applied sciences

MDPI

Article

Underwater Acoustic Target Recognition with a Residual Network and the Optimized Feature Extraction Method

Feng Hong [1,*], Chengwei Liu [1,2], Lijuan Guo [3], Feng Chen [1] and Haihong Feng [1]

[1] Shanghai Acoustics Laboratory, Chinese Academy of Sciences, Shanghai 201805, China; liuchengwei19@mails.ucas.ac.cn (C.L.); chenfeng@mail.ioa.ac.cn (F.C.); fhh@mail.ioa.ac.cn (H.F.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] The School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 200237, China; lijuanguohf@gmail.com
* Correspondence: hongfeng@mail.ioa.ac.cn

Abstract: Underwater Acoustic Target Recognition (UATR) remains one of the most challenging tasks in underwater signal processing due to the lack of labeled data acquisition, the impact of the time-space varying intrinsic characteristics, and the interference from other noise sources. Although some deep learning methods have been proven to achieve state-of-the-art accuracy, the accuracy of the recognition task can be improved by designing a Residual Network and optimizing feature extraction. To give a more comprehensive representation of the underwater acoustic signal, we first propose the three-dimensional fusion features along with the data augment strategy of SpecAugment. Afterward, an 18-layer Residual Network (ResNet18), which contains the center loss function with the embedding layer, is designed to train the aggregated features with an adaptable learning rate. The recognition experiments are conducted on the ship-radiated noise dataset from a real environment, and the accuracy results of 94.3% indicate that the proposed method is appropriate for underwater acoustic recognition problems and sufficiently surpasses other classification methods.

Keywords: ResNet; underwater acoustics; ShipsEar; embedding; SpecAugment; UATR; MFCC; combined features

Citation: Hong, F.; Liu, C.; Guo, L.; Chen, F.; Feng, H. Underwater Acoustic Target Recognition with a Residual Network and the Optimized Feature Extraction Method. *Appl. Sci.* **2021**, *11*, 1442. https://doi.org/10.3390/app11041442

Received: 8 January 2021
Accepted: 30 January 2021
Published: 5 February 2021

1. Introduction

As a key technology to promote the intelligence of the underwater acoustic equipment system, underwater acoustic target-radiated noise recognition is one of the most important research directions of underwater acoustic signal processing [1]. Some reasons behind this are that underwater environments are complex and changing, making it difficult to identify underwater objects [2,3]. Besides, although recent years have experienced an increasing number of water vessels, the labeled large scale of the data acquisition has been hindered a lot because of confidentiality and cost. Another drawback based on the data quality is that the background noise and the interference from other noise sources inevitably exist.

Generally, Underwater Acoustic Target Recognition (UATR) is performed by well-trained sonar men, which is inaccurate due to the long-time work and may be affected by weather conditions [4]. Hence, developing a robust recognizing system to replace humans' work of identifying ship-radiated noise is of great importance. From a technical perspective, efforts are consistently made to improve the classification accuracy in the aspects of feature extraction and classifier training [5].

Much work attempts to extract hand-crafted features from ship-radiated noise and feed them into different kinds of classifiers. On one hand, as for the traditional machine learning feature extraction process, Support Vector Machines (SVM) [6,7] and Principal Component Analysis (PCA) [8] methods are widely used. For example, Meng et al. proposes a method straightly using the wave structure with SVM [6]. Wei et al. [7] present an extraction method based on $1_{1/2}$D spectrum and PCA. Features derived from Mel filters

of Mel Frequency Cepstral Coefficients (MFCC) and Log-Mel Spectrogram (LM) are two widely used features in Environment Sound Classification (ESC) tasks [9,10] with acceptable performance. Although such features originate from the speech or sound field, the effect of MFCC and its first-order differential MFCC or second-order MFCC features are proven for underwater acoustic target recognition [8]. Besides, a considerable number of research works indicate that the fusion feature can give a more comprehensive representation of environment sounds [11]. For the recognition of underwater acoustic targets, Meng et al. [6] exploit the fusion feature of zero-crossing wavelength, peek-to-peek amplitude, and zero-crossing-wavelength difference. On the other hand, the design of the neural networks plays an important role in achieving a competitive performance together with the optimized feature extraction. For example, a time-delay neural network (TDNN) and convolutional neural network (CNN) are introduced for UATR in [12]. Testolin et al. [13] present an innovative method that allows to accurately detect and track underwater moving targets from the reflections of an active acoustic emitter. The system is based on a computationally and energy-efficient preprocessing stage carried out using a deep convolutional denoising autoencoder (CDA), whose output is then fed to a probabilistic tracking method based on the Viterbi algorithm. Testolin et al. [14] also have proven that transfer learning can be a viable approach in these scenarios where tagged data is often lacking and evaluate the feasibility of the model based on Recurrent Neural Networks (RNN) for the scenarios requiring on-line processing of the reflection sequence. Shen et al. [15] introduce the auditory inspired convolutional neural networks trained from raw underwater acoustic signal.

