

Wiener-Guided Spectral Line Enhancement for Vessel Recognition

Muyin Chen

Southeast University

Longyu Jiang

jlj@seu.edu.cn

Southeast University

Article

Keywords: deep learning, vessel recognition, spectral line enhancement, Wiener filtering, acoustic signal processing

Posted Date: September 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10981928/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Wiener-Guided Spectral Line Enhancement for Vessel Recognition

Muyin Chen¹ and Longyu Jiang^{1*}

^{1*}School of Computer Science and Engineering, Southeast University,
Nanjing, China .

*Corresponding author(s). E-mail(s): jly@seu.edu.cn;

Abstract

Deep-learning-based vessel-recognition systems infer vessel categories from acoustic recordings and may implicitly learn narrowband tonal cues, but label supervision does not explicitly impose a delayed-reference prediction criterion for extracting physically grounded tonal evidence. Delayed-reference line enhancement provides such a criterion but remains stand-alone preprocessing outside recognition optimization. We propose Wiener-guided spectral line enhancement (WG-SLE), a physics-guided trainable front end coupling delayed-reference prediction with recognition through role-separated optimization. Wiener-initialized class-indexed prototypes generate a delay-predictable spectral estimate that complements the observed spectrum; stop-gradient blocks direct classification gradients, and a sample-derived query filter enables inference-time routing. On DeepShip and ShipsEar, WG-SLE improves overall accuracy by up to 1.47 and 6.77 percentage points over matched Audio Spectrogram Transformer baselines, respectively, and exceeds least-mean-square adaptive line enhancement in overall accuracy across four AST variants on both datasets. The results show that the delay-predictable spectral estimate provides complementary recognition information.

Keywords: deep learning, vessel recognition, spectral line enhancement, Wiener filtering, acoustic signal processing

1 Introduction

Passive acoustic vessel recognition infers vessel categories from recordings of ship-radiated noise, supporting maritime surveillance and vessel monitoring [1, 2]. Periodic

machinery and propulsion processes generate narrowband tonal components, including shaft- and blade-rate components and their harmonics, many of which lie at low frequencies [3, 4]. In spectral representations, these components appear as discrete narrowband spectral lines [5]. Periodic modulation associated with shaft rotation and propeller-blade passage can further produce characteristic modulation patterns around these lines [6]. When sustained across successive time frames, the narrowband tonal components form line-like tonal traces in a time–frequency representation. Their temporal persistence can produce correlation across a selected delay, providing the acoustic basis for delayed-reference prediction. In measured recordings, underwater propagation and ambient noise can reduce the visibility and consistency of such traces [7, 8]. Deep-learning-based vessel-recognition systems may implicitly learn spectral patterns associated with vessel acoustics from labeled recordings. However, label supervision alone does not explicitly relate these learned patterns to known delayed-reference acoustic relationships or clarify how narrowband spectral lines contribute to recognition decisions. Delayed-reference predictability therefore provides an explicit acoustic criterion for emphasizing spectral components that remain correlated across a selected delay, providing a physically grounded basis for incorporating delay-predictable tonal evidence into recognition. Different spectral components can provide complementary information for vessel classification [9]; weaker delayed-reference predictability therefore does not imply irrelevance to recognition. The resulting delay-predictable spectral estimate should supplement, rather than replace, the observed spectrum.

These observations expose a broader mismatch in vessel recognition: delayed-reference enhancement provides an explicit acoustic criterion, whereas data-driven recognizers learn representations through label-driven objectives, and conventional pipelines keep the two processes separate. The central challenge is therefore to couple delayed-reference prediction with recognition within a trainable front end while preserving the delayed-reference prediction role and retaining the observed spectrum as an explicit term in the fused spectrum.

To examine how existing work addresses this mismatch, vessel-recognition methods can be grouped into three complementary routes: task-driven representation learning, explicit spectral-structure modeling, and acoustic front-end processing. Task-driven representation learning derives high-level recognition representations primarily from label supervision, whereas explicit spectral-structure modeling introduces prescribed spectral organization into the recognition model. The third route, acoustic front-end processing, comprises two relevant directions: learnable acoustic front ends, which parameterize low-level transformations as part of recognition training, and delayed-reference line enhancement, which applies an explicit delayed-reference prediction criterion.

Task-driven representation learning progressively broadens the scale at which discriminative spectrotemporal information is modeled from the observed spectrum. Convolutional and residual architectures construct local hierarchical patterns [10], while multi-resolution models combine information across complementary time–frequency scales [11]. Attention mechanisms then reweight informative regions [12], and transformer-based recognizers extend this modeling to broader dependencies

across spectrogram regions [13]. These approaches demonstrate that both local spectrotemporal patterns and broader contextual relations can be learned effectively from labeled observations. Although these models may implicitly encode narrowband tonal cues, their representations are learned primarily through label supervision without explicitly incorporating the delayed-reference prediction criterion, and the role of such cues in the final recognition decision therefore remains implicit.

Explicit spectral-structure modeling makes acoustically motivated frequency organization part of the recognition architecture. BSWNet begins from sub-band partitioning, preserving frequency positions while learning the relative importance of different bands [14]. Auditory-inspired deep convolution introduces a finer line-spectrum decomposition and associated energy distribution for vessel-timbre characterization [15], while SNANet organizes feature extraction and fusion around spectral components of ship-radiated noise and their corresponding frequency bands [16]. Collectively, these methods progress from band partitioning to line-spectrum decomposition and component-guided fusion. They make frequency structure explicit, but the emphasized spectral information is determined by prescribed design choices rather than by delayed-reference predictability. Because these operations are embedded within the recognizer, transferring the same prescribed spectral organization to another backbone may require architecture-specific modification rather than attachment through a common front-end interface, making direct reuse of broadly pretrained backbones less straightforward.

Acoustic front-end processing provides a separable interface by applying low-level transformations before high-level representation learning. Learnable front ends progressively extend the scope of this parameterization: SincNet constrains first-layer filters to parameterized sinc functions [17], whereas LEAF makes filtering, pooling, compression, and normalization jointly learnable [18]. Underwater acoustic systems adapt the same principle to vessel recognition. UALF places learnable Gabor filters before the classifier [19], while AGNet replaces fixed wavelet parameters with fine-grained trainable parameters [20]. Together, these studies establish that structured low-level acoustic transformations can be optimized as part of recognition training. Their responses, however, are optimized primarily through the classification objective rather than through the delayed-reference prediction criterion considered here.

The second direction within acoustic front-end processing is delayed-reference line enhancement, which applies an explicit acoustic criterion based on the relative predictability of narrowband tonal components across a selected delay. Classical adaptive line enhancement (ALE) exploits this correlation contrast through adaptive filtering [21, 22], while sparsity-based frequency-domain ALE further constrains the resulting estimate toward narrowband spectral components [23]. More directly, Yu et al. cascade least-mean-square adaptive line enhancement (LMS-ALE) with a pretrained VGGish encoder and a support vector machine classifier, demonstrating that delayed-reference line enhancement can benefit recognition from ship-radiated noise [24]. Their ALE, feature-extraction, and classification stages are optimized separately, and the line-enhanced signal is supplied as the input to the downstream recognizer. The method therefore remains an enhance-then-recognize cascade in which

delayed-reference line enhancement is separated from downstream recognition optimization, rather than a trainable spectral front end integrated with recognition. Because the enhancement stage is optimized independently of the downstream decision objective and its output becomes the recognizer’s sole input, its recognition benefit depends on whether other discriminative spectral information is retained.

Taken together, these routes provide complementary capabilities but leave their integration unresolved. Task-driven representation learning provides data-driven recognition representations, explicit spectral-structure modeling introduces prescribed spectral organization, learnable acoustic front ends integrate low-level transformations with recognition training, and delayed-reference line enhancement supplies an explicit acoustic criterion based on delayed-reference predictability but is conventionally separated from downstream recognition optimization. The remaining problem is therefore how to satisfy three requirements within the same spectral front end: to incorporate delay-predictable tonal evidence through delayed-reference prediction, to couple the prediction process with recognition optimization without allowing the classification objective to override the delayed-reference prediction role, and to keep the observed spectrum as an explicit term so that information beyond the delay-predictable spectral estimate is not discarded before recognition.

WG-SLE instantiates these requirements as a physics-guided, trainable spectral front end. Wiener-initialized, trainable class-indexed frequency-wise complex spectral prototypes parameterize the responses applied to the delayed spectrum, producing a delay-predictable spectral estimate. Scaled residual fusion uses the observed spectrum as the additive basis of the fused spectrum and introduces the estimate as a complementary term. Stop-gradient prevents the classification objective from directly updating the spectral prototypes through the enhancement path. Delayed-reference prediction and recognition are therefore coupled in one forward pipeline while their parameter-update roles remain separated. At inference, when class labels are unavailable, a sample-derived query filter produces soft routing weights over the prototype bank to form the routed spectral operator.

Figure 1 illustrates the redistribution of downstream classifier responses after introducing WG-SLE. The observed log-mel spectrogram contains three visible line-like tonal traces below approximately 1 kHz. In the matched baseline, the Grad-CAM response extends into middle- and high-frequency regions, including a marked response peak above 2 kHz without a comparably clear line-like tonal trace. With WG-SLE, the response becomes more concentrated around the visible sub-1 kHz line-like tonal traces. Consistent with this redistribution, the frequency-marginal response curves show a reduced response peak above 2 kHz and increased responses near the marked tonal-trace regions. This diagnostic contrast motivates evaluating whether delay-predictable tonal evidence can complement, rather than replace, the observed spectrum in downstream recognition. The maps therefore provide a response-level view of how WG-SLE changes the spectral regions emphasized by the classifier.

The subsequent evaluation distinguishes three settings: recognition without line enhancement, stand-alone time-domain LMS-ALE preprocessing, and the integrated WG-SLE front end. Least-mean-square adaptive line enhancement (LMS-ALE) [24]

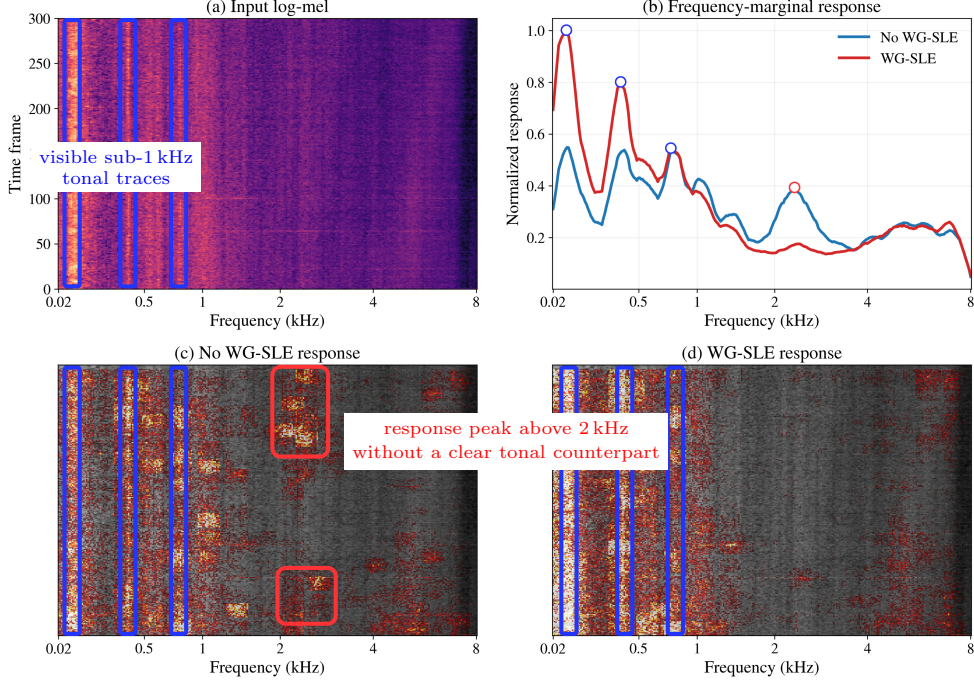


Fig. 1 Classifier-response redistribution for a representative DeepShip evaluation clip. (a) The observed log-mel spectrogram, with blue boxes marking visible tonal traces below 1 kHz. (b) Frequency-marginal Grad-CAM responses for the matched baseline (blue curve) and WG-SLE (red curve). (c) The matched-baseline Grad-CAM response map, with red boxes marking responses above 2 kHz without comparably clear tonal counterparts. (d) The WG-SLE Grad-CAM response map for the same clip. Both response maps target the same predicted-class logit.

serves as the conventional line-enhancement reference because it implements the classical delayed-reference principle through an independently updated adaptive filter. The comparison therefore contrasts two enhancement pathways: using a stand-alone line-enhanced signal as the sole recognizer input, or adding a delay-predictable spectral estimate to the observed spectrum. The AST-B/384 configuration of the Audio Spectrogram Transformer (AST) is selected as the principal downstream recognizer because it provides a standard patch-based spectrogram encoder without vessel-recognition-specific front-end modifications [25]. Accordingly, AST without line enhancement, AST with stand-alone LMS-ALE preprocessing, and AST with the WG-SLE front end provide a controlled comparison among the three settings. Additional convolutional and transformer-style recognizers, including ResNet-18 [26], are evaluated to examine the architectural compatibility of the same WG-SLE formulation with different downstream recognizers.

