Line spectrum extraction in spectrograms using a DNN with SAM2 encoding and reverse attention

REN Xiangyang[1,2] YU Tianxing[1,2] JIN Yiyang[2] WEI Yangjie[3] LUO Xinwei[1,2†]

1 Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing 211189, China
2 School of Information Science and Engineering, Southeast University, Nanjing 211189, China
3 Nanjing Research Institute of Electronic Engineering, Nanjing 210007, China

Received November 26, 2025; accepted January 16, 2026

Abstract To address the challenge of extracting weak line spectrum from spectrograms, this paper proposes a line spectrum extraction network (LSE-Net) with a simple structure and excellent performance. This network is built on a classic encoder-decoder architecture. To improve training efficiency with a small number of samples, this paper embeds adapters into each layer of the encoder to achieve parameter-efficient fine-tunin, enabling the model to complete training with only a small amount of data and attain excellent performance. To strengthen the line spectrum feature extraction capability, a multi-receptive field feature fusion module is integrated at the skip connections to capture line spectrum information at different scales; meanwhile, a reverse cross attention module is introduced in deep feature processing to accurately enhance the edge features of line spectra and effectively suppress background noise. Experimental results demonstrate that the proposed LSE-Net outperforms state-of-the-art methods (e.g., Deep Lofargram, Trans-UNet) in low signal-to-noise ratio (SNR) scenarios (ranging from −16 dB to −22 dB). Specifically, at −22 dB SNR, LSE-Net achieves a probability of detection of 0.965, a false alarm rate of 0.373, and a line localization accuracy of 0.786, which are 57.1%, 42.3%, and 85.8% higher than those of Deep Lofargram, respectively. The proposed network provides an efficient and reliable solution for weak line spectrum extraction in spectrograms.

Keywords Signal detection, Passive sonar, Line spectrum, Ship radiated noise, Deep neural network (DNN)

PACS: 43.30, 43.60 **DOI:** 10.3724/0217-9776.3025002 **CSTR:** 32050.14.11-2066.3025002

1 Introduction

Underwater target detection constitutes the core task of passive sonar systems. Ships are typical underwater acoustic targets, and the acoustic characteristics of their radiated noise form the basis for target detection and identification by passive sonar systems[1]. The power spectrum of ship radiated noise consists of a broadband spectrum and a discrete line spectrum. These noises mainly come from three sources: mechanical vibration generated by the operation of ship equipment, fluid dynamic noise formed by propeller rotation cavitation, and vibration radiation noise caused by ship structure vibration[2,3]. Line spectrum components in noise are excited by periodic vibration sources and exhibit stable frequencies. Owing to their concentrated energy in the frequency spectrum, these components can be detected over long distances, serving as reliable acoustic signatures for underwater targets[4]. The line spectrum characteristics of targets are of great significance for target detection, tracking, and recognition, and line spectrum detection is one of the research focuses in the field of passive sonar technology[5]. With the development of ship vibration and noise reduction technologies, the intensity of target radiated line spectra continues to decrease. Meanwhile, the increasing number of ships navigating the ocean has elevated the background noise level. These factors result in the line spectrum signals becoming increasingly weak relative to the ambient noise, posing significant challenges to line spectrum extraction methods.

Low-frequency analysis and recording (LOFAR) is a crucial method for analyzing the radiated noise of ship targets in sonar systems. Spectrograms of ship radiated noise can be obtained through LOFAR analysis, and various line spectrum detection methods attempt to extract spectral line

This work was supported by the National Natural Science Foundation of China (12174053) and the Fundamental Research Funds for Central Universities (2242025F20003).
† Corresponding author: LUO Xinwei, luoxinwei@seu.edu.cn

© The Author(s) 2026. This is an open access article under the CC-BY 4.0 License (https://creativecommons.org/licenses/by/4.0/).

trajectories corresponding to target characteristics from the spectrogram data.

Line spectrum detection, as one of the core issues in underwater acoustic signal processing, has been addressed through various technical approaches. Probability modelled approaches utilize Hidden Markov Models (HMM) to determine optimal state paths through maximum likelihood estimation, thereby providing a probabilistic framework for spectral line detection. Jauffret *et al*.[6] pioneered the application of HMM to line spectrum trajectory detection. They utilized the state transition probability matrix of HMM to precisely model the frequency state evolution of line spectra. Subsequently, through dynamic programming algorithms, they searched for and extracted line spectrum trajectories with maximum likelihood probability, offering a novel approach to detection. However, this approach exhibits certain limitations, being effective only for detecting a single line spectrum trajectory and thus struggling to meet the demands of multi-object scenarios. To address the challenge of detecting multiple line spectrum trajectories, Paris *et al*.[7] refined the methodology based on prior research. They incorporated the frequency change rate of the line spectrum signal into the state representation dimension, proposing a line spectrum tracker based on a two-dimensional HMM. This improved approach overcomes the constraints of traditional one-dimensional state modelling by simultaneously capturing both the frequency values and trend changes of acoustic spectra. Consequently, it achieves efficient and accurate detection of multiple acoustic spectrum trajectories, significantly enhancing the practical value of HMM in acoustic spectrum tracking. Luo *et al*.[8] proposed a multi-frequency line perception and tracking algorithm for hydroacoustic signals based on Markov models: first, the logarithmic coordinate spectrogram is partitioned into multiple subgraphs, with preliminary trajectory detection achieved through subgraph screening; subsequently, image stitching techniques and statistical models are integrated to merge frequency lines and suppress false alarms, effectively overcoming the application limitations of traditional Markov algorithms.

As the acoustic signal spectrum exhibits characteristic image features, researchers have incorporated image processing techniques into the task of extracting spectral trajectories. In early studies, Abel *et al*.[9] employed a dual-pass separation window to estimate broadband background noise in spectrograms, achieving spectral line detection through noise suppression. Gillespie[10] employed Gaussian convolution for spectral smoothing preprocessing, filtering out isolated high-amplitude noise points before combining edge detection with whale call spectral line structural features to complete detection. Addressing performance bottlenecks in low signal-to-noise ratio (SNR) scenarios, Lampert *et al*.[11] proposed a contour energy minimization approach. This method models spectral line trajectories as non-closed contours, detecting them by fitting and extracting high-intensity time-frequency points from LOFAR maps. However, this approach suffers from limitations in simultaneously processing multiple spectral lines and determining trajectory birth-death states. Consequently, Lampert *et al*.[12] introduced a feature detector to the CEM framework. By integrating narrowband windows with principal component analysis, they achieved birth-death discrimination of spectral line time-frequency points, significantly reducing false alarm rates (FARs). Additionally, Zhao *et al*.[13] constructed feature functions based on spectral line intensity, frequency continuity, and trajectory continuity. Through narrowband sliding windows, they screened for optimally featured spectral lines within each frequency band to determine spectral line trajectories.

Deep learning techniques have achieved widespread application in medical image segmentation. The U-Net architecture proposed by Olaf *et al*.[14], specifically designed for biomedical image segmentation, demonstrates outstanding performance in classification tasks. Since the advent of U-Net, researchers have implemented multi-dimensional enhancements to address diverse application requirements, yielding a series of architectures suitable for complex signal segmentation. Attention mechanisms, integrating modules such as embedded spatial attention and channel attention[15], enhance the model's ability to focus on key features while suppressing background noise interference. Models employing Transformers leverage their global attention mechanism to improve modelling capabilities for long-range correlated features.

In recent years, foundational models from the field of computer vision have demonstrated universality in image segmentation tasks, presenting researchers with an opportunity to explore their application in spectral line extraction within LOFAR spectra. The core requirement for acoustic spectral line segmentation is the precise separation of discrete spectral line components from broadband background noise and interference signals within the spectrum. This task fundamentally constitutes an image classification problem: neural network models must label pixels corresponding to spectral lines in LOFAR spectra as spectral line categories, while marking the remaining regions as noise. This requirement aligns closely with the objectives of computer vision image segmentation, driving the migration and application of various segmentation architectures into the field of acoustic signal processing[16]. Within LOFAR spectra, spectral lines manifest as small, weakly featured targets whose two-dimensional data structure is fully compatible with image data. This enables classical segmentation networks (such as U-Net) to operate without substantial modifications to their input interfaces. By feeding spectra processed via short-time Fourier transform (STFT) to the

U-Net, its feature extraction logic can be directly repurposed to achieve effective separation of spectral lines from broadband noise[17]. Han *et al*.[18] transformed single spectral line detection in LOFAR maps into a binary classification problem, proposing the Deep Lofargram model based on deep convolutional neural networks for end-to-end spectral line detection. While this method demonstrated excellent performance in single spectral line detection, it proved unsuitable for scenarios involving the dynamic appearance and disappearance of multiple spectral lines. Jiang *et al*.[15] proposed the FLD-Net model, which employs an attention mechanism to fuse feature modules and suppress noise interference within deep decoder features, thereby enhancing spectral reconstruction accuracy. This approach successfully achieved effective detection of multiple weak variable spectral lines in low-SNR underwater acoustic signals. Liang *et al*.[19] proposed an enhanced U-Net spectral line reconstruction algorithm based on temporal and harmonic attention mechanisms. This approach expands the model's receptive field while improving computational efficiency. It further optimizes detection performance by synchronously extracting spectral line amplitude and phase information through specific operators.

