frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 12 December 2023
DOI 10.3389/fmars.2023.1306229

OPEN ACCESS

EDITED BY
Xuebo Zhang,
Northwest Normal University, China

REVIEWED BY
Zhichao Lv,
Shandong University of Science and Technology, China
Irfan Hussain,
Khalifa University, United Arab Emirates
Zhiping Xu,
Jimei University, China

*CORRESPONDENCE
Jianxun Tang
alio@mails.guet.edu.cn

RECEIVED 03 October 2023
ACCEPTED 13 November 2023
PUBLISHED 12 December 2023

CITATION
Chen Z, Tang J, Qiu H and Chen M (2023) MGFGNet: an automatic underwater acoustic target recognition method based on the multi-gradient flow global feature enhancement network.
Front. Mar. Sci. 10:1306229.
doi: 10.3389/fmars.2023.1306229

MGFGNet: an automatic underwater acoustic target recognition method based on the multi-gradient flow global feature enhancement network

Zhe Chen[1,2], Jianxun Tang[3]*, Hongbin Qiu[1,2] and Mingsong Chen[1,3]

[1]School of Information and Communication, Guilin University of Electronic Technology, Guilin, Guangxi, China, [2]Cognitive Radio and Information Processing Key Laboratory Authorized by China's Ministry of Education Foundation, Guilin University of Electronic Technology, Guilin, Guangxi, China, [3]School of Ocean Engineering, Guilin University of Electronic Technology, Beihai, Guangxi, China

The recognition of underwater acoustic targets plays a crucial role in marine vessel monitoring. However, traditional underwater target recognition models suffer from limitations, including low recognition accuracy and slow prediction speed. To address these challenges, this article introduces a novel approach called the Multi-Gradient Flow Global Feature Enhancement Network (MGFGNet) for automatic recognition of underwater acoustic targets. Firstly, a new spectrogram feature fusion scheme is presented, effectively capturing both the physical and brain-inspired features of the acoustic signal. This fusion technique enhances the representation of underwater acoustic data, resulting in more accurate recognition results. Moreover, MGFGNet utilizes the multi-gradient flow network and incorporates a multi-dimensional feature enhancement technique to achieve fast and precise end-to-end recognition. Finally, a loss function is introduced to mitigate the influence of unbalanced data sets on model recognition performance using Taylor series. This further enhances model recognition performance. Experimental evaluations were conducted on the DeepShip dataset to assess the performance of our proposed method. The results demonstrate the superiority of MGFGNet, achieving a recognition rate of 99.1%, which significantly surpasses conventional methods. Furthermore, MGFGNet exhibits improved efficiency compared to the widely used ResNet18 model, reducing the parameter count by 51.28% and enhancing prediction speed by 33.9%. Additionally, we evaluated the generalization capability of our model using the ShipsEar dataset, where MGFGNet achieves a recognition rate of 99.5%, indicating its superior performance when applied to unbalanced data. The promising results obtained in this study highlight the potential of MGFGNet in practical applications.

KEYWORDS

underwater acoustic target recognition, underwater acoustic signal processing, feature enhancement, deep learning, feature fusion

1 Introduction

With the development of artificial intelligence, there is an increasing focus on utilizing AI-based methods to address research challenges in aquaculture. Fisheries and aquaculture constitute a global industry valued at $200 billion (Gladju et al., 2022). As this industry continues to expand, traditional processes involving essential technologies such as aquaculture environment monitoring, feeding, and fish behavior surveillance (Wu et al., 2022) incur significant costs. Hence, the urgent need arises to employ artificial intelligence technologies to enhance the economic, social, and environmental sustainability of the fish supply chain (Lim, 2022). AI-based aquaculture technologies primarily encompass environmental monitoring, intelligent feeding, biological behavior monitoring, and fishing vessel motion tracking (Setiyowati et al., 2022).

Environmental monitoring relies on water quality management systems to control the health of aquaculture water, preventing widespread diseases or issues such as slow growth in fish fry due to water quality problems (Hu et al., 2022). Koparan et al. (2018) developed an intelligent unmanned aerial vehicle to continuously monitor the water quality of a 1.1-hectare pond through intelligent sampling and analysis. Given that feed costs constitute over 60% of aquaculture expenses (Boyd et al., 2022), effective control of feed distribution is crucial. Lim and Whye, (2023) proposed a system that monitors fish behavior by detecting water wave vibrations caused by competitive feeding, thereby assessing fish hunger levels and significantly reducing feed consumption.

Biological behavior monitoring encompasses various aspects. Ahmed et al. (2022) and Darapaneni et al. (2022) employed computer vision and underwater optical imaging techniques, respectively, to obtain underwater images of fish activities for disease detection and prevention before widespread mortality. Fishing activities require strict control over timing and quantity globally. Bradley et al. (2019) and Kritzer, (2020) integrated automatic identification with artificial intelligence technology, utilizing underwater acoustic target recognition systems to track fishing vessel movements in real-time and predict their fishing activities, ensuring legitimacy.

In summary, due to the rapid development of computer vision technology effectively addressing the first three issues in aquaculture, our research focus shifts towards utilizing underwater acoustic target recognition technology for vessel motion monitoring.

Underwater acoustic target recognition involves collecting target radiated noise using hydrophones, analyzing and processing the data to discern target types (Ma et al., 2022). It holds significant importance in maritime vessel monitoring and underwater vehicle detection. Acoustic target recognition models typically consist of two modules: feature extraction and feature classification (Hong et al., 2021), and research in this field revolves around these modules.

Traditional methods of underwater acoustic target feature extraction can be categorized into signal physics-based and brain-like computing methods (Zhu et al., 2023). Signal physics-based methods rely on basic characteristics, temporal features, and non-Gaussian characteristics of underwater acoustic signals (Yao X. et al., 2023). This includes time-domain features like zero-crossing distribution, frequency-domain features like cepstral analysis (Zhu et al., 2022), and joint time-frequency domain features such as wavelet transforms (Han et al., 2022; Liu et al., 2022; Tian et al., 2023). Brain-like computing features for underwater acoustic signals include Mel-frequency cepstral coefficients (MFCC) simulating nonlinear processing of the human ear (Di et al., 2023) and Gammatone filtering simulating peripheral auditory processing (Zhou et al., 2022). Traditional classifier models include case-based reasoning (Ali et al., 2018) and perceptron neural networks (Linka and Kuhl, 2023). While traditional methods provide explicit directional analysis based on the physical meaning of underwater acoustic signals, they depend on prior knowledge and exhibit poor model generalization (Xiao et al., 2021).

Deep learning models, including Convolutional Neural Networks (CNN) (Yao Q. et al., 2023), provide new solutions for underwater acoustic target recognition (Jin and Zeng, 2023). Wang and Zeng (2015). demonstrated the feasibility of CNN models in underwater acoustic target recognition by testing them on three different measured acoustic targets. Studies have validated the applicability of deep learning in feature extraction. Huang et al. (2021) used autoassociative neural networks (AANN) to directly process mixed time-domain information of raw acoustic data without prior information, filtering ocean background noise, and obtaining effective spectral features of underwater acoustic targets. Additionally, research on deep learning-based classifiers is active. Li J. et al. (2022) designed AResNet to enhance feature extraction capability by increasing the width of the ResNet (He et al., 2016) residual network and incorporating channel attention mechanisms. Yang S. et al. (2023) developed LW-SEResNet10 to improve target recognition accuracy by reducing the number of ResNet residual structures and adding attention mechanisms. These classifiers operate similarly, performing feature extraction first and then inputting the features to obtain classification results.

Despite the advantages of existing deep learning-based underwater acoustic target recognition models in addressing some shortcomings of traditional methods, several challenges persist:

1. Existing models have independent feature extraction and classifiers (Zhufeng et al., 2022), failing to meet end-to-end underwater acoustic target recognition requirements.
2. Current feature extraction methods primarily use two-dimensional feature methods based on signal physics or brain-like computing features or their fusion methods (Li J. et al., 2022; Yang S. et al., 2023), overlooking the high-dimensional features of underwater acoustic data, resulting in insufficient representation capabilities of fused features.
3. Current classifiers mainly enhance feature extraction capabilities by stacking convolutional layers (Ji et al., 2023). However, due to the mixture of ocean environmental noise and partial information of underwater acoustic target features (Xu et al., 2019), standard convolutional operations tend to lose some effective features of underwater acoustic targets and erroneously retain ocean environmental noise (Li J. et al.,

2022), reducing the capability to extract effective features in underwater acoustic target recognition models. Thus, the model's parameter quantity and its recognition performance cannot achieve an effective balance, failing to meet the requirements of fast recognition speed and high accuracy in underwater acoustic target recognition.

4. As underwater acoustic data collection requires substantial financial and labor support, most existing publicly available underwater acoustic datasets exhibit imbalances in sample quantities across categories (Zhou et al., 2021). When training deep learning-based target recognition models, this can lead to overfitting phenomena (Li B. et al., 2022), suppressing model recognition performance.

To address these issues, we propose a novel underwater acoustic target automatic recognition network model based on a multi-gradient flow global feature enhancement network, referred to as MGFGNet.

Contributions of this work include:

1. Introducing a high-dimensional feature fusion method based on signal analysis and brain-like features.
2. Proposing a multi-gradient network to reduce model parameters and enhance feature extraction capabilities.
3. Presenting an adaptive feature fusion and enhancement module to enrich the physical, channel, and contextual information of pre-existing features.
4. Inventing a loss function, adding only three hyperparameters, and transforming the multi-classification task into multiple binary classification tasks, significantly improving the model's ability to suppress sample imbalances and recognition accuracy.

The following outlines the general structural framework of the remaining content in this article. Section 2 provides a detailed exposition of the Ship Radiated Noise Classification Method, known as MGFGNet. In Section 3, qualitative and quantitative experiments are conducted to compare MGFGNet with existing advanced underwater acoustic target recognition models, followed by an analysis of the experimental findings. Finally, Section 4 serves as the conclusion of this article.

2 Methods

This section primarily delineates MGFGNet. Section 2.1 provides an overview of its architectural framework. Sections 2.2 through 2.5 subsequently delve into its Feature Extraction and Fusion Module (FEFM), the Multi-gradient Flow Block with Attention (Multi-grad Block), the Context Augmentation and Fusion Module (CAFM), and the dynamic classification loss function known as Taylor-MCE Loss.

2.1 Proposed model

MGFGNet comprises two core modules: FEFM and the MGFGNet classifier. Figure 1 illustrates its detailed architecture.

FEFM utilizes various feature extraction algorithms based on signal analysis and brain-like features to extract multidimensional features from vessel radiated noise signals. Subsequently, multiple three-dimensional features are fused using the proposed feature fusion method to form high-dimensional fused features, which serve as inputs to the MGFGNet network.

The MGFGNet classifier primarily consists of the Multi-grad Block module and the CAFM module. The Multi-grad Block utilizes a multi-gradient flow network and residual modules to rapidly extract deep abstract features with different receptive fields from underwater acoustic target signals while reducing model parameters. Simultaneously, it leverages the multi-head self-attention mechanism (MHSA) (Han et al., 2021) to enhance the model's focus on foreground information, aiming to preserve the spatiotemporal characteristics of target line spectra in the acoustic energy spectrogram. This enhances the model's ability to extract effective information from sonar signals.

The CAFM module uses dilated convolutions with different dilation rates to adaptively fuse and enhance contextual information with a broad range of receptive fields, enriching the feature representation of physical, channel, and contextual information extracted by the preceding module. Finally, the Taylor-MCE Loss is employed to calculate prediction loss, addressing the issue of suppressing model recognition performance on imbalanced datasets. The Taylor-MCE Loss incorporates Taylor series (Gonzalez and Miikkulainen, 2021)

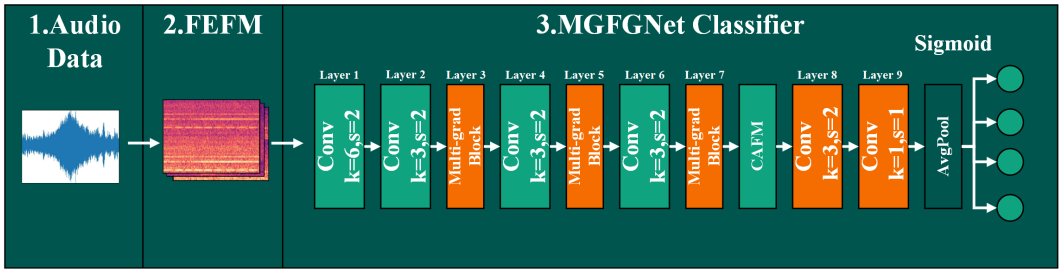

FIGURE 1
MGFGNet model architecture.

into binary cross-entropy loss (BCE) (Ruby and Yendapalli, 2020), including two components: one suppresses imbalances in sample components, and the other is a low-order term of the perturbation factor aimed at enhancing model recognition accuracy. Additionally, it transforms the multi-class classification task into multiple independent binary classification tasks.

