existing CNN structures. Ref.[14] proposed a scale-matching strategy to

[*] This work has been supported by the National Natural Science Foundation of China (Nos.62005196 and 61974104).

[**] E-mail: zhchen@email.tjut.edu.cn

crop the objects according to the target size, reduce the scale gap between different objects, and avoid losing small target information when downsampling the images. Ref.[15] proposed a novel object detection method based on shallow feature fusion and semantic information augmentation (FFSI). High-level semantic information is injected into low-level features to guide the enhancement of specific detail information. Ref.[16] proposed an extended feature pyramid network, which uses additional high-resolution pyramids specifically for small target detection.

In this paper, we propose an improved algorithm based on the you only look once version 7 (YOLOv7) algorithm applied to remote sensing image datasets for target detection. Because the YOLOv7 algorithm has poor detection effect on small-scale targets, the multi-scale learning ability of the network is improved by adding a small-scale detection layer. Then we integrate Bottleneck Transformers to strengthen the ability of global modeling and long-distance modeling. After that, convolutional block attention module (CBAM) mechanism is introduced to improve the model's ability to capture target features and improve the model's recognition performance for small targets. Finally, change weighted non-maximum suppressed (NMS) to distance intersection over union-non maximum suppression (DIOU-NMS), to solve the problem of low accuracy of the standard YOLOv7 algorithm for overlapping target detection. This experimental result shows that the improved YOLOv7 algorithm not only improves the detection performance of small and medium scales and dense categories in remote sensing images, but also increases the detection accuracy of remote sensing datasets.

The author team of YOLOv4[17] proposed YOLOv7[18]. The YOLOv7 network structure is composed of input, backbone network and neck network. The input module preprocesses the input image through Mosaic data enhancement and adaptive anchor frame calculation. The backbone network module mainly uses three structures: CBS, MP and ELAN, and the structure is shown in Fig.1. The ELAN structure continuously enhances the learning ability of the network by controlling the shortest and longest gradient path without destroying the original gradient path, so that the deeper network can effectively learn and converge. The neck module adopts the characteristic pyramid network FPN structure and the path aggregation network PAN structure. PAFPN structure can fuse the output multi-scale feature layer, so that the small-scale feature layer has rich semantic information, while the large-scale feature has more abundant feature information. For P3, P4 and P5 output by the neck module, the prediction module adjusts the number of channels through RepConv, and finally uses 1×1 convolution to predict the objectivity, class and box.

