IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 3, 2020, accepted September 12, 2020, date of publication September 18, 2020,
date of current version October 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3024813

Shipwreck Target Recognition in Side-Scan Sonar Images by Improved YOLOv3 Model Based on Transfer Learning

TANG YULIN, (Member, IEEE), SHAOHUA JIN , (Member, IEEE), GANG BIAN, AND YONGHOU ZHANG

Department of Hydrography and Cartography, Dalian Naval Academy, Dalian 116018, China

Corresponding author: Shaohua Jin (jsh_1978@163.com)

This work was supported by the National Science Foundation of China under Grant 41876103 and Grant 41576105.

ABSTRACT When used to recognize side-scan sonar images of shipwreck targets, the Faster R-CNN model is time-consuming and has low efficiency and a high missed detection rate for small targets. Considering that existing datasets of side-scan sonar images of shipwreck targets are small, we propose a YOLOv3 model that can automatically recognize side-scan sonar images of shipwreck targets based on transfer learning. Based on the Darknet-53 network, we froze part of the convolutional layer of the YOLOv3 model trained on COCO dataset images, and conducted transfer learning. Multi-scale training of shallow feature fusion was done based on multi-scale feature fusion with Feature Pyramid Networks (FPN) support, and the proportion of recognized shallow features of shipwreck targets increased. Meanwhile, the parameters and sizes of target anchor boxes were reset using K-means clustering, which allowed us to improve the speed of target recognition and the precision of smaller target recognition and positioning. Lastly, the binary classification cross entropy function was used to improve the loss function of the YOLOv3 algorithm. Experimental results show that under the same recognition target, the average precision (AP) value of the YOLOv3 model based on transfer learning reached 89.49%, which is an improvement of 0.31% and 1.77%, respectively, compared with the Faster R-CNN model and the traditional YOLOv3 model. Moreover, the YOLOv3 model based on transfer learning had the highest harmonic mean (F1), reaching 90.71%, which is 3.96% and 1.63% higher, respectively, than the harmonic means of the Faster R-CNN and the traditional YOLOv3 model. Lastly, the traditional YOLOv3 model takes an average 0.17s to identify a target. In contrast, the Faster R-CNN model takes an average of 2.8s to identify a target. Hence, our transfer learning YOLOv3 model greatly improves detection efficiency, meets the needs of real-time target recognition, and ultimately has better overall performance than existing methods.

INDEX TERMS Side-scan sonar shipwreck target, YOLOv3 model, transfer learning, shallow feature fusion, K-means clustering algorithm, binary classification cross entropy.

I. INTRODUCTION

In recent years, with the rapid development of the marine economy, human marine use and development activities have become more frequent, and the incidence of marine accidents has increased year by year. The importance of maritime safety has become more and more prominent. Among them, the use of side-scan sonar to search for wrecked ships is an important part of marine obstacle inspection and maritime search and rescue [1]–[4]. To address the issues with traditional manual interpretation, such as low efficiency, great time and resource consumption, strong subjective uncertainty, and excessive dependence on experiences, scholars at home and abroad have done a lot of research on side-scan sonar image classification and recognition and target detection [5]–[13], Suraj Kamal proposed a deep learning framework for underwater target recognition based on the DBN structure, which achieved 90.23% accuracy in 40 categories of classification problems [14]; Jason Rhinelande proposed a method for target recognition and classification of side-scan sonar images based on support vector machines [15]; Guo Jun proposed a side- scan sonar image classification research based on SVM

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

algorithm and GLCM [16]; Chen Qiang uses a simple BP neural network to classify and recognize underwater image targets, manually select features and send them to the neural network for classification training, the accuracy rate is 80% [17]; Although these methods have been automated to a large extent, there are problems in the design of feature extraction algorithms and the lack of generalization ability. It is difficult to grasp the validity and comprehensiveness of the extracted feature parameters. With the wide application of convolutional neural networks in the field of image recognition and target detection [18]–[20], TANG Yulin and others proposed the side-scan sonar image shipwreck target recognition method by using the Faster R-CNN model. This method maintains high recognition precision, but its RPN takes too much time and decreases the processing speed; therefore, it cannot meet the real-time requirements for maritime search and rescue. Meanwhile, the shipwreck target in a side-scan sonar image is generally small and occupies little of the image; thus, it is considered a small-scale target. Since the Faster R-CNN model performs regression forecasting in the deep feature map of the convolutional network, it loses some position information while obtaining rich semantic information. Moreover, its recognition of smaller targets is unsatisfactory due to its high missed detection rate [21]–[24]. In 2016, Joseph Redmon and others proposed the YOLO network [25]. This network uses an end-to-end training method instead of area selection. It directly divides an image into a fixed number of smaller areas and identifies each smaller area. A category confidence level is obtained, and finally the areas with confidence levels higher than the threshold are combined to identify the category and location of the target. Because there is no training in stages, the YOLO network is faster than the Faster R-CNN. Thus, the YOLO network meets real-time requirements. However, the detection result for smaller targets is not satisfactory because the image is directly divided into a fixed number of areas. YOLOv2 [26], [27], proposed in the same year, added batch normalization (BN), abandoned dropout, and hierarchical clustering to the boundary prediction. These improvements make the preliminary target boundary prediction more accurate. In target classification, a convolutional network is used instead of a fully connected layer to further reduce the number of parameters, thereby making target detection faster and more accurate. Expanding on this, the YOLOv3 model was built in 2018 by adding multi-scale prediction and a better basic classification network, i.e., Darknet-53. This model has faster speed and better precision in small target detection [28], and meets the real-time requirements of side-scan sonar shipwreck target recognition and small target detection.

Although convolutional neural networks have been widely used in various fields, their performance can only be demonstrated when the network structure is relatively complex and the number of training samples is sufficient. Convolutional neural networks often have millions of parameters. Therefore, many labeled samples are needed to train convolutional neural networks. However, there are generally few side-scan sonar sample images. During training, phenomena including overfitting, the personal best solution, and poor generalization ability are common. The pre-trained convolutional neural network model is often used to fix the above issues in transfer learning. Transfer learning applies the structure and parameters of the trained model to a model with similar issues, and then obtains a model that has already fixed the issues through retraining [29]–[36].