Existing deep learning-based spectral line extraction methods still encounter bottlenecks in practical applications: under low-SNR conditions (e.g., $R_{\mathrm{SN}} \leqslant -20$ dB), traditional deep convolutional neural networks struggle to effectively distinguish weak spectral line features from strong background noise. This leads to severe feature degradation, resulting in low detection rates and insufficient localization accuracy. Spectral line extraction methods reliant on manually designed features (such as peak detection and energy thresholding) exhibit sharp degradation in low SNR and multi-target overlapping scenarios due to their inadequate adaptability to non-stationary environments[20]. Existing deep learning models heavily depend on training with large-scale annotated real ocean data, yet the acquisition and annotation of hydroacoustic data are costly and time-consuming, limiting their generalization capability in complex real-world environments. The General Image Segmentation Model 2 (SAM2), proposed by Meta, is a general-purpose, efficient computer vision model specifically designed for image segmentation tasks. Its encoder employs the Hiera encoder architecture based on the layered Visual Transformer (ViT)[21]. As the core module processing image information, the encoder's performance directly determines segmentation outcomes. Prior to training the Hiera encoder, SAM2 undergoes pre-training via a Masked Autoencoder (MAE). The Hiera module maintains performance gains even after stripping away the additional modules and components of the original ViT[22]. SAM2's zero-shot segmentation capability enables the model to accurately segment complex structures with minimal input, accomplishing the segmentation of fine objects within images without requiring large amounts of manually annotated training data. This feature holds significant value for segmentation and detection within LOFAR acoustic signal spectra.

This research work proposes a line-spectrum extraction network (LSE-Net) that achieves precise extraction of line spectra from LOFAR spectrograms under low-SNR conditions. The network demonstrates significant advantages in both structural design and performance, with the following key features.

(1) A simple model architecture employing a classical encoder-decoder structure: This design ensures ease of use alongside excellent scalability, facilitating subsequent research and engineering applications.

(2) Training efficiency is enhanced through the incorporation of adapter modules within each encoder layer, enabling parameter-efficient fine-tuning (PEFT). This reduces reliance on large datasets, achieving robust model performance with minimal training data.

(3) Superior feature extraction is achieved by incorporating multi-receptive field (MRF) fusion modules into the network's skip connections, enabling effective integration of multi-scale feature information. Concurrently, the introduction of reverse attention modules in deep feature processing enhances the representation of spectral line edges, significantly improving the accuracy and completeness of spectral line extraction.

(4) Robust performance: Following training on simulated datasets, the network model can be directly transferred to spectral line detection tasks using real-world measurement data while maintaining excellent detection capabilities.

2 Network architecture

This paper proposes an LSE-Net based on SAM2 encoding and attention feature fusion. The network architecture of the proposed method is illustrated in Fig. 1, which adopts an encoder-decoder structure as a whole. For the input signal spectrogram, it is first fed into the Hiera encoding module for feature encoding. This module generates four feature layers with distinct scales from the signal spectrogram through a multi-scale hierarchical feature extraction mechanism. For the feature layer with the smallest scale, the reverse cross attention (RCA) mechanism is introduced to accurately capture the edge details of the line spectrum trajectory; the other three groups of feature layers are subjected to feature fusion via multi-scale receptive fields, which effectively alleviates the loss of spatial information during feature transmission. In terms of decoder design, the pixel-level localization advantage of the U-Net structure is fully utilized. Through the upsampling operation, the decoder features are correlated with the output features of each encoder layer, and

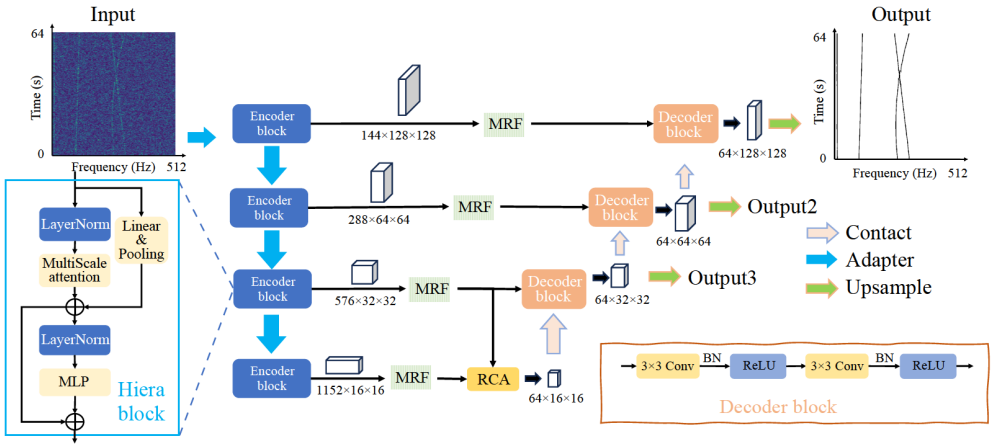


Fig. 1 Network architecture.

a multi-level joint loss function is integrated to optimize the model training process, which significantly improves the segmentation performance of line spectrum targets in the spectrogram.

The encoder component employs the Hiera module as its backbone network. As a multi-scale hierarchical Transformer architecture, the Hiera module exhibits inherent compatibility with the multi-layer network structure of the U-Net. Compared to other visual Transformer models (such as ViT and Swin Transformer), Hiera demonstrates significant advantages in spectral line extraction tasks. Hiera employs a hierarchical feature extraction mechanism combined with progressive downsampling, enabling the natural generation of multi-scale feature maps that align with spectral line features. In contrast, ViT utilizes a fixed patch size, lacking inherent multi-scale representation capabilities. The Hiera encoder undergoes pre-training via masked autoencoding, enhancing its capacity to learn discriminative features from noisy data[21]. Compared to Swin Transformer, Hiera possesses a more flexible field of view adjustment mechanism, enabling superior adaptation to the variable spatial distribution characteristics of spectral lines within LOFAR spectra.

For the input LOFAR spectrum $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$, after processing through four encoding units, it outputs four feature layers $\mathbf{X}_i \in \mathbb{R}^{3\times H/2^{i+1}\times W/2^{i+1}}$ at different scales, where $i = 1, 2, 3, 4$, corresponding to the four feature levels. The Hiera encoding module captures hierarchical feature information through a multi-scale attention mechanism, efficiently extracting multi-scale features to provide robust support for subsequent tasks. The specific structure of the Hiera module is illustrated in Fig. 2. During model training, its parameters are initialised using pre-trained model parameters, and these parameters are subsequently frozen to prevent degradation of the pre-trained feature representation capabilities.

The original Hiera encoder module has a large number

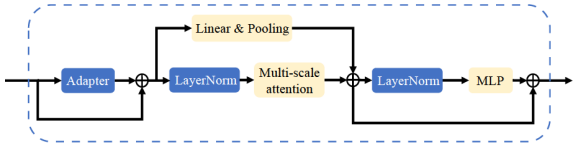


Fig. 2 Schematic diagram of the Hiera block and Adapter block.

of parameters, and direct full-scale training of it will significantly increase task complexity. Meanwhile, to preserve the generalization capability acquired by the pre-trained model, all parameters of the encoder are frozen during the training phase. For efficient parameter fine-tuning, a trainable adapter module is added before each encoder module, enabling the model to adaptively learn task-specific knowledgefor line spectrum extraction while retaining the pre-trained feature representation capability. By fine-tuning the output data of the encoding module through the adapter, the model fully reuses the feature advantages of the pre-trained encoder, achieving excellent line spectrum detection performance even in scenarios with limited training data. PEFT, as a widely adopted technique, enhances training efficiency by fine-tuning only a subset of parameters in pre-trained models. In this paper, PEFT is employed to facilitate effective training on small-sample datasets, providing a rapid and efficient means of model fine-tuning for this research. Methods such as low-rank adaptation (LoRA)[23] also demonstrated favorable outcomes in model fine-tuning and may serve as avenues for future research expansion.

