Comparative Analysis of Deep Learning Techniques for Passive Underwater Acoustic Target Recognition: Overview, Challenges, and Future Directions

Song Yifei[1], Mohamad Farhan Mohamad Mohsin[2]
University Utara Malaysia (UUM), 06010 Sintok, Kedah, Malaysia[1]
School of Computing, University Utara Malaysia (UUM), Sintok, Kedah, Malaysia[2]

***Abstract*—Passive underwater acoustic target recognition (UATR) involves analyzing acoustic waves captured by passive sonar to extract valuable information about submerged targets. The underwater acoustics community has increasingly turned its attention to deep learning techniques, owing to their remarkable success in image recognition tasks. This study presents a comprehensive overview of the evolution of UATR techniques, categorizing them into three distinct groups: early methods, conventional machine learning approaches, and modern deep learning-based techniques. Additionally, it provides an in-depth summary of the recognition process utilizing deep learning, detailing various deep network architectures, classifiers specifically designed for underwater acoustic target recognition, and different data input modalities. Finally, the study synthesizes current research findings and outlines potential future directions for advancements in this field, emphasizing opportunities for innovation across these three categories.**

Keywords*—Underwater acoustic target recognition; deep learning; deep network architecture; classifier

I. Introduction

Underwater acoustic target recognition (UATR) is grounded in the production, transmission, and reception of sound waves in water. This technology has extensive applications in military operations, marine exploration, sonar systems, marine life monitoring, and underwater communication through sound navigation and ranging systems [1]. Despite the influence of variables such as temperature, salinity, and pressure on sound propagation in the ocean, sound waves remain the most efficient medium for long-distance underwater detection, transmission, and communication [2],[3]. Passive-sonar systems, which operate by quietly detecting transmitted noise, are gaining popularity due to their enhanced stealth capabilities. Consequently, the analysis of passive sonar underwater target radiation noise data has emerged as an effective method for studying underwater objects. Renowned research institutes in this field include the Woods Hole Oceanographic Institution, the Chinese Naval University of Engineering, Harbin Engineering University, the Pakistan Naval Research Institute, the Institute of Acoustics of the Chinese Academy of Sciences, and the Applied Physics Laboratory at Washington University [4].

Lei et al. [5] highlight that advancements in technology have facilitated the extensive application of earlier methodologies, classical machine learning techniques, and deep learning-based network designs in UATR. Early approaches, though effective, relied on the recognition abilities of skilled sonar operators, leading to high costs and significant susceptibility to subjective factors. Conventional machine learning methods, often dependent on manually crafted features, face significant challenges such as loss of feature information and difficulty in extracting optimal features. In contrast, deep learning techniques are expected to achieve greater accuracy and robustness as computational power continues to grow [6]. By enabling end-to-end recognition processes, these techniques address many of the limitations posed by human involvement in early and traditional approaches.

II. Motivation

The UATR process is typically divided into three key components: the model's input, the model itself, and the model's output, as illustrated in Fig. 1. Based on the execution of each phase and the interconnections between them, recent review studies have categorized various approaches to underwater target identification [7],[8],[9]. For instance, in deep learning models, feature extraction and classification are integrated into a unified process, whereas in traditional approaches, feature design and extraction are performed manually or automatically before data is input into the model. However, a few studies provide comprehensive explanations of each component in a detailed, step-by-step manner. In this context, this study aims to review the latest research on passive UATR from three perspectives: model input, the model itself, and output types. By examining the use of different data input techniques, models, and task output formats, we seek to identify key challenges in current research and propose promising directions for future exploration.

III. Contribution

This study provides a novel perspective on the evaluation and analysis of UATR tasks through comprehensive research from three key aspects: model input, the model itself, and model output. Unlike conventional viewpoints that focus on different approaches to UATR tasks, this study emphasizes a structured analysis of the entire process.

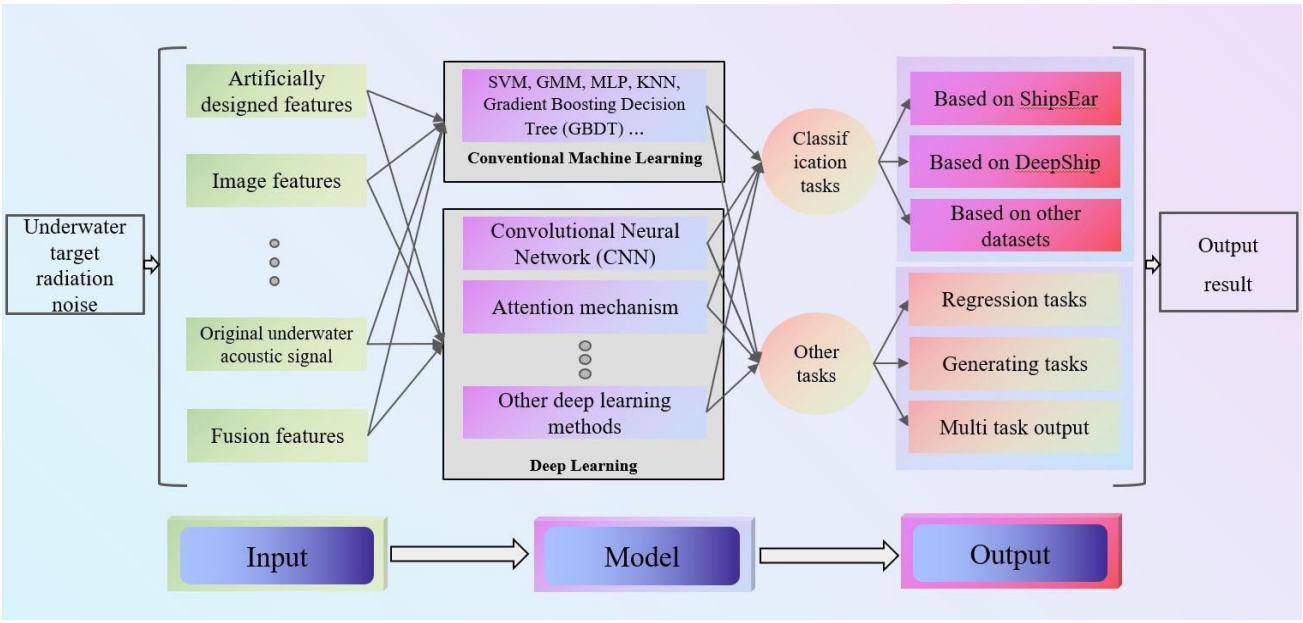

Fig. 1. General process and research architecture of UATR. Feature extraction, classification and recognition modeling, and model output constitute the three key stages in the fundamental process of both deep learning and conventional machine learning approaches for UATR.

Moreover, it highlights critical challenges that must be addressed in existing deep learning-based UATR tasks. By systematically examining and summarizing the input modes of underwater acoustic signals, model types, and model input strategies, this research provides a comprehensive synthesis of the strengths and limitations of various approaches. These insights serve as a valuable reference for researchers advancing UATR methodologies, facilitating future advancements in the field.

IV. ORGANIZATION

Section I introduces the comprehensive situation and basic issues in current research on UATR, serving as the foundation for this study. Section II presents the research motivation, emphasizing the significance of this study. By adopting a novel analytical and review approach based on the fundamental process of underwater target detection, it provides researchers with a broader perspective for understanding these challenges.

Sections III and IV outline the primary contributions and methodological framework of this work. Sections V, VI, and VII discuss different underwater acoustic signal input options, network topologies used in UATR, and the various task output types of the model, respectively. Furthermore, Sections VIII and IX explore the challenges of applying deep learning to UATR, propose potential directions for future research, and summarize the study's key insights. While previous studies have extensively examined the application of various technologies in underwater target detection, this review takes a more structured approach, delving into the specifics of each component. The overall organization of this study is illustrated in Fig. 2.

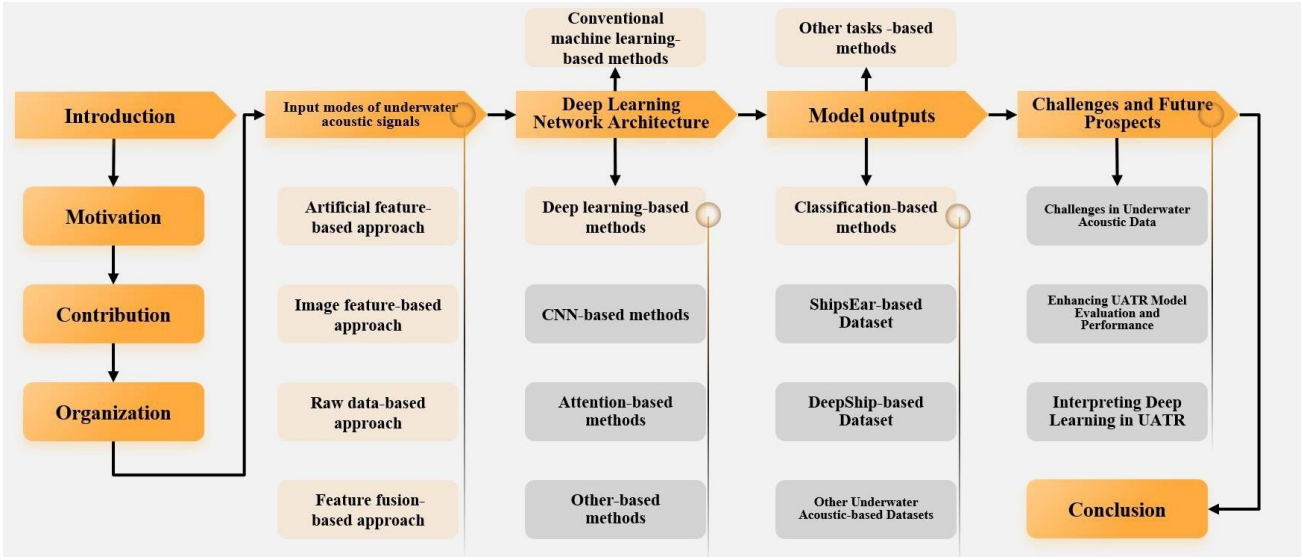

Fig. 2. Organization of the study.

