IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received January 7, 2022, accepted January 28, 2022, date of publication February 9, 2022, date of current version February 18, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3150339

Target Detection of Forward-Looking Sonar Image Based on Improved YOLOv5

HAOTING ZHANG, MEI TIAN, GAOPING SHAO, JUAN CHENG, AND JINGJING LIU
PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China
Corresponding author: Haoting Zhang (zht08251008@163.com)

ABSTRACT Forward-looking sonar is a commonly used underwater detection device at present, but the detection accuracy is poor due to the complex underwater environment, small target highlight area and fuzzy feature details. Therefore, this paper proposes a forward sonar image target detection model based on You Only Look Once Version 5 (YOLOv5) network using transfer learning method. First, the YOLOv5 network is pretrained with COCO data set. Then the pre-training model is fine-tuned according to the training set of forward-looking sonar images. Before fine-tuning, the traditional k-means clustering is improved. The intersection over union (IoU) value is used as the distance function to cluster the labeling information of the training set of the forward-looking sonar image. The results of clustering serve as the initial anchor frame of the training network. This operation greatly improves the detection speed. Second, due to the characteristics of weak echo intensity and small target area of forward-looking sonar image, an improved feature extraction method of CoordConv was proposed to give corresponding coordinate information to high-level features which improves the accuracy of network detection regression. Finally, the fine-tuned network is used to detect the target in the forward-looking sonar image. The experimental results show that the improved model based on YOLOv5 network is superior to the original YOLOv5 network and other popular deep neural networks for target detection in the forward-looking sonar image, which has a reference significance for underwater target detection. The CoordConv-YOLOv5 network based on transfer learning proposed in this paper shows the best performance in both detection accuracy and detection speed. Detection accuracy mAP@0.5:0.95 can reach 56.95%, and detection speed can reach 9ms.

INDEX TERMS YOLOv5, forward-looking sonar, target detection, transfer learning, IoU k-means, CoordConv.

I. INTRODUCTION

The detection of targets in sonar images is an emerging topic in the field of target detection. In civil and military fields, it is of great significance to submarine landform mapping, underwater search and rescue, salvage, oil exploration and submarine suspicious target detection. In addition, sonar image can directly reflect underwater scene information, which pro-vides strong support for unmanned underwater vehicle (UUV) automatic target recognition technology. But the sonar image resolution is low, the reverberation is serious, and the effective feature is fuzzy, resulting in poor detection accuracy.

In the past decades, sonar images have been detected by artificial feature extraction, mainly based on pixel [1], [2], feature [3], [4] and echo [5], [6]. Most of this traditional method of underwater target detection based on pixel values features, gray threshold, or prior information of corresponding targets. However, the underwater environment is complicated, and the echo is affected by self-noise, reverberation noise and environmental noise, which leads to the low resolution, fuzzy edge details, and the serious speckle noise of the sonar image. So, it is difficult to find the good characteristics of pixels and gray level threshold. On the other hand, it is too expensive to obtain some prior information artificially for the underwater target is uncertain. Therefore, the current traditional algorithm is not accurate in detecting underwater targets.

To solve the above problems, deep learning detection method can extract multi-layer abstract features, which solves the trouble of manual feature extraction in traditional methods. In recent years, optical image target detection based on deep convolutional neural network has performed well. Inspired by this, researchers gradually apply convolutional

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

neural network to sonar image target detection. Among them, Fan *et al.* [7] proposed a feature extraction network constructed from residual blocks to replace Residual Network (ResNet) [8] in Mask RCNN [9]. While ensuring the detection accuracy, the training parameters of the network are greatly reduced, which lays a solid foundation for future engineering. Wang *et al.* [10] combined the pretreatment technology of bilinear interpolation with You Only Look Once Version 3 (YOLOv3) network for target detection of sonar images and got better results. Later, Sheng and Huo [11] constructed a simulation model for sonar mine detection, and combined simulation samples with real samples to use YOLOv3 network to detect them, which solved the problem of insufficient sonar image data. The experimental results show the validity of the simulation data. And in the latest research, Jin *et al.* [12] proposed the detection model of significant region segmentation and pyramid pooling to reduce the influence of background noise on target feature extraction. The features extracted by pyramid pooling are integrated with multi-scale features, which makes up for the fuzzy details of sonar images. Finally, in most current studies, researchers basically use YOLOv3 network or refer to its idea [31]. This network first introduces feature pyramid fusion in single-stage detection network, which can fuse multi-scale feature information to enlarge the receptive field. However, there is room for improvement in the deep convolutional network at the front, which can extract richer feature information, further deepen the feature pyramid, and integrate richer multi-scale feature information. To this end, YOLOv5 improves both deep convolutional network and feature pyramid that improves detection speed and detection accuracy.

At present, YOLOv5 network shows excellent performance in optical image target detection task, but its application in acoustic image target detection is lacking [32], [34]. In addition, the lack of sonar image data makes it difficult to apply in deep neural network. Therefore, this paper proposes an improved YOLOv5 forward-looking sonar image target detection model based on transfer learning [13]. Aiming at the target detection task of forward-looking sonar image, IoU was introduced as a distance [15] function to improve the traditional k-means algorithm [15] in obtaining the prior anchor frame, so as to improve the detection performance. The training method of transfer learning is used to deal with the problem of insufficient image data set of forward-looking sonar. In addition, according to the target characteristics of forward-looking sonar image, CoordConv [16] is introduced to extract the features with coordinate information, which effectively improves the detection accuracy of small targets in the forward-looking sonar image. This paper first introduces YOLOv5 and some networks with good performance in the current target detection field. Then, the improved YOLOv5 forward looking sonar image target detection model based on transfer learning is introduced. And through experimental test and performance comparison, the improved YOLOv5 model has more efficient detection performance in sonar image target detection model than YOLOv5, YOLOv3, YOLOv4, Faster R-CNN, EfficientDet and some self-built models. Finally, the experimental summary and prospect of the present work are given to lay a foundation for underwater target detection in the future.

II. RELATED WORK

In this section, the current mainstream target detection networks are introduced, including the Faster R-CNN, EfficientDet and YOLO models. Among them, the YOLOv5 network is introduced in detail, which lays a foundation for the construction of the YOLOv5 network of forward-looking sonar target detection in the third part.

A. FASTER R-CNN

Faster R-CNN [20] is a two-stage target detection network proposed by Ross B. Girshick in 2016. It is a deep convolutional neural network. Features were extracted by deep convolutional network (ResNet50 [8]), and the extracted features were input into the Region Proposal Network [18] (RPN) to screen candidate regions. Softmax [19] was used to determine whether the candidate regions were targets, and the candidate regions were corrected by bounding box regression. Finally, the proposed regional location information and the last feature layer are input into return on investment (ROI) pooling [20] to form a feature layer of uniform scale, and then enter the fully connected layer for category prediction and location regression. Faster R-CNN uses convolutional network to generate suggestion frames by itself, and shares the convolutional network with target detection network, which reduces the number of suggestion frames from about 2000 to 300, and the quality of suggestion frames is also substantially improved.

