Ocean Engineering 343 (2026) 123234

Contents lists available at ScienceDirect

Ocean Engineering

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

Research paper

Lightweight underwater acoustic time-frequency separation network for efficient marine target recognition

Menghan Chen [a], Yuchen Lu [a], Liangliang Cheng [c], Rongxin Zhu [d], Kun Tao [b], Yifei Li [b,*], Magd Abdel Wahab [e,f]

[a] Yantai Research Institute, Harbin Engineering University, Yantai, 264005, China
[b] School of Engineering, Huzhou University, Huzhou, 313000, China
[c] Engineering and Technology Institute Groningen, Faculty of Science and Engineering, University of Groningen, Nijenborgh 4, 9747 AG, Groningen, the Netherlands
[d] Hainan University, Haikou, 570028, China
[e] Soete Laboratory, Department of Electrical Energy, Metals, Mechanical Constructions, and Systems, Faculty of Engineering and Architecture, Ghent University, Gent, Belgium
[f] Parallel and Distributed Systems Laboratory, Jožef Stefan Institute, Jamova Cesta 39, Ljubljana, SI-1000, Slovenia

ARTICLE INFO

Keywords:
Underwater acoustic target recognition
Time-frequency separation convolutions
Marine environment monitoring
Lightweight neural networks
Ocean engineering

ABSTRACT

Underwater acoustic target recognition is a core technology in ocean engineering, essential for marine security, underwater navigation, and resource development, with growing demand driven by expanding marine activities and autonomous platforms. Traditional underwater acoustic target recognition methods process time-frequency spectrograms using standard two-dimensional convolutions, failing to exploit signal physical properties due to coupled time-frequency processing. This paper presents our proposed lightweight Underwater Acoustic Time-Frequency Separation Network UATFSN implementing decoupled modeling through our designed Time-Frequency Separate Convolutions TFSC modules. TFSC splits feature maps into frequency path and time path, employing anisotropic convolution kernels respectively to extract harmonic features in frequency dimension and transient features in time dimension, incorporating adaptive residual normalization for enhanced stability. Experimental results demonstrate UATFSN achieves 98.23 % accuracy on DeepShip dataset, surpassing the strongest baseline, ResNet34, by 2.04 percentage points while requiring only 0.05M parameters, representing 99.77 % reduction compared to ResNet34. On ShipsEar dataset, UATFSN attains 98.75 % accuracy with 99.26 % computational reduction. Robustness evaluation confirms 84.20 % accuracy under 0 dB SNR and 78.17 % accuracy using 5 % training data. Our proposed UATFSN represents the first breakthrough achieving simultaneous accuracy and computational efficiency in underwater acoustic recognition, providing critical technological support for intelligent deployment on resource-constrained marine platforms. Our proposed UATFSN achieves a remarkable balance between accuracy and computational efficiency in underwater acoustic recognition, providing an effective technological solution for intelligent deployment on resource-constrained marine platforms.

1. Introduction

Underwater acoustic target recognition (UATR) represents one of the core technologies in ocean engineering (Luo et al., 2023; Tian et al., 2023a; Yao et al., 2024), which identifies target types through the analysis of underwater acoustic signals and plays a crucial role in key applications such as marine security monitoring, marine resource development, and marine environmental protection (Feng et al., 2024; Müller et al., 2024; Zhu et al., 2023a). With the rapid development of the global marine economy and the continuous expansion of marine space utilization, the application of underwater unmanned platforms in ocean engineering has become increasingly widespread, leading to an urgent and growing demand for efficient and reliable underwater acoustic target recognition technologies (Lu et al., 2025a; Xie et al., 2022). These ocean engineering platforms require real-time and accurate target recognition capabilities when performing tasks such as seabed resource exploration, marine structure inspection, and marine environmental monitoring to ensure operational safety and efficiency (Zhang et al.,

* Corresponding author.
E-mail address: 03354@zjhu.edu.cn (Y. Li).

https://doi.org/10.1016/j.oceaneng.2025.123234
Received 11 August 2025; Received in revised form 5 October 2025; Accepted 17 October 2025
Available online 21 October 2025

2023, 2025; Liu et al., 2024; Li et al., 2023).

In recent years, deep learning has become an important driving force for the advancement of UATR technology. Among various approaches, a mainstream method involves converting one-dimensional raw acoustic signals into two-dimensional time-frequency spectrograms, followed by feature extraction and classification using convolutional neural networks (CNNs) (Jiang et al., 2020; Yao et al., 2023; Lin et al., 2024; Xu et al., 2023; Zhu and Zhao, 2024; Zhu et al., 2024a). This method has been widely adopted in both academia and industry due to its intuitive nature and effectiveness (Zhou et al., 2024; Tian et al., 2021; Lu et al., 2025b; Zhu and Zhao, 2023). However, this approach suffers from representational limitations—standard two-dimensional convolution processes time and frequency dimensions in a unified isotropic manner, failing to adequately consider the differences in their physical meanings (Li and Yang, 2024; Chen et al., 2025; Ma et al., 2022). The frequency dimension primarily reflects steady-state characteristics such as the harmonic structure of targets, while the time dimension captures transient variations in signal intensity (Luo et al., 2021; Xie et al., 2024; Liu et al., 2025a). This difference in physical dimensions indicates that ideal feature extraction should employ an anisotropic processing mechanism, whereas the coupled processing approach of existing methods fails to fully exploit this characteristic. It is worth noting that while techniques exist in deep learning to factorize a 2D convolution into two sequential 1D convolutions for computational efficiency, such methods are fundamentally an approximation of a standard 2D receptive field. Their motivation is to reduce computational cost and they fail to address the core issue of physical heterogeneity in time-frequency features.

To improve recognition accuracy, researchers typically construct deeper and more complex network models, but this exacerbates the contradiction between model performance and ocean engineering deployment requirements (Feng and Zhu, 2022; Du et al., 2024; Li and Yang, 2021). In practical ocean engineering applications, underwater unmanned platforms face strict energy consumption constraints, limited computational resources, and harsh marine environment challenges. The high computational power and energy consumption demands of complex models form a sharp contrast with the resource constraints of these ocean engineering platforms, severely restricting the deployment of advanced recognition technologies in actual marine operations (Jin et al., 2025; Dong et al., 2025; Yang et al., 2024a; Zhao et al., 2025; Liu et al., 2025b). The traditional approach of trading complexity for accuracy focuses mainly on optimization at the model scale. It often overlooks the physical characteristics of underwater acoustic signal processing in the time–frequency domain, especially when viewed from the perspective of the unique requirements of ocean engineering applications (Zhu et al., 2023b, 2024b; Xu et al., 2024; Liu et al., 2025c). Therefore, it is urgent to re-examine the time–frequency domain processing mechanism from the perspective of architectural design. At the same time, efficient feature extraction methods need to be developed. These methods should reflect the physical characteristics of underwater acoustic signals and satisfy the deployment requirements of ocean engineering (Zeng et al., 2025; Xu et al., 2025; Lyu et al., 2024).

To address the aforementioned key technical challenges in ocean engineering applications, this paper proposes a lightweight underwater acoustic time-frequency separation network (UATFSN). This network is specifically designed for the deployment requirements of ocean engineering platforms, with its core innovation being the time-frequency separation convolution (TFSC) module, which effectively addresses the representational limitations of traditional two-dimensional convolution by decomposing feature processing into independent temporal and frequency paths and employing anisotropic convolutional kernels for decoupled modeling. This architecture achieves effective representation of underwater acoustic time-frequency characteristics while significantly reducing computational load, making it particularly suitable for resource-constrained underwater platform applications in ocean engineering. Additionally, the introduced adaptive residual normalization (ARN) mechanism further enhances the model's training stability and generalization capability under complex marine environments. The design philosophy of UATFSN reflects the urgent need for efficient, reliable, and practical technological solutions in ocean engineering, providing important technical support for the intelligent upgrade of underwater unmanned platforms. The main contributions of this paper are as follows.

1. A novel time-frequency separation convolution architectural module (TFSC) is proposed, which introduces a parallel dual-path processing paradigm. By splitting channels, this architecture decouples feature extraction into a dedicated frequency path for processing harmonic features and a time path for capturing transient features, thereby providing a modeling mechanism that is highly consistent with the physical characteristics of the signal.
2. An end-to-end lightweight network UATFSN oriented toward ocean engineering applications is constructed, which achieves excellent recognition performance while significantly reducing computational overhead. It achieves 98.23 % accuracy on the DeepShip dataset with only 0.05M parameters, representing a 99.77 % reduction in parameters compared to ResNet34, providing a practical technological solution for resource-constrained ocean engineering platforms.
3. Through comprehensive experimental validation on the DeepShip and ShipsEar datasets, the effectiveness and robustness of the proposed method under complex marine environmental conditions are demonstrated, including stable performance under different signal-to-noise ratios and training data proportion conditions, showcasing its deployment potential in practical ocean engineering applications.

2. Related work

1.1. Lightweight network architectures design

Lightweight networks reduce model parameters and computational complexity through structural optimization to meet edge device deployment requirements (Hu et al., 2025). Representative architectures such as MobileNet (Howard et al., 2017) and ShuffleNet (Zhang et al., 2018) employ techniques including depthwise separable convolutions and channel shuffling, successfully achieving significant improvements in computational efficiency. EfficientNet (Tan and Le, 2019) further optimizes network resource allocation by introducing a compound scaling strategy, realizing higher computational efficiency.

However, the optimization strategies of these general-purpose lightweight architectures primarily focus on improving the computational efficiency of traditional two-dimensional convolutions, with their core operators still based on isotropic convolution operations. Such models treat time-frequency spectrograms of underwater acoustic signals as ordinary images (Xu et al., 2024; Wu et al., 2024), employing uniform convolution kernels to simultaneously process temporal and frequency features, thereby neglecting the fundamental physical differences between these two dimensions. This design fails to effectively leverage the inherent physical heterogeneity prior of underwater acoustic signals, resulting in limited feature extraction efficiency. The design objectives of existing architectures emphasize "general computational efficiency" rather than "modeling efficiency for physical characteristics of underwater acoustic signals", making it challenging to achieve the high precision requirements of underwater acoustic tasks under extremely low computational overhead.

1.2. Time-frequency modeling in acoustic recognition

In the field of professional acoustic processing, researchers have explored more sophisticated time-frequency modeling techniques (Li et al., 2025; Yan et al., 2024). Convolutional Recurrent Neural Networks attempt to capture spatial features and temporal dependencies in time-frequency spectrograms by cascading Convolutional Neural Networks with Recurrent Neural Networks. However, since the front-end

Convolutional Neural Network has already performed coupled processing of the temporal and frequency dimensions, the subsequent Recurrent Neural Network can only conduct temporal modeling based on mixed features, making it difficult to effectively decouple and finely characterize the harmonic structures unique to the frequency dimension. This inherent processing sequence limits the method's capability to represent the complex time-frequency structures of underwater acoustic signals.