V. Input Modes of Underwater Acoustic Signals

Deep learning-based UATR technology effectively addresses the issue of feature loss caused by human intervention. It typically employs an end-to-end learning framework, where neural networks replace manually designed components [10],[11]. To ensure the generation of high-quality underwater signals, appropriate preprocessing is usually necessary before inputting the target signal into the model [12]. The following sections will explore four approaches: the artificial feature-based approach, the image feature-based approach, the raw data-based approach, and the feature fusion-based approach.

A. Artificial Feature-Based Approach

A set of feature values that are manually crafted and extracted to characterize the properties of the target signal is known as artificially designed features. These features are highly interpretable, possess clear mathematical and physical significance, and are rooted in prior knowledge and experience.

In the field of underwater target identification, Mel frequency cepstral coefficients (MFCC) and DEMON spectra are widely utilized. MFCC captures the characteristics of underwater acoustic signals through a series of steps, including pre-emphasis, frame windowing, fast Fourier transform (FFT), power spectrum computation, Mel filter bank filtering, logarithmic transformation, and discrete cosine transform (DCT). Similarly, DEMON extracts the features of underwater acoustic data through preprocessing, envelope extraction, FFT, feature extraction, and subsequent analysis. This process is illustrated in Fig. 3.

Mel-frequency cepstral coefficients (MFCCs) are a widely used example of manually crafted features, extensively applied in language and speech recognition due to their ability to reduce feature map size, suppress noise interference, and withstand temporal variations [13]. On the ShipsEar dataset, Liu et al. [14] achieved an impressive 99.34% identification accuracy using MFCCs as input features for the RACNN model. Additionally, other studies have confirmed that MFCCs outperform models like VGG16 and ResNet34 in terms of computational complexity and parameter efficiency [15],[16]. However, standard MFCCs primarily capture static spectral envelope information, making it difficult to represent dynamic signal properties effectively. To address this limitation, Chen et al. [17] proposed using 3D MFCCs as input features. By incorporating first-order and second-order differential features, 3D MFCCs can capture dynamic characteristics such as the rate of change and acceleration of signals over time. Using 3D MFCC features as input for the Attention Mechanism Residual Concatenate Network (ARescat) model, an accuracy of 95.80% was achieved on the ShipsEar dataset.

The improvement in model recognition accuracy can be attributed to two main factors: the extraction of features that better represent the original data and the design of models with more advanced performance. However, it remains unclear whether 3D MFCC, an enhancement of traditional MFCC features, can more effectively represent the characteristics of raw data and is better suited for designing deep learning models. It is worth noting that numerous studies have compared MFCC features with other alternatives, such as FMSE and GFCC [18], [19]. Fig. 4 illustrates the performance of various feature extraction techniques on the same model using the ShipsEar dataset.

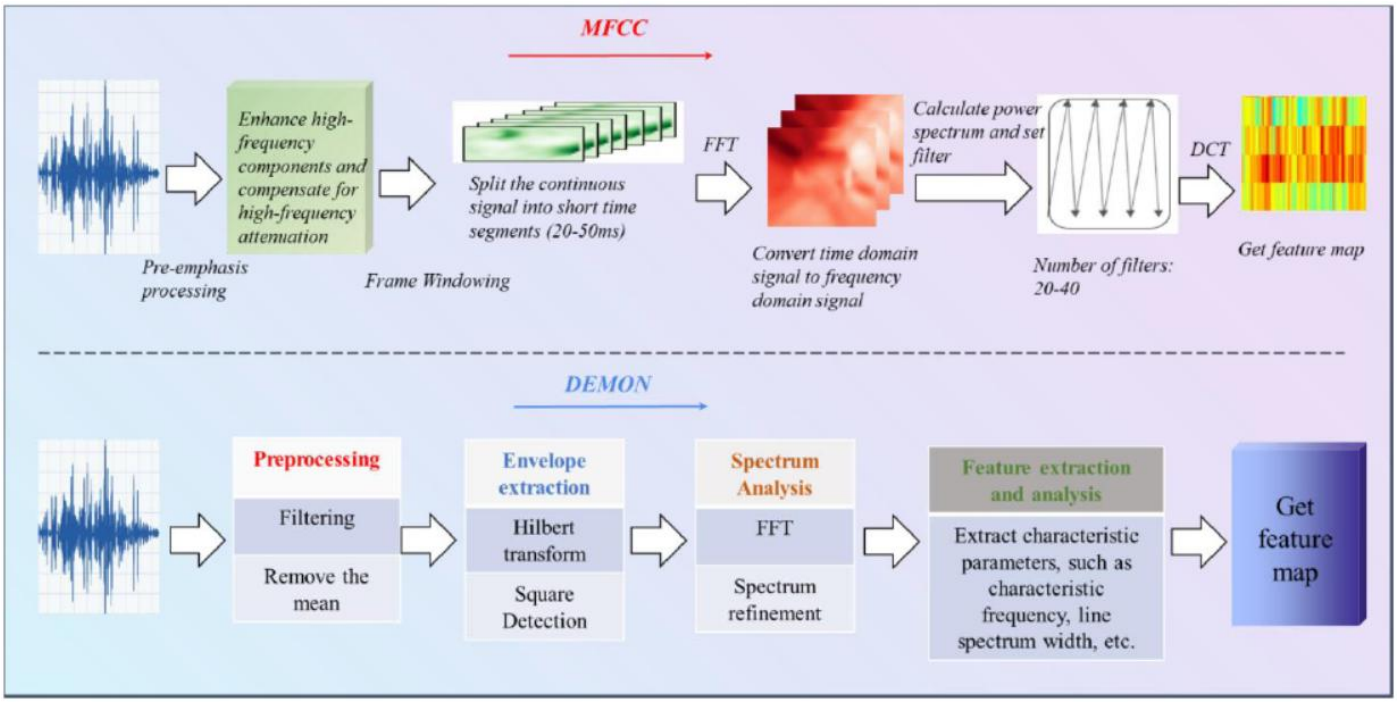

Fig. 3. MFCC and DEMON feature extraction process.

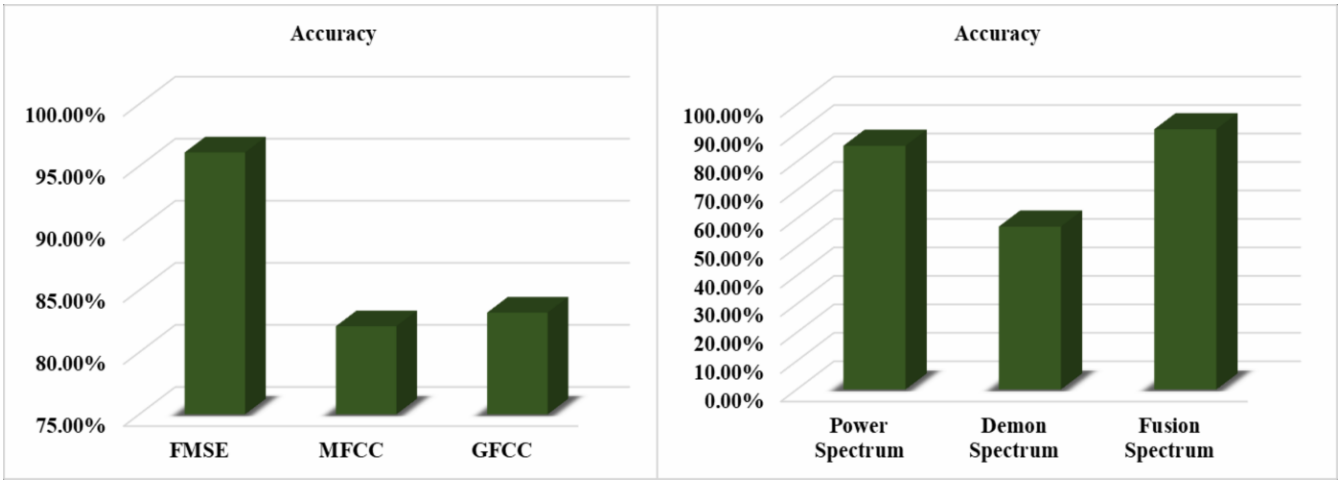

Fig. 4. Comparison of accuracy for different artificial and image features.

A broadband demodulation method, known as DEMON analysis, can distinguish between ship-radiated noise signals and the modulation envelope caused by propeller cavitation. By employing spectral analysis and line spectrum detection, it is possible to determine the number of propeller shafts, shaft frequencies, and blade counts [20]. Using the DEMON spectrum as input for the DEMONet model, Xie et al. [21] achieved an accuracy of 80.45 ± 0.67% on the DeepShip dataset. It is noteworthy that the feature extraction and model design in their study aimed for consistency. Furthermore, techniques such as Wavelet Packet Decomposition, Low-Pass Filtering, Gamma Frequency Cepstral Coefficients (GFCC), Multiscale Entropy Decomposition (MEMD), and LPS Feature Extraction, along with enhanced MFCC-based feature extraction methods, have been extensively studied and applied [22],[23],[24],[25]. Notably, Luo et al. [26] evaluated a range of feature extraction methods, including DEMON, LOFAR, MFCC, GFCC, EMD, and HHT, discussing their performance and suitability in different scenarios.

