Journal of *Marine Science and Engineering*

MDPI

Article

Model for Underwater Acoustic Target Recognition with Attention Mechanism Based on Residual Concatenate

Zhe Chen [1,2], Guohao Xie [3], Mingsong Chen [1,2] and Hongbing Qiu [1,2,*]

[1] School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, China; chenzhe2015@guet.edu.cn (Z.C.); cms@guet.edu.cn (M.C.)
[2] Cognitive Radio and Information Processing Key Laboratory Authorized by China's Ministry of Education Foundation, Guilin University of Electronic Technology, Guilin 541004, China
[3] School of Ocean Engineering, Guilin University of Electronic Technology, Beihai 536000, China; 22172301007@mails.guet.edu.cn
* Correspondence: qiuhb@guet.edu.cn

Abstract: Underwater acoustic target recognition remains a formidable challenge in underwater acoustic signal processing. Current target recognition approaches within underwater acoustic frameworks predominantly rely on acoustic image target recognition models. However, this method grapples with two primary setbacks; the pronounced frequency similarity within acoustic images often leads to the loss of critical target data during the feature extraction phase, and the inherent data imbalance within the underwater acoustic target dataset predisposes models to overfitting. In response to these challenges, this research introduces an underwater acoustic target recognition model named Attention Mechanism Residual Concatenate Network (ARescat). This model integrates residual concatenate networks combined with Squeeze-Excitation (SE) attention mechanisms. The entire process culminates with joint supervision employing Focal Loss for precise feature classification. In our study, we conducted recognition experiments using the ShipsEar database and compared the performance of the ARescat model with the classic ResNet18 model under identical feature extraction conditions. The findings reveal that the ARescat model, with a similar quantity of model parameters as ResNet18, achieves a 2.8% higher recognition accuracy, reaching an impressive 95.8%. This enhancement is particularly notable when comparing various models and feature extraction methods, underscoring the ARescat model's superior proficiency in underwater acoustic target recognition.

Keywords: SE attention mechanism; residual network (ResNet); underwater acoustic target recognition; feature extraction

Citation: Chen, Z.; Xie, G.; Chen, M.; Qiu, H. Model for Underwater Acoustic Target Recognition with Attention Mechanism Based on Residual Concatenate. *J. Mar. Sci. Eng.* **2024**, *12*, 24. https://doi.org/10.3390/jmse12010024

Academic Editor: Rouseff Daniel

Received: 10 November 2023
Revised: 10 December 2023
Accepted: 15 December 2023
Published: 20 December 2023

1. Introduction

Traditional underwater acoustic target recognition methods relying on signal analysis often face challenges; they demand heightened computational efficiency, rely on manual parameter tweaking, exhibit low reliability, have limited application scenarios, and lack strong generalizability.

However, the rise of deep learning, especially with its capacity to abstract and learn from heterogeneous features, offers promise. It not only enhances the adaptability of recognition models but also realizes end-to-end information flow and autonomous recognition. As long-term oceanic observations continue to amass passive data on underwater acoustic targets from diverse marine environments, it furnishes a robust foundation for deep learning explorations, particularly within the paradigm of deep ocean big data.

Deep learning techniques have increasingly found applications in underwater acoustic target recognition [1–3]. Two pivotal aspects underscore this approach: the workings of neural networks and feature extraction. For feature extraction in deep learning, techniques like Mel-frequency Cepstrum Coefficient (MFCC) [4], Constant-Q Transform (CQT) [5,6], wavelet features [7,8], Detection of Envelope Modulation on Noise (DEMON)

and Low-frequency Analyzer and Recorder (LOFAR) spectra [9], and advanced spectral features [10–13] are employed. Moreover, deep learning can directly assimilate features from raw signals. Such methodologies efficiently filter out extra data, thereby curtailing the computational demands of subsequent models. In a comparative analysis, Yang et al. revealed that 3D dynamic Mel-frequency Cepstrum Coefficient (3D MFCC) surpasses MFCC, Mel spectrograms, 3D dynamic Mel spectrograms, and CQT in terms of efficacy [14]. Optimal performance and efficient feature extraction in this domain are intricately tied to the neural network's architecture. Although large-scale deep models offer profound insights, they come at a significant computational cost [15]. A tangible trade-off exists between model accuracy and operational efficiency, with the latter emphasizing rapid results with fewer resources. In addition to methods for spectrogram processing explored in our study, recent years have seen a rise in alternative approaches. Physically based techniques, such as Yann Le Gall et al.'s [16] passive underwater acoustic filtering scheme, have been proposed to differentiate time-frequency fringes of moving vessels. On the source side, researchers like Kuz'kin V.M. [17] have analyzed the resolving power of source localization interferometry. Further contributions by Ehrhardt M. et al. [18] and Pereselkov S.A. et al. [19] delve into interference patterns in acoustic fields and hydroacoustic signal interference for source localization, respectively.

Interestingly, in the realm of computer vision, the ubiquity of ResNet has been challenged [20]. Gao et al. underscored the surplus layers in ResNet, showcasing that removing specific layers during training barely affected the algorithm's convergence or the outcomes [21]. In a related experiment, Tian et al. examined the network depth of a multiscale residual deep neural network (MSRDN) designed for underwater acoustic target recognition [22]. Their findings indicated the superfluity of several layers in the original MSRDN. This notion of redundancy in ResNet was echoed by Xue et al., who observed a drop in identification rates by adding more residual layers [23]. Looking ahead, Lei et al. highlight a pivotal research direction: minimizing computational expenses in underwater acoustic data processing [24]. Current methodologies often grapple with striking a harmonious balance between accuracy and efficiency in deep learning applications.

Advancements in attention mechanisms have showcased remarkable progress in mitigating interference challenges within residual networks. In a study by Chen et al., reverse attention was leveraged to accentuate residual details within the residual network, facilitated by top-down guidance, enhancing recognition performance [25]. Lu et al. explored spectral and spatial features through a tri-layer parallel residual network framework, subsequently integrating a 3D attention module to amplify component representation, yielding improved classification outcomes [26].

The loss function stands paramount in the model training paradigm, serving as a metric to gauge the divergence between the model's predictions and ground truth, steering model optimization. One pivotal role of the loss function is to mitigate target value variance. Hong et al. employed a joint loss function, honing in on the attributes of an array of underwater acoustic targets [27]. Nonetheless, target recognition efficacy diminishes when a universal loss function indiscriminately monitors every parameter, encompassing interference factors like aquatic background noise.

Drawing inspiration from the attention mechanism, we introduce the "SE attention mechanism residual concatenate network", engineered to address representation deficiencies stemming from the restricted effective receptive field intrinsic to residual networks. Our model ingests features distilled via the 3D MFCC operation, adeptly condensing the original time domain data of the target signal and extrapolating its dynamic temporal nuances. In the vein of feature extraction, we put forth the Attention Mechanism Residual Concatenate Specify Dimensions Block (ARCSB) structure, an ode to ResNet's principles. This architecture optimizes model parameters by curtailing ResNet's residual units and weaving them in a parallel concatenation framework, ensuring a harmonious equilibrium between accuracy and computational efficiency [28]. To further this, a streamlined attention mechanism approach, termed Squeeze-Excitement (SE) [29,30], is infused into our model.

This aids the network in attributing differential weights to input features, facilitating the extracting of paramount information. By embedding the SE attention mechanism blueprint into the ARCSB architecture, we intensify the interplay among elements, bolstering the convolutional network's prowess in isolating frequency traits from auditory visuals.

