remote sensing

MDPI

Article

Dual-Feature Fusion Learning: An Acoustic Signal Recognition Method for Marine Mammals

Zhichao Lü [1], Yaqian Shi [1], Liangang Lü [2], Dongyue Han [1], Zhengkai Wang [1] and Fei Yu [1,*]

[1] College of Ocean Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China; lvzhichao@hrbeu.edu.cn (Z.L.); shiyaqian@sdust.edu.cn (Y.S.); handongyue@sdust.edu.cn (D.H.); wzk18724791626@sdust.edu.cn (Z.W.)

[2] First Institute of Oceanography, Ministry of Natural Resources, Qingdao 266061, China; lvlg@fio.org.cn

* Correspondence: feiyu@sdust.edu.cn

Abstract: Marine mammal acoustic signal recognition is a key technology for species conservation and ecological environment monitoring. Aiming at the complex and changing marine environment, and because the traditional recognition method based on a single feature input has the problems of poor environmental adaptability and low recognition accuracy, this paper proposes a dual-feature fusion learning method. First, dual-domain feature extraction is performed on marine mammal acoustic signals to overcome the limitations of single feature input methods by interacting feature information between the time-frequency domain and the Delay-Doppler domain. Second, this paper constructs a dual-feature fusion learning target recognition model, which improves the generalization ability and robustness of mammal acoustic signal recognition in complex marine environments. Finally, the feasibility and effectiveness of the dual-feature fusion learning target recognition model are verified in this study by using the acoustic datasets of three marine mammals, namely, the Fraser's Dolphin, the Spinner Dolphin, and the Long-Finned Pilot Whale. The dual-feature fusion learning target recognition model improved the accuracy of the training set by 3% to 6% and 20% to 23%, and the accuracy of the test set by 1% to 3% and 25% to 38%, respectively, compared to the model that used the time-frequency domain features and the Delay-Doppler domain features alone for recognition.

Keywords: marine mammals; acoustic signals; feature fusion; target recognition; neural networks

Citation: Lü, Z.; Shi, Y.; Lü, L.; Han, D.; Wang, Z.; Yu, F. Dual-Feature Fusion Learning: An Acoustic Signal Recognition Method for Marine Mammals. *Remote Sens.* **2024**, *16*, 3823. https://doi.org/10.3390/rs16203823

Academic Editor: Marc Pinto

Received: 8 September 2024
Revised: 11 October 2024
Accepted: 11 October 2024
Published: 14 October 2024

1. Introduction

The ability of marine mammals to use sound for communication is one of the key features of their adaptation to the underwater living environment, and their acoustic signals exhibit remarkable diversity. American scholar Lilly categorized these acoustic signals into three main types based on their functions and parameters, Click, Whistle, and Burst Pulse, with each sound type serving distinct purposes [1]. Studying the acoustic signals of marine mammals holds great significance for understanding their biological behavior, rational utilization of marine resources, and conservation efforts to protect these species. Passive Acoustic Monitoring (PAM) of marine mammals is a widely employed biomonitoring method. Since acoustic waves propagate with lower energy attenuation in the ocean than light waves, they can travel longer distances, making them particularly suitable for monitoring and identifying marine mammals [2]. In the pursuit of developing PAM systems for detecting and classifying marine mammals, numerous algorithms have been devised specifically to analyze the acoustic characteristics of these animals. Notably, most of these detection and classification algorithms rely on the distinct acoustic features exhibited by marine mammals [3].

Feature extraction is a crucial step in acoustic signal processing. By employing feature extraction techniques, one can obtain the acoustic features of the target, which subsequently serve as a reliable basis for target identification. Ibrahim et al. demonstrated the effectiveness of using Mel-Frequency Cepstral Coefficients (MFCC) and Discrete Wavelet

Transformation Coefficients in classifying marine mammal calls using Support Vector Machines (SVM) [4]. Brown et al. pioneered the use of the Gaussian Mixture Model combined with the Hidden Markov Model (GMM-HMM) for recognizing MFCC features [5]. In their study, they achieved a classification consistency of over 90% for a set of 75 killer whale calls. Dugan, on the other hand, extracted time-frequency features from cetacean calls and classified them using three different models, achieving the highest assignment rate of 86.45% [6]. Maheen et al. acoustically classified six marine mammal species by fusing one-dimensional localized binary patterns with MFCC features. Their approach yielded a training test accuracy of 90.4% [7]. Meanwhile, Zhong Mingtuo et al. fused MFCC, Linear Cepstral Coefficients, and time-domain features from 61 species of marine mammals as feature parameters [8]. By utilizing SVM for classification, they managed to improve the recognition rate by 5.5% compared to traditional MFCC-based methods. Li Songbin et al. extracted six features—MFCC, FBanks, PNCC, PSRCC, GFCC, and MSRCC—from three marine mammal species for comparative analysis [9]. They employed a Convolutional Neural Network-Gated Recurrent Unit (CNN-GRU) structure for recognition and achieved a classification accuracy of 74%. Some of these studies use one feature to identify marine mammal acoustic signals, and some use multiple, but none of the features they use can be separated from the time-frequency domain. Since climatic factors, such as rainfall and typhoons, have a great impact on the marine acoustic environment, the recognition of marine mammal acoustic signals only by time-domain features or frequency-domain features has certain limitations. Zhang Xuebo et al. coherently synthesized signals in the range-Doppler domain associated with each receiver after performing range cell migration correction (RCMC) for each receiver, and then corrected the azimuth offset [10,11]. To improve the simulation efficiency of multi-receiver synthetic aperture sonar (SAS) echo signal, Zhang Xuebo et al. multiplied the spectrum of the transmitted signal with the phase shift related to the delay so that the spectrum of the echo signal can be accurately obtained [12,13]. Compared with the traditional echo simulation algorithm, this method significantly improves the simulation efficiency of the echo signal without losing performance. Although the MFCC features have achieved good recognition results in the current field, the Delay-Doppler (DD) domain features can respond to the speed information of the target and are not easily affected by environmental factors such as climate, so combining them with the MFCC features can increase the reliability of the recognition, and the use of the dual-feature fusion learning method may be beneficial to improve the accuracy of the recognition.

Convolutional Neural Networks (CNNs) are widely utilized in the field of speech recognition. Traditionally, acoustic signal recognition involves time-frequency analysis of the acoustic signal to generate a spectrogram, which is then used to identify the acoustic signal based on its unique patterns. Detailed information on machine learning and deep learning-based methods and underwater sound sources, features, classifiers, datasets, related techniques, challenges, and future trends for marine ship sound classification and fish sound classification are discussed by Aslam et al. [14]. Bianco et al. presented the development of machine learning in four acoustic research areas: source localization in speech processing, source localization in marine acoustics, bioacoustics, and environmental sounds in everyday scenes [15]. In the process of feature extraction for speech signals, the original acoustic features are often replaced with acoustic feature images, and CNN-based image recognition techniques are employed to recognize the acoustic signals. This approach has proven to achieve accuracy rates that are difficult to match using traditional methods, especially when dealing with large-sample datasets. Zhang Xuebo et al. focus on the application of a nonlinear chirp scaling algorithm in SAS and validate the proposed method through simulation and real data. The processing results show that the imaging efficiency is greatly improved compared with the phase center approximation (PCA) method [16]. Wang et al. proposed a comprehensive underwater image enhancement framework, the metalantis framework, which enhances state-of-the-art physical models of underwater imaging by utilizing virtually generated data for reinforcement learning [17,18]. Meanwhile,

they gave two examples in [19,20]. Shiu et al. explored the use of deep CNNs and Recurrent Neural Networks (RNNs) with spectrograms to detect vocalizations of North Atlantic Right Whales [21]. Their findings indicated that deep learning architectures can produce false positive rates several orders of magnitude lower than other algorithms. Griffiths et al. proposed a multivariate clustering method to identify distinct Click vocal clusters of Dall's Porpoise in the U.S.A., and the validity of the three clusters was verified using the Random Forest method [22]. Cai et al. designed a multichannel-based classification model with a parallel structure, fusing predictions and introducing data enhancement techniques to further improve classification accuracy [23]. Duan Dexin et al. trained a random forest classifier using time-frequency graph features to detect and distinguish echolocation signals, achieving higher recall and accuracy under low Signal-to-Noise Ratio (SNR) conditions [24]. Cominelli et al. combined pre-trained acoustic classification models (VGGish, NOAA, and Google Humpback Whale Detector), dimensionality reduction (UMAP), and balanced random forest algorithms to demonstrate how machine-learned acoustic features can capture different aspects of the marine acoustic environment [25]. The current state of research suggests that applying deep learning to marine mammal acoustic recognition has become a trend, but most of the methods proposed so far are based on a feature-based recognition model. Although neural network models for marine mammal acoustic recognition offer higher accuracy and reduce time and labor costs, traditional neural network models often suffer from complexity, computational costs, and other problems.

