Journal of *Marine Science and Engineering*

MDPI

Article

Underwater Acoustic Target Recognition Based on Deep Residual Attention Convolutional Neural Network

Fang Ji [1,*], Junshuai Ni [1], Guonan Li [1], Liming Liu [1] and Yuyang Wang [1,2]

[1] China Ship Research and Development Academy, Beijing 100192, China; nijunshuai2022@163.com (J.N.)
[2] College of Shipbuilding Engineering, Harbin Engineering University, Harbin 150009, China
* Correspondence: heujifang@163.com

Abstract: Underwater acoustic target recognition methods based on time-frequency analysis have shortcomings, such as missing information on target characteristics and having a large computation volume, which leads to difficulties in improving the *accuracy* and immediacy of the target recognition system. In this paper, an underwater acoustic target recognition model based on a deep residual attention convolutional neural network called DRACNN is proposed, whose input is the time-domain signal of the underwater acoustic targets radiated noise. In this model, convolutional blocks with attention to the mechanisms are used to focus on and extract deep features of the target, and residual networks are used to improve the stability of the network training. On the full ShipsEar dataset, the recognition *accuracy* of the DRACNN model is 97.1%, which is 2.2% higher than the resnet-18 model with an approximately equal number of parameters as this model. With similar recognition accuracies, the DRACNN model parameters are 1/36th and 1/10th of the AResNet model and UTAR-Transformer model, respectively, and the floating-point operations are 1/292nd and 1/46th of the two models, respectively. Finally, the DRACNN model pre-trained on the ShipsEar dataset was migrated to the DeepShip dataset and achieved recognition *accuracy* of 89.2%. The experimental results illustrate that the DRACNN model has excellent generalization ability and is suitable for a micro-UATR system.

Keywords: underwater acoustic target recognition; time-domain signal; convolutional neural network; channel attention mechanism; residual connections

Citation: Ji, F.; Ni, J.; Li, G.; Liu, L.; Wang, Y. Underwater Acoustic Target Recognition Based on Deep Residual Attention Convolutional Neural Network. *J. Mar. Sci. Eng.* **2023**, *11*, 1626. https://doi.org/10.3390/jmse11081626

Received: 18 July 2023
Revised: 8 August 2023
Accepted: 18 August 2023
Published: 20 August 2023

1. Introduction

Underwater acoustic target recognition (UATR) has always been a hot topic of research in the field of passive sonar and is also a technical problem that needs to be solved internationally, both in the civilian and military fields. Traditional UATR [1–5] still faces challenges such as unstable target features under complex conditions, interference from environmental noise and other targets, distortion of characteristics during acoustic propagation, etc. In recent years, with the development of artificial intelligence technology and the increase in public data on underwater acoustic signals [6,7], researchers have used deep learning models to establish the mapping relationship from the original data of the target category, using a data-driven network model which extracts non-linear features in a way that provides a new idea to solve the above problems.

Most UATR methods based on deep learning use the time-frequency map of underwater acoustic signals to characterize the target characteristic information, which is used as the input pattern of convolutional neural networks for feature extraction and recognition. Short-time Fourier transform is a method to obtain linear time-frequency spectrum maps represented by LOFAR spectrum [8], while Mel time-frequency spectrum [9] and logarithmic Mel time-frequency spectrum [10] can better describe the energy distribution pattern of underwater acoustic signals for introducing non-linear factors. It has been shown that the recognition performance of using ResNet or DenseNet models with residual connectivity to recognize the Mel time-frequency spectrum of underwater acoustic targets

is better than more mature models in image processing, such as VGG19 [11,12]. Hong fused the three feature maps and designed a ResNet-18 network model containing a central loss function and an embedding layer [13] for UATR. In order to solve the problems of a lack of realistic target data and interference from marine environmental noise, Li added an attention mechanism to the residual network to enhance the model's ability to detect line spectrum and transient signals [14]. Li introduced the transformer mechanism into the convolutional neural network and proposed a UATR model named STM [15], which improved the recognition *accuracy* by 1.8% compared with the ResNet-18 model. In order to expand the number of training samples, some methods use deep conditional generative adversarial networks (DCGAN) or Spec- Augment algorithms for data augmentation of real-world targets [16,17] to solve problems such as model overfitting caused by an insufficient number of samples and imbalance of training samples on categories, resulting in significant improvements in the target recognition performance of deep learning-based UATR methods with small samples.

The process of transforming an underwater acoustic target signal from the time domain into the time-frequency domain inevitably involves some information loss, and the high-quality and stable time-frequency spectrum relies on the accumulation of signal energy over a long period of time, which adversely affects the *accuracy* and real-time performance of UATR systems. The raw signals of underwater acoustic targets contain the richest information about target characteristics and have the potential to be recognized by deep learning models. Hu proposed a deep neural network model for UATR containing depth-separable convolution and time-discrete convolution by taking the raw signal of underwater acoustic targets as the model's input [18]. Li designed a timbre-aware deep convolutional neural network called ASTEM_DCNN for extracting line spectrums and fusing different frequency signal components to improve the interpretability of the feature extraction [19]. Song proposed a new method for UATR by integrating a one-dimensional convolutional neural network and a long short-term memory network (LSTM) [20]. Yang extracted acoustic features from ship-radiated noise time-domain signals and designed a set of weight-learning neurons to establish the relationship between deep features and target attributes to achieve sensing of underwater acoustic target parameters such as vessel size and working conditions [21]. Although some deep learning methods have been shown to achieve advanced recognition *accuracy*, the design of residual networks and the introduction of attention mechanisms to optimize feature extraction will further improve the overall recognition performance, including recognition *accuracy* and operating cost, thus meeting the needs of practical engineering applications.

In this paper, a UATR method based on deep residual attention convolutional neural network (DRACNN) is proposed, which preprocesses the underwater acoustic target radiated noise signals acquired by passive sonar and uses the preprocessed time-domain signals as the input and the target category labels as the output to the model, achieving end-to-end recognition. The residual structure is used to solve the problem of gradient disappearance that tends to arise in the deep network and to improve the convergence of the model fit during training. A channel attention mechanism is incorporated into each residual network unit to ensure that the model always focuses on extracting highly distinguishable features of the target during the information transfer from the input layer to the output layer, improving the model's ability in terms of target characteristic expression and resistance to noise interference. A convolution filter bank with four convolutional layers and a convolution step of two is used to connect two residual attention convolutional blocks (RACB) spaced apart for achieving data dimensionality reduction and feature fusion and improving the *accuracy* of hydroacoustic target recognition.

The rest of the paper is organized as follows: The underwater acoustic target characteristics and signal preprocessing methods are explained in Section 2. The basic process of the UATR methods and the DRACNN model proposed in this paper are specifically described in Section 3. The test results of the proposed method on publicly available datasets and

comparison results with other state-of-the-art models are given in Section 4. Section 5 summarizes the paper.

2. Under Water Acoustic Target Characteristics and Signal Preprocessing

2.1. Target Characteristics

Underwater acoustic target radiated noise consists of mechanical noise, propeller noise and hydrodynamic noise and is a broadband and non-smooth signal coupled by multiple sound sources. Hydrodynamic noise generated by the interaction of the ship with the water is irregular and contributes little to target recognition [22]. Mechanical and propeller noise has an audible rhythm of strong and weak periodic undulations, reflecting the rotation speed of the main engine, bearings, propellers and other devices. The vibration of mechanical structures is locally smooth, where the radiated noise generated by the ship mainframe ignition, blades cutting water flow field, collision and friction between bearing and nesting, propeller blade resonances, and this rich characteristic information is embedded in the time-domain signal of the underwater acoustic target, so they are an important source of features for target recognition [23].

