**Frontiers in Computing and Intelligent Systems**
ISSN: 2832-6024 | Vol. 15, No. 1, 2026

# MFD-MoE: Multi-scale Fusion Denoising Mixture of Experts Network for Stable Recognition of Underwater Acoustic Targets

**Haiyang Yao [1, 2, *], Daqing Guo [1], Haiyang Wang [1, 2, 3] and Fan Wu [3]**

[1] School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, Shaanxi 710021, PR China
[2] Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, Shaanxi 710021, PR China
[3] School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an Shaanxi 710072, PR China

*** Corresponding author:** Haiyang Yao

**Abstract:** Stable recognition of underwater acoustic targets is a key technology for ensuring continuous and efficient maritime defense, resource exploration, and environmental monitoring. However, in complex marine environments, background noise, propagation disturbances, and dynamic changes in target features often lead to unstable recognition results. To address this issue, this paper proposes the Multi-scale Fusion Denoising Mixture of Experts (MFD-MoE) Network, with a focus on improving recognition stability. We design the Denoised Time-Frequency Feature Fusion module (DTFF), in which first performs soft-threshold denoising, followed by a feature decomposition strategy that includes key steps such as channel splitting, multi-scale feature extraction, and upsampling. We then design a multi-scale feature fusion strategy that combines multi-modal and pooling results, enhancing the stable expression of features in the presence of interference and target state changes, resulting in a feature extraction architecture consisting of four DTFF layers. Furthermore, we introduce a mixture of experts model, and based on the characteristics of underwater acoustic information recognition, we design an expert mixture mechanism with residual connections to improve the model's stability under target motion or interference. Experimental results show that the MFD-MoE model significantly outperforms existing methods in multi-class underwater acoustic target recognition tasks, offering both high stability and high accuracy, thus providing a new solution for intelligent underwater target perception in complex environments.

**Keywords:** Underwater Acoustic Target Recognition; Attention Mechanism; Mixture of Experts.

## 1. Introduction

In the complex and ever-changing marine environment, stable recognition of underwater acoustic targets not only maintains high accuracy under interference conditions but also enhances adaptability to target motion and cross-scenario environments. This provides a reliable foundation for achieving sustained and efficient maritime defense, resource exploration and development, as well as long-term environmental monitoring and assessment. Therefore, studying and achieving stable recognition of underwater acoustic targets is not only of great scientific value but also holds profound significance for marine defense and resource utilization strategies[1][2]. However, environmental factors such as seawater temperature, depth, and ocean currents exhibit significant spatiotemporal non-stationarity. Disturbances like surface waves and seabed topography introduce random noise and strong reflections, causing propagation paths to dynamically change over time and space, resulting in signal feature distortion and energy attenuation. The target's own motion state and radiation characteristics are dynamic, with its radiated noise spectrum and amplitude changing according to variations in speed and posture, further increasing feature uncertainty. There are also significant differences in cross-regional background noise, including multi-source noise from ships, marine life, and industrial activities, which can easily interfere with target signals. The combination of these multiple factors leads to feature drift, decision conflicts, and insufficient robustness during recognition, making stable recognition challenging.

Intelligent recognition of underwater acoustic signals typically involves two key steps: feature extraction and classification recognition. In terms of feature extraction, research has evolved from manually designed features based on time-domain, frequency-domain, and time-frequency analysis to the use of deep neural networks that automatically learn multi-level deep features. The parametric approach has also developed from being fixed to adaptively optimized, significantly improving the ability to adapt to dynamic changes in features under complex acoustic environments. In terms of classification recognition models, there has been a gradual development from early rule-based manual classification to machine learning methods such as support vector machines and random forests, and further to deep learning models like convolutional neural networks (CNN), recurrent neural networks, and Transformers. This progression has continuously improved recognition accuracy and robustness[3][4]. The field of underwater acoustic signal recognition has undergone a profound shift from being experience-driven to being data-driven and adaptive learning-based. Early reliance on manual feature extraction limited performance due to environmental changes and noise interference. However, with the advancement of machine learning and deep learning, models are now able to automatically learn high-level feature correlations and demonstrate stronger generalization capabilities under conditions such as cross-sea areas and low signal-to-noise ratios. Furthermore, deep models with adaptive learning mechanisms can dynamically adjust parameters to cope with non-stationarity and feature drift, enabling continuous optimization. Underwater acoustic recognition has thus

transitioned from a static process relying on prior knowledge to a dynamic process of data and model co-evolution, laying the foundation for achieving stable, robust, and reliable recognition in complex environments.

In our previous work, we proposed a time-frequency-based method for detecting ship radiated noise, demonstrating the effectiveness of time-frequency features in noise signal detection [5]. However, in complex marine environments, underwater acoustic signals are easily affected by background noise, propagation interference, and changes in target motion states, leading to unstable feature representations, which in turn reduce recognition accuracy and robustness. To address this, we designed the Denoised Time-Frequency Feature Fusion (DTFF) module, which uses soft-threshold denoising to achieve noise suppression, and combines channel splitting, multi-scale feature extraction, and upsampling decomposition strategies to enhance the multi-level representation ability of time-frequency features. At the same time, we employed a fusion mechanism of multi-modal features and pooling results to further improve the model's stable expression ability under interference and feature drift conditions. On this basis, we introduced a residual-enhanced mixture of experts model, which leverages the complementarity between experts and the stability of residual learning to enhance the model's adaptability to target motion and attribute changes. The overall design aims to achieve robust feature extraction at the feature layer and robust fusion at the decision layer, significantly improving the stability and reliability of underwater acoustic target recognition in complex environments. Our model has demonstrated excellent performance on two public datasets, ShipEar and DeepShip, contributing to the accurate measurement of underwater acoustic signals.