Each adapter module is composed of a downsampling linear layer, an activation function, an upsampling linear layer, and an output activation function in cascade, with the core function of realizing adaptive mapping and dynamic adjustment of feature dimensions, as illustrated in Fig. 3. Specifically, the adapter corresponding to each encoder maps the output feature of the encoder layer to 64 channels through downsampling, and then maps it back to the channel

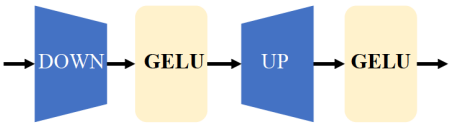


Fig. 3 Schematic diagram of the adapter module.

dimension of the original encoder feature, where the number of channels of the output feature layer of each encoder remains $C_i \in \{144, 288, 576, 1152\}$.

The skip-connection design of U-Net enables the network to fuse feature maps at corresponding positions in the encoder across channels during the up-sampling process at each level[14], allowing the network to retain more high-resolution detail information and thereby improving image segmentation accuracy. Whilst SAM2 incorporates a multi-scale feature representation and fusion mechanism through its Hiera module for hierarchical feature extraction, our proposed MRF module further enhances the capabilities of SAM2's native multi-scale feature representation and fusion mechanism. Building upon the role of skip connections, the MRF module further utilizes the rich detail information within the feature layers, thereby maximizing the utilization of feature information across all layers of the network. The multi-receptive field fusion mechanism not only extracts greater detail from features across different scales but also enhances computational efficiency by reducing the number of channels, which is crucial for spectral line extraction. It is worth emphasizing that the hierarchical multi-scale representation in SAM2 is realized at the inter-stage level through the Hiera encoder, whereas the proposed MRF module operates at the intra-stage level, enriching receptive-field diversity within each individual scale before skip-connection fusion.

The skip-connection section employs a multi-receptive field feature fusion module to uniformly map the multi-layer features output from each level of encoders, with its structure illustrated in Fig. 4. In the channel dimension, the feature map $\mathbf{F}$ output by each encoder is first divided into four parallel branches (denoted as Branch_1–Branch_4), which perform differentiated cross-scale feature synchronous extraction respectively.

Branch_1 uses 1×1 convolution to complete cross-channel feature fusion and channel alignment of residual connections, with its calculation shown in Eq. (1), and finally outputs the feature matrix $\mathbf{F}_{1*}$,

$$\mathbf{F}_{1*} = \text{Conv}_{1\times1}(\mathbf{F}_1). \tag{1}$$

Branch_1–Branch_4 adopt a unified structure: first, $1 \times k$ and $k \times 1$ strip convolutions (where k is the length of the convolution kernel) are used to reduce the number of parameters and enhance the nonlinear expression ability of features; then, 3×3 dilated convolution with a dilation rate of d is used to expand the receptive field and establish long-distance semantic associations between targets; finally, 1×1 convolution is used to reduce the number of channels and achieve feature alignment of each branch, outputting feature matrices $\mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4$. The specific calculation is shown in Eq. (2).This design can accurately capture detailed features of different scales in the LOFAR spectrogram, where $\text{Dilated_Conv}_{(k,d)}$ denotes the dilated convolution operation with a kernel size of $k \times k$ and a dilation rate of d, and $\text{Conv}_{(k\times k)}$ denotes the standard convolution operation with a kernel size of $k \times k$.

$$\mathbf{F}_{i*} = \text{Dilated_Conv}_{(3,k)}\left(\text{Conv}_{k\times1}\left(\text{Conv}_{1\times k}(\text{Conv}_{1\times1}(\mathbf{F}_i))\right)\right),\quad i = 2, 3, 4; k = 3, 5, 7. \tag{2}$$

Finally, the output features of the four branches are subjected to channel concatenation, and then 1×1 convolution is used to complete feature fusion under multi-receptive fields, outputting the final feature matrix $\mathbf{F}_{\text{fusion}}$:

$$\mathbf{F}_{\text{fusion}} = \text{Conv}_{1\times1}([\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4]), \tag{3}$$

where $[\cdot]$ denotes the channel concatenation operation.

Attention mechanisms have been extensively applied across various visual tasks by compensating for the limitations of standard convolutions through the computation of weighted tensors for input tensors. However, during the down-sampling process in the encoder, image information is prone to loss as network depth increases. Although traditional U-Net architectures preserve low-dimensional features

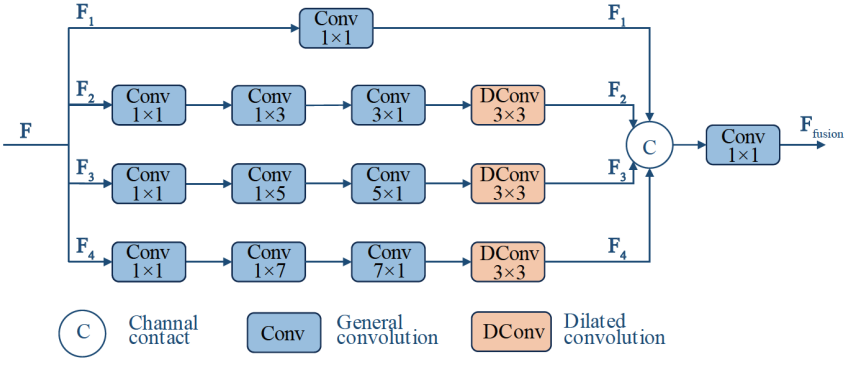


Fig. 4 Schematic diagram of the multi-receptive field feature fusion module.

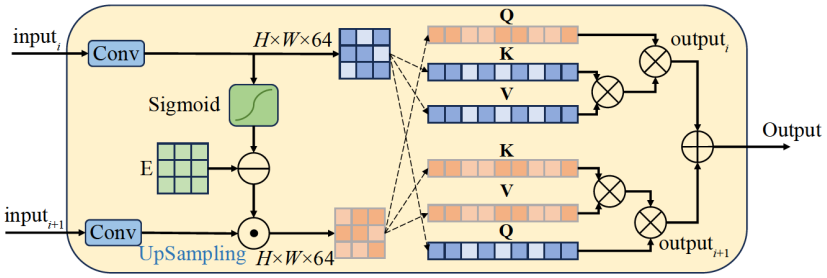


Fig. 5 Schematic diagram of the reverse cross attention module.

via skip connections, such features contain substantial redundant information, and the feature layers output by the backbone network lack clear target boundary information. To achieve effective fusion of high-dimensional features with low-dimensional information while preserving spectral edge details, this paper proposes a reverse attention module for spectral edge information retention, whose specific architecture is illustrated in Fig. 5.

The distinction between reverse attention and conventional attention lies in their processing logic for targets. Conventional attention focuses on discriminative target regions by computing feature weight matrices, simultaneously suppressing feature responses in background areas. The reverse attention module employs an inverted logic, refraining from directly enhancing features in target regions. By learning to suppress feature responses in target areas, it delineates target edges and the boundary regions between background and target, thereby achieving precise capture of edge features. In the line pattern detection task addressed herein, the edge features of line pattern targets directly determine both target morphology and localization accuracy. Consequently, the precise capture of line pattern edge information constitutes the core requirement for enhancing detection performance. To address the deficiency of edge information in the encoder's output feature map, the RCA module employs reverse weights to optimize features. The specific workflow is as follows.

Step 1: reverse attention weight generation and feature calibration. First, the Sigmoid activation function is used to compress the value of the feature input_i (output from the ith layer of the encoder), and then the "reverse operation" is implemented by element-wise subtraction with the all-ones matrix $\mathbf{E}$ to obtain the reverse attention weight ε:

$$\varepsilon = \mathbf{E} - \mathrm{Sigmoid}(\mathrm{input}_i), \tag{4}$$

where $\mathbf{E}$ is an all-ones matrix with the same size as $\mathrm{Sigmoid}(\mathrm{input}_i)$, used to construct the reverse feature space.

Subsequently, the high-dimensional output feature input_{i+1} (from the $(i+1)$th layer of the encoder) is element-wise multiplied by the reverse attention weight ε to complete the feature calibration of the target boundary, and the feature layer input_{i+1}^{m} with enhanced edge information is

$$\mathrm{input}_{i+1}^{m} = \varepsilon \odot \mathrm{input}_{i+1}, \tag{5}$$

where $\odot$ denotes element-wise multiplication.

