POLISH MARITIME RESEARCH 2 (130) 2026 Vol. 33; pp. 145-155
10.2478/pomr-2026-0028

Heterogeneous Multi-Branch Feature Fusion Architecture for Underwater Acoustic Target Recognition

Yuexiong Yang [1]
Zhenhao Chu [1,*]
Zilong Peng [1]
[1] School of Energy and Power Engineering, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu, China

* Corresponding author: *zhenhaochu@163.com (Zhenhao Chu)*

ABSTRACT

The task of underwater acoustic target recognition has significant value for both military and civilian applications involving complex marine environments. However, traditional methods often struggle to achieve robust performance due to strong noise interference, weak target signals, and insufficient complementarity among the different features. This paper therefore proposes a deep learning recognition framework based on multimodal spectrogram features, in which short-time Fourier transform magnitude spectra, mel-spectrograms, and continuous wavelet transform spectrograms are used as multi-branch inputs, and a heterogeneous convolutional neural network is employed to extract time-frequency feature representations from each modality. Tailored multi-scale convolutional kernels are designed according to the characteristics of the different features, thereby effectively capturing the energy textures, envelope characteristics, and multi-scale structural information. At the fusion stage, a structure-dependent fusion strategy selection method is introduced, which enables the optimal integration of multi-branch features through inverse-variance weighting. Experimental results demonstrate that the proposed method achieves an accuracy of 97.60% in underwater acoustic target recognition tasks, significantly outperforming single-modality and homogeneous multi-branch models, a finding that validates the effectiveness of the proposed heterogeneous multi-branch architecture and its adaptive fusion strategy.

Keywords: underwater acoustic target recognition; multimodal fusion; convolutional neural network (CNN); heterogeneous architecture; feature fusion

INTRODUCTION

Underwater acoustic target recognition (UATR) is an important technology in fields such as ocean exploration, intelligent monitoring, and underwater communications, in which the core objective is to effectively distinguish between and identify acoustic signals from various sources within a complex marine environment [1]. However, inherent challenges in real-world ocean settings such as unpredicTable background noise, strong signal non-stationarity, and the diverse characteristics of the targets severely constrain the robustness and generalisation capability of traditional methods [2]. To support research in this area, various platforms have been developed for underwater acoustic data acquisition, such as acoustic gliders for deep-sea noise profiling [3] and autonomous underwater vehicles equipped with sonar for seabed mapping(AVUs) equipped with sonar for seabed mapping [4]. Furthermore, an understanding of the radiation noise characteristics of underwater targets, such as unmanned underwater vehicles, forms the foundation for effective recognition systems [5].

Early research in this field relied heavily on the expertise of human sonar operators, who interpreted LOFAR and

DEMON spectrograms and performed auditory analysis for classification [6]. With the development of statistical learning theory, the focus shifted to a process dominated by handcrafted feature design; commonly used features included line spectra, cepstral parameters, and time-frequency energy distributions, which were combined with traditional classifiers such as hidden Markov models [7], support vector machines [8], and Gaussian mixture models. Although they are effective in simple scenarios with low levels of noise, these methods often fail to adequately capture the intrinsic discriminative characteristics of signals under low signal-to-noise ratio (SNR) conditions or when subtle inter-class differences exist, leading to performance degradation. To address these limitations, researchers proposed various improvements, such as multi-neural network classification systems [9], neural network-based sonar signal separation systems [10], more robust features based on modified gammatone frequency cepstral coefficients [11], and the introduction of semi-supervised learning (the IS3L framework) to handle weakly labelled data for underwater target detection [12].

In recent years, with the significant progress that has been made in computational hardware and signal processing techniques, deep learning-based approaches have achieved significant advancements in this domain by leveraging their powerful capability for hierarchical automatic feature learning [13]. Convolutional neural networks (CNNs), recurrent neural networks, and their variants have been widely applied, and have yielded substantially improved classification performance [14]; for instance, researchers have employed CNNs for robust ship noise recognition in shallow water under low-SNR conditions [11], while others have proposed a multi-attribute correlation perception method for underwater acoustic targets to enhance performance by exploiting inter-attribute correlations [15]. In some studies, a deep learning framework based on mel power spectra and data augmentation (including time-domain and time-frequency transformations) has been adopted to improve model generalisation, and additional work has used an auditory-inspired CNN trained directly on raw signals, which has achieved superior classification accuracy to traditional methods [16].

Despite the progress of single-modality deep learning methods, however, research indicates that there are limitations associated with relying solely on a single feature representation. Different time-frequency representations (e.g. short-time Fourier transform (STFT), mel, and wavelet spectra) have distinct advantages in terms of energy distribution, modulation characteristics, and time-frequency resolution. Dependence on a single spectrogram may lead to information loss or model overfitting [17]. Consequently, multi-feature fusion has emerged as an important research direction, and has been shown to have significant advantages. For instance, Da et al. [18] designed an ensemble neural network with three sub-networks by integrating STFT magnitude spectra with phase spectra and bispectral features, with improved recognition accuracy and noise robustness. Pan et al. [19] fused time-frequency domain features with frequency-domain image features and employed an attention mechanism for adaptive weighted fusion, achieving a recognition accuracy of 94.92%.

Recent advances in audio classification have also demonstrated the effectiveness of specific time-frequency representations across various domains. For instance, Seo et al. [20] proposed a CNN using log mel-spectrogram separation for audio event classification, with robust performance across unknown devices. In the specific context of underwater acoustics, Khalilabadi [21] applied a CNN to mel-spectrograms for ship-radiated noise (SNR) recognition, and reported promising results. Similarly, Sareen and Seeja [22] demonstrated the utility of mel-spectrograms and CNNs for speech emotion recognition, thereby further validating the effectiveness of this representation in audio analysis tasks. Taken together, these studies support the rationale for incorporating mel-spectrograms as one of the complementary feature representations in our heterogeneous multi-branch architecture.

Nevertheless, existing fusion schemes still exhibit noTable shortcomings in terms of the feature–network compatibility and the robustness of fusion mechanisms. Firstly, many approaches employ simple feature concatenation or homogeneous networks to process heterogeneous features, and fail to account adequately for the intrinsic properties of different time-frequency representations [23]. Comparative studies have also shown that different time-frequency representations interact differently with neural network architectures [24]. Furthermore, the use of multi-time-frequency fusion for tasks involving underwater or acoustic scenes indicates that improper handling of heterogeneous features may limit model performance [25]. Secondly, fusion strategies often struggle to strike a balance between the simplicity of the model and excellent performance. Prior work indicates that naïve fusion modules or single-stream architectures frequently lack robustness and adaptability across diverse feature domains [26]. In addition, studies of multi-spectrogram encoder–decoder or weighted-fusion frameworks have emphasised that the issue of architecture–fusion compatibility remains insufficiently explored [27].

