IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 29, 2019, accepted August 27, 2019, date of publication September 2, 2019, date of current version September 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2939005

Accurate Underwater ATR in Forward-Looking Sonar Imagery Using Deep Convolutional Neural Networks

LEILEI JIN, HONG LIANG, AND CHANGSHENG YANG, (Member, IEEE)
School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China

Corresponding author: Hong Liang (lianghong@nwpu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61971354, Grant 61379007, and Grant 61771398.

ABSTRACT Underwater automatic target recognition (ATR) is a challenging task for marine robots due to the complex environment. The existing recognition methods basically use hand-crafted features and classifiers to recognize targets, which are difficult to achieve ideal recognition accuracy. In this paper, we proposed a novel method to realize accurate multiclass underwater ATR by using forward-looking sonar—Echoscope and deep convolutional neural networks (DCNNs). A complete recognition process from data preprocessing to network training and image recognition was realized. Firstly, we established a real, measured Echoscope sonar image dataset. Inspired by the human visual attention mechanism, the suspected target region was extracted via the graph-based manifold ranking method in image preprocessing. Secondly, an end-to-end DCNNs model, named EchoNet, was designed for Echoscope sonar image feature extraction and recognition. Finally, a network training method based on transfer learning was developed to solve the problem of insufficient training data, and mini-batch gradient descent was used for network optimization. Experimental results demonstrated that our method can implement efficiently, and the recognition accuracy on a nine-class underwater ATR task reached 97.3%, outperforming traditional feature-based methods. The proposed method is expected to be a potential novel technology for the intelligent perception of autonomous underwater vehicles.

INDEX TERMS Automatic target recognition (ATR), forward-looking sonar, sonar image processing, deep convolutional neural networks (DCNNs), transfer learning.

I. INTRODUCTION

Accurate target recognition is a crucial basis for underwater exploration and ocean development. As early as the 1960s, underwater target recognition has been highly valued by naval departments [1]. In recent decades, with the recovery of the global economy, the demand for underwater target recognition technology has become increasingly urgent in civil and commercial fields [2]–[4], including tracking and protection of endangered aquatic organisms, salvage and rescue, aquaculture, and underwater archaeology. However, due to the changeable environment and limitations on the sensing of the marine, accurate multiclass underwater automatic target recognition (ATR) has not been commendably solved.

With the engineering application of sonar imaging system, especially side-scan sonar and synthetic aperture sonar (SAS), several works devoted to recognizing underwater targets through sonar images [5]–[8], as images are easier to reflect the underwater scenes. In general, we choose the type of imaging sonar according to the practical application. Side-scan sonar is a system used for generating an image of the sea bottom area, which is not suitable for the identification of floating objects, nor for the real-time recognition tasks. The main difficulty with SAS relates to micro navigation and platform trajectory estimation, the high requirement on the platform movement limits its application scenario. Real-time imaging sonar Echoscope is one of the most important innovations in underwater observation in recent years. In fact, the invention of a more delicate and efficient imaging system facilitates the generation of high-resolution images and the design of appropriate technologies to automatically

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

understand underwater scenes [9]. In this paper, we tackle the challenging task of underwater target recognition by utilizing Echoscope sonar.

Previously, hand-crafted features are employed for visual object classification tasks, such as Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradient (HOG) [10], and Fisher Vector. These hand-crafted features encode shape, texture, and color information followed by different classifiers such as the works of [5]–[7], [11], [12]. The features can perform well for specific data and tasks, but most have limited generalization capability, and feature extraction requires expertise and a lot of trial and error. Underwater targets are diverse in terms of size, shape, texture, and background even for the same class, which makes it extremely difficult for conventional methods to accurately perform multiclass target recognition tasks.

Deep learning is a research hotspot in the field of machine learning in recent years [13]–[15], which attempts to extract high-level features from mass data automatically through the learning process. In the 1990s, Lecun *et al.* [16] established the modern structure of convolutional neural networks called LeNet-5 to classify handwritten numbers. Since 2006, many studies focus on the improvement of CNNs. In 2012, Krizhevsky *et al.* [17] proposed a classic structure of deep convolutional neural networks (DCNNs) and shown its outstanding performance in the ImageNet 2012 Large Scale Visual Recognition Competition. Considering the excellent performance in diverse optical image recognition applications [18], [19], we expect DCNNs to solve the underwater target recognition problem as well. It's worth noting that the different imaging mechanisms result in the distinct characteristics between sonar image and optical image. Sonar image exhibits characteristics including incomplete boundary and weak contrast [20], while the optical image shows evident details that can be easily recognized by the human visual system.

The research of deep learning for underwater target recognition is far from enough. One reason is that the success of DCNNs in image processing depends on the use of large amounts of training data, while the costly and time-consuming underwater experiment results in the lack of sample images for DCNNs training. It has been observed that DCNNs is prone to overfitting in small samples, that is, high training accuracy and poor test results. Wang *et al.* [21] attempted to overcome network overfitting by a very complicated weight initialization method, specifically, they used the generated weights of the deep belief network (DBN) [22] to adaptively replace the random weights of the DCNNs. However, the classification accuracy of this method is only 85.5% on a six-class target recognition task.

In this paper, we propose a novel underwater ATR method based on deep learning to improve the accuracy of underwater multiclass target recognition tasks. A complete process from data preprocessing to network training and image recognition is realized. An end-to-end DCNNs model named EchoNet is designed, and the corresponding training strategy is developed to solve the problem of insufficient training data. Features are learned from the data itself, so domain knowledge of sonar image feature extraction is not needed. Besides, we construct an Echoscope sonar image dataset, which is available to the vision research community and can be used to test sonar image recognition algorithms. To our knowledge, this work is the first to consider using the Echoscope sonar image and DCNNs for underwater ATR. Experimental results show that our method can greatly improve the accuracy of underwater multiclass ATR compared to traditional feature-based classifiers.

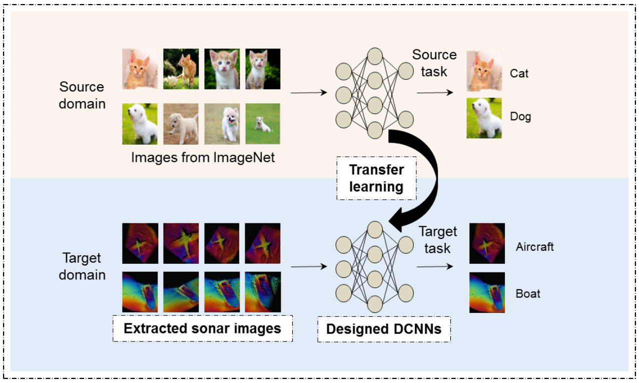


FIGURE 1. Overview of the proposed underwater target recognition method. Accurate multiclass target recognition is realized by using the Echoscope sonar image and deep convolutional neural networks, while network training method based on transfer learning is developed to solve the problem of insufficient sonar image data.