For ShipsEar [16], the machine learning method using a GMM-based classifier with the standard expectation maximization (EM) algorithm for training purposes could be used as a baseline method, whose best classification rate is 75.4%. As the results obtained by typical methods based on machine learning are not very high, the methods based on deep learning models are worthy of researching. Li et al. [5] introduce a feature optimization approach with Deep Neural Networks (DNN) and optimizing loss function and achieve an accuracy of 84%. Yang et al. [17] propose a so-called competitive Deep Belief Nets (cDBN) for UATR. Luo et al. [18] present a UATR method based on Restricted Boltzmann Machine (RBM), which achieves the accuracy of 93.17% on the dataset of ShipsEar. Ke et al. [4] propose a novel recognition method of four steps including preprocessing, pretraining, finetuning, and recognition, which achieves the recognition accuracy of 93.28%.

In this paper, we propose the three-dimensional fusion features along with the data augment strategy of SpecAugment and an 18-layer Residual Network (ResNet18) containing the center loss function with the embedding layer to achieve good accuracy. The remaining parts of this paper are organized as follows. Section 2 introduces the framework of the classification method for UATR. Section 3 describes the suitable feature aggregate scheme of the three-dimensional fusion features and an 18-layer Residual Network (ResNet18). Section 4 presents the experimental results of ShipsEar with the proposed method and other methods. Section 4 gives the conclusion.

2. Description of the Classification Method

As depicted in Figure 1, the classification method mainly contains several steps, i.e., preprocessing, feature extraction, Residual Network training, and embedding layer design. Note that the underwater acoustic waves have different time duration, therefore, the preprocessing steps of truncation and framing are necessary. Afterward, we perform the feature extraction process, which contains feature calculation, feature fusion, and feature augment, to obtain the three-dimensional feature. Instead of using the general structure of CNN or DNN, we design an 18-layer Residual Network named "ResNet18" together with the center loss function. As a complex network, ResNet18 could learn very informative presentations in the training process. To avoid overfitting, we adopt the tactic strategy to train the aggregated features with early stopping and adaptable learning rate. Besides, aiming to increase the distance for inter-class and decrease that for Intra-Class,

the embedding layers are well-tuned where the center loss function and softmax is used instead of the softmax.

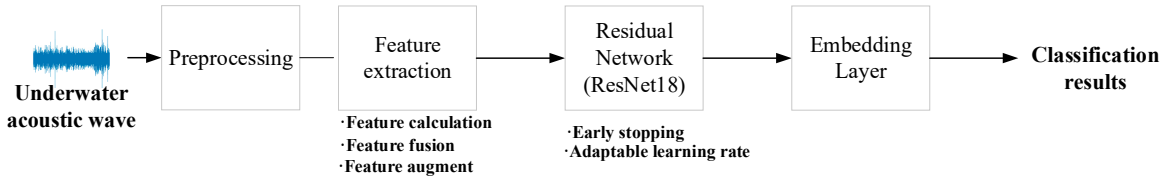


Figure 1. The depiction of the classification method.

2.1. Preprocessing

Although the practically captured data from the environment may have a variable duration, underwater acoustic processing methods based on CNN or DNN or ResNet networks generally need the length of the input sample to be fixed. Therefore, preprocessing is a prerequisite before feeding the signals of different lengths to the designed networks. One feasible way is to split the underwater acoustic signal into several frames of fixed length using a sliding designed window of appropriate width. Considering that an underwater acoustic signal can be seemed as stable in a very short time, the framing of the recorded signal could be performed. This naturally increases the number of samples as some parts of the underwater acoustic signal are reused and that can be viewed as some sort of data augmentation [19]. For simplicity, our framing method may do not take overlapping and no window is used in it. Note that each preprocessed signal must be labeled as the raw label. Besides, for underwater acoustic, a sampling rate of 20,480 Hz may be considered a good trade-off between the quality of the input sample and the computational cost of the model. We present the demonstration of the framing of the signal of the type of passengers into several framess and s + 1 with no overlapping in Figure 2.

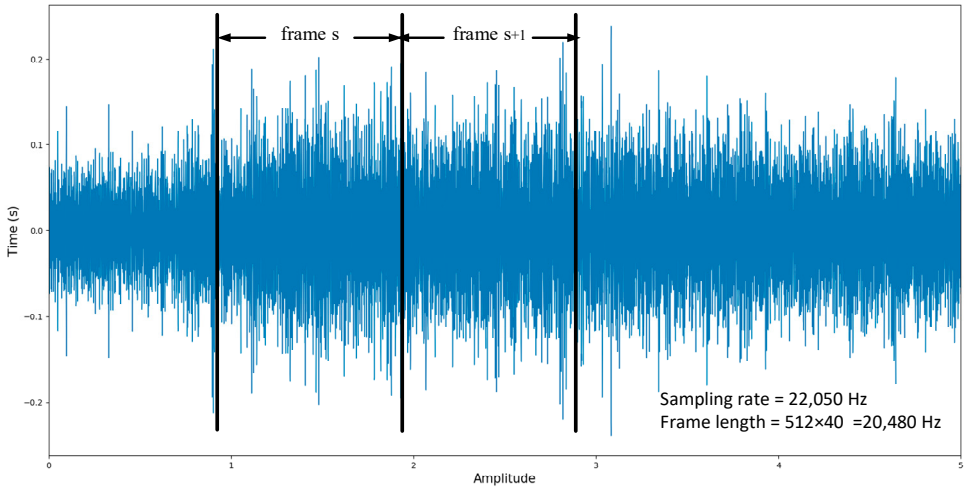


Figure 2. Framing the signal of the type of passengers into several frames s and s + 1 with no overlapping.

