Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19178-9

Underwater target recognition based on adaptive multi-feature fusion network

Xiaoying Pan[1,2,3] · Jia Sun[1,2,3] · TianHao Feng[1,2,3] · MingZhu Lei[1,2,3] · Hao Wang[1,2,3] · WuXia Zhang[1,2,3]

Received: 28 September 2023 / Revised: 6 February 2024 / Accepted: 2 April 2024

Abstract
Due to the complexity of the underwater environment, underwater acoustic target recognition is more challenging than ordinary target recognition, and has become a hot topic in the field of underwater acoustics research. In recent years, deep learning has been widely used in underwater recognition due to its powerful feature learning capabilities. However, most deep learning-based underwater target recognition methods use only the time-frequency domain features of signals, ignoring the frequency domain image features of signals. Additionally, they only use simple summation or splicing strategies to fuse and discriminate the extracted features, which leads to limited improvement in the accuracy of underwater target recognition. To address these issues, this paper proposes an underwater target recognition method based on an adaptive multi-feature fusion network. The method consists of a data pre-processing module, a multi-dimensional feature extraction module, and an adaptive multi-feature fusion module. In the multi-dimensional feature extraction module, the long short time memory network (LSTM) and the one-dimensional convolutional neural network (1DCNN) are used to extract the time-frequency features of the underwater signal. Furthermore, the two-dimensional convolutional neural network (2DCNN) is used to extract the image features in the frequency domain. The adaptive multi-feature fusion module uses an attention mechanism for adaptive weighted fusion to make full use of the learned features. The effectiveness of the proposed method is validated on the ShipEar dataset, and the recognition accuracy reaches 94.92%, which is higher than other existing methods.

Keywords Underwater target recognition · Deep learning · Adaptive weighting · Feature fusion

1 Introduction

Underwater target recognition is one of the most important research directions in underwater acoustics signal processing and a hot topic in the field of underwater acoustics [1]. It has significant importance in both the national economy and defense and military fields. Under-

✉ WuXia Zhang
wuxiazhang100@126.com

Extended author information available on the last page of the article

Published online: 24 April 2024

water acoustic signals are widely used in underwater detection, communication, rescue, and ocean development, among other fields [2]. In carrying out underwater warning defense and military attack activities, sonar needs to distinguish between true and false targets based on the received noise signals. Moreover, when detecting multiple targets simultaneously, it needs to recognize the types of each target and decide what actions to take against them, such as attack or avoidance, based on the results of the above two judgments [3].

Underwater target recognition technology can be mainly divided into two categories: active recognition and passive recognition [4]. Active sonar recognition uses active sonar to transmit sound signals and makes judgments about the target's properties based on the received echo signal characteristics. The advantage of this method is that the received echo signal carries a lot of information that is beneficial for classification and recognition, and reflects the essential features of the target. However, it has the disadvantage of being easily exposed and not conducive to self-protection. Passive sonar recognition technology uses passive sonar to receive target radiation noise for classification and recognition. It has the characteristics of high security and strong concealment, and is suitable for the classification and recognition of remote targets. It can provide strong support for detecting and effectively attacking enemies. This article mainly studies passive sonar target recognition, that is, recognizing targets by studying the signals radiated by ships.

The core of underwater target recognition lies in the processing of acoustic signals, which are inherently complex due to the source and propagation environment of the sound. The signals received by passive sonar are highly varied because of the various noise sources, considerable differences in the radiation noise, and the complexity, diversity, and strong time-varying characteristics of the ocean environment [5]. Therefore, how to extract target features with discriminative power is the key issue for passive underwater target recognition, and it is also the primary issue for achieving automatic target recognition [6].

Common methods for underwater target recognition can be roughly classified into three categories based on the type of classifier: traditional methods based on signal analysis, traditional machine learning methods, and deep learning methods. Traditional methods based on signal analysis usually rely on sonar operators to discriminate the types of targets. This method is constrained by manual experience and is relatively unstable, leading to low recognition accuracy. Traditional machine learning methods usually involve manually extracting features and then processing the extracted features before inputting them into a machine learning model. The model then automatically learns the correlation between features and target categories and makes decisions accordingly. This method can overcome the limitations of manual experience and achieve higher recognition accuracy than traditional methods based on signal analysis, while also achieving a certain degree of automation. However, the accuracy of this method mainly depends on complex feature engineering, making it increasingly difficult to meet the requirements of high-precision and high-intelligence underwater target recognition.

Deep learning is a special type of machine learning method that enables computers to automatically learn pattern features and incorporate them into the model building process, thereby reducing the incompleteness caused by human-designed features. Specifically, it involves constructing machine learning models with multiple hidden layers and large amounts of training data to automatically learn more useful features, ultimately improving classification or prediction accuracy. An increasing number of scholars are introducing deep learning into the field of underwater target recognition. Deep learning-based methods for underwater target recognition can satisfy the requirements of high accuracy and high intelligence, and

are becoming mainstream in the field of underwater acoustic recognition. Han X C et al. [7] proposed a method for underwater target recognition based on a one-dimensional convolutional neural network (1DCNN) and long short-term memory networks (LSTM). This method fully utilized the temporal characteristics of ship noise signals and used Mel-scale Frequency Cepstral Coefficients (MFCC) features as input to further extract deep temporal features for underwater target recognition. The recognition results showed that considering the temporal nature of underwater signals could effectively improve recognition accuracy. Zhang Q et al. [8] proposed a 2DCNN-based method for underwater target signal recognition. This method fully utilized the frequency domain information of ship noise signals. First, it extracted the Short-Time Fourier Transform (STFT) amplitude spectrum, STFT phase spectrum, and bicepstrum of the underwater signal. Second, it designed an ensemble neural network consisting of three 2DCNNs, each trained using a different spectrum. Finally, the Shuffled Frog Jumping Algorithm (SFLA) was used to combine the recognition results of the three networks. Experimental results demonstrated that considering the frequency domain information of underwater signals could effectively improve recognition accuracy. Although the above methods have achieved excellent results, most of them are based on extracting time-frequency domain features from sequence data and have not explored the image features of frequency domain maps. The time-frequency domain features of sequential underwater signals and the image features of two-dimensional frequency spectra can describe the characteristics of underwater targets from the perspectives of both signal and image, which can improve the accuracy of underwater target recognition. In addition, these methods usually use simple addition or concatenation strategies for feature fusion and discrimination, and do not fully utilize the learned deep features.

To address the aforementioned issues, we propose a method based on an adaptive multi-feature fusion network for underwater target recognition. The method consists of a data preprocessing module, a multi-dimensional feature extraction module, and an adaptive multi-feature fusion module. The data preprocessing module extracts the MFCC and two-dimensional time-frequency spectrogram of the underwater acoustic signal. The multi-dimensional feature extraction module utilizes LSTM and 1DCNN to extract the time-frequency features of the underwater acoustic signal from both the time and frequency domains. Additionally, 2DCNN are used to extract the image features of the frequency domain of the underwater acoustic signal, resulting in a more comprehensive feature set. The adaptive multi-feature fusion module employs channel attention mechanisms to adaptively weight and fuse the extracted multi-dimensional features, thereby improving the discriminability of the features and fully utilizing the learned features to achieve high-precision target recognition.

The contributions of this paper are as follows.

(1) Introducing a multi-dimensional feature extraction method that simultaneously exploits the complementary information of time-frequency features in the signal domain and image features in the two-dimensional frequency spectrum.
(2) Proposing an adaptive multi-feature fusion module that leverages a channel attention mechanism to intelligently weight and fuse diverse feature information, thereby maximizing the potential of learned features and significantly enhancing recognition accuracy.
(3) Conducting experiments on the standard ShipEar dataset, and our proposed method based on an adaptive multi-feature fusion network achieved a 2.78% improvement in recognition accuracy compared to the suboptimal results on the same dataset.

