

Time-Frequency Feature Fusion with Multi-Scale Attention for Underwater Acoustic Target Recognition

Gang Hu¹, Yifei Song², Mohamad Farhan Mohamad Mohsin³

School of Mathematics, Anshan Normal University, Anshan, Liaoning, China¹

University Utara Malaysia (UUM), 06010 Sintok, Kedah, Malaysia²

School of Computing, University Utara Malaysia (UUM), Sintok, Kedah, Malaysia³

Abstract—Traditional convolutional neural networks (CNNs) for underwater acoustic target recognition (UATR) are limited by restricted local receptive fields, and reliance on frequency-domain inputs alone can discard temporal information present in the raw waveform. This study proposes TFMSUATR, a model that fuses a raw-waveform branch and a log-magnitude spectrogram branch through a learned adaptive gate, refines the fused representation with a multi-scale attention module with squeeze-and-excitation re-weighting, and models long-range dependencies with a Swin Transformer encoder. Under this protocol, TFMSUATR achieves an accuracy of 97.41% on the six-group ShipsEar task and 94.87% on DeepShip under identical grouping, clip length, and split. Controlled input ablations, a capacity-controlled comparison, additive-noise robustness, cross-dataset transfer, and Grad-CAM saliency analysis are reported. The results indicate that time-waveform fusion and multi-scale attention improve recognition relative to single-domain and single-scale baselines.

Keywords—Underwater acoustic target recognition; time-frequency feature fusion; multi-scale attention; Swin Transformer; global dependency modeling

I. INTRODUCTION

Underwater acoustic target recognition (UATR) plays a vital role in marine environmental monitoring, underwater surveillance, and marine resource exploration. CNNs have achieved notable success in UATR [1, 2], but their locally confined receptive fields limit the capture of global contextual and long-range dependencies in ship-radiated noise [3]. Transformer-based architectures, and the Swin Transformer in particular, have therefore been introduced to model long-range dependencies efficiently [3-9].

Despite these advances, three issues remain. First, reliance on frequency-domain inputs alone can discard temporal information carried by the raw waveform. Second, high-frequency discriminative features at multiple temporal and spectral scales are underutilized by single-scale attention. Third, and crucially, many published UATR results split clips after segmentation, so overlapping windows from the same recording appear in both training and test sets; this inflates accuracy and makes comparisons non-transparent.

This study proposes TFMSUATR, a Swin-Transformer-based UATR model that 1) fuses a raw-waveform branch and a spectrogram branch through a learned adaptive gate, 2) refines

the fused representation with a multi-scale attention module, and 3) is evaluated under a strict recording-level split. The contributions are 1) an explicit time-waveform fusion design with a learnable gate; 2) a multi-scale attention module with dilated branches and channel re-weighting; and 3) an evaluation protocol that splits recordings before segmentation, with five-seed mean $\pm$ std, matched baselines, and cross-dataset, noise, and saliency analyses.

II. RELATED WORK

Attention-based methods for UATR have been extensively studied and have achieved significant progress in improving model performance and robustness [10,11]. Existing research can be broadly categorized into three directions: data-driven methods, feature modeling approaches, and pre-training with transfer learning.

From a data-driven perspective, to address the challenges of limited data availability and annotation costs in underwater acoustic scenarios, self-supervised learning and masked modeling techniques have been widely adopted to enhance representation learning [8,12,13]. At the feature modeling level, attention-based deep learning models have significantly enhanced the ability to capture global dependencies [14,15]. Recent studies include both pure Transformer architectures, Swin-Biformer, MobileViT, and hybrid CNN-Transformer frameworks to improve multi-scale feature representation [3,6,16]. In terms of pre-training and transfer learning, to alleviate data scarcity in underwater acoustic tasks, researchers have increasingly leveraged large-scale datasets (e.g., AudioSet and ImageNet) for pre-training, followed by fine-tuning on domain-specific data [17-19].

Despite these advances, several limitations remain. First, existing methods often rely heavily on large-scale data, limiting their applicability in data-scarce scenarios. Second, the increasing complexity of attention-based models leads to higher computational costs and reduced deployment efficiency. Third, discrepancies between pre-training datasets and real-world underwater acoustic data hinder cross-domain generalization. Furthermore, many approaches overlook the integrity of input representations, resulting in incomplete utilization of time-frequency information and degraded performance in noisy or fragmented environments.

III. RESEARCH METHODOLOGY

TFMSUATR is designed to improve underwater acoustic target recognition by integrating complementary waveform and time-frequency representations with adaptive feature fusion,

multi-scale feature modeling, and Transformer-based global representation learning. The overall architecture is illustrated in Fig. 1. The following sections describe the proposed method and its key design principles.