Based on this, the present paper intends to introduce the YOLOv3 model into the side-scan sonar shipwreck target detection and proposes an improved YOLOv3 model based on transfer learning to improve the detection performance. First, according to the characteristics of the side-scan sonar shipwreck data set, the traditional YOLOv3 model is improved. At the same time, in view of the problem of too few samples in the side-scan sonar shipwreck data set, a transfer learning method is proposed to train and optimize the network model. Specifically reflected in (1) Based on the original Feature Pyramid Networks-supported (FPN-supported) multi-scale feature fusion training, shallow features such as contours, textures, and grayscales learned by 4x and 2x downsampling are fused to enrich the image information used in algorithm learning. This allows us to improve the precision of the model to recognize and locate the smaller scale targets. (2) At the same time, the K-means clustering algorithm is used to reset the parameters and size of the anchor box to generate an anchor box that is more suitable for the features of the wreck dataset. Therefore, a better intersection over union (IOU) can be obtained between the predicted value and the real value, thereby further improving the precision of recognition and positioning of shipwreck targets. (3) Besides, the binary classification cross entropy function is used to improve the loss function of the YOLOv3 algorithm so that the model achieves a better convergence effect. (4) Lastly, freeze the partial convolutional layer of the YOLOv3 model based on training with the COCO dataset, we achieve transfer learning of a smaller sample side-scan sonar shipwreck dataset and the overall performance of the model is improved.

II. IMPROVED YOLOv3 MODEL AND METHODS

A. BASIC STRUCTURE OF YOLOv3 MODEL

Compared with the YOLO model, the main improvement of the YOLOv3 network is the adjustment of the network structure, the new backbone network Darknet-53, and the FPN-supported multi-scale features built for detection. The specific model structure is shown in Figure 1. The YOLOv3 model divides the input image into an $S \times S$ grid, and uses the Darknet-53 basic network deepened by the residual network to perform full-convolution feature extraction while using FPN-like upsampling and the fusion method to perform detection at multiple scales. Therefore, the new network improves the precision of detection while maintaining a speed advantage. Most of all, the detection of smaller targets is enhanced.

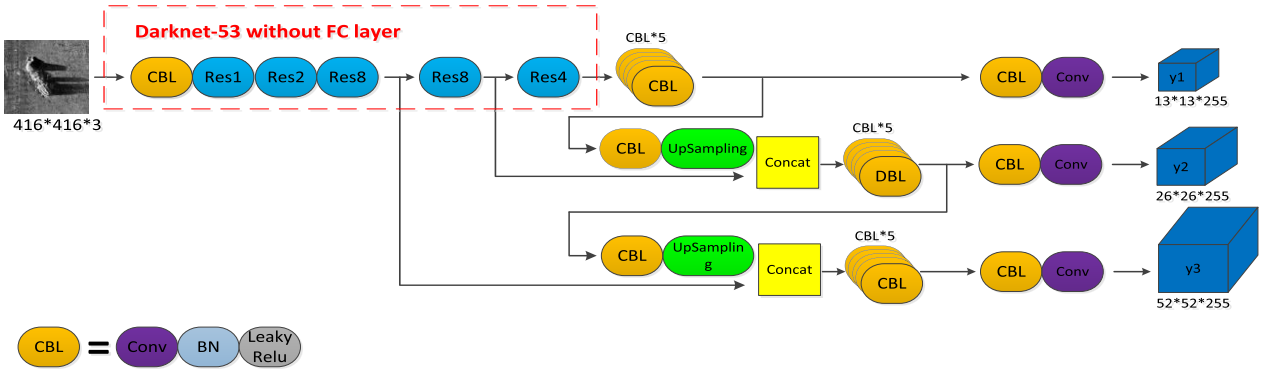


FIGURE 1. YOLOv3 Model structure diagram.

	Type	Filters	Size	Output
	Conv	32	3×3	416×416
	Conv	64	3×3/2	208×208
1×	Conv	32	1×1	
	Conv	64	3×3	
	Residual			208×208
	Conv	128	3×3/2	104×104
2×	Conv	64	1×1	
	Conv	128	3×3	
	Residual			104×104
	Conv	256	3×3/2	52×52
8×	Conv	128	1×1	
	Conv	256	3×3	
	Residual			52×52
	Conv	512	3×3/2	26×26
8×	Conv	256	1×1	
	Conv	512	3×3	
	Residual			26×26
	Conv	1024	3×3/2	13×13
4×	Conv	512	1×1	
	Conv	1024	3×3	
	Residual			13×13

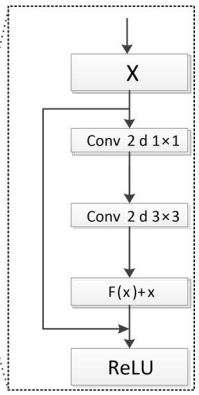


FIGURE 2. Darknet-53 structure.

YOLOv3 uses the Darknet-53 network structure for image feature extraction. As shown in Figure 2, this network is mainly composed of 53 convolutional layers, 1×1 and 3×3, which are located at the front of the residual layer [37]. Each convolutional layer is followed by a BN layer [38] and a LeakyReLU layer, which make up the DBL. As shown in Figure 1, this is the basic component of the YOLOv3 network structure. The YOLOv3 model adds a skip connection layer and an upsampling layer based on the Darknet-53 network, resulting in a total of 75 convolutional layers.

To fix the issue of gradient divergence caused by deepening the network model, the Darknet-53 network learns from the residual network. The digits 1, 2, 4, and 8 in the leftmost column indicate the numbers of repeated residual components. A schematic diagram of the Darknet-53 network is shown in Figure 2. The residual connection includes two convolutional layers. The first convolutional layer has a convolution kernel size of 1×1, and the second convolutional layer has a convolution kernel size of 3×3. The network adds output x to output f(x), and uses the ReLU activation function as the output of the final model. By directly passing input x to the output, the output result is $f(x) = x$; when $f(x) = 0$, then $H(x) = x$, the residual result is close to 0, and the training model converges. This structure can ensure that the network can still converge when there are numerous layers and that the model can be trained. The deeper the network gets, the better the expression of the features; moreover, the recognition precision will not decrease. At the same time, the 1×1 convolution in the residual network reduces the channel of each convolution, thereby greatly reducing the number of parameters. Furthermore, the amount of calculation is decreased to a certain extent, and the convergence of the model is accelerated.

