This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2951443, IEEE Access

Sonar Image detection based on Multi-Scale Multi-Column Convolution Neural Networks

ZHEN WANG, SHANWEN ZHANG

School of Information Engineering, Xijing University, Xi'an 710123, China

Corresponding author: Shanwen Zhang (zswstudent@163.com)

This work was supported in part by the National Neural Science Foundation of China under Grant 61671465

ABSTRACT Automatic detection of underwater objects by sonar images is an important and challenging topic in applications of Autonomous Underwater Vehicle (AUV) under the complex marine environment. A detection method is proposed based on Multi-Scale Multi-Column Convolution Neural Networks (MSMC-CNNs). Firstly, the Multi-Scale Multi-Column CNNs is used to form an encoder network for extracting multi-scale features of the sonar image. Secondly, the bicubic linear interpolation algorithm is used as the deconvolution process of the decoder networks to restore the sonar image size and resolution. Moreover, a novel transfer learning manner based on progressive fine-tuning to accelerate the model training. Finally, the proposed method is validated on the sonar image dataset and is compared with other existing detection methods. The pixel accuracy (PA) of MSMC-CNNs for different categories sonar image is over 95%. The experiment results show that the MSMC-CNNs model has better detection effect and more robustness to noise.

INDEX TERMS Underwater object detection, MSMC-CNNs, Bicubic linear interpolation algorithm, Deconvolution.

I. INTRODUCTION

In Europe, the United States, and China pay much attention to marine research and have a deep foundation, and have conducted much research on the detection and localization of the underwater target. In China, many scientific research institutes and universities, such as Chinese Academy of Sciences (CAS), Harbin Engineering University, Institute of Acoustics of Chinese Academy of Sciences and Northwestern Polytechnic University have implemented much research on target detection and localization. Currently, underwater imaging technology mainly includes optical imaging and sonar imaging [1]. Optical imaging has a better resolution, but its resolution poor under the drowning environment and the imaging distance is relatively close [2]. Sonar imaging has the advantage of long operating distance and strong penetration ability, which is especially suitable for underwater environment. Therefore, it is widely used in underwater geomorphological exploration, underwater lost object search, and mine detection [3]. However, because of the complex and changeable characteristics of the water medium and its boundary of the underwater acoustic channel, as well as the propagation loss and scattering of the acoustic wave itself. As a result, the collected underwater object often has the characteristics of the low contrast, strong speckle noise and blurred target edge, which brings great difficulties to the manual interpretation of underwater object [4][5].

Accurate segmentation of the underwater object is convenient for further analysis of the underwater object. Underwater object detection not only depends on the segmented target region but also has a close relationship with seafloor noise and the background region, so its segmentation is difficult and complicated. The purpose of underwater object detection is to extract the target and shadow from the complex seafloor reverberation region and retain the original edge information of the underwater objects. Underwater object detection algorithm divide into a supervised segmentation method and unsupervised segmentation method. Supervised segmentation methods include Bayesian framework and variation theory framework [6] [7]. The Bayesian framework uses the similarity between local pixel statistics and seabed prototype statistics to represent the conditional likelihood function [8]. The most widely used methods in the Bayesian are the maximum posterior method and the maximum boundary method. Moreover, the maximum boundary probability method has been proved to be more suitable for underwater object detection [9]. Different from the Bayesian framework, the segmentation method based on variation theory needs to minimize the similarity region function between texture statistics and predefined prototype statistics in the region of the variation model. Since most supervised segmentation algorithms require hypothetical training classifiers, and the structure design of the algorithm is complex, there are few pieces of research on supervised segmentation algorithm for underwater object [10]. The unsupervised segmentation algorithm is more widely used than supervised segmentation algorithm. Most underwater unsupervised segmentation algorithms need a learning stage to segment automatically, among which Markov Random Field (MRF) method and active contour method are widely used[11-15]. Generally, since the underwater object contains a lot of seafloor reverberation noise, it cannot be detected effectively by conventional image segmentation method. Specifically, the image segmentation based on MRF is a method of pixel classification by using the spatial correlation of pixels in an image [16]. The MRF method can accurately describe the category to which each pixel belongs and its dependence on the surrounding pixel category. In order to realize accurate image segmentation based on this method, it is necessary to clarify the distribution characteristics of pixels in different regions. In the segmentation method based on MRF, the most widely used is the Hamersley-Clifford theorem, which represents the relationship between the local features and the global features of the underwater object through the energy function of the physical system [17]. For example, According to the imaging characteristics of the sonar target, Xie et al. [18] establish the segmentation constraint condition, make use of the small gray mean ratio of the shadow to the target to carry on the initial segmentation, and then remove the false target according to the width difference between the segmented target and the shadow. This method takes into account the dependence between adjacent pixels and has the advantages of strong anti-noise and accurate segmentation. The segmentation method based on active contour model is combined with the relevant theory of partial differential equation, the problem of underwater object detection can classify as a minimum functional problem, and then the minimum functional problem is transformed into the problem of solving a partial differential equation by variation method. The active contour model can divide into parametric active contour model and geometric active contour model [19]. The parametric active contour model based on the local information of the image, which is easily affected by noise and the initial test curve must be close to the edge of the target in order to get the correct segmentation results [20]. Geometric active contours can efficiently deal with topological changes which are challenging to deal with by level set method [21]. Specifically, Huo et al. [22] proposed a segmentation method based on non-local speckle reduction and improved active contour model. In the method, the non-local speckle filtering method is used to eliminate underwater speckle noise to improve the accuracy, and the k-means clustering method is adopted for the initial segmentation of the underwater object. At the same time, an edge-driven constraint is added to the region fitting filtering model to accelerate the convergence speed and drive the active contour to obtain the desired boundary. Sang et al. [23] use the active contour level set method to segment the underwater object. In the process of segmentation, there is no prior hypothesis or statistical modeling, and its core idea is to obtain the minimum effect when the bottom image is segmented. It is robust to underwater noise and has excellent regularization performance. Image processing methods based on histograms are also widely used in the field of object detection. Lv et al. [24] proposed a novel multi-scale object histogram distance (MOHD) to detect the target region of the remote sensing image. The method first calculates the frequency histogram of the image, then uses the bin-to-bin distance to measure the target change, and uses the Otsu algorithm to complete the target area segmentation. Liu et al. [25] proposed a histogram trend similarity adaptive detection method for high-resolution remote sensing image object recognition. The method first quantitatively analyzes the adaptive histogram trend of the target region, then uses the improved bin-to-bin distance to detect the magnitude of the change of the object, and uses the binary threshold method to complete the segmentation of the image of the changed region.

Image detection and classification methods based on deep learning have achieved tremendous success in digital image processing for object detection and classification. Convolutional neural networks (CNNs) have widely used to solve the problem of image processing and achieved significant progress not only in the task of image classification but also in semantic segmentation. Long et al. [26] proposed an FCN (Fully Convolution Neural Networks, FCN) model for semantic segmentation, which could be trained end-to-end on pixel-wise prediction. Badrinarayanan et al. [27] proposed an end-to-end SegNet semantic segmentation network model. At present, the segmentation method based on FCN and SegNet are applied to a variety of segmentation scenes, including medical image segmentation, autopilot, underwater image segmentation, and satellite image segmentation. Zhang et al. [28] combined with several existing CNNs models, constructed the satellite image classification method, and applied it to the satellite image analysis system. Deng et al. [29] proposed a CNNs model for multi-scale remote sensing target detection and used it for small target recognition in remote sensing images. Ji et al. [30] propose an underwater image restoration method based on CNNs, which is used to complete the image restoration for many kinds of underwater images. Moniruzzaman et al. [31] systematically described base on deep learning underwater imagery analysis methods in recent years. These approaches are categorized according to the object of detection, and the features and deep learning architectures used are highlighted. It concludes that there is an excellent scope for automation in the analysis of digital seabed imagery using deep learning, especially for the detection and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2951443, IEEE Access

monitoring underwater object by sonar image. Inspired by FCN [26], SegNet [27], and U-Net [31], a novel CNNs model, and namely MSMC-CNNs is proposed for underwater sonar images detection. MSMC-CNNs is composed of encoder and decoder, in which the multi-scale multi-column CNNs is used as encoder structure for extracting multi-scale information of sonar image, and the bicubic convolution is used as a deconvolution for the decoder structure to restore the original image size. Moreover, a novel transfer learning manner based on progressive fine-tuning to accelerate the training. The experimental results in the sonar image dataset show that MSMC-CNNs are superior to other semantic segmentation methods. The main contributions of this paper are listed as follows,

1) We propose the MSMC-CNNs model based on an encoder-decoder structure for sonar image recognition.
2) Using nine-point bilinear interpolation convolution as a deconvolution operation for the decoder structure in MSMC-CNNs, which it can restore the input image size and resolution.
3) The transfer learning algorithm based on progressive fine-tuning is used in the training process of the model, which can effectively improve the training speed and detection accuracy.
4) The method is validated and compared with the state-of-the-art methods on the sonar image dataset.