The main contributions of this work are as follows:

- We develop WG-SLE as a physics-guided, trainable spectral front end that incorporates delayed-reference prediction into a shared training framework with role-separated optimization. Wiener-initialized class-indexed spectral prototypes

produce a delay-predictable spectral estimate; scaled residual fusion keeps the observed spectrum as an explicit additive term, and stop-gradient separates spectral-prototype refinement from direct updates by the classification loss. The estimate therefore complements, rather than replaces, the observed spectrum.

- To resolve the absence of class labels at inference, WG-SLE uses a sample-derived query filter to replace label-based prototype selection with soft routing over the prototype bank. This supporting mechanism selects the inference-time spectral operator while leaving vessel-category prediction to the downstream recognizer.
- We evaluate WG-SLE on DeepShip and ShipsEar through controlled comparisons with recognition without line enhancement and stand-alone LMS-ALE preprocessing. Cross-backbone comparisons, ablation experiments, additive-noise evaluations, and class-, representation-, prototype-, and response-level diagnostics further assess recognition effects, architectural compatibility, component roles, optimization behavior, and classifier-response redistribution.

2 Results

2.1 Main results

Matched front-end comparison

Tables 1 and 2 compare stand-alone time-domain LMS-ALE preprocessing with the integrated WG-SLE front end under a common 30 s AST-family protocol. The matched baselines without line enhancement provide the references for the superscripted changes, and the $\Delta_{\text{WG-LMS}}$ rows report WG-SLE minus LMS-ALE. This comparison contrasts a separately optimized line-enhanced signal used as the sole recognizer input with a delay-predictable spectral estimate incorporated as complementary information within the recognition framework.

In the DeepShip comparison, WG-SLE improves all four reported metrics relative to the corresponding matched baseline for every evaluated AST variant, whereas stand-alone LMS-ALE decreases all four metrics. WG-SLE also exceeds LMS-ALE across every configuration, with OA differences ranging from 10.27 to 13.43 percentage points; AA, κ , and MF1 show the same direction.

In the ShipsEar comparison, stand-alone LMS-ALE produces metric-dependent changes relative to the matched baselines: AA increases for all four AST variants, whereas OA decreases for AST-B/384 and is unchanged to the displayed precision for the other three; the changes in κ and MF1 are mixed. WG-SLE exceeds LMS-ALE in OA, κ , and MF1 across all four variants and in AA for three, with AST-B/224 as the sole AA exception. Taken together, these comparisons suggest that line enhancement optimized separately from the recognizer does not necessarily yield uniform improvements in downstream recognition performance.

Backbone dependence and compatibility

Beyond the AST-family front-end comparison, we examine the architectural compatibility of WG-SLE with different downstream recognizers. Tables 3 and 4 report the corresponding results under the common 30 s protocol for DeepShip and ShipsEar.

Table 1 DeepShip front-end comparison under the common 30 s protocol. Superscripts indicate changes relative to the corresponding matched baseline without line enhancement; $\Delta_{\text{WG-LMS}}$ denotes WG-SLE minus LMS-ALE. Differences are calculated from unrounded results before display. Arrows indicate numerical direction only, not statistical significance. OA, AA, κ , and MF1 denote overall accuracy, average accuracy, Cohen’s κ coefficient, and macro-F1, respectively.

Backbone	Front end	OA (%)	AA (%)	κ	MF1 (%)
AST-B/384	WG-SLE	79.51 ^{↑1.47}	79.33 ^{↑1.40}	0.7264 ^{↑0.020}	79.35 ^{↑1.35}
	LMS-ALE	66.08 ^{↓11.96}	65.61 ^{↓12.32}	0.5459 ^{↓0.161}	65.25 ^{↓12.76}
	$\Delta_{\text{WG-LMS}}$	+13.43	+13.71	+0.181	+14.10
AST-B/224	WG-SLE	77.26 ^{↑0.28}	77.29 ^{↑0.20}	0.6967 ^{↑0.004}	77.28 ^{↑0.29}
	LMS-ALE	65.24 ^{↓11.74}	65.33 ^{↓11.75}	0.5366 ^{↓0.156}	65.32 ^{↓11.67}
	$\Delta_{\text{WG-LMS}}$	+12.02	+11.96	+0.160	+11.96
AST-S/224	WG-SLE	75.06 ^{↑0.96}	74.80 ^{↑0.95}	0.6667 ^{↑0.013}	74.86 ^{↑0.93}
	LMS-ALE	62.13 ^{↓11.96}	61.98 ^{↓11.87}	0.4942 ^{↓0.160}	61.71 ^{↓12.22}
	$\Delta_{\text{WG-LMS}}$	+12.92	+12.82	+0.173	+13.16
AST-Ti/224	WG-SLE	74.83 ^{↑0.68}	74.81 ^{↑0.64}	0.6642 ^{↑0.009}	74.80 ^{↑0.65}
	LMS-ALE	64.56 ^{↓9.59}	64.24 ^{↓9.93}	0.5261 ^{↓0.129}	64.39 ^{↓9.77}
	$\Delta_{\text{WG-LMS}}$	+10.27	+10.57	+0.138	+10.42

Table 2 ShipsEar front-end comparison under the common 30 s protocol. Superscripts indicate changes relative to the corresponding matched baseline without line enhancement; $\Delta_{\text{WG-LMS}}$ denotes WG-SLE minus LMS-ALE.

Backbone	Front end	OA (%)	AA (%)	κ	MF1 (%)
AST-B/384	WG-SLE	82.29 ^{↑6.77}	79.35 ^{↑11.76}	0.7508 ^{↑0.101}	76.77 ^{↑9.75}
	LMS-ALE	73.96 ^{↓1.56}	70.75 ^{↑3.16}	0.6363 ^{↓0.014}	67.01 ^{↓0.02}
	$\Delta_{\text{WG-LMS}}$	+8.33	+8.60	+0.115	+9.76
AST-B/224	WG-SLE	80.73 ^{↑3.65}	72.76 ^{↑3.95}	0.7241 ^{↑0.052}	73.97 ^{↑2.69}
	LMS-ALE	77.08 ^{→0.00}	74.08 ^{↑5.28}	0.6743 ^{↑0.002}	70.31 ^{↓0.97}
	$\Delta_{\text{WG-LMS}}$	+3.65	-1.32	+0.050	+3.66
AST-S/224	WG-SLE	79.17 ^{↑5.73}	73.95 ^{↑5.55}	0.7055 ^{↑0.083}	75.76 ^{↑11.93}
	LMS-ALE	73.44 ^{→0.00}	69.06 ^{↑0.66}	0.6217 ^{↓0.001}	67.39 ^{↑3.55}
	$\Delta_{\text{WG-LMS}}$	+5.73	+4.89	+0.084	+8.37
AST-Ti/224	WG-SLE	73.96 ^{↑1.56}	71.36 ^{↑5.37}	0.6379 ^{↑0.030}	71.20 ^{↑4.44}
	LMS-ALE	72.40 ^{→0.00}	66.59 ^{↑0.60}	0.5973 ^{↓0.010}	66.73 ^{↓0.03}
	$\Delta_{\text{WG-LMS}}$	+1.56	+4.77	+0.041	+4.47

Each row gives the WG-SLE result, and the superscripts denote the change relative to the corresponding matched baseline.

Within the DeepShip backbone results, most evaluated downstream recognizers show positive changes relative to their corresponding matched baselines after introducing WG-SLE. All evaluated transformer-style recognizers improve across the reported

Table 3 DeepShip backbone compatibility under the common 30 s protocol.

Family	Backbone	OA (%)	AA (%)	κ	MF1 (%)
<i>CNN backbones</i>					
CNN	MobileNetV3-S [27]	63.71 ^{↑1.13}	63.49 ^{↑1.20}	0.5147 ^{↑0.015}	63.51 ^{↑1.10}
CNN	CRNN-9 [28]	74.04 ^{↑0.11}	73.89 ^{↑0.15}	0.6533 ^{↑0.002}	73.90 ^{↑0.16}
CNN	ResNet-18 [26]	78.89 ^{↑0.84}	76.10 ^{↓1.88}	0.6848 ^{↓0.022}	78.73 ^{↑0.78}
CNN	Xception [29]	74.04 ^{↑2.31}	73.83 ^{↑2.29}	0.6529 ^{↑0.031}	74.03 ^{↑2.39}
CNN	DenseNet-121 [30]	76.75 ^{↑0.17}	76.55 ^{↑0.15}	0.6894 ^{↑0.002}	76.67 ^{↑0.26}
CNN	EfficientNet-B0 [31]	74.49 ^{↑0.23}	74.59 ^{↑0.27}	0.6599 ^{↑0.003}	74.41 ^{↑0.17}
<i>Transformer-style backbones</i>					
Transformer	AST-B/384 [25]	79.51 ^{↑1.47}	79.33 ^{↑1.40}	0.7264 ^{↑0.020}	79.35 ^{↑1.35}
Transformer	MHT-UATR-Transformer [32]	69.84 ^{↑1.15}	69.66 ^{↑1.42}	0.5986 ^{↑0.016}	69.86 ^{↑1.12}
Transformer	MF-UATR [33]	75.34 ^{↑0.62}	75.15 ^{↑0.60}	0.6706 ^{↑0.008}	75.24 ^{↑0.61}
Transformer	EAT-base [34]	76.35 ^{↑1.02}	76.23 ^{↑1.03}	0.6841 ^{↑0.014}	76.38 ^{↑1.09}
Transformer	ConViT [35]	67.71 ^{↑0.52}	63.40 ^{↑0.11}	0.5573 ^{↑0.023}	59.13 ^{↑1.27}

Table 4 ShipsEar backbone compatibility under the common 30 s protocol.

Family	Backbone	OA (%)	AA (%)	κ	MF1 (%)
<i>CNN backbones</i>					
CNN	MobileNetV3-S [27]	78.65 ^{↑1.04}	72.45 ^{↑1.43}	0.6988 ^{↑0.016}	72.36 ^{↑1.71}
CNN	CRNN-9 [28]	77.60 ^{↑0.52}	73.28 ^{↑1.12}	0.6841 ^{↑0.014}	72.77 ^{↑1.25}
CNN	ResNet-18 [26]	82.29 ^{↑1.04}	77.42 ^{↑1.78}	0.7491 ^{↑0.014}	75.75 ^{↑2.57}
CNN	Xception [29]	73.96 ^{↑2.60}	73.20 ^{↑7.91}	0.6350 ^{↑0.037}	71.23 ^{↑7.25}
CNN	DenseNet-121 [30]	79.69 ^{↑1.04}	76.85 ^{↑2.95}	0.7124 ^{↑0.017}	74.56 ^{↑2.85}
CNN	EfficientNet-B0 [31]	78.65 ^{↑0.53}	75.38 ^{↑1.24}	0.6984 ^{↑0.006}	73.48 ^{↑1.15}
<i>Transformer-style backbones</i>					
Transformer	AST-B/384 [25]	82.29 ^{↑6.77}	79.35 ^{↑11.76}	0.7508 ^{↑0.101}	76.77 ^{↑9.75}
Transformer	MHT-UATR-Transformer [32]	74.22 ^{↑1.10}	68.03 ^{↑0.49}	0.6376 ^{↑0.013}	73.04 ^{↑0.27}
Transformer	MF-UATR [33]	76.56 ^{↑0.52}	68.64 ^{↑1.33}	0.6616 ^{↑0.010}	70.69 ^{↑0.82}
Transformer	EAT-base [34]	73.96 ^{↑3.65}	68.16 ^{↑7.33}	0.6305 ^{↑0.047}	67.60 ^{↑5.88}
Transformer	ConViT [35]	67.61 ^{↑0.85}	67.86 ^{↑1.14}	0.5684 ^{↑0.012}	67.64 ^{↑1.27}

metrics, and most CNN backbones also show consistent gains. ResNet-18 is the main exception, with increases in OA and MF1 but decreases in AA and κ .

Within the ShipsEar backbone results, WG-SLE yields positive changes across all reported metrics for all evaluated downstream recognizers relative to their corresponding matched baselines, although the magnitude of the improvement varies across backbones and metrics.

Table 5 DeepShip configuration-level ablation using AST-B/384. Δ OA is calculated relative to the full WG-SLE setting within each group. Parameter-freezing and gradient-path definitions are given in Methods.