B. METHOD OF IMPROVING YOLOv3 MODEL

1) MULTI-SCALE TRAINING OF SHALLOW FEATURE FUSION

In order to increase the ability to recognize smaller targets while ensuring acceptable detection speed, the YOLOv3 network can learn the deep and shallow features at the same time. This is inspired by the FPN [39], and the deep features are extracted through upsampling. The YOLOv3 network's dimensions are the same as those of the feature layer to be fused. The feature maps of different scales are fused and then predicted so that the model has fine-grained features and the ability to recognize smaller targets increases. As shown in Figure 1, the model is tested at 32X downsampling, 16X downsampling, and 8X downsampling. The 32X downsampling results in a prediction of $13 \times 13 \times 255$. The receptive field is large and suitable for detecting large-size objects. $13 \times 13 \times 255$ features are obtained by upsampling. A series of 3×3 and 1×1 convolution operations yields $26 \times 26 \times 255$ features with a medium-scale receptive field, which is suitable for detecting medium-scale objects. The process is similar for $52 \times 52 \times 255$ features with the smallest receptive field, which is suitable for detecting small objects.

Although high-power downsampling can obtain a higher receptive field, result in additional and deeper semantic features, and achieve a good classification effect, a higher receptive field or a larger downsampling factor loses certain position information, resulting in reduced positioning precision. Since the shallow scan features of side-scan sonar

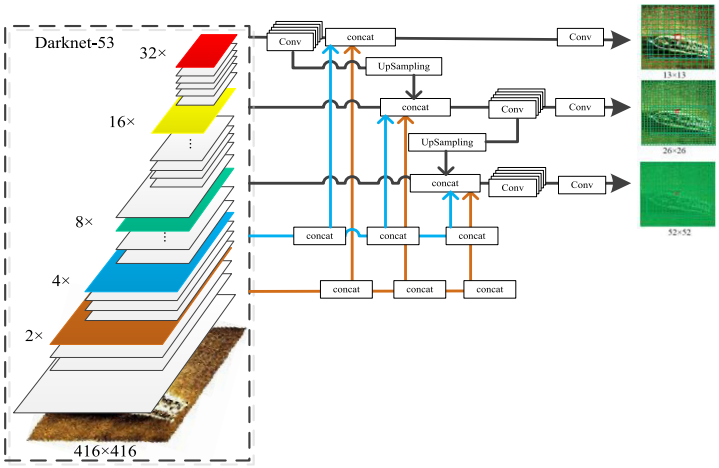


FIGURE 3. Improved multi-scale feature fusion structure.

images for shipwreck targets (such as profiles, grayscale gradient changes, and shadow distributions) play an important role in target recognition, they should take a greater weight. Although the traditional YOLOv3 model performs 32X downsampling, 16X downsampling, and 8X downsampling, and uses multi-scale feature fusion to merge shallow features with deep semantic features, 8X downsampling does not comprehensively learn shallow features. Hence, this paper conducted multi-scale training of shallow feature fusion. Specifically, as shown in Figure 3, the features learned by 4x downsampling and 2x downsampling were merged with the traditional three-scale features to learn the shallow features. Outline texture and other information of a layered wreck were merged with deep semantic abstract features to increase the proportions of certain features so that the image had more abundant information. Through multi-scale fusion training of shallow features, both the detection efficiency and the learning of shallow and deep features could be ensured, the degree of nonlinearity improved, the generalization ability increased, and the precision of network recognition and positioning of sub-scale targets improved.

2) K-MEANS CLUSTERING BOUNDARY PREDICTION STRATEGY

YOLOv3 uses the anchor box generated by the K-means clustering algorithm in YOLOv2, divides the current feature layer into S×S areas, and predicts three potential anchor boxes for each area through the K-means clustering method. The dashed rectangular box in Figure 4 is the anchor box, and the solid rectangular box is the predicted boundary box calculated by the network-predicted offset. (C_x, C_y) is the central coordinate of the anchor box on the feature map, (P_x, P_y) is the width and height of the anchor box on the feature map, (t_x, t_y) is the central offset of the boundary box predicted by the network, (t_w, t_h) is the width–height zoom ratio, and (b_x, b_y, b_w, b_h) is the final predicted target

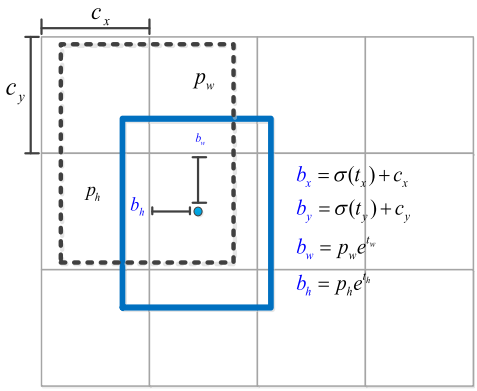


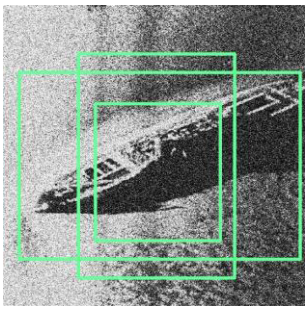
FIGURE 4. Schematic diagram of anchor box of YOLOv3 model.

boundary box. The conversion process from the anchor box to the final prediction boundary box is as follows:

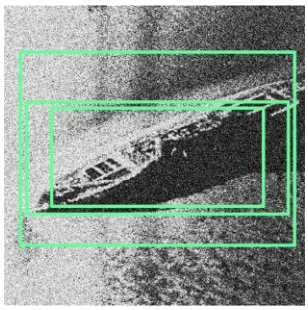
$$b_x = \sigma(t_x) + c_x \tag{1}$$
$$b_y = \sigma(t_y) + c_y \tag{2}$$
$$b_w = p_w e^{t_w} \tag{3}$$
$$b_h = p_h e^{t_h} \tag{4}$$

Among them, $P_w = \frac{w_{anchor}}{w_{image}}$, $P_h = \frac{h_{anchor}}{h_{image}}$, $t_w = \ln \frac{w_{pred}}{w_{anchor}}$, and $t_h = \ln \frac{h_{pred}}{h_{anchor}}$ functions are sigmoid functions with the purpose to scale the prediction offset to between 0 and 1 so that the central coordinates of an anchor box can be fixed in a cell to speed up network convergence.

