remote sensing

Article

Multi-Task Mixture-of-Experts Model for Underwater Target Localization and Recognition

Peng Qian [1], Jingyi Wang [2,3], Yining Liu [1], Yingxuan Chen [1], Pengjiu Wang [1], Yanfa Deng [1], Peng Xiao [1,*] and Zhenglin Li [1]

[1] School of Ocean Engineering and Technology, Sun Yat-sen University, Daxuelu Road, Tangjiawan Town, Zhuhai 519082, China; qianp@mail2.sysu.edu.cn (P.Q.); liuyn223@mail.sysu.edu.cn (Y.L.); chenyx688@mail2.sysu.edu.cn (Y.C.); wangpj25@mail2.sysu.edu.cn (P.W.); dengyf37@mail2.sysu.edu.cn (Y.D.); lizhlin29@mail.sysu.edu.cn (Z.L.)

[2] Shanghai Marine Electronic Equipment Research Institute, Shanghai 201108, China; jywang_tju@163.com

[3] Science and Technology on Underwater Acoustics Antagonizing Laboratory, Shanghai 201108, China

* Correspondence: xiaop36@mail.sysu.edu.cn

Abstract

The scarcity of underwater acoustic data in deep and remote sea environments poses a significant challenge to data-driven target recognition models, severely restricting their performance. To address this challenge, this study presents a ray-theory-based data augmentation method for generating synthetic ship-radiated noise datasets in oceanic environments at a depth of 3500 m—DS3500, encompassing both direct and shadow zones. Additionally, a novel MEG (multi-task, multi-expert, multi-gate) framework is developed to achieve simultaneous target localization and recognition by integrating relative positional information between the target and sonar, which dynamically partitions parameter spaces through multi-expert mechanisms and adaptively combines task-specific representations using multi-gate attention to simultaneously predict target localization and recognition. Experimental results on the DS3500 dataset demonstrate that the MEG framework achieves 95.93% recognition accuracy, a range localization error of 0.2011 km and a depth localization error of 20.61 m with a maximum detection range of 11 km and depth of 1100 m. This study provides a new technical solution for underwater acoustic target recognition in deep and remote seas, offering innovative approaches for practical applications in marine monitoring and defense.

Keywords: underwater target recognition; underwater target localization; data augmentation; ray theory; multi-task learning; multi-gate mixture-of-experts model

check for updates

Academic Editors: Prasad S. Thenkabail and Jorge Vazquez

Received: 30 March 2025
Revised: 20 July 2025
Accepted: 25 August 2025
Published: 26 August 2025

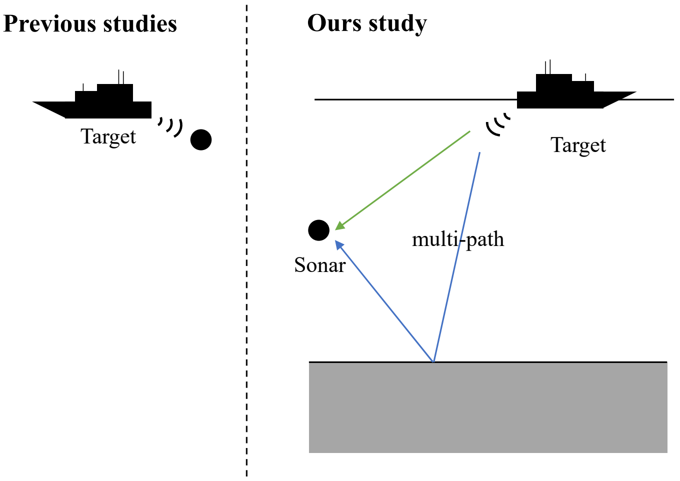
Citation: Qian, P.; Wang, J.; Liu, Y.; Chen, Y.; Wang, P.; Deng, Y.; Xiao, P.; Li, Z. Multi-Task Mixture-of-Experts Model for Underwater Target Localization and Recognition. *Remote Sens.* **2025**, *17*, 2961. https://doi.org/10.3390/rs17172961

1. Introduction

Underwater acoustic target recognition is a key area in marine acoustics [1]. It automatically identifies different target classes by analyzing the radiated noise of underwater targets. This technology is applied in various areas, including underwater monitoring, protection, and the enhancement of security and defense [2–4]. In recent years, deep learning has become the main technology used in underwater acoustic recognition systems.

However, underwater acoustic signals in the actual marine environment face numerous challenges, which are vividly illustrated in Figure 1. Traditional underwater acoustic detection methods predominantly rely on near-field datasets, as shown on the left side. Although data-driven acoustic recognition models based on deep learning demonstrate promising performance on public datasets, the complex time-varying characteristics of the ocean channel pose significant problems [5,6]. During propagation, sound waves are influenced by factors such as seawater temperature, salinity, depth, and ocean topography. These factors give rise to phenomena like signal attenuation, scattering, and multi-path propagation, resulting in blurred and distorted received underwater acoustic signals.

In view of this, we have employed the ocean acoustic channel simulation method based on the theory of ocean acoustic propagation to generate the far-field dataset of the ocean channel. This dataset comprehensively takes into account complex factors such as multi-path propagation, effectively overcoming the limitations of traditional near-field datasets.



	Previous studies	Ours study
Dataset	Near-field	Ocean channel far-field
Network	Classification	Classification and localization

Figure 1. Underwater acoustic detection evolution: shifting from past near-field to our far-field ocean channel-based model.

According to research, most of the existing underwater acoustic recognition literature only relies on class labels as supervision information. The latest progress in machine learning technology provides opportunities to utilize information other than class labels, among which multi-task learning (MTL) can be an attractive solution [7].

In this study, an MTL framework is used, with "estimating the relative position between the target and the sonar" as the localization task, enabling the recognition task to perceive the robust patterns related to the relative position between the target and the sonar in underwater signals. The labels for the localization task are marked simultaneously when creating the dataset based on the sound field model. Moreover, it can also prompt the model to learn acoustic features related to the target, such as comb filtering and interference fringes, thereby deepening the model's understanding of the signal.

Inspired by the mixture-of-experts model (MoE) and the multi-gate mechanism [8,9], an improved multi-task framework MEG (multi-task, multi-expert, multi-gate) [7] is adopted to fully exploit the potential of MTL. Specifically, MEG replaces the traditional output layer with multiple independent network layers (expert layers). The expert layers have the same architecture but different parameters, enabling them to specialize in different aspects and provide fine-grained knowledge through independent parameter spaces. In addition, MEG uses multiple gating layers to dynamically learn task-specific weights, allowing each task to linearly combine the outputs of the expert layers with unique weights to obtain task-specific representations. The top-k gating mechanism [10,11] is used to dynamically select the top k expert layers according to the importance of the task, improving the model's efficiency and performance.

In order to verify the superior performance of the multi-task framework and the MEG model, a data augmentation method based on ray theory was used to generate a synthetic dataset of ship-radiated noise in the direct arrival zone and the shadow zone at a sea depth of 3500 m. A series of experiments were then conducted on a DS3500 dataset. The results show that the multi-task model can achieve a high recognition accuracy and also performs well in the localization task. On the DS3500 dataset, MEG can reach an accuracy of 95.93% in the five-class recognition task. The main contributions of this study are as follows:

1. A ship-radiated noise data augmentation model based on ray theory is proposed for simulating ship-radiated noise signals received by sonar in the ocean channel. This model helps to address the issue of limited underwater acoustic data by generating additional training samples.
2. An identification framework named MEG is designed. By enabling the model to learn the relative position between the target and the sonar, its ability to capture robust patterns is enhanced.
3. The top-k gating mechanism is introduced to dynamically screen the top *k* expert layers according to the importance of the task, improving the model's efficiency and performance.
4. A large number of experiments were conducted to verify the superiority of the MEG model in localization and recognition. The model achieves excellent performance on the DS3500 dataset.

2. Related Research

In the field of underwater acoustic research, existing methods can be comprehensively categorized into three main aspects: underwater acoustic channel modeling methods, underwater acoustic target recognition based on deep learning, and multi-task learning in underwater acoustics. Each of these aspects plays a crucial role in advancing the understanding and application of underwater acoustic technologies.

2.1. Underwater Acoustic-Channel Modeling Methods

Underwater acoustic-channel modeling is closely related to the acoustic-field model and serves as the foundation for underwater acoustic research. It mainly encompasses two categories of methods: frequency-domain and time-domain approaches.

In frequency-domain modeling, the Fourier synthesis of CW (continuous wave) results is commonly used to confront pulse propagation problems [12]. This method calculates the spatial transfer function using existing models and then obtains the time-domain solution through FFT (fast Fourier transform). It is simple with respect to programming and highly compatible. However, issues such as windowing and aliasing effects may arise due to the truncation and discretization of the frequency integral [13]. These problems can be addressed by leveraging the physical knowledge of waveguides and special numerical techniques, such as appropriately selecting time windows and sampling parameters or using the complex frequency integration technique [14].

Time-domain modeling includes several techniques. Ray theory is relatively straightforward in the time domain. For narrow-band signals, the received signal is a scaled and delayed version of the source signal. However, when considering phase changes caused by boundary reflections and caustics, the received waveform becomes more complex [12,15]. The spectral integral technique obtains the governing equations of the time-domain FFP (fast field program) by applying the Hankel transform, omitting the Fourier transform with respect to frequency. After numerical solutions, the pressure field is obtained through the inverse Fourier–Bessel transform, and appropriate wavenumber sampling ranges need to be set [12]. The parabolic equation (PE) can obtain a time-domain solution by directly solving the time-domain wave equation or through Fourier synthesis, and its derivation is based on specific approximation conditions [16–18].

When the motion of the source or receiver is involved, a Doppler shift occurs. This can be analyzed using spectral and modal representation methods. The former incorporates motion into the wave equation to derive expressions and simplifies the calculation by transforming to the receiver's frame of reference. The latter obtains results by considering the characteristics of modal propagation based on the spectral representation [12,19].

In practical applications, these modeling methods have been verified through numerical simulations of various scenarios. For example, in scenarios such as the head-wave problem, mode dispersion in waveguides, and 3D wedge waveguide propagation, these methods are able to analyze complex propagation phenomena [12,20–24].

Underwater acoustic-channel modeling based on ray theory is a crucial research area. The ray-tracing technique can vividly illustrate multi-path propagation characteristics by precisely depicting the propagation paths of sound waves. It is especially applicable to the propagation modeling of high-frequency signals in deep-sea environments, as indicated by the authors of [25,26]. Thus, under deep-sea conditions, by leveraging underwater acoustic-channel modeling based on ray theory and integrating statistical models to enhance the target-radiated noise emitting through the channel, the modeling accuracy and system performance can be effectively improved.

2.2. Underwater Acoustic Target Recognition Based on Deep Learning

Over the past decade, the advent of deep learning technologies and the establishment of extensive underwater noise databases have propelled underwater acoustic target recognition (UATR) research. The application of deep neural networks (DNNs) in this field has become ubiquitous, and there has been an exponential growth in related studies [27].

A significant portion of the research body focuses on designing and improving acoustic features. A substantial amount of research has been dedicated to the design and improvement of acoustic features. The features of classification include those extracted based on the short-time Fourier transform [28,29]. Additionally, techniques such as low-frequency analysis and recording (LOFAR) and the detection of envelope modulation on noise (DEMON) have been widely applied in ship target identification [30–33]. Inspired by human auditory perception, Mel spectrograms, Mel-frequency cepstral coefficients (MFCCs), and Gammatone-frequency cepstral coefficients (GFCCs) have become fundamental elements for underwater acoustic target detection [27,34–36]. Multiscale spectrograms capture information across different frequency resolutions, while cyclic modulation spectra help analyze the periodicity in acoustic signals [37].

Researchers have dedicated significant efforts to optimizing DNN architectures [38]. To further bolster recognition capabilities, recent investigations, such as those carried out by Xie et al. [39], have delved into the application of MoECs (mixture-of-expert clusters). Their exploration has yielded remarkable outcomes in relevant recognition tasks.

Specifically, the learnable fine-grained wavelet transform, as proposed by Xie et al. [40], enables networks to perform adaptive feature extraction, automatically pinpointing relevant features. Moreover, interpretable contrastive learning, as explored by the authors of [41], enhances the models' robustness and generalization capabilities. Through the implementation of these innovative approaches, researchers achieved substantial improvement in recognition accuracy [38–41].

However, most existing UATR studies rely solely on class labels for supervision. This approach often fails to capture the complex underlying patterns of targets, especially when the dataset is limited [42]. Some recent studies have recognized the importance of incorporating additional factors, such as acoustic channels [43], source distances, channel depths, wind speeds [42], and target properties [7]. Nevertheless, the data utilization efficiency of training paradigms such as data augmentation and contrastive learning remains low, suggesting significant room for improvement.

2.3. Multi-Task Learning in Underwater Acoustics

Multi-task learning (MTL) is a powerful approach in machine learning that enables the simultaneous learning of multiple tasks by sharing parameters [44]. It can capture the commonalities and differences among tasks, enhancing data utilization efficiency and generalization capabilities for each individual task. The most basic MTL structure is hard parameter sharing [45], and based on this, improved algorithms, such as the cross-stitch network [46] and sluice network [47], have been proposed. In recent years, the mixture-of-experts (MoE) [48] has been used in MTL to implement dynamic weight sharing. Representative works include multi-gate MoE [9], progressive layered extraction (PLE) [49], and differentiable select-k (DSelect-k) [50], which can reduce the conflict between tasks and shared parameters.