To address the limitations of traditional recognition methods that rely solely on a single feature input, this paper proposes a dual-feature fusion learning target recognition model. To validate the effectiveness of the proposed model for marine mammal acoustic recognition, this study combines three common CNN recognition models with two signal features, respectively, for single-feature recognition, which is compared with two-feature recognition. The innovations of the dual-feature fusion learning method presented in this paper are as follows:

The marine mammal acoustic signal is preprocessed using adaptive filtering to enhance the SNR and mitigate the interference of environmental noise.

1. Delay-Doppler domain features are introduced into the acoustic feature recognition of marine mammals, effectively addressing the impact of seasonal changes in the marine environment on marine mammal acoustic signals.
2. A dual-feature fusion learning target recognition model is developed, capable of recognizing both MFCC features and Delay-Doppler domain features simultaneously. This model exhibits high recognition accuracy and strong generalization ability for mammal acoustic signal recognition in complex marine environments.

2. Theory

2.1. Least Mean Square Adaptive Filter

To mitigate the effects of environmental noise on marine mammal identification, this study employs the Least Mean Square (LMS) adaptive filter to process the acoustic signals of marine mammals, thereby enhancing the SNR [26]. LMS is an adaptive filter based on the minimum mean square error criterion. The core idea is to minimize the sum of squares of the errors between the output signal of the filter and the desired signal by continuously adjusting the coefficients of the filter. The LMS algorithm is based on the Wiener filtering theory, and adopts the algorithm of estimating the gradient vector by instantaneous values to update the adaptive filter weights coefficients by minimizing the energy of the error signal. The structure of the adaptive filter is shown in Figure 1, the input signal $x(n)$ is passed through a parameter-adjustable digital filter to produce an output signal $y(n)$, which is compared with the desired signal $d(n)$ to form an error signal $e(n)$, and the filter-parameters are adjusted by an adaptive algorithm to minimize the mean square value of $e(n)$ [27].

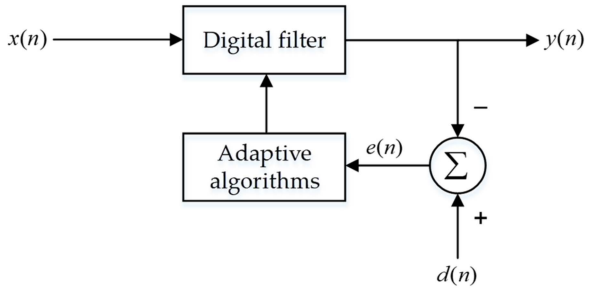


Figure 1. The structure of the LMS adaptive filter.

Let the input signal be a vector $x(n)$ and the output of the LMS adaptive filter be $y(n)$. Then we have [28]:

$$y(n) = \sum_{i=0}^{N} s_0 x(n-i) = S^T(n)X(n) \tag{1}$$

$$e(n) = d(n) - y(n) = d(n) - S^T(n)X(n) \tag{2}$$

where $S(n)$ is the weight coefficient of the filter and N is the order of the filter.

Adaptive filtering can utilize the results of the filter parameters that have been obtained at the previous moment to automatically adjust the filter parameters at the current moment to adapt to the unknown or time-varying statistical characteristics of the signal and noise, thus achieving optimal filtering to improve the SNR.

2.2. Mel-Frequency Cepstral Analysis

The Mel-Frequency Cepstral Coefficient is a cepstral parameter derived from the Mel scale frequency domain. By employing equally spaced band divisions on the Mel scale, MFCC more closely mimics the human auditory system compared to linearly spaced bands used in traditional logarithmic cepstral analysis. This results in superior recognition performance, particularly when the SNR is low [29]. Figure 2 illustrates the steps involved in extracting MFCC features from a signal.

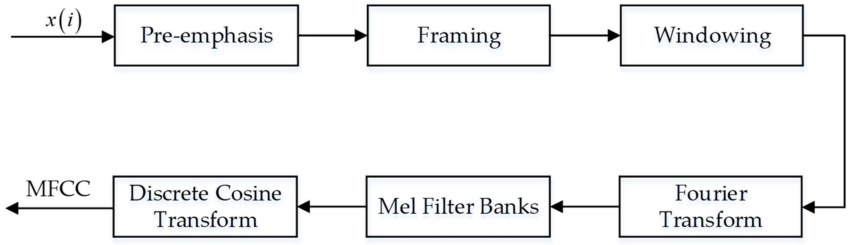


Figure 2. The procedure for extracting MFCC features.

(1) Preprocessing: Pre-emphasis, framing, windowing. The purpose of pre-emphasis is to boost the high-frequency part so that the gap between the peaks of the spectrum of the whole signal is reduced by passing the audio signal through a high-pass filter. Due to the non-smooth and short-time smooth characteristics of the speech signal, the speech signal is divided into frames, N samples are gathered into one frame, and there should be a section of overlapping area between two neighboring frames to avoid too large variations between two neighboring frames. To increase the continuity of the left and right ends of the frame, it is necessary to multiply the audio signal of each frame by a window function, and the Hamming window matrix C is multiplied by the post-split-frame matrix S to get the post-windowed matrix SC.

(2) Fast Fourier Transform (FFT) and calculation of energy spectrum. The spectrum of each frame is obtained by performing FFT on each frame of the signal after framing and windowing [30]:

$$X(i) = \sum_{n=o}^{N-1} x(n) e^{-j\frac{2\pi i}{n}}, 0 \le i \le N \tag{3}$$

where $x(n)$ is the input signal and N represents the number of points of the Fourier transform.

(3) Mel filter. The filter bank consists of 32 triangular filters, with dense filters and high thresholds at low frequencies and sparse filters and low thresholds at high frequencies. The conversion relationship between frequency and the Mel scale is [31]:

$$Mel(f) = 2595 lg\left(1 + \frac{f}{700}\right) \tag{4}$$

where f is the frequency in Hz.

(4) Discrete Cosine Transform (DCT). The filter bank coefficients are highly correlated, which may be problematic in machine learning algorithms; as such, the DCT is applied to de-correlate the filter bank coefficients to finally obtain the MFCC features [31]:

$$C(n) = \sum_{k=1}^{p} log X(k) cos\left(\frac{\pi(k-0.5)n}{p}\right), n = 1, 2, \ldots, L \tag{5}$$

where L is the MFCC coefficient order and p is the number of triangular filters.

2.3. Delay-Doppler Domain Analysis

The Delay-Doppler domain combines the properties of a signal in the time and frequency domains, mapping the signal into a two-dimensional time-frequency lattice by placing lattice points in both the time-delay dimension and the Doppler dimension. This representation allows the signal to maintain relatively stable transmission characteristics in a complex Doppler frequency shift environment. The relative motion between the target and the signal-receiving equipment induces a Doppler effect on the signal. The Delay-Doppler domain analysis of the signal can reveal the frequency shift caused by this effect. Since different marine mammals exhibit distinct average speeds, they generate unique Doppler shifts, allowing the Delay-Doppler domain features of the signal to reflect the movement characteristics of these animals.

There exists a defined relationship between the time-frequency domain and the Delay-Doppler domain of the signal. Specifically, the time-domain signal undergoes FFT to convert it into the frequency domain. Subsequently, two-dimensional sampling with specific periods and frequency intervals is performed to obtain the time-frequency domain features. Finally, the time-frequency domain signal is subjected to the Symplectic Finite Fourier Transform (SFFT) to derive Delay-Doppler domain features. The expression for SFFT is provided in [32]:

$$Y[p, q] = \frac{1}{\sqrt{PQ}} \sum_{n=0}^{Q-1} \sum_{m}^{P-1} X[l, k] e^{-j2\pi(\frac{qk}{Q} - \frac{pl}{P})} \tag{6}$$

where X [l, k] is the signal after Fourier transform and sampling of the time domain signal, M and N denote the number of time delay dimensions and Doppler dimensions, respectively, and $0 \le p \le P$, $0 \le q \le Q$.