The main contributions of this work are summarized as follows:

1) Proposing a multi-level feature extraction architecture for denoised time-frequency feature fusion: We designed the Denoised Time-Frequency Feature Fusion (DTFF) module, which combines soft-threshold denoising, channel splitting, multi-scale feature extraction, and upsampling strategies. This design mitigates the impact of noise on target characteristics and enables the extraction of noise-independent abstract semantic features from time-frequency features. This approach effectively enhances the robustness and expressive power of features, laying the foundation for stable recognition of underwater acoustic targets in complex marine environments.

2) Constructing a residual-enhanced mixture of experts model to improve recognition stability: To address the feature distribution differences in underwater acoustic signals during target motion and attribute changes, we propose a residual-optimized expert mixture mechanism. By fully leveraging the complementarity between different experts and the stability of residual learning, we enhanced the model's generalization ability and decision-layer robustness. Experimental results show that this method significantly outperforms existing approaches in multi-class underwater acoustic target recognition tasks, achieving high stability and accuracy in complex environments.

The remainder of this paper is organized as follows: Section 2 reviews the related research on underwater acoustic target recognition, focusing on analyzing the shortcomings of existing methods in complex marine noise environments. Section 3 provides a detailed description of the structure of the proposed MFD-MoE model and its key modules, including the multi-level feature extraction architecture for denoised time-frequency feature fusion and the design and implementation of the residual-enhanced mixture of experts model. Section 4 validates the performance advantages of the proposed method through comparative and ablation experiments across multiple evaluation metrics. Finally, Section 5 summarizes the main contributions of this paper and discusses future research directions.

# 2. Related Works

Underwater acoustic target recognition is a fundamental task in maritime defense, resource exploration, and environmental monitoring, where accurate and stable recognition remains highly challenging due to complex oceanic conditions. In recent years, deep learning has emerged as a powerful paradigm, enabling automatic extraction of hierarchical representations and significantly improving recognition performance compared with traditional handcrafted approaches. Against this background, a number of studies have focused on advancing underwater acoustic target recognition through novel network architectures, feature extraction strategies, and fusion mechanisms, which will be systematically reviewed below.

In our previous work, we proposed a Time-Frequency Swin-Transformer model (TFST) [6], which extracts more time-frequency features through a hierarchical self-attention module to improve recognition accuracy. Furthermore, adaptive parameters derived through self-distillation learning were found to partially mitigate the feature mixing observed in the results [7]. However, this method still lacks stability and robustness when dealing with background noise and target state changes in complex marine environments, and it is unable to fully cope with the non-stationarity of the environment and the rapid changes in target characteristics.

Early underwater acoustic target recognition was carried out using classical machine learning algorithms to process time-frequency features. For example, Wang and Zeng [8] employed bark-wavelet analysis in combination with the Hilbert-Huang transform to analyze signals, using support vector machines (SVM) as the classifier. However, in the complex marine environment, even with sufficient training data, traditional machine learning methods are unable to extract deeper features. Deep learning models, by constructing complex multi-layer structures and possessing powerful nonlinear expression capabilities, can automatically learn and extract key features relevant to the task from raw, complex data. CNN models have clear advantages in extracting local features from time-frequency features. Hong et al. [9] proposed an underwater acoustic target recognition method based on a residual network and optimized feature extraction methods. By using 3D fused features and data augmentation strategies, the recognition accuracy was improved. Lin et al. [10] proposed an underwater acoustic target recognition method based on iterative short-time Fourier transform. They extracted inherent frequency distribution features of energy in different frequency bands, combined time-frequency features to form a feature matrix, and introduced a fuzzy mixed feature enhancement mechanism. They trained the model using a reduced-dimensional RSNet-18 network and performed recognition using a stepwise voting strategy, which effectively improved recognition performance. Liu et al. [11] proposed an underwater acoustic target recognition method combining

Mel-frequency cepstral coefficients (MFCC) and residual attention convolutional neural network (RACNN). By introducing residual structures and attention mechanisms, the feature extraction ability and focus on key information were enhanced, which effectively improved the recognition accuracy of radiated noise in different frequency bands. Xie et al. [12] proposed an underwater target recognition method integrating adaptive gating mechanisms with attention-guided networks. By introducing a gating mechanism to dynamically adjust the importance of feature channels and combining an attention-guided module to strengthen the expression of key feature regions, they effectively alleviated feature redundancy and interference issues under complex marine backgrounds. Compared to CNN models, Transformer models can capture long-range dependencies in time-frequency features through global attention, enhancing their ability to understand complex structures and global semantics. Transformers have also been widely applied in underwater target recognition. Feng et al. [13] proposed a Transformer-based deep learning network for underwater acoustic target recognition. By introducing the multi-head self-attention mechanism from the Transformer structure, the recognition performance of ship radiated noise in complex marine environments was significantly improved. However, when facing redundant information in complex environments, the performance of the model is still limited. To address this, the attention mechanism, as a strategy for dynamically adjusting the importance of features, can effectively help the model focus on key feature regions, thereby improving recognition accuracy in complex scenarios.