Manual feature design has several drawbacks, including low flexibility, limited expressive capacity, and high labor costs [27]. Additionally, manually crafted features often suffer from insufficient automated feature learning, as they cannot effectively capture or adapt to the inherent structure of data. In contrast, deep learning automatically extracts hierarchical features from data and enables end-to-end optimization, whereas manual feature extraction typically requires a separate design for classifiers or regressors [28]. As a result, researchers are increasingly focusing on technologies that leverage visual features [17],[29].

B. Figures and Tables

UATR based on image features is achieved by converting underwater acoustic signals into images and applying pattern recognition theory and image processing techniques to extract and analyze image features. This approach involves three main stages: image feature extraction, feature selection and optimization, and classification recognition. Image feature extraction involves using feature extraction algorithms to obtain representative and discriminative features from the transformed images. Feature selection and optimization focus on choosing the most relevant subset of features for target identification, removing unnecessary and redundant features, reducing feature dimensionality, and ultimately improving recognition accuracy and efficiency. Classification recognition involves using classification algorithms in pattern recognition to compare the extracted and optimized image data with pre-existing target feature libraries, enabling the identification of unknown underwater acoustic targets.

In terms of expressive power, interpretability, and adaptability, the image feature-based approach has demonstrated significant advantages [16]. Common image features include the amplitude spectra derived from LOFAR and STFT. Chen et al. [30] extracted the LOFAR spectrum of underwater sound waves and identified key characteristics using a multi-step decision-based line spectrum augmentation method. They combined the reconstructed line spectrum with the original LOFAR spectrum using a dual-threshold calculation approach. A convolutional neural network (CNN) was then trained on the reconstructed LOFAR spectrum, achieving an average recognition accuracy of 95.22% on the ShipsEar dataset. Despite the high recognition accuracy of this method, data augmentation techniques not only improve recognition performance but also mitigate noise interference from the marine environment. However, further research is needed to evaluate their impact on the recognition performance of neural networks. Additionally, Xu et al. [31] applied Mel (or Bark) filter banks to the FFT-generated spectrum to compute the STFT amplitude spectrum through complex FFT calculations, achieving an identification accuracy of 82.97% on the ShipsEar dataset. Yao et al. [32] further compared the LOFAR spectrum with other feature extraction methods, such as MFCC and CQT, using statistical histograms to analyze feature information across different ship types. Experimental results revealed that LOFAR image features lack low-frequency information, highlighting a limitation in capturing certain aspects of feature information.

Tang et al. [33] utilized Mel spectrograms as feature representations for model input. After feeding the data into the Transformer model, they conducted experiments on a small sample dataset, achieving a 90% identification accuracy with the Swin Transformer. Similarly, Xie et al. [34] achieved identification accuracies of 77.14%, 74.85%, and 95.48% on

the ShipsEar, DeepShip, and DTIL datasets, respectively, using Mel spectrograms as input. Studies employing spectrograms as input features have also yielded positive results [25],[35],[36],[37]. In Fig. 5, the recognition accuracy based on image features in UATR tasks is presented.

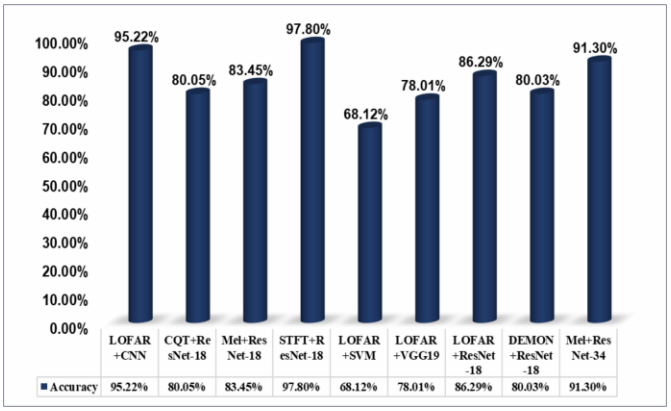

Fig. 5. Recognition accuracy based on image features.

Fig. 5 provides an overview of feature extraction methods based on image representations. Different deep learning architectures exhibit varying levels of recognition performance when using image features as model inputs. Notably, employing STFT features as input to ResNet-18 achieves a high recognition accuracy of 97.80%. Furthermore, due to the strong time-frequency correlation inherent in underwater acoustic signals, the use of raw time-domain signals as model inputs has been extensively investigated to preserve as much detailed information as possible. In contrast, image-based feature extraction methods may struggle to capture complex or high-dimensional patterns and are more susceptible to the loss of critical information [38], [39].

C. Raw Data-Based Approach

Raw signal-based UATR technology aims to identify underwater targets by directly utilizing the received raw underwater acoustic signals without performing complex transformations. This approach leverages various signal processing and pattern recognition algorithms. During the feature extraction stage, three main methods are employed: time domain, frequency domain, and time-frequency domain feature extraction. Time domain feature extraction involves directly obtaining amplitude information from the raw signal, such as peak values, mean, variance, and other statistical characteristics. Frequency domain feature extraction involves calculating the power spectral density of the raw signal using the Fourier transform and other operations, enabling target identification based on the distribution of signal power across different frequencies. Time-frequency domain feature extraction divides the raw signal into multiple short-time segments and performs a Fourier transform on each segment to obtain the distribution of signal components across both time and frequency domains. This approach provides a comprehensive view of how signal characteristics vary over time and frequency.

The intrinsic time-frequency correlation of underwater acoustic signals can lead to the loss of significant frequency domain information, even though image-based techniques can moderately enhance UATR [40]. Consequently, the direct use of raw data as input has become a prominent area of research to ensure the preservation and accuracy of information in underwater acoustic signals [41]. Hu et al. [42] investigated a technique for identifying ship-radiated noise using raw time-domain waveforms. They employed a depth-wise separable convolutional network to extract deep features that mirrored the auditory system's acoustic information processing. The model achieved an average classification accuracy of 90.9% when tested on a dataset of real acoustic signals from civilian ships. Additionally, extreme learning-based approaches have also shown promising results [43],[44].

However, using raw data directly as input for the model requires substantial computational resources, particularly for high-sampling-rate underwater acoustic signals [45]. Moreover, in the context of UATR with limited sample sizes, insufficient data can quickly lead to overfitting or poor model training.

To address the issue of inadequate underwater acoustic samples, several studies have explored segmenting and resampling original sound data. Yang et al. [46] resampled raw data at 16,000 Hz and divided it into 1-second non-overlapping intervals. Ji et al. [47] resampled underwater acoustic at 20 kHz, truncating each frame into 4096 samples with a 2048-sample overlap between successive frames. Similarly, Yang et al. [48] segmented each 5-minute WAV audio clip into 6-second intervals representing sound events and normalized the data for each segment. Li & Yang [49] split each audio recording into 3-second parts, normalizing each segment based on unprocessed raw time-domain data. In Fig. 6, the recognition accuracy with raw data as input is presented. From Fig. 6, it can be seen that the highest recognition rate can reach 95.30%.

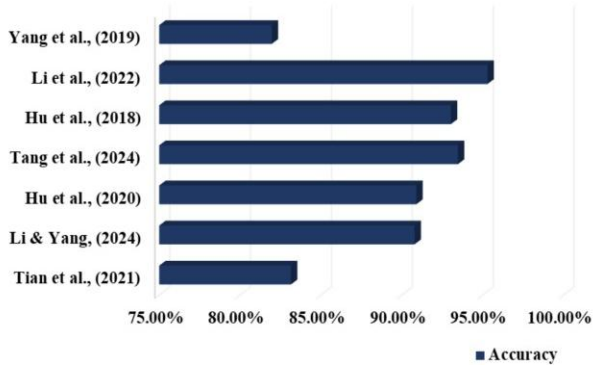

Fig. 6. Recognition accuracy based on raw data.

While resampling techniques can somewhat alleviate the issue of insufficient samples, they have limitations, including the potential amplification of noise and reduced diversity in the generated data. Therefore, methods based on fused features are becoming increasingly popular as a research direction.

D. Feature Fusion-Based Approach

Feature-based fusion is widely utilized in UATR technology to enhance accuracy and reliability by integrating various types of feature information. The main components of this approach include multi-domain feature fusion, multi-modal

feature fusion, and multi-scale feature fusion. Multi-domain feature fusion involves combining time domain, frequency domain, and time frequency domain characteristics of underwater acoustic signals to capture comprehensive signal properties. Multi-modal feature fusion integrates data collected from different sensors or based on different physical principles. For example, sonar signal features can be combined with underwater target features captured by optical sensors. Multi-scale feature fusion involves utilizing features at different scales, such as integrating fine local details with broader global target characteristics in image processing, to achieve a more robust and detailed representation of the target.

The fusion feature approach enhances identification accuracy by combining multiple feature types, such as time domain, frequency domain, and time frequency domain features, effectively addressing the limitations of single feature information. Additionally, integrating diverse features improves the system's robustness and adaptability to challenging conditions, such as interference and noise [50], [51]. Table I provides a summary of techniques for fused feature extraction.