In the field of underwater acoustics, MTL has extensive applications. For underwater acoustic communication-related tasks, it often uses additional tasks, such as channel estimation [51], channel tracking [52], channel equalization [53], and demodulation [54] to improve performance. In other areas such as synthetic aperture sonar (SAS) image classification [55] and sound source localization [56,57], MTL is also widely employed.

However, in underwater acoustic target recognition, research on MTL is still in its infancy. Zeng et al. [58] proposed a multi-task sparse feature learning method for underwater acoustic target recognition by recovering and enhancing prominent structures on spectra. Li et al. [59] designed an anti-noise task and frequency-selection task to optimize the acoustic feature extraction. Most previous studies in this field mainly used MTL to optimize acoustic feature extraction or learning without enabling the model to directly perceive knowledge related to the robust properties of targets. Moreover, they mostly adopted the basic MTL framework, leaving much room for improvement. In this study, we adopt "the relative position of the target to the sonar" as an additional task, to perceive the impact of the ocean acoustic channel, and implement an MoE-style MTL approach in underwater acoustic recognition.

3. Materials and Methods

3.1. Datasets

3.1.1. Ray-Theory-Based Data Augmentation

The ShipsEar [60] dataset was collected along the Atlantic coast of Spain from 2012 to 2013. The hydrophone had a sampling frequency of 52,734 Hz. This dataset contains the radiated noise of 11 different types of ships (e.g., motorboats, fishing boats, tugboats, etc.) with a total of 90 samples. The duration of each recording ranges from 15 s to 10 min. According to the original annotation, the dataset is divided into five categories: A, B, C, D, and E, where category E is ambient noise.

The ShipsEar dataset is comprised of actual ship radiated noise data and exhibits high representativeness. However, due to issues such as excessive noise and blank segments, they all need to be preprocessed. Most existing methods use data cleaning, denoising, and data augmentation operations to address these problems. For the ShipsEar dataset, except for simply removing blank segments, no other operations were performed. According to relevant papers, the data from the dataset were divided into 5-second segments to obtain a large number of short-time segments, thus expanding the data volume. The specific information of the datasets used in this study is shown in Table 1.

To ensure the effectiveness and reliability of model training and evaluation, we employed a 5-fold cross-validation strategy with sequential sampling. In detail, we adopted a systematic approach where, for each category of data, every fourth sample was selected as part of the test set. This sampling process was iterated five times with samples numbered 1, 2, 3, 4, and 5 serving as the starting points for each round of test set extraction, respectively. This method contributes to maintaining data order, balancing class distribution, and reducing potential evaluation biases.

Table 1. Description of the ShipsEar dataset.

Category	Quantity	Sample Length
A	345	5 s
B	235	5 s
C	785	5 s
D	395	5 s
E	188	5 s

The DS3500 dataset expands the ship-radiated noise data through the ocean channel using ray theory. First, the sound field environment is modeled, the sound field in this environment is calculated according to the BELLHOP model [61], and then, the sonar-received signals—after passing through the ocean's acoustic channel—are simulated based on the calculation results of the sound field.

The specific scenario is as follows: The sonar platform is located at 17.17°N latitude and 114.22°E longitude in the deep-sea area north of the Zhongsha Islands with a sea depth of approximately 3500 m. We assume that the target is located 55° due north of the sonar platform, moving at a distance of 1–11 km away from the sonar platform, and its draft depth is 10 meters.

The distance between the target and the sonar varies from 1 km to 11 km at intervals of 2 km, and the sonar depth varies from 100 m to 1100 m at intervals of 200 m. We simulated the 5 s ShipsEar data according to the order of the file names to generate an augmented dataset at 36 locations. The size of the augmented dataset is the same as that of the source dataset. This was carried out to avoid overly large datasets that could result in slow training.

The sound-speed profile is obtained using an empirical sound-speed profile formula based on the temperature data from the World Ocean Database 2018 (WOA18). The seabed's topography is approximated as a flat bottom. According to historical sediment survey data, the seabed's parameters are as follows: sound speed of 1601.9 m/s, density of 1.7 g/cm^3, and attenuation coefficient of $0.39f^{1.71}$ dB/m, where the unit of frequency f is kHz. The sound-speed profile at the sonar platform's location is shown in Figure 2, representing a typical non-waveguide channel.

The process [12] of generating the realistic scenarios dataset is as follows:

1. Extract 5 s audio segments from the original audio files.
2. Read the .wav data from the ship-radiated noise and then use the Fourier transform $S(f)$ to convert it to the frequency-domain representation.

$$S(f) = \int_{-\infty}^{\infty} s(t) \exp(-j2\pi f t) dt, \tag{1}$$

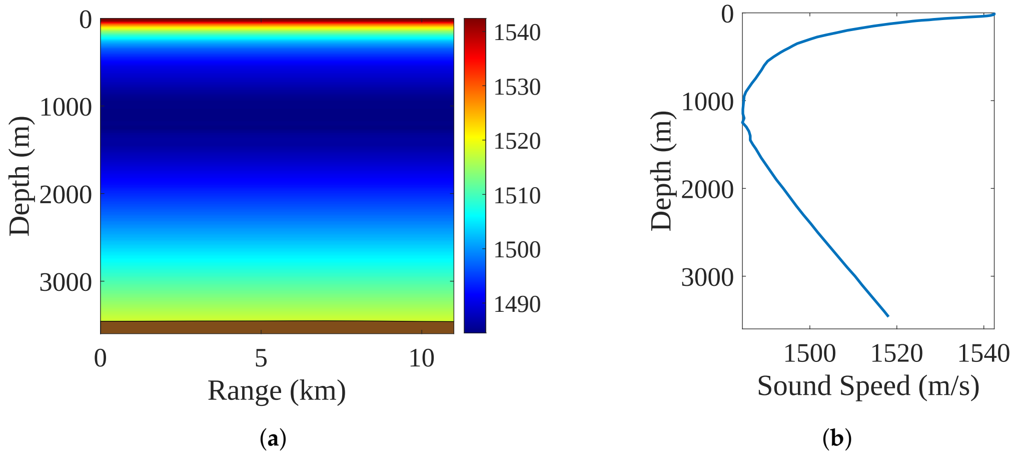


Figure 2. Sound-speed profile and seafloor topography: (**a**) 2D sound-speed profile and seafloor topography; (**b**) 1D sound-speed profile at the sonar position.

3. Calculate the arrival structure from the target to the sonar, including the amplitude A_n and delay information τ_n of each arrival path. To reduce the impact of the overall signal shift, we use the time delay of the sound ray corresponding to the maximum amplitude to perform reverse translation on the delay information τ_n.

$$\tau_n = \tau_n - \tau_{\arg\max_{i=1,2,...} A_i}, \tag{2}$$

4. Superimpose the multi-path effect on the signal by multiplying the frequency components in the frequency domain, as shown in Formula (3).

$$Y(f) = \sum_{n=1}^{N} S(f) A_n \exp(j2\pi f \tau_n), \tag{3}$$

where A_n and τ_n represent the amplitude and delay contributed by the n-th sound ray, respectively. $Y(f)$ represents the spectrum of the signal after passing through the ocean acoustic channel.

5. Perform the inverse Fourier transform on the frequency-domain output to obtain the time-domain output $y(t)$, and normalize the output to a .wav file.

$$y(t) = \int_{-\infty}^{\infty} Y(f) \exp(j2\pi f t) df. \tag{4}$$

The maximum-value normalization formula is

$$y(t) = \frac{0.9 \times y(t)}{\max(|y(t)|)}. \tag{5}$$

This formula is applied considering the storage limitations of .wav files. Since .wav files cannot store values with an absolute value greater than 1, to avoid amplitude clipping distortion during storage, we adopt the maximum-value normalization method. The factor of 0.9 is set as an engineering choice, providing a safety margin while making full use of the storage dynamic range of .wav files, ensuring that the normalized data remain within a suitable range for storage without exceeding the upper limit.

> Note: The dataset has been uploaded to two platforms. Interested researchers can download it from https://modelscope.cn/datasets/qianpeng897/DS3500 (accessed on 24 August 2025) or https://huggingface.co/datasets/peng7554/DS3500 (accessed on 24 August 2025). If the link is inaccessible, please use the alternative URL provided after 'or'—this alternative link has been verified to be accessible without regional restrictions.

By simulating the multi-path structure in the underwater acoustic channel, the impact on the characteristics of the target-radiated noise and recognition is analyzed.

The characteristics of the spectrum, time domain, and Log-Mel spectrum of the original signal and the signal passing through the ocean acoustic channel are analyzed. Figures 3 and 4 show the time-domain, frequency-domain, and Log-Mel spectrum of the original signal and the signal passing through the ocean acoustic channel, respectively.

The figures illustrate the distortion of target features that occurs after they pass through the ocean channel. Briefly, it will have a comb-filtering effect on the target noise, and there is a significant difference in the comb-filtering interval between the direct-arrival zone and the non-direct-arrival zone. The following sections explain this phenomenon based on the acoustic-ray interference analysis in the direct-arrival and non-direct-arrival zones.

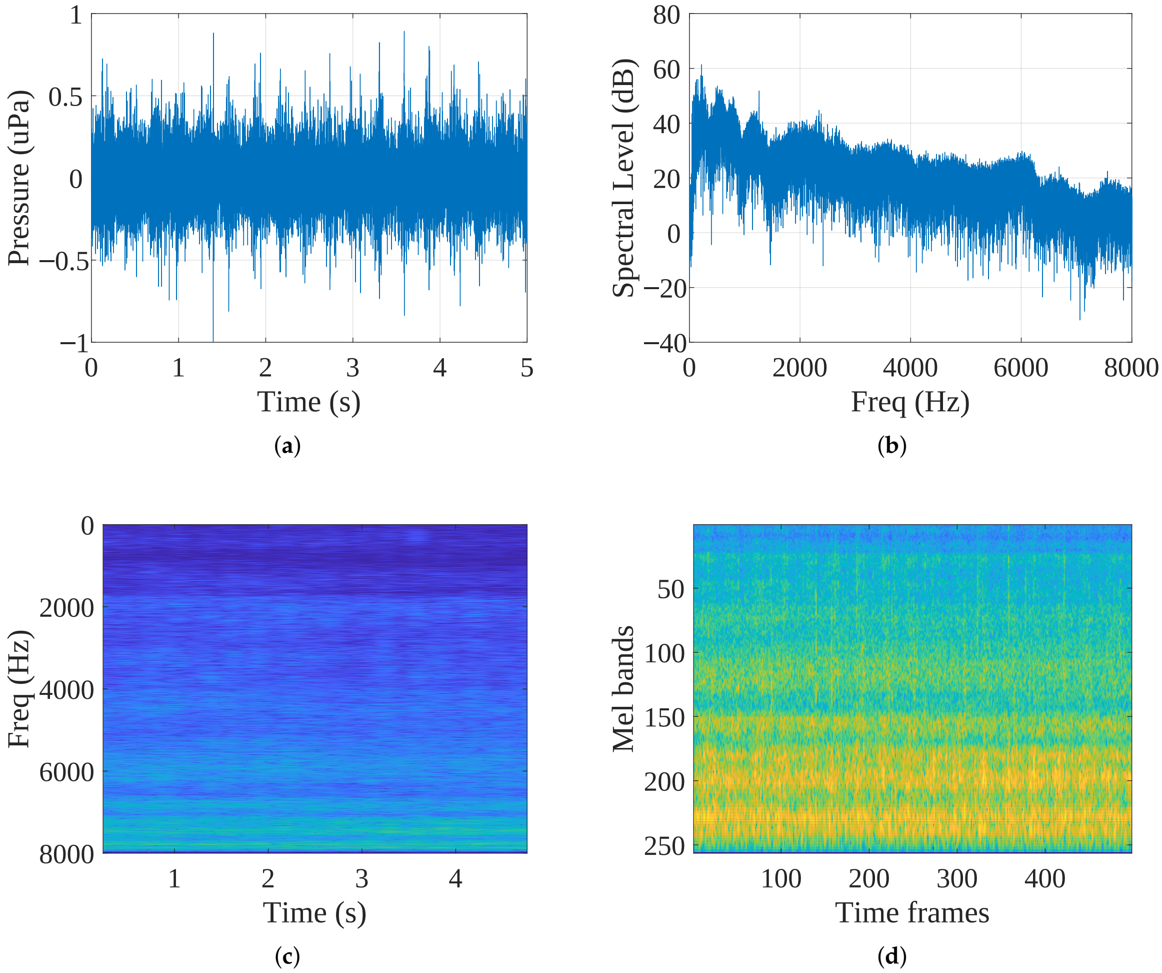


Figure 3. Spectrum analysis of the original signal: (**a**) original signal waveform; (**b**) Fourier transform result of the original signal; (**c**) LOFAR spectrum of the original signal; (**d**) Log-Mel spectrum of the original signal (yellow represents high energy, and blue represents low energy).