Figure 1 shows the time-domain waveforms and time-frequency spectrums of the radiated noise signals of three different hydroacoustic targets. Line spectrums of the mainframe ignition radiated noise shown in Figure 1b are generally distributed in the frequency range of 0~200 Hz, and there is a clear harmonic relationship between the line spectrums. The number, frequency and intensity of the line spectrums are related to the mainframe construction and working conditions; these characteristics exist in any underwater acoustic target powered by diesel engines. The line spectrums of shaft grinding noise shown in Figure 1b are generally distributed over the frequency range of 200–1000 Hz, and the main shaft rotation is modulated to produce line spectrums accompanying phenomenon, so the frequency difference between two adjacent line spectra is equal to the shaft frequency, most of the civil vessels such as fishing boats and cargo ships have this typical feature. The line spectrums of the propeller resonance noise shown in Figure 1f usually have frequencies greater than 1000 Hz and are commonly found in small boats with high propeller speeds. These differences in characteristics are reflected in the time domain signal in the difference in the structure of the waveforms shown in Figure 1a,c,e, which is the main basis for identifying underwater acoustic targets.

2.2. Signal Preprocessing

It requires resampling the underwater acquisition data to the same sampling frequency to ensure that samples of a given sampling point have the same data duration when we use the time-domain signals to recognize underwater acoustic targets. If the sampling rate is too high, the corresponding data duration is short and carries insufficient information about the target characteristics. If the sampling rate is too low, the high-frequency component is missing, which is not conducive to fine target identification. Considering that the energy of the underwater acoustic target radiated noise is mainly concentrated at low frequencies, and part of the high-frequency harmonic line spectrum is generally distributed within 10 kHz, we set the resampling frequency to 20 kHz. After resampling and weighing the two aspects of information and computation, we intercepted the data according to 4096 samples per frame, with two adjacent frames overlapping by 2048 samples, so that the signal duration of each frame is about 0.2 s. In order to improve the spectrum resolution in frequency, time-frequency analysis generally sets the time-domain signal duration of 3~5 s. In comparison, our method greatly reduces the required data duration, which means that the method in this paper will have a faster UATA speed. The segmentation method is shown in Figure 2.

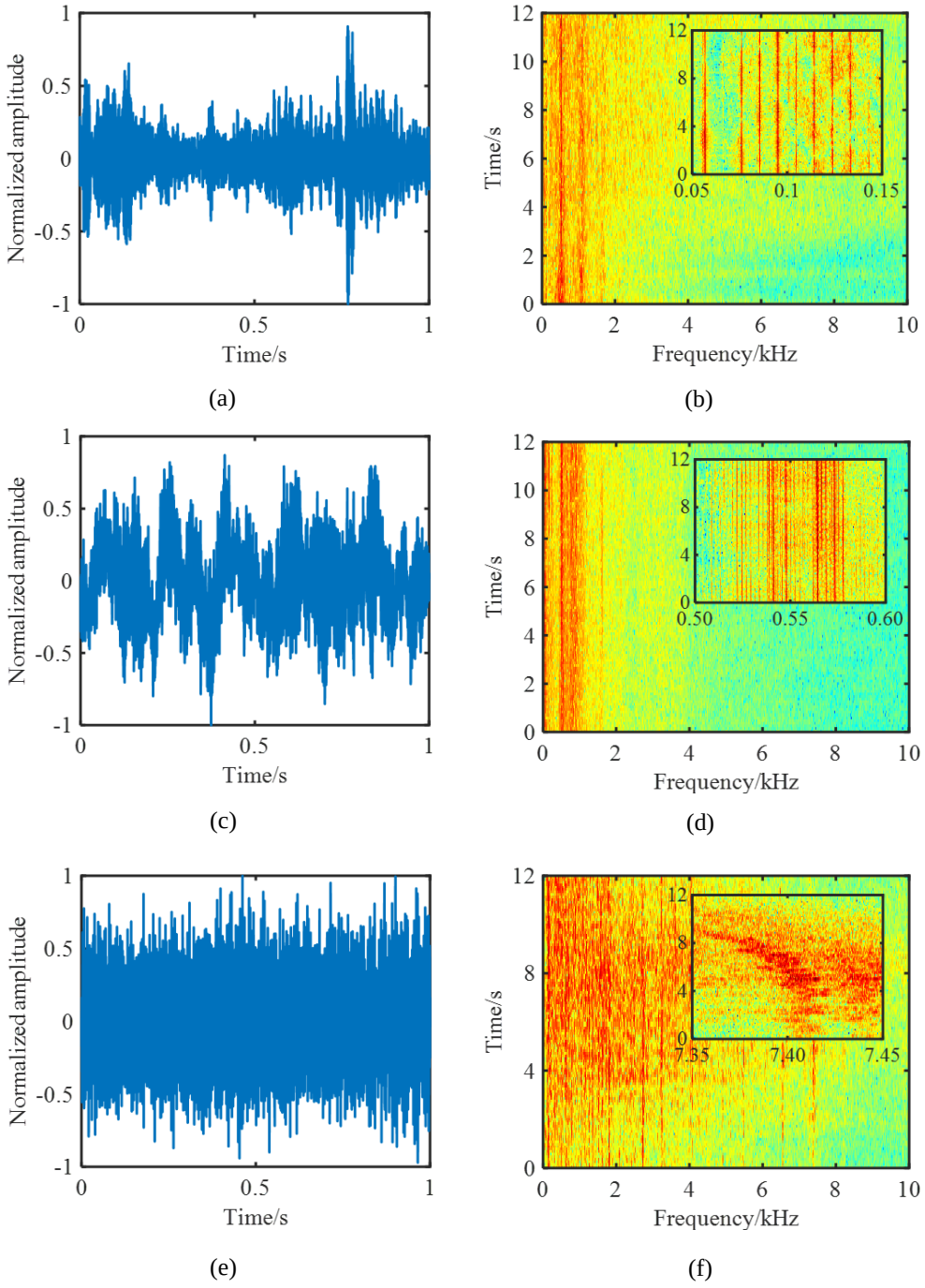

Figure 1. Time domain waveform and Spectrogram of underwater acoustic target radiated noise. (**a**) Time-domain waveform of target I; (**b**) Time-frequency spectrum of target I; (**c**) Time-domain waveform of target II; (**d**) Time-frequency spectrum of target II; (**e**) Time-domain waveform of target III; (**f**) Time-frequency spectrum of target III.

After each frame of data is zero-averaged and normalized, we assume the signal as $s(n)$, $n = 1, 2, \ldots, 4096$, and the treated sample can be written as:

$$s'(n) = \frac{s(n) - \frac{1}{4096}\sum_{j=1}^{4096} s(j)}{\max\left(\left|s(n) - \frac{1}{4096}\sum_{j=1}^{4096} s(j)\right|\right)}, n = 1, 2, \cdots, 4096 \quad (1)$$

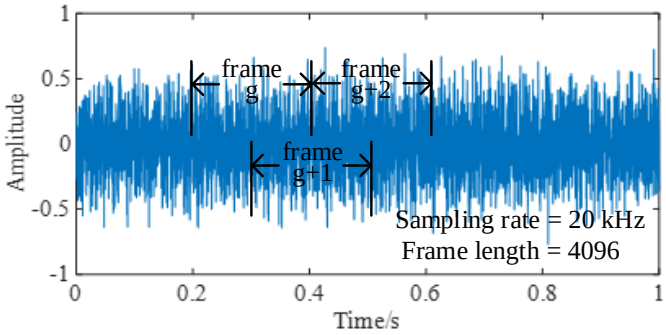

Figure 2. Schematic diagram of underwater acoustic target signal framing.

