Deep Covariance Denoising and Attention-Guided On-Grid Classification: A Dual-Stage Network for Robust Underwater DOA Estimation

Bin Zhang*, Jiawen He*, Peishun Liu*, Wenxu Wang*, Ruichun Tang*, and Liang Wang†‡
*Dept. of Computer Science and Technology, Ocean University of China, Qingdao, China
†Dept. of Marine Technology, Ocean University of China, Qingdao, China
Emails: {zhangbin9145}@stu.ouc.edu.cn; {hejiawen, liups, wangwenxu, tangruichun, wanglianger}@ouc.edu.cn

***Abstract*—The direction of arrival (DOA) estimation is a pivotal task in the domain of underwater acoustic signal processing. Some traditional existing methods and recently emerged deep learning algorithms still face issues such as sensitivity to low signal-to-noise ratio (SNR) and high model redundancy. Moreover, the models are often regarded as black boxes, lacking interpretability. To address the aforementioned challenges, this paper utilizes a uniform sonar linear array, employing the array covariance matrix as dual-channel input features. Under different SNR conditions, a denoising autoencoder (DAE) is designed as a pre-filter to reconstruct spatial spectrum features, enhancing the decoupling capability of SNR features. Heatmaps are applied to visualize and elucidate the reconstruction alignment process, enhancing model feature interpretability. Furthermore, a convolutional neural network with a channel attention mechanism is constructed as a post processing classification model to achieve underwater acoustic DOA estimation. The proposed dual-stage algorithmic framework demonstrates superior accuracy with reduced computational overhead across diverse signal scenarios. Evaluated using simulated signals (model size: 1.59 MB), it achieves peak accuracies of 94.95% with an Root Mean Square Error (RMSE) of 1.603 (frequency: 200 Hz, array elements: 10). Under real-world operating conditions with field measurement data (model size: 2.05 MB), the model attains 99.99% accuracy with an RMSE of 0.121, confirming its practical efficacy. We still conducted extensive sensitivity analyses, establishing that this performance—whether in terms of speed, size, or accuracy—is superior to that of traditional DOA estimation algorithms and other deep learning models, with an emphasis on interpretability.**

***Index Terms*—Dual-stage Neural Network, DOA estimation, Spatial Covariance Reconstruction, Adaptive Channel-attention Selector, Heatmap Interpretability**

I. Introduction

Array signal processing plays a crucial role in spatial sensing by enabling the extraction of directional information from multi-channel observations. As a fundamental task in this field, DOA estimation utilizes the propagation characteristics of acoustic waves across sensor arrays to determine the spatial location of sound sources. This capability is particularly important in underwater acoustic environments, where DOA estimation has been extensively employed in applications such as underwater target localization [1] and acoustic communication systems [2].

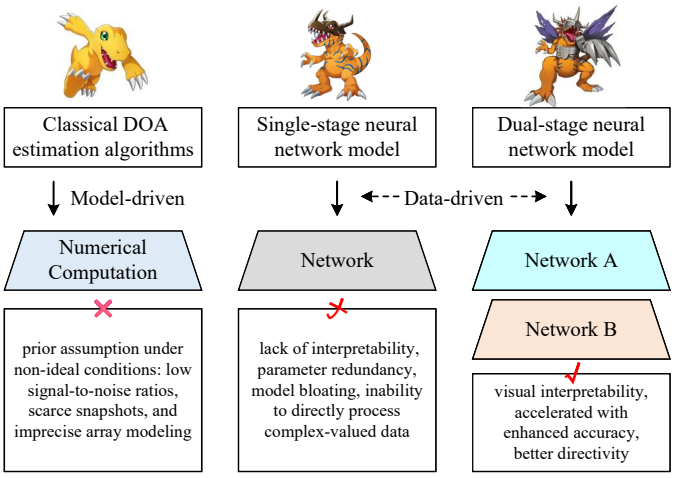

Fig. 1. Compared with traditional DOA-estimation algorithms, existing mature single-stage neural networks—thanks to their powerful nonlinear fitting capability—greatly reduce reliance on prior knowledge, whereas dual-stage architectures divide labor more explicitly and make decisions in a finer-grained manner.

Several classical DOA estimation algorithms have been widely investigated, such as conventional beamforming (CBF) [3], minimum variance distortionless response (MVDR) [4], and subspace-based methods including multiple signal classification (MUSIC) [5], ESPRIT [6], and ROOT-MUSIC [7]. CBF scans the spatial spectrum via fixed steering vectors, whereas MVDR improves resolution by minimizing output power subject to a distortionless constraint. Subspace methods exploit the orthogonality between signal and noise subspaces of the covariance matrix to achieve high-resolution estimation: MUSIC uses spectral peak search, ESPRIT employs rotational invariance to avoid scanning, and ROOT-MUSIC transforms the search into a polynomial rooting problem. Although effective under ideal assumptions—high SNR, sufficient snapshots, and accurate array modeling—these methods often degrade in

‡Corresponding author. This work was supported by the National Natural Science Foundation of China under Grants 12474444. Code is available at https://github.com/binzhangbin/LFD-ACSNet.

practical underwater acoustic environments. Factors such as signal attenuation [8], severe multipath [9], and ambient noise [10] frequently violate these assumptions, leading to notable performance loss, especially at low SNRs.

Recent deep learning methods for DOA estimation reduce the need for prior knowledge and show strong generalization across various environments. They achieve robust performance in low-SNR conditions using advanced architectures such as convolutional networks and attention mechanisms [11]–[14]. However, these methods often have high computational complexity, limited interpretability, and face challenges in real-time deployment. In low-SNR scenarios, non-orthogonality between signal and noise subspaces causes severe feature entanglement and structural distortion, degrading the reliability of learned representations. Therefore, effective representation learning strategies are needed to decouple informative features from noise and preserve the intrinsic data structure. Unsupervised representation learning models, especially autoencoders and their variants [15], offer a promising solution. By learning compact latent representations through encoder–decoder architectures, they enable accurate data reconstruction and robust feature extraction for downstream tasks [16], [17]. Owing to their theoretical grounding and architectural flexibility, autoencoders provide an efficient and interpretable framework for feature compression and reconstruction without explicit supervision.