The anchor box is obtained by K-means algorithm clustering. The clustering central digit is set to 9, the initial position is randomized, the distance between each callout box and the clustering central point in the label information is calculated, and the callout box is assigned to the nearest clustering central point. Having been assigned callout boxes, the clustering central point is recalculated until the clustering



(a) Original Anchor Boxes



(b) Anchor Boxes After Reclustering

FIGURE 5. Scope ofanchor boxes on a shipwrecktarget.

center does not change. The traditional K-means algorithm uses Euclidean distance as the similarity measure. However, in the detection algorithm, a reasonable presetting anchor box is required for us to obtain a better intersection over union (IOU) between the predicted value and the real value. Therefore, YOLOv3 uses the K-means algorithm that takes the IOU as the distance metric. The distance formulas are shown in equations (5) and (6).

$$d(b, o) = 1 - IOU(b, o) \tag{5}$$

$$IOU(b_{pt}, b_{gt}) = \frac{b_{pt} \cap b_{gt}}{b_{pt} \cup b_{gt}} \tag{6}$$

Here, $d(b, o)$ is the distance between the anchor box b and the clustering center o, b_{pt} is the anchor box, and b_{gt} is the actual box.

According to the network detection mechanism, the size of the anchor box has a direct impact on the precision of recognition. The nine preset anchor boxes in the original YOLOv3 are obtained by clustering the COCO dataset, which contains 80 evenly distributed types, scale shapes, and sizes. However, the side-scan sonar images of shipwreck targets in this experimental dataset are mostly flat and vertical. Therefore, continuing to use the anchor box of the COCO dataset will be unfavorable for the recognition of shipwreck targets. In this paper, the K-means clustering algorithm is used to recluster the wreck dataset, and the average results obtained through five sets of clustering are ((75, 55), (85, 30), (116, 76)); ((46, 24), (52, 17), (57, 25)); ((22, 13), (31, 12), (34, 41)). As shown in Figure 5, the original anchor boxes cannot adapt well to the side-scan sonar image of a submarine shipwreck target, but the reclustered anchor boxes are more in line with the shape features of the shipwreck target.

3) IMPROVEMENT ON LOSS FUNCTION

In view of the problem of few samples and large noise in datasets of wreck side-scan sonar images, it is difficult to choose a suitable initial learning rate during model training. If the learning rate is too small, the convergence speed will be very slow. If it is too large, the loss value will continue to oscillate or even deviate from the minimum value, and the same learning rate cannot be applied to the learning of each parameter. To allow the model to learn more elaborate image features and obtain the optimal parameter values, this paper uses the adaptive learning rate Adam algorithm combined with the Momentum and RMSProp algorithms [40]. The Adam algorithm comprehensively considers the first moment estimation (the mean of the gradient) and the second moment estimation (the non-central variance of the gradient), and calculates the update steps. Because the update of model parameters is not affected by the expansion and contraction of the gradient, it can process noise samples better and automatically adjust the learning rate. Hence, it provides stronger robustness to the parameters, makes the model achieve better convergence, and effectively prevents overfitting. The loss function of the model is shown in formula (7).

$$\begin{aligned} Loss &= \sum_{i=0}^{s^2}\sum_{j=0}^{B} l_{ij}^{obj}[(\sigma(t_x)_i^j - \sigma(\hat{t}_x)_i^j)^2 + (\sigma(t_y)_i^j - \sigma(\hat{t}_y)_i^j)^2] \\ &+ \sum_{i=0}^{s^2}\sum_{j=0}^{B} l_{ij}^{obj}[(\sigma(t_w)_i^j - \sigma(\hat{t}_w)_i^j)^2 + (\sigma(t_h)_i^j - \sigma(\hat{t}_h)_i^j)^2] \\ &+ \sum_{i=0}^{s^2}\sum_{j=0}^{B} G_{ij} L(C_i^j, \hat{C}_i^j) \\ &+ \sum_{i=0}^{s^2}\sum_{j=0}^{B}\sum_{C\in classes} l_{ij}^{obj} L(p_i^j(c) - \hat{p}_i^j(c)) \end{aligned} \tag{7}$$

When the prediction box predicts an object $l_{ij}^{obj} = 1$, $\hat{C}_i^j = 1$; when the prediction box predicts an object $l_{ij}^{obj} = 0$, $\hat{C}_i^j = 0$. When the prediction box is not responsible for predicting an object and the IOU with the actual box is greater than the set threshold (in this paper, IOU=0.5), $G_{ij} = 0$; otherwise, $G_{ij} = 1$. x, y, w, and h are the central coordinates as well as the width and height of the prediction box, respectively. S is the number of grids into which the feature map is

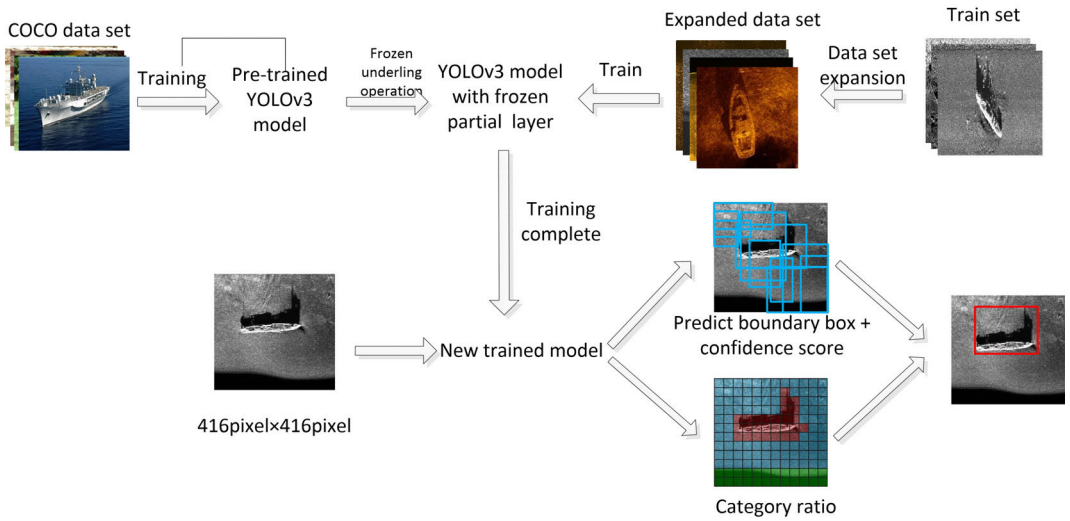


FIGURE 6. Flow chart describing transfer learning.

divided, B is the number of prediction boxes for each grid, C is the confidence of the prediction, and p is the probability of the category. The central coordinates, length, and width of the prediction box are the mean square deviation, and the error is calculated using the sigmoid function σ. Since the mean square deviation of the partial derivative of the parameter is multiplied by the derivative of sigmoid σ', when its variable value is large or small, σ' approaches 0. The gradient update amplitude is small, the parameter update speed is slow, and the convergence time is long. Therefore, the confidence and category error in this paper are calculated by using the binary classification cross entropy function (formula (8)), and through this we achieve a better convergence effect.

$$L = \frac{1}{N}\sum_{i=1}^{N}[(-x\log\hat{x}) - (1-x)\log(1-\hat{x})] \qquad (8)$$

C. TRANSFER LEARNING

This paper adopts the transfer learning strategy to train the network model. Retraining a complicated convolutional neural network requires massive data resources, a large amount of computing resources, and time resources. Considering that all tasks are correlative, the knowledge obtained in previous tasks can be directly applied to new tasks with minor transformations or even without any change. When it is difficult to obtain this knowledge using a small amount of data in new tasks, transfer learning can share the learned model parameters to the new model, thereby speeding up and optimizing the learning efficiency of the new model. This method reduces repetitive labor and reliance on target task training data, and improves the model's performance.

