Applied Acoustics 208 (2023) 109344

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Data augmentation method for underwater acoustic target recognition based on underwater acoustic channel modeling and transfer learning

Daihui Li [a], Feng Liu [a,*], Tongsheng Shen [a,*], Liang Chen [a,b], Dexin Zhao [a]

[a] National Innovation Institute of Defense Technology, Chinese Academy of Military Science, China
[b] Institute of Ocean Engineering and Technology, Zhejiang University, China

ARTICLE INFO

Article history:
Received 8 September 2022
Received in revised form 9 March 2023
Accepted 21 March 2023
Available online 12 April 2023

Keywords:
Underwater acoustics
Target recognition
Data augmentation
Transfer learning

ABSTRACT

Data augmentation methods as a critical technique in deep learning have not been well studied in the underwater acoustic target recognition, which leads difficult for recognition models to cope with data scarcity and noise interference. This study proposes a data augmentation method based on underwater acoustic channel modeling and Transfer learning to address these challenges. A underwater acoustic channel modeling approach is proposed to generate the augmented signal. A feature-based transfer learning method is presented to narrow the distribution differences between augmented and observed data, and the noise is randomly added to enhance model robustness during training. Dataset acquired in a real-world scenario is used to verify the proposed methods. The proposed methods' effectiveness is proved by utilizing data augmentation in the model training process, which effectively improves the accuracy and noise robustness of the recognition model, especially when observed data is scarce.

1. Introduction

As a data-driven technology, deep learning has made breakthroughs in many fields, and it is no exception in underwater acoustic target recognition [1]. Researchers have researched underwater acoustic target recognition based on neural network structure design, feature extraction and data augmentation. Cao et al. introduce a novel CNN model using second-order pooling to capture the temporal correlations for underwater target classification, which learns the temporary similarities of different CNN filters by computing the covariance matrix of the CNN feature maps along the time axis [2]. Van-Sang Doan et al. propose an approach using a dense CNN model for underwater target recognition, which cleverly uses all former feature maps to optimize classification rates under various impaired conditions while satisfying low computational costs [3]. Tian et al. explore the appropriate structure of deep convolution stacks for perceiving underwater acoustic radiated noise and give full play to the automatic feature learning and extraction capabilities of deep neural networks, a multiscale residual unit (MSRU) is proposed [4]. Depthwise separable convolution and time-dilated convolution are used for underwater acoustic target recognition for the first time by Hu et al. Its accuracy is significantly improved by 6.8% compared with the traditional method [5]. In addition, researchers hope to improve the feature extraction performance of neural networks. The one-dimensional convolution auto encoder-decoder model is proposed to extract features from the high-resonance components. Experiments showed that pre-trained with a large number of unlabeled data, the model can gradually extract invariant and more informative features for underwater acoustic target recognition with increasing layers [6]. Wang et al. proposed a dimension reduction method to obtain the multi-dimensional fusion features of the original underwater acoustic signal [7], which ensures the consistency of the time dimension. And the Gaussian mixture model (GMM) is used to modify the structure of the deep neural network (DNN) to obtain high accuracy and strong adaptability. Zhang et al. used the ensemble learning method to fuse the characteristics of STFT amplitude spectrum, STFT phase spectrum, and bispectrum through the neural network. Compared with a single feature, the fusion feature contains richer target information and has stronger noise robustness [8].

The groundbreaking achievements are attributed to the fact that deep learning technology can automatically learn representative features highly related to specific tasks from a large number of data, rather than the fixed features that may be lost in traditional manual extraction. However, the performance of deep learning automatic feature extraction is seriously affected by the number and quality of training samples, which usually require massive high-quality training samples as support. Obtaining many

* Corresponding authors.
E-mail addresses: liufeng_cv@126.com (F. Liu), shents_bj@126.com (T. Shen).

https://doi.org/10.1016/j.apacoust.2023.109344
0003-682X/

high-quality training samples in practical applications is not always possible. Because for the underwater acoustic target data, the difficulty of data acquisition is not only the high cost of underwater acoustic data acquisition but also closely related to the uncertainty of the target feature information, the changing marine noise environment, and the complexity of the underwater acoustic channel. The target feature information is the information that can accurately and simplify the target state and identity contained or extracted from the signal data received by the hydrophone, which depends on the structure, class, and navigation state of the target. The structure and class of the target may be related to the category of the classification task. However, the navigation state is irrelevant and changeable to the category of the classification task. In addition, the underwater acoustic channel and marine environmental noise are also complex and changeable. These factors affect the underwater acoustic target signal received by the hydrophone, making it difficult or even impossible to obtain complete and high-quality massive underwater acoustic target data. Therefore, it is difficult for the underwater acoustic target recognition model to extract useful features, resulting in low classification accuracy or poor robustness of the deep learning model. In order to solve the problem that the recognition model is hard to train due to the scarcity of underwater acoustic target data, researchers have begun to study the enhancement and augmentation of underwater acoustic target data. Liu et al. used specaugment technology to augmentation data on the time axis and frequency axis of time–frequency features, improving the model's recognition accuracy [9]. Satheesh Chandran C. et al. used a generative adversarial network to model the causal attributes of target labels so that the network can tolerate the distortion caused by environmental noise and channel artifacts. They systematically evaluated the target instances collected from different locations in the Indian Ocean and achieved encouraging results [10]. Luo et al. used the Boltzmann self-encoder to reconstruct the original signal. They generated the output signal with the same overall features as the input signal, effectively increasing the amount of data in the data set [11]. However, most of these methods are based on pure signal processing or conventional machine learning methods for data augmentation or enhancement. They do not propose solutions from the propagation characteristics of underwater acoustic signals or the root cause of incomplete underwater acoustic target data.