Further augmenting our model's precision, we incorporate a maximum pooling layer and the ASPP module. These components steer the model towards discerning pivotal facets of acoustic imagery, offering meticulous data extraction, purging extra data, and fortifying model robustness. For classification, a Multilayer Perceptron (MLP) discerns underwater acoustic targets. Notably, habitual reliance on the Cross-entropy Loss function can induce model underfitting, especially given the sample distribution skewness of the ShipsEar dataset. To enhance the network's adaptability to such imbalanced datasets, Focal Loss [31] is utilized, rectifying sample disparity.

In summary, our manuscript's salient contributions encompass the following:

(1) The unveiling of the ARescat network, adept at conserving target features amidst extraction and recognition processes. Infusing channel SE attention mechanism, it accentuates disparate channels during pivotal information extraction, adeptly negating ambient noise and upholding network adaptability amidst environmental flux.

(2) The introduction of Focal Loss as a remedy for dataset imbalance. Through its application, we can facilitate balanced feature supervision, including interference nuances like marine ambient noise. This methodology addresses sample disproportion and ensures balanced feature oversight, excavating salient details from merged components.

(3) The accuracy of the proposed model ARescat was verified to be higher than other models on the shipsEar dataset.

2. System Overview

This section introduces the classification framework for underwater acoustic target recognition. The ARescat network model is presented in Section 1, and the extraction technique for 3D MFCC features is shown in Section 2. Section 3 introduces the modules in the ARescat network, and Section 4 presents the Focal Loss idea to solve the sample imbalance problem.

2.1. Construction of the Proposed Model

ARescat Network Model (Refer to Figure 1)

The ARescat model bifurcates into two pivotal segments: feature extraction and classification. Within feature extraction, audio signal frames transform to be represented in 3D MFCC format, capturing both the original signal target and dynamic temporal details. The processed 3D MFCC is iterated using two ARCSB modules and a Maxpool module. This strips redundant frequency information. These features integrate through the ARCSB module, the ASPP module, and the Maximum Pooling Layer module. This amalgamation elevates the model's prowess in delineating underwater acoustic target features. The culminating step involves feeding these refined features into the MLP for categorization, facilitated by connecting two fully connected (FC) layers.

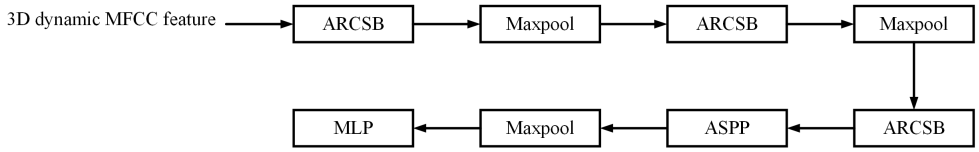

Figure 1. ARescat network model.

2.2. 3D MFCC Feature Extraction Methods

The proposed model is engineered to accept features gleaned through the 3D MFCC procedure (depicted in Figure 2). The extraction process unravels in a series of meticulous steps:

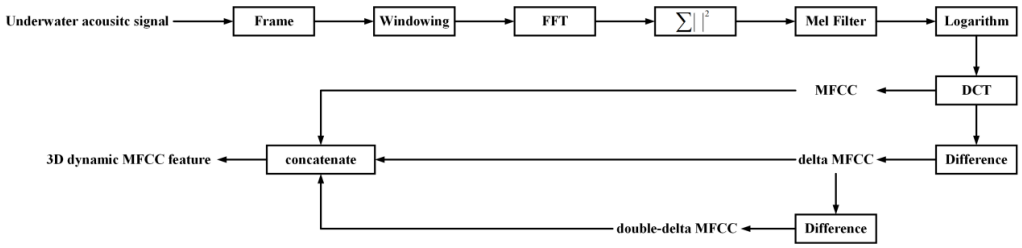

Figure 2. 3D dynamic MFCC feature extraction frame.

1. Frame Splitting and Windowing: MFCC features begin segmenting frames and leveraging the Hanning window for frequency domain exploration. This foundation is pivotal for speech segmentation and feature extraction. For analysis, a frame length of 2048 bits is established, intersecting with a 75% overlap.

2. FFT and Power Spectrum Computation: Each frame experiences a Fast Fourier Transform. Subsequent squaring and summing yield the power spectrum, crucial for aggregating both time- and frequency-domain data.

3. Mel Filter Bank and DCT: The power spectrum undergoes a logarithmic transformation through the Mel filter bank, yielding the log Mel spectra. Ultimately, discrete cosine transformation distills these spectra into MFCC features. For a representative 5 s audio snippet with a 22,050 Hz sampling rate, the MFCC assumes the shape of (128 × 216). As input to ARescat, *Mel_3D* with 128 × 216 × 3 (*n_mels* = number of mel filters = 128, N = number of frames = 216, C = channel = 3) is harvested.

Differential Features: A novel dimension to this method is integrating delta features, which is pivotal for accentuating recognition accuracy. Traditional spectral features are constrained, only capturing static properties. Differential operations on MFCC unlock the dynamic attributes, termed the first-order differential MFCC features. The incremental spectral feature D_t is defined as follows, presuming that the MFCC at frame t is c_t.

$$D_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{1}$$

Here, n is the number of adjacent frames, and c_t is the inverse spectral coefficient of the static Mel spectrogram for frame t. By the difference operation, D_t is the difference coefficient of the Mel spectrogram calculated from the static coefficients c_{t-n} and c_{t+n}. The time-frequency spectrogram's difference dimension is indicated by the letter N. The 3D spectrogram is recorded as:

$$Mel\_3D \in R^{F \times T \times C} \tag{2}$$

This dynamic representation of the spectrogram is three-dimensional, encapsulated by dimensions F (Mel filter bins), T (time), and C (spectrogram dimensions).

Figure 3a illustrates the time-domain waveforms of sailboat radiated noise from the ShipsEar database. Mel-Frequency Cepstral Coefficients (MFCC) are a staple in acoustic signal processing, especially in speech recognition. They emulate human auditory perception of sound frequencies, capturing key speech features. Traditional MFCC, however, primarily capture static spectral envelope information, lacking in dynamic signal representation. This gap is bridged by computing first and second-order differences of the MFCC (Delta and Double-delta MFCC), as depicted in Figure 3b,c. These derivatives provide temporal dynamics, enhancing the representation of speech signal variations in rate and articulation. The integration of MFCC, Delta MFCC, and Double-delta MFCC offers a more holistic view of acoustic features, as shown in Figure 3d, crucial for effective automatic speech recognition systems.