The core idea of the attention mechanism is to dynamically highlight key features during the information processing and suppress irrelevant or redundant information. In deep learning, the attention mechanism works by weighting the input features, enabling the model to focus more effectively on the parts most useful for the current task, thereby enhancing the representation ability and decision-making accuracy. Common types of attention mechanisms include channel attention mechanisms [14] and spatial attention mechanisms[15]. Channel attention dynamically adjusts the weights of each channel by adaptively modeling the importance of each channel. Spatial attention, on the other hand, focuses on different spatial positions in the feature map, weighting two-dimensional regions of the image or feature map to enable the model to better capture prominent spatial areas. Due to its powerful performance, it is also widely used in underwater target recognition. Chen et al.[16] proposed a deep learning framework combining a residual network with a channel attention module to address the problem of overlapping ship radiated noise caused by multi-ship interference in real-world environments, aiming for underwater acoustic multi-target recognition. Yang et al. [17] integrated channel, frequency, and time attention modules to propose a lightweight CFTANet model, significantly reducing the model complexity while maintaining high accuracy. However, as the types of underwater acoustic data and environmental complexities continue to increase, relying solely on the attention mechanism to enhance feature attention within a unified model still faces performance bottlenecks. We further introduce a mixture of experts structure that can adaptively select different sub-models based on the input features.

The concept of the Mixture of Experts (MoE) model was introduced by Jacobs et al. [18]. The entire system consists of a gating network and multiple experts, where the routing layer is responsible for dynamically selecting or assigning the appropriate experts based on the input features, and the expert layers consist of multiple sub-models, each handling the assigned tasks. Shazeer et al. [19] introduced a sparse gating expert mixture layer and applied it to natural language processing, reducing computational load and improving model efficiency. With further research, various models have applied the expert mixture model to neural network architectures. Xie et al. [20] were the first to apply the expert mixture model to underwater target recognition, introducing multiple expert layers as independent learners, along with a routing layer that determines the assignment of experts according to the characteristics of inputs.

# 3. Methodology

This chapter is divided into three sections. First, the overall architecture of the model is introduced. Then, the multi-level feature extraction architecture for denoised time-frequency feature fusion is presented to extract multi-scale features. Finally, the residual-enhanced mixture of experts model is introduced to enhance the model's generalization ability and decision-layer robustness.

## 3.1. Overall Model Architecture

The overall model of this paper is shown in Fig 1. The raw time-domain data received by the hydrophone is referred to as the raw features, which usually involve a large amount of data and cannot be directly used for recognition. To achieve effective classification and recognition, we first process the underwater acoustic signals by framing, windowing, and applying Fourier transform. The frequency-domain signals are then passed through a Mel filter to form the time-frequency representation. Next, we input the obtained time-frequency features into our MFD-MoE network, which includes four stages of feature extraction, the Denoised Time-Frequency Feature Fusion module, and the Mixture of Experts with Residual. Finally, we output the recognition results.

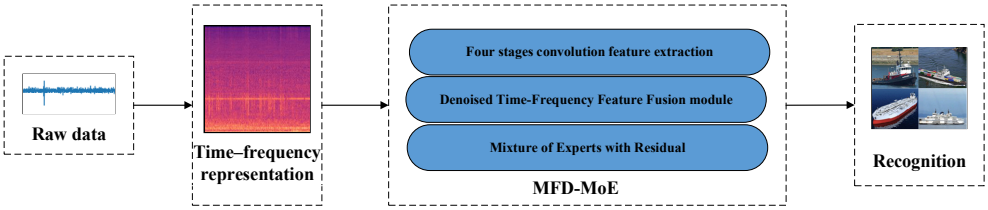



**Fig 1.** The MFD-MoE architecture

## 3.2. Front-end Backbone Network

The Front-end Backbone network structure consists of three parts: the ConvNeXt network, the Time-Frequency Feature Fusion module, and the soft-threshold denoising module. The Front-end Backbone network is responsible for

converting the input acoustic features into fixed-dimensional representations, which serve as the input to the expert layers and routing layers.

ConvNeXt [3] is based on the ResNet [21] architecture, and its design draws inspiration from the Swin-Transformer [4] concept. It uses LayerNorm (LN) as the normalization method and GELU as the activation function to simplify the network structure, reduce parameters and computational load, and enhance the model's generalization ability and performance. The input time-frequency representation matrix is 3*224*224. It first passes through convolutional layers and LN layers, followed by four stages implemented with downsampling modules and stacking multiple DTFF Blocks for feature extraction. The downsampling module consists of LN layers and convolutional layers, while the DTFF Block includes convolutional layers, LN layers, GELU layers, the Time-Frequency Feature Fusion (TFF) module, and the soft-threshold denoising module. We place the TFF module at the front of the DTFF Block to enhance the network's ability to extract key features. After adding the soft-threshold denoising module, we obtain a deep learning model with noise robustness. Finally, the deep features are output through a global average pooling layer. The deep features extracted by the front-end backbone network are then input to the subsequent layers, as shown in Fig 2.