In recent years, Transformer (Vaswani et al., 2017) and Conformer (Gulati et al., 2020) architectures based on self-attention mechanisms have achieved significant breakthroughs in the field of acoustic recognition (Gong et al., 2021). These architectures effectively capture long-range contextual information through global attention mechanisms. However, their performance advantages come at the cost of extremely high computational overhead, with computational complexity increasing quadratically with input sequence length. For resource-constrained environments such as underwater unmanned platforms with strict energy consumption limitations and real-time requirements, the computational burden of such methods is prohibitive. Although their design objectives aim to "maximize recognition performance", this is achieved at the expense of "computational efficiency", which is highly detrimental to practical deployment in marine engineering.

In summary, existing methods either lack the capability to effectively decouple the time-frequency physical characteristics of underwater acoustic signals, or are unable to be deployed on practical marine engineering platforms due to excessive computational complexity. To overcome the structural contradiction between "physical modeling efficiency" and "deployment feasibility", a novel architectural design is urgently needed. The UATFSN proposed in this paper introduces time-frequency separable convolutions to achieve feature decoupling from a physics-driven perspective, effectively addressing this core challenge with minimal computational cost.

2. Proposed method

2.1. Underwater acoustic signal processing and feature extraction

The key to underwater acoustic target recognition lies in effectively extracting target features from complex marine environments. Unlike terrestrial acoustic signals, underwater acoustic signals exhibit significant non-stationary characteristics in the time-frequency domain, with spectral structures that change dynamically over time and are subject to complex influences from multipath propagation, Doppler effects, and marine environmental noise. To convert raw underwater acoustic signals $s(t) \in \mathbb{R}^N$ into two-dimensional feature representations suitable for deep learning processing, this paper employs a standard time-frequency analysis pipeline: first generating a time-frequency matrix $S \in \mathbb{C}^{F\times T}$ via Short-Time Fourier Transform (STFT), then applying Mel filter banks to convert linear frequency to Mel scale, and finally employing logarithmic compression to enhance dynamic range. The Mel-spectrogram representation provides higher resolution in low-frequency regions, a characteristic that matches the physical property that ship radiated noise energy is primarily concentrated in low-frequency bands, providing effective input representation for subsequent network feature extraction.

2.2. Time-Frequency Separate Convolutions

2.2.1. Theoretical foundation and motivation

Traditional two-dimensional convolutional neural networks process spectrograms using unified convolutional kernels that capture information from both the time and frequency dimensions simultaneously. However, from the perspective of signal physics, the time and frequency dimensions carry fundamentally distinct types of information. The frequency dimension primarily reflects the stable harmonic structures generated by mechanical sources such as a ship's engine, while the time dimension captures transient dynamic changes caused by propeller rotation or shifts in the target's operational state.

Our proposed method does not ignore the inherent relationship between time-frequency features; rather, it aims to learn this relationship more efficiently. By designing specialized anisotropic convolutional kernels for the time and frequency dimensions, the network first extracts more refined harmonic and transient features. These precisely extracted features are then fused, enabling the deeper layers of the network to focus on modeling the complex dependencies between them. This structured approach, in contrast to methods like Transformers that learn all relationships from raw data, embeds clear physical prior knowledge into the architecture. This significantly reduces the model's learning difficulty and computational requirements, making high-precision recognition feasible even on resource-constrained platforms.

Specifically, the time-frequency characteristics of underwater acoustic signals can be represented as:

$$s(t) = \sum_{k=1}^{K} A_k(t)\cos(\phi_k(t)) \tag{1}$$

where $A_k(t)$ is the time-varying amplitude of the $k-th$ frequency component, and $\phi_k(t)$ is the time-varying phase. This representation reveals the dynamic nature of underwater acoustic signals in the time dimension and their structural characteristics in the frequency dimension. The limitation of traditional two-dimensional convolution lies in its fixed receptive field shape being unable to adapt to this time-frequency domain heterogeneity.

To achieve the aforementioned physically meaningful time-frequency decoupling, this paper proposes the Underwater Acoustic Time-Frequency Separation Network (UATFSN). To realize this, we introduce the TFSC module. It is important to emphasize that the design philosophy of TFSC fundamentally differs from conventional convolution factorization techniques. The latter typically utilizes a serial structure, performing a $1 \times k$ convolution followed by a $k \times 1$ convolution to approximate a $k \times k$ convolution, with the primary motivation being computational efficiency. In contrast, our TFSC module implements a parallel dual-path architecture. This design is not aimed at approximating any existing convolution but is instead driven by physical priors: by splitting feature channels into two dedicated processing pathways, it explicitly and independently models harmonic structures and transient variations. This structural shift from "efficiency approximation" to "physical modeling", and from "serial" to "parallel", is the core innovation of our method.

2.2.2. TFSC module design

The TFSC module splits the input feature map along the channel dimension into two independent processing branches, corresponding to the frequency path and time path respectively. This design enables the model to employ anisotropic convolutional kernels for specialized feature extraction targeting different characteristics of the time-frequency domain, achieving precise modeling of the physical differences in the time-frequency domain of underwater acoustic signals.

Let the input feature map be $X \in \mathbb{R}^{C\times F\times T}$, where C, F, and T represent the number of channels, frequency dimension, and time dimension, respectively. First, channel shuffle operations are employed to enhance information interaction between different channels, followed by precise splitting along the channel dimension:

$$X^{(f)} = X[0 : C/2, :, :] \in \mathbb{R}^{C/2\times F\times T} \tag{2}$$

$$X^{(t)} = X[C/2 : C, :, :] \in \mathbb{R}^{C/2\times F\times T} \tag{3}$$

The frequency path specifically processes frequency-domain features, employing anisotropic 3×1 depth convolution kernels to extract correlations between adjacent frequency components:

$$Z^{(f)} = \mathrm{DepthConv}_{3\times1}\left(X^{(f)}\right) \tag{4}$$

The depth convolution operation is performed independently within each channel, with its physical significance lying in capturing harmonic relationships and spectral structures within local frequency ranges.

$$V^{(f)} = \frac{1}{F}\sum_{i=1}^{F} Z^{(f)}[:, i, :] \tag{5}$$

This pooling operation is designed to generate a summary vector describing the signal's global temporal dynamics. The subsequent broadcast and residual connection then use this vector to adjust the original feature map, enabling the model to enhance specific frequency features synchronized with key temporal events based on the overall temporal context.

The compressed features learn nonlinear relationships between channels through 1×1 pointwise convolution:

$$\widehat{V}^{(f)} = \mathrm{PointConv}_{1\times1}\left(V^{(f)}\right) \tag{6}$$

The compressed features are expanded back to the original dimensions through broadcasting operations and residual connections are applied:

$$\widetilde{X}^{(f)} = X^{(f)} + \mathrm{Expand}\left(\widehat{V}^{(f)}\right) \tag{7}$$

where the broadcasting operation is defined as:

$$\mathrm{Expand}\left(\widehat{V}^{(f)}\right) = \widehat{V}^{(f)} \otimes \mathbf{1}_{F\times1} \tag{8}$$

Here $\otimes$ denotes tensor outer product, and $\mathbf{1}_{F\times1}$ is an all-ones tensor of size $F\times 1$. The broadcasting operation expands one-dimensional temporal features to two-dimensional time-frequency space, achieving injection of global temporal context information.

The time path adopts a symmetric processing approach, using anisotropic 1×3 depth convolution kernels to capture dynamic features in the time domain:

$$Z^{(t)} = \mathrm{DepthConv}_{1\times3}(X^{(t)}) \tag{9}$$

The global average pooling operation across the time dimension is as follows:

$$V^{(t)} = \frac{1}{T}\sum_{j=1}^{T} Z^{(t)}[:, :, j] \tag{10}$$

This operation aims to generate a summary vector representing the signal's average spectral profile. This vector is then used in subsequent processing to adjust the original feature map, thereby highlighting significant transient events that are highly discriminative, based on the signal's long-term spectral statistics.

The final output of the time path is obtained after pointwise convolution and broadcasting operations:

$$\widehat{V}^{(t)} = \mathrm{PointConv}_{1\times1}(V^{(t)}) \tag{11}$$

$$\widetilde{X}^{(t)} = X^{(t)} + \mathrm{Expand}(\widehat{V}^{(t)}) \tag{12}$$

where $\mathrm{Expand}(\widehat{V}^{(t)}) = \widehat{V}^{(t)} \otimes \mathbf{1}_{1\times T}$.

Finally, the outputs of the frequency path and time path are concatenated along the channel dimension to form a complete feature representation:

$$Y = \mathrm{Concat}\left(\left[\widetilde{X}^{(f)}, \widetilde{X}^{(t)}\right]\right) \tag{13}$$

This separation-fusion design paradigm based on anisotropic convolutions enables the model to independently model time-frequency domain features, significantly improving the representational capability for complex time-frequency structures of underwater acoustic signals.

2.2.3. Network architecture

UATFSN adopts a deep convolutional neural network architecture specifically optimized for underwater acoustic target recognition tasks. This convolutional neural network follows the progressive feature extraction principle from low-level features to high-level semantics, achieving a balance between accuracy and efficiency through reasonable hierarchical structure and channel configuration.

UATFSN begins feature extraction with two 3×3 convolutional layers, both with stride 2, achieving initial feature extraction and spatial dimension compression:

$$H_1 = \mathrm{Conv}_{3\times3}^{s=2}\left(X_{\mathrm{input}}\right) \tag{14}$$

$$H_2 = \mathrm{GroupConv}_{3\times3}^{s=2}(H_1) \tag{15}$$

where s denotes the convolution stride, and GroupConv represents group convolution operations. The first layer maps single-channel spectrograms to $\tau/2$ channel features, while the second layer employs group convolution to expand to 2τ channels, where τ is the base channel width parameter. This design provides a foundation for subsequent feature processing while maintaining computational efficiency.

UATFSN contains nine TFSC modules organized into four processing stages. The channel numbers follow a progressive expansion sequence of $2\tau\rightarrow\tau\rightarrow1.5\tau\rightarrow2\tau\rightarrow2.5\tau$, gradually enhancing feature representation capability while avoiding parameter redundancy. Between the first and second TFSC blocks, as well as between the second and third TFSC blocks, 2×2 max pooling layers are inserted to achieve intermediate feature downsampling, enhancing the convolutional neural network's capability to capture high-level semantic features

To enhance the training stability and generalization capability of the convolutional neural network, this paper introduces an Adaptive Residual Normalization (ARN) mechanism after each TFSC block:

$$\mathrm{ARN}(X) = \alpha\cdot\mathrm{BatchNorm}(X) + (1-\alpha)\cdot X \tag{16}$$

where $\alpha \in [0, 1]$ is a learnable parameter used to dynamically balance the relationship between normalization strength and original feature preservation. The theoretical basis of the ARN mechanism lies in the fact that batch normalization may excessively smooth feature distributions, losing detail information useful for classification, while original features retain important statistical characteristics. Through the learnable weight parameter α, UATFSN can adaptively select the optimal normalization degree. This mechanism improves training stability and convergence speed, performing well when processing complex underwater acoustic environment data.