Step 2: Cross-attention feature fusion. To strengthen the correlation between feature layers of different dimensions and improve the accuracy and consistency of line spectrum segmentation, the RCA module introduces the Cross-attention mechanism: the original feature input_i and the calibrated feature input_{i+1}^{m} are respectively used as Query matrices, and the other feature layer is used as Key and Value matrices. The correlation information between features is mined through cross-calculation:

$$\mathrm{output}_i = \text{Cross-attention}\left(\mathrm{input}_{i+1}^{\mathbf{Q}}, \mathrm{input}_i^{\mathbf{K}}, \mathrm{input}_i^{\mathbf{V}}\right), \tag{6}$$

$$\mathrm{output}_{i+1} = \text{Cross-attention}\left(\mathrm{input}_i^{\mathbf{Q}}, \mathrm{input}_{i+1}^{\mathbf{K}}, \mathrm{input}_{i+1}^{\mathbf{V}}\right), \tag{7}$$

where the superscripts **Q**, **K**, and **V**, respectively, represent the role mapping of the feature matrix in the attention mechanism.

Finally, the two cross-fused feature layers output_i and output_{i+1} are element-wise added to obtain the final output feature of the RCA module. This feature will be concatenated and fused with the feature of the corresponding level in the decoder, providing richer edge and semantic correlation information for subsequent line spectrum segmentation.

The decoder architecture of U-Net has been thoroughly validated for its effectiveness across various image segmentation tasks. The decoder employed in this paper also adopts the classic U-Net decoder architecture. Each decoder module receives inputs from two distinct feature layers: the skip-connection feature layer optimized via MRF, and the feature layer output from the lower-level decoder. Internally, it employs a dual-convolution structure, with the specific configuration illustrated in Fig. 6. The features output by the three decoder modules within the network must be restored to the dimensions of the input image through bilinear interpolation

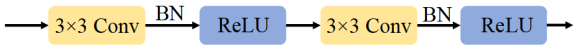


Fig. 6 Schematic diagram of the Decoder block.

operations, yielding three independent output results. Each output result is compared with the ground truth label to perform supervised training and compute the loss. The specific design of the relevant loss function will be detailed in subsequent sections.

3 Simulated data

Conducting hydroacoustic signal acquisition in real marine environments is constrained by cost and the randomness of environmental interference, making it challenging to construct large-scale, precisely annotated datasets. Existing publicly available hydroacoustic datasets generally lack signal data under varying SNR conditions, failing to meet the requirements for validating model generalization capabilities. To address this, this section constructs a standardized dataset by simulating multi-feature line spectrum signals in complex marine environments, providing reliable data support for training and validating line spectrum detection models. The simulated dataset achieves end-to-end construction through three core modules: acoustic signal modelling, LOFAR spectrogram generation, and automated label production. This dataset supports diverse line spectrum features (fractures, drifts, etc.), noise types, and spectrogram resolutions, adapting to the requirements of various detection scenarios. The final output comprises 512 × 512 LOFAR spectrograms with corresponding pixel-level segmentation labels, meeting model training and validation requirements. A "noise-added sine wave" serves as the fundamental signal model to simulate acoustic line spectrum characteristics, mathematically expressed as

$$x(t) = \sum_{n=1}^{N} A_n \cdot \sin\left(2\pi f_n t + \varphi_n\right) + n(t), \tag{8}$$

where N is the number of line spectra per sample (ranging from 2 to 6), t is the sampling time, A_n is the amplitude of the nth line spectrum, f_n is the instantaneous frequency of the nth line spectrum at time t, φ_n is the initial phase, and $n(t)$ is the superimposed noise component.

To simulate the acoustic spectrum characteristics of underwater sound, this section implements dynamic control over f_n. It emulates typical features of the simulated spectrum, including linear drift, non-linear drift, random jitter, and discontinuities. The frequency update model is

$$f_n(t) = f_n(t-1) + (1-\mu) f_{\text{lin}} + \mu f_{\text{nonlin}} + f_{\text{jitter}}, \tag{9}$$

where $f_n(t-1)$ is the main frequency of the nth line spectrum at time $t-1$, ensuring the continuity of frequency changes; μ is the nonlinear drift control factor following a Bernoulli distribution (probability of taking 1 is 50%), used to randomly trigger nonlinear drift; f_{lin} is the linear drift component with an adjustable drift rate set to $[-1.0, 1.0]$ Hz/s in this study; f_{nonlin} is the quadratic nonlinear drift component, simulating the bending characteristics of line spectra; f_{jitter} is the Gaussian random jitter component with a mean of 0 and a jitter intensity of 2 Hz, simulating random fluctuations of signals. In addition, to enhance data authenticity, a 20% probability of line spectrum breakage is set to randomly trigger local-time signal absence of a single line spectrum, simulating the discontinuity of line spectra in real marine environments.

To account for the complexity of real ocean noise, this section introduces three typical noise types: additive Gaussian white noise, ocean ambient noise, and impulse noise. For additive Gaussian white noise, the noise power is calculated based on signal power and target SNR. Gaussian noise with zero mean and corresponding power is generated and superimposed onto the clean signal at a 60% probability. Marine background noise is modelled using power spectral density, with the fast Fourier transform employed to convert noise from the frequency domain to the time domain. Impulse noise simulates sudden underwater disturbances by generating random impulse signals.

Although simulated data incorporate random jitter, multi-target superposition, and impulse noise to approximate real underwater acoustic signals more closely, differences from genuine underwater signals persist. Firstly, in real data, line spectra exhibit more random frequency drift accompanied by amplitude variations due to effects such as the Doppler effect and channel distortion. Secondly, the marine environment encompasses vessel interference, pulse noise, and multi-signal superposition interference, resulting in a more complex noise distribution.

Additional parameters for this dataset are as follows: signal sampling frequency of 2 kHz, duration of 64 s, and spectral frequency range of 0–512 Hz. Short-time Fourier transforms are applied using a Hann window (window length 1024 samples, 75% overlap) to generate LOFAR spectrograms. SNR covers four low-SNR scenarios: −16 dB, −18 dB, −20 dB, and −22 dB, with 500 samples per scenario. Spectrogram dimensions are uniformly set at 512 × 512 pixels, labelled as pixel-level binary segmentation maps (spectral regions marked as 1, background as 0). Figure 7 illustrates LOFAR spectrograms and corresponding labels for the same simulated signal under varying SNR conditions. The figure demonstrates that when SNR drops to −22 dB, the spectral line signal is entirely obscured by intense noise, rendering it difficult to discern by human vision alone. This effectively validates the model's detection capability under extremely low-SNR scenarios.