The signal preprocessing operation adjusts the varying lengths of underwater acoustic target radiated noise data acquired by the listening device into samples of each segment with the same time scale and amplitude scale. At this stage, we build a sample set by adding some labels according to the target category, which satisfies the sample requirements for data input and network training of the recognition model.

3. UATR Method

In general, the UATR method based on deep learning mainly contains four steps: signal acquisition, pre-processing, deep learning model training, and target classification recognition, as shown in Figure 3.

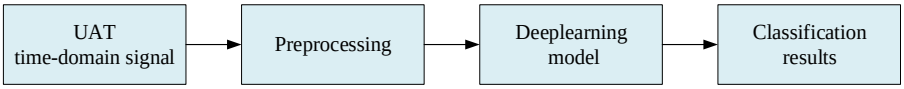

Figure 3. The flow of hydroacoustic target recognition method.

In this framework, we propose a DRACNN model in this paper for receiving pre-processed time-domain signals of underwater acoustic target radiated noise to produce classification results. This UATR method is named WAVE_DRACNN. Our model will be described in detail in the following part of this section.

3.1. Residual Attention Convolution Blocks

In order to perceive feature information from underwater acoustic target radiated noise time-domain signals, we first propose and design a residual attention convolution block. RACB is the basic unit that forms the DRACNN, which consists of two normal 1D convolutional layers, two ReLU activation layers, a MaxPooling layer, and a 1D convolutional layer with a channel attention mechanism. Its structure is shown in Figure 4.

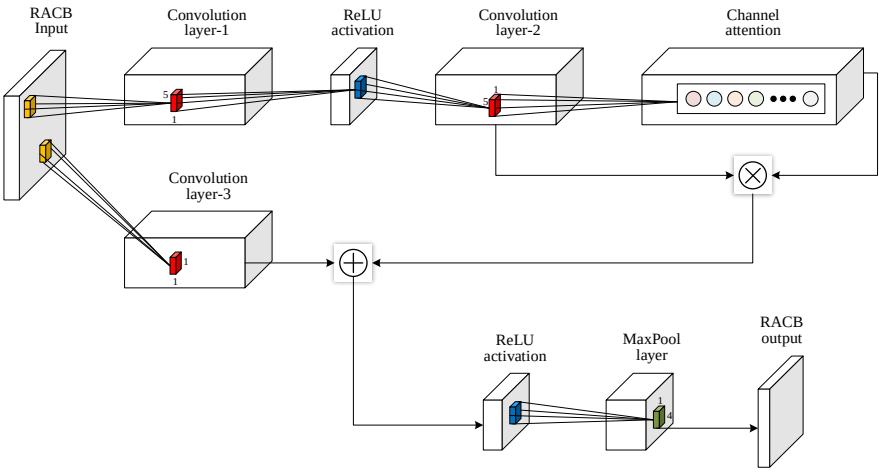

Figure 4. Residual attention convolution block structure.

We use the proposed residual attention convolution block to implement the following four functions:

(1) Feature extraction

Local features are extracted from the underwater acoustic target time-domain signal using the 1D convolutional layer-1 shown in Figure 4 with a specific convolutional kernel size and a number of filter channels to perform convolutional operations on the input data [24]. This process transforms the time-domain signal waveform into different frequency sub-band signals and outputs them through the convolutional layer channels. This is immediately followed by non-linear processing of the upper layer features using the ReLU activation function, which drives the model to learn more complex abstract features. The convolution layer operation is represented as:

$$y_i = w_i \cdot x + b_i \tag{2}$$

where x is the input data for convolution layer-1, w_i and b_i is the convolution kernel and bias corresponding to the output features y_i, respectively.

The activation function is defined as:

$$\mathrm{ReLD}(y_i) = max(y_i, 0) \tag{3}$$

where y_i is the input of the ith channel of the ReLU layer.

(2) Weighting feature maps by channel

Deep feature extraction is performed on ReLU output features using the 1D convolutional layer-2 with a certain number of channels, where the contribution of feature patterns learned by different filter channels to target recognition is different. So, we use the channel attention mechanism called Squeeze Excitation (SE) module [25] to adaptively train a weight value for each channel that measures the importance of the features and weights the channel features to act as a reinforcement for the key features.

In the SE model, the feature map is first downscaled using global pooling to output a set of vectors with the same dimensions as the number of channels, and then two dense layers, respectively, using RELU activation and Sigmoid activation, are added to learn the weights of each channel. The structure of the SE model is shown in Figure 5, and the process is represented as:

$$p_i^A = \sigma_i \cdot p_i \tag{4}$$

where p_i is the output feature map of convolution layer-2, σ_i is the channel weights, and p_i^A is the attention-weighted feature maps.

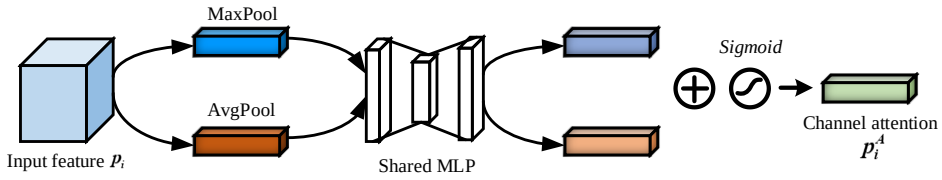

Figure 5. Diagram of channel attention model.

(3) Residual concatenation

The convolutional layer-3 with a kernel size of 1×1 and i filters is used to transform the input data into i channel feature maps and summed with the SE module's output feature maps by channel to achieve a residual connection between the input data and the depth-weighted features. In addition, it also serves as a multi-scale feature fusion, which is expressed as

$$q_i = w_i^R \cdot x + b_i^R + p_i^A \tag{5}$$

where w_i^R and b_i^R respectively are the convolution kernel and bias of convolution layer-3, and q_i is the residual output feature map.

Then a ReLU activation function is then used after the residuals to sparse the output features, allowing the model to better mine the target features.

(4) Feature map down sampling

The feature map is down-sampled to $1/l$ its original length by using a MaxPooling layer of kernel size $l \times 1$ to remove redundant information and reduce the model parameters. In terms of target characteristics, the MaxPooling operation serves to detect transient impact signals of underwater acoustic targets, such as host ignition noise and shaft bumper noise and helps the model to extract local features of the time-domain waveform of the target radiated noise signal and to recognize them using the differences in the short-time smooth processes of different targets.

3.2. DRACNN Model

In this paper, we propose a deep residual attention convolutional neural network model, which consists of two parts: a feature extraction module and a target classification module. The model structure is shown in Figure 6.