This paper proposes a data enhancement method based on underwater acoustic channel modeling. Therefore, the deep learning model can effectively use underwater acoustic channel modeling technology for data enhancement through transfer learning and ultimately effectively improve the recognition accuracy under scarce training data. The ship's radiated noise signal received by the hydrophone contains the features of the ship itself and the information brought about by the underwater acoustic channel. This paper aims to increase the number of training samples by introducing underwater acoustic channel knowledge in the training process so that the neural network can easily extract the Distinguishable feature information of the target. For underwater acoustic channels, scholars of underwater acoustic physics have conducted many studies on the propagation law of sound waves in the ocean and the simulation of underwater acoustic signals and proposed a series of typical simulation modeling methods and related tools for underwater acoustic propagation characteristics [12–14]. On the other hand, the transfer learning method has been widely recognized and applied in various fields of pattern recognition such as image and signal processing [15–18]. It has already developed into multiple subdivision research fields, including model transfer, domain adaptation, and domain generalization. Therefore, this paper focuses on the data enhancement of underwater acoustic target recognition based on underwater acoustic channel modeling and transfer learning methods. The simulation of acoustic propagation characteristics of underwater acoustic signals can use a numerical calculation to simulate the propagation attenuation and frequency response of underwater acoustic channels close to the actual situation according to the profile of sound velocity, conditions of the sea surface, seabed boundary, and the location of sound source and receiver. Based on the underwater acoustic channel response obtained by underwater acoustic channel modeling, this paper uses the underwater acoustic channel response to augment the underwater acoustic target data, and proposes a transfer learning method using augmented data to train the target recognition model.

The rest of this paper is organized as follows: in Section 2, the method of modeling the underwater acoustic channel and generating the augmented data is briefly presented. In Section 3, the developed transfer learning method and how they can be used to improve recognition accuracy based on the augmented data is described. The experimental results are analyzed in Section 4, and conclusions are derived in Section 5.

2. Time-domain signal augmentation based on underwater acoustic channel modeling

As a classical simulation method, underwater acoustic channel modeling can simulate the time domain signal received by the hydrophone through numerical calculation. It has been widely used in underwater acoustic signal processing. In general, the wave equation for an ideal medium is described as follows:

$$\nabla^2 p(x,y,z,t) - \frac{1}{c_0^2}\frac{\partial^2 p(x,y,z,t)}{\partial t^2} = 0, \tag{1}$$

where ∇^2 denotes the laplacian operator, $p(x,y,z,t)$ denotes the acoustic pressure in the angle coordinate system (x,y,z), t is the temporal variable, and c is acoustic velocity. Considering the acoustic excitation, the Helmholtz equation is as follows:

$$\left[\nabla^2 + k^2(\boldsymbol{r})\right]\psi(\boldsymbol{r},\omega) = f(\boldsymbol{r}_0,\omega), \tag{2}$$

where $k(\boldsymbol{r}) = \omega/c(\boldsymbol{r})$ denotes the wave number, $\psi(\boldsymbol{r},\omega)$ denotes the potential function, ω is the angular frequency of the acoustic source, c is the acoustic velocity of the medium, and $\boldsymbol{r}$ is the distance between the receiving point and the acoustic source. The acoustic pressure and underwater acoustic channel response at the receiving point can be obtained by adding the energy of all the acoustic rays reaching the receiving point. The function can be expressed as:

$$p(x,y,z) = \sum_{i=1}^{N} p_i(x,y,z), \tag{3}$$

$$\boldsymbol{R}_n = \sum_{i=1}^{N} A_{n,i} * \mathrm{e}^{-\boldsymbol{f} d_{n,i}}, \tag{4}$$

where $p_i(x,y,z)$ represents the acoustic pressure of the i-th acoustic ray, the total number of acoustic rays is N, the amplitude is A, and the delay is d. $\boldsymbol{f}$ is the frequency sequence and $\boldsymbol{R}$ is the response matrix of underwater acoustic channel. Based on the underwater acoustic channel response, the ship radiated noise signal collected by the hydrophone is transmitted through the underwater acoustic channel to generate a time domain simulation signal received at a specific location. The simulation signal is generated based on the frequency domain method. Transform the input signal into the frequency domain. The frequency domain is multiplied by the system's frequency response in broadband, and the time domain simulation signal is obtained by inverse Fourier transform. The function is described as follows:

$$Y(k) = X(k)H(k), \tag{5}$$

where $X(k)$ represents the Fourier transform of the input signal, $H(k)$ represents the channel frequency response, and $Y(k)$ represents the Fourier transform of the output signal. The valuable augmentation signals can be obtained by adding noise to the simulation signals. The time–frequency map of observed and augmented signals is shown in Fig. 1.

3. Data augmentation method based on transfer learning

3.1. Transfer learning method based on augmented data

Data augmentation processes the training data through a specific method, effectively increasing the sample size of the training data without destroying the target features so that the neural network can learn the classifiable features of the target. However, there is some difference between the augmentation data and the data collected by the hydrophone. It is not convenient for the neural network to search for the classifiable features of the target in the data with the difference. In this paper, a transfer learning method is designed so that the recognition model can reasonably utilize augmented data for training enhancement. Firstly, the underwater acoustic channel is modeled according to the actual scene. Secondly, the received time-domain augmented signals at multiple receiving positions are generated based on the observed ship radiated noise signal collected by the hydrophone and the underwater acoustic channel model. Finally, based on the widely concerned method of domain adaptation [19], the augmented and observed signals are utilized as the data of the source domain and target domain, respectively. Build effective learning models, network models, and losses to narrow the distribution differences between source and target domain data. The propose method is presented in Fig. 2.

In this paper, the main idea of transfer learning is to map the input data to a common feature space. By matching appropriate metrics in the feature space, the feature distribution of the source domain and the target domain data is as close as possible. In the training phase, the augmented and observed signals are source domain data and target domain data, respectively. We align the simulated signals at different receiving positions with the observed signals at the transmitter in the feature space to enhance the model's ability to weaken the simulated channel. The proposed method redesigns the training stage to achieve this process. The neural network is used to map the input data to the feature space, and the network learning feature distribution of three shared parameters is constructed. Shared parameters mean that the three networks use the same backbone network parameters, which are affected by the three networks during training and updated synchronously. After the training, the shared parameters can be directly used as the parameters of the backbone network in the test stage.