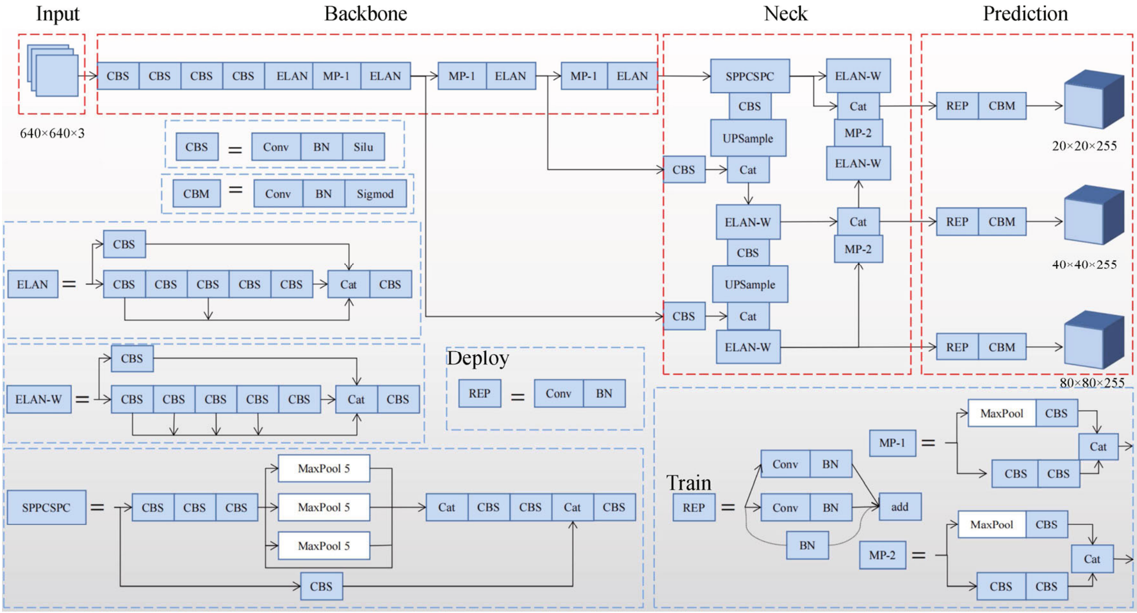

Fig.1 Main modules of YOLOv7 network

Due to the small size of the small target samples and YOLOv7's lower sample multiplier is larger, it is more difficult to learn the feature information of the small target in the deeper feature map, so the original model of YOLOv7 has poor detection ability for small targets. The input image size of the original model is 640×640, and the minimum detection scale is 80×80. The sensory field of each grid is 8×8, so if the height and width of the target in the original image are less than 8 pixels, it is difficult for the original network to recognize the feature information of the target in the grid. For this reason, this paper proposes to add a small-scale detection layer to the PAN structure to solve the above problem by adding a 160×160 detection scale on top of the previous three-scale detection layer to become a four-scale detection. The main process of adding a small-scale detection layer is as follows: the 80×80 feature map is upsampled to expand the feature map to 160×160, concat fused with the feature map of the second layer in the backbone network, and a set of anchors is added by the K-mean clustering algorithm as a way to perform small-scale detection at a larger feature map. Although the amount of computation is increased, the accuracy of small target detection in remote sensing images is really improved.

BotNet[19] is an exploration of the combination of Revolution+Transformer by researchers from Berkeley and Google. Using CNN+Transformer, we propose a Bottleneck Transformer to replace ResNet Bottleneck. That is, only the last three Bottleneck blocks in the ResNet framework use multi head self attention (MHSA) to replace 3×3 spatial convolutions. The MHSA layer includes relative position coding and MHSA model, which can enable the network model to learn more features and details of the image and improve the network performance, and 84.7% accuracy was achieved in ImageNet. The network structure of Bottleneck Transformer is shown in Fig.2.

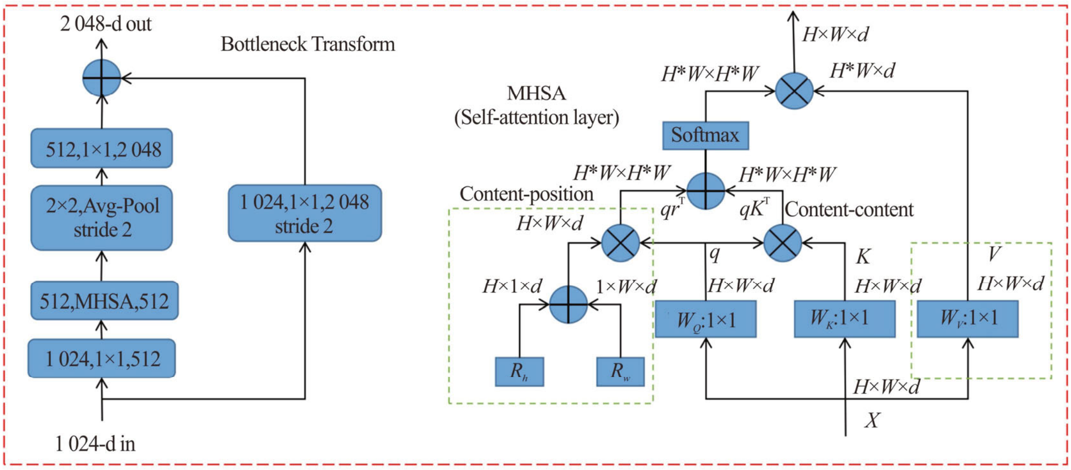

Fig.2 Bottleneck Transformer structure diagram

Bottleneck Transformer replaces 3 with MHSA×3 convolution, 3 in the first Bottleneck×3 convolutional stride=2, but MHSA module does not support stride operation, so BotNet adopts 2×2 average pooling for down sampling.

The purpose of adding attention mechanism is to tell the model which position and content to focus on. Through the difference in the way and position of attention weight, attention mechanism can be divided into three types: spatial domain attention, channel domain attention and mixed domain attention. In this experiment, the method of CBAM in the hybrid domain attention method[20] is used. It is an attention mechanism module that provides attention maps from the channel and spatial dimensions in order, mainly divided into channel attention module and spatial attention module, which can make the features extracted from the model more refined and effectively improve the classification effect of the model.

The CBAM schematic is shown in Fig.3, which inferred the attention weights from the intermediate feature map along the 2 dimensions of space and channel, and then multiplied the weights with the original feature map to adjust the features adaptively, so as to achieve the purpose of focusing on the target features. At the same time, CBAM is a lightweight general-purpose module that can be integrated into CNNs at a small cost and can be trained end-to-end together with the basic CNN.