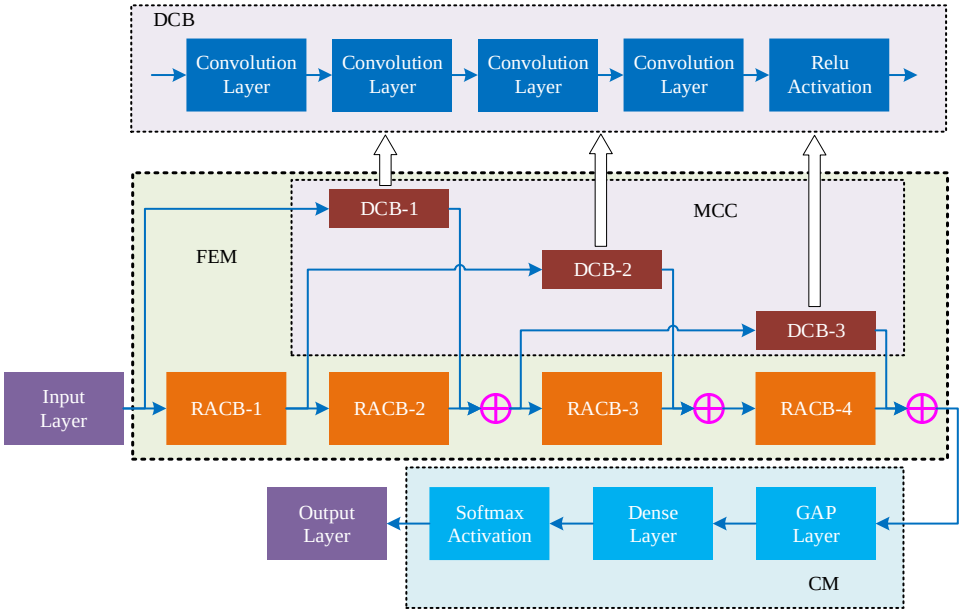

Figure 6. Structure of the DRACNN model.

The Feature extraction model (FEM) consists of 4 RACB units, with all convolution kernels on the model backbone of size 5×1, all convolution steps of 1, and all pooling kernels of size 4×1. In addition, we designed a deep convolutional block (DCB) consisting of four standard convolutional layers and one ReLU activation function layer and fused two adjacent RACBs by multi-convolutional layer jumper connection (MCC) and summation. The DCB achieves feature downsampling by means of four convolution kernels with a step size of 2. The number of channels in the DCBs' layers is the same as the number of RACB channels connected at the output. This kind of design effectively preserves the envelope characteristics of the underwater acoustic target signal and improves the model's ability to learn the essential characteristics of the target and fit the training data under complex operating conditions changes and background noise interference.

In the Classification Module (CM), we first use a Global Average Pooling (GAP) layer [26] to down-sample several channel feature maps generated by the Feature Extraction module by channel averaging to obtain a set of underwater acoustic target depth feature vectors with the same number of dimensions as the number of channels. Secondly, we add a fully connected layer after the GAP layer to produce one-hot encoded target category

labels and activate the output values of this layer using the Softmax function shown in Equation (6):

$$\text{Softmax}(z_k) = \frac{e^{z_k}}{\sum_k e^{z_k}} \tag{6}$$

where k is the number of target classes and z_k is the output value of the neuron of the fully connected layer.

The specific parameter settings for each layer (block) of the DRACNN model are shown in Table 1.

Table 1. Specific parameter settings of the DRACNN model.

Layer (Block)	Channels	Input Shape	Output Shape
Input Layer	1	\	(None, 4096, 1)
RACB-1	16	(None, 4096, 1)	(None, 1024, 16)
RACB-2	32	(None, 1024, 16)	(None, 256, 32)
DCB-1	32	(None, 4096, 1)	(None, 256, 32)
RACB-3	64	(None, 256, 32)	(None, 64, 64)
DCB-2	64	(None, 1024, 16)	(None, 64, 64)
RACB-4	128	(None, 64, 128)	(None, 16, 128)
DCB-3	128	(None, 256, 32)	(None, 16, 128)
GAP Layer	\	(None, 16, 128)	(None, 128)
Dense Layer	\	(None, 128)	(None, m)
Output Layer	\	(None, m)	\
Total params: 0.26 M		Flops: 5.12 M	

It is easy to see from Table 1 that the preprocessed underwater acoustic target time-domain signal of 4096 samples in length is received by the input layer with a number of channels of 1 and a number of nodes of 4096 and is then passed through four RACBs with a number of filter channels of 16, 32, 64 and 128 respectively to extract features and downsample them (all the convolutional layers within each RACB have the same number of filter channels) to obtain a feature map of size $16 \times 1 \times 128$. The GAP layer down samples the feature map into a set of 128-dimensional feature vectors, which are finally classified by a Softmax classifier in the fully connected layer and the classification result is given in the output layer, finally realizing the mapping from signal space to feature space to category space. The model has 0.26 M (millions) parameters and 5.12 M floating-point operations.

4. UATR Experiment and Analysis

4.1. Experimental Database

In this paper, the recognition performance of the DRACNN model is validated using the ShipsEar dataset available at http://atlanttic.uvigo.es/underwaternoise/ (accessed on 1 July 2023). The ShipsEar dataset was selected from audio recordings collected in the fall of 2012 and the summer of 2013 under different sea conditions in Vigo Harbor, Spain. This dataset contains 91 sound recordings of 11 vessel types and one background noise class, with a total duration of 3 h and 10 min and a sampling rate of 52,734 Hz. Targets in this dataset were categorized into five categories by vessel length, as shown in Table 2.

Table 2. ShipsEar data recognition details. (Duration in seconds).

Category	Type of Vessel	Files	Duration
Class A	Fishing boats, trawlers, mussel boats, tugboats, dredgers	17	1880
Class B	Motor boats, pilot boats, sailboats	19	1567
Class C	Passenger ferries	30	4276
Class D	Ocean liners, ro-ro vessels	12	2460
Class E	Background noise recordings	12	1145

As can be seen from the durations of the five categories of target data in Table 2, category C has the largest amount of data, with a duration of 4276 s, while Category E has the smallest amount of data, with a length of only 1145 s. Therefore, the ShipsEar dataset is seriously unbalanced in terms of the samples of each category.

4.2. Introduction to the Sample Set

If the ShipsEar dataset is cleaned to eliminate most of the samples that are not easily recognized according to some references, very impressive recognition results will be obtained, but the results under this operation are not convincing. In order to reflect the comparability of the recognition performance between our model and the reference methods, we use all the raw data in the ShipsEar dataset to conduct test and comparison experiments without any data filtering and data enhancement measures. The raw data were preprocessed to obtain 110,542 samples, of which 80% were randomly selected as the training set and the remaining 20% as the test set. The sample set details are shown in Table 3.

Table 3. Sample size of training and test set.

Category	Training Set	Test Set
Class A	14,708	3638
Class B	12,165	3114
Class C	33,247	8485
Class D	19,278	4733
Class E	9034	2140
Total	88,432	22,110

4.3. Experimental Results and Analysis

The DRACNN model proposed in this paper is built in a deep learning development environment with a Windows 10 operating system, Python 3.6.5, Keras 2.2.4, TensorFlow-1.14.0, Cuda 10.0.130, and trained on a workstation equipped with Nvidia GTX1660ti GPU, Core I5-10400F CPU, and 16 GB RAM. We have chosen appropriate hyperparameters for the model as follows: optimizer setting to Adam, batch-size setting to 256 and epochs setting to 100. In the iterations, we update the model parameters using the error back-propagation (BP) algorithm and optimize the model parameters using the Adam optimizer with an adaptive learning rate. The initial learning rate of the Adam optimizer is 0.001. The exponential decay rate of the first-order moment estimation is set to 0.9, and the exponential decay rate of the second-order moment estimation is set to 0.999. We use the joint loss function containing a multicategory cross-entropy loss function, and a central loss function to measure the error between the model predicted values and the actual values, minimizing the error through iteration.