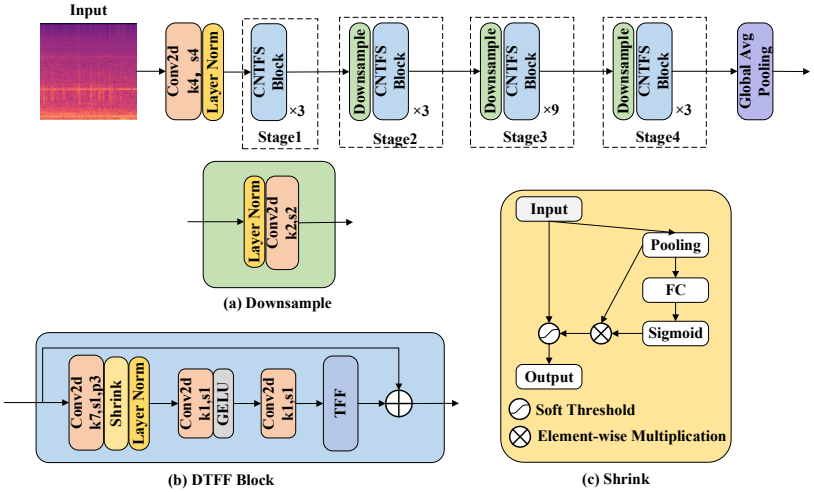



**Fig 2.** The front-end Backbone network architecture

## 3.3. Denoised Time-Frequency Feature Fusion module

The denoised Time-Frequency Feature Fusion module as shown in Fig 2(b), the Time-Frequency Feature Fusion (TFF) module is illustrated in Fig 3(a). First, the input feature $X \in \mathrm{R}^{B\times C\times H\times W}$ is evenly split into four parts along the channel dimension, i.e., $\left[X_0, X_1, X_2, X_3\right]$, where each feature $X_i \in \mathrm{R}^{B\times C/4\times H\times W}$ has the same spatial dimensions as $X$ but one-quarter of the channels. They are then fed into the Multi-Scale Feature Extraction Unit (MFEU). Among them, $X_0$ is directly processed by a $3\times 3$ depth-wise convolution for feature extraction, and the output is directly used as the feature at this scale. The remaining parts are downsampled via pooling operations, followed by $3\times 3$ depth-wise convolution, and finally restored to the original spatial size through upsampling, resulting in features $X_i^{\theta} \in \mathrm{R}^{B\times C/4\times H\times W}$. Given the input feature $X$, this process can be formulated as:

$$\left[X_0, X_1, X_2, X_3\right] = Split\left(X\right),$$

$$X_0^{\theta} = \mathrm{DW-Conv}_{3\times 3}\left(X_0\right),$$

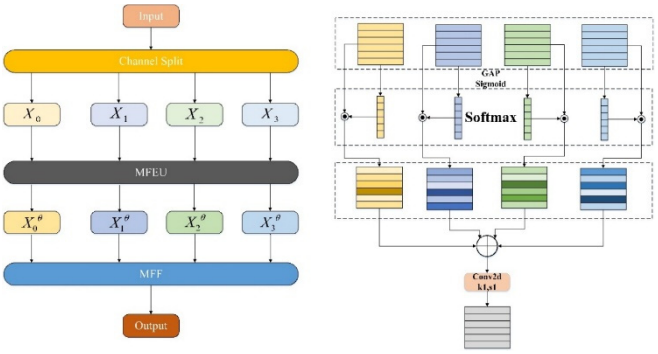
$$X_i^{\theta} = \uparrow_p \left(\mathrm{DW-Conv}_{3\times 3}\left(\downarrow \frac{p}{2^i}\left(X_i\right)\right)\right), 1 \le i \le 3,$$

where $Split\left(\cdot\right)$ corresponds to the channel splitting operation, DW-Conv $_{3\times 3}$ () is the $3\times 3$ depthwise convolution, $\uparrow_p \left(\cdot\right)$ denotes upsampling the features to the original resolution $p$ using nearest-neighbor interpolation, and $\downarrow \frac{p}{2^i}$ indicates downsampling the input features to size $\frac{p}{2^i}$ via pooling. With this design, features from local to global in the time-frequency map can be captured more comprehensively.

Subsequently, we input the obtained multi-scale features into the Multi-Scale Feature Fusion (MFF) module, as shown in Fig 3(b). MFF fuses features from different scales $x_i$. First, we perform a global average pooling (GAP) operation on the multi-scale features, then use the Sigmoid function to ensure that the channel weights for features of different scales fall within the range (0, 1). Next, the obtained multi-scale average channel weights are normalized using the Softmax operation. This design considers both the channel importance of individual scales and balances the weight distribution across different scales. Subsequently, the weight of each scale is multiplied element-wise with its corresponding input feature, and the weighted features from all scales are summed to obtain the fused multi-scale feature. Finally, a 1 × 1 convolution is applied to restore the number of channels.

The impact of marine environmental noise and disturbances during underwater acoustic propagation varies in degree across different time periods and frequency bands for signal features. The Soft-threshold module (STM) processing is an effective denoising method that mitigates the impact of complex interference by setting feature values below a threshold to zero [22]. We leverage deep learning techniques to automatically learn the optimal soft-threshold.

