Journal of *Marine Science and Engineering*

MDPI

Article

Enhancing Underwater Acoustic Target Recognition Through Advanced Feature Fusion and Deep Learning

Yanghong Zhao [1], Guohao Xie [2], Haoyu Chen [3], Mingsong Chen [1,2,*] and Li Huang [2]

[1] School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, China; 15977446488@163.com
[2] School of Ocean Engineering, Guilin University of Electronic Technology, Beihai 536000, China; 22172301007@mails.guet.edu.cn (G.X.); bhxq7507@163.com (L.H.)
[3] Guangxi Electrical Polytechnic Institute, Nanning 530299, China; chenhy970403@163.com
* Correspondence: cms@guet.edu.cn

Abstract: Underwater Acoustic Target Recognition (UATR) is critical to maritime traffic management and ocean monitoring. However, underwater acoustic analysis is fraught with difficulties. The underwater environment is highly complex, with ambient noise, variable water conditions (such as temperature and salinity), and multi-path propagation of acoustic signals. These factors make it challenging to accurately acquire and analyze target features. Traditional UATR methods struggle with feature fusion representations and model generalization. This study introduces a novel high-dimensional feature fusion method, CM3F, grounded in signal analysis and brain-like features, and integrates it with the Boundary-Aware Hybrid Transformer Network (BAHTNet), a deep-learning architecture tailored for UATR. BAHTNet comprises CBCARM and XCAT modules, leveraging a Kan network for classification and a large-margin aware focal (LMF) loss function for predictive losses. Experimental results on real-world datasets demonstrate the model's robust generalization capabilities, achieving 99.8% accuracy on the ShipsEar dataset and 94.57% accuracy on the Deepship dataset. These findings underscore the potential of BAHTNet to significantly improve UATR performance.

Keywords: underwater acoustic target recognition (UATR); feature fusion; deep learning network; dataset imbalance

check for updates

Academic Editor: Rouseff Daniel

Received: 8 December 2024
Revised: 18 January 2025
Accepted: 24 January 2025
Published: 31 January 2025

Citation: Zhao, Y.; Xie, G.; Chen, H.; Chen, M.; Huang, L. Enhancing Underwater Acoustic Target Recognition Through Advanced Feature Fusion and Deep Learning. *J. Mar. Sci. Eng.* **2025**, *13*, 278. https://doi.org/10.3390/jmse13020278

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Underwater Acoustic Target Recognition (UATR) is crucial in maritime traffic management and ocean monitoring. This facilitates the timely detection of underwater threats, including submarine activities. In addition to enhancing naval safety, the UATR contributes to marine research by safeguarding biodiversity, developing resources, and preserving underwater cultural heritage. As a result, UATR has attracted significant attention across multiple disciplines.

Sonar technology, a key aspect in underwater acoustic research, is currently a focus of intense investigation by researchers. Sonar is also currently being focused on by researchers. Different approaches and techniques within sonar systems are being explored and developed to improve the performance and effectiveness of UATR. For instance, the applied methods can be mixed or even changed. Multi-attention networks with feature transfer are among the techniques being examined. Additionally, combinations and alterations of existing methods are being considered to enhance the capabilities of sonar systems in underwater acoustic target recognition.

UATR relies on hydrophones to capture radiated noise from targets. The collected data were analyzed and processed to identify the type of target [1]. This technology is crucial for monitoring maritime vessels and for detecting underwater vehicles. Typically, acoustic target recognition models have two main components: feature extraction and classification [2]. Most research in this field has focused on these two components.

Traditional methods for extracting the features of underwater acoustic targets can be divided into two main categories: signal physics-based approaches and brain computing-inspired methods [3]. Signal physics-based methods focus on the fundamental characteristics of underwater acoustic signals [4], including their temporal and non-Gaussian features. The key attributes in the time domain include a zero-crossing distribution, whereas the features in the frequency domain often involve cepstral analysis [5]. Additionally, joint time-frequency domain features are analyzed using techniques, such as wavelet transform [6–8]. In contrast, brain computing-inspired methods utilize features, such as Mel-Frequency Cepstral Coefficients (MFCC) [9], which simulate the nonlinear processing of the human ear, and gamma tone filtering [10], which mimic peripheral auditory processing. The standard traditional classifier models employed in these methods include case-based reasoning [11] and perceptron neural networks [12].

Although traditional methods provide directional analysis based on the physical attributes of underwater acoustic signals, they rely heavily on prior knowledge, limiting the generalization capabilities of the models [13]. Current feature extraction techniques, predominantly two-dimensional, are based on signal physics, brain-computing features, or hybrid approaches [14–21] but often overlook the diverse features present in underwater acoustic data, reducing the symbolic power of integrated features.

In underwater acoustic target recognition (UATR) research, a prevalent method involves signal analysis and processing to extract features from labeled samples, which is a characteristic of supervised learning within the deep learning framework. Classifiers typically follow a two-step process: they initially perform feature extraction and subsequently employ these features to produce classification outcomes. Ren et al. [22] developed a learning-capable front-end for the UATR system that integrates adaptive feature extraction with classification, enhancing the system's robustness. Similarly, inspired by DenseNet, Santos-Domínguez et al. [23] created a UATR model designed for environments with low signal-to-noise ratios that require minimal computational resources. Khishe [24] utilized the ShipsEar [25] dataset to combine wavelet networks with autoencoders, effectively mitigating the marine environment's impact on underwater acoustic target recognition. Additionally, Feng et al. [26] were the first to apply a transformer network to UATR, achieving superior results compared to various convolutional neural networks (CNNs) on the DeepShip [27] dataset.

Numerous issues persist in underwater acoustic target recognition (UATR) research. Current models typically separate the feature extraction component from the classifier [28], which complicates the achievement of the end-to-end recognition requirements. Classifiers frequently improve feature extraction abilities by piling up convolutional layers [29], but they struggle to effectively differentiate between underwater acoustic target features [30] and marine environmental noise. This intermingling can lead to the loss of critical features while inadvertently retaining noise [31], thereby diminishing the model's ability to extract compelling features. Consequently, achieving a balance between the number of model parameters and recognition performance becomes challenging, undermining the high demand for speed and accuracy in the UATR. The acquisition of underwater acoustic data requires significant financial and human resources.

Moreover, many existing public datasets show imbalances in sample quantities among the different categories [32]. This imbalance may cause overfitting in target recognition

models based on deep learning [33]. Consequently, overfitting reduced the recognition performance of the model.

To address the challenge of insufficient representational power in feature fusion, we introduce a novel method called CQT MFCC FBank Feature Fusion (CM3F), which integrates the Constant-Q Transform (CQT), Mel-Frequency Cepstral Coefficients (MFCC), and Filter Bank (FBank) to provide a comprehensive acquisition of underwater signal information. We propose a Boundary-Aware Hybrid Transformer Network (BAHTNet), a deep learning architecture designed to enhance feature extraction and recognition accuracy while addressing sample imbalance. BAHTNet incorporates the change-boundary-aware channel prior convolutional attention ResNeXt module (CBCARM) and the XCAttention Transformer Encoder (XCAT) to improve feature extraction and representation capabilities.

The contributions of this study are as follows.

1. The introduction of a high-dimensional feature fusion method, named CM3F, is based on signal analysis and brain-motivated features.
2. The proposal of BAHTNet enriches existing features with physical, channel, and contextual information and enhances feature extraction capabilities.
3. Incorporating the large-margin aware focal (LMF) loss function, which determines weights based on the dataset's class distribution, significantly improves the model's ability to suppress sample imbalance and recognition accuracy.

The remainder of this paper is organized as follows. Section 2 describes the classification method for ship-radiated noise, namely BAHTNet. Section 3 presents qualitative and quantitative experiments that compare BAHTNet with current advanced underwater acoustic target recognition models and analyzes the results. Finally, Section 4 concludes the study by summarizing the main findings.

2. Methods

The complexity of underwater sonar target identification stems from diverse noise levels, which hamper the extraction of consistent features. These inconsistencies contribute to analytical challenges, such as skewed data distribution, model overfitting, and suboptimal recognition rates. To address these complexities, we introduce the Boundary-Aware Hybrid Transformer Network (BAHTNet), an innovative deep-learning framework. This section delineates the comprehensive framework of the BAHTNet. Section 2.1 offers a synoptic view of its architectural scaffolding, and subsequent Sections 2.2–2.6 dissect the individual components in detail. These include the novel feature extraction and fusion module (CM3F), an advanced convolutional attention mechanism within the ResNeXt framework (CBCARM), the XCAttention Transformer encoder (XCAT), a sophisticated pattern classification approach known as the Kolmogorov–Arnold Network (Kan), and the formulation of an efficacious large-margin focal (LMF) loss function designed to counteract the identified challenges.

2.1. Proposed Model

The BAHTNet model was bifurcated into two principal components: the Comprehensive Multimodal Feature Fusion Module (CM3F) and the BAHTNet classifier itself. A schematic representation of its architecture is presented in Figure 1, which illustrates the intricate interplay between its constituent elements.

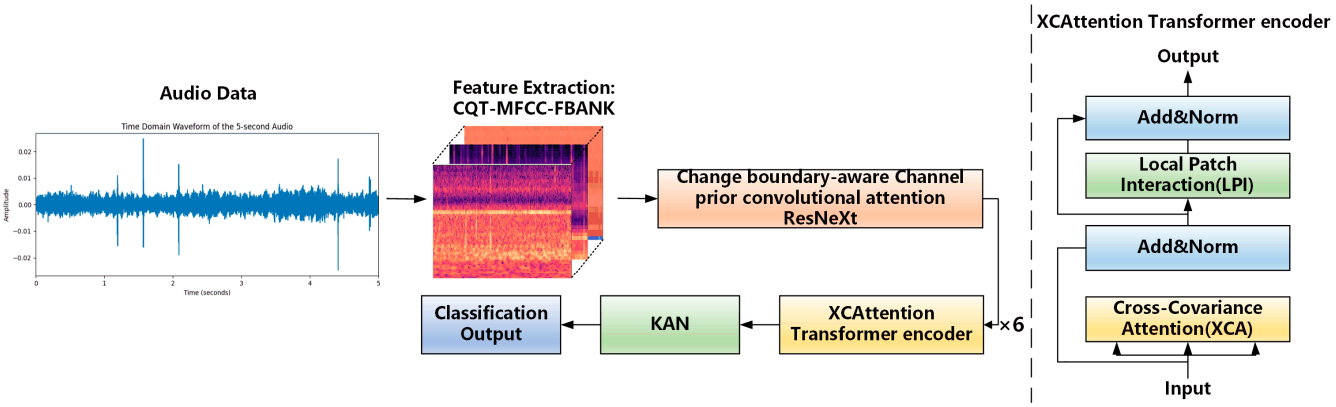

Figure 1. BAHTNet model architecture.

Figure 1 shows the BAHTNet model architecture diagram. The model mainly consists of the Comprehensive Multimodal Feature Fusion Module (CM3F) and the BAHTNet classifier. Among them, the BAHTNet classifier realizes the innovation of underwater acoustic target recognition by integrating advanced modules that enhance feature extraction and classification. Its core modules are the Channel-Priority and Boundary-Aware ResNeXt Module (CBCARM) and the XCAttention Transformer (XCAT). The Change Boundary-Aware Mechanism (CBM) and the Channel-Priority Convolutional Attention (CPCA) in CBCARM can dynamically identify and preserve critical features. At the same time, the components of XCAT utilize the transformer framework to achieve efficient feature capture and local patch interaction. The synergy of the two enables BAHTNet to effectively deal with the complexity brought by marine environmental noise, improve the accuracy of underwater acoustic target feature extraction and recognition, enrich the representation of physical, channel and context data, and significantly improve the model accuracy. In addition, BAHTNet adopts the large-margin focal (LMF) loss function to deal with the challenges brought by imbalanced datasets. This function strategically adjusts the weights based on the class distribution, focuses on improving the recognition accuracy and reducing the misclassification of minority classes. By integrating weight adjustment and margin constraints, it enhances the model's ability to handle imbalanced datasets, thereby improving the overall accuracy and stability.

The BAHTNet classifier revolutionizes underwater acoustic target recognition by integrating advanced modules that enhance the feature extraction and classification. The Channel-Priority and Boundary-Aware ResNeXt Module (CBCARM) and the XCAttention Transformer (XCAT) are at their core. CBCARM's Change Boundary-Aware Mechanism (CBM) and Channel-Priority Convolutional Attention (CPCA) of CBCARM dynamically identify and preserve critical features. Simultaneously, XCAT's components leverage the transformer framework for efficient feature capture and local patch interaction.

This synergistic integration allows BAHTNet to effectively manage the complexities introduced by marine environmental noise, thereby improving the accuracy of extracting and recognizing underwater acoustic target features. The collaborative effect of BAHTNet modules results in an enriched representation of physical, channel, and contextual data, significantly enhancing the precision of the model.