The DRACNN model is iterated on the ShipsEar dataset for 100 epochs, and Figure 7 shows the change curves of *accuracy* and loss during the whole process. As shown in Figure 7a, the training loss and validation loss decrease rapidly within about 20 epochs, after which they gradually decrease and converge to a certain fixed value in the absence of overfitting and underfitting. It is also shown in Figure 7b that the training and validation accuracies grow rapidly and quickly converge to a stable value, with an optimal *accuracy* of 0.999 on the training set and 0.970 on the validation set. In the experiments to test the recognition *accuracy* of the model, we conducted a total of 10 repetitions of the experiment, and each time, we reselected the samples of the training set and the test set randomly with fixed sample capacities. The recognition results are shown in Table 4.

The mean recognition *accuracy* of the 10 experiments is 97.1% with a standard deviation of 0.24, and the mean loss function value is 0.11 with a standard deviation of 0.01. In order to show the prediction information in detail, we give the recognition accuracies of the DRACNN model for each class of targets in the ShipsEar dataset as well as the proportion of samples that misidentify one class of targets as other classes in the form of confusion matrices, as shown in Figure 8.

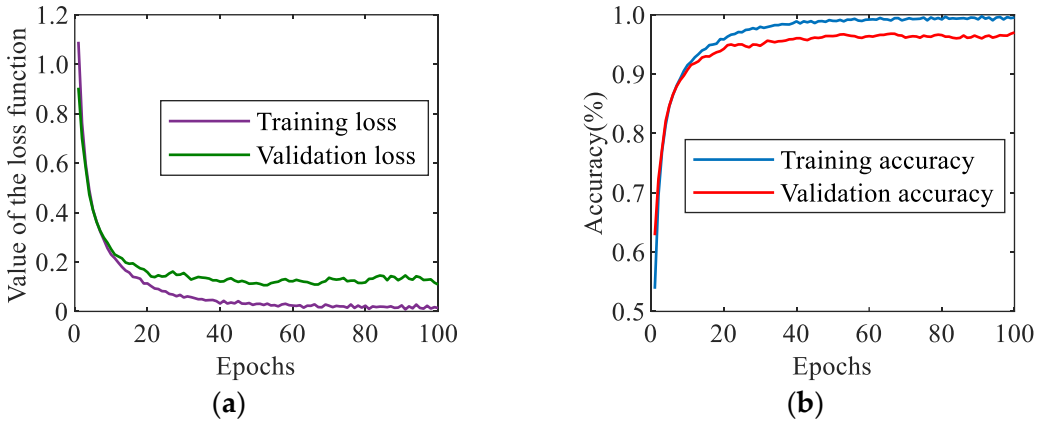

Figure 7. Training and validation curves of the DRACNN model. (**a**) Loss. (**b**) *Accuracy*.

Table 4. Recognition *accuracy* in ten experiments.

Experiment Times	*Accuracy* (%)	Value of Loss Function
01	97.0	0.11
02	96.8	0.12
03	97.2	0.10
04	97.6	0.09
05	96.9	0.12
06	97.1	0.11
07	97.2	0.11
08	97.0	0.13
09	97.4	0.10
10	96.8	0.12
Average	97.1	0.11
Std	0.24	0.01

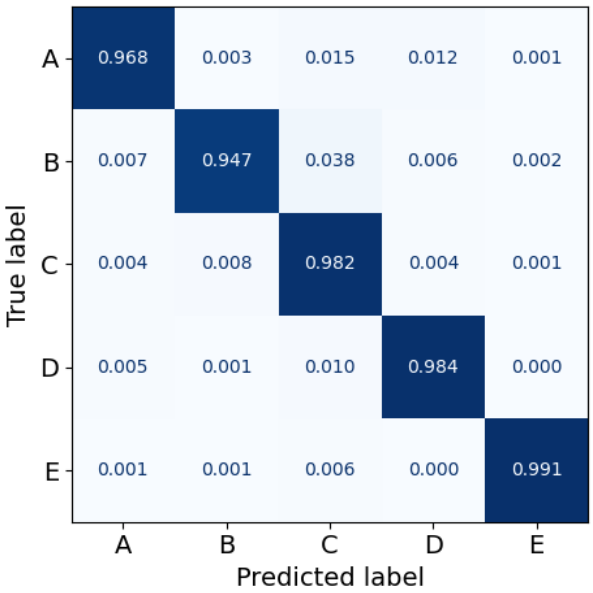

Figure 8. Confusion matrix for recognition results on ShipsEar dataset.

Targets in class E have the best recognition results with an *accuracy* of 99.1% despite the smallest number of training samples, mainly because these targets are marine background noise, which are easy to distinguish from ship targets in class A, B, C, and D, because of the significant difference in the noise characteristics. Among the four types of ship targets,

targets in class B have the worst recognition results, with a recognition *accuracy* of 94.7%. One of the possible reasons is that the training samples of this class of targets are the least, leading to the model training not being sufficient. Another possible reason is that targets (Motor boats, pilot boats, sailboats) in class B have similar characteristics to targets (Passenger ferries) in class C and are easily misidentified as class C. In addition to recognition *accuracy*, in the experimental results presentation, we use *precision*, *recall*, and *F1_score* to evaluate the recognition performance of the DRACNN model more comprehensively. The results are shown in Table 5, and the formula for each index is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1\_score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

where TP is true positive, FP is false positive and FN is false negative.

Table 5. The recognition results of each class.

Category	*Precision* (%)	*Recall* (%)	*F1_Score* (%)
Class A	97.7	96.8	97.3
Class B	97.1	94.7	95.9
Class C	97.3	98.2	97.8
Class D	97.8	98.4	98.1
Class E	99.4	99.1	99.2
Average	97.9	97.5	97.7

In order to better analyze the results, we compare some methods also validated using the ShipsEar dataset with our experimental results. Here, in addition to the recognition *accuracy*, we also compare the number of parameters and the amount of floating-point computation of these deep learning-based models given in the references, as shown in Table 6.

Table 6. Comparison of this method with other state-of-the-art methods. (Flops in Gigabit).

No.	Model	*Accuracy* (%)	Params (M)	Flops (G)
1	DenseNet-121 [27]	90.1	6.96	0.610
2	DarkNet-53 [28]	96.6	40.59	1.930
3	RepVGG-A0 [29]	97.0	7.83	0.420
4	CRNN-9 [30]	91.4	3.88	0.110
5	Autoencoder [31]	93.3	0.18	0.410
6	ResNet-18 [13]	94.9	0.33	0.110
7	AResNet [14]	98.0	9.47	1.460
8	UATR-Transformer [32]	96.9	2.55	0.230
9	MobileNet-V2 [33]	94.0	2.23	0.140
10	Our	97.1	0.26	0.005

The DRACNN model proposed in this paper achieves a target recognition *accuracy* of 97.1% on the ShipsEar dataset, which has the highest recognition *accuracy* except for the AResNet model, with a recognition *accuracy* of 98.0%. The UTAR-Transformer model, introducing the self-attention mechanism into CNN, has a recognition *accuracy* of 96.9%, which is roughly the same as the model recognition *accuracy* in this paper. However, it is

worth mentioning that unlike AResNet and other comparative methods, we did not conduct any data filtering and data enhancement in the process of model training and testing, which makes our results more credible and convincing. More importantly, our model has only 0.26 M parameters, which is about 1/36th and 1/10th of the AResNet model and UTAR-Transformer model, respectively, and has 5 M floating-point computations, which is about 1/292nd and 1/46th of the two models, respectively. So, our model has a smaller number of parameters and floating-point computations in comparison, which means that less memory and computational resources are required to run the model, facilitating the deployment of the model on a minicomputer system and the fast implementation of target recognition. On our computer, preprocessing consumes 2.9 ms per sample, and recognition consumes 0.2 ms per sample.