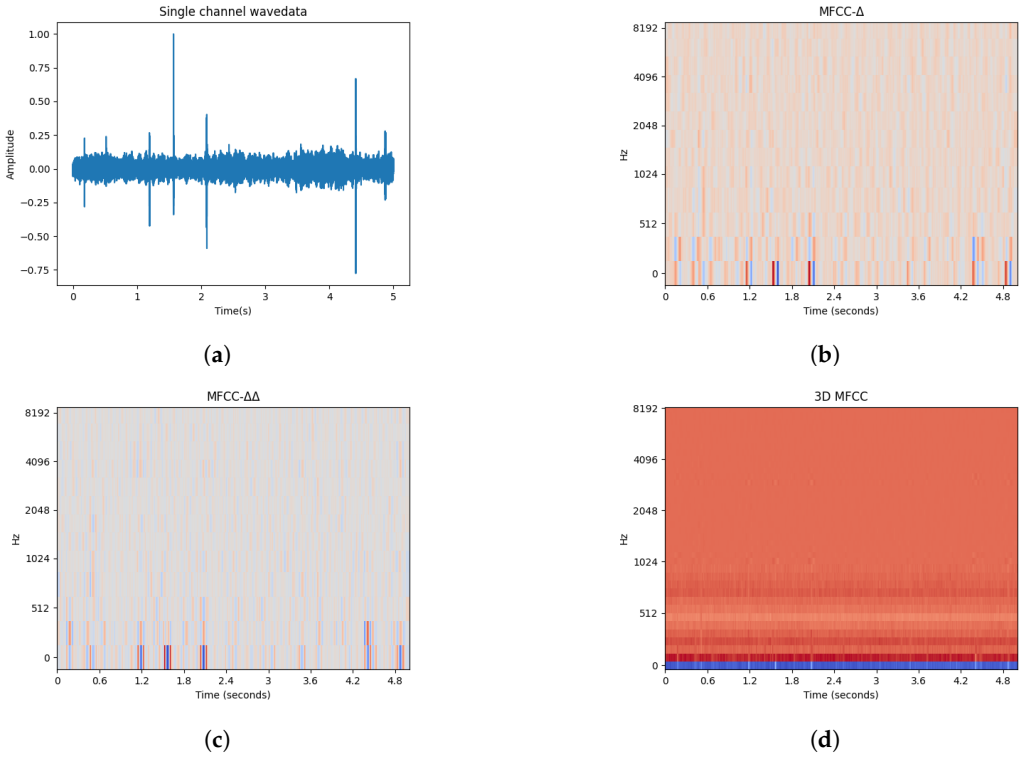

Figure 3. 3D_Mel spectrum generated from the original signal. (**a**) The single channel waveform; (**b**) delta MFCC feature; (**c**) double-delta MFCC frature; (**d**) 3D MFCC frature.

To visualize this process, Figure 3 illustrates the primal order differential MFCC, the secondary dynamic MFCC, and 3D MFCC consisting of MFCC and delta mfcc and double-delta mfcc, specifically from the time-domain waveforms of sailboat radiated noise sourced from the ShipsEar database.

2.3. ARCSB Network

2.3.1. Structure and Components

The ARCSB network, as illustrated in Figure 4, amalgamates three integral modules:

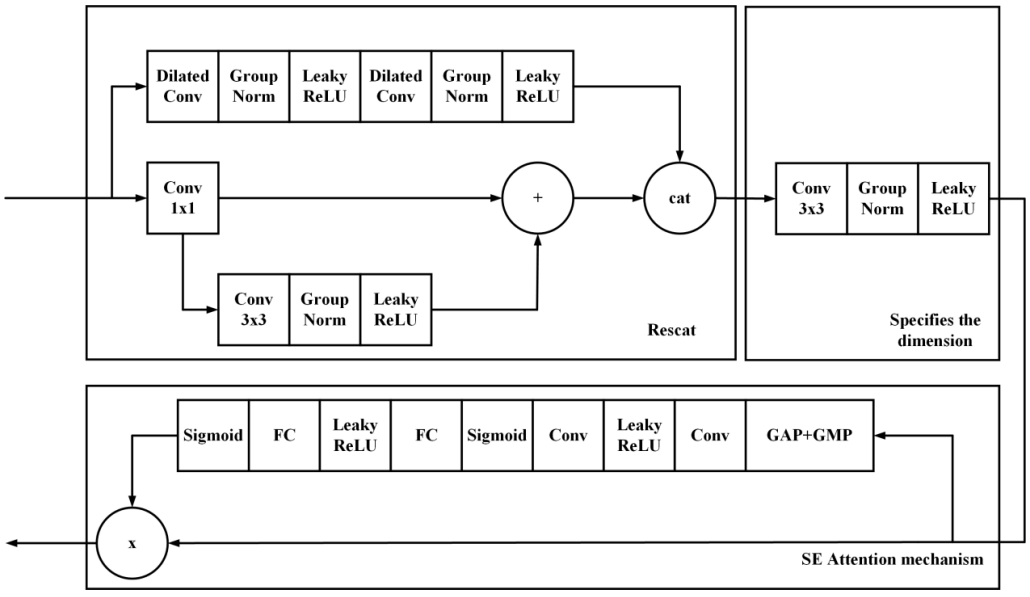

Figure 4. The ARCSB network.

1. The Residual concatenate (Rescat) module;
2. The Specified dimension module;
3. The SE attention mechanism module.

Their union is instrumental in robustly extracting feature information and approximating the categorical probabilities based on spectral frequencies.

1. Rescat Module:

Due to the high similarity of frequency features in acoustic images, the module uses superposition computation to obtain global details, obtain the information of a larger sensory field to represent the target features, increase its feature extraction, and fuse the spectrogram features extracted from different convolutions together to build a more fine-grained representation of the features and improve the model's expressive ability. Within the Rescat module, the incoming data, designated as x, are bifurcated into two pathways.

The upper trajectory processes x using dual 3×3 dilated convolutions. This not only broadens the receptive field but also maintains the resolution, ensuring that the contextual multi-scale data are effectively harnessed.

The subsequent pathway transforms the channel number of x. This entails initial processing with a 1×1 convolution, followed by a 3×3 convolutional layer. The residual network structure here ensures the network is more robust and adaptive, averting potential gradient-related issues and priming the features for subsequent module integration. The culmination of the module sees the two pathways conjoin, enhancing feature extraction.

Moreover, the spectrogram features derived from varied convolutions are amalgamated, generating a more intricate representation of features and bolstering the expressive capacity of the model. Hence, we can extract and fuse supplementary multi-scale detail information to attain more precise target features. The formula for the receptive field can be defined as follows:

$$l_k = l_{k-1} + \left[(f_k - 1) * \prod_{i=1}^{k-1} s_i \right] \tag{3}$$

where f_k is the size of the kth layer's convolution kernel, or the pooling dimension of the pooling layer, and I_{k-1} is the size of the receptive field corresponding to the $k-1$st layer. s_i is the stride size of layer i. Both s and k denote the number of layers.

2. Specified Dimension Module:

This module utilizes a 3×3 convolutional layer, group normalization, and a leaky ReLU activation function. The choice of a 3×3 convolution is pivotal for minimizing the model parameters, thereby enhancing training efficiency and generalizability. The activation function aids in abstracting spatial data, while the normalization layer processes batches of data. This coordination bolsters the model's feature extraction capacity, acting as a bridge between the prior Rescat module and the impending SE attention mechanism.

3. SE Attention Mechanism Module:

As depicted in Figure 5, this module focuses on the fusion of global context data. Core operations include:

(1) Squeeze: Uses global average pooling and global Max pooling to transform feature maps, concentrating on the global context.

(2) Middle Section: Bolsters the abstract representation capacity of the network's local section through dual convolution layers.

(3) Excitation: Harnesses a two-layer fully connected sequence to determine channel weights in the feature map. By emphasizing critical features and downplaying less reliable components, the module enhances the extraction capabilities of the network.

In essence, SE attention mechanism module permits the model to focus more intently on the core segments of the auditory image frequency. This targeted approach prevents potential losses of target features, ensuring a more refined and accurate feature extraction process.

While the "Squeeze" function focuses on spatial reduction through pooling operations, the "Excitation" mechanism globally and comprehensively characterizes the ship's acoustic signal. By leveraging fully connected layers, the Excitation mechanism gauges the significance of each channel, modeling inter-channel relationships with learned parameters.

The formula for the "Squeeze" operation is as follows:

$$z = \sigma\left(W_2\left(W_1\left(F^c_{\text{avg}}\right)\right) + W_2(W_1(F^c_{\text{max}}))\right) \tag{4}$$

where z is the extracted feature, F^c_{avg} is the global average pooling of the feature map to compress it into a feature vector, F^c_{max} is the global maximum pooling of the feature map to compress it into a feature vector, σ denotes the sigmoid activation function, W_1, W_2 both denote the dimension, $W_1 \in R^{\frac{C}{r}\times C} W_2 \in R^{C\times\frac{C}{r}}$.