The proposed method is different between the training and test stages. The training stage is set as two main modules to conduct a variety of learning modes and data loading methods to help the backbone network train. The test stage is a single main module, which can directly use the trained backbone network to perform forward operations to complete the classification of target categories. Since the architecture adopts a network structure with shared parameters, all trained parameters can be transferred from the training module to the testing module. Therefore, the test stage does not add any redundant computation compared to the method without transfer learning. The training stage consists of two main modules, module A and module B. Module A is used to reduce the distribution difference between the source domain and target domain, and module B is used to improve the model's performance to deal with signals disturbed by noise. During training, module A randomly selects two batches of data from the source and target domains, respectively, which is for gradient back-propagation one time. The source domain data is selected from the augmented signal received at a random receiving location, as shown in Fig. 3. In addition, noise is added to the data to simulate different signal-to-noise ratios (SNR). The source domain data from module A is randomly added with noise to a simulated SNR of 20 dB to 150 dB. The domain adaptation loss is used to minimize the covariance difference of features in different domains. In order to complete the learning of the classification task, the source domain label and the target domain label are respectively sent to the output of module A to construct the classification loss. At the same time, the input data of module B is augmented on the target domain data of module A, and the simulated SNR of the samples input to module B is −5 dB to 20 dB. Module B makes the neural network suitable for low SNR situations through distance and distribution constraints and learns classification through classification loss.

3.2. Backbone network of classification tasks

The backbone network mainly determines the performance of recognition system. This paper introduces the main structure of ECAPA-TDNN [20] as the backbone network for feature extraction. ECAPA-TDNN is a delayed neural network, and time-delay neural

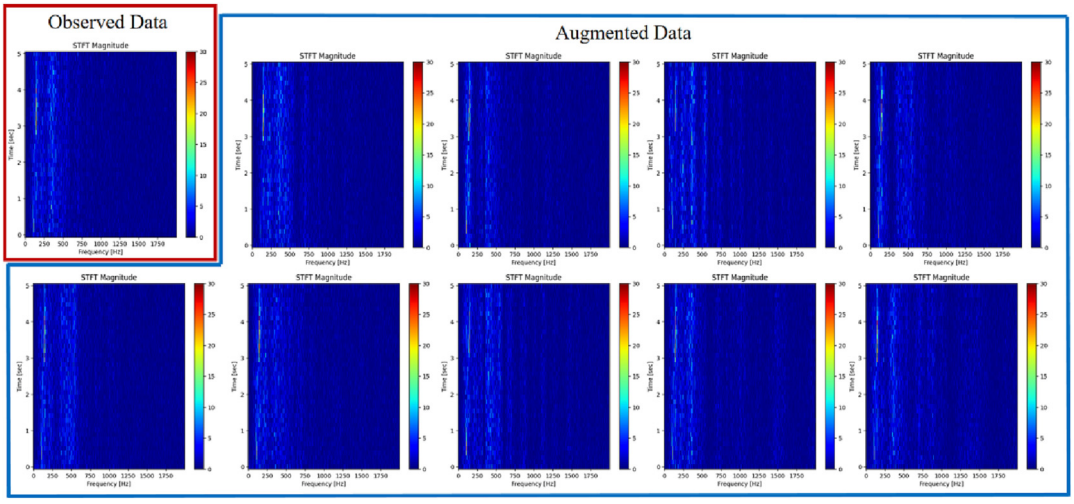

Fig. 1. The time–frequency map of observed and augmented signals.

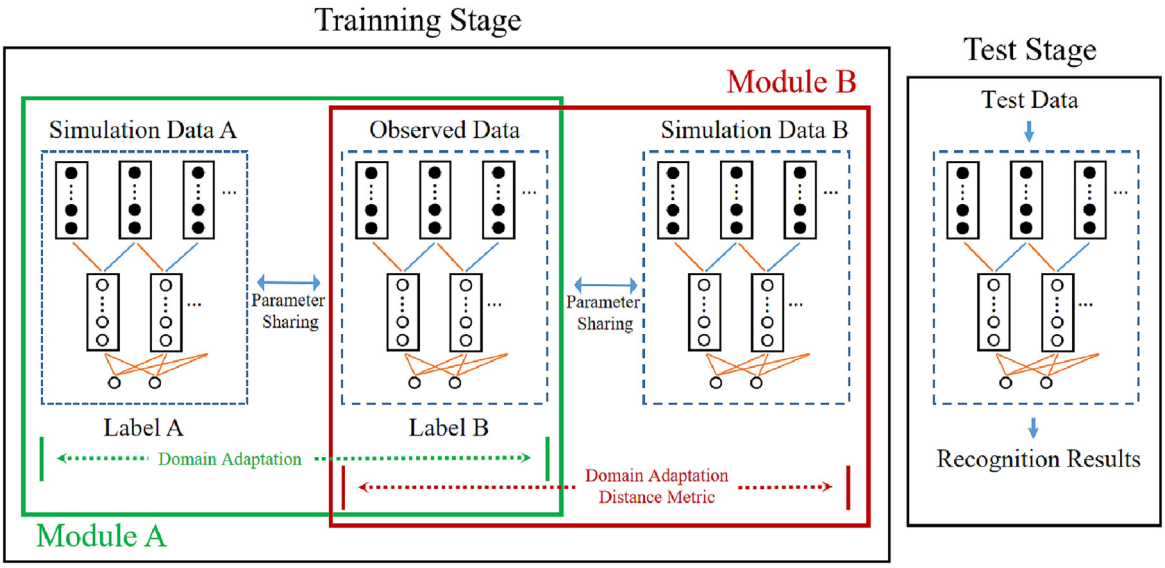

Fig. 2. The framework of the proposed method.