Due to the close distance from the targets to the hydrophones, the signal-to-noise ratio is high for the underwater acoustic target radiated noise signals in the ShipsEar dataset. We add Gaussian noise with different signal-to-noise ratios to these signals and repeat the previous experimental steps for recognition. Figure 9 shows the recognition results of our model with different SNRs, which shows the target recognition *accuracy* increases with the increase of SNR. When the SNR is −20 dB, the recognition *accuracy* of our method is 65.8%, and when the SNR is greater than 0 dB, the recognition *accuracy* of our method achieves more than 90.0%.

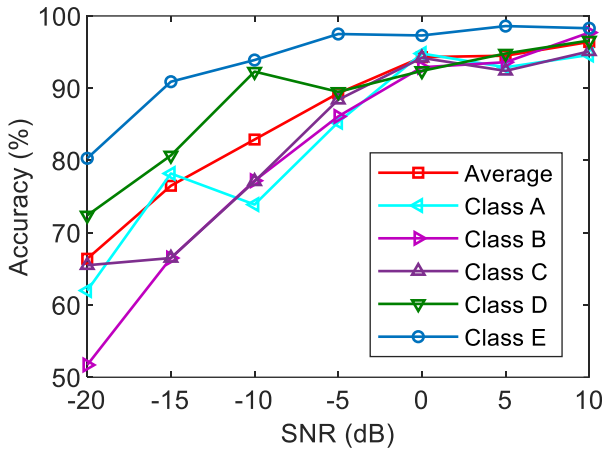

Figure 9. Recognition *accuracy* at different signal-to-noise ratios.

4.4. Generalization Ability of DRACNN Model

It is worth focusing on whether our method is still applicable to other datasets or application scenarios. For the DRACNN model proposed in this paper, the FEM in front of the GAP layer in CM is used to extract deep features from the underwater acoustic target radiated noise, and we use the DeepShip dataset to verify the generalization ability of the features extracted by FEM. A detailed description of the DeepShip dataset was given in the paper [7]. The DeepShip dataset records the radiated noise signals of four types of targets, and they are Cargo (Car), Passenger ship (Pas), Tanker (Tan), and Tug. However, The DeepShip dataset is too large for our model's training and test, so we only select part of the data in this dataset without losing representativity, and the selections are as follows:

(1) We have a subset of audio files with labeling no greater than 60 from the full set of each type of DeepShip dataset.

(2) We intercept signals with a duration of 10 s from the middle of each signal for making the sample set.

We obtained a total of 45,162 samples after preprocessing, 80% of the samples were randomly selected as the training set, and 20% of the samples were used as the test set. The detailed sample size is shown in Table 7.

Table 7. Sample size of training and test set for the DeepShip dataset.

Category	Training Set	Test Set
Cargo	9277	2363
Passenger ship	8872	2240
Tanker	8626	2144
Tug	9354	2286
Total	36,129	9033

The DRACNN model is trained using the ShipsEar dataset. On this basis, we input the test data from the ShipsEar dataset and DeepShip dataset into the pre-training DRACNN model and use the TSNE algorithm to downsize the input data and the deep features extracted from the GAP layer so that their dimensionality is all changed to 2, which can be conveniently used for visualization and analysis. The results are shown in Figure 10.

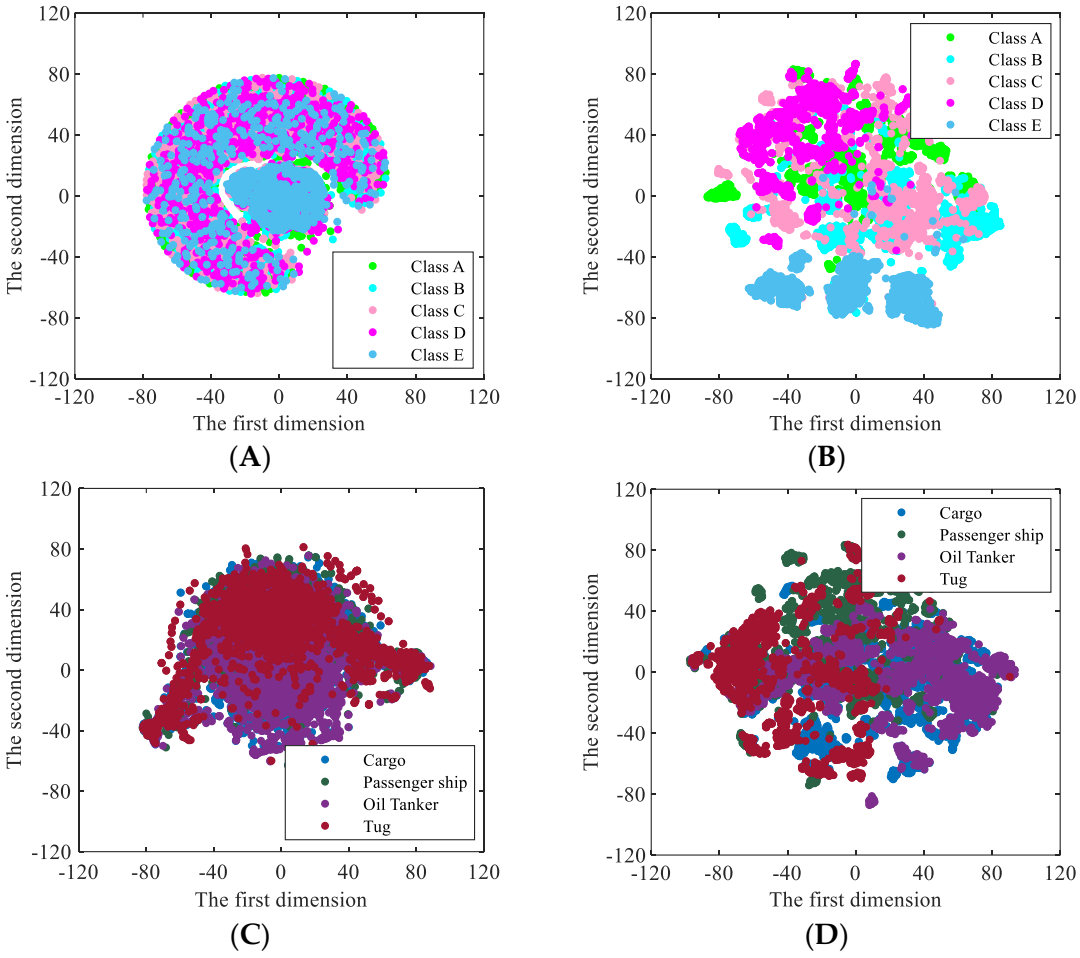

Figure 10. Downscaling and visualization of data using TSNE. (**A**) Raw signals of ShipsEar dataset. (**B**) Features of the ShipsEar dataset extracted by GAP layer. (**C**) Raw signals of DeepShip dataset. (**D**) Features of the DeepShip dataset extracted by GAP layer.