BAHTNet employs a large-margin focal (LMF) loss function to counter the challenges posed by imbalanced datasets. This function strategically adjusts weights based on class distribution, focusing on improving the recognition accuracy and reducing the misclassification of minority classes. By incorporating weight adjustment and margin constraints, the LMF loss function fortifies the model's ability to handle imbalanced datasets, thereby improving overall accuracy and stability.

BAHTNet's innovative design and strategic integration of modules offers a formidable underwater acoustic target recognition solution with potential applications in maritime surveillance, defense systems, and acoustic research. Its ability to overcome the limitations of conventional models and effectively manage dataset imbalances makes BAHTNet a cutting-edge tool in underwater acoustics.

2.2. CQT, MFCC, and FBank Feature Fusion Module (CM3F)

Traditional machine-learning techniques for underwater sonar target feature extraction often rely on handcrafted features that may not fully capture the complexity of underwater signal dynamics and can be computationally inefficient. However, although effective in feature extraction, deep learning approaches face challenges, such as the need for large datasets and high computational resources, which are not always feasible in underwater applications.

In the context of our study, the processing of audio data into spectrogram or Mel spectrogram models was not analyzed. This is because our work is based on previous research. In the original article, extensive experiments have been conducted, which clearly showed that the spectrogram and Mel spectrogram methods do not yield satisfactory results in underwater target recognition. Therefore, we focused our efforts on other methods of the research.

To address these challenges, we introduced the CQT MFCC FBank Feature Fusion (CM3F) module, as shown in Figure 2, a novel solution that combines the strengths of the Constant-Q Transform (CQT), Filter Bank (FBank), and Mel-Frequency Cepstral Coefficients (MFCC). This integration results in a multidimensional, three-channel feature set that captures the nuances of underwater signals, including frequency, energy distribution, and spectral shape.

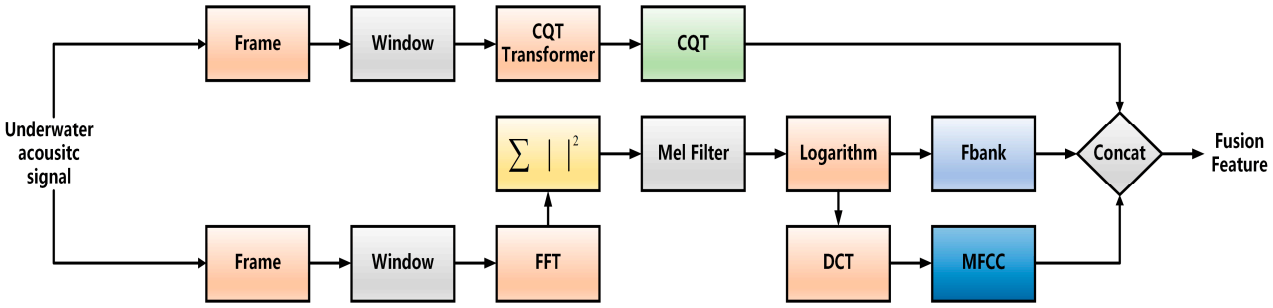

Figure 2. The feature extraction process for fused features.

The CM3F module's innovative color space representation technique transforms the traditional single-channel feature set into a more expressive format, enhancing the model's ability to understand and analyze underwater acoustic signals. This approach enriches the dataset for more precise insights and streamlines the classification process by reducing the computational burden on subsequent layers of the recognition network.

Integrating the CM3F module at the forefront of our recognition network is crucial for efficient end-to-end target recognition. It offers a robust solution for underwater sonar systems, where accurate feature extraction and classification are critical for effective monitoring and detection of underwater targets.

2.2.1. CQT Extraction Process

In the feature extraction process, the frame length is set to 2048, with a 75% overlap between frames. A Hanning window, matching the frame length, is applied to each frame signal.

The CQT transformation for a finite-length sequence $x(n)$ is:

$$X^{XQT}(k) = \frac{1}{N_k} \sum_{n=0}^{N_k-1} x(n) w_{N_k}(n) e^{-j\frac{2\pi Q}{N_k}n} \tag{1}$$

In this context, $w_{N_k}(n)$ represents a Hanning window of length N_k, Q denotes the constant factor in the CQT, and k indicates the number of CQT frequency bins, with k being related to N_k.

$$Q = \frac{1}{2^{\frac{1}{b}} - 1} \tag{2}$$

In this context, b represents the number of spectral lines.

$$f_k = f_{\min} \times 2^{\frac{k}{b}}, k = 0, 1, \ldots, k-1 \tag{3}$$

$$N_K = \left[Q \frac{f_s}{f_k} \right], k = 0, 1, \ldots, k-1 \tag{4}$$

In this case, the CQT information is stored in matrix $X^{CQT}(k, n)$, with $f_{\min}$ set to 1 and f_s to 22,050. The original underwater acoustic data has a sampling rate of 22,050 Hz over a duration of 5 s, resulting in a CQT shape of 128 × 216.

2.2.2. MFCC Extraction Process

The extraction of Mel-Frequency Cepstral Coefficients (MFCC) is a systematic process critical for various audio-processing applications. It begins with frame splitting and windowing, where the audio signal is segmented into 2048-sample frames with 75% overlap, followed by the application of a Hanning window to minimize spectral leakage. The Fast Fourier Transform (FFT) then translates these frames into the frequency domain, enabling analysis of the signal's frequency components.

The subsequent power spectrum calculation provides vital information regarding the energy distribution of the signal. Audio signal processing is further refined through Mel Filter Bank Processing, which applies a logarithmic scale to better align with human auditory perception, creating a log-Mel spectrum. The process culminates with the Discrete Cosine Transform (DCT), which condenses the spectrum into MFCCs that capture the spectral shape of the audio signal.

MFCCs are particularly advantageous in underwater acoustics because they represent the perceptual characteristics of sounds, such as human hearing. The ability of MFCCs to convey the perceptual attributes of sounds, such as human hearing, makes them well-suited for speech recognition and acoustic event detection in complex environments, including underwater applications.

Integrating MFCC extraction in practical systems often involves automated pipelines that process real-time or recorded audio data, thereby enhancing the capabilities of sonar and other acoustic monitoring technologies. The robustness of the MFCC features contributes to the reliability and accuracy of these systems under diverse acoustic conditions.

2.2.3. FBank Extraction Process

We employed parameters synchronized with MFCC extraction during the feature extraction phase to ensure a uniform signal-processing approach. The Fast Fourier Transform (FFT) application decomposes the signal into its fundamental frequencies. Following this, the Mel Filter Bank is applied to create the FBank feature set, representing the spectral attributes of the signal by overlaying a series of triangular filters on the power spectrum according to the Mel scale.

Figure 3 visually compares the ship-radiated noise waveforms and spectral plots derived from the ShipsEar dataset for the CQT, MFCC, and FBank features. These visual aids provide insights into different signal characteristics and their representations.

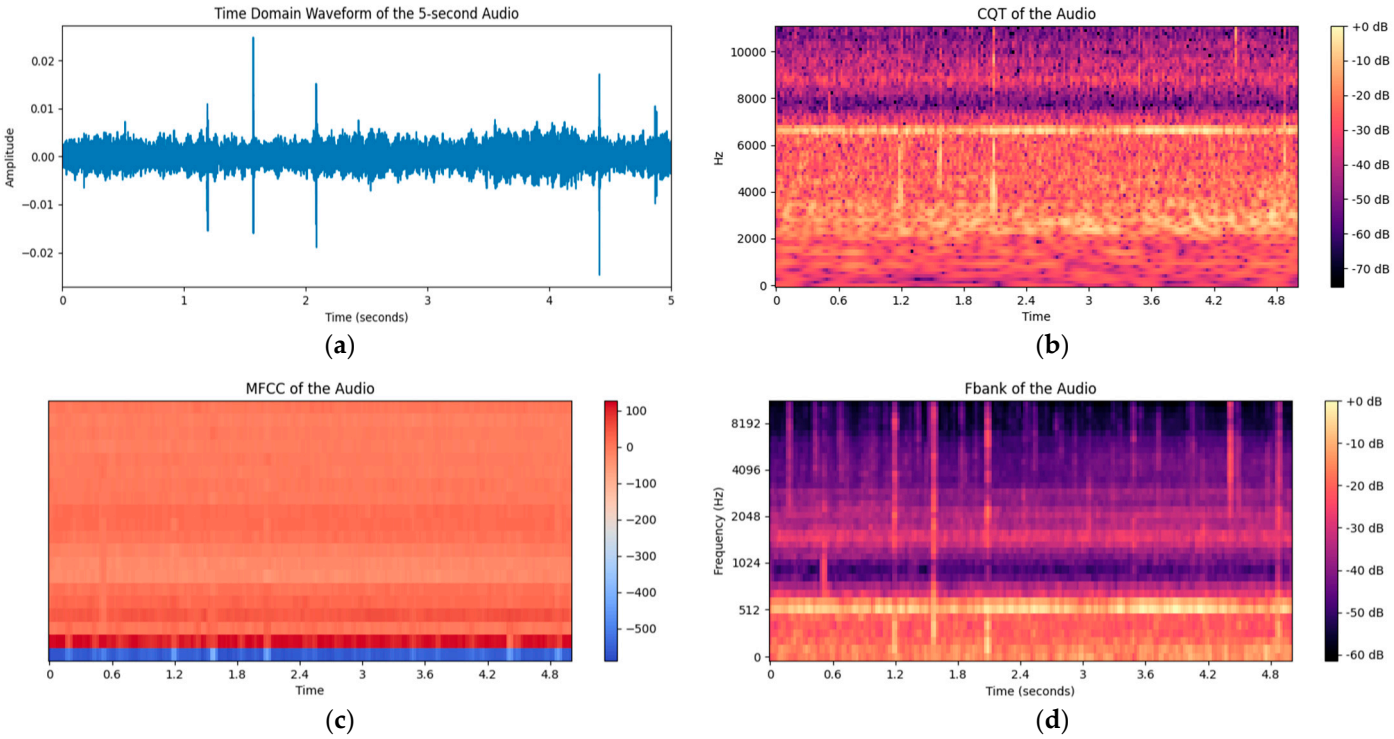

Figure 3. Spectral feature plots of radiated noise from ships in the ShipsEar database: (**a**) Time-domain waveform; (**b**) CQT; (**c**) MFCC; (**d**) FBank.

In the culmination of feature processing, we integrated the spectral data from CQT, MFCC, and FBank into a multichannel feature set. This integration along the channel dimension enriches the representative capacity of the model, allowing for a deeper comprehension of underwater acoustic signals. Such an integrated feature set is vital for precise target identification and classification, as it encapsulates a holistic view of the properties of the signal.

The methodology of feature fusion is elaborated in the following sections, detailing the techniques and rationale for combining distinct feature sets. This approach enhances the ability of the model to discern underwater objects and has practical implications for improving the reliability and accuracy of underwater sonar systems.

2.2.4. Feature Fusion

In traditional underwater acoustics, feature extraction often results in single-channel data that captures the essential dimensions of sound, such as time, frequency, and energy. Our method introduces a novel color-space transformation to enhance the representativeness of these features. This transformation expanded the single-channel features into a three-channel set, significantly enriching the descriptive power of the energy domain.

The resulting $3 \times 128 \times 216$ dimensional feature set was constructed by strategically sequencing Constant-Q Transform (CQT), Mel-Frequency Cepstral Coefficients (MFCC), and Filter Bank (FBank) features across channel dimensions. The CQT in the first channel provides a high-resolution frequency component analysis, which is crucial for low-frequency sound characteristics. In the second channel, the MFCC offers a human-like perception of

spectral shape and dynamics. FBank, assigned to the third channel, details the energy distribution across frequency bands, which is vital for sound spectral feature characterization.

This fused feature set harnesses the individual strengths of each feature extraction technique, offering a comprehensive representation of underwater acoustic signals. Such an enriched representation is critical for enhancing the accuracy of target recognition and classification in intricate marine environments, where signals can be subjected to various distortions and interferences.

The potential applications of this multichannel feature set extend to advanced sonar systems, where the ability to accurately discern and classify underwater objects is paramount. Enhanced feature representation could lead to improved detection capabilities and more reliable underwater monitoring, with implications for marine research and defense applications.

$$T = concat(\sum_{j=0}^{2} T_j^{CQT}, \sum_{j=0}^{2} T_j^{MFCC}, \sum_{j=0}^{2} T_j^{Fbank}) \tag{5}$$

Here, T_j^{CQT} denotes the feature map of layer j within the channel dimension of the CQT spectral feature matrix, while T_j^{MFCC} and T_j^{Fbank} carries the same meaning.

2.3. Change Boundary-Aware Channel Prior Convolutional Attention ResNeXt Module (CBCARM)

The CBCARM module represents a remarkable innovation within the ResNeXt architecture, purposefully designed to meet the demands of underwater sonar identification. It accomplishes this through the integration of two crucial sub-modules: the Change Boundary-Aware Module (CBM) and the Channel Prior Convolutional Attention Module (CPCA).