Inverse Symplectic Finite Fourier Transform (ISFFT) is the inverse transformation of SFFT, which is able to simultaneously transform the time-delay domain to the frequency domain and the Doppler domain to the time domain, so as to complete the transformation of the received signal from the Delay-Doppler domain to the time-frequency domain. The

transform relationships in the time-frequency domain and Delay-Doppler domain are shown in Figure 3.

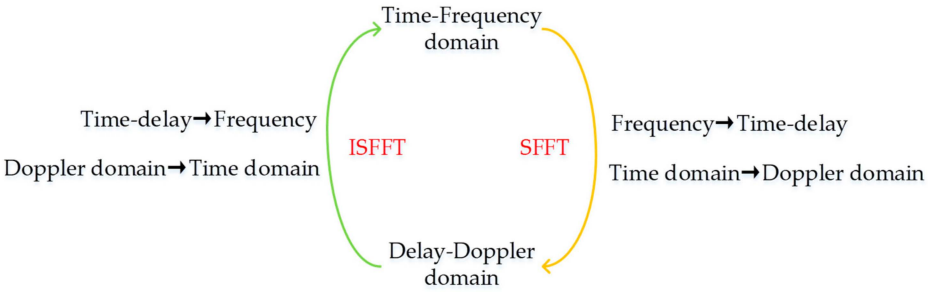


Figure 3. Transformation relations in the time-frequency and Delay-Doppler domains. The green arrow indicates Inverse Symplectic Finite Fourier Transform (ISFFT), the yellow arrow indicates Symplectic Finite Fourier Transform (SFFT). The red font indicates the two transformations.

2.4. Convolutional Neural Network Model

CNN is a deep learning model commonly used in fields such as image and audio recognition The core idea is to extract the features of the data such as images through convolutional operations to achieve the purpose of classification and recognition of the data. Figure 4 illustrates the typical architecture of a CNN, which typically encompasses convolutional layers, pooling layers, and fully connected layers.

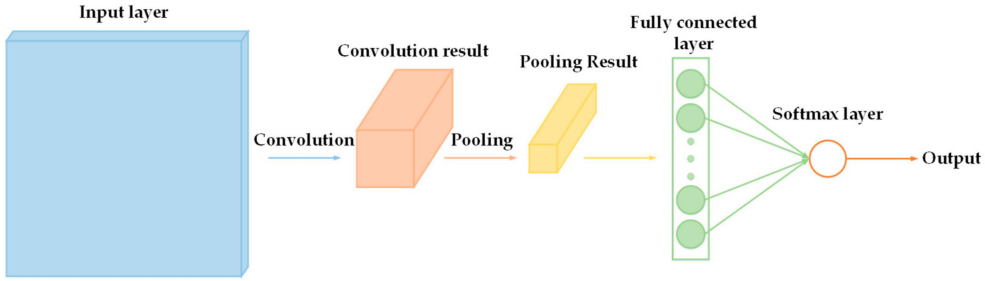


Figure 4. The architecture of CNN.

The convolutional layer consists of multiple convolutional kernels and the complete feature map is obtained by using different convolutional kernels. The eigenvalue of position (i, j) in the kth feature map of the lth layer is [33]:

$$v_{p,q,k}^{l} = W_{k}^{l^T} i_{p,q}^{l} + B_{k}^{l} \tag{7}$$

where W_k^{lT} is the weight vector of the kth filter of the lth layer, B_k^l is the bias term of the kth filter of the lth layer, and $i_{p,q}^l$ is the input segment centered at position (i, j).

As hardware technology and computational capabilities continue to advance at a rapid pace, so too do the design and application of convolutional neural network models. These models are tailored to specific tasks and application scenarios to optimize performance and efficiency, with notable examples including VGG16, GoogleNet, ResNet, and others.

The VGG16 model consists of 13 convolutional layers and 3 fully connected layers; the convolutional part uses a smaller 3×3 convolutional kernel and a convolutional operation with a step size of 1 [34]. This design approach allows the network to be deeper. Between every two convolutional layers, VGG16 also uses a 2×2 maximum pooling layer to reduce the size of the feature map and retain the most salient features. After the final convolutional layer, VGG16 uses 3 fully connected layers, each with 4096 hidden units, and the last fully connected layer outputs the predictions of the model.

The GoogleNet model has about 22 layers, including convolutional layers, pooling layers, fully connected layers, and the Inception module [35]. Its most important feature

is the use of the Inception module, which allows the simultaneous use of multiple convolutional kernels of different sizes and pooling layers to extract features, thus increasing the expressive power and accuracy of the network. GoogleNet also uses global average pooling instead of maximal pooling, which averages the entire feature map in an average operation to obtain a feature vector as the final output, reducing the model's number of parameters and preventing overfitting.

The ResNet model is proposed to solve the network degradation problem when there are too many hidden layers in a deep neural network [36]. While traditional neural networks try to learn a function of the input and target mapping, ResNet learns the residuals between the input and target. The residuals are obtained by comparing the input signal with the desired output signal and then learning the residuals; this approach helps to solve the problems such as gradient vanishing. The ResNet model can learn a deep network with 152 layers and can obtain higher accuracy than the VGG model and GoogleNet model.

3. Method

First, this research performs adaptive filtering on acoustic signals of marine mammals to enhance their SNR. Subsequently, it extracts the MFCC and DD domain features of marine mammals as the two input features of this recognition method. This study constructs a dual-feature fusion learning target recognition model, which can be trained by inputting two marine mammal acoustic signal features at the same time and can improve the target recognition accuracy. The overall idea of the paper is shown in Figure 5.

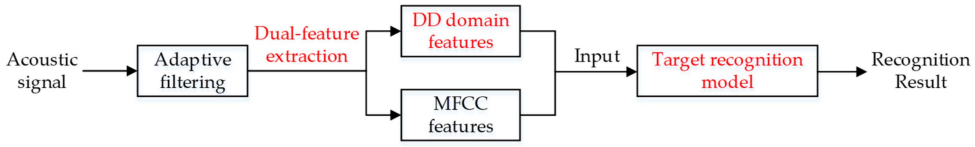


Figure 5. Overview of the experimental procedure. The red font indicates the innovations of this paper: dual-feature extraction, DD domain features and target recognition model.

3.1. Framing and Normalization

Framing: The purpose of framing is to extract a series of shorter, discrete time segments (frames) from a continuous signal so that each frame can be further analyzed and processed. A frame is a fixed-length sequence of samples extracted from an audio signal. The length of a frame is usually a power of 2, such as 256, 512, or 1024, because such a length makes subsequent Fast Fourier Transform (FFT) calculations more efficient. The frameshift is the number of samples between two consecutive frames. The frameshift determines the degree of overlap between frames. Smaller frameshifts provide higher temporal resolution, but increase computational effort; larger frameshifts reduce computational effort, but decrease temporal resolution.

For example, after inputting a segment of the signal, set the number of samples per frame to 1024 and the frameshift to 512, i.e., move 512 samples at a time to get a new frame, which ensures 50% overlap, and draw the data of the first three frames as shown in Figure 6.

Normalization: Min-Max Normalization is a method of scaling data features to a specific range (usually between 0 and 1). This method is implemented through the maximum and minimum values of each feature, using a linear transformation to map the data to the new range. The formula for this is [37]:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{8}$$

where x is the original data point, $min(x)$ is the minimum value in the data set, $max(x)$ is the maximum value in the data set, and x' is the normalized data point.

The advantage of Min-Max Normalization is that it is sensitive to outliers, since changes in the maximum and minimum values directly affect the normalized result, and the data can be easily scaled to any specified range.

