sensors

Article

Underwater Acoustic Target Recognition Based on Depthwise Separable Convolution Neural Networks

Gang Hu [1,2], Kejun Wang [1,*] and Liangliang Liu [1]

[1] College of Automation, Harbin Engineering University, Harbin 150001, China; hugang@hrbeu.edu.cn (G.H.); liuliangliang@hrbeu.edu.cn (L.L.)
[2] College of Business, Anshan Normal University, Anshan 114007, China
* Correspondence: wangkejun@hrbeu.edu.cn

Abstract: Facing the complex marine environment, it is extremely challenging to conduct underwater acoustic target feature extraction and recognition using ship-radiated noise. In this paper, firstly, taking the one-dimensional time-domain raw signal of the ship as the input of the model, a new deep neural network model for underwater target recognition is proposed. Depthwise separable convolution and time-dilated convolution are used for passive underwater acoustic target recognition for the first time. The proposed model realizes automatic feature extraction from the raw data of ship radiated noise and temporal attention in the process of underwater target recognition. Secondly, the measured data are used to evaluate the model, and cluster analysis and visualization analysis are performed based on the features extracted from the model. The results show that the features extracted from the model have good characteristics of intra-class aggregation and inter-class separation. Furthermore, the cross-folding model is used to verify that there is no overfitting in the model, which improves the generalization ability of the model. Finally, the model is compared with traditional underwater acoustic target recognition, and its accuracy is significantly improved by 6.8%.

Keywords: underwater acoustic target; ship radiated noise; deep learning; depthwise separable convolution; dilated convolution

Citation: Hu, G.; Wang, K.; Liu, L. Underwater Acoustic Target Recognition Based on Depthwise Separable Convolution Neural Networks. *Sensors* **2021**, *21*, 1429. https://doi.org/10.3390/s21041429

Academic Editor: Iren E. Kuznetsova

Received: 2 February 2021
Accepted: 14 February 2021
Published: 18 February 2021

1. Introduction

Underwater acoustic target recognition technology is used to analyze ship radiated noise received by sonar and to judge the classification of the target [1,2], which has important economic and military value. Because of the complex marine environment and application of acoustic stealth technology, underwater acoustic target recognition has always been an internationally recognized problem. Traditional underwater acoustic target recognition methods based on ship radiated noise classify ship types by using artificially designed features and shallow classifiers, focusing on feature extraction and the development of nonlinear classifiers [3–8]. The features of artificially designed ship-radiated noise include waveform [9], spectrum [10], and wavelet [11], etc., which are dependent on expert knowledge and prior knowledge and have weak generalization ability. Shallow classifiers such as support vector machine (SVM) [12] and the shallow neural network classifier [13] have weak fitting and generalization abilities when dealing with complex and large numbers of samples. Generally speaking, classifier design and feature extraction are conducted independently, which may lead to the design of the feature not being suitable for the classification task. For example, in the classification model based on auditory features, auditory filter banks designed based on perceptual evidence tend to focus only on the property of signal description rather than the purpose of classification [14,15].

The human brain has strong abilities in perception, reasoning, induction, learning, etc. Inspired by the human neural structure and brain information processing mechanism, researchers proposed deep neural networks (DNNs) which processed information and decision-making in a similar way to brain [16,17]. Due to the emergence of deep learning technology, a complete deep learning model can not only realize the mathematical modeling for the original signals but also predict targets. These findings of auditory system research show that the acoustic signals of an auditory system in the time domain can be decomposed into frequency components; different regions of the auditory system perceive information of different frequency components; the brain uses information from all these regions to analyze and classify acoustic signals. In addition, research on the plasticity of the auditory cortex has demonstrated that the adult brain can be reshaped in appropriate environments [18].

The language model method has been widely used in natural language processing [18–20]. Drossos et al. applied the language model to detect acoustic events and achieved good effects [21]. The advantage of the language model method was that the input of the recursive neural network (RNN) could be adjusted according to the previous prediction of the classifier, and RNN could be used to process long-term dependencies in temporal information and to model class activities in context so as to learn longer in-class and inter-class time models. When the language model method was used for underwater acoustic target recognition, the performance could be improved by modeling these class dependencies [21].

Inspired by auditory perception and the language model, in this paper, a new depthwise separable convolutional neural network for underwater acoustic target recognition is proposed, which consists of a series of depthwise separable convolutions, integration layers, and time-dilated convolutions. The depthwise separable convolutions with variable convolution kernel width are used to decompose original time-domain ship-radiated noise signals into different frequency components, and extract signal features based on auditory perception. Due to the use of a variety of convolution kernels of different sizes, the model can achieve frequency decomposition and feature extraction with a variety of frequency and time precision. Compared with the traditional method of feature extraction based on frequency data, this method solves the contradiction between time precision and frequency precision well, and preserves the phase information of the model input signal to the maximum extent in the process of feature extraction. In the fusion layer, the one-dimensional feature vectors extracted at several consecutive moments are fused to form a two-dimensional feature matrix, which adds time information to the one-dimensional feature vectors. Finally, we use the time-dilated convolution for the modeling of long time attention, which can make full use of the intra-class and inter-class information for underwater acoustic target recognition just like the language model.

The remainder of this paper is organized as follows. In Section 2, we introduce related work in underwater acoustic target feature extraction and recognition. In Section 3, the method is proposed in detail. We describe the evaluation process in Section 4. In Section 5, experimental results are given and discussed. We give our conclusions in the last section.

2. Related Work

All existing studies on passive underwater acoustic target feature recognition with deep learning were still in a preliminary stage and focused on theoretical exploration and small-scale experiments. Generally speaking, applications of deep learning should be combined with big data. However, due to the limitations of practical conditions, it was often difficult to collect sufficient data for model training, which greatly limited the performance of deep neural network. Nevertheless, urgent demands still promoted continuous development of deep learning in passive underwater acoustic target recognition. Convolutional Neural Network (CNN) had a variety of applications in passive underwater target recognition because it was suitable for processing the original underwater acoustic signals and could obtain the implicit correlation that is difficult to be found by conventional feature analysis methods to a certain extent [22–25]. With internal feedback mechanism, RNN can process time-series signals. The audio signal is a typical sequence signal, which is provided with a memory function by RNN through circumferential joints of internal neurons. Therefore, the correlation of acoustic signals in the time dimension can be utilized dynamically. In RNN, there is a typical structure called "Long Short Term Memory" (LSTM) [20], which has been applied to passive underwater target recognition [26–28].

In a paper previously published on audio classification, CNN was replaced by depthwise separable (DWS) convolution [29,30]. DWS was a decomposition form of standard convolution, which decomposed a standard convolution into one convolution and one 1×1 convolution (called pointwise convolution) [31]. It firstly learned spatial information and then processed to cross-channel mode [32]. This convolutional decomposition for typical CNN resulted in less trainable parameters and memory occupation, and the reduction in computation complexity was $K_o^{-1} + (K_h \cdot K_w)^{-1}$, where K_h and K_w were the height and width of the CNN kernel, respectively, and K_o was the output channel of the CNN [27]. Dilated convolution was considered a method of improving the long-term learning capacity of the CNN [33]. In short, the kernel of the dilated convolution was dilated and there was a distance between two elements. As a result, the kernel of dilated convolution could be used on elements of its input patch with interval N (dilation factor), increasing the receptive field of the kernel instead of its parameters [34,35]. The kernel dilation could be used in any combination (for example, dilation in time dimension or feature dimension only) or all combinations of its dimensions. Li et al. provided a method to combine dilated convolution with RNN in audio classification task [36], which clearly focused on the exploration and learning of long-term patterns. Drossos et al. proposed an improved Convolutional Recursive Neural Network (CRNN) structure [31] which used DWS and dilated convolution with dilation in the time dimension only, i.e., time-dilated convolution. With discrete wavelet and time-dilated convolution, this structure had 85% fewer parameters than CRNN, but achieved better performance on typical audio classification datasets. The improvement of the dilated convolution showed that these convolutions had similar functions with RNN and could be used effectively for long-term contextual modeling.

For human beings, acoustic perception and recognition are accomplished through the auditory system, including the auditory periphery and auditory center. Frequency receptive fields in the auditory center, auditory cortex, auditory midbrain and other structures can adjust the frequency receptive fields and the optimal frequency to complete the learning task [37,38]. These findings about the auditory system indicated that the acoustic time-domain signals could be decomposed according to the frequency components in the auditory system. The decomposition of frequency components could be explained as product filtering for acoustic frequency-domain signals. Since the product of frequency domain signals is equal to the convolution of the time-domain signal [39], the frequency-domain component could be quickly realized by parallel computation of time-domain convolution; different regions of the auditory system perceived different frequency components; the brain collected information of all areas for analysis and for classifying acoustic signals. In addition, studies on the shaping of auditory cortex have shown that the adult brain could be reshaped in an appropriate environment. The auditory experience could change the functions and even structure of the auditory system.