To address these problems, this paper proposes a UATR method based on a heterogeneous multi-branch convolutional neural network. The core concept of this method is to design dedicated feature extraction network structures that are tailored to the inherent characteristics of different time-frequency features. Three complementary time-frequency representations, the STFT magnitude spectrum, mel spectrum, and CWT spectrogram, are first selected to characterise the energy texture, auditory envelope, and multi-scale time-frequency structure of the signal, respectively. Following this, differentiated network architectures are designed according to the characteristics of each spectrogram, to form a heterogeneous three-branch feature extraction framework.

At the feature extraction level, a multi-scale convolutional kernel design is introduced to enhance the model's perception of features at different scales through differentiated receptive fields. Various fusion strategies were explored for the feature fusion stage; initially, widely used adaptive fusion schemes were tested, but it was found that the additional complexity that was introduced to the model prevented the expected performance gains from being achieved. Through a comprehensive

comparative analysis, an inverse-variance weighting method was ultimately established in which weights were assigned based on branch stability in a more concise manner, and this was verified as the best fusion scheme for this framework, as it achieved efficient integration of multi-source features.

The main contributions of this paper can be summarised as follows:

1. A heterogeneous multi-branch network architecture is proposed, which enables effective mining and complementary fusion of multi-source heterogeneous features through differentiated designs for STFT magnitude spectra, mel spectra, and CWT spectrograms.
2. A hierarchical multi-scale feature enhancement mechanism is designed to significantly improve the model's perception of multi-scale features in underwater acoustic signals through both intra-branch and cross-branch differentiated convolutional kernel designs.
3. Experimental validation demonstrates that the proposed heterogeneous multi-spectrogram fusion method significantly enhances UATR performance, thereby confirming the effectiveness of the multi-modal fusion strategy. Further systematic comparisons are conducted to show that inverse-variance weighting is the optimal fusion method for this framework.

METHODS

This paper presents a heterogeneous multi-branch convolutional neural network (H-MBCNN) architecture for UATR, which was developed to fully exploit the complementary characteristics of different time-frequency representations. The proposed architecture is illustrated in Fig. 1, and comprises three main components: multi-modal feature construction, branch-specific feature encoding, and modality fusion with classification. During the multi-modal feature construction phase, three distinct image-based features are extracted from the raw underwater acoustic signals, each of which emphasises different fundamental characteristics: the STFT magnitude spectrum excels in representing localised energy distribution and textural patterns, whereas the mel-spectrogram effectively simulates human auditory perception while preserving speech-like time-frequency envelope characteristics, and the CWT spectrogram provides superior multi-scale time-frequency representation capabilities, making it particularly suiTable for characterising non-stationary signals.

To accommodate the distinct physical properties of these three feature types, a heterogeneous design strategy is employed in the branch-specific feature encoding module. In the STFT magnitude spectrum branch, a deep network with small convolutional kernels is used to precisely extract fine-grained time-frequency energy patterns. The mel-spectrogram branch relies on larger convolutional kernels, which are better adapted to capturing broad spectral band characteristics. The CWT branch includes a multi-scale parallel convolution module to efficiently capture multi-resolution characteristics in the time-frequency domain. This customised architectural approach enables each branch to extract the discriminative features of underwater acoustic signals from different perspectives, thereby providing high-quality feature representations for subsequent fusion.

Several fusion strategies were systematically evaluated for use at the modality fusion and classification stage. Although the adaptive fusion scheme that was initially implemented successfully achieved feature integration, the additional complexity it introduced to the model meant that the expected performance improvements were not delivered. Through a comprehensive comparative analysis, an inverse-variance weighting method was ultimately identified as the optimal fusion strategy for this framework. In this approach, weights are assigned according to the output stability of each branch to achieve complementary integration of multi-source features through a computationally efficient mechanism. This scheme effectively preserves the energy distribution advantages of STFT magnitude spectra while incorporating the perceptual characteristics of mel-spectrograms and the non-stationary signal representation capabilities of CWT transforms, thereby establishing a synergistic recognition mechanism that significantly enhances the model's robustness and generalisation performance.

The core design concept leverages the complementary advantages of STFT magnitude spectra, mel-spectrograms, and CWT spectrograms for underwater acoustic signal characterisation. Through customised branch architectures and an optimised fusion mechanism, effective integration of multi-source information is achieved, yielding enhancements in both the recognition performance and the generalisation capability of the model.

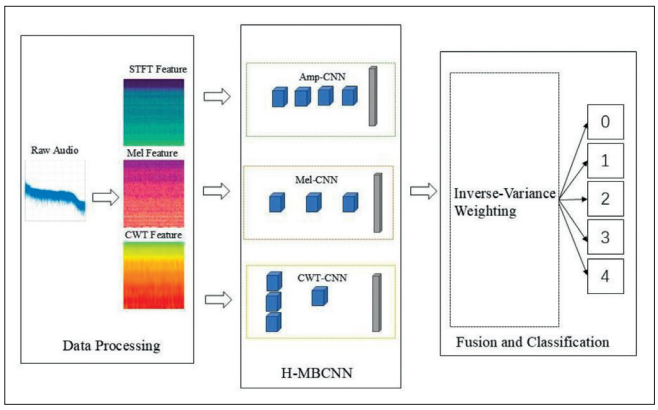

Fig. 1. Proposed H-MBCNN architecture

DATASET

The data used in this study were obtained from the publicly available DeepShip dataset [28]. It comprises 609 original recordings, including 240 samples of tankers, 69 of tugs, 191 of passenger ships, and 109 of cargo ships, covering four ship categories in total. Each recording has a duration of several minutes, with the total duration of all data being approximately 47 hours and 13 minutes. All signals are in monaural format, with a uniform sampling frequency of 32 kHz.

Although the original DeepShip dataset [28] contains these four classes of SRN, it does not include a background noise class.