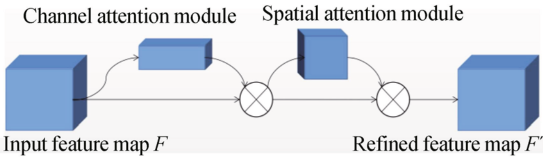

Fig.3 CBAM attention mechanism structure diagram

The channel attention module focuses on the meaningful information in the input image and can compress the spatial dimension while keeping the channel dimension unchanged. The structure of the channel attention module is shown in Fig.4, where the feature map $F(F\in R^{C\times H\times W})$ is input at the input side, and after average pooling and maximum pooling, the feature map of size $C\times H\times W$ is transformed into $C\times 1\times 1$ size, then they are fed into the neural network MLP, where the number of neurons in the first layer is C/r, r is the descent rate, the activation function is Relu, and the number of neurons in the second layer is C. The results are summed up after completion, then the new features are scaled by a Sigmoid function M_C. The weight coefficients are calculated as shown in Eq.(1), and multiplied by the initial input to obtain the new scaled features.

$$M_C(F)=\sigma(W_1(W_0(F^C_{\text{avg}}))+W_1(W_0(F^C_{\max}))), \qquad (1)$$

where σ denotes the Sigmoid function, avg denotes the global average pooling, max denotes the maximum pooling, $W_0\in R^{C\times\frac{C}{r}}$, $W_1\in R^{C\times\frac{C}{r}}$, F^C_{avg} denotes the average pooling feature of size $1\times1\times C$, and $F_{\max}^{C}$ denotes the maximum pooling feature of size $1\times1\times C$.

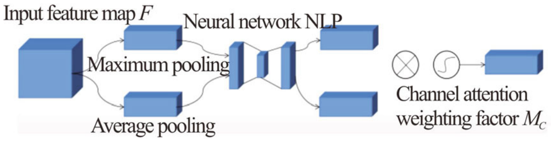

Fig.4 Channel attention module structure

The structure of the spatial attention module is shown in Fig.5. The results obtained in the previous step are divided into two channel descriptions of size $1\times H\times W$ by maximum pooling and average pooling, and then the tensor is stacked together by the concatenation operation, and then the weight coefficients M_S are obtained by the convolution operation and a Sigmoid. The weight coefficients are calculated as shown in Eq.(2), and the newscaled features are obtained by multiplying the weight coefficients by the input of the previous step, which completes the spatial attention operation.

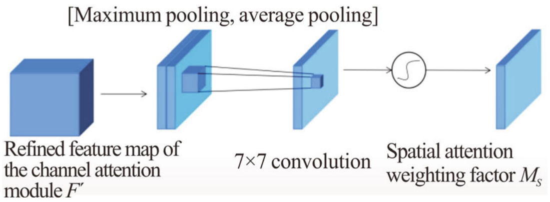
[Maximum pooling, average pooling]
Refined feature map of the channel attention module *F′*
7×7 convolution
Spatial attention weighting factor M_S

Fig.5 Spatial attention module structure

$$M_S(F)=\sigma(f^{7\times7}([F_{\text{avg}}^{S};F_{\max}^{S}])), \tag{2}$$

where $f^{7\times7}$ denotes a 7×7 convolution, F_{avg}^{S} denotes average pooling feature of size $1\times H\times W$, and $F_{\max}^{S}$ denotes maximum pooling feature, also of size $1\times H\times W$.

In the prediction phase of target detection, multiple prediction frames are generated, where multiple prediction frames overlap significantly, and multiple overlaps may all revolve around the same target, which requires the NMS function to continuously IOU the highest scoring frame with the other frames, and then remove the frames whose IOU values exceed a given threshold. In this process, a large number of anchor frames with low scores and high overlap will be suppressed, resulting in missing detection and affecting the detection accuracy. IOU_Loss has two serious problems: when two coordinate frames do not intersect, IOU is 0, IOU_Loss is not conductive; if two cases of the same IOU occur, IOU_Loss function does not distinguish the two. To address this problem, subsequent research has produced a series of algorithm improvements, proposing a method of NMS filtering for location priority, which adds IOU prediction branches to the network, but tends to increase the computational effort. Therefore, this paper uses the improved DIOU-NMS[21] as the evaluation criterion of NMS to improve the missed detection problem in the obscured scene. DIOU_Loss considers both the distance between the center points of two prediction boxes and the area between prediction boxes. DIOU-NMS is shown in Eqs.(3) and (4).

$$\mathrm{DIOU}=\mathrm{IOU}-(\rho^2(b,b^{8t}))/c^2. \tag{3}$$

The DIOU is based on IOU considering the distance between the two bounding boxes PB and GT, IOU is the intersection ratio of PB and GT. b and b^{8t} denote the center point of PB and GT. $\rho^2(b, b^{8t})$ denotes the Euclidean distance between the center point of PB and GT. c denotes the shortest diagonal length of the minimum bounding box of PB and GT.

$$s_i=\begin{cases} s_i, \mathrm{DIOU}(M,B_i)<\varepsilon \\ 0, \mathrm{DIOU}(M,B_i)\geq\varepsilon \end{cases}. \tag{4}$$

When the M with the highest prediction score and the prediction frame B_i are less than the DIOU non-maximum threshold ε, they are removed. When the distance is too far greater than the set threshold value, it is considered that another target is detected, which is helpful to solve the problem of missing detection when the target occludes each other.