The rest of this paper is organized as follows: Section II describes the details of the EchoNet for sonar image recognition; in Section III and Section IV, experimental results are presented and discussed; finally, concluding remarks and directions for future research are provided in Section V.

II. ACCURATE TARGET RECOGNITION

The framework of the proposed accurate underwater ATR method is shown in Fig. 1. The main part is to train our designed DCNNs applied for underwater target recognition using Echoscope sonar images (the bottom half of Fig. 1). We will discuss the three key points in detail below, including sonar image preparation, DCNNs model design, and network training based on transfer learning. The top half of Fig. 1 is to train a DCNNs using standard supervised learning with a large image dataset (e.g. ImageNet [23]), and the trained model plays the role of basic network in transfer learning. We can view the trained basic network as an analog to the prior knowledge a human learns from previous visual experiences, which is conducive to the learning task of the target network [24].

A. SONAR IMAGE PREPARATION

The Echoscope imaging sonar, developed by Coda Octopus, is the world's highest-resolution commercial real-time sonar. With a horizontal and vertical resolution of 0.4°, Echoscope can generate high definition images with a maximum range

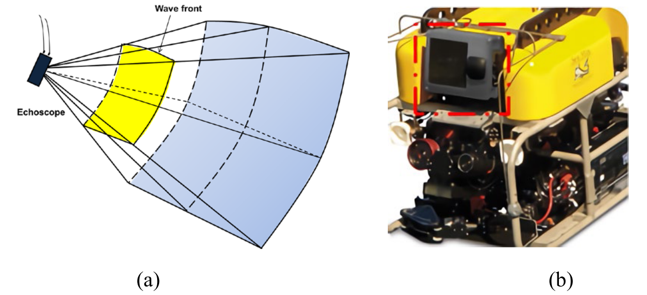


FIGURE 2. Overview of the Echoscope: (a) Beam energy distribution of the phased array imaging sonar system; (b) An underwater ROV equipped with an Echoscope (marked with a rectangle).

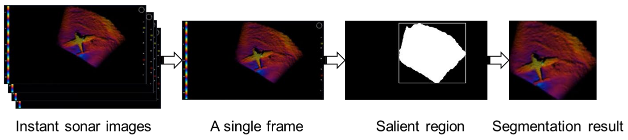


FIGURE 3. Extraction of the interested image region. Underwater scenes are first recorded by E-UIS, saliency detection method via graph-based manifold ranking is then applied to each single frame to get the interested image region. These segmentation results are finally collected for training and testing the EchoNet.

of 100 m. The imaging sonar uses phased-array techniques to generate more than 16,000 discrete beams each time, and the beams yield a range measurement obtaining data points with known position and intensity (x, y, z, i), with which to generate a complete sonar image [25]. With a ping rate up to 12 Hz, Echoscope can provide successive image frames similar to video images to monitor targets that are both moving and stationary. Fig. 2 shows a schematic diagram of the Echoscope imaging principle, and an Echoscope mounted on a remotely operated vehicle (ROV).

High-resolution sonar image provides an intuitive view of the underwater scene that makes automatic identification of suspect targets possible. The sonar data generated by Echoscope enables to create continuous pictures of the underwater scene, which can be viewed and recorded by the Echoscope-Underwater Inspection System (E-UIS). Yang *et al.* [26] proposed a saliency detection method for optical images via graph-based manifold ranking. Inspired by their method, we extract interested image regions from records of E-UIS; the whole process is shown in Fig. 3.

The preprocessing can eliminate the redundant background and make it easier for the ATR task. All sonar data used in this work were acquired at sea by Echoscope, and objects would be recognized from the segmentation results. With RGB channels normalized separately by min-max normalization, pixel values of the scene-level sonar images are normalized to [0, 1], to reduce the undesirable influence on target recognition. More details about the Echoscope sonar image dataset are described in Section III.

B. ECHONET ARCHITECTURE

DCNNs model is a deep learning architecture inspired by biological visual cognitive mechanisms [27]. The architecture of EchoNet, as depicted in Fig. 4, consists of 5 convolutional layers and 2 fully connected layers, whose structure is based on the AlexNet [17]. Starting with raw input, the output of the last fully connected layer is fed to Softmax to generate a probability distribution over the n class labels, and n corresponds to the category number of underwater targets of interest. AVE pooling is used rather than the commonly used MAX pooling after the first convolutional layer, since average pooling should be more robust when dealing with the speckle-like nature of sonar image. To reduce the number of connection parameters, we use only two fully connected layers, and dropout is only used in FC1.

1) CONVOLUTIONAL LAYER

Given an input feature map x_i and a convolution filter k_{ij}, the output feature map y_j may combine convolutions with multiple input maps and can be expressed as

$$y_j = f\left(\sum_i {x_i}^* k_{ij} + b_j\right) \tag{1}$$

where * is the two-dimensional discrete convolution operator and b_j is an additive bias. The activation function $f(x) = \max(0, x)$, called Rectified Linear Units (ReLU) is applied to each convolutional layer and FC1 layer. Both convolution filter weights and biases are model parameters that need to be learned.

2) POOLING LAYER

AVE pooling operation is used to compute the average value over a pixel's neighborhood region, while MAX pooling operation is to compute the maximum value. The pooling layer can reduce the dimension of the feature maps and introduce small translation invariance.

3) FULLY CONNECTED LAYERS

Acts as a classifier in the whole network. The fully connected layers are computed as $Y_6 = f(W_6 Y_5 + B_6)$ and $Y_7 = \psi(W_7 Y_6 + B_7)$, where W_l and B_l are matrixes of the trainable parameters, $\psi(X)[i] = e^{X[i]} \Big/ \sum_j e^{X[j]}$ is the Softmax function.

C. NETWORK TRAINING WITH TRANSFER LEARNING

The training process of EchoNet seeks to minimize the classification error on the training dataset, in other words, model parameters θ are learned to minimize the cross-entropy cost function:

$$J(\theta) = -\frac{1}{M}\left[\sum_{m=1}^{M}\sum_{c=1}^{n} 1\left\{h^{(m)} = c\right\} \cdot \log p_c(\theta; x^{(m)})\right] + \lambda \cdot R(\theta) \tag{2}$$

where $1\{\cdot\}$ is an indicator function defined as $1\{\text{true}\} = 1$ and $1\{\text{false}\} = 0$, $\left\{\left(x^{(1)}, h^{(1)}\right), \ldots, \left(x^{(M)}, h^{(M)}\right)\right\}$ is a training set of M labeled examples, labels $h^{(m)} \in \{1, 2, \ldots, n\}$,$p_c(\theta; x^{(m)})$ is the estimated probability of the c-th class, and $\lambda \cdot R(\theta)$ is a weight decay term (L2 regularization in this paper).

