SpecSlice-ViT: Preserving Spectral Integrity via Slicing Transformers for Underwater Acoustic Target Recognition

Ruiting Sun, Zhangjie Cai, Zhenhong Liao, Guanwen Zhang*, and Wei Zhou
School of Electronics and Information, Northwestern Polytechnical University
Xi'an, Shaanxi, 710129, China
{sunrt, czj, liaozh}@mail.nwpu.edu.cn, {*guanwen.zh, zhouwei}@nwpu.edu.cn

***Abstract*—To address the challenges of non-stationary signals and the disruption of spectral continuity by standard patch-based Transformers in underwater acoustic target recognition, we propose SpecSlice-ViT. This novel architecture preserves spectral integrity via an acoustic-aware Feature Slicing strategy that slices completely along the frequency dimension to maintain harmonic structure. We introduce a dual-view fusion mechanism, combining fine-grained Short-Time Fourier Transform (STFT) details with robust Mel-Frequency Cepstral Coefficients (MFCC) envelopes through cross-attention and adaptive gating. Additionally, a Hierarchical Encoder cascades window attention with token compression to efficiently model multi-scale temporal dependencies. Experimental results on the DeepShip dataset demonstrate that SpecSlice-ViT outperforms existing Convolutional Neural Networks and Transformer baselines, achieving competitive accuracy and efficient inference speeds, indicating potential for future deployment on underwater platforms.**

***Index Terms*—Underwater Acoustic Target Recognition, Vision Transformer, Spectral Integrity, Feature Slicing, Dual-View Fusion**

I. Introduction

Underwater acoustic target recognition (UATR) is a cornerstone technology for marine environmental monitoring and maritime defense. Unlike optical imaging, underwater acoustic signals captured by passive sonar are inherently non-stationary. They are heavily contaminated by ambient noise, multipath propagation, and the Doppler effect. These hydroacoustic complexities lead to severe feature aliasing and low Signal-to-Noise Ratios [1]. Existing approaches struggle to balance computational efficiency with the need to capture global dependencies. Convolutional Neural Networks are limited by local receptive fields [2]. Standard Vision Transformers (ViTs) [3] suffer from high computational redundancy when processing high-resolution spectrograms. Crucially, the direct application of patch-based tokenization disrupts the spectral continuity of acoustic signals [4]. Spectrograms differ fundamentally from natural images as the frequency dimension carries specific physical meanings. Randomly patching the spectrogram breaks the semantic consistency of frequency distributions within a time frame [5].

*Corresponding Author: Guanwen Zhang
This work was supported by the National Major Science and Technology Projects of China under Grant 2025ZD1401102.

To address these challenges, we introduce the concept of preserving spectral integrity. We propose a conflict resolution strategy named SpecSlice-ViT. This framework is designed to treat the frequency dimension as a unified entity rather than fragmenting it. By doing so, we aim to transform the recognition task into a robust temporal sequence modeling problem. We further combine this acoustic-aware design with window attention and token compression to ensure the global modeling is computationally affordable. Our framework also leverages the complementarity of heterogeneous acoustic features. We utilize Short-Time Fourier Transform and Mel-Frequency Cepstral Coefficients spectrograms. These features are combined to integrate granular spectral details with robust spectral envelopes. Finally, we propose **SpecSlice-ViT** (**Spec**tral **Slice Vi**sion **T**ransformer), a hybrid framework that synergizes acoustic-aware slicing with dual-view fusion. As illustrated in Fig. 1, our key contributions are summarized as follows:

- **Acoustic-Aware Slicing Tokenization:** We introduce a slicing strategy tailored to spectrograms. It replaces two-dimensional patching with full-frequency slicing. This preserves the complete harmonic structure within each time frame and stabilizes token semantics.
- **Dual-View Complementary Fusion:** We propose a fusion module for STFT and MFCC features. A cross-attention mechanism aligns the robust envelope of MFCCs with the fine-grained details of STFTs. A gating network adaptively weights the features to suppress environmental noise.
- **Hierarchical Efficient Encoder:** We devise an Encoder that cascades window attention with token compression. Window attention captures local transient details at a low computational cost. Token compression significantly reduces sequence length for efficient global modeling.