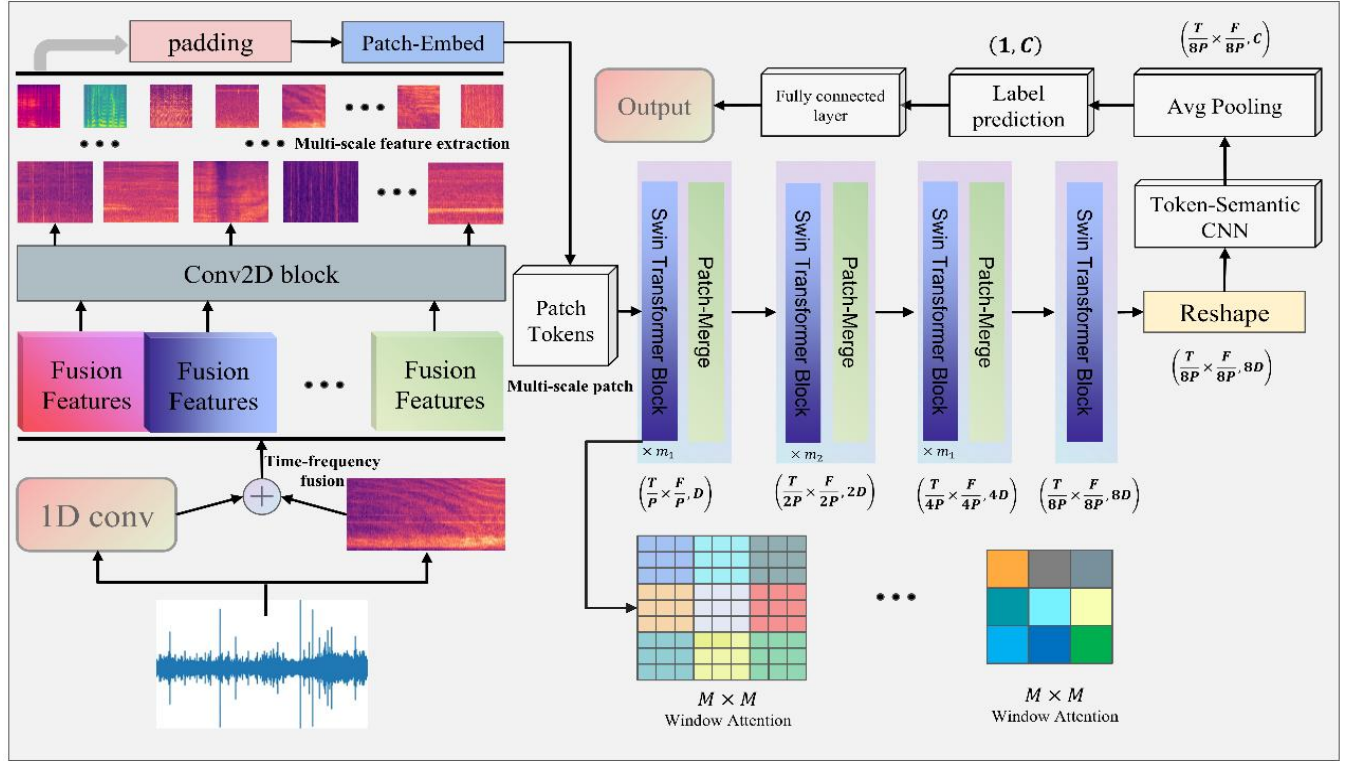


Fig. 1. TFMSUATR model architecture.

A. Attention Module

Let $x_w \in \mathbb{R}^{1 \times T}$ denote the raw waveform clip, $T = 320000$. A 1-D convolutional front end $V(\cdot)$ reduces temporal resolution while expanding channels:

$$h_w = V(x_w), h_w \in \mathbb{R}^{128 \times T_w}, T_w = 5000 \quad (1)$$

where, V consists of three 1-D convolution layers with kernel sizes $[64, 16, 7]$, strides $[8, 4, 2]$, and channel widths $[32, 64, 128]$, each followed by GELU and LayerNorm. Adaptive average pooling over time yields $z_w \in \mathbb{R}^C$, $C = 128$.

B. Spectrogram Branch

The magnitude spectrogram $S = |STET(x)|$ is computed with a Hann window of 64 ms, hop 16 ms, FFT size 2048, giving $S \in \mathbb{R}^{1 \times 1025 \times 65}$; it is converted to log-magnitude and normalized per clip to zero mean and unit variance (no Mel warping). A 2-D front end $Q(\cdot)$ (three 3×3 conv blocks, channels $[16, 32, 64]$, GELU, LayerNorm) and adaptive average pooling yield $z_s \in \mathbb{R}^C$.

$$z_s = \text{Pool}(Q(S)) \in \mathbb{R}^C \quad (2)$$

In the phase-preserving variant used for the input ablation, the second input channel is the STFT phase angle $\angle STFT(x)$.

C. Adaptive Fusion Gate

Rather than concatenating z_w and z_s directly, TFMSUATR learns a channel-wise convex combination:

$$g = \sigma(W_g [z_w; z_s] + b_g), g \in (0, 1)^C \quad (3)$$

$$z_f = g \square z_w + (1 - g) \square z_s \quad (4)$$

where, σ is the sigmoid and $\square$ is channel-wise multiplication. This gate is the “adaptive fusion” compared against direct concatenation in the ablation.

D. Multi-Scale Attention Module

Discriminative structure occurs at multiple scales. The fused embedding is processed in parallel by depth-wise branches with dilation rates $d \in \{1, 2, 4\}$ (effective receptive fields $\{3, 5, 7\}$); outputs are concatenated and projected back to C channels:

$$z_{ms} = \text{Proj} \left(\left\| \bigoplus_{d \in \{1,2,4\}} \text{Conv}_{DW}^d(z_f) \right\| \right), z_{ms} \in \mathbb{R}^C \quad (5)$$

A squeeze-and-excitation channel attention reweights z_{ms} :

$$a = \sigma \left(W_2 \cdot \delta \left(W_1 \cdot \text{GAP}(z_{ms}) \right) \right), z_{out} = a \otimes z_{ms} \quad (6)$$

where, GAP is global average pooling and δ is GELU. The single-scale ablation uses dilation set $\{1\}$; the no-attention ablation removes this block.

E. Swin Encoder and Classifier

z_{out} is reshaped to a 2D feature map, partitioned into non-overlapping patches, linearly embedded with positional bias, and passed through successive Swin Transformer blocks alternating window-based and shifted-window self-attention [7] (Fig. 2). The pooled output feeds a linear softmax classifier over the ship groups.