The rest of this paper is organized as follows. In Section 2, the related works are introduced, including FCN model structure, SegNet, and U-Net. In Section 3, MSMC-CNNs are described in detail. The experiments on the sonar image dataset are implemented in Section 4, and the conclusion is described in Section 5.

II. RELEVANT WORK

In this Section, the related works are introduced, including SegNet, U-Net, and FCN structure.

A. FCN

Since the traditional CNNs divides the pixel values contained in the image into different pixel blocks as the input of the network, the pixel-level image segmentation task cannot be completed. In order to solve the shortcomings of CNNs in the field of image segmentation, Long et al. [26] proposed the FCN model for image segmentation. The FCN is derived from the conversion of all fully connected layers of the original CNNs to a convolutional layer. The essential networks commonly used for conversion include AlexNet, GoogleNet, and VGG16. In these three basic network models, AlexNet is relatively simple, so it is inferior in effect [32]. GoogLeNet involves more training parameters, which makes the model less practical [33]. Taken together, the original FCN model was designed based on the VGG16 model [34]. The FCN uses VGG16 as the backbone networks and its structure is shown in the Fig.1, in which Conv represents the convolutional layer, Pool represents the max-pooling layer, and Upsampling represents the up-sampling layer. In the FCN model, it first replaces the fully convolutional layer of VGG16 with a convolutional layer and then obtains FCN-8s, FCN-16s, and FCN-32s by deconvolution and up-sampling operations.

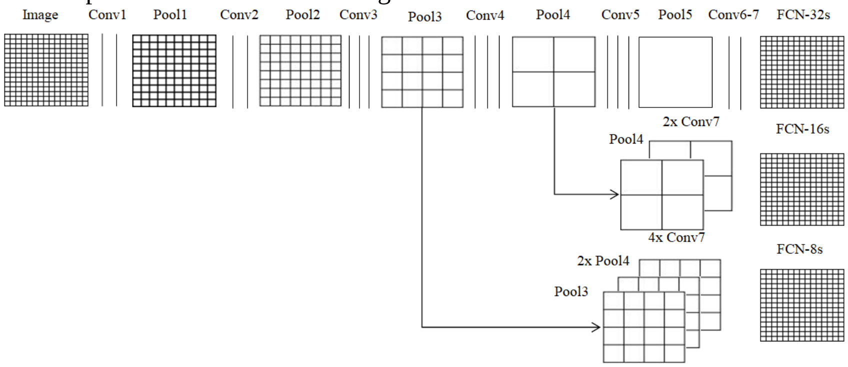


Fig.1 Architecture of FCN

B. SegNet

CNNs can automatically extract the in-depth features of the input images and has achieved the prominent performance in a series of image processing tasks, such as image classification, object detection, and segmentation. SegNet is an end-to-end semantic segmentation model, consisting of encoder network, decoder network, and pixel-wise classification layer, and each convolutional layer includes batch normalization and ReLU activation functions. Its architecture is shown in Fig.2, where the encoder network converts high-dimensional vectors into low-dimensional vectors and realizes the low-dimensional extraction of high-dimensional features. The decoder network uses the max-pooling index information of the corresponding feature layer saved by the encoder down-sampling to map the low-resolution feature map to the high-resolution feature map and realizes the reconstruction from low-dimensional vector to high-dimensional vector. The reuse of the max-pooling index in the decoder process can optimize the boundary profile, reduce the number of parameters, and carry out end-to-end model training.

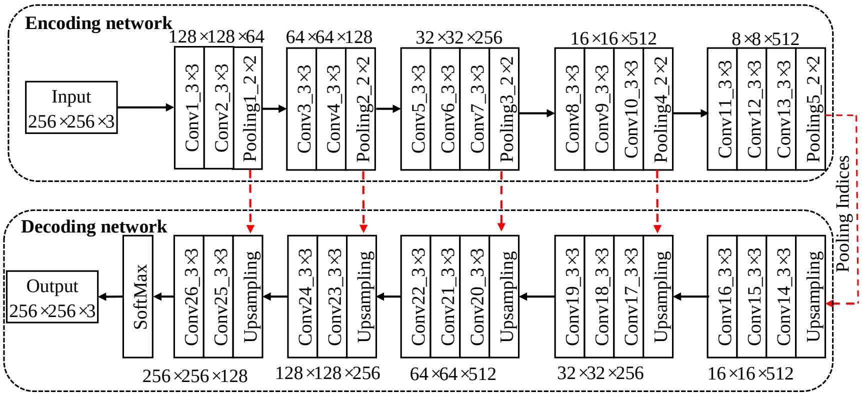


Fig.2 Architecture of SegNet

C. U-Net

The U-Net architecture is a U-shaped model with features of an image learned at different levels through a set of convolutional and max-pooling layers, as seen in Fig. 3.

U-Net contains contracting network and expanding network corresponding to each other to form a U-shaped structure. The contracting network is mainly used to capture the context information in the image, and the advertised network is symmetrical to precisely locate the part that needs to be segmented in the image. The characteristic of U-Net is that the contraction network and the expansion network are mutually mapped. In the process of expansion, the missing boundary information is complemented by the features of the merged map contract layers, and the accuracy of the predicted edge information is improved.

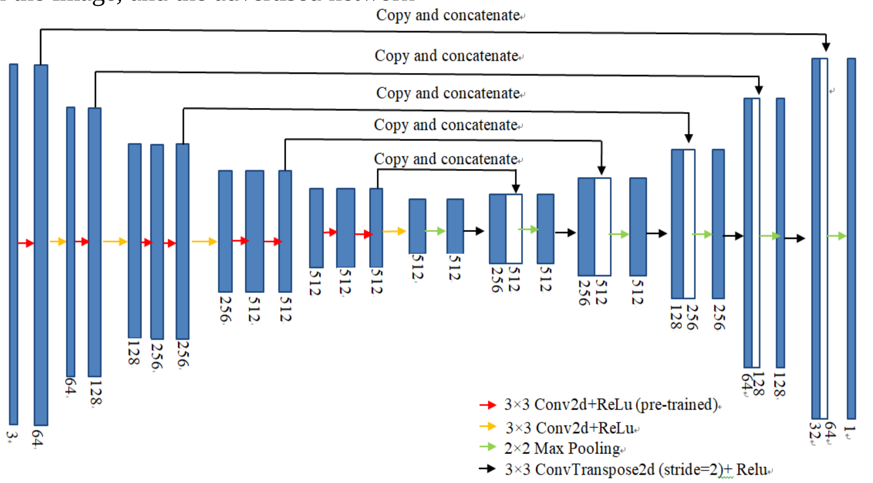


Fig.3 Architecture of U-Net

III. MSMC-CNNs

In the section, the MSMC-CNNs is constructed for underwater object detection, including architecture, nine-point bilinear interpolation algorithm, procedure, and performance evaluation.

A. Model Architecture

MSMC-CNNs is a novel encoder-decoder network structure, where the multi-scale multi-column CNNs is used to construct the encoder structure, which can extract multi-scale features of the sonar image, and using bicubic linear interpolation as the decoder network, which can effectively recover the lost information caused by the encoder network. The MSMC-CNNs model uses the Multi-Column CNNs model as the backbone structure and learns the concept of Multi-Scale connections in the SA-CNNs model [35]. Similar to the Multi-Column CNNs, which MSMC-CNNs is also composed of three parallel sub-convolution neural networks, each of which has the same structure except for the size and number of convolution kernels. In the traditional CNNs model, the pooling operation can compress the feature map and simplify the network computation complexity, but it also leads to the loss of features. Inspired by the SA-CNNs model, in each sub-network of MSMC-CNNs, multi-scale connections are made to the feature maps of different layers to adjust the scale and viewing angle of the sonar image. The connected feature map has the characteristics of multi-scale features, including low-level features and high-level features, corresponding to different targets in the sonar image. The overall structure of the MSMC-CNNs is shown in Fig.4.