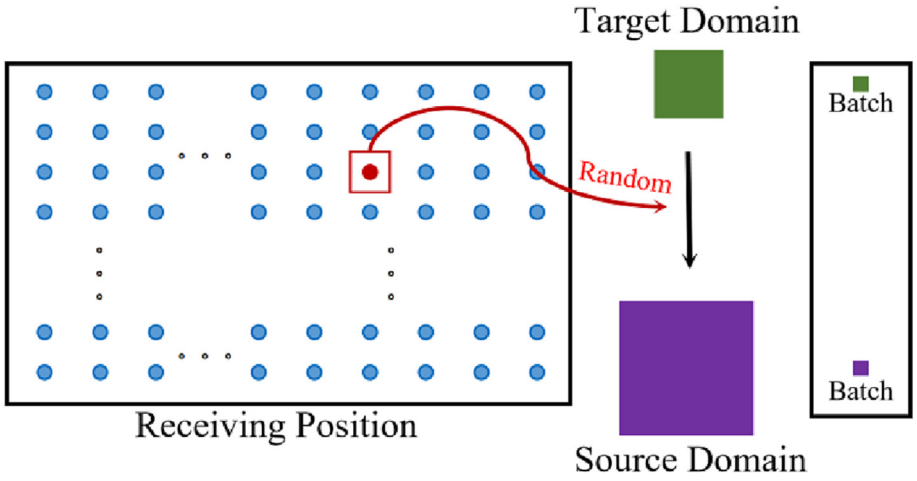

Fig. 3. Sampling strategy in model training.

networks are usually composed of multiple feedforward network layers. Time-delay neural networks model the feature relationship within the adjacent time frames through each network layer, and model the feature relationship across time frames through layers stacking. It is suitable for modeling the time dependence between features. The backbone network needs to extract features from the frequency and establish the time dependence between frequencies for the time–frequency features-based underwater acoustic target recognition algorithm. Therefore, it is suitable for modeling the radiated noise signal of underwater acoustic targets. As a time-delay neural network with outstanding performance, ECAPA-TDNN introduces many improvement schemes from advanced technologies in machine vision, speech processing, and other fields, including SE block[21], residual network [22], and temporal self-attention system [23,24]. ECAPA-TDNN applies statistics pooling to project variable-length signals into fixed-length target characterizing embeddings. Specifically, ECAPA-TDNN adopts a SE-RES2Block with a channel attention function to effectively improve the feature extraction performance of the neural network. In addition, the ECAPA-TDNN uses a multi-layer feature aggregation structure and dense connection network to generate statistics features. The statistics pooling of the ECAPA-TDNN projects variable-length signals into fixed-length target characterizing embeddings, which gives the final features extracted by the backbone network a strong expression performance. The network topology of ECAPA-TDNN is shown in Fig. 4, and the parameter setting is according to [20]. The parameters of the Conv layers and SE-Res2Block layers are shown in Table 1.

3.3. Loss function

This paper uses a multi-class classification problem as an example to describe the proposed method. A recognition system with excellent performance needs a strong feature extraction network. However, the performance of feature extraction depends not only on the backbone network but also on training methods. As mentioned above, we utilize a robust backbone network ECAPA-TDNN to extract depth features. We hope that the features not only have excellent class distinction but also adapt to the difference between augmented and observed data so that the algorithm can efficiently use the augmented data to assist the backbone network training. Therefore, we introduce a domain adaptation loss named CORAL Loss [25], which is defined as the covariance of source and target features:

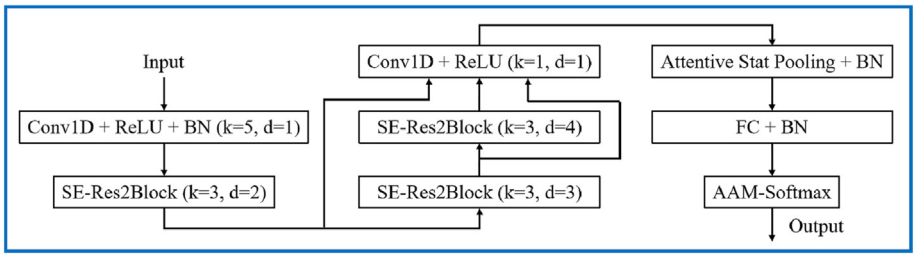

Fig. 4. The network topology of ECAPA-TDNN.

Table 1
Parameters of Conv layers and SE-Res2Block layers in ECAPA-TDNN.

Layers	channels	Kernel size	Stride/Dilation	Scale
Conv1	1024	5	1	–
SE-Res2Block	1024	3	2	8
SE-Res2Block	1024	3	3	8
SE-Res2Block	1024	3	4	8
Conv1	1536	1	1	–

$$l_{\text{coral}} = \frac{1}{4d^2} \| C_S - C_T \|_F^2, \tag{6}$$

where $\| \cdot \|_F^2$ is the squared matrix Frobenius norm, d denote size of source feature, the C_S and C_T denote the covariance matrices of the source and target data, which are given by:

$$C_S = \frac{1}{n_S - 1}(D_S^{\mathrm{T}} D_S - \frac{1}{n_S}(\mathbf{1}^{\mathrm{T}} D_S)^{\mathrm{T}}(\mathbf{1}^{\mathrm{T}} D_S)), \tag{7}$$

$$C_T = \frac{1}{n_T - 1}(D_T^{\mathrm{T}} D_T - \frac{1}{n_T}(\mathbf{1}^{\mathrm{T}} D_T)^{\mathrm{T}}(\mathbf{1}^{\mathrm{T}} D_T)), \tag{8}$$

where $\mathbf{1}$ denotes a column vector with all elements equal to 1.

