Archives of Acoustics Vol. 50, No. 2, pp. 187–198 (2025), https://doi.org/10.24425/aoa.2025.153659

Research Paper

Single Vector Hydrophone DOA Estimation: Leveraging Deep Learning with CNN-CBAM

Fanyu ZENG, Yaning HAN, Hongyuan YANG, Dapeng YANG, Fan ZHENG*

Key Laboratory of Geophysical Exploration Equipment, Ministry of Education, College of Instrumentation and Electrical Engineering, Jilin University
Changchun, China

*Corresponding Author e-mail: zhengfan@jlu.edu.cn

Received November 18, 2024; accepted March 9, 2025; published online May 5, 2025.

In recent years, single vector hydrophones have attracted widespread attention in target direction estimation due to their compact design and advantages in complex underwater acoustic environments. However, traditional direction of arrival (DOA) estimation algorithms often struggle to maintain high accuracy in non-stationary noise conditions. This study proposes the novel DOA estimation method based on a convolutional neural network (CNN) and the convolutional block attention module (CBAM). By inputting the covariance matrix of the received signal into the neural network and integrating the CBAM module, this method enhances the model's sensitivity to critical features. The CBAM module leverages channel and spatial attention mechanisms to adaptively focus on essential information, effectively suppressing noise interference and improving directional accuracy. Specifically, CBAM improves the model's focus on subtle directional cues in noisy environments, suppressing irrelevant interference while amplifying essential signal components, which is crucial for an accurate DOA estimation. Experimental results under various signal-to-noise ratio (SNR) conditions validate the method's effectiveness, demonstrating superior noise resistance and estimation precision, providing a robust and efficient solution for underwater acoustic target localization.

Keywords: single vector hydrophone; direction of arrival (DOA); convolutional neural network (CNN); convolutional block attention module (CBAM); noise resistance.

1. Introduction

In recent years, vector hydrophones have gained a wide range of research and applications in underwater target detection and localization. Compared with traditional scalar hydrophones, vector hydrophones can record acoustic pressure information and partially mitigate isotropic noise. Traditional direction of arrival (DOA) estimation algorithms primarily include high-resolution algorithms based on eigenvalue decomposition, such as multiple signal classification (MUSIC) and estimation of signal parameters via rotational invariance techniques (ESPRIT) (Tichavsky *et al.*, 2001), which perform well in idealized stationary noise environments. However, nonlinear effects, noise interference, and multipath effects in real marine environments often degrade the algorithm performance.

Advances in deep learning have facilitated the application of various neural networks in the underwater DOA estimation. Xiao *et al.* (2020) proposed a deep unfolding network called DeepFPC, which is based on a fixed-point algorithm, utilizes 1-bit quantization measurements for sparse signal recovery, and has been successfully applied to the DOA estimation, significantly improving estimation accuracy and computational efficiency. In parallel with deep learning advancements, Xu *et al.* (2022) developed a block sparse-based dynamic compressed sensing estimator for underwater acoustic communication, addressing challenges like impulsive noise. This method's adaptability to varying underwater channel conditions enhances the potential of neural network-based approaches for improving estimation accuracy in noisy, real-world marine environments. Liu *et al.* (2024) in-

troduced a deep learning-based method for graph similarity computation, which contributes to the field of signal processing and could further enhance DOA estimation techniques. In parallel, Xu *et al.* (2019) proposed the M-SIMMUKF algorithm for tracking underwater maneuvering targets, demonstrating its robustness under dynamic and noisy conditions. Moreover, Wajid *et al.* (2020; 2022) explored direction estimation and tracking methods using acoustic vector sensors, highlighting their ability to improve source localization in underwater environments, providing insights that complement deep learning methods for the robust DOA estimation in real-world conditions. Liu *et al.* (2021) proposed the DOA estimation method for underwater acoustic arrays based on a convolutional neural network (CNN), which significantly enhanced the direction estimation accuracy of underwater signals, with strong adaptability and excellent noise resistance. Varanasi *et al.* (2020) combined spherical harmonic decomposition with a deep learning framework to achieve the robust DOA estimation in complex environments, providing an effective solution for signal direction estimation under high-noise conditions. Numerous studies have shown that neural network-based methods can effectively improve the DOA estimation accuracy and adaptability in complex noise environments. For example, Yao *et al.* (2020) proposed a recursive neural network model that achieves the DOA estimation for unknown signal sources through the Toeplitz matrix reconstruction. Niu *et al.* (2017a; 2017b) investigated the performance of three machine learning methods – feedforward neural networks, support vector machines, and random forests – based on vertical arrays for source ranging and validated the feasibility and effectiveness of these methods at different signal-to-noise ratios (SNR). Progress has also been made in machine learning applications for underwater surface and subsurface target resolution in vertical arrays, and direction estimation with horizontal arrays. Chi *et al.* (2019) employed a feedforward neural network combined with early stopping for source ranging, which effectively enhanced the model's generalization ability, allowing it to maintain strong ranging performance across various environments. Choi *et al.* (2019) used supervised learning methods to classify surface and submerged vessels in the ocean, significantly improving classification accuracy, demonstrating the potential of their method for practical marine monitoring. Ozanich *et al.* (2020) employed a feedforward neural network for the DOA estimation, demonstrating the efficiency and accuracy of their method in underwater acoustics, further validating the potential of deep learning in this field. These methods demonstrate excellent performance not only in synthetic data but also show significant potential in practical ocean experiments. The application of neural networks and machine learning expands the possibilities for the DOA estimation with single vector hydrophones, particularly in terms of adaptability and real-time performance.

This paper proposes the CNN-CBAM-based DOA estimation method that uses a normalized covariance matrix as input and incorporates the convolutional block attention module (CBAM) to enhance key feature extraction (Woo *et al.*, 2018) this design is particularly advantageous for the underwater DOA estimation, where capturing subtle directional cues amidst noise is critical. The model is trained on a simulated dataset to improve generalization. Experimental results validate the performance advantages of this approach under varying signal-to-noise conditions, providing an efficient and robust solution for underwater target direction estimation.