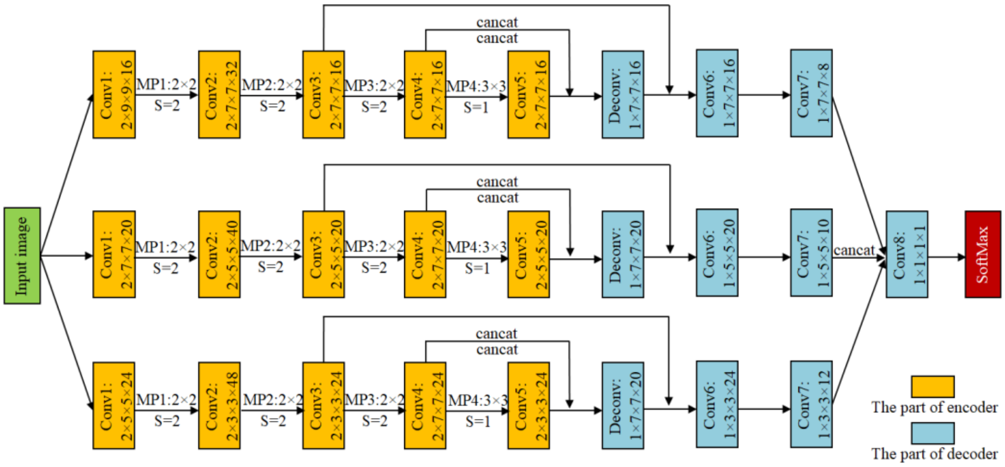


Fig.4 Structure of MSMC-CNNs

Each group of sub-networks has twelve convolutional layers, one deconvolution layer, four max-pooling layers, and two multi-scale connections. Referring to Fig.4, the following introduce the components of MSMC-CNNs,

1) Multi-Scale Connection. The multi-scale connection can connect feature maps of the same resolution output from different convolutional layers on a convolution channel. The purpose of multi-scale connection is to be able to share the same low-level parameters and feature maps, which can reduce the number of parameters, and speed up the training process.
2) Convolution Layer. The convolution layer is to extract features from the input image. Conv in Fig.4 represents the convolution layer. The p_1 in the parameter $p_1 \times p_2 \times p_3 \times p_4$ represents the number of the convolution layers, the $p_2 \times p_3$ represents the size of the convolution kernel, and p_4 represents the number of convolution kernel channels.
3) Max-Pooling Layer. The function of Max-Pooling layer is to reduce the channel numbers of convolution layers. In the show of Fig.4, the MP represents the max-pooling layers. With the exception of the pooling layer MP4, the pooling size of the other max-pooling layer is defined as 2×2 and the stride size is 2. In order to connect the feature maps output by Conv4 and Conv5 of multiple scales, MSMC-CNNs sets the pooling are size of MP4 is 3×3 and the stride size to 1.
4) Deconvolution Layer. Deconv in Fig.4 represents a deconvolution layer with a parameter form similar to that of a convolutional layer. This method uses a deconvolution layer to up-sampling the Conv4 and Conv5 multi-scale connections to a quarter of the input image resolution. Therefore, the feature map of the Deconv output and the feature map of the Conv3 output can be further multi-scale connection.

B. Bilinear Interpolation Algorithm

Traditional semantic segmentation network models such as FCN, SegNet, and U-Net use bilinear interpolation to recover image resolution during deconvolution. We are inspired by Robert G. Keys et al. [36], so we used the bicubic linear interpolation method in the MSMC-CNNs model to deconvolution and restored the input image resolution. In the process of deconvolution using bicubic linear interpolation, one-dimensional interpolation is performed on the convolution layer feature map in the vertical direction and the horizontal direction, thereby realizing two-dimensional bicubic linear interpolation. The size of the input convolution feature map is defined as $M \times N$, and the feature map is enlarged to size by the linear interpolation method. The implementation steps of the bicubic convolution interpolation method are as follows,

Step 1: Define the input image as F and the image size as $M \times N$.

Step 2: The input image is interpolated with the vertical and horizontal direction.

Step 3: The interpolation image is defined as G, and the interpolated image size as $S \times T$.

The expression of the bicubic convolution interpolation function is as follows,

$$g(x)=\frac{c_{i+2}-c_{i+1}+c_i}{2}s^2+\frac{-c_{i+2}+4c_{i+1}-3c_i}{2}s+c_i \qquad (1)$$

Where,

$$\begin{cases} c_i = f(x_i) \\ c_m = 3c_{m-1}-3c_{m-2}+c_{m-3} \\ c_{m+1} = 3c_m - 3c_{m-1} + c_{m-2} \end{cases} \qquad (2)$$

According to Eq. (2), calculating the value of the function at requires only calculating the values of the three sample points. In the image processing process, the two-dimensional bicubic convolution interpolation method can be expressed as follows,

$$g(x,y)=\sum_{a=-1}^{2}\sum_{b=-1}^{2}c_{i+a,j+b}u(s_1-a)u(s_2-b) \qquad (3)$$

where, $s_1 = \frac{x}{\triangle x} - i\,, i = \left|\frac{x}{\triangle x}\right|$; $s_2 = \frac{y}{\triangle y} - j\,, j = \left|\frac{y}{\triangle y}\right|$.

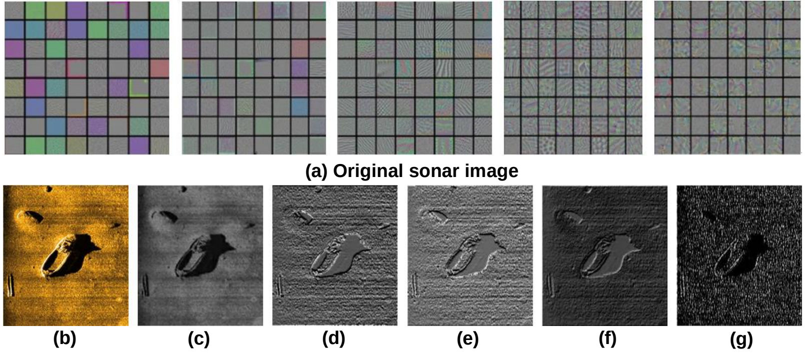
The two-dimensional cubic convolution interpolation method is similar to the one-dimensional bicubic linear interpolation convolution principle. Because of the third-order approximation of $g(x, y)$ and $f(x, y)$, the coefficient of the cubic term of $g(x, y)$ is zero, so the calculation coefficient relation of the two-dimensional cubic convolution interpolation method is as follows,

$$\begin{cases} c_{i,j} = f(x_i, y_i) \\ c_{i+a,j-1} = 3c_{i+a,j} - 3c_{i+a,j+1} + c_{i+a,j+2} \\ c_{i+a,j+2} = 3c_{i+a,j+1} - 3c_{i+a,j} + c_{i+a,j-1} \\ c_{i-1,j+b} = 3c_{i,j+b} - 3c_{i+1,j+b} + c_{i+2,j+b} \\ c_{i+2,j+b} = 3c_{i+1,j+b} - 3c_{i,j+b} + c_{i-1,j+b} \end{cases} \tag{4}$$

Therefore, Eq. (3) only needs to use nine sampled pixel values to predict the value of an interpolated pixel.

C. Feature Extraction

In the training process of the network, the input image is input into the first convolutional layer, and the feature maps are extracted from several convolutional and pooling layers. The extracted maps by different convolutional kernels are different. The shallow convolutional kernels extract the color and contour features of the disease image; the deeper convolutional kernels extract the texture and detail features of the image. Convolutional kernels and feature maps can be visualized to visually display features extracted from different convolutional kernels, as shown in Fig.5. As seen in Fig.5 (a), the convolutional kernels (Kernel1~Kernel3) displays coarser information, and the convolutional kernel (Kernel4~Kernel5) shows more detailed features of the image, the feature maps obtained by the convolutional layer (Conv1~Conv3) are mainly the contour feature of the sonar image, and the convolutional layer (Conv4~Conv5) mainly contains the texture features of the sonar image, which fully reflects the convolutional nerve. With the weight sharing technology of the network, as the number of convolutional layers increases, the model acquires more detailed features of the input image. It is shown that different convolutional kernels can obtain different features of the input image during the training process. Each convolutional kernel pays attention to different parts of the image, and can thoroughly study the salient image regions of the respective parts of interest, which lays a foundation for the accurate segmentation of subsequent images. As can be seen from the Fig.5 (c), the edge of the sonar image is highlighted by the convolutional neural network. When it comes to the output of the middle layer which is shown in Fig.5 (d), the feature map appears to be more abstract. This is because the middle layer of a neural network is difficult to interpret in general. In Fig.5 (e), the output of the convolutional neural network has no noticeable sharp edges. The edges of the sonar image gradually fade, which is expected because the model is supposed to pay more attention to the object region instead of the edge of the sonar image. In the output of the activation layer shown in Fig.5 (f) and Fig.5 (g), which is close to the output layer, the sonar object region becomes more concentrated.