The improved YOLOv7 network model is shown in Fig.6.

The experiment is based on the PyTorch framework, using graphics processing unit (GPU) for training, and the specific configuration of the experimental environment is shown in Tab.1.

Since the size of the detection target in the custom datasets differs from the public datasets, the computational idea of auto learning bounding box anchors is used in YOLOv7 in this experiment, and the K-means clustering algorithm is used to automatically learn from the marked target frame to obtain the appropriate anchor frame, which improves the recall rate of the model, which is obtained from the training data automatically. The standard YOLOv7 has 9 anchor frames based on the COCO datasets, namely (12, 16), (19, 36), (40, 28), (36, 75), (76, 55), (72, 146), (142, 110), (192, 243), and (459, 401). For example, in this paper, in the DOTA1.0 datasets, anchor frames are reassigned according to the detection layer scale, and statistical anchor frames are assigned as (21, 26), (28, 22), (24, 46), (41, 37), (57, 58), (79, 106), (148, 123), (163, 291), and (336, 315).

For small targets with obscure boundaries, three anchor frame sizes of (10, 11), (25, 12) and (13, 23) were used to increase the detection of small targets in remote sensing images to achieve smaller classes. The mosaic data enhancement method is used in the training process to enrich the sample background and enhance the robustness of the network model. The sample is positive when the IOU of anchor box and real box is >0.45, otherwise it is negative. The batch size is set to 32, the initial learning rate is 0.01, and the epochs are set to 300. After considering the category richness, annotation quality and the number of small targets, the DOTA1.0 datasets and the

NWPU-VHR10 datasets are used as the base datasets to train the network model.

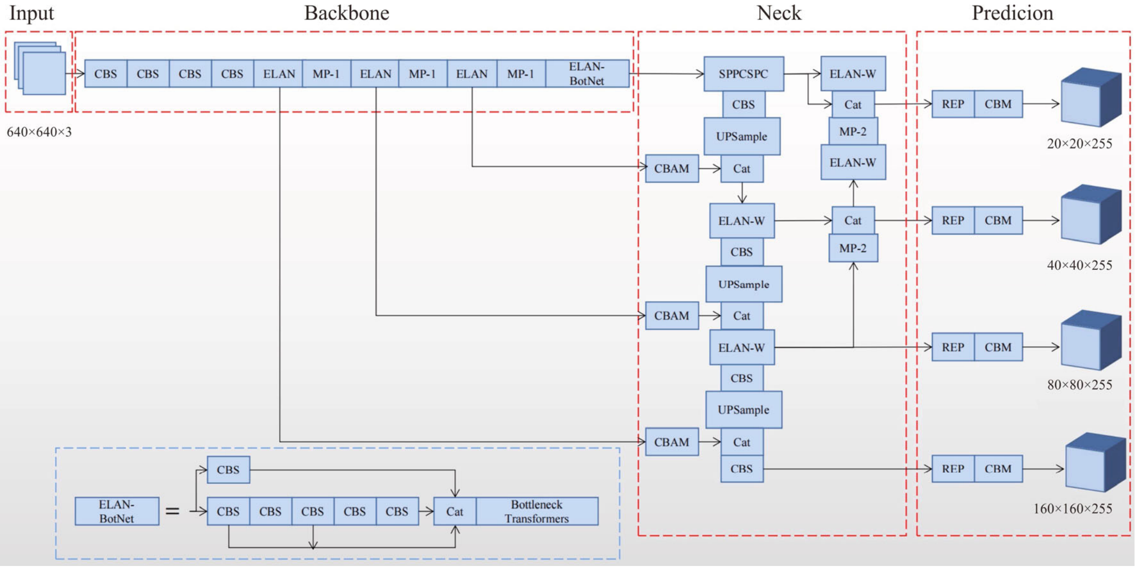

Fig.6 Improved YOLOv7 network structure diagram

Tab.1 Specific configuration of the experimental environment

Projects	Environment
Operating system	Windows 10 (x64)
CPU	IntelXeonE52680v4@2.40 GHz
GPU	NVIDIA Tesla P40 24 GB
PyTorch	10.1

When the external environmental conditions in the image change, such as changes in information, they can have an impact on the detection results. Target detection relies on the pixel values and features in the image to recognize and analyze the object, and when the external environmental conditions change, it may affect the visual information in the image, which in turn affects the detection results. In order to enhance the robustness of the model, various transformations are used to expand the training data during the training phase, such as randomly adjusting the brightness, contrast, and illumination, so that the model can better adapt to different environmental conditions. In this paper, splicing, cropping and scaling are used to expand the training data of remote sensing image datasets.

