electronics

MDPI

Communication

A Robust Feature Extraction Method for Underwater Acoustic Target Recognition Based on Multi-Task Learning

Daihui Li [1], Feng Liu [1,*], Tongsheng Shen [1,*], Liang Chen [1,2] and Dexin Zhao [1]

[1] National Innovation Institute of Defense Technology, Chinese Academy of Military Science, Beijing 100000, China
[2] Institute of Ocean Engineering and Technology, Zhejiang University, Zhoushan 316000, China
* Correspondence: liufeng_cv@126.com (F.L.); shents_bj@126.com (T.S.)

Abstract: Target classification and recognition have always been complex problems in underwater acoustic signal processing because of noise interference and feature instability. In this paper, a robust feature extraction method based on multi-task learning is proposed, which provides an effective solution. Firstly, an MLP-based network model suitable for underwater acoustic signal processing is proposed to optimize feature extraction. Then, multi-task learning is deployed on the model in hard parameter-sharing so that the model can extract anti-noise interference features and embed prior feature extraction knowledge. In the model training stage, the simultaneous training method enables the model to improve the robustness and representation of classification features with the knowledge of different tasks. Furthermore, the optimized classification features are sent to the classification network to complete target recognition. The proposed method is evaluated by the dataset collected in the real environment. The results show that the proposed method effectively improves recognition accuracy and maintains high performance under different noise levels, which is better than popular methods.

Keywords: underwater acoustics; target recognition; multi-task learning; robust feature extraction; intelligent algorithm

check for updates

Citation: Li, D.; Liu, F.; Shen, T.; Chen, L.; Zhao, D. A Robust Feature Extraction Method for Underwater Acoustic Target Recognition Based on Multi-Task Learning. *Electronics* **2023**, *12*, 1708. https://doi.org/10.3390/electronics12071708

Academic Editor: Miin-shen Yang

Received: 1 February 2023
Revised: 30 March 2023
Accepted: 31 March 2023
Published: 4 April 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Underwater acoustic target recognition based on ship-radiated noise received by a hydrophone is a research hotspot. Signal information received by a hydrophone depends on the target characteristics and the marine environment. Features of the signal are closely related to the route state and mechanical working state of the target, which are complex and challenging to describe. The marine environment is usually accompanied by different noise levels, which will weaken the target features and reduce the discrimination of the target. Therefore, the importance of robust feature extraction ability for recognition algorithms is ineffable.

In underwater acoustic target recognition, classical features include time-domain waveform features [1,2], frequency and time–frequency features [3–9], and auditory perception features [10–14]. However, features with larger dimensions are often redundant and difficult to process, and features with smaller dimensions cause a lot of information loss to varying degrees. In addition, the application scenarios and scope of different types of features are limited, and features with different dimensions and complexity have different requirements for the design of classifiers. These reasons lead to poor generalization and limitations of classification models. With the rapid development of deep learning technology, it is a foreseeable reality to complete high-quality feature extraction in large-dimensional features with rich information and even original signals. Deep learning technology promotes advanced intelligent algorithms. Its powerful data learning ability provides the model with strong feature extraction performance, ensuring that it obtains good results in

underwater acoustic target location [15–17] and recognition [18–25]. Many researchers have proposed corresponding algorithms from the perspective of improving feature extraction. Qi et al. [26] proposed an integrated neural network based on feature fusion learning for underwater acoustic target recognition. This method extracts the short-time Fourier transform (STFT) amplitude spectrum, STFT phase spectrum, and bispectrum features of underwater acoustic signals to form the network's input. It uses a shuffled frog leaping algorithm (SFLA) to train the weight coefficients of different networks, achieving higher recognition accuracy and stronger noise robustness. Luo et al. [27] used the restricted Boltzmann machine (RBM) to automatically encode the combined data of the power spectrum and demodulation spectrum of ship-radiated noise without supervision and extract the deep data structure layer by layer to obtain the signal feature vector. Tian et al. [24] proposed a multi-scale residual deep neural network (MSRDN) to construct a deep convolutional stack network. The problem of feature extraction using large convolution kernels in the initial stage of neural networks is improved to avoid the lack of depth and structural imbalance in the network. MSRDN can directly use the original signal waveform as the input and achieve high recognition performance after training. Doan et al. [25] proposed a dense model for underwater target recognition. The proposed model skillfully reuses all former feature maps to optimize recognition accuracy under various impaired conditions while satisfying low computational costs. Cao et al. [21] proposed a second-order pooling convolutional neural network (CNN) model to capture temporal correlation, which improved the performance of maximum pooling in CNN applied to underwater acoustic target recognition. Wang et al. [20] proposed a dimension reduction method to obtain the multi-dimensional fusion features of the original underwater acoustic signal, which ensures the time dimension's consistency. Additionally, the Gaussian mixture model (GMM) was used to modify the structure of the deep neural network (DNN) to obtain high accuracy and strong adaptability. Ke et al. [28] proposed a one-dimensional convolutional autoencoder-decoder model to extract features from high resonance components, proposed a supervised feature separation algorithm to separate further the features extracted in pre-trained, and finally increased the recognition rate. Most methods for improving feature extraction focus on two aspects: 1. improving the primary features of input by signal processing or feature fusion; and 2. improving the structure of the classification neural network by model optimization. Although these representative deep learning-based methods have achieved acceptable results in underwater acoustic target recognition tasks, directly inputting features into the black box classifier for training shields the internal working mechanism of the model and reduce the interpretability and performance.

This paper presents a robust feature extraction method (RFEM) that adds prior knowledge to improve feature robustness and model performance based on multi-task learning. The proposed method not only extracts high-level features using the classification task-driven model but also guides the model to extract anti-noise interference features and learn prior knowledge based on multi-task learning. RFEM learns to extract manual features based on prior knowledge while learning to resist noise interference, and uses this information to extract small-dimensional robust features to improve the recognition performance of the model. Specifically, the proposed method designs a multi-layer perceptron-based (MLP) module to suppress noise interference on the signal. It generates a robust feature based on multi-task learning that integrates time–frequency, hand-designed, and specific task requirement features. The proposed method is efficient, has practical value, and can be combined with other advanced deep learning-based methods after simple improvement.

The following sections are divided into three parts. Section 2 introduces the details of RFEM, including the components of the MLP module, the training method and inference stage of RFEM, and the loss function. Experiments and discussions are described in Section 3, including multiple ablation experiments and identification experiments under different signal-to-noise ratios. Finally, the conclusion is presented in Section 4.