(a) Original sonar image

(b) (c) (d) (e) (f) (g)

Fig.5 Convolutional Kernels and corresponding feature map

D. Sonar Image Recognition Procedure

The experimental procedure of the MSMC-CNNs based underwater object detection is shown in Fig.6, which is briefly described as follows,

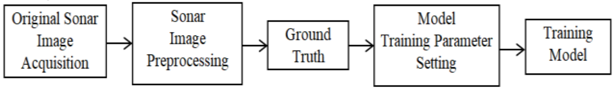

Original Sonar Image Acquisition
Sonar Image Preprocessing
Ground Truth
Model Training Parameter Setting
Training Model

Fig.6 Experimental steps of MSMC-CNNs based underwater object detection method

Suppose $S = \{(X_n, Y_n), n = 1, ..., N\}$ is the input dataset, where $X_n = \{X_{ij}^n; i = 1, ..., w, j = 1, ...h\}$ represent the original input image with width (w) and length (h) , and $Y_n = \{Y_{ij}^n; i = 1, ..., w, j = 1, ..., h; Y_{ij}^n \in \{0,1\}\}$ is the real ground truth binary map for image X_n. The entire set of network layer parameters in residual block and convolution block are represented as w_b and w. The functions of convolution block compute outputs y_i by $y_i = F(x_i, w_i)$, in which $\boldsymbol{F}$ represent the layer type of batch normalization layer, ReLU nonlinearity activation function, and matrix multiplication for convolution.

Suppose there are S output results in the network, Θ is the entire parameters of network layer. Each result is assigned with a classifier, and the corresponding weights can be defined as $(\theta^{(1)},...,\theta^{(N)})$. CNNs optimize network parameters by calculating the loss value during the training process. In the training process of MSMC-CNNs, loss function is defined as follows,

$$Loss(\Theta) = -\frac{1}{n}\left[\sum_{i=1}^{n}(\alpha y_i = 1 | X;\Theta) + (1-\alpha)(1-y_i)\log P(y_i = 0 | X;\Theta))\right] + \lambda R(\Theta) \quad (10)$$

Where n represent the number of training samples in each batch and α corresponds to the ratio of background pixels over all the pixels. The $LogP(y_i = 0 | X;\Theta)$ represent calculated using sigmoid functions on the activation value at pixel i , $\boldsymbol{R}$ is the regularization term, and λ is the hyper-parameters of regularization. In the process of the model training, the gradient descent algorithm is used to optimize the objective function. The objective function of MSMC-CNNs is defined by $L(\Theta, \theta^{(m)}) = \beta loss(\theta)$, in which $(\beta_1, \beta_2, ..., \beta_m)$ some hyper parameters are introduced for loss function from the output. The objective function below is minimized by $(\Theta, \theta^{(m)}) = argmin(L(\Theta, \theta^{(m)}))$, which is the stochastic gradient descent (SGD).

E. Evaluation Indication

Four evaluation metrics are adopted to quantify the segmentation performance: Pixel-classification Accuracy (PA) which is the average of the prediction accuracy over all categories, Mean Accuracy (MA), Mean Intersection over Union (MIU), and Frequency Weighted Intersection over Union (FWIU), are computed as follows,

$$\text{PA} = \sum_i n_{ii} / \sum_i t_{ii} \quad (6)$$

$$\text{MA}=1/n_{cl} \sum_i n_{ii} / \sum_i t_i \quad (7)$$

$$\text{MIU} = 1/n_{cl} \sum_i n_{ii} / (t_i + \sum_j n_{ji} - n_{ii}) \quad (8)$$

$$\text{FWIU} = (\sum_k t_k)^{-1} \sum_i t_i n_{ii} / (t_i + \sum_j n_{ji} - n_{ii}) \quad (9)$$

where n_{cl} is the number of categories of image pixels to be divided, i is the correct pixel class corresponding to a pixel, j is the pixel class to which a pixel is misclassified, t_i is the total number of pixels of category i in the standard segmentation result, and n_{ii} is the segmentation quantity of pixels in the result that are correctly labeled as category i , and n_{ji} is the number of pixels in the segmentation result that the pixel book belongs to category i but is misclassified into category j.

IV. EXPERIMENTS AND ANALYSIS

To validate the MSMC-CNNs based underwater object detection method, a lot of experiments are conducted on the side underwater object dataset, and compared with three state-of-art methods, i.e., FCN, SegNet and U-Net. Taking into account the memory constraints of the server, the model using trained in batch training. Each of the 650 images is input into the model as a batch. The training of all data in the dataset requires 35 batches. The model parameters are updated by the algorithms of gradient descent and back propagation. All experiments are conducted with Intel Xeon E5-2643v3 @3.40GHz CPU, 64GB RAM, NVidia Quadro M4000 GPU, 8GB of video memory, by CUDA Toolkit 9.0, CUDNN V7.0, Python 3.5.2, Tensorflow-GPU 1.8.0, Windows 7 64bit operating system.

A. Data Collection and Preprocessing

The sonar images of underwater objects were collected by dual-frequency side-scan sonar (Shark-S450D) in Qingdao, China. Some examples are shown in Fig.7. The bilinear interpolation method is used to enhance the image of the underwater object, and the results are displayed in Fig. 8.

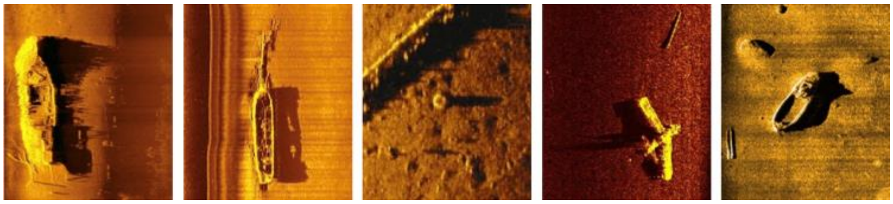


Fig.7 Examples of the side underwater objects

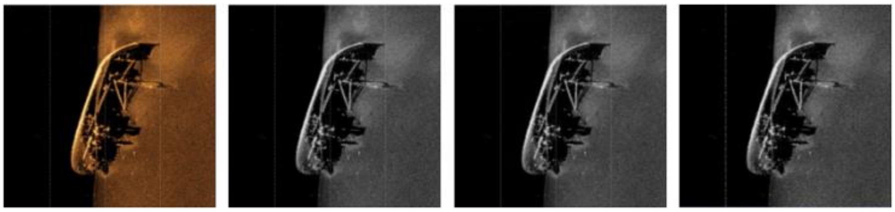
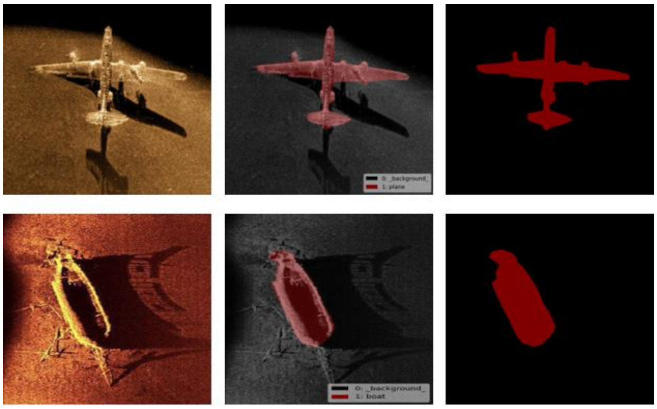


Fig.8 Image preprocessing of underwater

For MSMC-CNNs training, it is essential to label the training image to establish the benchmark of detection results. The label image used to train the weight parameters of the model and obtain the optimal training results. We used Label-Me is an image annotation tool developed by MIT to label images manually. The labeled images are shown in Fig.9, where the red represents the detection targets in the underwater that are airplane and shipwreck, while the black is the background.



(a) Original image (b) manually label (c) Label result

Fig.9 The sonar images of underwater and corresponding labeled images

In order to speed up the training time, each original image is cut with the size of 15000×8000. Then, 70% of the images are randomly selected as the training dataset and the rest as test dataset. In the training set, about 20% of all images are randomly selected as the validation dataset. Finally, the dataset consisted of 4120 training images, 932 validation images, and 1125 testing images.