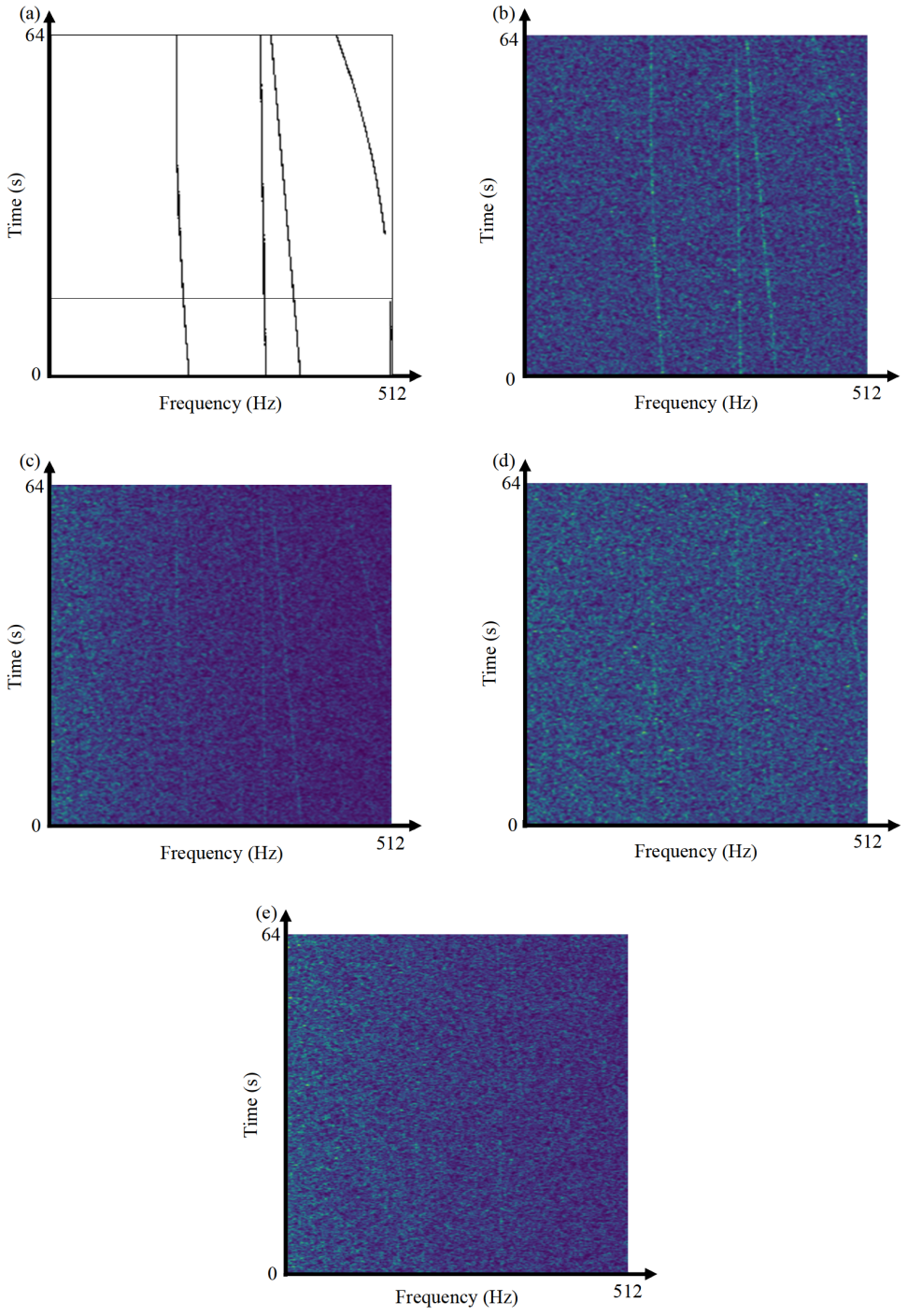


Fig. 7 Simulated spectrograms with different SNRs: (a) Label; (b) −16 dB; (c) −18 dB; (d) −20 dB; (e) −22 dB.

4 Loss function and implementation details

4.1 Loss function

Loss functions for image segmentation predominantly employ pixel-wise mean loss calculations, neglecting the disparity in pixel proportion between targets and backgrounds within the image. Taking the widely used binary cross-entropy (BCE) loss in binary classification tasks as an example, its formula is

$$L_{\mathrm{BCE}} = -\frac{1}{HW} \times \sum_{i=1}^{H}\sum_{j=1}^{W}\left[\mathbf{T}_{i,j}\log(\mathbf{P}_{i,j}) + (1-\mathbf{T}_{i,j})\log(1-\mathbf{P}_{i,j})\right], \tag{10}$$

where $\mathbf{T}_{i,j}$ denotes the ground truth label for pixel (i,j), and $\mathbf{P}_{i,j}$ represents the predicted value for the corresponding pixel.

For the spectral line segmentation task addressed herein, the input LOFAR spectrum label matrix $\mathbf{T} \in \mathbb{R}^{H\times W}$ denotes: the model outputs a prediction matrix $\mathbf{P} \in \{1,0\}^{H\times W}$, where "1" indicates the pixel is classified as a spectral line and "0" indicates it is classified as background noise. Since spectral line regions constitute an extremely small proportion of the entire spectrum, directly using L_{BCE} would cause the loss function to be dominated by background noise regions, making it difficult for the model to focus on learning features specific to spectral line regions.

To address this data imbalance, adaptive weights $\alpha_{ij}^{\mathrm{s}} \in [0,1]$ must be assigned to each pixel. The core principle is that pixels at spectral line edges exhibit significant contrast with their surrounding background and should thus be assigned higher weights to amplify their loss contribution. α_{ij}^{s} is defined as the proportion of target pixels within the local region containing pixel (i,j), calculated as

$$\alpha_{ij}^{\mathrm{s}} = \left| \frac{\sum_{m,n\in\mathbf{A}_{ij}} \mathbf{P}_{mn}^{\mathrm{s}}}{\sum_{m,n\in\mathbf{A}_{ij}} 1} \right|, \tag{11}$$

where $\mathbf{A}_{ij}$ denotes the local window centred on pixel (i,j) (set to 31 × 31 in this paper); $\mathbf{P}_{mn}^{\mathrm{s}}$ represents the ground truth value for pixel (m,n) in the label matrix; the numerator is the total number of spectral pixels within the window, and the denominator is the total number of pixels within the window. Embedding α_{ij}^{s} into the binary cross-entropy loss and introducing a weight adjustment factor γ yields the weighted binary cross-entropy loss L_{WBCE}, thereby enhancing the loss for spectral regions:

$$L_{\mathrm{WBCE}} = -\frac{\sum_{i=1}^{H}\sum_{j=1}^{W}\left(1+\gamma\alpha_{ij}^{\mathrm{s}}\right)\times L_{\mathrm{BCE}}}{\sum_{i=1}^{H}\sum_{j=1}^{W}\left(1+\gamma\alpha_{ij}^{\mathrm{s}}\right)}. \tag{12}$$

For the widely used intersection over union (IoU) loss in segmentation tasks, this section designs a weighted intersection over union loss L_{WIOU} to further strengthen the global constraint on target regions. Employing pixel weights α_{ij}^{s} increases the proportion of spectral line regions in the global loss, calculated as

$$L_{\mathrm{WIOU}} = 1 - \frac{\sum_{i=1}^{H}\sum_{j=1}^{W}\left(1+\gamma\alpha_{ij}^{\mathrm{s}}\right)\left(\mathbf{P}_{i,j}*\mathbf{T}_{i,j}\right)}{\sum_{i=1}^{H}\sum_{j=1}^{W}\left(1+\gamma\alpha_{ij}^{\mathrm{s}}\right)\left(\mathbf{P}_{i,j}+\mathbf{T}_{i,j}-\mathbf{P}_{i,j}*\mathbf{T}_{i,j}\right)}. \tag{13}$$

Combining L_{WBCE} with L_{WIOU} yields the fusion loss suitable for spectral extraction in this paper:

$$L = L_{\mathrm{WBCE}} + L_{\mathrm{WIOU}}. \tag{14}$$

To enhance gradient propagation during network training, a multi-level loss supervision strategy is adopted. For the N decoder outputs in the network, different weights are assigned to distinct decoders. The overall loss is

$$L = \sum_{i=1}^{N} \frac{1}{2^{i-1}} L_i, \tag{15}$$

where L_i denotes the fusion loss corresponding to the ith decoder output, with weights increasing as decoder layers ascend.

4.2 Experimental implementation details

The core parameter configuration of the experiment in this paper is as follows.

(1) Optimizer: AdamW optimizer is adopted, with an initial learning rate of 0.001, combined with a cosine decay strategy to stabilize the training process.

(2) Data augmentation: two strategies of random vertical flipping and random horizontal flipping are used to improve the generalization ability of the model.

(3) Pretrained model: the Hiera encoder uses the SAM2_Hiera_Large pretrained weights officially released by Meta.

(4) Input specification: the resolution of all input images is uniformly adjusted to 512 × 512.

(5) Training parameters: the batch size is set to 4, and the total number of training epochs is set to 100.

5 Experimental results and analysis

5.1 Evaluation metrics

To address issues such as target blurring and stringent localization accuracy requirements in line spectrum detection under low-SNR, this section employs multiple evaluation metrics to assess network performance. Core metrics include probability of detection (PD), FAR, and line localization accuracy (LLA). The PD and FAR evaluate the validity of detection outcomes, while LLA specifically addresses the evaluation requirements for line spectrum localization.

The PD and FAR are defined based on binary classification outcomes at the image pixel level. True positive (TP) represents the number of pixels predicted as line patterns that are indeed line patterns. False negative (FN) denotes the number of pixels that are actual line patterns but predicted as background. The calculation formulas for these metrics are

$$\mathrm{PD} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{16}$$

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}}. \tag{17}$$

The PD quantifies the model's ability to extract genuine line patterns, with a range of [0,1]. The PD closer to 1 indicates a lower probability of line pattern omission. The FAR quantifies the probability of noise being misclassified as line

patterns, also with a range of [0,1]. The FAR closer to 0 indicates stronger network resistance to interference.

Traditional pixel-level evaluation metrics (such as the Dice coefficient) focus on the pixel overlap between predictions and ground truth labels, failing to reflect positioning deviations along continuous line edges. Therefore, this section employs LLA to assess the degree of alignment in line spectrum positioning. The LLA quantifies the Euclidean distance between predicted line spectrum pixels and ground truth pixels, calculated as

$$\mathrm{LLA}=\frac{1}{\max(|\mathbf{P}|,|\mathbf{T}|)}\sum_{(l,m)\in\mathbf{P}}\frac{1}{1+\lambda\min_{(i,j)\in\mathbf{T}}\left([l,m]-[i,j]^2\right)}, \tag{18}$$

where $\mathbf{P}$ denotes the set of predicted line pattern pixels, with each element (l,m) representing the pixel coordinates of the predicted line pattern. $\mathbf{T}$ denotes the set of actual line pattern pixels, with each element (i,j) representing the pixel coordinates of the actual line pattern. $[l,m]-[i,j]^2$ represents the Euclidean distance between the predicted pixel and the nearest actual pixel, a smaller distance indicates more precise positioning. λ denotes the distance weighting coefficient, regulating the influence of distance on positioning accuracy. In this paper, its value is set to 1, following the performance comparison paper[18]. $\max(|\mathbf{P}|,|\mathbf{T}|)$ represents the normalisation term, mitigating bias arising from differences in predicted and actual pixel counts. The LLA spans the range of [0,1]. When $\mathbf{P}$ and $\mathbf{T}$ are identical, LLA = 1. As the discrepancy between prediction and reality increases, LLA approaches 0. This metric proves particularly effective in scenarios involving spectral line drift or discontinuities, precisely quantifying the model's ability to restore spectral edge details.