2. Proposed Method

Figure 1 shows the details of RFEM with a recognition system as an example, including the training method and inference stage. In short, RFEM utilizes a particular feature extraction network to extract robust features. The extracted features are sent to the subsequent classification network for high-level feature extraction and target discrimination. In the training stage of the model, the robust feature extraction network is trained on an anti-noise task, an a priori knowledge-based feature extraction task, and a classification task based on multi-task learning. The network finds a balanced feature in the three tasks to make it suitable for signal recognition under different noise levels. Given the complexity of the underwater acoustic target radiated noise signal, this paper designs an MLP model to extract features suitable for different tasks. In addition, various neural networks suitable for underwater acoustic target recognition can be adapted for the classification network. In the training stage, the damaged signal and the original signal are used to train the robust feature extraction network and the classification network. After the training, two networks can be cascaded to construct the recognition system.

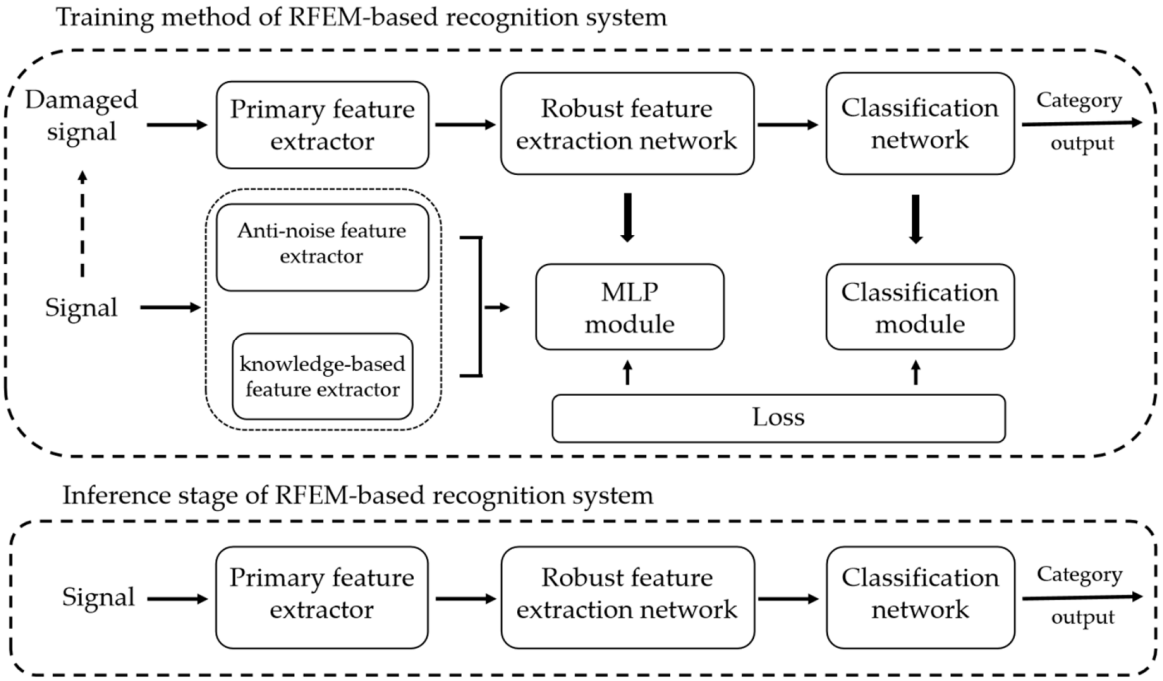

Figure 1. The illustration of an RFEM-based recognition system.

As shown in Figure 1, RFEM is introduced with a recognition system as an example. RFEM includes the design of the basic block of the MLP (BBM) module, the design of the robust feature extraction network, the loss and the training method. The three parts are described in detail in the following parts.

2.1. Basic Block of the MLP Module

The underwater acoustic channel is a complex time-varying space-varying channel, which makes various characteristics of ship-radiated noise time-varying. In addition, the mechanism of ship-radiated noise is complex, which increases the difficulty of feature extraction. Figure 2 shows the spectrogram of some underwater acoustic targets. It can be seen that some ship-radiated noise has stable characteristics at specific frequencies, and there is also a certain level of time variation, which is typical for a non-stationary signal. These spectrograms are often used as primary features in deep learning-based target recognition algorithms.

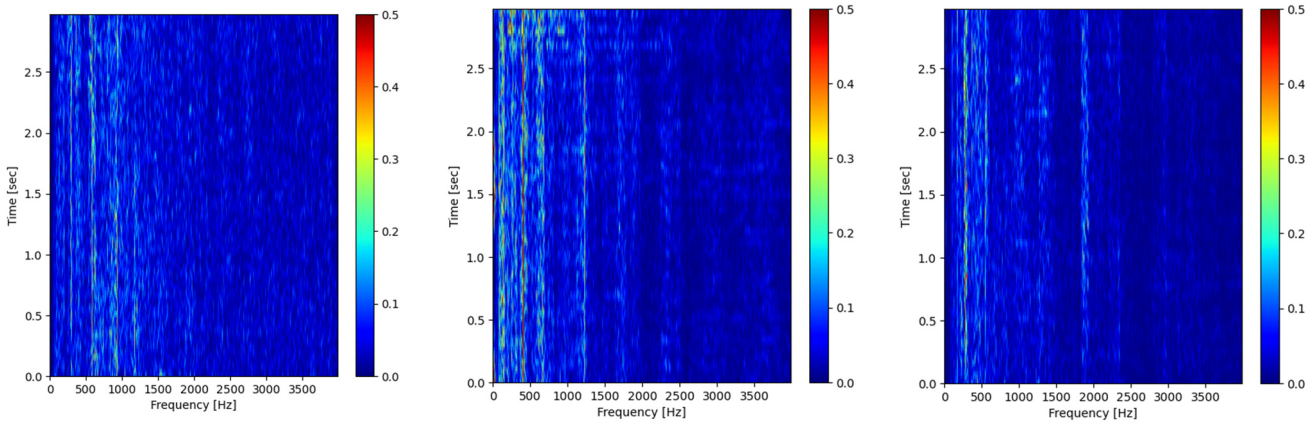

Figure 2. The spectrogram of underwater acoustic targets. (**a**) Passenger ship; (**b**) ocean liner; (**c**) fishing boat.

