

Self-supervised Learning-for Underwater Acoustic Signal Classification with Mixup

Qisheng Xu, Jingfei Jiang, Kele Xu, *Member, IEEE*, Yong Dou, Caili Gao, Boqing Zhu, Kang You, Qian Zhu

Abstract—Underwater acoustic signal classification is a critical task that involves identifying different types of signals in a complex and dynamic underwater environment, which is often contaminated by strong ambient noise. Recent studies have demonstrated that deep learning-based methods can achieve remarkable performance in this task by leveraging large-scale labeled data. However, obtaining labeled data in real-world scenarios can be challenging due to the labor-intensive and expert-dependent nature of the labeling process, especially for underwater scenarios. In this study, we propose a novel self-supervised learning framework combined with mixup-based augmentation that can learn discriminative representations from large-scale unlabeled data, thereby reducing the dependence on labeled data. Additionally, we propose a Test Time Augmentation module to further improve the model's robustness. Our proposed approach achieves a classification accuracy of 86.33% on the DeepShip dataset, surpassing previous competitive methods by a significant margin. Notably, our method demonstrates excellent generalization performance in few-shot scenarios and low signal-to-noise settings, highlighting its potential for practical applications.

Index Terms—Underwater Acoustic Signal Classification, Self-Supervised Learning, Data Augmentation, Mixup, Test Time Augmentation.

I. INTRODUCTION

THE classification of underwater acoustic signals plays a critical role in extracting meaningful information from underwater environments, with wide applications in the remote sensing field such as underwater target detection [1] and marine environment monitoring [2]. However, the complex and dynamic underwater environment introduces various interference, such as ocean ambient noise, multi-path effects, and acoustic propagation distortion, posing significant challenges for accurate classification.

To address these challenges, researchers have made numerous attempts [3], [4]. Traditional underwater acoustic signals classification primarily utilizes handcrafted features combined with shallow architecture-based classifier [5]–[7]. These handcrafted features include waveform features, wavelet features and spectrogram features [8]. Furthermore, advanced feature representation methods have been proposed to express more information about original signals, including Low-Frequency Analysis and Recording (LOFAR), Detection of Envelope

Modulation On Noise (DEMON), Gammatone frequency cepstrum coefficient (GFCC) [9], etc. Although these features have yielded promising results in many applications, the selection of representations relies heavily on specialized domain knowledge and expert experience. Besides, Shallow classifiers, like the support vector machine (SVM) and shallow neural network classifier, exhibit weak fitting and generalization capabilities due to their constrained capacity imposed by their shallow architectures [10].

In recent years, deep learning method has shown impressive performance in underwater acoustic signal classification tasks [11]–[13]. Typically, these methods utilize conventional supervised learning paradigm, which faces a significant challenge—the demand for substantial labeled data to train models. Furthermore, obtaining an adequate amount of labeled data presents a significant obstacle, as the labeling process proves to be labor-intensive and reliant on expert expertise. Fortunately, in many practical scenarios, large-scale unlabeled datasets are readily available, enabling the use of self-supervised learning (SSL) without the need for manual annotations becomes a promising alternative method. SSL usually leverages the inherent structure or patterns within the data itself to create supervisory signals, thereby encouraging the model learning and updating. Generally speaking, a SSL framework comprises two key stages: pre-text training (pre-training) and downstream task fine-tuning. The core objective of SSL is to acquire meaningful and useful information from an unlabeled dataset by formulating an effective pre-text task. By training on these self-generated tasks, the model learns to capture essential features and relationships present in the data, which can then be transferred to downstream tasks with limited labeled data. Recently, the Transformer-based SSL methods have emerged as a powerful technique that achieves state-of-the-art results in various fields by guiding model learning through the construction of labels derived from the data itself [14]. For general audio classification, several attempts have been made, such as SSAST [15], AudioMAE [16], and BEATs [17], which leverage pure self-attention methods for audio classification and yield remarkable performance. However, the related work of SSL techniques remains limited in the context of underwater acoustic signal classification. The underwater acoustic signal is more intricate due to the multi-path effect inherent in marine environment, making labeled data even scarcer. Consequently, the exploration of SSL methods in underwater acoustic classification becomes an urgent concern, which is the main focus of this paper.

In this paper, we propose a novel approach for underwater acoustic signal classification by utilizing SSL framework.

Manuscript received xx xx, xxxx; revised xx xx, xxxx (Corresponding author: Yong Dou).

Qisheng Xu, Jingfei Jiang, Kele Xu, Yong Dou, Caili Gao, Boqing Zhu, Qian Zhu are with School of Computer, National University of Defense Technology, Changsha, 410073, China (e-mail: yongdou@nudt.edu.cn; qishengxu@nudt.edu.cn).

Kang You is with Tongji University, Shanghai, China.

The core challenge in this attempt lies in effectively representing the acoustic signals, which has a significant impact on the performance of the classification task. To overcome this challenge, we present a two stage methodology. Initially, we adopt a hierarchical multi-scale mask modeling approach as pre-text task, which can capture both local and global information so that obtaining a excellent initial weights for the downstream classification task. Given the unpredictable and ever-changing of the marine environment, underwater acoustic signals often grapple with substantial interference. To attain a robust and generalizable representation, we introduce a mixup-based strategy during the pre-training stage. For the fine-tuning phase, we fine-tune the pre-trained model using a small amount of labeled underwater acoustic data. Furthermore, we integrate a Test Time Augmentation (TTA) strategy, generating an ensemble of predictions to further enhance model performance. Our approach demonstrates the efficacy of SSL paradigm in addressing challenges presented by intricate and ever-changing marine environments, significantly reducing the demand for labeled data when compared to existing methods. In summary, the primary contributions of this paper include the following key aspects:

- 1) A novel self-supervised learning framework is proposed for underwater acoustic signal classification. In this framework, we implement a hierarchical multi-scale mask modeling strategy during the pre-training phase, which enables the model to concurrently capture both local and global information from a large amount of unlabeled acoustic signals (AudioSet dataset [18]). By this way, we could obtain a robust and generic representations for downstream tasks, thereby significantly decreasing the requirement for labeled data.
- 2) Inspired by the multi-path effect inherent in marine environments, we develop a mixup-based strategy to enhance data diversity, thereby enabling the model to learn more comprehensive and discriminative representations. In the subsequent fine-tuning phase, we devise a Test Time Augmentation strategy to further enhance the model's robustness, which is inspired by the observation that different representations of the same sample exhibit intrinsic consistency. Finally, our comprehensive experiments demonstrate the effectiveness of our approach for classifying underwater acoustic signals.

The remainder of this article is organized as follows: Section II reviews the related work about underwater acoustic signal classification and self-supervised representation learning. Section III provides a detailed introduction to our proposed method, including both the pre-training and fine-tuning phases that constitute our approach. In Section IV, we conduct comprehensive experiments to evaluate our method's performance and analyze its key components. Finally, Section V presents the main conclusions drawn from our research and outlines potential directions for future work.

II. RELATED WORK

A. Underwater acoustic signal classification

The classification of underwater acoustic signals, due to its noteworthy economic and military value, has garnered escalating interest in recent years. In summary, prior studies can be broadly categorized into two primary approaches: the utilization of handcrafted features in conjunction with shallow classifiers, and the application of deep learning-based methods.

Wang *et al.* [19] integrated Bark-wavelet analysis and Hilbert–Huang transform to extract acoustic features and employed Support Vector Machine (SVM) as the classifier. Shi *et al.* [20] analyzed the features of underwater acoustic signals from time domain, frequency domain and cyclostationarity domain, combined these features and then adopt Decision Tree (DT) for classification. Li *et al.* [21] combined MFCC with Hidden Markov Model (HMM) for underwater target classification. Another work by Li *et al.* [6] proposed a fusion frequency feature extraction method based on variational mode decomposition (VMD), duffing chaotic oscillator (DCO) and a form of permutation entropy (PE).