The final stage of UATFSN employs a 1×1 convolutional layer to map high-dimensional features to dimensions corresponding to the number of target categories, followed by global average pooling operations to generate the final classification feature vector:

$$F_{\mathrm{final}} = \mathrm{GlobalAvgPool}(\mathrm{Conv}_{1\times1}(H_{\mathrm{last}})) \tag{17}$$

$$P = \mathrm{Softmax}\left(F_{\mathrm{final}}\right) \tag{18}$$

where P represents the predicted probability distribution for each underwater acoustic target category. Global average pooling reduces the number of parameters, improves model generalization capability, and effectively prevents overfitting phenomena.

2.2.4. Computational efficiency analysis

UATFSN achieves significant computational complexity optimization through separate convolution design. From a theoretical analysis perspective, traditional $k \times k$ two-dimensional convolution kernels require k^2 parameters, while the proposed combination of $k \times 1$ and $1 \times k$ one-dimensional convolution kernels requires only $2k$ parameters,

achieving a parameter compression ratio of $\frac{k^2}{2k} = \frac{k}{2}$. When the convolution kernel size $k = 3$, the parameter compression ratio is 1.5 times, and this advantage becomes increasingly pronounced as the convolution kernel size increases.

Combined with the adoption of depthwise separable convolutions, the computational overhead of UATFSN is further optimized. Let the input feature map size be $C \times H \times W$, the number of output channels be C', and the convolution kernel size be K. The floating-point operations (FLOPs) for standard convolution are:

$$\mathrm{FLOPs}_{\mathrm{standard}} = C \cdot C' \cdot K^2 \cdot H \cdot W \tag{19}$$

Depthwise separable convolution decomposes standard convolution into two steps: depthwise convolution and pointwise convolution. The computational complexity of depthwise convolution is:

$$\mathrm{FLOPs}_{\mathrm{depthwise}} = C \cdot K^2 \cdot H \cdot W \tag{20}$$

The computational complexity of pointwise convolution is:

$$\mathrm{FLOPs}_{\mathrm{pointwise}} = C \cdot C' \cdot H \cdot W \tag{21}$$

Therefore, the total computational complexity of depthwise separable convolution is:

$$\mathrm{FLOPs}_{\mathrm{separable}} = C \cdot K^2 \cdot H \cdot W + C \cdot C' \cdot H \cdot W \tag{22}$$

The computational complexity compression ratio relative to standard convolution is:

$$\frac{\mathrm{FLOPs}_{\mathrm{separable}}}{\mathrm{FLOPs}_{\mathrm{standard}}} = \frac{1}{C'} + \frac{1}{K^2} \tag{23}$$

When the number of output channels C' and convolution kernel size K are large, this compression ratio decreases significantly, demonstrating the computational advantages of depthwise separable convolution in large-scale convolutional neural networks.

Broadcasting operations further reduce the computational load of pointwise convolution through dimension compression and expansion mechanisms. Taking the frequency path as an example, global average pooling compresses the original $C/2 \times F \times T$ features to $C/2 \times 1 \times T$, reducing computational load by a factor of F. This design enhances inference speed while maintaining feature representation capability, making UATFSN suitable for resource-constrained underwater acoustic target recognition application scenarios.

Beyond computational efficiency improvements, another important advantage of this architectural design lies in its larger effective receptive field. Through time-frequency separation processing, UATFSN can capture richer time-frequency domain contextual information, thereby achieving more accurate target feature recognition in complex underwater acoustic environments. This design improves model recognition accuracy and enhances its robustness under different marine environmental conditions. Fig. 1 illustrates the proposed UATFSN network architecture.

3. Datasets and evaluations

3.1. Dataset description

This paper employs two publicly available underwater acoustic datasets for comprehensive evaluation experiments: the DeepShip dataset and the ShipsEar dataset. These datasets capture real radiated noise from different marine environments, providing authentic and diverse benchmarks for algorithm evaluation.

DeepShip (Irfan et al., 2021): The DeepShip dataset contains acoustic records from 265 different vessels, authentically capturing the radiated noise characteristics of ships. The dataset comprises 609 real recordings with a total duration of 47 h and 4 min. All recordings were constructed through sonar systems using single-channel sampling with a sampling frequency set at 32 kHz. The data originates from Ocean Networks Canada, covering cargo ships, passenger ships, tankers, and tugboats under different sea conditions and noise levels throughout the year. The dataset is characterized by the diversity and authenticity of its recordings, containing ship radiated noise under various marine environmental conditions, providing an important foundation for algorithm robustness evaluation. The data records cover real marine environments with different seasons, sea conditions, and noise levels, ensuring the representativeness and practicality of the dataset.

ShipsEar: The ShipsEar dataset collected real recordings of ship radiated noise from the Atlantic coastal waters of northwestern Spain during the period from autumn 2012 to summer 2013. The dataset contains 90 recordings from 11 different categories of vessels as well as one category of natural marine noise recordings. The dataset comprises 12 target categories, specifically: dredger, fish boat, motor boat, mussel boat, natural noise, ocean liner, passengers, pilot ship, RO-RO, sailboat, trawler, and tug boat. All audio recordings have a uniform sampling rate of 52,734 Hz. The unique feature of this dataset lies in its inclusion of natural marine noise as an interference category, providing important

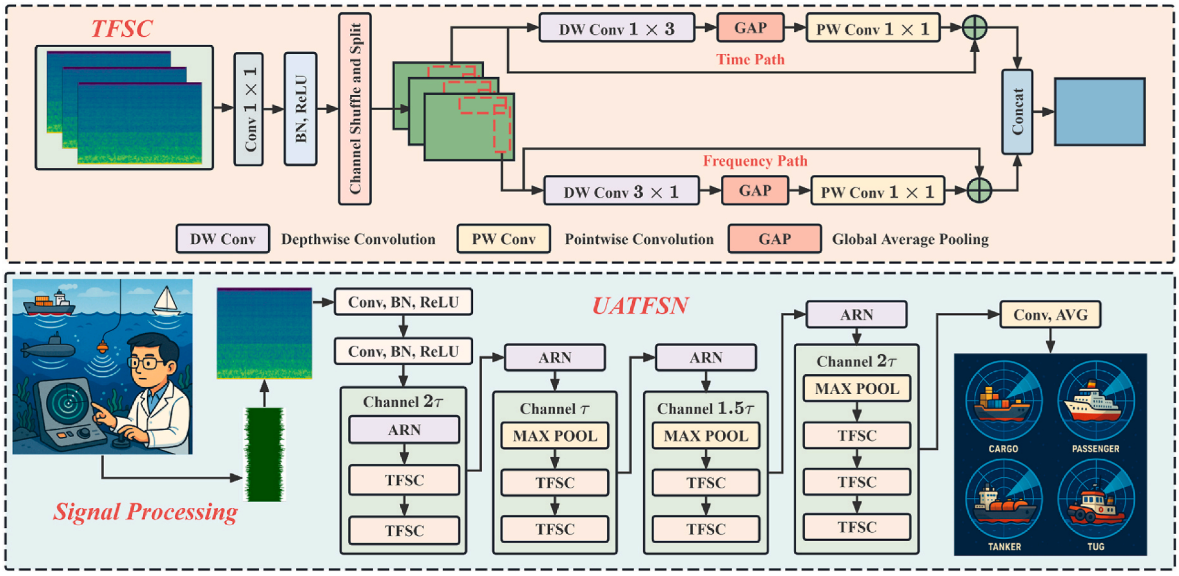


Fig. 1. The proposed UATFSN network architecture.

reference for evaluating algorithms' noise suppression capabilities in complex marine environments. The dataset covers various vessel types from small fishing boats to large passenger liners, providing comprehensive test benchmarks for multi-class classification algorithm performance evaluation.

3.2. Experimental setup

3.2.1. Hardware and software environment

Experiments were conducted on a computing platform equipped with an Intel Core i9-13980HX processor, NVIDIA GeForce RTX 4060 Laptop GPU, and 64 GB memory. The PyTorch 2.0.1 deep learning framework was employed with Python version 3.9.18 and CUDA version 11.8. Other key dependency libraries include: librosa 0.10.1 for audio processing, numpy 1.24.3 for numerical computation, and scikit-learn 1.3.0 for evaluation metric calculation.

3.2.2. Training configurations

This section provides detailed elaboration of the specific parameter settings employed in model training, as detailed in Table 1. These parameters are designed to ensure experimental reproducibility and result reliability. To mitigate class imbalance in the ShipsEar dataset, we employed a weighted cross-entropy loss, where class weights were set inversely proportional to their sample frequencies in the training set.

3.2.3. Data splitting strategy

To ensure the reliability and scientific validity of our experimental results, and to facilitate a fair and credible comparison with existing state-of-the-art methods, our strategies for data preprocessing, class grouping, and dataset splitting strictly adhere to the authoritative experimental setup established by Shuang Yang et al. (2024b). The specific strategies were as follows:

DeepShip dataset processing: Each original recording was divided into independent 3-s time frame segments, ultimately obtaining 56,468 acoustic samples. To ensure data independence and avoid the impact of recording duplication on experimental results, the same recording would not appear simultaneously in both training and test sets, thereby guaranteeing the scientific validity of dataset partitioning. The training and test sets were divided according to an 8:2 ratio.

ShipsEar dataset processing: The entire dataset was segmented according to 5-s durations, obtaining a total of 2223 sample segments. These segmented sample segments were treated as independent acoustic sample units. To facilitate classification research, the original 12 target categories were regrouped into four main categories: Class A (including fish boat, trawler, mussel boat, tug boat, and dredger), Class B (including motor boat, pilot boat, and sailboat), Class C (including passengers), and Class D (including ocean liner and RO-RO). The same 8:2 training-test set division ratio was adopted.

3.3. Evaluation metrics

This paper evaluates the UATFSN model from three dimensions: classification performance, uncertainty, and computational complexity. First, accuracy is employed to evaluate classification performance. Accuracy is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{24}$$

To ensure the reliability of experimental results, each experiment was executed independently 10 times, reporting arithmetic means to eliminate the effects of randomness. Based on this, standard deviation was used to quantify the stability of model performance:

$$\text{SD} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{25}$$

Furthermore, 95 % confidence intervals were calculated to assess statistical significance:

$$\text{CI}_{95\%} = \bar{x} \pm t_{0.025,9} \times \frac{\text{SD}}{\sqrt{n}} \tag{26}$$

where $n = 10$ and $t_{0.0,259} = 2.262$. Narrower confidence intervals indicate more stable model performance and more reliable results.

Besides performance metrics, computational complexity analysis is equally important. This paper employs parameter count and floating-point operations (FLOPs) to evaluate the computational efficiency of the model. Parameter count statistics include all trainable parameters in the network:

$$\text{Parameters} = \sum_{l=1}^{L}\left|\theta_l\right| \tag{27}$$

The FLOPs for convolutional layers are calculated as:

$$\text{FLOPs}_{\text{conv}} = K_h \times K_w \times C_{\text{in}} \times C_{\text{out}} \times H_{\text{out}} \times W_{\text{out}} \tag{28}$$

All complexity metrics were measured under a unified hardware environment, while also recording model inference time and memory usage to comprehensively evaluate the model's practicality.