B. EFFICIENDET

EfficientDet is a single-stage target detection model proposed by Tan *et al.* [21]. It is also a convolutional neural network. The single-stage target detection network does not need to extract candidate regions, only to extract features through deep convolutional network and then fuse the features of different scales in the neck to enrich feature information. Finally, the features of different scales are input into the detection module for detection. In this network, an efficient weighted bidirectional feature pyramid network (BiFPN [21]) was proposed, which introduced learnable weights to learn the importance of different scale features extracted from the backbone network of EfficientDets. BiFPN repeatedly applies top-down and bottom-up sampling methods to fuse multi-scale features. Finally, the fused features are input into the class prediction network and the box prediction network for detection. It is worth mentioning that the performance of the network will increase with the depth of the network based on the stacking of BiFPN.

C. YOLOv5

Compared with two-stage detection networks such as Faster R-CNN [22] and Faster R-CNN [17], YOLO [23]

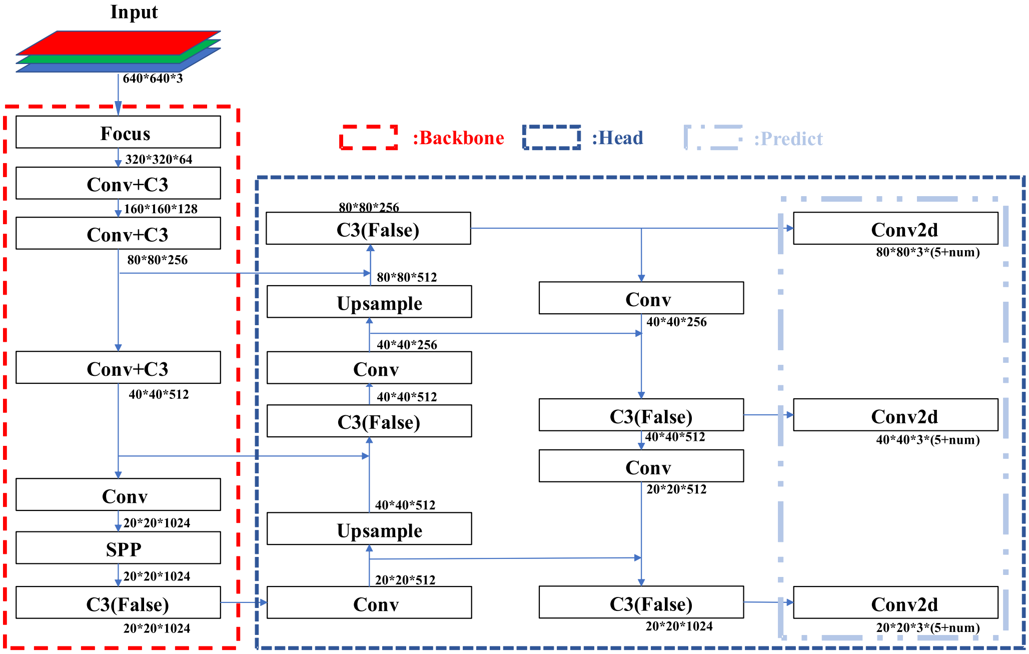


FIGURE 1. **YOLOv5 network model.**

(You only look once), as a kind of convolutional neural network of single-stage mode, does not need to generate candidate regions, conducts coordinate regression directly through the grid that greatly improve the speed of target detection. YOLOv1 model [23] combines with context information to predict targets based on the global information, which reduces the detection false alarm rate. It can provide good help for sonar images with serious reverberation noise. However, the grid resolution of YOLOv1 is not high, which leads to low prediction accuracy. What is more, each grid of the network can only predict one target, which results in the bad prediction performance of the network for small cluster targets.

From the perspective of target detection task of sonar image, the two improvement schemes of YOLOv2 [24] are helpful for us. Firstly, k-means clustering algorithm is used to obtain the prior anchor box to predict each grid. By predicting the offset between the anchor frame and the real frame, the learning task is simplified, and the recall rate of the target is effectively improved. Secondly, a feature layer is added to improve the resolution of the feature layer, which can increase the accuracy of small target detection. YOLOv2 uses Darknet19 [25] on the backbone network to reduce the amount of computation, but it limits the range of the receptive field and results in the limited information presented by the feature layer for it does not do feature fusion, which harms the detection performance to some extent finally.

Therefore, YOLOv3 [26] enhances the characteristics of the acceptance domain effectively through introduce the idea of spatial pyramid pooling (SPP) [26] into the backbone and improve Darknet19 into Darknet53. In addition, Feature Pyramid Networks (FPN) [27] module is referenced in the neck, which integrates low-level physical features into high-level semantic features that further improve the accuracy of recognition and location. This has achieved tremendous improvements for sonar image target detection.

However, the neck based on FPN module is immature. Path Aggregation Network (PANet) [28] module is used in YOLOv4 [29]. In addition, Darknet53 is improved into CSPDarknet53 in feature extraction network. For sonar image, feature layer fusion based on the same scale is more conducive to the target context information acquisition.

It is worth mentioning that the addition of a neck module in the test may improve the accuracy of the test, and the damage performance. According to current studies, the BiFPN module [21] proposed by Tan has excellent performance in VOC data set test. However, both BiFPN and PANet, introduced into YOLOv3 in this paper, have com-promised the final detection performance. So, the complexity of the neck feature fusion module needs to consider the degree of low-level semantic features extracted from the backbone and the scale of data volume.

According to the above analysis, the evolution of YOLO network has a certain effect on the characteristics of sonar images. At present, compared with YOLOv4, YOLOv5 has added Focus layer in Backbone, as shown in FIG. 1. which improves floating-point operation per second (FLOPS) without affecting accuracy. In addition, introduced C3 modules to both Backbone and PANet to promote feature fusion, which has stronger performance compared with content security policy (CSP) module in YOLOv4. Moreover, YOLOv5 network can adaptively change the depth and width of the network by changing parameters to adapt to its own data volume scale. This improvement is undoubtedly beneficial to sonar image target detection. It can extract more abundant global information to reduce false alarm rate, improve detection efficiency through data segmentation, and adjust model parameters according to the size of sonar image data set to achieve self-adaptation.

As shown in Fig. 1, YOLOv5 network model is divided into backbone and Head. The input is down sampled for five times through backbone of the network, and then takes the last three feature layers as the input of head. In backbone network, data first passes a focus layer to cut the width and thickness of data to half of the original, greatly speeding up the forward transmission speed of the network. Then it passes through four convolution layers plus a C3 layer. It proposed a new residual component (C3 layer) which is improved since CSP layer in YOLOv4 network. With the help of C3 layer, the backbone network can extract more detailed features. In addition, YOLOv5 replaced the activation function in the CSP layer following the improved activation function method in YOLOv4. Finally, it should be noted that a SPP module was added in the middle of the last Conv+C3 layer. This module was proposed from YOLOv3, which can expand the receptive field of feature layer and play a great role in feature fusion of Head module.