The CBM holds a strategic position within the residual connections. Its primary task is to detect and maintain essential boundary information of the signals. By doing so, it acts as a safeguard against the loss of critical signal details, which are vital for accurate identification. After the signals pass through the residual connections, the CPCA comes into play. It dynamically assigns attention weights to the significant regions of the signals. This enables the module to better distinguish the intricate features present in the underwater acoustic signals.

There is a highly coordinated interaction between the CBM and the CPCA. Initially, the CBM captures the boundary information, laying a solid foundation for the subsequent processing steps. Subsequently, the CPCA capitalizes on the information processed by the CBM and further hones in on the key regions. This seamless cooperation not only enhances the CBCARM module's proficiency in analyzing underwater acoustic signals but also enriches the depth and precision of feature capture.

Consequently, this integrated approach leads to a significant improvement in the fidelity and precision of the identification process. As a result, the CBCARM module can ensure the accurate classification of underwater sonar signals. Table 1 comprehensively presents the network architecture of the module and the progression of output dimensions across its various stages.

Table 1. CBCARM network structure.

Stage	Output	CBCARM
Conv1	216 × 128	7 × 7, 64, stride 2, C = 32
Conv2	108 × 64	3 × 3 max pool, stride 2
		$\begin{bmatrix} 3\times3, & 128 \\ 3\times3, & 128 \end{bmatrix} \times 2$, C = 32
Conv3	54 × 32	$\begin{bmatrix} 3\times3, & 256 \\ 3\times3, & 256 \end{bmatrix} \times 2$, C = 32
CBM	27 × 16	
Conv4	27 × 16	$\begin{bmatrix} 3\times3, & 512 \\ 3\times3, & 512 \end{bmatrix} \times 2$, C = 32
Conv5	14 × 8	$\begin{bmatrix} 3\times3, & 1024 \\ 3\times3, & 1024 \end{bmatrix} \times 2$, C = 32
CPCA	7 × 4	
	7 × 4	Global average pool

The innovative design of the CBCARM module effectively addresses the unique challenges posed by underwater acoustic data. It offers a reliable and robust solution for accurate and dependable sonar target identification. Its potential applications in maritime surveillance and defense systems clearly demonstrate the importance and significance of such architectural advancements in the field of underwater acoustics.

2.3.1. Change Boundary-Aware (CBM) Module

The CBCARM module, as illustrated in Figure 4, marks a notable evolution of the ResNeXt framework, with the Change Boundary-Aware Module (CBM) at its core. The CBM is meticulously designed to identify and maintain crucial boundary information within the transitional data segments, a feature critical for the nuances of underwater sonar identification.

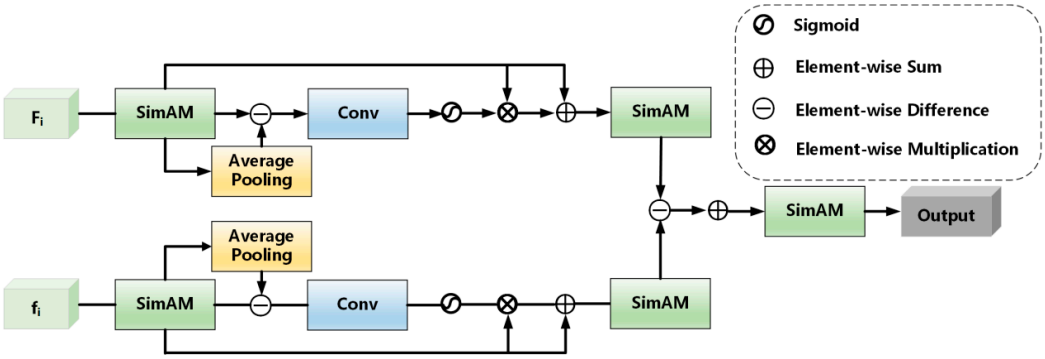

Figure 4. Structural diagram of the CBM module.

The CBM's strategic implementation is carried out through refined steps, each meticulously crafted to capture and enhance the boundary details essential for distinguishing features in underwater sonar signals. This systematic methodology is critical to the CBCARM module's advanced feature extraction capabilities and contributes to the heightened precision in sonar target identification.

Moreover, the CBM's integration within the CBCARM module refines the ResNeXt architecture's ability to process complex sonar data and enhances its overall performance. The CBM's ability to isolate and process pivotal boundaries enriches the model's capacity

to discern subtleties in underwater acoustic signals, thereby improving the reliability and accuracy of target identification.

The CBCARM module, with the CBM at its forefront, stands as a testament to the power of strategic architectural enhancements in advancing underwater sonar target identification. Its potential applications in maritime surveillance and defense systems underscore the significance of such innovations in the field.

1. Input Feature Extraction: Features f_i from the extraction network are fed into the CBM to enhance the edge information of dual-temporal features.

$$\hat{f} = SimAM(f_i) \tag{6}$$

Here, $SimAM$ represents the attention mechanism.

2. Edge Feature Extraction: Edge features are obtained through pooling, subtraction, and convolution operations. Average pooling smooths the features, while subtraction operations emphasize the edge areas. This process is followed by convolution, which, along with a sigmoid activation function, enhances the contrast and prominence of the extracted edge features.

$$f_e^t = Conv_{1\times1}(\hat{f} - AP(\hat{f})) \times \hat{f} + \hat{f} \tag{7}$$

Here, AP denotes the average pooling layer of 3×3, and $Conv_{1\times1}$ represents the convolutional block that uses the sigmoid activation function.

3. Feature Enhancement and Repetition: Edge features are further enhanced through multiplication and addition, followed by the reapplication of the attention mechanism $SimAM$ to identify significant areas, resulting in the final feature set $\hat{f}_t^e$.

$$\hat{f}_e^t = SimAM(f_e^t) \tag{8}$$

The Change Boundary-Aware Module (CBM) is a pivotal component in underwater sonar target recognition, adept at integrating edge information and performing differential feature operations to enhance the recognition of target boundaries. By leveraging existing knowledge, the CBM captures the intricate structure of targets, clarifying their boundaries and reducing the obscurity of edges.

The CBM's potential applications in future technologies could include further optimization for real-time sonar processing and integration with autonomous underwater vehicle systems, underscoring its significance in advancing underwater acoustics.

2.3.2. CPCA Attention Mechanism

The Channel Prior Convolutional Attention module, illustrated in Figure 5, sequentially executes channel attention followed by spatial attention. The overall attention process can be summarized as follows:

$$F_c = CA(F) \otimes F \tag{9}$$

$$\hat{F} = SA(F_c) \otimes F_c \tag{10}$$

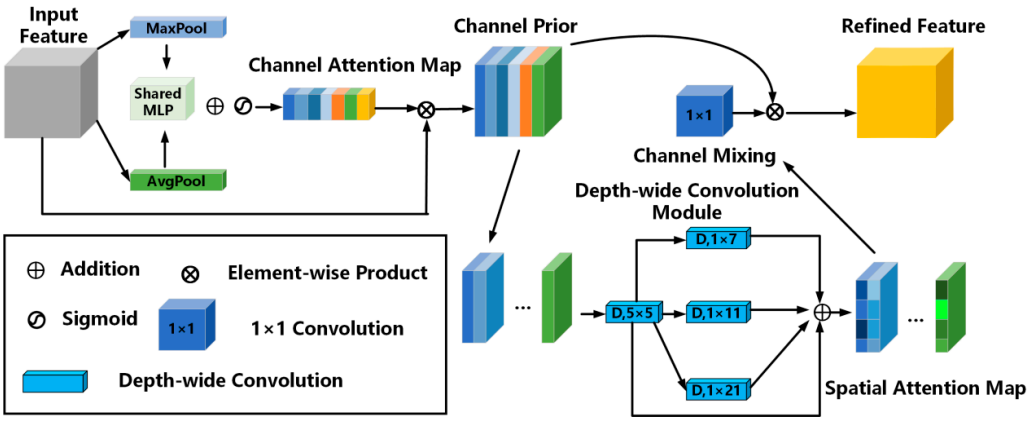

Figure 5. CPCA attention mechanism.

In this context, $\otimes$ denotes element-wise multiplication, $F \in R^{C\times H\times W}$ refers to the input, and $F_c \in R^{C\times H\times W}$ signifies the refined feature with a focus on channels. Additionally, C, H, and W represent the number of channels, height, and width, respectively.

The channel attention map is generated by combining the shared multi-layer perceptron (MLP) outputs through element-wise summation. The shared MLP is designed with a single hidden layer to minimize parameter overhead. The size of the hidden activation is denoted as $R^{c/r\times 1\times 1}$, while r represents the reduction ratio. The computation of channel attention can be summarized as follows:

$$CA(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{11}$$

where σ represents the sigmoid function.

The computation of spatial attention can be described as follows:

$$SA(F) = Conv_{1\times 1}(\sum_{i=0}^{3} Branch_i(DwConv(F))) \tag{12}$$

Here, DwConv refers to depth-wise convolution, while $Branch_i$ and $i \in \{0,1,2,3\}$ represent the i-th branch. $Branch_0$, known as an identity connection, denotes a method that transmits the input directly to the output without any transformation or processing.

The CPCA mechanism's ability to selectively emphasize critical data points is particularly beneficial in complex underwater environments where signal clarity can be compromised. Its strategic application within the CBCARM module positions BAHTNet as a robust underwater sonar target identification solution with potential maritime surveillance and defense systems applications.

2.4. XCAttention Transformer Encoder (XCAT)

In the field of underwater sonar target recognition, the performance of traditional Transformer encoders is often significantly affected when facing the complex underwater noise environment. This is mainly because they are sensitive to noise and tend to blur the target features during processing, thereby leading to a decrease in the recognition accuracy of the model. To effectively address this issue, we have developed the XCAT module, which is an innovative advanced transformer encoder that integrates two submodules, XCAttention and LPI, and improves the model performance through their collaborative work.

The XCAttention submodule introduces a cross-covariance mechanism in the attention calculation process, which is an important improvement over the traditional self-attention mechanism. To illustrate this more clearly, we have drawn the principle schematic diagram

as shown in Figure 1. It can be seen from the figure that when processing underwater signals, the traditional self-attention mechanism may fail to accurately focus on the key information due to the interference of noise, resulting in insufficient feature extraction. However, the XCAttention submodule, through the calculation of cross-covariance, can measure the correlation between features at different positions more finely, thereby extracting the information crucial for target recognition from the complex underwater signals more effectively. For example, when processing underwater sonar signals containing multiple targets, the XCAttention submodule can better distinguish the features of different targets, reduce the confusion between features, and improve the accuracy and effectiveness of feature extraction.

$$A_{XC}(K, Q) = soft\max(K^T Q/\tau) \tag{13}$$

$$XC - Attention(Q, K, V) = VA_{XC}(K, Q) \tag{14}$$

Here, the learnable temperature parameters τ and A are the attention weights.

The Local Patch Interaction (LPI) module, seamlessly integrated into the XCAT framework, is pivotal in enhancing the underwater sonar target identification process. By promoting interactions among local data patches, the LPI module enables the model to capture fine-grained details and nuances, leading to a more accurate understanding of the data's underlying structures.

In environments where signals are often complex and noisy, the LPI module's strategic design allows the model to discern targets and navigate intricate signal patterns effectively. Coupled with the XCAttention module, the LPI enhances the model's feature extraction capabilities, ensuring a robust response to signal mixtures and other challenges inherent in underwater acoustics.

Through in-depth analysis of Table 7, we can clearly gain insight into the outstanding performance of the XCAT module in terms of performance improvement. With its unique architectural design, this module integrates the XC attention and LPI components, forming a powerful synergy that greatly enhances the overall performance of the system. In practical applications, this integration advantage is fully demonstrated, not only significantly improving the recognition accuracy but also comprehensively enhancing the reliability and effectiveness of the underwater sonar target recognition process. It is precisely due to the close collaboration and complementary advantages among the components within the XCAT module that it has become an indispensable and powerful tool in application scenarios such as ocean and defense that have strict requirements for accurate target recognition, providing a solid technical support for the development of related fields.

2.5. Kolmogorov–Arnold Networks (Kan)

In underwater sonar target identification, traditional Multilayer Perceptrons (MLPs) often fall short of accurately extracting information from noisy and complex signals, resulting in overfitting and a lack of transparency in their decision-making processes. This study introduces Kan networks (as shown in Figure 6) to address these issues, demonstrating superior performance in feature classification tasks.

This adaptability significantly reduces the risk of overfitting and bolsters the model's ability to generalize to new data.

Moreover, Kan networks incorporate a mechanism for calculating regularization loss, essential for controlling model complexity and enhancing generalization performance. The visual characteristics of these networks also address the interpretability concerns associated with MLPs, providing a clearer understanding of the decision-making process.