Normally, the training process of DCNNs goes like this: first, the learning network is randomly initialized and fed

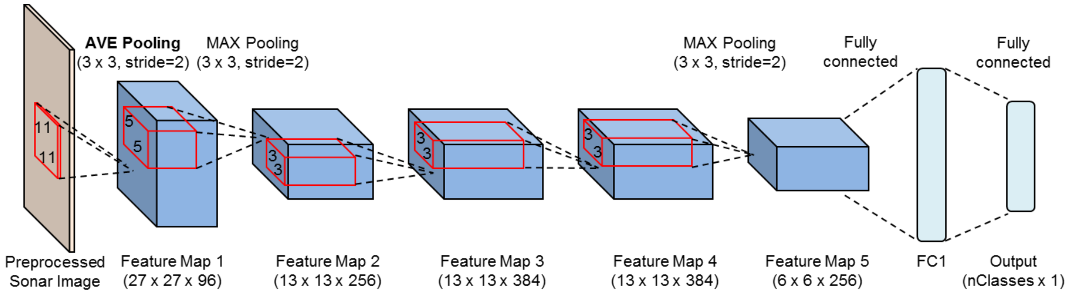

AVE Pooling (3 x 3, stride=2)
MAX Pooling (3 x 3, stride=2)
MAX Pooling (3 x 3, stride=2)
Fully connected
Fully connected
11
11
5
5
3
3
3
3
3
3
Preprocessed Sonar Image
Feature Map 1 (27 x 27 x 96)
Feature Map 2 (13 x 13 x 256)
Feature Map 3 (13 x 13 x 384)
Feature Map 4 (13 x 13 x 384)
Feature Map 5 (6 x 6 x 256)
FC1
Output (nClasses x 1)

FIGURE 4. The architecture of the EchoNet. The red cuboid represents the convolution filter, and the number next to it specifies the filter size; the blue cuboids are feature maps corresponding to the output of the first five layers, and the light green rectangles represent the fully connected layers.

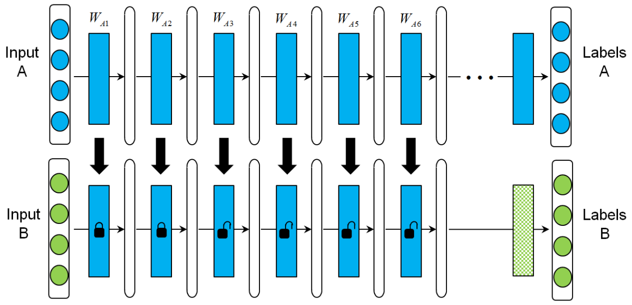
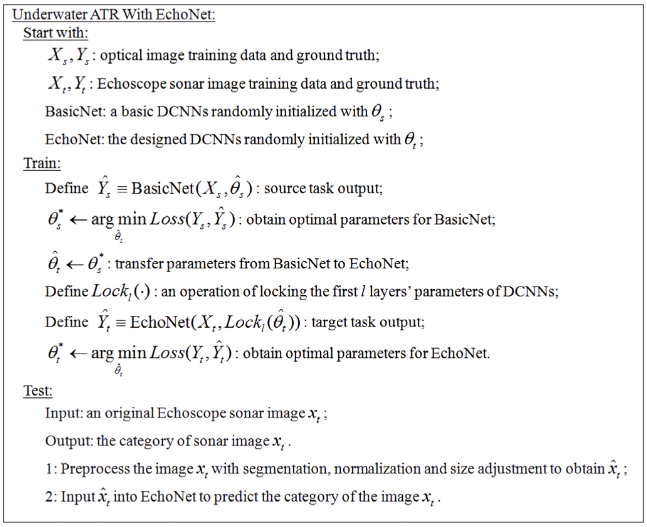


FIGURE 5. The training method of the EchoNet. The labeled rectangles (e.g. W_{A1}) represent weights learned for each layer, and color indicates which dataset the weights were originally trained on. The ellipsoids between rectangles represent the feature maps at each layer.



```
Underwater ATR With EchoNet:
  Start with:
    $X_s, Y_s$: optical image training data and ground truth;
    $X_t, Y_t$: Echoscope sonar image training data and ground truth;
    BasicNet: a basic DCNNs randomly initialized with $\theta_s$;
    EchoNet: the designed DCNNs randomly initialized with $\theta_t$;
  Train:
    Define $\hat{Y}_s \equiv \mathrm{BasicNet}(X_s, \hat{\theta}_s)$: source task output;
    $\theta_s^* \leftarrow \arg\min_{\hat{\theta}_s} Loss(Y_s, \hat{Y}_s)$: obtain optimal parameters for BasicNet;
    $\hat{\theta}_t \leftarrow \theta_s^*$: transfer parameters from BasicNet to EchoNet;
    Define $Lock_l(\cdot)$: an operation of locking the first $l$ layers' parameters of DCNNs;
    Define $\hat{Y}_t \equiv \mathrm{EchoNet}(X_t, Lock_l(\hat{\theta}_t))$: target task output;
    $\theta_t^* \leftarrow \arg\min_{\hat{\theta}_t} Loss(Y_t, \hat{Y}_t)$: obtain optimal parameters for EchoNet.
  Test:
    Input: an original Echoscope sonar image $x_t$;
    Output: the category of sonar image $x_t$.
    1: Preprocess the image $x_t$ with segmentation, normalization and size adjustment to obtain $\hat{x}_t$;
    2: Input $\hat{x}_t$ into EchoNet to predict the category of the image $x_t$.
```

FIGURE 6. Procedure for underwater ATR with EchoNet.

with a large number of images, then costs are computed by forward-propagation, and finally backpropagation algorithm is used to tune the network parameters.

Although DCNNs take advantage of weight sharing and local connection, the networks still have millions of weights that need to learn, which determines that the networks should be trained on a large dataset. Due to the limited amount of sonar images at hand, a network training method based on transfer learning is developed to avoid overfitting of the EchoNet, as depicted in Fig. 5.

Transfer learning is a tool in machine learning to solve the basic problem of insufficient training data [28], [29], which tries to transfer the knowledge from the source domain to the target domain. We present a flexible transfer learning method.