Owing to the intricate and ever-changing marine environment, coupled with diverse interfering factors, the performance of shallow architecture-based methods falls short of meeting the demand for high accuracy. Cao *et al.* [22] utilized convolutional neural network (CNN) for classification with a second-order pooling to capture the temporal correlations from the time-frequency (T-F) representation of the radiated acoustic signals. Wang *et al.* [23] adopt a hybrid series neural network to alleviate the effects losses for underwater acoustic signal modulation classification. However, it is noteworthy that these methods primarily rely on single representations, leading to performance limitations in complex scenarios due to the restricted diversity of information [24], [25]. Hong *et al.* [26] aggregated multiple acoustic representation and used SpecAugment technique to improve representation discrimination. Tian *et al.* [27] combined wave and T-F representation and proposed a joint model, thereby enhancing the discrimination of representations. Li *et al.* [28] proposed a transfer learning-based data augmentation method for underwater acoustic signal classification. Despite these methods have obtained promising performance, they can be time-consuming and labor-intensive to manually design or select optimal features. Actually, this challenge is pervasive across various domains. For instance, in hyper-spectral anomaly detection, Li *et al.* [29] replaced conventional manual parameter selection with trainable modules, resulting in significant performance improvements. Similarly, in acoustic classification fields, Yang *et al.* [30] have sought to simulate the human auditory perception system by combining it with a CNN for underwater acoustic target classification. Xie *et al.* [31] modified fixed wavelet parameters into learn-able parameters and utilized a CNN combined with attention mechanism for classification. Ren *et al.* [32] proposed a learn-able front-end based on the cognitive process of the human for intelligent underwater acoustic classification. These methods are good at capturing local patterns in the data and can extract high-level features by DNN learning from the training data. However, they

are poor at capturing long range dependencies information between time and frequency segments. Currently, transformer purely based on self-attention shows its great performance for capturing both local and global information in various tasks. Feng *et al.* [33] utilized log Mel-spectrogram and transformer for underwater target classification. Transformer-based self-supervised learning methods have shown great potential for underwater acoustic signal classification tasks, but have been under-explored in previous studies.

B. Self-supervised representation learning

Self-supervised representation learning (SSRL) is a methodology designed to yield robust, generic, and abstract representations. This approach capitalizes on inherent data characteristics to create supervised signals through pre-text tasks, eliminating the necessity for substantial labeled datasets and effectively addressing the limitations posed by annotation bottlenecks. Pre-text tasks commonly encompass context prediction [34], masked modeling [35], clustering-based methods [36], contrastive self-supervised learning [37], and information maximization [38]. Transformer, a network purely relying on self-attention mechanisms, has garnered significant attention in the researchers due to its remarkable performance. For example, Yao *et al.* [39] proposed an Extended Vision Transformer (ExViT) for Land Use and Land Cover Classification and achieved great progress. In computer vision domains, an illustrative contemporary example is the application of Masked Autoencoders (MAE) [40], which proficiently reconstructs partially masked sections of raw data, exhibiting excellent performance in downstream tasks. In the fields of underwater acoustic signal classification, CNNs are commonly adopted as encoders for automated extraction of acoustic features [41]. Notably adept at discerning local attributes like textures and contours, CNNs may nonetheless exhibit limitations in capturing substantial long-term dependencies pivotal to audio classification. To address this, the incorporation of transformer-based Deep Neural Networks (DNNs) has emerged as a pivotal factor in augmenting classification efficacy [33]. Gong *et al.* [15] introduced a novel approach named Self-Supervised Audio Spectrogram Transformer (SSAST), which integrates joint discriminative and generative masked spectrogram patch modeling to advance generic audio classification. The pioneering utilization of mask modeling SSL in the context of underwater acoustic classification was introduced by Xu *et al.* [42]. Their innovative strategy, involving the reconstruction of multiple representations, yielded substantial performance enhancements.

III. METHODOLOGY

A. Overall Architecture

Fig. 1 illustrates the proposed SSL-based framework for underwater acoustical signal classification, which comprises two primary phases: pre-training and fine-tuning. During the pre-training phase, we employ a mixup technique to augment the original audio signal and then convert it into a log Mel-spectrogram, followed by randomly masking a segment of the spectrogram. We then feed the unmasked patches into

the Swin-Transformer encoder, which encourages the neural network to predict the masked-out regions. For the fine-tuning stage, we integrate a TTA module by combining the pre-trained encoder with a linear classification head, aiming to enhance the subsequent task of underwater acoustic signal classification.

B. Pre-training phase

The primary goal of the pre-training phase is to establish highly effective initial weights that facilitate the extraction of discriminative representations for subsequent tasks. During this phase, we formulate the pre-text task as a masking and reconstruction procedure, achieved through a randomized masking strategy applied to the log Mel-spectrogram. The log Mel-spectrogram, constituted by half-overlapping triangular filters centered at frequency f_{mel} , can be mathematically defined as follows:

$$f_{mel} = 2595 * \log_{10}(1 + \frac{f}{700}) \quad (1)$$

where f (Hz) signifies the frequency of the signal.

Observing this pattern, it becomes evident that as the frequency f rises, the filter's bandwidth progressively expands, rendering it more adept at capturing low-frequency attributes within underwater acoustic signals. Furthermore, the log Mel-spectrogram encompasses a broader range of domain representations across time and frequency domains, enabling them to comprehensively capture the intrinsic structure of acoustic signals compared to the raw waveform signal. Guided by these insights, we opt for the log Mel-spectrogram as the chosen input representation. As illustrated in Fig. 1, the unmasked portions of the patches are subsequently subjected to the encoder for feature extraction. Then the output is fed into the decoder to reconstruct the masked parts.

1) *Network Structure*: As shown in Fig. 1, the pre-training phase of our proposed SSL-based underwater acoustic signal classification framework follows a standard encoder-decoder architecture. Diverging from the precedent CNN-based studies, which often concentrate on grasping localized patterns, transformer-based approaches have demonstrated remarkable prowess in encompassing both local and global information across diverse domains. Bearing this consideration in mind, we choose Swin-Transformer [43] as the foundational architecture for our encoder. As shown in Fig. 1, a Swin-Transformer block encompasses a normalization layer, a multi-head attention layer for capturing local information, and a shift window multi-head attention layer for capturing global information. At the end of the block, a multi-layer perception is employed to reduce the token dimension back to the input dimension. Our selection is anchored in the innate capability of Swin-Transformer to concurrently capture both local intricacies and broader interdependencies. This is accomplished through its integration of pooling and sliding window mechanisms. Furthermore, Swin-Transformer achieves the extraction of multi-scale insights by incorporating diverse window sizes, thereby enhancing the model's ability to derive more nuanced representations for underwater acoustic signals. This heightened

