Small Sample Underwater Acoustic Target Recognition Based on Full Dimensional Dynamic Feature Enhancement Network

Jianxun Tang[1], Zhe Chen[2], Mingsong Chen[1,2(✉)], Junyi Wang[2], and Xiaodong Ma[2]

[1] School of Ocean Engineering, Guilin University of Electronic Technology, Beihai 536000, Guangxi, China
cms@guet.edu.cn

[2] School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, Guangxi, China

Abstract. Due to the complex marine background noise, the existing underwater acoustic target recognition methods are mainly based on a single recognition network, which is difficult to fully consider the multi-faceted characteristics of underwater acoustic targets and to filter and identify target features from the underwater acoustic signals mixed with a large amount of marine noise. At the same time, the traditional underwater acoustic target recognition model is mainly a two-stage model, which cannot meet the existing demand for rapid recognition of underwater acoustic targets. In order to solve the above problems, this paper proposes a full-dimensional dynamic feature enhancement network for small sample underwater acoustic target recognition model (CAODCSP-Darknet), which mainly uses ODCf2MHSA to make the model acquire multiple receptive field gradient flow information in full dimension while focusing on the target, and PloyLoss to customize the classification loss function for small sample underwater acoustic dataset to avoid the model the overfitting problem caused by small samples and imbalance of samples in each category. The experimental results show that the proposed model outperforms the other three baseline models by more than 20% and achieves 98.3% accuracy in identifying four classes of underwater acoustic signals on the Deepship underwater acoustic dataset. The performance is better than existing underwater acoustic target recognition models.

Keywords: Underwater Acoustic Target Recognition · Deep Learning · ODConv · MHSA · PolyLoss

1 Introduction

As one of the core technologies for realising the intelligence of weapons and equipment, underwater automatic target recognition technology has important military significance. The problem of underwater acoustic target recognition is of great complexity, and the complex underwater environment can distort the radiation noise, making underwater

J. Wang (Ed.): MobiMedia 2023, LNICST 576, pp. 86–95, 2025.
https://doi.org/10.1007/978-3-031-60347-1_9

acoustic target recognition more challenging than conventional speech recognition. At the same time, as feature extraction and classifier are usually 2 relatively independent parts in traditional passive recognition systems for underwater acoustic targets, although feature extraction and classifier have been extensively researched, these methods still suffer from single recognition features, low recognition rate and low integration, etc. Therefore, how to design an efficient end-to-end method remains the main research problem for existing underwater acoustic target recognition models.

In recent years, deep learning has emerged in the field of machine learning. Unlike traditional methods, deep learning only requires input information containing rich target characteristics, which can be mined through layer-by-layer learning so that the most discriminative abstract information can be automatically extracted for recognition. From the perspective of system connectivity, feature extraction and pattern recognition in deep learning are an integral part of the system, avoiding the "coupling" effect of traditional methods due to step-by-step execution of feature extraction and pattern recognition, and further improving the performance of the recognition system. Wang et al. [1] proposed a new CNN structure for automatic recognition of underwater targets by adding a convolutional layer with a kernel of 1 to the network structure, which better preserves the time-domain information of the audio signal and effectively improves the model's underwater acoustic target recognition accuracy to 91.7%. [2] applied Auto-Encoder to the underwater target recognition task and achieved over 90% accuracy.

Although the existing deep learning based underwater acoustic target recognition can solve some of the shortcomings of traditional underwater acoustic target detection, there are still the following problems. Firstly, due to the difficulty of acoustic data collection, the existing public acoustic data sets are all small sample data sets, and the deep learning-based acoustic target recognition model is prone to overfitting in the feature extraction process of small samples, and secondly, as the marine background noise mainly exists in the low frequency band, the target information in the low frequency band is mixed with the marine background noise in the feature extraction process, and the traditional convolution operation can easily lose the target The effective information is lost.