B. MSMC-CNNs TRAINING PROCESS

Since the training of CNNs needs to be carried out on large data sets, but due to environmental and equipment constraints, it is impossible to collect massive sonar data. If the network model is trained directly on the sonar image dataset, will result in the network model over-fitting. Therefore, in the process of MSMC-CNNs training, the feature extraction ability learned on the Image-Net dataset needs to be transference to the sonar dataset as a priori knowledge. However, the similarity between sonar image and the original image is low, and it is difficult for CNNs to accurately summarize sonar image features through the learned feature knowledge, resulting in weak segmentation effect of the network model. Therefore, the pre-training network needs to be retrained on the sonar dataset, so that the network adaptively adjusts the network parameters for the target samples, and improves the feature extraction ability of the network model. In the feature extraction process of CNNs, shallow convolutional layers are used to extract low-level features such as color, edge, and shape of the input image. As the number of network layers increases, the network model can extract high-level features such as image hierarchies and textures. From low-level features to high-level features, CNNs have a feature-specific transition to the feature extraction process of images, while the traditional transference learning strategy does not further explore the relationship between feature gradation and sample data size and feature similarity. In order to improve the training efficiency of MSMC-CNNs, we propose a progressive fine-tuning strategy based on migration learning. The progressive fine-tuning strategy is shown in Fig. 10.

As shown in Fig.10, the steps of the progressive fine-tuning strategy are as follows,

Step 1: Only the new convolution layer with random initialization is trained.

Step 2: On the basis of the network nonlinear feature classifier, the convolution layer is released layer by layer, and the trainable layer is fine-tuned until the entire network is trained.

Step 3: Quantitatively analyze the change in the loss value after fine-tuning the layers, and then determine the sufficient depth of the fine-tuning.

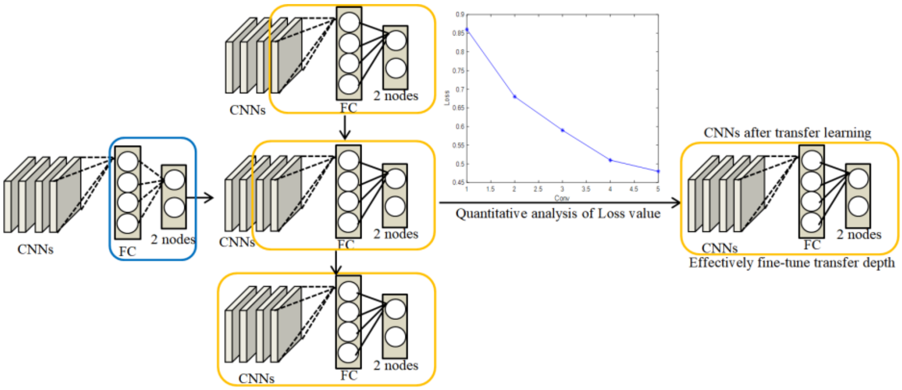


Fig.10 Progressive fine-tuning strategy

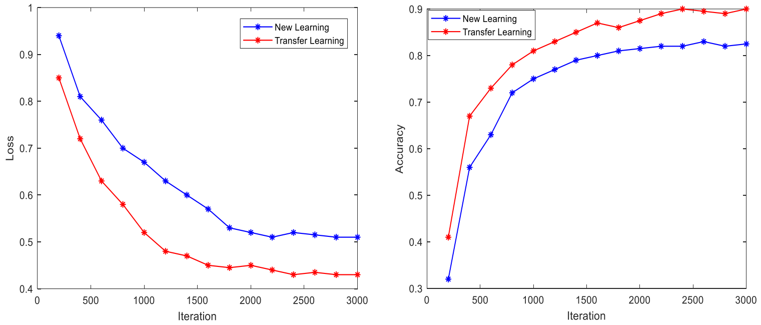


Fig.11 Comparison of two training methods

The initial learning rate of the network model is set to 0.001, the regular term coefficient is set to 0.001, and the number of iterations is set to 1000. The training effect of the two training models of the new learning and transfer learning is compared in the weed dataset. The effect of the different training methods is shown in Fig.11.

Using the transfer learning method to train the network model can effectively improve the training speed and recognition accuracy of the model. It can be seen from Fig.11 that in the new learning training mode, the initial recognition accuracy of the model is only 0.65. When the number of iterations reaches 500, the recognition accuracy reaches 0.86, but the model has a convergence trend. After the iterative training is completed, the model recognition accuracy is only 0.91, which indicates that the model has a weak training effect. When using transfer learning to train the model, the initial learning rate of the model reached 0.73, which has the ability to identify weed images. During the first 500 iterations, the model converges rapidly. When the number of iterations reaches 500, the accuracy of model recognition is the same as the accuracy of 1000 iterations in the new learning mode, indicating that transfer learning can save the model training times. After completing the iterative training, the model recognition accuracy rate reached 0.92, which is much improved in the recognition accuracy compared with the new learning mode.

C. Visualization of the MSMC-CNNs

In the underwater object detection, each training original sonar image is first manually labeled, where the target part of the image is marked as the foreground, and the rest of the image is marked as the background. Then the convolution layers are used to extract the labeled image, the pooling layers are used to down-sample the convolution feature maps, and finally, the activation layer is utilized to enhance the feature expression ability. The decoder includes up-sampling, convolution, and deconvolution. The up-sampling is used to recover the feature information loss caused by the pooling layers, and the deconvolution layers of decoder network are used to extract the in-depth semantic features of the feature map after the convolution layers are used to extract the significant area of the feature maps. After encoder and decoder, the SoftMax classifier is used to classify each pixel of the image, and outputs the detection result, as shown in Fig. 12.

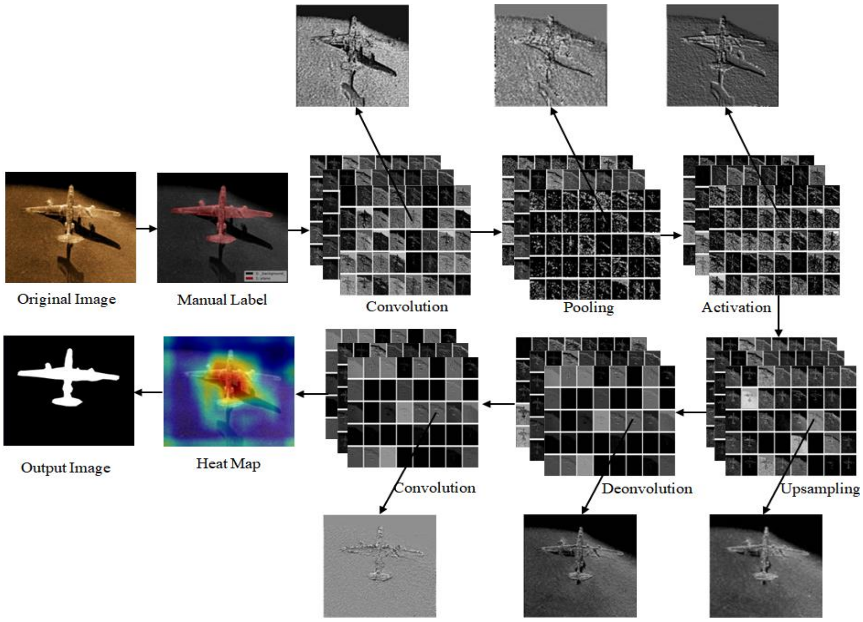


Fig. 12 Visualization of the underwater object detection process

D. DETECTION RESULTS

The proposed method is compared with FCN, SegNet, and U-Net. The detection images of the underwater object are shown in Fig.13. The detection results in term of PA, MA, MIU, and FWIU by Eqs. (6) to (9) are listed in Tab.1. From Fig.13 and Tab.1, it can be seen that the performance of MS-SegNet is the best, and its PA is over 95%, SegNet is better than U-Net and FCN. FCN can segment the contour region of the underwater object, but the effect of image detail detection is imperfect. U-Net is better than FCN, because its copy and concatenate operation can restore the detail features of the input images, so the detection effect is better, and its PA value is more than 93%.