3. Deep Convolution Neural Networks

3.1. The Structure of the Model

We propose a new deep convolution neural network model for feature extraction and classification of ship radiated noise, which includes a series of depthwise separable convolution, fusion layer and time-dilated convolution. The structure of the proposed model is shown in Figure 1.

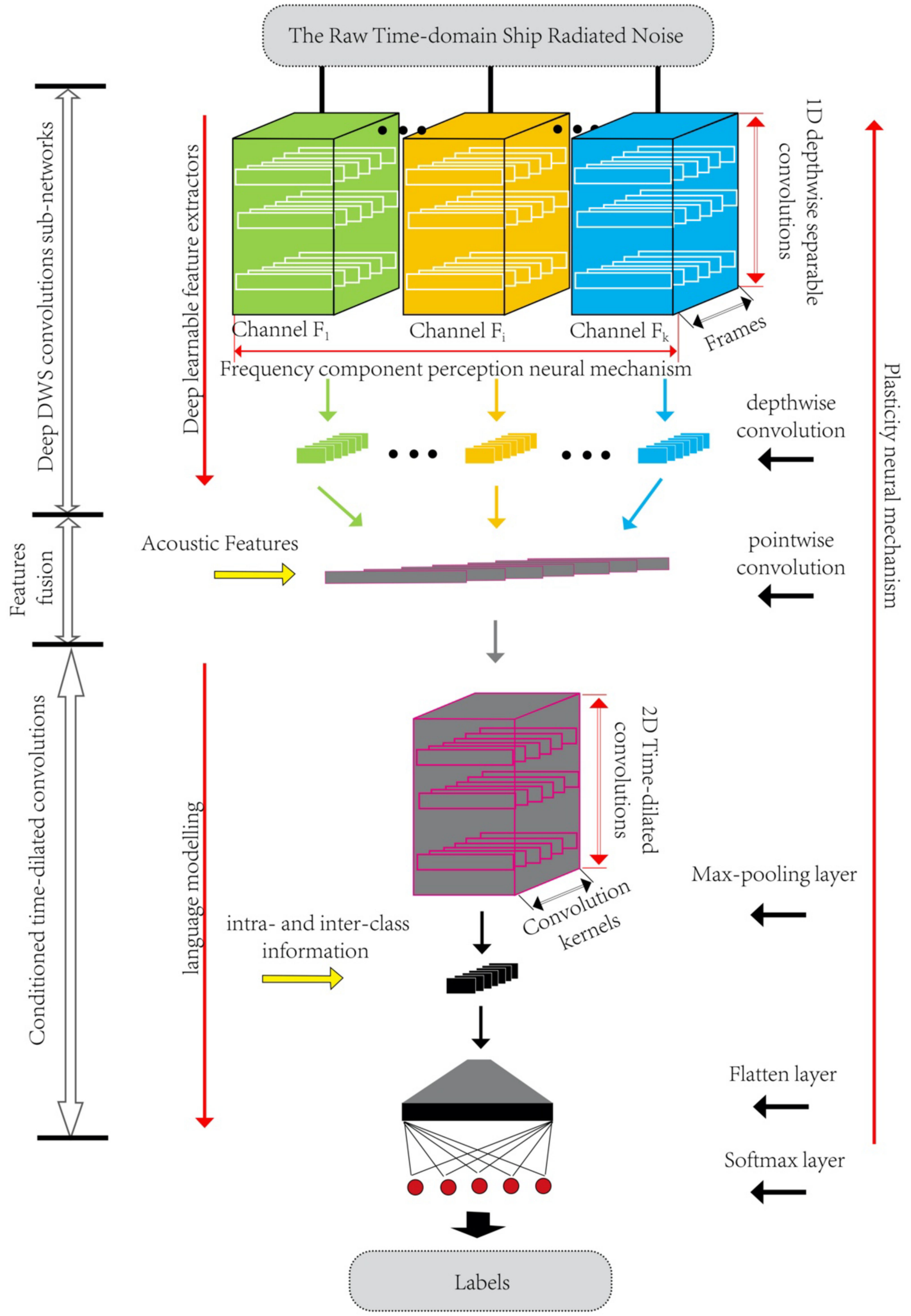

The Raw Time-domain Ship Radiated Noise
1D depthwise separable convolutions
Channel F1
Channel Fi
Channel Fk
Frames
Frequency component perception neural mechanism
Deep learnable feature extractors
Deep DWS convolutions sub-networks
depthwise convolution
Plasticity neural mechanism
Features fusion
Acoustic Features
pointwise convolution
Conditioned time-dilated convolutions
language modelling
2D Time-dilated convolutions
Convolution kernels
Max-pooling layer
intra- and inter-class information
Flatten layer
Softmax layer
Labels

Figure 1. The structure of depthwise separable convolutional neural networks.

The model proposed in this paper takes a sequence $X \in \mathbb{R}^{T \times N}$ of vectors T as input, and each vector T is composed of original underwater acoustic time-domain data with length N. A learnable feature extractor composed of depthwise separable convolution (DWS) and time-dilated convolution is used as a time pattern recognition The label vector of C classes corresponding to X is $Y = [y_1, \cdots, y_C]$, where $y_c \in \{0, 1\}$, represents whether the input belongs to class c. The model output is a vector $\hat{Y} = [\hat{y}_1, \cdots, \hat{y}_C]$, where $\hat{y}_c \in [0, 1]$ represents the prediction classification result of the model for underwater acoustic target when X is input. Inspired by the structure of extracting deep acoustic information of auditory system, we design a series of depthwise separable convolutions, takes the original underwater acoustic data as input and is completed by the one-dimensional depthwise separable convolution neural network. In addition, DWS convolution realizes the frequency decomposition for input signals and extract the features of decomposed signals.

The feature fusion layer realizes multi-channel feature information fusion by pointwise convolution at every moment. In the feature fusion layer, all one-dimensional acoustic features outputted by DWS filter at all T times are combined and analyzed comprehensively. The combined acoustic features are two-dimensional time mode features at T moments, which can be used as the input of the time-dilated convolution layer. The two-dimensional frequency convolution layer is applied to preserve locality and reduce frequency spectrum change ship-radiated noise. In the time-dilated convolution layer, in order to make use of the intra-class and inter-class activity mode, we adopt a time-dilated convolution like language modeling and use Softmax layer classifier to obtain a prediction probability for each ship of each sample. The classification of ship is taken as the target function, then driven by the original ship-radiated noise signal, and the learning and optimization are carried out in the whole training process. This optimization mechanism reflects the shaping neural mechanism of the auditory system.

This model can realize decomposition, feature extraction and classification for ship-radiated noise and be used for underwater acoustic target recognition.

3.2. *Depthwise Separable Convolution*

CNN is an artificial neural network signal (ANN) convolution which carries out a series of convolutions for input. The operation in CNN is equivalent to time-domain convolution in a traditional filter [40]. In this paper, multi-layer CNN is designed in each deep filter to realize filtering function, so we define it as a deep convolution filter . Through repeating the above process layer by layer, the multi-layer CNN construct can extract more abstract features from deep structure. However, deeper units may be indirectly connected to all or most of the signals. The receptive field of deep units in a depthwise convolution is larger than that of the shallow unit [41]. The parameters of the depthwise separable convolution filter are randomly initialized and learnt from ship-radiated noise. Driven by the time-domain signals of ship-radiated noise, the frequency decomposition ability of the depthwise separable convolution filter is learnable and adjustable. In addition, larger convolution kernels can contain longer wavelengths, which implies lower frequencies of components and vice versa. Thus, the learned filter is more suitable for the underwater acoustic target recognition task.

In order to learn spatial information and cross-channel information, we do not use the convolution of a single kernel, but the convolution of two different kernels in a series (that is, the output of the first is the input of the second). This decomposition technique is called deep separation convolution (DWN) and has been used in a variety of image processing structures (such as Exception, GoogleLeNet, Inception and MobileNets model). It has been proven that DWS convolution can reduce the number of parameters and improve performance [42–44]. DWS convolution consists of depthwise convolution and pointwise convolution. In the deep convolution layer, only one filter is applied per input channel. Pointise convolution is a simple 1 × 1 convolution that is then used to create linear combinations of the depthwise layer outputs. The learnable feature extractor consists of DWS convolution blocks. The lth DWS convolution block obtains output of the previous block as input, i.e., $H_{l-1} \in \mathbb{R}^{H_{l-1}^c \times H_{l-1}^h \times H_{l-1}^w}$, where H_{l-1}^c, H_{l-1}^h and H_{l-1}^w are the number, height and width of channels output by the $l-1$th DWS convolution block, respectively. $H_0 = X$ is the input time-domain signal of the ship-radiated noise, and $H_0^c = 1$, $H_0^h = 1$, $H_0^w = N$. The output of the l^{th} DWS convolution block is $H_l \in \mathbb{R}^{H_l^c \times H_l^h \times H_l^w}$, where H_l^c, H_l^h and H_l^w are the number, height and width of channel output by the l^{th} DWS convolution block, respectively. Each DWS convolution block includes one DWS convolution operation, one normalization, one down-sampling and one non-linear function.