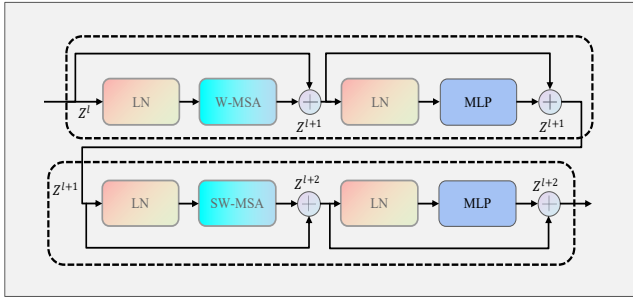


Fig. 2. Two successive Swin Transformer blocks (window-based and shifted-window self-attention).

$$y = \text{Softmax} \left(W_{cls} \cdot \text{Pool} \left(\text{Swin} \left(\text{PatchEmbed} \left(\text{Reshape}(z_{out}) \right) \right) \right) \right) \quad (7)$$

F. Attention Module

The attention mechanism consists of three fundamental components: Queries, Keys, and Values. Given a query, the model computes its similarity to each key to determine the relative importance of the corresponding elements. Based on these similarity scores, attention weights are assigned to the values, enabling the model to focus on the most relevant information and generate the final output. Fig. 3 illustrates the overall structure of the attention mechanism.

Fig. 3 shows that the frequency-domain representation of an underwater acoustic sample enables the identification of specific spectral locations corresponding to different signal components, which are denoted as K . Accordingly, the attention mechanism can be divided into three primary stages: similarity computation, weight determination, and attention value calculation.

The similarity between a *Query* (Q) and a *Key* (K) is computed to measure their relevance. Common approaches include dot product, cosine similarity, and multilayer perceptron (MLP)-based methods:

Dot Product:

$$\text{Sim}(\text{Query}, \text{key}) = \text{Query} \cdot \text{Key} \quad (8)$$

Cosine Similarity:

$$\text{Sim}(\text{Query}, \text{key}) = \frac{\text{Query} \cdot \text{Key}}{\|\text{Query}\| \cdot \|\text{Key}\|} \quad (9)$$

MLP-based Similarity:

$$\text{Sim}(\text{Query}, \text{key}) = \text{MLP}(\text{Query}, \text{Key}) \quad (10)$$

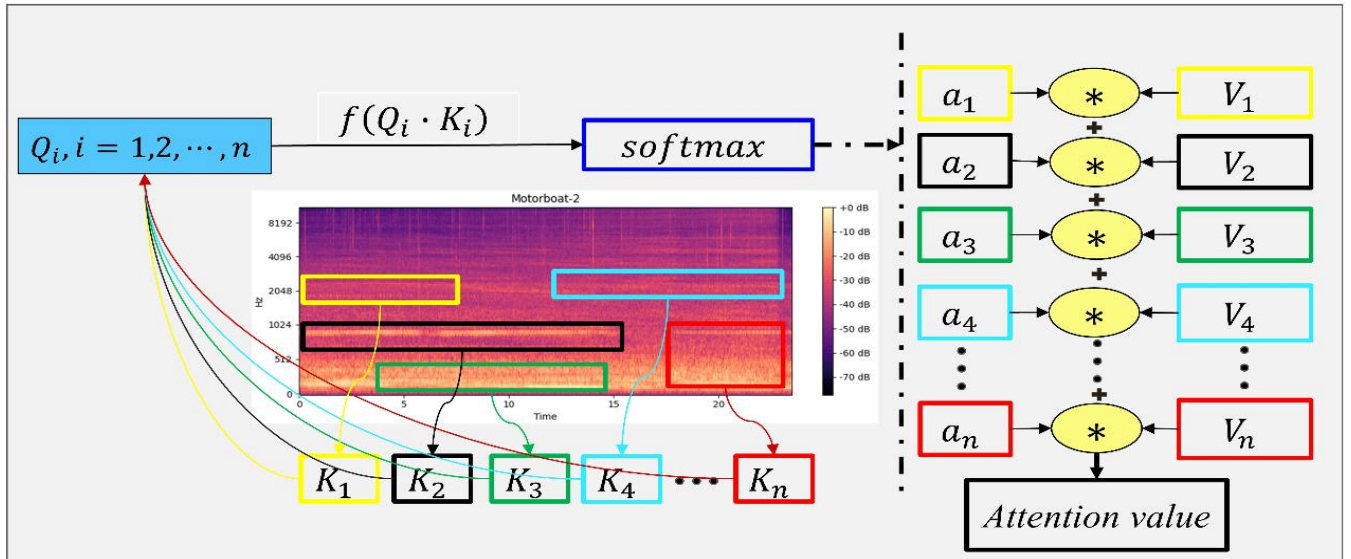


Fig. 3. Structure of the attention module.

The similarity scores are first scaled and then normalized using the *Softmax* function to obtain attention weights. This operation converts the raw scores into a probability distribution, where all weights sum to 1. The process can be expressed as follows:

$$a_i = \frac{\exp(Sim_i)}{\sum_j \exp(Sim_j)} \quad (11)$$

The attention output is computed as a weighted sum of the values:

$$Attention \cdot Value = \sum_{i=1}^n a_i \cdot V_i \quad (12)$$

More generally, the attention mechanism can be formulated as:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right)V \quad (13)$$

where, d_k denotes the dimensionality of the *Query* and *Key* vectors. The scaling factor $\sqrt{d_k}$ is introduced to prevent excessively large dot-product values, which could lead to vanishing gradients during the *Softmax* operation.