2.2 Feature extraction and fusion module

Although deep learning-based feature extraction methods can capture more profound abstract features compared to traditional signal processing methods, they also come with a substantial increase in computational costs (Aggarwal et al., 2022). Vessel radiated noise primarily consists of mechanical noise, hydrodynamic noise, and propeller noise (Yang et al., 2019). Additionally, different feature extraction methods express distinct signal characteristics, and using multiple features for fusion can yield improved recognition results (Li Y. et al., 2022). Therefore, this paper, based on the generation mechanism of ship radiated noise, employs a fusion feature extraction method grounded in signal physical characteristics and brain-like features to represent underwater acoustic signals in multiple dimensions.

The fusion features in this paper mainly comprise energy-enhanced features from three types of features: CQT (Singh et al., 2022), delta MFCC (Nouhaila et al., 2022), and double delta MFCC (Nouhaila et al., 2022).

Firstly, since vessel radiated noise carries a significant amount of valid information in the low-frequency subband (Zhang et al., 2023), CQT provides better frequency resolution in the low-frequency subband (Mateo and Talavera, 2020). Hence, CQT is utilized as one of the feature extraction methods.

Secondly, MFCC, as a static feature, can not only eliminate ocean background noise but also effectively represent the spectral information of underwater acoustic targets. However, it lacks dynamic temporal signal features (Yang S. et al., 2023). To introduce temporal dynamic information, this paper performs local estimation of the differential operation along the time axis for the MFCC feature, obtaining delta MFCC and double-delta MFCC features. Both of these feature extraction methods are incorporated into the extraction of underwater acoustic target features.

Furthermore, as the single-channel feature information (graphically represented as a grayscale image) formed by these feature extraction methods can only express three-dimensional information of underwater sound, such as time, frequency, and energy domains, this paper expands the single-channel energy domain digital features of the above feature extraction methods into three-channel energy domain features using a color space representation. The detailed expansion method is described as follows.

Finally, this feature extraction and fusion module are embedded in the front end of the target recognition network, significantly reducing the computational burden of the classifier while achieving end-to-end target recognition.

Figure 2 illustrates our raised feature extraction method. Its process consists of four main parts:

1. In the first step, CQT features and MFCC features are extracted.

2.2.1 CQT extraction process

In the feature extraction process, the frame length is 2048 and the frame overlap is the portion between two frames of size 75% of the frame length, then using a Hanning window with a window size equal to the frame length for each frame signal.

The CQT transform of a finite length sequence x(n) is

$$X^{CQT}(k) = \frac{1}{N_k} \sum_{n=0}^{N_k-1} x(n) w_{N_k}(n) e^{-j\frac{2\pi Q}{N_k}n} \tag{1}$$

where $w_{N_k}(n)$ is a Hanning window of length N_k; Q is a constant factor in the CQT; k is the CQT frequency number, and the value of N_k is related to the value of k.

$$Q = \frac{1}{2^{\frac{1}{b}} - 1}, \tag{2}$$

where b is the number of frequency spectral lines, the

$$f_k = f_{\min} \times 2^{\frac{k}{b}}, k = 0, 1, \ldots, K-1, \tag{3}$$

$$N_k = \left\lceil Q \frac{f_s}{f_k} \right\rceil, k = 0, 1, \ldots, K-1, \tag{4}$$

where CQT information are stored in a matrix X^{CQT} (k,n), $f_{\min} = 1, f_s = 22050$. Since the sampling rate of the raw underwater acoustic data is 22050Hz for 5s, the shape of the CQT is 128×216.

2.2.2 MFCC extraction process

In the feature extraction process, the frame length and frame overlap are set to be the same as in the CQT extraction process. A Hanning window with a window size equal to the frame length is

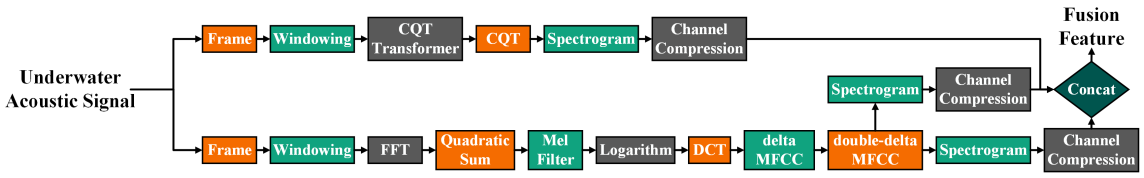

FIGURE 2
Feature extraction process for fused features.

then used for each frame. The short-time Fourier transform is then used to filter the noise and the sum of squares is used to obtain the power spectrum. Then 128 Mel filter banks were used to filter the information of each frame and logarithm was obtained to obtain Mel spectrum. Finally, MFCC was obtained by logarithm fitting the Mel spectrum to human hearing and discrete cosine transform (DCT). Since the sampling rate of the raw underwater acoustic data is a 5s signal at 22050Hz, the shape of the MFCC is 128×216.

2. The second step focuses on the extraction of delta MFCC and double-delta MFCC features by adding delta features and double-delta features to the MFCC features.

3. The third step focuses on transforming the above three features into spectrograms based on the size of 512, 12 and 0 for Hop length, bins per octave and tuning, respectively, with a preset image size of 3 × 640 × 480 for per image. Figure 3 illustrates the time-domain waveform diagram of radiated noise of a ship in the Deepship (Irfan et al., 2021) dataset and the spectrum diagram of CQT, delta MFCC and double-delta MFCC.

4. In the fourth step, the spectral graphs of CQT, delta MFCC and double-delta MFCC are fused respectively in channel dimension. The detailed fusion process is as follows.

2.2.3 Feature compression

as a result of the image pixel values reflected the important degree of information, so each spectrum diagram of three channel dimension values together to form a characteristic picture of 640 x 480.

2.2.4 Feature range mapping

Since the original pixel size range of each channel dimension is 0-255, the pixel value range of the feature map at this time is 0-765. To facilitate input for subsequent model calculations, map it to the range [0,255].

2.2.5 Feature fusion

Finally, the mapped features are in the order of CQT and two MFCC-derived features from top to bottom in the channel dimension to form a fusion feature with a shape of 3×640×480.

The formula of the fusion process above is expressed as:

$$T' = Map(concat(\sum_{j=0}^{2} T_j^{CQT}, \sum_{j=0}^{2} T_j^{deltaMFCC}, \sum_{j=0}^{2} T_j^{double-deltaMFCC})) \quad (5)$$

Where T_j^{CQT} represents the feature map of the J-th layer in the channel dimension of the CQT spectral graph feature matrix, $T_j^{deltaMFCC}$ and $T_j^{double-deltaMFCC}$ have the same meaning. Concat represents connecting matrices in the channel dimension. Map represents the range mapping of matrix data, a matrix T with data range of (x_{min},x_{max}), mapping its data to the range of (y_{min},y_{max}), and the mapped matrix is

$$Map = \frac{y_{max} - y_{min}}{x_{max} - x_{min}} \times (T - T_{x_{min}}) + T_{y_{min}}, \quad (6)$$

Where $T_{x_{min}}$, and $T_{y_{min}}$ both represent a constant matrix with the same latitude as T, and its content is the value represented by the Angle symbol.

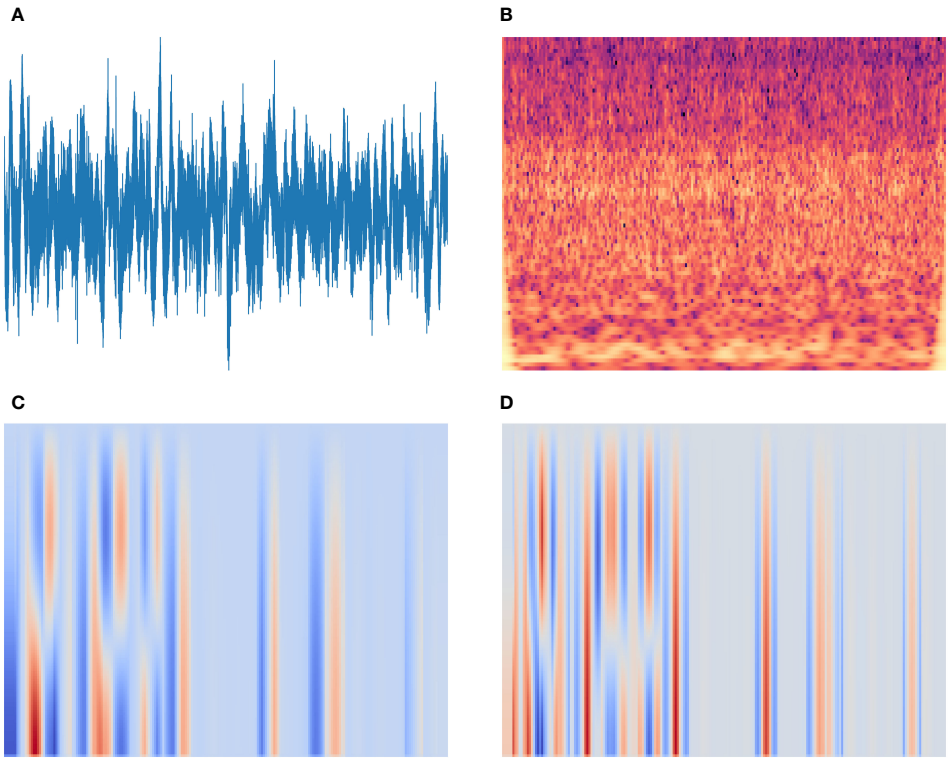

FIGURE 3
Spectrogram feature plots of radiated noise from a ship in the Deepship database: **(A)** Time-domain waveform; **(B)** CQT; **(C)** Delta MFCC; **(D)** Double-delta MFCC.

2.3 Multi-gradient flow with attention block

Existing models primarily increase the depth of the network to enhance feature extraction capabilities, but this leads to an increase in parameters while also losing a substantial amount of valuable information (Ji et al., 2023). In order to reduce the model's parameter count and enhance its ability to extract multidimensional features, this paper, inspired by the Cross Stage Partial Network (CSPNet) (Wang et al., 2020), which efficiently extracts effective feature information to alleviate model complexity, proposes the Multi-gradient flow bottleneck with attention Block (Multi-grad Block).

The Multi-grad Block concatenates multiple residual modules (Resblocks) to form a multi-gradient flow network. This structure enables the rapid acquisition of target information and gradient flow information from different receptive fields, accelerating the model's feature extraction speed while reducing the model's parameter count. Since traditional convolution operations lack sufficient discrimination between the spectra of multiple target lines and ocean background noise during feature extraction (Li J. et al., 2022), MHSA is introduced in the Resblock to increase the model's focus on targets rather than background noise or other irrelevant elements (Han et al., 2021). The detailed model structure is illustrated in Figure 4.

The detailed calculation process for the MHSA is as follows. MHSA is calculated as follows.

$$MH(A,B,C) = Concat(H_1, H_2, \ldots, H_h)W^O, \tag{7}$$

where A, B, C denote the query vector, key vector and value vector respectively, H_i illustrates the output of the i-th head, h is the number of headers, and W° is the output transformation matrix. The output of each header $head_i$ can be expressed as

$$head_i = Attention(QW_i^A, KW_i^B, VW_i^C) \tag{8}$$

where W_i^A, W_i^B, W_i^C are the A, B, and C transformation matrices for the i-th header, respectively, and Attention is a self-attentive calculation function with the following equation.

$$Attention(A_h, B_h, C_h) = soft\max\left(\frac{A_h B_h^T}{\sqrt{d_k}}\right)C_h, \tag{9}$$

Where d_k is the dimension of the key vector, softmax function mainly performs normalization, calculates the weight of each key vector, then multiplies the weight by the value vector, and finally performs weighted summation to get the attention output.

2.4 Context augmentation and fusion module

Due to the complex distribution of targets in the hybrid spectrogram generated by the feature extraction and fusion module of the original underwater acoustic signal, there are numerous small targets locally and larger, medium-sized targets globally (Wang B. et al., 2023). Using a single receptive field cannot fully capture the multidimensional features of the original signal, which reduces target classification accuracy (Wang Z. et al., 2023). To address these issues, this article introduces the Context Augmentation and Fusion Module (CAFM).