The raw signals of underwater acoustic targets are cluttered in the feature space, and we can hardly find any information that can distinguish the targets from Figure 10A,C, while from Figure 10B,D, we can see the distinguishability of the depth features extracted from the GAP layer is significantly improved. As can be seen from Figure 10B, the features of the four classes of ship targets show some separation, especially the marine environmental noise represented by category E, which is clearly distinguished from the data of other categories.

The class separability of the depth features of the DeepShip dataset samples extracted by DRACNN is significantly improved, which is shown in Figure 10D. These results indicate that the DRACNN model effectively learns the intrinsic properties of underwater acoustic targets and has good generalization ability. Finally, we connect two dense layers with 32 nodes and a relu activation function layer and a Softmax classifier with four nodes after the FEM to form a new model in sequence. The new model has only 0.005 M trainable parameters, which is equivalent to a sample deep neural network and is shown in Figure 11.

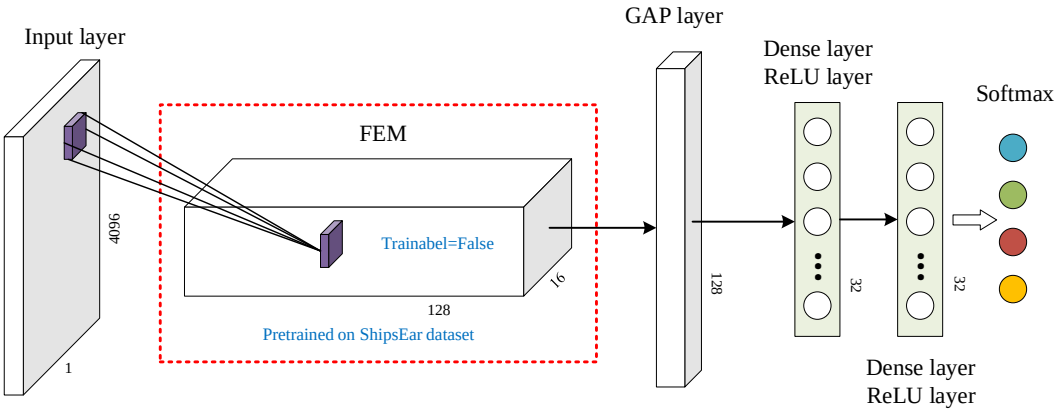

Figure 11. A new model whose FEM is untrainable.

The parameters in the FEM are frozen so that they are no longer involved in the training process. Several newly added layers are trained using training set samples included in the DeepShip dataset and then used to recognize the test set samples, achieving 89.2% recognition *accuracy*, exceeding that of most traditional methods and deep learning methods. The confusion matrix of the recognition results is shown in Figure 12.

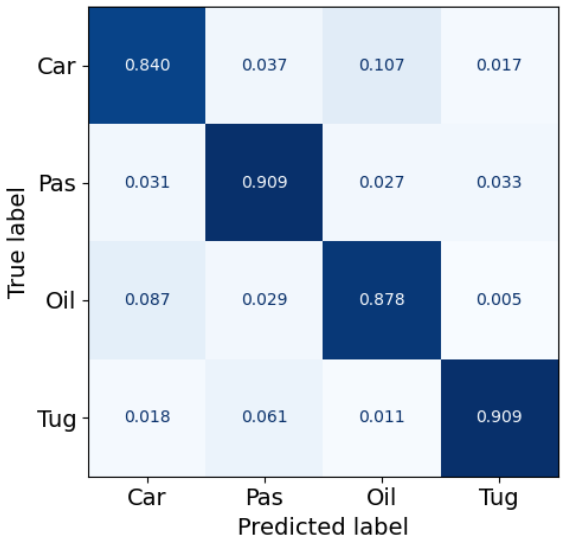

Figure 12. Confusion matrix for recognition results on DeepShip dataset.

It is easy to see that we only trained a classifier for the DRACNN model on the DeepShip dataset, while the feature extraction was done by the FEM trained on the ShipsEar dataset, and the new model respectively achieves 84.0%, 90.9%, 87.8%, and 90.9% recognition accuracies for the four types of targets. Another experimental result shows that the recognition *accuracy* of the new model without FEM is only 51.8%, which further illustrates that the deep features extracted by FEM can reflect the target characteristics well.

5. Conclusions

In this paper, a UATR method using the time-domain signals of the underwater acoustic target radiated noise is proposed, and a deep residual attention convolutional neural network-based UATR model DRACNN is designed. On this basis, the main contributions of this article are as follows:

(1) This method eliminates the step of time-frequency analysis and can achieve target recognition by the signal of a 0.2 s time duration, which greatly improves the immediacy of the system.

(2) The DRACNN model takes full advantage of the high stability of residual networks and consistently focuses on and effectively extracts features reflecting the essential characteristics of the underwater acoustic target by the SE model, achieving 97.1% recognition *accuracy* on the ShipsEar dataset, which has better-integrated performance than ResNet-18, AResNet, CRNN-9 and other current methods.

(3) This model exhibits good generalization performance on the DeepShip dataset and has an extremely smaller number of model parameters and floating-point operations, providing good technical support for the target classification and recognition function of a SONAR system.

At the same time, there is still a lot of worthwhile research to do based on our work, such as:

(1) Using simulation data or data augmentation to expand the dataset, improve the recognition and generalization ability under sample imbalance;

(2) Using auditory attention-inspired mechanisms to improve model interpretability and recognition performance under complex target working conditions.

These will be the research directions of our future works.

Author Contributions: Conceptualization, F.J. and J.N.; methodology, F.J. and J.N.; software, J.N. and F.J.; validation, J.N., F.J. and G.L.; formal analysis, J.N., F.J. and G.L.; investigation, J.N. and F.J.; resources, J.N. and Y.W.; data curation, J.N. and L.L.; writing—original draft preparation, J.N. and Y.W.; writing—review and editing, J.N., L.L. and Y.W.; visualization, J.N. and F.J.; supervision, F.J.; project administration, F.J.; funding acquisition, F.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under grant number 51409239.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available in a publicly accessible repository. Datasets are openly available at http://atlanttic.uvigo.es/underwaternoise/ (accessed on 1 July 2023) at 10.1016/j.apacoust.2016.06.008 in ref. [6] and at https://github.com/irfankamboh/DeepShip/ (accessed on 1 July 2023) at 10.1016/j.eswa.2021.115270 in ref. [7].