Group	Configuration	OA (%)	AA (%)	κ	MF1 (%)	Δ OA
Initialization	Zero initialization	77.31	77.03	0.6967	76.95	$\downarrow 2.20$
	Fixed Wiener-initialized front end	77.60	77.36	0.7006	77.50	$\downarrow 1.91$
	Wiener initialization + spectral-prototype refinement	79.51	79.33	0.7264	79.35	–
Training objective	CE loss only	78.05	77.93	0.7069	78.01	$\downarrow 1.46$
	CE + Wiener prediction	78.84	78.62	0.7171	78.72	$\downarrow 0.67$
	CE + Wiener prediction + ℓ_1	79.51	79.33	0.7264	79.35	–
Fusion path-way	Enhancement path only	66.31	65.10	0.5493	65.70	$\downarrow 13.20$
	Scaled residual fusion without stop-gradient	78.27	78.12	0.7098	78.24	$\downarrow 1.24$
	Scaled residual fusion + stop-gradient	79.51	79.33	0.7264	79.35	–

2.2 Ablation study

The main recognition results characterize the external behavior of WG-SLE. We next examine three questions in the evaluated DeepShip AST-B/384 configuration with 30s inputs: whether Wiener initialization benefits subsequent spectral-prototype refinement, whether the Wiener prediction objective and sparsity regularization improve upon classification-loss-driven front-end learning, and whether the delay-predictable spectral estimate should replace or supplement the observed spectrum. The initialization group compares zero initialization, a fixed Wiener-initialized front end whose WG-SLE parameters remain at their predefined values, and Wiener initialization followed by spectral-prototype refinement. The objective group begins with a CE-loss-only configuration in which the trainable front end is constrained directly by the classification loss without the Wiener prediction objective or sparsity regularization.

Table 5 addresses three distinct aspects of WG-SLE. In the initialization group, keeping the complete Wiener-initialized front end fixed yields only a small numerical advantage over zero initialization, whereas allowing spectral-prototype refinement from the Wiener anchors produces the highest values across OA, AA, κ , and MF1. The difference therefore lies primarily in trainable refinement rather than in the preset Wiener values alone.

In the objective group, the CE-loss-only setting provides the reference for classification-loss-driven front-end learning. Adding the Wiener prediction objective and then sparsity regularization is associated with progressively higher recognition metrics in the reported configurations. The fusion group addresses a separate question: the enhancement-path-only configuration performs substantially worse than either scaled-residual-fusion configuration, indicating that the delay-predictable spectral estimate is inadequate as the sole recognizer input in this setting. Stop-gradient gives the highest values within the fusion group, although the available configurations do not isolate its causal contribution.

Table 6 DeepShip paired noisy-evaluation changes averaged over three matched random seeds for each noise family. The same three seeds are used across SNRs and metrics within each family, and the table reports means only. OA, AA, and MF1 changes are in percentage points; the κ column reports $100\Delta\kappa$. Positive values favor WG-SLE. Individual observations and sample standard deviations are shown in Fig. 2.

Noise	SNR (dB)	Δ OA (pp)	Δ AA (pp)	$100\Delta\kappa$	Δ MF1 (pp)
White	-5	+6.84	+6.65	+9.01	+8.35
	0	+3.67	+3.60	+4.86	+4.24
	5	+3.81	+3.58	+5.00	+4.22
	10	+3.59	+3.35	+4.72	+3.75
Brown	-5	+1.81	+2.01	+2.48	+3.49
	0	+3.64	+3.62	+4.83	+5.01
	5	+3.02	+2.89	+3.97	+3.93
	10	+1.72	+1.60	+2.24	+2.22
Wenz-shaped [7]	-5	+0.96	+1.49	+1.66	+2.92
	0	+1.24	+1.37	+1.73	+2.57
	5	+2.29	+2.25	+3.00	+3.27
	10	+2.40	+2.21	+3.13	+3.19

2.3 Noisy evaluation and diagnostic analyses

The following analyses examine whether the recognition changes persist under synthetic perturbations and how they appear at complementary analytical levels. Paired confusion matrices localize class-level changes, t-SNE provides a qualitative projection of the learned embeddings, spectral-prototype alignment measures how optimization changes the Wiener-initialized spectral prototypes, and Grad-CAM maps characterize downstream response redistribution. Together, these diagnostics connect recognition outcomes to decision-, representation-, prototype-, and response-level behavior without treating any single visualization as causal evidence.

Noisy evaluation

We evaluate additive-noise sensitivity in the DeepShip AST-B/384 configuration using checkpoints trained on the original recordings and introducing synthetic perturbations only at evaluation. The three conditions comprise spectrally flat white Gaussian noise, brown noise with an inverse-square power spectrum, and a Wenz-shaped colored-noise proxy motivated by classical ocean ambient-noise spectra. Table 6 reports the paired WG-SLE-minus-baseline changes from -5 to 10 dB.

Across the three-seed means, WG-SLE shows positive paired differences for all tested noise families, SNR levels, and reported metrics. The dependence on SNR differs across noise spectra: the white-noise advantage is largest at -5 dB, the brown-noise advantage peaks at 0 dB, whereas the Wenz-shaped advantage increases across the tested SNR range. Absolute recognition performance nevertheless improves with increasing SNR for both the matched baseline and WG-SLE.

Class-level diagnostic

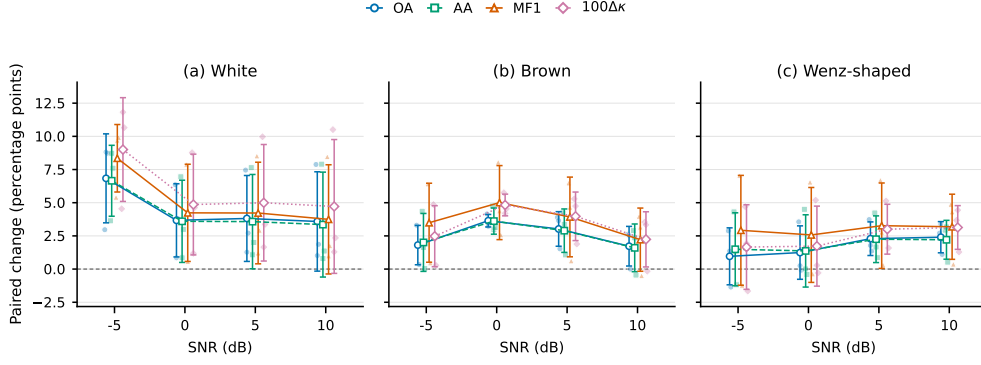


Fig. 2 Paired noisy-evaluation changes for DeepShip AST-B/384. Panels show (a) white Gaussian noise, (b) brown noise, and (c) Wenz-shaped colored noise. The same three matched random seeds are used at -5 , 0 , 5 , and 10 dB in each panel. Blue circles, green squares, orange triangles, and pink diamonds denote OA, AA, MF1, and $100\Delta\kappa$, respectively. Lines connect the mean paired changes at the four tested SNR levels; error bars indicate the sample standard deviation across the three seeds, and translucent markers show individual paired values. OA, AA, and MF1 are reported in percentage points. Positive values favor WG-SLE.

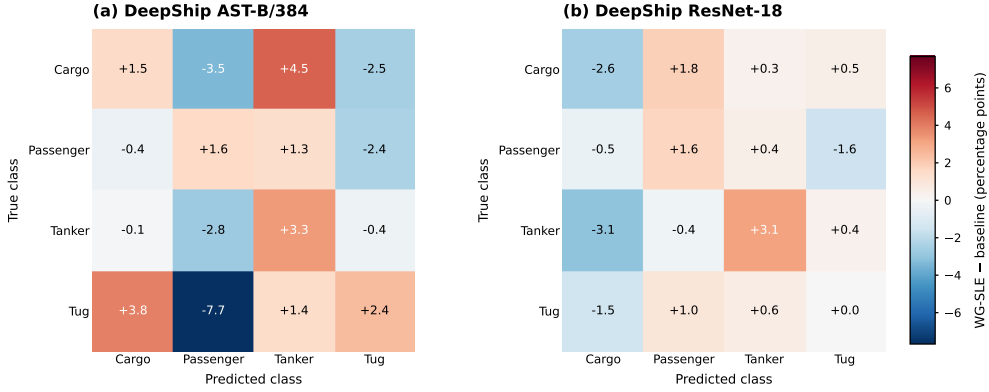


Fig. 3 Paired changes in row-normalized DeepShip confusion matrices for (a) AST-B/384 and (b) ResNet-18, averaged over three matched random seeds. Rows denote true classes and columns predicted classes; values are WG-SLE-minus-baseline changes in percentage points. Both panels use the same zero-centered, symmetric color scale.

Figure 3 summarizes the paired class-level changes for AST-B/384 and ResNet-18. AST-B/384 shows positive diagonal changes for all four DeepShip classes, whereas ResNet-18 exhibits a mixed class-wise pattern. These matrices are used as descriptive diagnostics; their backbone-dependent interpretation is considered in Discussion.

Embedding visualization

Figure 4 compares t-SNE projections of the matched baseline and WG-SLE embeddings using DeepShip and ShipsEar. Each point represents a held-out sample, and

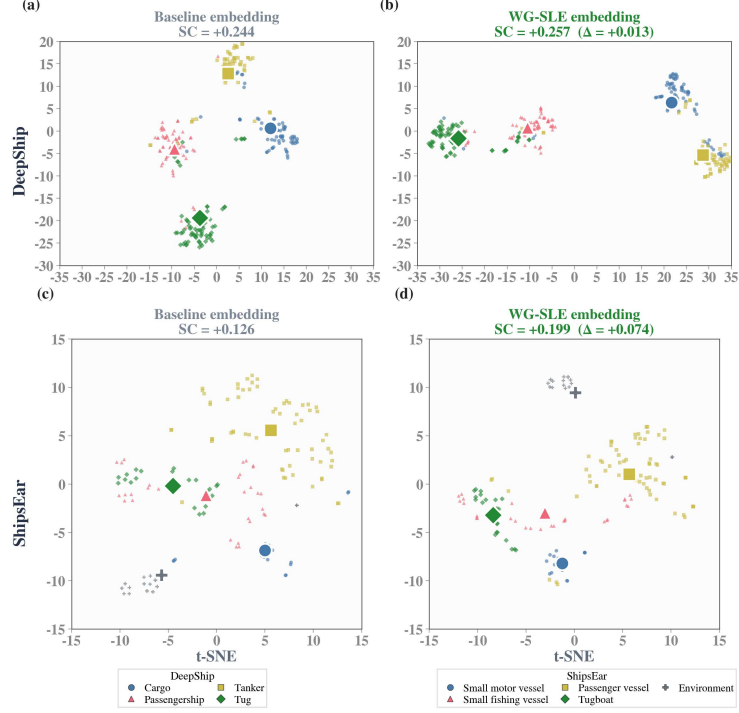


Fig. 4 Two-dimensional t-SNE projections of embeddings from (a) the matched baseline on DeepShip, (b) WG-SLE on DeepShip, (c) the matched baseline on ShipsEar, and (d) WG-SLE on ShipsEar. Each point represents an evaluation-partition sample, and larger markers denote the projected class centers.

the larger markers indicate the projected class centers. The projections are used as qualitative diagnostics of class structure in the learned embedding space.

Within the DeepShip projection, both embeddings contain distinct class clusters. The passenger and tug clusters overlap less with neighboring projected points after introducing WG-SLE, while the cargo and tanker clusters remain separated. The silhouette coefficient computed on this two-dimensional projection is 0.244 for the matched baseline and 0.257 for WG-SLE.

Within the ShipsEar projection, passenger vessel, small fishing vessel, tugboat, and small motor vessel samples occupy more compact projected regions after introducing WG-SLE, whereas the ambient background samples remain compact in both panels. The corresponding two-dimensional silhouette coefficients are 0.126 and 0.199.

Because t-SNE is a nonlinear projection, the displayed geometry should be interpreted qualitatively. The silhouette coefficients are computed on the two-dimensional projections and are therefore treated only as projection-dependent descriptive summaries. Their higher values are consistent with the reduced overlap visible in the plots, but do not directly measure geometry in the original embedding space.

Spectral-prototype alignment

We next examine how spectral-prototype refinement changes the Wiener-initialized prototypes. This diagnostic quantifies their directional agreement with the class-wise Wiener anchors and the relative magnitude of the resulting complex-valued updates.

Using the fixed STFT-domain analysis ranges defined in Eq. (53), let $\mathcal{B} \in \{\mathcal{B}_{40:8000}, \mathcal{B}_{40:1000}\}$. For class c , define the reference and trained prototype vectors as

$$\mathbf{w}_{c,\mathcal{B}}^{\text{ref}} = \left[\hat{W}_c(f) \right]_{f \in \mathcal{B}} \in \mathbb{C}^{|\mathcal{B}|}, \quad \mathbf{w}_{c,\mathcal{B}}^{\text{train}} = [W_c(f)]_{f \in \mathcal{B}} \in \mathbb{C}^{|\mathcal{B}|}. \quad (1)$$

For the nonzero reference and trained vectors in the evaluated checkpoints, directional agreement and relative update magnitude are measured by

$$\text{Cos}_{c,\mathcal{B}} = \frac{|\langle \mathbf{w}_{c,\mathcal{B}}^{\text{ref}}, \mathbf{w}_{c,\mathcal{B}}^{\text{train}} \rangle|}{\|\mathbf{w}_{c,\mathcal{B}}^{\text{ref}}\|_2 \|\mathbf{w}_{c,\mathcal{B}}^{\text{train}}\|_2}, \quad (2)$$

$$L_{2,c,\mathcal{B}}^{\text{rel}} = \frac{\|\mathbf{w}_{c,\mathcal{B}}^{\text{train}} - \mathbf{w}_{c,\mathcal{B}}^{\text{ref}}\|_2}{\|\mathbf{w}_{c,\mathcal{B}}^{\text{ref}}\|_2}. \quad (3)$$

Here, $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^H \mathbf{b}$, where $(\cdot)^H$ denotes conjugate transpose. The absolute value of the complex inner product makes $\text{Cos}_{c,\mathcal{B}}$ insensitive to a global complex-phase offset, whereas $L_{2,c,\mathcal{B}}^{\text{rel}}$ measures the magnitude of the complete complex-valued update relative to the corresponding Wiener anchor.