2. Vector signal model and data preprocessing

2.1. Single vector hydrophone signal model

Under the far-field plane wave assumption, a single vector hydrophone can measure the sound pressure p and the three velocity components, ν_x, ν_y, and ν_z, at a single point in the sound field. Under ideal conditions, the sensitivities of the sound pressure sensors and velocity sensors are identical, so the received signal for a single vector hydrophone can be represented as

$$\begin{aligned} p(t) &= \sum_{i=1}^{N} s_i(t) + n_p(t), \\ \nu_x(t) &= \sum_{i=1}^{N} s_i(t)\cos\theta_i\cos\varphi_i + n_x(t), \\ \nu_y(t) &= \sum_{i=1}^{N} s_i(t)\sin\theta_i\cos\varphi_i + n_y(t), \\ \nu_z(t) &= \sum_{i=1}^{N} s_i(t)\sin\varphi_i + n_z(t), \end{aligned} \tag{1}$$

where $s_i(t)$ represents the incident signal from the i-th source; θ_i and φ_i denote the horizontal and pitch angles, respectively; $n_p(t)$, $n_x(t)$, $n_y(t)$, and $n_z(t)$ indicate the noise in the sound pressure and three velocity channels. This expression can be rewritten in a matrix form for further processing:

$$x(t) = A \cdot s(t) + n, \tag{2}$$

$$x(t) = [p(t)\ \ v_x(t)\ \ v_y(t)\ \ v_z(t)]^{\mathrm{T}}, \tag{3}$$

$$A = \begin{bmatrix} 1 & 1 & & 1 & & 1 \\ \cos\theta_1\cos\varphi_1 & \cos\theta_2\cos\varphi_2 & \cdots & \cos\theta_i\cos\varphi_i & \cdots & \cos\theta_N\cos\varphi_N \\ \sin\theta_1\cos\varphi_1 & \sin\theta_2\cos\varphi_2 & & \sin\theta_i\cos\varphi_i & & \sin\theta_N\cos\varphi_N \\ \sin\varphi_1 & \sin\varphi_2 & & \sin\varphi_i & & \sin\varphi_N \end{bmatrix}^{\mathrm{T}}, \tag{4}$$

$$s(t) = [s_1(t)\ \ s_2(t)\ \ \ldots\ \ s_i(t)\ \ \ldots\ \ s_N(t)]^{\mathrm{T}}, \quad (5)$$

$$n = [n_p(t)\ \ n_x(t)\ \ n_y(t)\ \ n_z(t)]^{\mathrm{T}}. \quad (6)$$

2.2. Data preprocessing

Before the received signal $x(t)$ is input into the neural network, it requires preprocessing to enable the neural network to more effectively extract features. First, the covariance matrix $\mathbf{Rxx}$ of the received signal is computed:

$$\mathbf{Rxx} = E\left[x(t)x^{\mathrm{T}}(t)\right], \quad (7)$$

followed by normalization of $\mathbf{Rxx}$:

$$\mathbf{Rxx}_{i,j} = \begin{cases} 2 \times \dfrac{\mathbf{Rxx}_{i,j} - \min_{\mathrm{val}}}{\max_{\mathrm{val}} - \min_{\mathrm{val}}} - 1, & \mathbf{Rxx}_{i,j} \neq 0, \\ 0, & \mathbf{Rxx}_{i,j} = 0, \end{cases} \quad (8)$$

where $\mathbf{Rxx}_{i,j}$ represents the element at the i-th row and j-th column of the matrix $\mathbf{Rxx}$; $\min_{\mathrm{val}}$ and $\max_{\mathrm{val}}$ are the minimum and maximum elements of the matrix $\mathbf{Rxx}$, respectively. By normalizing only the non-zero elements, we retain the sparse structure of $\mathbf{Rxx}$, which is crucial for maintaining the integrity of the signal representation.

3. Network-based direction estimation of a single vector hydrophone

3.1. Convolutional neural network

A CNN is a feedforward deep neural network based on convolutional computations and consists of input, hidden and output layers. The hidden layers include convolutional layers, activation functions, pooling layers, and fully connected layers. The convolutional layer performs convolution operations on input data using kernels of various sizes, with each layer containing multiple kernels. Each kernel consists of weight coefficients and biases and is activated by an activation function. The convolutional layers are connected in sequence to extract higher-dimensional data features through multiple convolution operations. The formula for the convolutional layer is as follows:

$$f_j^{(l)} = g\left(\sum_{i=1}^{n} w_{ij}^{(l)} \cdot x_i^{(l-1)} + b_j^{(l)}\right), \quad (9)$$

where $f_j^{(l)}$ represents the feature value of the j-th feature in the l-th layer, capturing the output of the convolution operation for this feature; $w_{ij}^{(l)}$ is the weight coefficient connecting the i-th input feature in the $(l-1)$-th layer to the j-th feature in the current layer; $x_i^{(l-1)}$ denotes the feature value of the i-th input in the $(l-1)$-th layer, serving as the input to the current layer; $b_j^{(l)}$ represents the bias term corresponding to the j-th feature in the l-th layer, which offsets the weighted sum of inputs; $g(\cdot)$ is the activation function, unlike a sigmoid function, which is inherently nonlinear, ReLU is piecewise linear but still allows the network to model complex relationships through its composition across multiple layers. Lastly, n is the number of input features in the previous layer. The fully connected layer links the extracted features through neurons and uses ReLU as the activation function:

$$g(x) = \max(0, x). \quad (10)$$

In recent years, CNNs have shown significant potential for improving the DOA estimation accuracy through their feature extraction capabilities. However, underwater acoustic environments present unique challenges; complex noise and interference can hinder CNN's ability to focus on key features. These challenges necessitate an enhanced model structure that can effectively extract features while dynamically adjusting its focus to prioritize relevant information within high-dimensional data.

The CBAM addresses this issue by introducing an adaptive attention mechanism that refines feature selection based on channel and spatial importance. Integrating CBAM into the CNN architecture enables the model to selectively enhance informative features while suppressing irrelevant background noise. This design is particularly advantageous for the underwater DOA estimation, where capturing subtle directional cues amidst noise is critical. By leveraging the channel and spatial attention, CBAM integration not only enhances directional discrimination but also improves the robustness and accuracy of the DOA estimation process.

3.2. CBAM module

To fully leverage the CNN's capability for feature extraction from high-dimensional data matrices, this study improves the traditional neural network by introducing CBAM to the CNN structure, thereby enhancing the model's detail extraction capability for the DOA estimation. The CBAM spatial-channel attention module is illustrated in Fig. 1.

The channel attention module (CAM) adaptively adjusts channel weights in the feature map, enhancing feature selection. For instance, CAM effectively emphasizes subtle directional cues in underwater acoustic data, improving the model's focus amidst noise interference. First, the input feature map undergoes global average pooling and max pooling along the spatial dimension, resulting in two channel descriptors that represent global average and maximum features. Next, these descriptors pass through a shared fully connected layer sequence, including layers for dimensionality reduction and restoration, with a ReLU ac-