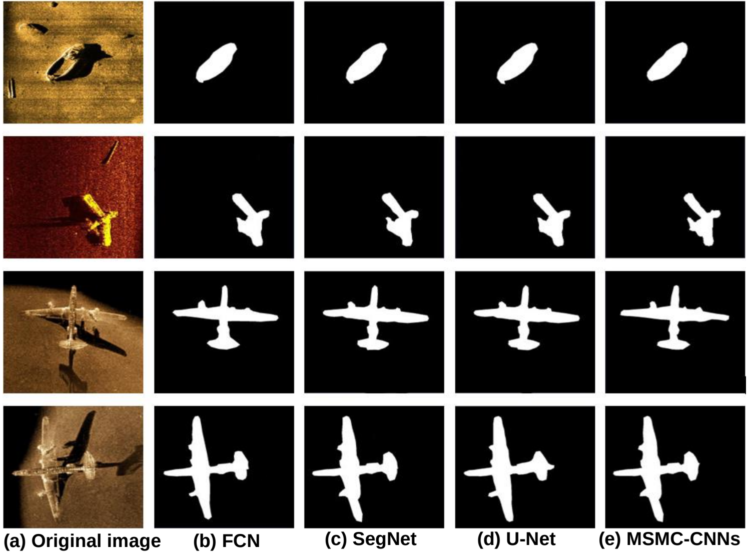


Fig.13 The detection results of the different methods

Tab.1 The detection results of the different methods

Model	PA (%)	MA (%)	MIU (%)	FWIU (%)
FCN	92.04	74.96	66.52	85.94
SegNet	92.85	75.96	66.81	86.92
U-Net	93.25	76.89	67.93	87.04
MSMC-CNNs	95.32	78.03	69.32	88.61

To further validate the performance of the proposed method, Fig.14 is the detection results of four networks versus the number of the iteration. From Fig.14, it is found that the convergence speed of the proposed method is fast in the process of training, and the training effect of the model is best under the condition of the same number of iterations. This is because the global average pooling layer is used in the design of MSMC-CNNs structure, which not only reduces the training parameters of the model but also speeds up the training speed. The convergence speed of FCN is the slowest and fluctuates significantly in the process of training, which not only restores the image resolution but also increases the parameter computation and training time of the model. The training process of U-Net and SegNet is similar because they have the same structure and are composed of the encoder network and decoder network. At the same time, the image resolution is restored by pooling index, but due to the direct and straightforward use of the superposition of the convolution layer and pooling layer. As a result, there were different amplitudes of shock in the process of training.

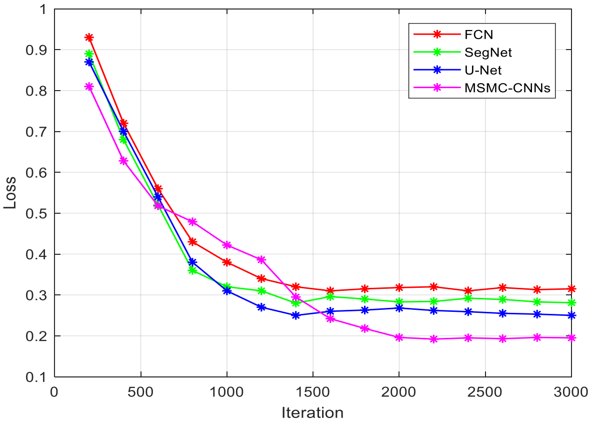


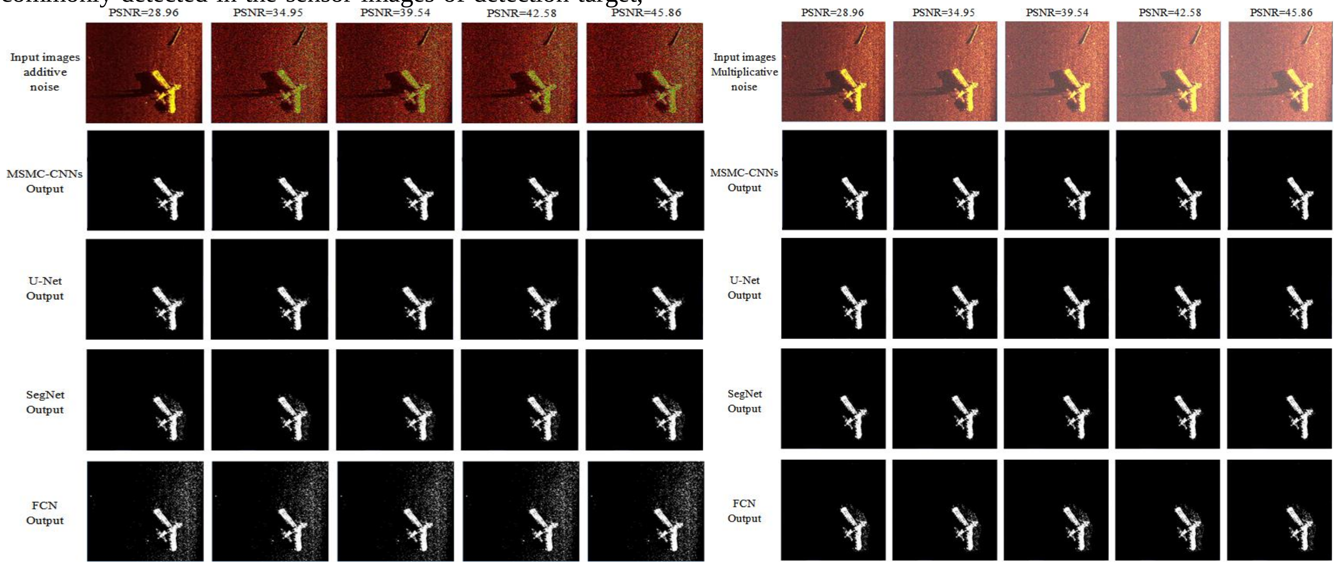
Fig.14 Training process for different network models

The calculation efficiency of the model is evaluated using the memory space, the number of parameters, training time, and testing time. Tab.2 compares the operational efficiencies of the four network models. In the comparison of training time and test time, MSMC-CNNs training time and testing time are the minimum values of four network models, indicating that the network model is more efficient. In the comparison of the number of parameters, the parameter quantity of MSMC-CNNs reaches 32,178,225. Since the global average pooling layer is used in MSMC-CNNs, the parameter calculation is minimal compared to the existing network model. Although there are many numbers of calculation parameters in MSMC-CNNs, the parameter quantity is significantly reduced compared with four network models. Because the MSMC-CNNs construct in a multi-scale manner, so MSMC-CNNs occupy less memory space.

Tab.2 Comparison of different architectures performance

Model	Memory Space	Training Time	Test Time	Parameters
FCN	8.45GB	8.3h	12.3 s	63,254,618
SegNet	6.13GB	6.2h	9.6 s	56,421,514
U-Net	5.42GB	5.8h	8.5s	48,163,254
MSMC-CNNs	3.28GB	3.6h	6.2s	32,178,225

Due to the influence of the seafloor environment, such as underwater acoustic channel, hydrological medium, and electromagnetic wave transmission, the imaging characteristics of sensor, high noise and weak boundaries are commonly detected in the sensor images of detection target, in which the speckle noise has the most considerable influence on the detection of sonar image. The first row is a sonar test sample with five different peak signal-to-noise ratios (PSNR).



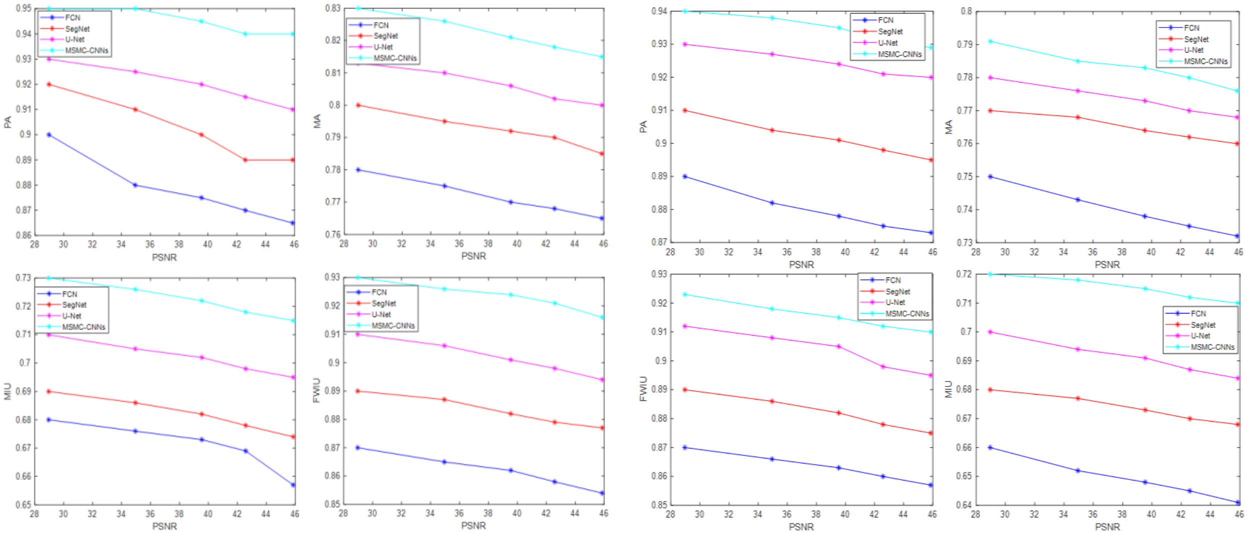
(a) Comparison of different segmentation model with five levels of additive noise