As the depth of the convolution layer increases, the convolutional neural network will learn a deeper abstract specific target feature. The texture, profile, and color of a shallow layer are universal shallow features obtained through learning and have high mobility. The convolutional layer with multi-scale feature fusion is a deep convolutional layer, and the extracted image features are more abstract and have lower mobility than the aforementioned shallow features. Considering these points, this paper freezes the weight parameters of the convolutional layer before multi-scale feature fusion, and initializes and retrains the 59th, 67th, and 75th convolutional layers, fully connected layers, and sigmoid output layer on the target dataset. A flow chart of this process is shown in Figure 6.

III. EXPERIMENT & ANALYSIS

The original experimental data consisted of 1,000 images provided by marine surveying departments and side-scan sonar manufacturers, as well as website screenshots. The images were labeled by using the open source application LabelImg. This paper first expands the original dataset by means of flip transformation, random trimming, color vibrancy, translation transformation, scale transformation, contrast transformation, noise perturbation, rotation transformation, etc., and then normalizes the pixels of the entire dataset. The images are fixed to 416×416 pixels. After preprocessing, the dataset contains 5000 images: 4000 are randomly selected to divide the training set and validation set at a ratio of 4:1, and perform 5-fold cross-validation, and 1000 are selected for the test set by balanced sampling. Furthermore, the small batch gradient descent method [41]–[45] is adopted: all images are divided into 88 batches for input model training, and each batch inputs 64 images for model training (i.e., the batch size is 64) with a total of 1,000 steps (epochs).

Experimental training and testing were conducted using Python programming based on TensorFlow. The experimental environment is Linux Ubuntu OS version 18.04, running on an Intel (R) Xeon (R) CPU E5-2678 v3@2.50GHz, with an NVIDIA TITAN RTX GPU and 24GB memory.

TABLE 1. 5-fold cross-validation results of models.

Model	1-fold	2-fold	3-fold	4-fold	5-fold	Mean
Traditional YOLOv3	89.21%	89.18%	89.26%	89.22%	89.24%	89.22%
Transfer learningYOLOv3	89.51%	89.50%	89.55%	89.47%	89.49%	89.50%

A. RESULTS EVALUATION

The evaluation criteria used in this experiment are average precision (AP) and harmonic mean (F1). AP is an indicator that reflects the performance of the entire model. It is the area of the precision–recall (P–R) curve, and represents the average precision. Precision (also called the precision rate) indicates how many targets are detected. Recall (also called recall rate) indicates how many targets are accurately detected, and thus measures the completeness of target detection. These two metrics are expressed in equations (9) and (10), respectively.

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

The detected samples can be divided into four categories according to the classification results: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). $TP + FP$ is the total number of correctly classified samples, and $TP + FN$ is the total number samples. The definition of AP is presented in equation (11).

$$AP = \int_0^1 P(R)dR \tag{11}$$

In order to better tune the model and find the parameter values that make the model's generalization performance the best, this experiment uses a 5-fold cross-validation method. The 4000 sample data used for training were randomly divided into 5 parts, and the model training and verification were performed 5 times. Each time 4 parts are selected as the training set and 1 part is used as the validation set. The AP values verified after each training are shown in Table 1.

It can be seen from Table 1 that the AP value of each fold training of transfer learning YOLOv3 model is higher than that of the traditional YOLOv3 model, and the mean of AP value of the 5-fold cross-validation is 89.50%, which is 0.28% higher than the traditional YOLOv3 model. It proves that the transfer learning YOLOv3 model has better robustness. Since this experiment is an experiment to find the optimal solution, and both the transfer learning and the traditional YOLOv3 model have the highest AP value after the 3-fold training, the model after the 3-fold training is selected for further analysis, including the change of the loss value during the training process and the performance evaluation of the model on the test set.

The loss values of the two YOLOv3 models (the traditional model and the model with transfer learning) are shown in Figure 7. The loss values of the two models decreases with

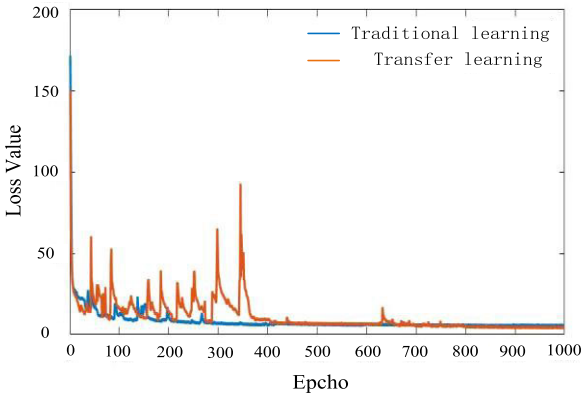


FIGURE 7. Loss values of two YOLOv3 models.

the increase in the number of epochs, but eventually becomes stable. The traditional YOLOv3 model tends to be stable after 600 epochs, and the loss value is finally maintained at around 5.5. The transfer learning YOLOv3 model extracts the parameters of partial shallow features because it learns from the model trained on our COCO-based dataset. The initial loss value is low and declines quickly. It is important to remember that our side-scan sonar wreck dataset is different from the COCO dataset: the learning of abstractive features uses a large number of parameters on the COCO dataset, and therefore the loss value fluctuates greatly in the first 350 steps of training. However, because the model can obtain the position information of the target well, and because it uses the binary classification cross entropy to calculate the error, its loss value tends to stabilize and converge to about 4.3 after 750 steps of training. This final loss value is lower than the loss value of the traditional YOLOv3 model, proving that the YOLOv3 model based on transfer learning has better generalization.