Extensive experiments on the DeepShip dataset [6] demonstrate the effectiveness of our method. SpecSlice-ViT achieves improved accuracy and high computational efficiency compared to state-of-the-art baselines.

II. Related Work

This section reviews advancements in underwater acoustic target recognition, specifically focusing on spectral feature

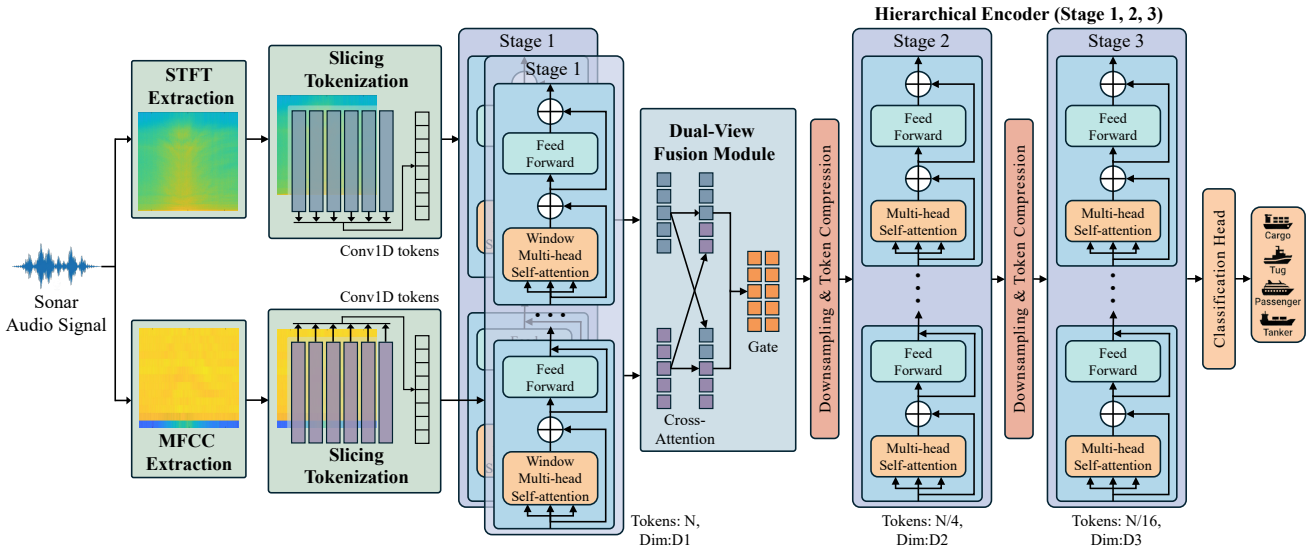

Fig. 1. The overall architecture of SpecSlice-ViT. Illustrates the three modules: slicing, fusion, hierarchical encoder. Stage 1 processes two views independently; Stages 2 and 3 are shared after fusion.

integrity, tokenization strategies, and efficient attention mechanisms.

A. Complementarity in Spectral Feature Extraction

Feature extraction is critical for reducing redundancy in raw hydroacoustic signals. Early research focused on physical attributes, utilizing methods like Low-Frequency Analysis Recording (LOFAR) [7] and Detection of Envelope Modulation on Noise (DEMON) [8] spectra to identify line spectra and propeller shaft frequencies. However, these features often struggle to characterize non-stationary signals in complex marine environments [2]. Consequently, time-frequency representations have gained prominence. The Short-Time Fourier Transform (STFT) is widely adopted to capture fine-grained transient details [9], as illustrated in Fig. 2, though it remains sensitive to ambient noise. Alternatively, Mel-Frequency Cepstral Coefficients (MFCC) are frequently employed to extract robust spectral envelopes [10], albeit at the cost of losing subtle textural information [8]. Existing deep learning approaches, such as standard CNNs, typically process these features in isolation [11], thereby failing to capitalize on their combined potential. To address this, we introduce a dual-view strategy that leverages the complementarity between the high-resolution details of STFT and the noise-robust envelopes of MFCC.