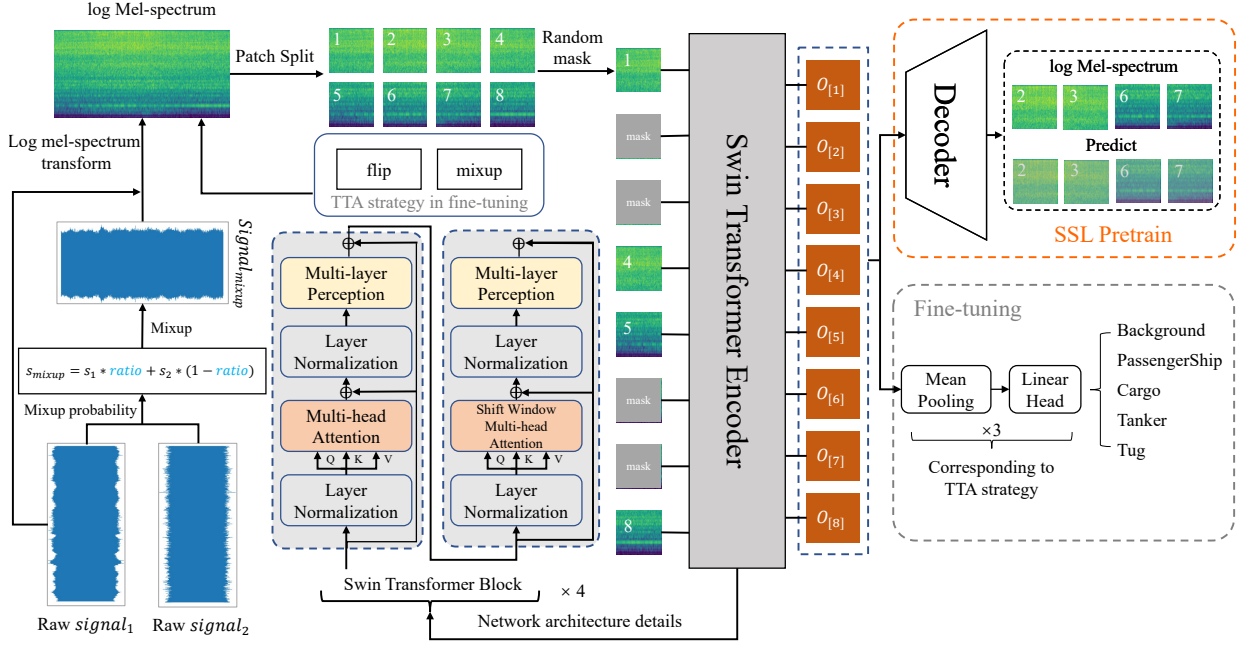


Fig. 1. The framework of the proposed approach combined with mixup and TTA strategies. The input signal is performed mixup augmentation with a certain probability and then converts to 2D log Mel-spectrogram, where the mixup ratio is generate by β -distribution. The log Mel-spectrogram is split into a sequence of 32×32 patches without overlap, and then random mask part of them. The unmasked patches are fed into Swin-Transformer encoder after adding a learn-able positional embedding. The output of encoder is used to 1) reconstruct masked patches in pre-training phase; and 2) classify underwater acoustic signal in fine-tuning phase.

capability, in turn, bolsters the model's effectiveness in subsequent classification tasks. In practice, our Swin-Transformer encoder is comprised of 4 blocks with depths of 2, 2, 18, and 2, respectively. Each block employs multi-head attention with sizes of 4, 8, 16, and 32, correspondingly. Additionally, we set the window size of the Swin-Transformer to 8×8 .

Regarding the decoder, we employ a very shallow network to reconstruct the masked patches. The decoder comprises a single-layer convolutional layer and PixelShuffle method [44]. This choice speeds up the training process without negatively impacting the classification accuracy of downstream tasks. Notably, the decoder is only used in the pre-training phase to reconstruct masked patches.

2) *Mask Modeling-based Pre-training with Mixup*: As mentioned earlier, SSL allows for the extraction of general representations from large-scale data without human annotations, which can be expensive and time-consuming. With this in mind, we proposes a pre-training strategy that utilizes a masking-reconstruction process. The process randomly masks a part of the log Mel-spectrogram, and the network is trained to reconstruct the masked parts. This approach helps the model capture more contextual information and extract more discriminative representations, as previous research has demonstrated [15]. To increase the diversity of the training data and improve the generalization and robustness of the pretext task, a mixup strategy is used to augment the training data. The mixup strategy is performed with a certain probability before converting the audio signal to log Mel-spectrogram. This process can be formulated as follows:

$$s_{mixup} = s_1 * ratio + s_2 * (1 - ratio) \quad (2)$$

where s_1 and s_2 denote the raw acoustic signals, while the ratio signifies the mixing ratio. This ratio is generated using the β -distribution, which is a continuous probability distribution defined on the interval $[0, 1]$. This can be formulated as follows:

$$\beta(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt \quad (3)$$

where a and b are the shape parameters of the distribution. Inspiration from [45], the mixup probability for our main experiment is 0.5 with a $\beta(10, 10)$ distribution.

Specifically, the mixed raw audio signal is then converted to a log Mel-spectrogram X . The log Mel-spectrogram is split into a sequence of 32×32 patches x , and parts of them are randomly masked (as shown in Fig. 1). The unmasked patches are fed into the Swin-Transformer-based encoder to extract generic features, after adding a trainable positional embedding. For each masked patch x_i , we obtain the corresponding Swin-Transformer encoder output O_i , which is subsequently input to a decoder to reconstruct the masked patch r_i . Our aim is to obtain a reconstructed patch r_i that closely matches the original x_i , enabling the model to correctly match (x_i, r_i) pairs. Accordingly, we employ the ℓ_1 loss to evaluate the reconstruction error:

$$loss = \frac{1}{n} \sum_{i=1}^N |r_i - x_i| \quad (4)$$

3) *Masking Strategy*: Masking strategy refers to a technique employed in various tasks, particularly in self-supervised learning, where specific portions or elements of data are intentionally hidden or masked during training [15], [40], [42].

This strategy encourages the model to predict or reconstruct the hidden parts, effectively guiding it to learn meaningful representations or features that capture important information within the data. As a practical implementation for the underwater acoustic signal classification task, the mask strategy in our proposed framework is determined by three factors: mask level, mask ratio, and mask size. The mask level can be divided into patch-level, frequency frame-level, and time frame-level. An illustrative instance of the masking strategy is depicted in Fig. 2. Notably, it is observable that frame-level masking (in both frequency and time domains) exhibits completed gaps, potentially limiting the capacity of SSL methods to learn solely the temporal frame structure. In contrast, patch-based masking empowers the model to capture both temporal and frequency spectrogram structures through analysis of adjacent patches. The choice of mask level hinges upon the particular type of acoustic signal. Generally, time frame-level masking proves more suitable for speech signals, while patch-level masking is found to be more effective for generic acoustic signals [15]. Thus, our approach employs patch-based masking. The mask ratio refers to the proportion of the masked part in the whole input, and the mask size determines the difficulty of reconstruction. A good mask strategy can enable the model to make full use of the contextual information from the unmasked parts, thereby extracting better representations. Drawing from the insights of [42] and our empirical analysis, we determined that employing a patch-level mask strategy, with a mask ratio of 0.6 and a mask size of 32×32 pixels (as exemplified in Fig. 1), yields enhanced performance in downstream tasks for our model. The underlying rationale behind masking is to curtail the model's over-reliance on particular features or patterns that might not exhibit strong generalization to novel data.

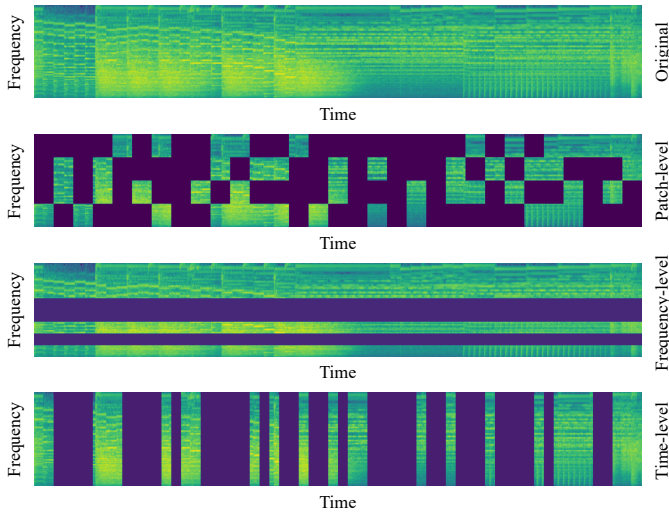


Fig. 2. Illustrated in the figure are examples of the mask strategy, encompassing patch-level, temporal frame-level, and frequency frame-level applications. The first row showcases the original log Mel-spectrogram, while the subsequent three rows offer a visual representation of the outcomes resulting from the application of various masking strategies.

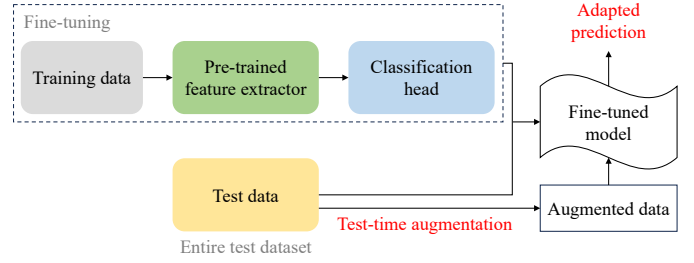


Fig. 3. The figure illustrates the setup of the fine-tuning phase. It depicts the main steps in this phase, including the integration of the TTA module. After fine-tuning, each batch of test data is augmented using the mixup strategy. Then, both the original and augmented data are input to the fine-tuned model. Finally, the predictions from these inputs are averaged and summed based on the mixup approach.