Motivated by the challenges of feature entanglement, structural distortion, and limited interpretability in low-SNR DOA estimation, we propose a lightweight dual-stage deep learning framework for robust and real-time DOA estimation across varying SNR conditions. Unlike mainstream single-stage networks that deepen blocks—increasing parameters and computations—our design first decouples and reconstructs features from the array covariance matrix, then performs attention-weighted classification. By explicitly addressing representation decoupling and reconstruction, the proposed method mitigates noise-induced feature degradation while maintaining low computational complexity. Detailed differences are illustrated in Fig. 1.

The principal contributions of this work can be summarized as follows: (1) A principled decomposition of the complex-valued covariance matrix into a dual-channel real-valued representation is proposed, enabling deep learning models to efficiently process complex acoustic data. (2) A novel lightweight dual-stage feature-decoupled classification network, termed LFD-ACSNet, is designed by integrating a DAE with a channel attention mechanism to achieve efficient and robust DOA estimation. (3) The proposed framework adopts a two-stage architecture, where the upstream Feature-Decoupling (FD) stage employs a CNN-based DAE to reconstruct the decoupled signal covariance matrix, and the downstream Adaptive Channel Selector (ACS) stage utilizes a channel-attention-weighted CNN (CACNN) to perform high-accuracy DOA estimation. (4) Extensive evaluations on simulated datasets are conducted to validate the robustness and effectiveness of the proposed LFD-ACSNet, and for the first time, heat-map-based visualizations are employed to provide an interpretable analysis of the learned feature representations in this field.

II. Related Work

A. Unsupervised Representation Learning

In recent years, autoencoders (AEs) and their variants have been increasingly adopted in array signal processing for noise suppression and latent feature extraction. Recent research has shifted from time–frequency compression to modeling the covariance structure of array observations. For example, Papageorgiou et al. [18] employed a DAE to recover reliable Hermitian covariance matrices from highly corrupted samples, reducing the performance degradation of covariance-based DOA estimators in low-SNR regimes. Chen et al. [19] combined a DAE with a deep neural network (DNN), where the DAE enhances the noisy received signals and the DNN maps the denoised features to a fine angular grid, achieving robust DOA estimation under low SNR and array imperfections. Ali et al. [20] extended AE-based covariance reconstruction to underwater acoustics using a convolutional autoencoder (CAE) to suppress noise and multipath effects. For sparse arrays, Guo et al. [21] introduced a deep convolutional autoencoder (DCAE) to reconstruct the sparse-array covariance matrix into its uniform linear array equivalent, increasing degrees of freedom, followed by conventional estimators such as MUSIC. They also proposed a DCNN-based direct-mapping method for high-accuracy DOA estimation. Despite these advances, most existing AE-based approaches tightly couple feature enhancement with estimation, limiting interpretability and task-specific flexibility.

B. Attention-Based Feature Learning

Attention mechanisms, which have proven effective in deep-learning-based image processing, have been increasingly applied to array signal processing and acoustic perception by adaptively highlighting informative features. Recently, several studies have introduced attention into deep learning-based DOA estimation. Mack et al. [22] designed a signal-aware wideband DOA network that uses attention to weight frequency sub-bands and array elements, improving accuracy and robustness under low SNR, uneven spectra, and multi-source interference. Liu et al. [23] proposed an attention-based network that focuses on array elements and feature channels less affected by noise, enhancing performance under unknown or non-uniform noise distributions. Wu et al. [24] developed a gridless DOA method with a Residual Attention Network and transfer learning, reducing off-grid errors and improving generalization across array geometries and SNRs. In another work, Wu et al. [25] integrated beamforming outputs with an attention-augmented ResNet, leveraging spatial priors to boost angular resolution and estimation precision in multi-source low-SNR scenarios. Overall, attention allows DOA models to selectively emphasize discriminative spatial and spectral features, thereby strengthening robustness and estimation accuracy. However, existing attention-based approaches are mostly

embedded within single-stage networks and lack systematic integration with structured feature-decoupling mechanisms.

C. Two-Stage Deep Learning Models

To improve robustness and interpretability, recent studies have increasingly adopted two-stage or multi-stage frameworks that decouple representation learning from task-specific inference. Cai et al. [26] introduced a two-stage deep CNN that learns feature extraction and direction regression separately, significantly enhancing DOA accuracy under strong impulsive noise. Zhang et al. [27] developed a two-stage MLP for high-resolution DOA estimation, extracting coarse and fine features in separate stages to improve resolving capability for closely spaced sources. Wang et al. [28] combined spatial whitening with normalized MUSIC in a two-stage framework to suppress interference and improve angular accuracy for GNSS signals under jamming. Zhang et al. [29] proposed a hierarchical sparse Bayesian two-stage method, using coarse-grid initialization followed by fine-grid refinement, enhancing resolution under low-SNR and overlapping source conditions. Despite these advances, most two-stage approaches focus on performance gains, with limited exploration of stage-specific roles and internal representation interpretability, restricting practical reliability.

III. Proposed Method

A. HLA Signal Model

Consider a horizontal linear array (HLA) of M sensors with inter-element spacing d. Under the narrow-band assumption, the complex base-band snapshot vector $\mathbf{X} \in \mathbb{C}^{M\times P}$ received from K far-field sources is modeled as

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{N}, \tag{1}$$

where $\mathbf{A} = [\mathbf{a}(\theta_1), \ldots, \mathbf{a}(\theta_K)] \in \mathbb{C}^{M\times K}$ is the array manifold matrix with

$$\mathbf{a}(\theta_k) = \left[1,\, e^{-j2\pi d\cos\theta_k/\lambda},\, \ldots,\, e^{-j2\pi(M-1)d\cos\theta_k/\lambda}\right]^{\mathrm{T}}, \tag{2}$$

$\lambda = c/f$ is the wavelength, $\mathbf{S}(t) = \exp(j2\pi ft)$, $\mathbf{S} \in \mathbb{C}^{K\times P}$ contains the P snapshots of the source signals, and $\mathbf{N} \in \mathbb{C}^{M\times P}$ is zero-mean spatially and temporally white Gaussian noise. The sample covariance matrix of the snapshots is computed as

$$\hat{\mathbf{R}} = \frac{1}{P}\mathbf{X}\mathbf{X}^{\mathrm{H}}. \tag{3}$$

Widely employed by CBF, MVDR, MUSIC, etc., Equation 3 serves as the standard signal model for DOA estimation. For brevity, we write $\hat{\boldsymbol{x}} = \hat{\mathbf{R}} \in \mathbb{C}^{M\times M}$ hereafter.