B. Tokenization for Audio Transformers

Transformers rely on discrete input tokens to perform self-attention [12]. The standard approach, adopted from Vision Transformers [3], divides images or spectrograms into fixed square patches. This strategy has been applied to underwater acoustic recognition in models like the Spectrogram Transformer Model [13]. However, recent studies suggest that rectangular patching may disrupt the inherent spectral continuity of hydroacoustic signals. To address this, newer methods such as those by Tang et al. [14] have explored slicing spectrograms along the frequency dimension to generate 1D tokens, aiming to better model temporal dynamics. Despite these advancements, efficiently balancing the preservation of high-resolution spectral integrity with the computational costs of global modeling remains a challenge. In this work, we propose an acoustic-aware feature slicing strategy combined with a hierarchical architecture to bridge this gap.

C. Hierarchical Efficiency in Audio Encoders

While Transformers offer superior global modeling to CNNs, their quadratic complexity ($O(N^2)$) poses a challenge for high-resolution sonar signals [4]. Standard global attention becomes prohibitive when processing long-duration underwater recordings capturing low-frequency periodicity. To mitigate this, hierarchical architectures adapted from vision, like the Swin Transformer [15], [16], employ window-based attention to limit computation to local regions. In the audio domain, the Hierarchical Token-Semantic Audio Transformer (HTS-AT) [17] refines this by cascading window attention with token reduction. This design allows the model to capture local details in shallow layers while efficiently modeling global dependencies in deep layers through progressively compressed token sequences. However, few UATR models integrate this hierarchical efficiency with multi-view fusion. Our framework addresses this by cascading window attention with token compression.

III. Methodology

A. Proposed Framework Overview

We define the underwater acoustic target recognition task as a supervised classification problem. The input is a raw hydroacoustic signal x. The objective is to predict the target vessel category $\hat{y}$ given the ground-truth label y. We process the signal to generate a dual-view input representation. This

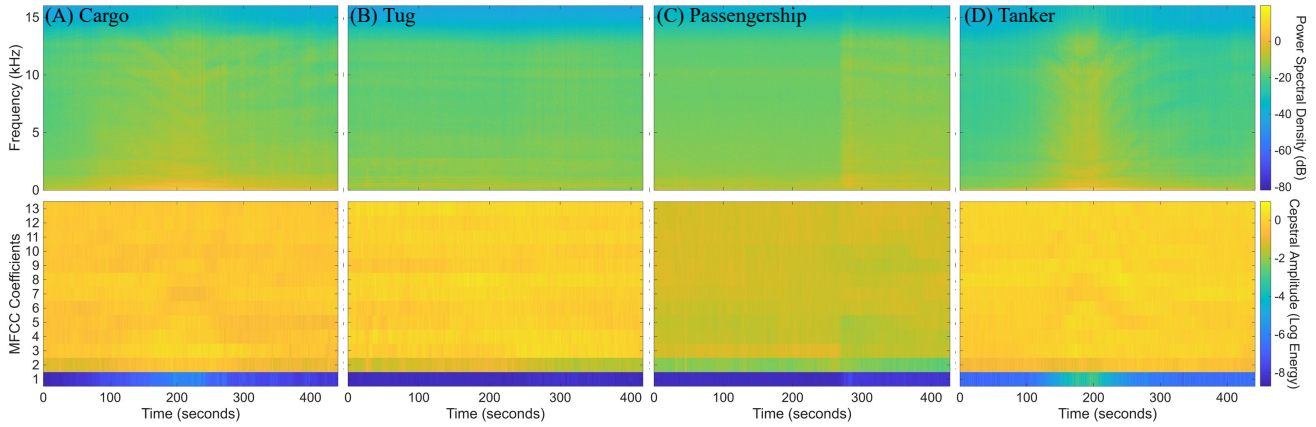

Fig. 2. Visualization of acoustic features for different ships. The top row presents STFT spectrograms, while the bottom row shows MFCC spectrograms.

representation consists of two distinct time-frequency features. The first view is the STFT, denoted as $X_{stft} \in \mathbb{R}^{F \times T}$. Here, F represents the frequency bins and T denotes the time frames. The second view is the MFCC, denoted as $X_{mfcc} \in \mathbb{R}^{C \times T}$. Here, C represents the cepstral coefficients.