2 Related work

Currently, commonly used methods for underwater target recognition can be roughly categorized into three types based on the classifier: traditional methods based on signal analysis, traditional machine learning-based methods, and deep learning-based methods.

Traditional methods based on signal analysis for underwater target recognition heavily rely on the subjective judgment of sonar operators, who make the final classification decision based on LOFAR and DEMON spectrograms and auditory analysis [9]. However, due to the constraint of human expertise, these methods exhibit poor stability and limited accuracy in recognition [10]. With the advance of machine learning techniques, methods based on traditional machine learning have gained popularity among researchers, as they demonstrate higher recognition accuracy than traditional signal analysis methods.

The methods based on traditional machine learning typically involve several components, including data acquisition, data preprocessing, feature extraction, feature selection, and classification decision. This approach can overcome the limitations of relying on manual experience and achieve a certain degree of automation. It has been found to have significantly higher recognition accuracy than traditional methods based on signal analysis. Farrokhrooz et al. [11] roposed a probabilistic neural network (PNN)-based method for ocean vessel classification. In this approach, the acoustic radiated noise of a vessel is modeled using an autoregressive (AR) model of appropriate order, and the model coefficients are used as classification features. Experimental evaluations demonstrated that using the AR model coefficients as discriminant features input to PNN can achieve high recognition rates. Meng Q et al. [3] constructed a nine-dimensional feature vector comprising statistical features such as zero-crossing wavelength, peek amplitude, zero-crossing wavelength difference, and beam area. These feature vectors were input to a support vector machine (SVM) to identify underwater acoustic targets, achieving an accuracy of 89.5% on the test set. H. Yang et al. [12] roposed a weighted sample and feature selection AdaBoost method (WSFSelect-SVME) for underwater acoustic target recognition. The AdaBoost method constructed an ensemble classifier that iteratively focused each new SVM classifier on the most difficult samples. By using a weighted immune clone sample selection algorithm and mutual information sequence forward feature selection algorithm, the number of training set samples and features was reduced while maintaining the performance of each new individual SVM classifier. The experimental results showed that the algorithm improved accuracy while reducing the spatial complexity of the ensemble. However, the recognition accuracy of these research methods mainly depended on complex feature engineering, and the classifiers adopted were mostly shallow classifiers such as SVMs and shallow neural classifiers, which had weaker fitting and generalization capabilities when dealing with complex and large samples. Therefore, it was becoming increasingly difficult to meet the requirements for high-precision underwater target intelligent identification.

In recent years, with the further development of computer hardware technology and signal processing technology, artificial intelligence technologies represented by deep learning have made great achievements in problems such as target recognition [13]. Deep learning, also known as deep machine learning, is a type of machine learning algorithm that effectively trains deep neural networks (DNNs) and can be used for high-level abstract modeling of data. Based on various models and algorithms, deep networks can learn suitable and effective features from large amounts of complex data. These features often achieve excellent results in solving practical problems, making deep learning widely favored by academia and industry. Therefore, more and more scholars are beginning to introduce deep learning into the field of underwater acoustic target recognition and have made breakthrough progress. Zhang et al.

[14] roposed an underwater target recognition method based on LSTM, which utilized LSTM to extract deep features from the MFCC of underwater targets and recognized. The recognition results showed that the method effectively differentiated different underwater targets. Abdoli et al. [15] roposed an end-to-end method for environmental sound classification based on 1DCNN, which directly recognized the original audio signal as the input target. Experimental results showed that the average accuracy of the method reached 89%. Mishachandar et al. [16] proposed an underwater target recognition method based on 2DCNN, which obtained the spectrogram of underwater target signals through short-time Fourier transform and then used it as the input for 2DCNN recognition. The results showed that combining the spectrogram with 2DCNN effectively reduced the influence of original signal noise, thereby improving the recognition accuracy.

3 Methodology

3.1 Overall analysis workflow

The overall framework of the underwater target recognition method based on the adaptive multi-feature fusion network proposed in this article is shown in Fig. 1. The method mainly consists of three modules: data preprocessing, extraction of multi-dimensional deep features, and adaptive multi-feature fusion. In the data preprocessing module, the original audio files are segmented into equal-length small audio segments, and the MFCC features and two-dimensional time-frequency spectrograms are extracted based on each small audio segment. In the multi-dimensional deep feature extraction module, 1DCNN and LSTM are used to extract deep time-frequency features of underwater signals, while 2DCNN is used to extract image features of underwater signal frequency domain. In the adaptive feature fusion module, the attention mechanism is applied to adaptively weight and fuse the features extracted by the three networks to fully utilize the features learned by each network.

3.2 Data pre-processing

3.2.1 Mel-scale frequency cepstral coefficients

Mel-scale Frequency Cepstral Coefficients (MFCC) is one of the most common speech features, which are obtained by extracting cepstral coefficients in the Mel-scale frequency domain. By combining the perceptual characteristics of the human ear with the mechanisms of speech production, MFCC are able to capture important information for speech recognition [17]. Owing to their resilience in acoustically challenging environments, MFCC have exhibited remarkable superiority in the domain of underwater target recognition. The process of extracting MFCC features is shown in Fig. 2.

In the pre-emphasis step, the high-frequency part of the audio signal is enhanced by a high-pass filter. To ensure the smoothness of the input signal, the speech signal needs to be segmented into small frames, which are usually 20-30ms long, and typically 25ms long. After the audio signal is framed, each frame needs to be windowed to increase the continuity of the ends of the frames and reduce spectral leakage. The most commonly used window function is the Hamming window. The transformed signal after windowing is usually difficult to observe its characteristics in the time domain. Therefore, it is generally converted into the energy distribution in the frequency domain by performing Discrete Fourier Transform

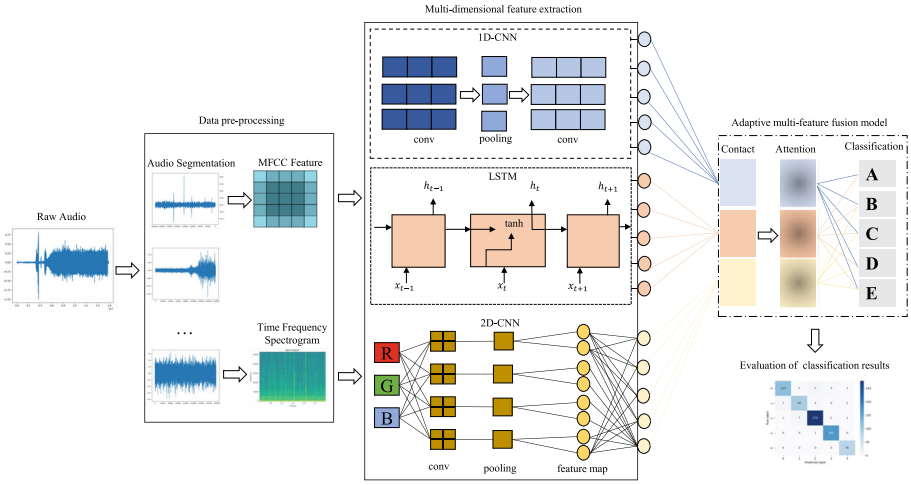

Fig. 1 The overall workflow of this study

(DFT) to further observe the characteristics of the signal. Mel filters consist of a bank of triangular band-pass filters that convert linear frequency into a nonlinear scale called Mel frequency. By passing the energy spectrum through the Mel filters, the spectrum can be smoothed and harmonic components can be removed. Additionally, data reduction can be achieved, which subsequently reduces the computational burden in the following processes. Since the human ear responds to sound on a logarithmic scale, the output of the Mel filters is typically logarithmically compressed to approximate the human auditory system. After the logarithmic operation, the signal values of different orders are correlated to a certain degree. Discrete cosine transform (DCT) can remove this correlation and reduce the dimensionality of the signal. Finally, the signal is mapped into a low-dimensional space to obtain MFCC features that can be used for recognition and classification.