(b) Comparison of different segmentation model with five levels of multiplicative noise

Fig. 15 Comparison of different segmentation model with multiplicative noise and additive noise

Correspondingly, the effectiveness of the method is proved by analyzing the experimental results. FCN, SegNet, and U-Net were tested in the same sonar image data set. Fig.15(a) shows four evaluation parameters (PA, MA, MIU, and FWIU) at five additive noise levels. Fig 15(b) shows Comparison of five multiplicative noise levels, where the X coordinates is PSNR. We can see the following trends,

1) MSMC-CNNs have high stability for additive noise and multiplicative noise in the process of sonar image segmentation. The four evaluation indexes (PA, MA, MIU, and FWIU) have a small fluctuation range with increasing signal-to-noise ratio, and the segmentation accuracy is the highest.
2) When the PSNR is less than 40 in Fig.15(a), all four indicators of U-Net have a significant downward trend, which indicates that U-Net is not as stable as MSMC-CNNs when additive noise is relatively high.
3) In Fig.15(a), FCN and SegNet are sensitive to additive noise, especially when the PSNR is less than 35. A comparison of the four evaluation indicators (PA, MA, MIU, and FWIU) shows that they are not as good as MSMC-CNNs and U-Net. Although these two methods seem to have similar performance to the U-Net and MSMC-CNNs in the PA indicator, this phenomenon should be related to the defect of the PA indicator.
4) In Fig.15(b), although the overall performance of FCN and SegNet is not as good as MSMC-CNNs, FCN and SegNet have high stability and tolerance to multiplicative noise. When PSRN is higher than 35, the PA of FCN and SegNet is significantly reduced.



(a) Four image segmentation standard value under five additive noise levels

(b) Four image segmentation standard value under five multiplicative noise levels

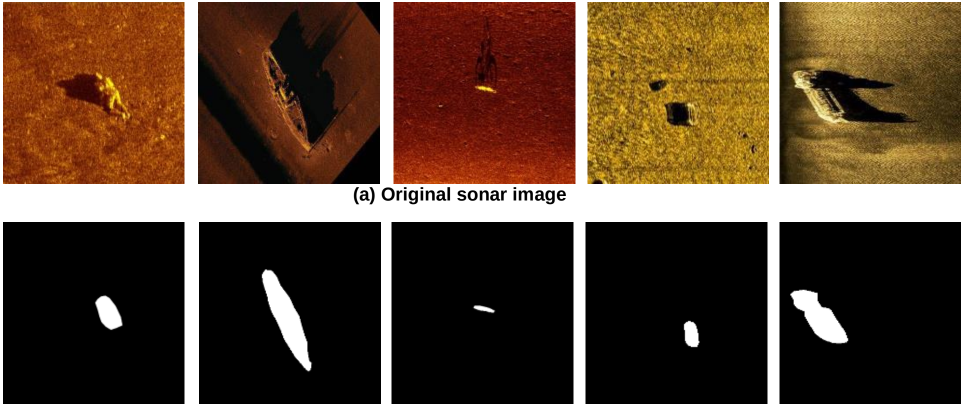
Fig.16 Four image segmentation standard value under five additive noise and multiplicative levels

In summary, the four parameters of additive and multiplicative noise (PA, MA, MIU, and FWIU) in Fig.16(a) and Fig.16(b) are compared to the FCN, SegNet, and U-Net semantic segmentation algorithms. Tab.3 shows the comparison of segmentation accuracy (PA, MA, MIU, and FWIU average) and run time (GPU times) using the same test dataset. MSMC-CNNs not only outperform other semantic segmentation models in terms of four evaluation indicators but also have an advantage in terms of runtime.

Tab.3 Comparison of the four segmentation methods

Model	Time (s)	PA	MA	MIU	FWIU
FCN	1.23h	0.88	0.82	0.78	0.82
SegNet	1.12h	0.93	0.89	0.80	0.91
U-Net	1.21h	0.95	0.91	0.88	0.87
MSMC-CNNs	1.08h	0.97	0.95	0.93	0.91

In the detection of underwater images, the change in image quality has a greater impact on the detection results[38]. The sonar image is a kind of underwater image, and its image quality also has interference with the detection result. In order to analyze the effect of MSMC-CNNs on different quality sonar images, it is used to detect sonar image of different pixel sizes (PS), and the experimental results are as follows,



(a) Original sonar image

(b) Detection result

Fig.17 Detection result of the different quality sonar image

Tab.4 Detection results of the different quality sonar image

Type	PA (%)	MA (%)	MIU (%)	FWIU (%)
PS=10000×7200	96.13	81.02	73.58	90.01
PS=8000×6400	94.16	79.38	71.06	82.56
PS=6000×4300	92.26	72.18	68.93	80.26
PS=2000×1200	90.36	70.28	65.89	78.51

It can be seen from Fig.17 and Tab.4. As the image quality decreases, the detection accuracy of MCMS-CNNs also decreases, indicating that the quality of sonar images has an effect on the detection effect of MSMC-CNNs. However, its PA value is maintained above 90%, indicating that MSMC-CNNs are less affected by changes in sonar image quality.

K-fold cross-validation can effectively avoid over-fitting and under-fitting. Therefore, we randomly divide the sonar image data set into K groups to verify the effectiveness of the training model. K-fold cross-validation uses each subset of the data set as a test set and the remaining K-1 sets as a training dataset. In the K-fold cross-validation, we use the mean and variance of PA and MA to verify the performance of the model. The mean and variance of PA and MA for the network model for K = 3, 5, 7, and 9 are shown in Tab.5. When K is increased to 7, the mean and variance of PA and MA tend to be stable. Here, K is set to 7 in consideration of statistical stability and calculation cost. Therefore, the ratio of the test set to the training set is 1:6, the total dataset has 4200 samples, the test dataset has 600 samples, and the training dataset contains 3600 samples.

Tab.5 Mean and variance of PA and MA based on K-fold cross-validation

K	PA (Average)	MA (Average)	PA (Variance)	MA (Variance)
3	0.965	0.921	0.013	0.0096
5	0.973	0.935	0.0024	0.0054
7	0.981	0.942	0.00051	0.0043
9	0.986	0.956	0.00036	0.0021

From Figs.13 to 16, Tab.2 to Tab.4, it can be seen that the proposed method outperformers the other networks. MSMC-CNNs can accurately detect the whole area of underwater objects. It can not only detect the contour area of the underwater object but also completely detect the edge details of the underwater object. SegNet and U-Net can also completely detect the target area of the underwater object, and the detection effect on the small area is better, but when there are shadows in the image, they cannot divide the shadows and target, resulting in more miss detection. The above results validate that the proposed method is feasible for underwater object detection by the sonar image.

V. CONCLUSION

Accurate detection of the underwater sonar image is an essential and challenging task. In this paper, we proposed novel neural network MSMC-CNNs for detection of underwater sonar image. MSMC-CNNs consist of encoder network and decoder network. To enhance the feature extraction capability of the encoder network, the multi-scale multi-column CNNs architecture is used to extract the features of sonar image. Since the encoder network decomposes the detailed features of the sonar image, the bicubic linear interpolation algorithm is used in the decoder network to restore the size and resolution of the image. Since there are few samples of the sonar image dataset, we propose a progressive fine-tuning transfer learning method to training MSMC-CNNs, which can effectively improve the training efficiency of the model. At the same time for verify the effectiveness of the proposed method that MSMC-CNNs are used to detect sonar images with different PSNR noises, and results show that MSMC-CNNs have strong robustness. In future work, we will concentrate on how to reduce its number of iterations and shorten its running time. In addition, we will consider how to compress network model parameters and apply them to the practical application of underwater portable devices.

REFERENCES

[1] Kocak, D.M. & Dalgleish, Fraser & Caimi, Frank & Schechner, Yoav. “A Focus on Recent Developments and Trends in Underwater Imaging,” Marine Technology Society Journal, vol.42, pp.52-67, March 2008. doi:10.4031/002533208786861209

[2] O’Byrne M, Ghosh B, Schoefs F and Pakrashi V. “Image-based Damage Assessment for Underwater Inspections: A Primer – From Theory to Implementation,” Taylor and Francis, August 2018.