The "excitation" equation looks like this:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2\delta(W_1 z)) \tag{5}$$

In the above equation, σ denotes the sigmoid activation function and δ denotes the leaky ReLU activation function, while W_1, W_2 both denote the dimension, $W_1 \in R^{\frac{C}{r}\times C} W_2 \in R^{C\times\frac{C}{r}}$. F_{ex} is the excitation operation, i.e., it is equivalent to recalibrating the weights W with the features z taken out earlier. W is the weights, z is the features taken out earlier, and s denotes the weights of each channel.

To reduce the complexity of the model and to generalize it, two FC layers are used to parameterize the pick-and-pass mechanism, i.e., a dimensionality reduction layer with a reduced dimensionality rate r, a leaky ReLU, followed by a dimensionality elevation layer, which then moves on to the channel dimensions of the output feature map. Subsequently, dimensionality is escalated to align with the output feature map's channel dimensions. The entire block's outcome is derived from the feature map's rescaling through the activation function. The final production of SE_Block is as follows:

$$\widetilde{X}_C = F_{\text{scale}}(u_c, s_c) = s_c u_c \tag{6}$$

where $\widetilde{X} = [\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_c]$ and $F_{\text{scale}} = (u_c, s_c)$ refers to channel-wise multiplication between the scalar s_c and the feature map $u_c = R^{H\times W} \cdot s_c$ is the activations obtained through interchanneling, and u_c is the output of feature extraction.

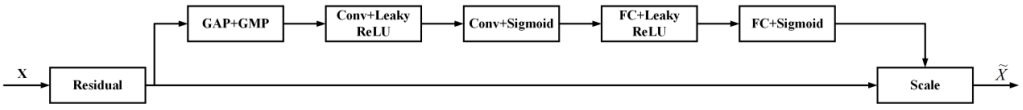

Figure 5. Structure of the SE attention mechanism module.

2.3.2. ASPP (Atrous Spatial Pyramid Pooling)

Designed to bolster the underwater acoustic target recognition capability, the ASPP module, depicted in Figure 6, enriches the network's feature representation and sensing prowess. Notably, it enlarges the receptive field without sacrificing the sampling rate. Its chief advantages include an enhanced capacity for multiscale context capture, precise multiscale feature extraction, and achieving global features with minimal resolution degradation.

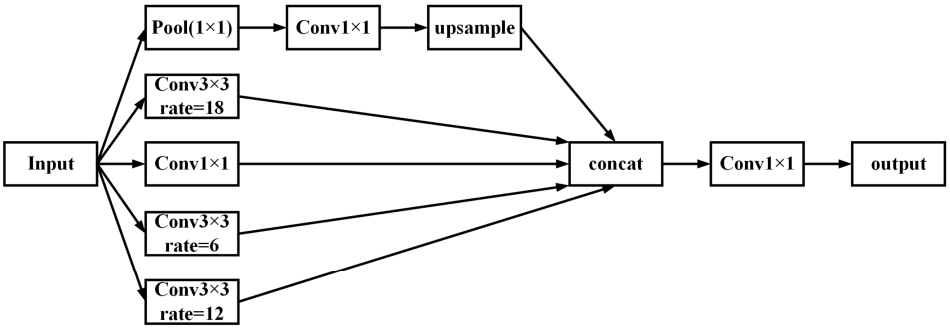

Figure 6. ASPP network structure.

2.4. MLP (Multilayer Perceptron)

Serving as a classifier, the MLP, as showcased in Figure 7, is optimized to handle deep learning features replete with higher-order terms and intricate correlations. These features, inherently aligned with target information, are processed by the MLP, which, for an input of 27,648 data points, connects these data to an array of neurons, culminating in the output layer. The output then presents five classification outcomes.

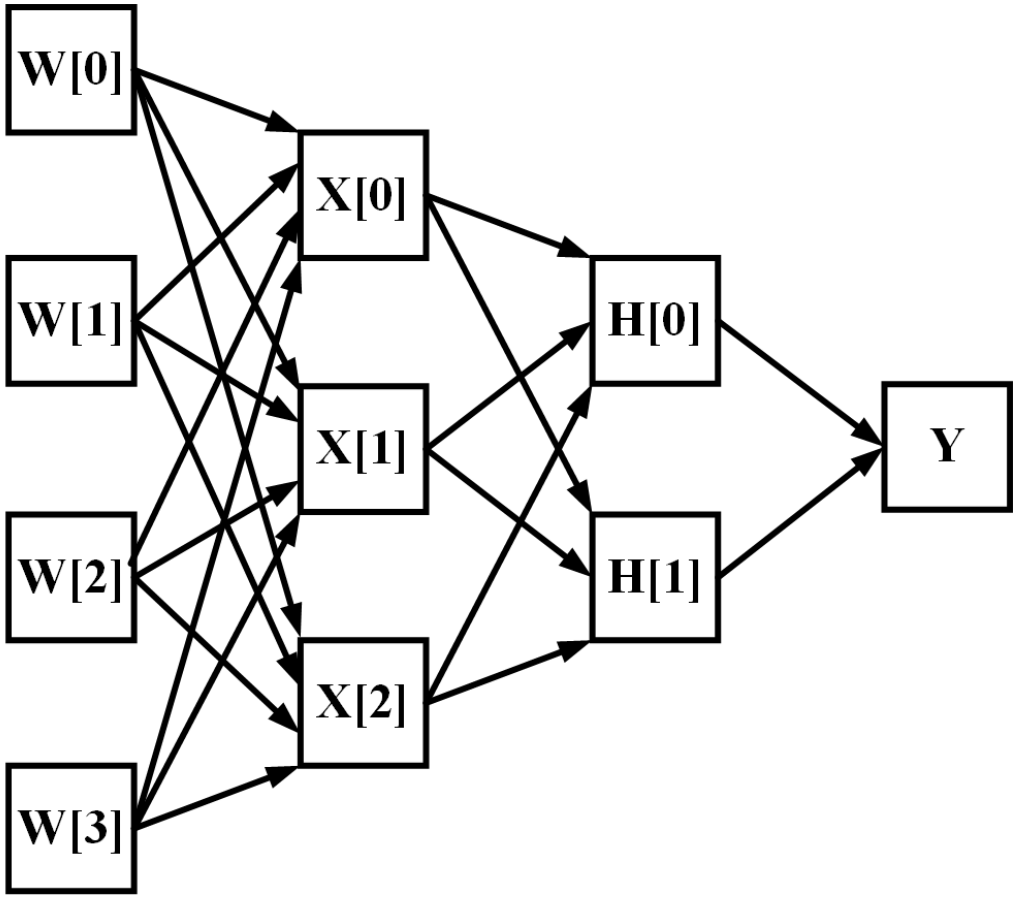

Figure 7. MLP model.

Figure 7 details the network architecture of the Multilayer Perceptron (MLP) used in our model. W[] denotes the input unit, and x[] and H[] represent the first and second hidden layers, respectively. Each hidden layer processes the input and relays it to the next layer. The output unit, Y, receives inputs from all hidden units, synthesizes them, and delivers the final classification result. The network structure follows a fully connected feedforward design, with each input unit connected to every hidden unit and vice versa.

Our MLP comprises two hidden layers, with an input of 27,648 features. The first hidden layer outputs 1000 features, which feed into the second hidden layer, resulting in an output of 100 features. These are then weighted and summed to compute the final results, leading to a five-category classification.

2.5. Focal Loss

Directly utilizing the Cross-entropy Loss function with the ShipsEar database (Figure 8) could lead to performance deficits and a heightened risk of model underfitting due to the dataset's uneven sample distribution. In addressing this imbalance, the Focal Loss method is introduced.