C. Fine-tuning for downstream tasks

In our method, the fine-tuning phase is to encourage the pre-trained model to perform well on the underwater acoustic signal classification task by using a small number of labeled data. The fine-tuning framework of our method is comprised of an encoder that extracts generic audio representations and a classification head for classifying the signals. In implementation, we replace the final layer of the pre-trained model by a random initialized linear layer, which is exactly the classification head, as shown in Fig. 3.

As previously mentioned, the mixup-based pre-training has been effective in enhancing data diversity to alleviate the multi-path effects inherent in marine environments (this can later be supported in experiments, see Table II and Fig. 5), enabling the model to acquire generic and discriminative representations. While mixup has been relatively successful, its granularity is somewhat coarse in this context. Furthermore, the complexities introduced by the unpredictable and ever-changing nature of marine environments necessitate a more refined approach. Recognizing the Swin-Transformer encoder's proficiency in capturing both local and global features, coupled with the observation that a single sample may exhibit similarities or near-identical characteristics when viewed from different perspectives, we introduce a TTA module to further enhance model robustness. Notably, since the mixup strategy is already employed during the pre-training phase, a natural extension of TTA emerges in the form of mixup. In this context, we apply mixup directly to the original log Mel-spectrogram, introducing an augmentation coefficient. This step could emphasize boundary information, a critical factor for classification. This augmentation process can be defined as follows:

$$spec = \lambda \times spec_0 \quad (5)$$

where $spec$ denotes the spectrogram resulting from the mixup augmentation, and $spec_0$ represents the original log Mel-spectrogram. In this context, λ signifies the mixup augmentation coefficient.

Additionally, drawing inspiration from computer vision, it's noteworthy that even when the same image undergoes flip or rotation transformations, the resulting images maintain remarkable intra-consistency. Expanding on this insight and

considering the temporal nature of acoustic signals, a vertical flip strategy has been integrated into the TTA module to preserve the segmentation of the signal. The finalized structure of the TTA module is as follows:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n p_i \quad (6)$$

where $\hat{p}$ represents the final adapted prediction result, n donates the number of transformations used in TTA (here, $n = 3$). p_i (where i ranges from 1 to n) corresponds to the predictions for the original test data and TTA with different transformations, respectively.

In summary, throughout the fine-tuning phase, we initiate the process by extracting the log Mel-spectrogram and inputting them into the pre-trained encoder for the extraction of hidden features. These acquired features are subsequently utilized to train the classification head. To further amplify model robustness, a TTA module is employed, generating an ensemble of predictions.

IV. EXPERIMENT RESULTS AND DISCUSSION

In this section, we elucidate the experimental setup and protocols that were employed to assess the efficacy of our proposed model. We provide an intricate account of the utilized dataset, preprocessing methodologies, evaluation metrics, and a comprehensive assessment of our method's performance across diverse scenarios. This encompassing evaluation encompasses sensitivity analysis of parameters, examination under conditions of low signal-to-noise ratio (SNR), and scrutiny in the context of few-shot scenarios.

A. Dataset description

In this paper, we present a pre-training and fine-tuning approach for our method using the AudioSet-2M dataset [18] and the DeepShip dataset [46], respectively. AudioSet-2M is a multi-label audio event classification dataset that contains 2 million 10-second audio clips from YouTube videos with 527 sound classes. Due to its diverse range of acoustic signal categories, including human sounds, animal sounds, and sounds of things, and its large amount of audio data volume, pretraining with AudioSet-2M enables our model to obtain a comprehensive perception of various acoustic signals and learn discriminative representations. The DeepShip dataset consists of 47 hours and 4 minutes of real-world underwater acoustic signals produced by 256 types of ships and divided into four classes. For this study, we added a fifth class of marine background noise¹ to form a new DeepShip dataset. We randomly divided the new dataset into training and test sets. To align with the 10-second audio clips used in the pretraining phase with AudioSet-2M, we further partitioned the acoustic recordings in both sets into segments of 10 seconds for fine-tuning. Table I presents comprehensive information about the fine-tuning dataset. In practical implementation, the audio was sampled at a rate of 32,000Hz. We proceeded to extract log Mel-spectrograms using time frames of 10ms, encompassing

128 frequency bins per frame. Subsequent to this extraction, a normalization procedure was conducted on the log Mel-spectrograms using the dataset's mean and standard deviation. The calculated values for the mean and standard deviation were -4.268 and 4.569, correspondingly.

TABLE I
THE DESCRIPTION OF DEEPSHIP DATASET.

name	Class	Training samples	Testing samples
DeepShip	Cargo	2505	1279
	Passengers	2810	1510
	Tankers	2710	1570
	Tug	2724	1289
	Background	1921	800

B. Implementation details

This part gives a details of our method's implementation, which can be divided into two phase: the pre-training phase on AudioSet and the fine-tuning phase on DeepShip dataset. During the pre-training phase, we leveraged the Swin-Transformer architecture, employing a configuration of 4 blocks for our encoder. To ensure a fair comparison with the Vision Transformer (ViT), we established the depths of these four Swin-Transformer blocks at 2, 2, 18, and 2 layers, respectively. Furthermore, we designated the number of multi-head attention mechanisms within each block as 4, 8, 16, and 32 heads, correspondingly. The decoder component takes the form of a streamlined network, encompassing a solitary convolutional layer and adopting the PixelShuffle method. Concerning the specific details of the pre-training process, we adopted a mixup ratio of 0.5, drawing inspiration from [47]. Subsequently, the preprocessed log Mel-spectrograms were fed into the model using a batch size of 128, and an individual iteration employed a learning rate of 0.003. The training protocol spans across 128 epochs, incorporating a warm-up epoch that extends over 2 iterations. This protocol is orchestrated using the AdamW optimizer, while the learning rate schedule is orchestrated through the implementation of the cosine decay strategy. Significantly, the entirety of the pre-training phase was accomplished within an approximate span of 3 days, harnessing the computational capabilities of 8 NVIDIA A40 GPUs.

For the subsequent downstream fine-tuning, we introduced a pooling layer and a linear classification head, which replaced the final layer of the pre-trained model, with parameters initialized randomly for the targeted classification task. This transition allowed the model's adaptation to the specific task. The fine-tuning phase extended over 100 epochs, employing a batch size of 256 and a learning rate set at 0.000625. A warm-up period spanning 10 epochs was integrated into the process, facilitated by the utilization of the AdamW optimizer. The learning rate adjustment followed a cosine decay strategy. In terms of the loss function, we employed the cross-entropy (CE) loss, which is well-suited for classification tasks, during the fine-tuning stage. To enhance the model's performance, a TTA module was incorporated. This strategy involved employing a mixup ratio of 1.1 along with a flip transformation, contributing to improved generalization and robustness.

¹<https://github.com/ZhuPengsen/Dataset-segmentation>

TABLE II
QUANTITATIVE COMPARISON BETWEEN DIFFERENT KINDS OF METHODS USING THE DEEPSHIP DATASET.

Type	Methodology	Venue	Supervised	Self-supervised	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Deep Learning Methods	Ours	–		✓	86.33	85.72	82.91	84.29
	SSLMM	JASA 2023		✓	80.22	80.81	79.94	80.07
	SNANet	Applied Acoustics 2023	✓		78.25	79.55	79.39	79.16
	SSAST	AAAI 2022		✓	77.70	78.13	78.25	78.19
	AudioMAE	NeurIPS 2022		✓	76.66	85.54	79.00	82.14
	SCAE	Expert Systems with Applications 2021	✓		77.53	77.75	77.41	77.58
	Residual		✓		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
Traditional Machine Learning Methods	SVM	Expert Systems with Applications 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

Consistent with numerous prior studies, we utilize four commonly employed metrics to assess the efficacy of our proposed approach. These metrics encompass accuracy, precision, recall, and F1-score.