Firstly, a basic network is trained using standard supervised learning with a large number of labeled images (top row). ImageNet provides such an ideal large image dataset, which contains over 10 million optical images with each image labeled. Then, pre-trained parameters ($W_{A1} \sim W_{A6}$) of the basic network are transferred to the EchoNet (bottom row), and the last layer of EchoNet is initialized randomly from Gaussian distribution. Finally, the EchoNet is fine-tuned on the Echoscope sonar image dataset. Particularly, the first 2 layers are locked and the remaining layers are allowed to learn during the fine-tuning process. Mini-batch gradient descent (MBGD) is used to update the model weight w as follows:

$$\begin{aligned} v_{t+1} &= \eta \cdot v_t - \alpha \cdot \lambda \cdot w_t - \alpha \cdot \left[\frac{\partial J}{\partial w} \Big|_{w_t} \right]_{B_t} \\ w_{t+1} &= w_t + v_{t+1} \end{aligned} \tag{3}$$

where t is the iteration index,η is the momentum variable,v_t is the previous weight update,α is the learning rate,λ is the weight decay variable, and $\left[\frac{\partial J}{\partial w} \big|_{w_t} \right]_{B_t}$ is the average over the t-th batch B_t of the derivative of the cost function with respect to w, evaluated at w_t.

Most of the transfer learning methods aim to cope with the same data type—optical image, while our work tries to transfer image representations from the optical images to sonar images.

From what have been described above, we outline our underwater ATR method in Fig. 6.

III. EXPERIMENTS AND RESULTS

A. ECHOSCOPE SONAR IMAGE DATASET

We evaluate the effectiveness of our underwater target recognition method on a real, measured sonar image dataset. Sea experiments have been conducted by Coda Octopus using the

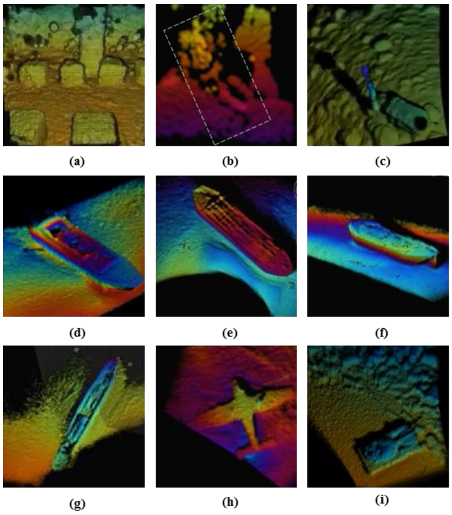


FIGURE 7. High-resolution sonar images of interested underwater targets. From top left to bottom right: (a) Cornerstone; (b) Diver (marked with a rectangle dotted line); (c) ROV; (d) Sunken barge1; (e) Sunken barge2; (f) Sunken barge3; (g) Shipwreck; (h) Sunken plane; (i) Sunken military tank.

TABLE 1. Details of the Echoscope sonar image dataset.

Target	Sea experiments		Number of
	Year	Location	extracted images
Cornerstone	2009	USA	335
Diver	2010	UK	287
ROV	2011	UK	311
Sunken barge1	2012	UK	223
Sunken barge2	2012	UK	394
Sunken barge3	2012	UK	195
Shipwreck	2012	UK	370
Sunken plane	2012	UK	583
Sunken military tank	2012	UK	217

high-resolution imaging sonar Echoscope between 2009 and 2012 in various geographical locations with diverse environmental conditions. In each experiment, a specific object, including cornerstone, diver, ROV, sunken barge, shipwreck, sunken plane, and sunken military tank was investigated as an interested underwater target, as shown in Fig. 7.

We collected these experimental results and extracted sonar images to establish a high-resolution sonar image dataset. Table 1 summarizes the detail of each sea experiment and the number of scene-level sonar images we obtained.

There are totally 2,915 verified target images of 9 classes, which have been preprocessed and manually labeled by us. The total sonar images are divided into 3 subsets: 900 images (100 images of each the 9 classes) for training, 450 images (50 images of each the 9 classes) for validation, and the rest 1565 for test. The original size of these false-color sonar images various from 150 × 150 × 3 to 240 × 240 × 3 pixels. Considering the average size of the images, all the images are resized to 227 × 227 × 3 when fed to the EchoNet. The images shown in Fig. 7 have been resized to the same size. Sonar images in the dataset vary significantly not only in target position, orientation and scale within each class, but also in colors and textures.

B. TRAINING AND TESTING THE ECHONET

In this subsection, we first describe training details and then show experimental results of the proposed target recognition method on the Echoscope sonar image dataset. We implement DCNNs models based on the efficient and practical open-source Caffe framework [30], the EchoNet is constructed as described in subsection II. B.

The training process of EchoNet has been introduced in detail in subsection II. C. AlexNet architecture is employed as the basic network, which is trained on a dataset obtained from the ImageNet-2012, a subset with 1000 optical images in each of 1000 categories, and obtains a final top-1 error on the validation set of 42.6%. Then, parameters ($W_{A1} \sim W_{A6}$) of the AlexNet are transferred to the EchoNet. We lock the first 2 layers of EchoNet and train the remaining layers on the Echoscope sonar image dataset. MBGD with a batch size of 45 is used to train the EchoNet by back-propagating the classification error. The learning rate is set to 0.001, and decreased by a factor of 2 every 200 iterations, in conjunction with weight decay of 0.005 and momentum of 0.9. The total number of training iterations is set to 1000, i.e., 50 epochs.

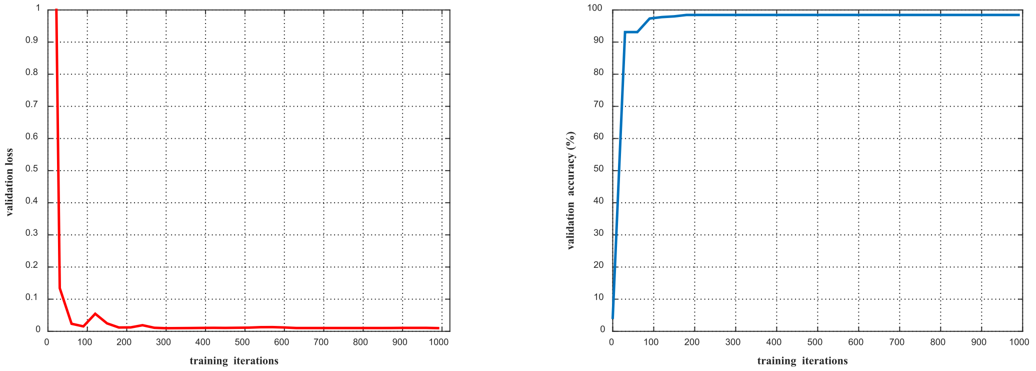
To illustrate the results precisely, we repeated the EchoNet training and testing experiment five times (each experiment takes about 1 hour), and got an average testing accuracy of 96.4% on the nine-class underwater target recognition task, with the highest accuracy reached 97.3% and the lowest accuracy reached 94.4%. Validation loss vs. training iterations curve and validation accuracy vs. training iterations curve that in one of the EchoNet training experiments are shown in Fig. 8. Overfitting does not occur despite the huge number of network parameters, which can be attributed to the special training method developed, as well as the reduction of fully-connection layers. Taking the best classification results as an example, the confusion matrix is given in Table 2, shows that only a few highly similar samples are misclassified.