The soft-threshold module (Shrink) is shown in Fig 1(c), and its denoising formula is as follows:




(a) (b)

**Fig 3.** (a)TFF model, (b)MFF model.

### 3.4. Mixture of Experts Module with Residual

The Mixture of Experts (MoE) Module consists of a multi-layer perceptron with two linear layers. Different expert layers focus on differentiated deep features, thereby enhancing the model's classification ability. Each expert has the same structure, but its parameters are independent. Each expert is only activated when it encounters the appropriate input. This structure, by capturing the latent characteristics of high-level features, adaptively learns diverse data, significantly improving the performance of underwater acoustic target recognition, The MoE module as shown in Fig 4.

We set the batch size to $n$, the input spectrograms to $\{x_i\}$ with corresponding labels $\{y_j\}$ (where $i=1,2,...,n$), the number of expert layers to $m$, and the expert layers to $E_1(\cdot),\ldots,E_m(\cdot)$. We input $x_i$ into the front-end backbone network to obtain deep features $r_i \in \mathrm{R}^{n\times 768}$. The deep features $r_i$ are then fed into the routing layer $S(\cdot)$ to obtain routing scores $s_i = S(r_i) \in \mathrm{R}^{b\times m}$. These scores are normalized into routing probabilities $p_i \in \mathrm{R}^{1\times m}$ using the softmax function, where $p_i$ represents the probability of assigning the sample to each expert.

The model inputs the deep feature $r_i$ to the expert layer with the highest probability. The entire routing assignment is described by the formula below, illustrated using an MFD-MoE model with 4 experts. For example, when $p_i = (0.2, 0.1, 0.2, 0.5)$, the deep feature $r_i$ will be assigned to $E_4(\cdot)$.

Furthermore, to prevent overfitting in routing assignments and model performance, we adopt the concept of residual connections and propose Residual RMoE. As shown in Fig 4 (bottom), RMoE retains the structure of MoE while adding an additional fixed expert layer $E_R(\cdot)$. Here, $r_i$ is directly fed into the fixed expert layer without routing computation, generating $\mathrm{log}its_{res} = E_R(r_i)$. Finally, $\mathrm{log}its_{res}$ is added to the outputs $\mathrm{log}its_{exp}$ from the other expert layers to obtain the overall $\mathrm{log}its$.

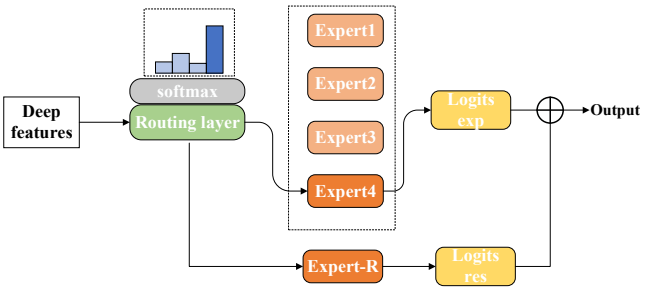



**Fig 4.** Mixture of Experts Module

## 4. Experiments

### 4.1. Dataset

This paper uses two publicly available underwater acoustic datasets, ShipEar[23] and DeepShip[24]. The ShipEar dataset classifies ships into four categories based on their size: medium and small ships, small ships, large passenger ships, and large ocean-going ships. These categories are labeled as Class A, Class B, Class C, and Class D, with an additional category for natural background noise in the marine environment, labeled as Class E, totaling five categories, with 90 noise samples. Details are shown in table 1. The DeepShip dataset is currently the largest open-source underwater dataset, containing over 47 hours of underwater acoustic signals from 265 different types of ships. These ships are labeled into four types: Cargo, Tug, Passenger, and Tanker.shown in table 2.

In our experiments, the ShipEar dataset is segmented into 5-second audio clips, while the DeepShip dataset is divided into 30-second segments. The datasets are then split into training and validation sets in a 7:3 ratio. The specific numbers for both datasets are shown in Table 2.

**Table 1.** Dataset types

| | Shipsear | Deepship |
|---|---|---|
| Class label | Ship type | Ship type |
| A | fishing boats, trawlers, mussel boats, tugboats and dredgers | Cargo |
| B | motorboats, pilot boats and sailboats | Passengership |
| C | passenger, ferries | Tanker |
| D | ocean liners and ro-ro vessels | Tug |
| E | Background noise | |

**Table 2.** Experiment dataset partition

| Dataset | Classes | Total size | Training size | Validation size |
|---|---|---|---|---|
| Shipsear | A,B,C,D,E | 2223 | 1556 | 667 |
| Deepship | Passenger, Tug, Tanker, Cargo | 5264 | 4212 | 1052 |

### 4.2. Setup and Parameter

We use the AdamW optimizer in combination with a cosine annealing learning rate scheduler to improve the training performance and convergence speed of the model. The initial learning rate is set to 0.001, and the weight decay coefficient is 0.01. The batch size is 64, and the number of training epochs is 70. The cross-entropy loss function is employed to measure the classification performance.