In order to solve the above problem, this paper proposes a small sample underwater acoustic target recognition model CAODCSP-Darknet based on a full-dimensional dynamic feature enhancement network. Firstly, inspired by the Cross Stage Partial Network (CSP) [3] module which achieves fast acquisition of image features by streaming the original convolutional feature extraction process, the proposed The ODCf2MHSA module is proposed to obtain underwater acoustic feature information quickly and efficiently through multi-gradient flow [4]. Then, PolyLoss [5] is introduced to change the loss function by adjusting the polynomial coefficients and dynamically adjust the loss function ratio according to the samples of the underwater acoustic target dataset to alleviate the overfitting problem of the classification model. Finally, in order to solve the problem of overfitting in training small samples of underwater acoustic targets, this paper trains the CAODCSP-Darknet model in the Imagenet dataset to obtain pre-training weights, and then uses migration learning to obtain classification results after convergence in the underwater acoustic target dataset.

2 The Proposed Programme

The model proposed in this paper, CAODCSP-Darknet, consists of two main components, the feature extraction module and the CAODCSP-Darknet classifier. In the feature extraction module, Constant Q transform (CQT) [6] is used to extract energy spectrum features from the raw underwater acoustic data, and then the data set is expanded using the SpecAugment data enhancement method and used as the input to the classifier. CAODCSP-Darknet mainly consists of several ODCf2MHSA modules. The main function of the ODCf2MHSA module is to quickly obtain effective deep abstract features of the target energy spectrum and multiple gradient flow information while filtering the ship radiation noise. Finally, this model uses the PolyLoss loss function to adjust the coefficients of the first polynomial of the Taylor expansion for different loss functions to perform differential supervision for different channels and to alleviate model overfitting of the classifier model due to large differences in the sample size of the underwater acoustic target dataset. The detailed architecture of CAODCSP-Darknet is shown in Fig. 1.

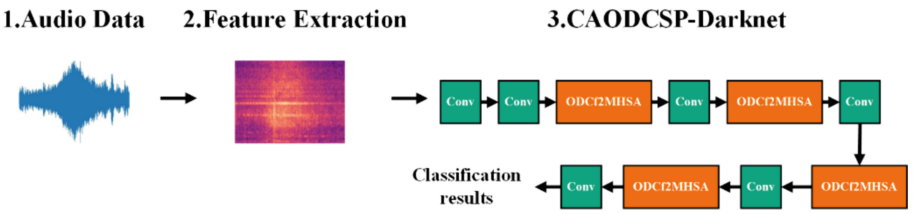

Fig. 1. CAODCSP-Darknet architecture

2.1 Feature Extraction Module

In this paper, the signal with a sampling frequency of 22050Hz is divided into frames with a window length of 43ms and each frame has 512 sampling points. To ensure local stability, 40 frames are chosen to extract the underwater acoustic signal features.

It is very difficult to extract target sound features directly from the original audio timing signal because there is a large amount of marine background noise in the original audio signal, and the frequency of marine background noise is similar to that of some ship radiation noise. However, the underwater acoustic target signal has a variety of physical features, and the target information can be extracted for its physical features with targeted features to filter the marine background noise and improve the target separability. As the underwater radiation signal carries very useful signal information in the low frequency sub-band, which is associated with the features of the propeller, and CQT provides better frequency resolution in the low frequency sub-band, and is inspired by a speech detection method in voice active detection (vad) (high energy is speech, low energy is noise), this paper adopts (CQT) the energy spectrum as the main feature extraction method for the

underwater acoustic dataset. Its detailed formulation is as follows:

$$X(m,n) = \sum_{i=n-\lfloor N_m/2 \rfloor}^{n+\lfloor N_m/2 \rfloor} x(i)(a_m^*)(i-n+N_m/2) \tag{1}$$

where $m = 1, 2, ..., M$ is the number of CQT frequency bins, $\lfloor . \rfloor$ is rounding to negative infinity, N_m is the window length, and a_m^* is complex conjugate of basis function.