B. LFD-ACSNet Model

1) FD-Stage: Unsupervised Learning-Based DAE: Inspired by the literature [30], we designed a DAE based on a CNN architecture to achieve the reconstruction and alignment process of the input complex covariance matrix $\hat{\boldsymbol{x}}$. Due to the fact that neural networks can only process real-valued data, we decompose the real and imaginary parts of the complex matrix $\hat{\boldsymbol{x}}$ into two separate channels of real-valued matrices as inputs, that is: $\hat{\boldsymbol{x}} = \{\mathrm{Re}\{\hat{\boldsymbol{x}}\}, \mathrm{Im}\{\hat{\boldsymbol{x}}\}\}$, where the dimensionality is changed from $\mathbb{C}^{M\times M}$ to $\mathbb{R}^{2\times M\times M}$.

The detailed architecture and comprehensive structural parameters are elaborated in Fig. 2. The noisy spectral spatial covariance matrix $\hat{\boldsymbol{x}} \in \mathbb{R}^{B\times 2\times M\times M}$, where B denotes the batch size, after passing through the encoder, is compressed into the latent factor feature space $\boldsymbol{z} \in \mathbb{R}^{B\times Q}$:

$$\boldsymbol{z} = f_\delta(\boldsymbol{W}^{(En)}\hat{\boldsymbol{x}} + \boldsymbol{b}^{(En)}). \tag{4}$$

Here, $\boldsymbol{W}^{(En)}$ and $\boldsymbol{b}^{(En)}$ represent the weight matrix and bias terms of the encoder, respectively, and f_δ denotes a series of nonlinear activation transformations.

The decoder primarily performs the inverse operation of the encoder, aiming to denoise and reconstruct the covariance matrix $\boldsymbol{v}$ from the latent factor feature space $\boldsymbol{z}$. The latent factor feature space $\boldsymbol{z}$ with dimension Q is restored to the reconstructed original spectral space $\boldsymbol{v} \in \mathbb{R}^{B\times 2\times M\times M}$ after passing through the decoder:

$$\boldsymbol{v} = h_\zeta(\boldsymbol{W}^{(De)}\boldsymbol{z} + \boldsymbol{b}^{(De)}). \tag{5}$$

Here, $\boldsymbol{W}^{(De)}$ and $\boldsymbol{b}^{(De)}$ represent the weight matrix and bias terms of the decoder, respectively, and h_ζ denotes a series of nonlinear activation inverse transformations.

2) ACS-Stage: Supervised Learning-Based CACNN: The ACS-stage CACNN employs convolutional layers [31] and channel attention modules [32], using small kernels of size 3 to preserve spatial details and enhance model discriminability and generalization while reducing parameters. These structures, termed Channel-wise Weighted Particle blocks, are stacked four times for progressive feature extraction and dynamic reweighting.

The decoder output $\boldsymbol{v} \in \mathbb{R}^{B\times 2\times M\times M}$ is processed by convolutional layers into $\boldsymbol{g} \in \mathbb{R}^{B\times C\times H\times H}$, where C denotes the number of convolutional channels in different layers, and H represents the feature dimension. Channel attention then generates weights $\boldsymbol{w} \in \mathbb{R}^{B\times C\times 1\times 1}$ via pooling and convolution, producing $\boldsymbol{o} = \boldsymbol{g} \otimes \boldsymbol{w}$ to emphasize informative channels. Here, '$\otimes$' means element-wise multiplication. This operates similarly to residual connections [33]. Four Channel-wise Weighted Particle blocks are stacked to progressively extract and filter spectral-semantic features, thereby enhancing the capacity of the model to dynamically concentrate on principal components within acoustic spatial spectra and further suppress interference from irrelevant background noise subspaces. After four blocks, features $\boldsymbol{u} \in \mathbb{R}^{B\times T\times L\times L}$ are flattened and classified via two fully connected layers and ReLU. The output vector $\hat{\boldsymbol{y}}$ yields the DOA estimate $\hat{\theta}$ as the argmax index.

3) Loss Function: The loss function of the proposed method comprises two components: reconstruction error minimization and class probability distribution alignment. The reconstruction loss $L_r(\boldsymbol{x}, \boldsymbol{v})$, based on mean squared error, facilitates covariance matrix recovery in spectral space under low SNR conditions. The distribution loss $L_d(\boldsymbol{y}, \hat{\boldsymbol{y}})$ employs binary cross-entropy to minimize the discrepancy between the