Table 1
Proposed training Configuration.

Parameter	Value
Number of Mel Bins	128
Number of MFCCs	13
FFT Window Size	2048
Batch Size	32
Learning Rate	0.01
Epochs	80
Loss Function	Cross-Entropy (Deepship)/Weighted Cross-Entropy (ShipsEar)

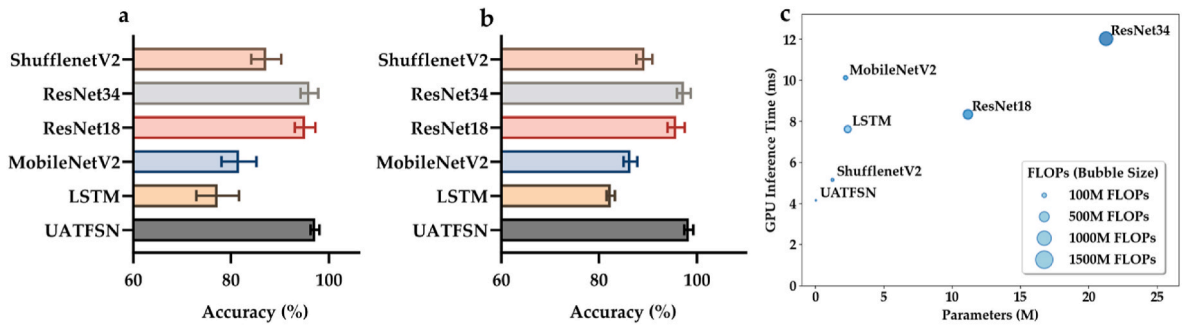
4. Results and discussions

4.1. Comparison with baseline models

To comprehensively evaluate the performance advantages of UATFSN, this paper selected five representative deep learning models as baseline comparisons: ResNet34, ResNet18, MobileNetV2, LSTM, and ShuffleNetV2. These models cover different architectural types including convolutional neural networks, recurrent neural networks, and lightweight networks, providing comprehensive comparison benchmarks for UATFSN performance evaluation. All comparison models were based on Mel-spectrograms for feature extraction, employing the same data preprocessing pipeline and training configurations to ensure fairness and comparability of experimental results.

As shown in Fig. 2(a), on the DeepShip dataset, UATFSN achieved a recognition accuracy of 97.26 %, significantly outperforming the sequence-modeling-based LSTM's 77.82 %, MobileNetV2's 81.60 %, and ShuffleNetV2's 87.16 %. Notably, UATFSN also demonstrated certain performance advantages compared to ResNet18's 95.13 % and ResNet34's 96.19 %, with improvements of 2.13 and 1.07 percentage points, respectively. On the ShipsEar dataset, as illustrated in Fig. 2(b), UATFSN achieved a recognition accuracy of 98.32 %, likewise surpassing all comparison models, with improvements of 15.64, 11.90, 2.60, 1.01, and 9.07 percentage points compared to LSTM, MobileNetV2, ResNet18, ResNet34, and ShuffleNetV2, respectively.

In terms of computational complexity analysis, the bubble chart in Fig. 2(c) clearly demonstrates the comprehensive performance of each model across three dimensions: parameter count, floating-point operations, and GPU inference time. UATFSN achieved exceptional recognition performance with only 0.05M parameters, representing a 99.77 % reduction compared to ResNet34's 21.28M parameters and a 99.55 %



d

Deepship	STD	95 CI	Significance
UATFSN	0.89	(96.62, 97.90)	
LSTM	4.35	(74.16, 80.39)	****
MobileNetV2	3.57	(79.05, 84.16)	****
ResNet18	2.11	(93.62, 96.64)	ns
ResNet34	2.09	(94.70, 97.68)	ns
ShufflenetV2	3.06	(84.97, 89.35)	****

e

ShipsEar	STD	95 CI	Significance
UATFSN	0.91	(97.67, 98.97)	
LSTM	0.91	(82.02, 83.33)	****
MobileNetV2	1.41	(85.42, 87.43)	****
ResNet18	1.76	(94.45, 96.98)	ns
ResNet34	1.43	(96.29, 98.33)	ns
ShufflenetV2	1.63	(88.08, 90.42)	****

Fig. 2. Comprehensive performance comparison between UATFSN and baseline models. (a) Accuracy on DeepShip dataset. (b) Accuracy on ShipsEar dataset. (c) Efficiency analysis showing parameters (x-axis), inference time (y-axis), and FLOPs (bubble size). (d–e) Statistical analysis with standard deviation, 95 % confidence intervals, and ANOVA significance tests on DeepShip and ShipsEar datasets, respectively. Error bars represent standard deviation across 10 runs. $p < 0.0001$, $p < 0.001$, ns: not significant.

reduction compared to ResNet18's 11.17M parameters. Regarding floating-point operations, UATFSN required only 13.06M operations, significantly lower than ResNet34's 1755.49M and ResNet18's 849.44M operations, achieving reductions of 99.26 % and 98.46 %, respectively. GPU inference time testing revealed that UATFSN's inference time was 4.14ms, notably faster than ResNet34's 12.02ms and ResNet18's 8.33ms, representing inference speed improvements of 2.90 × and 2.01 × , respectively.

To evaluate the stability and statistical significance of model performance, each model was independently executed 10 times on both datasets. Fig. 2(d) and (e) respectively provide the standard deviation, 95 % confidence intervals, and ANOVA statistical analysis results comparing each model with UATFSN on the DeepShip and ShipsEar datasets. Statistical analysis demonstrated that UATFSN possessed significant performance advantages over LSTM, MobileNetV2, and ShufleNetV2, with significance levels all reaching $p < 0.001$. Although the performance differences with ResNet18 and ResNet34 did not reach statistical significance levels, UATFSN achieved substantial reductions in parameter count and computational complexity while maintaining comparable recognition accuracy, demonstrating clear efficiency advantages.

Comprehensive performance analysis demonstrates that UATFSN exhibits exceptional overall performance in underwater acoustic target recognition tasks. The model not only leads most baseline models in recognition accuracy but, more importantly, achieves significant breakthroughs in computational efficiency. Compared to traditional ResNet series models, UATFSN attained comparable or even superior recognition performance with extremely low parameter count and computational overhead, making it particularly suitable for deployment scenarios on underwater platforms with limited computational resources. Compared to lightweight models MobileNetV2 and ShufleNetV2, UATFSN significantly improved recognition accuracy while further compressing model size, fully validating the effectiveness of the time-frequency separation convolution architecture in underwater acoustic signal processing.

4.2. Comparison with SOTA methods

To validate the advancement of UATFSN, this paper compared it with state-of-the-art methods on both datasets. These methods represent the latest technological levels in the current underwater acoustic target recognition field. On the DeepShip dataset, as shown in Table 2, UATFSN achieved a recognition accuracy of 98.23 %, with improvements of 7.63, 20.70, 15.29, and 18.73 percentage points compared to CFTANet, SCAE, MSLEFC, and LMSRDN, respectively, validating the superiority of the time-frequency separation convolution architecture. In terms of parameter efficiency, UATFSN achieved the highest accuracy using only 0.05M parameters, representing an 89 %–99 % reduction in parameter count compared to other methods. Regarding computational efficiency, UATFSN's floating-point operations were only 13.06M, achieving 88 %–99 % reductions compared to other methods.

On the ShipsEar dataset, as shown in Table 3, UATFSN achieved a recognition accuracy of 98.75 %, which, although slightly lower than AMNet's 99.40 %, showed improvements of 4.15 and 6.85 percentage points compared to CRNN and ConvNeXt, respectively. More importantly, UATFSN achieved significant efficiency advantages while approaching optimal performance. The parameter count was only 0.05M, representing an 88 %–99 % reduction compared to other methods. The floating-point operations were 13.06M, achieving 94 %–99 % reductions compared to other methods.

Comprehensive comparison results demonstrate that UATFSN exhibits exceptional overall performance in underwater acoustic target recognition. The model achieved substantial reductions in parameter count and computational complexity while maintaining high recognition accuracy. Particularly on the DeepShip dataset, UATFSN comprehensively surpassed all methods in recognition accuracy and achieved order-of-magnitude improvements in computational efficiency. On the ShipsEar dataset, UATFSN achieved near-optimal performance with extremely low computational overhead, making it particularly suitable for deployment on resource-constrained underwater platforms. These results fully validate the effectiveness of the time-frequency separation convolution architecture.

Table 2
Comparison with SOTA methods in deepship dataset.

	UATFSN	CFTANet (Yang et al., 2024b)	SCAE (Irfan et al., 2021)	MSLEFC (Zhang and Zeng, 2023)	LMSRDN (Tian et al., 2023b)
Accuracy	98.23 %	90.60 %	77.53 %	82.94 %	79.50 %
Param	0.05M	0.47M	1.80M	15.75M	11.74M
Flops	13.06M	118.61M	6001.00M	1559.60M	1397M

Table 3
Comparison with SOTA methods in ShipsEar dataset.

	UATFSN	CRNN (Liu et al., 2021)	AMNet (Wang et al., 2023)	ConvNeXt (Liu et al., 2022)
Accuracy	98.75 %	94.60 %	99.40 %	91.90 %
Param	0.05M	0.45M	5.47M	26.51M
Flops	13.06M	250.12M	370.00M	4590.00M

4.3. Comparison with advanced lightweight networks

To further validate the advantages of UATFSN, this paper compared it with several advanced lightweight networks, including ConvMixer (Trockman and Kolter, 2022), EfficientNet (Tan and Le, 2019), Conformer (Gulati et al., 2020), and Audio Spectrogram Transformer (Gong et al., 2021). These methods represent different technical approaches in current lightweight deep learning and spectral signal processing. As shown in Table 4, UATFSN comprehensively outperforms all comparison methods in accuracy while maintaining extremely low computational overhead. Notably, UATFSN's parameter count and computational complexity are significantly lower than all compared models, fully demonstrating the efficiency advantages of the time-frequency separation convolution architecture in underwater acoustic target recognition tasks (see Table 5).