[3] Guang, Y., Zheng, Y., Wang, S.T., Xiang, X., & Yu, Y.T. “Review on Detection and Localization of Underwater Target,” Applied Mechanics and Materials, vol. 170-173, pp. 2127-2131, May 2012.

[4] Jianguo, L., Weidong, L., Li-E, G.., “Detection and localization of underwater targets based on monocular vision,” 2017 2nd International Conference on Advanced Robotics and Mechatronics (ICARM 2017). doi: 10.1109/ICARM.2017.8273142

[5] Ye, X. F. , Zhang, Z. H. , Liu, P. X. , & Guan, H. L., “Sonar image segmentation based on GMRF and Level-Set models,” Ocean Engineering, vol. 37(10), pp. 891-901 July 2010.

[6] Wang, L. Y., Li, M., & Gong, Z. B., “On sonar image processing techniques for detection and localization of underwater objects,” Applied Mechanics and Materials, vol.236-237, pp.509-514. November 2012.

[7] Mignotte, M., Collet, C., Perez, P., & Bouthemy, P., “Sonar image segmentation using an unsupervised hierarchical MRF model,” IEEE Transactions on Image Processing, vol.9 (7), pp.1216-1231. October 2000.

[8] Ma, W., Yang, H., Wu, Y., et al., “Change Detection Based on Multi-Grained Cascade Forest and Multi-Scale Fusion for SAR Images,” Remote Sensing, vol.11 (142), pp.143-151. January 2019

[9] Xiufen, Y. E., Zhang, Y., University, H. E., “Unsupervised sonar image segmentation method based on Markova Random Field,” Journal of Harbin Engineering University, vol.36 (4), pp.516-521. April 2015.

[10] Celik, T., & Tjahjadi, T., “A novel method for sidescan sonar image segmentation,” IEEE Journal of Oceanic Engineering, vol.36 (2), pp.186-194. May 2011.

[11] Song, S. M., Si, B. L., et al., “Label field initialization for MRF-based sonar image segmentation by selective auto encoding,” Ocean IEEE, vol.12(6), pp.16-31. February 2016.

[12] Cheng Shi, Chi-Man Pun, “Adaptive Multi-scale Deep Neural Networks with Perceptual Loss for Panchromatic and Multispectral Images Classification,” Information Sciences, vol.24 (3), pp.18-28. March 2019, doi: 10.1016/j.ins.2019.03.055.

[13] Tao, W., Ping, X., Liu, X., & Lei, B., “TS-MRF sonar image segmentation based on the levels feature information,” MIPPR: Multispectral Image Acquisition, Processing & Analysis, vol.32 (6), pp.13-22. December 2015.

[14] Huo, G., Yang, Simon X., Li, Qingwu, & Zhou, Yan., “A robust and fast method for sides-scan sonar image segmentation using nonlocal de-speckling and active contour model,” IEEE Transactions on Cybernetics, vol.47(4), pp.855-872. April 2016.

[15] Wu, J. P., Guo, Hai Tao., “Sonar image segmentation based on an improved selection of initial contour of active contour model,” Applied Mechanics & Materials, Vol.709, pp.447-450, December 2015. doi:10.4028/www.scientific.net/AMM.709.447

[16] A. Zare, N. Young, D. Suen., et al., “Possibilistic fuzzy local information C-Means for sonar image segmentation,” 2017 IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, 2017, pp. 1-8. doi: 10.1109/SSCI.2017.8285358

[17] Tim, S., Sunehag, P., “Induced graph semantics: another look at the hammersley-clifford theorem,” International Workshop for Bayesian Inference and Maximum Entropy methods in science and engineering. November 2007.

[18] Xie, M., Gao, J., Zhu, C., & Zhou, Y., “A modified method for MRF segmentation and bias correction of MR image with intensity inhomogeneity,” Medical & Biological Engineering & Computing, vol.53 (1), pp.23-35. October 2015.

[19] Da, Y. "An Efficient and Reliable Segmentation Method Based on Active Contour Model," IEEE/RSJ International Conference on Intelligent Robots & Systems, vol.12 (7), pp.13-21. January 2015. doi:10.1109/IROS.2006.282393

[20] Mishra, A., & Wong, A., "KPCA: A Kernel-Based Parametric Active Contour Method for Fast Image Segmentation," IEEE Signal Processing Letters, vol.17 (3), pp.312-315. April 2010.

[21] Sumengen, B., & Manjunath, B. S., "Graph partitioning active contours (GPAC) for image segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.28 (4), pp.509-521. October 2018.

[22] Huo, G., Yang, S. X., Li, Q., & Zhou, Y. A., "robust and fast method for sidescan sonar image segmentation using nonlocal despeckling and active contour model," IEEE Transactions on Cybernetics, vol.1-18, pp.855-872. April 2016.

[23] Sang, E., Shen, Z., Chang, F., & Li, Y., "Sonar image segmentation based on implicit active contours," IEEE International Conference on Intelligent Computing & Intelligent Systems, vol.53 (1), pp.23-35. November 2009.

[24] Lv Z Y, Liu T F, Atli Benediktsson J, et al. "Multi-Scale Object Histogram Distance for LCCD Using Bi-Temporal Very-High-Resolution Remote Sensing Images," Remote Sensing, vol.10 (11), pp. 25-36. November 2018. doi: 10.3390/rs10111809.

[25] Lv Z Y, Liu T F, Zhang P, et al. "Novel Adaptive Histogram Trend Similarity Approach for Land Cover Change Detection by Using Bitemporal Very-High-Resolution Remote Sensing Images," IEEE Transactions on Geoscience and Remote Sensing, vol.11(10), pp.31-46. August 2019. doi:

[26] Long, J., Shelhamer, E., & Darrell, T., "Fully convolutional networks for semantic segmentation," IEEE Transactions on Pattern Analysis & Machine Intelligence, vol.39 (4), pp.640-651. October 2015. doi:10.1109/TPAMI.2016.2572683

[27] Badrinarayanan, V., Kendall, Alex, & Cipolla, Roberto. "SegNet: A deep convolutional encoder-decoder architecture for scene segmentation," IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. (12), pp.2481-2495. December 2017.

[28] Zhang, P., Ke, Y., Zhang, Z., Wang, M., Li, P., & Zhang, S., "Urban Land Use and Land Cover Classification Using Novel Deep Learning Models Based on High Spatial Resolution Satellite Imagery," Sensors, vol.18, pp.3717-3722. June 2018.

[29] Deng Z, Sun H, Zhou S, et al. "Multi-scale object detection in remote sensing imagery with convolutional neural networks," ISPRS journal of photogrammetry and remote sensing, vol.145, pp. 3-22. May 2017.

[30] Tingting Ji., Guoyu Wang., Xiaolong Cheng., Guangrong Ji., & Tianhong Ya., "A fourth order P-Laplace underwater image restoration method based on MSG," Oceans,pp.1-4 November 2014.

[31] Moniruzzaman M, Islam S M S, Bennamoun M, et al., "Deep Learning on Underwater Marine Object Detection: A Survey, vol.23" pp.150-160. November 2017. doi:10.1007/978-3-319-70353-4_13

[32] Ronneberger, O., Fischer, P., & Brox, T., "U-Net: convolutional networks for biomedical image segmentation," vol.21 (2), pp.40-51. November 2015.doi.10.1007/978-3-319-24574-4_28

[33] Krizhevsky, A., Sutskever, I., Hinton, G., "ImageNet Classification with Deep Convolutional Neural Networks," Advances in neural information processing systems, vol.25 (18). January 2012.

[34] Szegedy C, Liu W, Jia Y, et al., "Going deeper with convolutions," vol. 1(1), pp. 1-9. March 2014.doi: 10.1109/CVPR.2015.7298594.

[35] Simonyan, Karen, and A. Zisserman., "Very Deep Convolutional Networks for Large-Scale Image Recognition," Computer Science, vol.1 (1), pp.18-24. September 2014.

[36] Zhang, Lu, M. Shi, and Q. Chen., "Crowd counting via scale-adaptive convolutional neural network," IEEE Computer Society, vol.21 (13), pp.18-36. November 2017.

[37] Keys R., "Cubic convolution interpolation for digital image processing," IEEE Trans on Acoustics, Speech, and Signal Processing, vol.29 (6), pp.1153-1160.

[38] O'Byrne M, et al., "Semantic Segmentation of Underwater Imagery Using Deep Networks Trained on Synthetic Imagery.," Journal of Marine Science and Engineering., vol.6.,pp.93-102.December 2018.