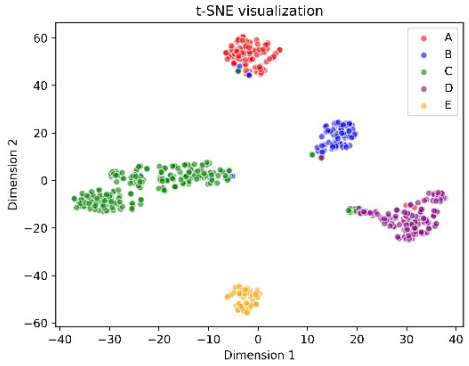
## 4.3. Evaluation Metrics

We use accuracy, F1 Score[25], and Kappa[26] coefficient as evaluation metrics. The F1 Score is a comprehensive metric that considers both Precision and Recall, used for model evaluation. The Kappa coefficient provides a more accurate reflection of the model's classification ability, especially in the case of class imbalance.

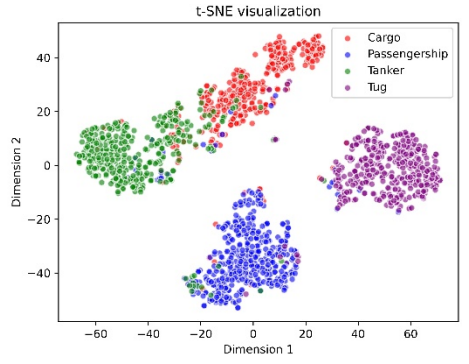
## 4.4. Recognition Performance

**Table 3.** Performance of MFD-MoE on two val sets.

| Dataset | Class | Precision (%) | Recall(%) | F1-score(%) |
|---|---|---|---|---|
| DeepShip | Cargo | 90.28 | 88.08 | 89.17 |
| | Passengership | 93.24 | 93.90 | 93.57 |
| | Tanker | 88.64 | 90.46 | 89.54 |
| | Tug | 97.45 | 96.95 | 97.20 |
| | Weighted avg | 92.40 | 92.35 | 92.37 |
| ShipEar | A | 96.30 | 94.55 | 95.42 |
| | B | 96.59 | 94.44 | 95.50 |
| | C | 98.78 | 96.03 | 97.39 |
| | D | 92.90 | 99.31 | 96.00 |
| | E | 98.53 | 100.00 | 99.26 |
| | Weighted avg | 96.62 | 96.87 | 96.71 |




(a)




(b)

**Fig 5.** t-SNE visualization of features in MFD-RMoE
(a) DeepShip val set, (b) ShipsEar val set.

We first tested the recognition performance of the Multi-scale Fusion Denoising Residual Mixture of Experts (MFD-RMoE) model on the ShipEar and DeepShip datasets. We chose Precision, Recall, and F1-score as evaluation metrics. table 3 lists the average metrics of MFD-RMoE for each category on the two test datasets. Among the three evaluation metrics, MFD-RMoE achieved a weighted average score of 96.62%, 96.87%, and 96.71% on ShipEar, with the highest F1-score for the recognition of Class E (background noise). On DeepShip, MFD-RMoE achieved a weighted average score of 92.40%, 92.35%, and 92.37%, with the highest F1-score for the recognition of the Tug class. Compared to ShipEar, the DeepShip dataset has more noise and is more complex, making recognition more difficult. However, all three metrics exceed 92%, indicating that MFD-RMoE is also applicable in complex noise environments.

Subsequently, we used t-SNE to visualize the classification results of the MFD-RMoE model. As shown in Fig 4, the feature distribution map of MFD-RMoE on the ShipEar validation set is shown in Fig 5(a). The data points from different categories show good clustering, with clear separation between the categories, demonstrating the model's good recognition performance. The feature distribution map of MFD-RMoE on the DeepShip validation set is shown in Fig 5(b), where there is some proximity between Cargo and Tanker, and they do not cluster well together. The Tug category shows a good clustering effect, which is consistent with the higher F1-score for the Tug class in table 3, while the Cargo and Tanker categories have lower F1-scores.

## 4.5. Comparative Experiments

We conducted comparative experiments with different models on the DeepShip and ShipEar datasets, comparing our model with the AMNet model[27], TFST model[6], CMoE model[20], as well as the InceptionNeXt model[28] and Swin Transformer model [4], which have achieved good performance in image classification tasks. The comparative experiments on the ShipEar dataset are shown in table 4, and those on the DeepShip dataset are shown in table 5.

Our method achieved the best results in the evaluation metrics on both datasets. In the comparative experiment on the ShipEar dataset, as shown in table 4, our model achieved Accuracy, F1-score, and Kappa of 96.68, 96.71, and 95.60, respectively. Compared to the second-best model, Accuracy improved by 1.65, F1-score increased by 1.4, and Kappa improved by 2.19. CMoE achieved the second-best results with scores of 95.03, 95.31, and 93.41 for the three evaluation metrics. AMNet and InceptionNeXt also produced similar results, while Swin Transformer and TFST showed less favorable performance.

On the DeepShip dataset, where the acoustic signals are interfered with by ocean noise, as shown in table 5, our model achieved Accuracy, F1-score, and Kappa of 92.45, 92.37, and 89.92, respectively. Compared to the second-best model, Accuracy improved by 2.54, F1-score increased by 2.56, and Kappa improved by 3.39. AMNet and TFST performed poorly when the data was affected by noise interference. In summary, our model demonstrated excellent performance on both datasets, highlighting its superior robustness.