The operation of DWS convolution itself includes two convolutions, one normalization process and one rectified linear unit (ReLU). The first convolution on the l^{th} DWS convolution block uses H_{l-1}^{c} number of kernels $K_l \in \mathbb{R}^{K_l^{dh} \times K_l^{dw}}$ and one bias deviation $a_l \in \mathbb{R}^{H_{l-1}^{c}}$ to learn spatial information input into H_{l-1}. K_l^{dh} and K_l^{dw} are the height and width of the kernel K_l; and s_l is the stride of the convolution kernel K_l. The depthwise convolution with a filter in each input channel (input depth) can be written as:

$$\begin{aligned} D_l^{h_{l-1}^c, d_l^h, d_l^w} &= \left(H_{l-1}^{h_{l-1}^c} \otimes K_l^{h_{l-1}^c} \right) \left(d_l^h, d_l^w \right), \\ &= \sum_{k^{dh}=1}^{K^{dh}} \sum_{k^{dw}=1}^{K^{dw}} H_l^{h_{l-1}^c, s_l \cdot d_l^h + k^{dh}, s_l \cdot d_l^w + k^{dw}} K_l^{h_{l-1}^c, k^{dh}, k^{dw}} + a_l^{h_{l-1}^c}, \end{aligned} \tag{1}$$

where $D_l \in \mathbb{R}^{H_{l-1}^c \times D_l^h \times D_l^w}$ is the output of the first convolution on the l^{th} DWS convolution block. D_l^h and D_l^w are the height and width of D_l, respectively. Figure 2 shows the operation process of the first convolution in DWS convolution.

Next, D_l is input into ReLU activation function after batch normalization (BN). The process is described as below:

$$D_l' = ReLU(BN(D_l)), \tag{2}$$

where $ReLU$ is the non-linear activation function of linear rectification; BN is the batch normalization; $D_l' \in \mathbb{R}^{H_{l-1}^c, D_l^h, D_l^w}$ is the output of $ReLU$. The second input of DWS convolution is D_l'.

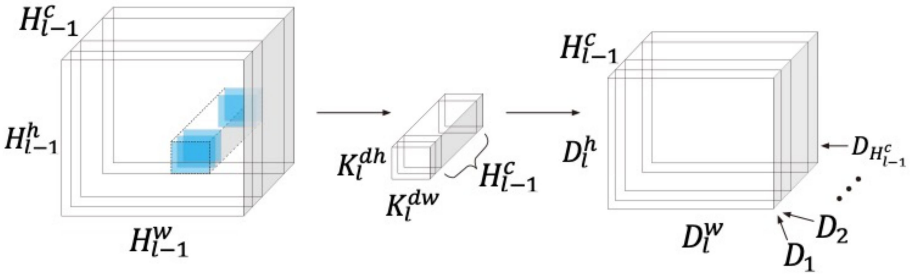


Figure 2. The first step of the depthwise separable convolution: learning spatial information, using H_{l-1}^c different kernels K_l, applied to each H_{l-1}.

In addition, H_l^c number of 1×1 convolution kernels $Z_l \in \mathbb{R}^{H_{l-1}^c}$ and a bias vector $b_l \in \mathbb{R}^{H_l^c}$ are used to learn cross-channel information, with the process as follows:

$$\begin{aligned} S_l^{h_l^c, d_l^h, d_l^w} &= \left(D_l'^{h_{l-1}^c} \otimes Z_l^{h_l^c} \right) \left(d_l^h, d_l^w \right) \\ &= \sum_{h_{l-1}^c=1}^{H_{l-1}^c} D_l'^{h_{l-1}^c, d_l^h, d_l^w} Z_l^{h_l^c, h_{l-1}^c} + b_l^{h_l^c}, \end{aligned} \tag{3}$$

where $S_l \in \mathbb{R}^{H_l^c, D_l^h, D_l^w}$ is the output of the second convolution block on the l^{th} DWS convolution. Figure 3 shows the operation process of the second convolution in DWS convolution.

The down-sampling is used on the application feature dimension after each H_l of the DWS convolution block, for example, maximum pooling. In Equations (1) and (3), $\mathrm{O}(H_{l-1}^c \cdot K_l^{dh} \cdot K_l^{dw} \cdot D_l^h \cdot D_l^w + H_l^c \cdot H_{l-1}^c \cdot D_l^h \cdot D_l^w)$ and $H_{l-1}^c \cdot K_l^{dh} \cdot K_l^{dw} + H_{l-1}^c \cdot H_l^c$ are the computation complexity and the number of parameters (neglecting deviation), respectively. Therefore, the computational complexity and the number of parameters are $(H_l^c)^{-1} + (K_l^{dh} \cdot K_l^{dw})^{-1}$ times lower than a standard convolution operation with the same functions. The final output of the DWS convolution block $H_L \in \mathbb{R}^{H_L^c \times H_L^h \times H_L^w}$ is compressed into one-dimensional vector $H' \in \mathbb{R}^F$, where H' is a one-dimensional feature vector that is extracted by learnable feature extractor inspired by auditory perception and F is the length of the feature vector.

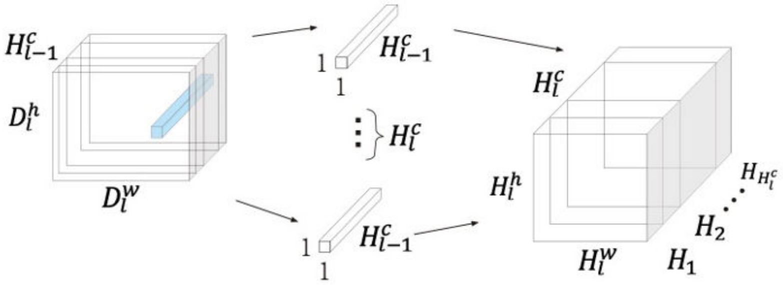


Figure 3. The second step of depthwise separable convolution: learning cross-channel information using H^c_l different kernels Z_l.

3.3. Time-Dilated Convolution

Dilated convolutions introduce a new parameter called "dilation rate" to the convolutional layer, which defines the spacing of values when the convolutional kernel processes data. The purpose of this structure is to provide a larger receptive field without the pooling layer (pooling layer will lead to information loss) and with the same amount of computation. Since the dilated convolutions can cluster and learn multi-scale information, they are widely used in the visual field of deep learning now and have achieved outstanding performances in target detection and image segmentation [35]. In addition, the feature extractor obtains the one-dimensional underwater acoustic feature vector $H' \in \mathbb{R}^F$ at continuous T times, then these H' into the two-dimensional matrix $I \in \mathbb{R}^{T\times F}$ in the fusion layer, where T and F respectively represent the height and width of the matrix, and I is the input of the two-dimensional time-dilated convolution. We use the two-dimensional time-dilated convolution to create a language model of the input, and the classifier is the linear layer of the Softmax activation function.

The time-dilated convolution network here used for long-term mode consists of J time-dilated convolution blocks. The j^{th} time dilated convolution block obtains the output of the previous block as input, i.e., $O_{j-1} \in \mathbb{R}^{O^c_{j-1}\times O^h_{j-1}\times O^w_{j-1}}$, where O^c_{j-1}, O^h_{j-1} and O^w_{j-1} are channel number, height and width output by the $j-1^{\text{th}}$ time-dilated convolution block, respectively. $O_0 = I$ is the feature matrix of the input ship-radiated noise, and $O^c_0 = 1$, $O^h_0 = T$, $O^w_0 = F$. The output of the jth time-dilated convolution block is $O_j \in \mathbb{R}^{O^c_j\times O^h_j\times O^w_j}$, where O^c_j, O^h_j and O^w_j are channel number, height and width of the jth time-dilated convolution block, respectively. The jth time-dilated convolution consists of O^c_j kernels $K'_j \in \mathbb{R}^{O^c_{j-1}\times K'^h_j\times K'^w_j}$ and bias vector $b'_j \in \mathbb{R}^{O^c_j}$, where K^{dh}_l and K^{dw}_l are the height and width of the kernel K'_j, respectively. Thus we have:

$$\begin{aligned} Q_j^{o^c_j,o^h_l,o^w_l} &= \left(O_{j-1}^{o^c_{l-1}} \otimes K_j^{\prime o^c_{l-1}}\right)\left(o^h_l, o^w_l\right) \\ &= \textstyle\sum_{o^c_{j-1}=1}^{O^c_{j-1}} \sum_{k^{dh}=1}^{K^{dh}} \sum_{k^{dw}=1}^{K^{dw}} O_{j-1}^{o^c_j,o^h_l+\xi^h\cdot k^{dh},o^w_l+k^{dw}} K_j^{\prime o^c_j,o^c_{j-1},k^{dh},k^{dw}} + b_j^{\prime o^c_j}, \end{aligned} \tag{4}$$

where ξ^h is the dilation rate of K'_j at K'^h_j dimension. It should be noted that dilation is conducted only in the time dimension in this paper. The dilation rate ξ^h multiplied by k^{dh} is used for visiting the element of O_{j-1}.This allows context information to be clustered in proportion at the output of the operation [35]. In fact, this means that the feature result calculated with the time-dilated convolution is calculated from a larger area. Consequently, a longer time context can be used to create the recognition model. The process described in Equation (4) is shown in Figure 4.

