Engineering Applications of Artificial Intelligence 133 (2024) 107983

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Underwater acoustic target recognition based on sub-band concatenated Mel spectrogram and multidomain attention mechanism

Shuang Yang [1], Anqi Jin [1], Xiangyang Zeng [*], Haitao Wang, Xi Hong, Menghui Lei

School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China

ARTICLE INFO

Keywords:
Underwater acoustic target recognition
Mel spectrogram
Attention mechanism
Residual network

ABSTRACT

Underwater acoustic target recognition is extremely challenging because of the pronounced background noise and intricate sound propagation patterns inherent to maritime environments. Herein, we propose a sub-band concatenated Mel spectrogram to amplify low-frequency ship-radiated noise. This method enhances features through multispectrogram concatenation. Furthermore, we introduce a multidomain attention mechanism to enhance the performance of a simple residual network to develop a lightweight CFTANet model. The recognition accuracies of the recognition system are 90.60% and 96.40% on two open datasets. On the DeepShip dataset, the recognition accuracy is 7.06% higher than those of previous state-of-the-art methods.

1. Introduction

The objectives and tasks of underwater acoustic-signal processing (UASP) can be broadly classified into four categories: detection, tracking, localization, and recognition. Underwater acoustic target recognition (UATR) is an information-processing technique that utilizes active target echoes from sonar, passive target-radiated noise, and other sensor data to extract features for determining the target type. As the ultimate goal of UASP, UATR has become increasingly important in national defense applications (Teng and Zhao, 2020). Additionally, owing to the continuous development of underwater unmanned platform technology, endurance and autonomous navigation capabilities continue to improve, and the overall design and manufacturing level of the all-ocean deep submarine standard has advanced significantly (Lin et al., 2023; Tijjani et al., 2022; Wang et al., 2020a; Hou et al., 2022). The use of machine learning to conduct automatic identification research has intensified, particularly for unmanned marine equipment (Song et al., 2023; Dzikowicz et al., 2023). Furthermore, unmanned equipment are increasingly being operated, which has resulted in the integration of UATR and unmanned equipment technology. However, to realize these application prospects, the pertinent research and development must be continued. This includes improving the accuracy of the algorithms involved, identifying new methods for data acquisition and processing, and conducting real-world testing and validation for various practical application scenarios.

In recent years, owing to the updated iteration of computer hardware and the continuous innovation of deep learning theory (Hinton and Salakhutdinov, 2006; Krizhevsky et al., 2012), deep learning has afforded satisfactory results in the fields of computer vision (Liu et al., 2021a), natural language processing (Vaswani et al., 2017), data enhancement (Goodfellow et al., 2014; Kim et al., 2021; Ge et al., 2021; Zhang et al., 2019a), data noise reduction (Park and Lee, 2017), etc. Several influential network models have emerged in the field of computer vision, among which the classic representatives are convolutional neural networks (CNNs), Transformers (Dosovitskiy et al., 2020), and graph convolutional neural networks (GCNs) (Kipf and Welling, 2016). Classic CNN architectures such as AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2014), and ResNet (He et al., 2016) have not only set new records in terms of accuracy and performance but have also resulted in breakthroughs in hardware acceleration using GPUs and TPUs. In recent years, the Transformer architecture, which has been shown effective in natural language processing, has gradually been extended to computer vision tasks. For example, the Vision Transformer model (Dosovitskiy et al., 2020) introduces self-attention mechanisms into image processing, thus enabling efficient representation learning from images. Additionally, GCNs provide a new approach for modeling objects and scenes in images (Xu et al., 2021). These advanced computer vision algorithms are beneficial to UATR. Computer vision requires

* Corresponding author.
E-mail addresses: yang_s@mail.nwpu.edu.cn (S. Yang), jinaq@mail.nwpu.edu.cn (A. Jin), zenggxy@nwpu.edu.cn (X. Zeng), wht@nwpu.edu.cn (H. Wang), xihong@mail.nwpu.edu.cn (X. Hong), leimenghui@mail.nwpu.edu.cn (M. Lei).
[1] Shuang Yang and Anqi Jin are co-first authors.

https://doi.org/10.1016/j.engappai.2024.107983
Received 25 August 2023; Received in revised form 13 December 2023; Accepted 24 January 2024
Available online 19 February 2024

efficient signal processing and complex relationship modeling, similar to the UATR. The application of Transformer models, including the Swin Transformer, to UATR has intensified (Feng and Zhu, 2022; Li et al., 2022; Xu et al., 2023; Chen et al., 2024). However, the large-scale parameters and high-performance GPU requirements of Transformers and Swin Transformer models pose deployment challenges. Additionally, the scarcity of data in underwater target recognition tasks, particularly in specific application scenarios, limits the performance of model algorithms and prevents them from achieving competitive performance levels. Therefore, further research is required to maintain high performances while reducing model complexity to facilitate broader applications. Jiang et al. (2022) proposed the S-ResNet model, which combines a CNN and SqueezeNet, to achieve good classification accuracy while significantly reducing the model complexity. Tian et al. (2023) used lightweight network design techniques to create a lightweight MSRDN, which reduced the number of parameters by 64.18%, reduced and the FLOPs by 79.45% from the original MSRDN with minimal accuracy loss, and achieved an accuracy of 79.50% on the DeepShips dataset. Yang et al. (2023) introduced a lightweight LW-SEResNet10 model that combined ResNet10 and the channel attention mechanism, which achieved a 97.7% classification accuracy on the ShipsEar dataset and afforded balance between classification accuracy and model complexity. These lightweight CNN models not only maintain high performance but also significantly reduce computational and resource requirements, thus rendering them suitable for practical applications and deployment. By adopting lightweight CNN models, UATR technology can be applied more widely to satisfy the requirements of different scenarios. This provides further possibilities for future research and practical applications to achieve more efficient, scalable, and practical UATR systems.

In general, the sound pressure channel of hydroacoustic signals comprises one-dimensional time-domain data from 20 Hz to 20 kHz. Unlike general pattern recognition, such as image recognition based on deep learning, which can directly use image data as the input to the network, for hydroacoustic signals with a long duration and high sampling rate, the network may be overburdened if raw data are directly used as the input without frame-splitting processing. Raw data typically contain noise and interference (Domingos et al., 2022). As ocean noise and interference information increase, the expression of the target becomes weaker, whereas the transform-domain features can express the target more efficiently and increase the class spacing between different classes. Therefore, owing to the specificity of hydroacoustic signals, most studies pertaining to hydroacoustic target recognition are based on feature extraction, and the manually extracted features are used as inputs for classification. Moreover, the use of artificial feature parameters such as Mel-frequency cepstral coefficients (MFCCs) (Liu et al., 2008), constant-Q transform (Irfan et al., 2021), wavelet features (Wei et al., 2011), DEMON and LOFAR spectra (Chen and Xu, 2017), and high-order spectral features (Liu et al., 2021b) has effectively reduced information redundancy and minimized computational costs in backend models. The Mel spectrogram reflects the human perceptual characteristics of speech and is a frequency feature extracted at a Mel-scale frequency. The Mel spectrogram is more consistent with the auditory characteristics of the human ear; therefore, it is widely used in speech recognition and UATR (Zhu et al., 2023a). Inspired by the radiated noise characteristics of ships and considering the disadvantages of low-frequency signal-focusing feature-extraction techniques, such as the Mel spectrogram, we propose a sub-band concatenated Mel (SC-Mel) spectrogram to solve the problem of feature extraction by jointly realizing feature signal enhancement via a multispectrogram. Subsequently, to efficiently extract time–frequency features, we construct simple residual networks to obtain lightweight networks that can be deployed easily. Finally, to combine the features in each frequency band and the information interaction of the features in each time frame as well as to enhance the features of each channel, we propose two modules, i.e., frequency attention and the CFTA block. The CFTA block comprises successive channel, frequency, and time attention mechanisms. The recognition model is named the CFTANet. Experimental results show that the proposed CFTANet achieves 90.60% optimal recognition results on DeepShip, which is a significant improvement over other classical methods. A high recognition rate of 96.40% is achieved on ShipsEar. Additionally, owing to its lightweight parameters and FLOPs, the CFTANet model is suitable for deployment to underwater unmanned platforms.

The remainder of the paper is organized as follows: Section 2 presents the relevant surveys. Section 3 describes the proposed SC-Mel feature method and introduces the proposed CFTANet. Section 4 describes the setup of the experimental datasets and the setting of the specific experimental parameters. Section 5 presents the experimental results and discussion, including the recognition performances of two datasets and details regarding ablation experiments. Finally, Section 6 presents the conclusions and directions for future research.

2. Related studies

2.1. Challenges and frontier research for UATR

Ferguson et al. (2017) proposed a CNN based on cepstrum inputs, which achieved an accuracy of 99.78% on a self-measured ship-radiated noise dataset. The proposed model successfully detected and ranged sources at long distances under different signal-to-noise ratios (SNRs). Choi et al. (2019) simulated acoustic data from multiple sources with multiple SNRs and uniformly generated samples in each depth region using a normal model and a reciprocal spectral covariance matrix. The experimental results demonstrated the superiority of the CNN as a feature extractor and classifier. In a four-stage method, the discrete wavelet transform (DWT) was used to classify underwater acoustic signals with noise robustness (Kim et al., 2021). The stages included white noise elimination based on the DWT, an imaging stage where the spectrogram of the discrete wavelet coefficients was obtained, data augmentation, and the final classification stage. With its classification accuracy of 99.7% and SNR of 0 dB, the wavelet transform demonstrated its effectiveness for improving the noise robustness of the system and surpassed six different CNN architectures. Vahidpour et al. (2015) performed preprocessing to reduce the effect of noise and provide signals for feature extraction. To extract useful features, a binary image was created based on the signals' segmentation spectrum. Third, the signals were classified using a neural classifier. Consequently, an accuracy rate of 95.1% was achieved on a five-class acoustic dataset. The feature-extraction and deep-learning methods above have achieved good results on different datasets; however, the research methods are difficult to replicate because of the confidentiality of the data. Therefore, studies based on public datasets are necessary.