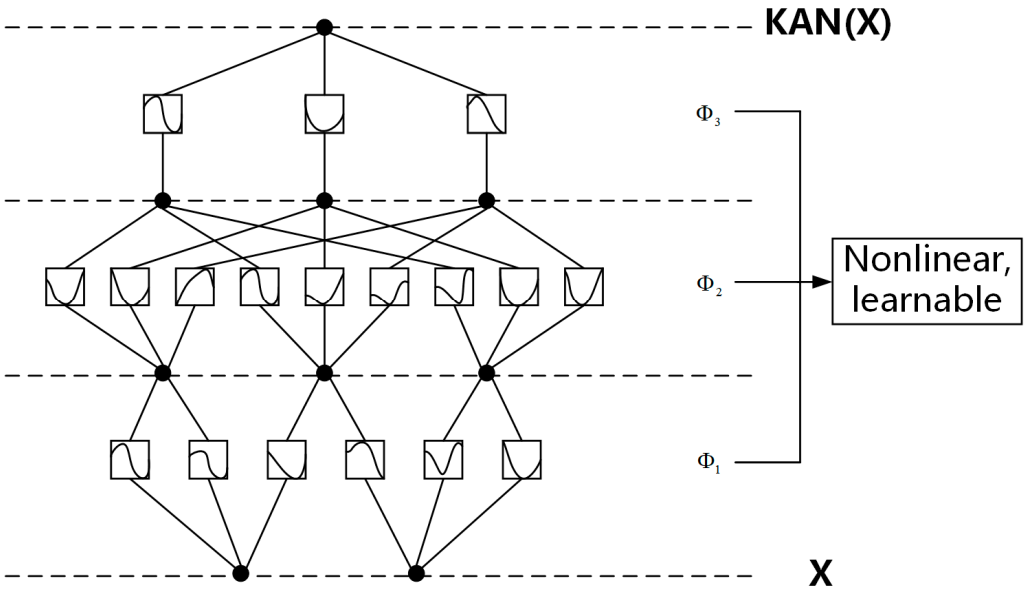

Figure 6. Structural diagram of the Kan network.

The introduction of Kan networks marks a significant advancement in underwater sonar target identification, offering a robust and interpretable alternative to traditional MLPs. Future work will explore the deployment of Kan networks in real-world applications and further refine their architecture to optimize performance.

$$KAN(X) = (\Phi_3 \circ \Phi_2 \circ \Phi_1)(X) \tag{15}$$

Integrating the Kan network into the BAHTNet system has become a pivotal enhancement over the traditional Multilayer Perceptron (MLP), serving as an adept classification layer. This strategic upgrade seamlessly complements the underwater sonar target recognition system's existing modules, enriching the extraction and representation of physical, channel, and contextual data details.

The sophisticated architecture of the Kan network facilitates a significant boost in overall recognition accuracy, demonstrating its proficiency in navigating the intricacies of underwater acoustic environments. By effectively managing noise and complexity, the Kan network endows the BAHTNet system with robustness, fortifying its ability to identify a spectrum of underwater targets accurately.

The transition from MLP to the Kan network has optimized performance metrics and enhanced the model's generalization capabilities, equipping BAHTNet to confront real-world challenges with precision and reliability. This advancement positions BAHTNet as a cutting-edge underwater sonar target identification solution with potential applications in various marine and defense sectors.

2.6. Large-Margin Aware Focal (LMF) Loss Function

The large-margin aware focal (LMF) loss function is an innovative approach designed to tackle the critical issue of sample imbalance in underwater sonar target identification. In imbalanced datasets, traditional classifiers often exhibit reduced accuracy due to their tendency to favor majority classes over minority ones.

The LMF loss function counters this bias by dynamically shifting the focus towards minority classes during training. This strategic adjustment improves the recognition of underrepresented classes and promotes a balanced learning environment. As a result, the classifier becomes more adept at generalizing from the training data, leading to a more precise and dependable recognition system.

The effectiveness of the LMF loss function lies in its ability to intelligently adjust the learning process according to the sample distribution, thereby enhancing the model's capacity to identify a diverse range of underwater sonar targets accurately. This approach sets a new precedent for addressing sample imbalance, with potential applications extending beyond underwater acoustics to other domains where imbalanced datasets are prevalent.

The LMF loss function formulation is as follows:

$$L_{LMF} = \alpha L_{LDAM} + \beta L_{FL} \tag{16}$$

Here, L_{LDAM} represents the label distribution aware marginal ($LDAM$) loss,L_{FL} denotes the focal loss, and α and β are adjustable hyperparameters.

The LDAM loss minimizes generalization error by applying more robust regularization to minority classes. This approach enables the loss function to prioritize minority classes while allowing the model to learn effectively from majority classes. The formula is as follows:

$$L_{LDAM}((x, y), f) = -\log \frac{u}{u + \sum\limits_{j \neq y} e^{z_y}} \tag{17}$$

Here, x is the sample, and y is the corresponding label, $z_y = f(x)y$ represents the model's output for a specific sample, $z_y = f(x)y$, $\Delta_j = \frac{C}{n_j^{\frac{1}{4}}}$, j are specific classes, n_j is the number of samples in that class, C is a constant.

Focal loss addresses class imbalance by down-weighting the majority class. This adjustment encourages the model to concentrate on hard-to-classify samples, often from the minority classes. As a result, the model enhances its attention to these minority classes. The formula for focal loss is as follows:

$$FL(p_t) = -(1 - p_t)^{\gamma} \log(p_t) \tag{18}$$

Here, p_t is the probability score predicted by the model, γ is the focusing parameter.

The integration of LDAM loss allows for a nuanced adjustment of the loss calculation, ensuring that each class's contribution is proportional to its representation in the dataset.

Empirical results have demonstrated that the LMF loss function mitigates the adverse effects of sample imbalance and improves overall model accuracy. LMF loss has shown superior performance in various classification tasks compared to traditional loss functions, mainly where datasets exhibit class imbalance.

The LMF loss function's intelligent adjustment of learning focus sets a new standard for developing robust recognition systems. Its potential applications extend beyond underwater sonar target identification to other domains where imbalanced datasets are standard, such as medical diagnostics and financial fraud detection.

3. Experimentation and Analysis

To evaluate the effectiveness of BAHTNet under actual underwater conditions, we employed a recognized public underwater acoustic dataset that enabled qualitative and quantitative comparative analyses. Our examination delved into the performance of BAHTNet across different configurations, considering the various depths and complexities of the network architecture.

Furthermore, we conducted a comparative study using several well-established models that are prevalent in underwater acoustic target recognition. We conducted a comparative analysis that included benchmarking against ResNet, EfficientNet, and DenseNet models to scrutinize the accuracy, robustness, and adaptability of BAHTNet to the nuances of underwater acoustic data.

This comprehensive assessment aimed to establish the standing of BAHTNet among existing technologies and to demonstrate its potential for superior performance in real-world underwater detection applications.

3.1. Experimental Dataset

3.1.1. DeepShip Dataset

To systematically evaluate the model's performance, we utilized the DeepShip dataset from Northwestern Polytechnical University, which is focused on underwater acoustic analysis. This comprehensive dataset includes a variety of 265 vessels, such as cargo ships, passenger ships, oil tankers, and tugs, recorded in the Georgia Strait Delta at depths ranging from 141 to 147 m, targeting vessels within a 2-km radius of the sonar equipment.

In the preprocessing stage, all the WAV audio files from the dataset were standardized to a uniform sampling rate of 22,050 Hz. The raw underwater acoustic data were then segmented into 5-s clips, resulting in over 30,000 annotated audio samples ready for analysis.

To ensure a balanced and efficient training process and to mitigate the risk of overfitting, the dataset was meticulously divided into training, validation, and test sets in an 8:1:1 ratio. This partitioning strategy ensures the model's practical learning and generalization capabilities, thus leading to reliable recognition accuracy.

The specifics of the dataset division are detailed in Table 2, providing a clear breakdown of the distribution across the different subsets, which is crucial for understanding the dataset's composition and impact on the model's training and validation processes.

Table 2. Particulars of the four classifications in the Deepship dataset after pre-processing.

Class	Target	Total	Training Set	Validation Set	Testing Set
1	Cargo ship	7621	6097	762	762
2	Passenger Ship	9211	7369	921	921
3	Oil Tanker	8776	7022	877	877
4	Tug	8085	6467	809	809

3.1.2. ShipsEar Dataset

The ShipsEar Database is a standard for benchmarking underwater acoustic target recognition and is extensively used in scientific studies. It comprises a collection of noise recordings from the Atlantic coast of Spain, featuring a variety of ship noises as well as a range of artificial and natural sounds. The database comprises 90 WAV recordings, divided into five distinct classes.

In the preprocessing phase, the audio was standardized to a sampling rate of 22,050 Hz and segmented into 5-s clips, yielding 2223 annotated audio samples. These were then allocated into training and test sets, with 1778 samples reserved for training and 445 for testing, following an 80:20 ratio to ensure a robust evaluation of the model's performance.

The detailed distribution of the dataset is outlined in Table 3, which provides an insightful account of the division, facilitating a clear understanding of the data composition and its allocation for training and testing purposes.

Table 3. The five classes are included in the ShipsEar database in detail.

Class	Target	The Number of Samples
Class A	Background noise recordings	224
Class B	Dredgers/Fishing boats/Mussel boats/Trawlers/Tugboats	52/101/144/32/40
Class C	Motorboats/Pilot boats/Sailboats	196/26/79
Class D	Passenger ferries	843
Class E	Ocean liners/Ro-ro vessels	186/300

Compared to the DeepShip dataset, the ShipsEar dataset presents challenges due to its smaller sample size and a more significant class imbalance. The pronounced imbalance can heighten the risk of overfitting, as the model might learn to predict the majority class more accurately at the expense of the minority classes, thus complicating the identification of less frequently occurring classes.

This research, therefore, prioritizes addressing these challenges by focusing on strategies to mitigate the effects of small sample size and class imbalance within the ShipsEar dataset. Techniques may include data augmentation, re-sampling methods, or advanced modelling approaches that enhance the model's ability to learn from imbalanced datasets.

Additionally, the DeepShip dataset, with its more extensive and more balanced collection of samples, will serve as a critical tool to evaluate the model's generalization capabilities. By testing the model on this dataset, we can gain insights into how well it performs on a broader range of underwater acoustic data and ensure that it can accurately recognize targets in various real-world conditions.

3.2. Experimental Setup

We utilized the Lion optimizer with a momentum value of 0.9 during the model training process. This choice effectively reduced noise interference in the samples. The model underwent 300 training epochs, beginning with an initial learning rate of 0.0004. We adjusted this learning rate using a cosine decay function to optimize the learning speed. The batch size was set to 16, and the LMF loss function served as the primary evaluation metric. In the LMF loss, we configured the parameters as follows: $\alpha = 1$, $\beta = 1$, $\gamma = 2$, and s = 30. Additionally, the temperature parameter factor τ in XCAttention was set to 1.

3.3. Experimental Environment and Evaluation Metrics

To assess the network's ability to recognize patterns within the dataset, we utilized precision, recall, and F1-Score as evaluation metrics. Their respective mathematical formulas are as follows:

To ensure the reliability and reproducibility of our experimental findings, we pursued a rigorous and systematic research methodology. The experimental setup involved a thorough training and evaluation process for various models, with a dual focus on qualitative and quantitative performance assessments. We engaged in exhaustive comparative analyses to rigorously test the efficacy of the algorithms in question.

The computational setup for our study was equipped with a robust configuration, including a Windows 11 operating system, an Intel Core i7-12700H processor, 32 GB of RAM, an NVIDIA GeForce GTX 3070TI GPU, and PyTorch version 1.4.0, ensuring a powerful platform for executing our complex model training and evaluation tasks.

In the subsequent sections, we delve deeper into the experimental outcomes, presenting the principal metrics for gauging model accuracy. Precision is the ratio of correctly predicted positive observations to the total number of optimistic predictions made by the model. Recall, also known as sensitivity, measures the proportion of actual positives that were correctly identified. The $F1 - score$, a critical metric in classification, provides a balance between precision and recall, offering a single score that encapsulates a model's performance in both dimensions. Support represents the count of actual occurrences of each class in the dataset.

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

$$Recall = \frac{TP}{TP + FN} \tag{20}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{21}$$

In this context, TP represents the correctly predicted outcomes, FP refers to instances where predictions are incorrectly assigned to this category from other categories, and FN indicates cases where the labels of this category are misclassified as labels of other categories. These metrics are integral to evaluating the model's proficiency in pattern recognition within the dataset, providing a comprehensive view of its classification capabilities.

3.4. Ablation Study

3.4.1. Feature Ablation Experiment

In this experiment, various features involved in the BAHTNet model were studied. Among them, MFCC (Mel-Frequency Cepstral Coefficients) is a feature that simulates the nonlinear processing of the human ear and is commonly used in the field of audio processing; 3D_MFCC is an extended form based on MFCC; FBank (Filter Bank) represents the spectral attributes of the signal by applying a series of triangular filters on the power spectrum; CQT (Constant-Q Transform) conducts frequency analysis on underwater acoustic signals under specific settings and can effectively capture the frequency components of the signal. These features were selected for the experiment because they have certain representativeness in traditional underwater acoustic target recognition research and each has unique advantages and characteristics.