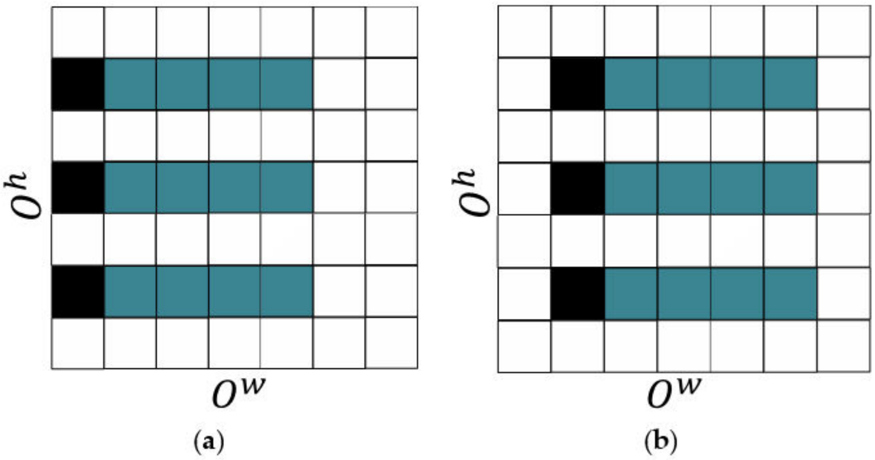


Figure 4. The illustration of the process described in Equation (4) using $\xi^h = 2$ and processing two consecutive patches of O. Squares coloured with cyan signify the elements participating at the processing of $O^{o_l^h,o_l^w}$, and coloured with grey are the elements of $O^{o_l^h,o_l^w+1}$. (**a**) Processing of $O^{o_l^h,o_l^w}$; (**b**) Processing of $O^{o_l^h,o_l^w+1}$.

Next, Q_j is input into the ReLU activation function after batch normalization (BN). The process is described as below:

$$O_j = ReLU\big(BN(Q_j)\big). \tag{5}$$

In the last formula, ReLU is the non-linear activation function of the linear rectification; BN is branch normalization; $O_j \in \mathbb{R}^{O_j^c,O_j^h,O_j^w}$ is the output of ReLU. The output of final time-dilated convolution block $O_J \in \mathbb{R}^{O_J^c,O_J^h,O_J^w}$ is compressed into a one-dimensional vector $O' \in \mathbb{R}^{O_J^c \times O_J^h \times O_J^w}$, then O' is input into the subsequent classifier $Cls(\cdot)$. The classification recognition for the underwater acoustic target is conducted as follows:

$$\hat{Y} = Cls(O'). \tag{6}$$

In Equation (6), $\hat{Y} = [\hat{y}_1, \cdots, \hat{y}_C]$ is the classification result predicted by the model for the underwater acoustic target. We use time-dilated convolution networks instead of RNN, which can effectively model long-term context, intra-class and inter-class activities for underwater acoustic target recognition. The model parameters are optimized through minimizing the cross-entropy loss between $\hat{Y}$ and Y.

4. Model Evaluation

In order to evaluate the proposed method, we used data samples from real civil ships, and used F1 score and precision as evaluation indexes. The artificially designed features, including waveform, wavelet, Mel-frequency cepstral coefficients (MFCC), Hilbert-Huang Transform (HHT), Mel frequency, non-linear auditory feature, spectral and cepstrum features are compared with those automatically extracted by the deep separable convolutional neural network. In addition, the histogram and (t-distributed stochastic neighbor embedding) t-SNE [45] are visualized the clustering performance of the proposed method.

4.1. Dataset and Data Pre-Processing

The dataset contains small ship, big ship, and ferry. The data are sampled at anchorage ground, and the frequency is 48,000 Hz. In the experiment, 80% of samples of each class are used as a training set, while the remaining 20% of samples are used as a testing set. Figure 5 shows a frequency domain diagram of underwater noise. Figure 6 shows a time domain diagram of underwater noise.

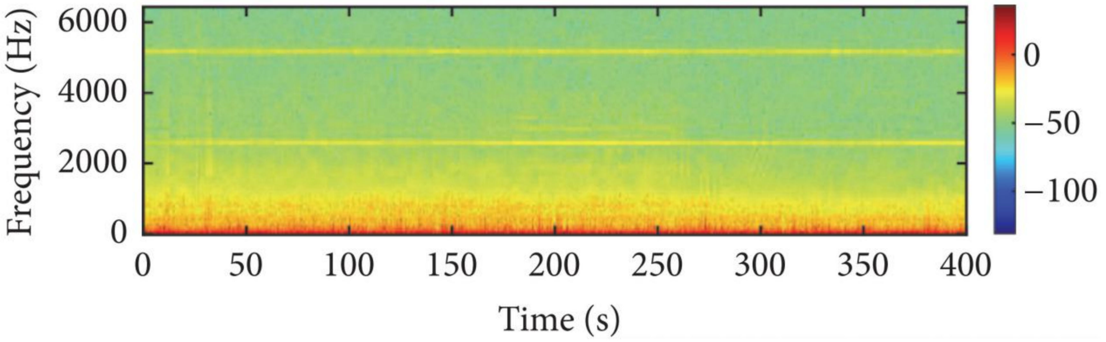


Figure 5. A frequency domain diagram of underwater noise.

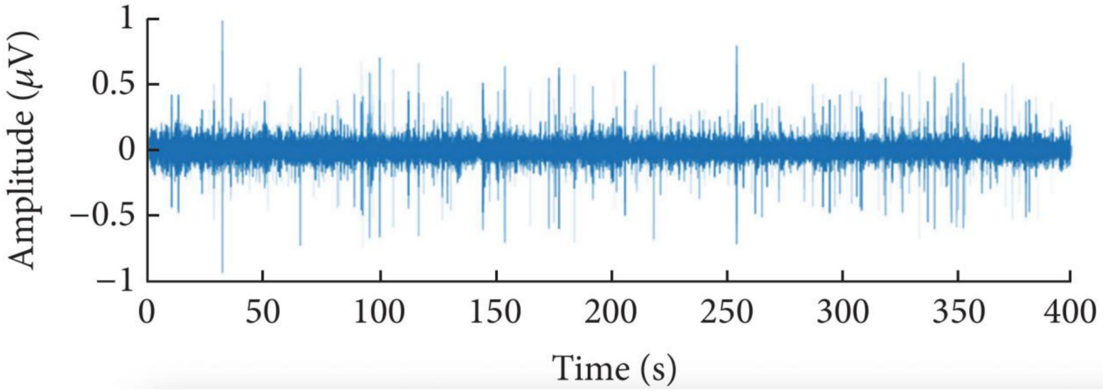


Figure 6. A time domain diagram of underwater noise.

Each record is generated by a WAV (Waveform Audio File Format) audio file. The records include training dataset and testing dataset; 80% of samples of each class are used as the training set and the remaining 20% of samples of each class are used as the testing set. Each record is divided into an audio segment of 10 s, the sampling time of the training samples and test samples are 45 ms and the sampling interval is 12.5 ms. The network training and testing are performed on the raw time domain data without any preprocessing. The total time and number of each type of samples in the training data set and the test data set are shown in Table 1.

Table 1. The experimental data description.

Data Set	Class	No. Segments	Total Time (Hour)	No. Samples	Percentage
Training	small ship	326	0.91	260,736	25.9%
	ferry	560	1.55	447,744	44.6%
	big ship	119	0.33	95,424	9.5%
Test	small ship	81	0.23	65,184	6.5%
	ferry	140	0.39	111,936	11.1%
	big ship	30	0.08	23,856	2.4%

It can be seen from Table 1 that the number of samples of each class is seriously unbalanced. The sample number of small ships accounts for 55.7% of the total sample number, more than half the total sample number, while the big ship samples only account for 11.9% of the total sample number. From the sample distribution of each class, the number of the category sample size is very uneven. For a different classification sample set, the F1 score is a better index than accuracy. To evaluate the performance of this model, we should not only look at the precision index but also look at the performance of classification and recognition. In this paper, F1 score and accuracy are adopted as the evaluation indexes.

4.2. Hyper-Parameters, Indexes and Evaluation Process

In order to evaluate the proposed method, this paper uses the dataset of real time-domain radiated noise, including three classes ($C = 3$), i.e., "small ship", "big ship" and "ferry". The sound fragments are divided into $T = 800$ vector sequences X of length $N = 2176$, and using a hamming window function with 75% overlap in the successive windows of 45 ms: first of all, doing the normalize to all of the input vectors of X sequence, then inputting X into the learnable feature extractor which outputs a radiated noise feature inspired by auditory perception; the length of feature vector is $F = 100$.