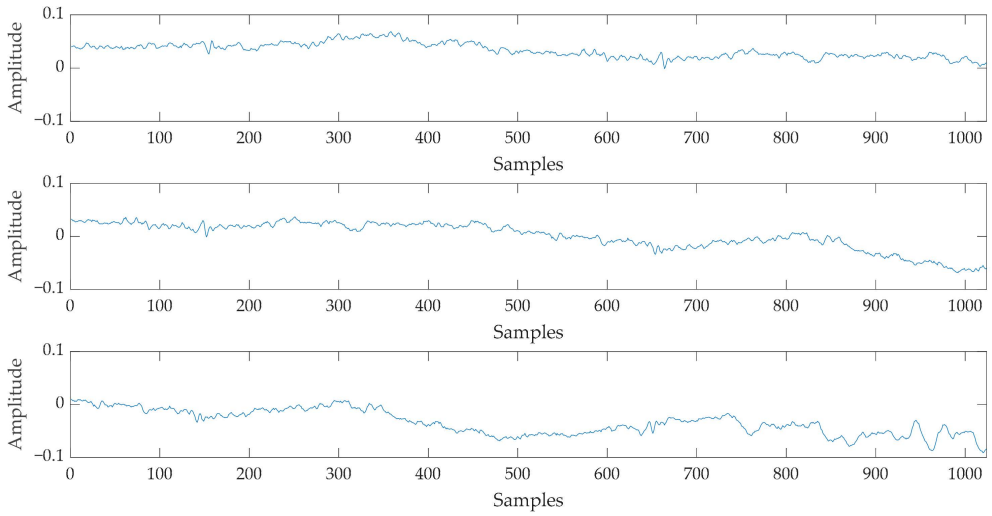


Figure 6. Time-domain plot of three frames.

3.2. Dual-Feature Extraction Analysis

Due to its excellent nonlinear perception ability, MFCC has been widely applied in acoustic signal recognition research. However, this feature remains a traditional time-frequency domain feature, susceptible to environmental changes. In this study, we extract the DD domain feature of marine mammals, which complements the MFCC feature and can reflect the motion characteristics of marine mammals. This approach addresses the limitations of recognizing marine mammals using a single feature.

Figure 7 depicts the MFCC features of the Fraser's Dolphin acoustic signal. The horizontal axis represents the frame rate, where each frame encompasses a specific number of samples with a certain overlap between neighboring frames. The more frames there are, the higher the temporal resolution, which can more accurately capture the dynamic properties of the signal and changes in short-term characteristics. This is because more frames mean that the signal is more finely segmented, which better reflects changes in the signal over time. The vertical axis indicates the MFCC parameter dimension, signifying the number of MFCCs extracted per frame. This number determines the dimensionality of the feature vector for each frame. For example, if the order of the DCT is 30, then 30 MFCC coefficients will be generated for each frame, but usually only the first 13 coefficients are retained because these lower-order coefficients contain the main spectral information, while the higher-order coefficients tend to be associated with noise.

Figure 8 illustrates the Delay-Doppler domain features of the Fraser's Dolphin acoustic signal. The horizontal axis represents the Doppler frequency shift, indicating the change in signal frequency resulting from the relative motion between the target and the receiver. The vertical axis depicts the time delay, which is the duration of signal propagation. The Delay-Doppler domain features effectively reflect the speed characteristics of organisms, as varying speeds among different organisms inherently leads to distinct Delay-Doppler domain features. As evident from the figure, the signal exhibits a pronounced frequency shift at approximately 8 Hz.

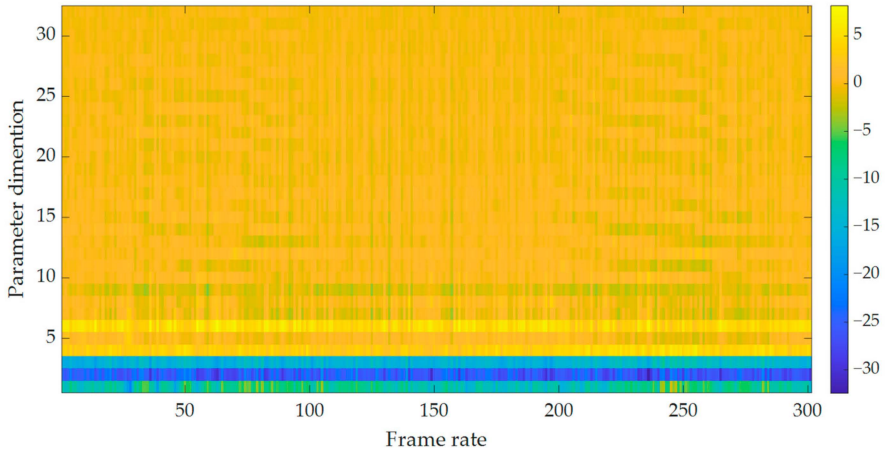


Figure 7. MFCC features.

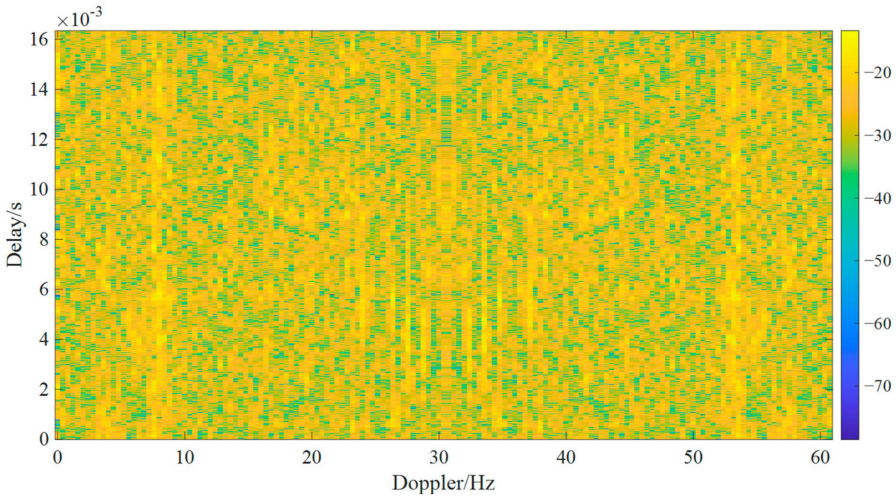


Figure 8. Delay-Doppler features.

3.3. Dual-Feature Fusion Learning Target Recognition Model

Traditional CNN models are typically based on a single feature and can only be trained for that specific feature, which often limits their generalization ability and robustness. Consequently, in this paper, we introduce a dual-feature fusion learning target recognition model that is capable of simultaneously inputting features from both the time-frequency and Delay-Doppler domains. The acoustic signals generated by different targets may have frequency overlap, so it is challenging to rely only on the time-frequency domain features of the signals to identify the targets. Since different targets move at different speeds, they are characterized differently in the DD domain. The target attributes in the other feature domain of the signal can be described by DD domain features, and the classification approach is equivalent to combining the MFCC features and the DD domain features as a joint feature and classifying the different marine mammals by using this joint feature as an axis, which corresponds to different points on this axis. Due to the increased dimensionality of the described signal, the target information can be reflected more comprehensively. We compare the recognition performance of this model with three widely used single-feature models: VGG16, GoogleNet, and ResNet.

As depicted in Figure 9, our dual-feature fusion learning target recognition model comprises nine convolutional layers and two fully connected layers. Notably, a max pooling layer is inserted between every two convolutional layers, and the recognition task is ultimately carried out by a SoftMax layer. The maximum pooling layer reduces the size of the feature map by selecting the maximum value of each region, which reduces the amount of computation and the number of parameters in the subsequent layers, helping to improve

the computational efficiency of the model and reduce the risk of overfitting. In addition to this, the maximum pooling layer provides a degree of translation invariance, so that even if there is a small translation of the image, the maximum pooling layer can extract the same features. With the pooling operation, the model is able to retain the most important features and ignore unimportant details, thus improving the robustness of the features. The SoftMax layer transforms the output of the neural network into a probability distribution, such that the output value of each category is between 0 and 1, and the sum of the output values of all the categories is 1. This allows the model to perform better on multi-classification tasks and improves the accuracy of the classification. The fully connected layer is able to integrate the features extracted from the convolutional and pooling layers and map these features to the sample labeling space, enhancing the feature integration capability of the model.

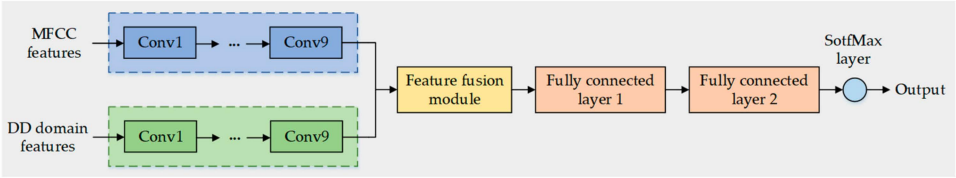


Figure 9. The architecture of the dual-feature fusion learning target recognition model.