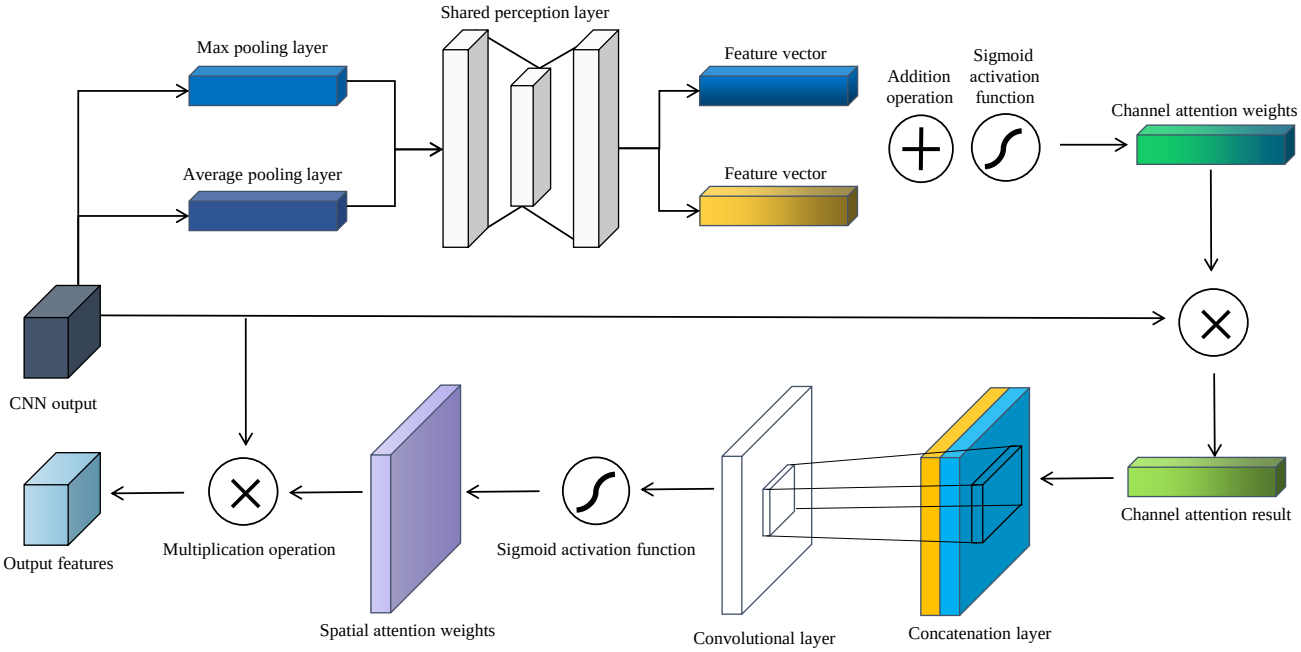

Fig. 1. Working principle of the CBAM attention mechanism in feature enhancement.

tivation function connecting the intermediate layers. Then, the two feature vectors are summed and passed through a Sigmoid activation function to obtain the attention weights for each channel. Finally, these weights are multiplied by the original input feature map on a per-channel basis, completing the channel weighting process. This design allows the network to adaptively allocate attention based on the global features of each channel, thereby effectively enhancing its focus on target features. The Sigmoid activation function is defined as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{11}$$

where x represents the input value, and $\sigma(x)$ represents the output of the Sigmoid function.

The spatial attention module (SAM) learns the weight distribution in the spatial dimension to highlight key area information, suppressing interference from background or irrelevant regions. First, the input feature map undergoes average pooling and max pooling along the channel dimension to produce two single-channel feature maps, representing spatial average and maximum information, respectively. Next, these two feature maps are concatenated along the channel dimension to form a two-channel feature map. This concatenated feature map is then processed by a convolutional layer with a kernel size of 7×7, capturing a broader range of spatial dependencies and producing a single-channel spatial attention weight map. Finally, this weight map is passed through a Sigmoid activation function and multiplied element-wise with the original input feature map to complete spatial weighting. Through this approach, the SAM can adaptively focus on key regions within the feature map, enhancing the model's spatial representation capability.

The CBAM combines channel attention and spatial attention to dynamically adjust the weights of key information within the feature map. Channel attention emphasizes key feature channels to enhance the role of different channel features in the network, while spatial attention focuses on critical regions within the feature map, thus capturing essential information required for accurate direction estimation.

3.3. Network structure

The overall network structure is illustrated in Fig. 2. This network model is a deep learning architecture based on a CNN combined with a CBAM, designed for the DOA estimation. The model includes two convolutional layers: the first layer increases the input feature channels from 1 to 32, and the second layer further increases the channels to 64. In the convolutional layers, '3×3' specifies the kernel size, and the third number indicates the number of kernels. Each convolutional layer is immediately followed by a CBAM module to enhance channel and spatial attention for the features. After processing by the convolutional layers and CBAM modules, the feature data is flattened and passed to two fully connected layers, containing 128 and 64 neurons, respectively, ultimately outputting two directional estimation values. Through the integration of convolution and attention mechanisms, this network structure can more effectively extract key features, thereby improving the accuracy of the DOA estimation.

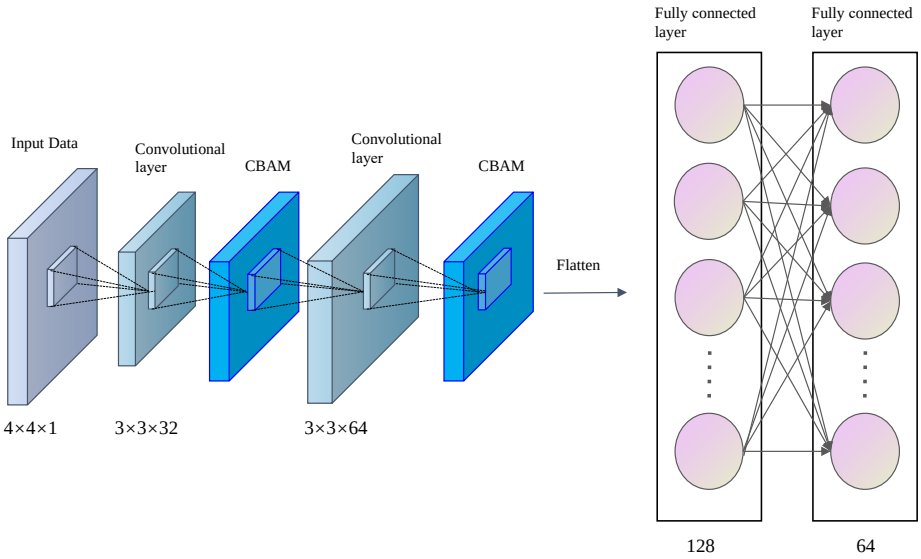

Fig. 2. Network architecture of CNN-CBAM.

3.4. Training process

Monte Carlo simulation is used to generate received signals $x(t)$ without added colored noise according to Eqs. (2)–(6); then, the covariance matrix **Rxx** of the received signal is computed according to Eq. (7) and normalized to a 4×4 matrix to serve as input for the neural network.

