Differentiable Architecture Search with Multi-domain Time-Frequency Features for Underwater Target Recognition

Yule Chen*, Hong Liang†, Zezhou Dai‡
*†‡School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China
Email: chenyule@mail.nwpu.edu.cn, lianghong@nwpu.edu.cn(corresponding author), zz_dai@mail.nwpu.edu.cn

***Abstract*—Underwater target recognition using active sonar echoes has gained significant research interest due to its critical role in marine surveillance and security. However, the sonar sensors encounter disturbances from seafloor reverberation and the complicated noise background. Traditional classification methods rely heavily on manual feature extraction and expert-defined systems, hindering their adaptability and robustness. In this paper, we propose a novel differentiable architecture search algorithm (DARTs-MTF) with multi-domain time-frequency features for underwater target recognition. Initially, we apply three different time-frequency transformation methods to the pre-processed active sonar echoes to extract robust echo features. We obtain the comprehensive feature dataset through the encoder-decoder network. By redesigning the search space of differentiable architecture search, the optimal network architecture is automatically identified from the extracted feature dataset. Extensive experiments conducted on South China Sea trial datasets demonstrate that the proposed DARTs-MTF approach achieves superior performance compared to existing manually designed classifiers and neural architecture search methods, achieving an overall accuracy of 83.08%.**

***Index Terms*—underwater target recognition, differentiable architecture search, multi-domain features, deep learning**

I. Introduction

Underwater Acoustic Target Recognition (UATR) [1] is a vital component of marine information processing, widely utilized for oceanographic research, naval defense, and underwater resource exploration. UATR can be broadly classified into active sonar-based and passive sonar-based methods [2]. The former transmits acoustic signals and detects echoes returning from targets, while the latter receives and detects sounds generated by targets. Due to its ability to detect targets even under low acoustic signal conditions, active sonar is particularly advantageous and has become extensively deployed [3].

The complex underwater environment poses challenges to traditional parameter-based active sonar target recognition methods [4]. Bernice Kubicek et al. [5] propose a multivariate statistical method, canonical correlation analysis, as a feature extraction technique prior to the multi-class classification of active sonar echoes. A target classification algorithm is proposed that can use power-normalized cepstral coefficients for target classification [6]. Nevertheless, these methods typically suffer from limited feature extraction capabilities, difficulties in iterative data updates, and reduced robustness against noise and environmental variability [7].

Recently, deep learning [8] has shown significant advantages over traditional methods by using data-driven approaches for feature learning. Despite their success, these approaches still heavily rely on manual network architecture designs, which involve extensive trial-and-error and expert-driven tuning processes [9], [10]. Moreover, deep learning methods for active sonar echo recognition often focus on a single feature domain, performing poorly when recognizing echo targets with high sample similarity.

Manually selecting from numerous potential model architectures often results in suboptimal performance or generalization [11]. Neural Architecture Search (NAS) [12] is a emerging method that automatically designs neural network architectures based on data-driven approaches. Despite significant progress in various fields [13], [14], NAS research in UATR remains scarce and challenging.

To address these issues, we apply the Short-Time Fourier Transform (STFT), Mel Transform (Mel), and Continuous Wavelet Transform (CWT) to extract and fuse multi-domain features, creating a comprehensive feature dataset. By redesigning the search space to automatically select the optimal network structure from the dataset, we propose a neural architecture search method tailored for UATR (DARTs-MTF). Experiments conducted on active sonar echo datasets collected during sea trials in the South China Sea validate the effectiveness of our approach.

II. Methods

The proposed method adapts the DARTS framework [15] for underwater target echoes by reconstructing the search space and improving strategies. Initially, pre-processed time-domain echo data are transformed using Short-Time Fourier Transform (STFT), Mel-Spectrogram (Mel), and Continuous Wavelet Transform (CWT) to create multi-domain time-frequency tensors. These tensors are processed through a redesigned encoder-decoder network incorporating an adaptive attention block to enhance feature extraction capabilities. The search process employs a directed acyclic graph (DAG) representing

This work is supported by National Natural Science Foundation of China under Grant 61971354 ,62371394. (Corresponding author:Hong Liang.)

