scientific reports

Check for updates

OPEN

Hybrid local-global representation learning with stochastic Gaussian classification for underwater acoustic target recognition

Cheng Yang[1,2,4], Qisheng Xu[1,2,4], Ming Feng[1,2], Hui Yang[1], Yulei Yuan[1], Yutao Dou[3], Junyi Zhao[1,2] & Kele Xu[1,2]✉

Underwater acoustic target recognition is critical for a broad spectrum of marine applications, yet its performance is often hindered by environmental variability, non-stationary propagation effects, and inherently low signal-to-noise ratio conditions. This work presents a novel hybrid deep learning framework that synergistically integrates global and local acoustic representations to enhance recognition robustness under such adverse conditions. The proposed approach employs a dual-branch encoder: a pre-trained self-supervised Audio Spectrogram Transformer branch to capture long-range temporal dependencies, and a multi-scale convolutional branch to extract fine-grained local spectral patterns. To further improve decision stability and mitigate uncertainty near classification boundaries, we introduce a Gaussian sampling-based classification module, which models class-specific weights as probabilistic distributions and performs Monte Carlo inference. Experiments on two representative underwater acoustic benchmark demonstrate that the proposed method not only achieves state-of-the-art recognition accuracy but also exhibits strong resilience to different environmental noise. Ablation analyses further validate the complementary advantages of local-global feature fusion and the probabilistic decision mechanism. These findings suggest that the proposed hybrid architecture offers a promising and practical solution for robust underwater acoustic classification in real-world operational scenarios.

Underwater acoustic target recognition (UATR) is a fundamental capability underpinning a wide range of maritime applications, including naval surveillance, autonomous underwater vehicle (AUV) navigation, ocean environment monitoring, and maritime situational awareness[1,2]. In the underwater domain, where electromagnetic wave propagation is severely constrained, acoustic sensing remains the most reliable long-range modality. However, robust recognition of underwater targets from acoustic signals remains challenging due to the highly complex and non-stationary nature of the acoustic channel. This complexity arises from multipath propagation, frequency-dependent attenuation, Doppler shifts, and environmental noise originating from both biological and anthropogenic sources[3], all of which degrade signal fidelity and hinder automatic target recognition (ATR)[4].

Convolutional neural networks (CNNs) have been widely adopted for underwater acoustic classification owing to their strong inductive bias for detecting localized patterns in time-frequency representations[5,6]. By exploiting local receptive fields, CNNs effectively capture salient spectral-temporal cues such as tonal harmonics and transient bursts[7]. Nevertheless, their fixed and limited receptive field hampers the mtation learning. Our dual-branch encoder integrates a pre-trained Transformer branch for long-range temporal modeling with a multi-scale convolutional branch for frequency-sensitive feature extraction, enabling robust joint learning of global and local acoustic cues. Beyond feature extraction, we introduce a Gaussian sampling-based stochastic classifier, which performs probabilistic ensembling at inference to enhance robustness and provide uncertainty-aware predictions under low-SNR and ambiguous conditions. Extensive evaluations on benchmark datasets demonstrate that our method consistently outperforms state-of-the-art approaches, with ablation studies confirming the complementary benefits of both the hybrid encoder and the stochastic classification module. This work should be of interest to the broad audiences that Scientific Reports wishes to reach, odeling of long-

[1]College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China. [2]National Key Laboratory of Parallel and Distributed Computing, National University of Defense Technology, Changsha 410073, China. [3]Hunan University, Changsha 410082, China. [4]Cheng Yang and Qisheng Xu have contributed equally to this work. ✉email: kele.xu@ieee.org

range temporal dependencies—an essential property for underwater acoustic signals that may exhibit slow-varying or intermittent modulation patterns.

Transformer-based architectures[8], initially developed for natural language processing, have recently demonstrated strong potential in acoustic modeling [9–14]. Their self-attention mechanism enables global contextual reasoning and the capture of long-range dependencies, both of which are critical for decoding complex underwater acoustic structures. However, Transformers often lack local inductive biases and demand large-scale labeled datasets to generalize effectively-limitations that are particularly pronounced in underwater applications, where annotated data are scarce and costly to obtain[15,16]. As illustrated in Fig. 1, SSAST treats the entire log-Mel spectrogram with uniform attention, thereby neglecting important local patterns.

To overcome these challenges, we propose a novel hybrid neural architecture that unifies the global modeling capacity of a pre-trained Transformer with the fine-grained spectral sensitivity of convolutional networks. Specifically, our dual-branch encoder comprises: (i) a self-supervised Audio Spectrogram Transformer (SSAST) branch to model long-range temporal structures, and (ii) a frequency-aware multi-scale convolutional branch to extract localized spectral features via residual encoding. This complementary design enables the joint learning of global and local acoustic representations, thereby improving discriminability in complex and noisy environments.

In addition to feature extraction, we address decision uncertainty by introducing a Gaussian sampling-based stochastic classification module. Unlike conventional deterministic classifiers, our method models each class-specific weight vector as a multivariate Gaussian distribution, performing Monte Carlo sampling at inference. This probabilistic formulation implicitly ensembles multiple decision boundaries, enhances robustness near class margins, and enables uncertainty-aware predictions—a particularly desirable property under low signal-to-noise ratio (SNR) conditions and in the presence of ambiguous targets.

We validate our approach on representative benchmark datasets, demonstrating consistent improvements over prior state-of-the-art methods and strong robustness against additive and environmental noise. Comprehensive ablation experiments confirm the complementary contributions of the local-global feature fusion and the stochastic classification mechanism. In summary, the main contributions of this work are as follows:

- We propose a hybrid local-global representation learning framework integrating a self-supervised SSAST Transformer for global contextual modeling and a multi-scale convolutional pathway for local feature extraction, tailored to the complex characteristics of underwater acoustic signals.
- A stochastic classification module based on Gaussian weight sampling, which enhances decision diversity, improves generalization, and enables uncertainty-aware prediction.
- Comprehensive evaluation and ablation studies on real-world underwater datasets, demonstrating superior recognition accuracy and robustness in noisy operational scenarios.

The remainder of this paper is organized as follows. Section 2 reviews relevant literature on UATR. Section 3 details the proposed framework. Section 4 describes the experimental setup and presents performance evaluation. Finally, Section 5 concludes the paper and discusses future research directions.

Related work

Traditional underwater acoustic target recognition methods

Early approaches to UATR primarily relied on *handcrafted feature engineering* combined with conventional machine learning classifiers. Feature extraction served as the cornerstone of these methods, transforming raw time-series signals into compact representations that preserved intrinsic signal characteristics while suppressing the effects of oceanic noise, reverberation, and propagation distortion. Commonly used acoustic descriptors included Mel-Frequency Cepstral Coefficients (MFCCs)[17], wavelet-based features[18], and time–frequency spectral statistics[19]. These features were typically paired with classifiers such as Support Vector Machines (SVMs)[20], Gaussian Mixture Models (GMMs)[21], *k*-Nearest Neighbors (KNN)[22], or Random Forests[23].

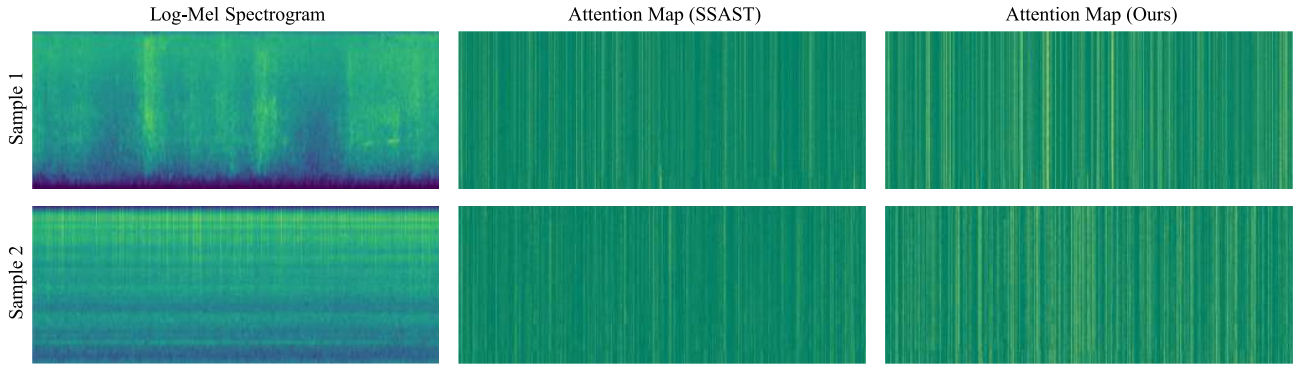


Fig. 1. Visualization of class activation maps on the log-Mel spectrogram, highlighting important time–frequency regions in comparison between the baseline SSAST and our hybrid local-global representation learning framework.

While traditional pipelines were computationally efficient and interpretable, they faced two main limitations: (i) high sensitivity to challenging underwater conditions, including variable SNR, multipath propagation, and non-stationary noise; and (ii) strong dependence on expert knowledge for the manual design and selection of discriminative features[24]. Feature extraction was typically categorized into time-domain, frequency-domain, and time-frequency representations, with classification accuracy highly dependent on the suitability of the chosen descriptors[25]. Although modern deep learning methods have largely outperformed handcrafted approaches in complex environments[26], the latter remain relevant in scenarios requiring real-time inference on resource-limited platforms or when labeled data are scarce.

Deep learning–based underwater acoustic target recognition methods

The limitations of handcrafted pipelines have driven a shift toward **end-to-end deep learning** methods capable of directly learning hierarchical representations from raw waveforms or spectrograms[1,27]. Among these, Convolutional Neural Networks (CNNs) have become the dominant paradigm in UATR, owing to their inductive bias toward localized time-frequency patterns[28,29]. CNNs automatically learn convolutional filters that capture key components such as tonal harmonics, spectral sweeps, and modulated frequency structures-features that previously required manual engineering[30,31].

Early work adopted shallow CNN architectures for ship-radiated noise classification[32], while later studies explored deeper and more expressive models, including ResNet variants[33] and attention-augmented CNNs[34,35]. Xiao *et al.*[35], for example, incorporated both channel-wise and spatial attention mechanisms to enhance feature discrimination. Additional improvements have included spectral pyramid encoding, frequency-band enhancement, and specialized loss functions that improve class separability in complex acoustic environments.