Figure 10 presents the specific parameters of the nine convolutional layers. For instance, the parameter "1 × 1 × 32" signifies the following: the convolutional kernel size is 1 × 1, allowing for cross-channel information integration without altering the spatial dimensions; furthermore, the number of convolutions is 32, indicating that the convolutional operation yields 32 feature maps that reflect the outcomes of convolving the input data with this kernel. Additionally, the activation function employed is ReLU (Rectified Linear Unit), renowned for its simplicity in computation, rapid convergence, and effectiveness in mitigating the gradient vanishing problem. The normalization method used is Min-Max Normalization, which is implemented by the maximum and minimum values of each feature, using linear transformations to map the data to new ranges.

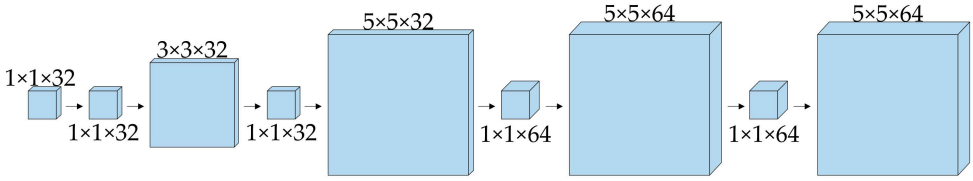


Figure 10. Parameters of the convolutional layers in the model.

4. Experiment and Analysis

In this section, the effectiveness of the aforementioned dual-feature fusion learning target recognition model is verified. First, the pre-processed signal undergoes MFCC and DD domain feature extraction. Subsequently, these two types of features are input into VGG16, GoogleNet, ResNet, and the dual-feature fusion learning target recognition model for training. The recognition performances of these different models are then analyzed and compared. The experimental environment configuration for this study is as follows: the GPU is NVIDIA GeForce RTX 4080 (NVIDIA Corporation, Santa Clara, CA, USA), the CPU is Intel i9-13900K (Intel Corporation, Santa Clara, CA, USA), and the neural network is trained to utilize the GPU. The operating system is Windows 10, with 128 GB of RAM (Samsung, Seoul, Republic of Korea). The Python version used is 3.9, and the neural network is constructed using the PyTorch framework. The development environment is PyCharm 2018.

4.1. Experimental Data and Evaluation Metrics

The marine mammal acoustic data used in this study were obtained from the Watkins Marine Mammal Sound Database open-source database, which was collected by William

Watkins, one of the founding fathers of marine mammal bioacoustics, and contains recordings from seven decades, from the 1940s to the 2000s. The database contains approximately 2000 unique recordings of more than 60 species of marine mammals, ranging in length from one second to several minutes, all in .wav format. Figure 11 shows three marine mammals used in the experiment [38].

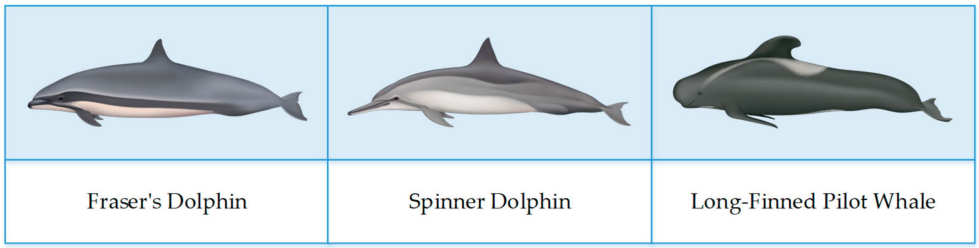


Figure 11. Types of marine mammals used in this study.

Due to the varying lengths of the audio data in the dataset and the short duration of marine mammal vocalizations, in this paper, the longer audio data are clipped into multiple short segments, which increases the amount of input data and improves the accuracy of the training. To ensure that each short segment after clipping contains at least one complete animal vocalization, this paper utilizes Adobe Audition 2024 software to analyze the time-frequency diagrams of the audio data before clipping. The dataset is divided into training sets and test sets in the ratio of 4:1, and the following experimental results are analyzed from these two aspects.

In this study, four metrics were selected to evaluate the experimental results: Accuracy, Precision, Recall, and F1 Score. These metrics are derived from the confusion matrix, which evaluates the model's accuracy by comparing the predicted category labels with the actual category labels. The structure of the confusion matrix is presented in Table 1 [39].

Table 1. Structure of the confusion matrix.

	Actual Category: Positive	**Actual Category: Negative**
Predicted category: Positive	*TP*	*FP*
Predicted category: Negative	*FN*	*TN*

True Positives (*TP*): the number of samples that the model correctly predicts as positive.
False Positives (*FP*): the number of samples that the model incorrectly predicts as positive.
True Negatives (*TN*): the number of samples that the model correctly predicts as negative.
False Negatives (*FN*): the number of samples that the model incorrectly predicts as negative.

Accuracy is the proportion of correctly predicted samples to the total number of samples. Precision is the proportion of samples that are positive classes among all samples that are predicted to be positive classes. Recall is the proportion of samples that are correctly predicted to be positive classes among all samples that are positive classes. The F1-score is the harmonic mean of precision and recall, which is used to measure the balanced performance of the model. They are calculated as follows [40]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

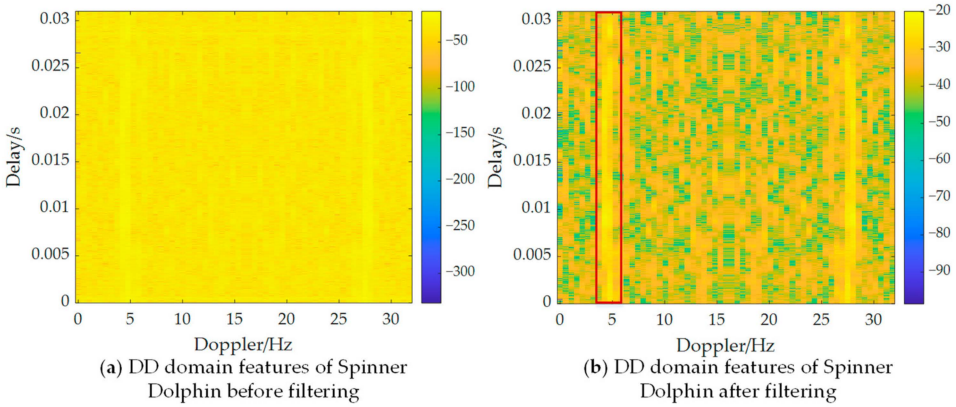
$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

4.2. Experimental Validation

In this paper, Section 2 mentions that a set of triangular filters is used for extracting MFCC features, which serve to improve the SNR of the original signal. The DD features are extracted without filtering the signal, whereas the signal is analyzed by LMS adaptive filtering before DD feature extraction. The analysis results are presented in Figure 12, which shows the DD domain feature extraction of the Spinner Dolphin before and after filtering, respectively. Notably, the results are significantly different, with the DD domain feature extraction after filtering making the visualization features more obvious and enhancing image contrast.



(**a**) DD domain features of Spinner Dolphin before filtering

(**b**) DD domain features of Spinner Dolphin after filtering

Figure 12. Comparison of DD domain features before and after filtering.

In this study, all marine mammal acoustic signals were minimum mean square filtered, and then the DD-domain features of the original signals and the DD-domain features of the filtered signals were input into the three single feature models of VGG16, GoogleNet, and ResNet as the target recognition features, respectively. The accuracy, precision, recall, and F1 scores of the two groups were obtained for comparison, and the results are shown in Tables 2 and 3:

Table 2. Recognition results of DD domain features before filtering.

Model-Feature	Acc (%)	Pre (%)	Re (%)	F1 (%)
VGG16-DD	62.47/64.71	62.76/75.87	62.47/64.71	62.31/64.42
GoogleNet-DD	72.60/54.90	73.13/55.90	72.59/54.90	72.55/54.01
ResNet-DD	74.82/57.84	75.15/57.84	74.82/57.84	74.69/57.78

Table 3. Recognition results for DD domain features after filtering.

Model-Feature	Acc (%)	Pre (%)	Re (%)	F1 (%)
VGG16-DD	75.56/59.80	75.71/67.08	75.56/59.80	75.55/59.22
GoogleNet-DD	79.26/72.55	79.29/73.23	79.26/72.55	79.24/72.74
ResNet-DD	77.28/59.80	77.44/60.48	77.28/59.80	77.33/59.82

The table contains the results of the training set and the test set; the two kinds of data are separated by "/", the data in front of "/" is the result of the training set, the data on the right side of "/" is the result of the test set. By comparing the recognition results of DD domain features from single-feature models before and after filtering, it becomes evident that the recognition accuracy of each model improves by 3% to 13% after filtering. Consequently, applying LMS adaptive filtering before feature extraction and recognition can effectively enhance recognition accuracy.