CAFM, as depicted in Figure 5, employs dilated convolution with varying rates to extract feature information from different receptive fields effectively (Gao et al., 2023). It enhances and fuses the multidimensional feature information obtained from the preceding gradient flow feature extraction module. Here's a breakdown of its structure:

1. The effective feature information obtained from the pre-gradient flow feature extraction module is rapidly processed using dilated convolution with three distinct rate values.

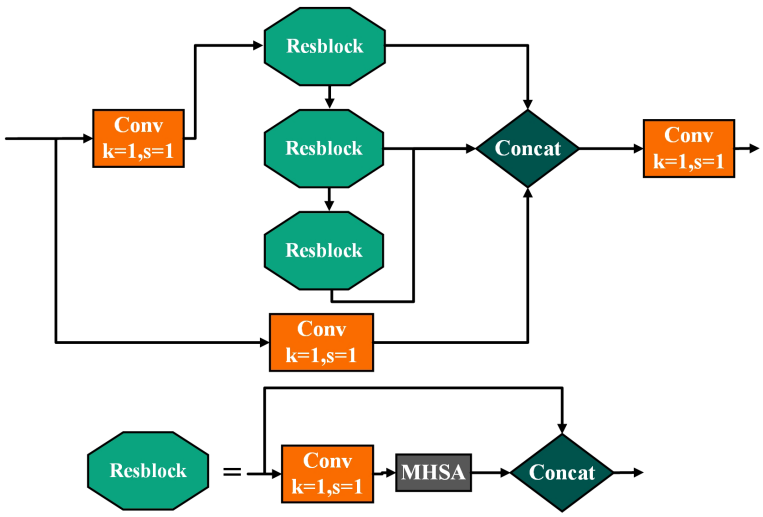

FIGURE 4
Model structure of Multi-grad Block.

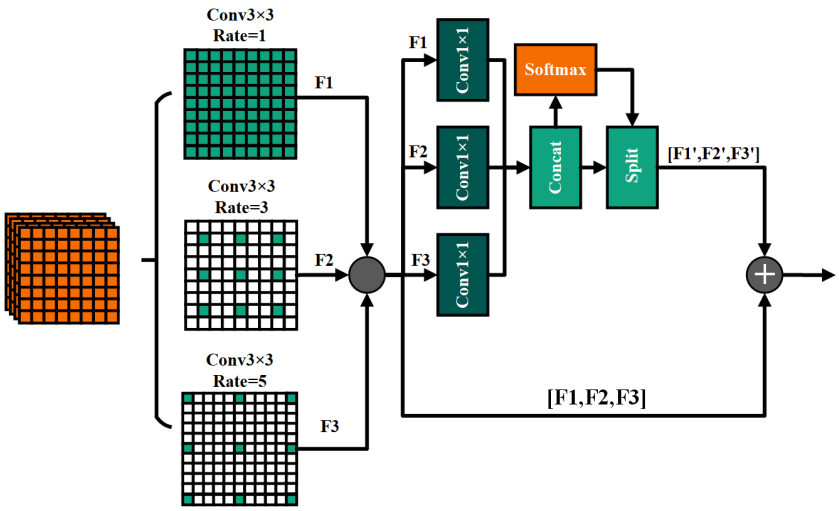

FIGURE 5
CAFM operation flows.

2. Target feature information is subsequently enhanced separately by the adaptive feature enhancement module and the cascade computing module.
3. The effective features derived from the adaptive feature enhancement module and the cascade computing module are then weighted and fused.

The former approach initially employs 1×1 convolution to compress and decrease the dimension of the pre-feature maps to single-channel feature maps. It then concatenates the feature maps in increasing rate order and calculates the weights for each channel using softmax. Finally, it enhances the channel dimension features through softmax-weighted multiplication.

The latter approach concatenates the feature maps obtained via expansion convolution at different rates to create a new feature map.

2.5 Taylor-MCE Loss

Existing mainstream classification loss functions primarily encompass the Cross-entropy Loss (CE) (Ho and Wookey, 2019) and its variations tailored for specific classification tasks. These adaptations include log loss (LL) (Lin et al., 2022) and BCE (Ruby and Yendapalli, 2020) for binary classification and focal loss (FL) (Lin et al., 2017) and categorical cross-entropy (CCE) (Ho and Wookey, 2019) for multi-class classification. However, the presence of a severe class imbalance among categories in underwater acoustic datasets poses a significant challenge (Zhou et al., 2021). Utilizing the aforementioned classification loss functions often leads to model overfitting (Leng et al., 2022), subsequently impacting recognition accuracy.

To tackle this challenge, this article introduces a novel loss function termed the Taylor-MCE Loss (Multiple Cross-Entropy Joint Loss Function based on Taylor Series). The Taylor-MCE Loss combines the polynomial terms derived from Taylor series expansion with BCE, FL, and low-order perturbation factors. It then transforms the multi-class classification task into a set of independent binary classification tasks, effectively resolving the issue of sample imbalance within the dataset and significantly enhancing the model's recognition performance.

The detailed design process is as follows:

1. Selection of the base loss function

Multi-class classification aims to calculate the likelihood of an object belonging to multiple categories, while binary classification seeks to identify whether an object is a specific category (e.g., discerning whether an object is a dog or not). Although these tasks may seem to differ only in the number of predicted categories, they have fundamental distinctions. In standard multi-class classification, CCE serves as the loss function, primarily relying on softmax to calculate the likelihood of an object belonging to multiple categories and selecting the category with the highest probability as the prediction. In contrast, binary classification tasks primarily employ BCE as the loss function, using sigmoid (output values between 0 and 1) to determine whether an object is closer to category 0 or category 1.

To select a more suitable base loss function and assess whether binary classification loss functions can be adapted for multi-class tasks, we conducted experiments in multi-class target recognition. The application of binary classification loss functions in multi-class tasks involved treating each category as an independent binary classification task. During our experiments, we made an intriguing observation: when inter-class sample sizes were balanced, CCE exhibited stable performance. However, in cases of sample imbalance, the use of BCE for multi-class tasks resulted in a significant improvement in accuracy compared to CCE (refer to Table 1 for details).

Treating each category as an individual binary classification task ensured that predictions for each category were mutually exclusive and independent (Ruby and Yendapalli, 2020), thereby addressing an issue. The problem when using CCE was that multiple categories

were predicted simultaneously [mutually exclusive but not independent (Ho and Wookey, 2019)]. Models employing CCE often favored categories with larger sample sizes, potentially overshadowing smaller categories during training (Leng et al., 2022). Additionally, BCE's core function was to enhance foreground weights while suppressing backgrounds (considering all other categories as backgrounds when predicting a single category) (Ruby and Yendapalli, 2020). This effectively balanced feature acquisition for different categories.

2. Exploring the relationship between loss functions using taylor series

The mutual constraints imposed by multiple categories can slow down model convergence. While combining multiple loss functions can enhance convergence speed and recognition accuracy (Li et al., 2019), it can also increase computational complexity. To minimize computational overhead while mitigating the impact of imbalanced datasets on the model, we drew inspiration from the Taylor series (Gonzalez and Miikkulainen, 2021) and explored the mathematical properties of BCE's polynomial form and loss functions designed to address imbalanced datasets. Our goal was to introduce minimal perturbation terms that retained the essential functionality of the loss function.

Since BCE can be represented as:

$$L_{BCE}(a,b) = -b_i \log(a) - (1-b_i)\log(1-a), \tag{10}$$

where $b_i \in \{0,1\}$ represents labels, and a represents predicted probabilities, and BCE is a special form of CCE, assuming

TABLE 1 Recognition accuracy, convergence time and number of parameters of CAFM at different locations of MGFGNet.

Model	Model convergence time (hours)	Parameters (M)	Accuracy
MGFGNet (-)	0.766	**5.576**	0.968
MGFGNet (+)	0.617	5.742	0.985
MGFGNet (1)	0.635	5.742	0.985
MGFGNet (2)	0.642	5.742	0.985
MGFGNet (3)	0.654	5.742	0.987
MGFGNet (4)	0.661	5.742	0.986
MGFGNet (5)	0.673	5.742	0.986
MGFGNet (6)	0.677	5.742	0.989
MGFGNet (7)	0.692	5.742	**0.991**
MGFGNet (8)	0.711	5.754	0.988

Bold font indicates the best-performing values within their respective columns.

$$a_t = \begin{cases} a, b = 1 \\ 1-a, otherwise, \end{cases} \tag{11}$$

CCE can be expressed as:

$$L_{CCE}(a,b) = -\log(a_t) \tag{12}$$

Applying Taylor series to CCE, the expression becomes:

$$L_{CCE} = -\log(a_t) = \sum_{i=1}^{\infty} \frac{1}{i}(1-a_t)^i \tag{13}$$

By observing the relationship between the Taylor expansion of CCE and FL, it is apparent that FL is equivalent to a horizontal shift (modulation factor) c of CCE. This is expressed as:

$$L_{FL} = -(1-a_t)^c \log(a_t) = (1-a_t)^c L_{CCE} \tag{14}$$

BCE is a special form of CCE; therefore, their physical properties are fundamentally consistent, differing mainly in the prediction process. To enhance the model's ability to address imbalanced datasets, we introduced an element to strengthen the suppression of imbalanced samples within the original BCE. This addition involved increasing the horizontal offset, resulting in the loss function:

$$\begin{aligned} L_{Taylor-MCE'} &= \alpha_1 L_{BCE} + \alpha_2 (1-a_t)^c L_{BCE} \\ &= [\alpha_1 + \alpha_2 (1-a_t)^c] L_{BCE} \end{aligned} \tag{15}$$

where $\alpha_1 + \alpha_2 = 1$ represents a scaling factor.

3. Analyzing the impact of gradient on loss functions

To enhance model recognition accuracy with minimal computational overhead, we compared the gradients of various loss functions and evaluated the influence of low-order and high-order terms on model recognition accuracy. The gradients of the aforementioned two loss functions (Eqs. 13 and 14) are expressed as follows:

$$-\frac{dL_{CCE}}{da_t} = \sum_{i=1}^{\infty}(1-a_t)^{i-1} = 1 + (1-a_t) + (1-a_t)^2 + \ldots \tag{16}$$

$$\begin{aligned} -\frac{dL_{FL}}{da_t} &= \sum_{i=1}^{\infty}(1+\frac{c}{i})(1-a_t)^{i+c-1} \\ &= (1+c)(1-a_t) + (1+\frac{c}{2})(1-a_t)^{1+c} + \ldots \end{aligned} \tag{17}$$

From the equations, it is evident that CCE possesses a fixed gradient term of 1. As i surpasses 1 and a_i approaches 1, the ith gradient tends towards zero. FL exhibits similar characteristics but introduces an additional perturbation factor (c). Consequently, the coefficients of high-order, low-order, and high-order terms collectively influence the outcomes of the loss function. The high-order parts primarily serve to suppress model errors, while the low-order components play a crucial role in fine-tuning the model to reach correct conclusions (Zhang et al., 2023). Therefore, we introduce a perturbation factor into the low-order term coefficients of CCE to enhance the model's recognition performance.

In summary, to mitigate the impact of sample imbalance on the model while minimizing the increase in parameter complexity, we

propose the Taylor-MCE Loss. The expression is as follows:

$$L_{Taylor-MCE} = \alpha_1 L_{BCE} + \alpha_2 (1 - a_t)^c L_{BCE} + \beta_1 (1 - a_t) \quad (18)$$

where $\beta_1 \in [-1,\infty)$ represents the perturbation factor.

3 Experimentation and analysis

To evaluate the performance of MGFGNet in a real underwater environment, we employ authentic underwater acoustic public datasets for both qualitative and quantitative comparisons. These comparisons involve MGFGNet and various versions with varying network depth and width of mainstream existing underwater acoustic target recognition models, including ResNet and EfficientNet (Mateo and Talavera, 2020).

3.1 Experimental dataset

3.1.1 Deepship

To assess the model's performance under ideal conditions, this study employed the Deepship dataset (Irfan et al., 2021), comprised of underwater acoustic data from vessels recorded by Northwestern Polytechnical University in the marine environment beneath the sea surface at depths ranging from 141 to 147 meters in the Georgia Strait Delta from 2016 to 2018. The data and time labels for this dataset were obtained by deploying sensors to locate vessel positions. Only singular vessel signals within a 2-kilometer range of the sonar device were considered, and recording ceased whenever a vessel exceeded this range. The dataset encompasses data from 265 vessels, including Cargo ships, Passenger Ships, Oil Tankers, and Tugs.