To improve generalization in noisy or low-resource settings, Fang *et al.*[36] introduced a momentum adversarial training strategy that enforces robustness under domain shifts such as vessel type variability and environmental changes. However, CNNs remain inherently limited by their local receptive fields, which constrain their ability to capture the long-range temporal dependencies characteristic of underwater acoustic signals, especially when modulation patterns evolve slowly or occur intermittently.

Transformer-based acoustic signal modeling

Transformer architectures[8], originally developed for natural language processing, leverage self-attention mechanisms to capture global temporal dependencies and model non-local relationships within sequences. Their success in speech and general audio tasks[10,12,13] has motivated their adoption in UATR, although the field remains relatively nascent. Li *et al.*[37] proposed the Spectrogram Transformer (STM), which outperformed ResNet and CRNN baselines, particularly for signals with extended temporal structures. Fan *et al.*[38] developed an end-to-end soft-threshold Swin Transformer (ESTMST-ST) incorporating a learnable dual filter module, soft-threshold mechanism, and a multi-loss self-distillation strategy, achieving substantial performance gains on the ShipsEar and DeepShip datasets. In the self-supervised domain, Feng *et al.*[39] introduced MHT-UATR, a hierarchical masked token learning framework for extracting structure-aware features from Mel-spectrograms without requiring manual labels, thereby improving robustness to occlusion and noise.

While these studies demonstrate the potential of Transformers for UATR, two challenges persist: (i) a lack of strong local inductive biases for capturing short-term spectral details, and (ii) high data requirements, which conflict with the scarcity of labeled underwater datasets[15,16].

Hybrid CNN–transformer architectures

To address the complementary weaknesses of CNNs and Transformers, hybrid architectures that integrate the local pattern recognition capabilities of CNNs with the global contextual modeling of Transformers have emerged in other domains such as computer vision[40,41] and natural language processing[42,43]. Such designs offer a promising pathway for UATR by leveraging CNNs to extract fine-grained spectral features while enabling Transformers to capture long-range temporal relationships. However, their adaptation to underwater acoustic recognition remains underexplored, particularly for scenarios characterized by high uncertainty, low SNR, and limited annotated data. These limitations highlight the need for architectures that can not only fuse complementary global and local representations, but also incorporate uncertainty-aware decision mechanisms to enhance robustness in real-world conditions.

Motivated by this gap, we propose a dual-branch hybrid framework that integrates a self-supervised Transformer for global context modeling with a multi-scale convolutional pathway for local feature extraction, coupled with a probabilistic classifier to improve decision reliability under adverse acoustic environments.

Methodology

Let $x(t)$ denote an input underwater acoustic signal. The objective of UATR is to assign $x(t)$ to one of C target classes $\mathscr{Y} = \{1, 2, \ldots, c\}$, under adverse conditions such as low signal-to-noise ratio (SNR), time-varying multipath propagation, and non-stationary background noise. These factors produce spectral smearing, temporal distortion, and high inter-class confusion, which pose two primary challenges: (*i*) learning robust representations that capture both long-range temporal structure and local spectral detail, and (*ii*) making reliable predictions under label ambiguity and noise-induced decision uncertainty.

To address these challenges, we design a hybrid architecture with three synergistic components:

1. *A dual-branch local-global feature extractor* that exploits a self-supervised Transformer to capture global temporal dependencies and a multi-scale CNN to preserve fine-grained spectral cues critical to UTAR.
2. *A feature fusion and alignment module* that adaptively integrates complementary embeddings from the two branches, balancing context modeling with local sensitivity.

3. *A Gaussian sampling-based stochastic classifier* that models each class boundary probabilistically to mitigate decision instability in low-SNR and ambiguous scenarios.

Given an input underwater acoustic signal $x(t)$, our method first converts it into a log-Mel spectrogram to capture perceptually relevant spectral-temporal patterns. The spectrogram is then processed by a *dual-branch feature extractor*: a self-supervised Audio Spectrogram Transformer branch models long-range temporal dependencies to capture global contextual information, while a multi-scale convolutional branch extracts fine-grained local spectral cues critical for distinguishing acoustically similar targets. The outputs of the two branches are *fused via an adaptive alignment module*, which balances global and local representations to produce a joint embedding that preserves both long-range context and detailed spectral features. Finally, a *Gaussian sampling-based stochastic classifier* models each class weight vector as a multivariate Gaussian distribution, performing Monte Carlo sampling during inference to generate robust and uncertainty-aware predictions. This end-to-end pipeline jointly learns complementary local and global representations while accounting for decision uncertainty, enabling accurate and reliable underwater acoustic target recognition under challenging low-SNR and noisy conditions. The overall pipeline is illustrated in Fig. 2.

Dual-branch local–global feature extractor

Log-mel spectrogram

We first convert $x(t)$ into a *log-mel spectrogram* $\mathbf{S} \in \mathbb{R}^{F \times T}$, which is perceptually motivated and compresses spectral dynamics in a way that is robust to small frequency shifts common in underwater propagation. The magnitude spectrum is obtained via STFT:

$$\mathbf{X}(f,\tau) = \left| \sum_{n=0}^{N-1} x[n] w[n - \tau H] e^{-j2\pi f n/N} \right|, \tag{1}$$

where $w[\cdot]$ is a Hamming window and H is the hop size. A Mel filterbank $\mathbf{M} \in \mathbb{R}^{F \times F_{\mathrm{STFT}}}$ projects $\mathbf{X}$ into the perceptual frequency scale:

$$\mathbf{S}_{\mathrm{Mel}}(m,\tau) = \sum_{f} \mathbf{M}(m,f)\, \mathbf{X}(f,\tau), \tag{2}$$

and the log operation with offset ϵ ensures numerical stability:

$$\mathbf{S}(m,\tau) = \log\left(\mathbf{S}_{\mathrm{Mel}}(m,\tau) + \epsilon\right), \tag{3}$$

The log-mel spectrogram is normalized to zero mean and unit variance.

Global branch: SSAST

The global branch leverages a Self-Supervised Audio Spectrogram Transformer (SSAST) pretrained on large-scale audio dataset, to learn contextualized embeddings with long-range temporal dependencies, which are important in UTAR for recognizing targets with slow or intermittent modulation patterns. Specifically, the spectrogram $\mathbf{S}$ is divided into M non-overlapping patches $\{\mathbf{p}_i\}$ of size $F_p \times T_p$, which are projected into a higher-dimensional embedding space through a patch embedding operation:

$$\mathbf{E}_i = \mathbf{w}_p\, \mathrm{vec}(\mathbf{p}_i) + \mathbf{b}_p, \quad \mathbf{E}_i \in \mathbb{R}^d, \tag{4}$$

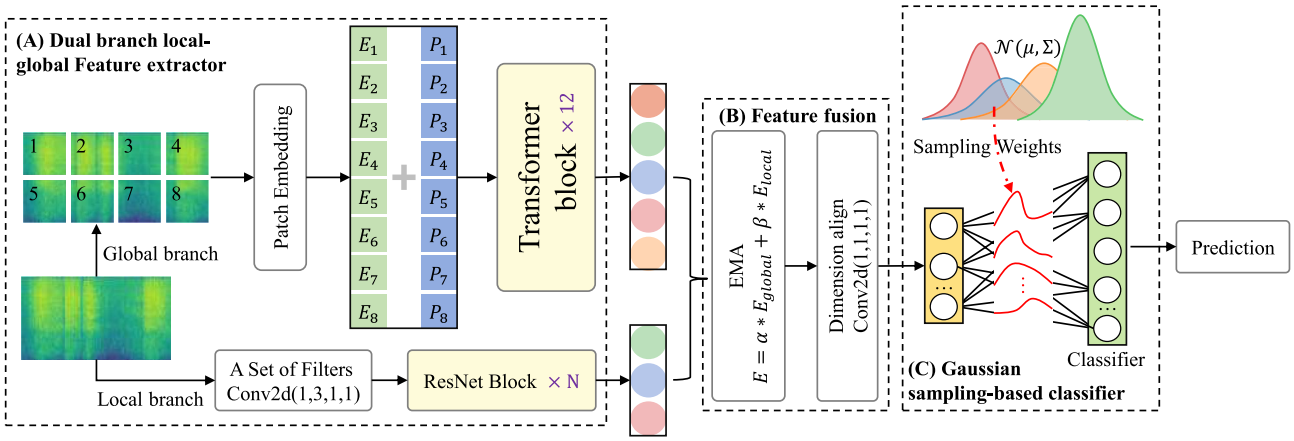


Fig. 2. Overview of the proposed framework for underwater acoustic target recognition. The architecture comprises (**A**) a dual-branch local–global feature extractor, (**B**) a feature fusion and alignment module, and (**C**) a Gaussian sampling-based classifier.

where $\mathbf{E}_i \in \mathbb{R}^d$ is obtained through patch embedding, enabling the model to learn richer representations and capture complex temporal-spectral relationships. And $\mathbf{w}_p$ denotes the patch embedding weights and $\mathbf{b}_p$ the corresponding bias. To preserve order information, a positional embedding $\mathbf{P}_i \in \mathbb{R}^d$ of the same dimension is added:

$$\mathbf{Z}_i = \mathbf{E}_i + \mathbf{P}_i, \tag{5}$$

after which the sequence $\mathbf{Z} = \{\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_M\}$ is fed into L_T Transformer encoder blocks.

$$\mathbf{Z}' = \mathbf{Z} + \mathrm{MHSA}(\mathrm{LN}(\mathbf{Z})), \tag{6}$$

$$\mathbf{Z}_{\mathrm{out}} = \mathbf{Z}' + \mathrm{FFN}(\mathrm{LN}(\mathbf{Z}')). \tag{7}$$

yielding global embeddings $\mathbf{E}_{\mathrm{global}} \in \mathbb{R}^{M \times d}$.