Table 4 by comparing the recognition accuracy rates of the model under different features, the degree of dependence and utilization effect of the BAHTNet model on different features can be reflected. For example, a higher accuracy rate indicates that the corresponding feature can better capture the key information of underwater targets and provide strong support for the accurate recognition of the model, while the differences in accuracy rates can reveal the performance of the model when fusing different features, as well as the influence of different feature combinations on the generalization ability and robustness of the model, thereby helping us deeply understand the characteristics and advantages of the BAHTNet model.

Table 4. Recognition accuracy of the BAHTNet model with different features on the ShipsEar dataset.

Model	MFCC	3D_MFCC	FBank	CQT	CM3F
RESNET18	91.46	92.82	95.31	94.62	96.65
RESNET34	90.60	92.63	95.08	94.82	96.43
RESNET50	90.37	91.94	95.08	93.52	96.19
RESNET101	88.53	91.72	94.86	93.06	94.87
Efficientnet_B0	91.04	90.82	95.74	94.40	96.19
Efficientnet_B1	90.82	90.60	95.30	94.63	95.97
Efficientnet_B2	89.70	90.37	94.64	94.40	95.08
Efficientnet_B3	89.58	89.91	94.18	93.29	94.17
Densnet121	90.17	92.14	96.19	94.63	97.30
BAHTNet	92.61	93.73	97.09	96.88	99.78

In the ShipsEar dataset evaluation, a noticeable variation in performance among models was observed when utilizing different feature sets. Specifically, using Mel-Frequency Cepstral Coefficients (MFCC) features resulted in moderate accuracy levels for the models, generally between 89% and 92%. The observation that the models achieved only moderate accuracy levels when using MFCC features alone indicates that relying solely on these features may not sufficiently capture the critical information for identifying underwater sonar targets.

The introduction of 3D MFCC features marginally improved the accuracy of the standard MFCC features, yet the improvement was not substantial. A more significant boost in accuracy, ranging from 90% to 95% for most models, was achieved with Filter Bank

(FBank) features, suggesting their excellent utility in conveying valuable information for target identification.

Furthermore, the Constant-Q Transform (CQT) features demonstrated even higher accuracy rates, typically between 93% and 95% for most models, highlighting their potential as a practical feature set for underwater sonar target identification.

Interestingly, a more profound architecture did not guarantee increased accuracy within the ResNet family of models, indicating that model depth may not linearly correlate with performance in this specific application. Similarly, the accuracy among the EfficientNet models did not show a consistent upward trend from EfficientNet_B0 to B3.

The RESNET18, EfficientNet_B0, DenseNet121, and BAHTNet models were selected for further evaluation based on their exceptional performance when using the Combined Multimodal Feature (CM3F) set. To assess these models' performance more deeply, we employed confusion matrices and t-SNE diagrams to visualize and analyze their classification capabilities. The results of this analysis are depicted in Figure 7, providing a detailed overview of how each model handles the classification of underwater acoustic signals using the CM3F features.

Significant variations in predictive performance were observed after a detailed examination of the confusion matrices and t-SNE diagrams for Densenet121, EfficientNet_B0, and ResNet18. Densenet121 showed inaccuracies in categorization, with its t-SNE visualization indicating category overlaps, suggesting limitations in distinguishing between similar classes. While EfficientNet_B0 showed robust discrimination for specific categories, it still needed enhancements to improve its overall performance. ResNet18 demonstrated a higher incidence of misclassification, and its t-SNE visualization lacked well-defined class separation, indicating weaker feature discrimination capabilities.

In contrast, the BAHTNet model showcased superior predictive accuracy across all categories, with a markedly lower rate of incorrect predictions. The t-SNE diagram for BAHTNet revealed well-defined and non-overlapping category distributions, indicating its proficiency in feature learning and classification. BAHTNet's ability to capture features effectively in complex data distributions translates to improved performance in tasks like underwater sonar target recognition, substantially reducing false positives and negatives and enhancing reliability in practical applications.

In the experimental evaluation, BAHTNet's performance with MFCC features reached an accuracy rate of 92.61%, outperforming other models. The utilization of 3D_MFCC features enhanced this accuracy to 93.73%. With FBank features, BAHTNet achieved an accuracy rate of 97.09%, and with CQT features, it attained 96.88%. Most notably, when employing the newly integrated CM3F features, BAHTNet realized an accuracy rate of 99.78%, underscoring its capability for robust feature extraction and recognition by effectively synthesizing information from multiple feature sets.

To elucidate the distinct advantage of CM3F features, a t-SNE diagram for the BAHTNet model under various feature extraction conditions was created, as depicted in Figure 8. This visualization further substantiates the effectiveness of the CM3F feature set in enhancing the model's classification accuracy and reliability in underwater sonar target recognition tasks.

Upon a thorough analysis of the t-SNE diagrams for MFCC, 3D MFCC, FBank, CQT, and CM3F features, distinct patterns in category discrimination emerged. The 3D MFCC showed only moderate performance with considerable overlap between categories. CQT features offered a more precise distribution, yet some marginal confusion remained. FBank features improved separation for some categories, but further enhancement was needed for overall discrimination. The MFCC features, while foundational, were the least distinct, lacking sharp category boundaries.

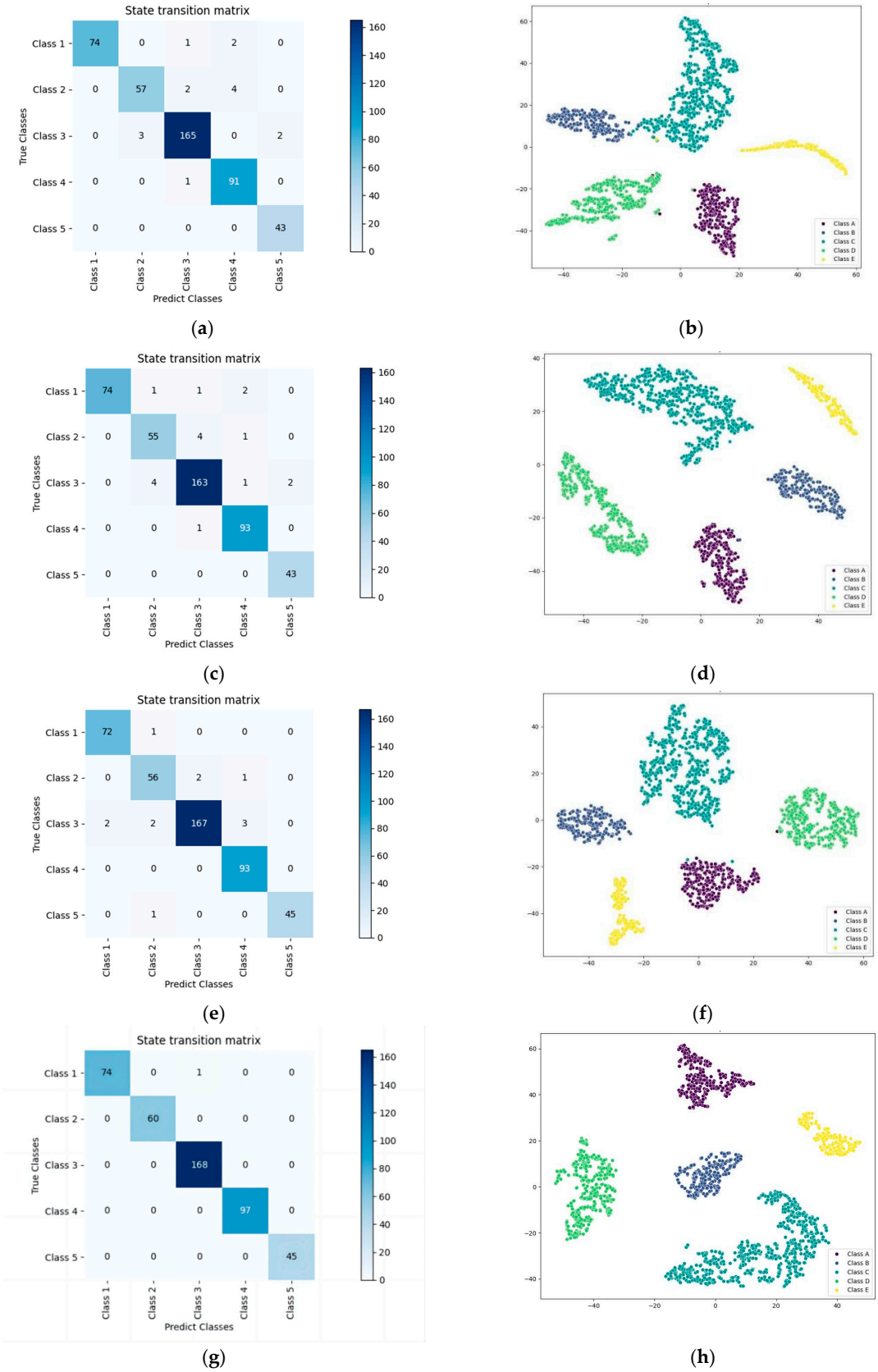

Figure 7. Confusion matrices and t-SNE diagrams of different models on the ShipsEar training dataset. (**a**) RESNET18 confusion matrix, (**b**) t-SNE visualization for the RESNET18 model, (**c**) the Efficientnet_B0 confusion matrix, (**d**) t-SNE visualization for the Efficientnet_B0 model, (**e**) Densnet121 confusion matrix, (**f**) t-SNE visualization for the Densnet121 model, (**g**) BAHTNet confusion matrix, (**h**) t-SNE visualization for the BAHTNet model.

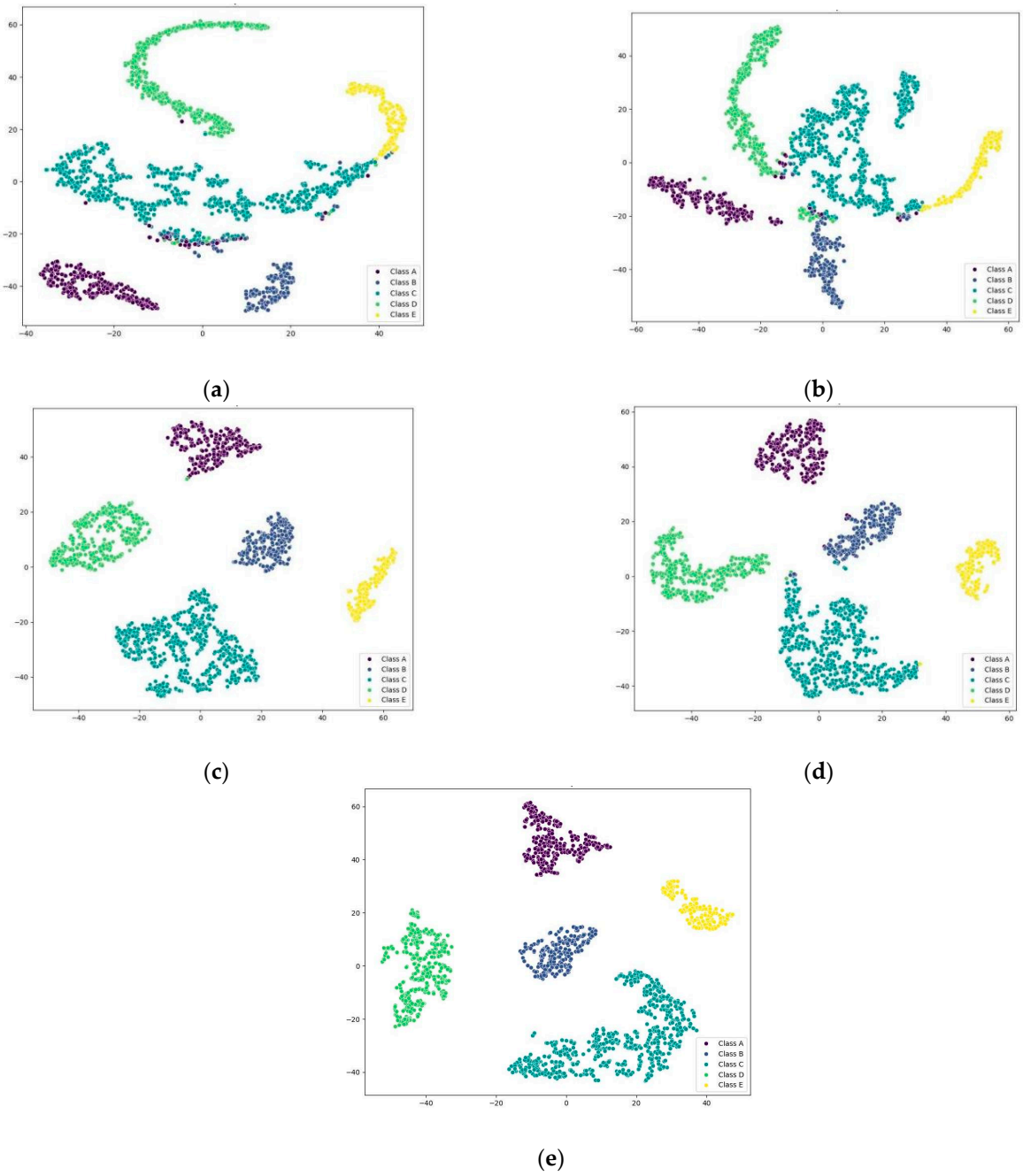

Figure 8. t-SNE visualization of output feature vectors from the BAHTNet model on different features of the ShipsEar training dataset. (**a**) MFCC, (**b**) 3D MFCC, (**c**) FBank, (**d**) CQT, (**e**) CM3F.