As illustrated in Fig. 1, the framework adopts a "early-stage parallel, late-stage shared" architecture. First, both X_{stft} and X_{mfcc} undergo Acoustic-Aware Feature Slicing (Φ_{slice}) and are independently processed by Stage 1 of the Hierarchical Encoder. The output features from Stage 1 are aggregated via the Dual-View Fusion module $\Psi_{fuse}(\cdot)$. The resulting fused representation F_{fused} is then passed through the deeper, shared stages (Stage 2 and 3) to extract high-level semantic features Z, which are projected to the class scores $\mathbf{s}$.

B. Acoustic-Aware Feature Slicing

We define an acoustic-aware Feature Slicing layer to directly transform the two-dimensional spectrogram into a one-dimensional temporal sequence. By treating the entire frequency axis as the feature channel, this layer maps raw spectral features into a high-dimensional latent space using a one-dimensional convolution.

We implement this slicing mechanism using a one-dimensional convolution to map raw spectral features into a high-dimensional latent space. Given an input 2D acoustic feature map $X \in \mathbb{R}^{F \times T}$ where F denotes the number of frequency bins and T denotes the number of time frames, we treat it as a time series with F input channels. We define a 1D convolutional layer with kernel weights $W_{conv} \in \mathbb{R}^{D \times F \times K}$ where D is the embedding dimension and K is the kernel size along the time axis. The Feature Slicing process can be expressed as a sliding window aggregation along the time axis:

$$Y[b, d, t'] = \sum_{f=0}^{F-1} \sum_{k=0}^{K-1} W_{conv}[d, f, k] \cdot X[b, f, t' \cdot S + k] + B_{conv}[d], \tag{1}$$

where b is the batch index, S is the convolutional stride, and B_{conv} is the bias term. This operation performs downsampling in the time dimension, outputting a feature map $Y \in \mathbb{R}^{B \times D \times T'}$, where the output sequence length T' is calculated as:

$$T' = \lfloor \frac{T - K}{S} \rfloor + 1. \tag{2}$$

To adapt to the input format of the Transformer Encoder, we further need to transpose the feature dimensions and add positional information. First, we transpose the dimensions of the convolution output Y from (B, D, T') to (B, T', D) to conform to the requirements of sequence modeling. Subsequently, to compensate for the loss of positional information due to the convolution operation and translation invariance, we introduce a learnable positional embedding parameter $P \in \mathbb{R}^{1 \times T_{max} \times D}$. The final embedding representation E_{slice} is obtained by element-wise addition of the feature projection and the positional embedding:

$$E_{slice} = \text{Transpose}(Y) + P_{adapt} \in \mathbb{R}^{B \times T' \times D}, \tag{3}$$

where P_{adapt} is the result of dynamically interpolating P according to the current sequence length T'. This E_{slice} serves as the robust input token sequence for the subsequent hierarchical processing stages.

C. Hierarchical Encoder with Token Compression

Resolving the conflict between the high temporal resolution required for capturing transient signals and the quadratic complexity of global Transformers constitutes a primary architectural challenge. We design a Hierarchical Encoder to address this issue by sequentially cascading Window Attention with a Token Compression module. This composite mechanism ensures that the model can process lengthy sequences efficiently while progressively abstracting high-level semantic information.

In the shallow stages of the network, acoustic features are primarily characterized by local textures and short-term transient events. We adopt Window Attention to restrict the computation of self-attention within non-overlapping local windows. This strategy effectively reduces the computational

complexity from quadratic to linear with respect to the sequence length. It enables the shallow layers to process high-sampling-rate sonar data with significantly lower memory consumption while maintaining high-resolution perception of local details.