TABLE I. SUMMARY OF FEATURE FUSION-BASED APPROACH

Author (s)	Feature	performance	Advantage	Disadvantages
Cao et al., (2019)	Axis frequency characteristics; LPS characteristics; WPCE characteristics	89.68% Acc	Adopting the SSAE model provides stronger discriminability; Constructing a joint feature set based on spectral and wavelet domain information	Underwater target classification tasks that require high real-time performance face challenges due to the complexity, cost, and time involved in obtaining a substantial amount of representative underwater target data.
Hong et al., (2021, May)	Log - Mel Spectrogram; MFCC; CCTZ	94.30% Acc	Using ResNet18, combined with embedding layers, early stopping, and adaptive learning rate strategies for training, Adopting data augmentation strategy SpecAugment	The universality of the method may be limited by the experimental validation, which was restricted to the ShipsEar dataset and did not include testing on a broader variety of datasets with different types and scales.
Domingos et al., (2022)	CQT; Gammatone Spectrograms; Mel Spectrograms	97.00% Acc	Compare various optimizers, neural network architectures, and preprocessing filters; Combine multiple filters into a three-channel signal	Accurately determining which scenario yields the best results under different testing conditions is challenging due to the varying number of instances in each case.
Chen et al., (2023)	CQT; delta MFCC; double - delta MFCC	99.10% Acc (DeepShip)	Propose a new loss function that only adds three hyperparameters to transform multi classification tasks into multiple binary classification tasks	In terms of model interpretability, the underlying mechanisms of each module are not thoroughly examined, leaving unclear how they influence the final outcomes.
Wu et al., (2023)	FBank; delta FBank; delta - delta FBank	90.50%	Combining cross domain pre training effectively improves recognition accuracy and saves a lot of training time	Insufficient exploration of innovative feature extraction techniques and potential feature combinations.
Pu et al., (2024)	Original wave; CQT; Mel spectrogram	96.37%	Design Scale ResNet module and RHAF module to improve information fusion efficiency and model adaptability to underwater acoustic data	In feature fusion, the lack of comparison with more advanced fusion techniques makes it difficult to determine its optimality across all scenarios.

Wu et al. [52] developed fusion features as model inputs by combining Mel filter-bank (F-Bank), delta F-Bank, and double delta F-Bank. To better capture the characteristics of underwater acoustic targets, they extracted neigh-boring dynamic features by computing delta features. Experimental results showed that this fusion approach improved identification accuracy by 0.9% and 1.2% on CNN and ResNet18 models, respectively. Chen et al. [53] proposed the FEFM method, which extracts multidimensional features from ship-radiated noise signals using various feature extraction techniques based on signal analysis and brain-inspired properties. These features are then fused using the proposed feature fusion technique to create high-dimensional fused features, which are used as inputs for the Multi-Gradient Flow Global Feature Enhancement Network (MGFGNet) network. Feature ablation tests were conducted on the Deepship dataset, comparing the fusion feature approach with other feature extraction methods. The fusion technique achieved a total accuracy of 99.1%.

Hong et al. [54] calculated various features, including the Log Mel Spectrogram (LM), Mel Frequency Cepstral Coefficients (MFCC), and CCTZ (which comprises the following features: Chroma, Contrast, Tonnetz, and Zero-cross ratio). These features were then fused and enhanced to create a three-dimensional feature matrix, which was used as input to the model. Domingos et al. [55] divided the raw dataset into one-second segments, padding shorter segments with zeros. They applied three preprocessing techniques—Mel spectrograms, Constant Q Transform (CQT), and Gammatone spectrograms—and combined them into a three-dimensional representation, referred to as "Complete", which was then be used as input to the model.

VI. DEEP LEARNING NETWORK ARCHITECTURE

The accuracy of UATR tasks is heavily influenced by the choice and design of the model architecture. With advancements in computing power, UATR technology has made significant progress. Conventional methods that relied on manual feature extraction and classifier design have been largely improved by deep learning approaches capable of automatic feature extraction and classification, resulting in substantial improvements in identification accuracy [56]. To clearly illustrate the development of these techniques, this study categorizes model architectures into two main groups: conventional machine learning-based methods and deep learning-based methods, depending on whether they employ an

end-to-end recognition process. A detailed discussion of these techniques follows below.

A. Conventional Machine Learning-Based Methods

UATR based on conventional machine learning methods refers to the process of classifying and identifying various underwater acoustic targets by extracting, analyzing, and evaluating the target information embedded in underwater acoustic signals using conventional machine learning theories and algorithms. Conventional machine learning methods rely on manually crafting features, which are then input into regression or classification models. Common approaches include multilayer perceptron (MLP), logistic regression, naive Bayes, support vector machines (SVM), and random forest classifiers [57], [58].

Honghui et al. [59] proposed the Multi-Attribute Correlation Perception (MCP) model, achieving 82.1% accuracy in ship detection tests by integrating the MCP module with a time-frequency feature extraction module for UATR. To enhance feature selection during classification, Fernandes et al. [60] applied neighborhood component analysis (NCA) for dimensionality reduction, combined with genetic algorithms and the K-nearest neighbor (KNN) method for object classification. Yu et al. [61] achieved an underwater target identification accuracy of 93.84% by employing the Gradient Boosting Decision Tree (GBDT) model in combination with two hydrophones and the VLA feature extraction approach. These approaches share a common characteristic: they all employ distinct feature extraction strategies prior to inputting features into the classification model to complete the classification task. However, separating feature extraction and classification introduces human variables due to the relative decoupling of these processes, potentially leading to a decline in recognition accuracy [18].

Sun et al. [19] achieved a recognition accuracy of 96.1% by using a support vector machine (SVM) classifier to process the retrieved features. Wang et al. [24] employed Gaussian Mixture Models (GMMs) to modify the topology of deep neural networks (DNNs), proposing the MFF-MDNN approach, which achieved an average recognition accuracy of 94.3%. While these methods reduce spurious variables to some extent and enhance recognition accuracy, the limited number of classifier layers in such models may constrain their ability to generalize effectively. Yang et al. [62] highlighted that the poor generalization ability of traditional recognition systems stems largely from the use of shallow classifiers, such as SVMs and shallow neural networks. Moreover, conventional methods, which predominantly rely on line and spectrum properties, often exhibit lower identification rates. Fortunately, deep learning-based approaches provide a more holistic framework for UATR, effectively addressing the limitations of traditional machine learning techniques and opening up promising directions for future research.

B. Deep Learning-Based Methods

Deep learning-based UATR technology is an advanced approach that automatically extracts, analyses, and recognizes target information within underwater acoustic signals by leveraging the powerful feature learning and pattern recognition capabilities of deep learning algorithms. The main process of this technology is illustrated in Fig. 7. The primary advantage of employing deep learning networks in UATR lies in their ability to bypass the complex feature engineering required by conventional machine learning approaches. Furthermore, they can partially address the problem of feature information loss associated with manual feature extraction [63]. The following will be discussed in detail.

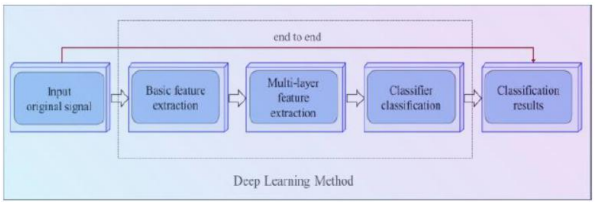

Fig. 7. General process of UATR based on deep learning methods.

1) CNN-based methods. The Convolutional Neural Network (CNN)-based UATR system leverages the powerful feature extraction and pattern recognition capabilities of convolutional neural networks to accurately identify targets in underwater acoustic data. CNNs are primarily used in tasks related to computer vision and speech recognition. Their architecture involves multiple processes that enable hierarchical feature learning and the use of automatically generated features for prediction [64]. Over time, CNNs have become a research hotspot in underwater target detection, driven by their remarkable performance advancements from AlexNet to GoogLeNet [39],[65].

ResNet is one of the most widely studied and popular deep learning models. Xu et al. [31] built their model on ResNet-18, incorporating a multi-head attention mechanism. The model was trained using two techniques: smooth induced regularization and the local masking and replication (LMR) approach. Its identification accuracy was evaluated using various input attributes, although several factors could potentially influence its performance. Xie et al. [34] adopted a dual-encoder framework with residual structured convolutional neural networks (ResNet) as the backbone to capture diverse acoustic features, specifically Mel spectra and continuous Q-transform spectra. By treating these networks as multi-view learners, they were able to extract more comprehensive features. Additionally, ResNet-based methods for developing and enhancing UATR have demonstrated promising results [54],[66],[67].

Significant advancements have been made in UATR techniques based on convolutional neural networks (CNNs) in recent years. Yang et al. [48] introduced a deep convolutional neural network (ADCNN) inspired by auditory perception, achieving an accuracy of 81.96%. Similarly, Hu et al. [44] proposed the auditory temporal convolutional neural network (ATCNN) model, which attained recognition accuracies of 95.9% on training data and 90.9% on test data. Meanwhile, more advanced CNN-based architectures, such as depth-wise separable convolutional neural networks, have also shown promising results. However, despite CNN's excellent feature extraction capabilities, the unique and complex nature of underwater acoustic signals poses several challenges. Specifically, CNN's convolutional kernels, limited by their

localized receptive fields, struggle to capture long-range correlations and global information inherent in underwater sound waves. Fortunately, techniques utilizing residual attention convolutional neural networks (RACNN) and deep residual attention convolutional neural networks (DRACNN) have opened new avenues for research. These methods have achieved impressive identification accuracies, addressing some of CNN's limitations in UATR [47], [68].

2) Attention-based methods. Attention-based UATR enhances the model's ability to extract and interpret critical information from underwater signals by incorporating attention mechanisms. These mechanisms efficiently capture the global characteristics of underwater acoustic signals by dynamically assigning attention weights based on signal features. They allocate different weights to various aspects of underwater acoustic signals across time, frequency, or spatial locations. The propagation of underwater acoustic signals is influenced by factors such as water temperature, salinity, and pressure. These factors can lead to attenuation, scattering, and refraction, impacting both the propagation distance and the quality of the signal. Consequently, the propagation of underwater acoustic signals is globally correlated, meaning that the signal's characteristics are shaped by the entire aquatic environment [69].