For each analysis range, the two measures are computed separately for the C class-indexed spectral prototypes and then averaged arithmetically across classes:

$$\overline{\text{Cos}}_{\mathcal{B}} = \frac{1}{C} \sum_{c=1}^C \text{Cos}_{c,\mathcal{B}}, \quad \overline{L}_{2,\mathcal{B}}^{\text{rel}} = \frac{1}{C} \sum_{c=1}^C L_{2,c,\mathcal{B}}^{\text{rel}}. \quad (4)$$

The differences reported in Table 7 are defined as the 40–1000 Hz value minus the corresponding 40–8000 Hz value:

$$\Delta \text{Cos} = \overline{\text{Cos}}_{\mathcal{B}_{40:1000}} - \overline{\text{Cos}}_{\mathcal{B}_{40:8000}}, \quad (5)$$

$$\Delta L_2^{\text{rel}} = \overline{L}_{2,\mathcal{B}_{40:1000}}^{\text{rel}} - \overline{L}_{2,\mathcal{B}_{40:8000}}^{\text{rel}}. \quad (6)$$

All differences are calculated from the unrounded class-averaged values.

Table 7 shows a consistent cross-range pattern across the four evaluated seed-42 AST-family checkpoints: $\overline{\text{Cos}}_{\mathcal{B}_{40:1000}}$ is higher, whereas $\overline{L}_{2,\mathcal{B}_{40:1000}}^{\text{rel}}$ is lower, than the corresponding values over $\mathcal{B}_{40:8000}$. The trained spectral prototypes therefore remain closer to their class-wise Wiener anchors within the 40–1000 Hz analysis range, while the nonzero relative updates in both ranges show that spectral-prototype refinement still occurs. Because these ranges are defined directly on the linear-frequency STFT axis, this analysis is distinct from the lowest-12.5%-mel-bin partition used in

Table 7 Alignment of trained spectral prototypes with their class-wise Wiener anchors on DeepShip. The two analysis columns report class-averaged values over STFT bins whose center frequencies lie within the corresponding frequency ranges. Differences are calculated from unrounded class-averaged values.

Backbone	Cosine similarity			Relative L_2 update		
	40–8000 Hz	40–1000 Hz	Δ	40–8000 Hz	40–1000 Hz	Δ
AST-B/384	0.973	0.993	+0.020	0.334	0.123	-0.212
AST-B/224	0.984	0.995	+0.011	0.283	0.100	-0.183
AST-S/224	0.974	0.993	+0.019	0.328	0.119	-0.209
AST-Ti/224	0.973	0.993	+0.020	0.333	0.122	-0.211

the Grad-CAM response summary; the two diagnostics respectively characterize prototype refinement and classifier-response redistribution. The interpretation of these cross-range differences is considered in Discussion.

Grad-CAM response analysis

We use gradient-weighted class activation mapping (Grad-CAM) to examine how WG-SLE changes the frequency distribution of downstream classifier responses. For each backbone, the quantitative analysis includes all DeepShip evaluation clips for which the matched baseline and WG-SLE predict the same category. The paired response maps are generated from the same audio clip and target the shared predicted-class logit. For AST-B/384, Grad-CAM is computed from the patch-token activations at the output of the final transformer block; for ResNet-18, it is computed from the final convolutional layer. The resulting nonnegative maps are bilinearly interpolated to the log-mel grid. Quantitative summaries are computed from the unclipped response maps, whereas clipping at the 99th percentile and min–max normalization are applied only for visualization.

For a nonnegative Grad-CAM response map $\mathbf{S} \in \mathbb{R}_{\geq 0}^{F_{\text{cam}} \times T_{\text{cam}}}$ with positive total response, where F_{cam} and T_{cam} denote its mel-frequency and time-axis sizes, respectively, the frequency-marginal response is defined as

$$p_u = \frac{\sum_{v=1}^{T_{\text{cam}}} S_{u,v}}{\sum_{u'=1}^{F_{\text{cam}}} \sum_{v=1}^{T_{\text{cam}}} S_{u',v}}, \quad u = 1, \dots, F_{\text{cam}}. \quad (7)$$

We summarize each response map using three complementary quantities:

$$H(\mathbf{S}) = -\frac{1}{\log F_{\text{cam}}} \sum_{u=1}^{F_{\text{cam}}} p_u \log p_u, \quad (8)$$

$$\text{Top10}(\mathbf{S}) = \sum_{\ell \in \mathcal{I}_{10\%}} \tilde{S}_\ell, \quad (9)$$

$$\text{LMB}_{12.5}(\mathbf{S}) = \frac{\sum_{u=1}^{K_{\text{LMB}}} \sum_{v=1}^{T_{\text{cam}}} S_{u,v}}{\sum_{u=1}^{F_{\text{cam}}} \sum_{v=1}^{T_{\text{cam}}} S_{u,v}}. \quad (10)$$

Table 8 DeepShip Grad-CAM response summaries over the backbone-specific prediction-agreement subsets. For each included clip, the matched baseline and WG-SLE maps target the shared predicted-class logit. ΔH is the mean paired change in normalized frequency-marginal entropy, whereas ΔTop10 and $\Delta \text{LMB}_{12.5}$ are mean paired changes in response share and are reported in percentage points (pp). The backbone superscript g is omitted from the column headings.

Backbone	ΔH	ΔTop10 (pp)	$\Delta \text{LMB}_{12.5}$ (pp)
AST-B/384	-0.0093	+1.29	+4.45
ResNet-18	-0.0008	+0.12	+0.42

Here, $\tilde{S}$ is the flattened response map normalized so that $\sum_{\ell=1}^{F_{\text{cam}} T_{\text{cam}}} \tilde{S}_{\ell} = 1$, and $\mathcal{I}_{10\%}$ indexes its $K = \max\{1, \lceil 0.10 F_{\text{cam}} T_{\text{cam}} \rceil\}$ largest entries. We use the convention $0 \log 0 = 0$. The normalized entropy H measures response dispersion along the mel-frequency axis. The Top10 metric measures the fraction of total response contained in the strongest 10% of time-frequency locations, irrespective of their frequency positions. The LMB_{12.5} metric uses $K_{\text{LMB}} = \max\{1, \lceil 0.125 F_{\text{cam}} \rceil\}$ and measures the fraction of total response assigned to the lowest 12.5% of mel-frequency bins across all time frames.

For backbone g , let $\mathcal{I}_{\text{eval}}$ denote the DeepShip evaluation partition. The prediction-agreement subset is

$$\mathcal{I}_{\text{agree}}^{(g)} = \left\{ i \in \mathcal{I}_{\text{eval}} : \hat{y}_{i,\text{base}}^{(g)} = \hat{y}_{i,\text{WG}}^{(g)} \right\}, \quad N_{\text{agree}}^{(g)} = \left| \mathcal{I}_{\text{agree}}^{(g)} \right|. \quad (11)$$

The three mean paired WG-SLE-minus-baseline changes are defined as

$$\Delta H^{(g)} = \frac{1}{N_{\text{agree}}^{(g)}} \sum_{i \in \mathcal{I}_{\text{agree}}^{(g)}} \left[H(\mathbf{S}_{i,\text{WG}}^{(g)}) - H(\mathbf{S}_{i,\text{base}}^{(g)}) \right], \quad (12)$$

$$\Delta \text{Top10}^{(g)} = \frac{1}{N_{\text{agree}}^{(g)}} \sum_{i \in \mathcal{I}_{\text{agree}}^{(g)}} \left[\text{Top10}(\mathbf{S}_{i,\text{WG}}^{(g)}) - \text{Top10}(\mathbf{S}_{i,\text{base}}^{(g)}) \right], \quad (13)$$

$$\Delta \text{LMB}_{12.5}^{(g)} = \frac{1}{N_{\text{agree}}^{(g)}} \sum_{i \in \mathcal{I}_{\text{agree}}^{(g)}} \left[\text{LMB}_{12.5}(\mathbf{S}_{i,\text{WG}}^{(g)}) - \text{LMB}_{12.5}(\mathbf{S}_{i,\text{base}}^{(g)}) \right]. \quad (14)$$

Here, $\hat{y}_{i,\text{base}}^{(g)}$ and $\hat{y}_{i,\text{WG}}^{(g)}$ denote the predicted categories of the matched baseline and WG-SLE, respectively. The paired maps $\mathbf{S}_{i,\text{base}}^{(g)}$ and $\mathbf{S}_{i,\text{WG}}^{(g)}$ are generated from the same evaluation clip and target the shared predicted-class logit. We report $\Delta H^{(g)}$ directly because H is a normalized entropy, whereas changes in Top10 and LMB_{12.5} are expressed in percentage points because both quantities are response fractions.

Within the prediction-agreement subsets, the three Grad-CAM summaries characterize complementary aspects of classifier-response redistribution. For both representative backbones, $\Delta H < 0$ indicates reduced response dispersion along the mel-frequency

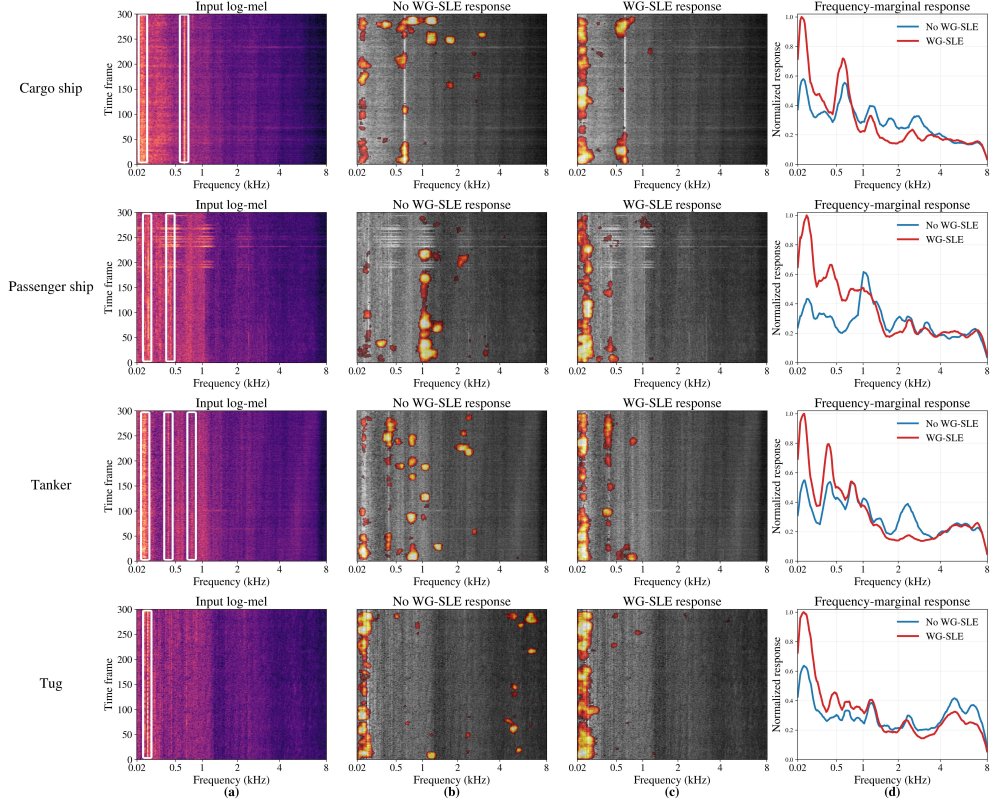


Fig. 5 DeepShip AST-B/384 Grad-CAM diagnostics for four fixed, class-balanced evaluation clips selected from the prediction-agreement subset. Columns show (a) the input log-mel spectrogram, (b) the matched-baseline response map, (c) the WG-SLE response map, and (d) the corresponding frequency-marginal response curves, with the matched baseline in blue and WG-SLE in red. Within each row, both response maps use the same clip and target the shared predicted-class logit. The top-to-bottom order is cargo, passenger, tanker, and tug.

axis, while positive ΔTop10 indicates greater concentration in the strongest time-frequency response locations. Positive $\Delta \text{LMB}_{12.5}$ indicates that a greater fraction of the total classifier response is assigned to the lowest 12.5% of mel-frequency bins. These changes are larger for AST-B/384 than for ResNet-18.

For qualitative visualization, Figs. 5 and 6 show four fixed, class-balanced evaluation clips for each backbone, with one clip selected from each DeepShip true class within the corresponding prediction-agreement subset. These examples illustrate class-specific response patterns and are not used to compute the aggregate values in Table 8.