We have noticed some other recent works about underwater target recognition based on sonar image. For example, Matheus *et al.* [12] proposed an object classification method by using forward-looking sonar and the best performance reached 93.6% on a five-class target recognition task. Myers *et al.* [6] presented a method called normalized shadow-echo matching (NSEM) and reached 94.8% accuracy on a four-class target recognition task using the SAS image. Zhu *et al.* [5] presented a classification system based on KELM and PCA to classify the underwater target image collected by side-scan sonar and got 94.2% accuracy of distinguishing between metal cylinders and rocks. These results further confirm the method we proposed is an effective way for accurate underwater ATR tasks.

C. METHODS COMPARISON

We carry out extensive experiments to evaluate the efficacy of the proposed EchoNet with four classic classifiers in pattern recognition and two state-of-the-art deep neural networks.

The four classic classifiers, including the k-nearest neighbor (KNN) classifier, multi-layer perceptron (MLP), nearest



(a)

(b)

FIGURE 8. Training curves of EchoNet: (a) Validation loss vs. training iterations. Loss declines sharply and tends to be 0 after about 300 iterations; (b) Validation accuracy vs. training iterations. Accuracy rises quickly at the first 100 iterations and tends to be stable after about 200 iterations.

TABLE 2. Confusion matrix of 9-class recognition results.

Category	Stone	Diver	ROV	Barge1	Barge2	Barge3	Ship	Plane	Tank	Accuracy
Stone	185	0	0	0	0	0	0	0	0	100.00%
Diver	0	137	0	0	0	0	0	0	0	100.00%
ROV	3	0	158	0	0	0	0	0	0	98.10%
Barge1	0	0	0	73	0	0	0	0	0	100.00%
Barge2	0	0	0	2	242	0	0	0	0	99.20%
Barge3	0	0	0	0	0	45	0	0	0	100.00%
Ship	0	0	0	0	0	0	220	0	0	100.00%
Plane	0	0	0	9	29	0	0	395	0	91.20%
Tank	0	0	0	0	0	0	0	0	67	100.00%
Accuracy					97.3%					

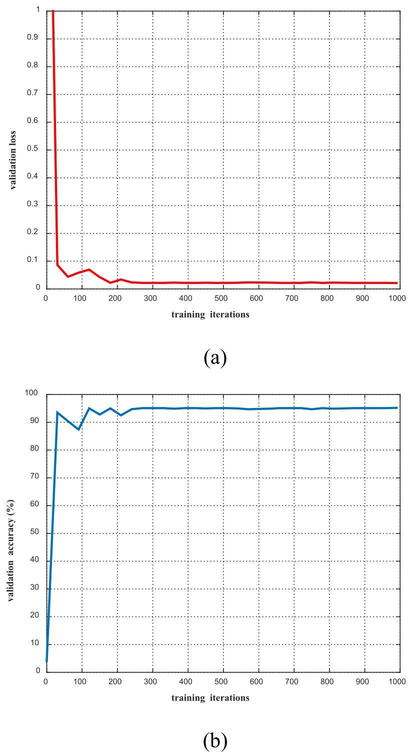
neighbor (NN) classifier and support vector machine (SVM), are mainly implemented by using the Scikit-learn machine learning module. The sonar image dataset is divided into 2 subsets: 1350 images (150 images of each the 9 classes) for training, and the rest 1565 images for the test. We separately use two kinds of features as the input of the classifiers, namely, the original pixel value of the image and the HOG feature.

Firstly, we define two preprocessing functions: The first one is to flatten an 227 × 227 × 3 image into a row of pixels. The second one is to extract the HOG feature from a resized 180 × 180 × 3 sonar image using ft.hog function (the image is divided into 15 × 15 blocks, each block contains 2 × 2 cells and each cell has 6 × 6 pixels). Then, we extract each image features and put them into arrays. Finally, the KNeighborsClassifier, MLPClassifier and SVC functions are applied to evaluate the data. For the KNN method, we change the number of neighbors and store the best result. In MLPClassifier, we set one hidden layer with 80 neurons and use stochastic gradient descent (SGD) to update the model weights. The learning rate is set to 0.1 and the maximum iteration is 1000. In SVC, the maximum iteration is 1000, and class weight is "balanced".

Using raw pixel values as input, the accuracy of the KNN classifier is 72.0% (with k = 10), and the accuracy of MLP is 89.3%. Using HOG features as input, the accuracy of the NN classifier is 91.4%. With a widely used baseline method HOG + SVM, we obtain an accuracy of 92.7%.

EchoNet is also compared with two well-known deep neural networks. We implement the AlexNet and GoogLeNet [31] on the Caffe framework, pre-trained models (code provided by Caffe) are employed and totally fine-tuned by using our Echoscope sonar image training dataset. We fine-tune the AlexNet with the same hyper-parameters as the EchoNet; for GoogLeNet, each iteration of MBGD used a batch size of 45, a momentum of 0.9, and a multiplicative weight decay of 0.005 per iteration. The learning rate started at 0.001, and the learning rate follows a polynomial decay with a power of 0.5. Learning stopped after 1200 iterations. After training, test data of the Echoscope sonar image dataset are feed to the fine-tuned AlexNet and GoogLeNet to test the networks. The experiment of each network is repeated five times, and the training curves of the AlexNet and GoogLeNet in one experiment are shown in Fig. 9 and Fig. 10 respectively.

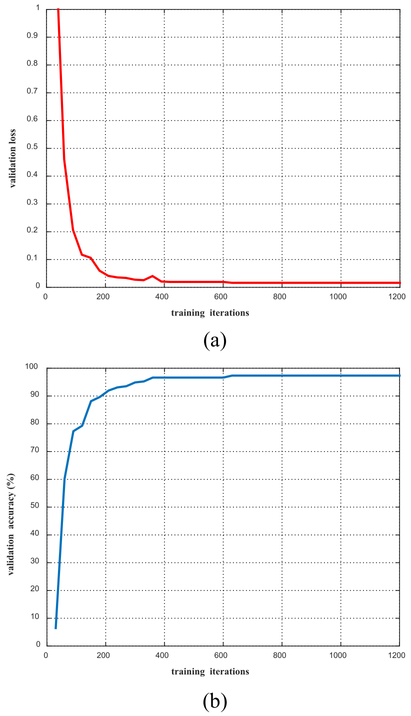
The experimental results of all methods are listed in Table 3, which shows the proposed EchoNet method outperforms all the comparison methods. Specifically, compared to the best baseline using traditional hand-crafted features (SVM_HOG), we achieve absolute increases of 4.6%. Compared to the deep-network-based methods, we achieve absolute increments of 3.2%, 0.3% respectively. Although



(a)

(b)

FIGURE 9. Training curves of the AlexNet: (a) Validation loss vs. training iterations; (b) Validation accuracy vs. training iterations.