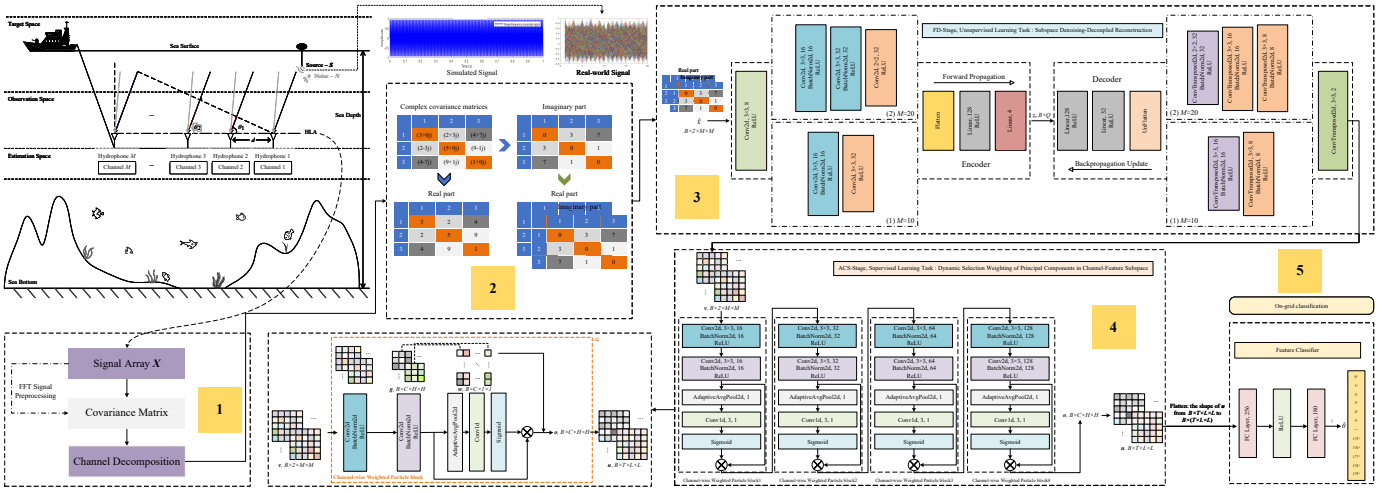

Fig. 2. Architecture of the LFD-ACSNet.

predicted and true azimuth angle probability distributions. The loss functions for each stage are defined as follows:

$$L_r(\boldsymbol{x}, \boldsymbol{v}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{v}_i)^2, \tag{6}$$

$$L_d(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\frac{1}{n}\sum_{i=1}^{n}[\boldsymbol{y}_i \log \hat{\boldsymbol{y}}_i + (1-\boldsymbol{y}_i)\log(1-\hat{\boldsymbol{y}}_i)]. \tag{7}$$

In this context, $\boldsymbol{x}$ denotes the clean spectral space matrix, $\boldsymbol{v}$ represents the spectral recovery matrix after the reconstruction output of the DAE. $\boldsymbol{y}$ denotes the real azimuth angle label vector, and $\hat{\boldsymbol{y}}$ represents the predicted azimuth angle vector.

IV. Experiments

The experiments were conducted on a workstation with an NVIDIA GeForce RTX 4090 GPU. Models were implemented in PyTorch with the following hyperparameters: in the FD-Stage, training used 40 epochs with a batch size of 128, and a reconstruction factor $Q = 4$; the ACS-Stage was trained for 200 epochs with a batch size of 2000 using the Adam optimizer (learning rate 1e-3). All weights were initialized from $N(0, 1)$ and biases to zero. Results are reported using the best model selected based on validation performance.

We evaluate the performance of the model using total parameters, parameter size, average inference time (AIT), accuracy, and RMSE. Among them, AIT indicates the inference time per sample, accuracy represents the percentage of correctly predicted positions out of all test samples, and RMSE represents the root mean square error between the predicted and true DOA, quantifying the average magnitude of angular prediction deviations. Mathematically, these metrics are defined as follows:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(\hat{\theta}_i = \theta_i\right) \times 100\%, \tag{8}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}, \tag{9}$$

where $\hat{\theta}_i$ and θ_i denote the predicted and true azimuth angles (in degrees) for the i-th sample, respectively, n is the total number of test samples, $\mathbb{I}(\cdot)$ represents the indicator function that equals 1 when the condition is satisfied and 0 otherwise.

A. Simulation Experiment Results

Simulated underwater acoustic data for a single target ($K = 1$) are generated using a 10-element ($M = 10$) array. The source frequency $f = 200$ Hz, sound speed $c = 1500$ m/s, element spacing $d = \lambda/2$, and sampling frequency $f_s = 1500$ Hz (satisfying the Nyquist criterion). The azimuth angle range is $[0°, 180°)$ with $1°$ intervals, yielding 180 classification targets. For each SNR level (-10, 0, 10, 20, 30 dB), 180,540 samples are allocated and split into training, validation, and test sets in a 7:2:1 ratio, resulting in 126,180 training, 36,180 validation, and 18,180 test samples.

1) Efficiency Metric Cross-Architecture Comparative Analysis: An ablation study is conducted to assess the denoising effect of the FD-stage by comparing LFD-ACSNet with its standalone ACS-stage (CACNN). As shown in the first row of Fig. 3, on the simulated dataset under SNR levels from -10 dB to 30 dB, LFD-ACSNet achieves accuracies of 80.73%, 95.63%, 98.73%, 99.67%, and 99.99%, respectively. These results outperform those of CACNN (79.71%, 94.03%, 96.33%, 96.89%, 98.88%) by 1.02%, 1.60%, 2.40%, 2.78%, and 1.11%, yielding an average improvement of 1.782% (a total gain of 8.91% across all SNR conditions). The RMSE of LFD-ACSNet (1.603) is 0.621 lower than that of CACNN (2.224). The consistent superiority of LFD-ACSNet is attributed to the FD-stage encoder–decoder, which reconstructs features and enhances generalization.

The first row of Fig. 3 also compares the results of the proposed LFD-ACSNet with several benchmarks, including the MUSIC algorithm, CNN [11], feedforward neural network (FNN) [34], and single-stage networks (ResNet18, ResNet34, ResNet50, ResNet101 [33], and ResNeXt [35]). Among the aforementioned models, the FNN