In UATR, the attention mechanism has gained significant traction for its ability to extract global features. Xiao et al. [70] introduced an attention-based neural network (ABN) model that integrated an attention module with a traditional deep neural network (DNN) composed of fully connected layers. By embedding the attention module within the DNN architecture, they developed a novel network designed to tackle UATR tasks. This model achieved an accuracy of 74.3% in multi-target resolution tasks, marking a 16.0% improvement over traditional DNNs. In contrast, Xie et al. [71] proposed the UART (Underwater Acoustic Recognition based on Templates) framework, which diverges from ABN by not depending on innovations in neural network design. Instead, this framework consists of an audio encoder, a spectrogram encoder, and a text encoder, offering an alternative approach to underwater object recognition.

The Transformer, a widely used attention mechanism, has been extensively applied in UATR using sound waves. Tang et al. [33] studied the Swin Transformer Biformer, a Transformer-based model. Similar to the four-layer pyramid structure of CNNs, the architecture of the Swin Transformer features a progressive reduction in the feature map's size as the feature layers deepen. The Swin Biformer enhances feature representation by combining the Swin Transformer with the Biformer, dynamically adjusting the partition window to refine regional feature extraction. Pu et al. [72] introduced the Attention Layer Supplementary Integration (ALSI) framework for UATR. This model demonstrated outstanding performance, effectively detecting underwater objects such as ship-radiated noise and achieving an identification accuracy of 96.39% on the ShipsEar dataset. Additionally, several other attention-based enhancement techniques have also yielded promising results [73], [74].

3) Other-based methods. The Restricted Boltzmann Machine (RBM), a fundamental component of deep learning, consists of visible and hidden layers of neurons. RBMs are widely used as energy-based models for tasks such as feature learning, dimensionality reduction, and model development [75]. In the field of UATR, RBM-based deep learning techniques have been effectively applied.

For instance, Luo and Feng [25], applied RBM and BP neural network techniques, achieving an accuracy of 93.17% on the ShipsEar dataset. Similarly, Luo et al. [18] developed a UATR system that demonstrated the effectiveness of BP neural networks and RBM autoencoders. However, despite RBM's advantages in feature learning, dimensionality reduction, and data generation, its shallow structure makes it difficult to model complex data distributions or high-dimensional features. To address these challenges, Yang et al. [35] proposed the GRU-CAE collaborative deep learning network. Their GRU-CAE-TM (Gated Recurrent Unit with Template Matching and Convolutional Autoencoder Collaborative Deep Learning Network) approach achieved an open-set recognition accuracy of 82.21%. While GRU effectively captures short-term time series dependencies, it faces limitations in processing underwater acoustic target signals with long-term dependencies. Moreover, deep learning techniques, including LSTM gradient-boosting decision trees, have been widely used to tackle issues related to few-shot learning in UATR [29],[61],[76].

VII. Model Outputs

The output of a deep learning model is typically defined by the result of forward propagation, depending on the task type and objective. Some common types of tasks include classification, regression, generation, segmentation, embedding or feature representation, and multitask output [77], [78],[79]. This section provides detailed explanations of UATR based on classification tasks and other related activities.

A. Classification-Based Methods

Classification-based methods for UATR typically involve steps such as data preprocessing, feature extraction and selection, and the application of classification algorithms to differentiate various underwater acoustic targets (e.g., submarines, fish schools, marine organisms) using sonar data. To provide a comprehensive overview of previous research, the following section will delve into various methods applied to different datasets.

1) ShipsEar-based dataset. The primary purpose of the ShipsEar dataset is to facilitate the development and testing of UATR algorithms and models [80]. By providing a substantial amount of underwater acoustic data, the ShipsEar dataset enables researchers to analyze and classify noise generated by various vessels, thereby enhancing environmental monitoring and maritime surveillance. The accuracy of the results varies depending on the model design, as illustrated in Table II.

TABLE II. ACCURACY OF UATR METHOD BASED ON SHIPSEAR DATASET

Author (s)	Methods	Accuracy
Fahad et al., (2024)	CNN + DenseNet	99.50%
Liu et al., (2024)	Residual Attention Convolutional Neural Network (RACNN)	99.45%
Yang et al., (2024)	1D CTN	96.84%
Pu et al., (2024)	Attention Layer Supplement Integration (ALSI)	96.37%
Luo et al., (2021)	Conditional Deep Convolutional Generative Adversarial Network (cDCGAN) Model	96.32%
Chen et al., (2023)	ARCSB modules (Rescat, SE,Maxpool, ASPP, MLP)	95.80%
Li et al., (2022)	VGGish	95.30%
Hong et al., (2021, May)	ResNet18	94.30%
Tang & Hu (2024)	DSCANet	93.00%
Ke et al., (2018)	1D convolution Autoencoder - Decoder model	93.28%
Luo et al., (2021)	RBM+BP neural network	92.60%
Du et al., (2024)	TF - DD - CNN	92.23%
Fernandes &Apolinário (2020)	KNN	71.10%
Chen et al., (2023)	ARCSB modules (Rescat, SE,Maxpool, ASPP, MLP)	95.80%

ShipsEar dataset consists of five categories: four ship types and one background noise category. The model's output can include accuracy or probability as performance metrics [60],[81]. For example, using the ShipsEar dataset, Pu et al. [72] reported a 96.39% identification accuracy for their model. This output format offers the advantage of being easily and directly comparable to the results of other studies. Moreover, Müller et al. [82] examined the strengths and weaknesses of each approach and proposed a method for integrating the ShipsEar and DeepShip datasets.

One critical factor that can influence object identification accuracy is how the dataset is divided into training and testing sets. Fahad et al. [83] pre-processed each signal into 1,000 observation frames, each containing 4,096 amplitude samples, and then randomly split the dataset, allocating 70% for training and 30% for testing. While this approach is effective, an 80/20 split between training and testing sets is more commonly used in similar studies [17], [76],[84].

In machine learning and pattern recognition, accuracy is a commonly used performance metric that measures the consistency between model predictions and actual outcomes. However, a model's accuracy is influenced by several factors, such as the criteria for splitting training and testing sets, the volume of training data, and the model's inherent performance. These factors collectively impact the overall accuracy to varying degrees. Moreover, accuracy on labeled datasets during training and validation does not reflect the model's ability to generalize data from unknown categories. Therefore, relying solely on accuracy to evaluate a model's performance is not an ideal approach. Instead, explaining the model's behavior—through mechanisms such as process visualization and interpretability—is increasingly recognized as an effective way to assess model performance comprehensively.

2) DeepShip-based dataset. Many UATR studies rely on private datasets, but the lack of public access to these datasets hampers the continuous improvement of UATR tasks [85]. The DeepShip underwater acoustic dataset, developed by Irfan et al. [86], was made publicly available to advance research in the field. Fig. 8 presents the recognition accuracy of various methods evaluated using the DeepShip dataset.

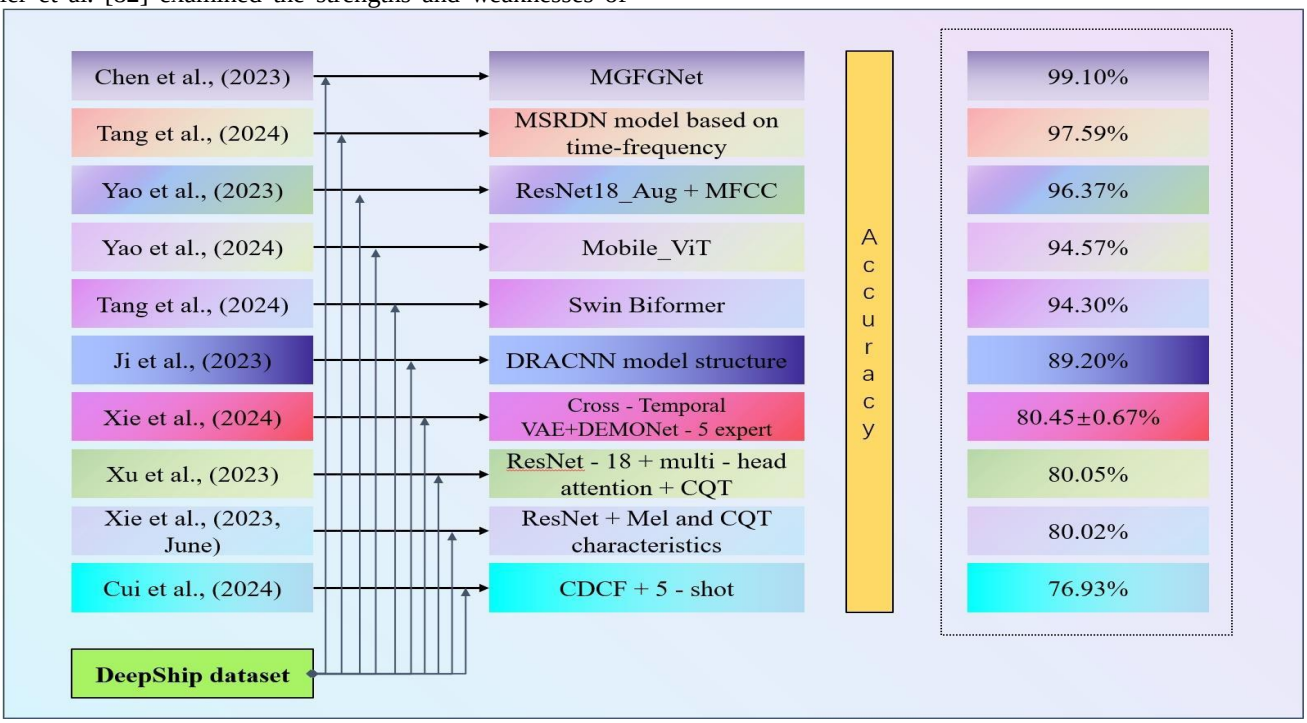

Fig. 8. Accuracy of UATR method based on DeepShip dataset.