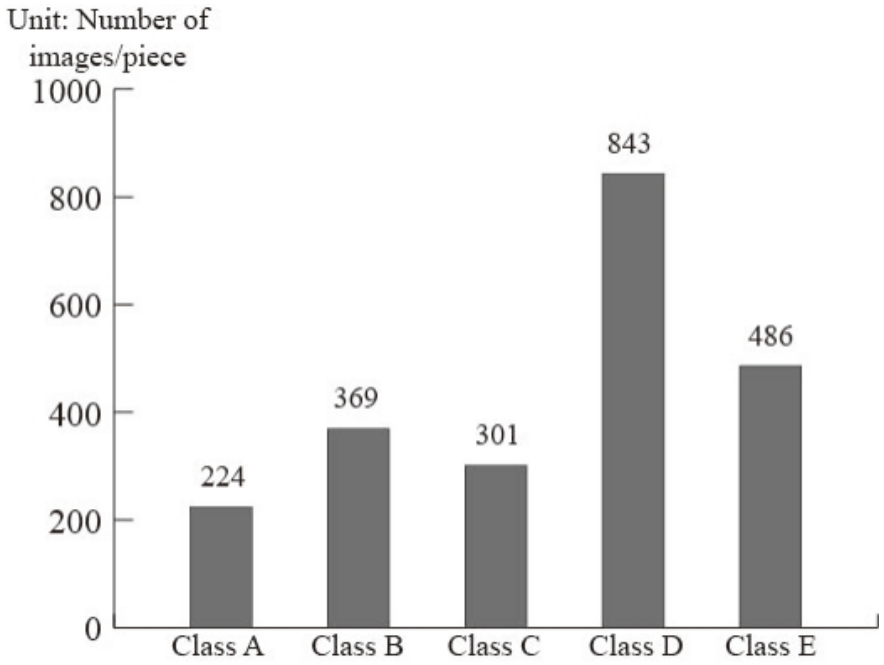

Figure 8. ShipsEar dataset sample categories and their numbers.

The definition of Focal Loss is as follows:

$$L_{fl} = \begin{cases} -\alpha(1-y')^{\gamma}\log y', y = 1 \\ -(1-\alpha)y'^{\gamma}\log(1-y'), y = 0 \end{cases} \quad (7)$$

where α is the adjustment of the positive and negative sample imbalance coefficients. There is a parameter that controls for imbalance in the difficulty sample, while y and y' represent the true label value and the predicted probability, respectively. The formula for Focal Loss incorporates the predicted sigmoid output y' and two parameters, α and γ, which, for all classification experiments in this context, are set to 0.25 and 2, respectively.

In essence, Focal Loss necessitates the proportion knowledge of each class dataset when training the model. Focal Loss makes the model more focused on hard-to-categorize samples during training by reducing the weight of easy-to-categorize samples. This strategic application permits the network to showcase commendable performance improvements for datasets with imbalanced samples, leading to a comprehensive enhancement in the model's overall performance.

These components collectively ensure that the ARCSB network and the related techniques can efficiently process underwater acoustic data, capture vital features, and make accurate classifications, even when confronted with challenges like sample imbalance.

3. Experimentation and Analysis

3.1. Dataset Description

The ShipsEar database is a prominent benchmark for underwater acoustic target recognition and has been employed in numerous scientific investigations. To gauge the effectiveness of our proposed model, we leveraged the ShipsEar database. This collection comprises recordings of ship-emitted noise sourced from the Spanish Atlantic coast, complemented by both human-generated and natural ambient noises. The database is organized into 90 WAV recordings, segmented into five categories, each encompassing one or multiple targets. The categorization and count of files are enumerated in Table 1.

Table 1. 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

To facilitate the study, the database underwent a preprocessing stage. All auditory recordings were standardized to a sampling rate of 22,050 Hz. Adopting a fixed time frame of 5 s, we extracted 2223 annotated audio samples. During data partitioning for model training, each 5-second audio segment was recognized as an individual sample. Of the cumulative pieces (2223), a substantial majority (1778) were allocated for training, with the remainder (445) reserved for testing, observing an 8:2 split ratio.

3.2. Hyperparameter Configuration and Loss Function Design

Our model's training capitalized on momentum (set at 0.9) coupled with the Lion's optimizer [32], resulting in effective mitigation of sample noise disruption. The entirety of the training regimen spanned 200 epochs. In a bid to expedite the learning process, the initial learning rate of 0.0004 was multiplied by the cosine decay function. As a result, we derived the learning rate for the entire learning procedure. The training was executed with a batch size of 4, and Focal Loss was designated as the primary loss function.

3.3. Performance Evaluation

Experiments were orchestrated using the ShipsEar dataset to appraise the proficiency of the devised model. The computational environment encompassed a Windows 11 OS backed by an Intel Core i7-12700H CPU, 32 GB RAM, an NVIDIA GeForce GTX 3070TI GPU, and the Pytorch 1.4.0 framework. The succeeding sub-sections dissect the experimental outcomes.

The following is the introduction of each indicator in the table. Precision refers to the proportion of positive samples that are judged positive by the classifier. Recall refers to the proportion of positive cases that are predicted to be positive as a percentage of the total number of positive cases. The F1-score, a measure of the classification problem, is the reconciled average of precision and recall. Support refers to the number of validation sets for each category.

We employed precision, recall, and F1-score metrics to evaluate the network's recognition capabilities on the provided dataset. The corresponding formulas for each metric are as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Re call} = \frac{TP}{TP + FN} \tag{9}$$

$$\text{F1}-\text{score} = \frac{2TP}{2TP + FP + FN} \tag{10}$$

Here, TP represents the prediction of the correct answer, FP stands for the mistaken prediction from other classes to this class, and FN indicates that this category of labels is predicted to be the other category of labels.

3.3.1. Model Performance Insights

We use the 3D MFCC extracted from the ShipsEar dataset as the data input to the model to evaluate our ARescat model. Since the test set is not involved in model training, the test accuracy can objectively evaluate the model performance. A comprehensive performance analysis encompassing recall, accuracy, and F1 scores is delineated in Table 2. Although specific categories did not surpass the 0.90 accuracy threshold, the overarching Precision, Recall, and F1-score, averaging at 0.958, offer optimism. In this scenario, the term "support" references the number of samples in a solitary test. For comparative analyses, we employ "average Precision" to measure classification accuracy. The classifier exhibits commendable prowess in discerning between Class A (ambient noise) and the distinct classes B–E. Notably, the vessel classification E, entailing ocean liners and roll-on/roll-off vessels, boasted a stellar classification rate of 1, translating to a 100% success quotient. Conversely, Class B, encompassing vessels like dredgers, fishing vessels, and tugboats, demonstrated suboptimal performance, potentially attributable to the pronounced background noise endemic to shallow water acoustics.

Table 2. Results of ARescat with three-dimensional features.

Class	Precision	Recall	F1-Score	Support
Class A	0.960	1.000	0.984	74
Class B	0.958	0.804	0.867	60
Class C	0.937	0.984	0.959	169
Class D	0.970	0.966	0.970	97
Class E	1.000	1.000	1.000	45
Average	0.958	0.958	0.958	445

A confusion matrix visualized in Figure 9 demystifies the classification outcomes on the ShipsEar dataset. The matrix employs a numerical nomenclature, with classes A–E represented by numbers 0–4. The diagonal elements signify accurate classification counts for the respective categories. The five elements from left to right in the diagonal of Figure 9 are 74, 48, 165, 94, and 45, which correspond to the number of samples correctly categorized in each of the five classes in class A to class E in the ShipsEar database. The elements in the non-diagonal line are the number of samples that were incorrectly categorized. Class B emerges as the primary challenge, while other classes are discerned with relative efficacy. It is pivotal to acknowledge that experimental outcomes offer insights into the optimal parameters—filter size, layer count, filter count, batch size, the initial learning rate, or early stopping criterion—for both training and validation stages. While a degree of latitude exists in parameter selection, deviations from the prescribed configurations might influence performance. Incorporating supplementary hyperparameters, however, is unlikely to yield significant performance shifts.