C. Quantitative comparison

We evaluated the effectiveness of our proposed method for underwater acoustic signal classification on the DeepShip dataset by comparing it against competitive baselines of traditional machine learning and deep learning. Specifically, we utilized Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN) algorithms as the traditional methods baselines, while SSLMM [42], SNANet [48], SSAST [15], Separable Convolutional Autoencoder (SCAE) [46], Convolutional Neural Networks (including the Residual Network and Inception architecture), and a standard Deep Neural Network (DNN) were selected as the deep learning baselines for comparison. SSLMM employs a Swin-Transformer based approach combined with two decoders for the purpose of reconstructing masked acoustic representations. This strategy effectively augments the model's capacity to learn and represent information. SNANet is a CNN-based approach that involves extracting spectrum features from each component in different frequency bands. SSAST is a standard transformer-based SSL method for audio classification. This method adopts a pre-text task centered around a masked spectrogram patch modeling strategy, integrating both discriminative and generative aspects. The fundamental procedure and core architecture of AudioMAE closely resemble that of SSAST. However, AudioMAE diverges in its approach by employing a higher mask ratio and exclusively inputting visual patches into the encoders. SCAE is a symmetric encoder-decoder framework where the encoder fuses a separable convolution block with an Xception block to transmute input data into an abstract discriminative representation. This representation is subsequently reassembled by the decoder, structured on the U-Net architecture [49]. ResNet uses residual connections to address the vanishing gradient problem in deep networks and typically consists of several residual blocks with multiple convolutional layers. Inception employs parallel

convolutional layers with varying filter sizes to capture features at different scales, enabling the network to effectively capture local and global patterns. In practice, we adopt ResNet50 and Inception_v4 as the backbones.

As illustrated in Table II, our method outperforms all other methods by achieving the highest accuracy score of 86.33%. The scores for other evaluation metrics, including precision, recall, and F1-Score, also demonstrate the superiority of our method over others, with scores of 85.72%, 82.91%, and 84.29% respectively. SSLMM remained at the second position by obtaining scores of 80.22%, 80.81%, 79.94% and 80.07% for classification accuracy, precision, recall and F1-Score. SNANet, SSAST, AudioMAE, SCAE, Residual-based network, Inception and DNN achieved a accuracy score of 78.25%, 77.70%, 76.66%, 77.53%, 76.98%, 76.16% and 73.11% respectively. Upon comparing the performance of our method, SSLMM, SSAST, and AudioMAE, both of which are transformer-based SSL approaches, it becomes evident that their performance surpasses that of other methods. This outcome serves as a validation of the effectiveness of the pre-training process. Notably, in the case of AudioMAE, there is a slight discrepancy between the accuracy score and precision. We believe this discrepancy may be attributed to the high mask ratio, which poses challenges for the model in accurately classifying negative samples, thus resulting in this outcome. Furthermore, in-depth analysis of the performance of models such as SCAE, Residual, Inception, and DNN reveals a notable enhancement in model performance owing to the inclusion of convolutional operations. Among machine learning methods, SVM achieved the highest accuracy score of 72.24% and outperformed other machine learning-based methods. RF ranked second with a classification accuracy score of 69.71%. Our findings indicate that DNN-based methods generally exhibit superior performance compared to traditional machine learning-based methods with shallow network architecture, which can be attributed to their increased capacity for learning more complex representations. It is observed that proposed method outperformed all other methods, which we attribute to three key factors: (1) the utilization of Swin-Transformer as the encoder, enabling extraction of both local contextual

information and global structure for more discriminative representations. (2) the incorporation of mixup strategy during pre-training to increase data diversity, thereby bolstering the model's robustness, and augment its ability to generalize across distinct data classes. and (3) the application of TTA module from various perspectives, enabling multiple views of the same sample. This is coupled with output averaging, which serves to mitigate the potential of misclassification.

D. Extended analysis

Parameter sensitivity: To ensure the attainment of a robust fine-tuned model, we conducted an additional assessment to gauge the model's responsiveness to the augmentation coefficient (λ) used within the TTA module. As depicted in Fig. 4, the integration of mixup in the TTA module effectively elevates classification task performance to a certain extent, with the optimal performance achieved when λ is set to 1.1. We attribute this enhancement to mixup's ability to enhance the sharpness of discriminative boundary information, a crucial element for successful classification. Conversely, it's apparent that the enhancement progressively diminishes when the value of λ surpasses 1.1, and even deteriorates beyond the performance of not employing mixup. This decline might stem from the fact that as boundaries become excessively sharp, the previously inconspicuous original boundaries are sharpened to a point where classification becomes confusing.

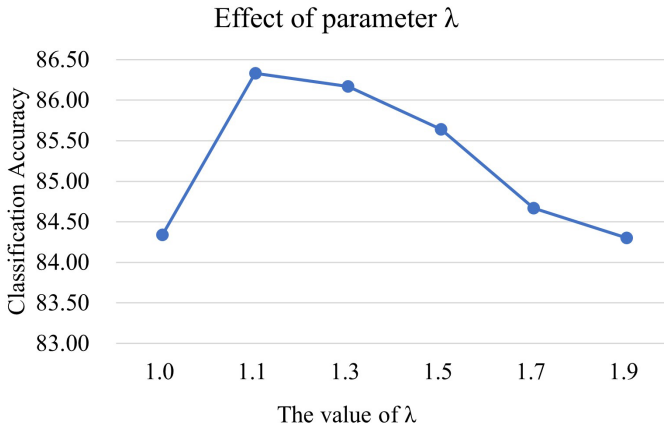


Fig. 4. The sensitivity of parameter λ .

Computational complexity: In general, the model's performance tends to improve as the number of model parameters increases. However, this often results in higher computational complexity. To provide a more comprehensive evaluation of our method compared to competitive baselines, we conducted a computational complexity analysis (focusing on DNN-based approaches). Specifically, we employed two widely-used metrics: the number of model parameters (Params) and floating-point operations per second (FLOPs) to assess computational complexity. The results are presented in Table III. It can be seen that the number of parameters in traditional DNN models varies significantly based on their architecture, whereas SSL-based methods tend to maintain an approximately consistent number of parameters. This distinction arises because traditional DNN methods often rely on structural changes to

enhance performance, whereas SSL methods prioritize the design of pre-text tasks. Additionally, DNN exhibits the lowest FLOPs, which aligns with their simpler structure. Residual network has the fewest parameters due to their convolution operations only related to kernel size and channel number, thereby greatly reduces parameters through receptive fields and weight sharing. Furthermore, SSL-based methods generally exhibit higher computational complexity than traditional DNN-based methods. This heightened complexity arises from the incorporation of self-attention mechanisms, enabling inputs from various locations to interact with one another, along with multi-headed attention mechanisms that enable models to prioritize different information within distinct scale. Besides, the absence of labels requires SSL models to learn implicit supervisory signals from the data itself, necessitating more parameters, such as the decoder used to reconstruct original masked patches. Notably, AudioMAE has lower FLOPs compared to other SSL methods, primarily because it only processes visual patches. In contrast, our method's FLOPs are roughly three times higher than SSLMM due to the integration of the TTA module.

TABLE III
THE COMPUTATIONAL COMPLEXITY OF CURRENT DNN-BASED METHODS FOR UNDERWATER ACOUSTIC SIGNAL CLASSIFICATION.

	Params(M)	FLOPs(M)	Accuracy (%)
DNN	67.24	1075.86	73.11
Inception	48.46	260665.81	76.16
Residual	23.51	169399.75	76.98
SCAE	-	-	77.53
AudioMAE	85.26	349679.77	76.66
SSAST	85.26	827537.63	77.70
SNANet	-	-	78.25
SSLMM	86.68	633621.15	80.22
Ours	86.68	1900863.46	86.33

Noise tolerance: It is widely acknowledged that data often undergo various forms of degradation, including noise effects and multi-path effects during propagation [42], [50], [51]. To assess the robustness of our method in the presence of such variabilities, particularly noise effects, we conducted comprehensive noise tolerance analyses. Inspired by the fact that the combination of multiple noises often approximates a Gaussian distribution in complex scenarios, we employed Gaussian white noise to simulate real-world noise generation. Subsequently, we introduced this noise into the original signals, varying the signal-to-noise ratios (SNR) from -5dB to 5dB. These modified signals were then utilized for training the model, with performance evaluation conducted using a clean testing dataset. The results are depicted in Figure 5. It can be seen that our method exhibits strong classification performance even in conditions with low SNR. Besides, our method achieves the best performance compared with the other competitive baselines. This reason can be attributed to the robust and versatile representations obtained through mixup-based pre-training, with further enhancement in robustness achieved through the application of the TTA module during fine-tuning.