Figure 5 shows the AST-B/384 examples. The displayed maps exhibit a redistribution toward lower mel-frequency regions containing visible line-like tonal traces, while class-dependent secondary response peaks remain.

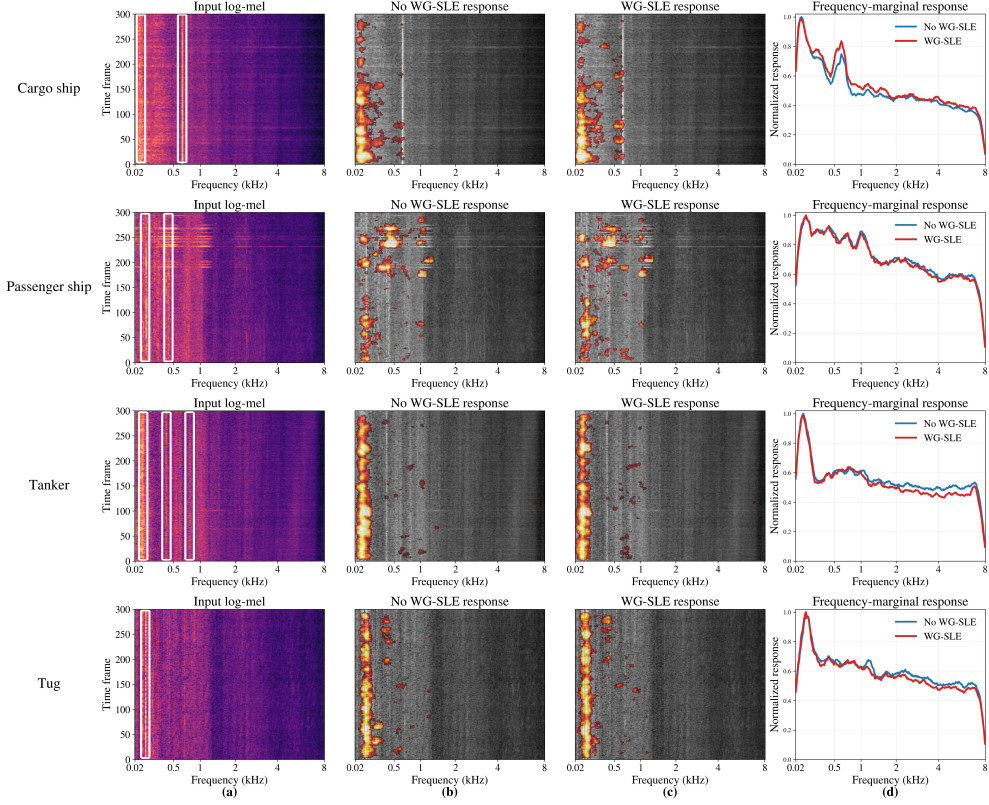


Fig. 6 DeepShip ResNet-18 Grad-CAM diagnostics for four fixed, class-balanced evaluation clips selected from the prediction-agreement subset. Columns show (a) the input log-mel spectrogram, (b) the matched-baseline response map, (c) the WG-SLE response map, and (d) the corresponding frequency-marginal response curves, with the matched baseline in blue and WG-SLE in red. Within each row, both response maps use the same clip and target the shared predicted-class logit. The top-to-bottom order is cargo, passenger, tanker, and tug.

Figure 6 shows the corresponding ResNet-18 examples. The displayed response redistribution toward lower mel-frequency regions is less pronounced than in the AST-B/384 examples, consistent with the smaller aggregate Grad-CAM changes for ResNet-18.

3 Discussion

The comparisons indicate that WG-SLE is most useful when the delay-predictable spectral estimate is integrated with, rather than substituted for, the recognition input. Conventional LMS-ALE optimizes the enhancement stage independently and supplies the line-enhanced signal as the recognizer input, whereas WG-SLE refines class-indexed spectral prototypes within the training framework and introduces the delay-predictable spectral estimate through scaled residual fusion. This distinction is important because delayed-reference predictability identifies correlation across a

selected delay rather than the complete set of spectral information relevant to vessel recognition. Retaining the observed spectrum as the additive base therefore allows the enhancement path to contribute complementary information without making it the sole recognition input.

The ablation study supports this interpretation. The degradation of the enhancement-path-only configuration shows that the delay-predictable spectral estimate is insufficient by itself, while the initialization and objective ablations indicate that Wiener initialization is most effective when followed by trainable refinement under the Wiener prediction objective and sparsity regularization. The stop-gradient design separates direct classification updates from spectral-prototype refinement, although the current ablations do not isolate its effect independently of the other components.

The diagnostic analyses provide a consistent but backbone-dependent view of the resulting representation. Grad-CAM indicates stronger redistribution of classifier response toward lower mel-frequency regions for AST-B/384 than for ResNet-18, while the t-SNE projections show reduced overlap only in the displayed two-dimensional embeddings. Spectral-prototype alignment further shows that the trained prototypes remain more closely aligned with their Wiener anchors in the 40–1000 Hz analysis range while still undergoing nonzero updates. Together, these observations are consistent with prediction-guided spectral-prototype refinement that retains the delayed-reference prediction role during optimization.

Class indexing does not assign a different physical rule to each vessel class; the same delayed-reference prediction criterion is used to summarize class-conditioned delay-to-current spectral statistics. However, training uses label-selected spectral operators, so the current experiments do not isolate whether class conditioning also provides auxiliary discriminative structure. A shared global prototype or training-time query-routing comparison would help separate this effect. The inference-time routing rule is also not evaluated independently, and the magnitude-based distance can reflect both response scale and shape.

The noisy-evaluation and cross-backbone experiments extend the analysis beyond the nominal matched setting. The former examines sensitivity to controlled additive perturbations using clean-trained checkpoints, while the latter shows that the same WG-SLE interface can be coupled to convolutional and transformer-style recognizers without modifying their internal classifiers. These analyses indicate that the magnitude of the WG-SLE effect remains dependent on the perturbation condition, backbone, and dataset.

Several limitations remain. Most main analyses use fixed-seed, analysis-specific checkpoints, whereas only the noisy-evaluation and confusion-matrix analyses use three matched seeds. In addition, label-selected training operators and soft-routed inference operators introduce a train–inference mismatch that warrants routing-specific baselines. Overall, WG-SLE is best interpreted as a physics-guided, trainable spectral front end in which a delay-predictable spectral estimate complements the observed spectrum rather than replacing it.

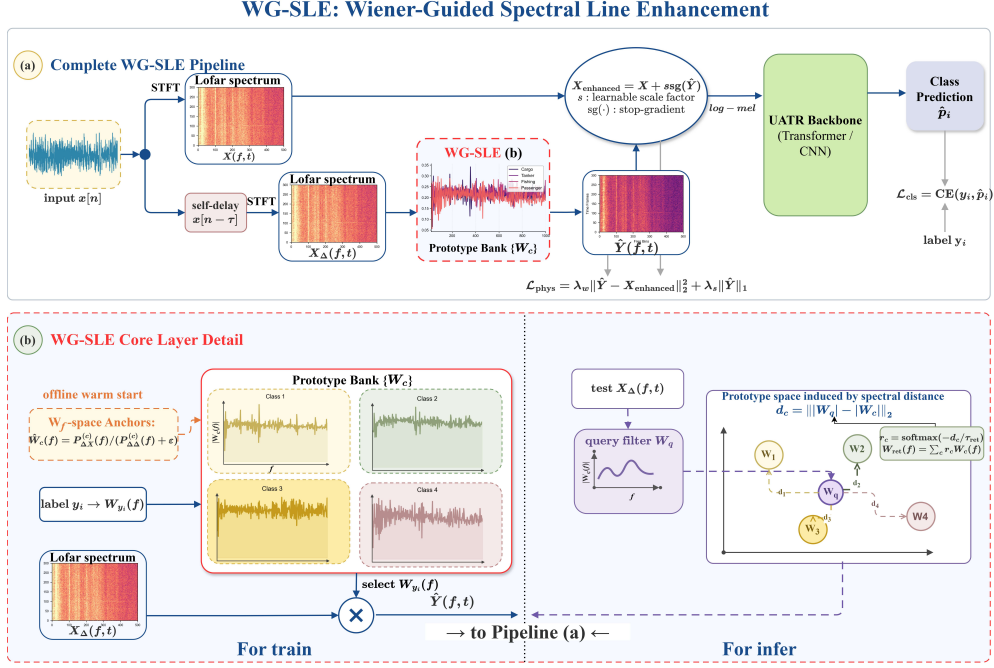


Fig. 7 Overview of WG-SLE, including Wiener-initialized class-indexed spectral prototypes, label-based training selection, inference-time query routing, and scaled residual fusion.

4 Methods

WG-SLE is a physics-guided, trainable spectral front end for downstream vessel recognition. As shown in Fig. 7, a delayed spectrum first passes through a frequency-wise complex operator to form a delay-predictable spectral estimate. During training, the class label selects this operator from a bank of Wiener-initialized spectral prototypes; at inference, a sample-derived query filter instead routes the same bank without access to the label. Class indexing applies a shared delayed-reference prediction criterion while allowing each prototype to summarize class-conditioned delay-to-current spectral statistics. The downstream recognizer receives the resulting fused log-mel representation rather than the class label. The estimate then enters a scaled residual path whose base term is the observed spectrum. This forward interface coordinates three distinct roles: delayed-reference prediction constructs the complementary estimate, residual fusion introduces it without making it the sole input, and the downstream network performs recognition. Stop-gradient preserves this role separation by blocking the direct gradient path from the classification loss to the spectral prototypes. The delayed-reference prediction criterion provides the acoustic guidance, the recording-level Wiener estimates provide an analytical initialization, and sparsity regularization promotes a sparse estimate. The resulting log-mel representation provides a common interface to different downstream recognizers.

4.1 Signal model and frequency-wise delayed-reference prediction

Let i index an acoustic clip, let $n \in \{0, \dots, N_i - 1\}$ be its discrete-time sample index, and let $y_i \in \{1, \dots, C\}$ be its label among C recognition classes. The observed waveform $x_i[n] \in \mathbb{R}$ is written conceptually as

$$x_i[n] = q_i[n] + r_i[n], \quad (15)$$

Here, $q_i[n]$ denotes a conceptual persistent narrowband contribution and $r_i[n]$ the remaining source, propagation, and ambient contributions. This decomposition is used only to formulate the delayed-reference correlation contrast; neither component is observed separately or used as a training target. Delay predictability is introduced by the assumption below.

Let $X_i \in \mathbb{C}^{F_s \times T_s}$ be the one-sided short-time Fourier transform (STFT) of $x_i[n]$, where F_s and T_s are the numbers of frequency bins and frames, respectively. The coefficient at frequency-bin index $f \in \{0, \dots, F_s - 1\}$ and frame index $t \in \{0, \dots, T_s - 1\}$ is denoted by $X_i(f, t)$. By linearity of the STFT,

$$X_i(f, t) = Q_i(f, t) + R_i(f, t), \quad (16)$$

where $Q_i(f, t)$ and $R_i(f, t)$ are the STFTs of $q_i[n]$ and $r_i[n]$, respectively.

The principle of adaptive line enhancement (ALE) is based on a relative correlation contrast across a selected delay. We therefore construct delayed references along the STFT-frame dimension:

With leading frames zero-padded when $t - \tau_\Delta < 0$, where $\tau_\Delta \in \mathbb{N}$ denotes the frame delay, the implementation uses $\tau_\Delta = 1$ STFT frame. Given the 1600-sample hop at 16 kHz, this frame shift is 0.100 s. Delayed-reference predictability is defined on the STFT representation used for recognition. The resulting predictor output is therefore interpreted as a delay-predictable spectral estimate rather than an isolated tonal source.

$$\begin{aligned} X_{i,\Delta}(f, t) &= \begin{cases} X_i(f, t - \tau_\Delta), & t \geq \tau_\Delta, \\ 0, & t < \tau_\Delta, \end{cases} \\ Q_{i,\Delta}(f, t) &= Q_i(f, t - \tau_\Delta), \quad t \geq \tau_\Delta, \\ R_{i,\Delta}(f, t) &= R_i(f, t - \tau_\Delta), \quad t \geq \tau_\Delta. \end{aligned} \quad (17)$$

Correlation-contrast assumption

The formulation permits a structured remainder and assumes a relative, rather than exclusive, delay-to-current correlation contrast. In Eqs. (18)–(22), $\mathbb{E}_{i,t}[\cdot]$ denotes expectation over the clip-frame distribution restricted to valid delayed frames $t \in \{\tau_\Delta, \dots, T_s - 1\}$; the finite-sample implementation is defined later through explicit recording- and clip-level accumulations. For frequency bins containing the conceptual persistent narrowband component, its delay-to-current correlation is assumed to exceed that of the remaining observed component:

$$|\mathbb{E}_{i,t}[Q_{i,\Delta}^*(f, t)Q_i(f, t)]| > |\mathbb{E}_{i,t}[R_{i,\Delta}^*(f, t)R_i(f, t)]|. \quad (18)$$

The cross-correlations between the targeted component and the remaining observed component are assumed to be weak:

$$\mathbb{E}_{i,t} [Q_{i,\Delta}^*(f,t)R_i(f,t)] \approx 0, \quad \mathbb{E}_{i,t} [R_{i,\Delta}^*(f,t)Q_i(f,t)] \approx 0. \quad (19)$$

Under this modeling assumption, the delay-to-current cross-power $\mathbb{E}_{i,t}[X_{i,\Delta}^*(f,t)X_i(f,t)]$ emphasizes components that remain correlated across the selected delay. It need not contain only the conceptual component Q_i , which is why the resulting quantity is described as a delay-predictable spectral estimate.