The CORAL loss introduces the network to extract invariant features between different data domains. These invariant features include invariant features between augmented data and observed data, and between high SNR data and low SNR data, as shown in Fig. 2. We use Euclidean distance to constrain the depth features of different SNR signals to improve the ability of the model to classify low SNR signals. In addition, only conducting CORAL loss and distance loss may lead to the collapse of features. The neural network may project all signal data to a single point, making the domain adaptive loss and distance loss zero. In this case, the extracted features are not identifiable, and the classifier with good performance cannot be constructed using these features. Therefore, this paper proposes a joint loss method, which utilizes CORAL loss, classification loss, and distance loss to jointly train the network so that the model can still learn valuable features even when the target domain data are scarce. The joint loss is given by:

$$l = \sum_{i=1}^{t} a_i l_{\text{coral}} + \sum_{j=1}^{k} b_j l_{\text{class}} + \sum_{n=1}^{p} c_n l_{\text{mse}}, \tag{9}$$

where t denotes the number of CORAL loss layers in a deep network, and a is a changeable weight with CORAL loss, the same as CLASS loss and MSE loss. These losses are balanced after training, and the trained model will perform well in the scarce real training data.

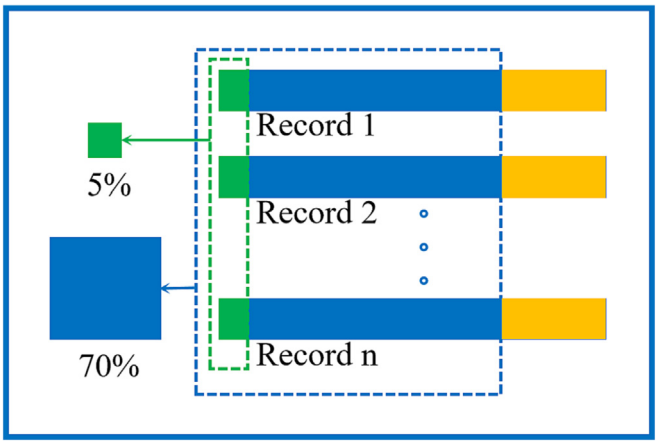

Fig. 5. Train the model with different sample sizes.

4. Experiments and results

4.1. Experimental data

In this paper, the proposed methods are verified based on a measured dataset named ShipsEar [26], made during the autumn of 2012 and summer of 2013 in different parts of the Spanish Atlantic coast in northwest Spain. The dataset comprises 90 records representing sounds from the natural environment noise and four types of ship-radiated noise. The types are merged from the original 11 vessel types. Most records were made in a river valley, 35 km long, 10 km at its widest point, and with a maximum depth of under 45 m. The simulation data receiving point is set every 1 m in depth and every 100 m in distance. The sound source depth is 6 m, with 25 reception points vertically and 50 reception points horizontally, totaling 1250 reception points. In the training stage, many samples of augmentation data are generated according to the method described above, and the receiving point is randomly selected.

The data is divided according to the proportion of training data in each record file to evaluate the performance of the algorithm on different amounts of training samples. We use 5% and 70% of the training samples to divide the records to evaluate the proposed methods, as shown in Fig. 5. In addition, the front of each record file is divided into the training set, and the tail is the test set.

After dividing the records, all samples are resampled at a sampling frequency of 16 kHz and divided into signals of 5 s in length. When 5% of the data is used as training samples, the training set contains 138 samples, and the test set contains 567 samples. When using 70% of the data as training samples, the training set contains 1434 samples, and the test set contains 567 samples.

4.2. Experimental methods

This paper proposes a data augmentation method to improve recognition accuracy. The proposed signal augmentation method generates many augmented signals through underwater acoustic channel modeling, effectively expanding the number of training samples. A transfer learning method is designed based on

Table 2
Comparison of recognition performance with 70% samples.

Method	Precision	Recall	F1-score	Accuracy
Baseline	0.776	0.814	0.789	0.800
ECAPA-TDNN	0.908	0.897	0.902	0.908
Spec	0.908	0.884	0.895	0.905
SRAN	0.906	0.913	0.909	0.915
AD	0.905	0.918	0.911	0.921
Trans-AD	0.917	0.921	0.919	0.921
Trans-AD + AG	0.925	0.932	0.929	0.928

Table 3
Comparison of recognition performance with 5% samples.

Method	Precision	Recall	F1-score	Accuracy
Baseline	0.557	0.579	0.546	0.564
ECAPA-TDNN	0.594	0.607	0.598	0.645
Spec	0.619	0.603	0.607	0.666
SRAN	0.604	0.600	0.597	0.666
AD	0.643	0.650	0.641	0.696
Trans-AD	0.706	0.705	0.705	0.739
Trans-AD + AG	0.724	0.682	0.690	0.756

parameter sharing to improve the utilization efficiency of augmentation data and the performance of the recognition model to recognize the low SNR signals. Therefore, we employ two groups of experiments to evaluate the performance of the proposed methods. The first group of experiments is to evaluate the classification performance of the proposed methods based on the observed data, which include: 1. The classification performance of ECAPA-TDNN as the backbone network. 2. The classification performance is based on mainstream data augmentation and ECAPA-TDNN. 3. The classification performance of the proposed signal augmentation method and transfer learning method. 4. Evaluation of a proposed methods combined with the existing mainstream data augmentation strategies. The second group of experiments is to add different noises to the original signal to simulate different SNRs and evaluate the proposed methods' anti-noise performance. Two types of noise are added to the experiment, which are colored noise and white noise. Colored noise is included in ShipsEar, and white noise is randomly generated. Different types of noise are used to train the models according to different interference factors in the evaluation stage, and the noise used in the training and test stages is isolated. The MFCC [27] and X-vector [28] be conducted as the Baseline, which is widely used in speech recognition and underwater acoustic target recognition. The weights of a, b, and c in the loss function are set to 10, 1, and 1, respectively. All methods are carried out under the two proportions of training data. F-Bank is used as the input feature, and ECAPA-TDNN is employed as the backbone network in all experiments except for Baseline. We use Accuracy, Recall, Precision, and F1-Score to characterize recognition performance.