3.2.2 Two-dimensional time-frequency spectrum

Two-dimensional time-frequency spectrogram is a spectrogram generated by performing a Fourier transform on the original audio signal, which contains abundant time-frequency domain information. Converting underwater acoustic signals into two-dimensional time-frequency spectrograms can effectively reduce the impact of noise, thus improving classification and recognition performance. The process of generating a two-dimensional time-frequency spectrogram is shown in Fig. 3.

1) Framing. Due to the non-stationarity and time-varying nature of ship signals, they are typically segmented into smaller frames for analysis, a process known as framing. To

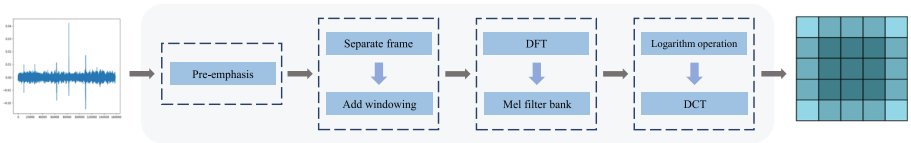

Fig. 2 The extraction process of MFCC features

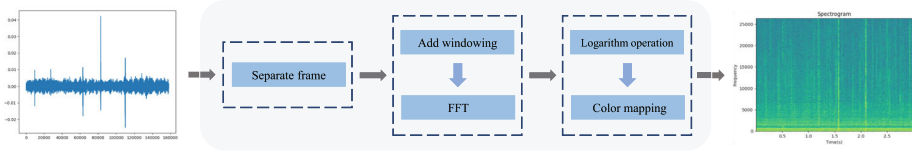

Fig. 3 The process of generating two-dimensional time-frequency spectrograms

ensure smooth transitions between signal frames, there is usually an overlap between adjacent frames.

2) Adding windows. Fourier transform of signals can suffer from spectral leakage due to non-periodic truncation. Adding windows can better meet the periodicity requirement for fourier transform processing and reduce spectral leakage. The most commonly used window function is the Hamming window, whose formula is:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\frac{2\pi n}{M-1}, 1 \ll n \ll M \\ 0, otherwise \end{cases} \quad (1)$$

Where M is the window length.

3) Fast Fourier Transform (FFT). Performing Fast Fourier Transform on the windowed signal frames can transform the signal from time domain to frequency domain. This allows for more efficient analysis and processing of the signal in the frequency domain.
4) Logarithmic calculation. Taking the logarithm of the spectral energy computed from the Fourier transform results in a more compact representation of the energy.
5) Color mapping. Color mapping is applied to discretized energy spectrum values in order to map them to corresponding RGB color values.
Stacking the discretized color values in chronological order produces the final two-dimensional time-frequency spectrogram, which can be used for applications such as target recognition.

3.3 Multidimensional feature extraction

The multi-dimensional feature extraction module includes deep temporal feature extraction, deep spatial feature extraction, and deep frequency domain image feature extraction. Audio MFCC feature data has both temporal and spatial continuity. Therefore, this paper uses LSTM network and 1DCNN network to further extract deep temporal information and deep spatial information from MFCC feature data, respectively. Based on the two-dimensional time-frequency spectrogram generated from the audio, which contains rich frequency domain information, this paper uses 2DCNN to further extract deep frequency domain image information.

3.3.1 Deep time-series feature extraction

Long Short-Term Memory Network (LSTM) [18] is an improved network based on Recurrent Neural Network (RNN) [19]. The main purpose of LSTM is to solve the problems of gradient vanishing and exploding in the training process of long sequences. Compared with ordinary RNN, LSTM can perform better in longer sequences [20]. Therefore, LSTM has gradually

become the mainstream model in the field of speech recognition. LSTM mainly designs a special structural unit on the original RNN structure, which mainly includes forgetting, selecting memory, and outputting three stages. Through these three stages, the information of each time node can be selectively added or removed. As the MFCC features of audio exhibit temporal continuity, we have utilized LSTM network to extract deep temporal features for identification based on the MFCC features of underwater acoustic signals in this paper. The constructed LSTM network structure is shown in Fig. 4.

The constructed LSTM consists of 4 layers, including an input layer, an LSTM layer, a dropout layer, and a fully connected layer. The input layer is a one-dimensional sequence vector of length 40. The number of hidden units in the LSTM layer is set to 128. To prevent overfitting of the LSTM on the training set, a dropout layer [21] is introduced to reduce the computational complexity during training, with a dropout rate of 0.2. The output from the dropout layer is a feature vector comprising 128 nodes. The output fully connected layer contains 5 nodes, representing the probabilities of the predicted samples being different underwater sound targets. Finally, the output of the dropout layer is extracted as the deep temporal feature set of the underwater sound signal.

3.3.2 Deep space feature extraction

Convolutional Neural Network (CNN) is a type of feedforward neural network that includes convolutional calculations and typically consists of three parts: a convolutional layer, a pooling layer, and a fully connected layer. The convolutional layer is used to extract features, the pooling layer is used to compress feature information, and the fully connected layer is used to output the final prediction results [22, 23]. Based on the dimension of kernel sliding, CNN can be divided into one-dimensional convolutional neural networks (1DCNN) and two-dimensional convolutional neural networks (2DCNN). Among them, the movement direction of the convolution kernel and pooling kernel of 1DCNN is one-dimensional, which is suitable for processing sequence data [24]. Due to the fact that the MFCC feature data exhibits both spatial continuity and temporal continuity, we have employed 1DCNN to process the MFCC features of underwater acoustic signals in this paper. By utilizing the spatial properties of 1DCNN, we are able to further extract the depth spatial features of underwater acoustic signals for identification. The 1DCNN network structure that we have designed for this purpose is shown in Fig. 5.

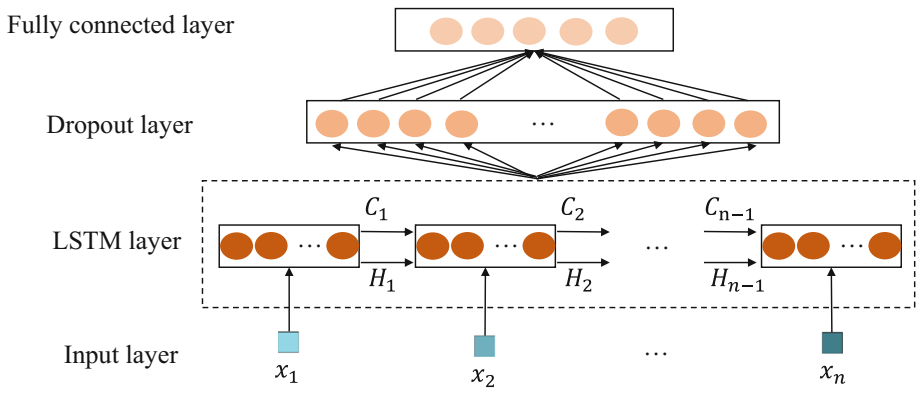

Fig. 4 The structure of the LSTM model

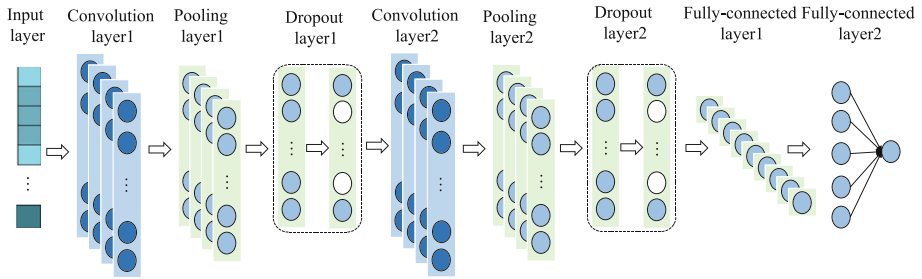

Fig. 5 The structure of the 1DCNN model

The designed 1D-CNN network consists of 9 layers, including one input layer, two convolutional layers, two pooling layers, two dropout layers, and two fully connected layers. The input layer accepts MFCC features of size 40×1, thus the input size is set to 40×1. The two convolutional layers are used to extract spatial features of the underwater sound signal, while one max pooling layer and one global max pooling layer are used for feature compression. The two dropout layers are added to prevent overfitting of the model. The output of the two fully connected layers predicts the probability of the input sample belonging to different underwater targets. Finally, the output of Fully-connected layer1 is extracted as the deep spatial feature set of the underwater sound signal. The detailed network parameters are shown in Table 1.