4.4. Deployment-oriented performance analysis

To systematically assess the deployment feasibility of UATFSN in practical engineering environments, this section presents a quantitative analysis of its inference latency, memory usage, and performance after INT8 quantization on a single-core CPU. All metrics were measured in a strictly controlled single-thread, single-core environment to ensure the accuracy and relevance of the evaluation. The baseline model of UATFSN is implemented in FP32 format, demonstrating efficient computational performance and fine resource control. As shown in Table 6, the average inference latency on the CPU is 4.783 ms, which meets the requirements for real-time processing applications. In terms of memory, the peak runtime memory usage of the model is 11.23 MB, consisting of 0.20 MB for static model parameters and 10.93 MB for peak activation memory. Through effective control of activation memory, UATFSN is able to operate within the constraints of embedded hardware with very limited memory resources. To validate its compatibility with standard deployment optimization processes, we performed INT8 quantization on the model. After quantization, the storage size of the model parameters was reduced to 0.05 MB, achieving a fourfold reduction. Additionally, the inference latency was further reduced to 4.675 ms. As a trade-off for the efficiency gains, the model's accuracy was slightly affected: the accuracy on the DeepShip and ShipsEar datasets decreased by 1.08 and 1.27 percentage points, respectively, but remained high at 97.15 % and 97.48 %. This excellent balance between computational efficiency, memory usage control, and model accuracy confirms the feasibility and advanced nature of UATFSN for engineering applications.

Table 4
Comparison with advanced lightweight networks.

	UATFSN	ConvMixer	EfficientNet	Conformer	AST
Deepship Accuracy	98.23 %	91.84 %	93.12 %	87.72 %	93.84 %
ShipsEar Accuracy	98.75 %	92.48 %	93.40 %	90.84 %	95.12 %
Param	0.05M	0.71M	7.03M	7.19M	10.78M
Flops	13.06M	337.14M	194.90M	352.59M	955.33M

Table 5
Detailed Deployment Performance of UATFSN in FP32 vs. INT8 Precision.

Category	Metric	FP32 Model	INT8 Quantized Model
Storage	Model Performance (MB)	0.20	0.05
Memory	Peak Memory (MB)	11.23 (0.20 + 10.93)	11.18 (0.05 + 10.93)
Speed	Avg. Latency (ms)	4.783	4.675
Performance	Accuracy-Deepship	98.23	97.15
	Accuracy-ShipsEar	98.75	97.48

Table 6
Ablation study results on DeepShip and ShipsEar datasets.

Configuration	DeepShip Acc(%)	ShipsEar Acc(%)	Drop (DeepShip)	Drop (ShipsEar)
Full UATFSN	98.19	98.64	–	–
w/o TFSC	89.42	91.35	−8.77	−7.29
w/o Time Path	87.63	88.71	−10.56	−9.93
w/o Frequency Path	85.28	86.45	−12.91	−12.19
w/o Channel Shuffle	94.87	95.92	−3.32	−2.72
w/o ARN	96.35	97.18	−1.84	−1.46

4.5. Hyper-parameter analysis

To gain deeper understanding of the impact of temporal and frequency domain convolution kernel sizes in the TFSC module on model performance, this paper conducted sensitivity analysis on these two key hyperparameters. The experiments set temporal convolution kernel sizes as 1×1, 1×3, and 1×5, and frequency domain convolution kernel sizes as 1×1, 3×1, and 5×1, conducting comprehensive evaluation on both datasets.

As illustrated in Fig. 3, the choice of convolution kernel size significantly affects model performance. On the DeepShip dataset, the model achieves its optimal performance of 97.15 % with a 1×3 kernel in the time domain and a 3×1 kernel in the frequency domain. In contrast, smaller kernel combinations, such as 1×1, yield only 71.71 % accuracy, while larger combinations, like 1×5 and 5×1, achieve 89.77 % performance. This suggests that moderate kernel sizes offer the best balance between capturing local time-frequency features and avoiding overfitting.

Similar trends were observed on the ShipsEar dataset, with the optimal configuration also being temporal convolution kernel 1×3 and frequency domain convolution kernel 3×1, achieving 98.83 % recognition accuracy. Notably, the impact of frequency domain convolution kernel size was more significant. When the frequency domain convolution kernel increased from 1×1 to 3×1, both datasets showed substantial performance improvements, indicating that appropriate frequency domain receptive fields are crucial for capturing the spectral structural features of underwater acoustic signals.

The experimental results validated the rationality of the TFSC module design, demonstrating that the combination of 3×1 and 1×3 convolution kernels can effectively balance the accuracy and computational efficiency of time-frequency domain feature extraction, providing optimal feature representation capability for underwater acoustic target recognition.

4.6. Robustness evaluation

To validate the practicality of UATFSN in complex marine environments, this paper conducted comprehensive robustness evaluation of the model from two dimensions: noise robustness and data scarcity. To conduct the noise robustness test, we simulated background interference of varying intensities by adding Gaussian White Noise to the original

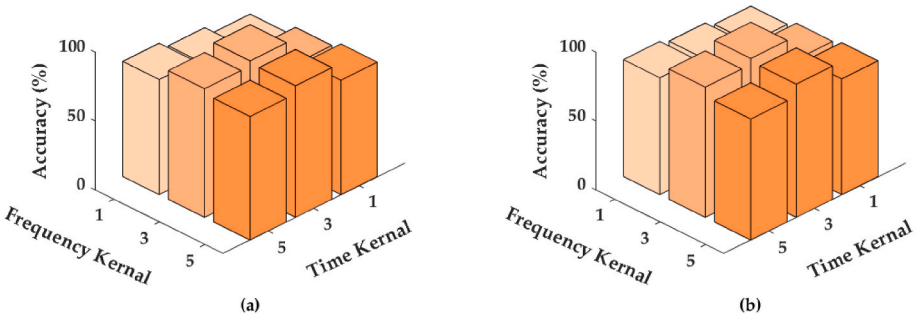


Fig. 3. Hyperparameter sensitivity analysis of TFSC module kernel sizes on (a) DeepShip and (b) ShipsEar da-tasets. Time kernel sizes: 1×1, 1×3, 1×5; Frequency kernel sizes: 1×1, 3×1, 5×1.

clean audio signals. The Signal-to-Noise Ratio (SNR) is defined as:

$$\mathrm{SNR}_{\mathrm{dB}} = 10 \log_{10}\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right) \tag{29}$$

where P_{signal} and P_{noise} are the average power of the signal and noise, respectively. For each original audio segment, we first calculated its signal power, P_{signal}. Then, for a target SNR (e.g., 0 dB), we calculated the required noise power as $P_{\mathrm{noise}} = P_{\mathrm{signal}}/10^{(\mathrm{SNR}_{\mathrm{dB}}/10)}$. A Gaussian White Noise signal of the same length was generated and scaled to match this target noise power before being added to the original signal. This procedure was used to create the test samples for each specified SNR level.

Noise robustness testing simulated background interference in real marine environments by adding Gaussian white noise of different intensities to the original signals, while data scarcity testing evaluated the model's generalization capability in small-sample scenarios by limiting the proportion of training data.

As shown in Fig. 4(a), UATFSN demonstrated good noise robustness under different signal-to-noise ratio conditions. Under extremely low signal-to-noise ratio of 0 dB, the model still achieved recognition accuracies of 84.20 % and 81.99 % on the DeepShip and ShipsEar datasets, respectively, demonstrating strong anti-noise capability. As the signal-to-noise ratio improved, model performance steadily enhanced, reaching recognition accuracies of 98.06 % and 98.19 % under 25 dB signal-to-noise ratio, approaching optimal performance under noise-free conditions. Notably, when the signal-to-noise ratio improved from 0 dB to 10 dB, both datasets showed significant performance improvements, indicating that UATFSN can effectively utilize useful information in signals to suppress noise interference.

As shown in Fig. 4(b), UATFSN demonstrated excellent generalization capability under different training data proportions. Even under extreme conditions using only 5 % training data, the model still achieved recognition accuracies of 78.17 % and 77.04 % on the DeepShip and ShipsEar datasets, respectively, proving the efficient feature learning capability of the time-frequency separation convolution architecture. As the proportion of training data increased, model performance continuously improved, reaching recognition accuracies of 96.48 % and 98.41 % under 40 % training data conditions. Particularly within the 10 %–20 % training data range, model performance improvement was most significant, indicating that UATFSN can rapidly converge and achieve good recognition performance with limited training samples.

The robustness evaluation results demonstrate that UATFSN possesses the capability to work stably in complex marine environments and data-constrained scenarios, providing reliable technical assurance for practical underwater acoustic target recognition applications.

4.7. Feature extraction analysis

To validate the advantages of Mel-spectrograms in underwater acoustic target recognition, this paper compared the performance of three mainstream time-frequency domain feature extraction methods: Short-Time Fourier Transform spectrograms (STFT), Mel-spectrograms (Mel), and Mel-Frequency Cepstral Coefficients (MFCC). These feature extraction methods represent different signal representation strategies, providing experimental basis for selecting optimal input features for the UATFSN architecture.

As shown in Fig. 5, Mel-spectrograms achieved optimal recognition performance on both datasets. On the DeepShip dataset, Mel-spectrograms achieved an average recognition accuracy of 98.02 %, representing improvements of 4.47 percentage points compared to STFT spectrograms' 93.55 % and 3.38 percentage points compared to MFCC's 94.64 %. On the ShipsEar dataset, Mel-spectrograms achieved an average recognition accuracy of 98.47 %, representing improvements of 4.53 percentage points compared to STFT spectrograms' 93.94 % and 4.66 percentage points compared to MFCC's 93.81 %.

The superior performance of Mel-spectrograms primarily stems from their alignment with human auditory perception characteristics. Mel

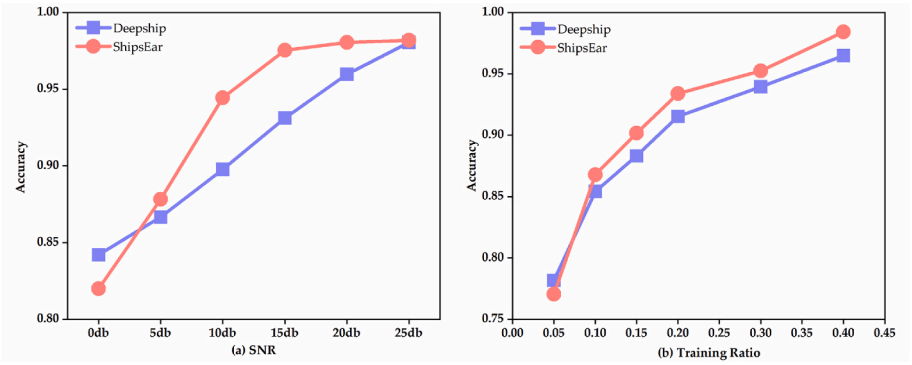


Fig. 4. Robustness evaluation of UATFSN on DeepShip and ShipsEar datasets. (a) Performance under different SNR conditions with Gaussian white noise. (b) Performance with different training data ratios.