4.3. Experimental results

4.3.1. Classification experiment results

This experiment is designed to demonstrate the performance of the proposed methods on the recognition task, and the test data include 4-category underwater acoustic targets. The recognition results of each model are shown in Table 2. Firstly, we compare the performance of Baseline and ECAPA-TDNN. The Baseline represents the recognition performance of mainstream method based on deep learning. ECAPA-TDNN is an excellent model in speech processing, and its applicability in underwater acoustic target recognition is evaluated in this section. Secondly, the performance of Specaugment (Spec) [9,29] and the random noise addition (SRAN) on the backbone model are compared. Specaugment (Spec) is one of the most popular data augmentation methods in sound event recognition. SRAN randomly adds noise to the observation samples and can be used as a data augmentation method. Finally, we evaluate the proposed methods separately, including directly using the proposed augmented data (AD) to train ECATA-TDNN, using the transfer learning strategy to improve the utilization of augmented data (Trans-AD), and using the Spec and SRAN strategy to enhance the training of Trans-AD (Trans-AD + AG). Except for Baseline, all

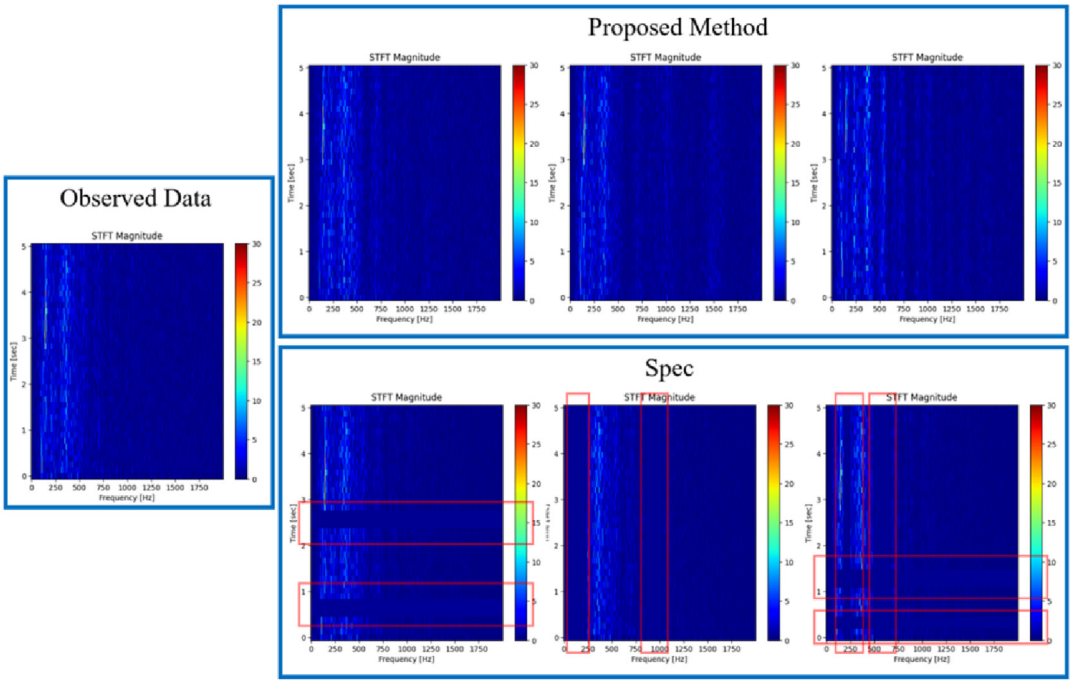

Fig. 6. The proposed data augmentation method is compared with Spec.

methods are based on ECAPA-TDNN as the backbone network. In addition, the SRAN strategy is equivalent to adding the same noise as AD based on ECAPA-TDNN. All experiments train the model using 70% or 5% samples, and the results are in Table 2 and Table 3.

Experiments in Tables 2 and 3 prove that ECAPA-TDNN not only has an excellent performance in speech processing but also in the field of underwater acoustic target recognition. Spec, as a widespread data augmentation method, brings limited performance gains, and the performance of Spec depends on factors such as mask width and the number of masks. Spec randomly conduct masks on the frequency axis and time axis of the signal spectrum to increase the number of training samples. However, the random mask may hide some necessary information and restrict the model's performance, as shown in Fig. 6.

The proposed methods AD, Trans-AD and Trans-AD + AG performed excellent results in all experiments, especially in few-shot training. The most important reason is that the proposed methods effectively increases the number of training samples. The proposed methods enable the neural network to extract deep features suitable for classification from massive data, and reduces the probability of overfitting the neural network in training. Compared with the Spec, the proposed signal augmentation method generates massive training data based on the underwater acoustic channel modeling that may be close to the observed signal. Spec uses a random mask strategy to generate training samples, and the generated training samples are quite different from the observed signals. In addition, Trans-AD + AG achieved the highest recognition accuracy in all experiments. The results show that the Trans-AD is not mutually exclusive with SRAN and Spec, and the recognition accuracy can be further improved by using SRAN and Spec technology. Next, the observed signal is added with different types of noise to evaluate the models.

4.3.2. Experimental results with different noise levels

Practical models usually have excellent robustness against certain noise disturbances. Underwater noise is complex, and anti-noise performance is crucial in underwater acoustic target recognition. In this paper, white noise and colored noise with different intensities is added into the observed signal which is used to evaluate the models. The first experiment is to add different intensities of the white noise interference recognition model, where the comparison results are plotted in Fig. 7.