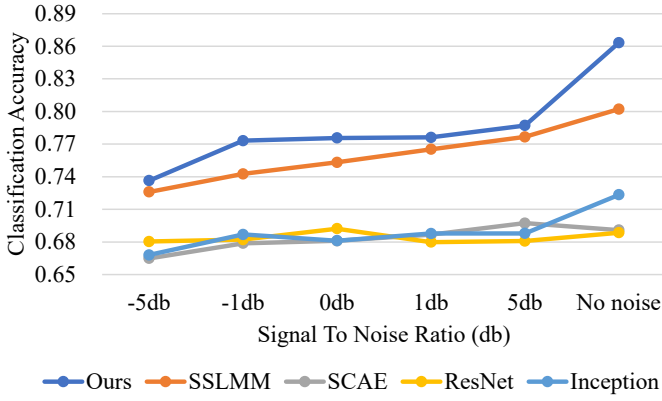


Fig. 5. Noise Tolerance Analysis: The quantitative comparison between our proposed method and competitive algorithms, within different SNR.

E. Results visualization

Reconstruction visualization: As mentioned before, the quality of a pre-trained model is closely tied to the success of downstream tasks. To further evaluate the effectiveness of the pre-training phase, we randomly selected an audio recording and performed patch-level masking on it. Subsequently, the pre-trained model was employed to reconstruct the masked patches, drawing on the knowledge gleaned from the unmasked counterparts. We visualize the outcomes of pre-training with and without the mixup strategy in Fig. 6. In the red region, it's apparent that the latter highlighted line more closely resembles the original. Moving to the yellow region, the latter reconstruction exhibits greater smoothness. Within the orange region, the latter reconstruction captures more information. Overall, these reconstructions exhibit greater resemblance to the original data. Based on these observations, it is evident that pre-training with mixup produces significant improvements. Our findings underscore that the pre-trained model, when integrated with mixup, achieves a more accurate reconstruction of masked patches. This underscores the model's successful assimilation of contextual information from the unmasked patches. Notably, in the red, yellow, and orange regions, the pre-trained model with mixup demonstrates enhanced capability in reconstructing intricate representations similar to the original data, as compared to the scenario without mixup.

Feature visualization: To provide a clear and intuitive presentation, we employ t-distributed stochastic neighbor embedding (t-SNE) to visualize the results of underwater acoustic signal classification. Recognizing the importance of hidden representations in classification, we utilize t-SNE to map the hidden representations generated by our proposed model's encoder after fine-tuning into a two-dimensional space, as illustrated in Fig. 7. The visual representation clearly shows that representations from the same class are closely clustered together, while those from different classes are separated. These observations affirm that our proposed method effectively captures discriminative representations, thereby enhancing its performance in downstream classification tasks.

F. Few-shot settings

Here, we conducted experiments to demonstrate the generalization ability of the proposed framework under few-shot settings. Specifically, we randomly sampled the labeled training data at sample ratios of 50%, 10%, 5%, and 1% from the whole DeepShip dataset and used these re-sampled data to train different methods. The accuracy results for both our proposed model and compared methods are presented in Fig. 8. Our method achieved accuracy scores of 54.43%, 78.03%, 80.79% and 83.04 with sample ratio of 1%, 5%, 10% and 50% respectively. Conversely, the second-ranking approach, SSLMM, attained percentages of 67.99%, 73.33%, 74.79%, and 79.62% for the respective metrics, indicating an overall inferior performance compared to our method. In the context of the 1% ratio, our method exhibited lower performance than SSLMM. This discrepancy might be attributed to the incorporation of the mixup strategy augments sample diversity, however, it's conceivable that the resulting discriminative representation might not have been sufficiently learned in the same limited samples. Notably, we observed that the classification accuracy curve of our method outperforms other competitive baselines, demonstrating that our approach achieves satisfactory classification performance even under limited data conditions. We attribute this improvement to the superior prior knowledge obtained from the pre-trained model in the AudioSet, which enables better classification accuracy on specific tasks despite having a small number of samples. Additionally, the pre-training and fine-tuning paradigm effectively leverages optimal model weights that were pre-trained on a large scale of unlabeled acoustic signals, enabling our method to quickly converge to specific tasks.

G. Ablation study

To demonstrate the validation of mixup and TTA strategies, We perform an ablation study to tease apart the effectiveness of each component with the DeepShip dataset. Specifically, we design three experiment settings: ① Without mixup in training phase and without TTA in fine-tuning phase; ② Using mixup in training phase and without TTA in fine-tuning phase; ③ Simultaneously using mixup in training phase and using TTA in fine-tuning phase. The relevant results are presented in Table IV. It can be seen that the utilization of all components within our proposed framework leads to best performance. Specifically, our method achieves an accuracy score of 86.33%, a precision score of 85.72%, a recall score of 82.91%, and an F1-score of 84.29%. These results surpass the case ① (which does not employ any strategies) by 6.88%, 5.85%, 3.74%, and 5.06%, respectively. Furthermore, our method outperforms case ② (which utilizes only the mixup strategy in the pre-training phase) with improvements of 1.99%, 1.42%, 1.70%, and 1.56%, respectively. Moreover, when comparing the improvements of case ① and ② against case ② and ③, it becomes evident that the mixup strategy employed during the pre-training phase is more effective than TTA used in the fine-tuning phase. This observation demonstrates that the mixup strategy during pre-training enriches the diversity of acoustic signals and effectively alleviates the multi-path

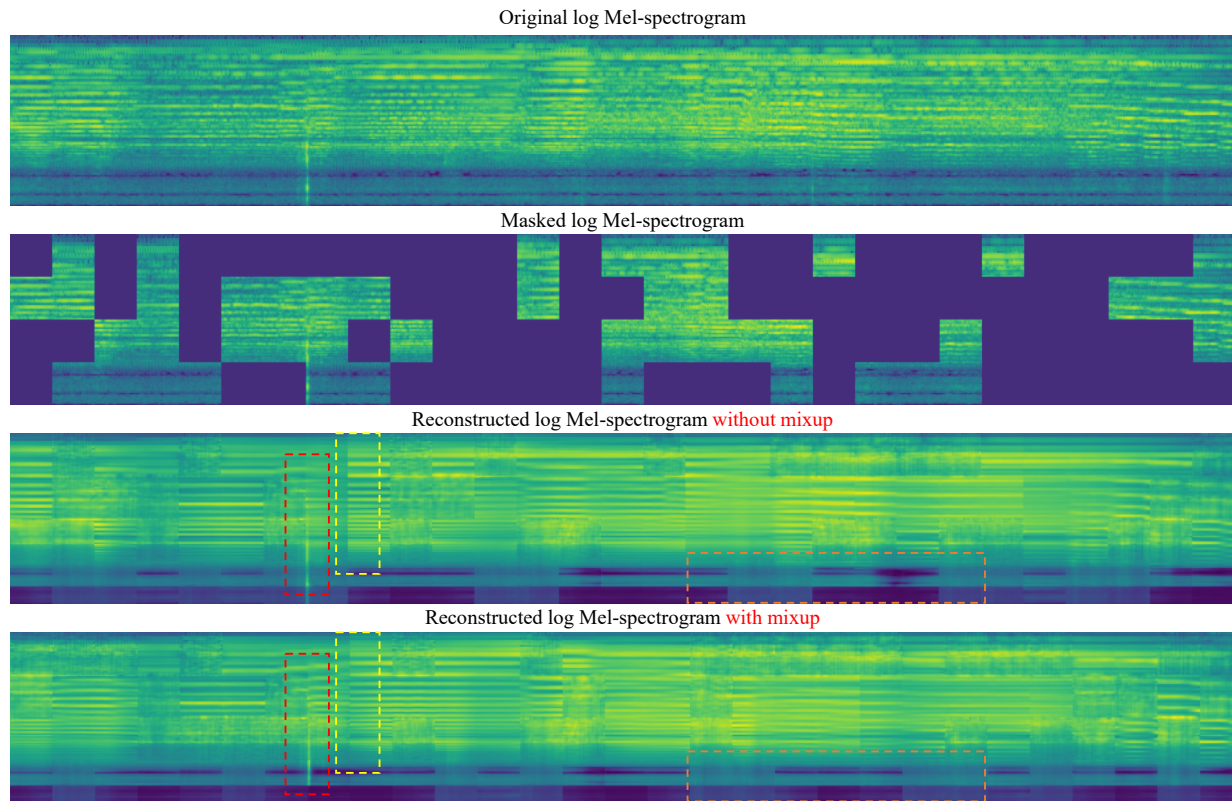


Fig. 6. Visualizing the Mask Reconstruction Results. Displayed are the original, masked, and reconstructed log Mel-spectrograms. Notably, the third row showcases the reconstruction by the pre-trained model without mixup, while the fourth row presents the reconstruction with mixup. In the red region, it's apparent that the latter highlighted line more closely resembles the original. Moving to the yellow region, the latter reconstruction exhibits greater smoothness. Within the orange region, the latter reconstruction captures more information. Based on these observations, it becomes evident that the pre-training with mixup yields valid improvements.