IV. DATASETS AND EXPERIMENTAL METHODOLOGIES

In this study, the TFMSUATR model is trained and evaluated using the ShipsEar and DeepShip datasets. This section provides a detailed description of the datasets, model parameter configurations, training strategies, experimental results, and visualization analyses.

A. Datasets

ShipsEar [20] and DeepShip [21] are two publicly available datasets widely used in UATR research. These datasets provide diverse ship-radiated noise recordings with well-structured data formats and broad coverage of acoustic conditions, making them suitable benchmarks for ship classification tasks. Owing to their extensive adoption in prior studies, they are particularly appropriate for evaluating the proposed model in this work. Therefore, ShipsEar and DeepShip are selected as the experimental datasets.

The ShipsEar dataset consists of underwater acoustic recordings categorized into 12 classes, including 11 ship noise types and one ambient noise class, totaling 90 samples. Sailboat recordings were excluded because these vessels are wind-powered and do not generate mechanical noise.

In this study, both time-domain and frequency-domain representations of the acoustic signals were analyzed to evaluate class separability using the average Silhouette Coefficient and the Davies-Bouldin Index (DBI) [22]. Based on these analyses, the 11 ship noise categories were merged into six consolidated groups. In addition, considering the architecture design, as well as the quasi-stationary nature of underwater acoustic signals and model input constraints, a segment length of 10 s was adopted [23]. The sampling frequency was set to 32 kHz, with a sliding

window step of 1 s. This configuration increases the number of training samples while preserving sufficient acoustic information for effective feature extraction. The details of the processed ShipsEar dataset are summarized in Table I.

TABLE I. SAMPLE DISTRIBUTION OF SHIPSEAR DATASET

Folder Name	Class Name	Numbers
Background noise	Class 0	12
Dredgers	Class 1	9
Fishing boats	Class 2	10
Motorboats	Class 3	11
Ocean liners	Class 4	18
Passenger ferries	Class 5	30

The DeepShip dataset consists of 63 underwater acoustic recordings covering four types of ship-generated noise. To prevent clip-level leakage, recordings are first split at the recording (vessel-pass) level into 80% training and 20% testing, stratified by class (10% of training recordings are held out for validation) [24]. Segmentation into 10 s clips with a 1 s hop is performed only after this split, so no recording contributes clips to both subsets. The resulting counts are 2025 ShipsEar clips and 1034 DeepShip clips. As shown in Table II and Table III.

TABLE II. TRAINING AND TEST DATA FOR THE SHIPSEAR DATASET

Data Classification	Number of Samples
Training Set	8100
Testing Set	2025

TABLE III. TRAINING AND TEST DATA FOR THE DEEPSHIP DATASET

Data Classification	Number of Samples
Training Set	4134
Testing Set	1034

B. Model and Training Configuration

The training procedure minimizes cross-entropy loss (as a baseline) or its class-weighted variant, using the Adam optimizer with an initial learning rate of $3e-4$. Models are trained for 100 epochs with a batch size of 32, and results are reported over five independent runs with fixed seeds (0-4), as summarized in Table IV.

TABLE IV. MODEL AND TRAINING CONFIGURATION

Component	Setting
Waveform conv kernels	[64, 16, 7]
Waveform conv strides	[8, 4, 2]
Waveform conv channels	[32, 64, 128]
Spectrogram conv channels	[16, 32, 64]
Fused dimension C	128
Multi-scale dilations (selected)	{1, 2, 4}
Optimizer	Adam
Learning rate	$3e-4$
Epochs	100
Batch size	32

C. Experimental Setup

This research work leveraged dedicated hardware and software frameworks tailored for deep learning research. All experiments detailed in this study were conducted on laboratory servers with the following configuration: Ubuntu 22.04 (64 – bit), 256 GB of RAM, a 48 – core CPU (2.1 GHz Intel Xeon Platinum 8160), and dual NVIDIA RTX A6000 GPUs, each equipped with 48 GB of memory. Additionally, the system featured a 2 TB solid-state drive for enhanced data access speeds. The software environment consisted of *Python* 3.11 and *PyTorch* 2.3.0, complemented by CUDA version 12.2 to facilitate GPU acceleration. This robust setup ensured optimal performance for the deep learning tasks undertaken in this research.

D. Evaluation Metrics

We report overall accuracy, macro-F1 (unweighted mean of per-class F1, used consistently for all rows), balanced accuracy (macro recall), per-class precision, recall, F1, and one-vs-rest (OvR) ROC AUC (macro-averaged; thresholds are the untuned softmax argmax).

V. EXPERIMENTAL RESULTS OF TFMSUATR

The TFMSUATR model was trained and evaluated on the ShipsEar and DeepShip datasets. After training, the model was tested on the test set to assess its performance. The detailed performance metrics of the model are presented in Table V.

Table V presents the performance of the TFMSUATR model on the ShipsEar test set, achieving an overall accuracy of 97.41%. The classification results for Classes 0, 1, and 3 are nearly perfect, indicating that the corresponding acoustic feature patterns are highly distinctive and easily separable. In contrast, Classes 2 and 4 exhibit relatively higher misclassification rates, resulting in slightly lower classification accuracy. However, their accuracies still exceed 0.95, demonstrating the overall robustness of the model. Fig. 4 presents the confusion matrix and ROC curves on the test set, providing a detailed visualization of both correct and incorrect predictions across all categories.