(a)

(b)

FIGURE 10. Training curves of the GoogLeNet: (a) Validation loss vs. training iterations; (b) Validation accuracy vs. training iterations.

GoogLeNet achieves similar recognition accuracy as the EchoNet, training GoogLeNet takes ten times as long as the EchoNet. Overall, the DCNNs based methods are better than the traditional hand-crafted features based methods.

TABLE 3. The results of comparison methods on the sonar image dataset.

Method	Accuracy	Inference time per image (ms)
KNN_raw_pixel	72.0%	341.2
MLP_raw_pixel	89.3%	1.1
NN_HOG	91.4%	311.4
SVM_HOG	92.7%	133.7
AlexNet	94.1%	73.1
GoogLeNet	97.0%	186.9
EchoNet	97.3%	60.6

Besides the accuracy of underwater multiclass target recognition, the real-time requirement is also an essential point in ATR. Therefore, we also examined the efficiency of each method listed in Table 3.

The design of a classifier usually consists of two parts: training and test (inference). In practice, we care more about the inference efficiency of a classifier. In Table 3, we show the average inference time each recognition method takes on a single sonar image, which is calculated by averaging the total test time of the 1565 test images. The test platform is based on a dual-core Intel processor. It should be noted that since the first four methods are developed in Python, and the last three methods run in the Caffe, these runtimes cannot be compared with each other directly, but only generally reflect the efficiency of each method. The Caffe framework is mainly based on C++, which is several times more efficient than Python.

In terms of details, the time complexity of the KNN and NN algorithm increase with the increase of training sample size and feature dimension. A test sample must be compared with all the training samples, which leads to heavy distance calculation. For the MLP method, the shallow structure and straightforward computation make it very efficient. For the SVM_HOG method, independent feature extraction (~50 ms) is required before using SVM for image recognition, resulting in reduced recognition efficiency. For the last three DCNNs based methods, their forward pass floating-point operations (FLOPs) are about 720M, 1550M, and 700M respectively. Although each model takes hours to train, it does not spend much time on test due to the advantages of end-to-end model structure and efficient numerical computation. As we can see from Table 3, EchoNet takes only 60.6 ms to test one sonar image. As the maximum refresh rate of the Echoscope is 12 frames per second, EchoNet is fast enough to process each frame in real-time.

IV. DISCUSSION

In this section, we mainly discuss the impact of image noise, network architecture and training method on the recognition performance, and further analyze the reason for the success of transfer learning through parameter visualization.

A. EFFECT OF IMAGE NOISE ON RECOGNITION

In Fig. 7, we have shown some sonar images generated by the Echoscope during sea experiments, which are of good quality. However, the underwater environment can sometimes be so

TABLE 4. **Recognition results of The image datasets polluted by noise.**

Method	Accuracy	
	Noise variance=0.01	Noise variance=0.1
MLP_raw_pixel	62.6%	30.9%
SVM_HOG	80.8%	56.0%
AlexNet	91.1%	89.3%
EchoNet	94.5%	92.6%

severe that sonar images may be heavily polluted by noise. In this subsection, we would like to test the effect of image noise on recognition accuracy of our method. We generate two kinds of zero-mean white Gaussian noises with variances of 0.01 and 0.1 respectively. By artificially adding the noises to the normalized real-measured sonar images, we obtain two simulated sonar image dataset contaminated by noise. Fig. 11 shows several representative sonar images.

We employ EchoNet and some comparison methods to conduct experiments on the polluted image datasets, and show experimental results in Table 4. The experimental settings are the same as described in subsection III. C and each data is the average value of five experimental results.

It can be seen from Table 4 that EchoNet and AlexNet can still achieve acceptable recognition accuracy at a high noise level, while the traditional feature based methods are susceptible to noise and significantly reduce their accuracy. This further demonstrates the advantage of the DCNNs. Besides, with the increase of noise level, the recognition accuracy of each method decreases. We can consider using image noise reduction method in sonar image preprocessing to obtain better application effect.

B. AVE POOLING VS. MAX POOLING

In Section II, we have described the architecture of the designed DCNNs, whose first pooling layer used average rather than the commonly used maximum approach. We would like to test a network that has a similar structure as the EchoNet, except that the first pooling layer is changed to MAX pooling. The same training method and hyper-parameters are employed to conduct the experiments, and the results are shown in Fig. 12. For the MAX pooling approach, the average testing accuracy of the five experiments is 92.6%, with the maximum is 96.2%, and the minimum is 89.3%. By contrast, AVE pooling approach gets an average testing accuracy of 96.4%, with the maximum is 97.3% and the minimum is 94.4%. It appears that using AVE pooling at the first pooling layer is more effective for Echoscope sonar image processing in our method.

C. TRAINING WITH DIFFERENT LOCKED LAYERS

In our training method, the first 2 layers of the EchoNet are locked and the remaining layers are allowed to learn during the training process. Now, we use N_L to indicate the number of locked layers and discuss the impact of N_L value on network performance. The values of N_L are chosen from $\{0,1,\ldots,6\}$, and corresponding new networks are trained. Note that all layers can participate in training when $N_L = 0$, whereas only the last fully connected layer is allowed to learn when $N_L = 6$. For each N_L value, we conduct four experiments. Except for different training strategies, the structure of each network is exactly the same as mentioned in Section II. All experiments draw their training data and test data from the same dataset, as specified in Table 1.

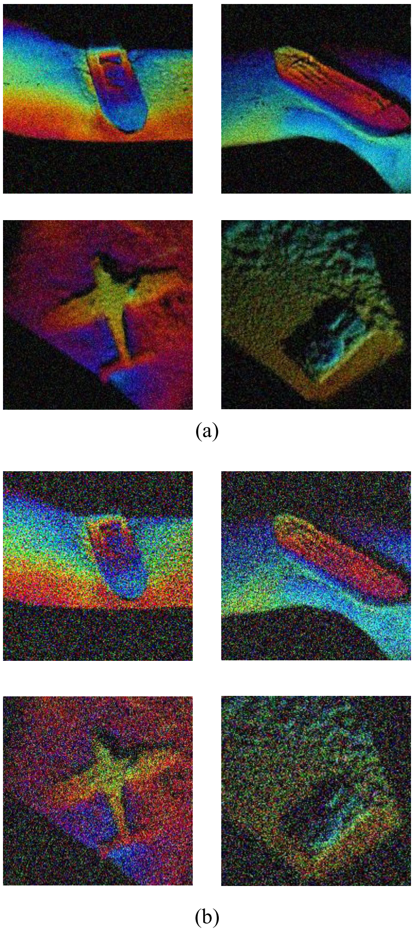


FIGURE 11. **Sonar images polluted by zero-mean white Gaussian noise: (a) Noise variance is 0.01; (b) Noise variance is 0.1.**