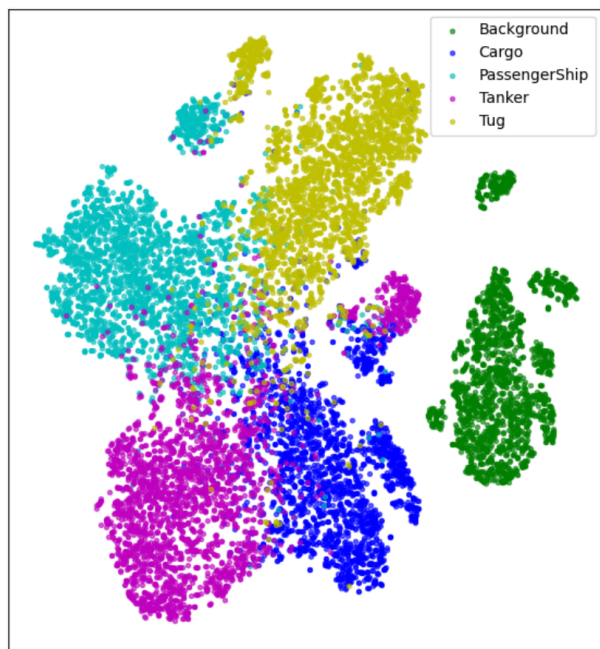


Fig. 7. t-SNE visualizing results of the hidden representations. Dots of different colors denote hidden representations of various classes.

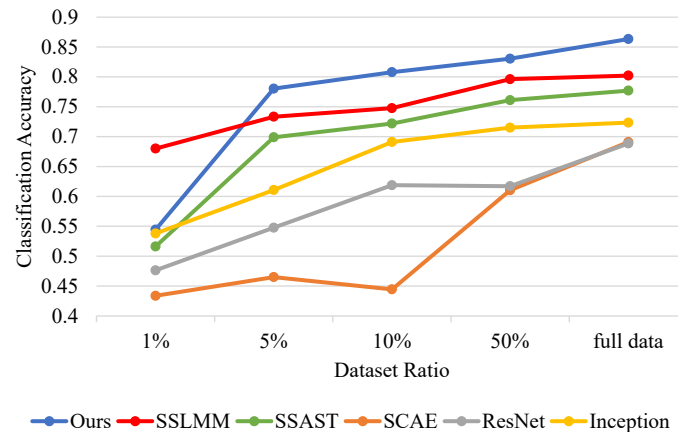


Fig. 8. The experimental results under few-shot settings. We compared with the competitive deep learning-based baselines.

effect present in marine environments, which encourages the model to capture robust and generic representations in turn. Additionally, it is worth noting that while the performance of case ① is slightly lower than the latest baseline SSLMM, this difference may be attributed to SSLMM's use of multi-representation mask modeling, whereas our method relies on a single representation for reconstruction.

TABLE IV

THE IMPACT OF THE MIXUP IN PRE-TRAINING AND TTA IN FINE-TUNING FOR THE CLASSIFICATION MODEL.

No.	Training with mixup	Fine-tuning with TTA	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
①	✗	✗	79.45	79.87	79.17	79.23
②	✓	✗	84.34	84.30	81.21	82.73
③	✓	✓	86.33	85.72	82.91	84.29

V. CONCLUSION

In this paper, we propose an underwater acoustic signal classification method based on self-supervised learning, combined with mixup, and TTA strategies. In our proposed method, we implement a mixup-based strategy during the pre-training phase to enhance the diversity of training data. This strategy effectively alleviates the multi-path effect characteristic of marine environments. Additionally, we design a hierarchical multi-scale mask modeling strategy as a pretext task during this phase, which allows us to derive robust and generic representations suitable for downstream tasks. In the subsequent fine-tuning phase, we introduce a TTA module to further bolster the robustness of our approach. The experimental results demonstrate the effectiveness of our proposed method, achieving a superior classification accuracy of 86.33% on the DeepShip dataset, which includes the challenging fifth class related to ocean environment noise. This performance surpasses that of competitive baseline models. Furthermore, we extended our experiments to few-shot settings, where we observed that our classification method consistently maintains impressive performance, even when faced with limited labeled data scenarios.

Notably, our approach incorporates a mixup-based strategy, inspired by the observed multi-path effect typical of marine environments. While our method has demonstrated effectiveness through high metric values, there remains room for improvement. For instance, the mixup-based strategy relies on artificial prior knowledge and design, which poses a limitation in terms of granularity.

In the future, we plan to conduct a more fine-grained analysis and modeling to better represent multi-path effects and other forms of interference in marine environments. This could involve exploring various data augmentation techniques to enhance our approach. Additionally, we are considering the application of Neural Architecture Search (NAS) techniques as an alternative solution for further optimization.

REFERENCES

- [1] V.-S. Doan, T. Huynh-The, and D.-S. Kim, "Underwater acoustic target classification based on dense convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2020.
- [2] T. Hemminger and Y.-H. Pao, "Detection and classification of underwater acoustic transients using neural networks," *IEEE Transactions on Neural Networks*, vol. 5, no. 5, pp. 712–718, Sep. 1994.
- [3] D. B. Kilfoyle and A. B. Baggeroer, "The state of the art in underwater acoustic telemetry," *IEEE Journal of oceanic engineering*, vol. 25, no. 1, pp. 4–27, 2000.
- [4] S. Jiang, "On securing underwater acoustic networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 729–752, 2018.