Convolutional neural networks (CNNs) are widely used in image and speech processing. The working mode of CNN provides it with an intrinsic advantage of establishing local spatial relations. For the complex cross-regional relations, CNN needs to rely on layer-by-layer stacking of convolutional layers to improve the receptive field of the neural network and establish it. For images or speech with strong local spatial relations, CNN or time-delay neural networks will be particularly suitable. However, there are no strong local spatial relations for underwater acoustic target signals, especially ship-radiated noise signals. Moreover, much of the available information may be lost in the local space due to the comb filtering characteristics of the underwater acoustic channel and ocean background noise. Therefore, the neural network that establishes local spatial relations layer by layer is unsuitable for underwater acoustic signal processing. The neural network needs to establish cross-regional relations to expand the range of feature searches. This paper proposes the BBM to establish a global cross-regional relation in each layer, so that the neural network can extract stable and reliable robust features as much as possible, especially for damaged signals. The overall architecture of the proposed BBM is shown in Figure 3.

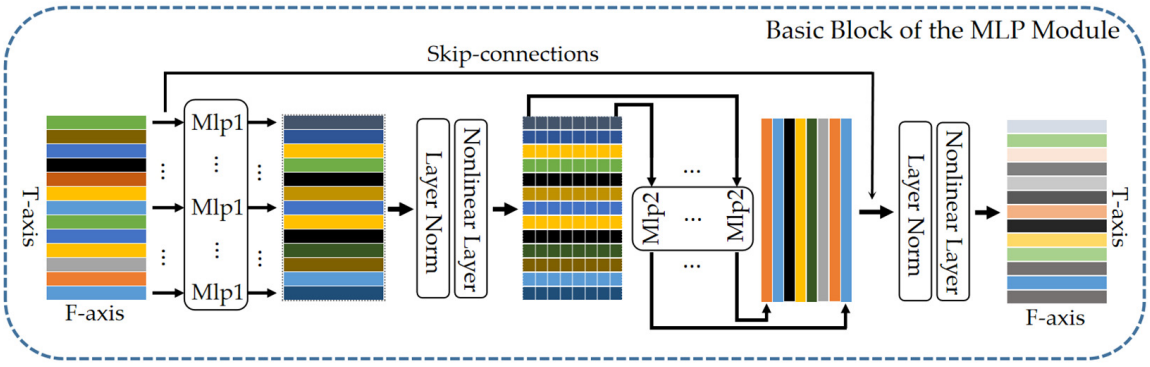

Figure 3. Schematic diagram of basic block of the MLP module.

The underwater acoustic target signal is subjected to time–frequency transformation to obtain a two-dimensional spectrogram. These two dimensions represent frequency and time, respectively, as shown in Figure 2. The two dimensions of input or output features of the BBM are defined as the time-axis (T-axis) and the frequency-axis (F-axis). If the spectrogram is directly input into BBM, the time dimension is represented by the T-axis, and the F-axis represents the frequency dimension. For the middle layer of the network, the T-axis still denotes the time dimension after the feature encoding, and F-axis denotes the frequency dimension after the feature encoding.

BBM divides features into patches based on the T-axis and F-axis. This division method enables neural networks to build long-distance relationships in a single network layer. Additionally, the division of patches is similar to VIT [29], but patches of BBM are divided according to the horizontal or vertical axis. The patches divided by the T-axis are called F-patches, and the patches divided by the F-axis are called T-patches. BBM consists of two MLPs [30], Mlp1 and Mlp2. As shown in Figure 3, BBM first sends F-patches of input features to Mlp1 for feature extraction, and Mlp1 will be encoded according to the F-patches. This encoding method enables the neural network to establish a unique frequency encoding, so that it can reconstruct the concept of frequency in its own way. Secondly, BBM sends the T-patches of the Mlp1 output to Mlp2. Mlp2 encodes a single frequency encoded value in the whole period, so the neural network can easily distinguish the frequency range with a significant response, and the frequency shielded by noise and other interference in a small range of time is compensated. Finally, the residual structure [31] is introduced into the BBM to avoid the disappearance of the gradient, and the features after nonlinear activation are output for the next block. BBM can establish a global feature relation and is suitable for processing ship-radiated noise signals. In the next section, this paper will construct a robust feature extraction network based on BBM, namely the MLP module.

2.2. Robust Feature Extraction Network

This section introduces the MLP module design based on the BBM constructed above, called robust feature extraction network. The construction of a robust feature extraction network aims to solve three defects in the classical deep learning-based underwater acoustic target recognition algorithm: 1. the classic recognition network is trained on a spectrogram, and the recognition accuracy is limited due to insufficient feature extraction performance; 2. the anti-noise ability of classical features is insufficient; and 3. the manual features extracted based on human prior are challenging to fuse with the features extracted automatically by the machine. The MLP module for extracting robust features is shown in Figure 4.

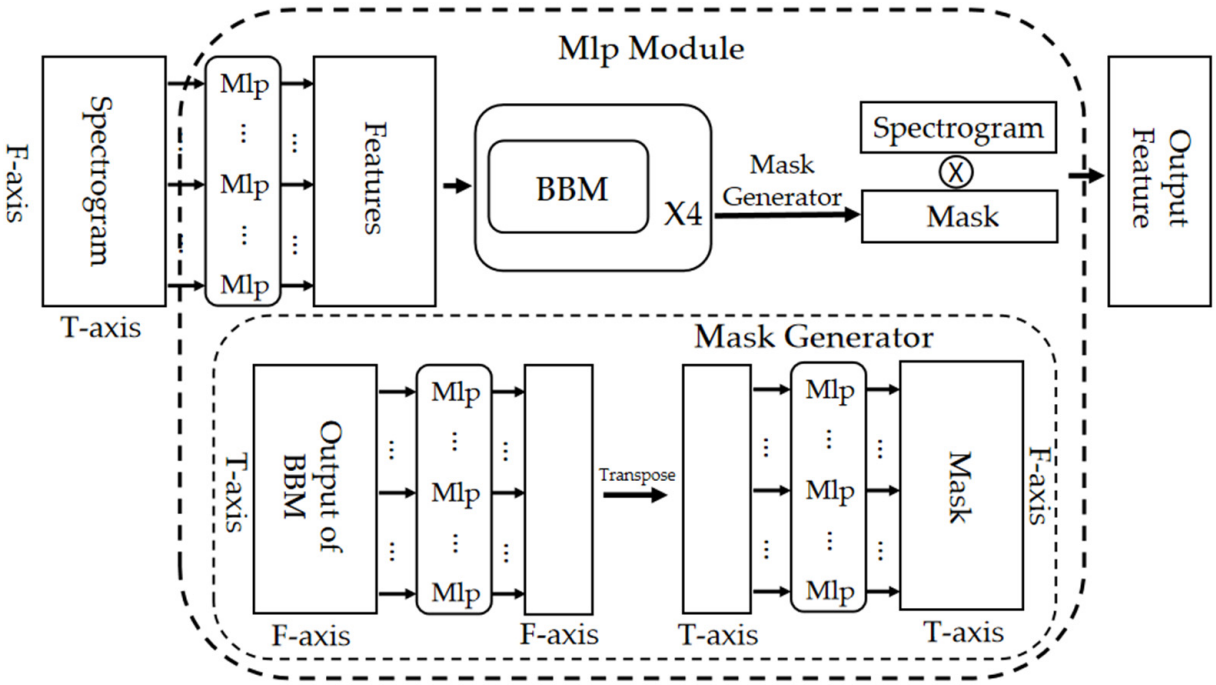

Figure 4. Schematic diagram of the MLP module.