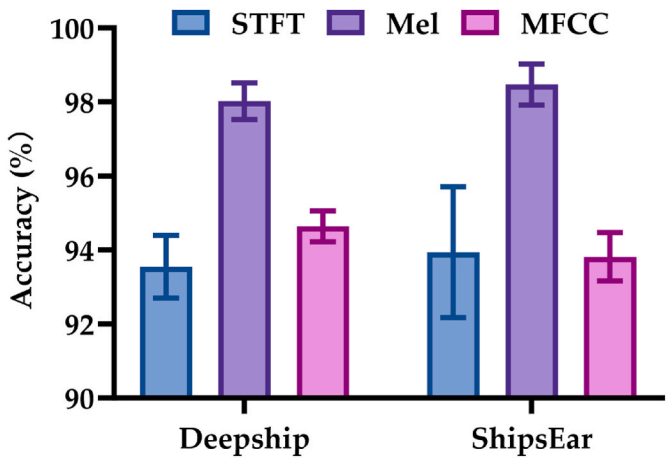


Fig. 5. Comparison of different feature extraction methods on DeepShip and ShipsEar datasets using UATFSN. STFT: Short-Time Fourier Transform spectrogram; Mel: Mel-spectrogram; MFCC: Mel-Frequency Cepstral Coefficients.

filter banks possess higher frequency resolution in low-frequency regions, a characteristic that highly matches the physical property that ship radiated noise energy is primarily concentrated in low-frequency bands. In contrast, STFT spectrograms employ linear frequency distribution with relatively low resolution in low-frequency regions, making it difficult to fully capture key features of ship radiated noise. Although MFCC is also based on Mel filter banks, its cepstral transformation process loses some time-frequency domain detail information, leading to slightly reduced recognition performance.

The experimental results demonstrate that Mel-spectrograms can provide the most suitable input feature representation for UATFSN, effectively enhancing the time-frequency separation convolution architecture's capability for underwater acoustic signal feature extraction, laying an important foundation for achieving high-precision underwater acoustic target recognition.

4.8. Ablation study

To validate the effectiveness of each component of UATFSN, this paper designed comprehensive ablation experiments. By systematically removing or replacing key components, the specific contribution of each module to model performance was analyzed, providing experimental support for the rationality of the architectural design.

As shown in Table 6, the removal of all components led to model performance degradation, fully validating the effectiveness of the UATFSN architectural design. Among these, the removal of the frequency domain path had the most severe impact on performance, causing performance drops of 12.91 % and 12.19 % on the DeepShip and ShipsEar datasets, respectively, indicating the crucial role of frequency domain feature extraction in underwater acoustic target recognition. The removal of the temporal path also caused substantial performance losses, with drops of 10.56 % and 9.93 %, respectively, proving the importance of temporal dynamic features.

Complete removal of the TFSC module resulted in significant performance drops of 8.77 % and 7.29 %, fully validating the tremendous advantages of the time-frequency separation convolution architecture compared to traditional two-dimensional convolution. This result indicates that specialized design targeting the time-frequency domain heterogeneity of underwater acoustic signals can significantly enhance feature extraction capability. The removal of channel shuffle operations caused performance drops of 3.32 % and 2.72 %, demonstrating the important contribution of enhancing information interaction between different channels to model performance.

The removal of the ARN mechanism caused performance losses of 1.84 % and 1.46 %, proving the important role of adaptive residual normalization in improving model training stability and recognition accuracy. This result indicates that dynamically balancing the relationship between normalization strength and original feature preservation through learnable parameters can further optimize model performance.

The ablation experiment results fully validate the necessity of each component of UATFSN, particularly the core role of the time-frequency separation mechanism. The collaborative work of frequency domain and temporal paths provides optimal feature representation for underwater acoustic target recognition, while other auxiliary components further enhance the overall performance of the model.

4.9. In-depth analysis and discussion

4.9.1. Interpretability analysis via Grad-CAM visualization

To thoroughly validate the TFSC module's capability in learning time-frequency physical features of underwater acoustic signals, we employ Gradient-weighted Class Activation Mapping to visualize the model's decision mechanism. Direct comparison between original Mel-spectrograms and attention response heatmaps intuitively reveals UATFSN's capacity for capturing time-frequency domain physical characteristics.

As shown in Fig. 6, visualization results on the DeepShip dataset demonstrate UATFSN's precise localization capability for critical time-frequency features. In cargo ship recognition, model attention concentrates on low-frequency steady-state harmonic regions, accurately corresponding to the physical characteristics of radiated noise from ship engines. Passenger ship recognition exhibits uniform broadband response patterns, precisely capturing stable broadband radiation features generated by complex power systems of large vessels. Tanker heatmaps show model focus on low-frequency continuous energy bands, closely aligning with spectral characteristics produced by propulsion systems during low-speed navigation. Tug visualization results reveal simultaneous capture of fundamental frequency components and mid-frequency impulse features, corresponding to transient impact noise from high-power propulsion systems. These findings confirm UATFSN's capability to adaptively extract discriminative time-frequency features targeting different vessel sound generation mechanisms.

As shown in Fig. 7, visualization results on the ShipsEar dataset further validate the model's feature learning capability. Fishing vessel class exhibits simultaneous strong responses in low-frequency and mid-frequency bands, reflecting bimodal spectral characteristics of small vessel engines and operational equipment. Motorboat class presents dispersed yet consistent response patterns across the entire time-frequency domain, corresponding to frequent transient variations caused by flexible maneuverability. Passenger class demonstrates broadband uniform response patterns consistent with the DeepShip dataset, revalidating robust recognition capability for steady-state broadband radiation features of passenger ships. Ocean liner class displays significant concentrated responses in mid-frequency bands, corresponding to high-energy mid-frequency components generated by large vessel propulsion systems. Although many categories present similar visual appearances in original Mel-spectrograms, Grad-CAM heatmaps clearly reveal differentiated feature regions attended by the model, indicating successful learning of deep physical structures beyond simple energy distributions.

Cross-dataset comparative analysis reveals three core characteristics of UATFSN. First, model attention regions highly align with physical mechanisms of ship radiated noise, demonstrating physics-oriented feature selectivity rather than random high-energy region responses. Second, similar vessel types activate similar frequency regions across different datasets, confirming the model learns generalizable physical features rather than dataset-specific patterns. Third, selective responses in the frequency dimension and transient capture in the time dimension operate independently yet synergistically, validating the effectiveness of time-frequency decoupling mechanisms. These visualization evidences

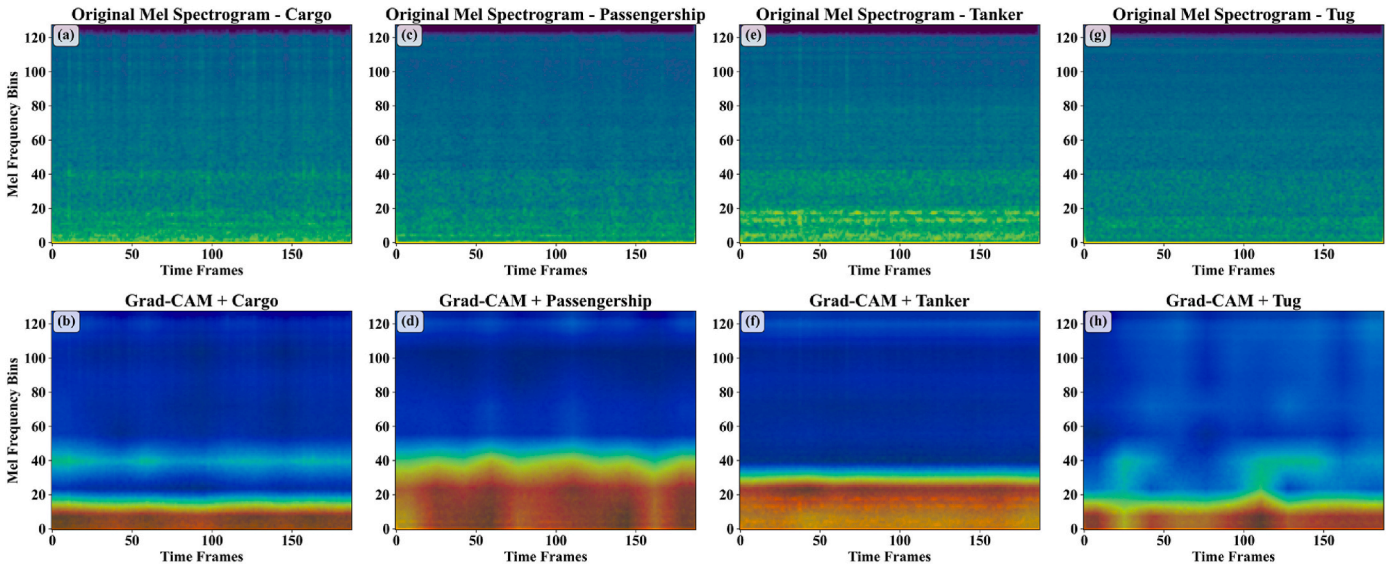


Fig. 6. Grad-CAM visualization results on DeepShip dataset showing model attention distribution across four vessel types.

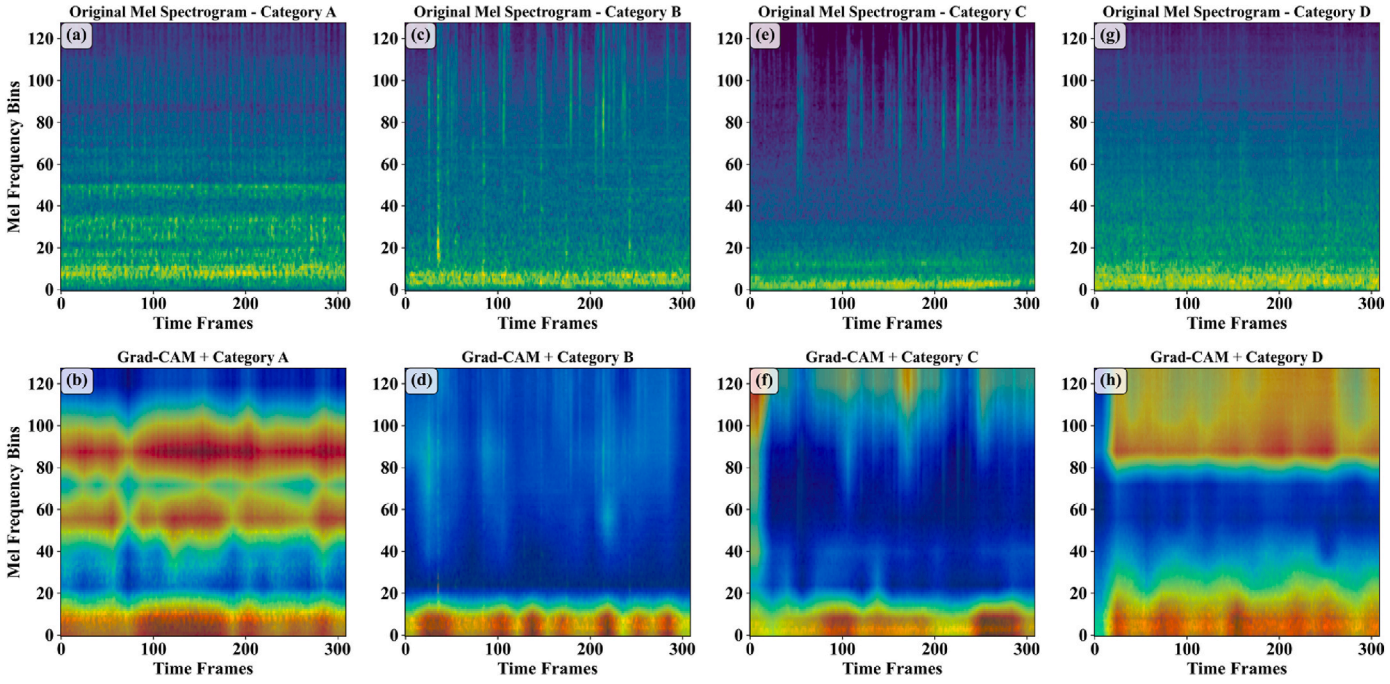


Fig. 7. Grad-CAM visualization results on ShipsEar dataset showing model attention distribution across four vessel categories.

strongly support the core assertion that UATFSN learns physically meaningful time-frequency features.