2.2 ODC2fMHSA

Since the underwater acoustic energy spectrogram has fewer pixel points, the ordinary convolution operation will lose a large number of effective features. Therefore, in order to obtain the effective feature information of the underwater acoustic target quickly and effectively, this paper introduces the Cf2 module of YOLOv8 [7], which can quickly obtain the rich gradient flow information by shunting the original input and effectively improve the feature extraction ability of the model. However, since the Cf2 module cannot effectively focus on underwater acoustic target features, Multi-headed Self-attention (MHSA) [8] is introduced to increase the model's focus on underwater acoustic target features. MHSA is a leading self-attention mechanism for detecting targets in the field of target recognition, so the 3×3 convolution in Bottleneck to form the ODBottleneckMHSA, which effectively improves the effective feature extraction capability of the model. Secondly, since the static convolution operation only focuses on the image characteristics of the line spectrum and ignores the four-dimensional information such as the number of convolution kernels, the size of convolution kernels, and the number of input and output channels. Therefore, in order to obtain more full-dimensional effective target feature information in the limited line spectrum energy map, this paper introduces Omni-Dimensional Dynamic Convolution (ODConv) [9] to replace the traditional static convolution operation and improve the effective feature extraction capability of the model by multi-dimensional attention and parallel strategy. In this paper, the Cf2 module that combines MHSA and ODConv is named ODCf2MHSA, and its detailed model structure is shown in Fig. 2.

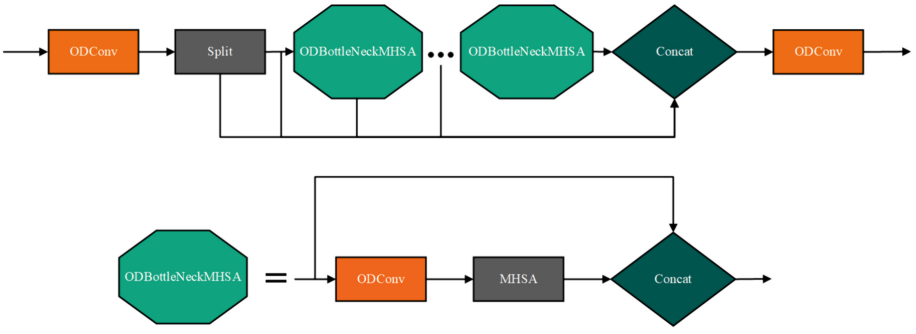

Fig. 2. Model structure of ODCf2MHSA

MHSA.
The multi-head mechanism of MHSA can effectively improve the expressiveness of the model, while also allowing the model to learn more diverse and complex features. Under the multi-head mechanism, the input sequence data is split into multiple heads, each of which performs independent computations to obtain different outputs. These outputs are finally stitched together to form the final output.

The detailed calculation steps are as follows, assuming the query is $q \in R^{d_q}$, the key is $k \in R^{d_k}$ and the value is $v \in R^{d_v}$. Each attention header can be expressed using the mathematical formula as

$$h_i = f\left(w_i^{(q)} q, w_i^{(k)} k, w_i^{(v)} v\right) \in R^{pv} \tag{2}$$

where the learnable parameters consist mainly of $w_i^{(q)}$, $w_i^{(k)}$ and $w_i^{(v)}$ and a function representing the convergence of attention f which can use either additive attention or scaled dot product attention. The output of a multi-headed attention requires another linear transformation, corresponding to the result of joining the h heads.

$$w_o \begin{bmatrix} h_1 \\ .... \\ h_h \end{bmatrix} \in R^{Po} \tag{3}$$

where w_o is the learnable parameter.

ODConv.
The model structure is shown in Fig. 3, where α_{si} is the attention of the k × k convolutional kernel space, α_{ci} is the attention of the input channel, α_{fi} is the attention of the output channel, and α_{wi} is the attention of the convolutional kernel W_i.