In this paper, the MLP module uses an MLP unit to encode the time information, and the encoded features are sent to the BBM sub-module for feature extraction. The BBM

submodule is composed of four BBMs in series, and its output is sent to a mask generator to generate a mask. The mask is multiplied by the input spectrum to filter noise and extract robust features suitable for the classification network. The MLP module provides a neural network suitable for processing underwater acoustic signals. The next section introduces the method of training the network.

2.3. Training Method of RFEM-Based Recognition Systems

There are no massive training samples for most underwater acoustic target intelligent recognition models. It is difficult for recognition algorithms to extract reliable features from noisy signals. In addition, many scholars have studied ship-radiated noise and proposed stable classical manual feature extraction methods, such as the frequency-selection method. Therefore, this paper proposes a multi-task learning method to train RFEM, so that RFEM learns to resist noise interference and learns to knowledge of frequency-selection to complete the extraction of manual features. The schematic diagram of the multi-task strategy is shown in Figure 5.

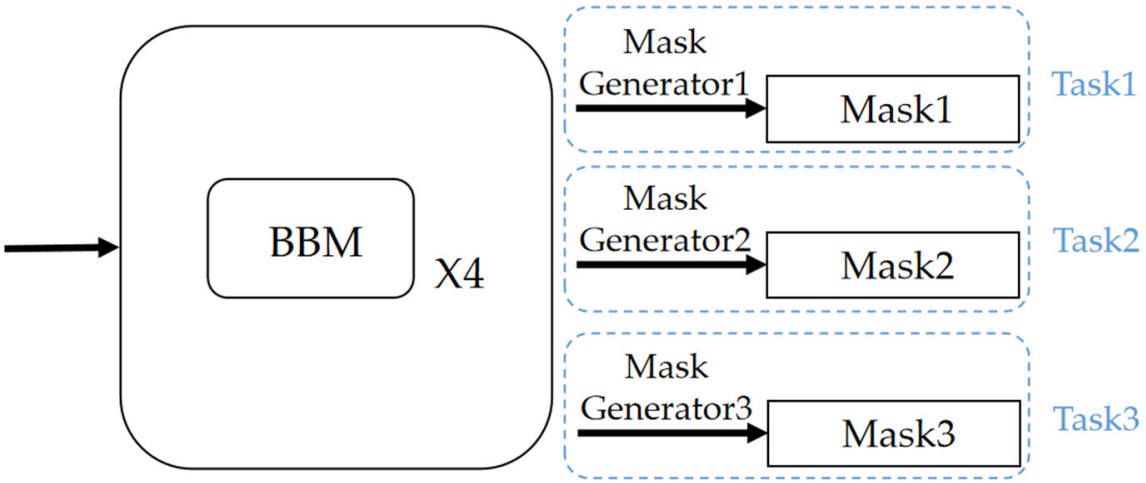

Figure 5. Schematic diagram of multi-task strategy.

For the MLP module, the single mask generator is replaced with three mask generators during the training stage. These three generators are used for different tasks: a robust feature extraction task for classification, an anti-noise task, and an optimized frequency-selection task. All three tasks employ the mask generators to generate masks, then use masks to shield interference information on the original spectrum and extract useful information. The three tasks are similar, so they have the effect of promoting each other.

2.3.1. Anti-Noise Task

The purpose of the anti-noise task is to make the model learn to resist different levels of noise and avoid interference in classification. Gaussian white noise with specified power is added to the original signal to form a damaged signal interfered by noise. The anti-noise ability of RFEM is trained on an anti-noise task based on these samples, as shown in Figure 6. The anti-noise task is called Task 1 in this paper.

The purpose of Task 1 is to optimize the damaged signal and extract features similar to the original signal. Gaussian white noise is added to the original signal according to Equation (1):

$$d = s + n \tag{1}$$

where n is Gaussian white noise, s is the original signal, and d denotes the damaged signal. The energy of the added signal is calculated according to Equation (2):

$$SNR = 10\lg\frac{E(s)}{E(n)} \tag{2}$$

where SNR represents the signal-to-noise ratio, $E(s)$ is the energy of the original signal, and $E(n)$ is the energy of the noise signal. It is worth noting that the signal-to-noise ratio here is not the actual signal-to-noise ratio of ship-radiated noise. The original and damaged signals are sent to the primary feature extractor to extract the primary features. The original signal is optionally autocorrelated before being fed into the primary feature extractor. The time–frequency feature used in this paper is the spectrum. After the primary feature extraction, the features of the damaged signal are input into RFEM, and the optimized features consistent with the input feature dimension are output. The model is trained to optimize the damaged signal and extract the same features as the original signal.

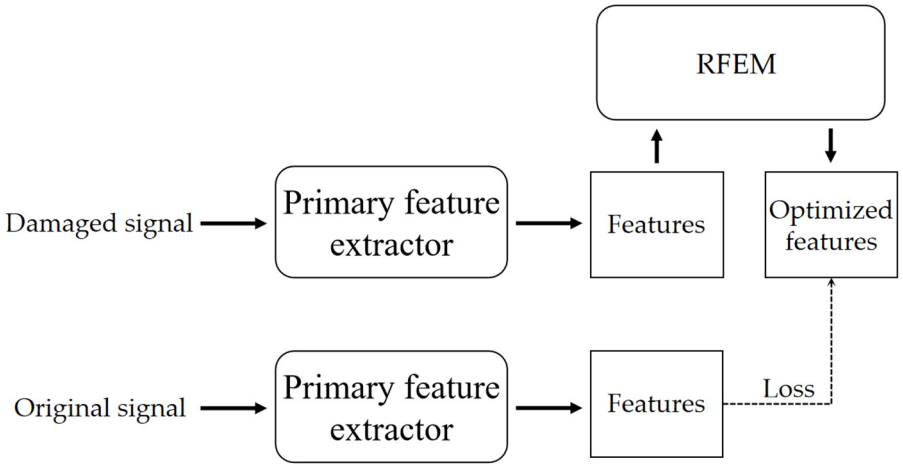

Figure 6. Schematic diagram of anti-noise task.