We partition the long sequence into multiple non-overlapping local windows of size W. Multi-Head Self-Attention (MSA) [15] is performed independently within each window to capture local transient details.

Following the local feature extraction, the sequence length typically remains substantial which hinders efficient global modeling. To facilitate the transition to deep feature extraction, we employ a Token Compression block that maps the high-resolution sequence into a compact latent space. This block utilizes stacked convolutional layers equipped with Batch Normalization (BN) and Gaussian Error Linear Unit (GELU) activation functions. By employing strided convolutions, we achieve a significant downsampling in the temporal dimension while simultaneously expanding the channel dimension:

$$F_{comp} = \text{GELU}(\text{BN}(\text{Conv}(F_{in}))). \tag{4}$$

This transformation reduces the computational burden for the subsequent global attention layers and enriches the semantic capacity of each token.

D. STFT–MFCC Fusion & Optimization Objectives

1) Dual-View Cross-Attention Fusion: To leverage the complementary advantages of heterogeneous representations, we place the fusion module immediately after the shallow feature extraction (Stage 1). The inputs to this module are the intermediate feature sequences output by Stage 1: $H^{(1)}_{mfcc}$ and $H^{(1)}_{stft}$. This block consists of two symmetric branches designed to enhance each feature modality using information from the other. Taking the MFCC enhancement branch as an instance, the inputs are the MFCC features from Stage 1, $H^{(1)}_{mfcc}$, and the STFT features, $H^{(1)}_{stft}$. We project these inputs to generate Query, Key, and Value vectors through linear layers. The cross-attention mechanism utilizes the MFCC features as the query to retrieve relevant details from the STFT features which serve as keys and values. This process aligns the granular texture of the STFT with the robust envelope of the MFCC:

$$\begin{aligned} M^{mfcc}_{enh} = \text{LayerNorm}\big(H^{(1)}_{mfcc} + \\ \text{Attn}(H^{(1)}_{mfcc}W_Q, H^{(1)}_{stft}W_K, H^{(1)}_{stft}W_V)\big), \end{aligned} \tag{5}$$

where $Attn$ denotes the standard attention operation. Similarly, the STFT enhanced feature M^{stft}_{enh} is obtained by swapping the roles of query and key/value.

We introduce a gating mechanism to dynamically adjust the fusion ratio between the enhanced features. The final fused representation is obtained via a weighted summation controlled by a learnable gate:

$$\begin{aligned} g &= \sigma(W_g[M^{mfcc}_{enh} \parallel M^{stft}_{enh}] + b_g), \\ F_{fused} &= g \odot M^{mfcc}_{enh} + (1-g) \odot M^{stft}_{enh}, \end{aligned} \tag{6}$$

where $\odot$ denotes element-wise multiplication, $\parallel$ denotes concatenation, and σ is the sigmoid activation function. This mechanism allows the model to adaptively select the most discriminative features and suppress noise interference effectively.

2) Optimization Objectives: The primary optimization target is the classification accuracy. We use the standard Cross-Entropy Loss function. Given the predicted class probability vector $\mathbf{p}$ and the one-hot ground-truth label $\mathbf{y}$, the loss is:

$$L_{ce} = -\sum_{i=1}^{N_{cls}} \mathbf{y}_i \log(\mathbf{p}_i). \tag{7}$$

The final prediction is derived from the fused feature representation F_{fused}. We also apply regularization techniques such as weight decay during training. This helps to prevent overfitting on the training data.

IV. Experiments

To rigorously evaluate the proposed SpecSlice-ViT and evaluate its effectiveness in handling non-stationary underwater acoustic signals, we conducted comprehensive experiments on the DeepShip benchmark dataset. This section details the experimental setup, comparative analysis against state-of-the-art baselines, ablation studies verifying component effectiveness, and a discussion on the model's internal mechanisms.