Fig.7 shows that data enhancement methods, such as mirroring, inverse color, scaling and adding noise, were used to expand the 650 data sets to 980, and randomly assigned to the training, test and validation sets in the ratio of 6: 2: 2.

The DOTA1.0 dataset contains 1 764 optical remote sensing images with label information, and the target categories are very representative. The varying image sizes within the datasets make it difficult to train the input network directly, so the DOTA1.0 dataset needs to be sliced and diced. Since the cut often leads to missing target information in the edge part of the image after the cut, a certain overlap region needs to be set. The data is expanded to 20 000 images by data processing, and the training and validation sets are assigned in a 3: 1 ratio.

The mean average accuracy (mAP) was used in this experiment to evaluate the performance of the model as shown in Eq.(5) and Eq.(6).

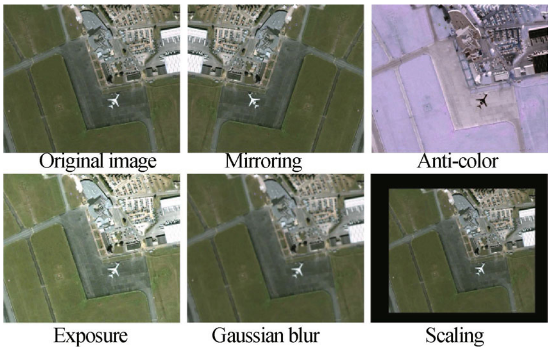

Fig.7 Image processing comparison chart

$$\text{precision} = \frac{TP}{TP+FP}, \tag{5}$$

$$\text{recall} = \frac{TP}{TP+FN}, \tag{6}$$

where *TP* denotes true case, *FP* denotes false positive case, and *FN* denotes false negative case. When the IOU between the real frame and the anchor frame of a single object is greater than 0.5, the anchor frame is defined as *TP*, otherwise it is defined as *FP*. The precision and recall are calculated according to the formula, and the precision curve is plotted, the step size is set to 0.1, and the precision value corresponding to recall of [0, 0.1, 0.2, ..., 1] is taken. The average of these precision values is AP, and the AP of each category is summed and averaged to obtain mAP.

In this paper, we perform comparison experiments on the NWPU-VHR10 dataset using RICNN, collection of part detectors (COPD)[22], rotation-insensitive and context-augmented object detection (RI-CAO)[23], and YOLOv5 with the standard YOLOv7 and improved YOLOv7 algorithms, while on the DOTA1.0 dataset using detection for small, cluttered and rotated objects (SCRDet)[24], feature-attentioned object detection (FADet)[25], RSDet[26], and YOLOv5 with the standard YOLOv7 and the improved YOLOv7 algorithm for comparison experiments. From the experimental results data in Tab.2 and Tab.3, we can see that the final mAP of the algorithm in this paper for NWPU-VHR10 dataset and DOTA1.0 dataset is 95.6% and 75.2%, respectively, which has good detection accuracy than other models as well. The improved YOLOv7 model has better detection performance than the standard YOLOv7 model on NWPU-VHR10 dataset and DOTA1.0 dataset, with an increase of 6.3% and 4.2%, respectively.

Tab.2 Accuracy table of training results of different algorithms on the data-enhanced NWPU-VHR10 dataset (%)

Method		RICNN	COPD	RI-CAO	YOLOv5	YOLOv7	Improved YOLOv7
Class	PL	83.4	62.2	99.7	99.4	99.5	99.5
	SH	77.3	69.3	90.8	90.9	90.1	91.2
	ST	85.2	64.5	90.6	75.5	73.4	94.3
	BD	88.1	82.1	92.9	99.3	98.9	99.0
	TC	40.8	34.1	90.3	94.1	93.2	97.5
	BC	58.5	35.2	80.1	87.8	86.7	93.2
	GTF	86.7	84.2	90.8	98.6	97.3	99.5
	HA	68.6	56.3	80.2	97.8	96.6	98.4
	BR	61.5	16.4	68.5	83.7	81.5	91.2
	VE	71.1	44.2	87.1	76.6	75.8	91.8
mAP (%)		72.1	54.9	87.1	90.4	89.3	95.6

Tab.3 Accuracy table of training results of different algorithms on DOAT1.0 dataset (%)