Table 2 lists the structure of the learnable feature extractor. The learnable feature extractor is built based on the above-mentioned one-dimensional DWS convolution, but the first layer is the standard one-dimensional convolution. All convolutions are followed by BN and ReLU activation functions. Figure 7 shows the comparison results between the standard convolution calculation process and the deep separation convolution calculation process.

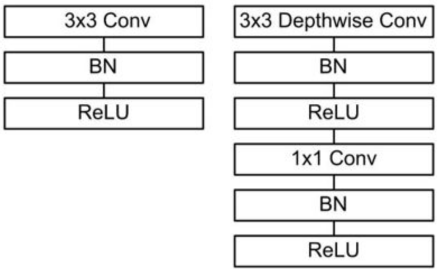


Figure 7. The depthwise separable convolutions with depthwise and pointwise layers followed by batch normalization (BN) and rectified linear unit (ReLU).

Table 2. The structure of the learnable feature extractor.

Type	Stride	Filter Shape	Input Size
Conv1D	50	204 × 1 × 64 dw	2176 × 1
Conv1D dw	2	12 × 64 dw	40 × 64
Conv1D	1	1 × 1 × 64 × 128	15 × 64
Conv1D dw	1	15 × 128 dw	15 × 128
Conv1D	1	1 × 1 × 128 × 100	1 × 100

The feature extractor network ultimately reduces the spatial resolution to 1. Our model structure puts a lot of computation into a 1 × 1 dense convolution, which can be achieved with a highly optimized Generic Matrix Multiplication (GEMM) function. However, an initial reordering called IM2COL is required in the memory to map it to GEMM. For example, this method is used in the popular Caffe package [46]. One by one convolution does not require reordering in memory; among all the deep separation convolution layers of the feature extractor network, 91% of the calculation time is spent in the 1 × 1 convolution, while 88% of the parameters are used; the computational complexity and number of parameters of each layer of deep separation convolution are shown in Table 3.

Table 3. The resource per depthwise separable layer.

Type	Mult-Adds	Parameters
Conv1D dw	11,520	832
Conv1D	122,880	8320
Conv1D dw	1920	2048
Conv1D	12,800	12,900

The structure of time-dilated convolutions inspired by the language model is shown in Table 4. The feature vector of one-dimensional underwater acoustic $H' \in \mathbb{R}^F$ of T time is combined to form the two-dimensional matrix $I \in \mathbb{R}^{T \times F}$ (where, $T = 800$ & $F = 100$) and I is a time-dilated convolution network inspired by the language model. The convolution layers are followed by BN and ReLU activation function, but the pooling layer is not provided with non-linear activation function. The final pooling layer is input into softmax layer for classification. There are five layers in the time-dilated convolution network.

Table 4. The structure of the time-dilated convolutions.

Type	Stride	Dilation	Filter Shape	Input Size
Conv2D Dilation	1 × 1	12 × 1	3 × 3 × 1 × 64	800 × 100 × 1
Max Pool	2 × 2	1 × 1	Pool 2 × 2	776 × 98 × 64
Conv2D Dilation	1 × 1	12 × 1	3 × 3 × 64 × 128	388 × 49 × 64
Max Pool	2 × 2	1 × 1	Pool 2 × 2	364 × 47 × 128
Conv2D Dilation	1 × 1	12 × 1	3 × 3 × 128 × 256	182 × 23 × 128
Avg Pool	2 × 2	1 × 1	Pool 2 × 2	158 × 21 × 256
Conv2D Dilation	1 × 1	12 × 1	3 × 3 × 256 × 512	79 × 10 × 256
Avg Pool	2 × 2	1 × 1	Pool 2 × 2	55 × 8 × 512
Conv2D Dilation	1 × 1	12 × 1	3 × 3 × 512 × 512	27 × 4 × 512
Avg Pool	2 × 2	1 × 1	Pool 2 × 2	3 × 2 × 512
Softmax			Classifier	1 × 1 × 3

In order to evaluate the performance of the proposed method, we use F1 score and accuracy as the indexes of evaluation. We compare the recognition model of artificially designed features, a one-dimensional depthwise convolution network [23] without time-dilated convolution and the proposed depthwise separable convolutional neural networks. These artificially designed features include waveform, wavelet, MFCC, HHT, Mel frequency, non-linear auditory feature, spectrum and cepstrum. The deep separation convolutional neural network model is implemented on the framework of MXNET, A flexible and efficient library for deep learning [47]. The MXNET Python library runs on a nvidia RTX graphic card, and an asynchronous gradient similar to Inception V3 [43] is used to decline the optimizer RMSProp [48]. Table 5 lists the hyper-parameters of the proposed model.

Table 5. The hyper-parameters of the proposed model.

Parameters	Values
Learning Rate	0.001
Batchsize	800
Epochs	100
Optimizer	RMSprop

Some researchers of neurosciences found that the brain can change its structure and functions to meet learning demands. In contrast to the large models of training, we use the techniques of less regularization and data processing, because the small models are not easy to overfit. Driven by the time domain signal of ship radiated noise, all parameters of the depthwise separable convolutional neural network are learned from the actual data. The frequency decomposition and perception ability of depthwise separable convolution networks are also learnable and adjustable.

5. Results and Discussion

The configuration of the server running the neural network is as follows: 64-bit Ubuntu 16.04 operating system, 64 GB memory, 52 CPU kernels and equipped with a TITAN RTX GPU accelerated computing card from NVIDIA (Computer systems design services company).

In this paper, the original time-domain ship-radiated noise data are used to train and test the model. The training parameters are 100 iterations, the size of the training batch is 800 and the training rate is 0.9. The detailed training process is shown in Figure 8:

As shown in Figure 8, in the process of model training, there is no over-fitting or under-fitting phenomenon, and there is no gradient disappearance or gradient explosion. By using the model with measured data, the final training result is that the recognition accuracy of the training data and the test data is 95.9% and 90.9%, respectively, which shows that the model has a high recognition accuracy.

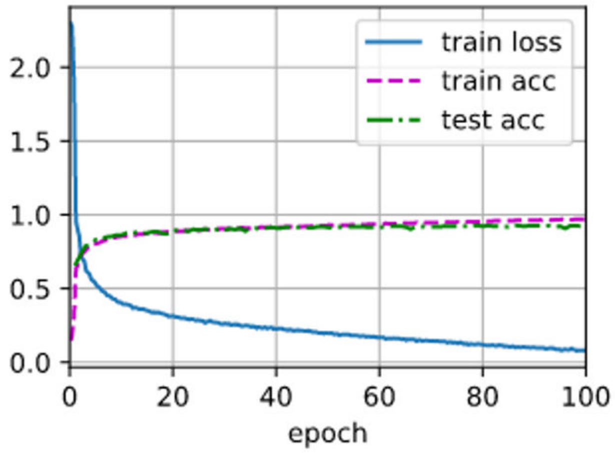


Figure 8. The training process of the model.

With 90.9% recognition accuracy, the model works well for underwater acoustic data with strong noise. Since the number of samples of each class varies greatly, we provide a confusion matrix for the recognition result of the proposed model, as shown in Figure 9. Each row of the confusion matrix corresponds to the real label and each column corresponds to the predicted label.

Figure 9 shows that the recognition results are very stable among the classes, indicating that the model has good recognition stability.

In order to more comprehensively evaluate the recognition performance of this method, on the basis of reflecting the recognition accuracy of the overall classification performance index of the model, the *F1*-score index reflecting the recognition performance of each class of the model is added. The *F1*-score for each class is calculated from a harmonic average of the accuracy and recall rates for that class, which is a better measure than accuracy for unbalanced datasets because both accuracy and recall rates are taken into account. The *F1*-scores for each class are "weighted" average and "micro" average. The accuracy rate, recall rate and *F1*-score of each category calculated by the recognition results in this paper are shown in Table 6:

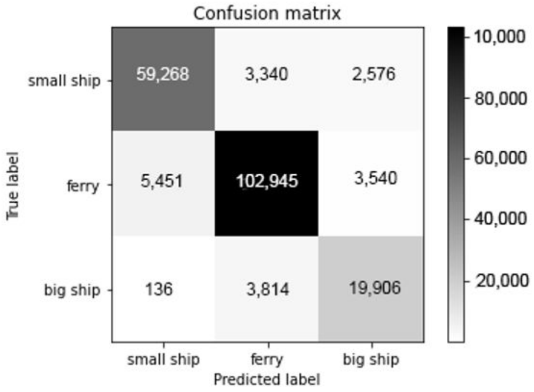


Figure 9. The confusion matrix of the proposed model obtained from testing data.

As shown in Table 6, the *F1* score of this model is very good, indicating that this model has good classification accuracy and stability. The results of small boat class and ferry class are better, with *F1* score of 0.91 and 0.93 respectively. The worst results are for the big ship class, with an accuracy of 0.76, a recall rate of 0.83, and an *F1* score of 0.80. It could be that the mechanical systems of the boats are similar to those of ferries, or some boats were passing by during the collection of ferry samples.

Table 6. The precision, recall and *F1* score for each class.