The data underwent preprocessing, with all WAV format audio files standardized to a 22,050Hz sample rate. Additionally, the underwater acoustic data were segmented into 5-second units, resulting in over 30,000 labeled sound samples. Recognizing that the model's recognition accuracy is proportional to the sample size of the training set, a significant number of samples were allocated for model training to mitigate the risk of overfitting. To prevent substantial fluctuations in the model's recognition accuracy due to a small sample size, a portion of the data was reserved for validating and testing the model's performance. Consequently, for optimal model parameter training, a large portion of the data was allocated to model training, with only a small amount used for validation and testing, following an 8:1:1 split ratio for the training, validation, and test sets. Table 2 provides details of the dataset division.

3.1.2 ShipsEar

So as to assess the model's capacity to adapt to diverse maritime environments, emphasizing its generalization capability, this study incorporated an additional authentic dataset of ship radiated noise collected in a real-world marine setting. The data collection took place along the Atlantic coast of Spain and encompasses recordings from 11 distinct ship types. These 11 ship categories were subsequently classified into four classes based on ship categorization, with the actual ocean background noise measurements, taken within these four categories, amalgamated to construct a five-class underwater acoustic dataset.

The dataset encompasses a total of 90 audio recordings, with individual recording durations varying from 15 seconds to 10 minutes. To ensure experimental precision, the "ShipsEar" dataset (Santos-Domínguez et al., 2016) underwent preprocessing identical to that applied to the "Deepship" dataset. A comprehensive class distribution is outlined in Table 3.

It is conspicuous that in the "Deepship" dataset, class proportions for classes 1-4 approximate ratios of 1:1.2:1.15:1.06. Conversely, the "ShipsEar" dataset presents imbalanced class proportions for classes 1-5, displaying a ratio of approximately 1.64:1.34:3.76:2.17:1. Consequently, when compared to the

TABLE 2 Details of the four categories in the Deepship dataset after pre-processing.

Class Number	Target	Total	Training set	Validation set	Testing set
1	Cargo Ship	7621	6097	762	762
2	Passenger Ship	9211	7369	921	921
3	Oil Tanker	8776	7022	877	877
4	Tug	8085	6467	809	809

TABLE 3 Details of the five categories of the ShipsEar dataset after pre-processing.

Class Number	Target	Total	Training set	Validation set	Testing set
1	Fishing boats, Trawlers, Mussel boats, Tugboats, Drafgers	369	296	37	36
2	Motoboats, Pilot boats, Sailboats	301	241	30	30
3	Passenger ferries	843	675	84	84
4	Ocean liner, Ro-Ro vessels	486	389	49	48
5	Background noise recordings	224	180	22	22

"Deepship" dataset, the "ShipsEar" dataset not only illustrates class imbalance but also contains significantly fewer samples, representing approximately 1/15 of the "Deepship" dataset. Such a dataset is highly susceptible to overfitting during the training process due to its limited sample size. Additionally, class imbalance can lead to notably reduced accuracy in recognizing classes with fewer samples.

3.1.3 SCTD

Synthetic Aperture Sonar (SAS) images (Huang and Yang, 2022; Wang and Huang, 2023; Yang, 2023; Zhang, 2023), known for their high resolution, significantly aid in target recognition in underwater acoustics. In order to assess the model's performance on a high-resolution underwater acoustic image dataset, this study introduces the SCTD dataset (Zhou et al., 2021). Since the original SCTD dataset is primarily designed for target detection tasks and its structure is not conducive to underwater acoustic target recognition models, certain modifications were implemented to adapt it to the classification task. Specifically, for the aircraft, human, and shipwreck categories within SCTD, the following steps were taken:

Firstly, multiple targets within a single image were individually cropped to ensure that each final image contains only one target, aligning it with the training sample format for target recognition.

Secondly, to augment the samples and balance the representation of each category, random cropping and flipping techniques were employed.

Finally, the dataset was partitioned into training, testing, and validation sets in an 8:1:1 ratio, as detailed in Table 4.

3.2 Hyperparameter setting

During the experimental process, the underwater acoustic target recognition model, MGFGNet, employed the Adaptive Moment Estimation optimizer (Adam) (Irfan et al., 2021) to mitigate sample noise interference. For this optimization process, the first-order momentum factor, second-order momentum factor, and Fuzz factor within Adam were configured at 0.9, 0.999, and 0.0000001, respectively. The initial learning rate was set to 0.001, with a weight decay coefficient of 0.0005, and a batch size of 32 was utilized. Finally, $a_1 = 0.5$, $a_2 = 0.5$, $c = 5$ and $\beta_1 = 5$ in Taylor-MCE Loss are set. The model was trained for 120 epochs (iterations) using the aforementioned parameters. Throughout the experimental process described below, unless otherwise specified, the experiment parameters mentioned above were consistently applied.

3.3 Experimental environment and performance indicators

The experiments were conducted on the PyTorch platform, running on the Windows 10. The hardware setup employed for these experiments is detailed in Table 5. To mitigate the potential influence of experimental variability, a systematic approach was taken. It involved the training and testing of various models, both qualitatively and quantitatively. Subsequently, a comparative analysis of algorithmic performance was performed.

Given that Accuracy can reflect the model's recognition capability across multiple classes, while Precision and Recall can indicate the overall classification performance of the model, these three evaluation criteria are employed to assess different models. Their formulas are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{19}$$

$$precision = \frac{TP}{TP + FP}, \tag{20}$$

$$recall = \frac{TP}{TP + FN}, \tag{21}$$

where TP represents instances that were originally true positive samples and were correctly predicted as positive samples by the underwater acoustic target recognition model. TN corresponds to instances that were originally true negative samples and were accurately predicted as negative samples by the model. FP signifies instances that were originally true negative samples but were erroneously classified as positive samples by the underwater acoustic target recognition model. FN stands for instances that were originally true positive samples but were incorrectly predicted as negative samples by the model.

3.4 Ablation experiments

3.4.1 Feature ablation experiments

To confirm the representational capabilities of the feature extraction approach raised in this study for original underwater acoustic signals, Table 6 presents an extensive comparison of diverse characteristics abstraction approachs on the Deepship. This comparison encompasses the original two-dimensional features, their corresponding three-dimensional counterparts, and the three-dimensional feature fusion approach introduced in Section 2.2 within the MGFGNet model. Notably, the recognition

TABLE 4 Details of the three categories of the SCTD dataset after pre-processing.

Class Number	Target	Total	Training set	Validation set	Testing set
1	Aircraft	575	459	58	58
2	Human	546	436	55	55
3	Shipwreck	488	390	49	49

TABLE 5 Details of the hardware environment for the experiment.

Hardware name	Parameters	Number
CPU	Intel Xeon Sliver 4310	2
GPU	NVIDIA Tesla A100 80G	1
RAM	SAMSUNG RECC DDR4 32GB	8

accuracy of spectral features for each feature extraction method surpasses that of the original two-dimensional features. Delta MFCC, owing to its ability to capture temporal correlations of MFCC, exhibits higher experimental accuracy than MFCC features, albeit with a modest 0.2% increase. Similarly, double-delta MFCC records a mere 0.4% improvement over delta MFCC since it primarily focuses on local estimations along the time axis for the differential operations of MFCC. CQT features, reflecting the frequency distribution patterns of underwater acoustic targets, outperform MFCC and its derivative features in terms of classification accuracy, thereby validating the superiority of CQT features over mel-spectrogram features (Domingos et al., 2022) in underwater acoustic target recognition. The horizontal comparison of spectral feature extraction methods among various feature extraction techniques exhibits similar characteristics as mentioned above.

It's worth mentioning that the overall accuracy of the fusion feature approach proposed in Section 2.2 surpasses that of other feature extraction techniques, achieving 99.1%. This represents a substantial increase of 3.9%, 3.7%, and 3.3% over the spectral features of MFCC, delta MFCC, and double-delta MFCC, respectively. Additionally, it outperforms CQT's spectral features by 1.6%. Moreover, there is a substantial increase in recognition accuracy across all categories compared to the spectral feature extraction methods of the remaining four features, thus validating the superiority of the fusion approach based on signal processing and brain-like features proposed in this study.

So as to provide a clearer illustration of the computational cost and efficiency of the feature extraction and fusion method introduced in this paper, we conducted additional experiments to assess the performance metrics of various feature extraction techniques. The testing dataset comprised 10 sets, each containing 10 noise data samples, and the experimental results represent the average of these 10 sets. Detailed experimental data is displayed in Table 6.

It is evident that the execution time for each feature extraction method's feature mapping technique increased by only approximately 0.0003 in comparison to the original method, with a memory consumption increment of around 20 MiB. Concurrently, the execution time of the feature extraction and fusion method proposed in this paper, which integrates three original feature components, remains within the same order of magnitude as their individual runtimes, indicating minimal additional time overhead.

Moreover, the memory consumption of the proposed method in this paper remains approximately at 350 MiB, aligning with the memory usage of all other feature extraction methods. This reaffirms the superiority and efficiency of the proposed method.

3.4.2 CAFM ablation experiment

So as to comprehensively evaluate the computational cost, convergence time, and performance of the feature extraction and fusion method presented in this paper at various positions within MGFGNet, a series of experiments were conducted. The experimental results on the Deepship dataset are presented in Table 1. In the model parameter nomenclature, the suffix indicates the layer within the target recognition model as depicted in Figure 1. For instance, "MGFGNet (1)" signifies the placement of the CAFM module after Layer 1 of MGFGNet, "-" indicates the absence of the CAFM module, and "+" denotes its placement at the beginning of MGFGNet, as illustrated in Figure 1.

Firstly, the integration of the CAFM module results in a modest 2.9% increase in model parameters compared to the original model. However, it significantly expedites the model's convergence speed. Furthermore, the convergence speed varies when the CAFM module is positioned at different locations within the model, and

TABLE 6 Recognition Accuracy of MGFG model on Deepship dataset using different features and the memory consumption and efficiency of each feature extraction method.

Feature	Cargo	Passenger Ship	Tanker	Tug	all	Time consumption (s)	Memory used (MiB)
MFCC	0.542	0.671	0.670	0.843	0.683	0.00033257	348.960938
MFCC Spec	0.930	0.950	0.951	0.975	0.952	0.00060603	368.828125
delta MFCC	0.629	0.655	0.623	0.849	0.687	0.00134368	352.488281
delta MFCC Spec	0.946	0.949	0.944	0.977	0.954	0.00166959	373.804688
double-delta MFCC	0.606	0.681	0.681	0.794	0.691	0.00137352	351.417969
double-delta MFCC Spec	0.946	0.957	0.957	0.97	0.958	0.00169537	372.640625
CQT	0.765	0.767	0.771	0.865	0.791	0.00353861	352.429688
CQT Spec	0.973	0.973	0.973	0.984	0.975	0.00356747	372.406250
Fusion Feature	0.929	0.929	0.977	0.993	0.957	0.00381105	357.812500
Fusion Feature of Spec	**0.984**	**0.985**	**0.994**	**1**	**0.991**	0.00384105	376.367188

Bold font indicates the best-performing values within their respective columns.

the speed is directly proportional to the sequence of the CAFM within the model. This is primarily due to the enhanced discriminability between background and target foreground in the feature maps when this module is applied, resulting in accelerated model convergence speed. Notably, after introducing the CAFM, the convergence time consistently remains between 0.6 and 0.7 hours, confirming the model's stability. In this experiment, convergence is defined as the point at which the loss remains unchanged in the thousandths place for three consecutive iterations.

Additionally, the incorporation of the CAFM module leads to a minimum 1.7% enhancement in recognition accuracy within MGFGNet, validating the CAFM module's capacity to boost model recognition accuracy through feature fusion and enhancement. However, the placement of the CAFM module also exerts an impact on recognition accuracy. For example, when the CAFM module is positioned at the head of MGFGNet and after Layer 1-2, the model exhibits improved recognition accuracy due to the fusion of multiscale acoustic target information and enhanced channel features. However, when the CAFM is placed at Layer 1, it leads to a rapid extraction of raw input features through a large convolutional kernel (kernel size of 6), resulting in the loss of significant valuable features and, consequently, inhibiting recognition accuracy. Furthermore, there is no subsequent feature enhancement in the feature extraction process, causing lower recognition accuracy compared to when the CAFM is placed after Layer 3-6.