Local branch: Multi-scale CNN

The global branch excels at capturing long-range temporal dependencies through the Transformer's self-attention mechanism, enabling effective modeling of global contextual patterns. However, it lacks inductive bias for short-term spectral details, which are critical for distinguishing vessels with similar overall structures but distinct tonal line patterns. To overcome this limitation, we introduce a local branch composed of L_C filters, where each filter applies parallel convolutions with multiple kernel sizes. Smaller kernels are effective for detecting fine-grained spectral details such as tonal lines, while larger kernels capture broader frequency-temporal cues, as demonstrated in[44]. Specifically, the spectrogram $\mathbf{S}$ is filtered by:

$$\begin{aligned} \mathbf{S}_c[i,j] &= \mathrm{Conv}(\mathbf{S} * h)[i,j] \\ &= \sum_{m=-M}^{M} \sum_{n=-N}^{N} \mathbf{S}[i-m, j-n]\, h[m,n] + b_k, \end{aligned} \tag{8}$$

where $\mathbf{S}_c[i,j]$ is the output at spatial position (i, j) for channel c, $h[m, n]$ denotes the convolution kernel, M and N specify the kernel's half-size along the two dimensions, and b_k is the bias term. This convolution captures spectral cues at different resolutions, complementing the global branch and enhancing the model's ability to discriminate subtle tonal variations.

To learn richer and more hierarchical feature representations, the output of the convolutional filter is further processed by a residual block. The skip connection preserves fine-grained spectral details while enabling the modeling of higher-level abstractions. When combined with parallel convolutions of different kernel sizes, the residual block facilitates multi-resolution feature extraction, capturing both short-term tonal lines and broader frequency-temporal patterns. This process can be formulated as:

$$\mathscr{F}(\mathbf{S}_c) = \sigma(\mathrm{Conv}_2(\sigma(\mathrm{Conv}_1(\mathbf{S}_c)))), \tag{9}$$

$$\mathbf{E}_{local} = \mathbf{S}_c + \mathscr{F}(\mathbf{S}_c). \tag{10}$$

where Conv_1 and Conv_2 are convolution operations, and $\sigma(\cdot)$ denotes the ReLU activation. The outputs are then projected to d channels and reshaped to $\mathbf{E}\mathrm{local} \in \mathbb{R}^{M \times d}$ for alignment.

Feature fusion

Although the global branch effectively captures long-range temporal-spectral context and the local branch emphasizes fine-grained spectral details, either branch alone remains insufficient for comprehensive representation. Specifically, the global branch may overlook subtle tonal variations, while the local branch may fail to encode broader contextual dependencies. To leverage their complementary strengths, we integrate their embeddings through a weighted feature fusion mechanism:

$$\mathbf{E} = \alpha \mathbf{E}_{global} + (1-\alpha)\mathbf{E}_{local}, \tag{11}$$

where α is an empirically chosen hyperparameter that enables the network to adaptively balance global context and local detail under varying task and noise conditions. The fused embedding $\mathbf{E}$ is then normalized into a fixed-dimensional vector $\mathbf{E} \in \mathbb{R}^d$. This fusion strategy enhances the model's ability to capture subtle vocalization patterns while preserving robustness against noise, thereby providing a more informative and reliable representation for downstream classification tasks.

Gaussian sampling–based stochastic classifier

Conventional UTAR classifiers typically employ a single linear layer or a shallow MLP, where each class is represented by a fixed weight vector. While simple, such classifiers are prone to overfitting noise artifacts and may produce brittle decision boundaries, as the weight vectors cannot adapt to variations in the input embedding distribution. To mitigate this, we model each class weight vector $\mathbf{w}_c$ as a Gaussian random variable:

$$\mathbf{w}_c \sim \mathscr{N}(\boldsymbol{\mu}_c, \mathrm{diag}(\boldsymbol{\sigma}_c^2)), \tag{12}$$

where $\boldsymbol{\mu}_c$ and $\boldsymbol{\sigma}_c$ represent the mean and standard deviation of the class weight vector, respectively, and both are learnable parameters. The standard deviation is parameterized as $\boldsymbol{\sigma}_c = \mathrm{softplus}(\boldsymbol{\rho}_c)$ to ensure positivity. Unlike conventional classifiers with fixed weights, this probabilistic formulation captures uncertainty in the class representations, enabling the model to more effectively handle noisy or ambiguous inputs.

During training, the mean and variance parameters $(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c)$ are optimized using standard cross-entropy loss, with weight sampling incorporated to approximate the expected logits over the Gaussian distribution. This encourages the model to learn weight distributions that are robust to input variations rather than relying on single deterministic weights.

During inference, K samples of each weight vector are drawn using the reparameterization strategy:

$$\mathbf{w}_c^{(k)} = \boldsymbol{\mu}_c + \boldsymbol{\sigma}_c \odot \mathbf{z}^{(k)}, \quad \mathbf{z}^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{13}$$

which are then used to compute the corresponding class logits and probabilities:

$$p_c^{(k)} = \frac{\exp(\mathbf{w}_c^{(k)\top}\mathbf{e})}{\sum_{c'} \exp(\mathbf{w}_{c'}^{(k)\top}\mathbf{e})}, \quad \bar{p}_c = \frac{1}{K}\sum_{k=1}^{K} p_c^{(k)}. \tag{14}$$

By averaging over multiple stochastic weight samples, this approach effectively forms an ensemble of classifiers without additional network parameters. The resulting stochastic ensembling reduces prediction variance, enhances robustness to spectral distortions, and provides a natural measure of predictive uncertainty, which is particularly valuable for operational decision-making in challenging acoustic environments.

Model architecture and training objective

The proposed model comprises two parallel branches: a Self-Supervised Audio Spectrogram Transformer (SSAST) branch and a CNN branch. The SSAST branch adopts the base architecture and is initialized with weights obtained from self-supervised pretraining, thereby providing strong prior representations. In contrast, the CNN branch is initialized using He initialization[45] to ensure stable gradient propagation. Specifically, the CNN branch consists of a 1×1 convolution, followed by a 3×3 convolution, a residual block composed of a 3×3 convolution with ReLU activation and another 3×3 convolution with ReLU, and a final 1×1 convolution.

The two branches are trained jointly in an end-to-end manner. The final prediction is obtained by averaging the outputs from both branches, and the model is optimized using the cross-entropy loss:

$$\mathscr{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \log \bar{p}_{i,c}, \tag{15}$$

where $\bar{p}_{i,c}$ denotes the averaged class probability of sample i for class c, $y_{i,c}$ is the corresponding one-hot label, N is the batch size, and C is the total number of classes.

Experiments

In this section, we evaluate the effectiveness of the proposed hybrid CNN-inserted Transformer model with stochastic classification on benchmark underwater acoustic datasets. We compare our method against state-of-the-art CNN and Transformer-based baselines under various conditions, including low signal-to-noise ratio (SNR) and limited training data. We also conduct ablation studies to assess the contribution of each architectural component.

Evaluation setup

We evaluate the proposed hybrid CNN-Transformer model on the DeepShip dataset and ShipsEar dataset, two widely used benchmark for underwater acoustic target recognition. These dataset consists of passive sonar recordings of various surface vessels under diverse environmental and operational conditions. To ensure fair comparison and robustness, all models are trained and tested using stratified splits, and results are averaged over three independent runs.

All models are implemented in PyTorch and trained on NVIDIA RTX 3090 GPUs. The log-mel spectrograms are extracted with a frame length of 1024 and hop length of 320, and normalized per utterance.

Training is conducted using the Adam optimizer with an initial learning rate of 5e-4, batch size of 32, and cosine annealing for 100 epochs. For the stochastic classifier, we use $K = 5$ classifiers with dropout rate $p = 0.3$. Early stopping is applied based on validation F1-score.

Benchmark datasets

This paper evaluates the proposed method on the DeepShip dataset. DeepShip[46] comprises approximately 47 hours of real-world underwater acoustic recordings captured from 265 distinct vessels under diverse sea states and varying ambient noise conditions, providing a challenging benchmark for underwater acoustic target recognition. The original dataset categorizes vessels into four classes; following prior studies[12,47], we augment the dataset by incorporating background noise as a fifth class, thereby enhancing its realism and suitability for evaluating model performance under noisy operational scenarios. ShipsEar[48] is an open-source underwater acoustic database containing real-world recordings collected near the Port of Vigo, Spain. The dataset consists of 90 raw underwater sound recordings spanning 11 representative vessel types, including fishing boats, trawlers, tugboats, dredgers, pilot boats, sailboats, ferries, and large ocean-going vessels, along with natural ambient

ocean noise. Following prior studies[48], we regroup these categories into five broader classes to ensure consistent labeling and sufficient sample coverage for model evaluation.

For experimental rigor, the dataset is randomly partitioned into training and test subsets in a 7:3 ratio, and five-fold cross-validation is employed to ensure a robust and statistically meaningful performance assessment. For feature representation, log-Mel spectrograms are computed using a 10 ms frame window and 128 frequency bins, serving as the input to our proposed hybrid learning framework. This configuration allows effective capture of both spectral and temporal characteristics of underwater acoustic signals, facilitating the evaluation of our model's capability in complex and noisy environments.

Evaluation metrics

To comprehensively assess model performance, we employ a set of widely used classification metrics. While these metrics are standard for binary classification, they can be naturally extended to multi-class settings. A common strategy is to treat each class as a one-vs-rest binary problem, compute the corresponding score per class, and then report either the macro-averaged (unweighted mean across classes) or micro-averaged (global aggregation) results. For example, the multi-class F1-score is typically computed as the average of per-class F1-scores.

Let *TP*, *TN*, *FP*, and *FN* denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. The metrics are formally defined as follows:

- **Accuracy (Acc)**: The proportion of correctly classified samples among all samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{16}$$

- **Precision (Prec)**: The fraction of true positive predictions among all positive predictions. Precision reflects the model's ability to avoid false alarms, which is particularly critical in safety-sensitive applications:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{17}$$

- **Recall (Rec)**: The fraction of actual positives that are correctly identified. Recall evaluates the model's ability to capture target events, which is crucial when missed detections are costly:

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{18}$$

- **F1-score (F1)**: The harmonic mean of precision and recall. F1 provides a balanced measure in scenarios with uneven trade-offs between false alarms and missed detections, and is especially useful under class imbalance:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{19}$$

Together, these metrics capture complementary aspects of classification performance, enabling a robust and fair evaluation across diverse operating conditions.

Baseline methods

To evaluate the effectiveness of our proposed model, we compare it against a comprehensive set of baseline methods, including traditional machine learning classifiers, conventional supervised deep neural networks, and recent self-supervised and Transformer-based architectures.

Traditional machine learning methods include Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN), all trained using log-mel spectrogram features. These approaches represent early-stage acoustic classification pipelines and serve as classical benchmarks in underwater acoustic target recognition.