3.3.3 Deep frequency domain image feature extraction

2DCNN consists of three parts: convolutional layers, pooling layers, and fully connected layers. It is primarily used for image-related tasks. The original image is dynamically extracted with rich image features through convolutional layers. The features are compressed through pooling layers to extract the main features, and the final prediction results are obtained through fully connected layers [25]. This paper employs 2DCNN to extract deep frequency domain image features from the two-dimensional spectrogram obtained from the original audio signal. The 2DCNN structure is shown in Fig. 6.

The designed 2DCNN network consists of 10 layers, including one input layer, three convolutional layers, three pooling layers, two dropout layers, and one fully connected layer.

Table 1 Parameters of 1DCNN Network Layers

Layer	Activation function	Kernel size	Padding	Output Shape
Input	-	-	-	(None, 40, 1)
Conv1	Relu	3×3	Valid	(None, 38, 64)
Max pool1	-	2×2	-	(None, 19, 64)
Droupt1	-	-	-	(None, 19, 64)
Conv2	Relu	3×3	Valid	(None, 17, 128)
Max pool1	-	2×2	-	(None, 128)
Droupt2	-	-	-	(None, 128)
Fully-connected layer1	Relu	-	-	(None, 64)
Fully-connected layer2	Softmax	-	-	(None, 5)

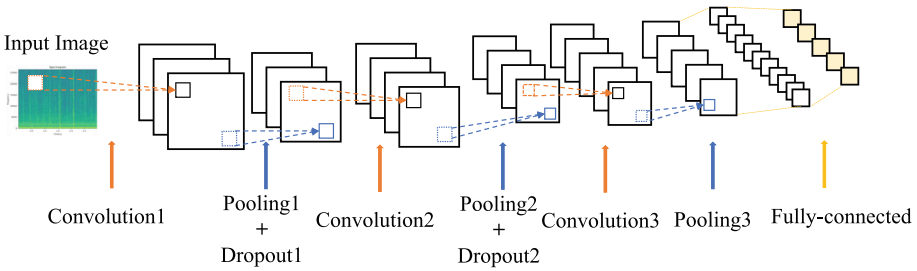

Fig. 6 The structure of the 2DCNN model

The input layer receives a time-frequency spectrogram with a size of 224 × 224 and three RGB channels. Therefore, the input size is set to 224 × 224 × 3. The three convolutional layers extract image features, and two max pooling layers and one global max pooling layer compress the feature information. Two dropout layers are used to prevent overfitting, and a fully connected layer is connected to output the probability of the predicted sample belonging to different underwater targets. Finally, the output of the max pool3 layer is extracted as the deep frequency-domain feature set of underwater sound signals. The detailed network parameters are shown in Table 2.

3.4 Adaptive multi-feature fusion

The adaptive multi-feature fusion module uses an attention mechanism to adaptively fuse the three features extracted by the multi-dimensional feature extraction module, providing more discriminative features for subsequent target recognition. The module adopts a channel attention mechanism, which adaptively weights the dependencies of each channel to improve the network's representation ability by assigning more weight to effective features, thereby solving the problem of inaccurate weight allocation for the extracted feature maps. With the channel attention mechanism, the network can selectively enhance useful feature information and suppress irrelevant features by learning from global information. The network structure of the adaptive multi-feature fusion module is shown in Fig. 7.

Table 2 Parameters of 2DCNN Network Layers

Layer	Activation function	Kernel size	Strides	Padding	Output Shape
Input	-	-	-	-	(None, 224, 224, 3)
Conv1	Relu	5×5	1	Same	(None, 224, 224, 8)
Max pool1	-	2×2	2	-	(None, 112, 112, 8)
Droupt1	-	-	-	-	(None, 112, 112, 8)
Conv2	Relu	5×5	1	Same	(None, 112, 112, 16)
Max pool2	-	2×2	2	-	(None, 56, 56, 16)
Droupt2	-	-	-	-	(None, 56, 56, 16)
Conv3	Relu	5×5	1	Same	(None, 56, 56, 32)
Max pool3	-	-	-	-	(None, 32)
Fully-connected layer	Softmax	-	-	-	(None, 5)

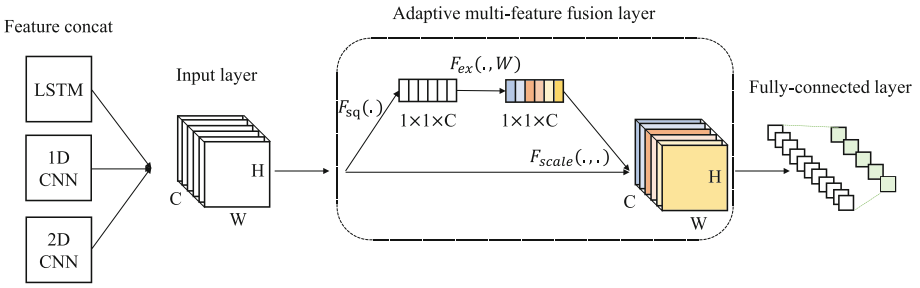

Fig. 7 The structure of adaptive multi-feature fusion module

The adaptive multi-feature fusion module consists of a total of four layers, including one input layer, one adaptive multi-feature fusion layer, and two fully connected layers. The input layer receives the mixed features after concatenating the three network features. Since the dimension of the mixed feature is 224×1, the input layer size is set to 224×1. To facilitate the adaptive multi-feature fusion layer's calculation of attention weights, the input is reshaped to $1 \times 1 \times 224$, which corresponds to $W \times H \times C$, where C is the number of input channels, and $W \times H$ is the feature dimension of each channel. The implementation of the adaptive multi-feature fusion layer mainly consists of three modules, namely, Squeeze, Excitation, and Scale [26]. Squeeze utilizes global average pooling (GAP) operation to compress the global spatial information of each channel, i.e., compress the two-dimensional feature of each channel ($W \times H$) to $1 \times 1 \times C$. The formula for global average pooling operation is:

$$z_c = F_{sq}(u_c) = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} u_c(i, j) \tag{2}$$

where z_c represents the weight parameter after the compression operation, $F_{sq}(\cdot)$ denotes the feature compression operation, u_c represents the c^{th} two-dimensional matrix in U, which is the set of multiple local feature maps; H is the height of the feature matrix, W is the width of the feature matrix. Excitation generates a range of (0,1) weights for each feature channel by means of the parameter w, which is learned to explicitly model the correlation between feature channels. Specifically, two fully connected layers (FC-ReLU-FC-Sigmoid) are used to calculate the weight values, and the weight calculation formula is:

$$s = F_{ex}(z, w) = \sigma(g(z, w)) = \sigma(w_2 \delta(w_1 z)) \tag{3}$$

where $\delta(w_1 z)$ represents the first fully connected operation, w_1 is a $C/r \times C$ dimensional matrix and r is a scaling parameter used to reduce the number of channels and computational complexity, set to 4 in this paper. The dimension of z is $1 \times 1 \times C$, thus the result of $w_1 z$ is $1 \times 1 \times C/r$, which is then passed through a ReLU layer with unchanged dimension. The result of $\delta(w_1 z)$ is multiplied with w2 for the second fully connected operation, where w_2 is a $C \times C/r$ dimensional matrix, resulting in a dimension of $1 \times 1 \times C$. Then, the sigmoid function is applied to obtain the final weight s. Scale weights the normalized weights obtained earlier by multiplying them with the features of each channel. In the Scale module, the normalized weights obtained earlier are used to weight the features of each channel. Finally, the weighted information from the adaptive multi-feature fusion layer is fed into two fully connected layers with 64 and 5 nodes respectively for underwater target recognition.