- [5] K. Feroze, S. Sultan, S. Shahid, and F. Mahmood, "Classification of underwater acoustic signals using multi-classifiers," in *2018 15th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*. IEEE, 2018, pp. 723–728.
- [6] Y. Li, X. Chen, J. Yu, and X. Yang, "A fusion frequency feature extraction method for underwater acoustic signal based on variational mode decomposition, duffing chaotic oscillator and a kind of permutation entropy," *Electronics*, vol. 8, no. 1, p. 61, 2019.
- [7] İ. G. Aksüren and A. K. Hocaoglu, "Automatic target classification using underwater acoustic signals," in *2022 30th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2022, pp. 1–4.
- [8] Z. Alouani, Y. Hmamouche, B. El Khamlichi, and A. E. F. Seghrouchni, "A spatio-temporal deep learning approach for underwater acoustic signals classification," in *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2022, pp. 1–7.
- [9] X. Wang, A. Liu, Y. Zhang, and F. Xue, "Underwater acoustic target recognition: a combination of multi-dimensional fusion features and modified deep neural network," *Remote Sensing*, vol. 11, no. 16, p. 1888, 2019.
- [10] G. Hu, K. Wang, and L. Liu, "Underwater acoustic target recognition based on depthwise separable convolution neural networks," *Sensors*, vol. 21, no. 4, p. 1429, 2021.
- [11] B. Wang, W. Zhang, Y. Zhu, C. Wu, and S. Zhang, "An underwater acoustic target recognition method based on amnet," *IEEE Geoscience and Remote Sensing Letters*, 2023.
- [12] C. Liu, F. Hong, H. Feng, and M. Hu, "Underwater acoustic target recognition based on dual attention networks and multiresolution convolutional neural networks," in *OCEANS 2021: San Diego–Porto*. IEEE, 2021, pp. 1–5.
- [13] D. Li, F. Liu, T. Shen, L. Chen, X. Yang, and D. Zhao, "Generalizable underwater acoustic target recognition using feature extraction module of neural network," *Applied Sciences*, vol. 12, no. 21, p. 10804, 2022.
- [14] D. Huang, W. Wu, W. Hu, X. Liu, D. He, Z. Wu, X. Wu, M. Tan, and E. Ding, "Ascnet: Self-supervised video representation learning with appearance-speed consistency," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8096–8105.
- [15] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "Ssast: Self-supervised audio spectrogram transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10 699–10 709.
- [16] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," *Advances in Neural Information Processing Systems*, vol. 35, pp. 28 708–28 720, 2022.
- [17] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, p. 16.
- [18] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.
- [19] S. Wang and X. Zeng, "Robust underwater noise targets classification using auditory inspired time–frequency analysis," *Applied Acoustics*, vol. 78, pp. 68–76, 2014.
- [20] H. Shi, J. Xiong, C. Zhou, and S. Yang, "A new recognition and classification algorithm of underwater acoustic signals based on multi-domain features combination," in *2016 IEEE/OES China Ocean Acoustics (COA)*. IEEE, 2016, pp. 1–7.
- [21] H. Li, P. Yue, and L. Jiangqiao, "Classification of underwater acoustic target using auditory spectrum feature and svdd ensemble," in *2018 OCEANS-MTS/IEEE Kobe Techno-Oceans (OTO)*. IEEE, 2018, pp. 1–4.
- [22] X. Cao, R. Togneri, X. Zhang, and Y. Yu, "Convolutional neural network with second-order pooling for underwater target classification," *IEEE Sensors Journal*, vol. 19, no. 8, pp. 3058–3066, 2018.
- [23] Y. Wang, H. Zhang, L. Xu, C. Cao, and T. A. Gulliver, "Adoption of hybrid time series neural network in the underwater acoustic signal modulation identification," *Journal of the Franklin Institute*, vol. 357, no. 18, pp. 13 906–13 922, 2020.
- [24] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4340–4354, 2021.
- [25] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE*

- Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5966–5978, 2021.
- [26] F. Hong, C. Liu, L. Guo, F. Chen, and H. Feng, “Underwater acoustic target recognition with a residual network and the optimized feature extraction method,” *Applied Sciences*, vol. 11.
- [27] S.-Z. Tian, D.-B. Chen, Y. Fu, and J.-L. Zhou, “Joint learning model for underwater acoustic target recognition,” *Knowledge-Based Systems*, vol. 260, p. 110119, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0950705122012151>
- [28] D. Li, F. Liu, T. Shen, L. Chen, and D. Zhao, “Data augmentation method for underwater acoustic target recognition based on underwater acoustic channel modeling and transfer learning,” *Applied Acoustics*, vol. 208, p. 109344, 2023.
- [29] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, “Lrr-net: An interpretable deep unfolding network for hyperspectral anomaly detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.
- [30] H. Yang, J. Li, S. Shen, and G. Xu, “A deep convolutional neural network inspired by auditory perception for underwater acoustic target recognition,” *Sensors*, vol. 19, no. 5, p. 1104, 2019.
- [31] Y. Xie, J. Ren, and J. Xu, “Adaptive ship-radiated noise recognition with learnable fine-grained wavelet transform,” *Ocean Engineering*, vol. 265, p. 112626, 2022.
- [32] J. Ren, Y. Xie, X. Zhang, and J. Xu, “Ualf: A learnable front-end for intelligent underwater acoustic classification system,” *Ocean Engineering*, vol. 264, p. 112394, 2022.
- [33] S. Feng and X. Zhu, “A transformer-based deep learning network for underwater acoustic target recognition,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.
- [34] C. Doersch, A. Gupta, and A. A. Efros, “Unsupervised visual representation learning by context prediction,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1422–1430.
- [35] Z. Li, Z. Chen, F. Yang, W. Li, Y. Zhu, C. Zhao, R. Deng, L. Wu, R. Zhao, M. Tang *et al.*, “Mst: Masked self-supervised transformer for visual representation,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 165–13 176, 2021.
- [36] C. Zhuang, A. L. Zhai, and D. Yamins, “Local aggregation for unsupervised learning of visual embeddings,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6002–6012.
- [37] O. Henaff, “Data-efficient image recognition with contrastive predictive coding,” in *International conference on machine learning*. PMLR, 2020, pp. 4182–4192.
- [38] A. Ermolov, A. Siarohin, E. Sangineto, and N. Sebe, “Whitening for self-supervised representation learning,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 3015–3024.
- [39] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, “Extended vision transformer (exvit) for land use and land cover classification: A multimodal deep learning framework,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.
- [40] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.
- [41] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, “Panns: Large-scale pretrained audio neural networks for audio pattern recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.
- [42] K. Xu, Q. Xu, K. You, B. Zhu, M. Feng, D. Feng, and B. Liu, “Self-supervised learning-based underwater acoustical signal classification via mask modeling,” *The Journal of the Acoustical Society of America*, vol. 154, no. 1, pp. 5–15, 2023.
- [43] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.
- [44] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.
- [45] D. Ng, R. Zhang, J. Q. Yip, C. Zhang, Y. Ma, T. H. Nguyen, C. Ni, E. S. Chng, and B. Ma, “Contrastive speech mixup for low-resource keyword spotting,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [46] M. Irfan, Z. Jiangbin, S. Ali, M. Iqbal, Z. Masood, and U. Hamid, “Deepship: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification,” *Expert Systems with Applications*, vol. 183, p. 115270, 2021.
- [47] K. Xu, D. Feng, H. Mi, B. Zhu, D. Wang, L. Zhang, H. Cai, and S. Liu, “Mixup-based acoustic scene classification using multi-channel convolutional neural network,” in *Advances in Multimedia Information Processing-PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part III 19*. Springer, 2018, pp. 14–23.
- [48] P. Zhu, Y. Zhang, Y. Huang, C. Zhao, K. Zhao, and F. Zhou, “Underwater acoustic target recognition based on spectrum component analysis of ship radiated noise,” *Applied Acoustics*, vol. 211, p. 109552, 2023.
- [49] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.
- [50] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, “An augmented linear mixing model to address spectral variability for hyperspectral unmixing,” *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1923–1938, 2019.
- [51] D. Hong, X. Wu, P. Ghamisi, J. Chanussot, N. Yokoya, and X. X. Zhu, “Invariant attribute profiles: A spatial-frequency joint feature extractor for hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 6, pp. 3791–3808, 2020.



Qisheng Xu Master student at the School of Computer Science, National University of Defense Technology. He received his bachelor's degree from Wuhan University in 2021. His research interest covers audio signal processing and parallel computing.



Jingfei Jiang Professor at the School of Computer Science, National University of Defense Technology. She was awarded B.S. (1997), M.S. (2000) and Ph.D. (2004) degrees in Computer Science by the School of Computer Science, National University of Defense Technology. Her research interest covers acceleration methods of machine learning algorithms and deep learning.



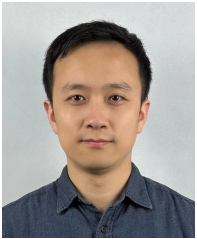
Kele Xu Associate Professor at the School of Computer Science, National University of Defense Technology. He received his doctorate from Paris VI University in 2017. His research interest covers audio signal processing, machine learning, and intelligent software systems.



Yong Dou Professor and Ph.D. supervisor with the National Key Laboratory of Parallel and Distributed Computing, National University of Defense Technology. His research interest covers high performance computing, intelligence computing, machine learning and deep learning.



Caili Gao Master student at the School of Computer Science, National University of Defense Technology. He Received his bachelor's degree from Nanchang University in 2021. His research interest covers face forgery detection and parallel optimization.



Boqing Zhu Ph.D. candidate at the School of Computer Science, National University of Defense Technology. He received his master's degree from National University of Defense Technology in 2019. His research interest covers multi-modal machine learning, incremental learning and acoustics model.



Kang You He received his bachelor's degree from Tongji University in 2023. His research interest covers audio signal processing and self-supervised learning.



Qian Zhu Master student at the School of Computer Science, National University of Defense Technology. He received his bachelor's degree from Naval University of Engineering in 2019. His research interest covers audio signal processing and machine learning.