2.3.2. Optimized Frequency-Selection Task

Ships are equipped with many complex types of machinery, including power systems and other auxiliary mechanical systems. These machines inevitably produce friction, collision, and vibration when they work or the ship moves, thus spreading into the ocean. These noises are likely to contain specific frequency components for target recognition [8]. This paper introduces an optimized frequency selection method to select important frequency components. The frequency selection task is designed in the training stage, called Task 2. Time–frequency analysis of a ship-radiated noise signal can obtain the change in the frequency component with time. For the information obtained by time–frequency analysis, the intensity of the frequency component is calculated according to Equation (3):

$$F_j = \frac{1}{T}\sum_{i=0}^{T} S_{ij} \tag{3}$$

where S denotes the optimized time–frequency spectrum matrix, T represents the length of the time dimension, and F is the frequency intensity vector. The frequency components with high intensity are screened according to Equation (4):

$$A_j = \begin{cases} |F_j - \overline{F}|, j \le \frac{f}{a} \\ F_j - \overline{F}, j > \frac{f}{a} \end{cases} \tag{4}$$

where $\overline{F}$ denotes the mean value of frequency intensity vector, f represents the maximum frequency value, a represents the frequency threshold, and A represents a frequency selec-

tion vector. After obtaining A, another optimized frequency selection vector is calculated according to Equation (5):

$$B_j = F_j - \frac{1}{N}\sum_{c=0}^{N} k_c(F_j) \tag{5}$$

where $k_c(\cdot)$ denotes the c-th kernel average smoother, N represents the total number of kernel average smoother, and B represents the another frequency selection vector. After obtaining two frequency selection vectors, the frequency intensity vector is optimized according to Equation (6):

$$L_j = \begin{cases} B_j, A_j > 0, B_j > 0 \\ 0, other \end{cases} \tag{6}$$

where L denotes optimized frequency intensity vector. The optimized frequency-selection task is designed according to the optimized frequency intensity vector, as shown in Figure 7.

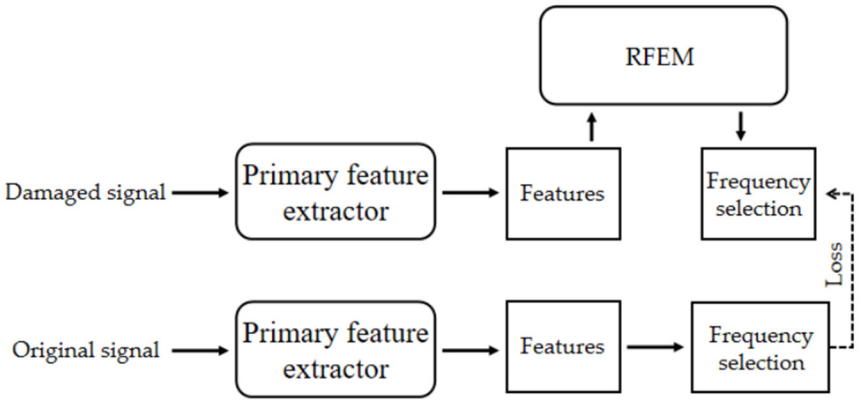

Figure 7. Schematic diagram of optimized frequency-selection task.

Based on the powerful feature extraction performance of RFEM, Task 2 guides the model learning to select the frequency suitable for classification. It is worth noting that when training the model based on Task 2, the frequency is still filtered based on the mask structure in Figure 5.

2.3.3. Training Strategy and Loss Function

The previous section introduces the anti-noise and optimized frequency-selection tasks in detail. This section introduces how to use these tasks to extract robust features and design loss functions. This paper presents a multi-task learning method that uses anti-noise tasks, optimized frequency-selection tasks as auxiliary tasks, and classification tasks as main tasks. The same optimized normalization approach is used to handle features in different tasks. Three tasks are collaboratively trained to search for common features automatically under the hard parameter-sharing mechanism. Driven by the training samples, the model is guided to extract robust features and complete target classification, as shown in Figure 8.

In the training stage, the damaged signal is first generated by original signal, and then the time–frequency analysis module is used to extract the time–frequency features. Finally, the time–frequency features of the damaged signal and the original signal are sent to RFEM to complete the three tasks. In this paper, Task 1 and Task 2 use the L1 loss training model, and Task 3 employs the negative likelihood loss training model. Although the three tasks are trained by independent loss, the three tasks are trained simultaneously according to Equation (7):

$$Loss = Loss1 + Loss2 + Loss3 \tag{7}$$

Simultaneous training and the hard parameter-sharing mechanism enable RFEM to extract robust features suitable for classification tasks.

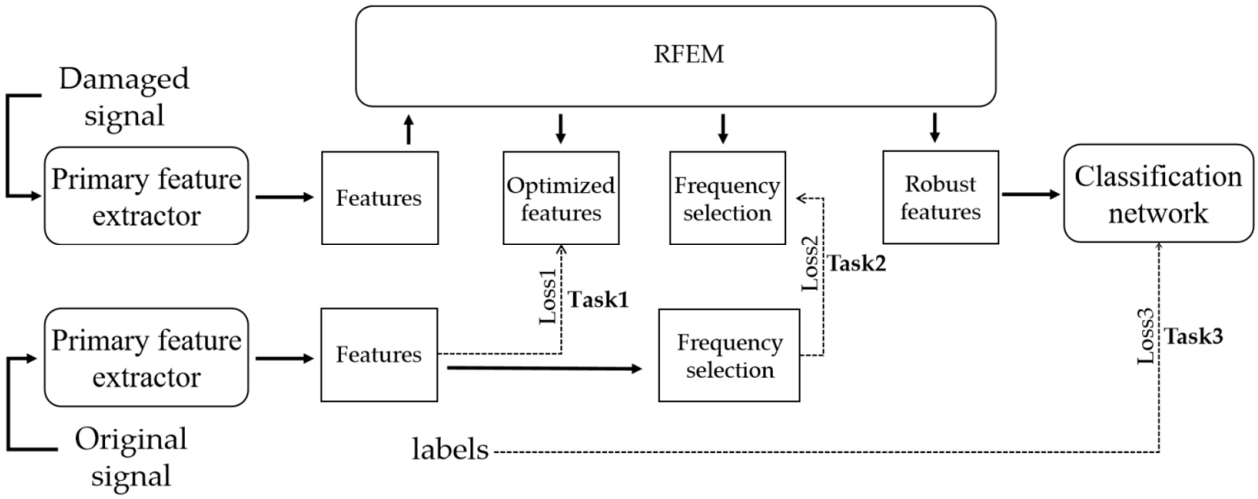

Figure 8. The training strategy of multi-task learning.