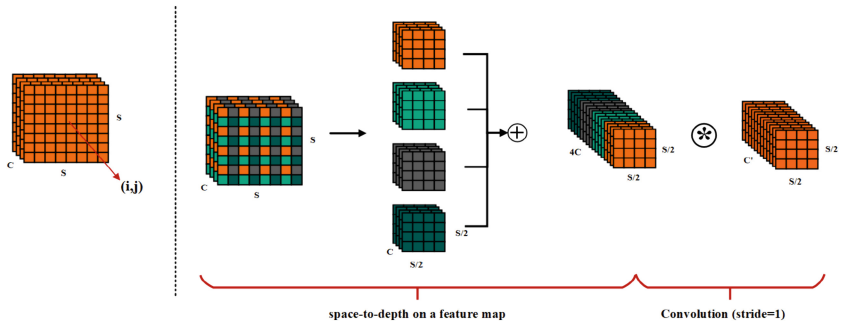

Fig. 3. Model structure of ODConv

When the convolution kernel is W_i, the details of the operations at α_{si}, α_{ci}, α_{fi} and α_{wi} are shown in the figure:

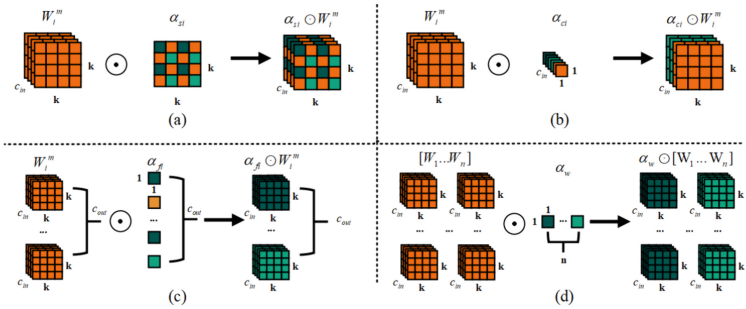

Fig. 4. Flow of the operations of α_{si}, α_{ci}, α_{fi} and α_{wi}

2.3 PolyLoss

In order to solve the problem of overfitting the model due to the large differences in the number of underwater acoustic data samples for each recognition category and the small number of samples, the model uses the PolyLoss loss function as the classification loss function for the underwater acoustic target recognition model. It is used to improve the adaptability of the model to the underwater acoustic target data set by changing the first polynomial coefficients of the Taylor expansions of the original Cross-entropy loss and focal loss. The formulation of PloyLoss used in this paper is expressed as follows.

$$L_{Poly-1} = -\log(P_t) + \varepsilon_1(1 - P_t) \tag{4}$$

where P_t is the target prediction probability, ε_1 is used as the improved dynamic parameter in this paper, $\varepsilon_1 \in [-1, \infty)$.

3 Experimentation and Analysis

The experiments are conducted on a real underwater acoustic public dataset containing four types of underwater acoustic target data and the effectiveness of the proposed approach is verified by comparing the proposed model qualitatively and quantitatively with CNN, DNN and ResNet18 network models.

3.1 Data Sets

The public dataset Deepship [10] used in this experiment contains 4 types of vessels: Cargo, Passenger Ship, Tanker and Tug. In the experiment, the 603 underwater acoustic data are divided into a training set (about 80%), a validation set (about 10%) and a test set (about 10%). The time domain waveform plots of some of these data sets with the corresponding CQT energy spectrum images are shown in Fig. 5.

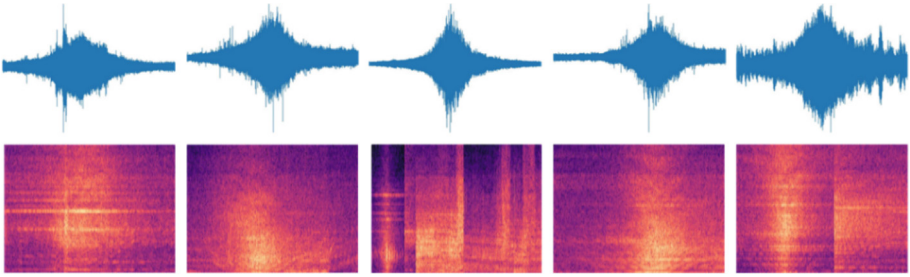

Fig. 5. Partial time domain waveforms of the four types of underwater acoustic data in Deepship with corresponding CQT energy spectra