Each sample includes the covariance matrix of noiseless signals generated at the specific azimuth and elevation angles. First, an angle conversion factor, the number of array elements, the sampling frequency, and the time sequence are set up to simulate the basic received signal. Sample angles are randomly generated within the specified azimuth and elevation ranges, and their corresponding array manifold vectors are calculated and multiplied with the basic signal to obtain the received data. The covariance matrix is then constructed from the received data, and its non-zero elements are normalized by mapping their values to the range $[-1, 1]$, resulting in a normalized covariance matrix. All generated covariance matrices form a dataset for a neural network input, with the array of the sample azimuth and elevation angles serving as output labels for training the deep learning model. The training data consists of noiseless, clean data, with target angles randomly selected between 0° and 359°. Figure 3 shows the time-domain waveforms of the

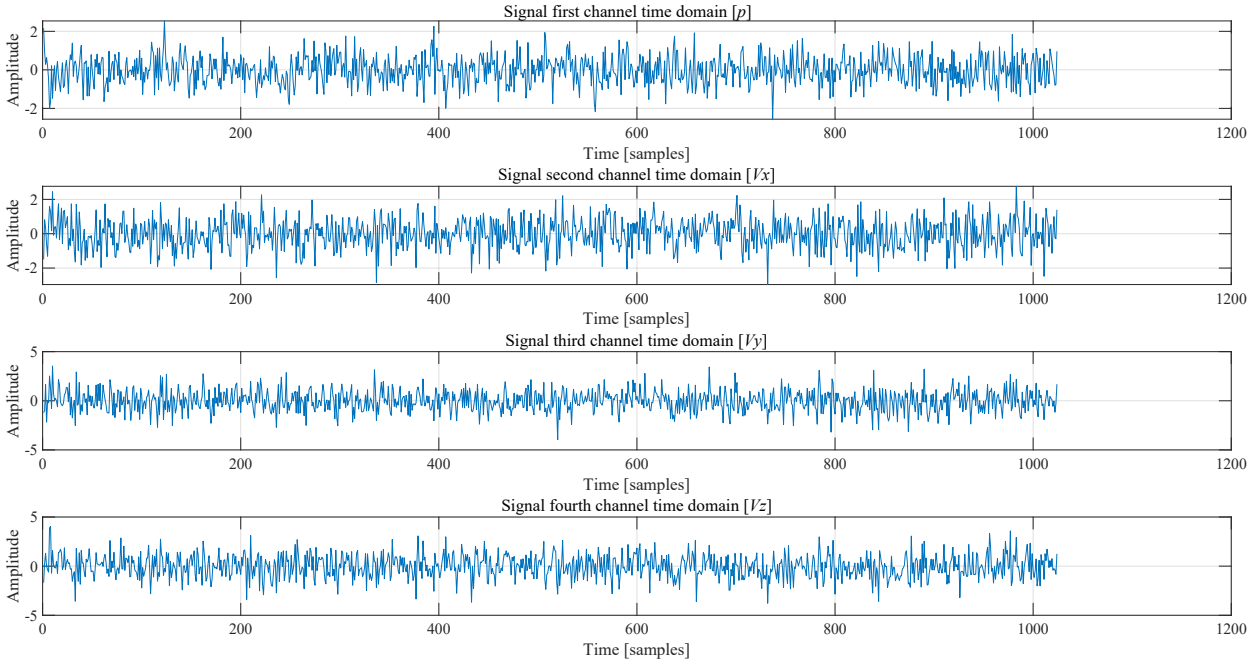

Fig. 3. Time-domain signals for four channels when the SNR is 0 dB.

received signal $x(t)$ for each of the four channels when the SNR is 0.

Training uses the mean squared error (MSE) loss function and the Adam optimizer. The model undergoes training for 180 epochs, with each epoch beginning by initializing the accumulated loss in training mode. Data is loaded in batches to the specified computation device (e.g., GPU), and the model outputs are obtained through forward propagation, with losses calculated between the output and true labels. Loss is backpropagated to update model parameters, and the mean loss for each epoch is accumulated. During training, a StepLR scheduler adjusts the learning rate dynamically every 100 epochs to enhance convergence. At the end of each epoch, the loss and current learning rate are recorded and displayed to monitor training progress. The training loss is shown in Fig. 4.

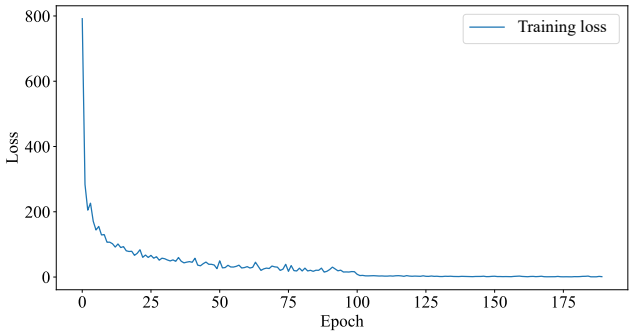

Fig. 4. Training loss variation.

To evaluate the potential overfitting issue, we conducted K-fold cross-validation (K = 4) and recorded the training and validation losses. The key parameters used in the validation process are as follows:

- K-value ($n_$splits): 4, indicating the dataset was divided into 4 subsets for cross-validation;
- batch size: 32, defining the number of samples processed in each iteration;
- number of epochs: 180, representing the total training iterations;
- optimizer: Adam, with a learning rate of 0.001;
- loss function: MSE, used to measure the discrepancy between predicted and true labels.

The results of the K-fold cross-validation are visualized in Fig. 5, which depicts the average training loss (blue line) and average validation loss (red line) over the epochs. The figure shows that both the training and validation losses exhibit a sharp decline in the initial epochs, followed by a gradual stabilization as the number of epochs increases. Notably, the training and validation loss curves remain closely aligned throughout the training process, the convergence of the loss curves to a low and stable level, along with the minimal gap between the training and validation losses, suggests that the model effectively avoids overfitting.

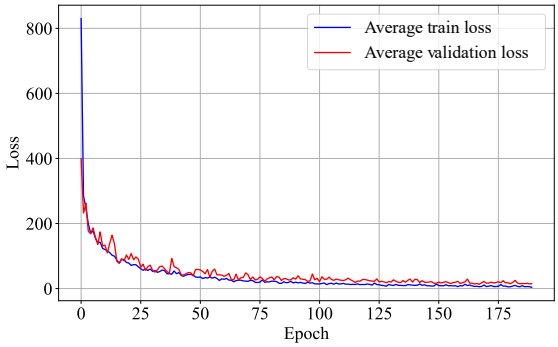

Fig. 5. Training and validation loss curves over epochs in K-fold cross-validation.

This demonstrates the model's ability to fit the data well.

4. Simulation results analysis

The study evaluates the CNN-CBAM model's single-target direction estimation performance across varying SNRs. In the simulation, the target azimuth and elevation angles are randomly selected within the range of 0° to 359°. SNR values are set to −5 dB, 0 dB, 5 dB, 10 dB, and 15 dB in 5 dB increments, with 10 000 data samples generated for each SNR, totaling 50 000 samples. The sampling frequency of the single vector hydrophone is 1 Hz, with each snapshot containing 1024 sample points and one direction estimated per snapshot. The x-axis and y-axis represent the azimuth and elevation angle errors relative to the target's true position, with blue points indicating errors within 10° for both angles. In Fig. 6, the left subfigure shows the histogram of azimuth estimation errors, while the right subfigure shows the histogram of elevation estimation errors. Each subfigure displays the error distributions under SNRs of 15 dB, 10 dB, 5 dB, 0 dB, and −5 dB. As SNR decreases, the error distribution gradually broadens, and errors increase. At higher SNRs, such as 15 dB, 10 dB, and 5 dB azimuth and elevation errors are primarily within 5°. At lower SNRs, like 0 dB and −5 dB, the CNN-CBAM model demonstrates reliable performance, with the majority of estimation errors not exceeding 15° in azimuth and 10° in pitch.