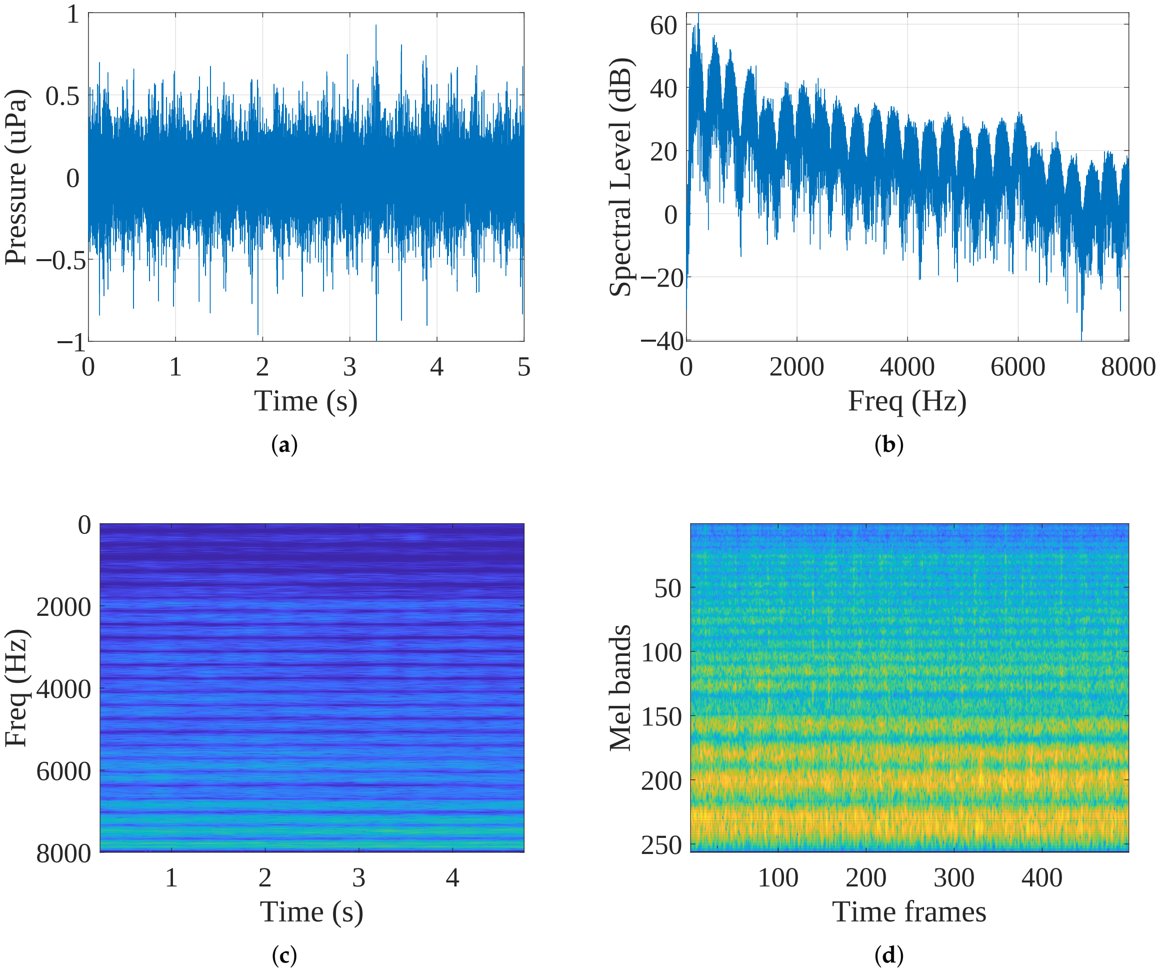


Figure 4. Spectrum analysis of the signal passing through the ocean acoustic channel. Sonar depth of 200 m and target distance of 1 km: (**a**) signal waveform after passing through the ocean acoustic channel; (**b**) Fourier transform result of the signal after passing through the ocean acoustic channel; (**c**) LOFAR spectrum of the signal after passing through the ocean acoustic channel; (**d**) Log-Mel spectrum of the signal after passing through the ocean acoustic channel.

3.1.2. Analysis of Interference Between Direct and Reflected Sound Paths

In underwater acoustic research, one of the main reasons for the generation of comb-filtering is the coherence between the direct sound ray and the top-reflected sound ray. Assuming that the sound speed in the medium is constant—that is, ignoring the refraction effect of the sound-speed gradient—there is a sound-path difference and a phase difference between the direct sound ray and the top-reflected sound ray reaching the receiver.

As shown in Figure 5, the propagation path lengths of the direct sound ray (red) and the sea-surface-reflected sound ray (green) are different, resulting in a difference in propagation distance. Due to the presence of the top reflection, there is also a phase difference of π.

Let the signal received from the direct sound ray be $h(t)$, and let the signal received from the reflected sound ray be $-h(t+\tau)$, where $\tau = \delta_x / c$ represents the time delay caused by the sound-path difference, and the "-" sign indicates the phase reversal caused by the sea-surface reflection. The total received signal $y(t)$ is as follows:

$$y(t) = h(t) - h(t+\tau). \tag{6}$$

According to the Fourier transform relationship,

$$\begin{aligned} H(\omega) &= \int h(t)\exp(-j\omega t)dt \\ -H(\omega)\exp(j\omega\tau) &= -\int h(t+\tau)\exp(-j\omega t)d\omega \end{aligned}. \tag{7}$$

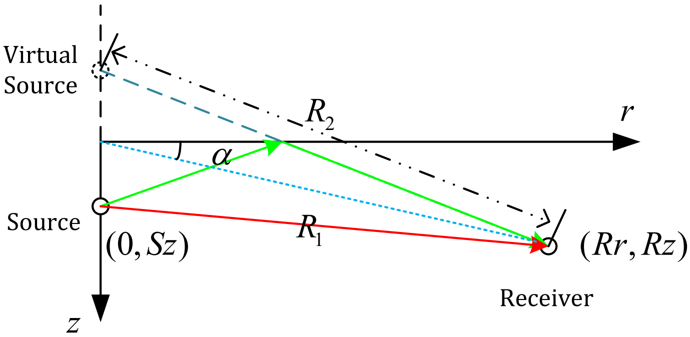


Figure 5. Schematic diagram of the direct sound ray (red) and the top-reflected sound ray (green).

It can be obtained that

$$Y(\omega) = \int_{-\infty}^{\infty} y(t)\exp(-j\omega t)dt = H(\omega)[1-\exp(j\omega\tau)]. \tag{8}$$

Therefore, when $1-\exp(j\omega\tau)=0$, $Y(\omega)=0$:

$$\omega\tau = 2n\pi \Rightarrow \delta_f = \frac{1}{\tau} = \frac{c}{\delta_x}, \tag{9}$$

where δ_f is the frequency interval (filter period) of the comb filter with the unit of Hz. In addition, using geometric relations, we can derive

$$\begin{aligned} \delta_x &= R_2 - R_1 \\ &= \sqrt{(R_z+S_z)^2+R_r^2} - \sqrt{(R_z-S_z)^2+R_r^2} \\ &= \frac{4R_zS_z}{\sqrt{(R_z+S_z)^2+R_r^2}+\sqrt{(R_z-S_z)^2+R_r^2}} \end{aligned}. \tag{10}$$

In a typical setting, given the source depth, receiver depth, and distance, with $R_r > R_z$ and $R_r \gg S_z$,

$$\delta_x \sim \frac{2R_zS_z}{R_r}. \tag{11}$$

Thus,

$$\delta_f = \frac{c}{\delta_x} = \frac{c\cdot\left(\sqrt{(R_z+S_z)^2+R_r^2}+\sqrt{(R_z-S_z)^2+R_r^2}\right)}{4R_zS_z} \sim \frac{cR_r}{2R_zS_z}. \tag{12}$$

This indicates that the frequencies of the comb filter approximately appear at multiples of $\frac{cR_r}{2R_zS_z}$, which is inversely proportional to the source and receiver depths S_z and R_z and directly proportional to the receiver range R_r.

3.1.3. Analysis of Interference Between Reflected Sound Rays

In deep-sea scenarios, due to the presence of a sound-speed gradient, the energy of direct sound rays is weak or even non-existent in certain regions. At this point, the coherence between the source–bottom-receiver path and the source–bottom-receiver–top-receiver path becomes dominant.

According to Figure 6, by using geometric relations, we can obtain

$$\begin{aligned}\delta_x &= R_2 - R_1 \\ &= \sqrt{(D-S_z+R_z)^2+R_r^2} - \sqrt{(D-S_z-R_z)^2+R_r^2}\ . \\ &= \frac{4R_zS_z}{\sqrt{(D-S_z+R_z)^2+R_r^2}+\sqrt{(D-S_z-R_z)^2+R_r^2}}\end{aligned} \tag{13}$$

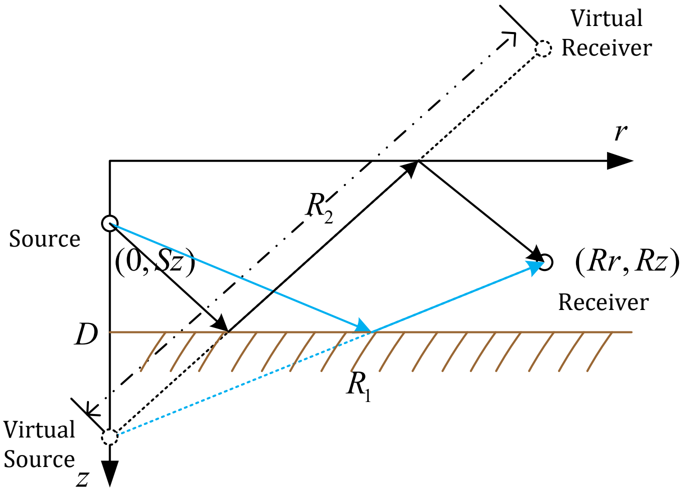


Figure 6. Schematic diagram of the bottom-reflected (blue) and bottom–top-reflected sound rays (black).

Since $R_r > R_z$ and $R_r \gg S_z$, we have the following:

$$\delta_x \sim \frac{2R_zS_z}{R_r}. \tag{14}$$

$$\begin{aligned}\delta_f &= \frac{c}{\delta_x} \\ &= \frac{c\cdot(\sqrt{(2D-S_z+R_z)^2+R_r^2}+\sqrt{(2D-S_z-R_z)^2+R_r^2})}{4R_zS_z}. \\ &\sim \frac{cR_r}{2R_z(2D-S_z)}\end{aligned} \tag{15}$$

This indicates that the frequencies of the comb filter approximately appear at multiples of $cR_r/[2R_z(2D-S_z)]$, and the filter's period is inversely proportional to $2R_z(2D-S_z)$ and directly proportional to the receiver range R_r.

As shown in Figure 7, in the ocean acoustic channel, when the sonar depth is 100 m, the received sound intensity changes with the target range and target frequency, as shown in Figure 6. From the figure, it can be observed that the frequency of the interference fringes increases with the range. The abscissa represents the range—ranging from 1 km to 11 km—and the ordinate represents the frequency—ranging from 100 Hz to 300 Hz. As the range gradually increases from 1 km to 11 km, the frequency corresponding to the interference fringes also gradually increases from a lower value to a higher value, indicating that the farther the range, the higher the frequency of the interfering sound waves.

In the deep-sea environment, when sound waves propagate, an interference structure is formed due to the superposition of sound waves from different paths. When sound waves are emitted by a sound source, part of them propagate in the form of direct sound, and part of them propagate after being reflected by boundaries, such as the seabed and the sea surface. When these sound waves from different paths meet in space, if their phase relationship meets certain conditions, an interference phenomenon occurs. The result of the interference is that the sound waves are enhanced in some positions and weakened in others, forming a complex interference pattern, which often appears as interference fringes in the frequency-range graph.

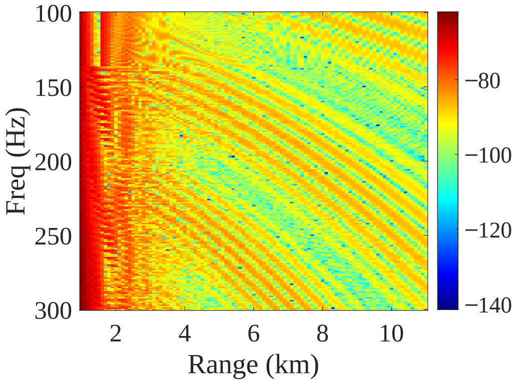


Figure 7. Deep-sea interference structure at a sonar depth of 100 m (with the application of acoustic-field reciprocity).

However, due to the relatively narrow frequency range of only 100 Hz to 300 Hz, the interference fringes are not obvious in the direct-arrival zone. The direct-arrival zone usually refers to the area where the sound waves emitted by the sound source reach the receiving point without reflection. In this area, the interference effect of the sound waves may be masked by the dominant role of the direct sound, or due to the limitation of the frequency range, clearly observing the interference fringes is difficult. This phenomenon is relatively common in deep-sea acoustic research. When analyzing relevant data, it is necessary to comprehensively consider various factors, such as the frequency range, propagation distance, and reflection and refraction of sound waves.

Furthermore, the above derivation of ray interference is actually closely related to the theory of the waveguide invariant β [62]. This invariant is commonly employed to characterize the acoustic interference patterns within waveguides.

$$\beta \equiv \frac{r}{\omega}\frac{d\omega}{dr} = -\frac{d(1/v)}{d(1/u)}, \tag{16}$$

where r is the range from the sound source, ω is the angular frequency, and u and v are the group velocity and phase velocity of the relevant acoustic modes, respectively.

From Equation (16), it can be seen that there is a close relationship between the comb-filter frequency and β. When other parameters remain relatively stable, as the range R_r increases, the comb-filter frequency interval δ_f will increase accordingly; that is, the comb-filter frequency will increase with the increase of the range. This result corroborates the above theoretical derivation process and also reflects the important role of the waveguide invariant β in explaining acoustic interference phenomena.