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviation

Abbreviation	Full name
UATR	Underwater acoustic target recognition
DRACNN	Deep residual attention convolutional neural network
DCGAN	Deep conditional generative adversarial network
LSTM	Long short-term memory
RACM	Residual attention convolution module
SE	Squeeze excitation
FEM	Feature extraction model
DCB	Deep convolutional block
MCC	Multi-convolutional layer jumper connection
CM	Classification module
GAP	Global average pooling

References

1. Xu, Y.C.; Cai, Z.M.; Kong, X.P. Improved pitch shifting data augmentation for ship-radiated noise classification. *Appl. Acoust.* **2023**, *221*, 109468. [CrossRef]
2. Li, G.H.; Bu, W.J.; Yang, H. Research on noise reduction method for ship radiate noise based on secondary decomposition. *Ocean. Eng.* **2023**, *268*, 113412. [CrossRef]
3. Esmaiel, H.; Xie, D.; Qasem, Z.A.; Sun, H.; Qi, J.; Wang, J. Multi-Stage Feature Extraction and Classification for Ship-Radiated Noise. *Sensors* **2021**, *22*, 112. [CrossRef] [PubMed]
4. Ni, J.S.; Zhao, M.; Hu, C.Q.; Lv, G.T.; Guo, Z. Ship Shaft Frequency Extraction Based on Improved Stacked Sparse Denoising Auto-Encoder Network. *Appl. Sci.* **2022**, *12*, 9076. [CrossRef]
5. Li, Y.X.; Tang, B.Z.; Jiao, S.B. Optimized Ship-Radiated Noise Feature Extraction Approaches Based on CEEMDAN and Slope Entropy. *Entropy* **2022**, *24*, 1265. [CrossRef]
6. Santos-Domínguez, D.; Torres-Guijarro, S.; Cardenal-López, A.; Pena-Gimenez, A. ShipsEar: An underwater vessel noise database. *Appl. Acoust.* **2016**, *113*, 64–69. [CrossRef]
7. Irfan, M.; Zheng, J.B.; Ali, S.; Iqbal, M.; Masood, Z. DeepShip: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification. *Expert Syst. Appl.* **2021**, *183*, 115270. [CrossRef]
8. Chen, J.; Han, B.; Ma, X.F.; Zhang, J. Underwater Target Recognition Based on Multi-Decision LOFAR Spectrum Enhancement: A Deep-Learning Approach. *Future Internet* **2021**, *13*, 265–285. [CrossRef]
9. Hong, G.; Suh, D. Mel Spectrogram-based advanced deep temporal clustering model with unsupervised data for fault diagnosis. *Expert Syst. Appl.* **2023**, *217*, 119511. [CrossRef]
10. Meng, L.X.; Xu, X.L.; Zuo, Y.B. Fault feature extraction of logarithmic time-frequency ridge order spectrum of planetary gearbox under time-varying conditions. *J. Vib. Shock.* **2020**, *39*, 163–169. [CrossRef]
11. Wen, L.; Li, X.; Li, X.; Gao, L. A New Transfer Learning Based on VGG-19 Network for Fault Diagnosis. In Proceedings of the 2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD), Porto, Portugal, 6–8 May 2019; pp. 205–209. [CrossRef]
12. Triyadi, A.B.; Bustamam, A.; Anki, P. Deep Learning in Image Classification using VGG-19 and Residual Networks for Cataract Detection. In Proceedings of the 2022 2nd International Conference on Information Technology and Education (ICIT&E), Malang, Indonesia, 22 January 2022; pp. 293–297. [CrossRef]
13. Hong, F.; Liu, C.W.; Guo, L.J.; Chen, F.; Feng, H.H. Underwater Acoustic Target Recognition with a Residual Network and the Optimized Feature Extraction Method. *Appl. Sci.* **2021**, *11*, 1442. [CrossRef]
14. Li, J.; Wang, B.X.; Cui, X.R.; Li, S.B.; Liu, J.H. Underwater Acoustic Target Recognition Based on Attention Residual Network. *Entropy* **2022**, *24*, 1657. [CrossRef]
15. Li, P.; Wu, J.; Wang, Y.X.; Lan, Q.; Xiao, W.B. STM: Spectrogram Transformer Model for Underwater Acoustic Target Recognition. *J. Mar. Sci. Eng.* **2022**, *10*, 1428. [CrossRef]
16. Luo, X.W.; Zhang, M.H.; Liu, T.; Huang, M.; Xu, X.G. An Underwater Acoustic Target Recognition Method Based on Spectrograms with Different Resolutions. *J. Mar. Sci. Eng.* **2021**, *9*, 1246–1265. [CrossRef]
17. Gao, Y.; Chen, Y.; Wang, F.; He, Y. Recognition Method for Underwater Acoustic Target Based on DCGAN and DenseNet. In Proceedings of the 2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC), Beijing, China, 10–12 July 2020; pp. 215–221. [CrossRef]
18. Hu, G.; Wang, K.J.; Liu, L.L. Underwater Acoustic Target Recognition Based on Depthwise Separable Convolution Neural Networks. *Sensors* **2021**, *21*, 1429–1448. [CrossRef] [PubMed]
19. Li, J.H.; Yang, H.H. The underwater acoustic target timbre perception and recognition based on the auditory inspired deep convolutional neural network. *Appl. Acoust.* **2021**, *182*, 108210. [CrossRef]
20. Song, X.P.; Cheng, J.S.; Gao, Y. A New Deep Learning Method for Underwater Target Recognition Based on One-Dimensional Time-Domain Signals. In Proceedings of the 2021 OES China Ocean Acoustics (COA), Harbin, China, 14–17 July 2021; pp. 1048–1051. [CrossRef]
21. Yang, H.H.; Li, J.H.; Sheng, M.P. Underwater acoustic target multi-attribute correlation perception method based on deep learning. *Appl. Acoust.* **2022**, *190*, 108644. [CrossRef]
22. Ni, J.S.; Hu, C.Q.; Zhao, M. Recognition method of ship radiated noise based on VMD and improved CNN. *J. Vib. Shock.* **2023**, *42*, 74–82. [CrossRef]
23. Yin, F.; Li, C.; Wang, H.B.; Nie, L.X.; Zhang, Y.L.; Liu, C.R.; Yang, F. Weak Underwater Acoustic Target Detection and Enhancement with BM-SEED Algorithm. *J. Mar. Sci. Eng.* **2023**, *11*, 357–373. [CrossRef]
24. Yao, Q.H.; Wang, Y.; Yang, Y.Y. Underwater Acoustic Target Recognition Based on Data Augmentation and Residual CNN. *Electronics* **2023**, *12*, 1206–1222. [CrossRef]
25. Woo, S.H.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Volume 11211, pp. 3–19. [CrossRef]
26. Malla, P.P.; Sahu, S.; Alotaibi, A.I. Classification of Tumor in Brain MR Images Using Deep Convolutional Neural Network and Global Average Pooling. *Processes* **2023**, *11*, 679–695. [CrossRef]
27. Huang, G.; Liu, Z.; Maaten, L.V.D.; Kilian, Q.W. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]

28. Pathak, D.; Raju, U. Shuffled-Xception-DarkNet-53: A content-based image retrieval model based on deep learning algorithm. *Comput. Electr. Eng.* **2023**, *107*, 108647. [CrossRef]
29. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. RepVGG: Making VGG-style ConvNets Great Again. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 13728–13737. [CrossRef]
30. Liu, F.; Shen, T.S.; Luo, Z.L.; Zhao, D.; Guo, S.J. Underwater target recognition using convolutional recurrent neural networks with 3-D Mel-spectrogram and data augmentation. *Appl. Acoust.* **2021**, *178*, 107989. [CrossRef]
31. Ke, X.Q.; Yuan, F.; Chen, E. Underwater Acoustic Target Recognition Based on Supervised Feature-Separation Algorithm. *Sensors* **2018**, *18*, 4318–4341. [CrossRef] [PubMed]
32. Feng, S.; Zhu, X.Q. A Transformer-Based Deep Learning Network for Underwater Acoustic Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1505805. [CrossRef]
33. Hsiao, S.F.; Tsai, B.C. Efficient Computation of Depthwise Separable Convolution in MobileNet Deep Neural Network Models. In Proceedings of the 2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), Penghu, Taiwan, China, 15–17 September 2021; p. 9602973. [CrossRef]