Two crucial open datasets were used in this study, i.e., DeepShip (Irfan et al., 2021) and ShipsEar (Santos-Domínguez et al., 2016). ShipsEar has been used more extensively because of its earlier public availability. In the convolutional recurrent neural network (CRNN) model established by Liu et al. (2021b), Mel spectrograms from three channels were utilized as network inputs and time-domain transforms, and time–frequency masking was employed to improve the data. A CRNN based on a CNN and LSTM was established to extract local and timing-related aspects of the signal. To improve recognition results, Hong et al. (2021) proposed a data-enhancement strategy with three-dimensional fusion features and SpecAugment and constructed an 18-layer residual network (ResNet18) on this basis; consequently, a correct recognition rate of 94.3% was achieved on the dataset. Other novel methods, such as deep recurrent wavelet autoencoders (Khishe, 2022) and AMNet (Wang et al., 2023), which couple attention and multibranching, have validated the effectiveness of these methods on ShipsEar. DeepShip is used less frequently compared with ShipsEar as it had only been proposed recently. Tian et al. (2023) proposed a lightweight MSRDN model that compensates for the absence of single modal features by combining one-dimensional time-domain modes and two-dimensional time–frequency-domain modes, which resulted in

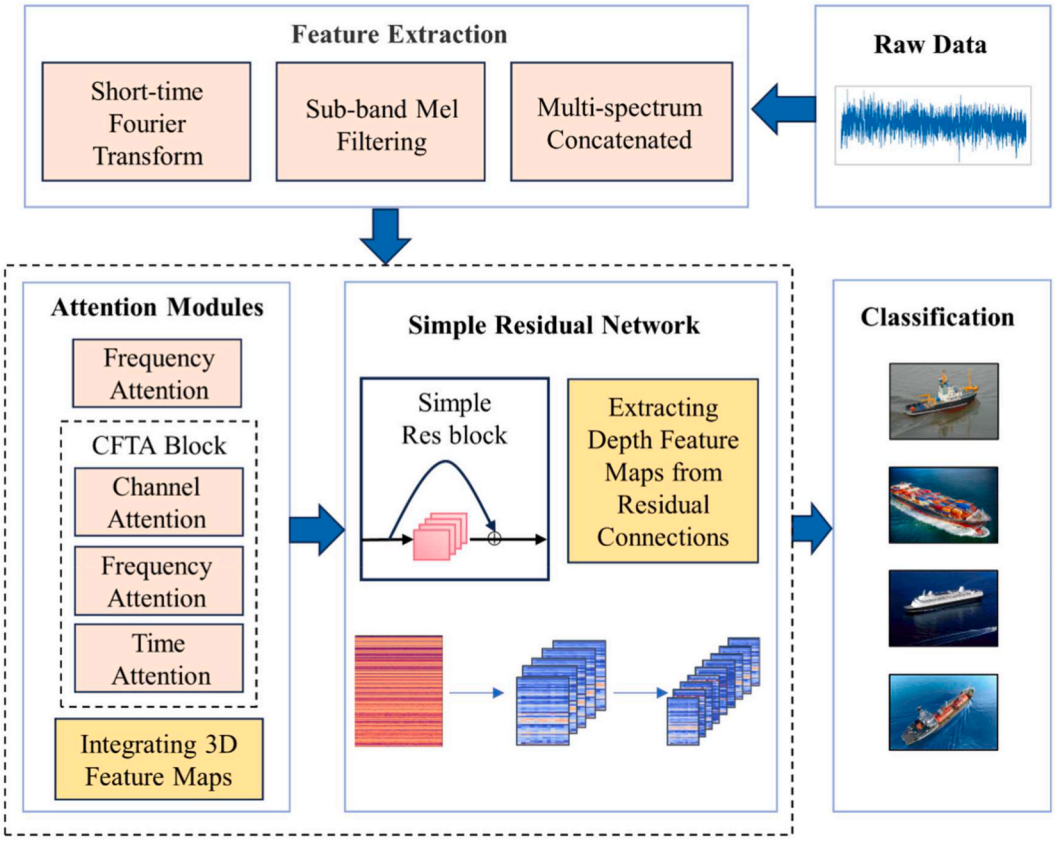


Fig. 1. Framework of proposed UATR system.

performance improvements on both the ONC and DeepShip datasets. Ren et al. (Ren et al., 2022; Xie et al., 2022) constructed several learnable model systems from different perspectives to improve the model's recognition performance. Zhang et al. (Zhang and Zeng, 2023) proposed the MSLEFC system, which combines multiscale time–frequency features, trapezoidal coding, and frequency attention. The method achieved improvement on several public datasets compared with most novel methods. In the current study, we use two open datasets, i.e., DeepShip and ShipsEar, to validate the performance of the proposed model.

2.2. Attention mechanisms

The method of focusing on the most critical areas within an image while disregarding irrelevant components is known as the attention mechanism. In the visual system, the attention mechanism is a dynamic selection process that adaptively weights the input features to reflect their importance. Guo et al. (2022) categorized attention methods based on the data domain into channel attention (Wang et al., 2020b), spatial attention (Dosovitskiy et al., 2020), temporal attention (Zhang et al., 2019b), and branch attention (Li et al., 2019). Different channels represent different objects in a deep neural network. In adaptive channel attention, where the importance of each channel is reweighted dynamically, objects are selected the area of focus is determined. Spatial attention mechanisms can be regarded as adaptive spatial-region selection mechanisms and are typically employed to capture global information. Meanwhile, temporal attention is considered a dynamic time-selection mechanism used to determine the time at which focus should be directed, which emphasizes the capture of short- and long-term interframe feature dependencies.

Furthermore, various joint attention techniques have been widely applied. To enhance information channels and emphasize critical areas, Woo et al. (2018) introduced a convolutional block attention module that concatenates channel attention and spatial attention. Song et al. (2017) proposed a joint spatial and temporal attention network based on LSTM to adaptively identify distinctive features and key frames. For UATR, Wang et al. (2023) utilized a convolutional attention network that combined channel and spatial attention modules to adaptively select effective features by weighting the global information of underwater time–frequency maps. Zhu et al. (2023b) integrated a joint attention module into DenseNet to enhance the ability of the model to recognize multiple target signals in mixed underwater signals through joint time, spatial, channel, and self-attention. Jin et al. (Jin and Zeng, 2023) incorporated channel and time attention mechanisms into ResNet and DenseNet to extract feature information from the channel and the time dimensions of underwater acoustics. Various joint attention methods have been successfully applied in UATR, thus demonstrating that multidomain joint attention can effectively focus on inherent features within nonstationary underwater signals, thereby improving target recognition performance.

3. Proposed method

The framework of the UATR system used in this study, which focuses on feature extraction, simple residual networks, and attention modules, is shown in Fig. 1. First, to maximize the utilization of raw data, we conducted a series of feature extraction steps on the raw data, including short-time Fourier transform (STFT), sub-band Mel filtering, and multispectrum concatenation. These feature extraction steps can amplify the low-frequency information of ship-radiated noise and allow the model to learn more dimensional Mel features. Subsequently, to efficiently extract time–frequency features, we established a simple residual network and classified the radiated noise features into different categories. Finally, to enhance the feature information of the channel, frequency, and time

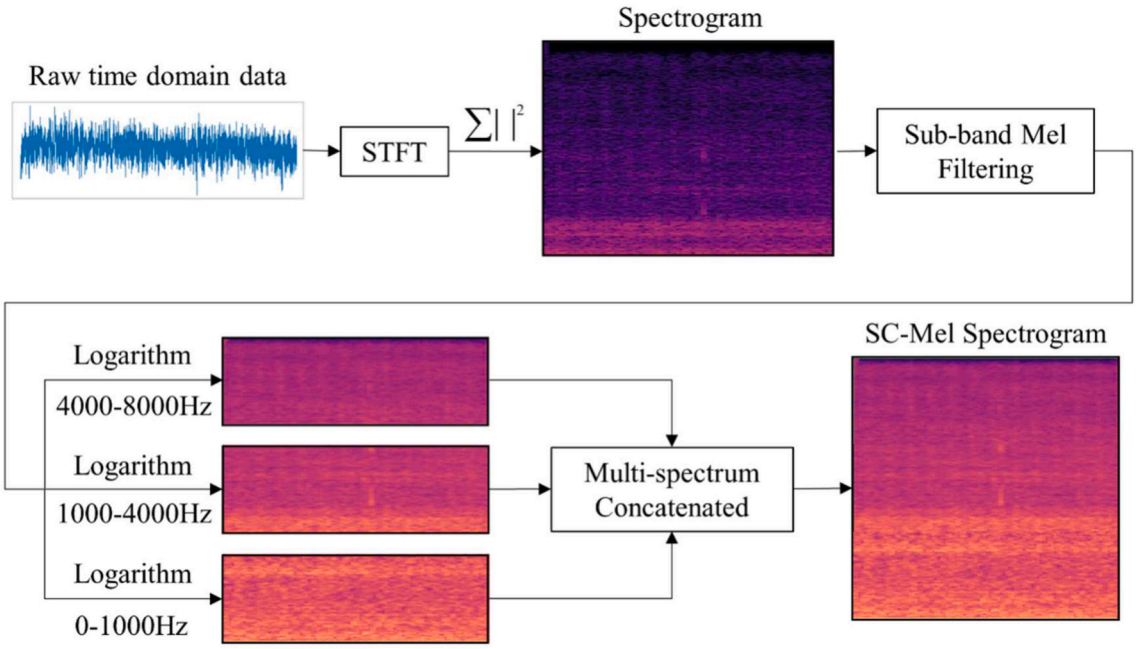


Fig. 2. Block diagram of feature-extraction method.

dimensions of the input features, we propose two attention modules: frequency attention and the CFTA block. The CFTA block comprises continuous channel, frequency, and time attention. Our experimental results show that the frequency attention and CFTA blocks significantly improved the performance of the proposed UATR system.

3.1. Feature extraction

Feature extraction is key in UATR. The raw time-domain data received by a hydrophone can be referred to as raw features, which generally contain a significant amount of data and cannot be used directly as recognition feature quantities. To achieve effective classification and recognition, one must select or transform the original features, obtain the features that best reflect the essence of classification, and form a feature matrix. For ship-radiated noise, feature extraction should satisfy three requirements: first, the number of extracted features should be small and may include slight redundancies; second, sufficient information should be available to best distinguish between different categories; and third, it should be insensitive to typically encountered changes and distortions, i.e., it must be highly robust. Generally, increasing the amount of feature information can improve the recognition performance but can cause feature redundancy. Therefore, balance must be achieved between recognition performance and redundancy. Fig. 2 illustrates the specific feature-extraction process used in our study.