3.3.2. Ablation Experiments on Network Modules

To validate the efficacy of our proposed framework, several ablation studies were undertaken. The intent behind these studies was to underscore the importance of the ARescat network. For illustrative purposes, three distinct models, namely "ARescat-1", "ARescat-2", and "ARescat-3", were pitted against our proposed model. Table 3 shows the test results for ARescat, ARescat-1, and ARescat-2

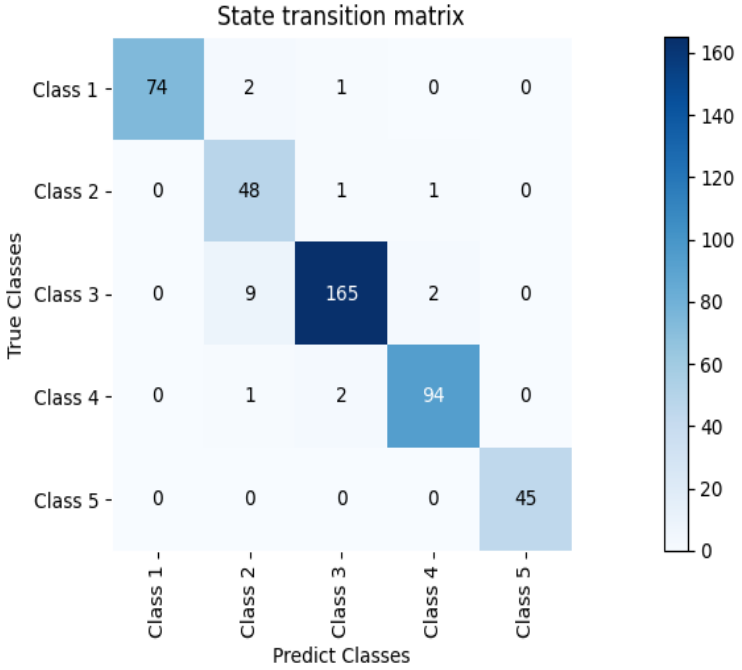

Figure 9. Confusion matrix of the proposed model.

Table 3. Test results for ARescat, ARescat-1, and ARescat-2.

Model	Class	Precision	Recall	F1-Score	Support
ARescat	Class A	0.960	1.000	0.984	74
	Class B	0.958	0.804	0.867	60
	Class C	0.937	0.984	0.959	169
	Class D	0.970	0.966	0.970	97
	Class E	1.000	1.000	1.000	45
	Average	0.958	0.958	0.958	445
ARescat-1	Class A	0.932	1.000	0.960	74
	Class B	0.921	0.754	0.825	60
	Class C	0.916	0.965	0.941	169
	Class D	0.979	0.930	0.954	97
	Class E	1.000	1.000	1.000	45
	Average	0.939	0.939	0.939	445
ARescat-2	Class A	0.753	0.908	0.819	74
	Class B	0.971	0.500	0.655	60
	Class C	0.806	0.952	0.883	169
	Class D	0.948	0.823	0.880	97
	Class E	1.000	0.956	0.981	45
	Average	0.857	0.857	0.857	445

As detailed in Table 3, the distinction between ARescat and ARescat-1 hinges on the absence of concatenation operations in the ARCSB module of ARescat-1, with the latter only executing residual functions on input data. ARescat-2, when juxtaposed against ARescat, omits the ASPP operation, leading to modifications in channel numbers and feature extraction. It is pertinent to note that the test accuracy assessment remains untainted by the model training, thus offering an unbiased performance metric.

From the comparison between ARescat and ARescat-1 in Table 3, the concatenate operation's prowess is evident. This operation amplifies feature extraction capabilities by synthesizing spectral features derived from disparate convolutions, thereby creating a comprehensive feature representation. This enhances the model's performance. Moreover, integrating multi-scale detailed data acts as auxiliary information, enabling a more precise

target feature extraction. When combined with the concatenate operation, a heightened accuracy is achieved. As shown by the comparison of ARescat and ARescat-2 experiments, ASPP is used to acquire multi-scale object information to make the target balanced in resolution and receptive field. The primary target of the ASPP module is to augment the network's receptive field, bolster its ability to grasp multi-scale contexts, refine the model's feature articulation, and adeptly abstract frequency characteristics from the spectrogram. Consequently, integrating the ASPP module and concatenate operation propels the model's performance metrics.

In Table 4, the distinction between ARescat-3 and ARescat rests on the integration (or lack thereof) of the self-attention mechanism within the ARCSB network. The intrinsic value of this mechanism is its capacity to recalibrate the original feature map by discerning its channel dependencies. Empirical evidence suggests that incorporating the self-attention mechanism substantially influences the model's accuracy. Given the acoustic images' high-frequency feature similarity, the feature extraction process is particularly vulnerable to target feature information loss. By integrating the self-attention mechanism, there is an accuracy uptick, and the network learns to prioritize informative features while simultaneously suppressing less pertinent ones. This prevents target feature attrition during the extraction and recognition phases, enhancing the network's feature extraction prowess.

Table 4. Test results for ARescat and ARescat-3.

Model	Class	Precision	Recall	F1-Score	Support
ARescat	Class A	0.960	1.000	0.984	74
	Class B	0.958	0.804	0.867	60
	Class C	0.937	0.984	0.959	169
	Class D	0.970	0.966	0.970	97
	Class E	1.000	1.000	1.000	45
	Average	0.958	0.958	0.958	445
ARescat-3	Class A	0.690	0.840	0.760	74
	Class B	0.781	0.702	0.744	60
	Class C	0.826	0.894	0.861	169
	Class D	0.953	0.754	0.834	97
	Class E	0.989	0.900	0.990	45
	Average	0.834	0.834	0.834	445

The primary distinction between ARescat and ARescat-3 models lies in the incorporation of the SE (Squeeze-Excitation) attention mechanism within the ARCSB network of ARescat. The SE attention mechanism recalibrates the original feature map by capturing channel dependencies, significantly impacting the model's accuracy. The comparative analysis between ARescat and ARescat-3, as presented in Table 4, demonstrates the substantial influence of the SE attention mechanism on model performance. Given the high-frequency feature similarity in acoustic images, there is a risk of losing crucial target feature information during the extraction process. The SE attention mechanism not only enhances accuracy but also strategically utilizes global information to highlight informative features while suppressing less reliable ones, thus mitigating the loss of target features and improving the overall feature extraction capability of the network.

3.3.3. Role of Loss Function in Model Performance

This subsection scrutinizes the repercussions of various loss functions on the model's performance. As illustrated in Table 5, there is a clear correlation between the choice of loss function and classification outcomes. Focal Loss, in particular, outshines its counterparts, boosting the average accuracy by 4.7% and 2.5% compared to Uniform Loss and Cross-entropy Loss, respectively. The data underscore the superior performance metrics achieved by the model incorporating Focal Loss relative to other models. Focal Loss is instrumental in optimizing the model's performance.

Table 5. Comparing various loss functions.