3. Experiments and Discussion

3.1. Experimental Dataset

This paper conducted experiments using ShipsEar [32], which was developed by the research group of the University of Diego on the Spanish Atlantic coast and is currently widely used by researchers. The dataset was recorded by hydrophones deployed from docks to capture different ship noises corresponding to docking or undocking maneuvers. The autonomous acoustic digitalHyd SR-1 recorder was used to record data. The recorder had a nominal sensitivity of −193.5 dB re 1V/1 uPa and a flat response in the 1 Hz-28 kHz frequency range. Annotated information includes recording technology and environmental and other conditions during collection. Finally, the dataset was made up of 90 recordings in wav. The recordings belonged to five categories: ocean noise and four different types of ship targets, as shown in Table 1.

Table 1. The type of noise contained in the dataset used in the experiments.

Category	Targets
Class-A	Ocean noise
Class-B	Fishing boats, trawlers, mussel boats, tugboats and dredgers
Class-C	Motorboats, pilot boats, sailboats
Class-D	Passenger liners
Class-E	Ocean liners and Ro-Ro vessels

Each category contained one or more targets, and the duration of each audio segment ranged from 15 s to 10 min. The data were pre-processed by removing the blank signal and segmenting all accords to a fixed duration of 3s, which resulted in 3626 labeled sound samples. Two dataset partitioning methods [33] were used in the experiments. The first method randomly sorted all samples, and the ratio of the training set to the test set was 4:1, named Dataset A. The other method involved taking only four types of target samples, and sorting the samples from the same record according to time. The samples in the front were test samples, and the samples in the back were training samples; the ratio of training data to test data was 3:1. The dataset divided in the second way was named Dataset B. Dataset B was more challenging than Dataset A in recognition tasks and is more suitable for practical applications, which is equivalent to using samples for a period of time to train the model to predict the category of targets for another time.

3.2. Experimental Setup

In this paper, three experiments were designed to verify the proposed method, including basic experiments, feature comparison experiments with observed data, and comparison experiments with published algorithms. The basic experiment compared the performance differences between the proposed method and the features extracted directly by the primary feature extractor. The primary features used here were classic short-time Fourier transform spectral features. The second experiment compared the proposed method with the popular feature-based method [34], including MFCC, F-Bank, and CQT. Additionally, the proposed method was compared with the published methods in the final experiment. The first two experiments were based on the challenging Dataset B to evaluate the proposed method comprehensively. Additionally, the last experiment used Dataset A to be consistent with the comparison method. Mfcc and F-bank had a window size of 2048, a jump length of 512, and the number of frequency bands was 40 and 128, respectively. In the experiment, we added different levels of Gaussian white noise to the observed data to simulate signals with different signal-to-noise ratios, and evaluated the robustness of the method to noise. Among them, noise was added to simulate different signal-to-noise ratio levels, ranging from -5 dB to 30 dB, to facilitate a more comprehensive demonstration of algorithm performance. The final experiment simulated signal-to-noise ratio levels from -10 dB to 5 dB, which was convenient to maintain consistency with the comparison methods.

The classic VGGish [35] was used as the classification network to construct the classifier. VGGish has been widely used, which can objectively reflect the classification performance brought about by feature improvement. All the raw audio recordings were resampled to 4 kHz. In the training stage, the Adam algorithm was used as an optimizer with the default parameters. The model had a total of 50,000 training steps, and the initial learning rate was 0.0001. When the number of training steps reached 30,000, the learning rate was reduced to one-tenth of the initial learning rate.

3.3. Basic Experiment

The basic experiment compared the performance differences between the proposed method and the features directly extracted by the primary feature extractor. In the experiment, the short-time Fourier spectrum extractor was used as the primary feature extractor. In other words, this part directly compared the performance between the robust features extracted by RFEM and the classical time-frequency features. During the experiment, classification networks were configured with the same parameters, and the following experiments also followed this rule. Precision, recall, f1-score, and accuracy were used as evaluation indicators. The experimental results are shown in Table 2.

Table 2. Classification accuracy of the basic experiment.

Model	STFT	Our
Precision	0.874	0.923
Recall	0.862	0.918
F1-score	0.867	0.920
Accuracy	0.875	0.926

Table 2 shows that the proposed method is ahead of STFT regarding various evaluation indicators, and the proposed method improves the accuracy by 5.1%, which proves that the proposed method can improve the recognition performance. In order to further verify the robustness of the proposed method, recognition experiments with different Gaussian noise levels were performed again on the dataset. In the experiment, Gaussian white noise with specified power was added to the data, and the results are shown in Figure 9.

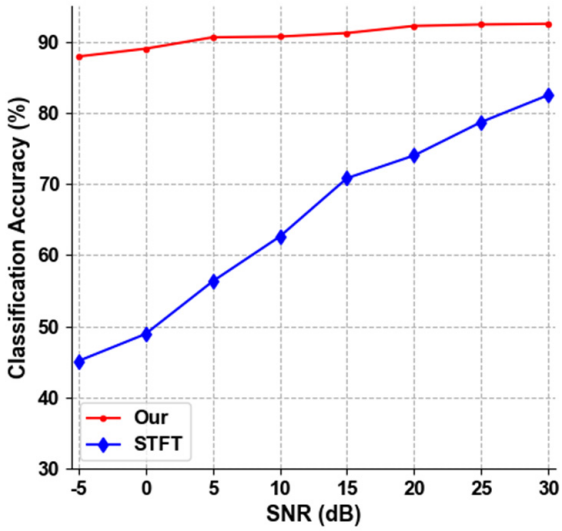

Figure 9. Experimental results with different noise levels.

Figure 9 shows that the proposed method significantly improves the anti-noise performance of the recognition model. Using the robust features generated by RFEM for classification, the accuracy is almost still the same when dealing with slight noise interference. Even under −5 dB, the accuracy is reduced by no more than 5%. In contrast, the classification model adapted to STFT features has feeble anti-noise performance. The two methods are based on the same primary feature extractor and classification network, but the performance is very different. The proposed method's RFEM and multi-task learning strategy effectively improve the feature extraction and classification performance.

3.4. Comparison of Popular Feature-Based Methods

In this experiment, the proposed method was compared with methods based on popular features, including MFCC, F-Bank, and CQT. These features are widely used in deep learning-based recognition methods. Experiments were conducted at different noise levels, and the results are shown in Figure 10.