3.2 Experimental Environment and Performance Indicators

The experiments in this paper were conducted on a Pytorch deep learning platform with two Intel Xeon Sliver 4310 CPUs, one NVIDIA Tesla A100 80G GPU, and 256GB RAM environment. In order to ensure the accuracy and authority of the experimental results, different models were trained and tested qualitatively and quantitatively during the experiments, and then the performance of the algorithms was compared and analysed.

The following equation is used in this paper to evaluate the performance metrics of the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$F1 - Score = 2 \times \frac{precision \times recall}{precision + recall} \tag{6}$$

$$precision = \frac{TP}{TP + FP} \tag{7}$$

$$recall = \frac{TP}{TP + FN} \tag{8}$$

where TP is a true positive, FP is a false positive and FN is a false negative.

3.3 Comparative Analysis of Model Performance Results

The variation of each performance metric during the training of CAODCSP-Darknet using the underwater acoustic dataset presented in Sect. 3.1 is shown in Fig. 6.

As shown in Fig. 6(a), the CAODCSP-Darknet model converges at 100 epochs of training and there is no fluctuating change subsequently. Also, according to the changes of other parameters in Fig. 6(b), it can be seen that the model is more adaptable and does not produce overfitting phenomenon even after convergence and then 150 epochs of training.

Tables 1 show the Accuracy, Precision, Recall and F1-Score results of CAODCSP-Darknet and CNN, DNN and Resnet18 for the test set in the Deepship underwater acoustic dataset. As can be seen from Table 1, the CAODCSP-Darknet model improved

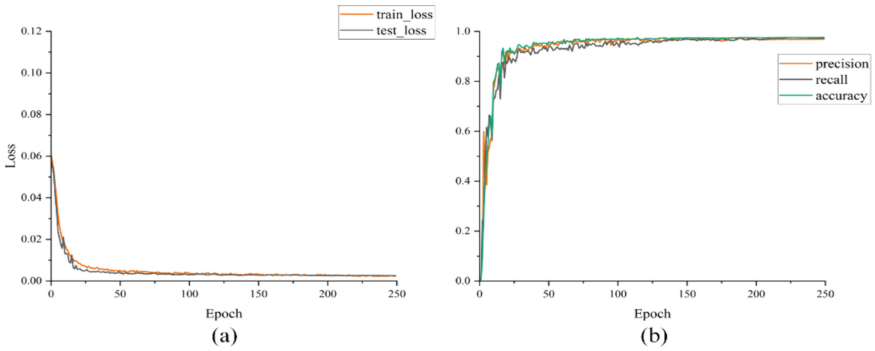

Fig. 6. Parameter changes during model training: a. Loss changes. b. Precision, recall and accuracy changes

Accuracy by 20% over Resnet18, and by 21.6% and 25% over CNN and DNN models respectively. It also outperformed the comparison models by more than 22% in Precision, Recall and F1-Score, thanks to the classifier model's full-dimensional feature extraction of target valid features, multi-dimensional feature collation and fusion before image classification, and a PolyLoss function customised to the underwater acoustic dataset.

Table 1. Identification Accuracy of different models on Deepship data sets.

Network Model	Accuracy	Precision	Recall	F1-Score
CAODCSP-Darknet	0.983	0.992	0.992	0.992
Resnet18	0.783	0.773	0.772	0.772
CNN	0.767	0.761	0.762	0.761
DNN	0.733	0.738	0.735	0.736

To further validate the performance of this model, the Accuracy of the four network models for different categories in the Deepship dataset was also compared, as shown in Table 2. The accuracies of Passenger, Tanker and Tug in CAODCSP-Darknet exceeded those of the comparison models by more than 21%, except for Cargo, for which the proposed model achieved 100%. Accuracy decreased compared to the other classes, but also reached 90.9%, exceeding the Resnet18, CNN and DNN models by 18.2%, 18.2% and 27.3%, respectively. Therefore, the recognition rate of CAODCSP-Darknet far exceeds that of the existing mainstream underwater acoustic classification models in all categories, demonstrating the performance of the proposed model for underwater acoustic target recognition.