4.9.2. Learning curve analysis and generalization assessment

To comprehensively evaluate UATFSN's generalization capability, we conduct learning curve analysis of the model training process. Training dynamics directly reflect the model's learning mechanism and generalization characteristics.

As shown in Fig. 8, learning curves on both datasets exhibit ideal convergence characteristics. In early training, training loss and testing loss decrease rapidly and synchronously, indicating effective learning of basic data patterns. On the DeepShip dataset, training and testing losses rapidly decrease from initial values to low-value ranges, with both curves maintaining synchronized descent and minimal separation. The ShipsEar dataset presents similar trends. This closely following pattern between training and testing losses serves as an important indicator of good generalization.

During mid-training, both curves continue steady descent and gradual convergence, with training and testing losses consistently maintaining aligned trends without separation. Stable convergence indicates continuous optimization of feature representations rather than memorization of training samples. In late training, loss curves stabilize and fluctuate in low-value regions. Critically, throughout the entire training process, testing loss never exceeds training loss, and the gap between them remains within an extremely small range, definitively ruling out overfitting possibilities.

The fundamental reason UATFSN achieves high accuracy with extremely low parameters lies in the physical rationality of its architectural design. Time-frequency separation convolution introduces explicit physical priors, significantly reducing the complexity of the

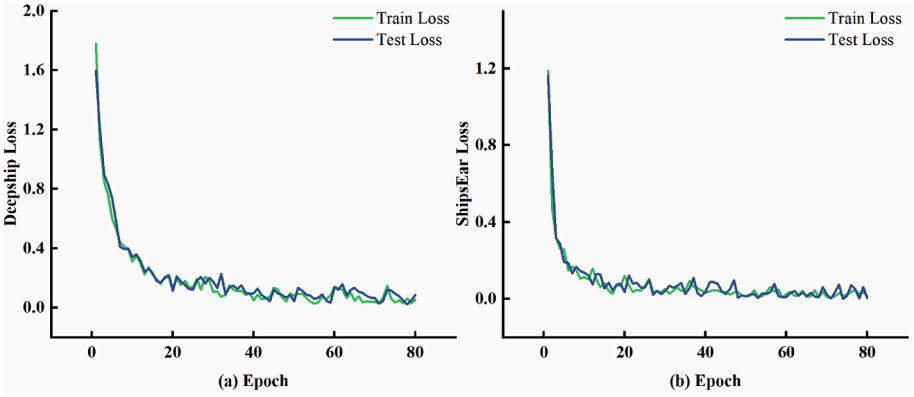


Fig. 8. Training and testing loss curves on DeepShip and ShipsEar datasets demonstrating model convergence and generalization characteristics.

model's hypothesis space. Compared to allowing numerous parameters to freely learn arbitrary feature mappings, UATFSN's structured design forces the model to perform feature extraction along physically meaningful directions. This inductive bias improves parameter efficiency, enabling a small number of parameters to sufficiently represent the feature space required for the task. Additionally, regularization mechanisms such as global average pooling and adaptive residual normalization further enhance model generalization capability. The ideal convergence characteristics of learning curves corroborate robustness experimental results, with stable performance under low SNR and small sample conditions further confirming strong generalization capability.

4.9.3. Confusion matrix and decision boundary analysis

To deeply understand UATFSN's classification decision mechanism, we conduct confusion matrix analysis on prediction results from both datasets. Confusion matrices reveal overall model performance while displaying confusion patterns between different categories, thereby evaluating the effectiveness of time-frequency separation mechanisms in establishing decision boundaries.

As shown in Fig. 9, on the DeepShip dataset, UATFSN demonstrates excellent classification performance for four vessel types. The confusion matrix presents highly diagonalized characteristics, indicating the model establishes clear decision boundaries between categories. Slight bidirectional confusion exists between cargo ships and tankers, which is physically reasonable—these two types of large commercial vessels share similarities in power system configurations and low-speed navigation patterns, with their radiated noise low-frequency harmonic structures exhibiting certain overlap. However, UATFSN effectively captures subtle differential features through time-frequency separation mechanisms, controlling confusion rates at extremely low levels.

In the four-category classification task on the ShipsEar dataset, UATFSN demonstrates even more exceptional performance. The confusion matrix presents a nearly perfect diagonal structure, indicating the model maintains strong discriminative capability in new classification tasks after multi-category reorganization. The extremely low confusion rate between passenger class and ocean liner class proves UATFSN can distinguish subtle feature differences between different-scale passenger vessels—although both belong to passenger ships, mid-frequency energy characteristics generated by large ocean liner propulsion systems exhibit identifiable differences from regular passenger ships, as verified in Grad-CAM analysis.

Synthesizing confusion matrix analysis from both datasets yields three key conclusions. First, independent time-frequency path processing extracts highly discriminative features, establishing clearly separated decision hyperplanes for different vessel types in feature space. Second, confusion patterns highly correlate with vessel physical characteristics, with slight confusion reflecting genuine acoustic similarities rather than model defects, proving the model learns physically meaningful features rather than superficial statistical patterns. Third, achieving highly diagonalized confusion matrices on two datasets with different sampling rates, category definitions, and data distributions fully demonstrates UATFSN possesses strong cross-dataset generalization capability. Combined with learning curve analysis and robustness evaluation, the high diagonalization characteristics of confusion matrices further confirm that the rationality of UATFSN achieving high accuracy with extremely low parameters stems from its physics-oriented architectural design.

5. Conclusion

This paper addresses the limitations of traditional 2D convolutions that fail to adequately consider the physical differences between time and frequency dimensions by proposing the Underwater Acoustic Time-Frequency Separation Network UATFSN. The network implements time-

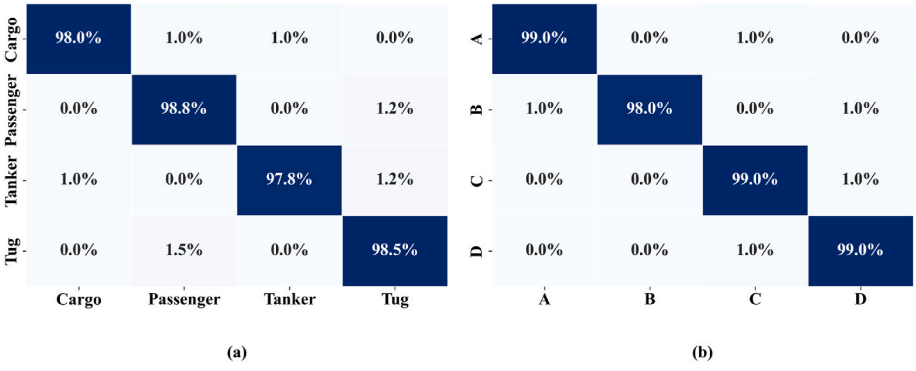


Fig. 9. Confusion matrices on (a)DeepShip and (b)ShipsEar datasets showing classification performance and decision boundaries.

frequency domain decoupled modeling through the Time-Frequency Separate Convolutions TFSC module, employing anisotropic convolution kernels and adaptive residual normalization mechanism to achieve dual optimization of recognition accuracy and computational efficiency. Experimental results fully validate the exceptional performance of UATFSN. On the DeepShip dataset, it achieves 98.23 % recognition accuracy, outperforming all baseline models including ResNet34, while on the ShipsEar dataset, it reaches 98.75 % recognition accuracy, outperforming existing methods by 6.85 percentage points. In computational efficiency, significant breakthroughs are achieved with only 0.05M parameters, representing over 99.7 % reduction compared to traditional ResNet architectures, and FLOPs reduced by over 99.2 %, realizing two orders of magnitude reduction in computational complexity. Robustness evaluation demonstrates strong performance under extreme conditions, achieving 84.20 % accuracy under 0 dB SNR and 96.48 % accuracy with 40 % training data. Ablation studies further validate the core role of the time-frequency separation mechanism, with frequency path removal causing 12.91 percentage points performance drop and temporal path removal causing 10.56 percentage points drop, fully proving the necessity and effectiveness of each component design. This paper effectively addresses the key challenge of coupled processing in underwater acoustic time-frequency domains, achieving industry-leading recognition performance with extremely low computational cost, providing a breakthrough solution for underwater acoustic target recognition in resource-constrained environments. Looking ahead, while UATFSN demonstrates excellent performance on the current datasets, its generalization to novel marine environments with distinct acoustic characteristics or to entirely unseen targets remains an important avenue for future research.

CRediT authorship contribution statement

Menghan Chen: Writing – original draft, Visualization, Software, Methodology, Formal analysis. **Yuchen Lu:** Visualization, Software, Formal analysis. **Liangliang Cheng:** Writing – review & editing, Validation. **Rongxin Zhu:** Visualization, Validation. **Kun Tao:** Visualization, Validation. **Yifei Li:** Writing – review & editing, Methodology, Funding acquisition, Conceptualization. **Magd Abdel Wahab:** Writing – review & editing, Supervision.

Funding

The authors gratefully acknowledge research funding from Zhejiang Key Laboratory of Industrial Solid Waste Thermal Hydrolysis Technology and Intelligent Equipment; Huzhou university Research Startup Project (RK33068).

Declaration of competing interest

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

Data availability

Data will be made available on request.