In the head part of YOLOv5, the main idea is the same as that of YOLOv4, which is improved based on PANet. It should be noted that YOLOv5 uses a different C3 layer pattern in head than backbone. In Backbone, the feature layer only carries out down sampling, so the C3 layer adopts residual combination. However, in head, PANet network integrates the feature layer of up-down sampling, so the residual combination is no longer used in C3 layer.

Finally, use the corresponding anchor frame for target detection in the feature layers of the three scales. Each grid of each feature layer can get $3 * (5 + num_{obj})$ detection results, where is the total number of categories of input data. The quantity 3 means that each grid is tested through 3 anchor frames, and the quantity 5 represents the five information of the target confidence predicted by each anchor frame, the coordinates of the center point and the width and height of the prediction frame.

The loss function of YOLOv5 consists of classification loss, confidence loss, and regression loss of target prediction. The total loss can be expressed as follows:

$$L_{total} = \alpha l_{box} + \beta l_{obj} + \lambda l_{cls} \tag{1}$$

where l_{box}, l_{obj} and l_{cls} respectively represent target regression loss, target prediction confidence loss and category loss. Restricted by the optimizer, corresponding gain should be added to each type of loss in target detection to scale and balance each type of loss [30]. In YOLOv5, the three gains α, β, and λ were determined to be 0.05, 1 and 0.5 by many trials.

Equation (2) is the calculation method of complete intersection over union ($CIoU$). And the predicted losses are obtained according to the $CIoU$ of target real box and predicted box. Compared with IoU and generalized intersection over union ($GIoU$), $CIoU$ not only contains all functions of them, but also can better handle the two cases in which the prediction box and the real box are completely separated or fully included.

$$CIoU = IoU - \frac{dis^2\left(a_{pre}, a_{tr}\right)}{c^2 + e} - \alpha v \tag{2}$$

where IoU represents the ratio of the intersection and union of the real frame and the prediction frame, a_{pre} and a_{tr} respectively represent the position of the center point of the prediction frame and the real frame, c represents the diagonal length of the minimum frame surrounding the real frame and the prediction frame, and e is a bias item. $CIoU$ added the punishment of αv compared with $DIoU$, where

$$v = \frac{4}{\pi^2} * \left(\arctan\frac{w_{tr}}{h_{tr}} - \arctan\frac{w_{pre}}{h_{pre}}\right)^2,$$

w_{tr}, h_{tr}, w_{pre} and h_{pre} are the width and height of the real box and the predicted box respectively, and $\alpha = v/(v - IoU + 1 + e)$.

So far, the regression loss of target positioning can be obtained as follows:

$$l_{box} = 1 - CIoU \tag{3}$$

The target prediction confidence is used to judge the probability of the existence of the target in the anchor frame, and the category prediction output the target category in the anchor frame. Directly the binary cross entropy calculation method is used to calculate the loss of the two for both are only one parameter. For confidence loss, there are:

$$l_{obj} = \frac{1}{n}\sum_{nt}\left(y_p * \ln\left(y_t\right) + \left(1 - y_p\right) * \ln\left(1 - y_t\right)\right) \tag{4}$$

where nt refers to the number of positive samples in a batch of images at a certain scale, y_p is a confidence obtained by the predicted target, and y_t is a confidence standard, which is defined as:

$$y_t = (1 - gr) + gr^* CIoU \tag{5}$$

where is the scale coefficient of a $CIoU$ with a range of [0, 1]. It is important that the confidence level is closer to 1 when g_r is too small, which will increase the difficulty of training. Finally, the confidence loss can be obtained through (4) and (5). For category loss, the calculation method is like confidence loss, and there is no need for standard, because we have real tags and target location information of each target.

Finally, by giving gain to each type of loss and summing it up, the loss of the image can be obtained from the output of a feature scale. The total loss of the image is obtained by adding and averaging the losses of the three scales. Then, the losses are multiplied by the batches to obtain the total loss value, which is passed to the stochastic gradient descent (SGD) optimizer to update the network weights.

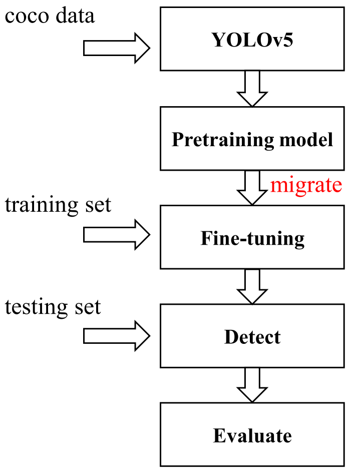


FIGURE 2. Training, detection, and evaluation process based on transfer learning.

The results show that the performance of these networks is better in the current optical target detection field. However, they are often ignored in the application of acoustic image target detection. Therefore, this paper uses these methods to test the target detection of the forward-looking sonar image and improves YOLOv5 accordingly.

III. PROPOSED METHOD

In this part, an improved YOLOv5 detection model based on transfer learning is proposed for the image target detection task of forward-looking sonar, which includes transfer learning training structure, *IoU* k-means clustering improved algorithm, CoordConv-YOLOv5 network and forward-looking sonar target detection model.

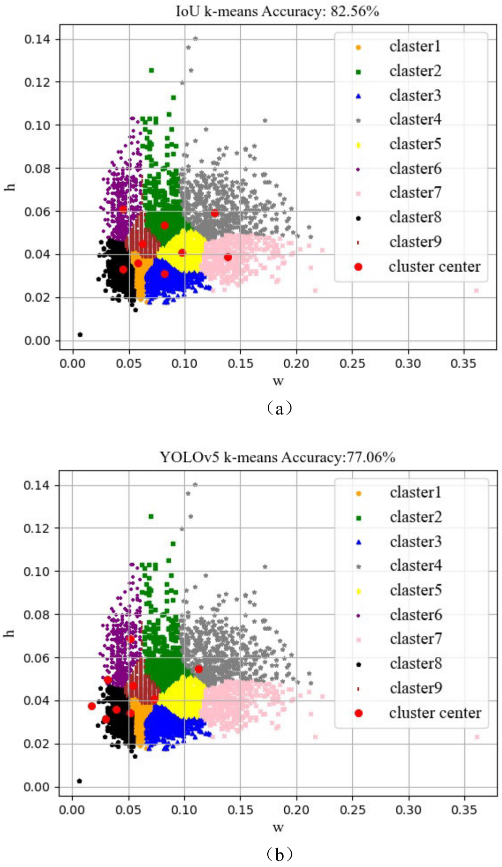
A. TRANSFER LEARNING TRAINING STRCTURE

It is difficult to obtain enough effective sonar image data because of the difficulty and high cost of underwater experiment. For deep convolutional neural networks, training fitting needs the support of many training samples. Among them, the number of YOLOv5 training samples on Pascal VOC and COCO public data sets was 20 categories, a total of 16,551 training images, and 80 categories, a total of 118,287 training samples. However, the data set of forward-looking sonar images available in this paper was only 3,240 training samples, a total of 8 categories. Regardless of whether the sample distribution of each category is uniform, it can be seen from statistics that the average amount of training image data of each category of available forward-looking sonar images is much lower than that of the other two data sets. This difficulty will greatly harm the convergence performance of the network.