P–R curves of the Faster R-CNN, traditional YOLOv3 model, and transfer learning YOLOv3 model in test set are shown in Figures 8(a)–(c), respectively. The larger the area of the curve and the coordinate, the better the model's effect. The AP values of the three models are 87.72%, 89.18%, and 89.49%, respectively. The AP of the traditional YOLOv3 model is significantly greater than that of the Faster R-CNN model, while the average precision of the YOLOv3 model based on transfer learning is 0.31% more than that of the traditional YOLOv3 model.

When the recall rate reaches 85%, the precision rate of the Faster R-CNN reaches 88%. If the recall rate is further improved, the precision rate drops significantly. The downward trend in precision rate of the traditional YOLOv3 model is slower, and the model can maintain a high recall rate

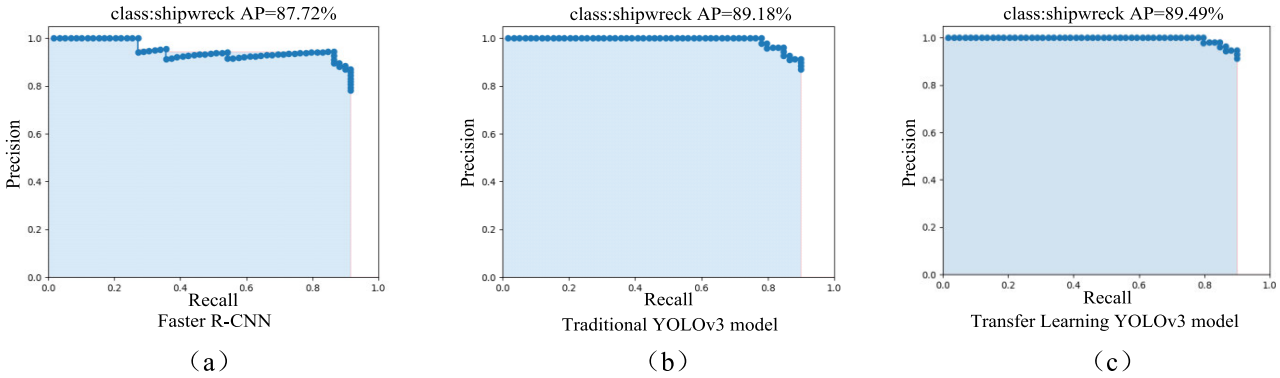


FIGURE 8. P–R Curves of the three tested models.

TABLE 2. Comparison of test results of the three models.

Model	Precision	Recall	AP	F1	Time
Faster R-CNN	89.26%	84.38%	87.72%	86.75%	2.8s
Traditional YOLOv3	86.89%	91.38%	89.18%	89.08%	0.17s
Transfer learningYOLOv3	89.78%	91.66%	89.49%	90.71%	0.17s

and high precision rate: the model has a precision rate of 89% when the recall rate reaches 90%. In contrast, the YOLOv3 model based on transfer learning has an even slower decline in the PR curve, and the area of the curve and the coordinate axis is larger; furthermore, the precision rate is 91% when the recall rate is 90%. These results prove that the YOLOv3 model based on transfer learning can best recognize side-scan sonar shipwreck targets.

F1 is the harmonic mean of the precision rate and recall rate. The F1 value shown in formula (12) is used to represent the overall performance of the algorithm. This paper sets both the confidence interval and IOU to 0.5. The test results of the three models are shown in Table 1.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{12}$$

It can be seen from Table 2 that although the recognition precision rate of the Faster R-CNN model is 2.37% higher than that of the traditional YOLOv3 model, its recall rate is 6% lower. Therefore, the Faster R-CNN is not as good as the YOLOv3 model in detecting smaller targets. Moreover, the F1 value of the Faster R-CNN is 2.33% lower. Considering the overall performance of the model, the traditional YOLOv3 model is superior to the Faster R-CNN model. Meanwhile, the precision rate, recall rate, AP value, and F1 of the YOLOv3 model based on transfer learning are all higher than those of the other two models. Its AP value is 1.77% and 0.31% higher than that of the Faster R-CNN model and the traditional YOLOv3 model, respectively. Moreover, F1 values are increased by 1.63% and 3.96%, respectively. The detection speed is also an important indicator of the overall performance of the model. It can be seen from the table that the Faster R-CNN takes 2.8 s to detect a target, while the traditional YOLOv3 model takes only 0.17 s. The detection efficiency is already greatly improved by using the traditional YOLOv3, and is only further enhanced when the YOLOv3 model based on transfer learning is employed.

B. RESULT ANALYSIS

Figure 9 compares the shipwreck targets detection results of the three models using side-scan sonar images. Figure 9(a) illustrates the different sizes of shipwreck targets and the diversity of scales. The Faster R-CNN model can recognize large-scale shipwreck targets well, but the recognition of smaller shipwreck targets is not satisfactory due to its high missed detection rate. For smaller targets, the traditional YOLOv3 model performs better than the Faster R-CNN model, but not perfectly. The red boxes in the figure indicate the missed targets. The transfer learning YOLOv3 model sees further improved performance in recognizing smaller targets. However, when the shipwreck targets are closely arranged, the positioning precision is decreased to a certain extent, and the two shipwreck targets are mistakenly detected as one target. In general, however, the YOLOv3 model can better recognize and distinguish smaller targets, and its missed detection rate is greatly reduced. As can be seen from Figures 9(b) and (c), the IOU of the detection frame and actual frame of the YOLOv3 model based on transfer learning is largest, and the positioning is most accurate. The IOUs in Figure 9(b) are 69.92%, 75.93%, and 86.09%, for the Faster R-CNN, traditional YOLOv3 model,