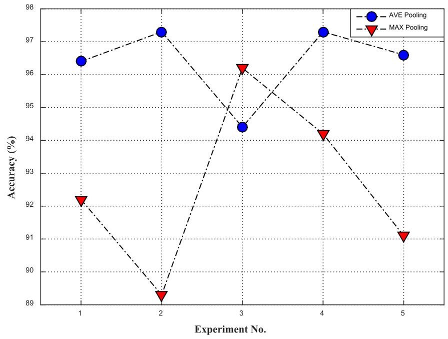


FIGURE 12. **Recognition results of AVE pooling approach and MAX pooling approach.**

Experimental results are shown in Fig. 13, we note that the average recognition accuracy changes with N_L. For our

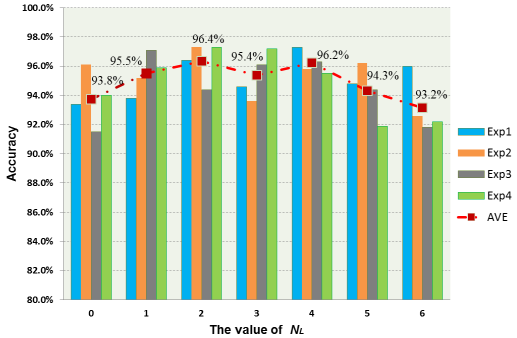
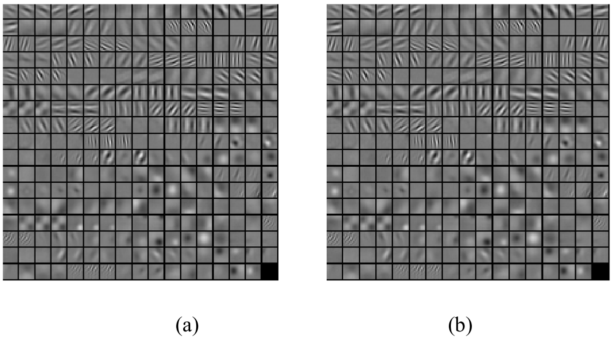


FIGURE 13. Experimental results of the EchoNet with different training strategies.



(a) (b)

FIGURE 14. Filter weights visualization of the first convolutional layers. (a) Filters only trained on optical images ($N_L = 2$); (b) Filters trained on optical images and fine-tuned by sonar images ($N_L = 0$).

designed DCNNs, one can see that $N_L = 2$ gives the best result, which suggests that completely fine-tuning the target network ($N_L = 0$) may not get the best performance. The choice of whether or not to lock the first l layers of the target network may depends on the size of the target dataset and the number of parameters in the first m layers [32].

D. VISULIZATION OF LEARNED FILTERS

Inspired by the parameter visualization method [33], we display the filter weights of the first convolutional layer respectively from the EchoNet with $N_L = 0$ and $N_L = 2$, as shown in Fig. 14.

Visually, there is little difference between the two sets of filters in Fig. 14. To describe the similarity of the two sets of weights quantitatively, the correlation coefficient of matrixes is introduced and calculated using the following formula:

$$r = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{\left(\sum_m \sum_n (A_{mn} - \bar{A})^2\right)\left(\sum_m \sum_n (B_{mn} - \bar{B})^2\right)}} \tag{4}$$

where A and B are two matrixes, $\bar{A}$ and $\bar{B}$ are corresponding mean values. The correlation coefficient $r \in [0, 1]$, the larger the r, the more similar the two matrixes are. For the two sets of filters, we put the parameters of each filter group into a large matrix respectively. The correlation coefficient of these two matrixes is 0.99, which is consistent with the results shown in Fig. 14.

Network parameters trained on optical images are similar to those used for sonar image recognition, which illustrates the generalization ability of DCNNs for image processing. The front layers of the DCNNs can be treated as a versatile feature extractor, so it is reasonable to transfer the knowledge of optical image recognition to sonar image recognition.

V. CONCLUSION

For improving the accuracy of underwater multiclass target recognition tasks, this paper proposes an ATR method in combining the forward-looking sonar image and deep convolutional neural networks. An end-to-end DCNNs model named EchoNet is designed and a corresponding training strategy is developed that could extract high-level features of sonar images automatically through the learning process, and perform target recognition. We also build a sonar image dataset contains a total of 2,915 sonar images, which can be used for testing sonar image recognition algorithms.

Through a series of experiments, the influences of network architecture and training method on the recognition performance are discussed, and the reason for the success of transfer learning is analyzed. Experimental results show that compared with traditional classifiers, the proposed method has high accuracy, good real-time performance and strong anti-noise ability. In particular, our method can achieve an accuracy of 97.3% on a nine-class underwater ATR task and surpasses four traditional classifiers and two deep neural networks.

Our research demonstrated the great prospect of using imaging sonar and DCNNs for underwater ATR, which is of great significance for underwater vehicles to perceive the ocean environment and navigate autonomously. It is believed that the deployment of DCNNs on unmanned platforms will not be a difficult problem in the future with the continuous optimization of network architecture and the development of hardware computing acceleration technology.

REFERENCES

[1] T. Fei, D. Kraus, and A. Zoubir, "Contributions to automatic target recognition systems for underwater mine classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 505–518, Jan. 2015. doi: 10.1109/TGRS.2014.2324971.

[2] S. Ntalampiras, "Hybrid framework for categorising sounds of Mysticete whales," *IET Signal Process.*, vol. 11, no. 4, pp. 349–355, Jun. 2017. doi: 10.1049/iet-spr.2015.0065.

[3] H. Qin, X. Li, J. Liang, Y. Peng, and C. Zhang, "DeepFish: Accurate underwater live fish recognition with a deep architecture," *Neurocomputing*, vol. 187, pp. 49–58, Apr. 2016. doi: 10.1016/j.neucom.2015.10.122.

[4] B. Bingham, B. Foley, H. Singh, R. Camilli, K. Delaporta, R. Eustice, A. Mallios, D. Mindell, C. Roman, and D. Sakellariou, "Robotic tools for deep water archaeology: Surveying an ancient shipwreck with an autonomous underwater vehicle," *J. Field Robot.*, vol. 27, no. 6, pp. 702–717, 2010. doi: 10.1016/j.neucom.2015.10.122.

[5] M. Zhu, Y. Song, J. Guo, C. Feng, G. Li, T. Yan, and B. He, "PCA and kernel-based extreme learning machine for side-scan sonar image classification," in *Proc. IEEE Underwater Technol. (UT)*, Busan, South Korea, Feb. 2017, pp. 21–24.