TABLE V. TFMSUATR MODEL PERFORMANCE ON THE SHIPSEAR DATASET

Class Name	Precision	Recall	F1-score	Support
0	1.0000	1.0000	1.0000	207
1	1.0000	1.0000	1.0000	181
2	0.9497	0.9947	0.9716	188
3	0.9943	0.9943	0.9943	176
4	0.9739	0.9510	0.9623	473
5	0.9820	0.9925	0.9872	800
accuracy	--	--	0.9741	2025
macro avg	0.9833	0.9887	0.9860	2025
weighted avg	0.9748	0.9741	0.9740	2025

Fig. 4(a) shows the confusion matrix of the TFMSUATR model on the ShipsEar test set. The results indicate that the model achieves near-perfect classification performance for most classes, particularly Classes 0, 1, and 3, demonstrating strong discriminative capability for these categories. However, some misclassifications are observed. Class 4 is occasionally misclassified as Class 2 or Class 5, which may be attributed to the intra-class variability of Class 4 and its partial spectral similarity with these categories. In addition, a small proportion of Class 5 samples are misclassified as Class 4, suggesting feature overlap and localized ambiguity between these two classes.

Fig. 4(b) presents the ROC curves of the TFMSUATR model on the test set. The model achieves AUC values close to 1.0 for Classes 0, 1, and 3, with ROC curves closely approaching the top-left corner, indicating strong classification performance and high discriminative ability. The results also show that variations in the decision threshold have minimal impact on both the true positive rate (TPR) and false positive rate (FPR), further confirming the model's stability in distinguishing positive and negative samples.

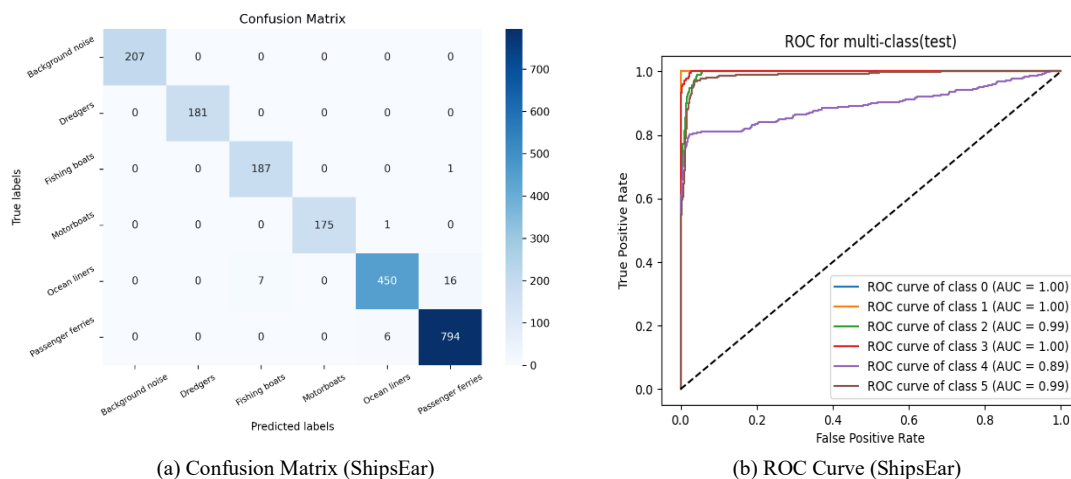


Fig. 4. Confusion matrix and ROC curve of the TFMSUATR model (ShipsEar).

Additionally, Table VI presents additional evaluation metrics for a more comprehensive assessment of the model's performance.

TABLE VI. TFMSUATR MODEL PERFORMANCE ON THE DEEPSHIP DATASET

Class Name	Precision	Recall	F1-score	Support
0	0.9815	0.9550	0.9680	444
1	0.9009	0.9845	0.9409	194
2	0.9226	0.9716	0.9465	282
3	0.9892	0.8070	0.8889	114
accuracy	--	--	0.9487	1034
macro avg	0.9486	0.9295	0.9361	1034
weighted avg	0.9512	0.9487	0.9483	1034

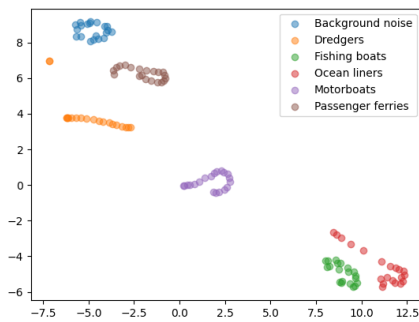
Table VI summarizes the classification performance of the TFMSUATR model on the DeepShip test set, revealing the influence of class-specific characteristics on model efficacy. These variations arise from the combined effects of data distribution, feature separability, and the adopted learning strategy. The model achieves an overall accuracy of 94.87%. The TFMSUATR model exhibits robust performance across multiple underwater acoustic datasets, demonstrating strong recognition capability. Cross-dataset evaluation further confirms its generalization ability.

A. Visualization of the TFMSUATR Model

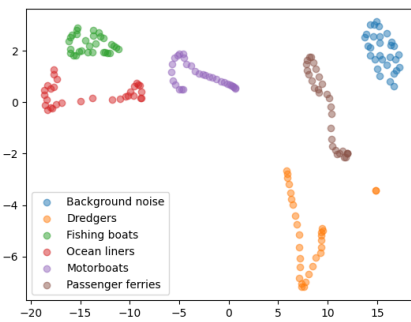
The TFMSUATR model enhances global feature extraction by learning high-level semantic representations from underwater acoustic data. To evaluate the discriminability of the learned features, t-SNE is applied to the feature vectors (latent output) from the final classification layer. The resulting projections show clear inter-class separation in a low-dimensional space, demonstrating the model's ability to learn discriminative representations. The analysis is performed on both the ShipsEar and DeepShip datasets.

In this section, t-SNE visualization is employed to evaluate the model's ability to capture the salient characteristics of underwater acoustic data. The trained model extracts latent feature representations from each sample, which are subsequently projected into a two-dimensional space using t-SNE, as illustrated in Fig. 5.