5.2 Experimental results

To validate the performance of the proposed network (LSE-Net), the line spectrum detection task was formulated as an image segmentation problem. Three line spectrum segmentation models—Deep Lofargram[18], FLD-Net[15], and Trans-UNet[24]—were selected for comparative experiments. To ensure fairness and objectivity, both model categories employed identical training/testing datasets, hyperparameter configurations, and evaluation metrics, with the sole difference being their architectural design. LSE-Net integrates core modules including adapter, MRF, and RCA. Deep Lofargram employs a traditional deep convolutional network architecture. FLD-Net augments the U-Net framework with attention mechanisms and novel loss functions. Trans-UNet utilizes Vit as the encoding module for feature extraction within U-Net. Performance comparisons across four SNR scenarios (−16–−22 dB) are presented in Table 1. Results demonstrate that LSE-Net exhibits significant advantages in both detection accuracy and robustness.

Indicator comparisons reveal that LSE-Net's performance advantage progressively increases as the SNR decreases. In high-SNR scenarios, both LSE-Net and Trans-UNet achieve markedly higher detection rates than deep network approaches, demonstrating LSE-Net's optimized feature extraction. However, within the SNR range of −18–−20 dB, both LSE-Net and Trans-UNet exhibit substantial improvements in detection rate and line localization accuracy. Under extremely low-SNR conditions, LSE-Net's performance improvement is particularly pronounced. While traditional deep neural network (DNN) methods exhibit comparable performance to LSE-Net at high SNR, our approach demonstrates a distinct advantage in low-SNR scenarios. The above comparisons conclusively demonstrate that LSE-Net resolves the issues of low detection rates and poor positioning accuracy encountered by conventional line spectrum segmentation models under low-SNR conditions. Compared to Deep Lofargram, which relies solely on deep convolutional processing, LSE-Net maintains highly efficient spectral line extraction capabilities in complex noisy environments through its optimization of hydroacoustic line spectrum features. This provides a more reliable technical solution for detecting ship radiated noise line spectra in practical marine scenarios.

This section selected four simulated spectrograms with differing SNRs for testing. The detection results of various methods are shown in Fig. 8. The first and second rows of Fig. 8 present detection outcomes at SNRs of −16 dB and −18 dB, respectively. LSE-Net, FLD-Net, and Trans-UNet accurately detected all spectral lines, whereas Deep Lofargram exhibited partial false alarms. The third row of Fig. 8 depicts results at −20 dB, where Deep Lofargram exhibited increased false positives and missed detections, while FLD-Net experienced spectral line discontinuities. Conversely, the performance of LSE-Net and Trans-UNet remained unaffected. As shown in the final row of Fig. 8, LSE-Net maintained robust detection performance even at an SNR of −22 dB. By contrast, Deep

Table 1 Detection performance of different networks

Network	−16 dB			−18 dB			−20 dB			−22 dB		
	PD	FAR	LLA	PD	FAR	LLA	PD	FAR	LLA	PD	FAR	LLA
Deep Lofargram[18]	0.801	0.323	0.790	0.779	0.356	0.752	0.617	0.419	0.653	0.394	0.647	0.423
FLD-Net[15]	0.941	0.328	0.814	0.922	0.328	0.819	0.761	0.404	0.739	0.485	0.542	0.545
Trans-UNet[24]	0.982	0.325	0.824	0.976	0.323	0.834	0.943	0.394	0.790	0.722	0.410	0.709
LSE-Net	0.984	0.324	0.828	0.985	0.319	0.832	0.973	0.326	0.825	0.965	0.373	0.786

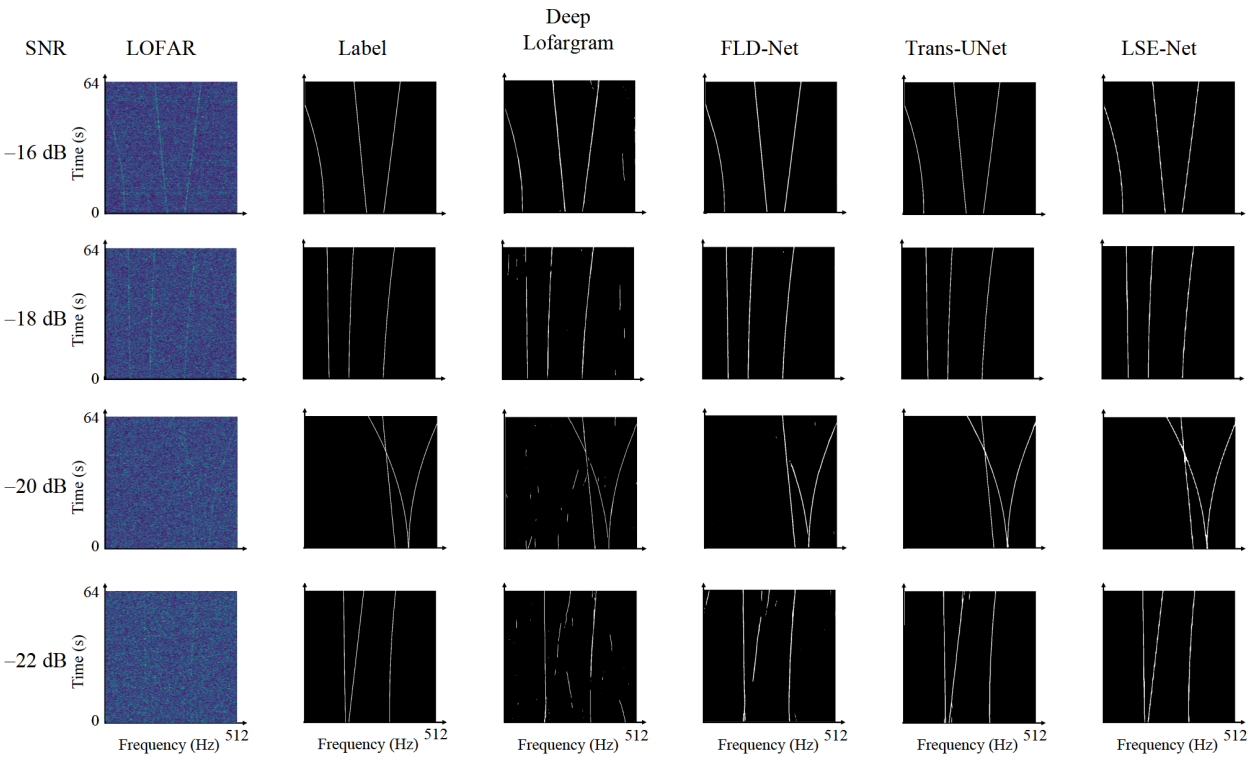


Fig. 8 Line spectrum extraction results under different SNRs.

Lofargram's results carry a significant risk of false positives, with spectral lines no longer discernible. FLD-Net detects some spectral lines, while Trans-UNet identifies spectral lines but also produces partial false alarms. Through the aforementioned comparative methods, the proposed LSE-Net achieves favorable detection performance across all four SNR conditions tested, consistent with the metric evaluations presented earlier.

5.3 Ablation experiment

To validate the extent to which each module enhances network performance, this section designs ablation experiments to evaluate individual components. The experiments use "Baseline" as the control benchmark, employing a Hiera encoder without an adapter, omitting the MRF module in skip connections, and removing the RCA for high-dimensional features in the decoder. Multiple comparative experiments were constructed by configuring the adapter, MRF module, and RCA module. The changes in model performance under the individual effects of each module are analyzed, with the experimental results presented in Table 2.

Following training on large-scale general data, the Hiera encoder demonstrates excellent general feature mapping capabilities. However, its domain adaptability to the spectral line features of ship radiation noise spectra still requires quantitative validation through experimentation. To clarify the efficacy of the adapter, this experiment established two comparative models: one serving as the baseline model, and another representing an improved model where the adapter was added before each encoder module based on the baseline. The experiment employed three key quantitative metrics—probability of detection, false alarm rate, and line localization accuracy—to evaluate the optimization effect of the adapter.