2.2. Feature Extraction Process

Figure 3 gives the graphical representation of the procedure for extracting the final features that contains several steps, i.e., feature calculation, feature fusion, and feature augment. In general, different features can capture different characteristics of the underwater acoustic signal. Therefore, multiple features can be combined to exploit the complementary information for further improvement of the feature design. At the stage of feature calculation and feature fusion, we extract the Log Mel (LM) as the first channel, the MFCC as the second channel, and the composition of Chroma, Contrast, Tonnetz, and Zero-cross

ratio called CCTZ as the third channel. The first two channels both have the feature size of 60 × 41, where 60 denotes the number of the bands and 41 denotes the number of frames. As for the third channel, the feature size of Chroma, Contrast, Tonnetz, and Zero-cross ratio is 24 × 41, 6 × 41, 6 × 41, and 1 × 41, respectively. To form a new feature set called CCTZ, the features are sequentially stacked with the remaining lines paddled by zero. Afterward, at the stage of feature augment, we apply the feature augment method of SpecAugment five times on the LM feature, leaving the other channels unchanged. SpecAugment originates from the data augmentation method for speech recognition and is applied directly to the feature inputs of a neural network (i.e., filter bank coefficients) [20]. The augmentation policy consists of warping the features, masking blocks of frequency channels, and masking blocks of time steps [20]. The resulting three-dimensional feature matrix is composed of the three channels as mentioned.

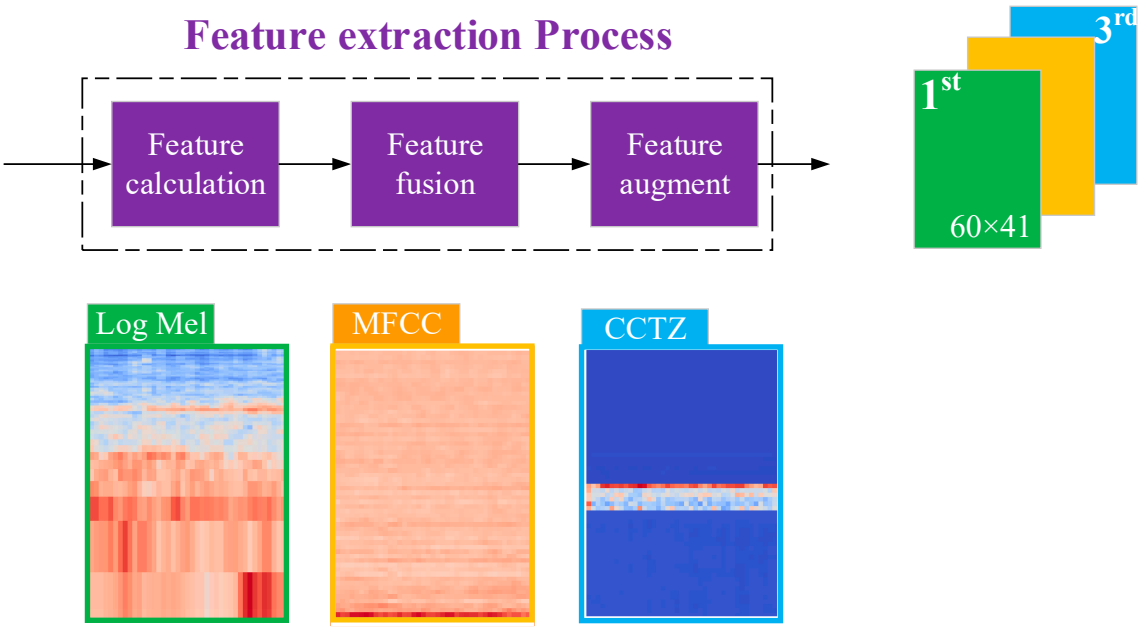


Figure 3. The depiction of the feature extraction process.

2.3. Structure of the ResNet18

The number of convolutional layers plays a key role in detecting high-level concepts [19]. Residual learning framework could decrease the difficulty of the training of deep neural networks with many layers [21]. Motivated by this, the architecture of ResNet18, which is a modified version of the general ResNet50, is shown in Figure 4. The details are listed as follows:

- Stage 0: The input layer is zero-paddled of 3 × 3. After the processing, the shape changes from 60 × 41 × 3 to 66 × 47 × 3.
- Stage 1: The first stage consists of a convolutional layer of 64 filters of the rectangular shape of 3 × 3 and stride of 2 × 2. The batch-normalization is applied, followed by a Rectified Linear Unit (ReLU) as the activation function and max-pooling of 3 × 3 and a stride of 2 × 2.
- Stage 2: The second stage consists of a convolutional block named "block-2a," an identity block named "block-2b," and an identity block named "block-2c." The structure of the convolutional block and the identity block is shown as Conv_block and Id_block in Figure 4, respectively.
- Stage 3: The third stage consists of a convolutional block named "block-3a," an identity block named "block-3b," and the average pooling layer of 2 × 2. After processing, the shape of the output of the flatten layer changes from 15 × 11 × 512 to 7 × 5 × 512.