The mechanism of ship-radiated noise signals is complex, and the underwater acoustic propagation channel is affected by the sound-velocity distribution, sea waves, seabed terrain, seabed sediment, seawater medium, and internal waves. It is a complex, time-varying, and spatially variable channel. Considering that the short-time characteristics of ship-radiated noise target signals can reflect the temporal dynamic characteristics, thus describing the time–frequency local characteristics of radiated noise signals more accurately, an STFT (Haykin and Van Veen, 2007) was first performed on the signal, which is defined as follows:

$$S(\omega, n) = \sum_{k=-\infty}^{\infty} s(k) w(n-k) e^{-j\omega k}, \tag{1}$$

where $w(n)$ is a window function that propagates with time n. We reduced the spectrum leakage by weighting the signal with the window function and balanced the time–frequency resolution by adjusting the frame length and frameshift. STFT first segmented the signal into frames and then performed a Fourier transform on each frame. When an appropriate time window function and frequency resolution are selected for the time–frequency analysis of the signals, various time-varying characteristics, including the low-frequency line spectrum and modulation spectrum, can be obtained (Boashash and O'shea, 1990). The amplitude spectrum was extracted from the Fourier transform result of each frame and then squared to obtain the power spectrum of each frame. Each frame signal can be regarded as intercepted from different relatively stable short-term waveforms, and the power spectrum of each frame ship-radiated noise signal an approximation of the power spectrum of each relatively stable waveform. The power spectra of all frames in the time sequence were concatenated to obtain a spectrogram. The spectrogram shows the energy distribution of the signals at different times and frequencies, which can intuitively show the changes in signal energy with time.

Enhancing the feature signal allows the depth model to learn more features and patterns, thus improving its generalization ability and performance. For UATR, Liu et al. expanded audio data by extracting a Mel spectrogram and its incremental features, thereby capturing the dynamic characteristics of ship-radiated noise (Liu et al., 2021b). Yang et al. proposed a similar method to extract the MFCC and incremental features of ship-radiated noise to increase the amount of dynamic feature information (Yang et al., 2023). Zhang et al. proposed a multi-scale STFT method to improve low-frequency information and maintain detailed information by increasing the number of channels (Zhang and Zeng, 2023). In this study, an SC-Mel spectrogram is proposed, and the spectrogram is enhanced by combining sub-band Mel filtering and multispectrum concatenation. The Mel spectrogram is a spectrum representation method typically used in audio signal processing and speech recognition. It describes the energy distribution of sound signals at the Mel scale. To perform calculations based on the Mel spectrogram, Mel filter banks are used to weight the spectrogram, which can map the high-dimensional features of the original spectrogram to a lower-dimensional Mel spectrogram representation and realize the compression of audio information. The Mel scale is a nonlinear pitch perception scale related to human auditory characteristics. The human ear is more sensitive to changes at lower frequencies than to those at

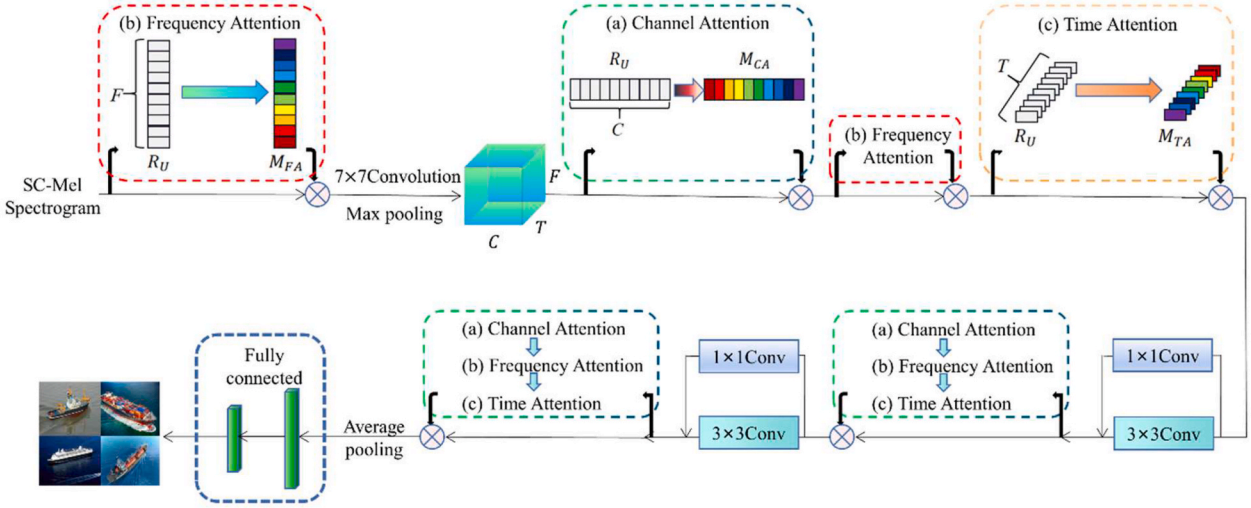


Fig. 3. Comprehensive structure of CFTANet model for identifying underwater acoustics target.

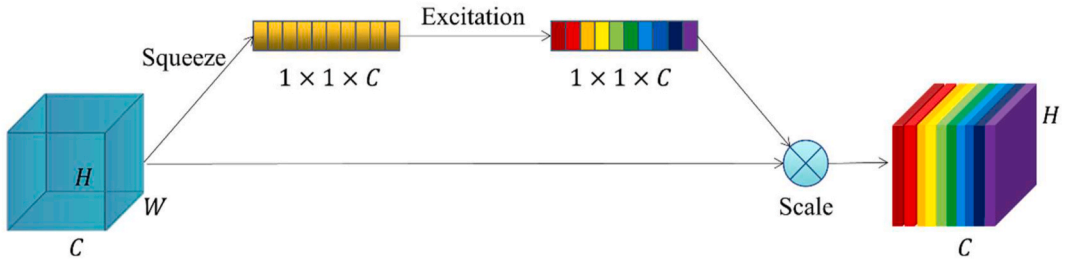


Fig. 4. Structural diagram of channel attention mechanism.

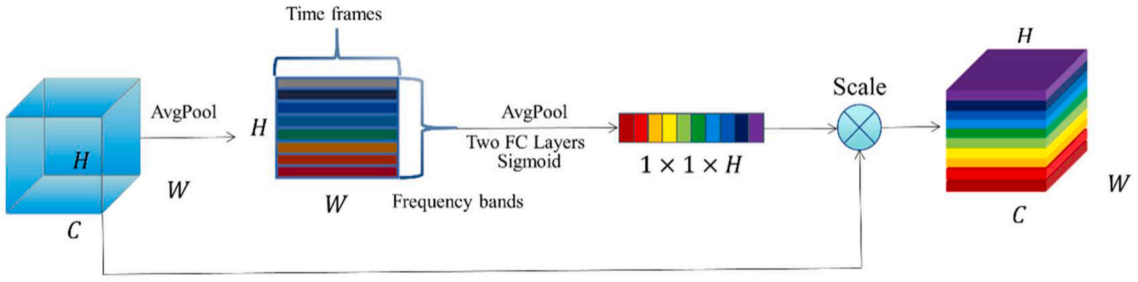


Fig. 5. Structural diagram of frequency attention mechanism.

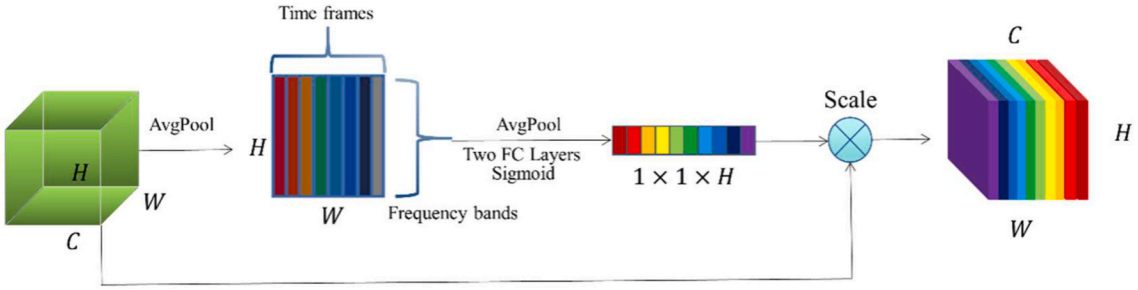


Fig. 6. Structural diagram of time attention mechanism.

higher frequencies. The Mel scale simulates the perceptual characteristics of the human ear accurately by converting frequency into the Mel frequency. The Mel frequency scale presents a logarithmic relationship with the actual frequency (Koening, 1949). The relationship between these parameters can be approximated as follows:

$$Mel(f) = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right), \tag{2}$$

Table 1
Detailed parameters of CFTANet.

Layer name	Output size	CFTANet
Feature	384 × 94 × 1	
Frequency Attention	384 × 94 × 1	
Convolution	192 × 47 × 64	7 × 7, 64, stride 2
Maxpool	96 × 24 × 64	3 × 3, stride 2
CFTA Block	96 × 24 × 64	[Channel Attention Frequency Attention Time Attention]
Residual Block	48 × 12 × 128	3 × 3 Conv, 128, stride 2
CFTA Block	48 × 12 × 128	[Channel Attention Frequency Attention Time Attention]
Residual Block	24 × 6 × 256	3 × 3 Conv, 256, stride 2
CFTA Block	24 × 6 × 256	[Channel Attention Frequency Attention Time Attention]
Classification layer	1 × 1 × 4 (number of classes)	Global average pool Fully connected LogSoftmax

Table 2
Performance of CFTANet on two test sets.