To solve the problem of insufficient training data, this paper uses the method of transfer learning to train the forward-looking sonar images. The weight trained by YOLOv5 network is put on the optical image data set as the pre-training weight of the target data set, and the target data set can be fine-tuned on the weight of the model to achieve better fitting effect. Among them, the fitting model of the reference optical image is reliable as a pre-training model of the acoustic image and the training will not fail because the low-level features in the deep network are common to different tasks. As shown in Fig. 2, the weights of YOLOv5 network model trained under coco data set are used to fine-tune the target's forward-looking sonar image data set. Finally, the training model is received for the forward-looking sonar image target detection task. Later, the test set of the forward-looking sonar image is detected by the training model of the task and evaluate the detection result.

B. IOU K-MEANS

Started from YOLOv3 network, the anchor frame is introduced based on the idea of Fast R-CNN. YOLOv3 and YOLOv4 both refer to k-means clustering method in the calculation of anchor frames. YOLOv5 transforms the manual calculation anchor frame mode of YOLOv4 into automatic calculation mode and combines k-means clustering with genetic algorithm (GA). However, the overall method is still the traditional k-means clustering mode, using Euclidian distance function as the clustering basis.



(a)

(b)

FIGURE 3. The class distribution and clustering results of the width-to-height ratio of the dataset object under different methods: (a) Improved IoU k-means clustering method. (b) Traditional k-means for YOLOv5.

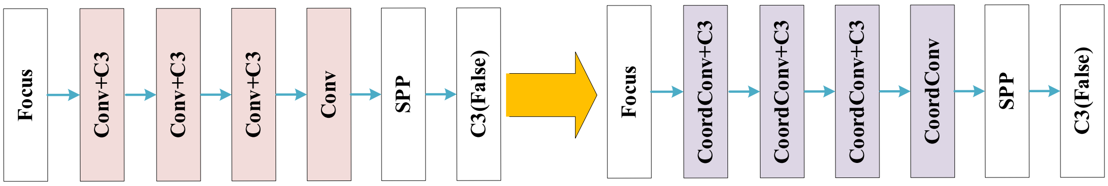


FIGURE 4. Introduce coordinate information to YOLOv5 backbone network.

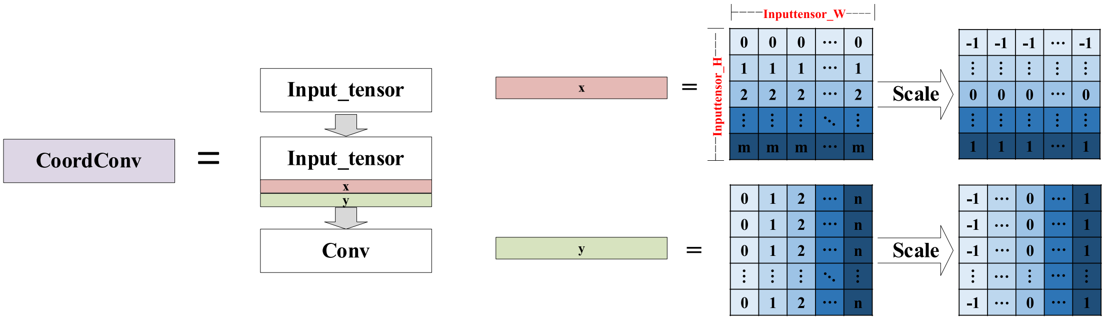


FIGURE 5. CoordConv workflow.

There is a one thing to note that the width-to-height ratio of training set annotation in this paper differs greatly from that of COCO dataset. Therefore, the initial anchor frame of the target data set needs to be redefined before the training of transfer learning. Firstly, the k-means algorithm combined with GA algorithm of YOLOv5 is considered.

However, using Euclidean distance function will cause larger target boxes to have larger deviations than smaller ones. What is more, the positioning accuracy of sonar image target detection task is a great consideration, this paper proposes to use IoU as the clustering function of k-means algorithm. Calculating the anchor frame closest to the target size by means of IoU distance function can further improve the accuracy and speed of regression prediction.

The overall process of the IoU k-means algorithm proposed in this paper is as follows:

Step1: All labeled coordinate information are counted in the training set and the width length and height length is normalized between 0 and 1.

Step2: Nine target widths and heights are randomly selected as the initial frame widths and heights.

Step3: Calculate the center of each target frame, and position nine initial anchor frames at the center to calculate the IoU value between the target frame and the nine anchor frames.

Step4: Record the IoU calculation results between all anchor frames and the current target frame, including the number, width, and length of anchor frames which has the maximum IoU value.

Step5: Calculate the median value of the width and height of all target frames corresponding to the number as the new width and height of the anchor frame.

Step6: Go to step3 and recalculate the IoU value until the target enclosure id in the last round is the same as that in the current round.

Step7: The update ends. The obtained anchor frame scale is converted to the corresponding scale size in the input image size.

The IoU k-means clustering algorithm proposed in this paper can effectively improve the accuracy of average IoU compared with the traditional k-means clustering algorithm. As shown in Fig. 3, the different colors represent the distribution of width and height of the training set labels. Here, the division of the overall category presents the clustering situation obtained by using IoU k-means, and the five-pointed star represents the width and height distribution of nine anchor frames obtained by different clustering methods. By evaluating the average IoU accuracy of the two algorithms, the introduction of IoU as a distance function does better than the traditional k-means used in YOLOv5 on the target clustering of the forward-looking sonar data set, and the average IoU accuracy is improved by 5.5%.

C. COORDCONV-YOLOv5

Since the imaging mechanism of forward-looking sonar is different from that of optical equipment, the echo intensity received by the receiver is greatly weakened by the complex underwater environment. the target presented in sonar image has the characteristics of low contrast and blurred edge. In addition, the most important is that under the influence of long-distance detection, beam Angle resolution and acoustic scattering, the target echo area in sonar image is small, which further aggravates the damage of regression positioning in the detection process.

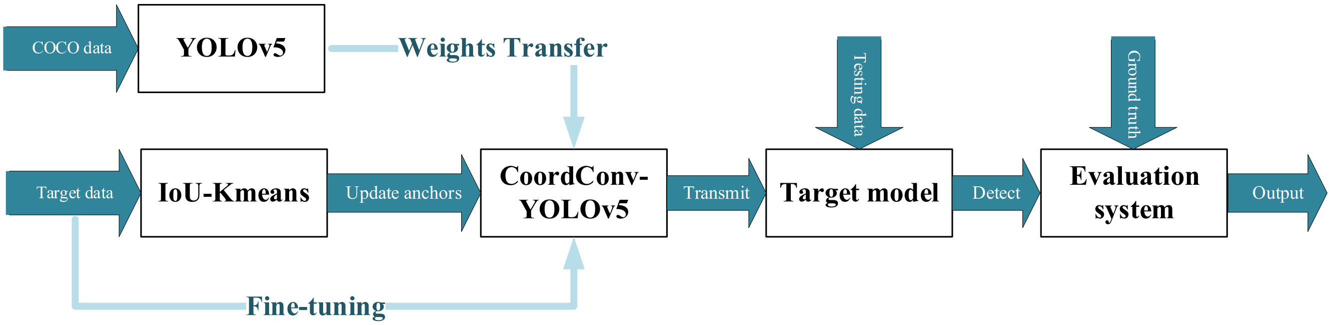


FIGURE 6. Training and test flow chart of re-initializing anchor frame.