Model	ARescat	ARescat	ARescat
Feature	**3D MFCC**	**3D MFCC**	**3D MFCC**
Loss function	**Uniform Loss**	**Cross-entropy Loss**	**Focal Loss**
Class A	0.896	0.931	0.984
Class B	0.739	0.786	0.867
Class C	0.917	0.933	0.959
Class D	0.973	0.974	0.970
Class E	1.000	1.000	1.000
Average	0.911	0.933	0.958

3.3.4. Characterization Ablation Experiments

MFCC, widely used in speech and audio processing, captures the short-term power spectrum envelope of a signal through a series of transformations, including Fourier transform, Mel filter bank processing, logarithmic transformation, and discrete cosine transform (DCT). The 3D dynamic MFCC extends this by incorporating the first- and second-order derivatives (delta and delta-delta features), representing the rate of change and acceleration over time, thus providing a deeper understanding of audio signal dynamics.

Similarly, the Mel-spectrogram offers a visual representation of sound signal variations in time and frequency. The 3D dynamic Mel-spectrogram extends this by adding dynamic changes in the time dimension, resulting in a richer and more contextual representation of sound signals.

Both 3D dynamic MFCC and 3D dynamic Mel-spectrogram transform audio signals into formats more suitable for machine processing, offering a comprehensive representation by including dynamic changes over time.

A horizontal comparison was initiated between stationary Mel-spectrogram features, MFCC attributes, and the corresponding 3D dynamic elements. As depicted in Table 6, a comparative analysis of ARescat model accuracy with diverse components was conducted, maintaining a consistent network structure. Empirical results suggest that the quartet of Mel-filtered time-frequency attributes adeptly mirrors the innate properties of target signals, facilitating precise target differentiation. The zenith of classification accuracy is realized through the amalgamation of the ARescat model and 3D dynamic MFCC features. The MFCC feature set, mirroring human auditory traits, delivers commendable classification outcomes. Given the multifaceted marine milieu, target signals often manifest non-uniform radiated noise. Both the first-order difference MFCC and second-order difference MFCC features adeptly capture the transient attributes inherent in the complex marine environment. These features extract correlated attributes from adjacent MFCC time frames. In summation, such features are pivotal in enhancing target classification accuracy.

Table 6. Accuracy of the ARescat model under different characteristics.

Feature	Accuracy
MFCC	0.944
3-D dynamic MFCC	0.958
Mel-spectrogram	0.921
3-D dynamic Mel-spectrogram	0.906

3.3.5. Comparison between Different Models

Audio signal feature extraction converts original audio signals into numerical values that describe their characteristics, such as frequency spectrum and energy distribution. Different extraction methods can significantly impact the accuracy of various models. Therefore, it is essential to compare models using the same feature extraction method.

The data in Table 7 indicate that for hydroacoustic target recognition, more complex ResNet models tend to yield lower accuracy. Our proposed ARescat model surpasses Resnet18, Resnet34, Resnet50, and CRNN by accuracy of 2.8%, 5.4%, 16.9%, and 8.8%, respectively. This comparison underscores the superior accuracy of the ARescat model among the evaluated models.

Table 7. Test accuracy of multiple network models with the same features.

Model	ARescat	Resnet18	Resnet34	Resnet50	CRNN
Feature	3D MFCC	3D MFCC	3D MFCC	3D MFCC	3D MFCC
Class A	0.984	0.939	0.878	0.709	0.800
Class B	0.867	0.837	0.752	0.522	0.681
Class C	0.959	0.935	0.911	0.864	0.923
Class D	0.970	0.931	0.939	0.816	0.886
Class E	1.000	1.000	1.000	0.859	1.000
Average	0.958	0.930	0.904	0.789	0.870

3.3.6. Accuracy Comparison of Various Models

In comparing accuracy across various models, Table 8 provides a direct juxtaposition of our proposed model with other prevailing ones, with the ShipsEar dataset as the benchmark. The ShipsEar study, considered the foundational model for autonomous underwater acoustic target recognition, boasts an accuracy of 0.754. Predominantly, the academic literature resorts to accuracy as the metric of choice for comparative analysis across models.

Table 8. Comparison between the proposed classifier using the Shipsear dataset and other existing literature models.

Num	Methods	Accuracy
1	Baseline ShipsEar [33]	0.754
2	ResNet18 + 3D [27]	0.948
3	CRNN-9 data_aug [10]	0.9406
4	Full-feature vector + DRW-AE [34]	0.9449
5	Cepstrum + average cepstrum + DCRA [35]	0.9533
6	Deep cepstrum-wavelet autoencoder [36]	0.948
7	ResNet18 + FBank [37]	0.943
8	Proposed method	0.958

Table 8 shows a comparison of the recognition accuracies of ship recognition models in recent years, with all data except the baseline having accuracies above 94%. It is crucial to acknowledge that while our methodology employs the standard reference dataset, the initial trio of models—namely ResNet 18 + 3D, CRNN-9 data_aug, and ResNet 18 + FBank—capitalizes on data augmentation to bolster the sample size. This data augmentation technique mitigates the risk of overfitting and augments model generalization. Consequently, the model's performance is enhanced by data augmentation. In scenarios devoid of data augmentation, our proposed model emerges as the frontrunner with an impressive accuracy of 95.8%, outperforming all the other models. While this comparison might lack stringency due to potential dataset segmentation discrepancies across models, purely from an accuracy standpoint, our ARescat model emerges as the most productive.

4. Conclusions

This research delves into the potential of an attention-centric residual concatenate network in underwater acoustic recognition. The hallmark of this model is its ability to distill high-order abstract data, streamlining the classification process. The proposed attention-based residual concatenate network focuses on refining target feature extraction paradigms. Following extensive experimentation and analysis, several salient conclusions emerge:

1. The Rescat component within the ARCSB framework leverages a unique residual concatenate operation, widening the model's horizon to assimilate more intricate details. This is achieved by broadening the receptive field and by amalgamating the spectrogram features derived from multiple convolutions, thereby crafting a sophisticated feature representation elevating the model's expressive prowess.

2. The ARCSB framework, by synergizing with the SE attention mechanism module, accentuates the differentiation between various channel features while ensuring comprehensive extraction of diverse sample attributes. By directing its focus on target components and mitigating multi-source interferences, it emphasizes pivotal data, bolstering the network's feature extraction capabilities.

3. The Focal Loss function is harnessed by the model to address challenges stemming from imbalanced feature supervision. This encompasses interference factors such as marine ambient noise. By ensuring unwavering attention to all facets during recognition, especially merged elements, the loss function adeptly tackles the uneven sample distribution challenge, subsequently optimizing the model's output.

4. Despite utilizing a smaller parameter set, this model outshines its counterparts in terms of classification accuracy. Experimental validation using the ShipsEar dataset vouches for efficacy, indicating a stellar average recognition accuracy of 95.8%.

In essence, this model holds promise for enhancing sonar systems' target identification and recognition proficiencies. Nonetheless, constraints surrounding dataset accessibility, primarily arising from confidentiality concerns, merit acknowledgment. Future explorations can delve deeper into the potential of this approach, especially its feature augmentation capabilities and target classification efficacy under varied signal-to-noise ratios and scales, to seamlessly adapt to the ever-evolving marine milieu.

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

Funding: This work is supported by the 2021 Open Fund project of the Key Laboratory of Cognitive Radio and Information Processing of the Ministry of Education, the Special Program of Guangxi Science and Technology Base and Talent under Grant No. AD21220098, and the Guangxi Natural Science Foundation, grant number 2022GXNSFDA035070 and the Innovation Project of Guangxi Graduate Education (YCSW2023329).