Under a diagonal per-frequency approximation, delayed-reference prediction becomes a scalar complex regression at each STFT frequency bin. This one-tap spectral formulation retains the delayed-reference prediction principle of ALE but does not reproduce the multi-tap adaptive-filter structure of conventional time-domain LMS-ALE. Given a spectral operator $W(f) \in \mathbb{C}$, its mathematical output is the delay-predictable spectral estimate

$$\hat{Y}_i(f,t) = W(f)X_{i,\Delta}(f,t), \quad (20)$$

and the corresponding mean-square criterion is

$$\mathcal{J}_f(W) = \mathbb{E}_{i,t} \left[|W(f)X_{i,\Delta}(f,t) - X_i(f,t)|^2 \right]. \quad (21)$$

This criterion is a frequency-wise complex least-squares regression from the delayed spectrum to the current spectrum. Applying Wirtinger differentiation with respect to $W^*(f)$ gives the theoretical unguarded least-squares solution when the delayed-reference power is nonzero:

$$W^*(f) = \frac{\mathbb{E}_{i,t} [X_{i,\Delta}^*(f,t)X_i(f,t)]}{\mathbb{E}_{i,t} [|X_{i,\Delta}(f,t)|^2]}, \quad (22)$$

where $(\cdot)^*$ denotes complex conjugation, the superscript $\star$ denotes the minimizer, and the denominator is assumed positive. Equation (22) is the unguarded least-squares solution; the recording-level initialization and sample-derived query filter use guarded empirical ratios with a numerical denominator floor $\varepsilon > 0$, as defined in Eqs. (35) and (36).

Wiener-anchor derivation

For a fixed frequency bin, write $X_\Delta = X_{i,\Delta}(f,t)$ and $X = X_i(f,t)$. The criterion $\mathcal{J}_f(W)$ in Eq. (21) can then be expanded as

$$\mathcal{J}_f(W) = \mathbb{E}_{i,t} [(W(f)X_\Delta - X)(W^*(f)X_\Delta^* - X^*)] \quad (23)$$

$$= \mathbb{E}_{i,t} [|W(f)|^2 |X_\Delta|^2 - W(f)X_\Delta X^* - W^*(f)X_\Delta^* X + |X|^2]. \quad (24)$$

Treating $W(f)$ and $W^*(f)$ as independent variables under Wirtinger differentiation gives

$$\frac{\partial \mathcal{J}_f}{\partial W^*(f)} = W(f) \mathbb{E}_{i,t} [|X_\Delta|^2] - \mathbb{E}_{i,t} [X_\Delta^* X]. \quad (25)$$

Setting Eq. (25) to zero gives Eq. (22). The implementation instead applies the numerical floor to finite-sample accumulated delayed-reference energy, first for each training recording and then for each inference query, as specified in Secs. 4.3 and 4.4. Both the theoretical expression and these guarded empirical estimates are observed-spectrum regressors because the current and delayed spectra can contain components from the conceptual remainder.

Equation (22) defines the theoretical frequency-wise complex operator for delayed-reference prediction wherever the delayed-reference power is nonzero. In WG-SLE, a guarded empirical counterpart provides the class-wise Wiener anchors that initialize the spectral-prototype bank; subsequent spectral-prototype refinement is controlled by the Wiener prediction objective and sparsity regularization rather than by an explicit distance-to-initialization penalty.

4.2 WG-SLE spectral operator layer

The frequency-wise Wiener form in Eq. (22) motivates representing the WG-SLE front end as a set of trainable frequency-wise complex operators. WG-SLE maintains the spectral-prototype bank

$$\mathcal{W} = \{W_c\}_{c=1}^C, \quad W_c = \{W_c(f)\}_{f=0}^{F_s-1} \in \mathbb{C}^{F_s}. \quad (26)$$

Each spectral prototype is a frequency-wise complex operator whose role is spectral enhancement before classification, distinct from an embedding centroid or class logit. The delayed-reference prediction criterion is shared across classes; class labels partition the training recordings so that each prototype summarizes class-conditioned delay-to-current spectral statistics. For vessel classes, these statistics may reflect machinery and propulsion patterns, whereas the ambient-background prototype summarizes persistent structure within the background class. The prototype bank supplies class-conditioned spectral operators, and the downstream recognizer produces the vessel-category logits.

To make the complex operators compatible with gradient-based optimization, each spectral prototype is represented by bounded real and imaginary parameters:

$$W_c(f) = B [\tanh(\theta_{c,\text{re}}(f)) + j \tanh(\theta_{c,\text{im}}(f))], \quad (27)$$

where $\theta_{c,\text{re}}(f)$ and $\theta_{c,\text{im}}(f)$ are real-valued trainable parameters for the real and imaginary responses, $j = \sqrt{-1}$, and $B > 0$ bounds each component of the complex operator. Hence $|\text{Re } W_c(f)| < B$, $|\text{Im } W_c(f)| < B$, and $|W_c(f)| < \sqrt{2}B$; B is not a direct bound on the complex magnitude. The spectral-prototype parameter tensor is

$$\Theta \in \mathbb{R}^{C \times F_s \times 2}. \quad (28)$$

Here, the first, second, and third tensor axes index recognition class, STFT frequency bin, and real/imaginary component, respectively. This parameterization keeps the complex response bounded while preserving a direct frequency-bin interpretation.

During training, the class label selects the spectral prototype used in the enhancement path:

$$W_i(f) = W_{y_i}(f). \quad (29)$$

The selected spectral operator is applied to the delayed spectrum as

$$\hat{Y}_i(f, t) = W_i(f)X_{i,\Delta}(f, t), \quad (30)$$

which follows the frequency-wise mapping introduced in Eq. (20). The initialization of the spectral prototypes from class-wise Wiener anchors is described in Sec. 4.3. At inference, when y_i is unavailable, inference-time query routing determines the spectral operator as described in Sec. 4.4.

4.3 Wiener initialization of spectral prototypes

The Wiener expression in Eq. (22) provides the initialization form for the spectral-prototype bank. The implementation first estimates one guarded Wiener filter per training recording and then takes the arithmetic mean of the recording-level filters within each class; this class-conditioned average is termed the class-wise Wiener anchor. The recording-equal construction prevents recordings that yield more clips from dominating the anchor.

Let

$$\mathcal{R}_c = \{\rho : y_\rho = c\} \quad (31)$$

denote the set of training recordings from class c , with ρ indexing a recording and y_ρ its class. Let $\mathcal{I}_\rho$ be the set of training clips extracted from recording ρ . The implementation accumulates the recording-level delay-to-current cross-product and delayed-reference energy as

$$\hat{S}_{\Delta X}^{(\rho)}(f) = \sum_{i \in \mathcal{I}_\rho} \sum_{t=0}^{T_s-1} X_{i,\Delta}^*(f, t) X_i(f, t), \quad (32)$$

and

$$\hat{S}_{\Delta \Delta}^{(\rho)}(f) = \sum_{i \in \mathcal{I}_\rho} \sum_{t=0}^{T_s-1} |X_{i,\Delta}(f, t)|^2. \quad (33)$$

The guarded recording-level filter and the resulting class-wise anchor are

$$\hat{W}^{(\rho)}(f) = \frac{\hat{S}_{\Delta X}^{(\rho)}(f)}{\max\{\hat{S}_{\Delta \Delta}^{(\rho)}(f), \varepsilon\}}, \quad (34)$$

$$\hat{W}_c(f) = \frac{1}{|\mathcal{R}_c|} \sum_{\rho \in \mathcal{R}_c} \hat{W}^{(\rho)}(f), \quad (35)$$

where $|\mathcal{R}_c|$ is the number of training recordings in class c , $\hat{W}^{(\rho)}(f)$ is the guarded Wiener estimate for recording ρ , and $\bar{W}_c(f)$ is their arithmetic mean within class c . Because the leading τ_Δ delayed frames are zero, their contributions to both accumulators are zero; summing all frames is therefore numerically equivalent to summing only valid delayed frames for these ratios. The implementation uses $\varepsilon = 10^{-10}$ for the accumulated-energy floor.

Because the averaging occurs after the recording-level ratios are formed, Eq. (35) is not generally equal to a single ratio of class-pooled cross-product and delayed-reference energy. Because the class anchor is formed by averaging recording-level ratios, it is generally different from a ratio computed after pooling all recordings in the class. The mapping from the complex anchors to the bounded parameters in Eq. (27) and their subsequent spectral-prototype refinement are described below.

4.4 Inference-time spectral-prototype routing

At inference, the class label is unavailable, and WG-SLE cannot directly select $W_{y_i}(f)$. A sample-derived query filter is therefore estimated from the test sample and compared with the spectral-prototype bank to produce soft routing weights.

For a sample i at inference, the sample-derived query filter $W_{\text{qry},i} \in \mathbb{C}^{F_s}$ is the guarded per-frequency ratio

$$W_{\text{qry},i}(f) = \frac{\sum_{t=0}^{T_s-1} X_{i,\Delta}^*(f, t) X_i(f, t)}{\max\left\{\sum_{t=0}^{T_s-1} |X_{i,\Delta}(f, t)|^2, \varepsilon\right\}}. \quad (36)$$

The sample-derived query filter is compared with the learned spectral prototypes using magnitude-response distances:

$$d_{i,c} = ||W_{\text{qry},i}| - |W_c|||_2, \quad c = 1, \dots, C. \quad (37)$$

The scalar $d_{i,c} \in \mathbb{R}_{\geq 0}$ is the Euclidean distance between the raw magnitude responses over all STFT frequency bins; no query clipping, magnitude normalization, frequency weighting, or phase alignment is applied. The query ratio is therefore not bounded by the prototype parameterization in Eq. (27), and the distance can reflect both response scale and shape.

These distances are converted into soft routing weights:

$$\alpha_{i,c} = \frac{\exp(-d_{i,c}/\tau_{\text{ret}})}{\sum_{c'=1}^C \exp(-d_{i,c'}/\tau_{\text{ret}})}, \quad (38)$$

where $\alpha_{i,c} \in (0, 1)$, $\sum_c \alpha_{i,c} = 1$, and $\tau_{\text{ret}} > 0$ is the routing temperature, fixed to 1.0 in the evaluated configuration. The routed spectral operator $W_{\text{ret},i} \in \mathbb{C}^{F_s}$ is then formed as the weighted combination

$$W_{\text{ret},i}(f) = \sum_{c=1}^C \alpha_{i,c} W_c(f). \quad (39)$$

The corresponding delay-predictable spectral estimate is

$$\hat{Y}_i(f, t) = W_{\text{ret}, i}(f) X_{i, \Delta}(f, t). \quad (40)$$

The magnitude response is used only to define the routing distance. The routed spectral operator remains complex-valued because the soft routing weights combine the complex spectral prototypes $W_c(f)$, including their learned real and imaginary components. This enables inference-time spectral-prototype routing when class labels are unavailable.

This routing step determines the spectral operator used for enhancement. Training uses the label-selected operator in Eq. (29), whereas inference uses the soft mixture in Eq. (39); the resulting operator-selection rules are therefore not identical, as considered further in Discussion.

4.5 Fusion and training objective

Let $\widetilde{W}_i(f)$ denote the spectral operator used for sample i . During training, $\widetilde{W}_i(f) = W_{y_i}(f)$; during inference, $\widetilde{W}_i(f) = W_{\text{ret}, i}(f)$. The corresponding delay-predictable spectral estimate is

$$\hat{Y}_i(f, t) = \widetilde{W}_i(f) X_{i, \Delta}(f, t). \quad (41)$$

WG-SLE incorporates this estimate through scaled residual fusion:

$$X_{i, \text{enh}}(f, t) = X_i(f, t) + s \text{sg}(\hat{Y}_i(f, t)), \quad (42)$$

where $s \in \mathbb{R}$ is a single unconstrained learnable scalar shared across samples, frequency bins, and frames and initialized to 1.0. The operator $\text{sg}(\cdot)$ preserves its argument in the forward pass and sets its derivative to zero in the backward pass. Equation (42) keeps the observed spectrum as an explicit additive base term and prevents the estimate from becoming the sole recognizer input. The observed spectrum is the additive base term in Eq. (42). Because fusion is performed in the complex STFT domain, the power of the fused spectrum also contains a cross term:

$$|X_i + s\hat{Y}_i|^2 = |X_i|^2 + s^2|\hat{Y}_i|^2 + 2s \text{Re}(X_i^* \hat{Y}_i), \quad (43)$$

where all terms are evaluated elementwise at (f, t) . The cross term permits constructive or destructive interference, and the unconstrained scale may become negative. Stop-gradient prevents the classification loss from directly updating the spectral prototypes through the enhancement path.