**Table 4.** Results of comparative experiment on ShipEar dataset

| Model | Accuracy (%) | F1-score(%) | Kappa(%) |
|---|---|---|---|
| TFST | 89.15 | 88.47 | 85.54 |
| AMNet | 93.97 | 94.34 | 91.99 |
| CMoE | 95.03 | 95.31 | 93.41 |
| Swin transformer | 89.90 | 89.62 | 86.58 |
| InceptionNeXt | 93.37 | 93.47 | 91.21 |
| MFD-RMoE | **96.68** | **96.71** | **95.60** |

**Table 5.** Results of comparative experiment on DeepShip datase

| Model | Accuracy (%) | F1-score(%) | Kappa(%) |
|---|---|---|---|
| AMNet | 83.76 | 83.67 | 78.34 |
| TFST | 81.23 | 81.09 | 74.95 |
| CMoE | 89.91 | 89.81 | 86.53 |
| Swin transformer | 84.40 | 84.28 | 79.19 |
| InceptionNeXt | 86.55 | 86.46 | 82.07 |
| MFD-RMoE | **92.45** | **92.37** | **89.92** |

## 4.6. Ablation Experiment

We validated the effectiveness of our residual optimization strategy, TFF module, STM module, and the combination of these strategies on the DeepShip dataset. Th e results are shown in table 6. From the experimental results, we can see that our residual optimization strategy improved Accuracy by 0.44, F1-score by 0.44, and Kappa by 0.58. The TFF module improved Accuracy by 1.41, F1-score by 1.14, and Kappa by 1.52. The STM module improved Accuracy by 1.96, F1-score by 1.92, and Kappa by 5.32. When combining the residual optimization strategy, TFF, and STM, Accuracy improved by 2.79, F1-score increased by 2.76, and Kappa improved by 3.71. Combining TFF and STM also yielded similar results.

**Table 6.** Results of ablation experiment on DeepShip datase

| Model | Accuracy (%) | F1-score (%) | Kappa(%) |
|---|---|---|---|
| CN_MOE | 89.66 | 89.61 | 86.21 |
| CN_RMOE | 90.10 | 90.05 | 86.79 |
| CN_MOE_TFF | 90.80 | 90.75 | 87.73 |
| CN_MOE_STM | 91.62 | 91.53 | 91.53 |
| MFD-MoE | 92.39 | 92.30 | 89.84 |
| MFD-RMoE | **92.45** | **92.37** | **89.92** |

## 4.7. Mixture of Experts Experiment

In this section, we investigate the impact of the number of experts in MFD-RMoE on underwater acoustic recognition. Increasing the number of experts allows for a more granular understanding of the data, but it also reduces the amount of data assigned to each expert. Therefore, selecting the optimal number of experts becomes a complex trade-off. table 7 and 8 present the comparison results for 2 and 4 experts on the ShipEar and DeepShip datasets. On the ShipEar dataset, the performance with 4 experts is better than with 2 experts, but on the DeepShip dataset, the performance with 2 experts is better than with 4 experts. This shows that the impact of the number of expert layers on performance depends on the characteristics of the data. Therefore, it is necessary to perform an extensive search to find the optimal number of experts. Fig 6 shows the expert assignment for the MFD-RMoE model with 4 experts on both datasets, where the colors in the Fig represent the proportion of samples assigned to each expert, validating the effectiveness of MoE in the MFD-RMoE model for underwater target recognition.

**Table 7.** Results of different number of experts experiment on DeepShip

| Expert num | Accuracy (%) | F1-score | Kappa (%) |
|---|---|---|---|
| 2 | **92.45** | **92.37** | **89.92** |
| 4 | 91.69 | 91.59 | 88.91 |

**Table 8.** Results of different number of experts experiment on ShipsEar

| Expert num | Accuracy (%) | F1-score | Kappa(%) |
|---|---|---|---|
| 2 | 96.23 | 96.50 | 95.01 |
| 4 | **96.68** | **96.71** | **95.60** |