In striking contrast, the CM3F features stood out with their distinct category distributions and clear boundaries, significantly reducing overlap and confusion. The CM3F's tighter clustering provided a more precise and consistent representation, allowing for more detailed information extraction and improved inter-category differentiation. The tighter clustering provided by the CM3F's features was crucial for bolstering the model's recognition and classification capabilities.

In conclusion, the adaptability of models to different features varied significantly in underwater sonar target identification. While most models performed well with CQT features, the integrated CM3F features offered a notable advantage. The BAHTNet model, in particular, excelled with various features, especially when using CM3F features, achieving remarkably high accuracy. This suggests that the BAHTNet model, in combination

with CM3F features, is likely to produce more precise and reliable recognition outcomes. These advancements are poised to strengthen underwater detection and marine monitoring applications, enhancing target identification precision, operational efficiency, and safety.

3.4.2. Module Ablation Study

We performed ablation studies to delve into the individual impacts and overall performance of the model's components. This systematic approach involved sequentially turning off certain modules, which enabled us to gauge the specific contributions of each to the model's performance, as detailed in Table 5. The insights gained from these studies illuminated the model's operational dynamics and pointed towards avenues for potential enhancement.

Table 5. Ablation study of different modules in the ShipsEar sample set.

Model	Class	Precision	Recall	F1-Score	Support
BAHTNet	Class A	0.99	1.00	0.99	74
	Class B	1.00	1.00	1.00	60
	Class C	1.00	0.99	0.99	169
	Class D	1.00	1.00	1.00	97
	Class E	1.00	1.00	1.00	45
	Average	0.998	0.998	0.996	445
S1	Class A	0.94	1.00	0.97	74
	Class B	0.96	0.92	0.94	60
	Class C	0.97	0.98	0.98	169
	Class D	1.00	0.96	0.98	97
	Class E	1.00	1.00	1.00	45
	Average	0.973	0.972	0.974	445
S2	Class A	0.96	1.00	0.98	74
	Class B	0.98	0.95	0.97	60
	Class C	0.98	0.99	0.99	169
	Class D	1.00	0.98	0.99	97
	Class E	1.00	0.98	0.99	45
	Average	0.984	0.980	0.984	445
S3	Class A	0.99	1.00	0.99	74
	Class B	1.00	0.97	0.98	60
	Class C	0.98	0.99	0.99	169
	Class D	1.00	0.98	0.99	97
	Class E	1.00	1.00	1.00	45
	Average	0.989	0.988	0.990	445
S4	Class A	0.99	1.00	0.99	74
	Class B	1.00	0.99	0.99	60
	Class C	0.99	0.99	0.99	169
	Class D	1.00	0.98	0.99	97
	Class E	1.00	1.00	1.00	45
	Average	0.994	0.990	0.991	445

We configured three distinct scenarios for the ablation exercises:

1. In scenario S1, the CBCARM module was excised from the model. The extracted features bypassed this module and were fed directly into the XCAT and Kan networks. In the S1 experiment, by excising the CBCARM module and directly feeding features into the XCAT and Kan networks, this action allowed us to isolate the influence of the CBCARM module on the model's performance.
2. Scenario S2 was designed to assess the role of the XCAT module by removing it. The features were processed through the CBCARM module, if present, and then forwarded to the Kan network for classification, thus highlighting the XCAT module's contribution.

3. Scenario S3 was crafted to evaluate the combined effect of the CBCARM and XCAT modules by inputting features into both simultaneously, followed by classification. This scenario aimed to explore the synergistic performance of these modules and their joint impact on the model's efficacy.
4. Scenario S4 assesses the effect of the classifier by replacing the Kan network with an MLP classifier.

Through these meticulously crafted experimental setups, we sought to deconstruct the model's performance to its core components, providing a micro-level analysis to inform macro-level optimization strategies. The results of these ablation studies are instrumental in understanding which modules are critical for the model's success and how these components interact to produce the best outcomes in underwater sonar target recognition tasks.

Analyzing the experimental setups presented in Table 5 sheds light on the distinct roles of each module within the model. In the S1 experiment, removing the CBCARM module and directly feeding features into the XCAT and Kan networks led to a notable decrease in Precision, Recall, and F1-Score, particularly in Recall. The significant drop in performance, especially in Precision, Recall, and F1-Score, in the experimental group C1 when the CBM module was absent strongly emphasizes the crucial contribution of the CBCARM module to the model's overall performance.

The CBCARM module is composed of the CBM and CPCA components. The CBM specializes in enhancing the detection and extraction of boundary information in variable regions, preventing data loss and facilitating more precise target identification. CPCA dynamically assigns attention weights across channels and spatial dimensions, focusing the model's gaze on salient areas and objects. The combined action of these components significantly bolsters the model's feature extraction and representation capabilities. Without the CBCARM module, the model's feature extraction and prioritization of crucial information are severely compromised, leading to diminished recognition accuracy.

In the S2 experiment, the XCAT module's removal and the subsequent change in processing sequence—where features were processed by the CBCARM module followed by classification with the Kan network—resulted in a less severe but still notable decline in the average metrics compared to the BAHTNet model. The XCAT module, consisting of the XCAttention and LPI components, is crucial in enhancing the traditional Transformer's feature extraction capabilities and improving processing performance through local patch interactions. The absence of the XCAT module impacts the model's ability to extract and process features effectively.

The S3 experiment, where features were input into the CBCARM and XCAT modules before final classification, showed an average F1-Score improvement over S1 and S2, approaching the BAHTNet model's performance. The fact that the average F1-Score in the S3 experiment improved over S1 and S2 when features were input into both the CBCARM and XCAT modules before final classification indicates that the collaborative effect of the CBCARM and XCAT modules is indispensable for optimal model performance. The CBCARM module's proficiency in feature extraction and enhancement, when combined with the XCAT module's efficient processing, creates a synergistic effect that enhances the model's adaptability to underwater sonar data and improves recognition capabilities.

By comparing BAHTNet and S4, the performance of the Kan network in BAHTNet is superior to the traditional MLP. MLP has difficulty extracting information from noisy and complex underwater sonar signals because it lacks Kan's dynamic grid adjustment mechanism. In cases of strong noise or complex frequencies, Kan can adapt but MLP cannot. Additionally, MLP's simple structure without an effective regularization loss calculation makes it prone to overfitting, especially with the limited and noisy datasets in underwater

acoustic target recognition. In contrast, Kan's regularization loss controls complexity and prevents overfitting for better generalization. Therefore, the combination of CBCARM, XCAT modules, and the use of Kan instead of MLP gives BAHTNet enhanced capabilities and higher accuracy than traditional MLP-based models.

In conclusion, the BAHTNet model's integration of the CBCARM and XCAT modules provides complementary strengths. This architecture elevates the model's performance and accuracy, enabling more effective processing of underwater sonar data and more precise recognition of underwater acoustic targets. The results highlight the integrated approach's distinct advantage in enhancing underwater sonar target identification.

3.4.3. Ablation Study of the CBCARM Module

An ablation study has been meticulously planned to deeply assess the impact of the CBM and CPCA components within the CBCARM module on the BAHTNet model's overall performance. This research endeavor is set to lay the groundwork for future model refinements. The methodology encompasses a series of systematic steps:

Initially, a control setup is established by maintaining the intact BAHTNet model. Utilizing the complete BAHTNet model as a control group is a benchmark under consistent experimental conditions against which variations will be compared. Subsequently, experimental group C1 is formulated by selectively omitting the CBM module. Consequently, the features extracted bypass the CBN's processing and are directly channeled into the XCAT and Kan networks. The resultant fluctuations in key performance metrics—such as accuracy, recall, and F1-Score—will be meticulously chronicled for group C1.

Subsequently, experimental group C2 is introduced by eliminating the CPCA module while preserving the remainder of the model's structure. The alterations in performance metrics for this configuration will be equally well-documented. The ablation study will be executed through this structured approach, with the findings succinctly detailed in Table 6.

Table 6. Analysis of the role of components in the CBCARM module of the BAHTNet model in underwater sonar target recognition.

Model	ShipsEar Dataset			
	Accuracy	**Precision**	**Recall**	**F1**
BAHTNet	0.998	0.994	0.998	0.996
X1	0.980	0.980	0.982	0.980
X2	0.984	0.984	0.980	0.984

This ablation study is designed to provide critical insights into the individual contributions of the CBM and CPCA components, informing targeted enhancements to the model's architecture and underwater sonar target identification performance.

The BAHTNet model has proven its mettle in underwater sonar target recognition, achieving high precision, recall, and F1-Scores, with particularly impressive scores of 0.998 for precision and recall. These results testify to the model's prowess in feature extraction and its ability to accurately classify underwater sonar targets.

However, in the experimental group C1, the absence of the CBM module led to a perceptible dip in performance, with precision, recall, and F1-Scores slipping to 0.980, 0.982, and 0.980, respectively. This drop highlights the CBM module's indispensable role in capturing boundary information, which is pivotal for preserving critical data and precisely identifying targets. Without the CBM module, the model's precision and recall suffered, affecting its overall effectiveness.

Group C2, with the CPCA module disabled, mirrored this decreased performance trend, with scores of 0.984 for precision, 0.980 for recall, and 0.984 for the F1-Score. The

CPCA module's contribution to dynamically focusing attention on critical areas and objects is evident, as it bolsters the model's feature extraction and representation, enhancing its overall performance.

In essence, the CBM and CPCA modules of the BAHTNet model are not just components; they are collaborative pillars that strengthen the model's accuracy and reliability in underwater sonar target recognition. The CBM's boundary perception and extraction, coupled with the CPCA's attention allocation and feature optimization, create a synergistic effect that propels the BAHTNet model to the forefront of underwater sonar target recognition technology.

3.4.4. Ablation Study of the XCAT Module

An ablation study has been meticulously planned to delve into the individual contributions of the components within the XCAttention Transformer Encoder (XCAT) module to underwater sonar target recognition. This study will unfold in a series of structured steps:

1. Establishment of a Control Group: The study begins by utilizing the complete BAHTNet model as a control group. This step lays down a uniform experimental baseline, ensuring that subsequent comparisons are made under consistent conditions.
2. Creation of Experimental Group X1: In this phase, the XCAttention component is excised from the XCAT module, forming experimental group X1. The performance metrics of this group will be closely monitored to evaluate the impact of the XCAttention component's removal on the model's overall performance.
3. Formation of Experimental Group X2: Experimental group X2 is subsequently established by removing the Local Patch Interaction (LPI) module from the XCAT module. Similar to group X1, the performance metrics of this group will be documented, allowing for an assessment of the LPI module's influence.
4. Performance Metrics Comparison: The final step involves a comparative analysis of the performance metrics between the control group and the two experimental groups (X1 and X2). This comparison will spotlight the distinctive contributions of both the XCAttention and LPI modules to the underwater sonar target recognition process.

The findings of this ablation study, presented in Table 7, are expected to provide valuable insights into the importance of each XCAT module component, thereby guiding further optimization and enhancement of the BAHTNet model's underwater sonar target recognition capabilities.

Table 7. The influence of XCAT module components in the BAHTNet model on underwater sonar target recognition.

Model	ShipsEar Dataset			
	Accuracy	**Precision**	**Recall**	**F1**
BAHTNet	0.998	0.994	0.998	0.996
X1	0.987	0.987	0.984	0.986
X2	0.988	0.988	0.988	0.988

The BAHTNet model distinguishes itself with its exceptional underwater sonar target recognition proficiency, achieving outstanding average precision and recall rates that approach 0.998. These stellar metrics testify to the model's formidable abilities in feature extraction and classification, ensuring the reliable identification of various underwater sonar targets.

The model demonstrates its versatility and accuracy by consistently recognizing targets across different categories, establishing itself as a dependable solution for underwater sonar applications. However, the experimental group X1, devoid of the XCAttention

module, experiences a modest decline in performance. The average precision, recall, and F1-Scores for X1 are 0.987, 0.984 and 0.986, respectively. This dip in performance underscores the XCAttention module's pivotal role in augmenting the Transformer's self-attention mechanism, thereby facilitating the efficient extraction of crucial information.