4 Experimental results

4.1 Dataset

The dataset used in this paper is sourced from the publicly available ShipEar dataset [27], which consists of oceanic recordings collected between 2012 and 2013 along the Spanish coastline using the MarSensing Lda company's (Portugal, Faroe Islands) autonomous acoustic digitalHyd SR-1 recorder. The dataset includes a total of 90 audio recordings, ranging from 15 seconds to 10 minutes in duration, and covers 11 categories of vessels and environmental noise. According to the dataset's source paper, it can be further classified into five categories: A, B, C, D, and E, where A, B, C, and D represent four major categories of vessel types and E represents environmental noise. The original audio data consists of only 90 recordings, and the significant differences in the number of recordings between different categories may lead to underfitting of the model. To address this issue, the original audio data was segmented into 3-second clips, resulting in an expanded dataset. After segmentation, the total number of audio clips reached 3824, as shown in Table 3.

Each audio is preprocessed as follows: MFCC features are extracted and a two-dimensional spectrogram is generated.

1). Extracting MFCC features: The dimension of the extracted MFCC features is (40, 309). The feature column vectors are compressed by taking their mean, resulting in an MFCC feature dimension of (40, 1).
2). Generating two-dimensional spectrogram: A two-dimensional spectrogram is obtained by performing fourier transform on the original audio. The size of the two-dimensional spectrogram is 569×435 with RGB three channels. For the network, a large input image size will increase the computational cost, while a small cropping size will cause significant information loss. Therefore, in this paper, the generated two-dimensional spectrogram is resized to $224 \times 224 \times 3$.

The 3824 samples in this paper are divided into a training set, validation set, and test set in a ratio of 6:2:2. There are 2292 samples in the training set, 765 samples in the validation set, and 767 samples in the test set. After 35 epochs, the loss and accuracy curves were obtained, as shown in Figs. 8 and 9. The blue curve represents the variations in the training set, while the orange dotted line depicts the changes in the test set.

4.2 Ablation experiments

In this paper, the code is written using the Keras platform and experiments are conducted on a server with an Intel(R) Xeon(R) Silver 4216 CPU@2.10GHz and an NVIDIA GeForce

Table 3 Classification results of the dataset

Target Category	Ship Category	Original number	Number after cutting
A	fishing boats, trawlers, mussel boats, tugboats and dredgers	17	635
B	motorboats, pilot boats and sailboats	19	532
C	passenger ferries	30	1441
D	ocean liners and ro-ro vessels	12	826
E	background noise recordings.	12	390

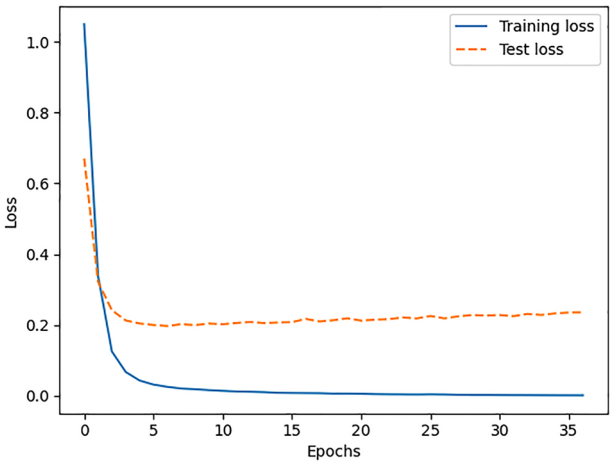

Fig. 8 Variation of loss

GTX 1080 Ti GPU. For the LSTM model, the learning rate is set to 0.001, the optimizer is set to Adam, the loss function is set to Categorical_crossentropy, the evaluation metric is set to Accuracy, the batch size is set to 64, and the number of training epochs is set to 200. For the 1DCNN, 2DCNN models, and the adaptive fusion model, the training parameters are the same as the LSTM model. To prevent overfitting, the early stopping strategy [28] is introduced during network training to monitor val_loss, with patience set to 20. To comprehensively evaluate the recognition performance of the network, recall, precision, and f1-score metrics are further used to evaluate the performance of the network on the test set.

To verify the effectiveness of the proposed adaptive multi-feature fusion network, ablation experiments are conducted on both the multi-dimensional feature extraction module and the adaptive multi-feature fusion module.

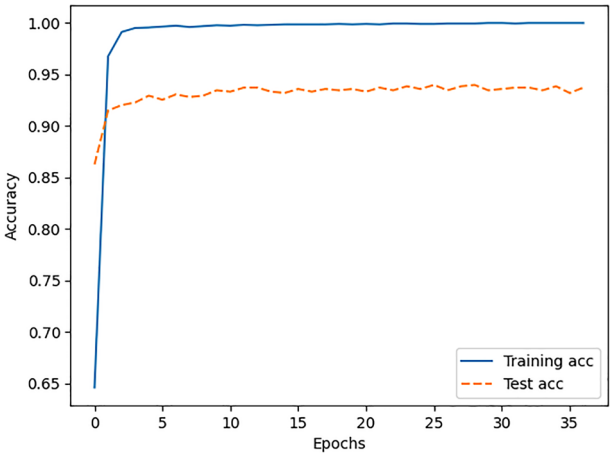

Fig. 9 Variation of accuracy

4.2.1 Validation of multidimensional feature extraction module

Firstly, the proposed adaptive multi-feature fusion network was compared with single sub-networks including 1DCNN, 2DCNN, and LSTM. Then, using the 2DCNN model as a baseline, the features extracted by 1DCNN and LSTM were concatenated and used for underwater target recognition. The experimental results are shown in Table 4. Finally, the features extracted by different networks were input into the adaptive multi-feature fusion module for underwater target recognition. The experimental results are shown in Fig. 10.

From Table 4, it can be seen that the single LSTM outperforms the 1DCNN and 2DCNN networks in terms of classification accuracy, recall, precision, and f1-score on the underwater sound dataset, with values of 0.9022, 0.9017, 0.8926, and 0.8967, respectively. As the underwater sound data is a time-series signal, the LSTM network pays more attention to temporal features, and thus performs better among the three single sub-networks. When the features extracted by different networks were grouped and fused, the recognition accuracy was higher than that of any single network. Among them, the recognition accuracy of fusing the features extracted by all three networks simultaneously was the highest, with values of 0.9348, 0.9296, 0.9336, and 0.9315 for accuracy, recall, precision, and f1-score, respectively. Compared to the performance of the single LSTM, it improved by 3.26%, 2.79%, 4.1%, and 3.48%, respectively. Compared to the second-best fusion feature set (2DCNN+LSTM), it improved by 1.31%, 0.82%, 2.03%, and 1.47%, respectively. It can be inferred that considering only the time-frequency domain features or the frequency domain image features of underwater sound signals in a single sub-network can lead to incomplete feature information, thus leaving room for improvement in recognition accuracy. The proposed model in this paper not only considers the time-frequency domain feature information of underwater sound signals but also explores the image feature information of frequency domain images, thus performing better than a single network and significantly improving recognition accuracy.