Class	Precision	Recall	*F1* Score	Support
small ship	0.91	0.91	0.91	65,184
ferry	0.94	0.92	0.93	111,936
big ship	0.76	0.83	0.80	23,856
Accuracy			0.91	
Macro avg	0.87	0.89	0.88	
Weighted	0.91	0.91	0.91	

In order to verify the non-unfitting and non-overfitting of the model, the *k*-fold cross-validation method is used. Cross validation is a statistical method to evaluate the generalization performance, which is more stable and comprehensive than the method of single partition of training set and test set. The data are divided many times and many models need to be trained. The most common cross validation is *k*-fold cross validation, where *k* is the number specified by the user. In this paper, we set *k* = *5*. When we perform a 5-fold cross validation, the data are first divided into five equal parts, each of which is called a fold. The first fold is used as the test set, and the other folds are used as the training set to train the first model. The model is constructed with 2~5 trade-off data, then the accuracy is evaluated on 1 trade-off. Then another model is built, where we use 2-fold as the test set and others folds as the training set. For the five times of dividing the data into training set and test set, the accuracy should be calculated each time. Finally, we get five accuracy values. The whole process is shown in Figure 10, and the confusion matrix of 5-fold cross validation is shown in Figure 11. The validation results of 5-fold cross validation are listed in Table 7.

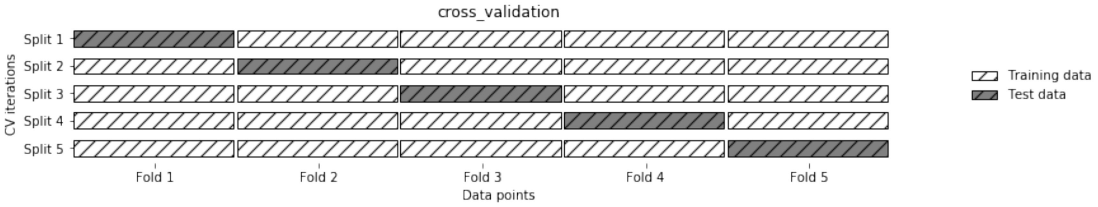
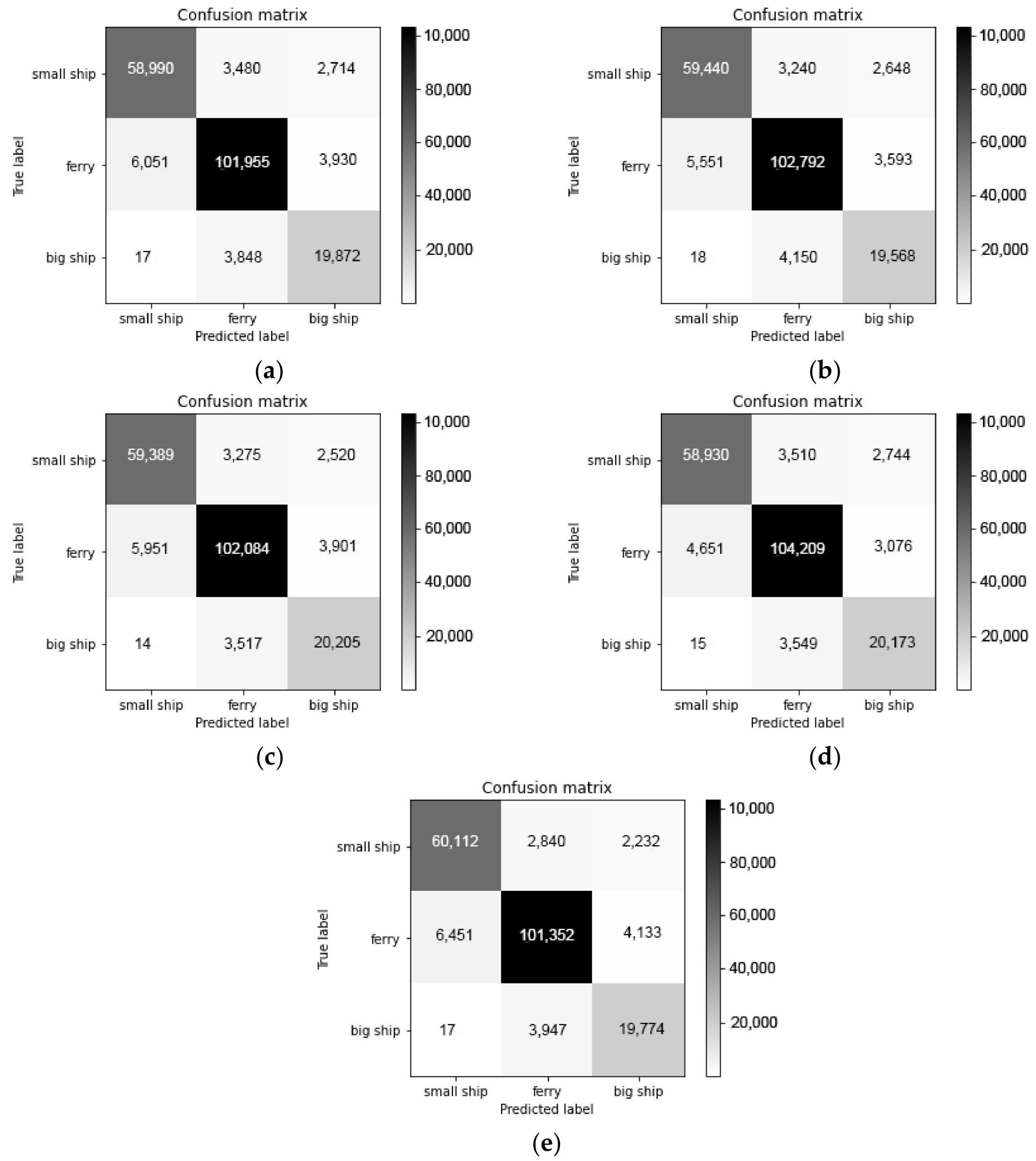


Figure 10. The whole process of 5-fold cross validation.



(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 11. The confusion matrix of 5-fold cross validation: (**a**) fold 1; (**b**) fold 2; (**c**) fold 3; (**d**) fold 4; (**e**) fold 5.

Table 7. The results of 5-fold cross validation.

Fold	Weighted Average *F1* Value
1	0.90
2	0.91
3	0.91
4	0.92
5	0.90

From the experimental results of Table 7, the proposed model has good generalization performance, and there is no serious unfitting and overfitting. In order to simulate practical applications of recognition for ship-radiated noise, the classification accuracy of each acoustic event is used to measure the classification performance of the model, which is defined as the percentage of all acoustic events that are correctly classified. The classification accuracy of the proposed model and the comparison model is shown in Table 8.

Table 8. The classification results of proposed model and compared models.

Input	Methods	Accuracy
HOS [49]	Support vector machine (SVM)	85.1%
Waveform [9,50]	SVM	78.9%
Wavelet [11]	SVM	84.3%
MFCC [51]	SVM	79.1%
Mel-frequency	SVM	84.6
Nolinear auditory	SVM	86.7
Spectral [27,52]	Deep neural network (DNN) [53]	87.0%
Cepstral [10,54]	DNN [44]	86.9%
Raw time domain data	Convoluted neural network (CNN) model [25]	88.4%
Raw time domain data	Convolution recursive neural network (CRNN) model [28]	89.2%
Raw time domain data	Proposed model	90.1%

HOS is high order statistics feature. MFCC is Mel-frequency cepstral coefficients. As shown in Table 8, compared with traditional underwater acoustic target recognition methods, the proposed model effectively improves the classification accuracy of the underwater acoustic target. Due to the complexity of the marine environment, it is also very important to improve the generalization performance and reduce the complexity of the model. Therefore, the regularization term of the first-order norm is added in the training process of the model to ensure good generalization performance of the model at the appropriate sacrifice of training classification accuracy. When the regularization term of the first-order norm is added, the classification accuracy of this model is 90.9%, which is 6.8% higher than that of the traditional recognition model, which is 85.1%, indicating that the classification recognition model significantly improves the classification accuracy of the traditional recognition method.

Next, the distribution of features extracted by the model is analyzed by visualization method. Here, a histogram is created for each extracted dimension feature, and the occurrence frequency (called bin) of the data points of a one-dimension feature in each class is calculated. This allows us to understand the distribution of each dimension's features in each class and how the eigenvalues are different between classes. In this paper, 100-dimensional features are extracted from underwater acoustic targets. Figure 12 is part of the histogram of feature results extracted from the model. As shown in Figure 11, the feature vectors extracted by the proposed model for all training samples have obvious distribution differences among different classes, while the feature distributions within the same class are stable and consistent, which indicates that the method of feature extraction by the proposed model is effective for classification.