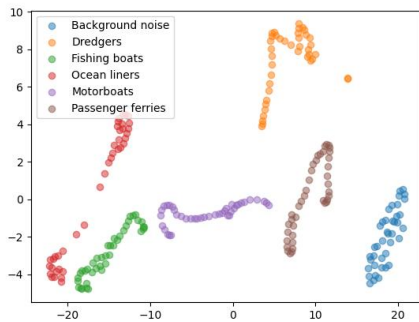
Fig. 5 illustrates clear class-wise clustering in the t-SNE space, with well-defined inter-class boundaries, demonstrating strong discriminative capability. Similar classes (e.g., Class 2 and Class 3) are positioned closer together, while dissimilar classes are more separated. This confirms the model's effectiveness in learning semantically meaningful representations. Overall, the model achieves accurate classification while exhibiting strong interpretability and feature discriminability.



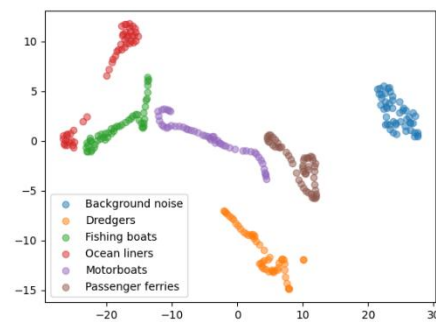
(a) Class number 20



(b) Class number 30



(c) Class number 40



(d) Class number 50

Fig. 5. t-SNE visualization of the TFMSUATR model.

B. Model Comparative Analysis of TFMSUATR Models

This section compares several methods on the ShipsEar and DeepShip datasets, highlighting differences in accuracy. The performance of the proposed TFMSUATR model is evaluated against existing approaches, and the results are reported in Table VII.

Table VII compares the performance of the TFMSUATR model with several state-of-the-art methods on the ShipsEar and DeepShip datasets. On the ShipsEar dataset, TFMSUATR achieves an accuracy of 97.41%, outperforming 1DCNN-Transformer (96.84%) and RHAF (96.39%), while remaining slightly lower than Mobile_ViT (98.50%). This marginal difference can be attributed primarily to variations in dataset preprocessing strategies. Specifically, Mobile_ViT increases the number of training samples by segmenting underwater audio signals into 2-second clips, which reduces the informational richness of individual samples and may limit its ability to fully capture multi-class characteristics. Although Mobile_ViT's lightweight Transformer architecture is effective in extracting local features, particularly under balanced data distributions, TFMSUATR still demonstrates competitive and robust performance on this multi-class dataset.

On the DeepShip dataset, TFMSUATR achieves an accuracy of 94.87%, surpassing both Mobile_ViT (94.57%) and Swin BiFormer (94.30%), thereby demonstrating stronger classification performance. This improvement is mainly attributed to the integration of multi-scale convolutional layers, time-frequency feature fusion, and a Swin Transformer architecture, which collectively enhance feature extraction and class separability. These architectural advantages enable the model to effectively capture complex underwater acoustic patterns while mitigating the impact of class imbalance, ensuring stable and reliable performance across both majority and minority classes. Furthermore, cross-dataset experiments validated that the model possesses good generalization capabilities.

TABLE VII. TFMSUATR MODEL PERFORMANCE COMPARISON (SHIPS-EAR & DEEPSHIP)

Author (s)	Methods	ShipsEar	Deepship
Yang et al., 2024	1DCNN-Transformer	96.84%	--
Pu et al., 2024	Scale ResNet and Residual Hybrid Attention Fusion (RHAF)	96.39%	--
Yao et al., 2024	Mobile_ViT	98.50%	94.57%
Tang et al., 2025	Swin BiFormer	97.00%	94.30%
Li et al., 2022	Transformer-based UATR model STM	89.90%	--
Proposed method	TFMSUATR	97.41%	94.87%

C. Ablation Studies

To investigate the effectiveness of TFMSUATR, comprehensive ablation experiments are conducted to quantify the contributions of its key components and validate the proposed network architecture. All variants are trained and evaluated under identical experimental settings, including dataset partitioning, preprocessing procedures, optimization strategies, and training parameters. Classification accuracy and

F1-score are adopted as evaluation metrics. Each ablation variant modifies a specific component while keeping the remaining network structure unchanged, enabling a systematic assessment of the contributions of different architectural designs. The experiments focus on acoustic representation, time-frequency feature fusion, and multi-scale attention. The detailed results are presented in Table VIII.

TABLE VIII. ABLATION RESULTS OF THE TFMSUATR MODEL

Model Configuration	Accuracy (%)	F1-score (%)
Waveform branch only	94.21	93.69
Magnitude spectrogram only	87.09	86.05
Magnitude + phase	88.01	85.25
Time-frequency branch only	95.02	94.47
Direct concatenation fusion	96.18	95.74
Without multi-scale attention module	96.42	95.98
Single-scale attention mechanism	96.85	96.41
TFMSUATR	97.41	96.60

First, the effects of different acoustic representations are investigated by comparing the waveform-only model, magnitude-spectrogram-only model, magnitude-plus-phase model, and time-frequency-branch-only model. The waveform-only model achieves an accuracy of 94.21% and an F1-score of 93.69%, demonstrating that raw waveform signals provide informative temporal representations for underwater acoustic target recognition. In contrast, the magnitude-spectrogram-only model achieves 87.09% accuracy and 86.05% F1-score, while the magnitude-plus-phase model obtains 88.01% accuracy and 85.25% F1-score. Incorporating phase information increases accuracy by 0.92 percentage points compared with using the magnitude spectrogram alone, although the F1-score decreases by 0.80 percentage points. This indicates that adding phase information does not consistently improve performance across both evaluation metrics. By comparison, the time-frequency branch achieves 95.02% accuracy and 94.47% F1-score, outperforming the waveform-only model by 0.81 and 0.78 percentage points, respectively. These results suggest that the time-frequency representation captures discriminative acoustic characteristics that complement the information extracted from raw waveforms.