- Stage 4: The stage contains the flatten layer and the designed module of center loss. The number of hidden units of the first fully connected layer is 17,920. As for the designed module of center loss, the second fully connected layer with 24 hidden units connects to the first layer, followed by the Parametric Rectified Linear Unit (PReLU) as the activation function. The last fully connected is with 5 hidden units.

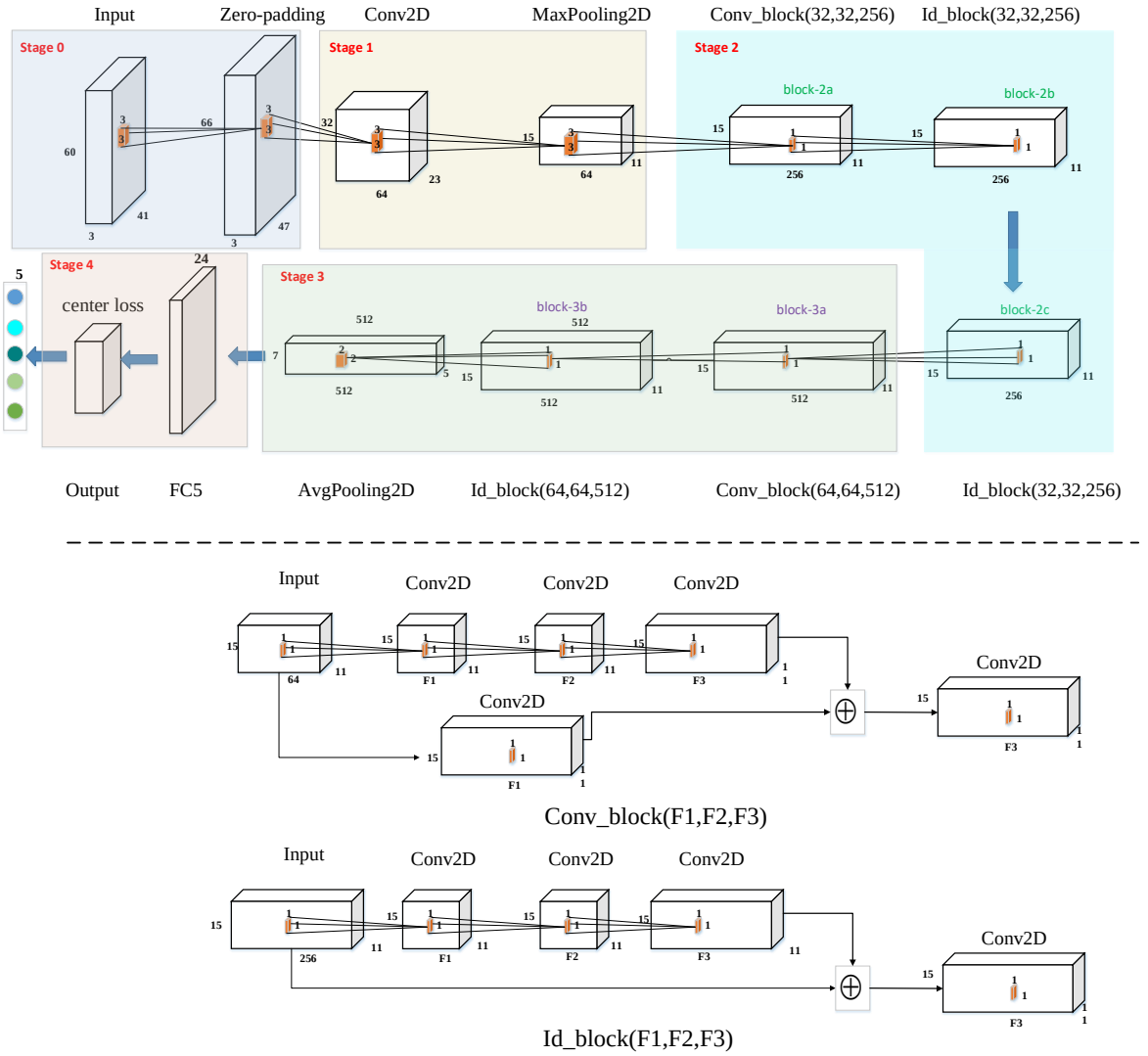


Figure 4. The overall framework of an 18-layer Residual Network (ResNet18).

We should mention that for the identity block, nonlinear activation helps the designed networks to learn more complex features. The activation function is defined as

$$PReLU(x_i) = \max(0, x_i) + \alpha_i \min(0, x_i), \quad (1)$$

where x_i is the input of the nonlinear activation on the ith channel and α_i is a learnable coefficient controlling the slope of the negative part.