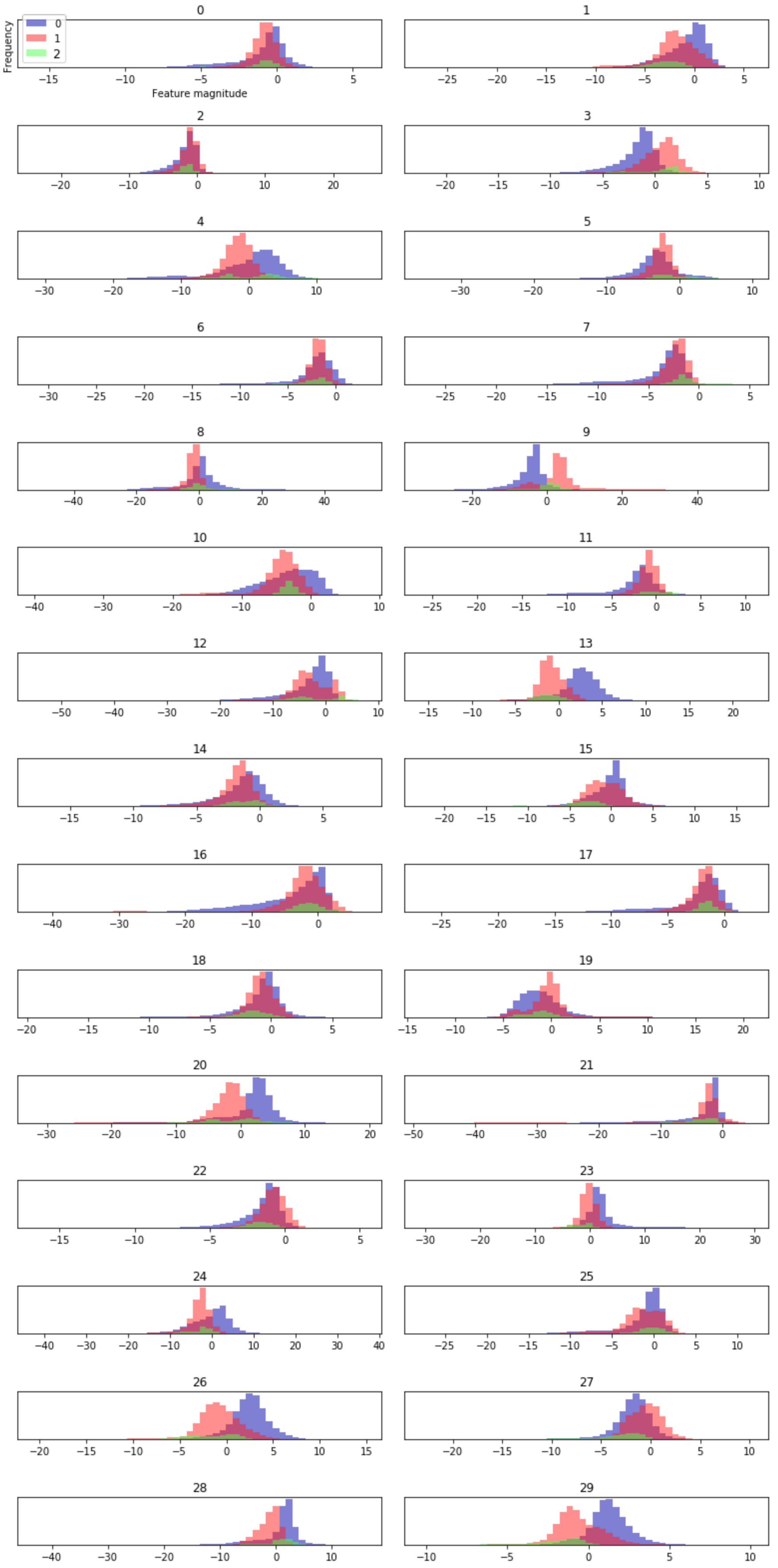


Figure 12. The histogram of features learned from the proposed model: 0 is small ship, 1 is big ship, 2 is ferry.

The manifold learning algorithm is used to carry out complex mapping of 100-dimensional feature vectors extracted from all samples, and a good visual 2-dimensional vector is obtained. We use the t-SNE algorithm to visualize the feature vectors of underwater acoustic targets. Figure 13 shows the scatter plot of the two-dimensional vector obtained from the complex mapping of 100 feature vectors. Here, we use the corresponding number of each class as a symbol to show the position of each class. It can be seen that each class is relatively well separated, indicating that the feature results extracted by the model have a good clustering effect, and visually proving that the feature extracted by the underwater acoustic target has good separability and stability.

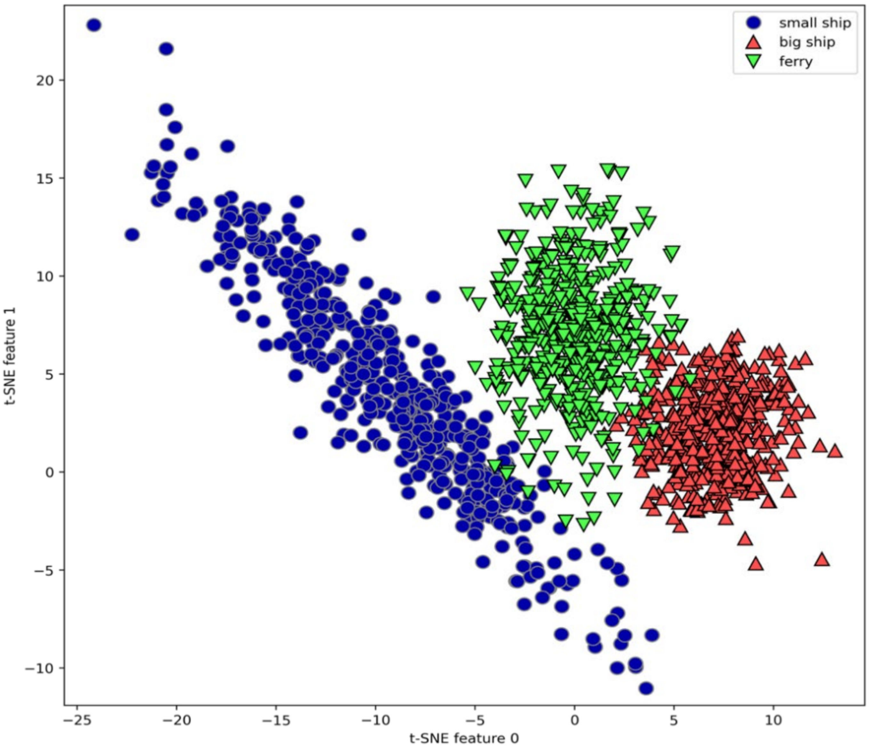


Figure 13. The scatter plot of features learned from proposed model using two components found by t-distributed stochastic neighbor embedding (t-SNE).

6. Conclusions

In this paper, a new depthwise separable convolutional neural network is proposed to identify ship radiated noise from original time-domain waveforms in an end-to-end mode. The deep features containing the internal information of the target are extracted by a DWS convolution network, which reflects the deep acoustic information extraction structure of the auditory system. By convolution decomposition of different frequency components of ship-radiated noise, the frequency distribution characteristics of ship radiated noise are revealed. The time-dilated convolution is used for modeling long time contexts, which can make full use of the intra-class and inter-class information for underwater acoustic target recognition just like the language model. Inspired by the plasticity neural mechanism, all the parameters in the model are learned and optimized under the drive of the time-domain ship radiated noise, so as to accomplish the underwater acoustic target recognition task. The average classification recognition rate reaches 90.9% when tested on a real civil ship acoustic signal set. Although the recognition rate is high, there is still a certain gap between it and the practical application, and the recognition rate needs to be further improved. The experimental results also show that the extracted 100-dimensional features of underwater acoustic target have good separability and stability, and the deep learning method based on auditory perception has great potential in improving the classification performance of underwater acoustic target recognition.

Author Contributions: G.H. contributed to the idea of the incentive mechanism and finished this manuscript. K.W. was responsible for some parts of the theoretical analysis. L.L. contributed some necessary experimental data. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Basic Scientific Research project (No. JCKY2017207B042) and the National Natural Science Foundation of China (No. 61573114).