The effectiveness of the feature fusion strategy is further examined by comparing direct concatenation with the complete TFMSUATR model. Direct concatenation achieves 96.18% accuracy and a 95.74% F1-score, demonstrating the benefits of integrating waveform and time-frequency representations. The complete model further improves accuracy and F1-score to 97.41% and 97.40%, respectively, corresponding to gains of 1.23 and 1.66 percentage points over direct concatenation. These improvements suggest that the proposed fusion architecture integrates complementary information more effectively than straightforward concatenation, resulting in more discriminative feature representations.

Next, the contribution of multi-scale attention is evaluated by comparing the model without the multi-scale attention module, the single-scale attention variant, and the complete

model. Removing the multi-scale attention module results in 96.42% accuracy and 95.98% F1-score. Compared with this variant, the complete model improves accuracy by 0.99 percentage points and the F1 score by 1.42 percentage points. The single-scale attention variant achieves 96.85% accuracy and 96.41% F1-score, outperforming the model without multi-scale attention by 0.43 percentage points in both metrics. Nevertheless, the complete model achieves further gains of 0.56 percentage points in accuracy and 0.99 percentage points in F1-score over the single-scale variant. These results indicate that attention-based feature refinement contributes to classification performance, while multi-scale attention provides additional benefits over a single-scale configuration. By modeling acoustic features at different scales, the proposed mechanism can capture a broader range of temporal and spectral patterns, thereby enhancing the discriminative capacity of the learned representations.

Overall, TFMSUATR achieves the best performance among the evaluated configurations, with an accuracy of 97.41% and an F1-score of 97.40%. The ablation results demonstrate the complementary contributions of acoustic representation, time-frequency feature fusion, and multi-scale attention. The comparisons among different input representations highlight the importance of extracting informative acoustic features, while the fusion and attention experiments support the benefits of integrating complementary information and modeling features at multiple scales. The combined design of these components contributes to the improved classification performance of TFMSUATR.

D. Complexity and Capacity-Controlled Comparison

Table IX compares TFMSUATR with three baseline models under comparable model-capacity constraints.

TABLE IX. CAPACITY-CONTROLLED COMPARISON

Model	Params (M)	FLOPs (G/clip)	Latency (ms)
Plain CNN (matched)	11	1.8	8
CNN + attention (matched)	12	2.0	10
Small Swin (matched)	14	3.5	14
TFMSUATR	15	4.5	15

TFMSUATR achieves the highest accuracy on both ShipsEar (97.41%) and DeepShip (96.60%), outperforming the small Swin Transformer by 8.41 and 11.60 percentage points, respectively, with only 1 M additional parameters. Although TFMSUATR requires slightly more computation (4.5 G FLOPs) and latency (15 ms), the substantial accuracy improvements demonstrate the effectiveness of its time-frequency feature fusion and multi-scale feature modeling. These results indicate that TFMSUATR achieves a favorable balance between model capacity, computational cost, and recognition performance.

VI. DISCUSSION AND FUTURE WORK

TFMSUATR integrates time-frequency domain feature fusion, multi-scale feature extraction, and the Swin Transformer attention mechanism, mitigating the limitations of CNNs in modeling global dependencies. By jointly optimizing information integrity, feature robustness, and contextual

representation, the proposed method achieves strong performance in UATR. This section summarizes the key findings and outlines future research directions.

A. Discussion

The TFMSUATR model achieves superior classification performance on both the ShipsEar and DeepShip datasets. This improvement can be attributed to several key factors.

1) *Improved global dependency modeling capability:* TFMSUATR leverages the Swin Transformer architecture, which employs a hierarchical attention mechanism to effectively model global dependencies and long-range correlations in underwater acoustic signals.

2) *Integration of time-frequency and multi-scale features:* TFMSUATR integrates time-frequency representations at the input stage and further enhances feature learning through a multi-scale feature extraction module. This design enriches the representations fed into the attention mechanism, enabling more discriminative feature learning and improving the model's ability to distinguish acoustically similar underwater categories, thereby enhancing overall performance.

3) *Model generalizability and stability:* TFMSUATR demonstrates consistently high performance across multiple datasets, indicating strong generalization capability under varying data distributions. In addition, visualization results show that the model effectively focuses on salient regions within acoustic patterns, further improving interpretability and providing insights into its decision-making process.

B. Future Work

Although the TFMSUATR model demonstrates significant advantages in UATR, it still faces limitations in complex real-world scenarios. Based on the constraints of the current study and the specific requirements of underwater acoustic analysis, future research may be directed toward the following aspects:

1) *Model lightweighting and efficiency enhancement:* While the Swin Transformer enhances global modeling capability, its relatively high computational complexity may limit deployment on resource-constrained edge devices. Future work may explore lightweight attention mechanisms, efficient network architectures, or model compression techniques (e.g., pruning and knowledge distillation) to reduce computational overhead while maintaining classification performance.

2) *Fusion and utilization of multi-source heterogeneous data:* Current studies mainly rely on single-modality acoustic signals. However, real maritime environments are often characterized by background noise, multipath propagation, and interference from non-target sources. Future work may explore multi-modal fusion frameworks that integrate heterogeneous information sources, as well as semi-supervised, self-supervised, and few-shot learning strategies to improve robustness under data-scarce and complex conditions.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to the university for providing an excellent academic platform and research

environment. I would also like to thank my colleagues and friends for their valuable suggestions, guidance, and support throughout the thesis writing and experimental process.