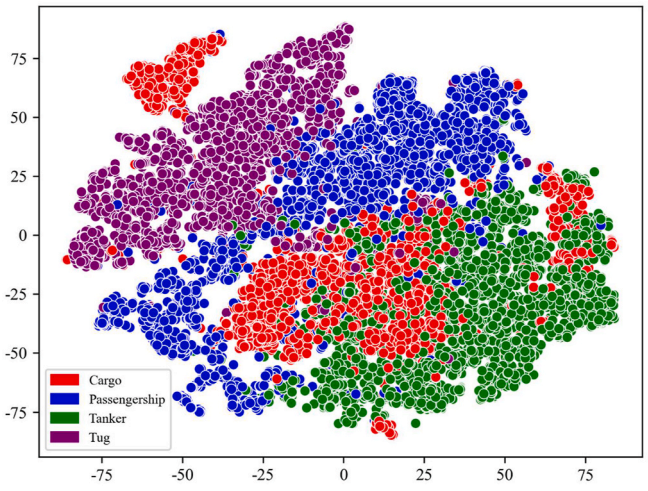
Dataset	Class	Precision (%)	Recall (%)	F1-score	Support
DeepShip	Cargo	86.79	86.23	0.8651	3841
	Passengership	89.66	95.93	0.9269	4591
	Tanker	90.05	87.28	0.8865	4459
	Tug	96.14	92.37	0.9422	4049
	Weighted avg	90.66	90.60	0.9059	16,940
ShipsEar	Class A	100.00	100.00	1.0000	67
	Class B	94.02	93.22	0.9362	118
	Class C	95.46	94.38	0.9492	89
	Class D	97.55	96.76	0.9715	247
	Class E	95.33	97.95	0.9662	146
	Weighted avg	96.41	96.40	0.9640	667

where f denotes the actual frequency (Hz), and. $Mel(f)$ is the perceptual frequency in Mel.

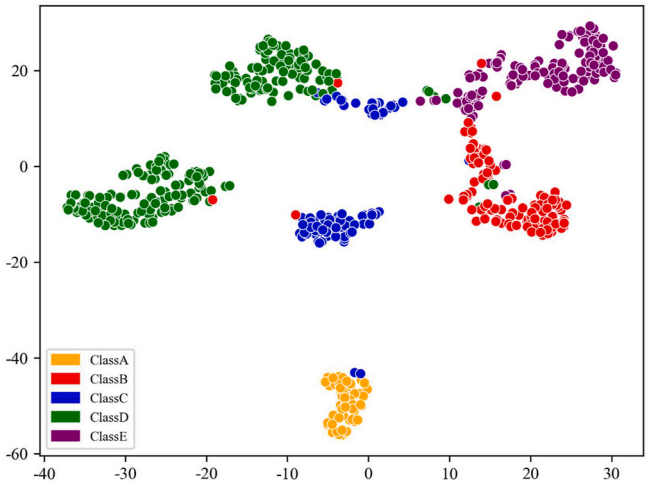
In our study, sub-band Mel filtering intercepted the extracted spectrogram in the sub-band, and Mel filtering was performed on multiple sub-bands to obtain Mel spectrograms in different frequency bands. Subsequently, the Mel spectrograms of different frequency bands were spliced along the frequency dimension to obtain the characteristics of the input model. Compared with withdrawing the Mel spectrogram directly with N Mel filters in the entire frequency band, our method uses independent N Mel filters in each frequency band, which increases the frequency resolution of the spectrogram and clearly displays the characteristics of each frequency band. In UATR, more detailed feature information is typically obtained at low frequencies; therefore, we adjusted the sub-bands to 0–1000, 1000–4000, and 4000–8000 Hz as appropriate to highlight the feature information in the low-frequency range more accurately.

3.2. Proposed CFTANet

In this study, a simple residual network was used as a feature-extraction network, and a frequency attention module and CFTA block were proposed for identifying underwater acoustic targets. In the CFTANet model, the successive channel attention, frequency attention, and time attention modules are defined as CFTA blocks. The CFTANet model is described based on two aspects: depth feature extraction using a simple residual network and feature enhancement using a CFTA block. As shown in Fig. 3, the framework of the proposed CFTANet model primarily comprises a convolution layer, two simple residual blocks, one frequency attention block, and three CFTA blocks. The process is described comprehensively in the following section.



(a)



(b)

Fig. 7. t-SNE visualization of features in CFTANet. (a) DeepShip test set; (b) ShipsEar test set.

Table 3
Comparison between proposed CFTANet and most advanced method on DeepShip dataset.

Methods	Accuracy (%)	Params (M)	FLOPs (G)
SCAE (Irfan et al., 2021)	77.53	1.80	6.001
Lightweight MSRDN (Tian et al., 2023)	79.50	11.74	1.397
UALF (Ren et al., 2022)	81.39	/	/
UART (Xie et al., 2022)	76.56	/	/
MSLEFC (Zhang and Zeng, 2023)	82.94	15.75	1.5596
Swin Transformer (Xu et al., 2023)	80.22	88	/
ConvNeXt (Liu et al., 2022)	83.54	26.51	2.5580
CFTANet (ours)	90.60	0.47	0.1186

3.2.1. Simple residual network

ResNet was proposed to solve the problem of deep neural network degradation (He et al., 2016). Its main feature is a "residual block," which allows the network to traverse directly across layers and train the network by capturing the residual, thus allowing the network to train extremely deep layers more easily. In this study, a simple residual structure was used to form a residual network. A simple residual block is

Table 4
Comparison between proposed CFTANet with most advanced method on ShipsEar dataset.

Methods	Accuracy (%)	Params (M)	FLOPs (G)
Baseline (Santos-Domínguez et al., 2016)	75.4	/	/
CRNN (Liu et al., 2021b)	94.6	0.45	0.25
ResNet18 (Hong et al., 2021)	94.3	11.19	1.88
DRW-AE (Khishe, 2022)	94.49	<0.1	0.02
AMNet (Wang et al., 2023)	99.4	5.47	0.37
STM + AudioSet (Li et al., 2022)	97.70	86	/
HUAT (Chen et al., 2024)	98.62	30.03	/
ConvNeXt (Liu et al., 2022)	91.90	26.51	4.59
CFTANet (ours)	96.40	0.47	0.20

a residual block in which one convolution layer is used as the primary component. A simple residual block is more suitable in certain shallow networks or cases involving limited resources as it introduces fewer parameters and calculations yet affords performance improvement. In our study, the main section of the simple residual block featured a convolution layer comprising a 3 × 3 convolution kernel with a stride of 2. The residual connection comprised a 1 × 1 convolution layer with a stride of 2. The feature map completes downsampling via a simple residual block, which can be expressed as

$$x^{'} = \mathrm{h} - \mathrm{swish}(w_1 x) \tag{3}$$

$$\begin{aligned} H(x) &= x^{'} + F(x) \\ &= x^{'} + \mathrm{h} - \mathrm{swish}(w_2 x), \end{aligned} \tag{4}$$

where x represents the input feature, w_1 the weight of the 1 × 1 convolution layer, w_2 the weight of the 3 × 3 convolution layer, $F(x)$ the residual mapping, and h-swish the nonlinear activation function (Howard et al., 2019). As shown in Fig. 3, the CFTANet uses two simple residual blocks to complete two downsampling steps.

3.2.2. CFTA block

In general, neural networks provide implicit attention to extract high-dimensional information from data. Incorporating an attention mechanism into a network allows valid information to be extracted explicitly. The CFTA block comprises successive channel, frequency, and time attention mechanisms in a sequential series. Next, we describe the channel, frequency, and time attention in the CFTANet based on the theory presented in (Guo et al., 2022).

3.2.2.1. Channel attention. In the channel attention mechanism, the network learns the importance of each channel by first compressing the space of the feature map and then learning the channel dimensions to determine the importance of each channel (Hu et al., 2018). As shown in Fig. 4, Squeeze utilizes the global average pooling (GAP) operation to extract the global receptive field (Eq. (5)), where all feature channels are abstracted to a single point; Excitation utilizes two fully connected layers to perform nonlinear feature transformations to display the construction of correlations between the feature maps (Eq. (6)); and Transform utilizes the sigmoid activation function to achieve feature recalibration (Eq. (7)), which strengthens the important feature maps and weakens the non-important feature maps.

First, global average pooling is performed to obtain a 1 × 1 feature map with global receptive fields.

$$y_c = \mathbf{F}_{sq}(m_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} m_c(i,j) \tag{5}$$

Subsequently, two fully connected layers are used to perform a nonlinear transformation.

$$n = \mathbf{F}_{ex}(y, \mathbf{W}) = \sigma(g(y, \mathbf{W})) = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 y)) \tag{6}$$

The importance weights obtained from Excitation are assigned to the original input features to obtain new features.

$$\widetilde{M}_c = F_{\text{scale}}\,(m_c, n_c) = n_c m_c \tag{7}$$

3.2.2.2. Frequency attention. The input to the network used in this study is a time–frequency spectrogram, which is different from a general image in that the horizontal and vertical axes of the time–frequency graph have specific physical meanings, i.e., they represent the frequency and time, respectively. In this study, the importance of different frequency bands for the recognition task is obtained by constructing frequency attention to improve the recognition correctness. As shown in Fig. 5, we first compress the channel and time axes of the feature map via GAP.

$$X_H = \mathbf{F}_{sq}(m_H) = \frac{1}{C \times W} \sum_{i=1}^{C} \sum_{j=1}^{W} m_H(i,j) \tag{8}$$

Next, a 1 × H × 1 feature map obtained after compression is incorporated into two fully connected layers for learning as follows:

$$K = \mathbf{F}_{ex}(X, \mathbf{W}) \tag{9}$$

Finally, the learned frequency weights are multiplied by the original feature map to obtain a frequency attention-enhanced feature map.

$$\widetilde{M}_H = F_{\text{scale}}\,(m_H, k_H) = k_H m_H \tag{10}$$

The frequency attention module learns the importance of each frequency band and improves the network performance.