Table 2. Identification Accuracy of different categories of Deepship.

Network Model	Cargo	Passenger	Tanker	Tug
CAODCSP-Darknet	0.909	1	1	1
Resnet18	0.727	0.789	0.783	0.857
CNN	0.727	0.789	0.739	0.857
DNN	0.636	0.737	0.739	0.857

4 Conclusion

In order to improve the shortcomings of a single neural network in classification recognition based on the overfitting problem caused by the limited number of samples in underwater acoustic target recognition, to make full use of the characteristics of ship noise signals as temporal signals, and to achieve a network structure with higher recognition accuracy, this paper proposes a full-dimensional dynamic feature enhancement underwater acoustic target recognition network CAODCSP-Darknet, through the ODCf2MHSA module quickly acquires the gradient flow information of different receptive fields while the full-dimensional dynamic convolution, while the MHSA makes the model pay more attention to the underwater acoustic target feature information, extracts the full-dimensional effective target feature information of the underwater acoustic target, and then improves the target recognition performance of the model.

After experimental demonstration, the recognition rate of the proposed underwater acoustic target recognition network has been greatly improved compared with traditional neural networks, and the network structure is simple, which provides a new research direction for underwater acoustic target recognition methods.

The shortcomings of this paper are that for some targets with high inter-class similarity, there are still a small number of false detections, and the data set used in this paper is single, so it is impossible to verify the performance of this network in the actual marine environment. Therefore, the future research directions are, firstly, to reduce the inter-class similarity from the perspective of feature fusion, and secondly, to expand the dataset by collecting more ship noise audio signals in the measured marine environment, and to enhance the universality of the network by increasing the number of datasets and continuously training and optimising the network parameters.

Acknowledgements. This study was supported in part by the National Natural Science Foundation of China (Grant No.91836301), the Special Program of Guangxi Science and Technology Base and Talents (Grant No.AD21220098) and the Innovation Project of GUET Graduate Education (Grant No. 2023YCXS021).

References

1. Xiaoyu, W., Fan, L., Lin, C., et al.: End to end underwater targets recognition using the modified convolutional neural network. J. Signal Process. **36**(6), 958–965 (2020)

2. Cao, X., Xiaomin Z., Yang Y., Letian N.: Deep learning-based recognition of underwater target. In: 2016 IEEE International Conference on Digital Signal Processing (DSP), pp. 89–93 (2016)
3. Wu, W., et al.: Application of local fully convolutional neural network combined with YOLO v5 algorithm in small target detection of remote sensing image. PLoS ONE **16**(10), e0259283 (2021)
4. Xiao, J., et al.: Context augmentation and feature refinement network for tiny object detection. (2022)
5. Leng, Z., et al.: Polyloss: a polynomial expansion perspective of classification loss functions. arXiv preprint arXiv:2204.12511 (2022)
6. Lidy, T., Alexander, S.: CQT-based convolutional neural networks for audio scene classification. DCASE, pp. 60–64 (2016)
7. Terven, Juan, and Diana Cordova-Esparza. A Comprehensive Review of YOLO: From YOLOv1 to YOLOv8 and Beyond. arXiv preprint arXiv:2304.00501 (2023)
8. Tan, H., Xiuping, L., Baocai, Y., Xin, L.: MHSA-Net: multihead self-attention network for occluded person re-identification. IEEE Trans. Neural Netw. Learn. Syst. **34**, 8210 (2022)
9. Li, C., Aojun, Z., Anbang, Y.: Omni-dimensional dynamic convolution. arXiv preprint arXiv:2209.07947 (2022)
10. 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. **183**, 115270 (2021)