MFCC and DD domain features are extracted from preprocessed data and individually input into VGG16, GoogleNet, and ResNet models for single-feature training. Subsequently, these two types of features are simultaneously input into the dual-feature fusion learning

target recognition model for training. Table 4 presents the recognition results for the two features using the four models.

Table 4. Recognition results (Fraser's Dolphin, Spinner Dolphin, Long-Finned Pilot Whale).

Model-Feature	Acc (%)	Pre (%)	Re (%)	F1 (%)
VGG16-MFCC	94.07/95.10	94.09/95.10	94.07/95.10	94.08/95.13
VGG16-DD	75.56/59.80	75.71/67.08	75.56/59.80	75.55/59.22
GoogleNet-MFCC	93.09/95.10	93.22/95.18	93.09/95.10	93.11/95.10
GoogleNet-DD	79.26/72.55	79.29/73.23	79.26/72.55	79.24/72.74
ResNet-MFCC	96.05/97.06	96.09/97.11	96.05/97.06	96.06/97.07
ResNet-DD	77.28/59.80	77.44/60.48	77.28/59.80	77.33/59.82
Dual-feature	99.26/98.04	99.28/98.04	99.26/98.04	99.26/98.04

In the final line, "Dual-feature" signifies that the dual-feature fusion learning target recognition model is employed to recognize both MFCC features and DD domain features. Analyzing the training results reveals that the accuracy rate stands at 59% to 80% for DD domain features alone, 93% to 98% for MFCC features alone, and a remarkable 98% to 99% when both MFCC features and DD domain features are utilized through the dual-feature fusion learning target recognition model. Therefore, the dual-feature fusion learning target recognition model is better than the other models in recognizing the acoustic signals of three marine mammals, namely, the Fraser's Dolphin, the Spinner Dolphin, and the Long-Finned Pilot Whale. Comparison with other models alone does not adequately demonstrate the superiority of the model proposed in this paper. As such, this study conducts generalizability experiments and ablation experiments to validate the model's ability to generalize and the reasonableness of the methodology.

4.3. Generalization Ability Analysis

To verify the generalizability of this dual-feature fusion learning target recognition model, two distinct marine mammal acoustic signals—the Ross seal and the Bearded seal—were chosen for training in this study. The training results are presented in Table 5.

Table 5. Recognition results (Ross seal, Bearded seal).

Model-Feature	Acc (%)	Pre (%)	Re (%)	F1 (%)
VGG16-MFCC	97.73/73.91	97.84/73.91	97.73/73.91	97.73/73.91
VGG16-DD	88.64/69.57	88.74/77.43	88.64/69.57	88.64/70.15
GoogleNet-MFCC	92.05/73.91	92.09/74.47	92.05/73.91	92.06/73.61
GoogleNet-DD	90.91/86.96	91.03/89.97	90.91/86.96	90.92/86.96
ResNet-MFCC	97.73/86.96	97.73/87.39	97.73/86.96	97.73/87.01
ResNet-DD	77.29/59.80	77.44/60.48	77.28/59.80	77.33/59.82
Dual-feature	98.86/91.30	98.89/92.55	98.86/91.30	98.87/91.20

When DD domain features are recognized in isolation, the accuracy ranges from 59% to 90%. When MFCC features are recognized alone, the accuracy lies between 73% and 97%. However, when the dual-feature fusion learning target recognition model recognizes both MFCC and DD domain features, the accuracy soars to 91% to 98%. The superior recognition accuracy achieved by the dual-feature fusion learning target recognition model, as compared to other models, underscores its excellent generalization capabilities.

4.4. Ablation Experiment

Ablation Study is an experimental design method commonly used in scientific research, especially in the fields of machine learning and deep learning. The core idea is to gain a deeper understanding of how the model works and how the components interact with each other by systematically removing or modifying certain parts of the model

(e.g., layers, nodes, features, parameters, etc.) and observing how such changes affect the model's performance.

An ablation experiment was conducted in this study to validate the efficacy of MFCC features, DD domain features, and LMS adaptive filtering within the proposed method for marine mammal acoustic signal recognition. By systematically removing each component from the model, we aimed to assess their contributions to the overall performance.

Ablation Experiment 1: MFCC features are removed and DD domain features are trained using CNN.

Ablation Experiment 2: Remove DD domain features and train on MFCC features using CNN.

Ablation Experiment 3: Remove LMS adaptive filtering and directly perform feature extraction on the original signal.

The training results of the three ablation experiments and the complete target recognition model are compared in Table 6, which shows that the removal of MFCC features decreases the target recognition accuracy by about 43%; the removal of DD domain features decreases the target recognition accuracy by 0–3%; and the removal of LMS adaptive filtering decreases the target recognition accuracy by 1–3%.

Table 6. Results of ablation experiments.

Ablation Experiment	Acc (%)	Pre (%)	Re (%)	F1 (%)
1	56.30/54.90	56.38/57.62	56.30/54.90	56.15/54.35
2	96.54/98.04	96.54/98.10	96.54/98.04	96.53/98.03
3	97.29/95.10	97.40/95.34	97.28/95.10	97.30/95.10
Dual-feature	99.26/98.04	99.28/98.04	99.26/98.04	99.260/98.04

These findings demonstrate that the inclusion of MFCC features, DD domain features, and LMS adaptive filtering is crucial for achieving optimal performance in the target recognition model. Removing any of these components leads to a decrease in model accuracy, highlighting their contributions to the overall performance.

4.5. Qualitative Validation

The loss function is a non-negative real-valued function used to quantify the difference between model predictions and true labels [41]. By calculating the value of the loss function, the accuracy of the model predictions can be quantified, and thus the performance of the model can be evaluated. By minimizing the value of the loss function during training, the parameters of the model can be optimized so that the model's predictions are closer to the true labels. To provide a more intuitive analysis of the processes and effects of each model and feature recognition, we have plotted the loss function curves during the training process and the recognition accuracies of each model in Figures 13 and 14.

An explanation of the legend follows:

VGG-MFCC. Recognition of MFCC domain features using the VGG16 model.

VGG-DD. Recognition of DD domain features using the VGG16 model.

GoogleNet-MFCC. Recognition of MFCC features using the GoogleNet model.

GoogleNet-DD. Recognition of DD features using the GoogleNet model.

ResNet-MFCC. Recognition of MFCC features using the ResNet model.

ResNet-DD. Recognition of DD domain features using the ResNet model.

MFCC-DD. Recognition of MFCC and DD domain features using the dual-feature fusion learning target recognition model.

Figure 13 illustrates the variation of the loss function during the training of the three single-feature models and the dual-feature fusion learning target recognition model. Figure 14, on the other hand, shows the accuracies of different models for two types of feature recognition, including both the training and test sets, enabling a more direct comparison of the recognition performance of each model.

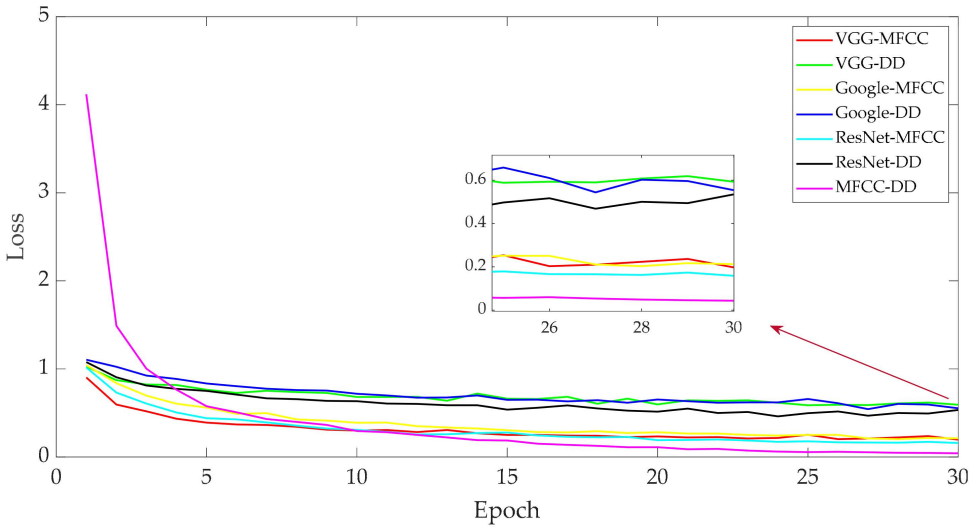


Figure 13. Loss function of each model during the training process.