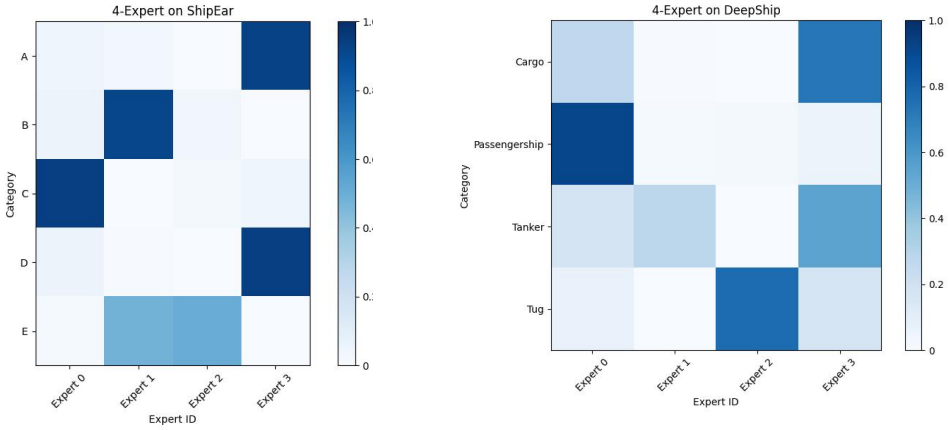



**Fig 6.** Assignment of the 4-Expert MFD-RMoE Model on the ShipEar and DeepShip Datasets

# 5. Conclusion

This paper addresses the problem of underwater acoustic signal recognition in complex marine environments by proposing the MFD-MoE model. First, the model introduces a multi-level feature extraction architecture for denoised time-frequency feature fusion, and designs the Denoised Time-Frequency Feature Fusion (DTFF) module to achieve stable feature representation under noise interference and changes in target states. This design effectively enhances the robustness and expressive power of features, laying the foundation for stable underwater target recognition in complex marine environments. In this paper, we also introduce the mixture of experts model and its residual learning optimization strategy. This strategy significantly improves the model's generalization ability to diverse features, effectively handling variations under different targets and environments, and enhancing the stability and accuracy of recognition. Through a series of comparative and ablation experiments on two public datasets, DeepShip and ShipEar, we demonstrate the robustness and stability of our model.

However, despite the good performance of our model, its internal mechanisms lack theoretical support. In the future, we will focus on introducing interpretability-driven techniques and methods to further analyze the internal mechanisms and decision-making processes of the model.

# References


[1] N. Jones, "The quest for quieter seas," Nature, vol. 568, no. 7751, pp. 158–161, 2019.

[2] C. M. Duarte, L. Chapuis, S. P. Collin, et al., "The soundscape of the Anthropocene ocean," Science, vol. 371, eaba4658, 2021.

[3] Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11976-11986.

[4] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[5] Haiyang Y, Zhichen Z, Haiyan W, et al. Narrow band time–frequency space matched passive detector for underwater signal[J]. Applied Acoustics, 2021, 183: 108287.

[6] Wu F, Yao H, Wang H. Recognizing the State of Motion by Ship-Radiated Noise Using Time-Frequency Swin-Transformer [J]. IEEE Journal of Oceanic Engineering, 2024, 49(3): 667-678.

[7] Fan W, Haiyang Y, Zhongda Z, et al. ESTMST-ST: An end-to-end soft threshold and multi-loss self-distillation based Swin-Transformer for underwater acoustic signal recognition[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024.

[8] Wang S, Zeng X. Robust underwater noise targets classification using auditory inspired time–frequency analysis [J]. Applied Acoustics, 2014, 78: 68-76.

[9] Hong F, Liu C, Guo L, et al. Underwater acoustic target recognition with a residual network and the optimized feature extraction method[J]. Applied Sciences, 2021, 11(4): 1442.

[10] Lin B, Gao L, Zhu P, et al. An underwater acoustic target recognition method based on iterative short-time fourier transform[J]. IEEE Sensors Journal, 2024.

[11] Liu D, Yang H, Hou W, et al. A novel underwater acoustic target recognition method based on MFCC and RACNN[J]. Sensors, 2024, 24(1): 273.

[12] Xie Y, Ren J, Xu J. Adaptive ship-radiated noise recognition with learnable fine-grained wavelet transform[J]. Ocean engineering, 2022, 265: 112626.

[13] Feng S, Zhu X. A transformer-based deep learning network for underwater acoustic target recognition[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.

[14] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[15] Sun L, Dong J, Tang J, et al. Spatially-adaptive feature modulation for efficient image super-resolution [C]// Proceedings of the IEEE/CVF international conference on computer vision. 2023: 13190-13199.

[16] Chen L, Luo X, Zhou H, et al. Underwater acoustic multi-target recognition based on channel attention mechanism[J]. Ocean Engineering, 2025, 315: 119841.

[17] Yang S, Jin A, Zeng X, et al. Underwater acoustic target recognition based on sub-band concatenated Mel spectrogram and multidomain attention mechanism[J]. Engineering Applications of Artificial Intelligence, 2024, 133: 107983.

[18] Jacobs R A, Jordan M I, Nowlan S J, et al. Adaptive mixtures of local experts[J]. Neural computation, 1991, 3(1): 79-87.

[19] Shazeer N, Mirhoseini A, Maziarz K, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer [J]. arXiv preprint arXiv:1701.06538, 2017.

[20] Xie Y, Ren J, Xu J. Unraveling complex data diversity in underwater acoustic target recognition through convolution-based mixture of experts[J]. Expert Systems with Applications, 2024, 249: 123431.

[21] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[22] Donoho D L. De-noising by soft-thresholding[J]. IEEE transactions on information theory, 2002, 41(3): 613-627.

[23] Santos-Domínguez D, Torres-Guijarro S, Cardenal-López A, et al. ShipsEar: An underwater vessel noise database[J]. Applied Acoustics, 2016, 113: 64-69.

[24] Irfan M, Jiangbin Z, Ali S, et al. DeepShip: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification[J]. Expert Systems with Applications, 2021, 183: 115270.

[25] Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation[C]//Australasian joint conference on artificial intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006: 1015-1021.

[26] Kraemer H C. Kappa coefficient[J]. Wiley StatsRef: statistics reference online, 2014: 1-4.

[27] Wang B, Zhang W, Zhu Y, et al. An underwater acoustic target recognition method based on AMNet[J]. IEEE Geoscience and Remote Sensing Letters, 2023, 20: 1-5.

[28] Yu W, Zhou P, Yan S, et al. Inceptionnext: When inception meets convnext[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 5672-5683.