[6] V. Myers and J. Fawcett, ''A template matching procedure for automatic target recognition in synthetic aperture sonar imagery,'' *IEEE Signal Process. Lett.*, vol. 17, no. 7, pp. 683–686, Jul. 2010. doi: 10.1109/LSP.2010.2051574.

[7] R. Fandos, A. M. Zoubir, and K. Siantidis, ''Unified design of a feature-based ADAC system for mine hunting using synthetic aperture sonar,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2413–2426, May 2014. doi: 10.1109/TGRS.2013.2260863.

[8] D. P. Williams, ''Underwater target classification in synthetic aperture sonar imagery using deep convolutional neural networks,'' in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Cancun, Mexico, Dec. 2016, pp. 2497–2502.

[9] V. Murino and A. Trucco, ''Three-dimensional image generation and processing in underwater acoustic vision,'' *Proc. IEEE*, vol. 88, no. 12, pp. 1903–1948, Dec. 2000.

[10] N. Dalal and B. Triggs, ''Histograms of oriented gradients for human detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, Jun. 2005, pp. 886–893.

[11] J. McKay, V. Monga, and R. G. Raj, ''Robust sonar ATR through Bayesian pose-corrected sparse classification,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5563–5576, Oct. 2017. doi: 10.1109/TGRS.2017.2710040.

[12] M. D. Santos, P. O. Ribeiro, P. Núñez, P. Drews-Jr, and S. Botelho, ''Object classification in semi structured enviroment using forward-looking sonar,'' *Sensors*, vol. 17, no. 10, pp. 2235–2250, Sep. 2017. doi: 10.3390/s17102235.

[13] G. E. Hinton and R. R. Salakhutdinov, ''Reducing the dimensionality of data with neural networks,'' *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006. doi: 10.1126/science.1127647.

[14] Y. LeCun, Y. Bengio, and G. Hinton, ''Deep learning,'' *Nature*, vol. 521, pp. 436–444, May 2015. doi: 10.1038/nature14539.

[15] M. Z. Alom, T. M. Taha, and C. Yakopcic, ''A state-of-the-art survey on deep learning theory and architectures,'' *Electron.*, vol. 8, no. 3, Mar. 2019. doi: 10.3390/electronics8030292.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, ''Gradient-based learning applied to document recognition,'' *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998. doi: 10.1109/5.726791.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''Imagenet classification with deep convolutional neural networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[18] Y. Sun, X. Wang, and X. Tang, ''Deep learning face representation from predicting 10,000 classes,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 1891–1898.

[19] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, ''Mastering the game of go without human knowledge,'' *Nature*, vol. 550, no. 7676, pp. 354–359, Oct. 2017.

[20] H. Cho and S. C. Yu, ''Real-time sonar image enhancement for AUV-based acoustic vision,'' *Ocean Eng.*, vol. 104, pp. 568–579, Aug. 2015. doi: 10.1016/j.oceaneng.2015.05.037.

[21] X. Wang, J. Jiao, J. Yin, W. Zhao, X. Han, and B. Sun ''Underwater sonar image classification using adaptive weights convolutional neural network,'' *Appl. Acoust.*, vol. 146, pp. 145–154, Mar. 2019. doi: 10.1016/j.apacoust.2018.11.003.

[22] G. E. Hinton, S. Osindero, and Y.-W. Teh, ''A fast learning algorithm for deep belief nets,'' *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006. doi: 10.1162/neco.2006.18.7.1527.

[23] D. Jia, D. Wei, S. Richard, L.-J. Li, K. Li, and L. Fei-Fei, ''Imagenet: A large-scale hierarchical image database,'' *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, Jun. 2009, pp. 248–255.

[24] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, ''Learning and transferring mid-level image representations using convolutional neural networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 1717–1724.

[25] A. Davis and A. Lugsdin, ''High speed underwater inspection for port and harbour security using coda echoscope 3D sonar,'' in *Proc. OCEANS*, Washington, DC, USA, Sep. 2005, pp. 2006–2011.

[26] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, ''Saliency detection via graph-based manifold ranking,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Portland, OR, USA, Jun. 2013, pp. 3166–3173.

[27] H. Qiao, Y. Li, F. Li, X. Xi, and W. Wu, ''Biologically inspired model for visual cognition achieving unsupervised episodic and semantic feature learning,'' *IEEE Trans. Cybern.*, vol. 46, no. 10, pp. 2335–2347, Oct. 2015. doi: 10.1109/TCYB.2015.2476706.

[28] S. J. Pan and Q. Yang, ''A survey on transfer learning,'' *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010. doi: 10.1109/TKDE.2009.191.

[29] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, ''A survey on deep transfer learning,'' Aug. 2018, *arXiv:1808.01974*. [Online]. Available: https://arxiv.org/abs/1808.01974

[30] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, ''Caffe: Convolutional architecture for fast feature embedding,'' Jun. 2014, *arXiv:1408.5093*. [Online]. Available: https://arxiv.org/abs/1408.5093

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ''Going deeper with convolutions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[32] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, ''How transferable are features in deep neural networks?'' Nov. 2014, *arXiv:1411.1792*. [Online]. Available: https://arxiv.org/abs/1411.1792

[33] E. Protas, J. D. Bratti, J. F. O. Gaya, P. Drews, and S. S. C. Botelho, ''Visualization methods for image transformation convolutional neural networks,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 2231–2243, Jul. 2019. doi: 10.1109/TNNLS.2018.2881194.

LEILEI JIN received the B.E. and M.E. degrees in electronic information engineering from the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China, in 2012 and 2015, respectively, where he is currently pursuing the Ph.D. degree in information and communication engineering. His research interests include sonar image processing, pattern recognition, and machine learning.

HONG LIANG received the B.E. degree from Xidian University, Xi'an, China, in 1990, and the M.E. and Ph.D. degrees from the School of Marine Science and Technology, Northwestern Polytechnical University (NPU), Xi'an, in 1995 and 2004, respectively. She has been a Faculty Member with NPU, since 1995 and has also been a Professor, since 2011. She teaches and conducts research at NPU in the areas of modern signal processing and signal detection and estimation. Her current research interests include weak signal detection, machine learning, and Chinese remainder theorem and its application.

CHANGSHENG YANG received the B.E., M.E., and Ph.D. degrees from the School of Marine Science and Technology, Northwestern Polytechnical University (NPU), Xi'an, China, in 1999, 2002, and 2006, respectively. He has been a Faculty Member with NPU, since 2006 and has also been an Associate Professor, since 2010. He teaches and conducts research at NPU in the areas of digital signal processing and tracking and locating of maneuvering targets. His current research interests include modern signal processing, array signal processing, and bionic intelligent perception and its application.