In real-world underwater acoustic target recognition applications, it is essential for a system to be able to distinguish between signals containing vessel noise and those consisting only of ambient noise. To make the classification task more realistic and comprehensive, we introduced an additional background class (Class 0). This background noise was obtained from an established underwater acoustic database [29], with a total duration of approximately 5 hours and 27 minutes. These recordings were processed using the same segmentation and preprocessing pipeline as the ship signals, resulting in a five-class dataset (four vessel classes + one background class). The sample distribution is presented in Table 1. All signals were then segmented into 3 s clips to serve as independent samples. A length of 3 s was chosen for the signal segments as this is consistent with the processing method used in the original paper describing the DeepShip dataset [28]; this duration can effectively capture the time-frequency characteristics of SRN while ensuring an adequate number of samples. Only complete segments with a duration of exactly 3 s were retained, while incomplete terminal segments were discarded. This resulted in a total of 76,103 audio samples, of which 80% were randomly allocated as the training set, 10% as the validation set, and 10% as the test set. The sample distribution across the four categories is presented in Table 1.

Tab. 1. Sample distribution of the dataset

Ship category (class)	Training set	Validation set	Test set	Total
Background (Class 0)	13,745	3,927	1,963	19,635
Cargo (Class 1)	8,961	2,560	1,280	12,801
Passenger ship (Class 2)	10,787	3,082	1,541	15,410
Tanker (Class 3)	10,334	2,952	1,476	14,762
Tug (Class 4)	9,447	2,699	1,349	13,495
Total	53,274	15,220	7,609	76,103

DATA PROCESSING

When the raw underwater acoustic signals had been segmented into fixed-duration segments of 3 s to ensure temporal consistency across all input samples, each segment was normalised as shown in Eq. (1):

$$\hat{x}_i = \frac{x_i - \mu(x_i)}{\sigma(x_i) + \varepsilon} \tag{1}$$

where $\mu(x_i)$ and $\sigma(x_i)$ represent the mean and standard deviation of the segment, respectively, and ε is a small constant included to prevent division by zero. In this study, ε was set to 1×10^{-6} for numerical stability purposes, including preventing division by zero in Eq. (1) and avoiding a log of zero in Eqs. (4), (6), and (8). This value is sufficiently small (six orders of magnitude smaller than the typical signal magnitudes) to have no practical impact on the feature distributions, while effectively preventing numerical errors.

The time-frequency distribution of the normalised signal for each segment was calculated using STFT, as shown in Eq. (2):

$$X(t,f) = \sum_{n=0}^{N-1} x[n] \cdot w[n-t] \cdot e^{-j2\pi fn/N} \tag{2}$$

where $x[n]$ denotes the discretised input signal, $w[n-t]$ is the window function (e.g. a Hamming window) employed to achieve short-time localisation, the complex exponential term corresponds to the Fourier basis at frequency f, and N is the window length. This expression is used to compute the complex spectrum $X(t,f)$ of the signal at time t and frequency f via a sliding window approach.

Sensitivity analysis of segment length

To justify the choice of 3 s segments, we conducted a sensitivity analysis to compare segment lengths of 1, 2, 3, and 5 s. These experiments were performed using the proposed H-MBCNN under identical training conditions. As shown in Table 2, a length of 3 s yielded the best trade-off between classification accuracy and temporal resolution: shorter segments (1 s, 2 s) may not capture sufficient temporal structure, while longer segments (5 s) increase the computational cost without significant performance gains, and reduce the number of training samples.

Tab. 2. Performance with different segment lengths

Segment length	Accuracy	Recall	Precision	F1
1 s	95.82%	95.66%	95.76%	95.41%
2 s	96.73%	96.61%	96.52%	96.28%
3 s	**97.60%**	**97.41%**	**97.62%**	**97.39%**
5 s	95.47 %	95.36%	95.57%	95.47%

FEATURE EXTRACTION

In this study, three complementary types of time-frequency image features (STFT magnitude spectra, mel-spectrograms, and CWT spectrograms) were constructed to fully exploit the underlying time-frequency characteristics of underwater acoustic signals. Of these features, the STFT magnitude spectrum was used to capture local energy texture patterns in SRN, whereas the mel-spectrogram was employed for effective characterisation of speech-like envelope features inherent in SRN, and the CWT spectrogram was used to reveal intrinsic broadband and multi-scale patterns under low-SNR conditions. The following sub-sections describe the feature extraction methods and present some examples of the results.

Amp spectrogram

The magnitude spectrum is obtained by taking the modulus of the complex-valued STFT, as shown in Eq. (3). It is then converted to a logarithmic (dB) scale, as shown in Eq. (4). Fig. 2 shows an example of the STFT magnitude spectrum for a segment of an underwater acoustic signal.

$$S_{\mathrm{Amp}}(t,f) = |X(t,f)| \tag{3}$$

$$S_{\mathrm{Amp}}^{\mathrm{dB}} = 20\log_{10}(S_{\mathrm{Amp}} + \varepsilon) \tag{4}$$

where ε is a small constant to prevent $\log_0$.

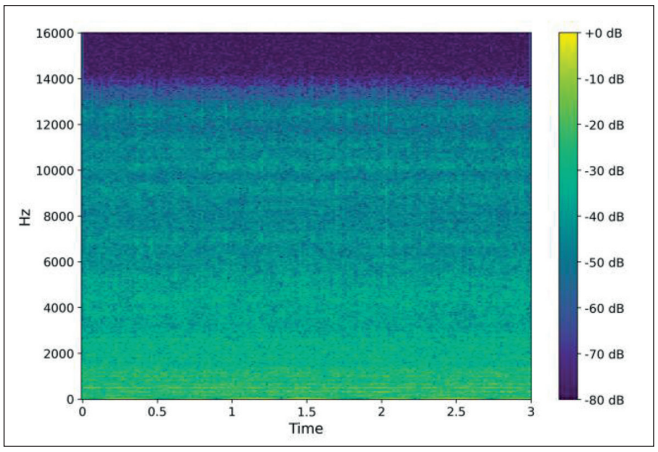

Fig. 2. Example of an STFT magnitude spectrum

Mel-spectrogram

The mel-spectrogram is obtained by passing the STFT power spectrum through a mel filter bank and then applying a logarithmic compression, as shown in Eqs. (5) and (6). In Eq. (5), the mel filter bank employed is a 128-band configuration. Fig. 3 presents an example of a mel-spectrogram for an underwater acoustic signal segment.

$$S_{\mathrm{Mel}(m,\,t)} = \sum_{f} |X(t,f)|^2 \cdot H_m(f) \tag{5}$$

where $|X(t,f)|^2$ is the power spectral density at time-frequency bin (t,f), and $H_m(f)$ is the frequency response of the -th mel filter. The logarithm is then taken to obtain:

$$S_{\mathrm{Mel}}^{\mathrm{dB}} = 10\log_{10}(S_{\mathrm{Mel}} + \varepsilon) \tag{6}$$