In view of the above problems, this section proposes to use CoordConv to improve the feature extraction module of YOLOv5 network. It not only extracted the multi-scale high level features, but also introduced the corresponding coordinate information for different scale features. The addition of coordinate information can improve the accuracy of positioning regression of detection module.

As shown in Fig. 4 CoordConv is adopted in this paper to replace the original convolution form of YOLOv5 backbone network, and the details of CoordConv are explained in Fig. 5. Before extracting parameterized features each time by convolution, each pixel of the input tensor is given corresponding coordinate information. There is x and y coordinate information. Since the tensors vary in size during feature extraction each time, x and y are further scaled to between [-1,1]. The generated two two-dimensional coordinate matrices are assigned to the corresponding tensor channel layers. Through conv layer operation, the corresponding high-level features with pixel-level coordinate information are further obtained, which provides more abundant feature information for locating regression in Head layer. Importantly, the extracted high-level features have corresponding coordinate information, which is of great help to improve the detection and positioning accuracy of small targets in the forward-looking sonar image.

D. FORWARD-LOOKING SONAR IMAGE TARGET DETECTION MODEL

In this section, an improved model based on YOLOv5 is proposed for the target detection task of forward-looking sonar image.

As shown in Fig. 6, firstly, the improved IoU k-means algorithm is used for clustering the forward-looking sonar training set annotations to obtain the anchor frame information required for training. Secondly, the feature coordinate information is added in the feature extraction module, and the convolution with coordinates is carried out to improve the regression accuracy. Then, we fine-tune the pre-training model, and receive the target model with the best training performance. And the testing set is used to test the training model. Finally, the performance of the detection results is evaluated and the evaluation indexes such as mAP of the detection results are obtained.

IV. EXPERIMENTS

This section conducts target detection on forward-looking sonar images based on YOLOv5 model.

The data set used is URPC2021, and the website address is https://code.ihub.org.cn/projects/14186, in which there are 4000 forward-looking sonar images, which are respectively composed of eight types of targets: human body, ball, round cage, square cage, tire, bucket, cube and cylinder. The data was obtained by tritech's forward-looking sonar, which detected eight types of targets in the real ocean. Sonar image data is often acquired together with water depth detection and bottom detection data, which enables us to observe the shallow structure of the seabed. This data set was launched by Pengcheng Laboratory, which is currently the largest and most extensive acoustic image data set in the industry.

The experiment in this paper is conducted on Pytorch deep learning platform under the environment of Intel(R) Core (TM) I7-10710U CPU and Quadro P5000 GPU. During the experiment, different models are trained and tested based on transfer learning method, and then compared and analyzed the performance of the algorithm. For the data set division of each algorithm, uniformly used a seed to randomly select and divide the data set into training set, validation set, and test set in a ratio of (9:1):1 to ensure the authority of the comparison of algorithm results.

A. EVALUATION INDEX OF DETECTION PERFOMANCE

In this paper, the average accuracy (AP) is used to evaluate the performance of each network in forward-looking sonar image target detection. AP is controlled by the following two parameters: P (Precision) and R (Recall). First, calculate the IoU value between the model detection result and the label. When the IoU ratio is greater than or equal to the set threshold and the category judgment is correct, judge the prediction target to be True Positive (TP). Then, all the predicted targets of this class are sorted by the confidence score, and count the

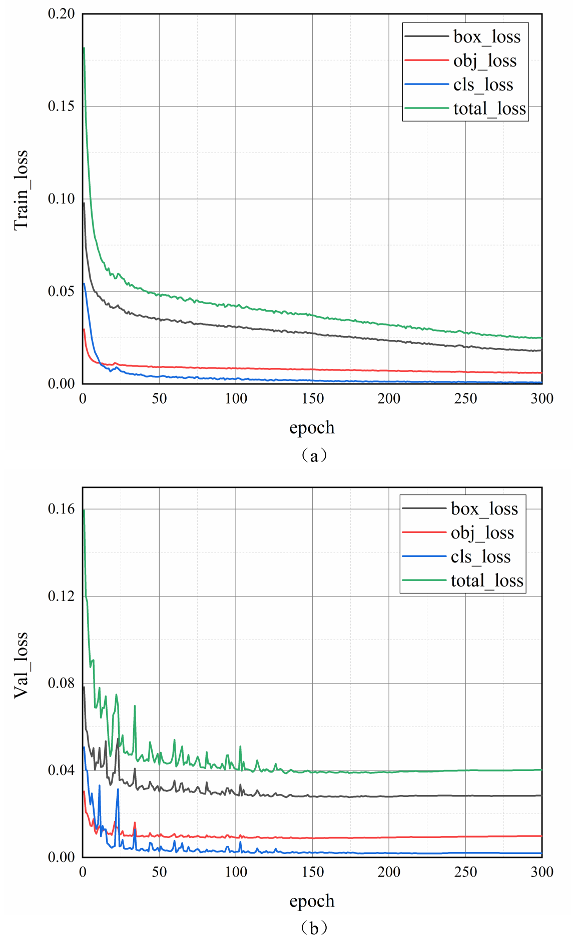


FIGURE 7. Four types of loss curves for each data set: (a) Training set; (b) Validation set.

TP values obtained into P-R curve, which is defined as:

$$P = \frac{TP}{TP+FP}$$
$$R = \frac{TP}{TP+FN} \tag{6}$$

where FP (false positive) represents the number of targets whose predicted categories are inconsistent with the real ones, which may be detection errors or false alarms. FN (false negative) is the number of real targets that are not detected, which is the representation of alarm leakage. According to the above, P-R curve of each type of target under a certain *IoU* index can be obtained. The AP value of the corresponding class is the area between P-R curve, P axis and R axis. With the improvement of *IoU* index, the localization regression standard of target detection becomes stricter. In this experiment, standard *IoU* standards of 0.5 and 0.75 were taken respectively to compare the accuracy. In addition, calculate the average value of ten *IoU* thresholds from 0.5 to 0.95 to measure the average accuracy of each model. And the time required for each image detection by the model is used for comparison and measurement for the efficiency of sonar image detection by the model.

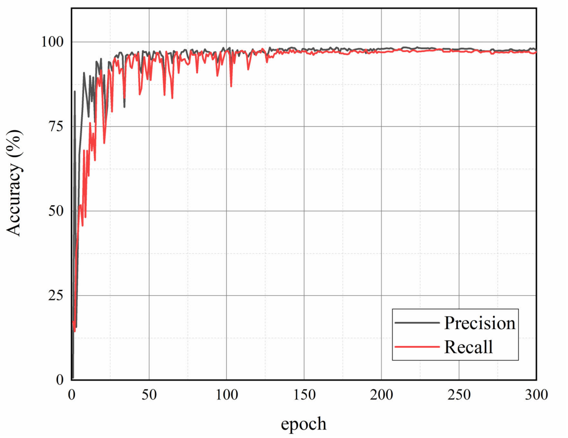


FIGURE 8. Precision and recall curves of validation sets.