3.2.2.3. Time attention. For underwater acoustic signals, multiple samplings in the time domain are important for providing multiple types of statistical information. As time progresses, more information will be obtained, thus facilitating target categorization. We determine the importance of different timeframes by constructing the time attention. As shown in Fig. 6, similar to the frequency attention, we obtain the time weights by compressing the frequency and channel axes via GAP.

$$Z_W = \mathbf{F}_{sq}(m_W) = \frac{1}{C \times H} \sum_{i=1}^{C} \sum_{j=1}^{H} m_W(i,j) \tag{11}$$

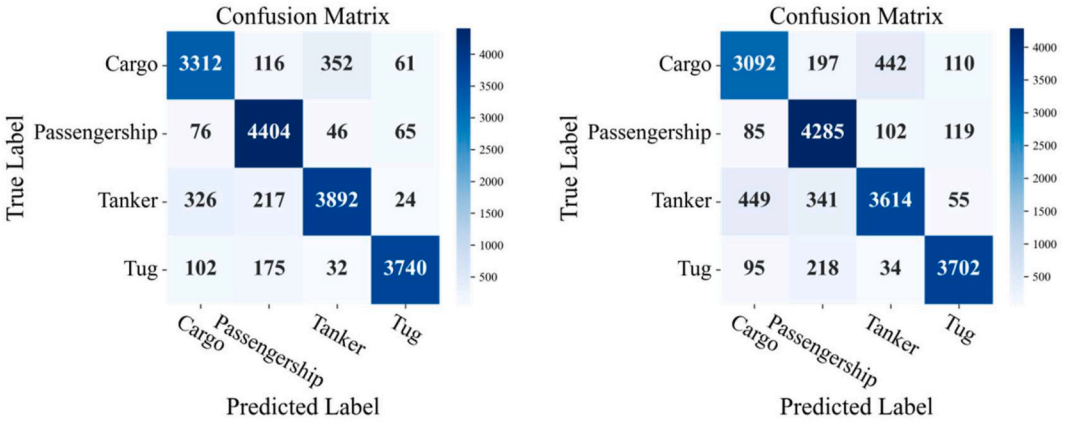
Subsequently, the weights are learned using two fully connected layers.

$$G = \mathbf{F}_{ex}(Z, \mathbf{W}) \tag{12}$$

Finally, a time-weighted feature map is obtained by multiplying the weights with the original feature map:

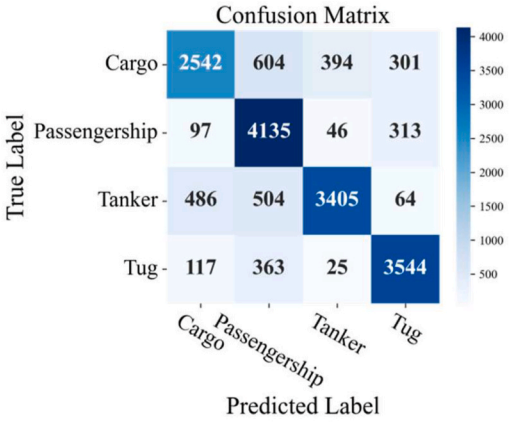
Table 5
Performance comparison of different methods on DeepShip test set.

Methods	Feature	Attention modules	Accuracy (%)	F1-score	Params(K)	FLOPs(M)
1 (Proposed)	SC-Mel	Use all	90.60	0.9059	467.726	118.608
2	Mel	Use all	86.74	0.8667	432.065	39.542
3	SC-Mel	No	80.44	0.8026	405.690	118.548
4	SC-Mel	Frequency Attention	89.15	0.8913	442.113	118.583
5	SC-Mel	The CFTA block after max pool layer	88.32	0.8832	409.284	118.551
6	SC-Mel	Two CFTA blocks after Simple Residual Block layers	87.88	0.8783	427.711	118.569

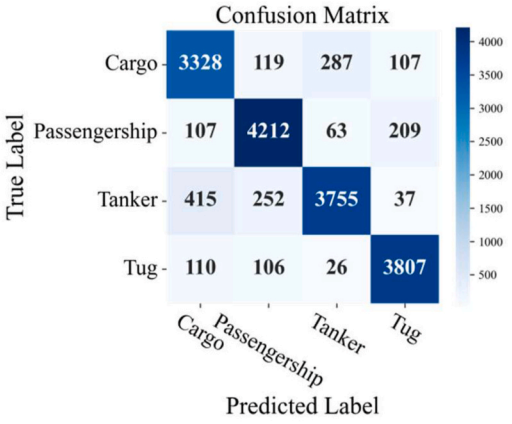


(a) Method 1

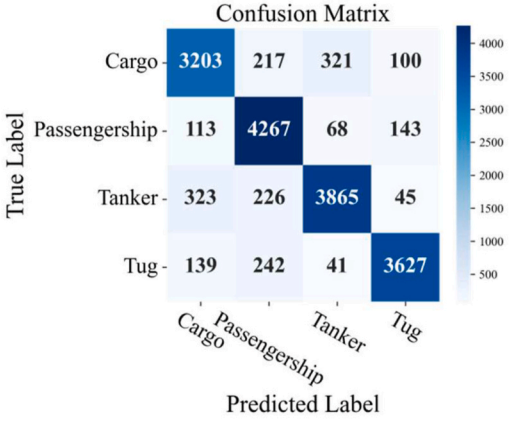
(b) Method 2



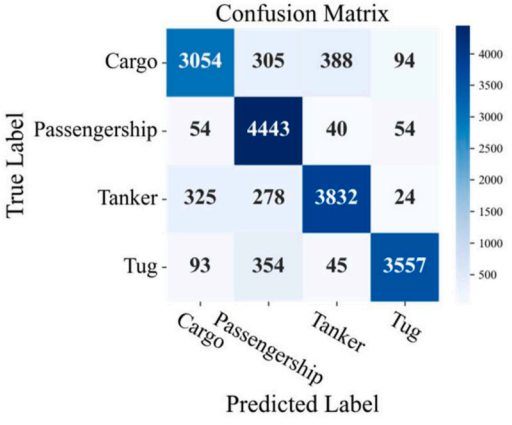
(c) Method 3



(d) Method 4



(e) Method 5

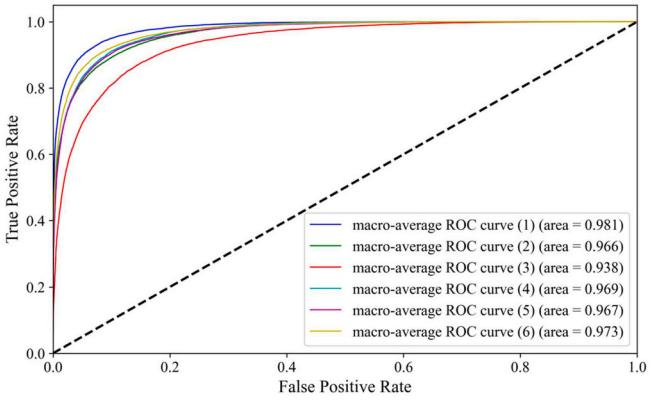


(f) Method 6

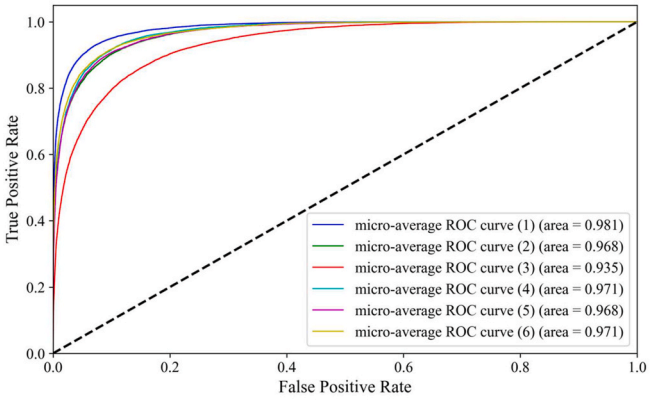
Fig. 8. Confusion matrices of different methods on test set. (a) Method 1, (b) Method 2, (c) Method 3, (d) Method 4, (e) Method 5, and (f) Method 6.

Table 6
Kappa values of different methods on test set.

Methods	Kappa values
1	0.8744
2	0.8227
3	0.7382
4	0.8552
5	0.8440
6	0.8377



(a)



(b)

Fig. 9. ROC curves of different methods on test set. Here, (1) to (6) in legend refer to Methods 1 to 6, respectively.

$$\widetilde{M}_W = F_{\text{scale}}\left(m_W, g_W\right) = g_W m_W \tag{13}$$

3.2.3. CFTANet

To enhance the input features adaptively, we first input the SC-Mel spectrogram features into the frequency attention module. The frequency attention module explicitly models the interdependence of frequency dimensions and transmits effective features through squeezing and excitation to enhance the subsequent convolution feature learning. Therefore, adding a frequency attention module before the first convolution can effectively improve the utilization rate of the frequency information in the feature map and accelerate the convergence of the model. As shown in Table 1, the convolution layer after the frequency attention layer, the maximum pooling layer, and the two simple residual blocks down-sampled the input feature map four times. After the first convolution layer and the convolution layers in the two simple residual blocks, batch normalization was applied to normalize the data batches, and then nonlinear mapping was performed using an h-swish activation function. The maximum pool layer and two simple residual blocks were followed by the CFTA block. The CFTA block comprised continuous channel, frequency, and time attention. The feature map enhances the channel, frequency, and time dimension information through these three attention layers while suppressing meaningless feature information. In the CFTANet, a unified squeezing excitation operation is adopted in the frequency attention module and the attention modules in the CFTA block, and the information is recalibrated via two fully connected layers. After the first fully connected layer, the aggregated feature map is compressed to $1/n$ of the input, and then nonlinear mapping is performed using the h-swish activation function. After the second fully connected layer, the learned channel, frequency, and time information are mapped to 0–1 using a sigmoid activation function. The n in frequency attention is set to 8. The n in the CFTA block after the maximum pool layer is set to 8. The n in the CFTA block after the two simple residual blocks is set to 4. After the aforementioned depth feature-extraction process is completed, the feature map is classified and recognized through the classification layer, which comprises global average pooling, a fully connected layer, and a LogSoftmax classifier.

4. Experiments

4.1. Dataset

The experiment was conducted on two published datasets pertaining to ship-radiated noise: DeepShip (Irfan et al., 2021) and ShipsEar (Santos-Domínguez et al., 2016). These two datasets have been widely used in UATR, and their application in UATR will be further investigated in this study. Further details regarding the dataset are provided below.

DeepShip comprises the data of 265 different ships from four different categories, totaling 47 h and 4 min. An icListen AF hydrophone was used to obtain radiated noise at a sampling frequency of 32,000 Hz. Ocean Networks Canada is the source of this information. The four categories are Cargo, Passengership, Tanker, and Tug. In our experiment, the sample length was 3 s, the samples did not overlap, and the total number of samples was 56,468. We randomly segmented the dataset into training and test sets at a ratio of 7:3. The training and test sets contained 39,528 and 16,940 samples, respectively.