Conversely, placing the CAFM module after Layer 3-6 introduces the Multi-gradient Block in front of the CAFM module, enriching the fused and enhanced features with a substantial amount of multi-gradient flow contextual information compared to the original information. This, in turn, enhances target feature information, leading to improved recognition accuracy. The highest recognition accuracy is achieved when the CAFM module is placed after Layer 7, as the model has undergone all the Multi-gradient Blocks by this stage, resulting in feature maps rich in multi-gradient flow, physical features, and numerous feature details. When the CAFM module is employed for feature fusion and enhancement at this stage, it effectively increases the importance of target information, thereby enhancing recognition accuracy.

However, due to the feature enhancement process preserving a substantial amount of suppressed background features, direct utilization of these feature maps for predictions can compromise experimental accuracy (Hu et al., 2018; Hou et al., 2021). Therefore, after employing the feature enhancement module, it is necessary to conduct further feature extraction on the enhanced feature maps using convolutional or feature extraction modules. This step helps discard numerous non-target features. For instance, attention mechanisms (AM) (Yang S. et al., 2023) and channel attention modules (CAM) (Li J. et al., 2022) both serve as feature enhancement modules. Ablation experiments have demonstrated that utilizing feature extraction or convolutional modules after feature enhancement enhances model recognition accuracy (Li J. et al., 2022; Yang S. et al., 2023). This substantiates why placing CAFM after Layer 7 results in higher recognition accuracy compared to after Layer 8.

3.4.3 Classification loss function ablation experiments

To assess the impact of the Taylor-MCE Loss on MGFGNet, this paper compared the recognition results of MGFGNet with various loss versions, including BCE, CCE, FL, and Taylor-MCE Loss, utilizing the Deepship dataset. The numbers 1, 2, and 3 following Taylor-MCE represent $\alpha_1 L_{BCE}$, $\alpha_2(1-a_t)^c L_{BCE}$, and $\beta_1(1-a_t)$, respectively. Notably, α_1 and α_2 have real values only when coexisting; otherwise, both are set to 1. A comprehensive summary of the experimental results is presented in Table 7.

Firstly, it is evident that Taylor-MCE Loss outperforms CCE, FL, and BCE in terms of recognition accuracy, demonstrating improvements of 2.4%, 2.2%, and 1.9%, respectively. The recognition accuracy of CCE and FL is quite similar. FL is derived from CCE through lateral shifting, aimed at mitigating the issue of sample imbalance. However, within the context of the Deepship dataset, where various classes exhibit a good balance, its effectiveness in addressing sample imbalance is reduced, resulting in a modest improvement of 0.2% compared to CCE. BCE, serving as a special form of CCE for binary classification, achieves a recognition accuracy improvement of 0.5%. This is primarily because BCE transforms multi-class classification into multiple binary classification tasks, where the predictions for each class are mutually exclusive and independent. This approach addresses a problem present in CCE where multiple classes are predicted simultaneously, leading the model to favor classes with larger sample sizes. This imbalance gradually drowns out smaller classes during training, providing a key rationale for choosing BCE as the base loss function for Taylor-MCE Loss. Taylor-MCE (1,3), inclusive of low-order perturbation terms ($\beta_1(1-a_t)$), contributes to the model's improved recognition accuracy, resulting in a significant advantage over Taylor-MCE (1,2), which only encompasses the component for addressing imbalance ($\alpha_2(1-a_t)^c L_{BCE}$). This finding reinforces the conclusion that low-order terms enhance recognition accuracy (Zhang et al., 2023). Taylor-MCE (2,3) achieves similar recognition accuracy to (1,3), primarily due to the relatively balanced distribution of class samples in the Deepship dataset, rendering the influence of (2,3) insufficient to significantly alter recognition accuracy.

TABLE 7 Recognition accuracy of MGFGNet with different classification loss functions on Deepship and ShipsEar.

Loss Function	Accuracy (Deepship)	Accuracy (ShipsEar)
CCE	0.967	0.937
FL	0.969	0.953
BCE	0.972	0.959
Taylor-MCE (1,2)	0.977	0.982
Taylor-MCE (1,3)	0.985	0.972
Taylor-MCE (2,3)	0.983	0.976
Taylor-MCE	**0.991**	**0.995**

Bold font indicates the best-performing values within their respective columns.

As the Deepship dataset comprises a substantial number of samples with a relatively balanced class distribution, it does not effectively validate the loss function's ability to suppress small samples and enhance recognition accuracy in unbalanced datasets. To further confirm the adaptability of Taylor-MCE Loss to imbalanced, small-sample underwater sound datasets, we conducted experiments using different classification loss functions on the ShipsEar dataset, characterized by class imbalance and limited sample sizes. A detailed overview of the experimental results is provided in Table 7. The unique sample characteristics of ShipsEar, featuring fewer samples and imbalanced class distributions, result in notable differences in model recognition accuracy when employing various loss functions. CCE, due to its lack of optimization for class imbalance, exhibits lower recognition accuracy compared to other loss functions. Both FL and BCE, which address class imbalance using different approaches (FL introduces horizontal shifting on top of CCE, while BCE transforms multi-class into multiple binary classification tasks to mitigate imbalance), yield similar and significantly improved recognition accuracy compared to CCE. In contrast, the results of the various versions of Taylor-MCE Loss are entirely opposite to those observed in the Deepship dataset. Given that the ShipsEar dataset has fewer samples and imbalanced class distributions, it necessitates substantial suppression of the imbalance component. When utilizing only the low-order perturbation term to enhance recognition accuracy, specifically Taylor-MCE (1,2), its recognition accuracy surpasses BCE by 2.3%, compared to the mere 0.9% improvement. Taylor-MCE effectively balances recognition accuracy and mitigates model overfitting attributed to class imbalance during training, ultimately yielding a recognition accuracy of 99.5%. This figure is 5.8%, 4.2%, and 3.6% higher than CCE, FL, and BCE, respectively.

These experiments affirm the adaptability of Taylor-MCE Loss to small-sample, imbalanced datasets, significantly enhancing model recognition accuracy.

3.5 Performance analysis

In this section, we compare the performance of MGFGNet with existing state-of-the-art target recognition models [such as ResNet (He et al., 2016), EfficientNet (Koonce, 2021), DenseNet (Iandola et al., 2014), etc.] under the same experimental conditions, examining various aspects.

3.5.1 Model identification accuracy and parameter analysis

To validate whether MGFGNet outperforms existing mainstream target recognition models, we trained and validated MGFGNet and other mainstream models under the experimental conditions described in Sections 3.2 and 3.3. The parameters of each model and their experimental accuracy on the Deepship test set are presented in Table 8. It is noteworthy that, to reduce the training time for various models, we modified the training epochs for all models on the Deepship dataset to 90. This decision is supported by the observation, as depicted in Figure 6, that MGFGNet exhibits a tendency toward convergence in loss before 90 epochs, with the optimal model being formed around the 71st epoch.

TABLE 8 Details of the number of parameters and the recognition accuracy on the Deepship dataset for various models.

Model	MFCC	delta MFCC	double-delta MFCC	CQT	Fusion Feature	Parameters (M)
ResNet18 (He et al., 2016)	0.939	0.942	0.947	0.963	0.970	11.7
ResNet34 (He et al., 2016)	0.929	0.937	0.942	0.966	0.971	21.8
ResNet50 (He et al., 2016)	0.921	0.933	0.937	0.952	0.965	25.6
ResNet101 (He et al., 2016)	0.913	0.931	0.937	0.947	0.953	44.5
EfficientNet_b0 (Koonce, 2021)	0.931	0.941	0.949	0.964	0.971	**5.3**
EfficientNet_b1 (Koonce, 2021)	0.930	0.938	0.945	0.967	0.968	7.8
EfficientNet_b2 (Koonce, 2021)	0.917	0.935	0.939	0.955	0.959	9.1
EfficientNet_b3 (Koonce, 2021)	0.915	0.931	0.934	0.945	0.951	12.2
DenseNet (Iandola et al., 2014)	0.866	0.871	0.878	0.913	0.931	1.1
CSPDenseNet (Wang et al., 2020)	0.889	0.896	0.913	0.937	0.951	0.9
CSPResNet18 (Wang et al., 2020)	0.938	0.945	0.953	0.966	0.973	5.6
MobileNetV1 (Howard et al., 2017)	0.759	0.787	0.793	0.822	0.841	3.2
MobileNetV2 (Sandler et al., 2018)	0.876	0.888	0.893	0.907	0.921	2.2
MobileNetV3-S (Howard et al., 2019)	0.732	0.747	0.752	0.773	0.806	1.5
MobileNetV3-L (Howard et al., 2019)	0.820	0.822	0.829	0.877	0.894	4.2
ViT (Dosovitskiy et al., 2020)	0.871	0.875	0.879	0.882	0.889	86.6
MGFGNet	**0.952**	**0.954**	**0.958**	**0.975**	**0.991**	5.7

Bold font indicates the best-performing values within their respective columns.

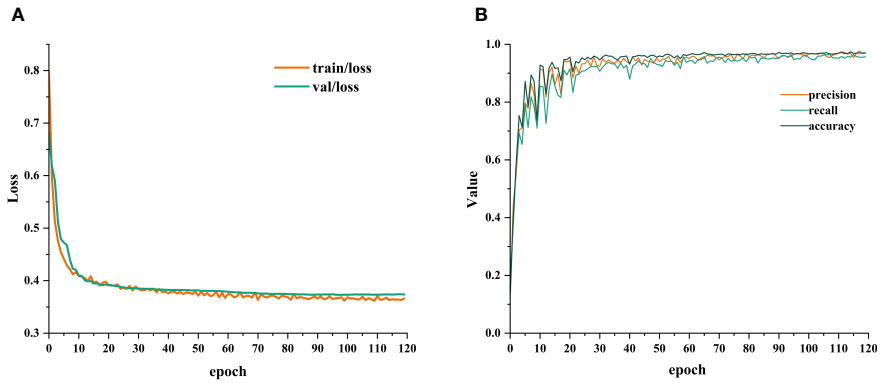

FIGURE 6
Variation of parameters during MGFGNet training: **(A)** Loss variation plot; **(B)** Precision, recall, and accuracy variation plot.

Clearly, MGFGNet demonstrates experiment accuracy superior to existing mainstream target recognition models across various feature input scenarios. This validates the robust feature extraction capability of MGFGNet in diverse experimental environments.

Furthermore, CSPNet (Wang et al., 2020) not only reduces model parameters but also effectively promotes the model's feature extraction capability. For instance, testing the original versions of ResNet18 (He et al., 2016) and DenseNet (Iandola et al., 2014), along with their versions incorporating CSPNet, reveals a noticeable reduction in parameters and an improvement in model performance under various feature inputs. MGFGNet, based on the CSPNet philosophy with the multi-gradient flow module as a primary component, successfully achieves an effective balance between recognition accuracy and parameter count.

The size of the model's parameters also influences recognition accuracy. Models with either too many or too few parameters yield suboptimal experimental accuracy. For example, ViT (Dosovitskiy et al., 2020) has significantly more parameters than other models, yet its recognition rate is lower than most models. In contrast, the MobileNet (Howard et al., 2017; Sandler et al., 2018; Howard et al., 2019) series, characterized by smaller parameter counts as lightweight models, generally exhibits lower recognition accuracy compared to other models. However, DenseNet and CSPDenseNet (Wang et al., 2020), despite having fewer parameters, achieve high recognition accuracy. This is mainly attributed to the dense connectivity in DenseNet, ensuring low-dimensional feature information and a stronger gradient flow (Iandola et al., 2014).

Within the same model, variations in recognition accuracy due to changes in model depth show a negative correlation with the number of parameters. As the number of parameters increases from ResNet18 to ResNet101 in the ResNet model, the recognition accuracy gradually decreases. A similar trend is observed in the EfficientNet (Koonce, 2021) model. However, different network models do not exhibit this phenomenon due to diverse feature extraction methods. For example, ResNet18 and EfficientNet_b3 have similar parameter counts, but ResNet18 outperforms EfficientNet_b2 in recognition accuracy across various feature extraction methods. Different versions of MobileNet do not show this phenomenon because new feature extraction or enhancement modules are introduced in each version.

3.5.2 Analysis of computational load, training time, and prediction time

To assess the training and inference efficiency of MGFGNet, this section analyzes the time consumption for training and inference of MGFGNet and its comparative models. Detailed comparative results are provided in Table 9. It is noteworthy that, to reduce the training time of the models, we maintained consistency with Section 3.5.1 and modified the training epochs for all models on the Deepship dataset to 90. The convergence definition for this experiment is when the value of the thousandth loss percentile remains unchanged for three consecutive times during the training process, indicating model convergence.