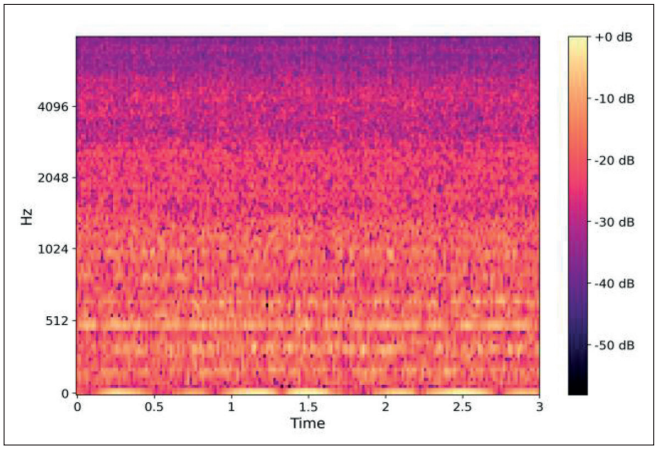

Fig. 3. Example of a mel-spectrogram

CWT spectrogram

The complex Morlet wavelet (cmor1.5-1.0) is employed to perform a multi-scale analysis of the signal, and gives its local energy distribution in the time-frequency plane, as shown in Eq. (7). The amplitude energy is converted into a logarithmic energy spectrogram using Eq. (8). Fig. 4 presents the CWT spectrogram for an exemplary underwater acoustic signal segment.

$$W(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t)\Psi^{*}\left(\frac{t-b}{a}\right)dt \tag{7}$$

where a denotes the scale parameter, b is the time shift parameter, and Ψ represents the mother wavelet. In this study, the complex Morlet wavelet is adopted, and the amplitude energy is converted into a logarithmic energy spectrogram as follows:

$$S_{CWT} = 10\log_{10}\left(|W(a,b)|^4 + \varepsilon\right) \tag{8}$$

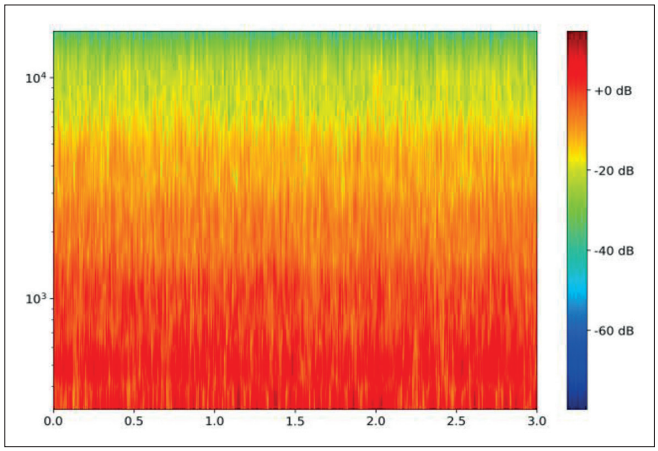

Fig. 4. Example of a CWT spectrogram

H-MBCNN MODEL

To fully exploit the complementary feature information contained in these different spectrograms, a framework called H-MBCNN was developed. This network consists of three structurally distinct specialised branches, each of which is custom-designed according to the physical characteristics of the STFT magnitude spectrum, mel-spectrogram, and CWT spectrogram. In the magnitude spectrum branch, a deep small-kernel convolutional structure is employed to extract fine-grained time-frequency energy patterns, whereas in the mel branch, larger convolutional kernels are used to accommodate wide-band spectral features. The CWT branch effectively captures multi-resolution characteristics in the time-frequency domain using a parallel multi-scale convolutional module. Through this heterogeneous architecture, each branch can achieve deep mining of the discriminative features of underwater acoustic signals from different dimensions. The following sub-sections give the details of the structures of each branch.

Amp-CNN branch

The amplitude branch is designed to extract time-frequency energy distribution features from the STFT magnitude spectrum. This spectrogram has a high frequency resolution and distinct energy peak structures, making it suiTable for extracting localised features via small convolutional kernels in a hierarchical manner.

The proposed Amp-CNN consists of eight layers, consisting of an input layer, four convolutional layers, three max-pooling layers, and a fully connected layer. All of the convolutional layers employ 3×3 kernels, which are stacked hierarchically to capture local energy texture patterns in the spectrogram.

The max-pooling layers compress the feature dimensions and alleviate overfitting, while dropout layers are incorporated to further enhance the model's generalisation capability. The final fully connected layer outputs a 128-dimensional feature representation. The details of the architecture of this branch are given in Table 3.

Tab. 3. Layers and parameters of the Amp-CNN network

Layer	Activation function	Kernel size	Padding	Output shape
Input	–	–	–	(B, 3, 64, 64)
Conv1	ReLU	3×3	1	(B, 32, 64, 64)
MaxPool1	–	2×2	–	(B, 32, 32, 32)
Conv2	ReLU	3×3	1	(B, 64, 32, 32)
MaxPool2	–	2×2	–	(B, 64, 16, 16)
Conv3	ReLU	3×3	1	(B, 128, 16, 16)
MaxPool2d_3	–	2×2	–	(B, 128, 8, 8)
Conv4	ReLU	3×3	1	(B, 128, 8, 8)
AdAvgPool1	–	–	–	(B, 128, 4, 4)
Fc1	ReLU	–	–	(B, 128)

Mel-CNN branch

The mel branch processes mel-spectrograms generated using the mel filter bank. These spectrograms provide higher resolution in lower frequency bands, aligning more closely with the perceptual characteristics of underwater acoustic signals. Consequently, this branch employs larger convolutional kernels to cover broader frequency regions, thereby facilitating the extraction of global structural features.

The designed Mel-CNN consists of eight layers: an input layer, three convolutional layers, three max-pooling layers, and a fully connected layer. The convolutional layers use 5×5 kernels, enabling the network to capture overall spectral trends through larger receptive fields. The fully connected layer in this branch also outputs a 128-dimensional feature vector, providing uniformly dimensioned inputs for subsequent fusion. The architecture of this branch is summarised in Table 4.