OCEANS 2025 Brest | 979-8-3315-3747-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/OCEANS58557.2025.11104626

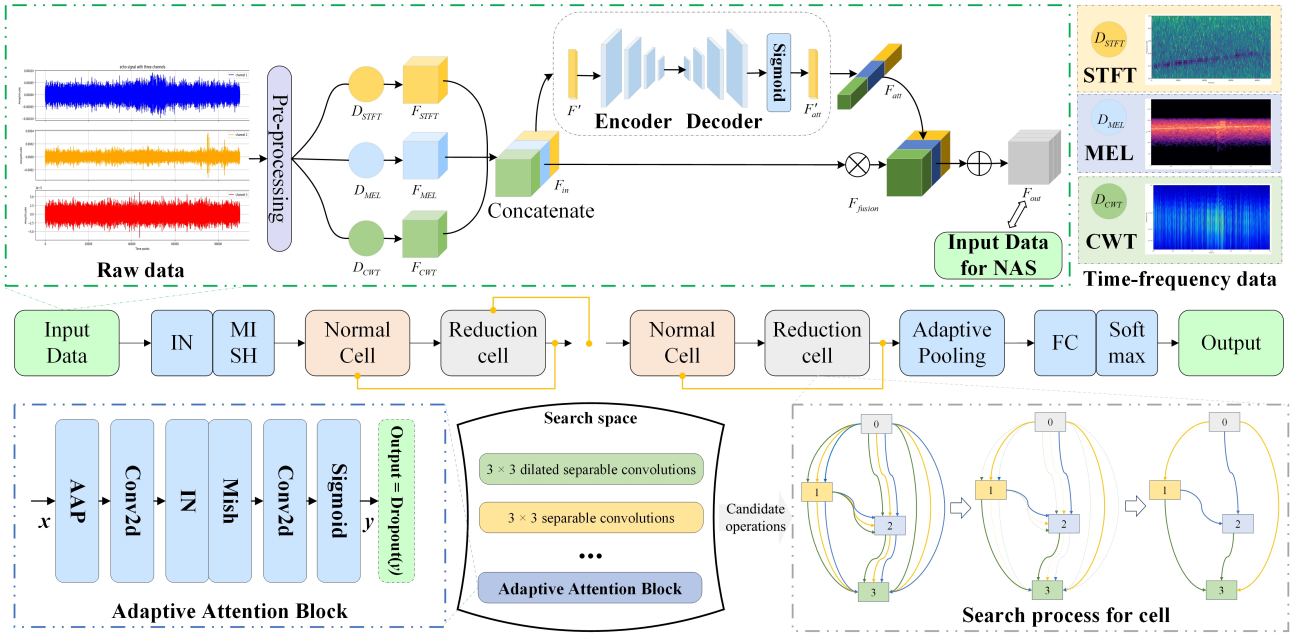


Fig. 1: The diagram depicts a pipeline for processing underwater echoes using our DARTs-MTF. The top layer of the figure is for input data processing, the middle layer is the model architecture, and the bottom layer illustrates the search space and the cell search process.

mixed operations between node pairs. The overall framework is shown in Fig. 1.

A. Multi-Domain Time-Frequency Feature Extraction

We extract robust features by applying STFT, Mel, and CWT methods to echo data. Each method captures unique aspects of the signals, collectively improving robustness and discriminability.

1) Short-Time Fourier Transform (STFT): STFT decomposes the signal into time-frequency components as follows:

$$\mathbf{X}_{\text{STFT}}(\tau,\omega)=\int_{-\infty}^{\infty}x(t)w(t-\tau)e^{-j\omega t}\,dt \tag{1}$$

where $x(t)$ is the time-domain signal, $w(t)$ is the window function, τ is the time shift, and ω represents the frequency.