The ShipsEar database was derived from several vessels cruising along the Atlantic coast of northwest Spain. A DigitalHYD SR-1 recorder was used to obtain radiated noise at a sampling frequency of 52,734 Hz. The dataset contains data from 11 types of ships and background noise. Among them, 11 types of ships are classified into four types of ship-radiated noise based on the ship size. Therefore, the dataset is finally classified into five categories as follows:

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

The sample length of the ShipsEar dataset is 5 s, and the samples do not overlap, thus resulting in 2223 samples. We randomly segmented the dataset into training and test sets at a ratio of 7:3. The training and test sets contained 1556 and 667 samples, respectively.

In the experiment, the sampling rate was set to 16,000 Hz. When the signal was Fourier transformed, the length of the FFT window was set to 2048, the frameshift was set to 512, and the Hanning window function was used. For sub-band Mel filtering, our method used three independent groups of 128 Mel filter banks in frequency bands of 0–1000, 1000–4000, and 4000–8000 Hz; subsequently, the three obtained sub-band Mel spectrograms were concatenated along the frequency direction to obtain 384-dimensional Mel features. For the DeepShip dataset, the number of frames for each dimension of the Mel feature was 94, and the feature shape measured 384 × 94. For the ShipsEar dataset, the

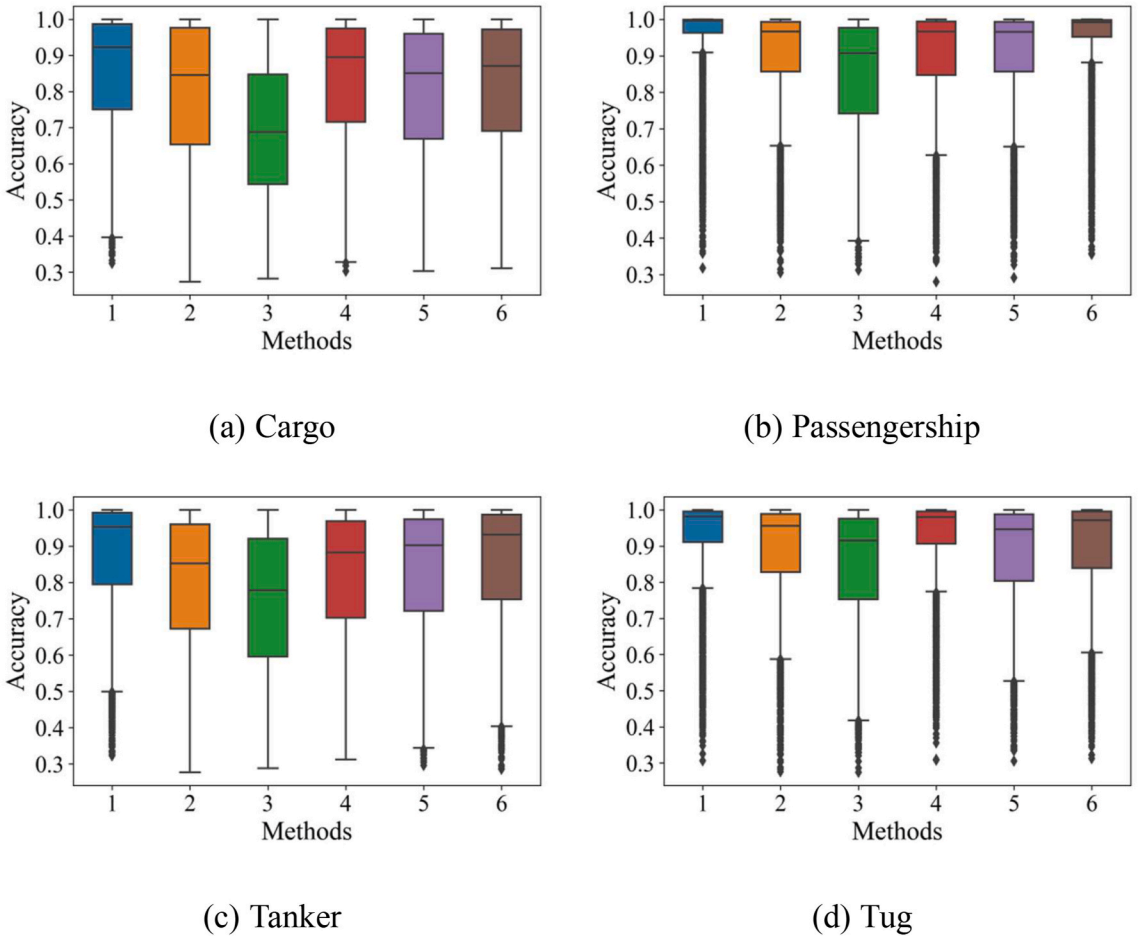


Fig. 10. Accuracy of different methods on various categories on test set. Abnormal samples are represented by black diamonds.

number of frames for the Mel feature in each dimension was 157, and the feature shape measured 384 × 157.

4.2. Setup and parameter

The adaptive moment estimation (Adam) optimizer (Kingma and Ba, 2015) with L2 regularization (set to 4×10^{-5}) was employed. The initial learning rate (set to 0.001) multiplied by the cosine learning rate-decay function (He et al., 2019), which accelerates the training process, determines the learning rate of the training process. The batch size was 64, 50 training epochs were used, and the cost function was the cross-entropy error (Goodfellow et al., 2016).

5. Results and discussion

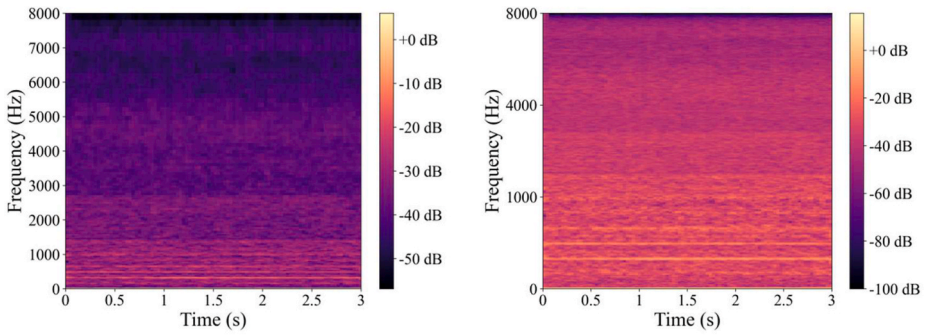
All experiments were performed using Pytorch 1.7.1 and Python version 3.8.5, and verified using computers with an NVIDIA GeForce RTX 3090 GPU and a Core i9-10900 K CPU.

5.1. Recognition performance

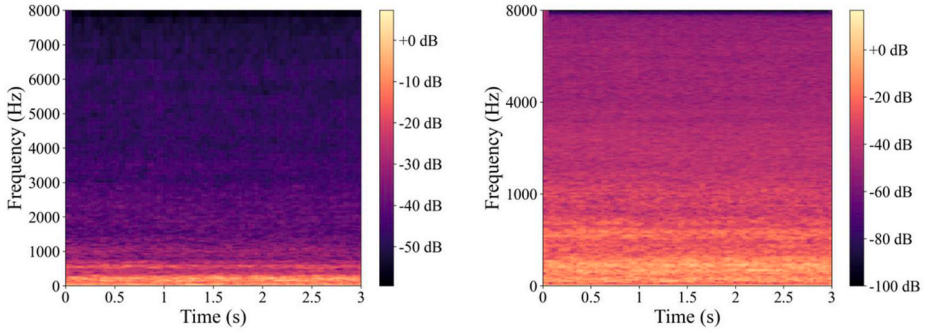
First, the recognition performance of the CFTANet model was tested on the DeepShip and ShipsEar datasets. We conducted 10 repeated experiments on the two datasets to ensure the reliability of the experiments. In the repeated experiments, 10 random seeds were used for data segmentation. Table 2 lists the average indicators of the CFTANet for each class on the two test datasets, which describe the classification system in terms of the precision, recall, and F1-score. The supporting information in Table 2 refers to the number of samples. In terms of the three evaluation indicators above, the weighted average scores of the CFTANet on DeepShip were 90.66%, 90.60%, and 0.9059, respectively, whereas those of the CFTANet on ShipsEar were 96.41%, 96.40%, and 0.9640, respectively. For the CFTANet results on DeepShip, the Tug class indicated the best F1-score. For the results of CFTANet on ShipsEar, Class A (background noise) indicated the best F1-score. Compared with the DeepShip test set, the ShipsEar test set indicated a greater difference in sample size between classes; however, its precision, recall, and F1-score for each class exceeded 90%. Therefore, the CFTANet is applicable under imbalanced sample categories.

Subsequently, t-SNE visualization (Van der Maaten and Hinton, 2008) was applied to observe the distribution of different categories in a high-dimensional feature space. We selected the t-SNE visualization result with the highest accuracy from the 10 repeated experiments. A feature distribution map on the DeepShip test set is shown in Fig. 7 (a). The data points of the Tug class were clustered well in the feature space, whereas dataset overlap occurred among the Cargo, Passengership, and Tanker categories. Meanwhile, the feature data points of the Cargo class did not cluster well together, and their intraclass consistency was unsatisfactory, as confirmed by the higher F1-score of the Tug class and the lower F1-score of the Cargo class (see Table 2). Fig. 7(b) shows the feature distribution diagram on the ShipsEar test set. Compared with the feature map on the DeepShip test set, the feature map on the ShipsEar test set contained category feature information that was easier to distinguish.