The fused spectrum is converted to the log-mel input required by the downstream recognizer. Let $\mathbf{M} \in \mathbb{R}_{\geq 0}^{F_m \times F_s}$ be the F_m -bin mel filterbank matrix, with $F_m = 300$. The mapping used in all paired comparisons is

$$\Phi(X) = [10 \log_{10} (\max\{\mathbf{M}|X|^2, 10^{-10}\})]^\top, \quad (44)$$

where $|X|^2 \in \mathbb{R}_{\geq 0}^{F_s \times T_s}$, $\mathbf{M}|X|^2 \in \mathbb{R}_{\geq 0}^{F_m \times T_s}$, and the scalar floor is applied elementwise. The transpose $(\cdot)^\top$ produces the $T_s \times F_m$ time-by-mel layout. Before input to a downstream backbone, this log-mel representation is resized to the backbone’s nominal spatial resolution by the preprocessing utilities of the `timmm` library. The same resizing

pipeline is used for the matched unenhanced, LMS-ALE, and WG-SLE configurations of a given backbone. The classifier logits are

$$\mathbf{z}_i = h_\psi(\Phi(X_{i,\text{enh}})), \quad (45)$$

where h_ψ denotes the recognition backbone and classification head, parameterized by ψ , and $\mathbf{z}_i \in \mathbb{R}^C$ contains the C unnormalized class logits. Let $\mathcal{I}_{\text{mb}}$ be a mini-batch index set of size N_{mb} . The softmax probability assigned to class c is

$$p_{i,c} = \frac{\exp(z_{i,c})}{\sum_{c'=1}^C \exp(z_{i,c'})}, \quad c = 1, \dots, C, \quad (46)$$

and the predicted category is

$$\hat{y}_i = \arg \max_{c \in \{1, \dots, C\}} z_{i,c}. \quad (47)$$

The cross-entropy (CE) classification loss, averaged over the mini-batch, is

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N_{\text{mb}}} \sum_{i \in \mathcal{I}_{\text{mb}}} \log p_{i,y_i}. \quad (48)$$

The Wiener prediction objective and sparsity regularization refine the spectral prototypes while retaining their delayed-reference prediction role. For each mini-batch, the implementation takes arithmetic means over clips, all F_s frequency bins, and all T_s frames:

$$\mathcal{L}_{\text{wiener}} = \frac{1}{N_{\text{mb}} F_s T_s} \sum_{i \in \mathcal{I}_{\text{mb}}} \sum_{f=0}^{F_s-1} \sum_{t=0}^{T_s-1} \left| \widetilde{W}_i(f) X_{i,\Delta}(f,t) - X_i(f,t) \right|^2. \quad (49)$$

This term implements the delayed-reference prediction principle derived in Sec. 4.1. The zero-padded leading frames are included in the mean; because $X_{i,\Delta} = 0$ there, they add an operator-independent term and provide no spectral-prototype gradient. Under the correlation-contrast assumption, the regression objective favors components that remain predictable across the selected delay. Its target is the observed spectrum $X_i(f,t) = Q_i(f,t) + R_i(f,t)$, so it can also retain delay-predictable content from the remaining observed component.

WG-SLE also applies an ℓ_1 magnitude penalty to the delay-predictable spectral estimate:

$$\mathcal{L}_{\text{sparse}} = \frac{1}{N_{\text{mb}} F_s T_s} \sum_{i \in \mathcal{I}_{\text{mb}}} \sum_{f=0}^{F_s-1} \sum_{t=0}^{T_s-1} \left| \hat{Y}_i(f,t) \right|. \quad (50)$$

This term is a differentiable sparsity-promoting surrogate that shrinks coefficients of the delay-predictable spectral estimate toward zero, while the Wiener term maintains the delayed-reference prediction objective.

The full training objective is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_w \mathcal{L}_{\text{wiener}} + \lambda_s \mathcal{L}_{\text{sparse}}, \quad (51)$$

Table 9 Dataset composition and partitioning protocol.

Characteristic	DeepShip	ShipsEar
Recordings	509	86
Duration	47 h 4 min	Approx. 3 h
Recognition labels	cargo, passenger, tanker, tug	Passenger vessel, small fishing vessel, tugboat, small motor vessel, ambient background
Split	7:3 recording-level	7:3 recording-level
Segmentation	30.00 s, no overlap	30.00 s, 50% overlap

where λ_w and λ_s weight the Wiener prediction objective and sparsity regularization, respectively. Because Eq. (42) applies stop-gradient to $\hat{Y}_i(f, t)$, $\mathcal{L}_{\text{cls}}$ provides no direct gradient to the spectral prototypes through the enhancement path. The Wiener prediction objective and sparsity regularization refine the spectral prototypes, whereas the classification loss optimizes the downstream recognizer and residual scale. Thus, WG-SLE uses a shared training framework with role-separated optimization rather than classification-driven spectral-prototype refinement.

Conventional ALE often updates its filter with an LMS-type online adaptive rule. WG-SLE instead represents each frequency-wise complex spectral operator as a differentiable neural layer. The operator $\widetilde{W}_i(f)$ is generated from trainable real-valued parameters through the bounded complex parameterization in Eq. (27), and $\hat{Y}_i(f, t) = \widetilde{W}_i(f)X_{i,\Delta}(f, t)$ is differentiable with respect to those parameters. The Wiener prediction objective and sparsity regularization can therefore refine the spectral prototypes through standard gradient-based optimization. At inference, the routed spectral operator is applied directly to the delayed spectrum without online LMS adaptation.

4.6 Experimental setup

4.6.1 Datasets and experimental protocol

The evaluation uses DeepShip [36] and ShipsEar [37]. DeepShip contains four commercial-vessel classes—cargo, passenger, tanker, and tug—and serves as the primary benchmark. Within ShipsEar, the original vessel-type taxonomy is organized into five broader recognition labels used in this study. These five labels include an ambient background class in addition to the four vessel-related labels listed in Table 9. The dataset and partitioning settings are summarized in Table 9.

Recordings are partitioned before segmentation, such that all clips derived from a given recording are assigned exclusively to either the training or evaluation partition. An approximate 7:3 recording-level split is used for both datasets. DeepShip recordings are divided into nonoverlapping 30.00 s clips, whereas ShipsEar recordings are segmented into 30.00 s clips with 50% overlap within each partition. Consequently, no recording contributes clips to both partitions, preventing direct clip- or recording-level leakage between training and evaluation. The checkpoint-selection procedure applied to the evaluation partition is described below.

Table 10 Core preprocessing and optimization settings.

Parameter	Setting	Parameter	Setting
Sampling rate	16 kHz	Input duration	30.00 s
			480000 samples
FFT size	4096	Window length	0.20 s
			3200 samples
Hop length	0.10 s	Mel bins	300
	1600 samples		
STFT window	Hamming	STFT centering	Enabled
Optimizer	AdamW	Learning rate	2×10^{-5}
Weight decay	1×10^{-5}	Batch size	16
Epochs	50	Multi-seed diagnostics	42, 43, 44

The core preprocessing and optimization settings are listed in Table 10. These settings are kept identical within each paired comparison between the matched baseline and WG-SLE.

Multichannel recordings are averaged to mono and mean-centered in the waveform domain. No additional peak, root-mean-square, or variance normalization is applied on the standard input path. The centered one-sided STFT uses a Hamming window and produces $F_s = 2049$ frequency bins and $T_s = 301$ frames for each 30.00 s clip. After log-mel conversion, resizing to each backbone’s nominal spatial resolution is performed by the preprocessing utilities of the `timm` library and is held fixed within each matched comparison. All main comparisons use the same data split, input representation, downstream recognizer, optimization settings, training budget, random seed, and checkpoint-selection rule within each matched pair. The front-end, cross-backbone, ablation, spectral-prototype alignment, and Grad-CAM results use seed-42 checkpoints, whereas the noisy-evaluation and confusion-matrix analyses use seeds 42, 43, and 44 as stated in their captions.

No independent validation partition is used. For each run, the checkpoint with the highest OA on the evaluation partition is reported; the resulting values are therefore checkpoint-selected results on that partition.

For the spectral-prototype alignment diagnostic, let $f_{\text{samp}} = 16 \text{ kHz}$ denote the sampling rate and $N_{\text{FFT}} = 4096$ the FFT size. The center frequency of one-sided STFT bin f is

$$\nu_f = \frac{f f_{\text{samp}}}{N_{\text{FFT}}}, \quad f = 0, \dots, F_s - 1. \quad (52)$$

Under the stated sampling and FFT configuration, the one-sided STFT spans 0–8 kHz, with $8 \text{ kHz} = f_{\text{samp}}/2$ corresponding to the Nyquist frequency. For spectral-prototype alignment, we define two fixed STFT-domain analysis ranges:

$$\begin{aligned} \mathcal{B}_{40:8000} &= \{f \in \{0, \dots, F_s - 1\} : 40 \text{ Hz} \leq \nu_f \leq 8000 \text{ Hz}\}, \\ \mathcal{B}_{40:1000} &= \{f \in \{0, \dots, F_s - 1\} : 40 \text{ Hz} \leq \nu_f \leq 1000 \text{ Hz}\}. \end{aligned} \quad (53)$$

These fixed ranges are used only for spectral-prototype alignment; WG-SLE operates over the complete one-sided STFT representation. The STFT-domain range

$\mathcal{B}_{40:1000}$ is distinct from the lowest 12.5% of mel-frequency bins used in the Grad-CAM summary.

For noisy evaluation, three matched DeepShip AST-B/384 checkpoint pairs trained on the original DeepShip recordings are frozen and evaluated at -5 , 0 , 5 , and 10 dB, with none of the synthetic noise considered here introduced during training. White Gaussian noise is used as a spectrally flat reference. Let ξ_k denote the positive Fourier frequency associated with noise-synthesis bin k . For the colored-noise conditions, the DC-bin amplitude is set to zero. Brown-noise amplitudes at positive-frequency bins are weighted in proportion to ξ_k^{-1} , corresponding to a power spectrum proportional to ξ_k^{-2} , whereas the Wenz-shaped condition uses amplitude weights proportional to $\xi_k^{-0.85}$ as a simplified colored-noise approximation motivated by classical frequency-dependent ocean ambient-noise spectra [7]. Each generated noise sequence is mean-centered and scaled according to waveform mean-square power to achieve the specified SNR before addition. For each random seed and evaluation record, the matched baseline and WG-SLE models receive the same perturbed waveform in the same order.

The main front-end, backbone, and ablation comparisons use a fixed random seed for controlled method-to-method evaluation under matched training and evaluation conditions. The noisy evaluation additionally examines whether the paired WG-SLE advantage remains stable under stochastic additive perturbations; therefore, three matched random-seed pairs are used for this analysis. For each noise family, the same three seed pairs are evaluated consistently across all four SNR levels and metrics. Table 6 reports the mean paired WG-SLE-minus-baseline differences across the three seeds, whereas Fig. 2 additionally shows the individual paired values and their sample standard deviations.

For the class-level diagnostic, paired baseline and WG-SLE predictions are examined on DeepShip using AST-B/384 [25] and ResNet-18 [26], with three matched random seeds evaluated for each backbone. Each panel uses the corresponding evaluation manifest and reports the mean over the three matched seeds. Within each matched pair, sample identifiers, ordering, and targets are required to agree exactly, and the corresponding training and evaluation manifests contain no clip- or recording-level overlap.

For each random seed, confusion counts are row-normalized within each DeepShip true class before computing the WG-SLE-minus-matched-baseline difference. Each displayed matrix reports the mean of these paired differences across the three matched random seeds. Individual clips are not treated as independent replications; the matrices are used as descriptive class-level diagnostics of the matched checkpoint pairs.

For Grad-CAM analysis, each backbone-specific quantitative set comprises all evaluation clips for which the matched baseline and WG-SLE predict the same category. Both maps in a pair target this shared predicted-class logit. Maps are bilinearly interpolated to the log-mel grid; quantitative summaries are calculated from nonnegative, unclipped maps using the total-response normalization defined in the corresponding metrics. Clipping at the 99th percentile and min-max normalization are used only for visual display. The qualitative figures use four fixed, class-balanced clips per backbone, one from each DeepShip true class, selected from the corresponding prediction-agreement subset.