Supervised deep learning baselines comprise several representative architectures, such as Deep Neural Network (DNN), Residual CNN, Inception, Stacked Convolutional Autoencoder (SCAE), and Ship radiated Noise spectrum component Analysis-based Network (SNANet, Applied acoustics 2023[49]). These models are fully trained with labeled data and emphasize the ability of deeper networks to capture hierarchical representations from spectrogram inputs.

Self-supervised and advanced learning methods are also included in our comparison. These encompass recent audio pre-training models such as SSAST (AAAI 2022[10]), AudioMAE (NeurIPS 2022[50]), and SSLMM (JASA 2023[12]), which leverage unlabeled data through masked prediction or contrastive objectives. MIXUP (IEEE JSTARS 2023[47]) applies data-level augmentation to improve robustness. TR-Tral (IEEE/ACM TASLP 2024[16]) represents a strong Transformer-based architecture trained with time-frequency masking strategies.

Quantitative results and analysis

Table 1 presents the quantitative evaluation of all baseline methods alongside our proposed approach on the DeepShip dataset. Several key trends emerge across traditional, supervised, and self-supervised approaches. Traditional machine learning methods such as SVM and RF achieve moderate performance, with SVM reaching

an F1-score of 72.28%. These models, although efficient, are inherently limited in capturing complex temporal and spectral structures present in underwater acoustic signals, particularly under multipath propagation and strong background interference. KNN performs the worst among traditional models due to its sensitivity to noisy or overlapping features.

Supervised deep learning models exhibit a clear performance improvement over traditional baselines. Architectures such as SCAE and Residual CNN achieve F1-scores of 77.58% and 76.92%, respectively, highlighting the benefits of hierarchical and residual feature extraction. Inception-based variants provide further gains by incorporating multi-scale convolutional filters. Nevertheless, these approaches remain constrained by their reliance on local receptive fields and the requirement for large volumes of labeled data, which hinders generalization in data-scarce scenarios.

Self-supervised learning approaches further advance performance by leveraging large-scale pretraining. SSAST, AudioMAE, and SSLMM consistently outperform supervised models, confirming the utility of label-efficient pretraining for underwater acoustics. Notably, AudioMAE yields high precision (85.54%) but relatively lower recall, reflecting confident yet conservative predictions. More recent methods incorporating data augmentation and Transformer architectures, such as MIXUP and TR-Tral, push performance further, with TR-Tral attaining an F1-score of 87.50%. Our proposed model achieves the best overall performance, establishing new state-of-the-art results on DeepShip: **88.48%** accuracy, **89.42%** precision, **89.41%** recall, and **89.41%** F1-score.

To further demonstrate the cross-dataset generalization of our method, we conducted additional experiments on the ShipsEar dataset, which contains similar maritime sound classes but under different recording conditions compared to DeepShip. The quantitative results are summarized in the Table 2. Our method achieves 98.62% accuracy, 98.36% precision, 98.76% recall, and 98.56% F1-score, significantly outperforming both traditional machine learning approaches (e.g., SVM 83.10%, RF 81.35%) and recent deep learning methods, including supervised and self-supervised models such as SSAST (92.62%), AudioMAE (89.38%), and SNANet (93.13%). These results demonstrate that our approach can achieve strong performance across different datasets, highlighting its robustness and adaptability to varying maritime acoustic conditions. We think this consistent improvement across all metrics arises from two key innovations:

- *Hybrid representation learning*: The combination of a self-supervised pretrained SSAST branch with a CNN branch allows the model to jointly capture global contextual patterns and localized spectral–temporal structures, thereby enhancing feature diversity and complementarity.
- *Stochastic prediction head*: The incorporation of a random perturbation–based classification head reduces reliance on deterministic decision boundaries, improving generalization to unseen conditions and robustness against noisy or imbalanced labels.

Overall, these findings demonstrate that effective UTAR requires modeling both local spectral detail and global temporal context. While SSL methods benefit significantly from pretraining, our hybrid architecture, trained end-to-end in a supervised manner, still outperforms them, underscoring the structural advantages of combining convolutional and attention mechanisms for robust underwater acoustic target recognition.

Ablation study

To validate the effectiveness of the proposed designs, we conduct an ablation study on two key components: the stochastic classification (SC) head and the hybrid CNN–Transformer structure (hybrid branch). The results are

Type	Method	Venue	Sup.	Self-sup.	Acc.	Prec.	Rec.	F1
Traditional ML	SVM	ESA (2021)	✓		72.24	72.49	72.08	72.28
	RF		✓		69.71	69.79	69.86	69.82
	KNN		✓		62.71	63.61	63.10	63.35
Deep Learning	SCAE	ESA (2021)	✓		77.53	77.75	77.41	77.58
	Residual CNN		✓		76.98	77.05	76.81	76.92
	Inception		✓		76.16	76.03	76.12	76.08
	DNN		✓		73.11	72.98	73.08	73.03
	SSAST	AAAI 2022		✓	77.70	78.13	78.25	78.19
	AudioMAE	NeurIPS 2022		✓	76.66	85.54	79.00	82.14
	SSLM-M	JASA 2023		✓	80.22	80.81	79.94	80.07
	SNANet	AA 2023	✓		78.25	79.55	79.39	79.16
	MIXUP	JSTARS 2023		✓	86.33	85.72	82.91	84.29
	TR-Tral	TASLP 2024		✓	87.26	87.45	87.80	87.50
	Ours	—		✓	**88.48**	**89.42**	**89.41**	**89.41**

Table 1. Quantitative Comparison on DeepShip Dataset. Models are grouped by method category and by supervised or self-supervised training. All values are reported in percentage (%). Bold values denotes the best performance for this metric.

Type	Method	Venue	Sup.	Self-sup.	Acc.	Prec.	Rec.	F1
Traditional ML	SVM	ESA (2021)	✓		83.10	83.39	83.10	83.24
	RF		✓		81.35	81.54	81.35	81.44
	KNN		✓		73.72	76.03	73.72	74.86
Deep Learning	SCAE	ESA (2021)	✓		–	–	–	–
	Residual†		✓		84.71	83.92	83.05	83.48
	Inception		✓		–	–	-	–
	DNN		✓		83.48	83.31	82.53	82.92
	SSAST	AAAI 2022		✓	92.62	92.94	94.00	93.47
	AudioMAE	NeurIPS 2022		✓	89.38	90.16	87.95	89.04
	SSLMM	JASA 2023		✓	–	–	–	–
	SNANet	AA 2023	✓		93.13	93.02	91.97	92.49
	MIXUP	JSTARS 2023		✓	88.70	90.20	86.80	88.47
	TR-Tral	TASLP 2024		✓	–	–	–	–
	Ours			✓	**98.62**	**98.36**	**98.76**	**98.56**

Table 2. Quantitative comparison on ShipsEar dataset. Models are grouped by method category and by supervised or self-supervised training. All values are reported in percentage (%).† indicates results cited from the original reference, and - denotes unavailable results due to the code not being publicly accessible. Bold values denotes the best performance for this metric.

	Accuracy	Precision	Recall	F1-Score
baseline	84.03	86.98	85.51	86.24
+ SC	88.10	89.12	89.07	89.10
+ hybrid branch	**88.48**	**89.42**	**89.41**	**89.41**

Table 3. The validation of proposed two key design. All values are reported in percentage (%). Bold values denotes the best performance for this metric.

summarized in Table 3. We select SSAST with Mixup augmentation as the baseline, as it demonstrates strong performance with relatively low reliance on labeled data.

Incorporating the stochastic classification head consistently improves performance across all metrics, confirming its ability to introduce adaptive decision boundaries that enhance robustness under noisy and imbalanced label conditions. Building on this, integrating local convolutional branch with Transformer encoder yields further gains, reflecting the benefit of hybrid local–global modeling that captures fine-grained spectral cues while preserving long-range temporal dependencies. These results indicate that the stochastic classification head primarily drives the robustness and adaptability of the model, while the hybrid CNN–Transformer design complements it by strengthening feature representations, ultimately leading to the best overall performance.

Parameter sensitivity analysis

Temperature in random classifier

To assess the stability and robustness of the proposed model under varying hyperparameter configurations, we conducted a sensitivity analysis on the temperature parameter T used in the auxiliary random classifier. This parameter modulates the sharpness of the softmax output, thereby affecting the gradient signal propagated to the backbone during training. We systematically varied T from 8 to 24 and evaluated model performance on the DeepShip dataset.

As shown in Table 4, model performance is generally stable across a wide range of temperature values, demonstrating the robustness of the auxiliary classifier design. However, the results also reveal a clear peak at $T = 14$, where the model achieves the highest accuracy (88.10%) and F1-score (89.10%). This suggests that moderate softmax smoothness helps balance gradient flow from the auxiliary head and avoids overconfident supervision.

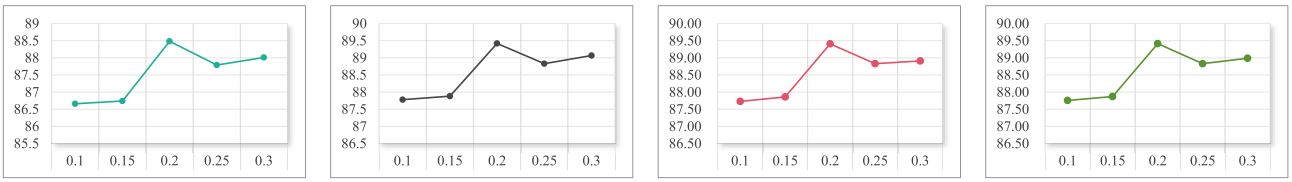
When the temperature is too low (e.g., $T = 8$), the auxiliary logits become sharper, which may result in unstable gradients and overfitting to noisy random targets. Conversely, when the temperature is too high (e.g., $T = 24$), the supervision signal becomes overly diffused, reducing its regularization effect. The proposed framework exhibits strong resilience to the temperature setting, while $T = 14$ provides an empirically optimal balance between training stability and representational diversity. We adopt this setting in all subsequent experiments unless otherwise specified.