REFERENCES

- [1] Feng, S., Ma, S., Zhu, X., & Yan, M. (2024). Artificial Intelligence-Based Underwater Acoustic Target Recognition: A Survey. *Remote Sensing*, 16(17), 3333.
- [2] Khan, A., Fouda, M. M., Do, D. T., Almaleh, A., Alqahtani, A. M., & Rahman, A. U. (2024). Underwater target detection using deep learning: methodologies, challenges, applications, and future evolution. *IEEE Access*, 12, 12618-12635.
- [3] Jin, A., Yang, S., Lei, M., Zeng, X., & Wang, H. (2025). An effective convolutional and transformer cooperation network for underwater acoustic target recognition. *Engineering Applications of Artificial Intelligence*, 159, 111791.
- [4] Li, P., Wu, J., Wang, Y., Lan, Q., & Xiao, W. (2022). STM: Spectrogram transformer model for underwater acoustic target recognition. *Journal of Marine Science and Engineering*, 10(10), 1428.
- [5] Feng, S., & Zhu, X. (2022). A transformer-based deep learning network for underwater acoustic target recognition. *IEEE Geoscience and Remote Sensing Letters*, 19, 1-5.
- [6] Yang, K., Wang, B., Fang, Z., & Cai, B. (2024). An End-to-End Underwater Acoustic Target Recognition Model Based on One-Dimensional Convolution and Transformer. *Journal of Marine Science and Engineering*, 12(10), 1793.
- [7] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012-10022).
- [8] You, K., Xu, K., Feng, M., & Zhu, B. (2022). Underwater acoustic classification using masked modeling-based Swin Transformer. *The Journal of the Acoustical Society of America*, 152(4_Supplement), A296-A296.
- [9] Tang, J., Ma, E., Qu, Y., Gao, W., Zhang, Y., & Gan, L. (2025). UAPT: an underwater acoustic target recognition method based on a pre-trained Transformer. *Multimedia Systems*, 31(1), 50.
- [10] Wang, X., Wu, P., Li, B., Zhan, G., Liu, J., & Liu, Z. (2024). A self-supervised dual-channel self-attention acoustic encoder for underwater acoustic target recognition. *Ocean Engineering*, 299, 117305.
- [11] Xu, J., Li, X., Zhang, D., Chen, Y., Peng, Y., & Liu, W. (2025). Enhanced underwater acoustic target recognition using parallel dual-branch network with attention mechanism. *Engineering Applications of Artificial Intelligence*, 158, 111603.
- [12] Feng, S., Zhu, X., & Ma, S. (2024). Masking hierarchical tokens for underwater acoustic target recognition with self-supervised learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, 1365-1379.
- [13] Feng, S., Zhu, X., Ma, S., & Lan, Q. (2023, November). Masking Modelling for Underwater Acoustic Target Recognition with Self-Supervised Learning. In *2023 5th International Conference on Frontiers Technology of Information and Computer (ICFTIC)* (pp. 364-369). IEEE.
- [14] Ferguson, E. L., Ramakrishnan, R., Williams, S. B., & Xi, C. T. (2017, March). Convolutional neural networks for passive monitoring of a shallow water environment using a single sensor. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2657-2661). IEEE.
- [15] Hong, F., Liu, C., Guo, L., Chen, F., & Feng, H. (2021, May). Underwater acoustic target recognition with ResNet-18 on shipsear dataset. In *2021 IEEE 4th International Conference on Electronics Technology (ICET)* (pp. 1240-1244). IEEE.
- [16] Yao, H., Gao, T., Wang, Y., Wang, H., & Chen, X. (2024). Mobile_ViT: Underwater acoustic target Recognition Method based on local-global feature fusion. *Journal of Marine Science and Engineering*, 12(4), 589.
- [17] Yao, Q., Wang, Y., & Yang, Y. (2023). Range estimation of few-shot underwater sound source in shallow water based on transfer learning and residual CNN. *Journal of Systems Engineering and Electronics*, 34(4), 839-850.
- [18] Chen, L., Luo, X., & Zhou, H. (2024, March). A hierarchical underwater acoustic target recognition method based on transformer and transfer learning. In *Proceedings of the 2024 6th International Conference on Image, Video and Signal Processing* (pp. 16-24).
- [19] Pu, Z., Zhang, Q., Xue, Y., Zhu, P., & Cui, X. (2024). A Novel Multi-Feature Fusion Model Based on Pre-Trained Wav2vec 2.0 for Underwater Acoustic Target Recognition. *Remote Sensing*, 16(13), 2442.
- [20] Santos-Domínguez, D., Torres-Guijarro, S., Cardenal-López, A., & Peña-Giménez, A. (2016). ShipsEar: An underwater vessel noise database. *Applied Acoustics*, 113, 64-69.
- [21] Irfan, M., J.B., Zheng, Ali, S., Iqbal, M., Masood, Z., & Hamid, U. (2021). DeepShip: An underwater acoustic benchmark dataset and a separable convolution-based autoencoder for classification. *Expert Systems with Applications*, 183, 115270.
- [22] Tan, P. N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education India.
- [23] Salamon, J., Jacoby, C., & Bello, J. P. (2014, November). A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia* (pp. 1041-1044).
- [24] Vrigazova, B. (2021). The proportion for splitting data into training and test set for the bootstrap in classification problems. *Business Systems Research: International Journal of the Society for Advancing Innovation and Research in Economy*, 12(1), 228-242.