2) Mel-Spectrogram(Mel): Inspired by human auditory perception, Mel-Spectrogram computation includes:

$$E(i,k)=|X_i(k)|^2=|FFT[X_i(m)]|^2 \tag{2}$$

Apply Mel filter bank to the power spectrum to obtain the Mel frequency spectrum:

$$\mathbf{X}_{\text{Mel}}(i,m)=\sum_{k=0}^{N-1}E(i,k)H_m(k),\quad 0\le m<M \tag{3}$$

where $H_m(k)$ is the m th Mel filter.

3) Continuous Wavelet Transform (CWT): CWT captures multi-scale features and is defined as:

$$\mathbf{X}_{\text{CWT}}(a,b)=\int_{-\infty}^{\infty}x(t)\frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right)dt \tag{4}$$

where a is the scale parameter, b is the translation parameter, and $\psi(t)$ is the mother wavelet function.

B. Feature Fusion and Encoder-Decoder Network

1) Feature Fusion: Outputs from STFT, Mel, and CWT are fused into a tensor $\mathbf{F}_{\text{in}}$:

$$\mathbf{F}_{\text{in}}=[\mathbf{X}_{\text{STFT}},\mathbf{X}_{\text{Mel}},\mathbf{X}_{\text{CWT}}] \tag{5}$$

2) Encoder and Decoder Networks: The encoder network [16] extracts high-level features:

$$\mathbf{F}'=\sigma(\mathbf{W}\cdot\mathbf{F}_{\text{in}}+\mathbf{b}) \tag{6}$$

where $\mathbf{F}'$ is the activation of the layer, $\mathbf{W}$ and $\mathbf{b}$ are the weights and biases, and σ is the activation function (e.g., ReLU). The decoder network maps these features back to the original space through deconvolution layers, restoring the structural information of the input features to obtain $\mathbf{F}_{\text{att}}$.

C. Adaptive Attention Block

The proposed adaptive attention block (*A_A_Block*) enhances feature extraction robustness, comprising adaptive pooling operation (AAP), convolutional layer, instance normalization (IN), Mish activation function, sigmoid activation function, and dropout, as detailed in Fig. 1.

D. DARTs-MTF Search Space Reconstruction

The search space is structured as a DAG with nodes as feature mappings and edges as operations:

$$o^{(i,j)}(x)=\sum_{o\in\mathcal{O}}\frac{\exp(\alpha_o^{(i,j)})}{\sum_{o'\in\mathcal{O}}\exp(\alpha_{o'}^{(i,j)})}o(x) \tag{7}$$

The optimization goal of DARTs-MTF is to determine the optimal network architecture α by minimizing the validation

loss function, thus maximizing the model's classification accuracy. Optimizing architecture involves minimizing validation loss:

$$\begin{aligned} \min_{\alpha} \quad & \mathcal{L}_{\text{val}}(\omega^*(\alpha), \alpha) \\ \text{s.t.} \quad & \omega^*(\alpha) = \arg\min_{\omega} \mathcal{L}_{\text{train}}(\omega, \alpha) \end{aligned} \tag{8}$$

E. Loss Functions and Optimization

To effectively train the model, we use an optimized loss function to optimize training process. We combine cross-entropy and contrastive losses to enhance feature separability:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) + \lambda \frac{1}{2N} \sum_{i,j} \left(y_{ij} D_{ij}^2 + (1 - y_{ij}) \max(0, m - D_{ij})^2 \right) \tag{9}$$

where λ is the weight of the contrastive loss.