4.2.2 Validation of the effectiveness of the adaptive multi-feature fusion module

Common decision weighting strategies include direct averaging, simple weighted averaging, and inverse variance weighting. These methods assign different weights to each sub-model through different calculation methods, and then multiply each model's prediction result by its corresponding weight and sum them up to obtain the final prediction result. Direct averaging sets the reciprocal of the current number of sub-models as the weight for each sub-model, and each model has the same weight. Simple weighted averaging ranks the models based on their prediction errors ($i = 1, 2, \ldots, n$), with models with larger errors ranked first, and each model's weight is calculated as $i/(1 + 2 + \ldots n)$. The larger the model's prediction error,

Table 4 Experimental results of model structure ablation

Network Type	Accuracy	Recall	Precision	F1-score
2DCNN	0.8396	0.8449	0.8344	0.8374
1DCNN	0.8604	0.8378	0.8675	0.8473
LSTM	0.9022	0.9017	0.8926	0.8967
2DCNN+1DCNN	0.9087	0.9097	0.9067	0.9076
2DCNN+LSTM	0.9217	0.9214	0.9133	0.9168
2DCNN+1DCNN+LSTM	**0.9348**	**0.9296**	**0.9336**	**0.9315**

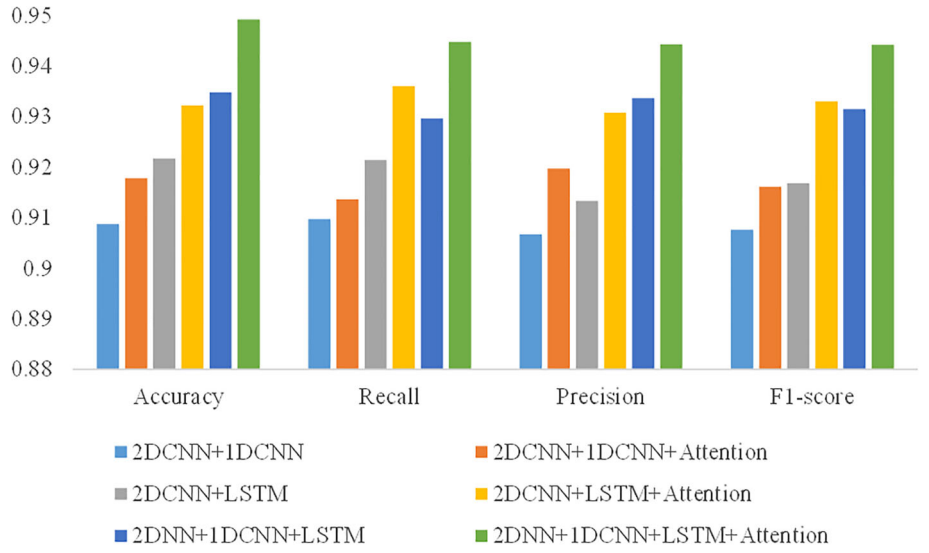

Fig. 10 Results of weighted strategy ablation experiments

the smaller the weight it will be assigned. Inverse variance weighting calculates a prediction error square for each sub-model, and the weighted weight for each model is the proportion of its own error square inverse in the sum of all models' error squares inverse. Models with smaller error squares will be assigned higher weights.

This paper proposes an adaptive weighting strategy to fuse the features extracted by the three sub-networks and discriminate the target categories for prediction. To verify the effectiveness of the adaptive weighting strategy, we compared it with three other weighting strategies, namely direct average weighting, simple average weighting, and reciprocal of variance weighting. Firstly, 1DCNN, 2DCNN, and LSTM were used to identify underwater acoustic signals. Then, the identification results of the three networks were separately weighted and fused using the direct average weighting, simple average weighting, and inverse of variance weighting methods. Finally, the features of the three networks' intermediate layers were extracted and adaptively weighted for identification through a channel attention mechanism. The experimental results are shown in Table 5.

As shown in Table 5, the accuracy, recall, precision, and f1-score of the adaptive weighted fusion recognition using features extracted from the three sub-networks were 0.9492, 0.9448, 0.9443, and 0.9442, respectively. Compared to the direct average weighting method, they were improved by 0.62%, 0.71%, 0.36%, and 0.56%, respectively. Compared to the simple average weighting method, they were improved by 1.83%, 1.99%, 1.86%, and 1.98%, respectively.

Table 5 Comparison experimental results of different integration strategies

Integration Strategy	Accuracy	Recall	Precision	F1-score
direct average weighting method	0.9430	0.9377	0.9407	0.9386
simple weighted average method	0.9309	0.9249	0.9257	0.9244
inverse of variance method	0.9295	0.9232	0.9221	0.9219
adaptive weighting method	**0.9492**	**0.9448**	**0.9443**	**0.9442**

Compared to the inverse variance weighting method, they were improved by 1.97%, 2.16%, 2.22%, and 2.23%, respectively. This indicates that traditional weighting strategies that rely on the decision-making of sub-networks' recognition results cannot adaptively utilize the learned information to improve recognition accuracy. Instead, by calculating the weights of features extracted from different sub-networks using channel attention and integrating them, important feature information can be better highlighted, thereby effectively improving the accuracy of underwater target detection.

To further verify the effectiveness of the proposed adaptive weighting strategy, we conducted the following ablation experiments. First, we extracted the features from the intermediate layers of the three networks and then combined and concatenated them (2DCNN+1DCNN, 2DCNN+LSTM, 2DCNN+1DCNN+LSTM) for underwater target recognition. Then, we added attention to each concatenated feature set (2DCNN+1DCNN+Attention, 2DCNN+LSTM+Attention, 2DCNN+1DCNN+LSTM+Attention) for adaptive weighting and underwater target recognition. The experimental results are shown in Fig. 10.

Figure 10 shows the results of the weighted strategy ablation experiment. It can be seen from the figure that the recognition accuracy of underwater targets is generally improved by weighting and fusing the features extracted from different networks using channel attention. Among them, the multi-feature fusion model proposed in this paper achieved the highest classification accuracy, recall, precision, and f1-score on the underwater dataset, which were 0.9492, 0.9448, 0.9443, and 0.9442, respectively. Compared with the performance before adding attention, they were improved by 1.44%, 1.52%, 1.07%, and 1.27%, respectively. This indicates that simple fusion of features extracted from three different networks can indeed take into account the complementary information of feature extraction from both the signal domain and the image domain, thereby improving the recognition accuracy. However, this simple feature fusion method did not consider that different features from different sources have different effects on the final recognition. The multi-feature fusion model proposed in this paper weights and fuses the features extracted by 2DCNN, 1DCNN, and LSTM using channel attention, which can allocate more weight to important features and better leverage their role. Therefore, it can significantly improve the recognition accuracy.

4.3 Comparison experiments

4.3.1 Comparison algorithm

1. Mishachandar et al. [16] proposed an underwater target recognition method based on 2DCNN. This method uses short-time Fourier transform to obtain the spectrogram of underwater target signals, which is then used as the input for 2DCNN recognition.
2. Zhang Q et al. [8] proposed an integrated neural network-based underwater target recognition method using 2DCNN. This method fully utilized the frequency-domain information of ship noise signals, firstly extracting the short-time Fourier transform (STFT) amplitude spectrum, STFT phase spectrum, and bicepstrum of underwater acoustic signals. Secondly, an ensemble neural network consisting of three 2D-CNNs was designed to train with different spectra. Finally, the shuffled frog jumping algorithm (SFLA) was employed to weight the recognition results of the three networks for decision-making.
3. Han X C et al. [7] proposed a water acoustic target recognition method based on 1DCNN and LSTM, which fully utilizes the temporal characteristics of ship noise signals. The method uses the MFCC feature of the audio as input and further extracts deep temporal features for water acoustic target recognition.