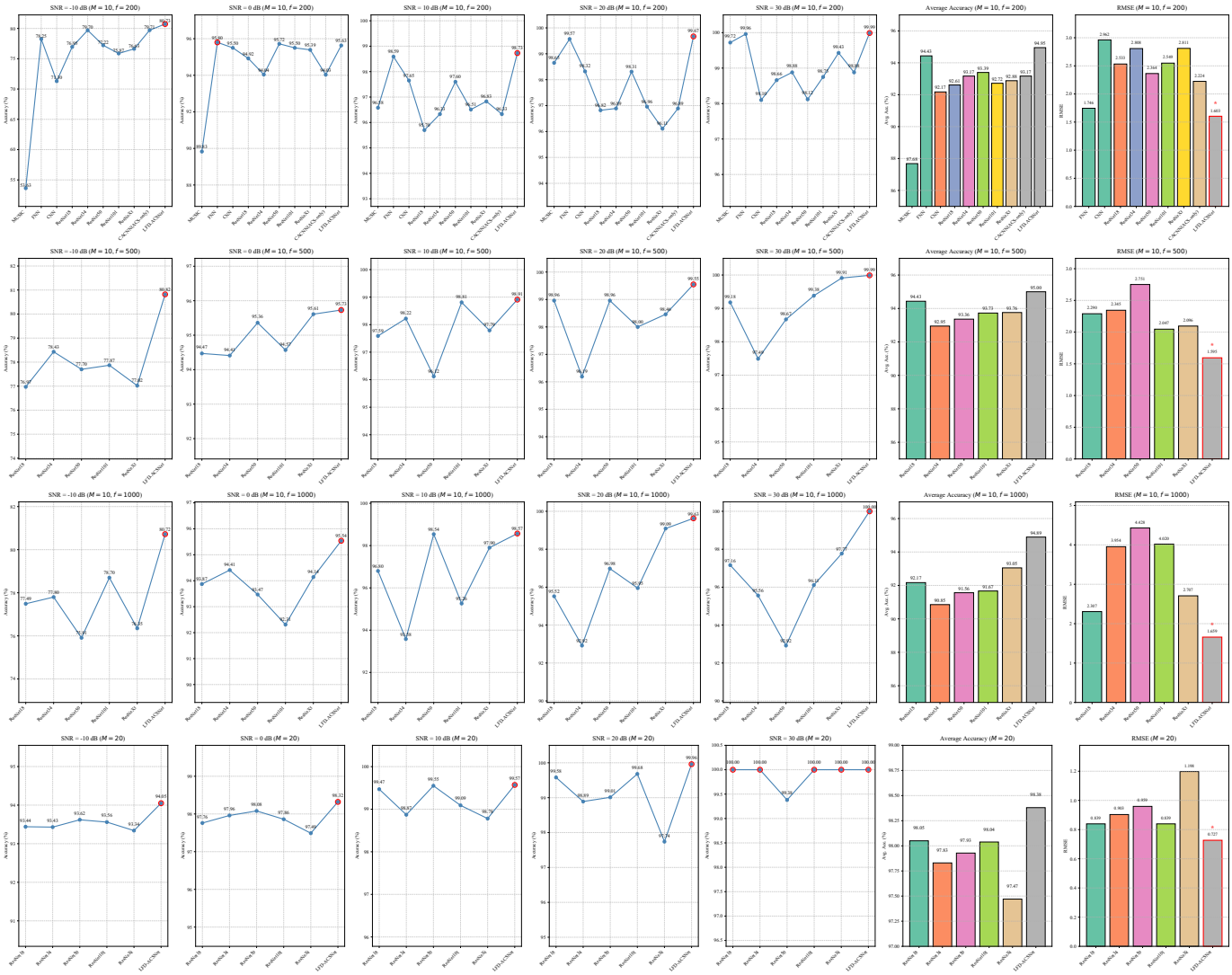

Fig. 3. Prediction accuracy, comprehensive average accuracy, and RMSE results under different SNRs, frequencies, and array aperture lengths. Here, the average accuracy denotes the mean accuracy across different SNR levels. The first row represents the benchmark simulation test results, while the second, third, and fourth rows represent the test results for sound source frequencies of 500 Hz, 1000 Hz, and an aperture length of 20 array elements, respectively.

achieved the highest average accuracy (94.43%), followed by ResNet50 (93.39%), ResNet34 (93.17%), ResNeXt (92.88%), ResNet101 (92.72%), and ResNet18 (92.61%), indicating that deeper networks do not consistently improve accuracy. MUSIC performed poorest overall (87.68%), while CNN attained 90.69%. In contrast, LFD-ACSNet reached an average accuracy of 94.95%, exceeding the second-best FNN by 0.52% and outperforming all other listed methods by 1.56–7.27%. Notably, with the lowest RMSE of 1.603, the proposed model exhibits minimal directional prediction deviation compared to other methods. Even at -10 dB, where noise disrupts the covariance matrix and degrades performance, LFD-ACSNet maintains a higher accuracy owing to its FD-stage reconstruction and Channel-wise Weighted Particle blocks.

As shown in Table I ($M = 10$), LFD-ACSNet reduces both total parameters and parameter size by approximately 3 to 100 times compared to other mainstream one-stage networks. With only 1.59 MB of parameters, it achieves notably higher efficiency than the benchmark models: FNN (11.81 MB), CNN (4.93 MB), ResNet18 (42.98 MB), ResNet34 (81.53 MB), ResNet50 (91.07 MB), ResNet101 (163.52 MB), and ResNeXt (89.06 MB). Furthermore, LFD-ACSNet attains an AIT of merely 226.11 ms, significantly faster than other deep neural networks and thus better suited for real-time applications. Overall, combined with earlier analysis, LFD-ACSNet demonstrates superior cost-effectiveness and comprehensive performance in terms of both test accuracy and inference speed.

2) Beam Directivity Visualization Analysis: Here, we conduct a directivity visualization analysis comparing the proposed algorithm with classical beamforming methods. While high-resolution algorithms such as CBF, MVDR, MUSIC, ESPRIT, and ROOT-MUSIC perform well under high-SNR conditions, their resolution degrades significantly at low SNR due to noise interference, which distorts signal features and

TABLE I
COMPARISON OF MODEL SPECIFICATIONS AND PERFORMANCE METRICS FOR SIMULATED AND REAL-WORLD ARRAY CONFIGURATIONS.