Method		SCRDet	FADet	RSDet	YOLOv5	YOLOv7	Improved YOLOv7
Class	SV	68.3	72.6	69.6	62.6	65.3	**76.4**
	LV	60.3	68.2	70.1	63.9	86.8	89.9
	PL	89.8	90.2	89.8	68.2	92.5	95.3
	ST	86.8	84.7	83.4	72.4	75.1	82.8
	SH	72.4	79.6	70.3	58.8	88.6	89.5
	HA	66.2	74.2	65.6	64.3	83.8	86.9
	GTF	68.3	76.4	65.2	60.7	67.2	69.8
	SBF	65.0	53.4	62.5	49.1	64.7	68.7
	TC	90.8	90.8	90.5	74.1	94.2	95.2
	SP	68.2	69.7	67.2	59.9	61.4	64.9
	BD	80.6	79.6	82.9	67.3	73.1	74.2
	RA	66.6	65.4	65.9	39.5	50.8	58.4
	BC	87.9	82.4	85.7	65.7	68.8	71.1
	BR	52.0	45.5	48.6	46.9	45.2	**49.4**
	HC	65.2	63.9	68.2	47.7	47.5	54.4
mAP (%)		72.5	73.1	72.4	60.1	71.0	75.2

But the detection effect of different categories varies greatly due to the very unbalanced target types in the DOTA1.0 dataset and the large differences in target size morphology. The detection effect for small vehicles (SV) is the most significant, while the detection effect for categories with more common shapes like bridges (BR) is not ideal, indicating that the huge differences in the datasets can lead to different detection effects between categories.

From Fig.8, we can see the improved training curve of YOLOv7, where Box is presumed to be the mean value of the DIOU_Loss function, the smaller the Box, the more accurate; Objectness is the mean value of the target detection loss, the smaller the target detection, the more accurate; Classification is the mean value of the classification loss, the smaller the classification, the more accurate; val Box is the validation set bounding box loss; val Objectnes is the mean value of target detection loss in the validation set; val Classification is the mean value of classification loss in the validation set; mAP@0.5: 0.95 denotes the average mAP at different IOU thresholds; mAP@0.5 denotes the average mAP at thresholds greater than 0.5; the main point is to observe the fluctuation of precision and recall. When the fluctuation is not very big, the training effect is better.

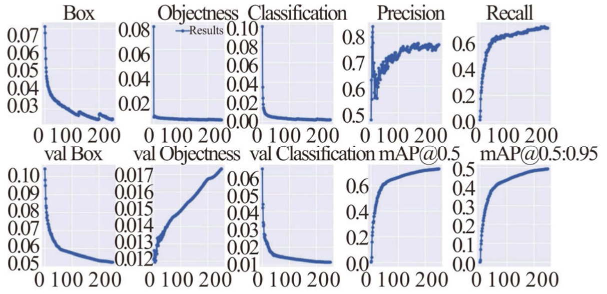

Fig.8 Improved YOLOv7 network model training curve

From Fig.9, the improved algorithm based on YOLOv7 in this paper can better identify small cars in-remote sensing images, has significantly improved the overall target detection accuracy, and has good detection of dense targets.

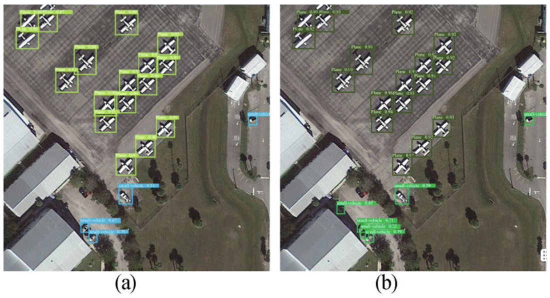
(a) (b)

Fig.9 Comparison of (a) the standard YOLOv7 algorithm and (b) the improved YOLOv7 algorithm

In order to verify the four improvements proposed in this paper for YOLOv7 methods, we add a small-scale detection layer (A-YOLOv7), fuse Bottleneck Transformers (B-YOLOv7), the attention module CBAM (C-YOLOv7), and use DIOU-NMS (D-YOLOv7) on the standard YOLOv7 model in order to judge the effectiveness of each improvement point, using the data-enhanced NWPU-VHR10 dataset, while keeping the same experimental configuration. The experimental results are shown in Tab.4.

Tab.4 Ablation experiments on the NWPU-VHR10 dataset based on the improved YOLOv7 algorithm (%)

Method		A-YOLOv7	B-YOLOv7	C-YOLOv7	D-YOLOv7	YOLOv7 -tiny	YOLOv7
Class	PL	99.4	99.4	99.6	99.5	99.4	99.5
	SH	82.9	83.5	86.0	83.7	83.9	90.1
	ST	99.4	99.4	99.5	99.0	98.2	73.4
	BD	99.1	99.0	98.6	99.1	98.5	98.9
	TC	94.8	94.9	95.1	97.0	96.2	93.2
	BC	94.2	80.7	94.7	92.8	48.1	86.7
	GTF	97.2	99.5	97.3	93.7	96.2	97.3
	HA	96.7	97.6	96.9	93.9	97.0	96.6
	BR	90.5	92.0	92.8	85.1	91.7	81.5
	VE	86.4	88.6	87.6	84.9	77.7	75.8
mAP (%)		94.1	93.5	94.8	92.9	88.7	89.3

The analysis of the results shows that adding a small-scale detection layer, fusing Bottleneck Transformers, introducing the attention mechanism, and using DIOU-NMS can improve 4.8%, 4.2%, 5.5%, and 3.6%, respectively, compared with the standard YOLOv7 network, indicating that adding the CBAM attention mechanism and fusing Bottleneck Transformers improves the feature extraction capability and enhances the multi-scale feature fusion of the network. Improving non-maximal suppression and adding multi-scale feature detection contribute to improving the fitting ability of the network.