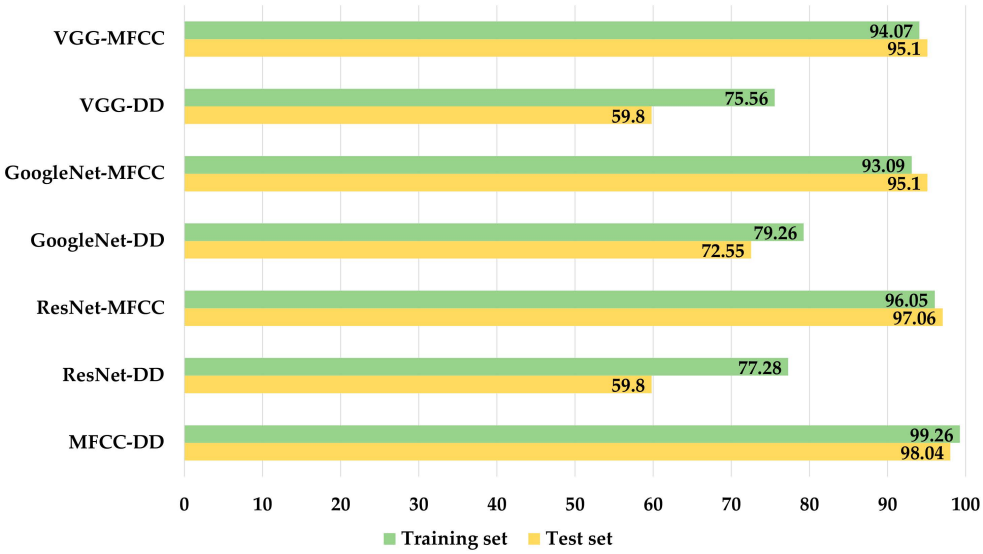


Figure 14. Recognition accuracy of each model.

(1) When recognizing DD domain features alone, the loss function decreases slowly and exhibits significant fluctuations. In contrast, recognizing MFCC features alone results in a relatively faster decrease in the loss function with fewer fluctuations. Notably, the dual-feature fusion learning target recognition model, which incorporates both MFCC and DD domain features, achieves an even faster decrease in the loss function with a smoother curve.

(2) Compared to the model using MFCC features alone, the dual-feature fusion learning target recognition model improves the accuracy of the training set by 3% to 6% and the accuracy of the test set by 1% to 3%. When compared to the model using DD domain features alone, the improvement in accuracy is even more pronounced, with an increase of 20% to 23% for the training set and 25% to 38% for the test set. Additionally, compared to models utilizing single features such as VGG16, GoogleNet, and ResNet, the structure of the dual-feature fusion learning target recognition model is simpler, contributing to an overall enhancement in model recognition efficiency.

5. Conclusions

In this paper, we propose a dual-feature fusion learning method that mainly consists of two parts: feature extraction and target recognition.

(1) Feature Extraction: The MFCC and DD domain features of marine mammals are extracted as input features. The MFCC, being closer to the human auditory system than other spectral features, captures the auditory characteristics of the marine mammals' vocalizations. Meanwhile, the DD domain features reflect the motion characteristics of these animals, providing complementary information. By combining these two features, the model ensures robust recognition performance even under low SNR conditions.

(2) Dual-Feature Fusion Learning Target Recognition Model: We introduce a novel dual-feature fusion learning target recognition model that can simultaneously input both features into a convolutional neural network for target recognition. In addition, generalizability experiments and ablation experiments are carried out in this study, which prove that the model has good generalization ability.

Compared with the traditional single-feature recognition model, the method proposed in this paper simplifies the model structure, improves the recognition accuracy and training efficiency, and has good generalization ability, which can provide some references for research in marine mammal acoustic recognition and other related fields. The method is based on the recognition of marine mammals by passive sonar, so it is suitable for the recognition of marine mammals that can actively emit sound, but it is less effective for the recognition of some fish. The acoustic signals emitted by underwater targets are transmitted through acoustic channels in the ocean, and their signal-to-noise ratio is bound to be greatly reduced, so noise in the ocean is an important issue affecting target recognition. With the increasing changes in the marine environment, including the impacts of climate change and human activities, the habitats and behavioral patterns of marine mammals are also changing [42]. Effective acoustic signal recognition techniques can help monitor these changes and study and protect animals and their habitats non-invasively and at ecologically relevant temporal and spatial scales. The future of this research can be applied to (1) the combination of active and passive sonar for the identification of marine organisms; (2) the combination of marine environmental noise filtering technology for detection; and (3) the upgrading of hardware and software technology, which will be mounted on a variety of marine observation platforms to observe organisms in the ocean in real-time, allowing us to achieve the goal of monitoring and protecting the marine biological environment.

In this paper, three models, VGG16, GoogleNet, and ResNet, are used as references in the comparison experiments, and future research will consider using more up-to-date algorithms for comparisons to improve the recognition efficiency of other models and to find more appropriate models. This study mainly relies on marine mammal acoustic signal data from the Watkins Marine Mammal Sound Database open-source database, and future studies will consider using more diverse datasets, including real-world data, to explore the performance of the target recognition model with dual-feature fusion learning in different environments.

Author Contributions: Conceptualization, Y.S. and Z.L.; methodology, Y.S.; validation, Y.S., Z.L. and Z.W.; formal analysis, Y.S.; investigation, Y.S. and Z.W.; writing-original draft preparation, Y.S.; writing-review and editing, Y.S., Z.L. and D.H.; supervision, Z.L., L.L., Z.W., F.Y. and D.H.; and project administration, Y.S. and Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (2022YFC2808003) and the Natural Science Foundation of Qingdao Municipality (23-2-1-100-zyyd-jch).

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to laboratory confidentiality regulations.

Acknowledgments: The authors would like to express their gratitude for the data used in this study: the Watkins Marine Mammal Sound Database. This valuable database was instrumental in conducting the research and validating the proposed model.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