	Simulated Data ($M = 10$)			Simulated Data ($M = 20$)			Real-world Data ($M = 32$)				
	Total Params	Params Size	AIT	Total Params	Params Size	AIT	Total Params	Params Size	AIT	Accuracy	RMSE
FNN [34]	3096244	11.81 MB	266.41 ms	-	-	-	10582708	40.37 MB	337.95 ms	97.60%	0.813
CNN [11]	1231413	4.93 MB	264.43 ms	-	-	-	-	-	-	-	-
ResNet18 [33]	11265716	42.98 MB	295.74 ms	11265716	42.98 MB	356.35 ms	11265716	42.98 MB	397.57 ms	99.96%	0.113
ResNet34 [33]	21373876	81.53 MB	499.35 ms	21373876	81.53 MB	490.01 ms	21373876	81.53 MB	544.81 ms	99.97%	0.187
ResNet50 [33]	23873716	91.07 MB	580.76 ms	23873716	91.07 MB	538.12 ms	23873716	91.07 MB	618.54 ms	99.96%	0.212
ResNet101 [33]	42865844	163.52 MB	810.78 ms	42865844	163.52 MB	818.50 ms	42865844	163.52 MB	862.09 ms	99.62%	0.193
ResNeXt [35]	23345588	89.06 MB	512.98 ms	23345588	89.06 MB	515.54 ms	23345588	89.06 MB	556.25 ms	98.80%	0.513
LFD-ACSNet	**416588**	**1.59 MB**	**226.11 ms**	**487309**	**1.96 MB**	**242.41 ms**	**510182**	**2.05** MB	**262.48** ms	**99.99%**	**0.121**

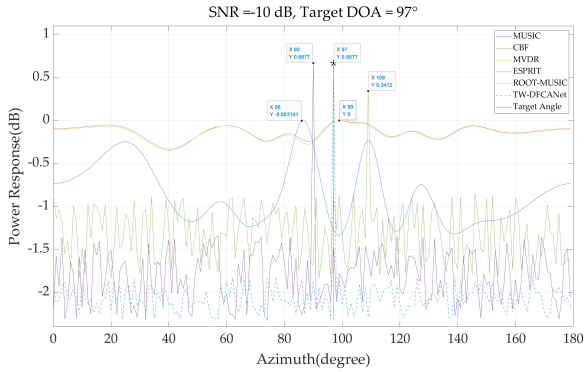

Fig. 4. Beam steering spatial response at 97° azimuth under -10 dB SNR.

limits practical utility. Thus, we evaluate LFD-ACSNet alongside these methods at an SNR of -10 dB. Taking the target source at 97° as an example, Fig. 4 presents the spatial spectrum results. It is observed that under low SNR conditions, the MUSIC algorithm yields numerous spurious peaks, with its highest peak occurring at approximately 86°, corresponding to an estimation error of 11°. Both CBF and MVDR produce relatively broad sidelobes, providing an estimated azimuth of 99° with an error of 2°. The Root-MUSIC and ESPRIT algorithms estimate the azimuth as 109° and 90°, with errors of 12° and 7°, respectively. In comparison, the proposed LFD-ACSNet outperforms all the conventional methods in terms of both angular resolution and estimation accuracy.

Furthermore, heatmaps of the spatial covariance matrix at 97° azimuth under -10 dB SNR are provided for detailed visual analysis, as shown in Fig. 5. Furthermore, we employ heatmaps of the spatial covariance matrix for a more detailed visual analysis, as shown in Fig. 5. As the noise-free reference, Fig. 5(c) exhibits clear diagonal structures (element self-power) in the real part and deterministic phase-difference patterns in the imaginary part. Fig. 5(b) shows severe noise contamination, resulting in an almost uniformly random distribution. At this low SNR, the reconstructed input features $\hat{\boldsymbol{x}}$ shown in Fig. 5(a) are closer to the clean values and show clearer separation, successfully recovering the diagonally dominant structure and phase consistency. This demonstrates the algorithm's effectiveness in suppressing noise and reconstructing signal subspace characteristics. Such reconstruction visualization enhances the interpretability of LFD-ACSNet, addressing the black-box limitation common in existing algorithms.

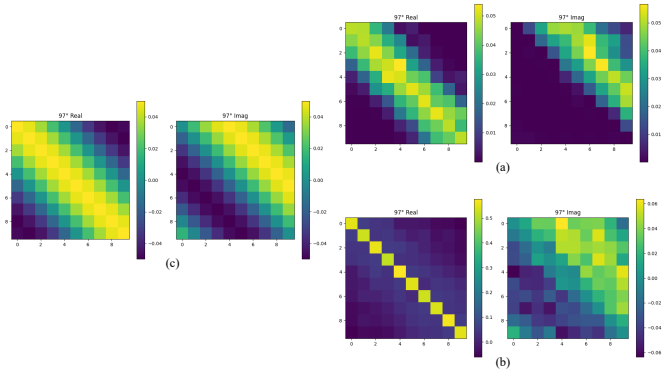

Fig. 5. (a) The reconstruction result. (b) The noisy result. (c) The noise-free spectrum. In (a)-(c), each cell (i, j) denotes the covariance between signals received by the i-th and j-th array elements $(i, j = 0, \ldots, 9)$, with the real and imaginary parts representing the in-phase and quadrature components, respectively; the colorbars indicate the corresponding magnitudes.

B. Sensitivity Analysis Experiment Results

TABLE II
TEST RESULTS WITH DIFFERENT RECONSTRUCTION FACTORS Q.

SNR (dB)	-10	0	10	20	30	Average	RMSE
Q=2	72.93%	95.45%	98.72%	99.65%	99.98%	93.35%	2.1439
Q=4	**80.73%**	**95.63%**	**98.73%**	**99.67%**	**99.99%**	**94.95%**	**1.603**
Q=8	80.36%	95.53%	**98.73%**	**99.67%**	99.98%	94.85%	1.6487

1) Differentiated Reconstruction Factors Analysis: The reconstruction factor Q influences the reconstruction capability of the model. While increasing its dimension can improve the learning capacity and enable the capture of more complex abstract features [36], an excessively high dimension may cause overfitting and degrade generalization. To identify the optimal Q, we evaluate the model performance across different dimensions. As shown in Table II, $Q = 4$ achieves the highest average accuracy of 94.95% across all SNR levels, with an RMSE of 1.603, consistently outperforming other settings. This indicates that $Q = 4$ offers the best balance between reconstruction fidelity and predictive performance.

2) Differentiated Frequency and Array Configuration Analysis: For narrowband signals, the frequency of different sound

sources and the number of array elements also critical influence algorithm performance [37]. Such array-scale vari tions, which are independent of model architecture, may affe different algorithms differently.