Tab.5 and Tab.6 below list the training set results for all categories on the NWPU-VHR10 and DOTA1.0 data sets. It can be found that before and after the model improvement, the training results of the same category vary greatly. Especially when the detection target is large, the performance of the model can be improved.

Such as helicopters, roundabout, and large vehicle on DOTA1.0 dataset, basketball courts, tennis courts and storage tanks on NWPU-VHR10 dataset, the detection effect of small targets is also improved significantly, such as small vehicle on DOTA1.0 dataset and vehicle on NWPU-VHR10 dataset.

It can be seen from Tab.7 that the accuracy of improved YOLOv7 on the NWPU-VHR10 dataset has increased by 6.3% compared with YOLOv7, but the number of parameters has increased by 8.8M, and the number of floating-point calculations has increased by 22G. This is also the deficiency of this paper, which needs to be further optimized.

Tab.5 Training results for all categories in the NWPU-VHR10 dataset

Method		Precision (%)		Recall (%)		mAP50 (%)	
		YOLOv7	Improved YOLOv7	YOLOv7	Improved YOLOv7	YOLOv7	Improved YOLOv7
Class	PL	99.0	97.7	98.2	99.0	95.5	99.5
	SH	94.6	91.5	83.1	79.4	90.1	83.7
	ST	78.3	99.7	78.6	95.8	73.4	99.0
	BD	95.7	90.4	98.4	1	98.9	99.1
	TC	92.3	97.5	83.9	97.5	93.2	97.0
	BC	87.9	87.4	84.6	85.7	86.7	92.8
	GTF	1	93.6	88.9	1	96.8	93.7
	HA	98.2	95.4	89.1	93.6	96.9	93.9
	BR	94.4	92.1	66.7	80.0	81.8	85.1
	VE	91.4	95.0	60.0	82.6	76.0	84.9

Tab.6 Training results for all categories in DOAT1.0 dataset

Method		Precision (%)		Recall (%)		mAP50 (%)	
		YOLOv7	Improved YOLOv7	YOLOv7	Improved YOLOv7	YOLOv7	Improved YOLOv7
Class	SV	53.7	69.0	78.7	77.2	65.3	76.4
	LV	82.0	86.4	85.8	85.4	86.8	89.9
	PL	91.6	92.3	89.9	90.5	92.5	95.3
	ST	89.8	89.3	66.8	71.9	75.1	82.8
	SH	89.5	90.0	87.4	87.7	88.6	89.5
	HA	84.5	84.7	82.3	83.8	83.8	86.9
	GTF	79.2	82.7	60.4	63.6	67.2	69.8
	SBF	72.5	75.4	57.0	58.7	64.7	68.7
	TC	95.7	95.1	92.4	92.5	94.2	95.2
	SP	68.1	66.5	65.5	71.2	61.4	64.9
	BD	82.4	79.4	66.5	65.3	73.1	74.2
	RA	74.2	84.6	44.1	43.8	50.8	58.4
	BC	76.6	75.8	69.0	68.5	68.8	71.1
	BR	67.1	66.7	39.7	43.8	45.2	49.4
	HC	37.2	63.9	54.5	57.5	47.5	54.4

Tab.7 Test results of different models on NWPU-VHR10 dataset

Method	mAP (%)	#Param	FLOPs
YOLOv7	89.3	37.3 M	105.1G
YOLOv7-tiny-silu	86.8	6.1 M	13.2G
Improved YOLOv7	95.6	46.1 M	127.1G

In this paper, the improved YOLOv7 model is applied to the detection task of remote sensing images with a relatively high proportion of small targets and complex targets and different target scales. By adding small-scale detection layers and obtaining anchor frames for linear scaling operation through K-means algorithm clustering, the missed detection rate of remote sensing small-scale targets is reduced with better detection effect. Fusing Bottleneck Transformers in the backbone network significantly improves the baseline and also reduces the parameters with minimum delay overhead. Adding CBAM in the neck attention mechanism to make the algorithm locate and identify remote sensing image targets more accurately and reduce the influence of background on remote sensing target detection, so as to reduce the false detection rate of the network model on remote sensing image targets. Finally, the non-maximum suppression function is modified to improve the recognition effect of the network model on remote sensing image dense targets. The experimental results show that the optimized YOLOv7 model effectively detects remote sensing images, and the mean average accuracy and small target class accuracy are significantly improved.