3.2. MEG Model

3.2.1. Multi-Task Mixture-of-Experts Model

The multi-task mixture-of-experts (MTM) is a multi-task learning method based on the mixture-of-experts (MoE) model [7]. It combines multiple expert models to achieve parameter sharing for different tasks, thereby improving the generalization ability and performance of the model.

The basic idea of MTM is to combine multiple expert models with each expert model responsible for handling different tasks. During the training process, each expert model has its own parameters. During testing, the input data are assigned to different expert models, and each expert model processes the input data according to its own parameters to finally obtain the outputs of different tasks.

As shown in Figure 8, the design idea and framework of the multi-task expert model MEG (multi-task, multi-expert, multi-gate) proposed in this paper mainly revolve around multi-task learning. By combining expert networks and gating networks, it realizes adaptive weighted processing for different tasks.

Note: The network code is provided at https://gitee.com/open-ocean/UWTRL-MEG (accessed on 24 August 2025) or https://github.com/Perry44001/UWTRL-MEG (accessed on 24 August 2025). The trained weight file can be found at https://modelscope.cn/models/qianpeng897/UWTRL-MEG (accessed on 24 August 2025) or https://huggingface.co/peng7554/UWTRL-MEG (accessed on 24 August 2025). If the link is inaccessible, please use the alternative URL provided after ’or’—this alternative link has been verified to be accessible without regional restrictions.

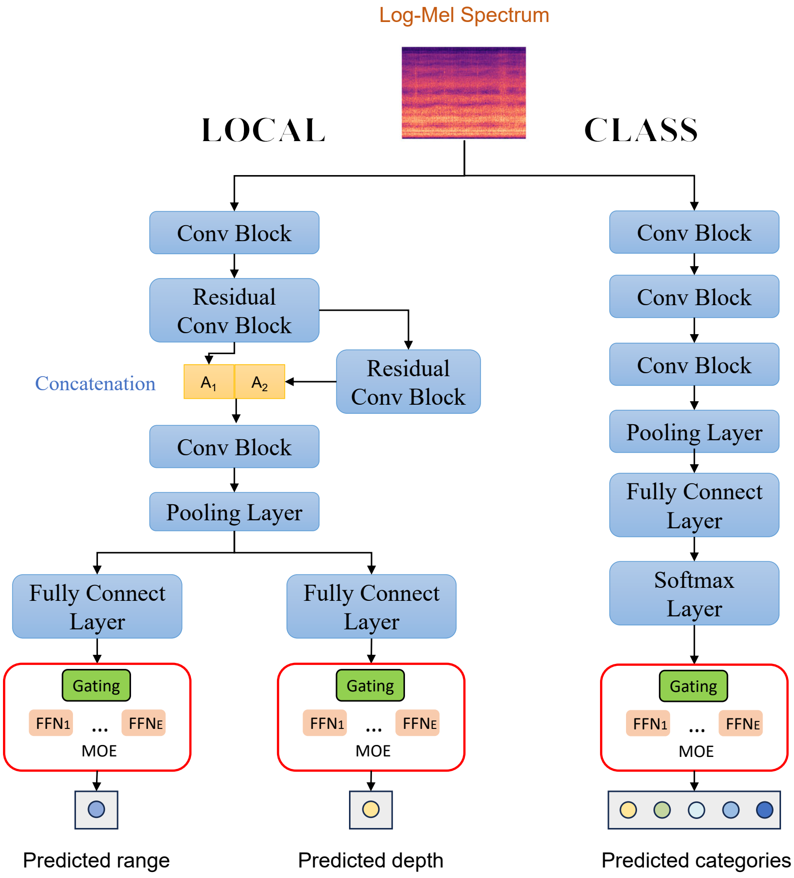


Figure 8. Multi-task mixture-of-experts MEG model.

3.2.2. Localization Task Head

The localization task head borrows the network model of MSRANET [6]. Originally, this network has proven highly effective in extracting time-frequency features for underwater acoustic recognition. We have innovatively adapted it to the task of extracting localization features, leveraging its unique architecture and feature extraction capabilities to address the specific requirements of the localization task, thereby exploring new application directions for this network in related research fields.

In Figure 8, the localization task head has two convolutional modules, which are the basic components of a convolutional neural network. By sliding the convolutional kernel on the input data for convolutional operations, local features in the data can be extracted. For example, in audio data, feature patterns at different frequencies and times can be extracted.

There are two residual convolutional modules shown in the figure. The introduction of residual connections can effectively alleviate the problem of gradient vanishing when the network depth is increased, allowing the model to learn more complex features. The branch structure therein can retain the features learned by previous layers and simultaneously enhance the model's ability to learn new features.

Between the two residual convolutional modules, there is a connection operation that splices the features from different paths (A1 and A2 in the figure), fusing feature information at different levels and paths to provide a more abundant feature representation for subsequent layers.

The pooling layer is usually used for downsampling data, reducing data dimension, and decreasing the computational load. It can also prevent overfitting to a certain extent. It retains important features and filters out some unimportant details by performing operations such as taking the maximum or average value of a local area. Through the fully connected layers of the range and depth tasks, the features are further compressed and output to the mixture-of-experts model to prepare for the final range and depth localization.

3.2.3. Classification Task Head

The classification task head also draws on the network of MSRANET [6]. Overall, the classification task head has a simpler structure compared to the localization task head. It replaces the residual modules with 2 convolutional modules. After passing through the pooling layer and the fully connected layer, the feature vector is flattened, and then the classification head features are obtained through the softmax layer. Finally, the features are output to the mixture-of-experts model to prepare for the final classification task.

3.2.4. Mixture-of-Experts Model

The expert network consists of a set of linear transformations. Each linear layer maps the input feature vector to the category space or localization space (with a dimension of N_c or 1). The number of expert networks (denoted as N_e) is set to 5.

We introduce a linear mapping to construct the gating network. This gating network takes a one-dimensional feature vector as input, and its function is to calculate the weights of each expert network for different tasks. The output of the gating network is the weights of each expert network, which are normalized by the softmax function. After normalization by the softmax function, the weight values of each expert network corresponding to the current input are obtained. On this basis, the top-k method is used to select 3 expert networks with the largest weight values—that is, 3 expert networks that are most closely related to the current input data are selected from a total of N_e expert networks.

The details of MEG are shown in Table 2.

Table 2. Detailed network architecture of the model.

Module	Detailed Network Architecture
Task Model	Localization Head (Input Feature Dimension = batchsize × feature height × feature width, Output Feature Dimension = batchsize × 192) Recognition Head (Input Feature Number = batchsize × feature height × feature width, Output Feature Dimension = batchsize × 192)
Expert Layer	Fully Connected Layer (Input Channel Number = 192, Output Channel Number for Localization Task = 1, Output Channel Number for Recognition Task = 5), with the number of 'num of experts' (5 here)
Gating Layer	Linear (Input Feature Number = 192, Output Feature Number = Number of Experts)
Loss Function	Multi-Task Loss Function

Note: "num of experts" represents the number of experts, which is a hyperparameter, and it is set to 5 in this case. The term "Linear" represents a linear layer; and "Flatten" represents a flattening layer.

3.2.5. Multi-Task Loss Function

Determining the multi-task loss function is a crucial step in model training. It comprehensively considers the loss calculations of the target category and the relative position between the target and the sonar. For the target localization loss, the mean squared error (MSE) loss function is used to measure the difference between the true and predicted values of the relative position between the target and the sonar. The mean squared error loss function calculates the average of squared differences between the predicted position coordinates (range and depth) of the target and the true position coordinates to evaluate the localization accuracy of the range and depth.

Its calculation formulas are shown below:

$$L_r = \frac{1}{N}\sum_{i=1}^{N}(r_i^{pred} - r_i^{true})^2, \tag{17}$$

$$L_z = \frac{1}{N}\sum_{i=1}^{N}(z_i^{pred} - z_i^{true})^2, \tag{18}$$

where N is the number of samples, (r_i^{pred}, z_i^{pred}) denotes the predicted values of the target range and sonar depth of the i-th sample, and (r_i^{true}, z_i^{true}) denotes the true values of the target range and sonar depth of the i-th sample. By minimizing the mean squared error loss, the prediction results can approach the true values as closely as possible, improving the accuracy of localization.

The loss of the target category confidence is calculated using the cross-entropy loss function. The cross-entropy loss function is used to measure the difference between the predicted category probability distribution and the true category label, and it can effectively reflect the accuracy of the classification task. In multi-category classification, for each sample, the true category label can be represented as a one-hot vector—that is, only the position corresponding to the true category is 1, and the other positions are 0. The predicted category probability distribution is a probability vector obtained through the softmax function of the category prediction branch output. The calculation formula of the cross-entropy loss function is as follows:

$$L_c = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}^{true}\log(y_{ij}^{pred}), \tag{19}$$

where N is the number of samples, C is the number of categories, y_{ij}^{true} is the true label of the i-th sample in the j-th category, and y_{ij}^{pred} is the predicted probability of the i-th sample in the j-th category. By minimizing cross-entropy loss, the predicted category probability distribution of the model can be made to approach the true category label as closely as possible, improving the accuracy of target recognition.

Considering the losses of the above 3 tasks, the multi-task loss function L can be expressed as follows:

$$L = w_r L_r + w_z L_z + w_c L_c, \tag{20}$$

where w_r, w_z, and w_c are the weight coefficients of the target range localization loss, sonar depth localization loss, and target category confidence loss, respectively. These weight coefficients can be adjusted according to the importance of the tasks and the characteristics of the data. For example, if the accuracy of target localization is more important in practical applications, the values of w_r and w_z can be appropriately increased; if the accuracy of target recognition is the key factor, the weight of w_c can be increased.

To address the differences in variance and offset among single-task losses, instead of manually adjusting c_τ, the relevant coefficients are incorporated into the learnable network parameter $\omega_T = (\theta_T, c_T)$. However, to prevent the model from obtaining trivial solutions, a regularization term $R(c_\tau)$ needs to be added to the combined loss L_{comb}. Based on the $R_{log}(c_\tau) = \log(c_\tau^2)$ proposed by Kendall et al., a slight modification is carried out [63], adjusting it to $R_{pos}(c_\tau) = \ln(1 + c_\tau^2)$ to ensure that the regularization value is always positive. This measure prevents the generation of negative loss values when c_τ is reduced to $c_\tau^2 < 1$ [64]. These factors are considered in the final multi-task loss function:

$$L_{\mathcal{T}}\left(x, y_{\mathcal{T}}, y'_{\mathcal{T}}; \omega_{\mathcal{T}}\right) = \sum_{\tau \in \mathcal{T}} \frac{1}{2 \cdot c_\tau^2} \cdot L_\tau\left(x, y_\tau, y'_\tau; \omega_\tau\right) + \ln\left(1 + c_\tau^2\right) \tag{21}$$

3.2.6. Evaluation Criteria

To better evaluate the recognition task, we compare the model's prediction results with the true labels and use accuracy, precision, recall, and $F1$ score as performance indicators. The specific calculation methods are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{22}$$

$$Precision = \frac{TP}{TP + FP}, \tag{23}$$

$$Recall = \frac{TP}{TP + FN}, \tag{24}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{25}$$

where TP represents the number of true positives—that is, the number of samples that are actually positive and predicted as positive; FP represents the number of false positives—that is, samples that are actually negative but predicted as positive; FN represents the number of false negatives—that is, samples that are actually positive but predicted as negative; and TN represents the number of true negatives—that is, samples that are actually negative and correctly predicted as negative.

When calculating the average value of each target-type indicator, the macro-calculation method is used to treat each category equally. The calculation method of each indicator is as follows:

$$P_{macro} = \frac{1}{n} \sum_{i=1}^{n} P_i, \tag{26}$$

$$R_{macro} = \frac{1}{n} \sum_{i=1}^{n} R_i, \tag{27}$$

$$F1_{macro} = 2 \times \frac{P_{macro} \times R_{macro}}{P_{macro} + R_{macro}}, \tag{28}$$

where i is the target category, n represents the total number of categories, and P_i and R_i represent the precision and recall of the i-th category, respectively.

The indicators for the localization task are the mean absolute error (MAE) and the mean absolute percentage error (MAPE), and they are defined as follows:

$$E_{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \widehat{y}_i|, \tag{29}$$

$$E_{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \frac{|y_i - \widehat{y}_i|}{y_i}, \tag{30}$$

where y represents the true value of the distance or depth, $\widehat{y}$ is the predicted value, and N is the number of data samples.

In the context of visualization, an accuracy rate of 100% represents the best-case scenario, while a mean absolute error (MAE) of 0 is considered ideal. To facilitate intuitive comparison and analysis within the same visualization framework, we convert the MAE into a scored MAE. Specifically, we transform the MAE, which ranges from its maximum value (MAX) to 0, into a score within the 0–100% range. The conversion method is defined as follows:

$$P_{MAE\text{-}R} = \left(1 - \frac{E_{MAE\text{-}R}}{R_{MAX}}\right) \times 100\%, \tag{31}$$

$$P_{MAE\text{-}D} = \left(1 - \frac{E_{MAE\text{-}D}}{D_{MAX}}\right) \times 100\%. \tag{32}$$

Here, $E_{MAE\text{-}R}$ and $E_{MAE\text{-}D}$ denote the mean absolute error (MAE) values associated with the parameters "Range" and "Depth", respectively. R_{MAX} and D_{MAX} represent the maximum possible values of the "Range" and "Depth" parameters. $P_{MAE\text{-}R}$ and $P_{MAE\text{-}D}$ are the converted percentages, which are analogous to the accuracy rate.