4.6.2 Evaluation metrics

Recognition performance is evaluated using overall accuracy (OA), average accuracy (AA), Cohen’s κ coefficient [38], and macro-F1 (MF1). Let M_{ab} denote the number of samples whose true class is a and predicted class is b , C the number of classes, and $N = \sum_{a=1}^C \sum_{b=1}^C M_{ab}$ the total number of evaluation samples. The four metrics are defined as follows:

$$\text{OA} = \frac{\sum_{c=1}^C M_{cc}}{N}, \quad (54)$$

$$\text{AA} = \frac{1}{C} \sum_{c=1}^C \frac{M_{cc}}{\sum_{b=1}^C M_{cb}}, \quad (55)$$

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad p_o = \text{OA}, \quad p_e = \frac{1}{N^2} \sum_{c=1}^C \left(\sum_b M_{cb} \right) \left(\sum_a M_{ac} \right), \quad (56)$$

$$\text{MF1} = \frac{1}{C} \sum_{c=1}^C \frac{2 \text{Prec}_c \text{Rec}_c}{\text{Prec}_c + \text{Rec}_c}, \quad (57)$$

where

$$\text{Prec}_c = \frac{M_{cc}}{\sum_a M_{ac}}, \quad \text{Rec}_c = \frac{M_{cc}}{\sum_b M_{cb}}. \quad (58)$$

OA measures overall classification accuracy, whereas AA and MF1 give equal weight to individual classes. Cohen’s κ further accounts for agreement expected by chance. If a precision or recall denominator is zero, the corresponding class-wise quantity is set to zero; if $\text{Prec}_c + \text{Rec}_c = 0$, the class-wise F1 term is also set to zero. OA, AA, and MF1 lie in $[0, 1]$ and are multiplied by 100 when reported as percentages in tables. Superscript changes for these metrics are percentage-point differences, whereas superscript changes for κ are absolute differences; arrows indicate numerical direction and do not denote statistical significance.

Additional quantities are reported only for diagnostic analysis. The mean silhouette coefficient is computed on the two-dimensional t-SNE projections in Fig. 4 and is used only as a projection-dependent descriptive summary. Complex cosine similarity and relative L_2 update measure the alignment between trained spectral prototypes and their Wiener anchors. Grad-CAM maps are characterized using three complementary summaries in Eqs. (8)–(10). The normalized frequency-marginal entropy H measures the dispersion of classifier responses along the mel-frequency axis. The Top10 metric measures the fraction of total response contained in the strongest 10% of time-frequency locations, irrespective of their frequency positions. The LMB_{12.5} metric instead measures the fraction of total response assigned to the lowest 12.5% of mel-frequency bins across all time frames and therefore characterizes low-mel-frequency response allocation rather than overall response concentration. These quantities are supporting diagnostics rather than primary recognition metrics.

4.6.3 WG-SLE implementation

WG-SLE uses the same spectral configuration as the downstream recognizer. The delayed reference is constructed with $\tau_{\Delta} = 1$ STFT frame, corresponding to one 1600-sample hop or 0.100 s under the 16 kHz protocol. The recording-level estimates and class-wise means in Sec. 4.3 initialize the spectral-prototype bank with the denominator floor $\varepsilon = 10^{-10}$. For the bounded complex parameterization in Eq. (27), $B = 2.0$. The real and imaginary components of each anchor $\hat{W}_c(f)$ are divided by B , clipped to $[-1 + \delta, 1 - \delta]$ with $\delta = 10^{-3}$, and mapped to the unconstrained parameter space using $\operatorname{arctanh}(\cdot)$. The resulting parameters are refined by gradient-based optimization with $\lambda_w = 0.5$ and $\lambda_s = 0.01$ in Eq. (51). In inference-time routing, $\tau_{\text{ret}} = 1.0$. Scaled residual fusion follows Eq. (42). The scalar $s \in \mathbb{R}$ is initialized to 1.0 and optimized directly without a sigmoid mapping. Across the seed-42 AST-family WG-SLE checkpoints selected using the evaluation partition reported in the main front-end tables, the learned values range from 1.0000 to 1.0111.

Ablation configurations

The ablation settings distinguish fixed front-end values from classification-loss-driven front-end learning. In the fixed Wiener-initialized configuration, all WG-SLE front-end parameters, including the spectral prototypes and residual scale, remain at their predefined values, and only the downstream recognizer is optimized using the CE loss. In the CE-loss-only configuration, the Wiener prediction objective and sparsity regularization are omitted and stop-gradient is removed, allowing the classification loss to update the trainable WG-SLE front end directly. Unless otherwise stated, this configuration retains the Wiener initialization, bounded spectral-prototype parameterization, learnable residual scale, and scaled residual fusion of the full model. The remaining configurations modify the indicated initialization, objective, or fusion component relative to the full WG-SLE setting.

4.6.4 Baselines and compared front ends

The experiments use three comparison settings for controlled evaluation. First, recognition without line enhancement uses the matched baseline with the same log-mel input, downstream recognizer, optimization settings, and evaluation protocol, thereby isolating the contribution of WG-SLE. Second, stand-alone time-domain least-mean-square adaptive line enhancement (LMS-ALE) serves as the conventional delayed-reference line-enhancement reference. It is implemented as time-domain LMS adaptive filtering: a delayed waveform drives a linear adaptive predictor, and the filter coefficients are updated sample by sample to minimize the instantaneous squared prediction error. The predicted component is used as the line-enhanced signal and is supplied as the sole recognizer input; the adaptive filter is optimized independently of the downstream recognition loss [21, 22, 24]. Third, the integrated WG-SLE front end is evaluated under the same matched data, backbone, optimization, resizing, and checkpoint-selection protocol. Unlike stand-alone LMS-ALE, WG-SLE uses the class-indexed spectral-prototype bank, Wiener prediction objective, sparsity regularization, and scaled residual fusion within the recognition training framework.

Architectural compatibility of the WG-SLE front-end interface is further examined using six CNNs—MobileNetV3-S [27], CRNN-9 [28], ResNet-18 [26], Xception [29], DenseNet-121 [30], and EfficientNet-B0 [31]—and five transformer-style recognizers—AST-B/384 [25], MHT-UATR-Transformer [32], MF-UATR [33], EAT-base [34], and ConViT [35]. These comparisons assess whether the enhancement-and-residual-fusion interface can precede different downstream recognizers. The same WG-SLE front end is evaluated with six convolutional and five transformer-style recognizers.

Data availability

The DeepShip dataset is available from the dataset authors at <https://github.com/irfankamboh/DeepShip>, and the ShipsEar dataset is available through the Universidade de Vigo underwater-noise database at <https://underwaternoise.atlanttic.uvigo.es/>, subject to the access conditions and terms of use specified by the original providers. The task3 recording-level training and evaluation partitions used in this study, together with the class mappings and partition manifests, are available at <https://github.com/we4237/wg-sle>. The original DeepShip and ShipsEar audio recordings are not redistributed in this repository.

Code availability

The source code supporting this study is not publicly available at the time of submission but will be made available to the editors and reviewers upon reasonable request to the corresponding author during peer review. The code will be released publicly at <https://github.com/we4237/wg-sle> upon publication. The public repository currently provides only the task3 dataset partition files described in the Data availability statement.

References

- [1] Sutin, A. *et al.* Stevens passive acoustic system for underwater surveillance. *2010 International WaterSide Security Conference* 1–6 (2010).
- [2] Hummel, H. I., van der Mei, R. & Bhulai, S. A survey on machine learning in ship radiated noise. *Ocean Engineering* **298**, 117252 (2024).
- [3] Ross, D. *Mechanics of Underwater Noise* (Pergamon Press, New York, 1976).
- [4] Urick, R. J. *Principles of Underwater Sound* 3rd edn (McGraw-Hill, New York, 1983).
- [5] Arveson, P. T. & Vendittis, D. J. Radiated noise characteristics of a modern cargo ship. *The Journal of the Acoustical Society of America* **107**, 118–129 (2000).

- [6] Zhou, H., Luo, X. & Chen, L. The weak modulation line spectrum enhancement and detection method for ship radiated noise in the absence of prior information. *Applied Acoustics* **226**, 110211 (2024).
- [7] Wenz, G. M. Acoustic ambient noise in the ocean: Spectra and sources. *The Journal of the Acoustical Society of America* **34**, 1936–1956 (1962).
- [8] Jensen, F. B., Kuperman, W. A., Porter, M. B. & Schmidt, H. *Computational Ocean Acoustics* (AIP Press, New York, 1994).
- [9] Ke, X., Yuan, F. & Cheng, E. Integrated optimization of underwater acoustic ship-radiated noise recognition based on two-dimensional feature fusion. *Applied Acoustics* **159**, 107057 (2020).
- [10] Hong, F., Liu, C., Guo, L., Chen, F. & Feng, H. Underwater acoustic target recognition with a residual network and the optimized feature extraction method. *Applied Sciences* **11**, 1442 (2021).
- [11] Luo, X., Zhang, M., Liu, T., Huang, M. & Xu, X. An underwater acoustic target recognition method based on spectrograms with different resolutions. *Journal of Marine Science and Engineering* **9**, 1246 (2021).
- [12] Xiao, X., Wang, W., Ren, Q., Gerstoft, P. & Ma, L. Underwater acoustic target recognition using attention-based deep neural network. *JASA Express Letters* **1**, 106001 (2021).
- [13] Feng, S. & Zhu, X. A transformer-based deep learning network for underwater acoustic target recognition. *IEEE Geoscience and Remote Sensing Letters* **19**, 1–5 (2022).
- [14] Wang, Y., Ren, J., Xie, Y., Xu, J. & Zhang, P. An underwater acoustic target recognition system with band splitting and sub-band weighting. *The Journal of the Acoustical Society of America* **158**, 1526–1537 (2025).
- [15] Li, J. & Yang, H. The underwater acoustic target timbre perception and recognition based on the auditory inspired deep convolutional neural network. *Applied Acoustics* **182**, 108210 (2021).
- [16] Zhu, P. *et al.* Underwater acoustic target recognition based on spectrum component analysis of ship radiated noise. *Applied Acoustics* **211**, 109552 (2023).
- [17] Ravanelli, M. & Bengio, Y. Speaker recognition from raw waveform with SincNet. *2018 IEEE Spoken Language Technology Workshop (SLT)* 1021–1028 (2018).
- [18] Zeghidour, N., Teboul, O., de Chaumont Quitry, F. & Tagliasacchi, M. LEAF: A learnable frontend for audio classification. *International Conference on Learning Representations* (2021).

- [19] Ren, J., Xie, Y., Zhang, X. & Xu, J. UALF: A learnable front-end for intelligent underwater acoustic classification system. *Ocean Engineering* **264**, 112394 (2022).
- [20] Xie, Y., Ren, J. & Xu, J. Adaptive ship-radiated noise recognition with learnable fine-grained wavelet transform. *Ocean Engineering* **265**, 112626 (2022).
- [21] Widrow, B. & Stearns, S. D. *Adaptive Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1985).
- [22] Haykin, S. *Adaptive Filter Theory* 4 edn (Prentice Hall, Upper Saddle River, NJ, 2002).
- [23] Liang, G., Hao, Y., Zou, N. & Qiu, L. Sparsity-based frequency-domain adaptive line enhancer. *The Journal of the Acoustical Society of America* **146**, 2799 (2019).
- [24] Yu, X., Chi, C., Li, S. & Li, D. Underwater acoustic target recognition using line enhancement and deep neural network. *Acta Acustica* **49**, 656–663 (2024).
- [25] Gong, Y., Chung, Y.-A. & Glass, J. AST: Audio spectrogram transformer. *Proceedings of Interspeech* 571–575 (2021).
- [26] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016).
- [27] Howard, A. *et al.* Searching for MobileNetV3. *Proceedings of the IEEE/CVF International Conference on Computer Vision* 1314–1324 (2019).
- [28] Liu, F., Shen, T., Luo, Z., Zhao, D. & Guo, S. Underwater target recognition using convolutional recurrent neural networks with 3-D Mel-spectrogram and data augmentation. *Applied Acoustics* **178**, 107989 (2021).
- [29] Chollet, F. Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1251–1258 (2017).
- [30] Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4700–4708 (2017).
- [31] Tan, M. & Le, Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the International Conference on Machine Learning* 6105–6114 (2019).
- [32] Feng, S., Zhu, X. & Ma, S. Masking hierarchical tokens for underwater acoustic target recognition with self-supervised learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **32**, 1365–1379 (2024).

- [33] Fu, B., Nie, J., Wei, W. & Zhang, L. Constructing a multi-modal based underwater acoustic target recognition method with a pre-trained language-audio model. *IEEE Transactions on Geoscience and Remote Sensing* **63**, 1–14 (2025).
- [34] Chen, S. *et al.* EAT: Self-supervised pre-training with efficient audio transformer. *arXiv preprint arXiv:2401.03497* (2024).
- [35] d’Ascoli, S. *et al.* ConViT: Improving vision transformers with soft convolutional inductive biases. *Proceedings of the International Conference on Machine Learning* 2286–2296 (2021).
- [36] Irfan, M. *et al.* DeepShip: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification. *Expert Systems with Applications* **183**, 115270 (2021).
- [37] Santos-Dominguez, D., Torres-Guijarro, S., Cardenal-Lopez, A. & Pena-Gimenez, A. ShipsEar: An underwater vessel noise database. *Applied Acoustics* **113**, 64–69 (2016).
- [38] Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**, 37–46 (1960).

Acknowledgements

The authors thank the creators and maintainers of the DeepShip and ShipsEar datasets for making these underwater acoustic recordings available to the research community. This research received no specific funding.

Author contributions

M.C. conceived the study, developed the methodology and software, conducted the experiments, performed the formal analysis and validation, curated the data, prepared the visualizations, and wrote the original manuscript. L.J. contributed to the conceptualization and methodology, supervised the study, contributed to validation and project administration, and reviewed and edited the manuscript. Both authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Longyu Jiang (jly@seu.edu.cn).