Feature fusion hyperparameter analysis

To evaluate the impact of the fusion hyperparameter α on model performance, we conducted experiments by varying its value from 0.1 to 0.3 and reporting accuracy, precision, recall, and F1-score. The results are

Temp	Accuracy	Precision	Recall	F1-score
8	86.79	87.84	87.93	87.88
10	86.32	87.48	87.43	87.46
12	85.10	86.39	86.20	86.29
14	**88.48**	**89.42**	**89.41**	**89.41**
16	87.48	88.49	88.49	88.49
18	86.03	87.26	87.12	87.19
20	86.88	87.93	87.93	87.93
22	87.28	88.35	88.36	88.35
24	85.79	86.97	87.03	87.00

Table 4. Temperature sensitivity analysis for the auxiliary random classifier. All values are reported in percentage (%). Bold values denotes the best performance for this metric.



(a) Impact of α on accuracy (b) Impact of α on precision (c) Impact of α on recall (d) Impact of α on F1-score

Fig. 3. Performance variation with respect to the fusion hyperparameter α across accuracy, precision, recall, and F1-score.

Local branch architecture	Accuracy	Precision	Recall	F1-Score
baseline	84.03	86.98	85.51	86.24
ResNet18	86.74	87.81	87.91	87.27
ResNet34	87.44	88.49	88.43	87.96
Custom architecture version 1	**88.48**	**89.42**	**89.41**	**89.41**
Custom architecture version 2	86.96	88.18	87.93	87.57

Table 5. Comparison of different local branch architectures. Residual block configurations with varying channel sizes are compared in terms of classification performance. All values are reported in percentage (%). Bold values denotes the best performance for this metric.

summarized in Fig. 3. From the results, it can be observed that the fusion hyperparameter significantly influences model performance. When the hyperparameter is set to **0.2**, the model achieves the best overall performance across all metrics, with an accuracy of 88.48% and an F1-score of 89.41%. This demonstrates that 0.2 provides the most effective balance between the fused components. While the performance at 0.3 also appears competitive, the improvements over lower values are less consistent and slightly weaker compared to the gains observed at 0.2. Thus, 0.2 can be regarded as the optimal setting among the tested values.

Analysis of local branch design variants and integration position

Evaluation of local branch design variants

Table 5 presents a comparative evaluation of several local branch design variants, including standard backbones (ResNet18 and ResNet34) and two customized residual configurations with lightweight convolutional layers. The first custom architecture version is `nn.Conv2d(1, 3) + ResidualBlock(3, 1) + nn.Conv2d(1, 1)`, and the second version is `nn.Conv2d(1, 6) +ResidualBlock(6, 1) +nn.Conv2d(1, 1)`. All models are evaluated on the same underwater acoustic classification task using Accuracy, Precision, Recall, and F1-Score as metrics.

As expected, all hybrid CNN–Transformer variants outperform the baseline models. In addition, as expected, all hybrid CNN–Transformer variants outperform the baseline models. In addition, ResNet34 outperforms ResNet18, achieving an accuracy of 87.44% versus 86.74%, which can be attributed to the deeper architecture's increased capacity for modeling hierarchical features. However, the improvement is modest, suggesting that simply increasing depth does not guarantee substantial gains for underwater acoustic signals, where fine-grained spectral patterns are critical.

Interestingly, the first custom configuration, which employs `Conv2d(1, 3)` in the first layer followed by a lightweight residual block, achieves the best overall performance, surpassing both ResNet18 and ResNet34. This

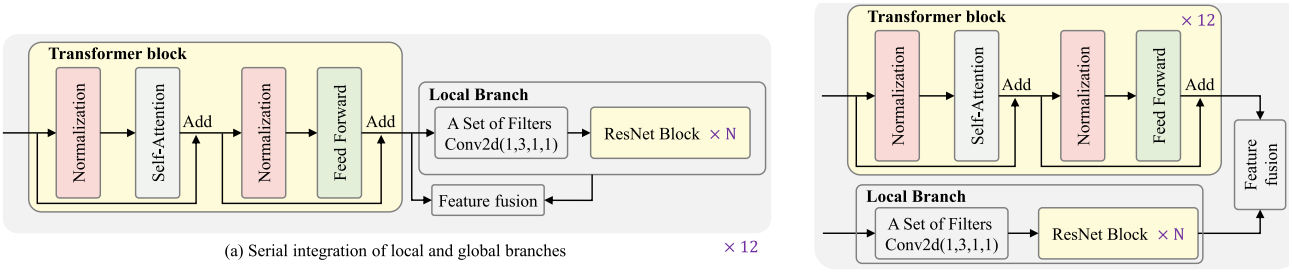


Fig. 4. Illustration of local–global branch integration strategies: (**a**) serial integration, where the local and global branches are fused sequentially; and (**b**) parallel integration, where both branches operate concurrently before fusion.

	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
①	86.14	87.28	87.14	87.21
②	**88.48**	**89.42**	**89.41**	**89.41**

Table 6. The impact of local branch integration position. Bold values denotes the best performance for this metric.

demonstrates that network depth is not the sole determinant of performance. Instead, the superior results stem from the carefully tailored convolutional design: the smaller kernel focuses on fine-grained local spectral cues, and the residual connections efficiently capture hierarchical features without over-parameterization. In contrast, larger kernels (e.g., `Conv2d(1, 6)`) or standard ResNet blocks may dilute these critical local features, limiting performance despite deeper or wider architectures.

These observations highlight that, for underwater acoustic target recognition, a shallow yet structurally optimized residual design can more effectively balance local feature extraction and overall feature hierarchy, outperforming deeper generic networks while maintaining low model complexity. This insight motivates the integration of such custom convolutional blocks into our hybrid local–global representation learning framework.

Evaluation of local branch integration position

To determine the optimal integration strategy for the local branch, we evaluate two alternatives, as illustrated in Fig. 4: ① embedding the local branch within each Transformer block, and ② incorporating it as a parallel Transformer encoder.

The results are summarized in Table 6. When the local branch is integrated inside every Transformer block (①), the model achieves an accuracy of 86.14% and an F1-score of 87.21%. While this design allows local convolutional cues to interact directly with global attention at each layer, it may also introduce redundancy and interfere with the Transformer's capacity to capture long-range dependencies, leading to suboptimal performance.

In contrast, positioning the local branch as an independent encoder (②) yields a substantial performance boost, achieving the highest accuracy (88.48%) and F1-score (89.41%). This configuration enables the branch to specialize in modeling fine-grained spectral structures, while the Transformer encoder focuses on global temporal–spectral dependencies. The complementary representations are then fused effectively at a higher level, producing more balanced and discriminative features.

Overall, the results highlight that integrating the local branch as a standalone encoder (②) is more effective than embedding it within each Transformer block, confirming that separating local and global modeling streams leads to superior feature representation and classification performance.

Noise robustness evaluation

To assess robustness under adverse acoustic conditions, we introduced additive white Gaussian noise (AWGN) at different signal-to-noise ratios (SNRs: -5, -1, 0, 1, and 5 dB). The resulting accuracy curves are shown in Fig. 5. Across all SNR levels, the proposed model consistently outperforms the baselines. Under the most challenging condition (-5 dB), it achieves 77.33% accuracy, clearly surpassing MIXUP (73.64%) and SSLMM (72.60%). At 0 dB, the advantage remains evident with 80.06% accuracy compared to 77.57% (MIXUP) and 75.33% (SSLMM). In moderate-to-clean scenarios (1–5 dB), our method continues to lead, reaching 83.39% at 5 dB. In the noise-free case, it attains the highest accuracy of 88.48%, outperforming MIXUP (86.33%) and SSLMM (80.22%). Conventional CNN-based baselines (SCAE, ResNet, Inception) exhibit pronounced performance degradation under noise. For instance, ResNet yields only 68.05% at -5 dB and remains below 70% even without noise, underscoring the difficulty of relying solely on convolutional filters in such conditions.

Beyond additive Gaussian noise, underwater acoustic signals are frequently corrupted by non-Gaussian and environment-specific interference, such as impulsive transient disturbances, biological ambient noise,

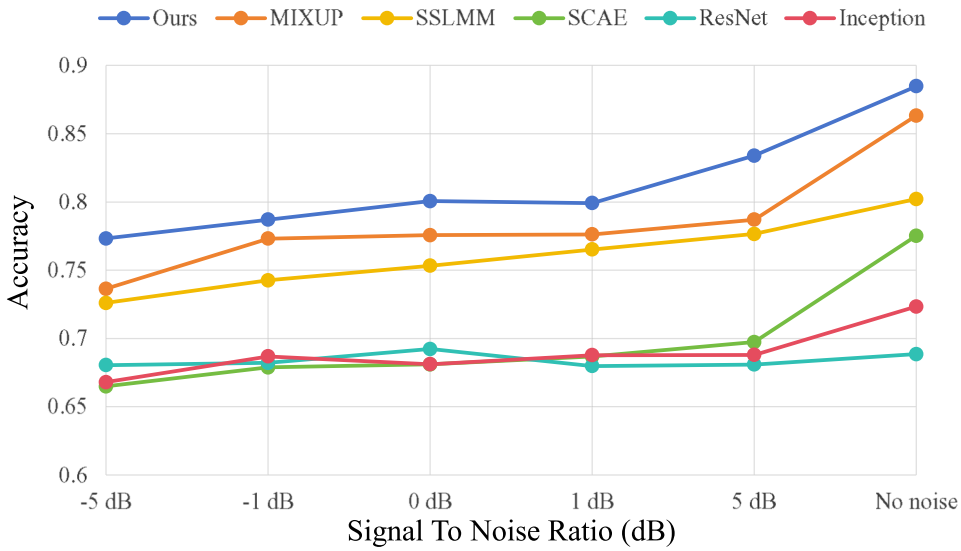


Fig. 5. Accuracy vs. SNR on Deepship dataset under additive white noise.