Yao et al. [32] utilized acoustic data from three ships, extracting 1,487,488 samples for each ship type to analyze acoustic characteristics from the DeepShip database. To prevent information leakage, Xie et al. [21] divided each signal recording into 30-second segments with a 15-second overlap between adjacent segments, ensuring that data from the same track was not included in both the training and testing sets. Similarly, Chen et al. [53] pre-processed two datasets by standardizing all WAV-format audio files to a sampling rate of 22,050 Hz. They segmented the underwater acoustic data into 5-second intervals, resulting in over 30,000 sound samples, which were then categorized into training, validation, and testing sets.

3) Other underwater acoustic-based datasets. Both private and publicly accessible observation datasets have been utilized in the UATR mission. Examples include the Seabed-Objects dataset, PondEx09/PondEx10 dataset, Ocean Networks Canada data (https://oceannetworks.ca), and various private datasets [68],[87],[88]. Table III presents the recognition accuracy of several techniques evaluated using these private datasets.

TABLE III. ACCURACY OF UATR METHOD BASED ON PRIVATE DATASET

Author (s)	Methods	Accuracy
Xue et al., (2022)	CamResNet (ResNet with Channel Attention Mechanism)	98.20%
Jin & Zeng, (2023)	CTA - RDNet	97.69%
Christensen et al., (2024)	Sparse MvDA + SVM	97.30%
Sun et al., (2024)	SVM + MFCC, FMSE, GFCC	96.10%
Wang et al., (2019)	MFF - MDNN	94.30%
Yu et al., (2024)	Gradient Boosting Decision Tree (GBDT)	93.84%
Hu et al., (2018)	CNN + ELM	93.04%
Hu et al., (2020)	ATCNN	90.90%
Li & Yang, (2024)	ASTEM - DCNN	90.79% (SNR-6dB)
Cao et al., (2019)	Stacked Sparse Autoencoder (SSAE)	89.68%
Tian et al., (2021)	MSRDN（Multiscale Residual Deep Neural Network)	83.15%
Yang et al., (2022)	GRU - CAE - TM	82.21%
Honghui et al., (2022)	MCPM2	82.10%
Yang et al., (2019)	ADCNN	81.96%
Xiao et al., (2021)	ABNN (In multi target resolution task)	74.30%

Domingos et al. [55] developed a dataset using the Ocean Network Canadian data by randomly selecting audio samples from each category. To ensure all records were distinct, they continued the selection process until the required total duration was achieved. Each item in the original dataset was divided into 1-second segments, with any segment shorter than 1 second padded with zeros. The dataset was then split into training, validation, and testing sets in proportions of 85%, 10%, and 5%, respectively. Algorithms based on this dataset have demonstrated strong recognition results [48], [49],[56]. Additionally, Hu et al. [42] utilized a private dataset for UATR tasks that included underwater noise data from ferries, large boats, and small boats, collected at an anchorage with a sampling frequency of 48,000 Hz. In their experiment, 80% of the samples from each category were allocated to the training set, while the remaining 20% were used for testing. Each record was derived from a WAV audio file divided into 10-second segments, with a sampling period of 45 milliseconds and a sampling interval of 12.5 milliseconds for both training and testing samples. Several UATR studies have shown strong recognition performance using such private datasets. At the same time, some studies have achieved impressive recognition accuracy [89], [90].

B. Other Tasks-Based Methods

UATR based on regression tasks typically achieves target tracking and recognition by predicting continuous target properties such as location, velocity, and depth. These methods, which excel in localization, tracking, and dynamic detection of underwater acoustic targets, primarily rely on regression models, machine learning algorithms, and data from underwater sensors.

Zhu et al. [76] proposed a feature selection (FS) technique for underwater sound source localization based on principal component regression. This approach employs a convolutional autoencoder to extract latent features and a multi-layer perceptron for source localization. The framework demonstrates high accuracy and robustness to unseen data while achieving a 95% reduction in training time after FS. To address the limitations of existing methods in accounting for multi-attribute correlations, Honghui et al. [59] developed a deep learning-based multi-attribute correlation perception (MCP) technique for UATR.

Generative task-based approaches provide significant advantages for UATR, especially in noisy or data-scarce environments. Luo et al. [67] trained a DCGAN model using the original sample set to develop an appropriate generator. This generator can increase the number of examples for each category in the dataset by generating samples across various categories based on input labels. In addition, multi-task output-based methods for UATR have gained significant attention in recent years. These methods are capable of simultaneously addressing multiple related tasks, effectively leveraging the relationships between multimodal features, and thereby enhancing both the accuracy and robustness of recognition.

VIII. CHALLENGES AND FUTURE PROSPECTS

The methodology, techniques, and current research on UATR indicate that factors such as training methods, model architecture, and underwater acoustic data significantly impact model performance. Additionally, the model's ability to address real-time challenges in practical applications plays a crucial role. The following sections will explore these aspects from three perspectives: challenges in underwater acoustic data, enhancing UATR model evaluation and performance, interpreting deep learning in UATR.

A. Challenges in Underwater Acoustic Data

Gathering underwater acoustic data presents significant challenges due to high costs and substantial resource demands. Most UATR studies depend on a limited number of publicly accessible datasets or data collected from relatively simple maritime environments. This reliance restricts the variety and richness of the data, making it difficult to generalize findings across diverse underwater scenarios. As a result, acquiring high-quality underwater acoustic data has become a critical issue, with insufficient sample sizes exacerbating the problem. The lack of diverse datasets not only hinders model training and validation but also limits advancements in research and practical applications in more complex and variable underwater environments. Addressing these challenges is essential for enhancing the reliability and effectiveness of UATR systems.

B. Enhancing UATR Model Evaluation and Performance

Deep learning models utilized for UATR often exhibit high complexity, which can lead to significant challenges in training effectiveness. When the complexity of a model exceeds the available quantity and quality of training data, achieving desired outcomes becomes increasingly difficult. While improving model accuracy is a primary objective, relying solely on accuracy as the sole metric for evaluating model performance is insufficient.

To enhance model evaluation, it is crucial to incorporate process visualization and develop a deeper understanding of the model's mechanisms. This approach not only aids in identifying the strengths and weaknesses of the model but also fosters greater transparency. By integrating physical variables from the input data—such as temperature, salinity, depth, and geographic location—into the evaluation process, researchers can create more robust models that better reflect real-world conditions. This integration facilitates a more holistic assessment of model performance, ultimately leading to improved adaptability and reliability in practical applications. Moreover, such comprehensive evaluations can guide future research directions, informing the development of models that are not only accurate but also interpretable and adaptable to varying underwater environments.

C. Interpreting Deep Learning in UATR

Despite their significant potential, deep learning models based on neural networks encounter notable limitations in real-world applications primarily due to their "black-box" nature. This opacity makes it challenging to interpret or modify these models in real time, which is crucial for practical deployment. One of the key factors contributing to this limitation is the insufficient incorporation of physical variables into the model's outputs and the underlying mechanisms governing those outputs. To overcome these challenges, it is essential to fully integrate physical variables into the modelling process. This integration would not only enhance the interpretability of the models but also enable real-time adaptations based on changing environmental conditions, ultimately improving their efficacy in practical applications. By bridging the gap between complex neural computations and tangible physical realities, it is possible to unlock the full potential of deep learning in dynamic real-world scenarios.

In summary, addressing the challenges associated with data collection, balancing model complexity with the availability of training data, and ensuring the interpretability and adaptability of models are essential for advancing the field of underwater target recognition. These steps are crucial because effective data collection is foundational for training robust models; without diverse and high-quality datasets, models may fail to generalize across various underwater environments. Balancing model complexity with training data is equally important; overly complex models can lead to overfitting, particularly when trained on limited data. Thus, achieving the right balance allows for improved performance while maintaining generalizability.

Moreover, enhancing model interpretability is vital for fostering trust and facilitating real-time adjustments in dynamic underwater scenarios. By ensuring that models can adapt based on physical variables—such as temperature, salinity, and depth—researchers can create systems that not only perform accurately but are also responsive to changing conditions. Collectively, these advancements will pave the way for more effective and reliable UATR systems, ultimately enhancing their practical applications in marine research, naval operations, and environmental monitoring.

IX. Conclusion

This study systematically elaborates on the background and inherent challenges of Underwater Acoustic Target Recognition (UATR). A comprehensive overview is presented of the key factors influencing UATR performance, including class imbalance, environmental fluctuations, and data noise. Through the analysis of existing literature and the examination of the unique characteristics of various methodologies, recent advancements in UATR are classified according to their applied techniques and underlying properties. The role of deep learning technologies—such as transfer learning, deep convolutional networks, and temporal modelling—is explored in enhancing the accuracy and robustness of UATR systems. Particular attention is given to the challenges of model training, notably the substantial demand for annotated data, highlighting the urgent need for methods that reduce dependence on large-scale labeled datasets. Additionally, the potential of multimodal data fusion is also emphasized, as the integration of multiple data sources—such as underwater acoustic signals and optical imagery—can significantly improve recognition accuracy. Furthermore, ensemble learning approaches and attention-based mechanisms are examined as promising strategies for advancing performance in the field. Despite recent progress, current research indicates a continuing need to improve model interpretability and to better understand the mechanisms contributing to high recognition accuracy. The interpretation and visualization of these models remain largely unexplored, pointing to crucial directions for future research. Overall, the advancement of technologies like unsupervised learning and multimodal data fusion is likely to enhance UATR performance in increasingly complex and demanding application scenarios.