2.4. The Embedding Layer with the Center Loss Function and Softmax

To address the face verification problem, center loss [22] is first exploited to compensate for softmax loss in 2016. For the UATR task, we improve generalization by leveraging such a simple but useful cost function, since it learns a center for the features of each class and meanwhile tries to pull the deep features of the same class close to the corresponding center [5]. Basically, for classification task $\{(x_i, y_i)\}_{i=1}^{N}$ consisting of N samples x_i and their

corresponding labels $y_i \in \{1, 2, \ldots, Y\}$. Here, x_i is embedding into a new vector $f(x_i)$ with a DNN [17]. The center loss function used here is defined as:

$$L_c(class = j) = -ln\left(\frac{e^{W_j^T x+b}}{\sum_{i=1}^{N} e^{W_i^T x+b_i}}\right) + \lambda\sum_{i=1}^{N} D(f(x_i), c_i), \quad (2)$$

where the first term denotes the softmax loss, the second term denotes the penetration term for clustering, and λ is a balance parameter within [0, 1]. Here,c_i is the center of class i and the function $D(\cdot)$ stands for the distance function. During training, the center loss will encourage instances between samples of the same classes to be closer to their learnable class center. Here, j is the class number, x is the input, W_j is weight, b_j is the bias, and N is the total number of the classes.

3. Experiments and Analysis

3.1. Dataset Description and Preparation

The detailed description of the dataset of the ship-radiated noise called ShipsEar (available at http://atlanttic.uvigo.es/underwaternoise/), which contains a total of 91 records of 11 vessel types and one background noise class, is presented in the literature [16]. During 2012 and 2013, the researchers recorded the sounds of many different classes of ships on the Spanish Atlantic coast. The recordings were made with autonomous acoustic digitalHyd SR-1 recorders, manufactured by MarSensing Lda (Faro, Portugal). This compact recorder includes a hydrophone with a nominal sensitivity of −193.5 dB re 1 V/1 µPa and a flat response in the 1 Hz to 28 kHz frequency range.

To keep consistent with other classification methods, the 11 vessel types are merged into four experiment classes and the background noise type is as one class, as shown in Table 1.

Table 1. A detailed description of the five classes.

Class A	Class B	Class C	Class D	Class E
Background noise	Fishing boats, trawlers, mussel boats, tugboats, and the dredger	Motorboats, pilot boats, and sailboats	Passenger ferries	Ocean liners and ro-ro vessels

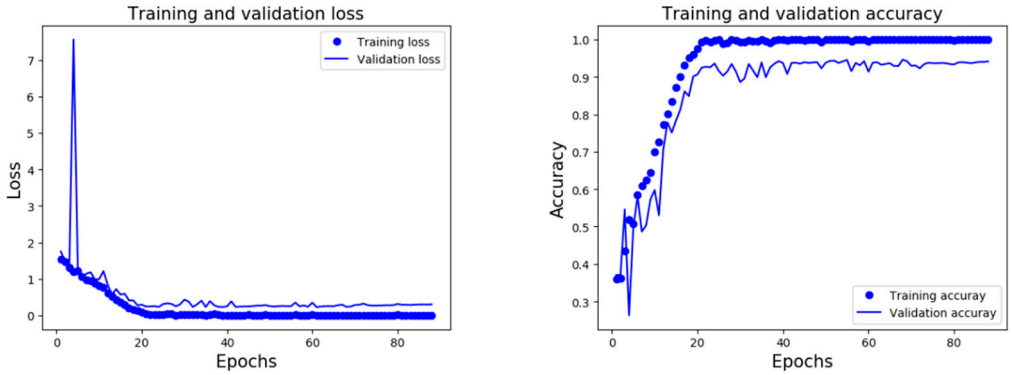
Before preprocessing, the number of the recorded sound clips with a duration of 5 s is 1956 by truncating the original records. The original signals are recorded at different sample rates. By preprocessing, each sound clip is further separated into 41 frames with an overlap of zero.

3.2. Experimental Result

The proposed method is verified by a computer with four GPUs of Nvidia GeForce RTX 2080Ti and Core i7-6900K CPU. Furthermore, the deep learning framework of the proposed model is implemented using Keras 2.2.4 with TensorFlow 1.12.0 as a backend.

When training the model, the batch size and the maximum number of epochs (each epoch includes one training cycle on all training data) are set to be 128 and 200, respectively. To accelerate the training process, the early stopping strategy is that the training will be stopped if the validation loss is reduced by larger than 0.00005 in 20 successive epochs. Besides, the adaptable strategy of learning rate is adopted, where the initial value is 0.001, and the value will be 60% of the former value every 20 epochs. Using such a strategy, the practical number of epochs used is 88. As shown in Figure 5a, the training loss and validation loss are rapidly reduced within about 20 epochs, and they are gradually decreased without overfitting due to the fact of the design of the adaptable learning rate. Meanwhile, Figure 5b also shows that the accuracy is improved at a varied speed. Only

small improvements are obtained after the turn and the best accuracy of the validation data is 0.946. We can observe that Figure 5a, b describes the same process.



(**a**) The loss (**b**) The accuracy

Figure 5. The training and validation accuracy and loss when training the model.