During the experiments, variations in the training and prediction times of models were observed in different operating environments. To ensure the accuracy of experimental data, each model, during its runtime, had the host free of other GPU-intensive deep learning tasks, preventing interference with the experimental results. Additionally, to mitigate random factors, all experimental data are the averages of results obtained from five repeated experiments. Floating Point Operations per Second (FLOPs) are used to measure the computational complexity of the model. Training time refers to the total time for model training and validation. Inference time denotes the total time required for predicting 3369 individual samples from the validation set of Deepship.

Notably, MGFGNet exhibits superior inference time compared to all comparative models, especially the lightweight MobileNet series commonly used in embedded systems, demonstrating its practical utility. This is primarily attributed to the inference time reduction effect of CSPNet (Wang et al., 2020). For instance, the inclusion of CSPNet in ResNet and DenseNet also significantly reduces inference time. MGFGNet, incorporating the CSPNet philosophy through the Multi-grad Block, outperforms EfficientNet_b0 in prediction time, despite having similar parameter counts and FLOPs values.

Furthermore, MGFGNet achieves convergence in the fewest epochs, indicating a faster convergence rate. This is mainly due to Taylor-Loss suppressing the rate of model variation under different numbers of input categories, thereby accelerating model

TABLE 9 Floating-point computation vs. training and predicting time.

Model	FLOPs@224(B)	Training Time (hours)	Epochs at convergence	Predicting Time(s)	Support
ResNet18	3.7	**1.075**	52	56	3369
ResNet34	7.4	1.391	68	59	3369
ResNet50	8.5	1.868	75	67	3369
ResNet101	15.9	4.975	98	104	3369
EfficientNet_b0	1.0	3.423	83	91	3369
EfficientNet_b0	1.5	5.121	91	102	3369
EfficientNet_b0	1.7	5.368	97	104	3369
EfficientNet_b0	2.4	5.753	102	107	3369
DenseNet	1.6	4.792	62	135	3369
CSPDenseNet	1.4	4.872	57	119	3369
CSPResNet18	0.5	1.397	48	42	3369
MobileNetV1	0.6	4.693	96	66	3369
MobileNetV2	0.4	4.401	72	63	3369
MobileNetV3-S	**0.1**	2.661	65	55	3369
MobileNetV3-L	0.2	3.100	78	75	3369
ViT	17.6	7.295	107	139	3369
MGFGNet	0.7	1.779	**41**	**37**	3369

Bold font indicates the best-performing values within their respective columns.

convergence. It is observed that within models of the same architecture, parameters and convergence epochs exhibit a positive correlation, as seen in ResNet and EfficientNet series. While MobileNet has a smaller parameter count, its frequent occurrence of gradient vanishing during training, mainly due to the use of depthwise separable convolution, leads to extensive time spent correcting and updating the model, resulting in an increased number of convergence epochs.

Additionally, while MGFGNet's training time is lower than that of most target recognition models, it still exceeds that of ResNet18, ResNet34, and CSPResNet18. This is mainly because the Multi-grad Block module, based on the CSPNet philosophy, invented in MGFGNet, reduces the number of parameters but increases the computational workload for backward gradient updates (Wang et al., 2020), thus extending the model's training time. The increased training time for CSPDenseNet and ResNet18 with CSPNet also validates this characteristic. However, since practical applications primarily require low prediction times for rapid target recognition, this drawback has minimal impact in real-world scenarios.

Finally, upon contrasting Tables 8 and 9, it becomes evident that there is no inherent correlation between the training time, model parameters, and FLOPs for the models. For instance, when compared to ResNet18, DenseNet, CSPDenseNet, and the MobileNet series all exhibit smaller parameter counts and FLOPs. However, these models demonstrate longer training times than ResNet18. A similar experimental pattern is observed between the EfficientNet and ResNet series.

3.5.3 Model stability validation

Figure 6 presents the loss variation chart as well as the precision, recall, and accuracy variation charts on the validation set during the same 120-epoch training process on the Deepship dataset.

From the loss curve, it can be observed that the network gradually stabilizes after the 60th epoch. By examining the changes in precision, recall, and accuracy on the validation set during the training process, with smooth variations and the absence of overfitting, it can be concluded that the proposed underwater acoustic target recognition model, MGFGNet, demonstrates stability.

3.5.4 Robustness analysis of models.

To assess the robustness of MGFGNet, i.e., the extent to which the model is affected by small variations in the data, we utilized spectrogram features of MFCC and its derived characteristics. Due to the high similarity between spectrograms of MFCC and its derived features (Yang S. et al., 2023), this study thoroughly compared the dependency of various models on different input conditions based on spectrogram features of MFCC and its derivatives, as illustrated in Figure 7.

It is evident that MGFGNet exhibits a relatively small disparity in experimental accuracy when considering spectrogram features of MFCC and its derived characteristics. However, there is still some improvement, indicating that MGFGNet can capture minor variations in the derived features of MFCC without causing significant predictive differences due to slight changes. This validates the robustness of the model.

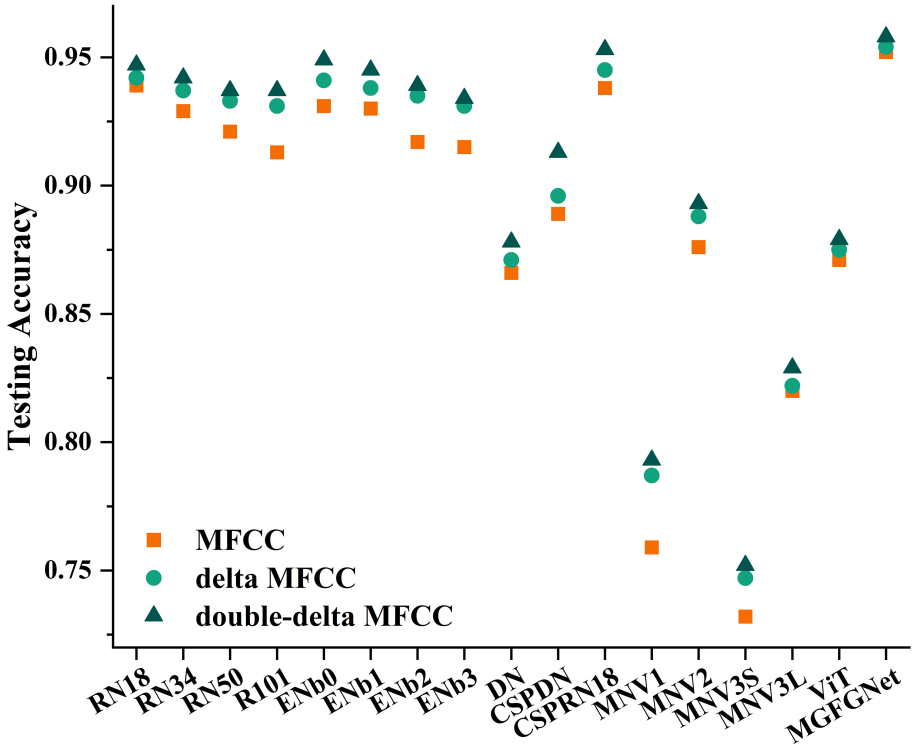

FIGURE 7
Recognition accuracy of multiple models under MFCC, delta MFCC and double-delta MFCC features.

ResNet18, DenseNet, and ViT models demonstrate comparable recognition accuracy across these three different feature extraction methods. In contrast, other models exhibit significant variations in model responses under these three feature extraction methods, indicating their reliance on features with high separability.

3.5.5 Generalizability analysis of the model

Due to varying predictive capabilities of models across different variable domains (distinct real underwater acoustic datasets), it is essential to assess the generalization performance of MGFGNet on additional datasets. This study conducts experiments placing each model under the experimental conditions defined in Sections 3.2 and 3.3, utilizing the shipsEar dataset. Detailed experimental results are presented in Table 10.

Evidently, MGFGNet exhibits a recognition accuracy surpassing all comparative models, achieving 99.5%. Furthermore, it is observed that MGFGNet achieves a 100% recognition rate for all categories except Class 1. This fact indicates a robust generalization capability of the model.

Additionally, on the shipsEar dataset, the ResNet series, Efficient series, and DenseNet also demonstrate strong performance, with recognition accuracies exceeding 93%. It is noteworthy that, with the involvement of CSPNet, DenseNet and ResNet18 show improved recognition accuracy, exceeding 96%, validating their enhancement in model feature extraction capabilities (Wang et al., 2020).

Finally, the MobileNet series performs poorly, with MGFGNet surpassing the highest recognition accuracy within its series, MobileNetV2, by 12.2%, and outperforming the lowest accuracy in MobileNetV3Small by 35%.

3.5.6 Scalability analysis

In order to further validate the scalability of MGFGNet on high-resolution sonar images, this study conducted experimental analyses, comparing MGFGNet with 16 other underwater acoustic target recognition models on the high-resolution sonar dataset SCTD. The recognition accuracy of each model is depicted in Figure 8.

Clearly, MGFGNet's recognition rate continues to surpass that of all other models, confirming the model's scalability. Due to the interference of underwater background noise, which results in poor separability between targets and background in sonar images (Huang and Yang, 2022), the multi-gradient flow model proposed in MGFGNet, based on CSPNet and MHSA, enhances the model's attention to targets (Wang et al., 2020; Han et al., 2021), ensuring the retention of a substantial amount of relevant target information during the feature extraction process. Additionally, further feature enhancement and fusion through CAFM contribute to an improved distinction between target foreground and background, effectively enhancing the model's recognition accuracy.

It is noteworthy that, as indicated in the performance and parameter analysis in Section 4.2, under the same model architecture, depth and recognition rate exhibit a proportional relationship in underwater sonar image target recognition. For instance, the recognition rates of the ResNet series and EfficientNet

10.3389/fmars.2023.1306229

TABLE 10 Details of the models's recognition Accuracy on the shipsEar dataset.

Model	Class 1	Class 2	Class 3	Class 4	Class 5	All
ResNet18	0.944	0.933	0.952	0.979	0.955	0.954
ResNet34	0.917	0.933	0.952	0.979	0.955	0.950
ResNet50	0.917	0.933	0.952	0.979	0.955	0.950
ResNet101	0.889	0.900	0.952	0.979	0.955	0.941
EfficientNet_b0	0.889	0.933	0.988	0.958	0.909	0.950
EfficientNet_b1	0.889	0.933	0.976	0.979	0.909	0.950
EfficientNet_b2	0.861	0.933	0.988	0.958	0.909	0.945
EfficientNet_b3	0.833	0.867	0.964	0.979	0.935	0.931
DenseNet	0.944	0.933	0.976	0.958	0.864	0.950
CSPDenseNet	0.972	0.9	0.988	0.958	0.955	0.964
CSPResNet18	0.972	1	0.964	0.979	0.909	0.968
MobileNetV1	0.861	0.8	0.905	0.792	0.818	0.85
MobileNetV2	0.889	0.867	0.917	0.854	0.727	0.873
MobileNetV3-S	0.667	0.6	0.655	0.625	0.682	0.645
MobileNetV3-L	0.917	0.867	0.786	0.854	0.909	0.845
ViT	0.944	0.933	0.929	0.938	0.909	0.875
MGFGNet	**0.972**	**1**	**1**	**1**	**1**	**0.995**

Bold font indicates the best-performing values within their respective columns.

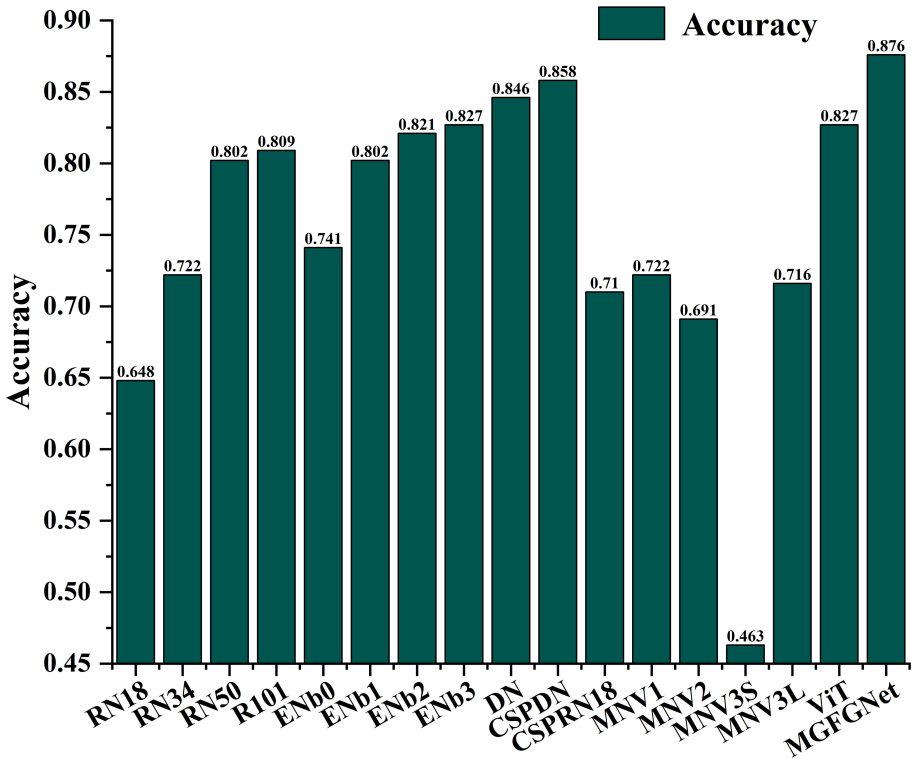

FIGURE 8
Details of the recognition Accuracy of the model on the SCTD dataset.

series validate this conclusion. Conversely, lightweight models such as MobileNet perform poorly, with recognition rates not exceeding 73%, once again confirming that MobileNet is not well-suited for underwater acoustic target recognition scenarios.