Algorithm 1 DARTs-MTF Framework

Require: Initial architecture parameters $\alpha_{(i,j)}$, network model weight parameters ω, learning rate ϵ, number of layers l_k, initial channels c_k, number of epochs $e \in \{1, 2, \ldots, E\}$

Ensure: The final architecture based on the learned α

1: Initialize the network weights ω, learning rate ϵ

2: **for** each epoch $e \in \{1, 2, \ldots, E\}$ **do**

3: Approximate $\pi(\tau)$ via k sampled architectures $\{\alpha_1, \ldots, \alpha_K\}$ drawn from $\pi(\alpha|\tau_0)$

4: Calculate $\nabla_w \mathcal{J}(w)$ according to:

$$\nabla_w \mathcal{J}(w) \leftarrow \frac{1}{N} \sum_{\tau_0 \sim \pi(\tau)} \left[\sum_k \alpha_i \sim \pi(\alpha|\tau_0) \nabla_w \gamma(\alpha_i) \mathcal{L}_{\text{train}}(w, \alpha_i) \right]$$

5: Update architecture α by descending:

$$\alpha \leftarrow \nabla_\alpha \mathcal{L}_{\text{val}}(w', \alpha) - \xi \nabla_\alpha^2 \mathcal{J}(w) \nabla_w \mathcal{L}_{\text{val}}(w', \alpha)$$

6: Update weights ω by descending:

$$\omega \leftarrow \nabla_\omega \mathcal{L}_{\text{train}}(w(\alpha), \alpha)$$

7: **end for**

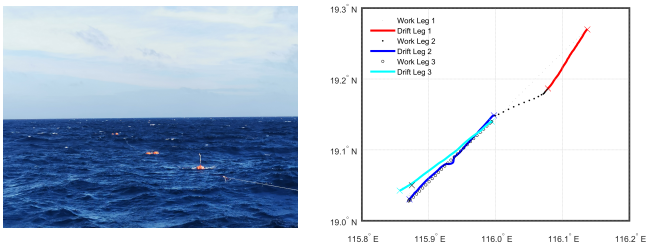
8: Extend the network depth l_k and the initial channels c_k (network width)

9: Derive the final architecture based on the learned α

III. EXPERIMENTS AND ANALYSIS

To demonstrate the innovation of our algorithm for UATR, we have conducted sea trial tests in the South China Sea. The details are shown in the Fig. 2. As the number of echo samples for the three types of targets is 254, 159, and 245. We give an example of one set of the echo data as shown in Fig. 3.

The experimental results are summarized in Table I. The manually designed CNN model (ConvNext-V2-A) achieved the highest overall accuracy (OA) of 77.53% . In comparison, the model derived using the DARTS framework attained an OA of 79.29%. Notably, the optimal architecture obtained through DARTS-MTF further improved performance, reaching an OA of 83.08% after full training.



(a) (b)

Fig. 2: Experimental implementation details. (a) Experimental conditions in the South China Sea. (b) The drift trajectory of the experimental ship.

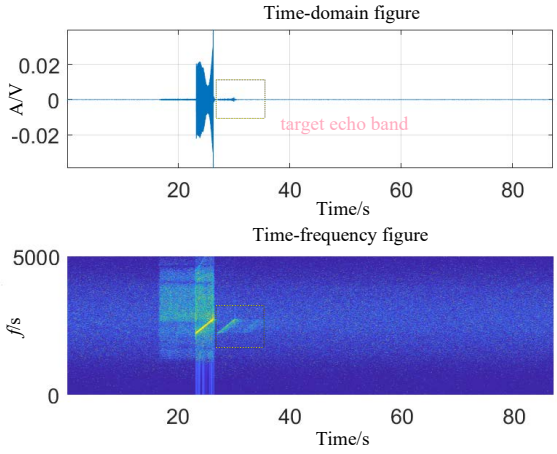


Fig. 3: The figure presents the time-domain waveform and corresponding time-frequency representation of echo data acquired during sea trials. The target echo segment is highlighted by a yellow box, while transmission leakage and reverberation components are also depicted.

TABLE I: Comparison of classification accuracy results.