We evaluate our model of ResNet18 with the three-dimensional feature. Twenty percent of the data is used as the validation set and 10% of the data is used as the test set, while each network is trained with 70% of the dataset with the augmentation. Table 2 depicts the detailed results of the classification system at the aspect of precision, recall, and f1-score. It is clear that all the accuracy of the recognition of each class is higher than 0.90 and the average precision or recall or f1-score is 0.943, where the support denotes the number of each test class. For the convenience of comparison, classifier performance is measured using the classification accuracy, defined as the average precision. The ability of the described classifier to identify different vessels is indicated by the fact that there is no confusion between background noise class E and four vessel classes A–D [16]. The vessel classes with the best results are A (background noise) and B (fishing boats, trawlers, mussel boats, tugboats, and the dredger), with classification rates of 0.970 and 0.958, respectively. The poorest results are obtained for C (motorboats, pilot boats, and sailboats). Although the acoustic dataset contains high background noise in shallow water, the overall performance is still satisfactory.

Table 2. The results of an 18-layer Residual Network (ResNet18) with the three-dimensional feature.

	Precision	Recall	F1-Score	Support
Class A	0.970	0.958	0.964	166
Class B	0.958	0.936	0.947	220
Class C	0.917	0.914	0.915	243
Class D	0.945	0.956	0.950	501
Class E	0.935	0.941	0.938	337
Average	0.943	0.943	0.943	1467

Figure 6 shows the confusion matrix of the proposed method on the ShipsEar dataset, where classes 0–4 denote classes A–E, respectively. Values along the diagonal indicate the number of samples classified correctly for each specific class. It shows that class C is the hardest class for the proposed classifier, while all other classes are well separated.

It is worth noting that the parameters such as the number of filters, filter size, and the number of layers and the hyperparameters for training such as the batch size, the initial learning rate, and the patience in early stopping are optimal choices according to the train and validation process in the experiment. Selecting other parameters is also feasible, however, using different parameters could encounter the performance loss and using different hyperparameters does not have significant impact but exhibit slightly inferior performance.

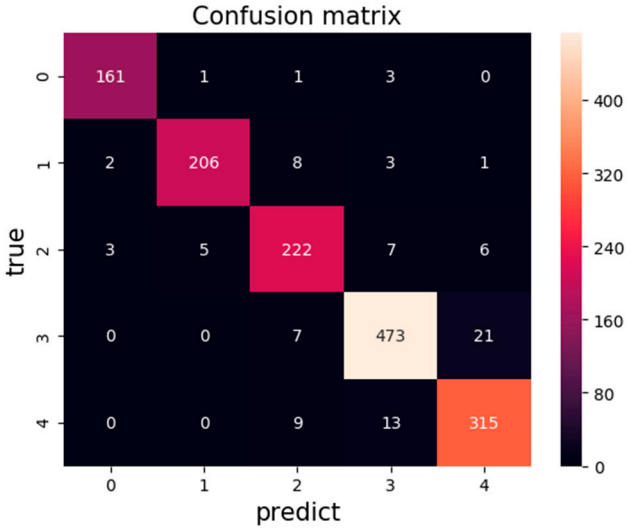


Figure 6. The confusion matrix of the proposed method.

3.3. Experimental Analysis

The effectiveness of three parts of the proposed approach is assessed, i.e., (1) the advantage of the feature extraction method, (2) the advantage of the ResNet18 model, and (3) the contribution of the embedding layer with the center loss function and softmax. Besides, the comparison of the performance of the described method and that of references is presented.

3.3.1. Experiment A: The Advantage of the Feature Extraction Method

This experiment is designed to verify that the three-dimensional features used in the ResNet18 model could yield better results compared with other feature extraction methods. As mentioned in [17], MFCC features have been experimentally proved to be the best hand-crafted features for the recognition, we use MFCC features and its deltas(Δ) and deltas-deltas(Δ^2) for comparison. Besides, considering that LM is a simple but effective feature, we also analyze its performance. The results of different feature extraction methods are shown in Table 3.

Table 3. The comparison of different feature extraction methods with ResNet18.

	MFCC	**MFCC + Δ + Δ^2**	**LM**	**LM + MFCC**	**LM + MFCC + CTZZ**
Class A	0.955	0.982	0.966	0.967	0.964
Class B	0.821	0.844	0.914	0.899	0.947
Class C	0.792	0.821	0.851	0.840	0.915
Class D	0.924	0.919	0.913	0.934	0.950
Class E	0.898	0.845	0.900	0.907	0.938
Average	0.884	0.882	0.906	0.911	0.943

From the results of MFCC and MFCC + Δ + Δ^2, we can see that deltas or deltas-deltas do not contribute in improving the average accuracy, even if the number of parameters increases due to the extra two dimensions. Using the LM feature extracted from the raw signal, we can achieve an average accuracy of 0.906. Therefore, we can see that such a feature is distinguishable with different classes similar to MFCC features. Compared with MFCC or LM, the recognition accuracy of LM + MFCC improves at 0.027 and 0.005, respectively. An important remark is that LM + MFCC + CTZZ achieves the highest recognition accuracy of 0.943, which surpasses 0.032 of the average accuracy of LM + MFCC, although the extraction of such a three-dimension feature increases the computation time.

3.3.2. Experiment B: The Advantage of the ResNet18 Model

This experiment is designed to compare the performance of the described classifier with other typical models. Different networks, named "CNN-1, CNN-2, LSTM, CRNN, and ResNet18," are exploited for comparison.