Experimental results demonstrate that incorporating the adapter into the baseline network model significantly reduces the false alarm rate while simultaneously improving the probability of detection. This indicates that the adapter effectively enhances the network model's capability to extract and localize spectral line features. As the SNR decreases, the performance enhancement brought by the adapter becomes more pronounced. This validates the module's adaptive advantage in scenarios with increased noise interference. Through minor parameter fine-tuning, this module enables pre-trained models to rapidly adapt to hydroacoustic spectral line features, effectively mitigating the performance degradation of general pre-trained models in specific domains.

Differences in feature distribution across different levels in skip-connections reduce the decoder's spectral line reconstruction accuracy. The MRF module achieves adaptive mapping of multi-layer features by unifying feature channel dimensions, with its optimization performance requiring quantitative validation. Experiments used the Baseline as a control benchmark, introducing the MRF module on top of

Table 2 Results of each module operating independently and different combinations of modules

Network	−16 dB			−18 dB			−20 dB			−22 dB		
	PD	FAR	LLA	PD	FAR	LLA	PD	FAR	LLA	PD	FAR	LLA
BaseLine	0.969	0.294	0.834	0.941	0.322	0.823	0.778	0.505	0.648	0.477	0.660	0.452
BaseLine + Adapter	0.971	0.342	0.805	0.958	0.341	0.814	0.888	0.436	0.734	0.648	0.575	0.580
BaseLine + MRF	0.973	0.268	0.847	0.960	0.274	0.853	0.875	0.403	0.747	0.629	0.556	0.587
BaseLine + RCA	0.977	0.281	0.839	0.964	0.292	0.843	0.862	0.403	0.755	0.621	0.549	0.589
BaseLine + Adapter + MRF	0.975	0.264	0.849	0.964	0.269	0.856	0.880	0.381	0.768	0.619	0.530	0.603
BaseLine + MRF + RCA	0.982	0.269	0.846	0.974	0.266	0.856	0.919	0.356	0.788	0.673	0.531	0.613
BaseLine + Adapter + RCA	0.975	0.264	0.849	0.963	0.270	0.856	0.875	0.379	0.769	0.625	0.543	0.592
LSE-Net	0.982	0.259	0.851	0.973	0.262	0.858	0.915	0.336	0.805	0.679	0.534	0.614

the Baseline. Analyzing scenarios with gradient SNRs ranging from −16 dB to −22 dB, the MRF module's enhancement of feature fusion quality was examined. Results indicate that the MRF module's optimization effect on model performance progressively strengthens as the SNR decreases. In high-SNR scenarios, the model's performance improvement over the Baseline is limited. Conversely, within the low-SNR range, the optimization effect is pronounced, with significant enhancements observed in PD and LLA, alongside a concurrent reduction in FAR. Concurrently, the experimental findings indicate that the performance gain from employing MRF alone is weaker than that achieved using adapter alone, suggesting that feature mapping optimization must be built upon a robust feature mapping foundation. Experimental results demonstrate that incorporating the MRF module significantly improves the model's performance under complex noisy backgrounds, with particularly pronounced advantages in multi-spectral line detection tasks.

The reverse attention module within the decoder achieves noise suppression and spectral edge enhancement through weight redistribution. To validate its function independently, interference from other modules must be excluded. This experiment constructs a "Baseline + RCA" network model to analyze the mechanism's contribution to feature extraction. Experimental results indicate that the RCA module's optimization focuses on false alarm reduction and improved localization accuracy. Across various SNRs, this network model exhibits a significantly lower FAR compared to the Baseline model. The LLA metric also shows slight improvement, though the relative gain in PD remains limited. These findings suggest the RCA module's core function lies in precisely identifying spectral line edge features to filter noise channels unrelated to the spectral line.

To evaluate functional complementarity among modules, cross-combination models were constructed, with the three-module fusion model LSE-Net serving as the final architecture. Performance metrics across the full SNR range demonstrate the synergistic effects of module integration. The experimental results are presented in Table 2. They demonstrate that the module combinations exhibit functional complementarity, with LSE-Net achieving overall optimal performance across all three metrics. The synergistic effect between the adapter and MRF is particularly pronounced in the LLA metric, indicating that the adapter's feature extraction and the MRF's feature mapping optimization form a closed-loop system, reducing the loss of line spectrum features during jump connections. The MRF-RCA combination significantly enhances PD while reducing FAR compared to either method alone. This demonstrates synergistic feature optimization and noise filtering, where MRF-processed spectral features provide a foundation for RCA's precise extraction. The adapter-RCA collaboration exhibits strong FAR suppression, validating the fusion of feature extraction with high-dimensional feature denoising.

LSE-Net, as the final model integrating three modules, demonstrated optimal performance across all SNRs, fully validating the rationale behind the three components: spectral feature extraction, feature optimization, and noise suppression for high-position features. Notably, all combination models incorporating an adapter significantly outperformed their adapter-free counterparts. This indicates that the MRF and RCA modules only function with enhanced precision when the generalized pre-trained model is adapted to acoustic spectral features via the adapter.

5.4 Results on real trial data

The real experimental data in this paper were collected using hydrophones deployed at a depth of 15 m underwater, with a sampling rate set to 4 kHz. The data acquisition-scenario during the lake trial phase is shown in Fig. 9, in the horizontal dimension, a fixed sound source was placed 1 km away from the hydrophones, while a moving ship, serving as the detection target, navigated within a 2 km range from the hydrophones. The first set of experiments selected a segment of time-varying spectral (TVS) signals containing an accelerating ship, primarily consisting of three line spectra with frequency ranges of 104–178, 150–196, and 50–88 Hz. The second set of experiments involved signals from a fixed

sound source emitting multiple continuous wave (CW) signals, with line spectrum frequencies of 314, 574, 624, and 664 Hz.

The spectrograms in simulated data are typically generated using simplified models. Compared to the target noise in real-world environments, the spectral features in the simulated data are more ideal. The signal structure is relatively stable, and the noise characteristics are simpler. However, in real data, the underwater environment is much more complex, with signals potentially influenced by multipath propagation, oceanic environmental changes, and other factors, which increases the complexity of the spectral features. The spectrograms in real data often exhibit stronger dynamic changes and nonlinear characteristics, which impose higher demands on the performance of line spectrum extraction models.

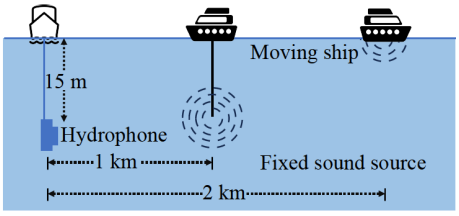


Fig. 9 Schematic diagram of lake trial experiments.

Figure 10 shows the results of line spectrum extraction using LSE-Net on the real experimental data. The frequency drift of the first segment of signals differs from the simulated signal. The frequency change of the ship's signal is influenced by the ship's movement state at the time, resulting in more significant frequency fluctuations compared to the smooth frequency changes seen in the simulated data. This makes the extraction of line spectra in the first segment more challenging, with line spectrum breakage occurring during periods of drastic frequency changes, although the overall extraction performance remains good. The second segment of the signal is relatively stable and matches the frequency changes in the simulated data, so LSE-Net performs well in terms of detection.