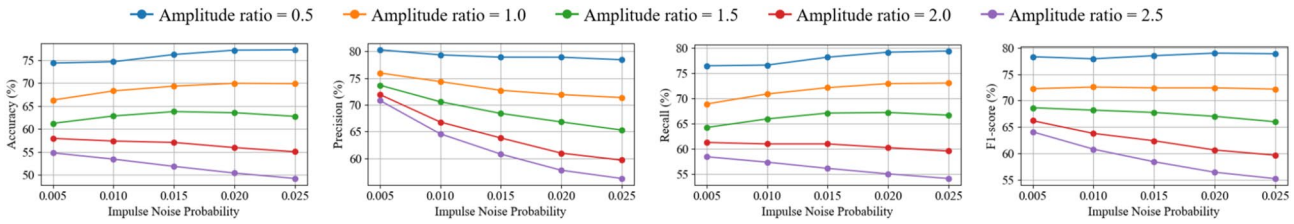


Fig. 6. Performance variations of Accuracy, Precision, Recall, and F1-score on the DeepShip dataset under impulse noise. The x-axis indicates the impulse noise probability, while each colored line represents a different impulse noise amplitude.

and Doppler-induced spectral distortions. To comprehensively evaluate robustness under realistic operating conditions, we further introduce three representative types of noise: impulse noise, biological noise, and Doppler frequency shifts.

Impulse noise, characterized by short-duration high-amplitude spikes, is common in underwater sensing due to mechanical impacts, communication glitches, and sensor instability. We vary both the occurrence probability (0.005–0.025) and relative amplitude (0.5–2.5) to assess the model's resilience under increasingly challenging transient disruptions. As shown in Fig. 6, across all amplitude groups, performance remains stable with respect to probability, reflecting the model's tolerance to sparse impulsive disturbances. When amplitude increases, a gradual yet controlled degradation is observed, indicating that the model's feature extraction pipeline maintains robustness even when strong outliers appear in the waveform. At low distortion levels (amplitude = 0.5), the model consistently achieves 74–77% accuracy and 78–79% F1-score, with the best performance obtained at probability = 0.020 (accuracy = 77.17%, F1 = 79.00%). At medium distortion (amplitude = 1.0–1.5), results remain reasonable; for example, amplitude = 1.0 yields accuracy 68.3–69.96% and F1-score around 72% across all probabilities. Even under severe impulsive interference (amplitude = 2.5), where signal waveforms are heavily distorted, the model maintains 49.22% accuracy and 55.19% F1-score at the worst case (probability = 0.025). This demonstrates that the proposed method retains a meaningful level of discrimination ability against strong, burst-type perturbations—conditions under which conventional CNN models typically collapse.

Biological noise is a naturally occurring and persistent interference in shallow-water environments, introduced by snapping shrimp, cetaceans, and various marine organisms. To simulate its masking effect, we mix biological noise at amplitude ratios ranging from 0.1 to 0.9. As shown in Fig. 7, across all tested conditions, the model exhibits highly stable performance, with mean accuracy of 80.61%, mean precision of 82.75%, mean recall of 82.00%, and mean F1-score of 82.37%, accompanied by small standard deviations (3.01–3.69%). These results confirm that biological ambient noise leads to only modest performance variations. Notably, even at the strongest biological noise ratio (0.9), the model still achieves 76.61% accuracy and 78.95% F1-score, indicating that the proposed representation is robust to natural underwater interference that typically overlaps spectrally with vessel signatures. These findings align with the expectation that biological noise, though broadband and fluctuating, does not severely disrupt the learned temporal–spectral structure of vessel acoustics.

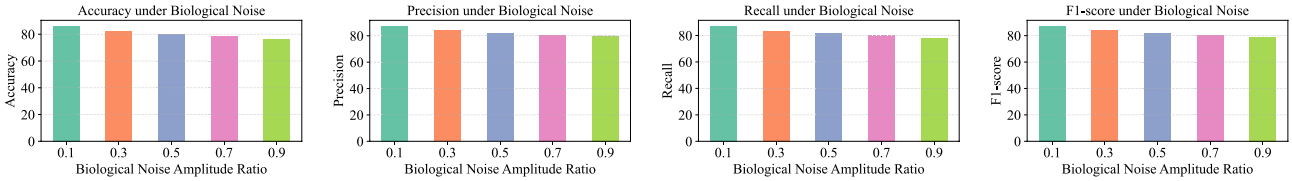


Fig. 7. Variations in Accuracy, Precision, Recall, and F1-score on the DeepShip dataset under biological noise. Each color represents a different biological noise amplitude ratio under the same experimental configuration.

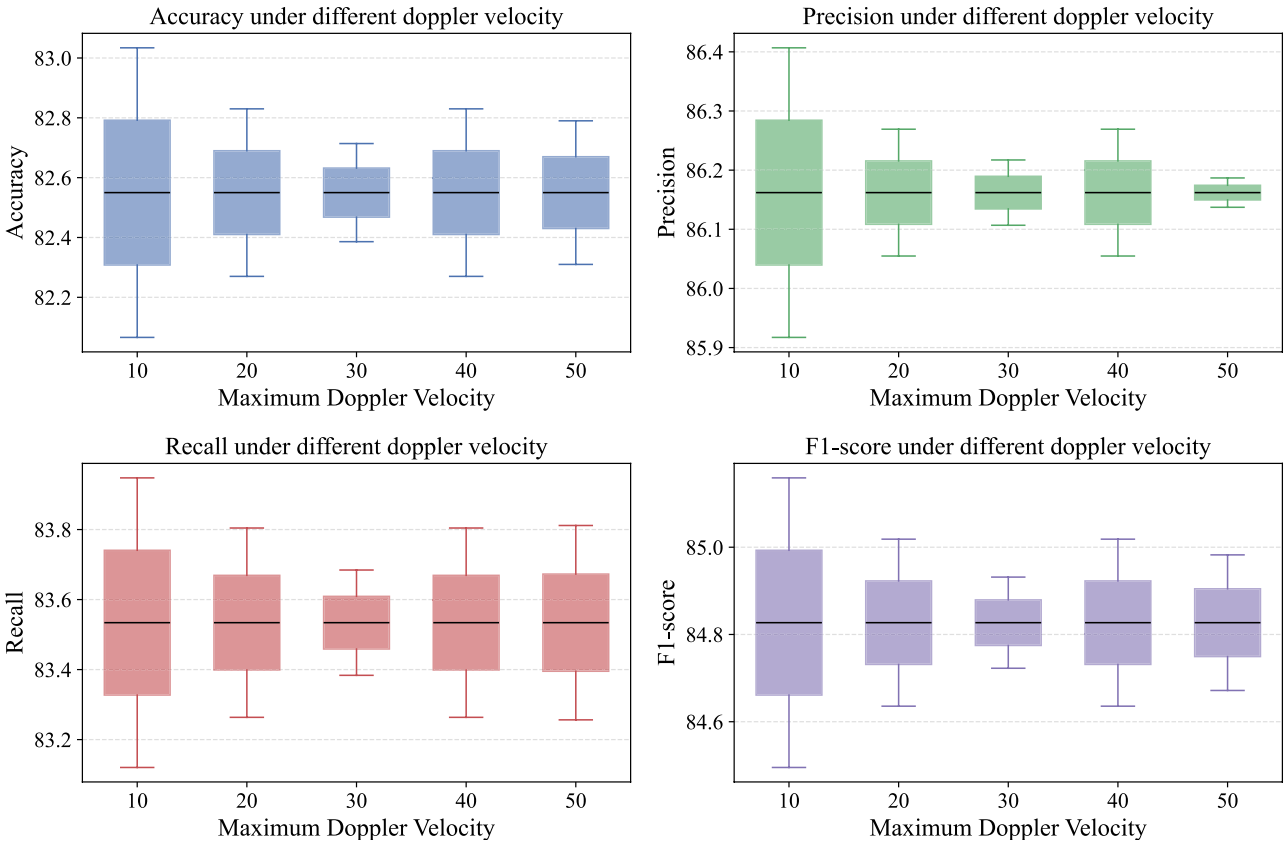


Fig. 8. Variations in Accuracy, Precision, Recall, and F1-score on the DeepShip dataset under Doppler-induced spectral distortion. Each boxplot visualizes the distribution of performance metrics obtained under different maximum target velocities ($v_{\max} = 10$–50), where Doppler shifting is simulated by time-scaling the waveform to emulate frequency compression or expansion caused by source-receiver relative motion. The boxes indicate the interquartile range (IQR), the central line denotes the median, and whiskers represent non-outlier ranges. Across all Doppler velocities, the metrics exhibit minimal spread and consistent central tendencies, demonstrating that the proposed model is highly robust to spectral deformation induced by Doppler effects.

Underwater acoustic signals are also subject to Doppler-induced spectral distortion when the source moves relative to the receiver, leading to frequency scaling and temporal warping that can substantially modify the observed waveform. To assess the robustness of the proposed method under such motion-induced effects, we simulate Doppler frequency shifts corresponding to maximum target velocities ranging from $v_{\max} = 10$ to 50. As shown in Fig. 8, across all tested velocities, the model exhibits remarkably stable performance, indicating that its learned representations remain resilient to elastic frequency variations. The accuracy consistently stays within a narrow range of approximately 81–83%, while the F1-score fluctuates only slightly around 84–85%, yielding very low standard deviations (0.71% for accuracy and 0.63% for F1-score). These results demonstrate that the model's spectral–temporal encoding effectively preserves discriminative structure even when frequency components are compressed or stretched due to Doppler effects. Overall, the minimal performance variation across increasing velocities confirms that the proposed method maintains strong robustness to realistic Doppler-induced distortions commonly encountered in dynamic maritime environments.

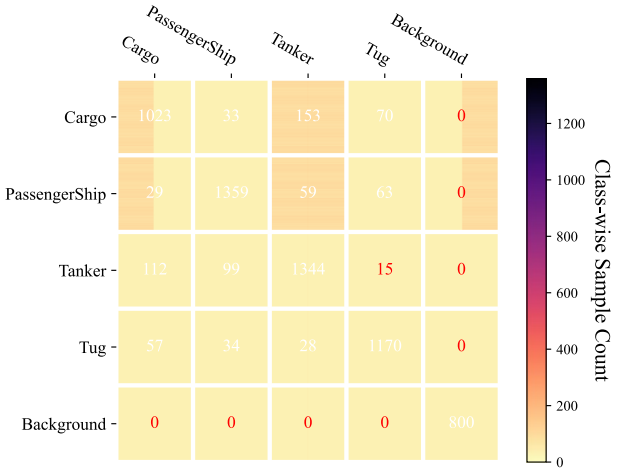


Fig. 9. Confusion matrix for proposed model on deepship dataset.