3.2.7. Training Parameter Settings

This research employed the PyTorch 2.4.1 deep-learning framework to configure the training environment. The initial learning rate was initialized to 0.001 with a 20-epoch linear warm-up. During the training phase, the cosine annealing learning rate adjustment strategy was adopted. When the rectified linear unit (ReLU) served as the activation function, the He initialization method [65] was applied to initialize network parameters.

The batch size was set to 64, and the cross-entropy function was adopted as the loss function for network training. To prevent overfitting and avoid local minima, the adaptive moment estimation (Adam) optimization algorithm [66] was utilized to iteratively optimize the model's parameters. The exponential decay rates of the first-moment and second-moment estimates were set to 0.9 and 0.999, respectively. Table 3 lists the detailed configurations of the network training hyperparameters.

Table 3. Hyperparameters of the proposed model.

Parameter	Value
Epochs (training rounds)	150
Batchsize (batch size)	64
Learning rate	0.001
Optimizer	Adam
First-moment estimates	0.9
Second-moment estimated exponential decay rate	0.999
Initialization method	He Initialization
Loss function	Multi-task loss

4. Results

4.1. Training and Test

We used the features extracted by the Short-Time Fourier Transform (STFT) as the input and utilized the DS3500 device to conduct 150 epochs of training on the MEG network. During the training process, we continuously recorded key parameters such as training loss and classification accuracy, and we visualized these data to more clearly observe the training dynamics and performance of the model.

Figure 9 depicts the loss curve, recognition accuracy, normalized range and depth localization error throughout the training process. As is evident, the loss curve steadily declines during training. This downward trend demonstrates that the network is effectively learning and making progress.

Simultaneously, the recognition accuracy steadily increases. This indicates that the network is successfully capturing relevant features, resulting in improved performance on the recognition task. Concurrently, both the range and depth localization errors gradually decrease. This improvement shows that the network can accurately estimate positions, signifying its effectiveness in the localization task.

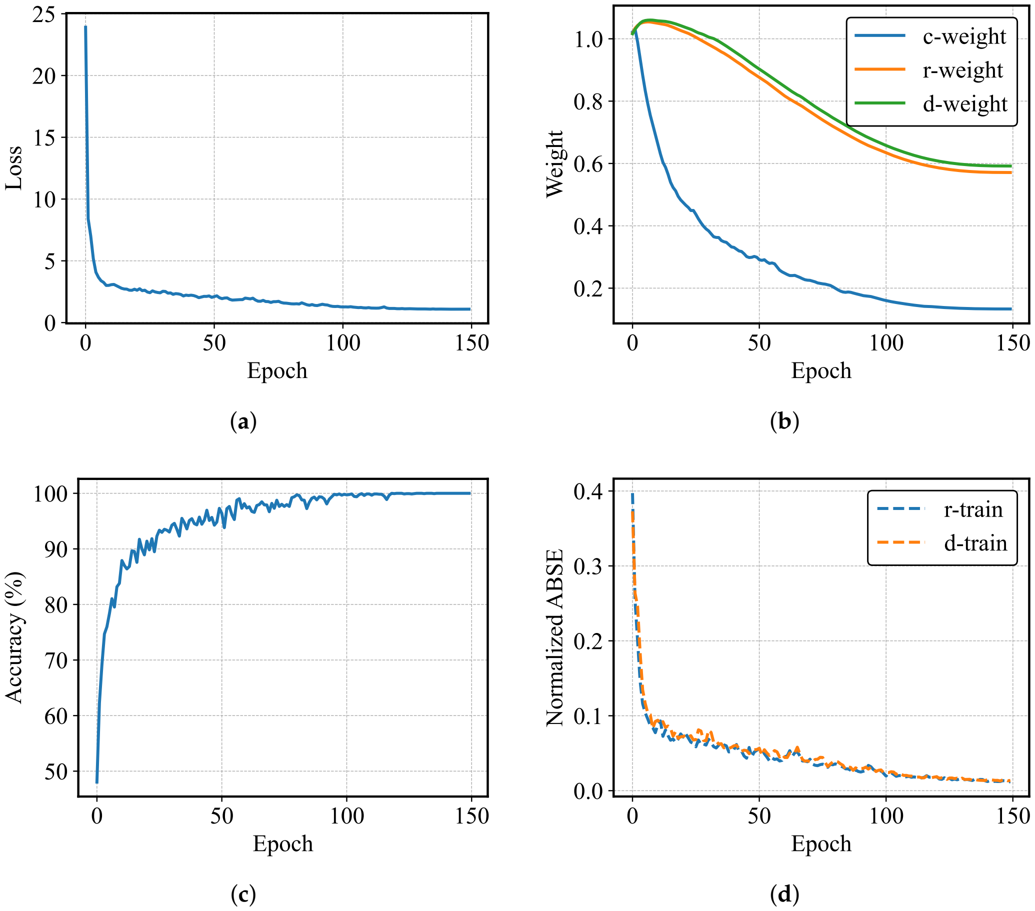


Figure 9. Training process. (**a**) Loss curve. (**b**) The trends in the weights of different tasks. (**c**) Classification accuracy. (**d**) Normalized range and depth localization MAE.

Upon completion of the training process, the network performance was evaluated using the test set. Figure 10 illustrates the receiver operating characteristic (ROC) curves and the confusion matrix. Figure 10a shows ROC curves with the AUC (area under curve) for each class, reflecting outstanding discriminative ability for every class. Figure 10b presents the confusion matrix, with the vertical axis as true labels and the horizontal axis as predicted labels, clearly depicting the classification performance for each class.

Figures 11 and 12 present the classification accuracies for the five types of tasks after the MEG network has been trained along with the range and depth localization results on the first 36 samples. The bar chart shows that there are high class accuracies (with accuracies greater than 0.9 for all classes). The 4th class achieves the highest accuracy, reaching 100%, while the 1st class has the lowest accuracy, at 91.67%. The range and depth plots demonstrate that the predicted values (red dashed lines) closely track the true values (blue solid lines), indicating that the localization network can accurately estimate the range and depth.

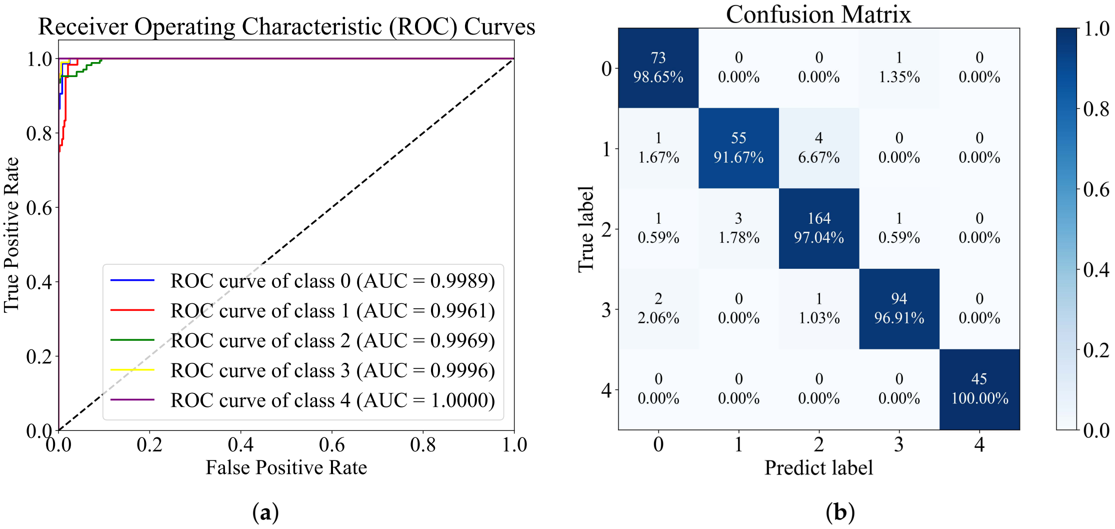


Figure 10. Receiver operating characteristic (ROC) curves and confusion matrix. (**a**) ROC curves for each class. (**b**) Confusion matrix with vertical-axis as true labels and horizontal-axis as predicted labels.

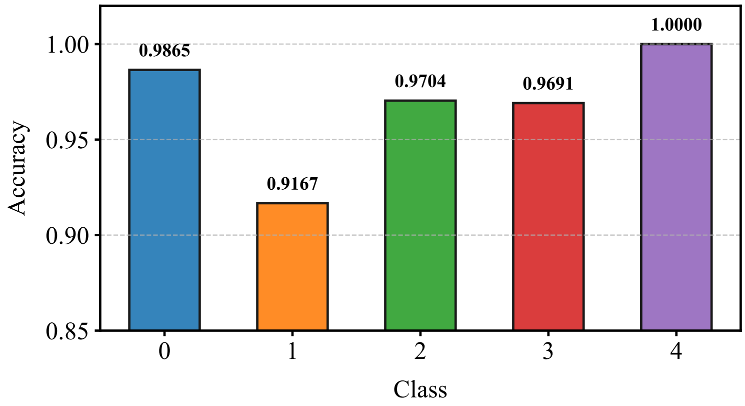


Figure 11. Classification accuracy of different classes.

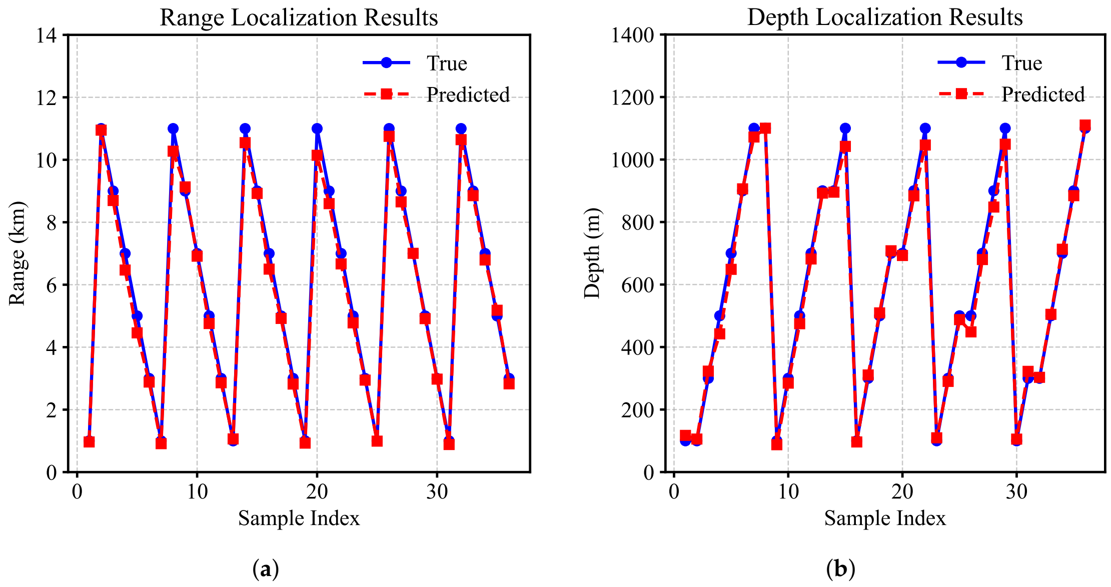


Figure 12. Range and depth localization results. (**a**) Range localization results (where the blue line represents true values and the red dashed line with squares represents predicted values). (**b**) Depth localization results (same as above).

4.2. Comparison Experiments of Different Features

To explore the impact of different feature parameters on the performance of the MCL (multi-task classification and localization network, i.e., MEG without MoE module), comparative experiments were designed and conducted. Five different feature types, namely Mel, STFT, MFCC, GFCC, and CQT, were selected for the experiment. The corresponding parameter settings for these feature types are shown in Table 4.

Table 4. Parameter settings of different feature types.

Feature Type	Number of Filter Banks	Feature Dimension
Mel	$\mathrm{n_{fft}} = 1024, \mathrm{n_{mels}} = 200, \mathrm{hop\_length} = 160, \mathrm{win\_length} = 400$	$[200, 501]$
STFT	$\mathrm{n_{fft}} = 1024, \mathrm{hop\_length} = 160$	$[513, 301]$
MFCC	$\mathrm{n_{mfcc}} = 40, \mathrm{n_{fft}} = 1024, \mathrm{hop\_length} = 160, \mathrm{n_{mels}} = 200$	$[40, 301]$
GFCC	$\mathrm{n_{mels}} = 200, \mathrm{f_{min}} = 50, \mathrm{f_{max}} = 8000$	$[200, 157]$
CQT	$\mathrm{bins\_per\_octave} = 24$	$[84, 157]$

The experimental results are summarized in Table 5 and Figure 13. Evidently, the data reveal that distinct feature parameters exert a substantial influence on the MCL network's performance. Among the evaluated features, the STFT feature emerges as the top performer in this experiment. It attains an accuracy (ACC) of 96.07%, a mean absolute error for range localization (MAE-R) of merely 0.26 km, and a mean absolute error for depth localization (MAE-D) of 27.68 m. This outstanding performance implies that the STFT feature can proficiently extract the crucial information associated with the tasks, empowering the network to achieve remarkable results in both classification and localization tasks.