4.3.2 Comparison of the experimental results

To further demonstrate the superiority of our proposed method in underwater target recognition tasks, comparative experiments were conducted with the methods proposed by Mishachandar B, Zhang Q, Han X C, and others. To ensure the fairness and validity of the comparative experiments, all the data processing steps involved in the comparisons adhered to a uniform standard, wherein the original data was consistently segmented into multiple 3-second audio clips. The comparative experimental results are shown in Table 6.

From Table 6, it can be seen that the multi-feature fusion network proposed in this paper has higher classification accuracy, recall, precision, and f1-score on the underwater dataset than other existing methods. Compared with the methods proposed by Mishachandar B, Zhang Q, and Han X C, our proposed method has improved the accuracy by 8.78%, 4.57% 2.78%, the recall by 8.46%, 4.55%, 2.53%, the precision by 8.53%, 4.22%, 2.12%, and the f1-score by 8.74%, 4.46%, 2.52%, respectively. Mishachandar B and Zhang Q's methods only considered the frequency-domain 2D spectrum of underwater signals, and Han X C's method only considered the temporal characteristics of underwater signals. Our proposed multi-feature fusion model combines the features extracted by 2DCNN, 1DCNN, and LSTM through adaptive weighting to jointly consider the temporal-frequency features of underwater signals and the image features of the 2D spectrum. However, it should be acknowledged that the parameter count of the proposed method in this study is indeed higher compared to other existing approaches, potentially leading to increased model complexity, prolonged training durations, and elevated computational resource consumption. Nevertheless, this increment in parameter quantity has been intentionally strategized to facilitate more nuanced feature extraction and more efficacious feature fusion. Although our model's parameters exceeds that of Mishachandar B and Han X C methods, it remains less than that required by Zhang Q method. This highlights our model's relative efficiency in utilizing computational resources while achieving significant improvements in accuracy. Therefore, compared with existing methods that can only consider the temporal-frequency features of underwater signals, our proposed adaptive multi-feature fusion network can significantly improve recognition accuracy.

To provide a more intuitive comparison of the performance of each method on the test set, the confusion matrices of the recognition results on the test set were visualized, as shown in Fig. 11. In the figure, the horizontal and vertical coordinates of 0 5 represent the predicted target categories A-E, respectively. It can be seen from the figure that Mishachandar B's method is prone to mistakenly identifying C-class targets as A-class targets, while Zhang Q and Han X C's methods are prone to mistakenly identifying B-class targets as C-class targets. In contrast, our adaptive multi-feature fusion network can effectively alleviate these issues.

The specific recognition accuracies of the four methods for the five target categories were further calculated using the confusion matrix, as shown in Fig. 12. Mishachandar B, Zhang

Table 6 Comparison of experimental results with other methods

Network Type	Accuracy	Recall	Precision	F1-score	Parameters
Mishachandar B et al [16]	0.8614	0.8602	0.8590	0.8568	**42610**
Zhang Q et al [8]	0.9035	0.8993	0.9021	0.8996	174170
Han X C et al [7]	0.9214	0.9153	0.9231	0.9190	45733
Ours	**0.9492**	**0.9448**	**0.9443**	**0.9442**	132116

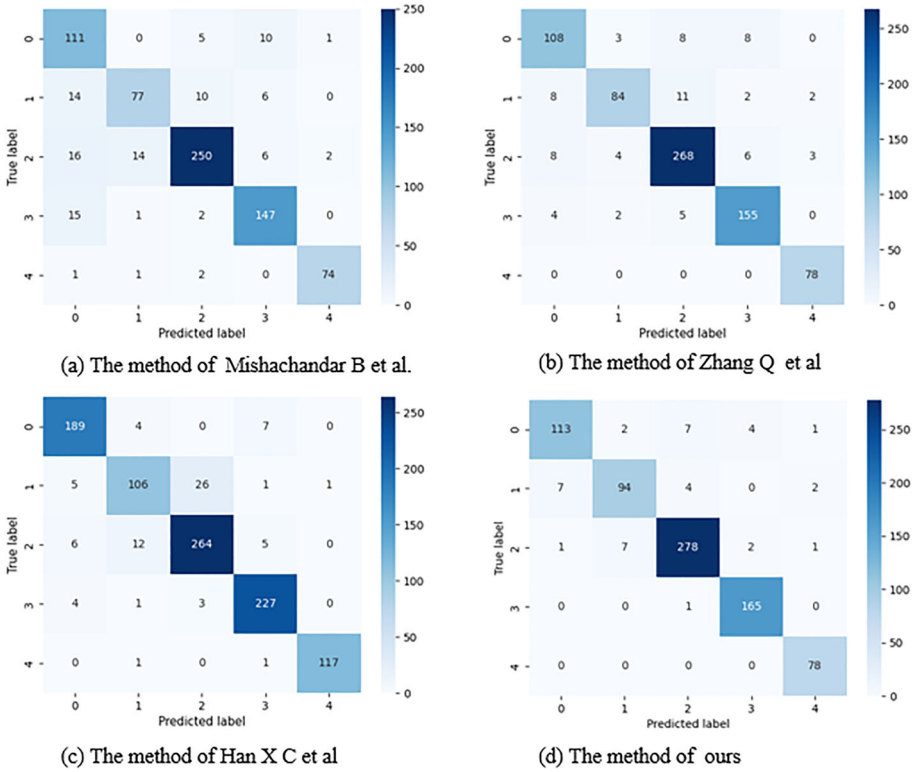

Fig. 11 Comparison of confusion matrix with other methods

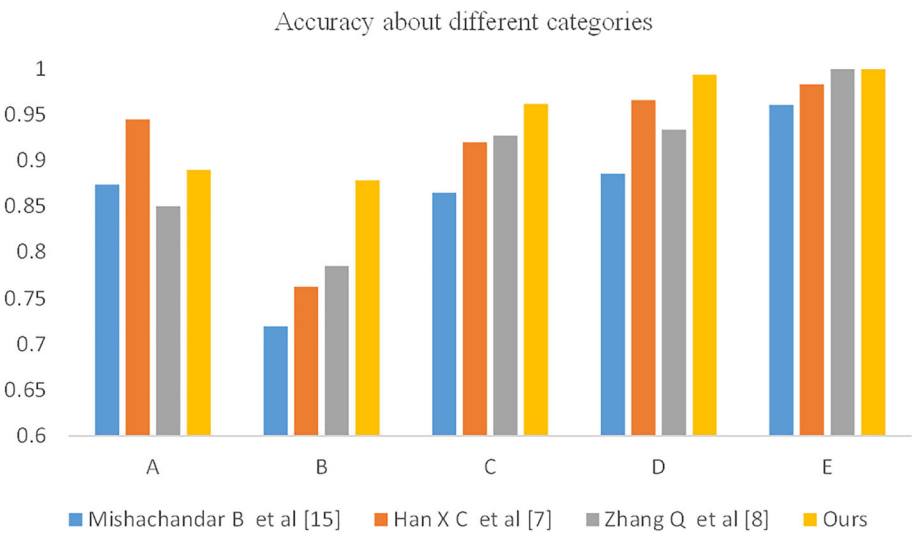

Fig. 12 Comparison of confusion matrix with other methods

Q, and Han X C's methods all achieved the highest recognition rate for the E-class target and the lowest for the B-class target. In comparison, our proposed adaptive multi-feature fusion network achieved recognition rates of 88.97%, 87.85%, 96.19%, 99.39%, and 100% for categories A, B, C, D, and E, respectively. Compared with Mishachandar B's method, our method improved the recognition rates of categories A, B, C, D, and E by 1.57%, 15.89%, 9.66%, 10.84%, and 3.9%, respectively. Compared with Zhang Q's method, our method improved the recognition rates of categories A, B, C, and D by 3.67%, 9.35%, 3.46%, and 6.02%, respectively. Compared with Han X C's method, our method improved the recognition rates of categories B, C, D, and E by 11.59%, 4.2%, 2.79%, and 1.68%, respectively. The comparison results demonstrate that considering only the time-frequency domain features or the frequency-domain image features of signals alone can result in one-sided feature information extraction, and using simple addition or concatenation strategies for feature fusion and discrimination may lead to inadequate utilization of learned feature information, resulting in room for improvement in the accuracy of underwater target recognition. Simultaneously considering both the time-frequency domain features and the frequency-domain image features of signals can form a more comprehensive feature set. Further, using an attention mechanism to adaptively weight the feature set can enhance the discriminability of features, fully utilizing the learned features, and effectively improving the accuracy of underwater target recognition.