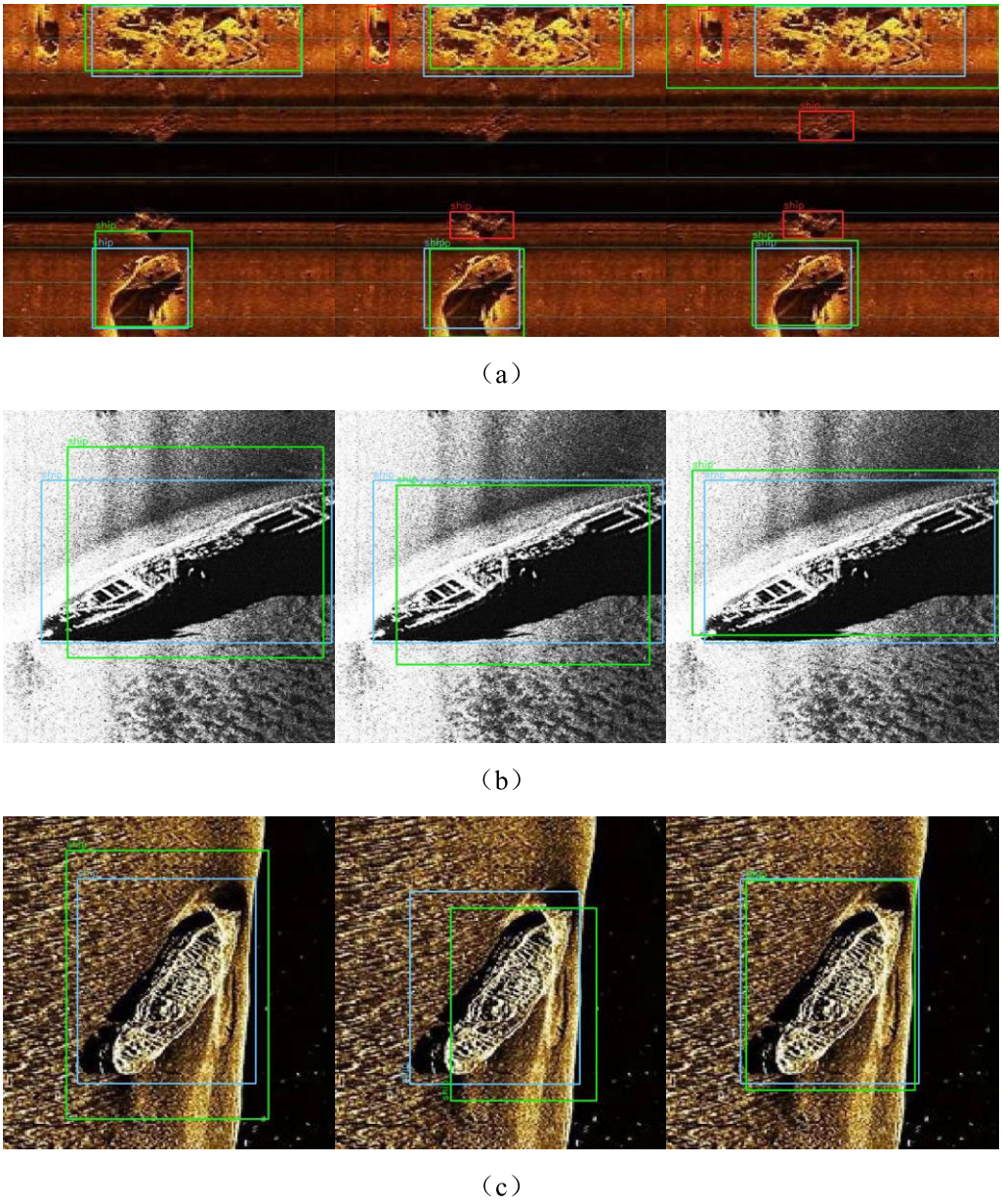


FIGURE 9. Comparison of partial target detection results of the three models. From left to right: the detection results of the transfer learning YOLOv3 model, traditional YOLOv3 model, and Faster R-CNN.

and transfer learning YOLOv3 model, respectively; the IOUs in Figure 9(c) are 77.03%, 69.32%, and 91.15%, respectively. Meanwhile, the confidence levels of the three models in Figure 9(b) are 98.88%, 98.97%, and 99.07%, respectively; in Figure 9(c), the confidence levels are 96.51%, 94.45%, and 99.42%, respectively. Obviously, the YOLOv3 model based on transfer learning has better recognition precision, positioning precision, and overall performance than the other two models.

IV. CONCLUSION

When applied to recognition of side-scan sonar images of shipwreck targets, the Faster R-CNN model is time-consuming and has low efficiency and a high missed detection rate for small targets. Considering the limitation that existing datasets of side-scan sonar images of shipwreck targets are small, we propose a YOLOv3 model that can automatically recognize side-scan sonar images of shipwreck targets based on transfer learning. According to the characteristics of the side-scan sonar wreck data set, 4x and 2x downsampling was done based on multi-scale feature fusion with FPN support by multi-scale training of shallow feature fusion; the target prediction was done by resetting parameters and size of anchor box using the K-means clustering and multi-scale feature fusion; lastly, the binary classification cross entropy function was used to improve the loss function of the YOLOv3 algorithm. Experimental results show that the AP value of the YOLOv3 model based on transfer learning reached 89.49%,

which is an improvement of 0.31% and 1.77%, respectively, compared with the Faster R-CNN model and the traditional YOLOv3 model; the harmonic mean F1 reached 90.71%, which is 3.96% and 1.63% higher, respectively, and it proves that the proposed model has better precision. The YOLOv3 model takes an average of 0.17 seconds to identify a picture, which is only 3/50 of R-CNN model. It greatly improves the detection efficiency and the overall performance of the model, and also meets the needs of real-time target recognition.

REFERENCES

[1] P. K. LeHardy and C. Moore, "Deep ocean search for Malaysia airlines flight 370," in *Proc. Oceans-St. John's*, Sep. 2014, pp. 1–4.

[2] Y. Liu, F. Xiao, and J. Bao, *Introduction to Hydrography*. Beijing, China: Serveying Mapping Publishing House, 2006, pp. 1–3.

[3] J. Zhao, J. Li, and M. Li, "Progress and future trend of hydrographic surveying and charting," *J. Geomatics*, vol. 34, no. 4, pp. 25–27, 2009.

[4] H. Cao, "Development status of deep-sea exploration technology and equipment in China," *Ship Supplies Marketing*, vol. 2, no. 2, pp. 19–22, 2005.

[5] Z. Wu, Y. Zheng, and Y. Chun, "Research status and prospect of sonar-detecting techniques near submarine," *Adv. Earth Sci.*, vol. 11, no. 11, pp. 58–65, 2005.

[6] H. Li, H. Teng, H. Song, D. Duan, and Y. Huang, "Technology on the extraction of seabed target based on high resolution side-scan sonar," *Hydrographic Survering Charting*, vol. 30, no. 6, pp. 71–73, 2010.

[7] M. Xiong, Z. Wu, and S. Lin, "Wavelet neural network identification and classification of sediment seabed sonar images based on genetic algorithms," *Acta Oceanologica Sinica*, vol. 36, no. 5, pp. 90–97, 2014.