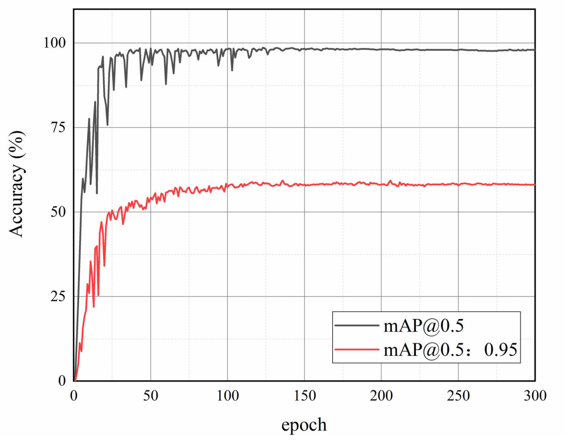


FIGURE 9. mAP test curve of validation set.

B. RESULTS

Fig. 7 shows the loss reduction curves of the training set and verification set of sonar images based on the improved YOLOv5s model based on transfer learning. It can be seen from the loss curve that the network gradually tends to be stable after the 100th turn. According to the network output, the output dimension of the three anchor frames on each grid of each layer is 13 dimensions, including eight categories, four position information and one confidence loss. Category loss and confidence loss are calculated by cross entropy, which contain fewer parameters. So, the loss can be reduced compared to the regression predicted loss calculated using *CIoU*.

Fig. 8 shows the statistical change process of accuracy rate and recall rate of verification set in the training process. By observing the stability curve and loss curve of accuracy rate and recall rate of verification set, the final stability of network training can be judged. As shown in Fig. 9, mAP for the verification set was obtained by testing after each training round, and two performance indexes mAP@0.5 and mAP@0.5:0.95 were recorded respectively. When the network is stable, mAP@0.5 and mAP@0.5:0.95

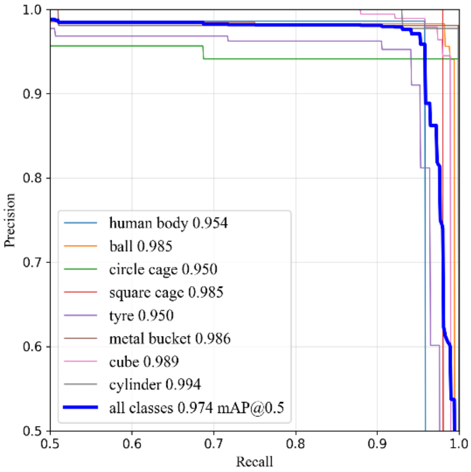
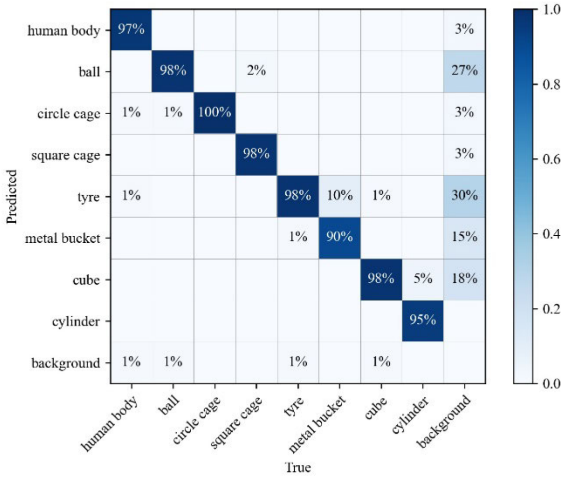


FIGURE 10. Test accuracy and recall curves.



Predicted \ True	human body	ball	circle cage	square cage	tyre	metal bucket	cube	cylinder	background
human body	97%								3%
ball		98%		2%					27%
circle cage	1%	1%	100%						3%
square cage				98%					3%
tyre	1%				98%	10%	1%		30%
metal bucket					1%	90%			15%
cube							98%	5%	18%
cylinder								95%	
background	1%	1%			1%		1%		

FIGURE 11. Confusion matrix of test results.

can reach 98.64% and 57.94% respectively. Finally, we can obtain the training model corresponding to the optimal index mAP@0.5:0.95 of the validation set.

In the process of testing the test set, to reduce a certain false alarm rate, the confidence threshold of detection was set as 0.6, and finally the test results were visualized under the *IoU* index of 0.5. Fig. 10 shows the change curves of accuracy rate and recall rate of detection results of eight types of targets under the forward-looking sonar. According to statistics, the overall detection result mAP@0.5 is 98%. In addition, as shown in Fig. 11, the detection results of each type of target are visualized in the form of confusion matrix.

For each type of target, the detection accuracy is relatively high, and the target miss rate is relatively low. However, network misjudged part of the background as the target because of the influence of background noise in the sonar image. So, the noise reduction part of the sonar image needs to be further solved in the subsequent research.

Secondly, the loss curve shows that the improved CoordConv-YOLOv5 model can achieve better performance than the baseline YOLOv5 model in Fig. 12. By observing the comparison curve of training loss, the improved CoordConv-YOLOv5 can achieve a lower loss stability than YOLOv5. In addition, the loss stability achieved by CoordConv-YOLOv5 was nearly 0.007 lower than the baseline model in terms of validation set losses. This result further verifies that the improved algorithm has better convergence effect. Therefore, compared with the baseline model, the target regression positioning results of forward-looking sonar images with high indicators can achieve better performance.

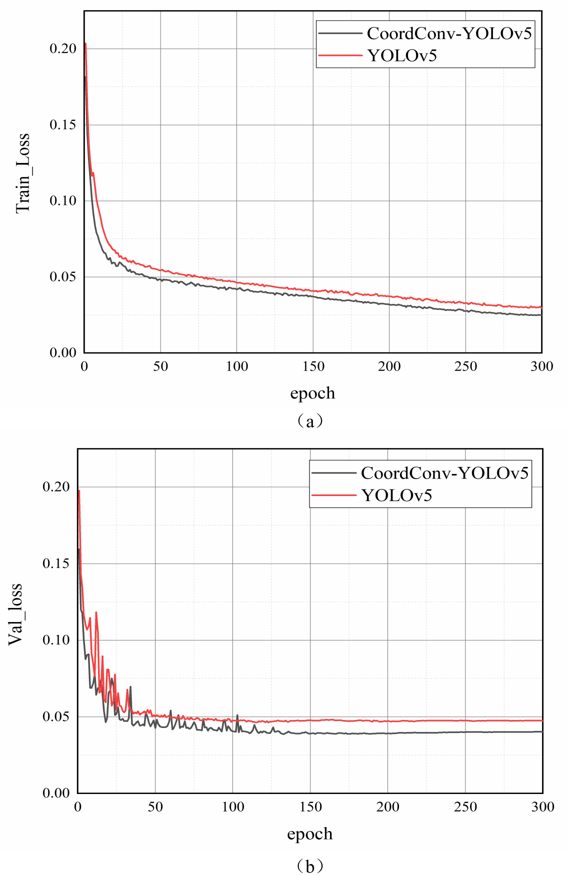


FIGURE 12. Comparison of loss change curves of different data sets of YOLOv5 and improved network: (a) Training set; (b) Validation set.