In this simulation, the target azimuth and elevation angles were set to [45°, 45°], with other simulation parameters remaining consistent with previous settings. The target direction was estimated using the weighted histogram method, MUSIC, Capon, the fourth-order cumulant method (Guo *et al.*, 2018), SBL (Liang *et al.*, 2021), and the CNN-CBAM model. Figure 7 illustrates the CNN-CBAM model's estimation results under various SNR conditions, where the x-axis and y-axis represent the azimuth and elevation angle errors relative to the true target position, with blue points

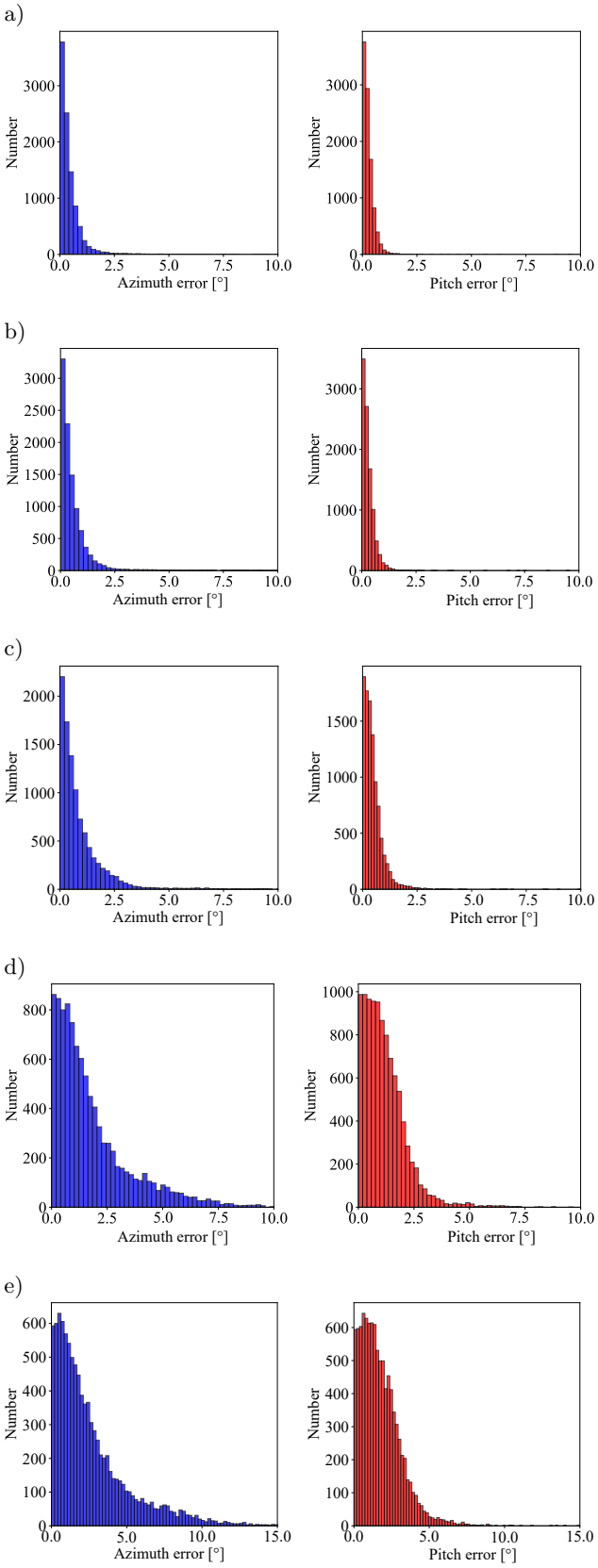

Fig. 6. Random direction estimation results of the CNN-CBAM method under different SNRs: a) SNR = 15 dB; b) SNR = 10 dB; c) SNR = 5 dB; d) SNR = 0 dB; e) SNR = −5 dB.

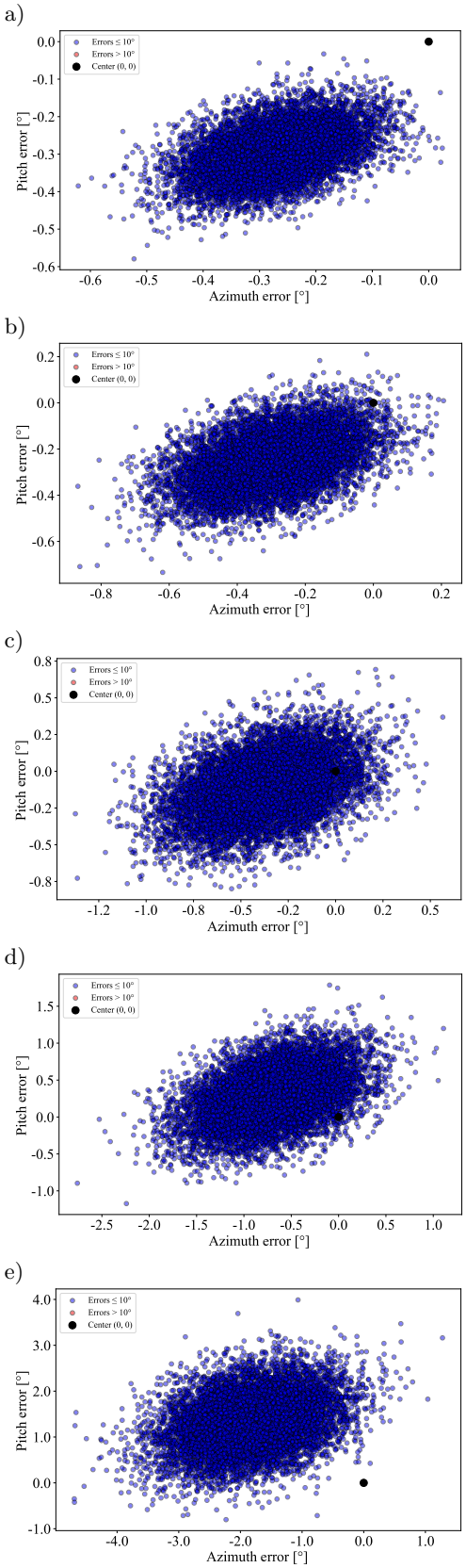

Fig. 7. Estimation results of the CNN-CBAM method for [45°, 45°] under different SNRs: a) SNR = 15 dB; b) SNR = 10 dB; c) SNR = 5 dB; d) SNR = 0 dB; e) SNR = −5 dB.

indicating errors within 10° for both angles. The results show that under SNR conditions of 5 dB, 10 dB, and 15 dB, the CNN-CBAM model achieves effective direction estimation with azimuth and elevation errors not exceeding 2°. Even under SNR conditions of −5 dB and 0 dB, the model successfully estimates the source's direction with errors in both azimuth and elevation within 7°.