Acknowledgment

Thanks to the School of Computing, University Utara Malaysia (UUM), especially my supervisor Dr. Farhan.

REFERENCES

[1] H. Yang, K. Lee, Y. Choo, and K. Kim, "Underwater acoustic research trends with machine learning: Passive SONAR applications," J. Ocean Eng. Technol., vol. 34, no. 3, pp. 227–236, 2020.

[2] C. C. Leroy, S. P. Robinson, and M. J. Goldsmith, "A new equation for the accurate calculation of sound speed in all oceans," J. Acoust. Soc. Am., vol. 124, no. 5, pp. 2774–2782, 2008.

[3] D. J. Sun, C. E. Zheng, J. C. Zhang, Y. F. Han, and H. Y. Cui, "Development and prospect for underwater acoustic positioning and navigation technology," Bull. Chin. Acad. Sci., vol. 34, no. 3, pp. 331–338, 2019.

[4] Q. Zhang, L. L. Da, C. Wang, Y. H. Zhang, and J. H. Zhuo, "An overview on underwater acoustic passive target recognition based on deep learning," J. Electron. Inf. Technol., vol. 45, no. 11, pp. 4190–4202, 2023.

[5] Z. F. Lei, X. F. Lei, N. Wang, and Q. Y. Zhang, "Present status and challenges of underwater acoustic target recognition technology: A review," Front. Phys., vol. 10, p. 1044890, 2022.

[6] G. Menghani, "Efficient deep learning: A survey on making deep learning models smaller, faster, and better," ACM Comput. Surv., vol. 55, no. 12, pp. 1–37, 2023.

[7] D. Neupane and J. Seok, "A review on deep learning-based approaches for automatic sonar target recognition," Electronics, vol. 9, no. 11, p. 1972, 2020.

[8] S. Feng, S. Ma, X. Zhu, and M. Yan, "Artificial intelligence-based underwater acoustic target recognition: A survey," Remote Sens., vol. 16, no. 17, p. 3333, 2024.

[9] A. Khan, M. M. Fouda, D. T. Do, A. Almaleh, A. M. Alqahtani, and A. U. Rahman, "Underwater target detection using deep learning: Methodologies, challenges, applications, and future evolution," IEEE Access, vol. 12, pp. 12618–12635, 2024.

[10] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, et al., "Deep Speech 2: End-to-end speech recognition in English and Mandarin," in Proc. Int. Conf. Mach. Learn., Jun. 2016, pp. 173–182.

[11] T. P. Marques, A. Rezvanifar, M. Cote, A. B. Albu, K. Ersahin, T. Mudge, and S. Gauthier, "Detecting marine species in echograms via traditional, hybrid, and deep learning frameworks," in Proc. 25th Int. Conf. Pattern Recognit. (ICPR), Jan. 2021, pp. 5928–5935.

[12] A. C. Singer, J. K. Nelson, and S. S. Kozat, "Signal processing for underwater acoustic communications," IEEE Commun. Mag., vol. 47, no. 1, pp. 90–96, 2009.

[13] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am., vol. 50, no. 2B, pp. 637–655, 1971.

[14] D. Liu, H. Yang, W. Hou, and B. Wang, "A novel underwater acoustic target recognition method based on MFCC and RACNN," Sensors, vol. 24, no. 1, p. 273, 2024.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint, arXiv:1409.1556, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.

[17] Z. Chen, G. Xie, M. Chen, and H. Qiu, "Model for underwater acoustic target recognition with attention mechanism based on residual concatenate," J. Mar. Sci. Eng., vol. 12, no. 1, p. 24, 2023.

[18] X. Luo, Y. Feng, and M. Zhang, "An underwater acoustic target recognition method based on combined feature with automatic coding and reconstruction," IEEE Access, vol. 9, pp. 63841–63854, 2021.

[19] Y. Sun, W. Chen, C. Shuai, Z. Zhang, P. Wang, G. Cheng, and W. Yu, "Feature extraction methods for underwater acoustic target recognition of divers," Sensors, vol. 24, no. 13, p. 4412, 2024.

[20] A. Pollara, A. Sutin, and H. Salloum, "Improvement of the detection of envelope modulation on noise (DEMON) and its application to small boats," in Proc. OCEANS 2016 MTS/IEEE Monterey, Sept. 2016, pp. 1–10.

[21] Y. Xie, X. Zhang, J. Ren, and J. Xu, "DEMONet: Underwater acoustic target recognition based on multi-expert network and cross-temporal variational autoencoder," arXiv preprint, arXiv:2411.02758, 2024.

[22] X. Cao, X. Zhang, R. Togneri, and Y. Yu, "Underwater target classification at greater depths using deep neural network with joint multiple-domain feature," IET Radar, Sonar & Navigation, vol. 13, no. 3, pp. 484–491, 2019.

[23] J. Jiang, Z. Wu, J. Lu, M. Huang, and Z. Z. Xiao, "Interpretable features for underwater acoustic target recognition," Measurement, vol. 173, p. 108586, 2021.

[24] X. Wang, A. Liu, Y. Zhang, and F. Xue, "Underwater acoustic target recognition: A combination of multi-dimensional fusion features and modified deep neural network," Remote Sensing, vol. 11, no. 16, p. 1888, 2019.

[25] X. Luo and Y. Feng, "An underwater acoustic target recognition method based on restricted Boltzmann machine," Sensors, vol. 20, no. 18, p. 5399, 2020.

[26] X. Luo, L. Chen, H. Zhou, and H. Cao, "A survey of underwater acoustic target recognition methods based on machine learning," Journal of Marine Science and Engineering, vol. 11, no. 2, p. 384, 2023.

[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, 2015.

[28] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, vol. 61, pp. 85-117, 2015.

[29] X. Cui, Z. He, Y. Xue, K. Tang, P. Zhu, and J. Han, "Cross-domain contrastive learning-based few-shot underwater acoustic target recognition," Journal of Marine Science and Engineering, vol. 12, no. 2, p. 264, 2024.

[30] J. Chen, B. Han, X. Ma, and J. Zhang, "Underwater target recognition based on multi-decision LOFAR spectrum enhancement: A deep-learning approach," Future Internet, vol. 13, no. 10, p. 265, 2021.

[31] J. Xu, Y. Xie, and W. Wang, "Underwater acoustic target recognition based on smoothness-inducing regularization and spectrogram-based data augmentation," Ocean Engineering, vol. 281, p. 114926, 2023.

[32] Q. Yao, Y. Wang, and Y. Yang, "Underwater acoustic target recognition based on data augmentation and residual CNN," Electronics, vol. 12, no. 5, p. 1206, 2023.

[33] J. Tang, E. Ma, Y. Qu, W. Gao, and L. Gan, "UAPT: An underwater acoustic target recognition method based on pre-trained Transformer," 2024.

[34] Y. Xie, J. Ren, and J. Xu, "Guiding the underwater acoustic target recognition with interpretable contrastive learning," in OCEANS 2023-Limerick, pp. 1-6, June 2023.

[35] H. Yang, K. Zheng, and J. Li, "Open set recognition of underwater acoustic targets based on GRU-CAE collaborative deep learning network," Applied Acoustics, vol. 193, p. 108774, 2022.

[36] E. L. Ferguson, R. Ramakrishnan, S. B. Williams, and C. T. Jin, "Convolutional neural networks for passive monitoring of a shallow water environment using a single sensor," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Mar. 2017, pp. 2657-2661.

[37] L. Du, Z. Wang, Z. Lv, D. Han, L. Wang, F. Yu, and Q. Lan, "A method for underwater acoustic target recognition based on the delay-Doppler joint feature," Remote Sensing, vol. 16, no. 11, p. 2005, 2024.

[38] P. Ashok and B. Latha, "An improving recognition accuracy of underwater acoustic targets based on gated recurrent unit (GRU) neural network method," in Proc. 1st Int. Conf. Comput. Sci. Technol. (ICCST), Nov. 2022, pp. 1-6.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Adv. Neural Inf. Process. Syst., vol. 25, 2012.

[40] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), pp. 421-425, Mar. 2017.

[41] D. Li, F. Liu, T. Shen, L. Chen, X. Yang, and D. Zhao, "Generalizable underwater acoustic target recognition using feature extraction module of neural network," Appl. Sci., vol. 12, no. 21, p. 10804, 2022.

[42] G. Hu, K. Wang, and L. Liu, "Underwater acoustic target recognition based on depthwise separable convolution neural networks," Sensors, vol. 21, no. 4, p. 1429, 2021.

[43] G. Hu, K. Wang, Y. Peng, M. Qiu, J. Shi, and L. Liu, "Deep learning methods for underwater target feature extraction and recognition,"Computational Intelligence and Neuroscience, vol. 2018, no. 1, p. 1214301, 2018.

[44] G. Hu, K. Wang, and L. Liu, "A features extraction and recognition method for underwater acoustic target based on ATCNN," arXiv preprint arXiv:2011.14336, 2020.

[45] T. C. Oliveira, Y. T. Lin, and M. B. Porter, "Underwater sound propagation modeling in a complex shallow water environment," Frontiers in Marine Science, vol. 8, p. 751327, 2021.