3.5.7 Computational bottleneck analysis

Due to the challenges associated with acquiring underwater acoustic datasets, the currently available open datasets are primarily limited to two ship radiated noise datasets: Deepship and ShipsEar. Given that the ShipsEar dataset comprises multiple ship types within each category and has a limited data volume, we conducted experiments with a substantial sample dataset extracted from Deepship to examine MGFGNet's recognition accuracy in relation to dataset size and to identify potential computational bottlenecks. This dataset, which was subject to preprocessing, included a total of 33,693 samples. Figure 9 presents the model accuracy of MGFGNet for various training set sizes sourced from Deepship.

The numerical values in the dataset version indicate the quantity of samples randomly chosen from each category in the Deepship dataset to form the training set for model training, while the test set configuration remained consistent with that presented in Table 2. The results clearly show that as the training set sizes for each category range from 100 to 800, the network model's recognition accuracy experiences rapid growth. Beyond the 800 mark, recognition accuracy tends to plateau, although there is still noticeable improvement as the dataset size increases. Importantly, no indications of encountering computational bottlenecks were observed.

4 Conclusion

An underwater acoustic object identification model MGFGNet based on multi-gradient flow global feature enhancement network is raised in this article. Firstly, by embedding feature extraction module into the target recognition network, the whole target recognition network forms an end-to-end model with underwater acoustic signal as input and classification result as output. Secondly, the invention of Muti-grad block uses multi-gradient flow network to obtain underwater acoustic signal features quickly and effectively, reducing the quantity of model parameters and feature extraction time. Then the CAFM module is used for multi-dimensional feature fusion and feature enhancement to improve the effective characteristic weight of underwater sound. Finally, the Taylor-MCE Loss function is introduced, which enhances model recognition accuracy and mitigates sample imbalance issues within the binary cross-entropy loss. This is achieved by incorporating low-order perturbation terms into the binary cross-entropy loss to suppress sample imbalance components. Consequently, the multi-class classification task is transformed into a set of independent binary classification tasks, effectively addressing the problem of dataset sample imbalance and improving model recognition performance.

The experimental results show that on the Deepship and ShipsEar underwater acoustic data sets, the feature extraction and fusion methods raised in this article have better ability to represent the original underwater acoustic signals. Compared with mainstream underwater acoustic target recognition models such

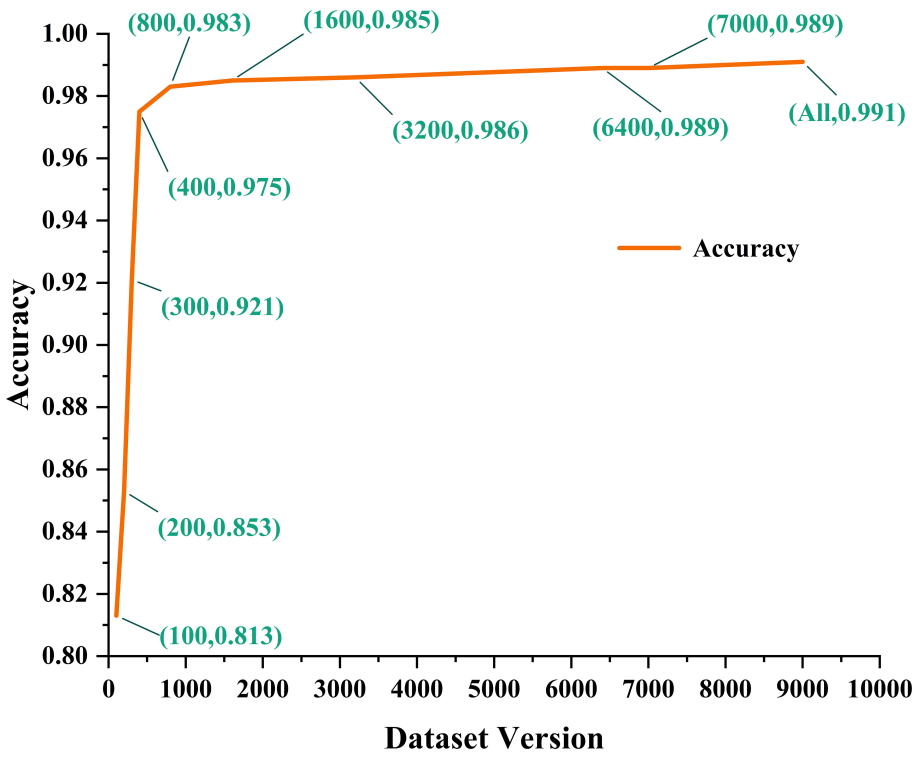

FIGURE 9
Details of the recognition Accuracy of the model on different versions of the Deepship dataset.

as ResNet and EfficientNet, the recognition accuracy of MGFGNet was greatly improved, and the inference time was greatly reduced. MGFGNet network has simple structure and few parameters, which can meet the requirements of end-to-end high precision and low latency in underwater acoustic target identification.

The experimental results of the model proposed in this article have the following potential implications for current underwater target recognition models:

Firstly, our feature extraction and fusion methods have demonstrated that traditional spectrogram-based feature extraction methods are better suited for representing the raw underwater acoustic features. This suggests that current methods for extracting underwater target features, such as those based on signal analysis and bio-inspired features, can be effectively combined with computer vision's feature enhancement techniques (e.g., channel and spatial feature enhancement methods) to further enhance feature representation.

Secondly, the design and experimentation with the Multi-grad block in our proposed classifier have shown that multi-gradient flow networks can better extract deep abstract features of the model while reducing the number of model parameters. This enables underwater target recognition models to depart from the mainstream design pattern of extracting effective features for underwater targets solely through convolution and residual network stacking.

Furthermore, the design of the CAFM in our proposed classifier has demonstrated that incorporating feature fusion and enhancement modules before the classification module in the classifier can significantly enhance the model's recognition accuracy. This enhancement may be related to feature loss during the extraction of effective features before classification and the numerical loss during the normalization process, as this process lacks specific loss control. This can lead to similarities between foreground and background values, making it difficult for the model to effectively recognize the target foreground. Subsequent research can focus on designing feature fusion and feature enhancement modules to improve the distinguishability between target foreground and background.

Lastly, the loss function designed in this paper was explored using the Taylor series, revealing factors influencing the loss function's functionality, such as lateral shifting to address sample imbalance and boosting model recognition accuracy through low-order terms in the Taylor expansion of the function. This enables future research to introduce fewer hyperparameters while gaining more benefits, providing a reference for subsequent studies and better explaining the underlying physical meaning of the loss function.

Additionally, the experimental results of the model proposed in this paper open up potential avenues for future research:

1. Deep learning-based underwater target recognition models have encountered certain bottlenecks, primarily due to their reliance on convolution and residual network stacking, which can lead to limitations in accuracy. The model design approach presented in this paper transforms traditional underwater target recognition into underwater image target recognition, broadening the model construction methods. In the future, lightweight module design methods from computer vision and ideas for feature enhancement and fusion based on the characteristics of underwater feature images can be introduced to enhance the model's efficiency and recognition accuracy, enabling real-time applications.
2. Existing underwater target recognition models mainly employ multi-class classification, but empirical evidence suggests that converting traditional multi-class tasks into multiple binary classification tasks is more suitable for underwater target recognition. Therefore, future research can delve into designing more effective underwater target recognition models based on multiple binary classification tasks that align with the physical characteristics of underwater sound.
3. Current research primarily focuses on building underwater target recognition models, with limited attention to loss function research. Traditional classification loss functions are primarily designed for object image classification and may not be highly adaptable to underwater target feature. Future research can focus on designing loss functions that better align with underwater target features based on the physical characteristics of underwater targets.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Author contributions

JT: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. ZC: Writing – original draft, Writing – review & editing. HQ: Funding acquisition, Writing – original draft. MC: Methodology, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Special Program of Guangxi Science and Technology Base and Talents, grant number AD21220098, the Guangxi Natural Science Foundation, grant number 2022GXNSFDA035070, and the Innovation Project of Guangxi Graduate Education, grant number YCSW2023329.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Aggarwal, A., Srivastava, A., Agarwal, A., Chahal, N., Singh, D., Alnuaim, A. A., et al. (2022). Two-way feature extraction for speech emotion recognition using deep learning. *Sensors* 22 (6), 2378. doi: 10.3390/s22062378

Ahmed, M. S., Aurpa, T. T., and Azad, M. A. K. (2022). Fish disease detection using image-based machine learning technique in aquaculture. *J. King Saud University-Computer Inf. Sci.* 34 (8), 5170–5182. doi: 10.1016/j.jksuci.2021.05.003

Ali, S., Iqbal, N., and Hafeez, Y. (2018). Towards requirement change management for global software development using case base reasoning. *Mehran Univ. Res. J. Eng. Technol.* 37, 639–652. doi: 10.22581/muet1982.1803.17

Boyd, C. E., McNevin, A. A., and Davis, R. P. (2022). The contribution of fisheries and aquaculture to the global protein supply. *Food Secur.* 14 (3), 805–827. doi: 10.1007/s12571-021-01246-9

Bradley, D., Merrifield, M., Miller, K. M., Lomonico, S., Wilson, J. R., and Gleason, M. G. (2019). Opportunities to improve fisheries management through innovative technology and advanced data systems. *Fish. Fish.* 20 (3), 564–583. doi: 10.1111/faf.12361

Darapaneni, N., Sreekanth, S., Paduri, A. R., Roche, A. S., Murugappan, V., Singha, K. K., et al. (2022). "AI based farm fish disease detection system to help micro and small fish farmers," in *2022 Interdisciplinary Research in Technology and Management (IRTM)* (Kolkata, India: IEEE), 1–5.