A. Experimental Setup

1) Dataset and Preprocessing: We utilized the DeepShip dataset [6], a large-scale real-world underwater acoustic benchmark collected in the Strait of Georgia delta. The dataset contains approximately 47 hours of passive sonar recordings from 265 distinct vessels, covering four categories: Cargo, Tug, Passengership, and Tanker. The diversity of sea states and background noise levels in DeepShip makes it a challenging standard for evaluating robustness. Detailed statistics of the dataset are provided in Table I.

TABLE I
DeepShip Benchmark Dataset Summary

Ship Type	Ships	Total Time	Recordings	Duration (s)
Cargo	69	10h 40m	110	180–610
Tug	17	11h 17m	70	180–1140
Passengership	46	12h 22m	193	6–1530
Tanker	133	12h 45m	240	6–700

To construct the dual-view input for our framework, all audio recordings were resampled to 16 kHz. We employed a sliding window strategy to segment the continuous signals into non-overlapping 5-second clips. Following common sample-level partitioning practices in previous studies (e.g., UATR-Transformer [5], STM [13], and 1DCTN [14]), the segmented samples were randomly split into training, validation, and testing sets with a ratio of 7:1.5:1.5.

TABLE II
PERFORMANCE COMPARISON ON THE DEEPSHIP DATASET

Method	OA (%)	Kappa	Samples/s	Infer. Time (ms)	Parameters (M)
CRNN [18]	88.4	0.846	108	9.22 ± 0.56	3.88
ResNet34 [19]	91.3	0.882	951	1.05 ± 0.19	21.29
ViT [3]	91.8	0.893	337	2.97 ± 0.22	86.24
Swin Trans. [15]	92.8	0.903	643	1.56 ± 0.11	27.50
MobileNetV2 [20]	93.5	0.913	140	6.90 ± 0.28	2.23
UATR-Trans. [5]	95.3	0.937	231	4.33 ± 0.40	2.55
SpecSlice-ViT (Ours)	**96.3**	**0.950**	**2194**	**0.46 ± 0.11**	88.73

2) Evaluation Metrics: We adopted a multi-dimensional evaluation protocol focusing on both classification accuracy and computational efficiency:

- **Overall Accuracy (OA):** The percentage of correctly classified samples on the test set.
- **Kappa Coefficient (κ):** A statistical measure of inter-rater agreement that accounts for chance agreement, offering a more robust metric for imbalanced classes.
- **Inference Throughput (samples/s):** The number of samples processed per second, indicating the model's efficiency in high-volume real-time scenarios.
- **Inference Time:** The average time required to process a single 5-second sample (in milliseconds).

3) Implementation Details: All experiments were implemented on an Intel Xeon Silver 4214R CPU and an NVIDIA GeForce RTX 3090 GPU. We optimized the model using the Adam optimizer with an initial learning rate of 2×10^{-4} and a batch size of 32. The training process spanned 300 epochs, employing Cross-Entropy Loss as the objective function. The model checkpoint achieving the highest accuracy on the validation set was selected for final testing.

B. Main Results

We compared SpecSlice-ViT with six representative baselines: convolutional architectures (CRNN [18], ResNet34 [19]), standard vision transformers (ViT [3], Swin Transformer [15]), and domain-specific methods (MobileNetV2 [20], UATR-Transformer [5]). The results are summarized in Table II.

1) Classification Performance Analysis: SpecSlice-ViT achieved a leading Overall Accuracy of 96.3% and a Kappa coefficient of 0.950, surpassing the strong domain-specific baseline UATR-Transformer by 1.0%. This performance gain is primarily attributed to the proposed slicing mechanism which maintains spectral integrity (as detailed in Sec. III-B), allowing SpecSlice-ViT to significantly outperform the patch-based ViT (91.8%). The proposed fusion module effectively mitigates the noise sensitivity of STFT by leveraging the robust spectral envelopes of MFCC, resulting in superior discriminability.