Tab. 4. Layers and parameters of the Mel-CNN network

Layer	Activation function	Kernel size	Padding	Output shape
Input	–	–	–	(B, 3, 64, 64)
Conv1	ReLU	5×5	2	(B, 32, 64, 64)
MaxPool1	–	2×2	–	(B, 32, 32, 32)
Conv2	ReLU	5×5	2	(B, 64, 32, 32)
MaxPool2	–	2×2	–	(B, 64, 16, 16)
Conv3	ReLU	5×5	2	(B, 128, 16, 16)
AdAvgPool1	–	–	–	(B, 128, 4, 4)
Fc1	ReLU	–	–	(B, 128)

CWT-CNN branch

-The proposed CWT-CNN consists of nine layers: an input layer, three parallel convolutional modules (each containing multi-scale convolutions), two max-pooling layers, a dropout layer, and two fully connected layers. The final fully connected layer again outputs a 128-dimensional feature vector. The architecture of this branch is detailed in Table 5.

Tab. 5. Layers and parameters of the CWT-CNN network

Layer	Activation function	Kernel size	Padding	Output shape
Input	–	–	–	(B, 3, 64, 64)
Conv3×3	–	3×3	1	(B, 32, 64, 64)
Conv5×5	–	2×2	2	(B, 32, 64, 64)
Conv7×7	–	3×3	3	(B, 32, 64, 64)
Concat	–	–	–	(B, 96, 64, 64)
BatchNorm	–	–	–	(B, 96, 64, 64)
ReLU	ReLU	–	–	(B, 96, 64, 64)
MaxPool1	–	2×2	–	(B, 96, 32, 32)
Conv4	ReLU	3×3	1	(B, 128, 32, 32)
AdAvgPool1	–	–	–	(B, 128, 4, 4)
Fc1	ReLU	–	–	(B, 128)

FUSION AND CLASSIFICATION MODULE

When the three front-end branches have completed the feature extraction process, each generates a 64-dimensional feature vector. These vectors are concatenated to form a 192-dimensional hybrid feature representation. Although an adaptive fusion scheme based on channel attention was initially explored, it was found that the additional complexity of the model meant that the expected performance gains were not achieved. An inverse variance weighting was therefore adopted as the fusion strategy, in which the importance of different features is effectively distinguished while the efficiency of the model is maintained.

Inverse variance weighting is a fusion strategy in which weights are assigned based on the prediction variance of each branch. The core principle is that a branch with a smaller variance in the predictions has more stable and reliable outputs, and thus should be assigned a higher weight.

For each heterogeneous branch, the output feature for the validation set is denoted as f_i, where i represents the branch index. To quantify the stability of each branch, we compute the variance of its output features. Specifically, for the i-th branch, we first calculate the variance σ_i^2 of its output feature map f_i on the validation set, as shown in Eq. (9). The weight w_i for this branch is then calculated as shown in Eq. (10). This method effectively enhances the stability and reliability of the fused features by assigning higher weights to branches with lower variance.

$$\omega_i = \frac{\frac{1}{\sigma_i^2}}{\sum_{j=1}^{N} \frac{1}{\sigma_j^2}} \tag{9}$$

where N is the number of branches. The normalised weights w_i are used to perform a weighted summation of the output features from the heterogeneous branches, yielding the fused features:

$$f_{fused} = \sum_{i=1}^{N} w_i \cdot f_i \tag{10}$$

RESULTS AND DISCUSSION

ABLATION STUDY

Experiments were conducted using the following hardware: an Intel Core i7-12700F processor, an NVIDIA GeForce GTX 1660 graphics card (6 GB dedicated GPU memory), and 14 GB system memory. The software environment included Python 3.11.9, PyTorch 2.2.2, and scikit-learn 1.3.0.

For training of the model, the Adam optimiser was used with an initial learning rate of 0.001, a batch size of 16, and a training duration of 30 epochs. A dynamic learning rate scheduling strategy was implemented in which the learning rate was halved if no improvement in validation accuracy was observed over five consecutive epochs. The loss function was a combination of the cross-entropy loss and triplet loss, with weighting coefficients of $\alpha = 1.0$ and $\beta = 0.5$, respectively.

To enable a comparative analysis, a homogeneous three-branch model (CNN3) was implemented with three identical branches, each consisting of a three-layer CNN structure with 5×5 convolutional kernels.

The initial experiments involved an adaptive fusion module based on channel attention, in which weights were allocated through fully connected layers and the softmax activation function. However, the results demonstrated that this scheme introduced significant model complexity while yielding limited performance gains. Building on these findings, an inverse-variance weighting strategy was explored in which fusion weights were assigned according to the stability of each branch's output features on the validation set, by calculating the variance of each branch's output features and normalising the weights using their reciprocal values. This approach achieved effective feature fusion with lower computational complexity, and subsequent experiments verified its superiority.

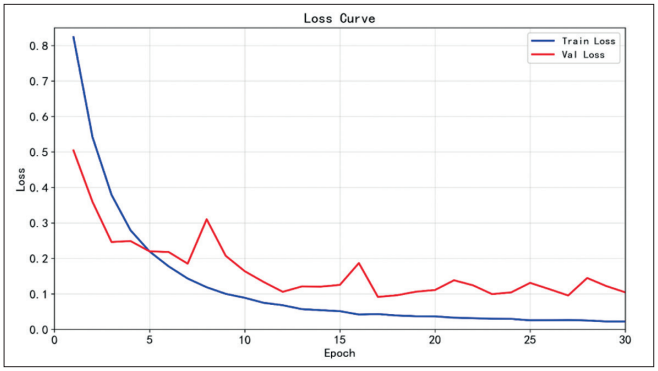

Fig. 5. Loss curves for the training and validation sets

Fig. 5 shows the loss function curves during the training process. It can be observed that both the training and validation losses exhibit favourable convergence trends: during the initial epochs, the losses decrease rapidly, indicating effective feature learning by the model, whereas in the latter epochs, the losses decline more gradually and stabilise, suggesting the model is approaching an optimal solution.

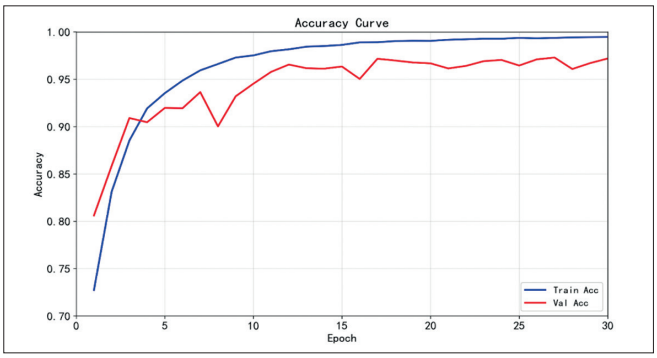

Fig. 6. Accuracy curves for the training and validation sets

Fig. 6 illustrates the change in accuracy during training: this increases rapidly during the first 17 epochs, after which it enters a relatively sTable phase. We also note that further improvement is observed around epochs 27 or 28, where the validation accuracy reaches its peak of 97.19%. This late-stage improvement suggests that the model continues to refine its feature representations even after the initial convergence. The close alignment between training and validation curves throughout the process indicates strong generalisation without overfitting.