References

Chen, L., Luo, X., Zhou, H., Shen, Q., Chen, L., Huan, C., 2025. Underwater acoustic multi-target recognition based on channel attention mechanism. Ocean Eng. 315, 119841.
Dong, W., Fu, J., Zou, N., Zhao, C., Miao, Y., Shen, Z., 2025. CAF-ViT: a cross-attention based transformer network for underwater acoustic target recognition. Ocean Eng. 318, 120049.
Du, L., Wang, Z., Lv, Z., Han, D., Wang, L., Yu, F., Lan, Q., 2024. A method for underwater acoustic target recognition based on the delay-doppler joint feature. Remote Sens.
Feng, S., Zhu, X., 2022. A transformer-based deep learning network for underwater acoustic target recognition. IEEE Geosci. Rem. Sens. Lett. 19, 1–5.
Feng, H., Chen, X., Wang, R., Wang, H., Yao, H., Wu, F., 2024. Underwater acoustic target recognition method based on WA-DS decision fusion. Appl. Acoust. 217, 109851.
Gong, Y., Chung, Y.-A., Glass, J.J.a.p.a., 2021. Ast: Audio Spectrogram Transformer.
Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y.J.a.p.a., 2020. Conformer: Convolution-Augmented Transformer for Speech Recognition.
Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.J.a.p.a., 2017. Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications.
Hu, R., Chu, X., Dou, D., Liu, X., Liu, Y., Qi, B., 2025. Toward real-world applicability: lightweight underwater acoustic localization model through knowledge distillation. IEEE J. Ocean. Eng. 50, 1429–1442.
Irfan, M., Jiangbin, Z., Ali, S., Iqbal, M., Masood, Z., Hamid, U., 2021. DeepShip: an underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification. Expert Syst. Appl. 183, 115270.
Jiang, J., Shi, T., Huang, M., Xiao, Z., 2020. Multi-scale spectral feature extraction for underwater acoustic target recognition. Measurement 166, 108227.
Jin, A., Yang, S., Lei, M., Zeng, X., Wang, H., 2025. An effective convolutional and transformer cooperation network for underwater acoustic target recognition. Eng. Appl. Artif. Intell. 159, 111791.
Li, J., Yang, H., 2021. The underwater acoustic target timbre perception and recognition based on the auditory inspired deep convolutional neural network. Appl. Acoust. 182, 108210.
Li, J., Yang, H., 2024. Deep learning method with auditory passive attention for underwater acoustic target recognition under the condition of ship interference. Ocean Eng. 302, 117674.
Li, D., Liu, F., Shen, T., Chen, L., Zhao, D., 2023. Data augmentation method for underwater acoustic target recognition based on underwater acoustic channel modeling and transfer learning. Appl. Acoust. 208, 109344.
Li, J., Han, Y., Wang, X., Zhang, X., Zhang, L., Wei, W., Yu, P., Tan, H., Yang, K., 2025. Clairaudience: a lightweight attentional residual neural network with data augmentation and feature fusion for underwater acoustic target recognition. Comput. J. bxaf099.
Lin, B., Gao, L., Zhu, P., Zhang, Y., Huang, Y., 2024. An underwater acoustic target recognition method based on iterative short-time fourier transform. IEEE Sens. J. 24, 26199–26210.
Liu, F., Shen, T., Luo, Z., Zhao, D., Guo, S., 2021. Underwater target recognition using convolutional recurrent neural networks with 3-D Mel-spectrogram and data augmentation. Appl. Acoust. 178, 107989.
Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A ConvNet for the 2020s. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR), pp. 11966–11976.
Liu, D., Yang, H., Hou, W., Wang, B., 2024. A novel underwater acoustic target recognition method based on MFCC and RACNN. Sensors.
Liu, H., Lu, Y., Cheng, W., Qiu, X., Li, X., 2025a. Marine pipeline corrosion rates prediction via adversarial cloud data synthesis and pipeline medium similarity graph neural networks. Ocean Eng. 342, 122832.
Liu, P., Li, S., Jin, H., Tian, X., Liu, G., 2025b. Shape parameterization method and hydrodynamic noise characteristics of low-noise toroidal propeller. Ocean Eng. 328, 121088.
Liu, P., Li, S., Jin, H., et al., 2025c. Shape parameterization method and hydrodynamic noise characteristics of low-noise toroidal propeller. Ocean Eng. 328, 121088.
Lu, Y., Zhu, Z., Liu, H., Chen, M., Qiu, X., Xu, H., Qu, X., 2025a. End-to-End graph neural network framework for precise localization of internal leakage valves in marine pipelines based on intelligent graphs. Adv. Eng. Inform. 68, 103716.
Lu, Y., Chen, M., Qiu, X., Ren, W., Zhao, C., Liu, H., 2025b. Offshore platform pipeline leakage valve localization using DCEEMDAN and ATSFN. J. Offshore Mech. Arctic Eng. 1–20.
Luo, X., Feng, Y., Zhang, M., 2021. An underwater acoustic target recognition method based on combined feature with automatic coding and reconstruction. IEEE Access 9, 63841–63854.
Luo, X., Chen, L., Zhou, H., Cao, H., 2023. A survey of underwater acoustic target recognition methods based on machine learning. J. Mar. Sci. Eng.
Lyu, C., Hu, X., Niu, Z., Yang, B., Jin, J., Ge, C., 2024. A light-weight neural network for marine acoustic signal recognition suitable for fiber-optic hydrophones. Expert Syst. Appl. 235, 121235.
Ma, Y., Liu, M., Zhang, Y., Zhang, B., Xu, K., Zou, B., Huang, Z., 2022. Imbalanced underwater acoustic target recognition with trigonometric loss and attention mechanism convolutional network. Remote Sens.
Müller, N., Reermann, J., Meisen, T., 2024. Navigating the depths: a comprehensive survey of deep learning for passive underwater acoustic target recognition. IEEE Access 12, 154092–154118.
Tan, M., Le, Q., 2019. Efficientnet: rethinking model scaling for convolutional neural networks. International Conference on Machine Learning. PMLR, pp. 6105–6114.
Tian, S., Chen, D., Wang, H., Liu, J., 2021. Deep convolution stack for waveform in underwater acoustic target recognition. Sci. Rep. 11, 9614.
Tian, S., Bai, D., Zhou, J., Fu, Y., Chen, D., 2023a. Few-shot learning for joint model in underwater acoustic target recognition. Sci. Rep. 13, 17502.
Tian, S.-Z., Chen, D.-B., Fu, Y., Zhou, J.-L., 2023b. Joint learning model for underwater acoustic target recognition. Knowl. Base Syst. 260, 110119.
Trockman, A., Kolter, J.Z., 2022. Patches Are all You Need?.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention Is all You Need, p. 30.

Wang, B., Zhang, W., Zhu, Y., Wu, C., Zhang, S., 2023. An underwater acoustic target recognition method based on AMNet. IEEE Geosci. Rem. Sens. Lett. 20, 1–5.
Wu, J., Guan, D., Yuan, W., 2024. Echo lite voice fusion network: advancing underwater acoustic voiceprint recognition with lightweight neural architectures. Appl. Intell. 55, 112.
Xie, Y., Ren, J., Xu, J., 2022. Underwater-art: expanding information perspectives with text templates for underwater acoustic target recognition. J. Acoust. Soc. Am. 152, 2641–2651.
Xie, Y., Ren, J., Xu, J., 2024. Unraveling complex data diversity in underwater acoustic target recognition through convolution-based mixture of experts. Expert Syst. Appl. 249, 123431.
Xu, J., Xie, Y., Wang, W., 2023. Underwater acoustic target recognition based on smoothness-inducing regularization and spectrogram-based data augmentation. Ocean Eng. 281, 114926.
Xu, Y., Kong, X., Cai, Z., 2024. Cross-validation strategy for performance evaluation of machine learning algorithms in underwater acoustic target recognition. Ocean Eng. 299, 117236.
Xu, J., Li, X., Zhang, D., Chen, Y., Peng, Y., Liu, W., 2025. Enhanced underwater acoustic target recognition using parallel dual-branch network with attention mechanism. Eng. Appl. Artif. Intell. 158, 111603.
Yan, C., Yan, S., Yao, T., Yu, Y., Pan, G., Liu, L., Wang, M., Bai, J., 2024. A lightweight network based on multi-scale asymmetric convolutional neural networks with attention mechanism for ship-radiated noise classification. J. Mar. Sci. Eng.
Yang, H., Zhou, T., Jiang, H., Yu, X., Xu, S., 2024a. A lightweight underwater target detection network for forward-looking sonar images. IEEE Trans. Instrum. Meas. 73, 1–13.
Yang, S., Jin, A., Zeng, X., Wang, H., Hong, X., Lei, M., 2024b. Underwater acoustic target recognition based on sub-band concatenated Mel spectrogram and multidomain attention mechanism. Eng. Appl. Artif. Intell. 133, 107983.
Yao, Q., Wang, Y., Yang, Y., 2023. Underwater Acoustic target recognition based on data augmentation and residual CNN. Electronics.
Yao, Q., Wang, Y., Yang, Y., 2024. Underwater Acoustic target recognition based on hilbert–huang transform and data augmentation. IEEE Trans. Aero. Electron. Syst. 60, 7336–7353.
Zeng, D., Yan, S., Yang, J., Pan, X., 2025. An efficient deep learning approach with frequency and channel optimization for underwater acoustic target recognition. Sci. Rep. 15, 27369.
Zhang, Y., Zeng, Q., 2023. MSLEFC: a low-frequency focused underwater acoustic signal classification and analysis system. Eng. Appl. Artif. Intell. 123, 106333.
Zhang, X., Zhou, X., Lin, M., Sun, J., 2018. Shufflenet: an extremely efficient convolutional neural network for mobile devices. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856.
Zhang, Y., Adin, V., Bader, S., Oelmann, B., 2023. Leveraging acoustic emission and machine learning for concrete materials damage classification on embedded devices. IEEE Trans. Instrum. Meas. 72, 1–8.
Zhang, Y., Pullin, R., Oelmann, B., Bader, S., 2025. On-Device fault diagnosis with augmented acoustic emission data: a case study on carbon fiber panels. IEEE Trans. Instrum. Meas. 74, 1–12.
Zhao, C., Liu, H., Li, M., Liu, W., Lu, Y., Qu, X., 2025. Nonperiodic inspection and maintenance optimization for floating wind turbine electric control systems based on an improved salp swarm algorithm. J. Offshore Mech. Arctic Eng. 148.
Zhou, Y., Xia, H., Yu, D., et al., 2024. Outlier detection method based on high-density iteration. Inf. Sci. 662, 120286.
Zhu, K., Zhao, C., 2023. Dynamic graph-based adaptive learning for online industrial soft sensor with mutable spatial coupling relations. IEEE Trans. Ind. Electron. 70, 9614–9622.
Zhu, K., Zhao, C., 2024. Towards the disappearing truth: fine-grained joint causal influences learning with hidden variable-driven causal hypergraphs in time series. Proc. AAAI Conf. Artif. Intell. 38, 17167–17174.
Zhu, P., Zhang, Y., Huang, Y., Lin, B., Zhu, M., Zhao, K., Zhou, F., 2023a. SFC-Sup: robust two-stage underwater acoustic target recognition method based on supervised contrastive learning. IEEE Trans. Geosci. Rem. Sens. 61, 1–23.
Zhu, R., Boukerche, A., Feng, L., et al., 2023b. A trust management-based secure routing protocol with AUV-Aided path repairing for underwater acoustic sensor networks. Ad Hoc Netw. 149, 103212.
Zhu, R., Boukerche, A., Long, L., et al., 2024a. Design guidelines on trust management for underwater wireless sensor networks. IEEE Commun. Surve Tutorials 26 (4), 2547–2576.
Zhu, R., Boukerche, A., Yang, Q., 2024b. An efficient secure and adaptive routing protocol based on gmm-hmm-lstm for internet of underwater things. IEEE Internet Things J. 11 (9), 16491–16504.