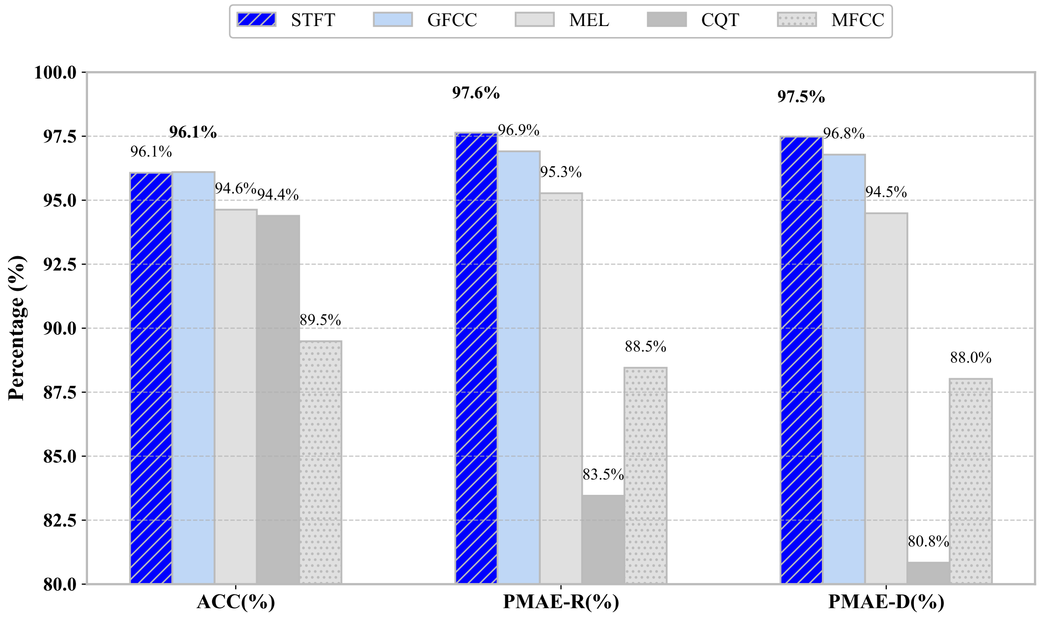


Figure 13. Comparison of five feature extraction methods on accuracy and normalized MAE (all metrics scaled to 100% as best and the bold data indicate the best results within the same group—similarly hereinafter).

The GFCC feature secures the second position. It has an ACC of 96.10%, an MAE-R of 0.34 km, and an MAE-D of 35.42 m. Even though its accuracy is marginally higher than that of the STFT feature, the STFT feature demonstrates superior performance in range and depth localization.

The Mel feature shows an ACC of 94.63%, an MAE-R of 0.52 km, and an MAE-D of 60.61 m. Although it lags slightly behind the STFT and GFCC features in localization metrics, it still sustains a relatively high level of recognition performance, suggesting its potential utility in the MCL network.

On the contrary, the MFCC and CQT features exhibit relatively suboptimal performance. The MFCC feature has an ACC of 89.49%, an MAE-R of 1.27 km, and an MAE-D of 131.78 m. The CQT feature records an ACC of 94.39%, an MAE-R of 1.82 km, and an MAE-D of 210.76 m. It is speculated that these two features may not fully capture the effective task-related information during the extraction process, which consequently impacts the network's performance to a certain extent.

Table 5. Experimental results of different feature parameters (the bold data indicate the best results within the same group—similarly hereinafter).

Feature	ACC (%)	MAE-R (km)	MAE-D (m)
STFT	96.07	**0.26**	**27.68**
GFCC	**96.10**	0.34	35.42
Mel	94.63	0.52	60.61
CQT	94.39	1.82	210.76
MFCC	89.49	1.27	131.78

4.3. Comparison Experiments of Different Network

To compare the performance of the MEG network with that of other networks, a series of comparative experiments were carried out. In this experiment, nine representative network models were selected: MEG (STFT), MCL (STFT), MEG-C (STFT), MEG-L (STFT), DenseNet121 [67], ResNet18 [68,69], MobileNetV2 [70], ResNet50 [69], and Swin-Transformer [71]. Here, 'network (STFT)' denotes that STFT is employed as the input feature of the network. MEG-C is the classification-task branch network of MEG, and MEG-L is the localization-task branch network of MEG. These two branch networks are used to compare the results between multi-task and single-task scenarios.

DenseNet121, ResNet18, MobileNetV2, ResNet50, and Swin-Transformer are all general-purpose architectures and are employed for the classification task. Given that the input feature dimension of the Short-Time Fourier Transform (STFT) is relatively large, a convolutional layer with a stride of 2 is added at the beginning of these general-purpose network frameworks to reduce the image dimension. These models vary in network depth, structural design, and parameter scale, enabling a comprehensive evaluation of their suitability for multi-task learning and single-task scenarios.

The experimental results are summarized in Table 6 and Figure 14. As indicated by the data, the network architecture significantly impacts both classification accuracy and localization precision with distinct performance patterns observed between multi-task and single-task networks.

Among the models under test, the MEG (STFT) network shines brightly, achieving a classification accuracy of 95.93%. It also demonstrates reliable localization performance, with an MAE-R reaching 0.2011 km and an MAE-D of 20.61 m. This fully reflects its ability to effectively integrate classification and localization tasks when handling multi-task learning. The MCL (STFT) network has a slightly higher classification accuracy, hitting 96.07%. However, there is a rather noticeable difference in its performance of the localization task compared to the MEG network with an MAE-R of 0.2565 km and an MAE-D of 27.68 m. This somewhat reveals the design concept of the MCL (STFT) network for multi-task scenarios. Although it shows a certain advantage in classification performance, there may still be room for optimization in terms of balancing the localization accuracy.

In comparison to networks optimized for joint localization and classification, single-task architectures designed exclusively for either localization or classification exhibit distinct performance trade-offs. Specifically, while single-task networks (e.g., MEG-C (STFT) for classification and MEG-L (STFT) for localization) achieve respectable performance—with a classification accuracy of 95.76% and localization metrics of MAE-R (mean absolute error for range) 0.2013 km and MAE-D (mean absolute error for depth) 20.79 m, respectively—the multi-task MEG network outperforms them in both effectiveness and efficiency. Notably, MEG delivers superior performance with only a single training process, demonstrating that multi-task architectures like MEG not only yield better results but also reduce computational overhead through the cross-task integration of feature extraction.

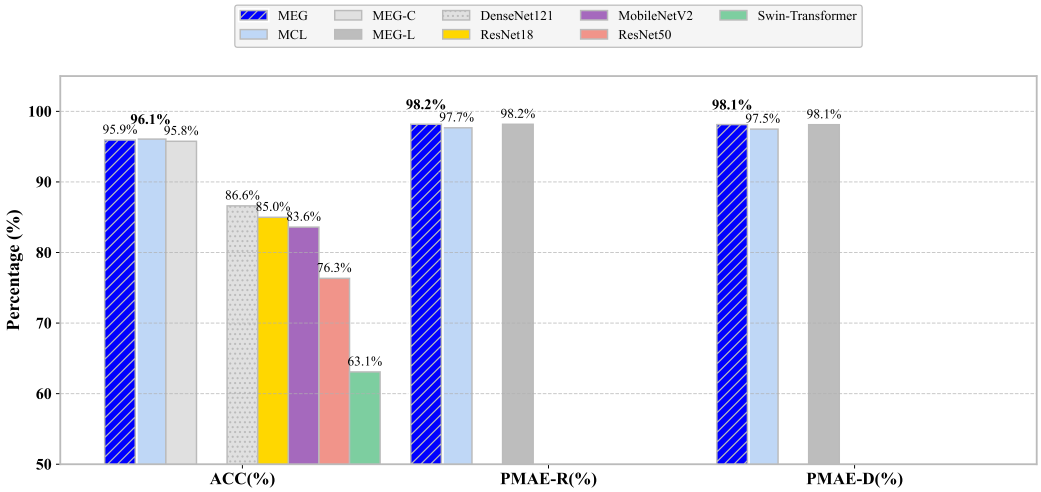


Figure 14. Comparison of different neural network models on accuracy and normalized MAE (all metrics scaled to 100% as best).

Table 6. Performance comparison of different network architectures (the bold data indicate the best results within the same group—similarly hereinafter).

Network	ACC (%)	MAE-R (km)	MAE-D (m)
MEG (STFT)	95.93	**0.2011**	**20.61**
MCL (STFT)	**96.07**	0.2565	27.68
MEG-C (STFT)	95.76	-	-
MEG-L (STFT)	-	0.2013	20.79
DenseNet121	86.61	-	-
ResNet18	84.99	-	-
MobileNetV2	83.60	-	-
ResNet50	76.34	-	-
Swin-Transformer	63.08	-	-

General architectures such as DenseNet121, ResNet18, MobileNetV2, ResNet50, and Swin-Transformer mainly focus on the classification task. Compared with MEG, their performance shows a significant gap. DenseNet121 has a classification accuracy of 86.61%, ResNet18 84.99%, MobileNetV2 83.60%, ResNet50 76.34%, and Swin-Transformer only 63.08%. These models struggle to achieve performance comparable to MEG, which is perhaps because they fail to extract useful features for classification as effectively as MEG does.

In summary, experimental results underscore that architectural design must align with task requirements. Multi-task networks outperform single-task models in overall performance, while single-task counterparts remain competitive in their specific domains.

4.4. Analysis of Network Convergence

To further understand the impact of different network architectures on the convergence process, we analyze the convergence epochs and training costs of several representative networks. The networks under consideration are MEG, DenseNet121, ResNet18, MobileNetV2, ResNet50, and Swin-Transformer. To ensure the fairness of the comparison, this study restricts its analysis to the recognition task alone.

The relevant data are presented in Table 7 and can also be visually observed from Figures 15 and 16.

Table 7. Convergence-related performance of different network architectures.

Network	ACC (%)	Params (MB)	Convergence Epoch	Training Cost (min)
MEG (STFT)	**95.93**	65.9	68	22
MEG (GFCC)	95.75	59.8	**2**	**10**
DenseNet121	86.61	53.2	31	43
ResNet18	84.99	85.3	20	17
MobileNetV2	83.60	**17.0**	44	21
ResNet50	76.34	179.4	46	32
Swin-Transformer	63.08	210.0	-	54

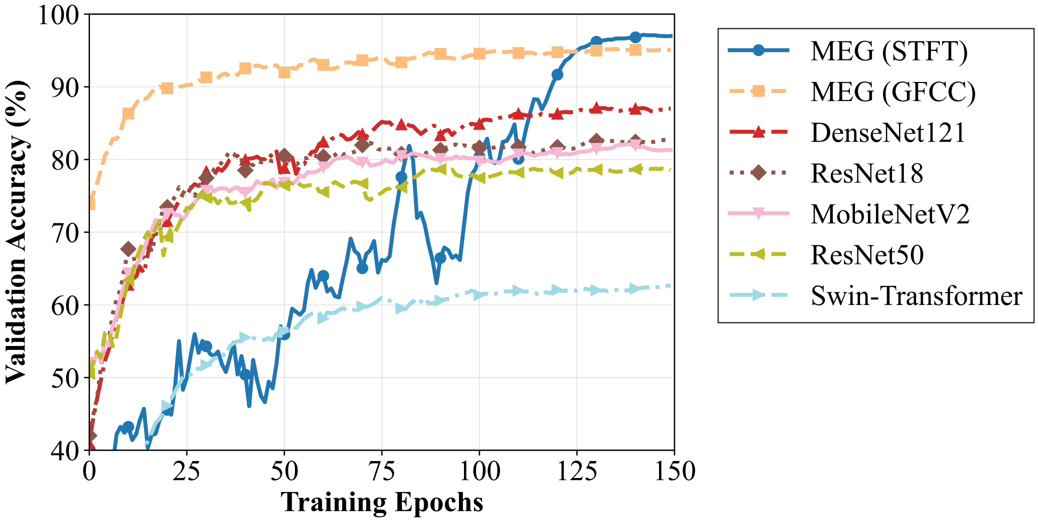


Figure 15. Convergence curves of different networks (after smoothing processing).

As can be seen from the data in the figures, there are significant differences in the convergence situations of different networks. MEG (STFT) requires 68 epochs to converge (the first epoch at which an accuracy of 80% is achieved); in contrast, MEG (GFCC) takes only 2 epochs. This improvement can be attributed to the GFCC features' reduced dimensionality in the channel domain (GFCC: 200, STFT: 513), which significantly reduces both the network size and computational complexity of convolution operations within the MEG architecture.

ResNet18 converges in 20 epochs. Its relatively simple architecture gives it an advantage in convergence speed, but when combined with the validation accuracy curve, there may be a trade-off in terms of precision. DenseNet121 converges in 31 epochs, MobileNetV2 in 44 epochs, and ResNet50 in 46 epochs. The different convergence epochs of these networks reflect the differences in the efficiency of feature learning and parameter optimization of their respective architectures.

During the training process, the validation accuracy of Swin-Transformer is always lower than 80%, and there is no convergence trend (no convergence epochs are shown), indicating that in the current task, due to factors such as poor adaptability between its architecture and the task, it is difficult to effectively improve the precision, and its cost - effectiveness is not high.