On the other hand, group X2, which lacks the LPI module, also records a slight performance decrease relative to the BAHTNet model, with average scores of 0.988 for precision, recall, and F1. The fact that group X2, which lacks the LPI module, also records a slight performance decrease relative to the BAHTNet model suggests that the LPI module is instrumental in seizing local information and nuances through its sliding window approach, thereby bolstering the model's overall recognition prowess.

In essence, the XCAttention and LPI modules within the BAHTNet model operate in concert to heighten the precision and reliability of underwater sonar target recognition. The attention module is vital for distilling critical information, while the LPI module enriches this process by capturing localized details. This collaborative functionality renders the BAHTNet model significantly superior for accurately recognizing underwater sonar targets.

3.4.5. Ablation Study on Loss Function

This study aims to quantify the influence of the Large Margin Focal (LMF) Loss on the BAHTNet model's underwater sonar target recognition performance by leveraging the ShipsEar dataset. To this end, we have compared the outcomes of different loss functions, namely cross-entropy (CE), contrastive loss (CB), focal loss (FL), and the LMF loss itself.

For this comparative analysis, we have developed three variant models of BAHTNet, each employing a distinct loss function: BAHTNet-CE with the cross-entropy loss, BAHTNet-FL with the focal loss, and BAHTNet-CB with the contrastive loss. These models were pitted against the original BAHTNet, equipped with the LMF loss, to evaluate their comparative effectiveness.

The ShipsEar dataset underwent a meticulous preprocessing phase, standardizing the audio to uniform sampling rates and chopping it into 5-s segments. Uniformity in training was maintained across all models, adhering to the BAHTNet's original training protocol. The uniform training across all models included consistent parameters for momentum, optimizer, training epochs, learning rate schedule, and batch size, ensuring an equitable comparison.

We meticulously tracked the models' loss values, accuracy, and other pertinent metrics throughout the training. Post-training, the performance of each model was gauged using a shared test set, focusing on metrics, such as classification accuracy, recall, and the F1-Score. This rigorous evaluation was designed to provide a clear and unbiased comparison of how each loss function impacts the model's ability to recognize underwater sonar targets.

The impact of different loss functions on the BAHTNet model's performance, as evaluated on the ShipsEar dataset, is evident in Table 8. The BAHTNet-CE model, which employs the conventional cross-entropy loss function, reports Precision, Recall, and F1-Scores of 0.935, 0.920, and 0.932, respectively. These scores indicate that while cross-entropy is widely utilized, it may not be optimal for handling imbalanced datasets. The BAHTNet-FL model, incorporating the focal loss function, achieves higher metrics by focusing more on challenging samples, with a Precision of 0.949, a Recall of 0.946, and an F1-Score of 0.950. The fact that the BAHTNet-FL model, incorporating the focal loss function, achieves higher metrics with a Precision of 0.949, Recall of 0.946, and an F1-Score of 0.950 suggests that the focal loss function is better at enhancing the model's ability to classify difficult instances.

Table 8. BAHTNet model metrics under different loss functions on the ShipsEar dataset.

Model	ShipsEar Dataset			
	Accuracy	Precision	Recall	F1
BAHTNet-CE	0.935	0.938	0.920	0.932
BAHTNet-FL	0.949	0.952	0.946	0.950
BAHTNet-CB	0.980	0.978	0.982	0.980
BAHTNet	0.998	0.994	0.998	0.996

The BAHTNet-CB model, which applies the CB loss function, further improves these metrics to a precision and recall of 0.980 and 0.982, respectively, along with an F1-Score of 0.980. The performance of the BAHTNet-CB model, with a precision and recall of 0.980 and 0.982, respectively, along with an F1-Score of 0.980, indicates that the CB loss function effectively adjusts for sample imbalance by adapting weights according to the sample distribution.

Most notably, the BAHTNet model equipped with the LMF loss function substantially outperforms the other models, with Precision and Recall reaching 0.998 and an F1-Score of 0.996. The LMF loss function's ability to assign weights dynamically based on the dataset's class distribution is critical to its success in reducing the impact of sample imbalance and enhancing recognition accuracy.

In conclusion, the LMF loss function's distinct advantages are underscored by its ability to significantly boost the BAHTNet model's accuracy and reliability in underwater acoustic target recognition.

3.5. Performance Evaluation

In this section, we comprehensively compare the BAHTNet model's performance with that of other state-of-the-art target recognition models. The evaluation is conducted under uniform experimental conditions to ensure a fair and accurate assessment. Key performance metrics are examined, including accuracy, precision, recall, and F1-Score. Additionally, we analyze various aspects of the models' capabilities, such as their generalization to new data, handling of class imbalance, computational efficiency, and robustness in real-world applications. This comparative study aims to provide insights into each model's relative strengths and potential areas for improvement.

3.5.1. Parameter Analysis

BAHTNet's innovative architecture has been widely recognized for its exceptional performance in target recognition tasks. The loss function is pivotal in evaluating this performance, quantifying the difference between the model's predictions and the actual labels. This function is indispensable for understanding the model's learning trajectory and identifying areas for improvement. Figure 9 visually represents the loss function's performance over training iterations, highlighting BAHTNet's convergence towards minimizing prediction errors. The graph illustrates the model's ability to adjust its parameters effectively, reflecting its robust learning process and the potential for high accuracy in practical applications.

Analysis of the loss function graph for the BAHTNet model yields insightful observations regarding its performance. The graph displays a continuous and significant reduction in training and validation loss, underscoring the model's robust optimization and learning capabilities. This downward trend indicates the model's ability to generalize well to new data, a crucial characteristic for its application in real-world scenarios.

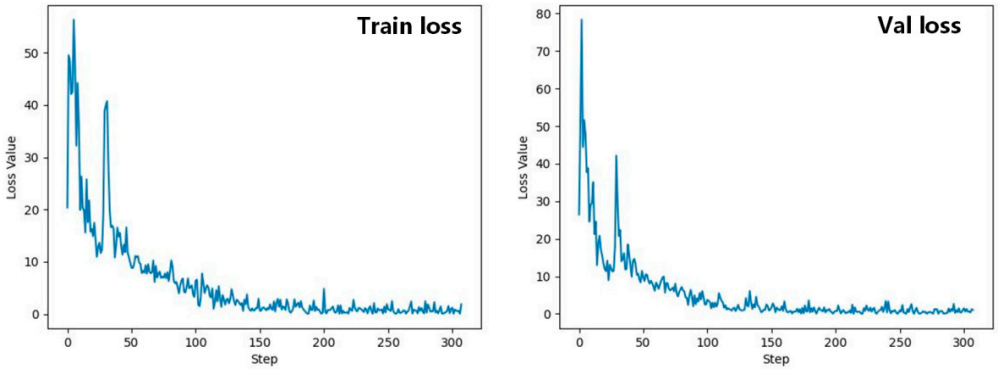

Figure 9. BAHTNet loss function variation chart.

Notably, the graph does not exhibit excessive overfitting or underfitting, suggesting that the model has struck a balance between capturing the nuances of the training data and maintaining its ability to perform on unseen data. The absence of overfitting is significant, as it ensures that the model does not become too tailored to the training dataset, thus preserving its predictive power on new information.

The model demonstrates remarkable stability, with the loss function graph reflecting a consistent progression towards minimal error. This stability is a testament to the model's effectiveness and provides a strong foundation for ongoing optimization efforts. Moreover, the model's generalization ability is a significant asset, making it a reliable tool for addressing various real-world challenges.

The BAHTNet model's performance metrics and generalization capabilities suggest that it is well-positioned for further refinement and application in diverse domains. Future work could enhance the model's interpretability, expand its scalability, and explore its integration with other data modalities to solve more complex problems.

3.5.2. Computational Bottleneck Analysis

Securing extensive and diverse underwater acoustic datasets is a notable challenge in the field, with the Deepship and ShipsEar datasets being the most prominent public resources available. The ShipsEar dataset, while featuring a range of ship types, is constrained by its modest size. We augmented our sample collection to mitigate this limitation, enabling a more robust investigation into the relationship between dataset size and model performance.

Our experiments were designed to assess how the volume of data affects the recognition accuracy of BAHTNet and other leading models. Furthermore, we sought to uncover any computational constraints that might arise from the dataset's size. By understanding these factors, we aim to contribute to developing more efficient models and inform the construction of more prominent and more representative datasets for future research.

Our analysis, as depicted in Table 9, reveals a clear trend in recognition accuracy improvement for models, such as RESNET18, EfficientNet_B0, and DenseNet121, with the growth of dataset size. Notably, BAHTNet distinguishes itself by achieving a higher baseline accuracy even at the smallest sample label percentage of 1% and remarkably scales up to an accuracy of 99.78% with the entire dataset.

The superior performance of BAHTNet is mainly due to its innovative architecture that integrates the CBCARM and XCAT modules, complementing each other's capabilities and enhancing the model's overall feature extraction and pattern recognition. This synergistic integration is pivotal for BAHTNet's exceptional accuracy, especially in limited data, like the ShipsEar dataset.

Table 9. Recognition accuracy of different models at various training dataset label ratios.

Model	Sample Label Percentage			
	1%	10%	50%	100%
RESNET18	44.11%	75.08%	92.59%	96.65%
Efficientnet_B0	44.73%	75.55%	94.19%	96.19%
Densnet121	47.23%	80.94%	93.29%	97.30%
BAHTNet	49.87%	85.43%	94.38%	99.78%

BAHTNet's demonstrated adaptability and accuracy present significant implications for underwater acoustic target recognition, suggesting its potential for real-world applications where data scarcity is common. Future work could further refine the model's architecture to improve computational efficiency and explore strategies for expanding and diversifying available datasets to bolster the field's research and practical applications.

3.5.3. Model Recognition Accuracy Analysis

We conducted a rigorous experimental comparison to evaluate the performance of BAHTNet against existing mainstream target recognition models. BAHTNet was trained and validated alongside other models using the same datasets, evaluation metrics, and training protocols to ensure a fair and controlled assessment.

Preliminary results indicate that BAHTNet performs remarkably, outperforming other models across various metrics, such as accuracy, precision, and recall. The unique architectural features of BAHTNet, including its advanced feature extraction and loss function mechanisms, contribute to its enhanced capabilities.

While BAHTNet shows promise, the experimental comparison also provides insights into areas where further optimization may be beneficial. Ongoing research will focus on refining the model's architecture, exploring additional datasets, and potentially integrating complementary technologies to bolster its performance further.

As detailed in Table 10, our comparative analysis unveils a stark contrast in the performance of various target recognition models when subjected to the same experimental conditions. Traditional models, including Yamnet, VGGish, ADCNN, CRNN9, and LSTM-based architectures, consistently underperformed across critical metrics, signifying their limitations in accuracy and reliability.

Table 10. Comparative results of different models on the ShipsEar dataset.

Model	Accuracy	Precision	Recall	F1-Score
Yamnet	0.7872	0.6864	0.8358	0.7264
VGGish	0.8675	0.8330	0.8582	0.8427
ADCNN	0.9358	0.9236	0.9749	0.9469
CRNN9	0.9117	0.7764	0.9543	0.8463
LSTM-based	0.9477	0.9132	0.9814	0.9449
VFR	0.9850	0.9850	0.9850	0.9850
Mobile_Vit	0.9850	0.9837	0.9840	0.9838
Hybrid model	0.9878	0.9839	0.9949	0.9878
BAHTNet	0.9980	0.9940	0.9980	0.9960

While showing improvement, advanced models, like VFR, Mobile_ViT, and Hybrid, still could not compete with BAHTNet's outstanding results. With an accuracy, precision, and recall rate of 0.9980 and an F1-Score of 0.9960, BAHTNet's performance is superior and underscores its robustness in target recognition tasks.

The secret behind BAHTNet's dominance lies in its innovative architecture, particularly the seamless fusion of the CBCARM and XCAT modules. This integration not only

leverages the strengths of each module but also enhances the model's overall capability to learn and generalize from data.

These findings are not just numerical victories; they translate to tangible benefits in real-world applications. BAHTNet's high performance ensures it can be trusted to make accurate predictions, even in complex and noisy environments. Future work will explore how these advantages can be harnessed to solve pressing problems in various domains, further solidifying BAHTNet's position as a leading model in target recognition.

3.6. Generalization Experiment

To determine the efficacy of various models across diverse domains, evaluating their performance on a range of underwater acoustic datasets is essential. This assessment is particularly important for gauging the generalization capabilities of BAHTNet. Our study selected the Deepship dataset as a benchmark for this evaluation, ensuring a controlled and standardized experimental setup.

The purpose of this assessment is twofold: to ensure that the models can accurately predict outcomes on new, unseen data and to identify potential areas for design improvement. By conducting these experiments, we aim to provide a comprehensive understanding of BAHTNet's performance relative to other models and to contribute to its ongoing development.