Institutional Review Board 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 conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ARescat	Attention Mechanism Residual Concatenate Network
SE	Squeeze-Excitation
ResNet	Residual Network
MFCC	Mel-frequency Cepstrum Coefficient
CQT	Constant-Q Transform
DEMON	Detection of Envelope Modulation On Noise
LOFAR	Low-frequency Analyzer and Recorder
3D MFCC	3D dynamic Mel-frequency Cepstrum Coefficient
MSRDN	Multiscale Residual Deep Neural Network
ARCSB	Attention Mechanism Residual Concatenate Specify Dimensions Block
ASPP	Atrous Spatial Pyramid Pooling
MLP	Multilayer Perceptron

FC	Fully Connected
Rescat	Residual Concatenate
CRNN	Convolutional Recurrent Neural Network

References

1. Kamal, S.; Mohammed, S.K.; Pillai, P.S.; Supriya, M. Deep learning architectures for underwater target recognition. In Proceedings of the 2013 Ocean Electronics (SYMPOL), IEEE, Kochi, India, 23–25 October 2013; pp. 48–54. [CrossRef]
2. Cao, X.; Zhang, X.; Yu, Y.; Niu, L. Deep learning-based recognition of underwater target. In Proceedings of the 2016 IEEE International Conference on Digital Signal Processing (DSP), IEEE, Beijing, China, 16–18 October 2016; pp. 89–93. [CrossRef]
3. Li, C.; Liu, Z.; Ren, J.; Wang, W.; Xu, J. A feature optimization approach based on inter-class and intra-class distance for ship type classification. *Sensors* **2020**, *20*, 5429. [CrossRef] [PubMed]
4. Yang, H.; Shen, S.; Yao, X.; Sheng, M.; Wang, C. Competitive deep-belief networks for underwater acoustic target recognition. *Sensors* **2018**, *18*, 952. [CrossRef] [PubMed]
5. Irfan, M.; Jiangbin, Z.; 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]
6. Permana, S.D.H.; Bintoro, K.B.Y. Implementation of Constant-Q Transform (CQT) and Mel Spectrogram to converting Bird's Sound. In Proceedings of the 2021 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT). IEEE, Purwokerto, Indonesia, 17–18 July 2021; pp. 52–56. [CrossRef]
7. Wei, X.; Gang-Hu, L.; Wang, Z. Underwater target recognition based on wavelet packet and principal component analysis. *Comput. Simul.* **2011**, *28*, 8–290.
8. Azimi-Sadjadi, M.R.; Yao, D.; Huang, Q.; Dobeck, G.J. Underwater target classification using wavelet packets and neural networks. *IEEE Trans. Neural Netw.* **2000**, *11*, 784–794. [CrossRef] [PubMed]
9. Chen, Y.; Xu, X. The research of underwater target recognition method based on deep learning. In Proceedings of the 2017 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). IEEE, Xiamen, China, 22–25 October 2017; pp. 1–5. [CrossRef]
10. Liu, F.; Shen, T.; Luo, Z.; Zhao, D.; Guo, S. Underwater target recognition using convolutional recurrent neural networks with 3-D Mel-spectrogram and data augmentation. *Appl. Acoust.* **2021**, *178*, 107989. [CrossRef]
11. Luo, X.; Zhang, M.; Liu, T.; Huang, M.; Xu, X. An underwater acoustic target recognition method based on spectrograms with different resolutions. *J. Mar. Sci. Eng.* **2021**, *9*, 1246. [CrossRef]
12. 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* **2021**, *24*, 22. [CrossRef]
13. Zhang, L.; Wu, D.; Han, X.; Zhu, Z. Feature extraction of underwater target signal using mel frequency cepstrum coefficients based on acoustic vector sensor. *J. Sens.* **2016**, *2016*, 7864213. [CrossRef]
14. 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]
15. Gou, J.; Yu, B.; Maybank, S.J.; Tao, D. Knowledge distillation: A survey. *Int. J. Comput. Vis.* **2021**, *129*, 1789–1819. [CrossRef]
16. Le Gall, Y.; Bonnel, J. Separation of moving ship striation patterns using physics-based filtering. In Proceedings of the Meetings on Acoustics, Montreal, QC, Canada, 2–7 June 2013; AIP Publishing: Melville, NY, USA, 2013; Volume 19. [CrossRef]
17. Kuz'kin, V.; Kuznetsov, G.; Pereselkov, S.; Grigor'ev, V. Resolving power of the interferometric method of source localization. *Phys. Wave Phenom.* **2018**, *26*, 150–159. [CrossRef]
18. Ehrhardt, M.; Pereselkov, S.; Kuz'kin, V.; Kaznacheev, I.; Rybyanets, P. Experimental observation and theoretical analysis of the low-frequency source interferogram and hologram in shallow water. *J. Sound Vib.* **2023**, *544*, 117388. [CrossRef]
19. Pereselkov, S.A.; Kuz'kin, V.M. Interferometric processing of hydroacoustic signals for the purpose of source localization. *J. Acoust. Soc. Am.* **2022**, *151*, 666–676. [CrossRef] [PubMed]
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778. [CrossRef]
21. Huang, G.; Sun, Y.; Liu, Z.; Sedra, D.; Weinberger, K.Q. Deep networks with stochastic depth. In *Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 646–661. [CrossRef]
22. Tian, S.; Chen, D.; Wang, H.; Liu, J. Deep convolution stack for waveform in underwater acoustic target recognition. *Sci. Rep.* **2021**, *11*, 9614. [CrossRef] [PubMed]
23. Xue, L.; Zeng, X.; Jin, A. A novel deep-learning method with channel attention mechanism for underwater target recognition. *Sensors* **2022**, *22*, 5492. [CrossRef] [PubMed]
24. Zhufeng, L.; Xiaofang, L.; Na, W.; Qingyang, Z. Present status and challenges of underwater acoustic target recognition technology: A review. *Front. Phys.* **2022**, *10*, 1044890. [CrossRef]
25. Chen, S.; Tan, X.; Wang, B.; Lu, H.; Hu, X.; Fu, Y. Reverse attention-based residual network for salient object detection. *IEEE Trans. Image Process.* **2020**, *29*, 3763–3776. [CrossRef]
26. Lu, Z.; Xu, B.; Sun, L.; Zhan, T.; Tang, S. 3-D channel and spatial attention based multiscale spatial–spectral residual network for hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4311–4324. [CrossRef]

27. 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]
28. Jiang, K.; Wang, Z.; Yi, P.; Chen, C.; Huang, B.; Luo, Y.; Ma, J.; Jiang, J. Multi-scale progressive fusion network for single image deraining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8346–8355. [CrossRef]
29. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19. [CrossRef]
30. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
31. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988. [CrossRef]
32. Chen, X.; Liang, C.; Huang, D.; Real, E.; Wang, K.; Liu, Y.; Pham, H.; Dong, X.; Luong, T.; Hsieh, C.J.; et al. Symbolic discovery of optimization algorithms. *arXiv* **2023**, arXiv:2302.06675. [CrossRef]
33. 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]
34. Khishe, M. Drw-ae: A deep recurrent-wavelet autoencoder for underwater target recognition. *IEEE J. Ocean. Eng.* **2022**, *47*, 1083–1098. [CrossRef]
35. Kamalipour, M.; Agahi, H.; Khishe, M.; Mahmoodzadeh, A. Passive ship detection and classification using hybrid cepstrums and deep compound autoencoders. *Neural Comput. Appl.* **2023**, *35*, 7833–7851. [CrossRef]
36. Jia, H.; Khishe, M.; Mohammadi, M.; Rashidi, S. Deep cepstrum-wavelet autoencoder: A novel intelligent sonar classifier. *Expert Syst. Appl.* **2022**, *202*, 117295. [CrossRef]
37. Wu, J.; Li, P.; Wang, Y.; Lan, Q.; Xiao, W.; Wang, Z. VFR: The Underwater Acoustic Target Recognition Using Cross-Domain Pre-Training with FBank Fusion Features. *J. Mar. Sci. Eng.* **2023**, *11*, 263. [CrossRef]