As shown in Table 4, CNN-1 and CNN-2 share the same network structure, while the features are MFCC and LM + MFCC + CCTZ, respectively. The first layer exploits 32 filters of the size of 3×3 with the activation function of the hyperbolic tangent function named "tanh," followed by the 2×2 max-pooling processing. The second layer uses 64 filters of the size of 3×3 with the activation function of tanh, followed by max-pooling of 2×2. The last layer utilizes 128 filters of the size of 3×3 with the activation function of tanh, followed by max-pooling of 2×2 and batch normalization. Afterward, the fully connected dense layers consist of 1024, 128, and 5 nodes, respectively. The optimizer is Adam with a learning rate of 0.001 and a decay factor of 1×10^{-6}. The batch size and the number of epochs are set to be 128 and 50, respectively. With CNN-2 fed by the optimized feature of LM + MFCC + CCTZ, we can achieve the average accuracy of 0.906, which surpasses that of CNN-1 of 0.845.

Table 4. The comparison of different models.

Model	CNN-1	CNN-2	LSTM	CRNN	ResNet18
Feature	MFCC	LM + MFCC + CCTZ	LM	LM + MFCC + CCTZ	LM + MFCC + CCTZ
Class A	0.926	0.973	0.831	0.936	0.964
Class B	0.806	0.902	0.805	0.867	0.947
Class C	0.744	0.853	0.869	0.832	0.915
Class D	0.881	0.919	0.862	0.908	0.950
Class E	0.832	0.897	0.875	0.876	0.938
Average	0.845	0.906	0.852	0.885	0.943

The LSTM network with the LM feature only achieves an average accuracy of 0.852. The first two LSTM layers have 256 and 128 outputs with batch normalization, respectively. The dropout is set to be 0.2. Afterward, the fully connected part consists of 256 and 5 nodes, respectively. The optimizer is Adam.

As for the CRNN network, the first layer has 64 filters of the size of (5,5) and the regularization method is L2 with the lambda of 0.01. The activation function is ReLU. Batch Normalization is also used and the dropout rate is 0.25. The second and third layers are LSTM, which have 64 units and are regularized L2 with the lambda of 0.01. The dropout rate of recurrent is 0.5, the activation function is tanh, and the dropout rate is 0.25. The fourth and fifth layers are time distributed layers with 128 and 64 nodes. The activation functions are ReLU and the dropout is 0.25. The sixth layer is a dense layer with 5 nodes. The optimizer is Adam with a learning rate of 0.001 and a decay factor of 1×10^{-6}. Such a method could achieve an average accuracy of 0.885.

From Table 4, we could have a better insight into the performance from the detailed accuracy. The ResNet18 with the proposed feature extraction method has an advantage over other methods.

Besides, we also compare the proposed method with other methods. The baseline design is based on the study [16], which shows that using the basic machine learning method, the accuracy is 0.754. Besides, the accuracy achieved by the ResNet18 model as well as that achieved by other state-of-the-art approaches of RBM + BP [18] and RSSD [4] described in the literature are presented in Table 5. Our method achieves an accuracy of 0.943. These results indicate that our ResNet18 has achieved significant improvement in UATR with three-dimensional features as input. Given that the split of the dataset could be different for different methods, the comparison is not rigorous. However, to our knowledge, the best results are obtained by the proposed method for ShipsEar from the accuracy perspective.

Table 5. The comparison of recognition accuracy with other models on ShipsEar.

Model	Accuracy
Baseline [15]	0.754
RBM + BP [14]	0.932
RSSD [4]	0.933
ResNet18 + 3D	0.943

3.3.3. Experiment C: The Contribution of the Embedding Layer with the Center Loss Function and Softmax

In this section, the effect of different loss functions is investigated. Although softmax is the frequently used cost function, the simple utilization of softmax is very easy to become overfitting. For the training process, it tends to increase the value of the outputs before being fed to softmax to reduce the cost. To penalize such behavior, another selectable loss function named "uniform loss" could be a choice that also tries to fit the uniform distribution and is defined as:

$$L_u(class = j) = -\frac{\lambda}{N} ln\left(\frac{e^{W_j^T x + b_j}}{\sum_{i=1}^{N} e^{W_i^T x + b_i}}\right) - (1-\lambda) ln\left(e^{W_j^T x + b_j}\right), \quad (3)$$

where the first term denotes the softmax loss and the second term denotes the uniform distribution. Note that λ is a balance parameter within [0, 1] and is set to be 0.3 in all the classification experiments.

Table 6 summarizes the impacts of different loss functions on the mean accuracy. The results indicated that selecting a good loss function can give a better result for classification. Using the center loss has led to another 3.3% and 0.9% improvement in the average accuracy compared with softmax and uniform loss, respectively. It can be seen in Table 6 that although the accuracy of class A by the model with the center loss is slightly worse than the others, the average performance is better. We can conclude that the performance has been improved by utilizing the embedding layer with the center loss function and softmax.

Table 6. The comparison of the different loss functions.