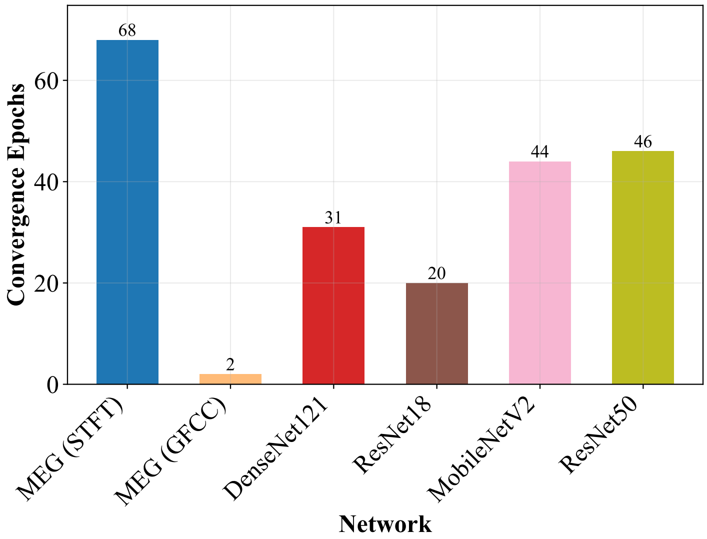


Figure 16. Convergence speed of different networks.

In terms of training time, as shown in Figure 17, MEG (GFCC) only takes 10 min, which is significantly advantageous in terms of training time consumption. Combined with its extremely fast convergence epochs, it reflects the good performance of this network in training efficiency when GFCC is used as the input. MEG (STFT) takes 22 min for training. Although the number of convergence epochs is large, the time consumption is still within an acceptable range. ResNet18 has a training duration of 17 min, which matches its relatively fast convergence, demonstrating the characteristics of a simple architecture in terms of training efficiency. DenseNet121 requires 43 min for training, MobileNetV2 takes 21 min, ResNet50 takes 32 min, and Swin-Transformer takes as long as 54 min. Swin-Transformer not only has low accuracy and no sign of convergence but also has the longest training time. Considering both training efficiency and effectiveness, it is not practical in the current task. Overall, the input features and network architecture jointly affect the number of convergence epochs, training duration, and precision performance.

The dual advantages of MEG (GFCC) in convergence efficiency and training time consumption provide a reference for network design and optimization, and they also make the differences in task adaptability among different networks more clear, which is helpful for the subsequent targeted selection or improvement of networks.

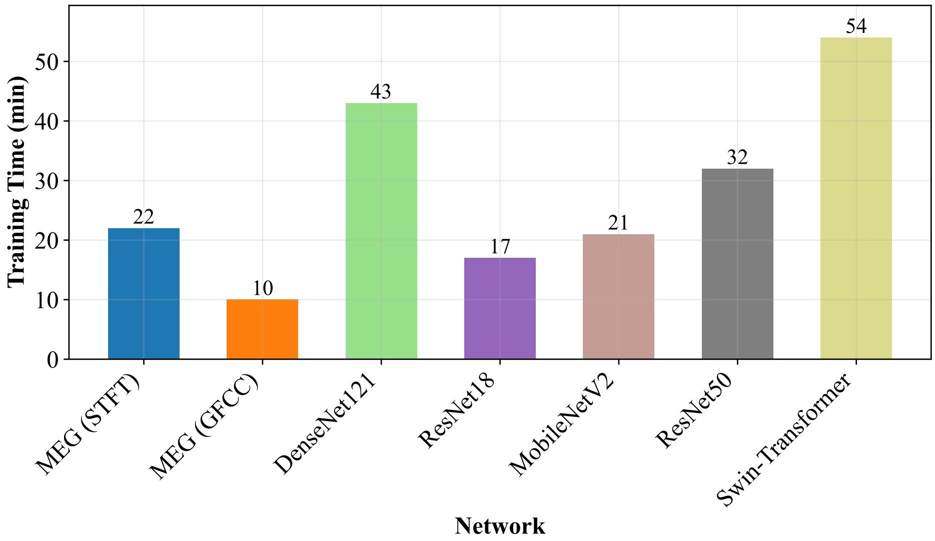


Figure 17. Training time comparison of different networks.

4.5. k-Fold Cross-Validation Results Analysis

To evaluate the generalization ability and robustness of different feature inputs combined with network architectures, we conducted k-fold cross-validation experiments. The performance metrics include classification accuracy (ACC), standard deviation (std) of accuracy, MAE-R, standard deviation of MAE-R, MAE-D, and standard deviation of MAE-D. The experimental results for MEG and MCL networks with STFT and GFCC features are summarized in Table 8 and Figure 18.

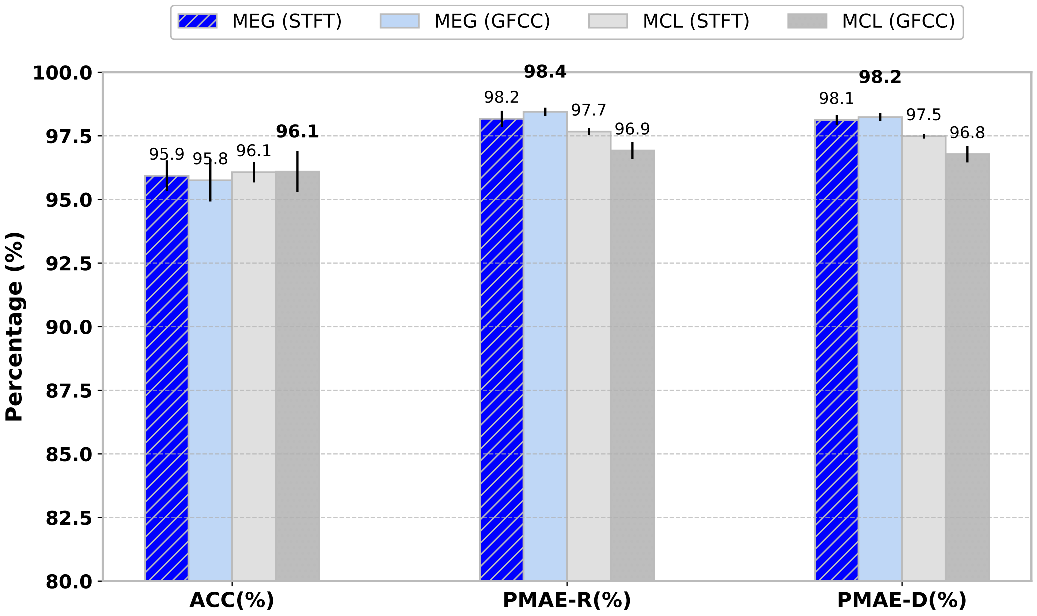


Figure 18. Comparative trends of k-fold cross-validation performance for different feature-network combination (the bold data indicate the best results within the same group).

Table 8. Detailed k-fold cross-validation performance metrics for different feature-network combinations (the bold data indicate the best results within the same group).

Model	ACC (%)	STD	MAE-R (km)	STD	MAE-D (m)	STD
MEG (STFT)	95.93	0.6067	0.2011	0.03481	20.61	2.141
MEG (GFCC)	95.75	0.8315	**0.1707**	0.01769	**19.43**	1.707
MCL (STFT)	96.07	0.4005	0.2565	0.01572	27.68	1.001
MCL (GFCC)	**96.10**	0.8039	0.3384	0.03687	35.42	3.566

Note: “ACC”, “MAE-R”, and “MAE-D” denote the average values of accuracy, range error, and depth error, respectively; “STD” represents the standard deviation of the corresponding cross-validation results.

The analysis shows that different models exhibit varying strengths in classification and localization. MCL (GFCC) achieves the highest classification accuracy at 96.10%, but its localization performance is relatively poor. MCL (STFT) outperforms MCL (GFCC) in localization, yet it has a marginal advantage in classification. MEG (GFCC) demonstrates the best localization performance, with MAE-R and MAE-D as low as 0.1707 km and 19.43 m, respectively, though its classification accuracy is slightly lower. MEG (STFT) offers a balanced performance, achieving a classification accuracy of 95.93%, localization errors of 0.2011 km and 20.61 m.

5. Discussion

The relationship between localization and recognition tasks bears resemblance to recognizing an object under diverse weather conditions [64]. In an underwater context, the ocean acoustic channel impacts the recognition features in a similar way that weather affects visual features.

The signal received by the sonar can be precisely described by the following equations. In the time domain, it is expressed as shown below:

$$y(t) = h(t) * s(t) + n(t), \tag{33}$$

where $y(t)$ denotes the received signal, $h(t)$ represents the ocean channel function, $s(t)$ is the target signal, $n(t)$ stands for the noise, and $*$ symbolizes convolution.

In the frequency domain, the equation is

$$Y(\omega) = H(\omega) \cdot S(\omega) + N(\omega). \tag{34}$$

Here, $Y(\omega)$ is the spectrum of the received signal, $H(\omega)$ is the frequency-domain representation of the ocean channel function, $S(\omega)$ is the spectrum of the target, $N(\omega)$ is the spectrum of the noise, and $\cdot$ indicates multiplication.

From these equations, it is clear that $H(\omega)$ serves for target localization, while $S(\omega)$ is utilized for target recognition. Typically, $H(\omega)$ and $S(\omega)$ exhibit no correlation.

Moreover, a conventional multi-task network aims to share network features to enhance multi-task learning effectiveness. However, in the ocean channel scenario, the features of the ocean channel and those of the target are uncorrelated. To explore this further, two shared one-dimensional convolutional layers were added to the front-end of the network. The recognition results are as follows.

Table 9 clearly shows that the MEG network with shared layers has a significantly lower recognition rate and larger localization errors, indicating its suboptimal performance in both localization and recognition tasks. Compared with other configurations, the shared-layer design proves relatively ineffective in enabling efficient feature extraction and task integration for these two tasks, failing to fully realize the original intention of the multi-task design.

Table 9. Experimental results (the bold data indicate the best results within the same group).

Network	Acc (%)	MAE-R (km)	MAE-Z (m)
MEG	**95.93**	**0.2011**	**20.61**
MEG with Shared Layers	95.35	0.2394	22.97

Nevertheless, the concept of using a multi-task network still holds promise. Although the features learned by the two tasks may seem unrelated, this approach allows us to perform both localization and recognition within a single network, making the training process more lightweight and resource-efficient. Moreover, we regard this as a foundation for future research. There may be hidden connections between the two tasks that, once discovered, could enhance overall performance.

6. Conclusions

This research centers on two persistent bottlenecks in underwater acoustic target recognition: scarce data resources and the suboptimal generalization capacity of existing models. To address these challenges, we presented a data augmentation model grounded in ray theory and the MEG (multi-task, multi-gate, multi-expert) recognition framework. Rigorous experimental evaluations were conducted to validate the effectiveness of the proposed methodologies:

1. Data Augmentation:
   Our ray-theory-based ship-radiated noise augmentation model expands the ShipsEar dataset, creating the DS3500 dataset covering 3500-meter deep-sea direct and shadow zones. The model effectively simulates sonar-captured ship-radiated noise in the ocean channel, enriching data diversity for better model training.
2. Model Design:
   The MEG framework is innovative and efficient. By incorporating the localization task, it learns relative target–sonar positions, enhancing robust pattern extraction. MEG integrates multi-expert and multi-gate mechanisms, allocating distinct parameter spaces for different underwater signals. Our task-specific expert-assignment strategy reduces task interference.
3. Experimental Validation:
   Experiments on the DS3500 dataset clearly demonstrate the superiority of the MEG model when compared with networks such as ResNet and Swin-Transformer. The MEG model not only achieves accurate target recognition but also shows excellent localization performance.

Overall, this study provides a practical solution for underwater target localization and recognition. The proposed data augmentation model and MEG framework contribute both theoretically and practically. Future research could optimize the MEG model's structure and parameters for more complex ocean conditions and integrate it with advanced techniques like reinforcement and transfer learning to expand its application.

Author Contributions: Conceptualization, P.Q. and P.X.; methodology, P.Q., Y.L., J.W. and Y.C.; software, P.Q., J.W. and Y.C.; validation, P.Q., Y.L., J.W. and Y.C.; formal analysis, P.Q., Y.L., J.W. and Y.C.; investigation, none specified; resources, P.X. and Z.L.; data curation, none specified; writing—original draft preparation, P.Q., J.W. and Y.C.; writing—review and editing, P.W., Y.D., P.X. and Z.L.; visualization, P.Q. and Y.D.; supervision, P.W. and P.X.; project administration, P.W.; funding acquisition, P.X. and Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China grant numbers U22A2012 and 12474446 and by the Guangdong Basic and Applied Basic Research Foundation grant number 2024A1515030149.