EFFECTS OF BRANCH SELECTION STRATEGIES

This section provides a systematic comparative analysis of several multi-feature fusion strategies based on experimental results from single-branch, dual-branch, homogeneous three-branch, and heterogeneous three-branch models. Table 6 presents a performance comparison of different model configurations on both the validation and test sets.

The results show that among the single-branch models, the magnitude spectrum branch achieved the best performance, with a test accuracy of 94.31%. Of the dual-branch fusion models, the Amp-Mel combination yielded optimal performance, with a test accuracy of 96.86%. Notably, the homogeneous three-branch model (CNN3) achieved a test accuracy of 95.95%, which, while superior to the single-branch models, was significantly lower than that of the heterogeneous three-branch model. Our proposed heterogeneous three-branch model (H-MBCNN) achieved the best performance, with a test accuracy of 97.40%, representing improvements of 1.45% over the homogeneous three-branch model and 0.54% over the best dual-branch model. These results indicate that simply increasing the number of branches without considering the characteristics of the features cannot yield significant performance improvements, and that a heterogeneous design approach is essential to achieve effective feature fusion.

A detailed analysis reveals that the heterogeneous model performed optimally across all evaluation metrics, achieving a macro-average F1 score of 96.98%. Compared to the homogeneous three-branch model, the proposed heterogeneous model had values for the accuracy, recall, and F1-score that were improved by 1.45%, 1.42%, and 2.34%, respectively, thus demonstrating the effectiveness of the customised network architecture.

Following the fusion of the three features via the heterogeneous branch network, the model achieved favorable performance with high accuracy. Compared to the homogeneous three-branch model, the advantages of the proposed heterogeneous model were manifested not only in the overall values for the metrics but also in refinements to the error distribution and a balanced performance across the different categories, which significantly enhanced the robustness of the model. This indicates that the use of the magnitude spectrum, mel-spectrogram, and CWT spectrogram in our customised heterogeneous network architecture achieves deep feature complementarity: the mel-spectrogram can effectively enhance the representation capability of time-frequency features and provide crucial support for the improvement of model performance, while the CWT branch captures subtle time-frequency patterns that are overlooked by the other branches. The synergistic interaction among the three modalities promotes comprehensive performance optimisation.

Tab. 6. Results of ablation experiments on the model structure

Network type	Accuracy	Recall	Precision	F1
Amp	94.42%	94.00%	94.31%	93.95%
Mel	94.08%	93.65%	94.08%	93.58%
CWT	84.27%	83.27%	83.59%	82.96%
Amp-CWT	94.53%	93.98%	93.71%	93.89%
Amp-Mel	96.25%	95.87%	96.86%	95.94%
CNN3	95.03%	94.58%	95.95%	94.64%
H-MBCNN	**97.19%**	**97.00%**	**97.40%**	**96.98%**

The performance improvement stems not only from the feature extraction capability of each branch but more essentially from the heterogeneous design itself, which facilitates diverse feature representations. The homogeneous three-branch model, in which identical network structures are employed to process different features, tends to produce homogenised feature representations, limiting the effectiveness at the fusion stage. In contrast, the distinct network structures of the heterogeneous model (e.g. convolutional kernel sizes, depths) naturally guide each branch to interpret data from unique perspectives, generating features with lower redundancy and stronger complementarity. The results for the Amp-Mel dual-branch model showed some synergistic effects, and when the unique time-frequency analysis capability of the CWT branch had been incorporated through customised modules, the three branches collectively provided a more robust and comprehensive feature representation space.

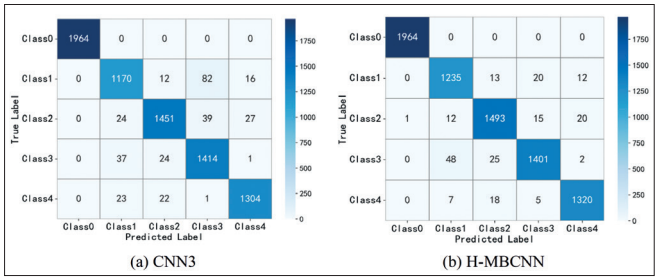

Fig. 7. Confusion matrices for the experimental results from CNN3 and H-MBCNN

A comparison of the confusion matrices shown in Fig. 7 indicates that the proposed H-MBCNN outperforms the baseline CNN3 in the five-class classification task, achieving effective improvement in most categories. The numbers of correctly classified samples for Classes 1, 2, and 4 were increased by 65, 42, and 16, respectively. Notably, the confusion between Classes 1 and 3 was significantly suppressed, with the number of misclassifications decreasing from 82 to 20. Although the accuracy for Class 3 was slightly decreased (with 13 fewer correct samples), this can be regarded as a local trade-off resulting from the enhanced discrimination boundary between Classes 1 and 3 during the optimisation process. Moreover, the misclassification of samples in Class 3 as Class 4 was nearly eliminated. These results demonstrate that H-MBCNN significantly improves inter-class differentiation by enhancing feature discriminability, thus validating the effectiveness of the proposed network architecture.

Feature visualisation analysis: As an intuitive evaluation of the feature learning capability of the model, we employed the t-SNE algorithm to reduce the 64-dimensional high-level features to two dimensions for visualisation. To ensure reproducibility, the t-SNE parameters were set as follows: perplexity=30, learning rate=200, number of iterations=1000, and random state=42. As shown in Fig. 8, the features extracted by H-MBCNN form well-separated clusters, with samples from the same category tightly grouped and those from different categories clearly separated.

Quantitative analysis: To enable a objective assessment of the feature separability, we calculated the silhouette coefficient for the high-dimensional features. This metric represents a comprehensive measurement of intra-class compactness and inter-class separation, and takes values ranging from [−1, 1] with higher values indicating better clustering. Experimental results show that the average silhouette coefficient for the features extracted by H-MBCNN is 0.63, significantly higher than for the homogeneous three-branch model (CNN3), with a value of 0.56. This quantitative result confirms that the heterogeneous multi-branch design of H-MBCNN allows it to learn features with better intra-class compactness and inter-class separation, which provides solid evidence and lays a reliable foundation for the 97.60% classification accuracy achieved by the model.