The analysis of azimuth and pitch angle error distributions under varying SNR conditions (SNR = 15, 10, 5, 0, −5) reveals a systematic increase in distribution non-uniformity and measurement inaccuracy as SNR decreases. At high SNR (15 dB, 10 dB), the majority of error points are clustered in the southwest direction relative to the origin, indicating high measurement precision with minimal deviation. As SNR reduces to 5 dB, the error distribution shifts predominantly to the west, reflecting a moderate decline in accuracy. In low SNR conditions (0 dB and −5 dB), the error points are concentrated in the northwest direction, demonstrating significant dispersion and the emergence of systematic errors. This directional non-uniformity in the error distribution is attributed to noise interference and system instability, which are exacerbated under low SNR conditions. To quantify this systematic deviation, we introduce the concept of bias (B), defined as the mean difference between the estimated angles ($\widehat{y_i}$) and the true angles (y_i):

$$B = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i), \tag{12}$$

where n is the total number of measurements. This bias term captures the systematic error component, which becomes increasingly significant as SNR decreases, highlighting the need for robust error correction strategies in low SNR environments.

The bias in azimuth and pitch measurements refer to the systematic deviation of the estimated values from their true values. In this study, the bias is quantified as the mean error of azimuth and pitch measurements under different SNR conditions. As shown in Table 1, the mean azimuth errors exhibit a consistent negative bias across all SNR levels, ranging from −0.27° at 15 dB to −1.82° at −5 dB. This indicates that the azimuth estimates are systematically lower than the true values, and the magnitude of this bias increases with decreasing SNR.

Table 1. Biases of azimuth and pitch angles under different SNR conditions.

SNR [dB]	Azimuth error mean [°]	Pitch error mean [°]
15	−0.27095	−0.29417
10	−0.29320	−0.24869
5	−0.37225	−0.11144
0	−0.68524	0.28555
−5	−1.82368	1.33251

Similarly, the mean pitch errors demonstrate a transition from negative to positive bias as the SNR decreases. At higher SNR levels (e.g., 15 dB), the pitch errors show a negative bias of −0.29°, suggesting that the pitch estimates are slightly lower than the true values. However, as the SNR decreases, the bias shifts towards positive values, reaching 1.33° at −5 dB. This indicates that the pitch estimates become systematically higher than the true values under low SNR conditions.

The observed biases in both azimuth and pitch measurements highlight the influence of SNR on the accuracy of the estimation process. The increasing negative bias in azimuth and the transition from negative to positive bias in pitch suggest that the estimation algorithms may be more susceptible to noise in certain directions or dimensions. These findings emphasize the need for bias correction techniques, particularly in low SNR environments, to improve the accuracy of azimuth and pitch measurements.

Figure 8 shows the estimation results of different methods under a −5 dB SNR. MUSIC, Capon, weighted histogram, and fourth-order methods use a spectral peak search step size of 0.1°, while SBL employs a grid step size of 0.1°. CNN-CBAM and SBL stand out as the most effective methods for the DOA estimation, offering high accuracy. While weighted histogram and fourth-order cumulant methods remain competitive. MUSIC and Capon methods are more sensitive to noise and exhibit higher estimation errors. However, a notable limitation of CNN-CBAM in this scenario is that a significant portion of its estimates do not uniformly distribute around the origin, as observed in the error distribution plot. This deviation indicates that while CNN-CBAM achieves high accuracy in many cases, its estimates can be biased or skewed under low SNR conditions, leading to occasional instability. This limitation highlights the need for further refinement of the method to ensure more consistent and uniform performance across all scenarios. Future work could focus on enhancing the noise resilience of CNN-CBAM by optimizing its attention mechanisms, incorporating additional noise suppression techniques, or integrating it with probabilistic frameworks like those used in SBL to address this issue and improve its robustness in highly noisy environments.

This study adopts the root mean square error (RMSE) as the performance metric for direction estimation, where $\widehat{y_i}$ represents the estimated data and y_i represents the actual data:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}. \tag{13}$$

Figure 9 illustrates the RMSE of six DOA estimation methods, weighted histogram, MUSIC, Capon, fourth-order, CNN-CBAM, and SBL, across SNR levels ranging from −5 dB to 15 dB. As SNR increases,

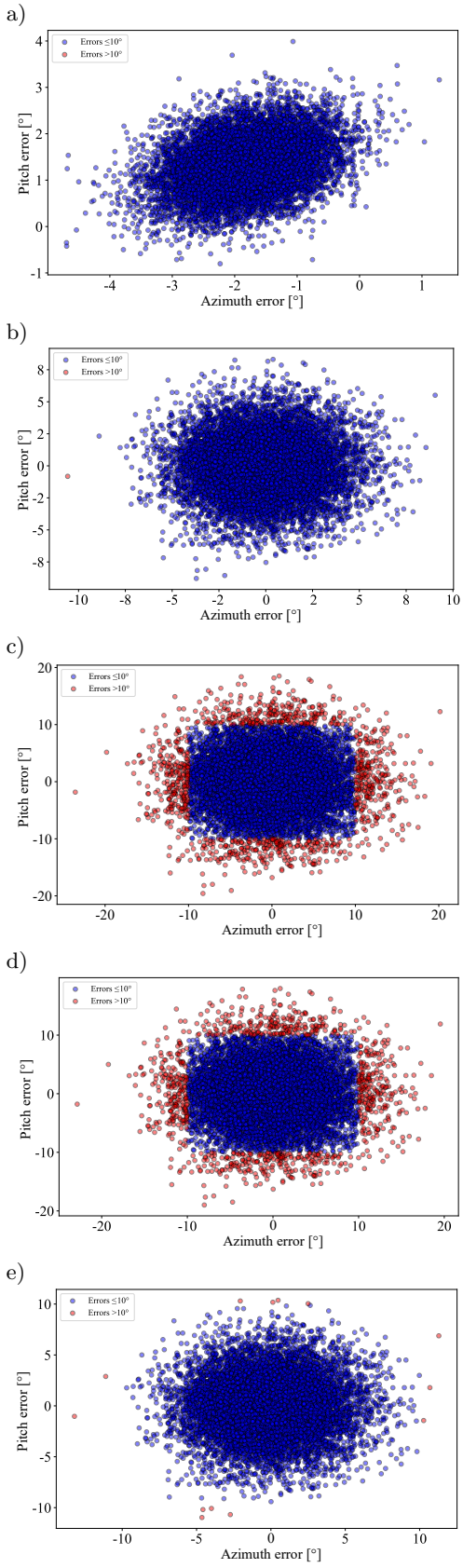

Fig. 8. Estimation results of various methods for [45°, 45°] when the SNR is −5 dB: a) CNN-CBAM; b) weighted histogram; c) MUSIC; d) Capon; e) fourth-order cumulant; f) SBL.