Model	ResNet18	ResNet18	ResNet18
Feature	LM + MFCC + CCTZ	LM + MFCC + CCTZ	LM + MFCC + CCTZ
Loss function	Softmax	Uniform loss	Center loss
Class A	0.967	0.979	0.964
Class B	0.892	0.938	0.947
Class C	0.874	0.894	0.915
Class D	0.922	0.945	0.950
Class E	0.923	0.924	0.938
Average	0.910	0.934	0.943

4. Conclusions

In this paper, we designed a Residual Network called ResNet18 and the optimizing feature extraction method. The three-dimensional fusion features along with the data augmentation strategy of SpecAugment showed its advantages. Besides, the center loss function and softmax with the embedding layer also contributed to the best performance with the aggregated features using an adaptable learning rate. The accuracy results of 94.3% on the ShipsEar dataset indicated that the proposed method achieved state-of-the-art accuracy. The results showed that the proposed method can extract high-level abstract information that was beneficial to classification. The proposed method provided good technical support for the target classification and recognition function of the Sonar system.

The feature augmentation and target classification of the proposed method under the conditions of different sizes and different SNR are worthy of further study.

Author Contributions: Conceptualization, F.H.; methodology, F.H.; validation, F.H.; investigation, C.L.; writing—original draft preparation, F.H., and C.L.; visualization, C.L., L.G., and F.C.; project administration, F.H. and H.F.; funding acquisition, F.H. and H.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work reported herein was funded jointly by the National Natural Science Foundation of China for Young Scholar (Grant No. 61801471), National Key R&D Program of China (Grant No. 2016YFC0302000), Youth Innovation Promotion Association CAS, the development fund for Shanghai talents, and Jiading Youth Talents Program.

Institutional Review Board Statement: The study did not require ethical approval.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data available in a publicly accessible repository. The data presented in this study are openly available in available at http://atlanttic.uvigo.es/underwaternoise/ at 10.1016/j.apacoust.2016.06.008 in ref [16].

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

References

1. Chen, C.; Fu, J. Recognition method of underwater target radiated noise based on convolutional neural network. *Acoust. Technol.* **2019**, *38*, 424.
2. Pezeshki, A.; Azimi-Sadjadi, M.R.; Scharf, L.L.; Robinson, M. Underwater target classification using canonical correlations. In Proceedings of the Oceans Celebrating Past Teaming Toward Future, San Diego, CA, USA, 22–26 September 2003; pp. 1906–1911.
3. Pezeshki, A.; Azimi-Sadjadi, M.R.; Scharf, L.L. Undersea target classification using canonical correlation analysis. *IEEE J. Ocean. Eng.* **2007**, *32*, 948–955. [CrossRef]
4. Ke, X.; Yuan, F.; Cheng, E. Underwater Acoustic Target Recognition Based on Supervised Feature-Separation Algorithm. *Sensors* **2018**, *18*, 4318. [CrossRef] [PubMed]
5. 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] [PubMed]
6. 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]
7. Jian, L.; Yang, H.; Zhong, L.; Ying, X. Underwater target recognition based on line spectrum and support vector machine. In Proceedings of the International Conference on Mechatronics, Control and Electronic Engineering (MCE2014), Shenyang, China, 29–31 August 2014; Atlantis Press: Paris, France, 2014; pp. 79–84.
8. 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]
9. Zhang, Z.; Xu, S.; Cao, S.; Zhang, S. Deep Convolutional Neural Network with Mixup for Environmental Sound Classification. *arXiv* **2018**, arXiv:1808.08405.
10. Li, J.; Dai, W.; Metze, F.; Qu, S.; Das, S. A comparison of deep learning methods for environmental sound detection. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 126–130.
11. Su, Y.; Zhang, K.; Wang, J.; Madani, K. Environment Sound Classification Using a Two-Stream CNN Based on Decision-Level Fusion. *Sensors* **2019**, *19*, 1733. [CrossRef] [PubMed]
12. 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*, 1214301. [CrossRef] [PubMed]
13. 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]
14. Testolin, A.; Kipnis, D.; Diamant, R. Detecting Submerged Objects Using Active Acoustics and Deep Neural Networks: A Test Case for Pelagic Fish. *IEEE Trans. Mob. Comput.* **2020**. [CrossRef]
15. 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] [PubMed]
16. 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]
17. Yang, H.; Shen, S.; Yao, X.; Sheng, M.; Wang, C. Competitive deep-belief networks for underwater acoustic target recognition. *Sensors* **2018**, *18*, 952. [CrossRef] [PubMed]

18. Luo, X.; Feng, Y. An Underwater Acoustic Target Recognition Method Based on Restricted Boltzmann Machine. *Sensors* **2020**, *20*, 5399. [CrossRef] [PubMed]
19. Abdoli, S.; Cardinal, P.; Lameiras Koerich, A. End-to-end environmental sound classification using a 1D convolutional neural network. *Expert Syst. Appl.* **2019**, *136*, 252–263. [CrossRef]
20. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *arXiv* **2019**, arXiv:1904.08779.
21. He, K.; Zhang, X.; Ren, S. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
22. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A discriminative feature learning approach for deep face recognition. In *ECCV*; Springer: Cham, Switzerland, 2016; pp. 499–515.