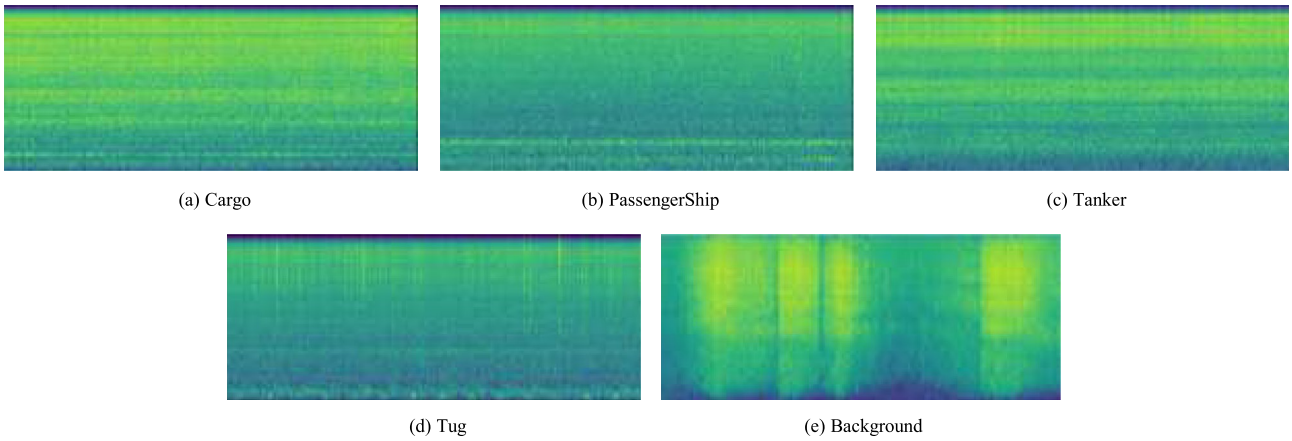


Fig. 10. Visualization of spectrograms across different classes. Cargo ships, tankers, and tugs exhibit similar acoustic patterns, making them prone to mis-classification, whereas passenger ships and background noise show more distinct characteristics, resulting in fewer errors.

These results show that the proposed hybrid CNN–Transformer equipped with stochastic classification maintains strong robustness under diverse acoustic degradations, including additive noise, impulsive disturbances, biological noise and Doppler-induced spectral distortion. Across all noise conditions and intensity levels, the model consistently exhibits higher stability. This makes it particularly well suited for real-world noisy underwater acoustic scenarios.

Class-wise performance and confusion analysis

To provide a detailed assessment of the model's discriminative capability across different vessel types, we analyzed the confusion matrix on the DeepShip dataset. Figure 9 shows that the proposed model achieves balanced performance across all vessel categories and maintains perfect separation of the Background class, demonstrating its ability to reliably distinguish ship-generated signals from ambient noise.

Most vessel classes are recognized with high accuracy, although misclassifications primarily occur among acoustically similar types, as illustrated in Fig. 10. Cargo ships are occasionally confused with Tankers or Tugs, reflecting overlapping low-frequency spectral patterns. Similarly, Tankers exhibit confusion with both Cargo and PassengerShips, consistent with partially shared broadband noise characteristics. Misclassifications of Tugs are less frequent but distributed across other ship classes, likely due to variable operational conditions.

In contrast, PassengerShips are identified with particularly high reliability, benefiting from distinct tonal structures and stable harmonic features. The stochastic classification head further enhances robustness: by averaging predictions across perturbed decision boundaries, it suppresses spurious misassignments and produces more consistent outputs, especially among closely related vessel types. The confusion matrix demonstrates that the hybrid CNN-Transformer with stochastic classification provides strong per-class discrimination and effectively handles inter-class similarities, resulting in consistently reliable performance across all categories.

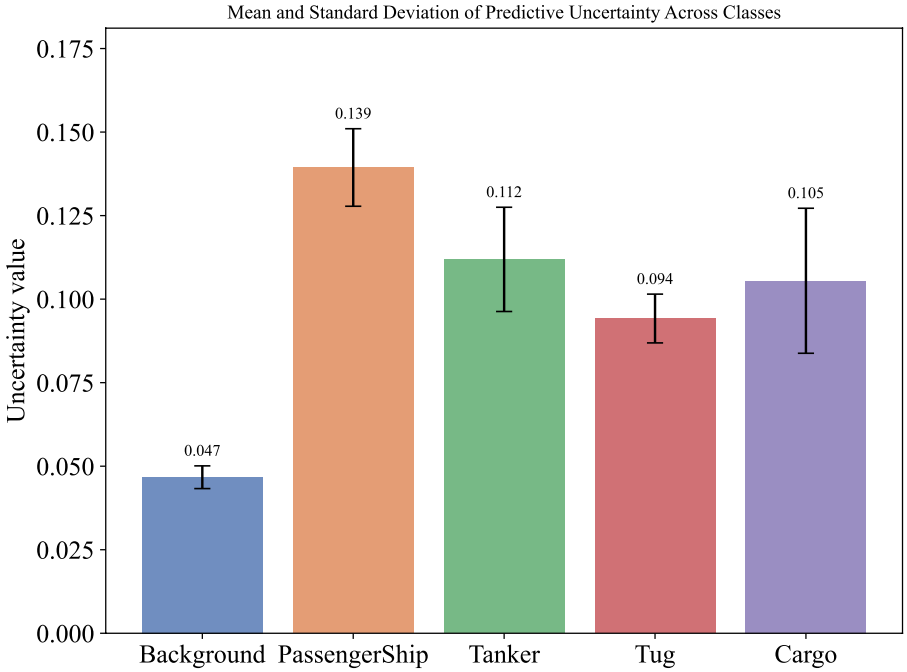


Fig. 11. Uncertainty analysis using five stochastic forward passes. Samples with low predictive variance correspond to confident and correct decisions, while high-variance predictions align with misclassifications or ambiguous cases, offering a reliable cue for uncertainty-aware decision-making.

Model	params(MB)	Gflops	Accuracy
DNN	67.24	**1.08**	73.11
Inception	48.46	260.67	76.16
Residual	**23.51**	169.40	76.98
AudioMAE	85.26	349.68	76.66
SSAST	85.26	827.54	77.70
SSLMM	86.68	633.62	80.22
MIXUP	86.68	1900.86	86.23
Ours	87.74	833.31	**88.48**

Table 7. Model efficiency comparison on Deepship, using single RTX 3090. Bold values denotes the best performance for this metric.

To further understand how uncertainty quantification supports practical decision-making, we analyze the predictive distributions produced by the stochastic classifier across multiple sampled classifier instances. We randomly selected a batch of samples and performed multiple stochastic forward passes (5 in our experiments) to compute prediction uncertainty. As shown in Fig. 11, the results show a clear correspondence between uncertainty levels and decision reliability: samples with concentrated predictive distributions (low variance) consistently achieve high classification accuracy, indicating that the model can confidently commit to a decision. In contrast, high-uncertainty cases strongly correlate with misclassifications or class ambiguities, providing an effective signal for triggering alternative actions such as requesting additional observations, switching to higher-resolution sensing, or deferring classification.

Parameter efficiency and inference speed

Despite its hybrid design, the proposed CNN-Transformer model remains computationally efficient. Table 7 summarizes the parameter size, computational cost (Gflops), and classification accuracy of major architectures on the DeepShip dataset, evaluated on a single NVIDIA RTX 3090 GPU. The integration of local convolutions within the Transformer introduces only a modest 12% parameter overhead compared to a pure Transformer while delivering substantial improvements in accuracy and stability. This demonstrates that the architecture achieves a favorable trade-off between performance and efficiency, making it suitable for real-time or resource-constrained underwater systems.

Among lightweight models, the standard DNN achieves the lowest computational cost (1.08 GFLOPS) with a modest parameter size of 67.24 MB, yet its accuracy is limited to 73.11%, reflecting its constrained capacity

for modeling complex temporal–spectral patterns. The Residual network provides a favorable balance, with only 23.51 MB of parameters and 169.40 GFLOPS, achieving 76.98% accuracy. Inception, while slightly larger, reaches 76.16% accuracy at a substantially higher computational cost (260.67 Gflops).

Self-supervised Transformer-based models (SSAST, AudioMAE, SSLMM, MIXUP) demand considerably more computation, ranging from 349.68 to 1900.86 Gflops, with parameter sizes around 85–87 MB. These models achieve higher accuracy than lightweight architectures, with SSLMM and MIXUP attaining 80.22% and 86.23%, respectively. MIXUP incurs the highest computational cost due to extensive data augmentation and transformer operations.

Our hybrid CNN-Transformer surpasses all baselines, achieving 88.48% accuracy with moderate parameter size (87.74 MB) and manageable computational cost (833.31 GFLOPS). This performance reflects the complementary strengths of local convolutions for fine-grained spectral feature extraction and global attention for long-range contextual modeling. Overall, the results indicate that the proposed architecture provides an effective balance between accuracy, parameter efficiency, and inference speed, making it highly suitable for practical underwater acoustic applications.

Conclusion

In this work, we proposed a novel hybrid deep learning framework for underwater acoustic target recognition that effectively integrates local and global feature modeling capabilities. By integrate convolutional module with Transformer encoder, the model captures both fine-grained spectral patterns and long-range temporal dependencies, addressing the limitations of CNN-only and Transformer-only architectures. Additionally, we introduced a stochastic classifier ensemble to enhance the model's robustness, particularly under low-SNR and ambiguous signal conditions frequently encountered in real-world underwater environments. Extensive experiments on two benchmark datasets DeepShip and ShipsEar, demonstrated the superior performance of the proposed approach. The hybrid CNN-inserted Transformer significantly outperformed conventional baselines in terms of accuracy, F1-score, and robustness to acoustic noise. Ablation studies confirmed the individual effectiveness of both the CNN-insertion mechanism and the stochastic classification module. Furthermore, evaluations under multiple noise conditions and cross-region testing demonstrate that the model maintains strong generalization across diverse acoustic propagation environments. This study provides new insights into architectural design for underwater acoustic analysis, highlighting the importance of multi-scale representation learning and uncertainty-aware classification in hostile acoustic environments.

Future work. Building upon the proposed hybrid local-global representation framework and Gaussian sampling-based stochastic classifier, future research will focus on enhancing model generalization and practical applicability. Specifically, we plan to explore self-supervised pretraining strategies tailored to underwater acoustic signals to further reduce reliance on labeled data, and to investigate extensions of the dual-branch architecture for *multi-modal fusion*, incorporating complementary visual, sonar, or bathymetric cues to strengthen feature representations. Furthermore, we aim to optimize the framework for real-time inference and efficient deployment on embedded underwater platforms, ensuring that the method maintains robustness and predictive reliability under operational constraints.