These experiments indicate that while the LSE-Net model trained on simulated data experiences performance fluctuations due to the complexity of spectral features and noise types in real data, it still shows good adaptability and extraction performance

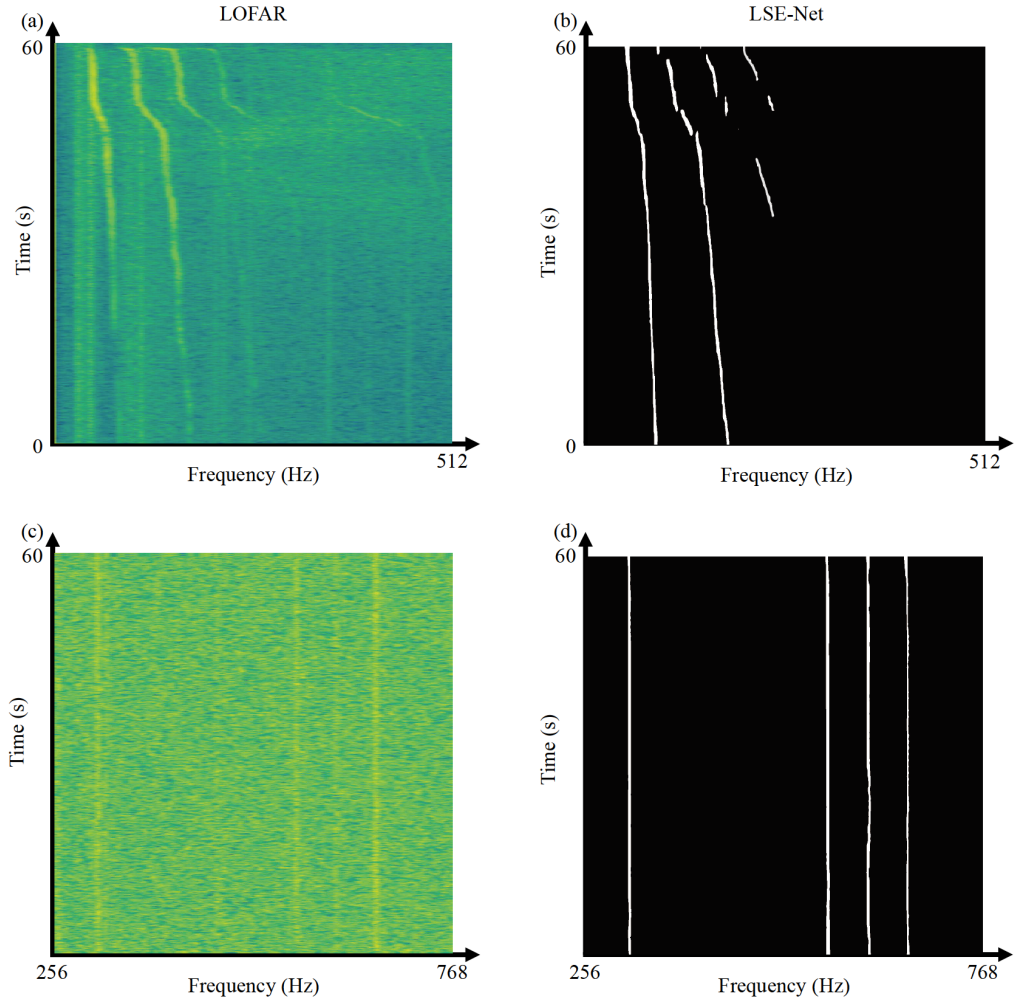


Fig. 10 Spectrograms and results of real trial data: (a) First set of experimental signal spectra; (b) The results of the first experiment; (c) Second set of experimental signal spectra; (d) The results of the second experiment.

for line spectrum extraction tasks in real-world scenarios, demonstrating its effectiveness in complex environments.

6 Conclusion

This paper addresses the challenge of line spectrum extraction from LOFAR spectrograms under low SNRs by designing and implementing LSE-Net. The primary research methodology is as follows.

(1) The network architecture is designed for ease of use and scalability. Adopting a classical encoder-decoder structure reduces the complexity of model deployment and debugging. A modular design provides ample scope for integrating new feature processing modules and adapting to diverse data types. This resolves the poor generalizability inherent in traditional line spectrum extraction.

(2) An efficient parameter-efficient fine-tuning strategy is employed. By incorporating adapters before each encoder level, the model achieves rapid convergence and learns spectral line features even on small sample datasets. This reduces reliance on large-scale annotated observational data, effectively mitigating the challenges of high acquisition and annotation costs for observational data while enhancing the feasibility of technological implementation.

(3) The spectral line extraction capability is enhanced through feature augmentation mechanisms. Multi-field feature fusion modules capture diverse spectral line information, while reverse attention modules identify edge regions to further improve feature discrimination. These dual mechanisms enable precise spectral line localization even under low SNRs, delivering robust extraction accuracy.

In summary, the proposed spectral line extraction network effectively resolves spectral line extraction challenges in low-SNR LOFAR spectra through its robust architecture, optimized parameter training strategy, and feature enhancement design. It demonstrates superior spectral line extraction performance on both simulated and real datasets. Future research may explore multimodal feature fusion schemes, leveraging the spatio-temporal correlations inherent in LOFAR data to enhance network performance under extreme low-SNR conditions. Concurrently, the network's applicability could be extended to the analysis of spectral maps from other types of array signals.

References

[1] Yu L, White L B. Broadband passive sonar detection using rational orthogonal wavelet filter banks. Seventh International Conference on Intelligent Sensors, Sensor Networks and Information Processing, IEEE, Adelaide, SA, Australia, 2011: 461–466

[2] Zhu C, Gaggero T, Makris N C, *et al*. Underwater sound characteristics of a ship with controllable pitch propeller. *J. Mar. Sci. Eng.*, 2022, **10**(3): 328

[3] Parsons M, Meekan M. Acoustic characteristics of small research vessels. *J. Mar. Sci. Eng.*, 2020, **8**(12): 970

[4] Hu L, Wang X, Wang S. Decentralized underwater target detection and localization. *IEEE Sens. J.*, 2020, **21**(2): 2385–2399

[5] Casselman F L, Freeman D F, Kerrigan D A, *et al*. A neural network-based passive sonar detection and classification design with a low false alarm rate. IEEE Conference on Neural Networks for Ocean Engineering, IEEE, Washington, DC, USA, 1991: 49–55

[6] Paris S, Jauffret C. Frequency line tracking using hmm-based schemes [passive sonar]. *IEEE Trans. Aerosp. Electron Syst.*, 2003, **39**(2): 439–449

[7] Paris S, Jauffret C. A new tracker for multiple frequency line. 2001 IEEE Aerospace Conference Proceedings (Cat. No. 01TH8542), IEEE, Big Sky, MT, USA, 2001, **4**: 1771–1782

[8] Luo X, Shen Z. A space-frequency joint detection and tracking method for line-spectrum components of underwater acoustic signals. *Appl. Acoust.*, 2021, **172**: 107609

[9] Abel J S, Lee H J, Lowell A P. An image processing approach to frequency tracking (application to sonar data). Proceedings of ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE, San Francisco, CA, USA, 1992, **2**: 561–564

[10] Gillespie D. Detection and classification of right whale calls using an "edge" detector operating on a smoothed spectrogram. *Can. Acoust.*, 2004, **32**(2): 39–47

[11] Lampert T A, O'Keefe S E M. An active contour algorithm for spectrogram track detection. *Pattern Recognit. Lett.*, 2010, **31**(10): 1201–1206

[12] Lampert T A, O'Keefe S E M. On the detection of tracks in spectrogram images. *Pattern recognit.*, 2013, **46**(5): 1396–1408

[13] Zhao L L, Li B Z. A novel method for extracting frequency line on Lofargram based on feature function. Proceedings of the 2022 4th International Conference on Image, Video and Signal Processing, New York, NY, USA, 2022: 25–31

[14] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, *et al*. (eds). Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015: 234–241

[15] Jiang D L, Luo X, Shen Q. Frequency line detection in spectrograms using a deep neural network with attention. *J. Acoust. Soc. Am.*, 2024, **156**(5): 3204–3216

[16] Hong F, Liu C, Guo L, *et al*. Underwater acoustic target recognition with a residual network and the optimized feature extraction method. *Appl. Sci.*, 2021, **11**(4): 1442

[17] Ashraf H, Jeong Y, Lee C H. Underwater ambient-noise removing GAN based on magnitude and phase spectra. *IEEE Access*, 2021, **9**: 24513–24530

[18] Han Y, Li Y, Liu Q, *et al*. DeepLofargram: a deep learning based fluctuating dim frequency line detection and recovery. *J. Acoust. Soc. Am.*, 2020, **148**(4): 2182–2194

[19] Liang G, Chen Y, Wang J, *et al*. Enhanced noise resilience in passive tone detection via broad-receptive field complex-valued convolutional neural networks. *J. Acoust. Soc. Am.*, 2024, **155**(6): 3968–3982

[20] Luo X, Zhang M, Liu T, *et al*. An underwater acoustic target recognition method based on spectrograms with different resolutions. *J. Mar. Sci. Eng.*, 2021, **9**(11): 1246

[21] Ryali C, Hu Y T, Bolya D, *et al*. Hiera: a hierarchical vision transformer without the bells-and-whistles. Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA, 2023: 29441–29454

[22] Ravi N, Gabeur V, Hu Y T, *et al*. SAM 2: segment anything in images and videos. arXiv preprint: 2408.00714, 2024

[23] Hu E J, Shen Y, Wallis P, *et al*. LoRA: low-rank adaptation of large language models. The Tenth International Conference on Learning Representations (ICLR 2022), Virtual Event, OpenReview.net, 2022

[24] Chen J, Mei J, Li X, *et al*. TransUNet: rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Med. Image Anal*, 2024, **97**: 103280