2) Computational Efficiency Analysis: In terms of efficiency, our method demonstrated exceptional real-time capability, reaching an inference speed of 2194 samples/s. This is approximately $9.5\times$ faster than UATR-Transformer and $2.3\times$ faster than ResNet34. The significant speedup stems from the Hierarchical Encoder (detailed in Sec. III-C), which effectively reduces the computational complexity from quadratic to linear. These results demonstrate the model's computational efficiency, suggesting it is a strong candidate for future validation on underwater platforms with limited computational resources.

TABLE III
ABLATION OF FEATURE SLICING (SLICING), HIERARCHICAL ENCODER (H-ENC), AND DUAL-VIEW FUSION (FUSION). THE BASELINE USES PATCH-BASED STFT.

Slicing	H-Enc	Fusion	OA (%)	Kappa	Samples/s
×	×	×	92.7	0.903	4301
✓	×	×	95.6	0.941	5283
✓	✓	×	95.9	0.946	1968
✓	✓	✓	**96.3**	**0.950**	**2194**

C. Ablation Studies

To validate the contribution of individual components, we conducted an ablation study using the STFT feature stream as the baseline. We incrementally integrated Feature Slicing, the Hierarchical Encoder, and finally the Cross-View Fusion with MFCC. Results are shown in Table III.

1) Impact of Feature Slicing: Replacing the standard patch embedding with our Feature Slicing strategy increased accuracy by 2.9% (from 92.7% to 95.6%). Inference throughput also rose from 4301 to 5283 samples/s. This joint gain in accuracy and throughput shows that slicing extracts more robust features than patching while staying computationally efficient, validating the efficiency of the 1D convolution design.

2) Efficacy of Hierarchical Encoder: Adding the Hierarchical Encoder further raised accuracy to 95.9%. Although cascaded attention stages reduced throughput to 1968 samples/s, the gain confirms the need for multi-scale modeling. Despite this added complexity, the model maintains high throughput and strikes a practical balance between precision and speed.

3) Analysis of Cross-View Fusion: Final integration with MFCC achieved the best performance, with 96.3% accuracy and a Kappa coefficient of 0.950. The extra 0.4% gain over the single-view STFT model shows that dual-view input provides complementary discriminative information unavailable to a single modality, while the efficient fusion design keeps inference speed competitive at 2194 samples/s.

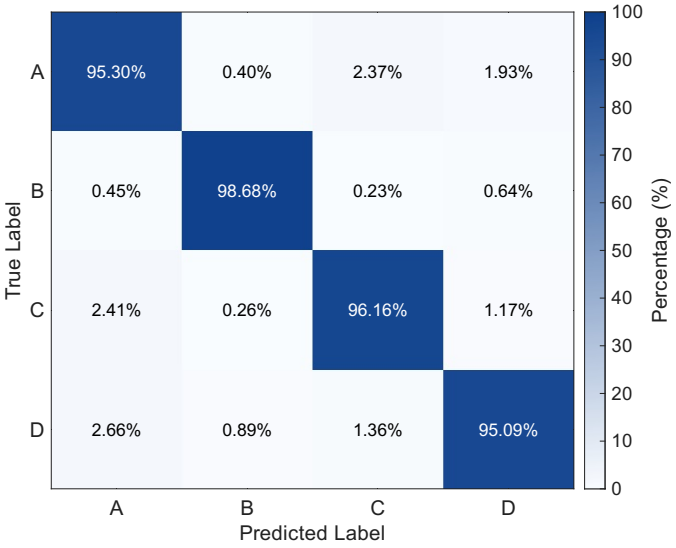

Fig. 3. Confusion matrix of SpecSlice-ViT on the DeepShip test set. The diagonal values represent the classification accuracy for each class. The class labels correspond to: A - Cargo, B - Tug, C - Passengership, and D - Tanker.

D. Visualization and Discussion

The confusion matrix (Fig. 3) shows that SpecSlice-ViT maintains high classification consistency across all classes, with diagonal elements above 95%. Analysis shows that the main challenge is distinguishing the acoustically similar Cargo and Passengership classes, which share overlapping low-frequency diesel signatures. The observed 2.37% misclassification of Cargo as Passengership highlights this physical ambiguity. However, compared with the patch-based ViT baseline in Table II (91.8% accuracy), SpecSlice-ViT markedly reduces this confusion. These results confirm that our Acoustic-Aware Feature Slicing preserves the harmonic integrity needed for such fine-grained discrimination, whereas patching disrupts continuous spectral lines. The high classification consistency for Tanker (95.09%) and Tug (98.68%) also validates the robustness of the Dual-View Fusion.