It can be noticed that the recognition accuracy of almost all methods generally rises along with the increment of SNR. The proposed data augmentation method introduces a large number of valuable signals. It improves the utilization of these augmented

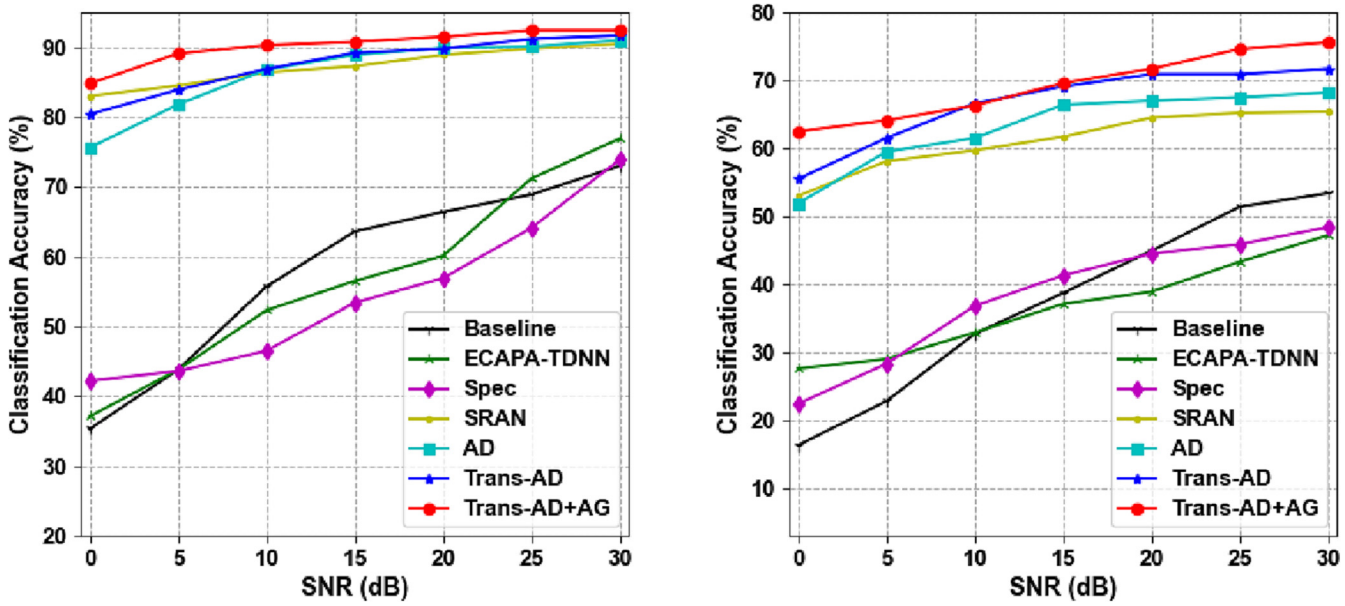

Fig. 7. Comparison of anti-white noise performance. The training data on the left is 70%, and on the right is 5%.

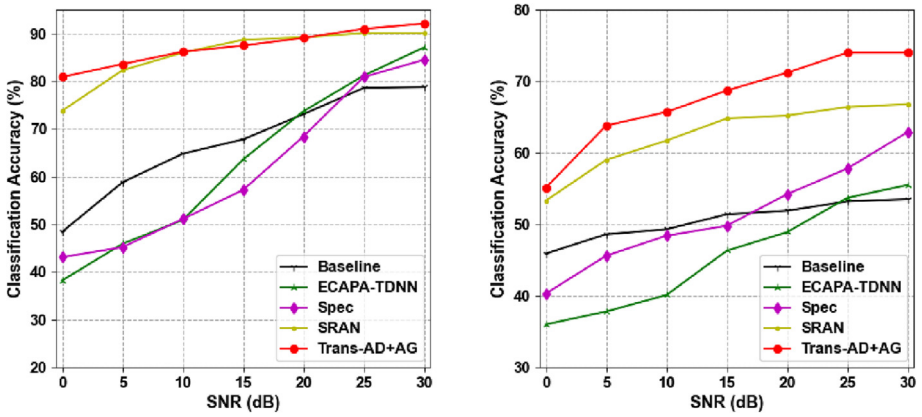

Fig. 8. Comparison of anti-colored noise performance. The training data on the left is 70%, and on the right is 5%.

signals based on the transfer learning method to perform strongly in identifying signals impaired by additive noise. Experiments demonstrate the effectiveness of the proposed methods, especially when training data is scarce, which is greater than the Baseline by approximately 39.2% at 0 dB SNR. The comparison results of recognition accuracy disturbed by colored noise are shown in Fig. 8. We only selected the Trans-AD + AG that performed best in the previous experiments to compare with other methods.

The experimental results of anti-colored noise show a similar trend to the first group of experimental results. It is worth noting that the ocean noise is complex, and the colored noise used in the experiment was collected on the same sea area as the experimental dataset. Therefore, this experiment can only demonstrate the model with strong anti-colored noise performance in a specific environment. This paper provides a developmental approach, especially when training data is scarce.

5. Conclusion

In this paper, an underwater acoustic target signal augmentation method is proposed based on underwater acoustic channel modeling, and a transfer learning method is proposed to improve the utilization efficiency of augmented data. The proposed methods consider the interference caused by the underwater acoustic channel to the target features, effectively utilizing these factors to augment the training data for the first time. In this way, the neural network can efficiently extract the essential features of the target. In addition, underwater acoustic target recognition models are usually impaired by the ubiquitous noise, which is taken into account when generating augmented data. The proposed methods effectively improve the anti-interference and recognition performance of the model and is applied to practical neural network model modeling problems, especially when the training data is scarce.

Acknowledgements

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

CRediT authorship contribution statement

Daihui Li: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft. **Feng Liu:** Methodology, Funding acquisition, Investigation, Writing – original draft. **Tongsheng Shen:** Project administration, Resources, Supervision, Investigation. **Liang Chen:** Visualization, Software, Writing – review & editing. **Dexin Zhao:** Data curation, Validation, Writing – review & editing.