5 Conclusion

To fully leverage the time-frequency information of ship signals and improve the accuracy of underwater acoustic target recognition, this paper proposes an adaptive multi-feature fusion network-based method for underwater acoustic target recognition. This method takes a dual approach from both the signal and image domains and explores the complementary information between the time-frequency features in the signal domain and the image features corresponding to the two-dimensional frequency spectrum in the image domain. In the data preprocessing module, the original audio files are segmented into equal-length audio segments, and MFCC features are extracted and two-dimensional time-frequency spectrograms are generated to obtain the fundamental time-frequency information of the underwater acoustic signal. In the multi-dimensional feature extraction module, 1DCNN and LSTM networks extract the deep time-frequency features in the signal domain, and a 2DCNN network extracts the deep image features corresponding to the two-dimensional frequency spectrum in the image domain, establishing a more comprehensive feature set. In the adaptive multi-feature fusion module, the deep time-frequency information extracted by the three networks is adaptively weighted, assigning more weight to important features, and thus better utilizing the time-frequency information, which significantly improves the recognition accuracy. The performance of the proposed method is validated on the Shipear dataset, and its recognition accuracy is higher than other existing methods, which not only fully demonstrates the superiority of this method in solving underwater acoustic recognition tasks but also provides new ideas for the development of underwater acoustic target recognition methods.

Acknowledgements This research was supported by the Innovation Fund for Marine Defense Technology Innovation Center (No.JJ-2021-705-03), Shaanxi Provincial Key R&D Programme General Project 2024SF-YBXM-572, National Natural Science Foundation of China (INo.62001380) and the General Special Scientific Research Program of Shaanxi Provincial Education Department (20JK0910).

Author Contributions Xiaoying Pan: Conceptualization, Formal analysis, Investigation, Methodology, Funding acquisition, Supervision, Writing - original draft, Writing - review & editing. Jia Sun: Methodology, Data curation, Software, Validation, Visualization, Writing - review & editing. Tianhao Feng: Conceptualization, Methodology, Project administration, Resources, Supervision. Mingzhu Lei: Resources, Writing - review & editing. Hao Wang: Resources, Writing - review & editing. Wuxia Zhang: Formal analysis, Supervision, Writing - review & editing.

Data Availability The datasets are publicly available at https://doi.org/10.1016/j.apacoust.2016.06.008.

Declarations

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. Jian M, Liu X, Luo H, Lu X, Yu H, Dong J (2021) Underwater image processing and analysis: A review. Signal Process Image Commun 91:116088
2. Wu Y, Li X, Wang Y (2018) Extraction and classification of acoustic scattering from underwater target based on wigner-ville distribution. Appl Acoust 138:52–59
3. Meng Q, Yang S (2015) A wave structure based method for recognition of marine acoustic target signals. J Acoust Soc Am 137(4):2242–2242
4. Zak A (2008) Neural classification of ships hydroacoustic signatures. J Acoust Soc Am 123(5):3953
5. Malyshkin G, Sidel'nikov G (2014) Optimal and adaptive methods of processing hydroacoustic signals. Acoust Phys 60:570–587
6. Hu G, Wang K, Peng Y, Qiu M, Shi J, Liu L (2018) Deep learning methods for underwater target feature extraction and recognition. Comput Intell Neurosci 2018
7. Han XC, Ren C, Wang L, Bai Y (2022) Underwater acoustic target recognition method based on a joint neural network. Plos one 17(4):0266425
8. Zhang Q, Da L, Zhang Y, Hu Y (2021) Integrated neural networks based on feature fusion for underwater target recognition. Appl Acoust 182:108261
9. Maksym JN, Bonner AJ, Dent CA, Hemphill GL (1983) Machine analysis of acoustical signals. Pattern Recog 16(6):615–625
10. Lourens J (1988) Classification of ships using underwater radiated noise. In: COMSIG 88@ m_Southern African conference on communications and signal processing. Proceedings, pp. 130–134. IEEE
11. Farrokhrooz M, Karimi M (2005) Ship noise classification using probabilistic neural network and ar model coefficients. In: Europe Oceans 2005, vol 2, pp. 1107–1110. IEEE
12. Yang H, Gan A, Chen H, Pan Y, Tang J, Li J (2016) Underwater acoustic target recognition using svm ensemble via weighted sample and feature selection. In: 2016 13th International Bhurban conference on applied sciences and technology (IBCAST), pp. 522–527. IEEE
13. Chen Y, Xu X (2017) The research of underwater target recognition method based on deep learning. In: 2017 IEEE International conference on signal processing, communications and computing (ICSPCC), pp. 1–5. IEEE
14. Zhang S, Xing S (2018) Intelligent recognition of underwater acoustic target noise by multi-feature fusion. In: 2018 11th International symposium on computational intelligence and design (ISCID), vol 1, pp. 212–215. IEEE
15. Abdoli S, Cardinal P, Koerich AL (2019) End-to-end environmental sound classification using a 1d convolutional neural network. Expert Syst Appl 136:252–263
16. Mishachandar B, Vairamuthu S (2021) Diverse ocean noise classification using deep learning. Appl Acoust 181:108141
17. Tiwari V (2010) Mfcc and its applications in speaker recognition. Int J Emerg Techn 1(1):19–22
18. Yu Y, Si X, Hu C, Zhang J (2019) A review of recurrent neural networks: Lstm cells and network architectures. Neural Comput 31(7):1235–1270
19. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv:1406.1078
20. Sherstinsky A (2020) Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. Physica D: Nonlinear Phenom 404:132306

21. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
22. Lei X, Pan H, Huang X (2019) A dilated cnn model for image classification. IEEE Access 7:124087–124095
23. O'Shea K, Nash R (2015) An introduction to convolutional neural networks. arXiv:1511.08458
24. Huang S, Tang J, Dai J, Wang Y (2019) Signal status recognition based on 1dcnn and its feature extraction mechanism analysis. Sensors 19(9):2018
25. Li S, Song W, Fang L, Chen Y, Ghamisi P, Benediktsson JA (2019) Deep learning for hyperspectral image classification: An overview. IEEE Trans Geosci Remote Sens 57(9):6690–6709
26. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141
27. Santos-Domínguez D, Torres-Guijarro S, Cardenal-López A, Pena-Gimenez A (2016) Shipsear: An underwater vessel noise database. Applied Acoustics 113:64–69
28. Prechelt L (1998) Early stopping-but when? In: Neural information processing systems

Authors and Affiliations

Xiaoying Pan[1,2,3] · Jia Sun[1,2,3] · TianHao Feng[1,2,3] · MingZhu Lei[1,2,3] · Hao Wang[1,2,3] · WuXia Zhang[1,2,3]

Xiaoying Pan
panxiaoying@xupt.edu.cn

Jia Sun
cici@stu.xupt.edu.cn

TianHao Feng
1598018796@qq.com

MingZhu Lei
leimingzhu@stu.xupt.edu.cn

Hao Wang
hwdking@163.com

1 Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an 710121, China

2 Xi'an Key Laboratory of Big Data and Intelligent Computing, Xi'an 710121, China

3 School of Computer Science and Technology, Xi'an University of Post & Telecommunications, Xi'an 710121, China