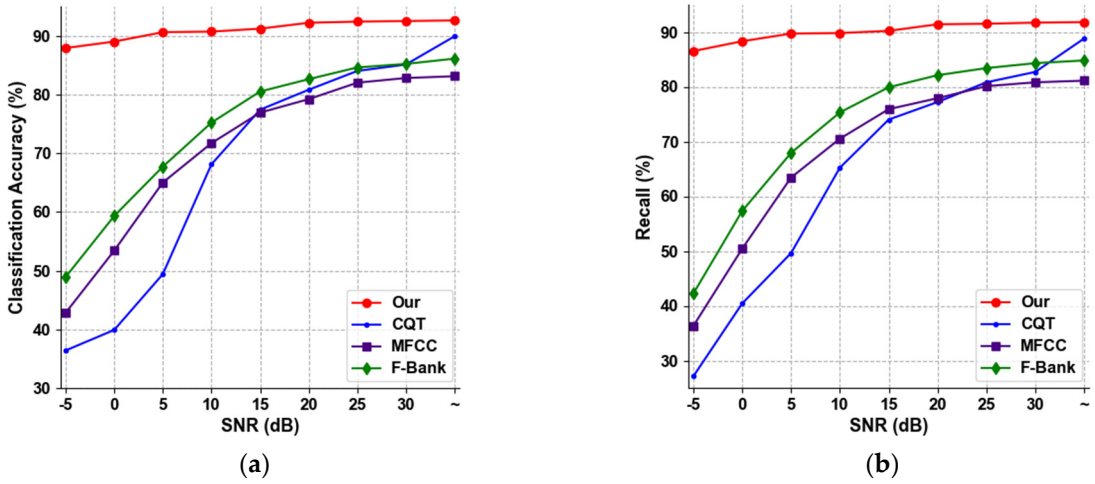

Figure 10. Experimental results with different features. (**a**) Classification accuracy; (**b**) Recall.

It can be seen from Figure 10 that the proposed method has better classification performance than several classical features without additional noise. For different levels of noise, different features show different performances. CQT performs well in the case of high SNR, but it cannot resist noise well. MFCC and F-Bank have similar anti-noise performance,

but F-Bank is slightly ahead of MFCC. It is worth mentioning that the proposed method is entirely ahead of these classical features.

3.5. Comparison of Published Methods

In order to better measure the performance of the proposed method, the method was compared with the recently published results. Experiments were arranged on Dataset A to maintain consistency with the experimental conditions of the comparison methods. We evaluated the method proposed in this paper five times, taking the appropriate average value and giving the error range. The experimental results are shown in Table 3.

Table 3. The recognition accuracy of the proposed method is compared with SVM [36,37], Simple-CNN [36,38,39], and MR-CNN-A [36].

SNR/dB	SVM	Simple-CNN	MR-CNN-A	Our
5	0.790	0.921	0.985	0.989 ± 0.003
0	0.752	0.868	0.955	0.986 ± 0.004
−5	0.695	0.738	0.917	0.978 ± 0.008
−10	0.643	0.726	0.884	0.944 ± 0.016

Compared with the recently published results, the proposed method presents an improved performance. At −10 dB, the recognition accuracy is 30.2%, 21.9%, and 6.1% higher than that of SVM, Simple-CNN, and MR-CNN-A, respectively. The recognition results further prove the effectiveness of RFEM and multi-task learning strategies.

4. Conclusions

In this paper, a neural network block suitable for underwater acoustic signal processing is designed, and a robust feature extraction method based on a multi-task strategy is proposed. The proposed neural network block establishes a global cross-regional relationship in a single block. Compared with the traditional convolutional neural network, it is more efficient and easy to establish a global receptive field, which is conducive to modeling signals using the neural network. Based on the proposed neural network block, a multi-task learning strategy is designed to learn anti-noise and prior knowledge-based feature extraction, which improves robust feature extraction and accuracy. Several experiments were conducted based on a public dataset. The results show that the proposed method presents an improved performance in anti-noise and accuracy compared with mainstream methods. In addition, this paper presents the concept of embedding prior knowledge into neural networks, which helps to promote the development of underwater acoustic target recognition methods.

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

Funding: This research was supported by the project of the National Natural Science Foundation of China (Grant No. 62201608).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository. The data presented in this study are openly available on 10.1016/j.apacoust.2016.06.008 in ref [32].