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

References

1. Xu, J.; Huang, Z.; Li, C. Advances in underwater target passive recognition using deep learning. *J. Sig. Proc.* **2019**, *35*, 1460–1475. [CrossRef]
2. Testolin, A.; Diamant, R. Combining denoising autoencoders and dynamic programming for acoustic detection and tracking of underwater moving targets. *Sensors* **2020**, *20*, 2945. [CrossRef] [PubMed]
3. Lourens, J.G. Classification of ships using underwater radiated noise. *Conf. Commun. Signal Process.* **1988**, *161*, 130–134.
4. Rajagopal, R.; Sankaranarayanan, B.; Rao, P.R. Target classification in a passive sonar—An expert system approach. *IEEE ICASSP-90* **2002**, *5*, 2911–2914.
5. Maksym, J.N.; Bonner, A.J.; Dent, C.A.; Hemphill, G.L. Machine analysis of acoustical signals. *Pattern Recognit.* **1983**, *16*, 615–625. [CrossRef]
6. Amab, D.; Arun, K.; Rajendar, B. Feature analyses for marine vessel classification using passive sonar. *UDT* **2005**, *7*, 21–23.
7. Farrokhrooz, M.; Karim, M. Ship noise classification using probabilistic neural network and AR model coefficients. *Oceans* **2005**, *2*, 1107–1110.
8. Nii, H.; Feigenbaum, E.; Anton, J. Signal-to-symbol transformation: Reasoning in the HASP/SIAP program. *IEEE Int. Conf. Acoust. Speech Signal Process.* **1984**, *9*, 158–161.
9. Meng, Q.; Yang, S. A wave structure based method for recognition of marine acoustic target signals. *J. Acoust. Soc. Am.* **2015**, *137*, 2242. [CrossRef]
10. Das, A.; Kumar, A.; Bahl, R. Marine vessel classification based on passive sonar data: The cepstrum-based approach. *IET Radar, Sonar Navig.* **2013**, *7*, 87–93. [CrossRef]
11. Wei, X.; Li, G.; Wang, Z.Q. Underwater target recognition based on wavelet packet and principal component analysis. *Comput. Simul.* **2011**, *28*, 8–290. [CrossRef]
12. Yang, H.; Gan, A.; Chen, H.; Yue, P.; Tang, J.; Li, J. Underwater acoustic target recognition using SVM ensemble via weighted sample and feature selection. In Proceedings of the 13th International Bhurban Conference on Applied Sciences and Technology, Islamabad, Pakistan, 12–16 January 2016. [CrossRef]
13. Filho, W.; De Seixas, J.; De Moura, N. Preprocessing passive sonar signals for neural classification. *IET Radar Sonar Navig.* **2011**, *5*, 605. [CrossRef]
14. Sainath, T.N.; Kingsbury, B.; Mohamed, A.R.; Ramabhadran, B. Learning filter banks within a deep neural network framework. In Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Olomouc, Czech Republic, 8–12 December 2013; IEEE: Piscataway, NJ, USA, 2014; pp. 297–302.
15. Gazzaniga, M.; Ivry, R.B.; Mangun, G.R. *Cognitive Neuroscience the Biology of the Mind*; W. W. Norton & Company: New York, NY, USA, 2018.
16. Weinberger, N.M. Experience-dependent response plasticity in the auditory cortex: Issues, characteristics, mechanisms, and functions. In *Springer Handbook of Auditory Research*; Springer: Cham, Switzerland, 2004. [CrossRef]
17. Shen, S.; Yang, H.; Li, J.; Xu, G.; Sheng, M. Auditory inspired convolutional neural networks for ship type classification with raw hydrophone data. *Entropy* **2018**, *20*, 990. [CrossRef]
18. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems—Volume 2*; MIT Press: Montreal, Canada, 2014; pp. 3104–3112, arXiv preprint.
19. Auli, M.; Galley, M.; Quirk, C.; Zweig, G. Joint language and translation modeling with recurrent neural networks. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 13–21 October 2013; Association for Computational Linguistics: Stroudsburg, PA, USA; pp. 1044–1106. Available online: https://www.aclweb.org/antholo-gy/D13-1106 (accessed on 17 February 2021).
20. Bengio, S.; Vinyals, O.; Jaitly, N.; Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; MIT Press: Cambridge, MA, USA, 2015; Volume 1, pp. 1171–1179. Available online: http://dl.acm.org/citation.cfm?id=2969239.2969370 (accessed on 17 February 2021).

21. Drossos, K.; Gharib, S.; Magron, P.; Virtanen, T. Language modelling for sound event detection with teacher forcing and scheduled sampling. In Proceedings of the Workshop on Detection and Classifification of Acoustic Scenes and Events (DCASE), New York, NY, USA, 25–26 October 2019.
22. Cheng, J.; Du, X.; Zeng, S. Research on audio feature extraction and recognition of underwater targets using deep learning method[C]//Chinese acoustic society. In Proceedings of the National Acoustics Conference 2018, Patras, Greece, 8–9 October 2018. (In Chinese).
23. Hu, G.; Wang, K.; Peng, Y.; Qiu, M.; Shi, J.; Liu, L. Deep learning methods for underwater target feature extraction and recognition. *Comput. Intell. Neurosci.* **2018**, *2018*, 1–10. [CrossRef] [PubMed]
24. Lang, Z. *Research on Underwater Target Feature Extraction Based on Convolutional Neural Network*; Harbin Engineering University: Harbin, China, 2017. (In Chinese)
25. Wang, N.; He, M.; Wang, H. Fast dimensionality reduction convolution model for underwater target recognition. *J. Harbin Eng. Univ.* **2019**, 1–6. (In Chinese)
26. Lu, A. *Underwater Acoustic Classification Based on Deep Learning*; Harbin Engineering University: Harbin, China, 2017. (In Chinese)
27. Cao, X.; Zhang, X.; Yu, Y.; Niu, L. Deep learning-based recognition of underwater target. In Proceedings of the 2016 IEEE International Conference on Digital Signal Processing (DSP), Beijing, China, 16–18 October 2016; pp. 89–93.
28. Zhang, S.; Tian, D. Intelligent classification method of Mel frequency cepstrum coefficient for under-water acoustic targets. *J. Appl. Acoust.* **2019**, 267–272. (In Chinese)
29. Drossos, K.; Mimilakis, S.; Gharib, S.; Li, Y.; Virtanen, T. Sound event detection with depthwise separable and dilated con-volutions. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020.
30. Fonseca, E.; Plakal, M.; Font, F.; Ellis, D.; Serra, X. Audio tagging with noisy labels and minimal supervision. In Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), New York, NY, USA, 25–26 October 2019.
31. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
32. Sifre, L. Rigid-Motion Scattering for Image Classification. Ph.D. Thesis, CMAP Ecole Polytechnique, Palaiseau, France, October 2014.
33. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016.
34. Holschneider, M.; Kronland-Martinet, R.; Morlet, J.; Tchamitchian, P. A real-time algorithm for signal analysis with the help of the wavelet transform. In *Wavelets*; Combes, J.-M., Grossmann, A., Tchamitchian, P., Eds.; Springer: Berlin/Heidelberg, Germany, 1990; pp. 286–297.
35. Shensa, M.J. The discrete wavelet transform: Wedding the a trous and Mallat algorithms. *IEEE Trans. Signal Process.* **1992**, *40*, 2464–2482. [CrossRef]
36. Li, Y.; Liu, M.; Drossos, K.; Virtanen, T. Sound event detection via dilated convolutional recurrent neural networks. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 286–290.
37. Robertson, D.; Irvine, D.R.F. Plasticity of frequency organization in auditory cortex of guinea pigs with partial unilateral deafness. *J. Comp. Neurol.* **2010**, *282*, 456–471. [CrossRef] [PubMed]
38. Weinberger, N.M. Learning-induced changes of auditory receptive fields. *Curr. Opin. Neurobiol.* **1993**, *3*, 570–577. [CrossRef]
39. Zhang, Y.; Zhang, J. Understanding convolution from filter. *Elec. Pro.* **2019**, *11*, 46–47. (In Chinese)
40. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
41. Lecun, Y.; Bengio, Y.; Hinton, G.E. Deep learning. *Nature* **2015**, *521*, 436. [CrossRef]
42. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.
43. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception structure for computer vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2818–2826.
44. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
45. Hinton, G.E. Visualizing high-dimensional data using t-SNE. *Vigiliae Christ.* **2008**, *9*, 2579–2605.
46. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional structure for fast feature embedding. *arXiv* **2014**, arXiv:1408.5093.
47. Chen, T.; Li, M.; Li, Y.; Lin, M.; Wang, N.; Wang, M.; Xiao, T.; Xu, B.; Zhang, C.; Zhang, Z. MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems. *arXiv* **2015**, arXiv:1512.01274. Available online: tmxnet.apache.org (accessed on 17 February 2021).
48. Tieleman, T.; Hinton, G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA Neural Netw. Mach. Learn.* **2012**, *4*, 26–31.
49. Chen, F.; Lin, Z.; Peng, Y. High order statistics feature extraction and feature compression of ship radiated noise. *Appl. Acot.* **2010**. [CrossRef]

50. Meng, Q.; Yang, S.; Piao, S. The classifification of underwater acoustic target signals based on wave structure and support vector machine. *J. Acoust. Soc. Am.* **2014**, *136*, 2265. [CrossRef]
51. 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]
52. Kamal, S.; Mohammed, S.K.; Pillai, P.R.S.; Supriya, M.H. Deep learning structures for underwater target recognition. In Proceedings of the 2013 International Symposium on Ocean Electronics, Kochi, India, 23–25 October 2013; IEEE: Piscataway, NJ, USA, 2014; pp. 48–54.
53. Yue, H.; Zhang, L.; Wang, D.; Wang, Y.; Lu, Z. The classification of underwater acoustic targets based on deep learning methods. *Adv. Intell. Syst. Res.* **2017**, *134*, 526–529.
54. 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]