Ethics declarations

Conflicts of interest

The authors declare no conflict of interest.

References

[1] LU X, ZHENG X, YUAN Y. Remote sensing scene classification by unsupervised representation learning[J]. IEEE transactions on geoscience and remote sensing, 2017, 55(9): 5148-5157.

[2] AFAQ Y, MANOCHA A. Analysis on change detection techniques for remote sensing applications: a review[J]. Ecological informatics, 2021, 63: 101310.

[3] ZHAO Z Q, ZHENG P, XU S, et al. Object detection with deep learning: a review[J]. IEEE transactions on neural networks and learning systems, 2019, 30(11): 3212-3232.

[4] SHAFIQUE A, CAO G, KHAN Z, et al. Deep learning-based change detection in remote sensing images: a review[J]. Remote sensing, 2022, 14(4): 871.

[5] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014, 978: 580-587.

[6] REN S Q, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis & machine intelligence, 2017, 39(06): 1137-1149.

[7] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//Computer Vision-ECCV 2016: 14th European Conference, October 11-14, 2016, Amsterdam, Netherlands. Berlin, Heidelberg: Springer International Publishing, 2016: 21-37.

[8] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 779-788.

[9] REDMON J, FARHADI A. Yolo9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 7263-7271.

[10] REDMON J, FARHADI A. Yolov3: an incremental improvement[EB/OL]. (2018-04-08) [2023-01-23]. https://arxiv.org/abs/1804.02767.

[11] ZHANG H, CISSE M, DAUPHIN Y N, et al. Mixup: beyond empirical risk minimization[EB/OL]. (2017-10-25) [2023-01-23]. https://arxiv.org/abs/1710.09412.

[12] WANG C, SHI J, YANG X, et al. Geospatial object detection via deconvolutional region proposal network[J]. IEEE journal of selected topics in applied earth observations and remote sensing, 2019, 12(8): 3014-3027.

[13] CHENG G, ZHOU P, HAN J. 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] YU X, GONG Y, JIANG N, et al. Scale match for tiny person detection[C]//2020 IEEE Winter Conference on Applications of Computer Vision (WACV), March 1-5, 2020, Snowmass, CO, USA. New York: IEEE, 2020: 1257-1265.

[15] LUO H, WANG P, CHEN H, et al. Object detection method based on shallow feature fusion and semantic information enhancement[J]. IEEE sensors journal, 2021, 21(19): 21839-21851.

[16] DENG C, WANG M, LIU L, et al. Extended feature pyramid network for small object detection[J]. IEEE transactions on multimedia, 2021, 24: 1968-1979.

[17] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: optimal speed and accuracy of object detection[EB/OL]. (2020-04-23) [2023-01-23]. https://arxiv.org/abs/2004.10934.

[18] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. Yolov7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 17-24, 2023, Vancouver, BC, Canada. New York: IEEE, 2023: 7464-7475.

[19] SRINIVAS A, LIN T Y, PARMAR N, et al. Bottleneck transformers for visual recognition[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE, 2021: 16514-16524.

[20] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Computer Vision-ECCV 2018: 15th European Conference, September 8-14, 2018, Munich, Germany. Berlin, Heidelberg: Springer International Publishing, 2018: 3-19.

[21] ZHENG Z, WANG P, LIU W, et al. Distance-iouloss: faster and better learning for bounding box regression[EB/OL]. (2019-11-19) [2023-01-23]. https://arxiv.org/abs/1911.08287.

[22] CHENG G, HAN J, ZHOU P, et al. Multi-class geospatial object detection and geographic image classification based on collection of part detectors[J]. ISPRS Journal of photogrammetry and remote sensing, 2014, 986: 119-132.

[23] LI K, CHENG G, BU S, et al. Rotation-insensitive and context-augmented object detection in remote sensing images[J]. IEEE transactions on geoscience and remote sensing, 2017, 56(4): 2337-2348.

[24] YANG X, YANG J, YAN J, et al. SCRDet: towards more robust detection for small, cluttered and rotated objects[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27, 2019, Seoul, Korea (South). New York: IEEE, 2019: 8231-8240.

[25] LI C, XU C, CUI Z, et al. Feature-attentioned object detection in remote sensing imagery[C]//2019 IEEE International Conference on Image Processing (ICIP), September 22-25, 2019, Taipei, China. New York: IEEE, 2019, 978: 3886-3890.

[26] QIAN W, YANG X, PENG S, et al. Learning modulated loss for rotated object detection[EB/OL]. (2019-11-19) [2023-01-23]. https://arxiv.org/abs/1911.08299.