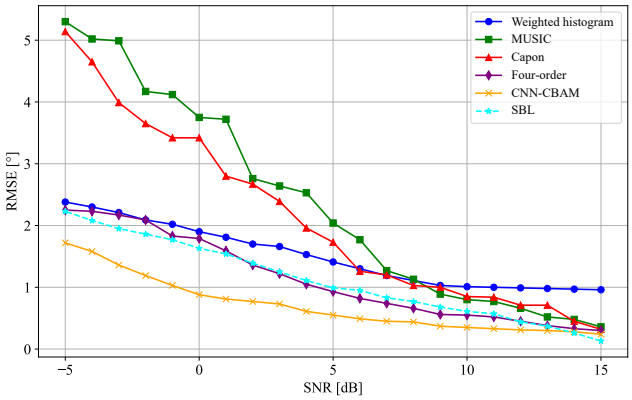

Fig. 9. RMSE of various methods for estimating [45°, 45°] under different SNRs.

the RMSE for all methods decreases, reflecting improved estimation accuracy. CNN-CBAM consistently achieves the lowest RMSE values, demonstrating high accuracy across all SNR conditions, particularly excelling at higher SNR levels. The fourth-order also performs well, closely following CNN-CBAM and SBL, while MUSIC and Capon show moderate performance with higher RMSE values at lower SNR. Overall, CNN-CBAM, the fourth-order and SBL stand out as the most effective methods, offering high accuracy and reliability in the DOA estimation.

Underwater environments are characterized by complex noise conditions, including not only Gaussian noise but also other types of noise such as impulse noise, ambient noise, and biological noise. To evaluate the adaptability of the CNN-CBAM model to such environments, we conducted experiments by adding impulse noise to the data at a SNR of 0 dB, with impulse noise ratios ranging from 0.05 to 0.25. We compared the RMSE of six DOA estimation methods. Figure 10 illustrates the RMSE performance of six DOA estimation methods, the RMSE of all methods generally rises, with MUSIC and Capon showing the most significant degradation in performance. In contrast, CNN-CBAM performs well under low impulse

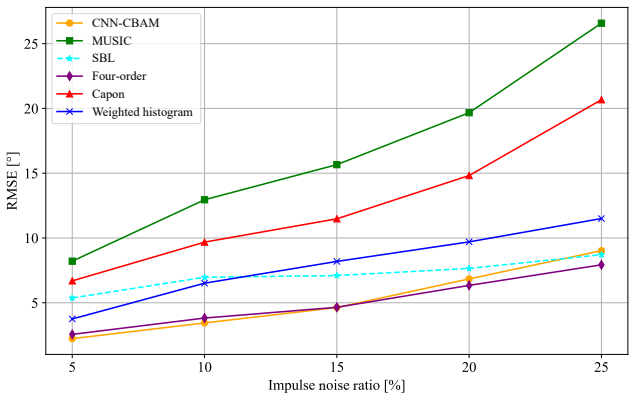

Fig. 10. RMSE of various methods under varying impulse noise ratios.

noise ratios (5 % and 10 %), achieving lower RMSE values, however, as the noise ratio increases, its performance degrades rapidly, with RMSE rising significantly, highlighting its sensitivity to higher levels of impulse noise, the fourth-order cumulant demonstrates greater resilience, with a slower increase in RMSE. The results suggest that CNN-CBAM, despite its advanced architecture, may require enhancements such as noise suppression techniques or hybrid approaches to improve its performance in environments dominated by impulse noise. Future work could focus on integrating traditional signal processing methods with deep learning models to achieve better adaptability to complex underwater noise conditions.

Table 2 presents the computation time of six DOA estimation methods for single and multiple (10) targets, MUSIC, Capon, weighted histogram, and fourth-order methods use a spectral peak search step size of 1°, while SBL employs a grid step size of 1°, revealing that CNN-CBAM, despite its slightly longer computation time (0.125 s) for single-target estimation compared to Capon (0.136 s) and weighted histogram (0.031 s), demonstrates superior scalability and efficiency for multiple targets, requiring only 0.228 s for 10 targets. This advantage stems from its parallel processing capability, attention mechanism, and optimized framework, which minimize computational overhead in complex scenarios. In contrast, methods like the fourth-order and SBL exhibit significantly longer computation times (6.722 s and 13.279 s, respectively) for multiple targets, making them less practical. Thus, CNN-CBAM emerges as an efficient choice for the real-time DOA estimation, particularly in applications involving continuous estimation.

Table 2. Comparison of methods in processing time.

Method	Time for single target [s]	Time for 10 targets [s]
CNN-CBAM	0.125	0.228
MUSIC	0.226	3.306
Capon	0.136	1.437
Weighted histogram	0.031	0.359
Fourth-order	0.583	6.722
SBL	1.247	13.279

5. Discussions

The proposed CNN-CBAM model represents a significant advancement in the DOA estimation for single vector hydrophones, particularly in complex underwater acoustic environments. By integrating the CBAM into a CNN, the model achieves superior noise resistance and estimation accuracy across a wide range of SNRs. This innovative approach addresses the limitations of traditional methods such as MUSIC and Capon, which often struggle with non-stationary noise and multipath effects. The model's ability to adaptively focus on critical features through channel and spatial attention mechanisms establishes it as a robust solution for real-time underwater target localization.

However, several challenges remain to be addressed. A notable limitation is the model's performance in multi-source environments or scenarios with overlapping signal sources. While CNNs excel in one-to-one mapping tasks, their performance deteriorates when handling multiple concurrent sources. This degradation is primarily attributed to the inherent complexity of disentangling overlapping signals, which demands more sophisticated feature extraction and separation techniques. Future work should prioritize enhancing the model's capability to handle multi-source scenarios, potentially through the integration of advanced signal separation algorithms or hybrid architectures that combine CNNs with other machine learning paradigms.

The computational efficiency of the CNN-CBAM model is another critical consideration. As demonstrated in Table 2, the model exhibits competitive processing times for single-target estimation and demonstrates superior scalability for continuous estimation. This efficiency is largely due to the parallel processing capabilities of CNNs and the optimized attention mechanisms of CBAM. Nevertheless, computational requirements may escalate significantly in multi-source environments, necessitating further optimization of the network architecture and training process. Future research should explore techniques such as model pruning, quantization, and distributed computing to enhance scalability and reduce computational overhead.

Integrating the CNN-CBAM model into existing underwater acoustic systems presents additional challenges. A key issue is compatibility with legacy hardware and software, which may require substantial modifications to accommodate the deep learning framework. Moreover, the model's reliance on large datasets for training poses logistical challenges in data collection and preprocessing. To address these issues, future work should focus on developing modular and adaptable frameworks that can be seamlessly integrated into existing systems, as well as exploring transfer learning techniques to reduce dependency on extensive training data.

The CNN output, while not always precise, is 'precisely wrong' in the sense that it consistently deviates from the true values in a predictable manner. This systematic bias, particularly evident in low SNR conditions, underscores the need for robust error correction strategies. Future research should investigate methods to mitigate this bias, such as incorporating probabilistic frameworks or ensemble learning techniques, to improve the model's reliability and accuracy. By addressing these challenges and advancing the proposed ap-

proach, the CNN-CBAM model has the potential to significantly enhance the state of the art in the DOA estimation, providing a robust and efficient solution for underwater acoustic target localization in real-world applications.