Types	Models	OA(%)	Parameters	time(s)
Manual	ResNet-18 [17]	72.22	11.69M	159.1
	ConvNeXt V2-A [18]	77.53	3.70M	195.4
NAS	DARTs [15]	79.29	4.13M	271.3
	DARTs-MTF(ours)	83.08	6.90M	312.5

To evaluate the efficiency of these methods, we consider the number of parameters and training time as key indicators. ResNet-18 and ConvNext-V2-A contain 11.69M and 3.70M parameters, respectively, while the DARTS-derived model has 4.13M parameters. Our proposed DARTS-MTF model includes 6.90M parameters, a moderate increase compared to DARTS due to the integration of the adaptive attention block. NAS-based models require longer training times than manually designed counterparts, which can be attributed to variations in library calls and code execution during training.

Figures 4 and 5 illustrate the structures of the optimal model's normal and reduction cells, derived from sea trial data. The incorporation of the *A_A_Block* in both cells enhances feature extraction and robustness, significantly improving recog-

nition accuracy.

As shown in Fig. 4, the normal cell preserves the spatial resolution of the input feature map while incorporating various operations such as separable convolutions (*sep_conv*) and dilated convolutions (*dil_conv*). Inputs from previous cells (c_{k-2} and c_{k-1}) are processed through these operations, and their outputs are aggregated to form c_k.

Fig. 5 depicts the reduction cell, which reduces the spatial resolution by halving both the height and width. This cell processes inputs from c_{k-2} and c_{k-1} using operations such as skip connections (*skip_connect*), separable convolutions, and dilated convolutions. The resulting output, c_k, retains critical feature representations while effectively reducing dimensionality, facilitating deeper network architectures with improved efficiency.

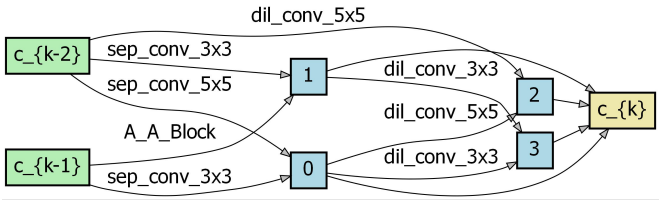


Fig. 4: The normal cell in our DARTs-MTF maintains the spatial resolution of the input feature map.

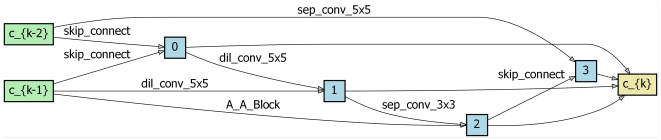
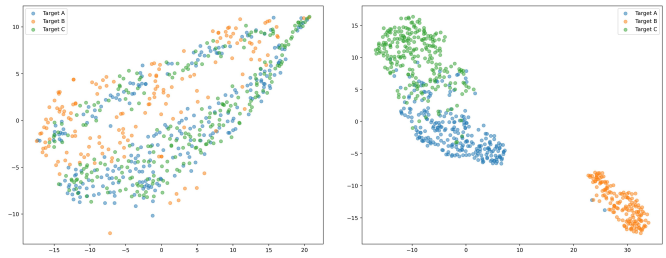


Fig. 5: The reduction cell downscales the spatial resolution by halving the height and width.

We further visualize the results using t-distribution stochastic neighbor embedding (t-SNE) on the dataset in Fig. 6. Through the advanced feature extraction of the model obtained via DARTs-MTF, the data points progressively evolve into well-defined, distinct clusters. This highlights the robust feature extraction capability of our method.



(a) (b)

Fig. 6: Visualization before and after using DARTs-MTF for feature extraction. (a) Before the feature extraction. (b) After the feature extraction.