In addition, Fig. 13 verifies the above results from the accuracy change curves under different indexes of the verification set. In the figure, the performance of the improved CoordConv-YOLOv5 and YOLOv5 is evaluated by mAP@0.75 and mAP@0.5:0.95 respectively. In mAP@0.75 index, the improved model is 9.26% higher than the original model on average after the verification set reaches stability. In mAP@0.5:0.95 index, it was 6.04% higher on average.

Through the visualization of the above two training results, it can be seen that compared with the original model, the improved model has improved the detection accuracy of the forward-looking sonar image to a certain extent.

Finally, each typical target detection model is trained and tested, and the performance of each model is compared and analyzed. As shown in Table 1.

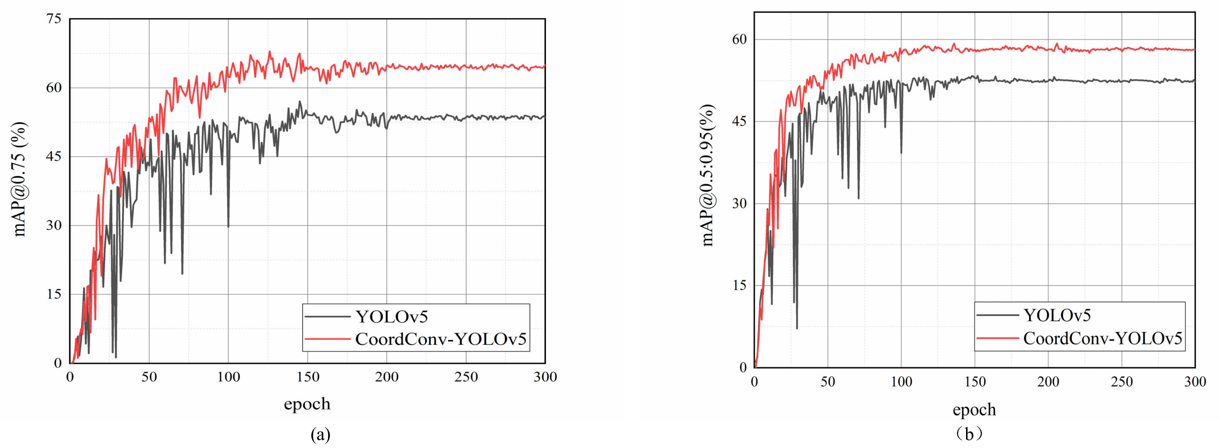


FIGURE 13. Comparison of different performance indexes between YOLOv5 and improved network on verification set: (a) mAP@0.75; (b) mAP@0.5:0.95.

TABLE 1. Detection results of eight types of targets in forward-looking sonar images by different models.

Method	Image size	mAP@0.5 (%)	mAP@0.75 (%)	mAP@0.5:0.95(%)	Detection speed(s)
YOLOv3	608*608	93.91	46.34	43.562	0.025
YOLOv3-PAN	608*608	94.04	47.26	43.952	0.035
YOLOv4	640*640	95.64	54.82	49.361	0.030
Faster R-CNN(ResNet-50)	608*608	53.53	24.23	20.213	0.275
EfficientDet-d0	608*608	60.02	32.04	30.341	0.118
EfficientDet-d4	608*608	86.56	38.90	40.864	0.418
YOLOv5s	640*640	98.00	56.30	54.21	0.013
YOLOv5m	640*640	97.95	56.52	54.23	0.025
YOLOv5l	640*640	**98.31**	56.60	54.25	0.047
YOLOv5x	640*640	98.30	56.62	54.30	0.071
YOLOv5s-IoU k-means	640*640	97.12	56.56	54.22	**0.008**
CoordConv-YOLOv5	640*640	97.40	**62.12**	**56.95**	0.009

Through the comparison of the training and testing performance of each model, our method is found to exhibit the best performance in the target detection process of the forward-looking sonar image, and the detection accuracy mAP@0.5:0.95 reaches 56.95%.

Firstly, compared with the model trained by the pre-training model of YOLOv5s only, the model trained by the anchor frame obtained by the improved *IoU* k-means algorithm improves the detection speed by one-third. However, the detection accuracy is not improved because the network framework is not improved.

Secondly, the improved target detection network of YOLOv5 forward looking sonar image proposed in this paper introduces coordinate information into the backbone network and gives corresponding coordinate information to the extracted deep parameterized features, which effectively improves the detection regression positioning accuracy. At the same time, this paper improved the traditional k-means algorithm. Compared with the original anchor frame, the anchor frame obtained by clustering algorithm with *IoU* as the distance function has a similar size to the target, which helps the detection module to achieve faster regression positioning and effectively improves the detection speed. In addition, by fine-tuning the pre-training model, the problem of insufficient data is well solved, and the effect of fitting is improved.

Compared with the test results of the original YOLOv5 network, it can be concluded that the CoordConv-YOLOv5 forward-looking sonar image target detection network based on transfer learning proposed in this paper maintains the performance of mAP@0.5 under low indicators. Compared with the original YOLOv5, mAP@0.75 has been improved by 5.82%, mAP@0.5:0.95 has been improved by 2.74%, and detection speed has been improved by 30.7%.

In addition, the YOLOv3 network is improved in the experiment. By modifying the Neck part of the YOLOv3 network, the FPN module is improved with PANet module, which makes the feature fusion more sufficient, and the detection accuracy is improved to some extent. However, compared with YOLOv5 network, although the overall idea is the same, the detection accuracy of YOLOv3 network is not as good as that of YOLOv5 because the backbone features are not sufficiently extracted.

EfficientDet is also an outstanding network in the current target detection field. In the experiment, efficientdet-d0 and efficientdet-d4 networks were trained and tested respectively. The difference between d0 and d4 network sizes lies in the use of 3-layer and 12-layer BiFPN feature fusion layers at the neck. The detection results show that Efficientdet-d4 does better than Efficientdet-d0 performance, but there is a big difference compared with YOLOv5 detection performance.

Furthermore, the single-stage detection network model is obviously superior to the two-stage network model in the forward-looking sonar image target detection task. Through experimental tests, the detection accuracy of mAP@0.5:0.95 of YOLOv5 is 36.74% higher than that of Faster R-CNN, and the detection speed is 27 times faster than that of Faster R-CNN. This is because in the detection process, YOLOv5 directly performs regression calculation of target location, without extracting candidate regions. However, Faster R-CNN and other two-stage detection networks need to extract candidate regions and then predict these candidate regions. Therefore, the detection efficiency is obviously inferior to that of single-stage detection network.