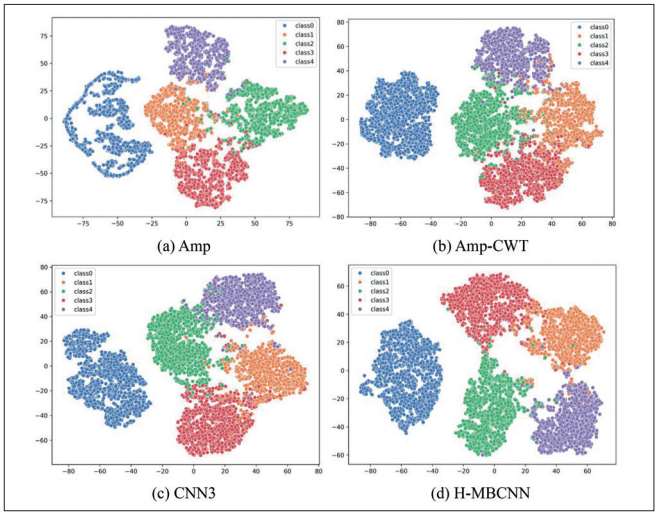

Fig. 8. t-SNE visualisations of feature representations across different model structures

EFFECTS OF FUSION STRATEGIES

To determine the optimal feature fusion strategy, three methods were systematically compared: adaptive weighting fusion, average weighting fusion, and inverse-variance weighting fusion. As shown in Tables 7, 8, and 9, these different fusion strategies significantly impacted the performance of the models, and their effectiveness was closely related to the model architecture.

In dual-branch models, average weighting fusion achieved excellent performance. The accuracy for the Amp-Mel model reached 97.08% under this strategy, slightly higher than the 96.86% achieved with inverse-variance weighting. This indicates that when dealing with a small number of branches with balanced performance, simple average fusion can effectively integrate the feature information. However, for the more complex three-branch heterogeneous networks, inverse-variance weighting offered distinct advantages. The proposed H-MBCNN model achieved the highest accuracy of 97.60% with this strategy, significantly outperforming the other fusion methods. This fusion method also yielded an accuracy of 95.76% in the CNN3 homogeneous three-branch model, illustrating its universal applicability to multi-branch fusion scenarios.

This disparity in the performance of the models stems from the compatibility between different fusion strategies and model architectures. Dual-branch models feature relatively simple feature relationships, and average weighting therefore suffices for the effective integration of information. In contrast, three-branch models involve more complex feature interactions that require more a refined approach to weight allocation. The advantage of inverse-variance weighting lies in its ability to achieve dynamic adaptation to the stability differences among branches by assigning greater weights to more sTable branches, thereby optimising the complex inter-branch relationships. Unlike adaptive weighting fusion, this method does not introduce additional complex parameters and reduces the risk of overfitting.

In summary, this research indicates that the choice of fusion strategy should be aligned with the model architecture: average weighting fusion is suiTable for dual-branch models, whereas inverse-variance weighting is recommended for three-branch and more complex heterogeneous networks. This hierarchical recommendation strategy provides refined methodological guidance for multi-modal feature fusion, suggesting that after sufficient optimisation of the feature extractors, the selection of appropriate simple fusion strategies based on the structural characteristics of the model often yields optimal results.

Tab. 7. Experimental results for models with adaptive weighting fusion

Type	Accuracy	Recall	Precise	F1
Amp-CWT	94.53%	93.98%	93.71%	93.89%
Amp-Mel	96.25%	95.87%	96.86%	95.94%
CNN3	95.03%	94.58%	95.95%	94.64%
H-MBCNN	**97.19%**	**97.00%**	**97.40%**	**96.98%**

Tab. 8. Experimental results for models with average weighting fusion

Type	Accuracy	Recall	Precise	F1
Amp-CWT	95.05%	94.73%	94.67%	94.63%
Amp-Mel	97.08%	96.79%	97.42%	96.81%
CNN3	94.59%	94.03%	95.44%	94.16%
H-MBCNN	**97.21%**	**97.08%**	**97.53%**	**96.99%**

Tab. 9. Experimental results for models with inverse-variance weighting fusion

Type	Accuracy	Recall	Precise	F1
Amp-CWT	92.34%	91.76%	94.56%	91.69%
Amp-Mel	96.86%	96.61%	97.42%	96.60%
CNN3	95.76%	95.53%	95.80%	95.39%
H-MBCNN	**97.60%**	**97.41%**	**97.62%**	**97.39%**

CONCLUSIONS AND FURTHER RESEARCH

This paper has addressed the challenge of UATR in complex marine environments by presenting an H-MBCNN framework based on multi-modal spectrogram features. Systematic experimentation was conducted to validate the effectiveness of the proposed approach, and the results confirmed the significant advantages of the heterogeneous multi-branch architecture: it was found that the specialised networks designed for STFT magnitude spectra, mel-spectrograms, and CWT spectrograms effectively exploited the discriminative information embedded in each feature type, enabling the model to achieve a test set accuracy of 97.40% and demonstrating the value of specialised design in processing multi-source heterogeneous features.

This study also revealed important compatibility patterns between model structures and fusion methods: dual-branch models are well-suited to average weighting fusion, whereas three-branch heterogeneous networks benefit from inverse-variance weighting fusion. For the model presented here, the latter strategy improved the accuracy to 97.60%, and feature visualisation analysis demonstrated that the model learned highly discriminative feature representations, forming clear category separations in the feature space.

The main contributions of this study are threefold: firstly, we have proposed a heterogeneous multi-branch network architecture tailored to multi-spectrogram representations, leveraging the complementary advantages of different time-frequency features through customised branch designs; secondly, we have developed multi-scale convolutional kernel structures adapted to specific feature modalities, which enhanced the model's ability to capture both the local details and global patterns inherent in underwater acoustic signals; and thirdly, we have established structure-dependent selection criteria for fusion strategies, providing a methodological reference for matching fusion methods with network architectures in multi-modal tasks. This work provides an effective technical solution for UATR, and offers valuable insights for other multi-modal information processing scenarios. Future work will focus on expanding the types of feature representation,

deepening cross-modal fusion mechanisms, and optimising the computational efficiency to further enhance the method's practicality and generalisation capability.