Data Availability Statement: The dataset can be downloaded from https://modelscope.cn/datasets/qianpeng897/DS3500 (accessed on 24 August 2025) and https://huggingface.co/datasets/peng7554/DS3500 (accessed on 24 August 2025). The network code is available at https://gitee.com/open-ocean/UWTRL-MEG (accessed on 24 August 2025) and https://github.com/Perry44001/UWTRL-MEG (accessed on 24 August 2025). The trained weight file can be accessed at https://modelscope.cn/models/qianpeng897/UWTRL-MEG (accessed on 24 August 2025) and https://huggingface.co/peng7554/UWTRL-MEG (accessed on 24 August 2025). If the link is inaccessible, please use the alternative URL provided after 'or'—this alternative link has been verified to be accessible without regional restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rajagopal, R.; Sankaranarayanan, B.; Rao, P.R. Target classification in a passive sonar-an expert system approach. *Proc. Int. Conf. Acoust. Speech Signal Process.* **1990**, *5*, 2911–2914.
2. Sutin, A.; Bunin, B.; Sedunov, A.; Sedunov, N.; Fillinger, L.; Tsionskiy, M.; Bruno, M. Stevens Passive Acoustic System for underwater surveillance. In Proceedings of the 2010 International WaterSide Security Conference, Carrara, Italy, 3–5 November 2010; pp. 1–6.
3. Vaccaro, R.J. The past, present, and the future of underwater acoustic signal processing. *IEEE Signal Process. Mag.* **1998**, *15*, 21–51. [CrossRef]

4. Fillinger, L.; de Theije, P.; Zampolli, M.; Sutin, A.; Salloum, H.; Sedunov, N.; Sedunov, A. Towards a passive acoustic underwater system for protecting harbours against intruders. In Proceedings of the 2010 International WaterSide Security Conference, Carrara, Italy, 3–5 November 2010; pp. 1–7.
5. Lv, Z.; Du, L.; Li, H.; Wang, L.; Qin, J.; Yang, M.; Ren, C. Influence of Temporal and Spatial Fluctuations of the Shallow Sea Acoustic Field on Underwater Acoustic Communication. *Sensors* **2022**, *22*, 5795. [CrossRef] [PubMed]
6. Wang, J.; Qian, P.; Chen, Y.; Zhou, S.; Li, Z.; Xiao, P. Adaptive underwater acoustic target recognition based on multi-scale residual and attention mechanism. *Digit. Signal Process.* **2025**, *163*, 105193. [CrossRef]
7. Xie, Y.; Ren, J.; Li, J.; Xu, J. Advancing robust underwater acoustic target recognition through multitask learning and multi-gate mixture of experts. *J. Acoust. Soc. Am.* **2024**, *156*, 244–255. [CrossRef]
8. Mistral AI Team. Mixtral of Experts. *arXiv* **2023**, arXiv:2310.06825v1.
9. Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; Chi, E.H. Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts. In Proceedings of the KDD '18: 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018.
10. Nguyen, H.; Akbarian, P.; Yan, F.Q.; Ho, N. Statistical Perspective of Top-K Sparse Softmax Gating Mixture of Experts. In Proceedings of the ICLR, Vienna, Austria, 7–11 May 2024.
11. Lu, E.; Jiang, Z.; Liu, J.; Du, Y.; Jiang, T.; Hong, C.; Liu, S.; He, W.; Yuan, E.; Wang, Y.; et al. MoBA: Mixture of Block Attention for Long-Context LLMs. *arXiv* **2025**, arXiv:2502.13189v1.
12. Jensen, F.B.; Kuperman, W.A.; Porter, M.B.; Schmidt, H. *Computational Ocean Acoustics*; Springer Science+Business Media, LLC: New York, NY, USA, 2011.
13. Oppenheim, A.V.; Schafer, R.W. *Discrete-Time Signal Processing*; Prentice Hall: Englewood Cliffs, NJ, USA, 1989.
14. Mallick, S.; Frazer, L.N. Practical aspects of reflectivity modeling. *Geophysics* **1987**, *52*, 1355–1364. [CrossRef]
15. Tolstoy, I. Phase changes and pulse deformation in acoustics. *J. Acoust. Soc. Am.* **1968**, *44*, 675–683. [CrossRef]
16. McDonald, B.E.; Kuperman, W.A. Time domain formulation for pulse propagation including nonlinear behavior at a caustic. *J. Acoust. Soc. Am.* **1987**, *81*, 1406–1417. [CrossRef]
17. Collins, M.D. The time-domain solution of the wide-angle parabolic equation including the effects of sediment dispersion. *J. Acoust. Soc. Am.* **1988**, *84*, 2114–2125. [CrossRef]
18. Collins, M.D. Applications and time-domain solution of higher-order parabolic equations in underwater acoustics. *J. Acoust. Soc. Am.* **1989**, *86*, 1097–1102. [CrossRef]
19. Schmidt, H.; Kuperman, W.A. Spectral and modal representations of the Doppler-shifted field in ocean waveguides. *J. Acoust. Soc. Am.* **1994**, *96*, 386–395. [CrossRef]
20. Porter, M.B. The time-marched fast-field program (FFP) for modeling acoustic pulse propagation. *J. Acoust. Soc. Am.* **1990**, *87*, 2013–2083. [CrossRef]
21. Sturm, F. Numerical study of broadband sound pulse propagation in three-dimensional oceanic waveguides. *J. Acoust. Soc. Am.* **2005**, *117*, 1058–1079. [CrossRef]
22. Schmalfeldt, B.; Rauch, D. *Explosion-Generated Seismic Interface Waves in Shallow Water: Experimental Results*; Rep. SR-71; SACLANT Undersea Research Centre: La Spezia, Italy, 1983.
23. Dougherty, M.E.; Stephen, R.A. Seismic energy partitioning and scattering in laterally heterogeneous ocean crust. *Pure Appl. Geophys.* **1988**, *128*, 195–229. [CrossRef]
24. Ferris, R.H. Comparison of measured and calculated normal-mode amplitude functions for acoustic waves in shallow water. *J. Acoust. Soc. Am.* **1972**, *52*, 981–988. [CrossRef]
25. Qarabaqi, P.; Stojanovic, M. Statistical Characterization and Computationally Efficient Modeling of a Class of Underwater Acoustic Communication Channels. *IEEE J. Ocean. Eng.* **2013**, *38*, 701–717. [CrossRef]
26. Cuji, D.A.; Stojanovic, M. Transmit Beamforming for Underwater Acoustic OFDM Systems. *IEEE J. Ocean. Eng.* **2024**, *49*, 145–161. [CrossRef]
27. Feng, S.; Zhu, X. A Transformer-Based Deep Learning Network for Underwater Acoustic Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1505805. [CrossRef]
28. Xiao, X.; Wang, W.; Ren, Q.; Gerstoft, P.; Ma, L. Underwater acoustic target recognition using attention-based deep neural network. *JASA Express Lett.* **2021**, *1*, 106001. [CrossRef]
29. Sun, Q.; Wang, K. Underwater single-channel acoustic signal multitarget recognition using convolutional neural networks. *J. Acoust. Soc. Am.* **2022**, *151*, 2245–2254. [CrossRef]
30. van Haarlem, M.P.; Wise, M.W.; Gunst, A.W.; Heald, G.; McKean, J.P.; Hessels, J.W.; de Bruyn, A.G.; Nijboer, R.; Swinbank, J.; Fallows, R.; et al. LOFAR: The LOw-Frequency ARray. *A&A* **2013**, *556*, A2. [CrossRef]
31. Wu, H.; Song, Q.; Jin, G. Deep Learning based Framework for Underwater Acoustic Signal Recognition and Classification. In Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence Xi'an, China, 8–10 December 2018; ACM: Shenzhen, China, 2018; pp. 385–388. [CrossRef]

32. Chung, K.W.; Sutin, A.; Sedunov, A.; Bruno, M. DEMON Acoustic Ship Signature Measurements in an Urban Harbor. *Adv. Acoust. Vib.* **2011**, *2011*, 952798. [CrossRef]
33. Ma, Y.; Liu, M.; Zhang, Y.; Zhang, B.; Xu, K.; Zou, B.; Huang, Z. Imbalanced Underwater Acoustic Target Recognition with Trigonometric Loss and Attention Mechanism Convolutional Network. *Remote Sens.* **2022**, *14*, 4103. [CrossRef]
34. Liu, F.; Shen, T.; Luo, Z.; Zhao, D.; Guo, S. Underwater target recognition using convolutional recurrent neural networks with 3-D Mel-spectrogram and data augmentation. *Appl. Acoust.* **2021**, *178*, 107989. [CrossRef]
35. Lim, T.; Bae, K.; Hwang, C.; Lee, H. Classification of underwater transient signals using MFCC feature vector. In Proceedings of the 2007 9th International Symposium on Signal Processing and Its Applications, Sharjah, United Arab Emirates, 12–15 February 2007; IEEE: Sharjah, United Arab Emirates, 2007; pp. 1–4. [CrossRef]
36. Lian, Z.; Xu, K.; Wan, J.; Li, G. Underwater acoustic target classification based on modified GFCC features. In Proceedings of the 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 25–26 March 2017; IEEE: Chongqing, China, 2017; pp. 258–262. [CrossRef]
37. Jiang, J.; Shi, T.; Huang, M.; Xiao, Z. Multi-scale spectral feature extraction for underwater acoustic target recognition. *Measurement* **2020**, *166*, 108227. [CrossRef]
38. Lu, A. Underwater Acoustic Classification Based on Deep Learning. Master's Thesis, Harbin Engineering University, Harbin, China, 2017.
39. Xie, Y.; Huang, S.; Chen, T.; Wei, F. Moec: Mixture of expert clusters. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; pp. 13807–13815.
40. Xie, Y.; Ren, J.; Xu, J. Adaptive ship-radiated noise recognition with learnable fine-grained wavelet transform. *Ocean Eng.* **2022**, *265*, 112626. [CrossRef]
41. Xie, Y.; Ren, J.; Xu, J. Guiding the underwater acoustic target recognition with interpretable contrastive learning. In Proceedings of the OCEANS 2023 Limerick, Limerick, Ireland, 5–8 June 2023; pp. 1–6.
42. Xie, Y.; Ren, J.; Xu, J. Underwater-art: Expanding information perspectives with text templates for underwater acoustic target recognition. *J. Acoust. Soc. Am.* **2022**, *152*, 2641–2651. [CrossRef]
43. Xie, Y.; Chen, T.; Xu, J. Advancing underwater acoustic target recognition via adaptive data pruning and smoothness-inducing regularization. *arXiv* **2023**, arXiv:2304.11907. [CrossRef]
44. Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv* **2017**, arXiv:1706.05098. [CrossRef]
45. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
46. Misra, I.; Shrivastava, A.; Gupta, A.; Hebert, M. Cross-stitch networks for multi-task learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3994–4003.
47. Ruder, S.; Bingel, J.; Augenstein, I.; Søgaard, A. Sluice networks: Learning what to share between loosely related tasks. *arXiv* **2017**, arXiv:1705.08142.
48. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive mixtures of local experts. *Neural Comput.* **1991**, *3*, 79–87. [CrossRef]
49. Tang, H.; Liu, J.; Zhao, M.; Gong, X. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In Proceedings of the 14th ACM Conference on Recommender Systems, Online, 22–26 September 2020; ACM: New York, NY, USA, 2020; pp. 269–278.
50. Hazimeh, H.; Zhao, Z.; Chowdhery, A.; Sathiamoorthy, M.; Chen, Y.; Mazumder, R.; Hong, L.; Chi, E. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 29335–29347.
51. Liang, Y.; Yu, H.; Ji, F.; Chen, F. Multitask sparse bayesian channel estimation for turbo equalization in underwater acoustic communications. *IEEE J. Ocean. Eng.* **2023**, *48*, 946–962. [CrossRef]
52. Zhang, Y.; Wang, H.; Li, C.; Meriaudeau, F. Complex-valued deep network aided channel tracking for underwater acoustic communications. In Proceedings of the OCEANS 2022-Chennai, Chennai, India, 21–24 February 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–5.
53. Stojanovic, M.; Catipovic, J.; Proakis, J.G. Adaptive multichannel combining and equalization for underwater acoustic communications. *J. Acoust. Soc. Am.* **1993**, *94*, 1621–1631. [CrossRef]
54. Zhang, Y.; Wang, H.; Li, C.; Chen, D.; Meriaudeau, F. Metalearning-aided orthogonal frequency division multiplexing for underwater acoustic communications. *J. Acoust. Soc. Am.* **2021**, *149*, 4596–4606. [CrossRef]
55. Williams, D.P. Transfer learning with sas-image convolutional neural networks for improved underwater target classification. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 78–81.
56. Liu, Y.; Niu, H.; Li, Z. A multi-task learning convolutional neural network for source localization in deep ocean. *J. Acoust. Soc. Am.* **2020**, *148*, 873–883. [CrossRef]
57. Wu, Y.; Ayyalasomayajula, R.; Bianco, M.J.; Bharadia, D.; Gerstoft, P. Sound source localization based on multi-task learning and image translation network. *J. Acoust. Soc. Am.* **2021**, *150*, 3374–3386. [CrossRef]

58. Zeng, X.; Lu, C.; Li, Y. A multi-task sparse feature learning method for underwater acoustic target recognition based on two uniform linear hydrophone arrays. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*; Institute of Noise Control Engineering: Washington, DC, USA, 2020; pp. 4404–4411.
59. Li, D.; Liu, F.; Shen, T.; Chen, L.; Zhao, D. A robust feature extraction method for underwater acoustic target recognition based on multi-task learning. *Electronics* **2023**, *12*, 1708. [CrossRef]
60. Santos-Domínguez, D.; Torres-Guijarro, S.; Cardenal-López, A.; Pena-Gimenez, A. ShipsEar: An underwater vessel noise database. *Appl. Acoust.* **2016**, *113*, 64–69. [CrossRef]
61. Acoustics Toolbox. Available online: http://oalib.hlsresearch.com/AcousticsToolbox/ (accessed on 25 December 2024).
62. Rouseff, D.; Spindel, R.C. Modeling the Waveguide Invariant as a Distribution. *AIP Conf. Proc.* **2002**, *621*, 137–150. [CrossRef]
63. Cipolla, R.; Gal, Y.; Kendall, A. Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7482–7491. [CrossRef]
64. Liebel, L.; Körner, M. Auxiliary Tasks in Multi-task Learning. *arXiv* **2018**, arXiv:1805.06334. [CrossRef]
65. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1026–1034. [CrossRef]
66. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:1412.6980. [CrossRef]
67. Huang, G.; Liu, Z.; Weinberger, K.Q.; van der Maaten, L. Densely Connected Convolutional Networks. *arXiv* **2016**, arXiv:1608.06993.
68. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
69. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 630–645.
70. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *arXiv* **2018**, arXiv:1801.04381.
71. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.