[46] K. Yang, B. Wang, Z. Fang, and B. Cai, "An end-to-end underwater acoustic target recognition model based on one-dimensional convolution and transformer," Journal of Marine Science and Engineering, vol. 12, no. 10, p. 1793, 2024.

[47] F. Ji, J. Ni, G. Li, L. Liu, and Y. Wang, "Underwater acoustic target recognition based on deep residual attention convolutional neural network," Journal of Marine Science and Engineering, vol. 11, no. 8, p. 1626, 2023.

[48] H. Yang, J. Li, S. Shen, and G. Xu, "A deep convolutional neural network inspired by auditory perception for underwater acoustic target recognition," Sensors, vol. 19, no. 5, p. 1104, 2019.

[49] J. Li and H. Yang, "Deep learning method with auditory passive attention for underwater acoustic target recognition under the condition of ship interference," Ocean Engineering, vol. 302, p. 117674, 2024.

[50] H. Yao, T. Gao, Y. Wang, H. Wang, and X. Chen, "Mobile_ViT: Underwater Acoustic Target Recognition Method Based on Local–Global Feature Fusion," Journal of Marine Science and Engineering, vol. 12, no. 4, p. 589, 2024.

[51] S. Zhang, C. Wang, and Q. Sun, "Underwater target noise recognition and classification technology based on multi-classes feature fusion," Xibei Gongye Daxue Xuebao/Journal of Northwestern Polytechnical University, vol. 38, no. 2, pp. 366-376, 2020.

[52] J. Wu, P. Li, Y. Wang, Q. Lan, W. Xiao, and Z. Wang, "VFR: The underwater acoustic target recognition using cross-domain pre-training with fbank fusion features," Journal of Marine Science and Engineering, vol. 11, no. 2, p. 263, 2023.

[53] Z. Chen, J. Tang, H. Qiu, and M. Chen, "MGFGNet: An automatic underwater acoustic target recognition method based on the multi-gradient flow global feature enhancement network," Frontiers in Marine Science, vol. 10, p. 1306229, 2023.

[54] F. Hong, C. Liu, L. Guo, F. Chen, and H. Feng, "Underwater acoustic target recognition with ResNet18 on Shipsear dataset," in Proceedings of the 2021 IEEE 4th International Conference on Electronics Technology (ICET), May 2021, pp. 1240-1244.

[55] L. C. Domingos, P. E. Santos, P. S. Skelton, R. S. Brinkworth, and K. Sammut, "An investigation of preprocessing filters and deep learning methods for vessel type classification with underwater acoustic data," IEEE Access, vol. 10, pp. 117582-117596, 2022.

[56] S. Tian, D. Chen, H. Wang, and J. Liu, "Deep convolution stack for waveform in underwater acoustic target recognition," Scientific Reports, vol. 11, no. 1, p. 9614, 2021.

[57] T. L. Hemminger and Y. H. Pao, "Detection and classification of underwater acoustic transients using neural networks," IEEE Transactions on Neural Networks, vol. 5, no. 5, pp. 712-718, 1994.

[58] H. Niu, E. Reeves, and P. Gerstoft, "Source localization in an ocean waveguide using supervised machine learning," The Journal of the Acoustical Society of America, vol. 142, no. 3, pp. 1176-1188, 2017.

[59] Y. Honghui, L. Junhao, and S. Meiping, "Underwater acoustic target multi-attribute correlation perception method based on deep learning," Applied Acoustics, vol. 190, p. 108644, 2022.

[60] R. P. Fernandes and J. A. Apolinário Jr, "Underwater target classification with optimized feature selection based on genetic algorithms," in Proc. Simpósio Brasileiro de Telecomunicações e Processamento de Sinais, 2020.

[61] Q. Yu, W. Zhang, M. Zhu, J. Shi, Y. Liu, and S. Liu, "Surface and underwater acoustic target recognition using only two hydrophones based on machine learning," The Journal of the Acoustical Society of America, vol. 155, no. 6, pp. 3606-3614, 2024.

[62] H. Yang, A. Gan, H. Chen, Y. Pan, J. Tang, and J. Li, "Underwater acoustic target recognition using SVM ensemble via weighted sample and feature selection," in Proc. 13th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST), Jan. 2016, pp. 522-527.

[63] Y. Chen and X. Xu, "The research of underwater target recognition method based on deep learning," in Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC), Oct. 2017, pp. 1-5.

[64] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., "Generative adversarial nets," in Adv. Neural Inf. Process. Syst., vol. 27, 2014.

[65] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1-9, 2015.

[66] Y. Xie, J. Xu, J. Ren, and J. Li, "Adversarial multi-task underwater acoustic target recognition: Toward robustness against various influential factors," J. Acoust. Soc. Am., vol. 156, no. 1, pp. 299-313, 2024.

[67] X. Luo, M. Zhang, T. Liu, M. Huang, and X. Xu, "An underwater acoustic target recognition method based on spectrograms with different resolutions," J. Mar. Sci. Eng., vol. 9, no. 11, p. 1246, 2021.

[68] X. Liu, H. Zhu, W. Song, J. Wang, L. Yan, and K. Wang, "Research on improved VGG-16 model based on transfer learning for acoustic image recognition of underwater search and rescue targets," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 2024.

[69] M. Li, K. Liu, H. Li, Y. Sun, X. Chen, and K. Mao, "Quantitative analysis on the influence of the oceanic front on underwater acoustic detection with investigated marine data," J. Mar. Sci. Eng., vol. 11, no. 8, p. 1574, 2023.

[70] X. Xiao, W. Wang, Q. Ren, P. Gerstoft, and L. Ma, "Underwater acoustic target recognition using attention-based deep neural network," JASA Express Lett., vol. 1, no. 10, 2021.

[71] Y. Xie, J. Ren, and J. Xu, "Underwater-ART: Expanding information perspectives with text templates for underwater acoustic target recognition," J. Acoust. Soc. Am., vol. 152, no. 5, pp. 2641-2651, 2022.

[72] Z. Pu, Q. Zhang, Y. Xue, P. Zhu, and X. Cui, "A novel multi-feature fusion model based on pre-trained Wav2vec 2.0 for underwater acoustic target recognition," Remote Sens., vol. 16, no. 13, p. 2442, 2024.

[73] J. Tang, W. Gao, E. Ma, X. Sun, and J. Ma, "Deep learning based underwater acoustic target recognition: Introduce a recent temporal 2D modeling method," Sensors, vol. 24, no. 5, p. 1633, 2024.

[74] P. Li, J. Wu, Y. Wang, Q. Lan, and W. Xiao, "STM: Spectrogram transformer model for underwater acoustic target recognition," Journal of Marine Science and Engineering, vol. 10, no. 10, p. 1428, 2022.

[75] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," Neural Computation, vol. 18, no. 7, pp. 1527-1554, 2006.

[76] M. Zhu, X. Zhang, Y. Jiang, K. Wang, B. Su, and T. Wang, "Hybrid Underwater Acoustic Signal Multi-Target Recognition Based on DenseNet-LSTM with Attention Mechanism," in Chinese Intelligent Automation Conference, Singapore: Springer Nature Singapore, pp. 728-738, Sep. 2023.

[77] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015.

[78] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[79] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[80] D. Santos-Domínguez, S. Torres-Guijarro, A. Cardenal-López, and A. Pena-Gimenez, "ShipsEar: An underwater vessel noise database," Applied Acoustics, vol. 113, pp. 64-69, 2016.

[81] X. Ke, F. Yuan, and E. Cheng, "Underwater acoustic target recognition based on supervised feature-separation algorithm," Sensors, vol. 18, no. 12, p. 4318, 2018.

[82] N. Müller, J. Reermann, and T. Meisen, "Navigating the Depths: A Comprehensive Survey of Deep Learning for Passive Underwater Acoustic Target Recognition," IEEE Access, 2024.

[83] T. O. Fahad, A. H. Miry, A. Al-Gizi, M. H. Miry, and A. T. Razzooqee, "Recognition of Underwater Acoustic Radar Signals Based on Multiresolution and Dense Convolutional Neural Network," Journal of Engineering and Sustainable Development, vol. 28, no. 6, pp. 793-800, 2024.

[84] C. Tang and G. Hu, "DSCANet: Underwater Acoustic Target Classification Using the Depthwise Separable Convolutional Attention Module," Earth Science Informatics, vol. 17, no. 6, pp. 6123-6135, 2024.

[85] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," arXiv preprint arXiv:1511.06434, 2015.

[86] M. Irfan, J.B. Zheng, S. Ali, M. Iqbal, Z. Masood, and U. Hamid, "DeepShip: An Underwater Acoustic Benchmark Dataset and a Separable Convolution Based Autoencoder for Classification," Expert Systems with Applications, vol. 183, p. 115270, 2021.

[87] A. Christensen, A. Sen Gupta, and I. Kirsteins, "Underwater Small Target Classification Using Sparse Multi-View Discriminant Analysis and the Invariant Scattering Transform," Journal of Marine Science and Engineering, vol. 12, no. 10, p. 1886, 2024.

[88] M. R. Azimi-Sadjadi, D. Yao, Q. Huang, and G. J. Dobeck, "Underwater target classification using wavelet packets and neural networks," IEEE Transactions on Neural Networks, vol. 11, no. 3, pp. 784-794, 2000.

[89] L. Xue, X. Zeng, and A. Jin, "A novel deep-learning method with channel attention mechanism for underwater target recognition," Sensors, vol. 22, no. 15, p. 5492, 2022.

[90] A. Jin and X. Zeng, "A novel deep learning method for underwater target recognition based on res-dense convolutional neural network with attention mechanism," Journal of Marine Science and Engineering, vol. 11, no. 1, p. 69, 2023.