V. Conclusion

This study reconciles the persistent conflict between high-fidelity spectral modeling and computational efficiency in underwater acoustic target recognition. We demonstrate that maintaining harmonic continuity via frequency-dimension slicing, rather than standard patching, is essential for processing non-stationary sonar signals. Crucially, by achieving high accuracy with inference speeds exceeding 2000 samples/s, SpecSlice-ViT demonstrates the potential for migration to underwater edge platforms, such as Autonomous Underwater Vehicles (AUVs), subject to further embedded verification. This establishes a robust baseline for real-time, high-precision marine monitoring. However, the current reliance on fully supervised learning paradigms limits the adaptability of the model to extreme low-signal-to-noise ratio conditions or unseen open-set targets. Future research will investigate self-supervised pre-training strategies to unlock greater robustness in highly dynamic and unlabelled marine environments.

References

[1] S. Feng, S. Ma, X. Zhu, and M. Yan, "Artificial intelligence-based underwater acoustic target recognition: A survey," *Remote Sensing*, vol. 16, no. 17, p. 3333, 2024.

[2] J. Lu, S. Song, Z. Hu, and S. Li, "Fundamental frequency detection of underwater acoustic target using demon spectrum and cnn network," in *2020 3rd International Conference on Unmanned Systems (ICUS)*. IEEE, 2020, pp. 778–784.

[3] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[4] S. Feng, X. Zhu, and S. Ma, "Masking hierarchical tokens for underwater acoustic target recognition with self-supervised learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1365–1379, 2024.

[5] S. Feng and X. Zhu, "A transformer-based deep learning network for underwater acoustic target recognition," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[6] M. Irfan, Z. Jiangbin, S. Ali, M. Iqbal, Z. Masood, and U. Hamid, "Deepship: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification," *Expert Systems with Applications*, vol. 183, p. 115270, 2021.

[7] J. Chen, B. Han, X. Ma, and J. Zhang, "Underwater target recognition based on multi-decision lofar spectrum enhancement: A deep-learning approach," *Future Internet*, vol. 13, no. 10, p. 265, 2021.

[8] X. Luo, L. Chen, H. Zhou, and H. Cao, "A survey of underwater acoustic target recognition methods based on machine learning," *Journal of Marine Science and Engineering*, vol. 11, no. 2, p. 384, 2023.

[9] F. Hong, C. Liu, L. Guo, F. Chen, and H. Feng, "Underwater acoustic target recognition with a residual network and the optimized feature extraction method," *Applied Sciences*, vol. 11, no. 4, p. 1442, 2021.

[10] D. Liu, H. Yang, W. Hou, and B. Wang, "A novel underwater acoustic target recognition method based on mfcc and racnn," *Sensors*, vol. 24, no. 1, p. 273, 2024.

[11] Q. Sun and K. Wang, "Underwater single-channel acoustic signal multitarget recognition using convolutional neural networks," *The Journal of the Acoustical Society of America*, vol. 151, no. 3, pp. 2245–2254, 2022.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[13] P. Li, J. Wu, Y. Wang, Q. Lan, and W. Xiao, "Stm: Spectrogram transformer model for underwater acoustic target recognition," *Journal of Marine Science and Engineering*, vol. 10, no. 10, p. 1428, 2022.

[14] J. Tang, W. Gao, E. Ma, X. Sun, and J. Ma, "Deep learning based underwater acoustic target recognition: Introduce a recent temporal 2d modeling method," *Sensors*, vol. 24, no. 5, p. 1633, 2024.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[16] L. Chen, X. Luo, and H. Zhou, "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*, 2024, pp. 16–24.

[17] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 646–650.

[18] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.