Data availability

The datasets generated and analysed during the current study are available in the https://github.com/irfankamboh/DeepShip repository and the fifth class Background noise is in the https://github.com/ZhuPengsen/Method-for-Splitting-the-DeepShip-Dataset repository.

Received: 27 October 2025; Accepted: 9 December 2025
Published online: 16 December 2025

References

1. Feng, S., Ma, S., Zhu, X. & Yan, M. Artificial intelligence-based underwater acoustic target recognition: a survey. *Remote Sens.* **16**, 3333 (2024).
2. Luo, X., Chen, L., Zhou, H. & Cao, H. A survey of underwater acoustic target recognition methods based on machine learning. *J. Marine Sci. Eng.* **11**, 384 (2023).
3. Etter, P. C. Advanced applications for underwater acoustic modeling. *Adv. Acoustics Vibrat.* **2012**, 214839 (2012).
4. Ashok, P. & Latha, B. Develop an improved mel-frequency cepstral coefficients signal processing algorithms for enhancing underwater acoustic signal through wireless network. *Measurement* **243**, 116414 (2025).
5. Galusha, A., Dale, J., Keller, J. & Zare, A. Deep convolutional neural network target classification for underwater synthetic aperture sonar imagery. In Isaacs, J. & Bishop, S. (eds.) *Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXIV*. SPIE (SPIE, Baltimore, MD, USA, 2019).
6. Ke, X., Yuan, F. & Cheng, E. Underwater acoustic target recognition based on supervised feature-separation algorithm. *Sensors* **18**, 4318 (2018).
7. Mu, W., Yin, B., Huang, X., Xu, J. & Du, Z. Environmental sound classification using temporal-frequency attention based convolutional neural network. *Sci. Rep.* **11**, 21552 (2021).
8. Vaswani, A. et al. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008 (2017).
9. Song, Y., Liu, F. & Shen, T. Method of underwater acoustic signal denoising based on dual-path transformer network. *IEEE Access* **12**, 81483–81494 (2022).
10. Gong, Y., Chung, Y.-A. & Glass, J. Ast: Audio spectrogram transformer. In *Proceedings of Interspeech 2021*, 571–575 (ISCA, 2021).
11. Feng, S. & Zhu, X. A transformer-based deep learning network for underwater acoustic target recognition. *IEEE Geosci. Remote Sens. Lett.* **19**, 1–5 (2022).
12. Xu, K. et al. Self-supervised learning-based underwater acoustical signal classification via mask modeling. *J. Acoustic. Soc. Am.* **154**, 5–15 (2023).
13. Tang, J. et al. Uapt: an underwater acoustic target recognition method based on pre-trained transformer. *Multimedia Syst.* **31**, 50 (2025).

14. Zhu, Q. et al. Ssast-adapter: A parameter-efficient incremental learning algorithm for underwater acoustic target recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5 (IEEE, 2025).
15. Fu, B., Nie, J., Wei, W. & Zhang, L. Constructing a multi-modal based underwater acoustic target recognition method with a pre-trained language-audio model. *IEEE Trans. Geosci. Remote Sens.* **63**, 1–14 (2025).
16. Feng, S., Zhu, X. & Ma, S. Masking hierarchical tokens for underwater acoustic target recognition with self-supervised learning. *IEEE/ACM Trans. Audio, Speech Language Process.* **32**, 1365–1379 (2024).
17. Davis, S. B. & Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.* **28**, 357–366 (1980).
18. David, P. M. & Chapron, B. Underwater acoustic signal analysis with wavelet process. *J. Acoust. Soc. Am.* **87**, 2118–2121 (1990).
19. Jiang, J., Shi, T., Huang, M. & Xiao, Z. Multi-scale spectral feature extraction for underwater acoustic target recognition. *Measurement* **166**, 108227 (2020).
20. Yang, H. et al. Underwater acoustic target recognition using svm ensemble via weighted sample and feature selection. In *2016 13th International Bhurban conference on applied sciences and technology (IBCAST)*, 522–527 (IEEE, 2016).
21. Wang, X., Liu, A., Zhang, Y. & Xue, F. Underwater acoustic target recognition: a combination of multi-dimensional fusion features and modified deep neural network. *Remote Sens.* **11**, 1888 (2019).
22. Shamshad, N. et al. Advanced KNN-based cost-efficient algorithm for precision localization and energy optimization in dynamic underwater sensor networks. *Sci. Rep.* **15**, 2182 (2025).
23. Fang, T., Wang, Q., Zhang, L. & Liu, S. Modulation mode recognition method of non-cooperative underwater acoustic communication signal based on spectral peak feature extraction and random forest. *Remote Sens.* **14**, 1603 (2022).
24. Stanković, I., Ioana, C., Daković, M. & Stanković, L. Analysis of off-grid effects in wideband sonar images using compressive sensing. In *OCEANS 2018 MTS/IEEE Charleston*, 1–6 (IEEE, 2018).
25. Xu, K. et al. General audio tagging with ensembling convolutional neural networks and statistical features. *J. Acoustical Soc. Am.* **145**, EL521–EL527 (2019).
26. Zhu, B., Xu, K., Kong, Q., Wang, H. & Peng, Y. Audio tagging by cross filtering noisy labels. *IEEE/ACM Trans. Audio Speech Language Process.* **28**, 2073–2083 (2020).
27. Hummel, H. I., van der Mei, R. & Bhulai, S. A survey on machine learning in ship radiated noise. *Ocean Eng.* **298**, 117252 (2024).
28. Wang, Y. et al. Underwater communication signal recognition using sequence convolutional network. *IEEE Access* **9**, 46886–46899 (2021).
29. Zhang, Y. et al. Deep learning-based signal detection for underwater acoustic OTFS communication. *J. Marine Sci. Eng.* **10**, 1920 (2022).
30. Sun, Q. & Wang, K. Underwater single-channel acoustic signal multitarget recognition using convolutional neural networks. *J. Acoustical Soc. Am.* **151**, 2245–2254 (2022).
31. Doan, V.-S., Huynh-The, T. & Kim, D.-S. Underwater acoustic target classification based on dense convolutional neural network. *IEEE Geosci. Remote Sens. Lett.* **19**, 1–5 (2020).
32. Li, C., Huang, Z., Xu, J. & Yan, Y. Underwater target classification using deep learning. In *Proceedings of OCEANS 2018 MTS/IEEE Charleston* (Charleston, SC, USA, 2018).
33. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
34. Huang, W. et al. A multi-task learning balanced attention convolutional neural network model for few-shot underwater acoustic target recognition. arXiv preprint arXiv:2504.13102 (2025).
35. Xiao, X., Wang, W., Ren, Q., Gerstoft, P. & Ma, L. Underwater acoustic target recognition using attention-based deep neural network. *JASA Express Lett.* **1**, 105001 (2021).
36. Fang, W., Zhongda, Z., Xiaobo, Z., Yuzhang, Z. & Haiyan, W. A momentum-based adversarial training approach for generalization in underwater acoustic target recognition: An individual-vessel perspective. *J. Acoustical Soc. Am.* **157**, 3508–3523 (2025).
37. Li, P., Wu, J., Wang, Y., Lan, Q. & Xiao, W. Stm: Spectrogram transformer model for underwater acoustic target recognition. *J. Marin. Sci. Eng.* **10**, 1428 (2022).
38. Fan, W. et al. Estmst-st: An end-to-end soft threshold and multi-loss self-distillation based swin-transformer for underwater acoustic signal recognition. *IEEE Transactions on Geoscience and Remote Sensing* (2024).
39. Feng, Y., Li, H. & Sun, X. Self-supervised masked hierarchical transformer for few-shot underwater acoustic target recognition. *IEEE/ACM Trans. Audio Speech Language Process.* **32**, 284–293 (2024).
40. Khan, A. et al. A survey of the vision transformers and their CNN-transformer based variants. *Artif. Intell. Rev.* **56**, 2917–2970 (2023).
41. Li, C. et al. Bossnas: Exploring hybrid cnn-transformers with block-wisely self-supervised neural architecture search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12281–12291 (2021).
42. Rafiepour, M. & Sartakhti, J. S. Ctran: Cnn-transformer-based network for natural language understanding. *Eng. Appl. Artif. Intell.* **126**, 107013 (2023).
43. Kalyan, K. S., Rajasekharan, A. & Sangeetha, S. Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint* arXiv:2108.05542 (2021).
44. Zeghidour, N., Teboul, O., de Chaumont Quitry, F. & Tagliasacchi, M. Leaf: A learnable frontend for audio classification. In *International Conference on Learning Representations*.
45. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision, 1026–1034 (2015).
46. Irfan, M. et al. Deepship: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification. *Expert Syst. Appl.* **183**, 115270 (2021).
47. Xu, Q. et al. Self-supervised learning-for underwater acoustic signal classification with mixup. *IEEE J. Select. Top. Appl. Earth Observat. Remote Sens.* **17**, 3530–3542 (2023).
48. Santos-Domínguez, D., Torres-Guijarro, S., Cardenal-López, A. & Pena-Gimenez, A. Shipsear: An underwater vessel noise database. *Appl. Acoust.* **113**, 64–69 (2016).
49. Zhu, P. et al. Underwater acoustic target recognition based on spectrum component analysis of ship radiated noise. *Appl. Acoustics* **211**, 109552 (2023).
50. Huang, P.-Y. et al. Masked autoencoders that listen. *Adv. Neural. Inf. Process. Syst.* **35**, 28708–28720 (2022).

Author contributions

Cheng Yang and Qisheng Xu: Model design and implementation. Cheng Yang, Qisheng Xu and Kele Xu: Model design and result analysis. Ming Feng, Hui Yang, Yulei Yuan, Yutao Dou: Model design and implementation. Qisheng Xu and Junyi Zhao: Result analysis. All authors contributed to the manuscript.

Funding

This work is supported by National Science and Technology Major Project (2023ZD0121101), National University of Defense Technology (ZZCX-ZZGC-01-04) and Major Fundamental Research Project of Hunan Province

(2025JC0005).

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to K.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.