This study has several limitations that should be acknowledged. For example, the DeepShip dataset used in our experiments lacks detailed metadata on the vessels (e.g. gross tonnage, length, speed, propulsion type). Such additional information would be beneficial for further analyzing the model behavior. In future work, we aim to address this limitation by: (i) integrating auxiliary data sources such as automatic identification systems to obtain vessel metadata; (ii) collecting or curating a dataset with richer annotations; and (iii) investigating the relationship between acoustic features and specific vessel parameters through multi-task learning or attribute prediction frameworks to enhance the interpretability of the model.

REFERENCES

1. Thorp W H. Analytic description of the low-frequency attenuation coefficient. Journal of the Acoustical Society of America 1967. https://doi.org/10.1121/1.1910566.

2. Luo X, Chen L, Zhou H, Cao H. A survey of underwater acoustic target recognition methods based on machine learning. Journal of Marine Science and Engineering 2023. https://doi.org/10.3390/jmse11020384.

3. Sun Q, Zhou H. An acoustic sea glider for deep-sea noise profiling using an acoustic vector sensor. Polish Maritime Research 2022. https://doi.org/https://doi.org/10.2478/pomr-2022-0006.

4. Zieja M, Wawrzyński W, Tomaszewska J, Sigiel N. A method for the interpretation of sonar data recorded during autonomous underwater vehicle missions. Polish Maritime Research 2022. https://doi.org/https://doi.org/10.2478/pomr-2022-0038.

5. Zhang C, Xu Q, Yang H, Peng Z, Li J, Zhou J. Experimental study and numerical simulation of radiated noise from unmanned underwater vehicle. Polish Maritime Research 2024. https://doi.org/https://doi.org/10.2478/pomr-2024-0057.

6. Chen Y, Xu X. The research of underwater target recognition method based on deep learning. Proc. 2017 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), 2017. https://doi.org/10.1109/ICSPCC.2017.8242464.

7. Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 2002. https://doi.org/10.1109/5.18626.

8. Wang Q, Zeng X, Wang L, Wang H, Cai H. Passive moving target classification via spectra multiplication method. IEEE Signal Processing Letters 2017. https://doi.org/10.1109/LSP.2017.2672601.

9. Gent C, Sheppard C. Special feature: Predicting time series by a fully connected neural network trained by back propagation. Computing and Control Engineering 1992. https://doi.org/10.1049/cce:19920031.

10. Hinton G E, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. Neural Computation 2006. https://doi.org/10.1162/neco.2006.18.7.1527.

11. Lian Z, Xu K, Wan J, Li G. Underwater acoustic target classification based on modified GFCC features. Proc. 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2017. https://doi.org/10.1109/IAEAC.2017.8054017.

12. Wang X, Zhang Y, Xiao Z, Huang M. IS3L: An integrated self-training semi-supervised learning strategy for underwater acoustic target detection. Applied Acoustics 2023. https://doi.org/10.1016/j.apacoust.2023.109477.

13. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015. https://doi.org/10.1038/nature14539.

14. Piczak K J. Environmental sound classification with convolutional neural networks. Proc. 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), 2015. https://doi.org/10.1109/MLSP.2015.7324337.

15. Honghui Y, Junhao L, Meiping S. Underwater acoustic target multi-attribute correlation perception method based on deep learning. Applied Acoustics 2022. https://doi.org/10.1016/j.apacoust.2022.108644.

16. Shen S, Yang H, Li J, Xu G, Sheng M. Auditory inspired convolutional neural networks for ship type classification with raw hydrophone data. Entropy 2018. https://doi.org/10.3390/e20120990.

17. Han Y, Kim J, Lee K. Deep convolutional neural networks for predominant instrument recognition in polyphonic music. IEEE/ACM Transactions on Audio, Speech, and Language Processing 2016. https://doi.org/10.1109/TASLP.2016.2632307.

18. Zhang Q, Da L, Zhang Y, Hu Y. Integrated neural networks based on feature fusion for underwater target recognition. Applied Acoustics 2021. https://doi.org/10.1016/j.apacoust.2021.108261.

19. Pan X, Sun J, Feng T, Lei M, Wang H, Zhang W. Underwater target recognition based on adaptive multi-feature fusion network. Multimedia Tools and Applications 2025. https://doi.org/10.1007/s11042-024-19178-9.

20. Seo S, Kim C, Kim J-H. Convolutional neural networks using log mel-spectrogram separation for audio event classification with unknown devices. Journal of Web Engineering 2022. https://doi.org/https://doi.org/10.13052/jwe1540-9589.21216.

21. Khalilabadi M R. Underwater ship-radiated acoustic noise recognition based on mel-spectrogram and convolutional neural network. International Journal of Coastal, Offshore and Environmental Engineering (IJCOE) 2023. https://doi.org/https://doi.org/10.22034/ijcoe.2023.166732.

22. Sareen V, Seeja K. Speech emotion recognition using mel spectrogram and convolutional neural networks (CNN). Procedia Computer Science 2025. https://doi.org/https://doi.org/10.1016/j.procs.2025.04.624.

23. Huzaifah M. Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. arXiv preprint arXiv:1706.07156 2017. https://doi.org/10.48550/arXiv.1706.07156.

24. Bi F, Yang L. Research on acoustic scene classification based on time–frequency–wavelet fusion network. Sensors 2025. https://doi.org/10.3390/s25133930.

25. Meng X et al. A multi-time-frequency feature fusion approach for marine mammal sound recognition. Journal of Marine Science and Engineering 2025. https://doi.org/10.3390/jmse13061101.

26. Pham L, Phan H, Nguyen T, Palaniappan R, Mertins A, McLoughlin I. Robust acoustic scene classification using a multi-spectrogram encoder-decoder framework. Digital Signal Processing 2021. https://doi.org/10.1016/j.dsp.2020.102943.

27. Zheng W, Mo Z, Xing X, Zhao G. CNNs-based acoustic scene classification using multi-spectrogram fusion and label expansions. arXiv preprint arXiv:1809.01543 2018. https://doi.org/10.48550/arXiv.1809.01543.

28. Irfan M, Jiangbin Z, Ali S, Iqbal M, Masood Z, Hamid U. DeepShip: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification. Expert Systems with Applications 2021. https://doi.org/10.1016/j.eswa.2021.115270.

29. Zhu P, Zhang Y, Huang Y, Zhao C, Zhao K, Zhou F. Underwater acoustic target recognition based on spectrum component analysis of ship radiated noise. Applied Acoustics 2023. https://doi.org/10.1016/j.apacoust.2023.109552.