Table 11 provides a detailed account of the experimental results, offering insights into each model's ability to generalize from the training data to the Deepship dataset. These findings are instrumental in validating the robustness of BAHTNet and guiding future refinements to enhance its predictive accuracy in real-world applications.

Table 11. Comparative results of different models on the Deepship dataset.

Model	Accuracy	Precision	Recall	F1-Score
Yamnet	0.6953	0.6959	0.6930	0.6929
VGGish	0.6685	0.6675	0.6670	0.6656
ADCNN	0.9023	0.9019	0.9046	0.9017
CRNN9	0.8614	0.8649	0.8590	0.8606
MobilenetV2	0.9018	0.9038	0.9010	0.9012
VFR	0.9380	0.9440	0.9430	0.9430
Mobile_Vit	0.9107	0.9010	0.9103	0.9097
CA_MobilenetV2	0.9350	0.9347	0.9355	0.9349
BAHTNet	0.9457	0.9454	0.9458	0.9456

Our thorough analysis of the Deepship dataset's experimental results unveils a notable disparity in the generalization performance of various target recognition models. Notably, BAHTNet emerged as a standout performer, achieving an accuracy of 0.9457, a precision and recall of 0.9454 and 0.9458, respectively, and an F1-Score of 0.9456. These results testify to BAHTNet's ability to maintain high accuracy and reliability, even when faced with real-world underwater acoustic data complexities.

The underperformance of models, like Yamnet and VGGish, indicates their limitations in capturing the nuances of underwater acoustic environments. Meanwhile, models, such as ADCNN, CRNN9, MobilenetV2, VFR, Mobile_Vit, and CA_MobilenetV2, despite showing some promise, still lag behind BAHTNet's capabilities and require further refinement.

BAHTNet's success is primarily attributed to its innovative design, including integrating CBCARM and XCAT modules and its sophisticated optimization strategies. These elements enhance the model's learning process and its adaptability to diverse datasets.

The BAHTNet model's demonstrated superiority in underwater acoustic target recognition suggests its potential for broader marine monitoring and defense systems applications. Future work may focus on further enhancing its computational efficiency and exploring its integration with other sensor modalities for multimodal analytics.

3.7. Discussion and Analysis

Although the BAHTNet model has achieved impressive results in recognizing underwater acoustic targets, it has challenges. The high cost and effort associated with collecting underwater acoustic data, coupled with unbalanced sample sizes in existing datasets, can impede model training and lead to overfitting.

To counter these issues, future research directions for BAHTNet should include the development of advanced data-augmentation techniques. Specifically, Generative Adversarial Networks can be utilized to synthetically expand the dataset, whereas data transformation and the introduction of controlled noise can enhance its diversity and balance.

Model optimization is another critical area of focus. Streamlining the BAHTNet architecture through compression and quantization techniques is essential for reducing its computational footprint, making it more viable for deployment in resource-limited settings.

Moreover, integrating multimodal data sources such as underwater imagery and sensor information can significantly boost the recognition accuracy and generalization of the model through a multimodal fusion approach.

Field testing in real-world environments will provide invaluable data on the practical performance of the BAHTNet, allowing for iterative refinement based on direct feedback and observed outcomes. Advanced machine-learning methodologies, including transfer and federated learning, can further adapt the model to various acoustic environments and enhance its robustness.

The future advancement of BAHTNet hinges on addressing these data-related challenges, optimizing its structural complexity, and broadening its application scope to solidify its standing as a leading model for underwater acoustic target recognition.

4. Conclusions

The BAHTNet model presented in this study addresses several pivotal challenges in underwater acoustic target recognition. It introduces a novel feature fusion method, CM3F, which combines CQT, MFCC, and FBank to capture comprehensive underwater signal information, thereby enhancing the identification and classification processes. Integrating CBCARM and XCAT modules within BAHTNet's architecture enriches feature extraction, whereas the Kan network substitution for traditional MLPs optimizes classification, captures critical information, and reduces overfitting.

The LMF loss function is a significant innovation that dynamically adjusts weights to counteract sample imbalances, thereby boosting recognition accuracy. The empirical results confirm the superior performance of BAHTNet, with high accuracy and strong generalization capabilities across various datasets, including smaller datasets, where its advantages are particularly evident.

The BAHTNet's consistent downward trend in loss during training and its exceptional accuracy and reliability in real-world evaluations highlight its potential for practical deployment in underwater detection and marine monitoring. It promises to significantly enhance the target recognition precision, operational efficiency, and safety.

The focus will be on further optimizing the BAHTNet's performance and exploring its integration into practical scenarios, such as autonomous underwater vehicles or sonar systems, to harness its full potential in enhancing underwater acoustic target recognition.

Author Contributions: Conceptualization, Y.Z. and M.C.; methodology, Y.Z. and G.X.; investigation, Y.Z. and G.X.; writing—original draft preparation, Y.Z.; writing—review and editing, M.C.; funding acquisition, M.C.; validation G.X.; software, H.C.; formal analysis, H.C. and L.H.; data curation, H.C. and L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Guangxi Science and Technology Base and Talent Project (No. GuikeAD21220098 and AD23026298). We are also thankful for the support of the 2021 Open Fund project of the Key Laboratory of Cognitive Radio and Information Processing of the Ministry of Education (No. CRKL210102), the Beihai City Science and Technology Bureau Project (No. Bei ke He 2023158004), and the Innovation Project of Guangxi Graduate Education (No. YCSW2024344), and the Innovation Project of Guangxi Graduate Education (Optimization of Underwater Acoustic Information Extraction Model Based on Graph Neural Network and Semi-supervised Target Recognition).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

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

References

1. 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]
2. Hong, F.; Liu, C.; Guo, L.; Chen, F.; Feng, H. Underwater acoustic target recognition with a residual network and the optimized feature extraction method. *Appl. Sci.* **2021**, *11*, 1442. [CrossRef]
3. Zhu, C.; Cao, T.; Chen, L.; Dai, X.; Ge, Q.; Zhao, X. High-order domain feature extraction technology for ocean acoustic observation signals: A review. *IEEE Access* **2023**, *11*, 17665–17683. [CrossRef]
4. Yao, Q.; Wang, Y.; Yang, Y. Underwater acoustic target recognition based on data augmentation and residual CNN. *Electronics* **2023**, *12*, 1206. [CrossRef]
5. Zhu, X.; Dong, H.; Rossi, P.S.; Landrø, M. Time-frequency fused underwater acoustic source localization based on contrastive predictive coding. *IEEE Sens. J.* **2022**, *22*, 13299–13308. [CrossRef]
6. Han, R.; Jia, N.; Huang, J.; Guo, S. Joint time-frequency domain equalization of MSK signal over underwater acoustic channel. *Appl. Acoust.* **2022**, *189*, 108597. [CrossRef]
7. Liu, Z.; Peng, D.; Zuo, M.J.; Xia, J.; Qin, Y. Improved Hilbert–Huang transform with soft sifting stopping criterion and its application to fault diagnosis of wheelset bearings. *ISA Trans.* **2022**, *125*, 426–444. [CrossRef]
8. Tian, C.; Zheng, M.; Zuo, W.; Zhang, B.; Zhang, Y.; Zhang, D. Multi-stage image denoising with the wavelet transform. *Pattern Recognit.* **2023**, *134*, 109050. [CrossRef]
9. Di, N.; Sharif, M.Z.; Hu, Z.; Xue, R.; Yu, B. Applicability of VGGish embedding in bee colony monitoring: Comparison with MFCC in colony sound classification. *PeerJ* **2023**, *11*, e14696. [CrossRef]
10. Zhou, C.; Wu, Y.; Fan, Z.; Zhang, X.; Wu, D.; Tao, Z. Gammatone spectral latitude features extraction for pathological voice detection and classification. *Appl. Acoust.* **2022**, *185*, 108417. [CrossRef]
11. Ali, S.; Iqbal, N.; Hafeez, Y. Towards requirement change management for global software development using case base reasoning. *Mehran Univ. Res. J. Eng. Technol.* **2018**, *37*, 639–652. [CrossRef]
12. Linka, K.; Kuhl, E. A new family of Constitutive Artificial Neural Networks towards automated model discovery. *Comput. Methods Appl. Mech. Eng.* **2023**, *403*, 115731. [CrossRef]
13. 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] [PubMed]
14. Li, J.; Wang, B.; Cui, X.; Li, S.; Liu, J. Underwater acoustic target recognition based on attention residual network. *Entropy* **2022**, *24*, 1657. [CrossRef]
15. Yang, S.; Xue, L.; Hong, X.; Zeng, X. A lightweight network model based on an attention mechanism for ship-radiated noise classification. *J. Mar. Sci. Eng.* **2023**, *11*, 432. [CrossRef]
16. Qi, Z.; Li, T.; Chen, J.; Yam, J.C.; Wen, Y.; Huang, G.; Zhong, H.; He, M.; Zhu, D.; Dai, R.; et al. A deep learning system for myopia onset prediction and intervention effectiveness evaluation in children. *NPJ Digit. Med.* **2024**, *7*, 206. [CrossRef]
17. Sheng, B.; Pushpanathan, K.; Guan, Z.; Lim, Q.H.; Lim, Z.W.; Yew, S.M.E.; Goh, J.H.L.; Bee, Y.M.; Sabanayagam, C.; Sevdalis, N.; et al. Artificial intelligence for diabetes care: Current and future prospects. *Lancet Diabetes Endocrinol.* **2024**, *12*, 569–595. [CrossRef]
18. Li, J.; Guan, Z.; Wang, J.; Cheung, C.Y.; Zheng, Y.; Lim, L.L.; Lim, C.C.; Ruamviboonsuk, P.; Raman, R.; Corsino, L.; et al. Integrated image-based deep learning and language models for primary diabetes care. *Nat. Med.* **2024**, *30*, 2886–2896. [CrossRef]

19. Ali, S.G.; Wang, X.; Li, P.; Li, H.; Yang, P.; Jung, Y.; Qin, J.; Kim, J.; Sheng, B. EGDNet: An efficient glomerular detection network for multiple anomalous pathological feature in glomerulonephritis. *Vis. Comput.* **2024**, 1–18. [CrossRef]
20. Dai, L.; Wu, L.; Li, H.; Cai, C.; Wu, Q.; Kong, H.; Liu, R.; Wang, X.; Hou, X.; Liu, Y.; et al. A deep learning system for detecting diabetic retinopathy across the disease spectrum. *Nat. Commun.* **2021**, *12*, 3242. [CrossRef]
21. Zhang, W.; Yu, P.; Wang, Z.; Wang, Z.; Xu, M. A hybrid offline-online method for sound event localization and detection. *Appl. Intell.* **2024**, *54*, 11357–11372. [CrossRef]
22. Ren, J.; Xie, Y.; Zhang, X.; Xu, J. UALF: A learnable frontend for intelligent underwater acoustic classification system. *Ocean Eng.* **2022**, *264*, 112394. [CrossRef]
23. Doan, V.-S.; Huynh-The, T.; Kim, D.-S. Underwater acoustic target classification based on dense convolutional neural network. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1500905. [CrossRef]
24. Khishe, M. DRW-AE: A deep recurrent-wavelet autoencoder for underwater target recognition. *IEEE J. Ocean. Eng.* **2022**, *47*, 1083–1098. [CrossRef]
25. 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]
26. Feng, S.; Zhu, X. A transformer-based deep learning network for underwater acoustic target recognition. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1505805. [CrossRef]
27. Irfan, M.; Zheng, J.; Ali, S.; Iqbal, M.; Masood, Z.; Hamid, U. DeepShip: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification. *Expert Syst. Appl.* **2021**, *183*, 115270. [CrossRef]
28. Zhu, F.L.; Xiao, F.L.; Na, W.; Qing, Y.Z. Present status and challenges of underwater acoustic target recognition technology: A review. *Front. Phys.* **2022**, *10*, 1044890. [CrossRef]
29. Ji, F.; Ni, J.; Li, G.; Liu, L.; Wang, Y. Underwater acoustic target recognition based on deep residual attention convolutional neural network. *J. Mar. Sci. Eng.* **2023**, *11*, 1626. [CrossRef]
30. Xu, J.; Huang, Z.; Li, C. Advances in underwater target passive recognition using deep learning. *J. Signal Process.* **2019**, *35*, 1460–1475. [CrossRef]
31. Li, Y.; Gao, P.; Tang, B.; Yi, Y.; Zhang, J. Double feature extraction method of ship-radiated noise signal based on slope entropy and permutation entropy. *Entropy* **2022**, *24*, 22. [CrossRef]
32. Zhou, Y.; Chen, S.; Wu, K.; Ning, M.; Chen, H.; Zhang, P. SCTD 1.0: Sonar common target detection dataset. *Comput. Sci.* **2021**, *48*, 334–339. [CrossRef]
33. Li, B.; Yao, Y.; Tan, J.; Zhang, G.; Yu, F.; Lu, J.; Luo, Y. Equalized focal loss for dense long-tailed object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6990–6999.

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.