Di, N., Sharif, M. Z., Hu, Z., Xue, R., and Yu, B. (2023). Applicability of VGGish embedding in bee colony monitoring: comparison with MFCC in colony sound classification. *PeerJ* 11, e14696. doi: 10.7717/peerj.14696

Domingos, L. C., Santos, P. E., Skelton, P. S., Brinkworth, R. S., and Sammut, K. (2022). An investigation of preprocessing filters and deep learning methods for vessel type classification with underwater acoustic data. *IEEE Access* 10, 117582–117596. doi: 10.1109/ACCESS.2022.3220265

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv* abs/2010.11929. doi: 10.48550/arXiv.2010.11929

Gao, X., Xie, J., Chen, Z., Liu, A. A., Sun, Z., and Lyu, L. (2023). Dilated convolution-based feature refinement network for crowd localization. *ACM Trans. Multimedia Computing Commun. Appl.* 19 (6), 1–16. doi: 10.1145/3571134

Gladju, J., Kamalam, B. S., and Kanagaraj, A. (2022). Applications of data mining and machine learning framework in aquaculture and fisheries: A review. *Smart Agric. Technol.* 2, 100061. doi: 10.1016/j.atech.2022.100061

Gonzalez, S., and Miikkulainen, R. (2021). "Optimizing loss functions through multivariate taylor polynomial parameterization," in *Proceedings of the Genetic and Evolutionary Computation Conference*. (Lille, France: ACM) 305–313. doi: 10.1145/3449639.3459277

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., and Wang, Y. (2021). Transformer in transformer. *Adv. Neural Inf. Process. Syst.* 34, 15908–15919. doi: 10.48550/arXiv.2103.00112

Han, R., Jia, N., Huang, J., and Guo, S. (2022). Joint time-frequency domain equalization of MSK signal over underwater acoustic channel. *Appl. Acoustics* 189, 108597. doi: 10.1016/j.apacoust.2021.108597

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. (Las Vegas, USA: IEEE) 770–778. doi: 10.1109/CVPR.2016.90

Ho, Y., and Wookey, S. (2019). The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access* 8, 4806–4813. doi: 10.1109/ACCESS.2019.2962617

Hong, F., Liu, C., Guo, L., Chen, F., and Feng, H. (2021). Underwater acoustic target recognition with a residual network and the optimized feature extraction method. *Appl. Sci.* 11, 1442. doi: 10.3390/app11041442

Hou, Q., Zhou, D., and Feng, J. (2021). "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13713–13722. doi: 10.1109/CVPR46437.2021.01350

Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., et al. (2019). Searching for mobilenetv3. *Proc. IEEE/CVF Int. Conf. Comput. Vision* (Seoul, Korea: IEEE), 1314–1324. doi: 10.1109/ICCV.2019.00140

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv. abs/1704.04861*. doi: 10.48550/arXiv.1704.04861g

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. (Utah, USA: IEEE) 7132–7141. doi: 10.1109/CVPR.2018.00745

Hu, W. C., Chen, L. B., Huang, B. K., and Lin, H. M. (2022). A computer vision-based intelligent fish feeding system using deep learning techniques for aquaculture. *IEEE Sens J.* 22 (7), 7185–7194. doi: 10.1109/JSEN.2022.3151777

Huang, P., and Yang, P. (2022). Synthetic aperture imagery for high-resolution imaging sonar. *Front. Mar. Sci* 9, 1049761. doi: 10.3389/fmars.2022.1049761

Huang, C., Yang, K., Yang, Q., and Zhang, H. (2021). Line spectrum extraction based on autoassociative neural networks. *JASA Express Lett.* 1, 016003. doi: 10.1121/10.0003038

Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., and Keutzer, K. (2014). Densenet: Implementing efficient convnet descriptor pyramids. *ArXiv.abs/1404.1869*. doi: 10.48550/arXiv.1404.1869

Irfan, M., Jiangbin, Z., Ali, S., Iqbal, M., Masood, Z., and Hamid, U. (2021). DeepShip: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification. *Expert Syst. Appl.* 183, 115270. doi: 10.1016/j.eswa.2021.115270

Ji, F., Ni, J., Li, G., Liu, L., and Wang, Y. (2023). Underwater acoustic target recognition based on deep residual attention convolutional neural network. *J. Mar. Sci. Eng.* 11, 1626. doi: 10.3390/jmse11081626

Jin, A., and 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, 69. doi: 10.3390/jmse11010069

Koonce, B. (2021). *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization* (Berkeley (CA): Apress).

Koparan, C., Koc, A. B., Privette, C. V., and Sawyer, C. B. (2018). *In situ* water quality measurements using an unmanned aerial vehicle (UAV) system. *Water* 10 (3), 264. doi: 10.3390/w10030264

Kritzer, J. P. (2020). Influences of at-sea fishery monitoring on science, management, and fleet dynamics. *Aquacult Fisheries* 5 (3), 107–112. doi: 10.1016/j.aaf.2019.11.005

Leng, Z., Tan, M., Liu, C., Cubuk, E. D., Shi, X., Cheng, S., et al. (2022). Polyloss: A polynomial expansion perspective of classification loss functions. *ArXiv. abs/2204.12511*. doi: 10.48550/arXiv.2204.12511

Li, Y., Gao, P., Tang, B., Yi, Y., and Zhang, J. (2022). Double feature extraction method of ship-radiated noise signal based on slope entropy and permutation entropy. *Entropy* 24, 22. doi: 10.3390/e24010022

Li, J., Wang, B., Cui, X., Li, S., and Liu, J. (2022). Underwater acoustic target recognition based on attention residual network. *Entropy* 24, 1657. doi: 10.3390/e24111657

Li, B., Yao, Y., Tan, J., Zhang, G., Yu, F., Lu, J., et al. (2022). "Equalized focal loss for dense long-tailed object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6990–6999.

Li, C., Yuan, X., Lin, C., Guo, M., Wu, W., Yan, J., et al. (2019). "Am-lfs: Automl for loss function search," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (Seoul, South Korea: IEEE) 8410–8419. doi: 10.1109/ICCV.2019.00850

Lim, L. W. K. (2022). Implementation of artificial intelligence in aquaculture and fisheries: deep learning, machine vision, big data, internet of things, robots and beyond. *J. Comput. Cogn. Eng*, 1-7. doi: 10.47852/bonviewJCCE3202803

Lim, K., and Whye, L. (2023). Blended Learning in Animal Biotechnology during Pre-COVID-19, COVID-19 and Post COVID-19 Recovery Phase Periods across the Globe: a Step Forward or Backward? *Int. J. Zool. Anim. Biol.* 6 (2), 1–5. doi: 10.23880/izab-16000451

Lin, T. Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*. (Venice, Italy: IEEE), 2980–2988. doi: 10.1109/ICCV.2017.324

Lin, S., Zheng, H., Han, B., Li, Y., Han, C., and Li, W. (2022). Comparative performance of eight ensemble learning approaches for the development of models of slope stability prediction. *Acta Geotech.* 17, 1477–1502. doi: 10.1007/s11440-021-01440-1

Linka, K., and Kuhl, E. (2023). A new family of Constitutive Artificial Neural Networks towards automated model discovery. *Comput. Methods Appl. Mechanics Eng.* 403, 115731. doi: 10.1016/j.cma.2022.115731

Liu, Z., Peng, D., Zuo, M. J., Xia, J., and Qin, Y. (2022). Improved Hilbert–Huang transform with soft sifting stopping criterion and its application to fault diagnosis of wheelset bearings. *ISA Trans.* 125, 426–444. doi: 10.1016/j.isatra.2021.07.011

Ma, Y., Liu, M., Zhang, Y., Zhang, B., Xu, K., Zou, B., et al. (2022). Imbalanced underwater acoustic target recognition with trigonometric loss and attention mechanism convolutional network. *Remote Sens.* 14, 4103. doi: 10.3390/rs14164103

Mateo, C., and Talavera, J. A. (2020). Bridging the gap between the short-time Fourier transform (STFT), wavelets, the constant-Q transform and multi-resolution STFT. *Signal Image Video Process.* 14, 1535–1543. doi: 10.1007/s11760-020-01701-8

Nouhaila, B., Taoufiq, B. D., and Benayad, N. (2022). An intelligent approach based on the combination of the discrete wavelet transform, delta delta MFCC for Parkinson's disease diagnosis. *Int. J. Adv. Comput. Sci. Appl.* 13 (4), 562-571. doi: 10.14569/IJACSA.2022.0130466

Ruby, U., and Yendapalli, V. (2020). Binary cross entropy with deep learning technique for image classification. *Int. J. Adv. Trends Comput. Sci. Eng.* 9 (4), 5393-5397. doi: 10.30534/ijatcse/2020/175942020

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L. C. (2018). "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. (Salt Lake City, USA: IEEE) 4510–4520. doi: 10.1109/CVPR.2018.00474

Santos-Domínguez, D., Torres-Guijarro, S., Cardenal-López, A., and Pena-Gimenez, A. (2016). ShipsEar: An underwater vessel noise database. *Appl. Acoustics* 113, 64–69. doi: 10.1016/j.apacoust.2016.06.008

Setiyowati, H., Thalib, S., Setiawati, R., Nurjannah, N., and Akbariani, N. V. (2022). An aquaculture disrupted by digital technology. *Austenit* 14 (1), 12–16. doi: 10.5281/zenodo.6499775

Singh, P., Waldekar, S., Sahidullah, M., and Saha, G. (2022). Analysis of constant-Q filterbank based representations for speech emotion recognition. *Digital Signal Process.* 130, 103712. doi: 10.1016/j.dsp.2022.103712

Tian, C., Zheng, M., Zuo, W., Zhang, B., Zhang, Y., and Zhang, D. (2023). Multi-stage image denoising with the wavelet transform. *Pattern Recognition* 134, 109050. doi: 10.1016/j.patcog.2022.109050

Wang, M., and Huang, P. (2023). A multireceiver SAS imaging algorithm and optimization. *IEEE Access* 11, 75112-75120. doi: 10.1109/ACCESS.2023.3297138

Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., and Yeh, I. H. (2020). "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (Virtual: IEEE) 390–391. doi: 10.1109/CVPRW50498.2020.00203

Wang, Z., Wang, Z., Zeng, C., Yu, Y., and Wan, X. (2023). High-quality image compressed sensing and reconstruction with multi-scale dilated convolutional neural network. *Circuits Systems Signal Process.* 42 (3), 1593–1616. doi: 10.1007/s00034-022-02181-6

Wang, Q., and Zeng, X. Y. (2015). Deep learning methods and their applications in underwater targets recognition. *Tech. Acoust* 34, 138–140.

Wang, B., Zhang, W., Zhu, Y., Wu, C., and Zhang, S. (2023). An underwater acoustic target recognition method based on AMNet. *IEEE Geosci. Remote Sens. Lett.* 20, 1–5. doi: 10.1109/LGRS.2023.3235659

Wu, Y., Duan, Y., Wei, Y., An, D., and Liu, J. (2022). Application of intelligent and unmanned equipment in aquaculture: A review. *Comput. Electron. Agric.* 199, 107201. doi: 10.1016/j.compag.2022.107201

Xiao, X., Wang, W., Ren, Q., Gerstoft, P., and Ma, L. (2021). Underwater acoustic target recognition using attention-based deep neural network. *JASA Express Lett.* 1 (10), 106001. doi: 10.1121/10.0006299

Xu, J., Huang, Z., and Li, C. (2019). Advances in underwater target passive recognition using deep learning. *J. Signal Process.* 35, 1460–1475. doi: 10.16798/j.issn.1003-0530.2019.09.003

Yang, P. (2023). An imaging algorithm for high-resolution imaging sonar system. *Multimedia Tools Appl.* doi: 10.1007/s11042-023-16757-0

Yang, H., Xu, G., Li, J., Shen, S., and Yao, X. (2019). Summary of passive underwater acoustic target recognition. *Unmanned Syst. Technol.* 2, 1–7.

Yang, S., Xue, L., Hong, X., and Zeng, X. (2023). A lightweight network model based on an attention mechanism for ship-radiated noise classification. *J. Mar. Sci. Eng.* 11, 432. doi: 10.3390/jmse11020432

Yao, Q., Wang, Y., and Yang, Y. (2023). Underwater acoustic target recognition based on data augmentation and residual CNN. *Electronics* 12, 1206. doi: 10.3390/electronics12051206

Yao, X., Yang, H., and Sheng, M. (2023). Automatic modulation classification for underwater acoustic communication signals based on deep complex networks. *Entropy* 25, 318. doi: 10.3390/e25020318

Zhang, X. (2023). An efficient method for the simulation of multireceiver SAS raw signal. *Multimedia Tools Appl.* doi: 10.1007/s11042-023-16992-5

Zhang, R., He, C., Jing, L., Zhou, C., Long, C., and Li, J. (2023). A modulation recognition system for underwater acoustic communication signals based on higher-order cumulants and deep learning. *J. Mar. Sci. Eng.* 11, 1632. doi: 10.3390/jmse11081632

Zhou, Y., Chen, S.-c., Wu, K., Ning, M.-q., Chen, H.-k., and Zhang, P. (2021). SCTD 1.0:Sonar common target detection dataset. *Comput. Sci.* 48 (11A), 334–339. doi: 10.11896/jsjkx.210100138

Zhou, C., Wu, Y., Fan, Z., Zhang, X., Wu, D., and Tao, Z. (2022). Gammatone spectral latitude features extraction for pathological voice detection and classification. *Appl. Acoustics* 185, 108417. doi: 10.1016/j.apacoust.2021.108417

Zhu, C., Cao, T., Chen, L., Dai, X., Ge, Q., and Zhao, X. (2023). High-order domain feature extraction technology for ocean acoustic observation signals: a review. *IEEE Access* 11, 17665-17683. doi: 10.1109/ACCESS.2023.3244782

Zhu, X., Dong, H., Rossi, P. S., and Landrø, M. (2022). Time-frequency fused underwater acoustic source localization based on contrastive predictive coding. *IEEE Sens J.* 22, 13299–13308. doi: 10.1109/JSEN.2022.3179405

Zhufeng, L., Xiaofang, L., Na, W., and Qingyang, Z. (2022). Present status and challenges of underwater acoustic target recognition technology: A review. *Front. Phys.* 10, 1044890. doi: 10.3389/fphy.2022.1044890