IV. Conclusion

In this paper, we present DARTs-MTF, a novel differentiable architecture search algorithm that integrates multi-domain time-frequency features for UATR. We apply three distinct time-frequency transformation methods on pre-processed active sonar echoes and utilized an encoder-decoder network to construct feature sets. An optimal model is then constructed within a redesigned search space. Extensive experiments on a sea trial dataset demonstrat that DARTs-MTF outperforms existing manual design methods and other NAS approaches. However, our work is not yet perfect. Future research should focus on improving search efficiency and feature matching to further enhance the performance of NAS-based on underwater target recognition systems.

References

[1] J. Jiang, T. Shi, M. Huang, *et al.*, "Multi-scale spectral feature extraction for underwater acoustic target recognition," *Measurement*, vol. 166, Art. no. 108227, Dec. 2020.

[2] Y. Choo, K. Lee, W. Hong, *et al.*, "Active underwater target detection using a shallow neural network with spectrogram-based temporal variation features," *IEEE J. Ocean. Eng.*, vol. 49, no. 1, Jan. 2024, pp. 279–293.

[3] K. T. Hjelmervik, K. E. Andreassen, E. M. Bohler, *et al.*, "Bayesian occupancy grid for active sonar detection and localization of moving targets," in *Proc. OCEANS 2020 Singapore – U.S. Gulf Coast*, Biloxi, MS, USA, Nov. 2020, pp. 1–9.

[4] Z. Sun, *et al.*, "Bio-inspired covert active sonar detection method based on the encoding of sperm whale clicks," *IEEE Sens. J.*, vol. 22, no. 2, Jan. 2022, pp. 1449–1460.

[5] B. Kubicek, A. S. Gupta, and I. Kirsteins, "Canonical correlation analysis as a feature extraction method to classify active sonar targets with shallow neural networks," *J. Acoust. Soc. Am.*, vol. 152, no. 5, Nov. 2022, pp. 2893–2904.

[6] S. Lee, I. Seo, J. Seok, *et al.*, "Active sonar target classification with power-normalized cepstral coefficients and convolutional neural network," *Appl. Sci.*, vol. 10, no. 23, 2020, Art. no. 8450.

[7] Y. Hwang, G. Kim, S. Shin, *et al.*, "Attention-based complementary learning for active target classification with limited sonar data," *IEEE Access*, vol. 12, 2024, pp. 79787–79801.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, May 2015, pp. 436–444.

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

[10] Y. Chen, H. Liang, and S. Pang, "Study on small samples active sonar target recognition based on deep learning," *J. Mar. Sci. Eng.*, vol. 10, no. 8, 2022, Art. no. 1144.

[11] G. Kim and Y. Choo, "Enhancing Generalization of Active Sonar Classification Using Semisupervised Anomaly Detection With Multisphere for Normal Data," *IEEE J. Ocean. Eng.*, vol. 49, no. 4, Oct. 2024, pp. 1530–1548.

[12] P. Ren, Y. Xiao, X. Chang, *et al.*, "A comprehensive survey of neural architecture search: Challenges and solutions," *ACM Comput. Surv.*, vol. 54, no. 4, 2021.

[13] Y. Chen, H. Liang, and S. Jiao, "NAS-MFF: NAS-Guided Multiscale Feature Fusion Network With Pareto Optimization for Sonar Images Classification," *IEEE Sens. J.*, vol. 24, no. 9, May 2024, pp. 14656–14667.

[14] C. Peng, Z. Zeng, J. Gao, *et al.*, "PNAS-MOT: Multi-Modal Object Tracking With Pareto Neural Architecture Search," *IEEE Robot. Autom. Lett.*, vol. 9, no. 5, May 2024, pp. 4377–4384.

[15] H. Liu, K. S. Karen, and Y. Yang, "DARTS: Differentiable architecture search," in *ICLR Conf. Blind Submission*, 2018.

[16] Z. Qiao, Y. Zhou, D. Yang, *et al.*, "SEED: Semantics enhanced encoder-decoder framework for scene text recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 13528–13537.

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

[18] S. Woo, S. Debnath, R. Hu, *et al.*, "ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders," in *Proc. 2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, 2023, pp. 16133–16142.