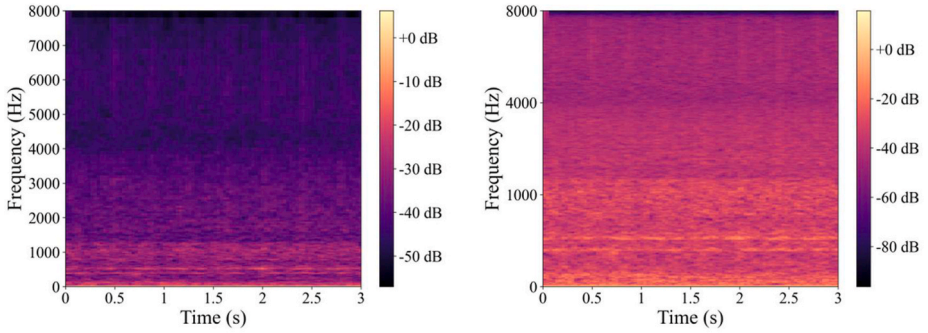
Moreover, we compared the recognition performance of the CFTANet model with those of previous models. For the DeepShip dataset, Table 3 show a comparison between the proposed CFTANet model with a method based on deep learning (Irfan et al., 2021; Tian et al., 2023;



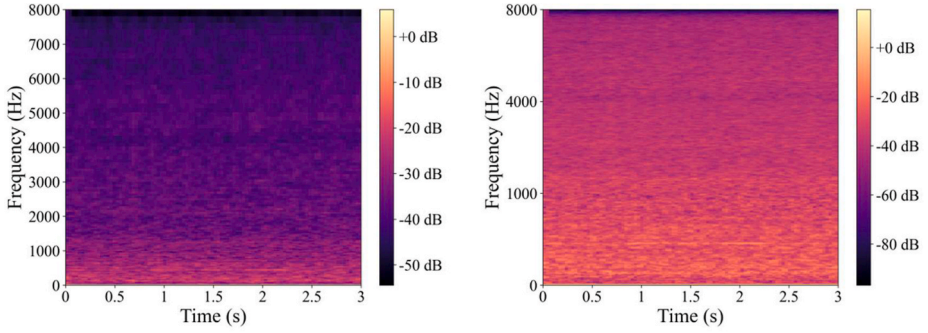
(a) Cargo



(b) Passengership



(c) Tanker



(d) Tug

Fig. 11. Mel-based features on various samples on test set. Left and right images for each category show Mel and SC-Mel spectrograms, respectively.

Ren et al., 2022; Xie et al., 2022; Zhang and Zeng, 2023; Xu et al., 2023; Liu et al., 2022). As shown, the accuracy of the proposed CFTANet model on the DeepShip dataset was 13.07%, 11.10%, 9.21%, 14.04%, 7.66%, 10.38%, and 7.06% higher than those of SCAE, Lightweight MSRDN, UALF, UART, MSLEFC, Swin Transformer, and ConvNeXt, respectively. Meanwhile, the parameters and FLOPs of the proposed CFTANet model were significantly lower than those of the other models.

For the ShipsEar dataset, Table 4 shows a comparison of the CFTANet model with methods based on deep learning (Santos-Domínguez et al., 2016; Liu et al., 2021b; Hong et al., 2021; Khishe, 2022; Wang et al., 2023; Li et al., 2022; Chen et al., 2024; Liu et al., 2022). As shown in Table 4, the accuracy of our model on the ShipsEar dataset was 21.00%, 1.80%, 2.10%, 1.91%, and 4.5% higher than those of the baseline, CRNN, ResNet18, DRW-AE, and ConvNeXt, respectively. The parameters and FLOPs of the DRW-AE network architecture were the lowest; however, its recognition accuracy was lower than that of CFTANet. The parameters and FLOPs of the CFTANet model and CRNN model were similar, whereas the accuracy of the CFTANet was higher than that of the CRNN August data model. Although the accuracy of the CFTANet model was lower than that of the AMNet, STM + AudioSet, and HUAT models, its parameters and FLOPs were lower than those of the three models. Therefore, the CFTANet achieved a good trade-off between accuracy and model size on the ShipsEar dataset.

5.2. Ablation experiment

In this section, we analyze and evaluate the effects of the feature and attention modules on the UATR system based on the DeepShip dataset. The comprehensive performances of the different methods are listed in Table 5. We labeled these methods as Methods 1–6. The accuracy and F1-score were used to measure the recognition effect of the recognition system, and the parameters and FLOPs values were used to measure the model size. In Method 1, the UATR system proposed in Section 2, namely, the SC-Mel spectrogram feature in Section 2.1 and the CFTANet model in Section 2.2, was used. In contrast to the feature-extraction method described in Section 2.1, Method 2 directly used the Mel spectrogram of the entire frequency band of the signal with 128 Mel filters. Therefore, the input data size of the CFTANet model in Method 2 was 128×94. Using the SC-Mel spectrogram, the number of parameters increased by 8.25%, the accuracy gain was 3.86%, and the F1-score gain was 0.0392; however, the FLOPs increased by 199.95%.

In Methods 3–6, the effects of the frequency attention module and CFTA block on the entire recognition system were evaluated by removing some attention modules from the CFTANet model. In Method 3, all attention modules in the CFTANet model were removed. In Method 4, only the frequency attention module before the first convolution layer was retained. In Method 5, only the CFTA block module after the maximum pool layer was retained. In Method 6, only the two CFTA block modules behind the two simple residual blocks were retained. Compared with Method 3, Method 1 increased the parameter by 15.29% and the FLOPs by 0.05%, as well as improved the accuracy by 10.16% and the F1-score by 0.1033. Compared with Method 3, Method 4 increased the parameters by 8.98% and the FLOPs by 0.03%, as well as improved the accuracy by 8.71% and the F1-score by 0.0887. Compared with Method 3, Method 5 increased the number of parameters by 0.89% and the FLOPs by 2.53×10^{-5}, as well as improved the accuracy by 7.88% and the F1-score by 0.0806. Compared with Method 3, Method 6 increased the parameters by 5.43% and the FLOPs by 0.02%, as well as improved the accuracy by 7.44% and the F1-score by 0.0757. In summary, Method 1 afforded the best recognition effect. The frequency attention module and the three CFTA block modules in the CFTANet model contributed to the final recognition effect, and the increase in the parameters and FLOPs was acceptable.

Fig. 8 shows the confusion matrix of the six methods on the test set. Table 6 lists the kappa values corresponding to the confusion matrices obtained using various methods. The kappa value was used to calculate the actual accuracy of the algorithm and the expected accuracy under random conditions. The higher the kappa value, the greater was the difference between the accuracy of the classification algorithm and random classification. The kappa value of Method 1 was the highest, which shows that the difference between the classification results of the CFTANet model in Method 1 and random classification was the most significant, and that the classification results were more reliable.

In addition, we used the receiver operating characteristic (ROC) metric to evaluate the classifier output quality. The shape of the ROC curve and its proximity to the upper-left corner reflect the classifier performance. The closer the curve is to the upper left corner, the better is the classifier performance. The area under the ROC curve (AUC) was used to quantitatively evaluate the classifier performance. Fig. 9 shows the ROC curves and AUC values obtained using the macro- and micro-averages, respectively. As shown, Method 1 yielded the highest AUC value under the macro- and micro-averages. This indicates that the overall performance of Method 1 and its performance in each category were the best.

To better understand the performance distribution of each method in different categories, we compared the accuracy changes in various categories of the samples on the test set, as shown in Fig. 10. Comparing the median and upper and lower quartile scores of Methods 1 and 2 on the four types of test sets, we discovered that the SC-Mel features effectively improved the median and upper and lower quartile scores and that the data distribution was more concentrated; in other words, the intraclass consistency of the SC-Mel features was greater than that of the Mel features. Comparing Method 1 with Method 3, the frequency attention and CFTA block modules in the CFTANet model effectively improved the median, upper, and lower quartile scores for the data pertaining to the four types of test sets. In addition, the median and upper and lower quartile scores of Method 1 were the highest for the four types of test set data, whereas the score range of the outliers was wider, thus indicating that the performance of the SC-Mel features combined with the CFTANet model was relatively better in most samples; this indicates that the proposed method has good generalization ability but does not perform well on some specific samples, which may include extreme data points.

In addition, we constructed Mel and SC-Mel spectrograms to visualize feature extraction. As shown in Fig. 11, the low-frequency line spectrum characteristics presented in the SC-Mel spectrogram are more prominent than those in the Mel spectrogram. The frequency range of the target signal in the ship-radiation noise was primarily concentrated in the low-frequency range of 0–1000 Hz, whereas the high-frequency band (e.g., 4000–8000 Hz) contained noise or other environmental sounds. The SC-Mel spectrogram uses the same number of Mel filters as the high-frequency broadband in the narrow low-frequency band; thus, the frequency resolution of the low-frequency spectrogram is higher, and more detailed feature information regarding the low-frequency band can be obtained.

5.3. Discussion

In this study, a UATR system based on the SC-Mel spectrogram and CFTANet model was proposed, and the effectiveness and superiority of the recognition system were verified on the DeepShip and ShipsEar datasets. Previous studies have primarily focused on the contribution of the SC-Mel spectrogram and multidomain attention mechanism to UATR. Simultaneously, the recognition system was analyzed in terms of feature complexity, model complexity, and performance differences for different categories. Frequency attention and CFTA blocks are lightweight and independent, and can be widely embedded in various network frameworks, thus providing more possibilities for future research and practical applications.

Although we verified our recognition system for open datasets in different sea areas, the discrepancy among underwater acoustic propagation channels in different sea areas affects the recognition performance. In UATR, we typically encounter situations in which the training

and testing samples are from different sea areas. When the sea areas are inconsistent, the recognition performance inevitably deteriorates. To improve the performance of the recognition system under mismatched sea conditions, a more robust recognition system should be established.

6. Conclusion

Herein, a UATR system combining an SC-Mel spectrogram and the CFTANet model was proposed. A sub-band Mel filter was used to extract the spectrogram features of multiple sub-bands, and a SC-Mel spectrogram was obtained by concatenating the Mel spectrograms of multiple sub-bands, which effectively yielded more detailed feature information from low-frequency bands. Subsequently, the frequency attention module and CFTA block module were proposed, and they were integrated into a simple residual network to realize the efficient learning and recognition of target features.

We evaluated the effectiveness of the proposed UATR system on the DeepShip and ShipsEar datasets. Compared with other deep-learning methods, the CFTANet model is more advantageous. In addition, because the parameters and FLOPs of the CFTANet model are lightweight, they are highly promising for practical applications, particularly on underwater unmanned platforms.

Ablation experiments were performed on the DeepShip dataset to evaluate the effectiveness of the SC-Mel spectrogram and attention modules. The experimental results demonstrated that the SC-Mel spectrogram, frequency attention module, and CFTA block facilitated the performance improvement of the proposed UATR system. In addition, the proposed UATR system effectively improved the prediction accuracy and generalization ability, although it presented a wide distribution of outliers. The reduction of outliers must be addressed in future studies.