Finally, the performance of several YOLO networks with different depths and widths is compared. Although YOLOv5s network has the smallest scale, it still performs better than deeper and wider networks. The main reason is that the deep network does not highlight its superior performance due to the limited amount of sonar data.

The above experimental data show that the CoordConv-YOLOv5 network proposed in this paper has better performance in the target detection task of the forward-looking sonar image, which provides certain help for the subsequent research on the target detection of forward-looking sonar image based on deep learning.

V. CONCLUSION

In this paper, a forward sonar image target detection model based on transfer learning CoordConv-YOLOv5 is proposed for the task of forward sonar image target detection.

First, IoU k-means algorithm is proposed to recalculate the initial detection frame of YOLOv5 network for the difference between optical image target and sonar image target. Second, the detection model of Coordconv-YOLOv5 is established on the basis of the target characteristics of the forward-looking sonar image. Finally, the idea of transfer learning is used to fine-tune the pre-training model and a detection model suitable for the forward-looking sonar image target detection is obtained.

In the experiment, the convolutional network models YOLOv3, YOLOv3-PAN, YOLOv4, Faster R-CNN, EfficientDet and YOLOv5 were used for training and testing. According to the comparison of detection performance, the CoordConv-YOLOv5 network based on transfer learning proposed in this paper shows the best performance in both detection accuracy and detection speed. Detection accuracy mAP@0.5:0.95 can reach 56.95%, and detection speed can reach 9ms.

However, according to the current research results, due to the complex underwater environment, the background speckle noise of the forward-looking sonar image is serious. Secondly, the echo intensity of the target is weak, and the attitude of the target is unstable underwater. As a result, there are some false alarms in the detection results of the current algorithm. To solve this problem, the future research will focus on the preprocessing of the forward-looking sonar image.

REFERENCES

[1] C. Zhe, H. Wang, S. Jie, and D. Xin, "Underwater object detection by combining the spectral residual and three-frame algorithm," *Lect. Notes Electr. Eng.*, vol. 279, pp. 1109–1114, Jan. 2014.

[2] S. A. Villar, G. G. Acosta, and F. J. Solari, "OS-CFAR process in 2-D for object segmentation from sidescan sonar data," in *Proc. 16th Workshop Inf. Process. Control (RPIC)*, Oct. 2015, pp. 1–6.

[3] K. Mukherjee, S. Gupta, A. Ray, and S. Phoha, "Symbolic analysis of sonar data for underwater target detection," *IEEE J. Ocean. Eng.*, vol. 36, no. 2, pp. 219–230, Apr. 2011.

[4] O. Midtgaard, R. E. Hansen, T. O. Saebo, V. Myers, J. R. Dubberley, and I. Quidu, "Change detection using synthetic aperture sonar: Preliminary results from the larvik trial," in *Proc. OCEANS MTS/IEEE KONA*, Sep. 2011, pp. 1–8.

[5] X. Yan, L. Jianlong, and H. Zhiguang, "Measurement of the echo reduction for underwater acoustic passive materials by using the time reversal technique," *Chin. J. Acoust.*, vol. 40, no. 1, pp. 110–116, Jan. 2016.

[6] D. S. Raghuvanshi, I. Dutta, and R. J. Vaidya, "Design and analysis of a novel sonar-based obstacle-avoidance system for the visually impaired and unmanned systems," in *Proc. Int. Conf. Embedded Syst. (ICES)*, Jul. 2014, pp. 238–243.

[7] Z. Fan, W. Xia, X. Liu, and H. Li, "Detection and segmentation of underwater objects from forward-looking sonar based on a modified mask RCNN," *Signal, Image Video Process.*, vol. 15, no. 6, pp. 1135–1143, Jan. 2021.

[8] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.

[9] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, Oct. 2017, pp. 2961–2969.

[10] X. Wang, Z. Guan, J. Wang, and Y. Wang, "Target detection of color image sonar based on convolutional neural network," *J. Comput. Appl.*, vol. 39, no. S1, pp. 187–191, Jul. 2019.

[11] Z. SHENG and G. HUO, "Detection of underwater mine target in sidescan sonar image based on sample simulation and transfer learning," *CAAI Trans. Intell. Syst.*, vol. 16, no. 2, pp. 385–392, Mar. 2021, doi: 10.11992/tis.202101030.

[12] L. Jin, H. Liang, and C. Yang, "Sonar image recognition of underwater target based on convolutional neural network," *Xibei Gongye Daxue Xuebao/J. Northwestern Polytechnical Univ.*, vol. 39, no. 2, pp. 285–291, Apr. 2021.

[13] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[14] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 34, no. 7, Feb. 2020, pp. 12993–13000.

[15] G. Hamerly and C. Elkan, "Learning the k in k-means," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 16, Mar. 2004, pp. 281–288.

[16] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and J. Yosinski, "An intriguing failing of convolutional neural networks and the CoordConv solution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2018, pp. 9628–9639.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2016.

[18] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with attention-RPN and multi-relation detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4013–4022.

[19] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Netw.*, vol. 107, pp. 3–11, Nov. 2018.

[20] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2874–2883.

[21] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787.

[22] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[24] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[25] C. Fachkha and M. Debbabi, "Darknet as a source of cyber intelligence: Survey, taxonomy, and characterization," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1197–1227, 2nd Quart., 2016.

[26] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[27] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[28] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[29] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[30] K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas, "Imbalance problems in object detection: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3388–3415, Oct. 2021.

[31] W. Yanchen, "Sonar image target detection and recognition based on convolution neural network," *Mobile Inf. Syst.*, vol. 2021, no. 6, pp. 1–8, Mar. 2021.

[32] Y. Yu, J. Zhao, Q. Gong, C. Huang, G. Zheng, and J. Ma, "Real-time underwater maritime object detection in side-scan sonar images based on transformer-YOLOv5," *Remote Sens.*, vol. 13, no. 18, pp. 2–28, Sep. 2021.

[33] W. Li, J. Wang, X. F. Zhao, Z. Wang, and Q. J. Zhang, "Target detection in color sonar image based on YOLOV5 network," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Aug. 2021, pp. 1–5.

[34] J. Zhou, M. Yan, C. Luo, and X. Xing, "Underwater sonar target detection based on YOLOv5," in *Proc. Int. Conf. Electron. Inf. Eng. Comput. Sci. (EIECS)*, Sep. 2021, pp. 729–732.

HAOTING ZHANG was born in 1998. He received the B.S. degree from PLA Information Engineering University, in 2020, where he is currently pursuing the master's degree. His research interest includes sonar image processing.

MEI TIAN received the M.S. degree from Zhengzhou University, in 2008. Her research interest includes communication signal processing.

GAOPING SHAO received the Ph.D. degree from the Beijing Institute of Technology, in 2009. His research interest includes communication signal processing.

JUAN CHENG received the Ph.D. degree from PLA Information Engineering University, in 2009. Her research interest includes image processing.

JINGJING LIU received the Ph.D. degree from Southeast University, in 2020. Her research interest includes image processing.