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

References

1. Meng, Q.; Yang, S.; Piao, S. The classification of underwater acoustic target signals based on wave structure and support vector machine. *J. Acoust. Soc. Am.* **2014**, *136*, 2265. [CrossRef]
2. Meng, Q.; Yang, S. A wave structure based method for recognition of marine acoustic target signals. *J. Acoust. Soc. Am.* **2015**, *137*, 2242. [CrossRef]
3. Rajagopal, R.; Sankaranarayanan, B.; Rao, P.R. Target classification in a passive sonar-an expert system ap-proach. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, NM, USA, 3–6 April 1990.
4. Boashash, B.; O'Shea, P. A methodology for detection and classification of some underwater acoustic signals using time-frequency analysis techniques. *IEEE Trans. Acoust. Speech Signal Process.* **1990**, *38*, 1829–1841. [CrossRef]
5. Ferguson, B. Time-frequency signal analysis of hydrophone data. *IEEE J. Ocean. Eng.* **1996**, *21*, 537–544. [CrossRef]
6. Ou, H.; Allen, J.S.; Syrmos, V.L. Automatic classification of underwater targets using fuzzy-cluster-based wavelet signatures. *J. Acoust. Soc. Am.* **2009**, *125*, 2578. [CrossRef]
7. Zeng, X.-Y.; Wang, S.-G. Bark-wavelet Analysis and Hilbert–Huang Transform for Underwater Target Recognition. *Def. Technol.* **2013**, *9*, 115–120. [CrossRef]
8. Liu, J.; He, Y.; Liu, Z.; Xiong, Y. Underwater Target Recognition Based on Line Spectrum and Support Vector Machine. In Proceedings of the 2014 International Conference on Mechatronics, Control and Electronic Engineering (MCE-14), Shenyang, China, 29–31 August 2014. [CrossRef]
9. Jahromi, M.S.; Bagheri, V.; Rostami, H.; Keshavarz, A. Feature Extraction in Fractional Fourier Domain for Classification of Passive Sonar Signals. *J. Signal Process. Syst.* **2018**, *91*, 511–520. [CrossRef]
10. Tucker, S. Auditory Analysis of Sonar Signals. Ph.D. Thesis, University of Sheffield, Sheffield, UK, 2001.
11. Li-Xue, Y.; Ke-An, C.; Bing-Rui, Z.; Yong, L. Underwater acoustic target classification and auditory feature identification based on dissimilarity evaluation. *Acta Phys. Sin.* **2014**, *63*, 134304. [CrossRef]
12. Wang, S.; Zeng, X. Robust underwater noise targets classification using auditory inspired time–frequency analysis. *Appl. Acoust.* **2014**, *78*, 68–76. [CrossRef]
13. Mohankumar, K.; Supriya, M.H.; Pillai, P.S. Bispectral Gammatone Cepstral Coefficient based Neural Network Classifier. In Proceedings of the IEEE Underwater Technology, Chennai, India, 23–25 February 2015; pp. 1–5. [CrossRef]
14. Zhang, L.; Wu, D.; Han, X.; Zhu, Z. Feature Extraction of Underwater Target Signal Using Mel Frequency Cepstrum Coefficients Based on Acoustic Vector Sensor. *J. Sens.* **2016**, *2016*, 7864213. [CrossRef]
15. Niu, H.; Gong, Z.; Ozanich, E.; Gerstoft, P.; Wang, H.; Li, Z. Deep-learning source localization using multi-frequency magnitude-only data. *J. Acoust. Soc. Am.* **2019**, *146*, 211–222. [CrossRef]
16. Liu, Y.; Niu, H.; Li, Z. A multi-task learning convolutional neural network for source localization in deep ocean. *J. Acoust. Soc. Am.* **2020**, *148*, 873–883. [CrossRef]
17. Cao, H.; Wang, W.; Su, L.; Ni, H.; Gerstoft, P.; Ren, Q.; Ma, L. Deep transfer learning for underwater direction of arrival using one vector sensor. *J. Acoust. Soc. Am.* **2021**, *149*, 1699–1711. [CrossRef]
18. Li, C.; Huang, Z.; Xu, J.; Yan, Y. Underwater target classification using deep learning. In *OCEANS 2018 MTS/IEEE Charleston*; IEEE: Piscataway, NJ, USA, 2018; pp. 1–5. [CrossRef]
19. Oikarinen, T.; Srinivasan, K.; Meisner, O.; Hyman, J.B.; Parmar, S.; Fanucci-Kiss, A.; Desimone, R.; Landman, R.; Feng, G. Deep convolutional network for animal sound classification and source attribution using dual audio recordings. *J. Acoust. Soc. Am.* **2019**, *145*, 654–662. [CrossRef]
20. Wang, X.; Liu, A.; Zhang, Y.; Xue, F. Underwater Acoustic Target Recognition: A Combination of Multi-Dimensional Fusion Features and Modified Deep Neural Network. *Remote Sens.* **2019**, *11*, 1888. [CrossRef]
21. Cao, X.; Togneri, R.; Zhang, X.; Yu, Y. Convolutional Neural Network with Second-Order Pooling for Underwater Target Classification. *IEEE Sens. J.* **2018**, *19*, 3058–3066. [CrossRef]
22. Wang, N.; He, M.; Sun, J.; Wang, H.; Zhou, L.; Chu, C.; Chen, L. ia-PNCC: Noise Processing Method for Underwater Target Recognition Convolutional Neural Network. *Comput. Mater. Contin.* **2019**, *58*, 169–181. [CrossRef]
23. Li, C.; Liu, Z.; Ren, J.; Wang, W.; Xu, J. A Feature Optimization Approach Based on Inter-Class and Intra-Class Distance for Ship Type Classification. *Sensors* **2020**, *20*, 5429. [CrossRef]
24. Tian, S.; Chen, D.; Wang, H.; Liu, J. Deep convolution stack for waveform in underwater acoustic target recognition. *Sci. Rep.* **2021**, *11*, 9614. [CrossRef]
25. Doan, V.-S.; Huynh-The, T.; Kim, D.-S. Underwater Acoustic Target Classification Based on Dense Convolutional Neural Network. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 1–5. [CrossRef]
26. Zhang, Q.; Da, L.; Zhang, Y.; Hu, Y. Integrated neural networks based on feature fusion for underwater target recognition. *Appl. Acoust.* **2021**, *182*, 108261. [CrossRef]
27. Luo, X.; Feng, Y.; Zhang, M. An Underwater Acoustic Target Recognition Method Based on Combined Feature with Automatic Coding and Reconstruction. *IEEE Access* **2021**, *9*, 63841–63854. [CrossRef]
28. Ke, X.; Yuan, F.; Cheng, E. Underwater Acoustic Target Recognition Based on Supervised Feature-Separation Algorithm. *Sensors* **2018**, *18*, 4318. [CrossRef]

29. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929. [CrossRef]
30. Tolstikhin, I.O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. MLP-Mixer: An all-MLP Architecture for Vision. *arXiv* **2021**, arXiv:2105.01601. [CrossRef]
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
32. 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]
33. Li, P.; Wu, J.; Wang, Y.; Lan, Q.; Xiao, W. STM: Spectrogram Transformer Model for Underwater Acoustic Target Recognition. *J. Mar. Sci. Eng.* **2022**, *10*, 1428. [CrossRef]
34. Irfan, M.; Jiangbin, Z.; Ali, S.; Iqbal, M.; Masood, Z.; Hamid, U. DeepShip: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification. *Expert Syst. Appl.* **2021**, *183*, 115270. [CrossRef]
35. Hershey, S.; Chaudhuri, S.; Ellis, D.P.W.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B.; et al. CNN architectures for large-scale audio classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 131–135. [CrossRef]
36. Ma, Y.; Liu, M.; Zhang, Y.; Zhang, B.; Xu, K.; Zou, B.; Huang, Z. Imbalanced Underwater Acoustic Target Recognition with Trigonometric Loss and Attention Mechanism Convolutional Network. *Remote Sens.* **2022**, *14*, 4103. [CrossRef]
37. Escalera, S.; Pujol, O.; Radeva, P. Separability of ternary codes for sparse designs of error-correcting output codes. *Pattern Recognit. Lett.* **2009**, *30*, 285–297. [CrossRef]
38. Wu, H.; Song, Q.; Jin, G. Deep Learning based Framework for Underwater Acoustic Signal Recognition and Classification. In Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence, Shenzhen, China, 8–10 December 2018; pp. 385–388. [CrossRef]
39. Jiang, J.; Shi, T.; Huang, M.; Xiao, Z. Multi-scale spectral feature extraction for underwater acoustic target recognition. *Measurement* **2020**, *166*, 108227. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.