References

1. Lilly, J.C. Sonic-ultrasonic emissions of the bottlenose dolphin. *Whales Dolphins Porpoises* **1966**, *165*, 503–507.
2. Fleishman, E.; Cholewiak, D.; Gillespie, D.; Helble, T.; Klinck, H.; Nosal, E.; Roch, M.A. Ecological inferences about marine mammals from passive acoustic data. *Biol. Rev.* **2023**, *98*, 1633–1647. [CrossRef] [PubMed]
3. Cauchy, P.; Heywood, K.J.; Merchant, N.D.; Risch, D.; Queste, B.Y.; Testor, P. Gliders for passive acoustic monitoring of the oceanic environment. *Front. Remote Sens.* **2023**, *4*, 1106533. [CrossRef]
4. Ibrahim, A.K.; Zhuang, H.; Erdol, N.; Ali, A.M. A new approach for North Atlantic right whale upcall detection. In Proceedings of the 2016 International Symposium on Computer, Consumer, and Control (IS3C), Xi'an, China, 4–6 July 2016; IEEE: New York, NY, USA, 2016; pp. 260–263.
5. Brown, J.C.; Smaragdis, P. Hidden Markov and Gaussian mixture models for automatic call classification. *J. Acoust. Soc. Am.* **2009**, *125*, EL221–EL224. [CrossRef] [PubMed]
6. Dugan, P.J.; Rice, A.N.; Urazghildiiev, I.R.; Clark, C.W. North Atlantic right whale acoustic signal processing: Part I. Comparison of machine learning recognition algorithms. In Proceedings of the 2010 IEEE Long Island Systems, Applications, and Technology Conference, Farmingdale, NY, USA, 7 May 2010; IEEE: New York, NY, USA, 2010; pp. 1–6.
7. Nadir, M.; Adnan, S.M.; Aziz, S.; Khan, M.U. Marine mammals classification using acoustic binary patterns. *Arch. Acoust.* **2020**, *45*, 721–731.
8. Zhong, M.; Cai, W. Marine Mammal Sound Recognition Based on Feature Fusion. *Electron. Sci. Tech.* **2019**, *32*, 32–37.
9. Li, S.; Liu, P.; Yan, J.; Wang, K.; Gan, W.; Wang, J. A marine mammal classification method based on acoustic features. In Proceedings of the 2021–2022 Academic Conference of the Hydroacoustics Branch of the Acoustical Society of China, Qingdao, China, 15 August 2022; South China Sea Research Station, Institute of Acoustics, Chinese Academy of Sciences: Beijing, China, 2022; pp. 432–435.
10. Zhang, X.; Yang, P.; Wang, Y.; Shen, W.; Yang, J.; Wang, J.; Ye, K.; Zhou, M.; Sun, H. A novel multireceiver sas rd processor. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4203611. [CrossRef]
11. Zhang, X.; Yang, P.; Sun, M. Experiment results of a novel sub-bottom profiler using synthetic aperture technique. *Curr. Sci.* **2022**, *122*, 461. [CrossRef]
12. Zhang, X.; Yang, P.; Sun, H. Frequency-domain multireceiver synthetic aperture sonar imagery with Chebyshev polynomials. *Electron. Lett.* **2022**, *58*, 995–998. [CrossRef]
13. Zhang, X. An efficient method for the simulation of multireceiver SAS raw signal. *Multimed. Tools Appl.* **2023**, *83*, 37351–37368. [CrossRef]
14. Aslam, M.A.; Zhang, L.; Liu, X.; Irfan, M.; Xu, Y.; Li, N.; Zhang, P.; Jiangbin, Z.; Yaan, L. Underwater sound classification using learning based methods: A review. *Expert Syst. Appl.* **2024**, *255*, 124498. [CrossRef]
15. Bianco, M.J.; Gerstoft, P.; Traer, J.; Ozanich, E.; Roch, M.A.; Gannot, S.; Deledalle, C.-A. Machine learning in acoustics: Theory and applications. *J. Acoust. Soc. Am.* **2019**, *146*, 3590–3628. [CrossRef] [PubMed]
16. Zhang, X.; Yang, P.; Feng, X.; Sun, H. Efficient imaging method for multireceiver SAS. *IET Radar Sonar Navig.* **2022**, *16*, 1470–1483. [CrossRef]
17. Wang, H.; Zhang, W.; Ren, P. Self-organized underwater image enhancement. *ISPRS J. Photogramm. Remote Sens.* **2024**, *215*, 1–14. [CrossRef]
18. Wang, H.; Zhang, W.; Bai, L.; Ren, P. Metalantis: A Comprehensive Underwater Image Enhancement Framework. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5618319. [CrossRef]
19. Wang, H.; Sun, S.; Bai, X.; Wang, J.; Ren, P. A reinforcement learning paradigm of configuring visual enhancement for object detection in underwater scenes. *IEEE J. Ocean. Eng.* **2023**, *48*, 443–461. [CrossRef]
20. Wang, H.; Sun, S.; Chang, L.; Li, H.; Zhang, W.; Frery, A.C.; Ren, P. INSPIRATION: A reinforcement learning-based human visual perception-driven image enhancement paradigm for underwater scenes. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108411. [CrossRef]
21. Shiu, Y.; Palmer, K.J.; Roch, M.A.; Fleishman, E.; Liu, X.; Nosal, E.-M.; Helble, T.; Cholewiak, D.; Gillespie, D.; Klinck, H. Deep neural networks for automated detection of marine mammal species. *Sci. Rep.* **2020**, *10*, 607. [CrossRef]
22. Griffiths, E.T.; Archer, F.; Rankin, S.; Keating, J.L.; Keen, E.; Barlow, J.; Moore, J.E. Detection and classification of narrow-band high frequency echolocation clicks from drifting recorders. *J. Acoust. Soc. Am.* **2020**, *147*, 3511–3522. [CrossRef]
23. Cai, W.; Zhu, J.; Zhang, M.; Yang, Y. A parallel classification model for marine mammal sounds based on multi-dimensional feature extraction and data augmentation. *Sensors* **2022**, *22*, 7443. [CrossRef] [PubMed]
24. Duan, D.X.; Jiang, Y.; Liu, Z.W.; Yang, C.M.; Lv, L.G. Echolocation signal detection method based on image processing. *Adv. Mar. Sci.* **2022**, *40*, 145–153.
25. Cominelli, S.; Bellin, N.; Brown, C.D.; Rossi, V.; Lawson, J. Acoustic features as a tool to visualize and explore marine soundscapes: Applications illustrated using marine mammal passive acoustic monitoring datasets. *Ecol. Evol.* **2024**, *14*, e10951. [CrossRef] [PubMed]
26. Widrow, B.; McCool, J.; Ball, M. The complex LMS algorithm. *Proc. IEEE* **1975**, *63*, 719–720. [CrossRef]
27. Zhang, X.; Yang, P.; Wang, Y.; Shen, W.; Yang, J.; Ye, K.; Zhou, M.; Sun, H. LBF-based CS algorithm for multireceiver SAS. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1502505. [CrossRef]
28. Liu, J.T.; Xi, B.; Jiang, H. Performance analysis of adaptive filters based on LMS algorithm. *Nav. Electron. Eng.* **2021**, *41*, 36–39.

29. Brummund, M.K.; Sgard, F.; Petit, Y.; Laville, F. Three-dimensional finite element modeling of the human external ear: Simulation study of the bone conduction occlusion effect. *J. Acoust. Soc. Am.* **2014**, *135*, 1433–1444. [CrossRef]
30. Rajaby, E.; Sayedi, S.M. A structured review of sparse fast Fourier transform algorithms. *Digit. Signal Process.* **2022**, *123*, 103403. [CrossRef]
31. Zheng, F.; Zhang, G.; Song, Z. Comparison of different implementations of MFCC. *J. Comput. Sci. Technol.* **2001**, *16*, 582–589. [CrossRef]
32. Xing, W.; Tang, X.; Zhou, Y.; Zhang, C.; Pan, Z. A review of delay-Doppler domain channel estimation methods for OTFS. *J. Commun.* **2022**, *43*, 188–201.
33. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. [CrossRef]
34. Lavanya, G.; Teja, U.S.M.P.; Dussa, V.K.; Reddy, A.L.; Nitish, Y. Classification of Underwater Fish Species Using Custom-Built Deep Learning Architectures. In Proceedings of the International Conference on Data Science and Applications, Jaipur, India, 14–15 July 2023; Springer Nature Singapore: Singapore, 2023.
35. Altarez, R.D. Faster R–CNN, RetinaNet and Single Shot Detector in different ResNet backbones for marine vessel detection using cross polarization C-band SAR imagery. *Remote Sens. Appl. Soc. Environ.* **2024**, *36*, 101297. [CrossRef]
36. Manikandan, D.L.; Santhanam, S.M. Parallel desires: Unifying local and semantic feature representations in marine species images for classification. *Mar. Geophys. Res.* **2024**, *45*, 16. [CrossRef]
37. Cabello-Solorzano, K.; Ortigosa de Araujo, I.; Peña, M.; Correia, L.; JTallón-Ballesteros, A. The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis. In Proceedings of the International Conference on Soft Computing Models in Industrial and Environmental Applications, Salamanca, Spain, 5–7 September 2023; Springer Nature Switzerland: Cham, Switzerland, 2023.
38. Sayigh, L.; Daher, M.A.; Allen, J.; Gordon, H.; Joyce, K.; Stuhlmann, C.; Tyack, P. The Watkins marine mammal sound database: An online, freely accessible resource. In *Proceedings of Meetings on Acoustics*; AIP Publishing: Melville, NY, USA, 2016; Volume 27.
39. Du, L.; Wang, Z.; Lv, Z.; Han, D.; Wang, L.; Yu, F.; Lan, Q. A Method for Underwater Acoustic Target Recognition Based on the Delay-Doppler Joint Feature. *Remote Sens.* **2024**, *16*, 2005. [CrossRef]
40. Deng, X.; Liu, Q.; Deng, Y.; Mahadevan, S. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Inf. Sci.* **2016**, *340–341*, 250–261. [CrossRef]
41. Xu, P.; Zhu, X.; Clifton, D.A. Multimodal learning with transformers: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12113–12132. [CrossRef]
42. Ghani, B.; Denton, T.; Kahl, S.; Klinck, H. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Sci. Rep.* **2023**, *13*, 22876. [CrossRef]