CRediT authorship contribution statement

Shuang Yang: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft. **Anqi Jin:** Formal analysis, Investigation, Resources, Writing – original draft. **Xiangyang Zeng:** Funding acquisition, Supervision, Writing – review & editing. **Haitao Wang:** Writing – review & editing. **Xi Hong:** Resources. **Menghui Lei:** Resources.

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.

Data availability

The authors do not have permission to share data.

Acknowledgments

This research is supported by the National Natural Science Foundation of China under grant number 52271351.

References

Boashash, B., O'shea, P., 1990. A methodology for detection and classification of some underwater acoustic signals using time-frequency analysis techniques. IEEE Trans. Acoust. Speech Signal Process. 38 (11), 1829–1841.
Chen, Y., Xu, X., 2017. The research of underwater target recognition method based on deep learning. In: Proceedings of the 2017 IEEE International Conference on Signal Processing, Communications and Computing. ICSPCC), Xiamen, China, 22–25 October.
Chen, L., Luo, X., Zhou, H., 2024. A ship-radiated noise classification method based on domain knowledge embedding and attention mechanism. Eng. Appl. Artif. Intell. 127, 107320.
Choi, J., Choo, Y., Lee, K., 2019. Acoustic classification of surface and underwater vessels in the ocean using supervised machine learning. Sensors 19 (16), 3492.
Domingos, L.C.F., Santos, P.E., Skelton, P.S.M., et al., 2022. A survey of underwater acoustic data classification methods using deep learning for shoreline surveillance. Sensors 22 (6), 2181.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al., 2020. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale arxiv preprint arxiv:2010.11929.
Dzikowicz, B.R., Yoritomo, J.Y., Heddings, J.T., et al., 2023. Demonstration of spiral wavefront navigation on an unmanned underwater vehicle. IEEE J. Ocean. Eng. 48 (2), 297–306.
Feng, S., Zhu, X., 2022. A transformer-based deep learning network for underwater acoustic target recognition. Geosci. Rem. Sens. Lett. IEEE 19, 1–5.
Ferguson, E.L., Ramakrishnan, R., Williams, S.B., Jin, C.T., 2017. Convolutional neural networks for passive monitoring of a shallow water environment using a single sensor. In: Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March, pp. 2657–2661.
Ge, Q., Ruan, F., Qiao, B., Zhang, Q., Zuo, X., Dang, L., 2021. Side-scan sonar image classification based on style transfer and pre-trained convolutional neural networks. Electronics 10, 1823.
Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., et al., 2014. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Canada, Montreal.
Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning[M]. MIT Press, Cambridge, MA, USA, p. 800. https://doi.org/10.1007/s10710-017-9314-z. ISBN: 0262035618.
Guo, M.H., Xu, T.X., Liu, J.J., et al., 2022. Attention mechanisms in computer vision: a survey. Computational visual media 8 (3), 331–368.
Haykin, S., Van Veen, B., 2007. Signals and systems[M]. John Wiley & Sons.
He, K., Zhang, X., Ren, S., 2016. Deep residual learning for image recognition. In: Las Vegas: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR), pp. 770–778.
He, T., Zhang, Z., Zhang, H., 2019. Bag of tricks for image classification with convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 558–567. https://doi.org/10.1109/CVPR.2019.00065.
Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504–507. https://doi.org/10.1126/science.1127647.
Hong, F., Liu, C., Guo, L., et al., 2021. Underwater acoustic target recognition with a residual network and the optimized feature extraction method. Appl. Sci. 11 (4), 1442.
Hou, S., Jiao, D., Dong, B., et al., 2022. Underwater inspection of bridge substructures using sonar and deep convolutional network. Adv. Eng. Inf. 52, 101545.
Howard, A., Sandler, M., Chu, G., et al., 2019. Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314–1324.
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.
Irfan, M., Jiangbin, Z., Ali, S., et al., 2021. DeepShip: an underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification. Expert Syst. Appl. 183, 115270.
Jiang, Z., Zhao, C., Wang, H., 2022. Classification of underwater target based on S-ResNet and modified DCGAN models. Sensors 22 (6), 2293.
Jin, A., Zeng, X., 2023. A novel deep learning method for underwater target recognition based on res-dense convolutional neural network with attention mechanism. J. Mar. Sci. Eng. 11 (1), 69.
Khishe, M., 2022. Drw-ae: a deep recurrent-wavelet autoencoder for underwater target recognition. IEEE J. Ocean. Eng. 47 (4), 1083–1098.
Kim, K.I., Pak, M.I., Chon, B.P., Ri, C.H., 2021. A method for underwater acoustic signal classification using convolutional neural network combined with discrete wavelet transform. Int. J. Wavelets, Multiresolut. Inf. Process. 19, 2050092.
Kingma, D., Ba, J., 2015. Adam: A Method for Stochastic Optimization. ICLR.
Kipf, T.N., Welling, M., 2016. Semi-supervised Classification with Graph Convolutional Networks arxiv preprint arxiv:1609.02907.
Koening, W., 1949. A New Frequency Scala for Acoustic Measurements. Bell Lab Rec., pp. 299–301
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA.
Li, X., Wang, W., Hu, X., et al., 2019. Selective kernel networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 510–519.
Li, P., Wu, J., Wang, Y., et al., 2022. STM: spectrogram transformer model for underwater acoustic target recognition. J. Mar. Sci. Eng. 10 (10), 1428.
Lin, C., Cheng, Y., Wang, X., et al., 2023. Transformer-based dual-channel self-attention for UUV autonomous collision avoidance. IEEE Trans. Intell. Veh.
Liu, G., Sun, C., Yang, Y., 2008. Target feature extraction for passive sonar based on two cepstrums. In: Proceedings of the 2008 2nd International Conference on Bioinformatics and Biomedical Engineering, Shanghai, China, 16–18 May, pp. 539–542.
Liu, Z., Lin, Y., Cao, Y., et al., 2021a. Swin transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, pp. 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986.
Liu, F., Shen, T., Luo, Z., et al., 2021b. Underwater target recognition using convolutional recurrent neural networks with 3-D Mel-spectrogram and data augmentation. Appl. Acoust. 178, 107989.

Liu, Z., Mao, H., Wu, C.Y., et al., 2022. A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986.

Park, S.R., Lee, J., 2017. A fully convolutional neural network for speech enhancement. Interspeech 2017. In: 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden. https://doi.org/10.21437/Interspeech.2017-1465.

Ren, J., Xie, Y., Zhang, X., et al., 2022. UALF: a learnable front-end for intelligent underwater acoustic classification system. Ocean Eng. 264, 112394.

Santos-Domínguez, D., Torres-Guijarro, S., Cardenal-López, A., et al., 2016. ShipsEar: an underwater vessel noise database. Appl. Acoust. 113, 64–69.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition arxiv preprint arxiv:1409.1556.

Song, S., Lan, C., Xing, J., et al., 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. Proc. AAAI Conf. Artif. Intell. 31 (1).

Song, X., Wu, C., Stojanovic, V., et al., 2023. 1 bit encoding–decoding-based event-triggered fixed-time adaptive control for unmanned surface vehicle with guaranteed tracking performance. Control Eng. Pract. 135, 105513.

Teng, B., Zhao, H., 2020. Underwater target recognition methods based on the framework of deep learning: a survey. Int. J. Adv. Rob. Syst. 17 (6), 1729881420976307.

Tian, S.Z., Chen, D.B., Fu, Y., et al., 2023. Joint learning model for underwater acoustic target recognition. Knowl. Base Syst. 260, 110119.

Tijjani, A.S., Chemori, A., Creuze, V., 2022. A survey on tracking control of unmanned underwater vehicles: experiments-based approach. Annu. Rev. Control.

Vahidpour, V., Rastegarnia, A., Khalili, A., 2015. An automated approach to passive sonar classification using binary image features. J. Mar. Sci. Appl. 14, 327–333.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11).

Vaswani, A., Shazeer, N., Parmar, N., et al., 2017. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA.

Wang, G., Yang, Y., Wang, S., 2020a. Ocean thermal energy application technologies for unmanned underwater vehicles: a comprehensive review. Appl. Energy 278, 115752.

Wang, Q., Wu, B., Zhu, P., et al., 2020b. ECA-Net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534–11542.

Wang, B., Zhang, W., Zhu, Y., et al., 2023. An underwater acoustic target recognition method based on AMNet. Geosci. Rem. Sens. Lett. IEEE.

Wei, X., Gang-Hu, L.I., Wang, Z.Q., 2011. Underwater target recognition based on wavelet packet and principal component analysis. Comput. Simulat. 28, 8–290.

Woo, S., Park, J., Lee, J.Y., et al., 2018. Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19.

Xie, Y., Ren, J., Xu, J., 2022. Underwater-art: expanding information perspectives with text templates for underwater acoustic target recognition. J. Acoust. Soc. Am. 152 (5), 2641–2651.

Xu, K., Huang, H., Deng, P., et al., 2021. Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing. IEEE Transact. Neural Networks Learn. Syst. 33 (10), 5751–5765.

Xu, K., Xu, Q., You, K., et al., 2023. Self-supervised learning–based underwater acoustical signal classification via mask modeling. J. Acoust. Soc. Am. 154 (1), 5–15.

Yang, S., Xue, L., Hong, X., Zeng, X., 2023. A lightweight network model based on an attention mechanism for ship-radiated noise classification. J. Mar. Sci. Eng. 11 (2), 432.

Zhang, Y., Zeng, Q., 2023. MSLEFC: a low-frequency focused underwater acoustic signal classification and analysis system. Eng. Appl. Artif. Intell. 123, 106333.

Zhang, J., Ding, W., He, L., 2019a. Data augmentation and prior knowledge-based regularization for sound event localization and detection. In: DCASE 2019 Detection and Classification of Acoustic Scenes and Events 2019 Challenge.

Zhang, R., Li, J., Sun, H., et al., 2019b. Scan: self-and-collaborative attention network for video person re-identification. IEEE Trans. Image Process. 28 (10), 4870–4882.

Zhu, P., Zhang, Y., Huang, Y., et al., 2023a. Underwater acoustic target recognition based on spectrum component analysis of ship radiated noise. Appl. Acoust. 211, 109552.

Zhu, M., Zhang, X., Jiang, Y., et al., 2023b. Hybrid underwater acoustic signal multi-target recognition based on DenseNet-LSTM with attention mechanism. In: Chinese Intelligent Automation Conference. Springer Nature Singapore, Singapore, pp. 728–738.