Data availability

The authors do not have permission to share data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Domingos LCF, Santos PE, Skelton PSM, Brinkworth RSA, Sammut K. A Survey of Underwater Acoustic Data Classification Methods Using Deep Learning for Shoreline Surveillance. Sensors 2022;22(6):2181.
[2] Cao Xu, Togneri R, Zhang X, Yu Y. Convolutional Neural Network With Second-Order Pooling for Underwater Target Classification. IEEE Sens J 2019;19(8):3058–66.
[3] Doan V, Huynh-The T, Kim D-S. Underwater Acoustic Target Classification Based on Dense Convolutional Neural Network. IEEE Geosci Remote Sens Lett 2022;19:1–5. https://doi.org/10.1109/LGRS.2020.3029584.
[4] Tian S, Chen D, Wang H, Liu J. Deep convolution stack for waveform in underwater acoustic target recognition. Sci Rep 2021;11(1). https://doi.org/10.1038/s41598-021-88799-z.
[5] Hu G, Wang K, Liu L. Underwater Acoustic Target Recognition Based on Depthwise Separable Convolution Neural Networks. Sensors 2021;21(4):1429. https://doi.org/10.3390/s21041429.
[6] Ke X, Yuan F, Cheng E. Underwater Acoustic Target Recognition Based on Supervised Feature-Separation Algorithm. Sensors 2018;18(12):4318. https://doi.org/10.3390/s18124318.
[7] Wang X, Liu A, Zhang Y, et al. Underwater Acoustic Target Recognition: A Combination of Multi-Dimensional Fusion Features and Modified Deep Neural Network. Remote Sens (Basel) 2019;11(16):1888. https://doi.org/10.3390/rs11161888.
[8] Zhang Qi, Da L, Zhang Y, Hu Y. Integrated neural networks based on feature fusion for underwater target recognition. Appl Acoust 2021;182:108261.
[9] Liu F, Shen T, Luo Z, Zhao D, Guo S. Underwater target recognition using convolutional recurrent neural networks with 3-D Mel-spectrogram and data augmentation. Appl Acoust 2021;178:107989.
[10] Chandran SC, Kamal S, Mujeeb A, et al. Generative adversarial learning for improved data efficiency in underwater target classification, Engineering. Sci Technol 2022;30:101043.
[11] Luo X, Feng Y, Zhang M. An underwater Acoustic target recognition method based on combined feature with automatic coding and reconstruction. IEEE Access 2021;9:63841–54. https://doi.org/10.1109/ACCESS.2021.3075344.
[12] Porter MB, Bucker HP. Gaussian beam tracing for computing ocean acoustic fields. J Acoust Soc Am 1987;82:1349–59. https://doi.org/10.1121/1.395269.
[13] Porter MB. The KRAKEN normal mode program. Washington D C: Naval Research Laboratory; 1992.
[14] Gilbert KE, Evans RB. A Green's Function Method for One-Way Wave Propagation in a Range-Dependent Ocean Environment. In: Akal T, Berkson JM, editors. Ocean Seismo-Acoustics. Boston, MA: Springer US; 1986. p. 21–8.
[15] Shao L, Zhu F, Li X. Transfer Learning for Visual Categorization: A Survey. IEEE Trans Neural Networks Learn Syst 2015;26(5):1019–34. https://doi.org/10.1109/TNNLS.2014.2330900.
[16] Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, et al. A Comprehensive Survey on Transfer Learning. Proc IEEE 2021;109(1):43–76.
[17] Mittal S, Srivastava S, Jayanth JP. A Survey of Deep Learning Techniques for Underwater Image Classification. IEEE Transactions on Neural Networks and Learning Systems, 2022, Early Access Article, https://doi.org/10.1109/TNNLS.2022.3143887.
[18] Nguyen CT, Van Huynh N, Chu NH, Saputra YM, Hoang DT, Nguyen DN, et al. Transfer Learning for Wireless Networks: A Comprehensive Survey. Proc IEEE 2022;110(8):1073–115.
[19] Wang M, Deng W. Deep visual domain adaptation: A survey. Neurocomputing 2018:135–53. https://doi.org/10.1016/j.neucom.2018.05.083.
[20] Desplanques B, Thienpondt J, Demuynck K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. Interspeech 2020:3830–4. https://doi.org/10.21437/Interspeech.2020-2650.
[21] Hu J, Shen Li, Albanie S, Sun G, Wu E. Squeeze-and-Excitation Networks. IEEE Trans Pattern Anal Mach Intell 2020;42(8):2011–23.
[22] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition 2016:770–8. https://doi.org/10.1109/CVPR.2016.90.
[23] Okabe K, Koshinaka T, Shinoda K. Attentive statistics pooling for deep speaker embedding. Interspeech 2018:2252–6. https://doi.org/10.21437/Interspeech.2018-993.
[24] Zhu Y, Ko T, Snyder D, et al. Self-attentive speaker embeddings for text-independent speaker verifi-cation. Interspeech 2018:3573–7. https://doi.org/10.21437/Interspeech.2018-1158.
[25] Sun B, Saenko K. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. European Conference on Computer Vision 2016;9915:443–50. https://doi.org/10.1007/978-3-319-49409-8_35.
[26] Santos-Domínguez D, Torres-Guijarro S, Cardenal-López A, Pena-Gimenez A. ShipsEar: An underwater vessel noise database. Appl Acoust 2016;113:64–9.
[27] Wang W, Li S, Yang J, et al. Feature extraction of underwater target in auditory sensation area based on MFCC. IEEE/OES China Ocean Acoustics (COA) 2016;2016:1–6. https://doi.org/10.1109/COA.2016.7535736.
[28] Zeinali H, Burget L, Cernocky J. Convolutional neural networks and x-vector embedding for DCASE2018 acoustic scene classification challenge. Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), 2018.
[29] Park DS, Chan W, Zhang Y, et al. Specaugment: A simple data augmentation method for automatic speech recognition. Interspeech 2019:2613–7. https://doi.org/10.21437/Interspeech.2019-2680.