We therefore conducted comparative experiments at f 500 Hz ($f_s = 2500$ Hz) and $f = 1000$ Hz ($f_s = 4000$ Hz) evaluate the robustness of the proposed model;the results a shown in the second and third rows of Fig. 3. To avoid bi introduced by network depth, ResNet is used as the baseli in these comparisons.

As seen in the second and third rows of Fig. 3, LFD-ACSN achieves the most noticeable accuracy improvement und low-SNR conditions (-10 dB), outperforming the second-be model by 2.39% at 500 Hz and by 2.02% at 1000 H Strikingly, the advantage of LFD-ACSNet increases monotor cally with SNR, demonstrating strong generalization. Overa the proposed model achieves average accuracies of 95.00 (RMSE: 1.595) and 94.89% (RMSE: 1.659) at the two fr quencies, consistently surpassing other methods. These resul confirm the model's stable performance across different source frequencies.

We further evaluated the models with a larger array configuration ($M = 20$). As illustrated in the last row of Fig. 3, a larger array aperture improves the prediction accuracy of all models, especially under low SNR. LFD-ACSNet reaches an average accuracy of 98.38% (RMSE: 0.727), clearly exceeding the second-best result of 98.05% (RMSE: 0.839). Table I ($M = 20$) summarizes the computational efficiency and complexity metrics. Despite the increased input dimension, LFD-ACSNet maintains the best accuracy-efficiency trade-off, achieving the smallest model size (1.96 MB) and the fastest inference time (242.41 ms).

C. Real-world Measured Experiment Results

The actual measurement was conducted from 10:35 AM to 12:20 PM on November 4, 2020, in Jiaozhou Bay, Qingdao, China. A HLA with $M = 32$ elements and an inter-element spacing of 1 m was used as the receiving array. The array was deployed at a depth of approximately 17 ± 3 m underwater, where elements exhibited vertical fluctuations due to sea tides, with adjacent element fluctuations ranging from 0 m to 1.29 m. Four target sound sources were present, each with a bandwidth of 500 Hz to 650 Hz and an azimuth angle of $45°$. The sampling frequency was 6000 Hz, and the in-situ sound speed was 1466 m/s. The available field-measured data used in this experiment comprise frequency-domain spatial spectra obtained through FFT, totaling 242 samples.

1) Progressive Transfer Learning Strategy: To address distribution shift, input dimension discrepancies, and real-data scarcity, this paper proposes a progressive transfer learning framework [38]. First, quasi-measured data aligned with real-world distribution are generated based on acoustic environmental characteristics, eliminating simulation-modality discrepancies while maintaining the same volume and partitioning as simulated data. Subsequently, a model pre-trained on simulated data initializes network parameters to capture prior

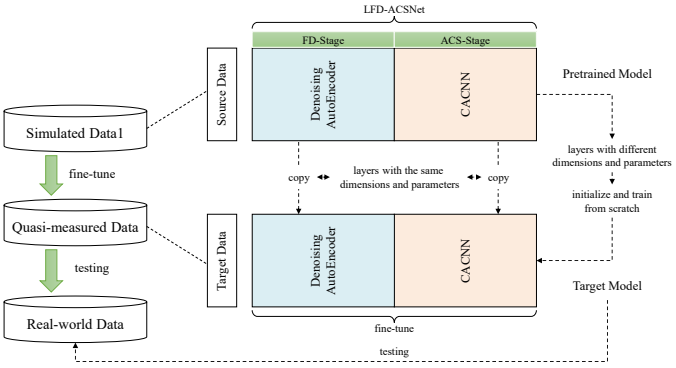

Fig. 6. Flowchart of the progressive transfer learning strategy based on fine-tuning.

knowledge of fundamental physical patterns. Fine-tuning for 10 epochs on quasi-measured data corrects distribution bias and enables cross-domain knowledge transfer. Finally, the framework is evaluated on an independent real-world dataset. The overall pipeline is illustrated in Fig. 6.

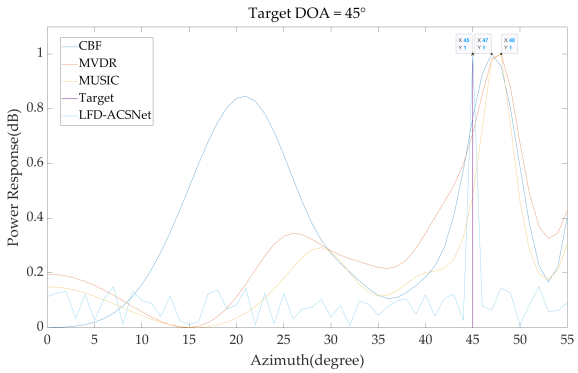

Fig. 7. Comparative analysis results of beam patterns.

2) Comprehensive Performance Verification Analysis: Following the aforementioned learning strategy, a detailed analysis of the azimuthal directivity is conducted. The temporal evolution of the test results at 1-second intervals is illustrated in the subfigure of Fig. 7, allowing for precise quantification of directional deviation patterns. Owing to environmental interference and fluctuations in the actual SNR, the main lobe directions of the MVDR, MUSIC, and CBF algorithms exhibit an average deviation of approximately 2°–3°. In comparison with these conventional algorithms, LFD-ACSNet demonstrates significantly stronger robustness in localization performance.

Comparative results with state-of-the-art one-stage deep learning methods are summarized in Table I ($M = 32$). LFD-ACSNet achieves the highest accuracy of 99.99% (RMSE: 0.121), outperforming all counterparts. Remarkably, it attains this performance with only 2.05 MB parameters, which is 19.7–79.8× smaller than other models (40.37–163.52 MB). In terms of inference speed, LFD-ACSNet reduces the AIT by 22.4% compared to FNN and is 1.5–3.3× faster than ResNet

variants (337.95–862.09 ms). These results validate that LFD-ACSNet optimally balances accuracy, efficiency, and cross-domain generalization.

V. Conclusion

This paper proposes LFD-ACSNet, a novel deep learning model for DOA estimation based on a dual-stage strategy. It outperforms conventional algorithms such as MUSIC, CBF, MVDR, ESPRIT, and ROOT-MUSIC in precision and directional accuracy, while maintaining favorable cost-effectiveness. Additionally, the model offers enhanced interpretability through feature heatmaps, alleviating the black-box limitations common in neural networks. Experimental results validate the efficiency of the model, speed, accuracy, and architectural rationale.