[8] M. Xiong, Z. Wu, and S. Li, "Seafloor sonar sediment image recognition with the support vector machine," *Mar. Sci. Bull.*, vol. 31, no. 4, pp. 409–414, 2012.

[9] F. Yang, J. Li, J. Zhao, and Z. Du, "Seabed texture classification using BP neural network based on GA," *Sci. Surveying Mapping*, vol. 2, no. 2, pp. 111–114 and 118, 2006.

[10] F. Yang, Z. Du, Y. Wu, J. Li, and F. Chun, "Object recognizing on sonar image based on histogram and geometric feature," *Mar. Sci. Bull.*, vol. 25, no. 5, pp. 64–69, 2006.

[11] L. Lv, C. Zhou, C. Chen, and S. Jin, "Real-time detection of underwater target using side-scan sonar based on false alarm function," *Hydrographic Survering Charting*, vol. 33 no. 4, pp. 35–38, 2013.

[12] F. Yang, Z. Du, J. Li, and Z. Wu, "Side-scan sonar imagery segmentation based on Markov random field model," (in Chinese), *Acta Oceanologica Sinica*, vol. 28, no. 4, pp. 43–48, 2006.

[13] D. Zhu and H. Bian, "Research of side-scan sonar target auto detection," *J. Jilin Univ. (Inf. Sci. Ed.)*, vol. 26, no. 6, pp. 627–631, 2008.

[14] S. Kamal, S. K. Mohammed, P. R. S. Pillai, and M. H. Supriya, "Deep learning architectures for underwater target recognition," in *Proc. Ocean Electron. (SYMPOL)*, Kochi, India, Oct. 2013, pp. 48–54.

[15] J. Rhinelander, "Feature extraction and target classification of side-scan sonar images," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2016, pp. 1–6.

[16] J. Guo, J. Ma, and A. Wang, "Study of side-scan sonar image classification based on SVM and gray level co-cccurrence matrix," *Geomatic Spatial Inf. Technol.*, vol. 38, no. 3, pp. 60–63, 2015.

[17] Q. Chen, "Research-based underwater acoustic images underwater target recognition," Harbin Eng. Univ., Harbin, China, Tech. Rep. 1, 2012.

[18] D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. X. Zhu, "Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 193–205, Jan. 2019.

[19] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, "ORSIm detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5146–5158, Jul. 2019.

[20] X. Wu, D. Hong, J. Chanussot, Y. Xu, R. Tao, and Y. Wang, "Fourier-based rotation-invariant feature boosting: An efficient framework for geospatial object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 302–306, Feb. 2020.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

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

[23] J. Yu, D. Huang, L. Wang, X. Liu, and W. Li, "On-board ship targets detection method based on multi-scale salience enhancement for remote sensing image," in *Proc. IEEE 13th Int. Conf. Signal Process. (ICSP)*, Nov. 2016, pp. 252–256.

[24] 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. 2017.

[25] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015, *arXiv:1506.02640*. [Online]. Available: http://arxiv.org/abs/1506.02640

[26] L. Wei and D. Anguelov, "Single shot multibox detector," 2015, *arXiv:1512.0325v2*. [Online]. Available: http://arxiv.org/abs/1512.0325v2

[27] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. CVPR*, Jul. 2017, pp. 6517–6525.

[28] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," in *Proc. CVPR*, 2018, pp. 1–6.

[29] A. Krizhevsky, I. Isutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[30] Y. Bengio, F. Bastien, A. Bergeron, N. Boulanger-Lewandowski. T. Breuel, Y. Chherawala, M. Cisse, D. Erhan, J. Eustache, X. Glorot, R. Pascanu, R. F. Savard, and G. Sicard, "Deep learners benefit more from out-of-distribution examples," in *Proc. 14th Int. Conf. Artif. Intell. Statist.* Ft. Lauderdale, FL, USA: MIT Press, 2011, pp. 164–172.

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

[32] L. Y. Pratt and S. Thrun, "Special issue on inductive transfer," *Mach. Learn.*, vol. 28, no. 1, pp. 1–4, 1997.

[33] B. Chuong, A. Y. Ng, "Transfer learning for text classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 299–306.

[34] L. Mihalkova, T. Huynh, and R. J. Mooney, "Mapping and revising Markov logic networks for transfer learning," in *Proc. 22nd AAAI Conf. Artif. Intell.*, 2007, pp. 608–614.

[35] A. Niculescu-Mizil and R. Caruana, "Inductive transfer for Bayesian network structure learning," in *Proc. 11th Int. Conf. Artif. Intell. Statist.*, 2007, pp. 339–346.

[36] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks," *Comput. Sci.*, vol. 7, pp. 3320–3328, 2014.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* New York, NY, USA: IEEE Press, Jun. 2016, pp. 770–778.

[38] S. Ioffe and C. Szegedy, "Batch normalization accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.* Heidelberg, Germany: Springer, 2015, pp. 448–456.

[39] 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.* New York, NY, USA: IEEE Press, Jul. 2017, pp. 936–944.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[41] Z. Jiang, *Tensorflow Deep Learning Algorithm Principle and Programming Practice*. China Water Power Press, 2019, p. 1.

[42] X. Wei, M. Tu, and X. Zhang, *Deep Learning and Image Recognition Principle and Practice*. Beijing, China: China Machine Press, 2019, p. 6.

[43] F. Chollet, *Deep Learning With Python*. Beijing, China: POST&TELECOM PRESS, 2018, p. 8.

[44] Y. Li and T. Zhang, *Deep Learning Mastering Convolutional Neural Networks From Beginner*. Beijing, China: China Machine Press, 2018, p. 7.

TANG YULIN (Member, IEEE) is currently pursuing the master's degree with the Dalian Naval Academy. His main research direction is side-scan sonar image processing and computer vision. He is good at deep learning algorithm research and the application of CNN in image recognition and target detection.

SHAOHUA JIN (Member, IEEE) received the Ph.D. degree from the Second National Institute of Oceanography, in 2019. He is currently an Associate Professor with the Department of Hydrography and Cartography, Dalian Naval Academy. His main research direction is the marine and maritime geodetic survey data processing. He is good at the classification and recognition of marine sediment.

GANG BIAN is currently pursuing the Ph.D. degree with the Dalian Naval Academy. He is also a Lecturer with the Dalian Naval Academy. His main research direction is ocean gravity and magnetic measurement. He is good at the marine magnetic survey data processing.

YONGHOU ZHANG is currently pursuing the master's degree with the Dalian Naval Academy. He specializes in the theories and methods of hydrographic data processing.