Future research directions should focus on advancing the multi-source DOA estimation through the integration of signal separation techniques or hybrid architectures, enhancing the model's capability in complex environments. Systematic biases in the model's output, particularly under low SNR conditions, must be addressed through robust error correction strategies to ensure reliable and accurate estimations. Computational efficiency and scalability can be further optimized via techniques such as model pruning and distributed computing, enabling real-time applications. To facilitate seamless integration into existing underwater acoustic systems, modular frameworks should be developed, overcoming compatibility and logistical challenges. Additionally, leveraging transfer learning techniques can reduce dependency on extensive training datasets while improving adaptability to diverse operational scenarios. Furthermore, real-world experiments will be conducted to validate the method's effectiveness in practical underwater environments, ensuring its robustness and applicability in real-world scenarios. By addressing these critical areas, the CNN-CBAM model is poised to significantly advance the state of the art in the DOA estimation, offering a robust and efficient solution for underwater acoustic target localization in real-world applications.

6. Conclusion

This study proposes a CNN-CBAM-based approach for the DOA estimation using a single vector hydrophone, enhancing accuracy in complex underwater environments. By integrating the CBAM with a CNN, the model processes normalized covariance matrices to focus on critical channels and spatial features. Experimental results demonstrate robustness across varying SNRs, with azimuth and elevation errors within 5° at higher SNRs (15 dB, 10 dB, 5 dB) and within 15° in azimuth and 10° in pitch at lower SNRs (0 dB, −5 dB).

The CNN-CBAM model outperforms traditional methods such as MUSIC and Capon in precision and noise resistance, addressing limitations of eigenvalue decomposition-based methods in non-stationary noise and multipath environments. Challenges remain in multi-source environments, where overlapping signals degrade performance. Future work will focus on enhancing the multi-source DOA estimation, optimizing computational efficiency, leveraging transfer learning for practical deployment, and conducting real-world experiments to validate the method's effectiveness. These advancements will solidify the CNN-CBAM model as a robust solution for real-time underwater target localization.

Fundings

This work was supported by the Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) (grant no. SML2022SP201).

Conflict of interest

The authors declare no conflict of interest.

References

1. Chi J., Li X., Wang H., Gao D., Gerstoft P. (2019), Sound source ranging using a feed-forward neural network trained with fitting-based early stopping, *The Journal of the Acoustical Society of America*, **146**(3): EL258–EL264, https://doi.org/10.1121/1.5126115.

2. Choi J., Choo Y., Lee K. (2019), Acoustic classification of surface and underwater vessels in the ocean using supervised machine learning, *Sensors*, **19**(16): 3492, https://doi.org/10.3390/s19163492.

3. Guo Y., Han J., Wang C. (2018), Multi-acoustic source localization algorithm based on fourth-order moments for single vector hydrophone [in Chinese], *Journal of Sichuan University Natural Science Edition*, **55**: 733.

4. Liu B., Wang Z., Zhang J., Wu J., Qu G. (2024), DeepSIM: A novel deep learning method for graph similarity computation, *Soft Computing*, **28**: 61–76, https://doi.org/10.1007/s00500-023-09288-1.

5. Liu Y., Chen H., Wang B. (2021), DOA estimation based on CNN for underwater acoustic array, *Applied Acoustics*, **172**: 107594, https://doi.org/10.1016/j.apacoust.2020.107594.

6. Liang G., Shi Z., Qiu L., Sun S., Lan T. (2021), Sparse Bayesian learning based direction-of-arrival estimation under spatially colored noise using acoustic hydrophone arrays, *Journal of Marine Science and Engineering*, **9**(2): 127, https://doi.org/10.3390/jmse9020127.

7. Niu H., Ozanich E., Gerstoft P. (2017a), Ship localization in Santa Barbara Channel using machine learning classifiers, *The Journal of the Acoustical Society of America*, **142**(5): EL455–EL460, https://doi.org/10.1121/1.5010064.

8. Niu H., Reeves E., Gerstoft P. (2017b), Source localization in an ocean waveguide using supervised machine learning, *The Journal of the Acoustical Society of America*, **142**(3): 1176–1188, https://doi.org/10.1121/1.5000165.

9. Ozanich E., Gerstoft P., Niu H. (2020), A feedforward neural network for direction-of-arrival estimation, *The Journal of the Acoustical Society of Amer-*

ica, **147**(3): 2035–2048, https://doi.org/10.1121/10.0000944.

10. Tichavsky P., Wong K.T., Zoltowski M.D. (2001), Near-field/far-field azimuth and elevation angle estimation using a single vector hydrophone, *IEEE Transactions on Signal Processing*, **49**(11): 2498–2510, https://doi.org/10.1109/78.960397.

11. Varanasi V., Gupta H., Hegde R.M. (2020), A deep learning framework for robust DOA estimation using spherical harmonic decomposition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **28**: 1248–1259, https://doi.org/10.1109/TASLP.2020.2984852.

12. Wajid M., Kumar A., Bahl R. (2020), Direction estimation and tracking of coherent sources using a single acoustic vector sensor, *Archives of Acoustics*, **45**(2): 209–219, https://doi.org/10.24425/aoa.2020.132495.

13. Wajid M., Kumar A., Bahl R. (2022), Microphone-based acoustic vector sensor for direction finding with bias removal, *Archives of Acoustics*, **47**(2): 151–167, https://doi.org/10.24425/aoa.2022.141646.

14. Woo S., Park J., Lee J.Y. (2018), CBAM: Convolutional block attention module, [in:] *Computer Vision – ECCV 2018*, Ferrari V., Hebert M., Sminchisescu C., Weiss Y. [Eds], Cham: Springer, pp. 3–19, https://doi.org/10.1007/978-3-030-01234-2_1.

15. Xiao P., Liao B., Deligiannis N. (2020), DeepFPC: A deep unfolded network for sparse signal recovery from 1-Bit measurements with application to DOA estimation, *Signal Processing*, **176**: 107699, https://doi.org/10.1016/j.sigpro.2020.107699.

16. Xu L., Chen L., Li Y., Jiang W. (2022), A block sparse-based dynamic compressed sensing channel estimator for underwater acoustic communication, *Journal of Marine Science and Engineering*, **10**(4): 536, https://doi.org/10.3390/jmse10040536.

17. Xu L., Ma Y., Yang Z., Gao T. (2019), Tracking of underwater maneuvering target via M-SIMMUKF algorithm, [in:] *Proceedings of the 6th International Conference on Information Science and Control Engineering (ICISCE)*, pp. 630–634, https://doi.org/10.1109/ICISCE48695.2019.00131.

18. Yao Y., Lei H., He W. (2020), A-CRNN-based method for coherent DOA estimation with unknown source number, *Sensors*, **20**(8): 2296, https://doi.org/10.3390/s20082296.