References

[1] J. Li, F. Tong, Y. Zhou, Y. Yang, and Z. Hu, "Small size array underwater acoustic doa estimation based on direction-dependent transmission response," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 12, pp. 12 916–12 927, 2022.

[2] T. Li, F. Zhou, L. Ma, X. Liu, and M. Muzzammil, "Multipath doa based mimo beamforming receiver scheme for high-rate underwater acoustic communications," *Applied Acoustics*, vol. 221, p. 109994, 2024.

[3] J. Benesty, J. Chen, and Y. Huang, *Conventional Beamforming Techniques*. Springer Berlin Heidelberg, 2008, pp. 39–65.

[4] M. Wolfel and J. McDonough, "Minimum variance distortionless response spectral estimation," *IEEE Signal Process Mag.*, vol. 22, no. 5, pp. 117–126, 2005.

[5] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[6] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, 1989.

[7] B. D. Rao and K. V. S. Hari, "Performance analysis of root-music," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 12, pp. 1939–1949, 1989.

[8] M. Jia, Y. Wu, C. Bao, and C. Ritz, "Multi-source doa estimation in reverberant environments by jointing detection and modeling of time-frequency points," *IEEE-ACM T. AUDIO. SPE.*, vol. 29, pp. 379–392, 2021.

[9] X. Han, M. Liu, S. Zhang, R. Zheng, and J. Lan, "A passive doa estimation algorithm of underwater multipath signals via spatial time-frequency distributions," *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3439–3455, 2021.

[10] A. Liu, S. Shi, and X. Wang, "Robust doa estimation method for underwater acoustic vector sensor array in presence of ambient noise," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023.

[11] Y. Liu, H. Chen, and B. Wang, "Doa estimation based on cnn for underwater acoustic array," *Appl. Acoust.*, vol. 172, p. 107594, 2021.

[12] W. Nie, X. Zhang, J. Xu, L. Guo, and Y. Yan, "Adaptive direction-of-arrival estimation using deep neural network in marine acoustic environment," *IEEE Sens. J.*, vol. 23, no. 13, pp. 15 093–15 105, 2023.

[13] G. K. Papageorgiou, M. Sellathurai, and Y. C. Eldar, "Deep networks for direction-of-arrival estimation in low snr," *IEEE Trans. Signal Process.*, vol. 69, pp. 3714–3729, 2021.

[14] H. Chu, C. Li, H. Wang, J. Wang, Y. Tai, Y. Zhang, L. Zhou, F. Yang, and Y. Benezeth, "Mtsa-net: A multiscale time self-attention network for ship radiated self-noise reduction," *Ocean Eng.*, vol. 292, p. 116566, 2024.

[15] K. Berahmand, F. Daneshfar, E. S. Salehi, Y. Li, and Y. Xu, "Autoencoders and their applications in machine learning: a survey," *Artif. Intell. Rev.*, vol. 57, no. 2, p. 28, 2024.

[16] D. Jana, J. Patil, S. Herkal, S. Nagarajaiah, and L. Duenas-Osorio, "Cnn and convolutional autoencoder (cae) based real-time sensor fault detection, localization, and correction," *Mech. Syst. Sig. Process.*, vol. 169, p. 108723, 2022.

[17] Q. Wu, T. Yang, Z. Liu, B. Wu, Y. Shan, and A. B. Chan, "Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14 561–14 571.

[18] G. K. Papageorgiou and M. Sellathurai, "Direction-of-arrival estimation in the low-snr regime via a denoising autoencoder," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2020, pp. 1–5.

[19] D. Chen, S. Shi, X. Gu, and B. Shim, "Robust doa estimation using denoising autoencoder and deep neural networks," *IEEE Access*, vol. 10, pp. 52 551–52 564, 2022.

[20] M. Ali, K. Hameed, and K. Nathwani, "Enhancing traditional underwater doa estimation techniques using convolutional autoencoder-based covariance matrix reconstruction," *IEEE Sensors Letters*, 2024.

[21] S. Guo, Q. Zhang, X. Fu, G. Zheng, and H. Zhou, "Doa estimation method for sparse arrays based on deep convolutional autoencoder and deep convolutional neural network," *Digital Signal Processing*, p. 105627, 2025.

[22] W. Mack, U. Bharadwaj, S. Chakrabarty, and E. A. Habets, "Signal-aware broadband doa estimation using attention mechanisms," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4930–4934.

[23] K. Liu, X. Wang, J. Yu, and J. Ma, "Attention based doa estimation in the presence of unknown nonuniform noise," *Applied Acoustics*, vol. 211, p. 109506, 2023.

[24] X. Wu, J. Wang, X. Yang, and F. Tian, "A gridless doa estimation method based on residual attention network and transfer learning," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 6, pp. 9103–9108, 2024.

[25] L. Wu, Y. Fu, X. Yang, L. Xu, S. Chen, Y. Zhang, and J. Zhang, "Research on the multi-signal doa estimation based on resnet with the attention module combined with beamforming (rab-doa)," *Applied Acoustics*, vol. 231, p. 110541, 2025.

[26] R. Cai and Q. Tian, "Two-stage deep convolutional neural networks for doa estimation in impulsive noise," *IEEE Transactions on Antennas and Propagation*, vol. 72, no. 2, pp. 2047–2051, 2023.

[27] Y. Zhang, Y. Huang, J. Tao, S. Tang, H. C. So, and W. Hong, "A two-stage multi-layer perceptron for high-resolution doa estimation," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 7, pp. 9616–9631, 2024.

[28] C. Wang, X. Cui, G. Liu, and M. Lu, "Two-stage spatial whitening and normalized music for robust doa estimation of gnss signals under jamming," *IEEE Transactions on Aerospace and Electronic Systems*, 2024.

[29] X. Zhang, Z. Chen, X. Yu, B. Lin, Z. Ma, J. Zhu, P. Wang, and J. Li, "A novel two-stage doa estimation of sound sources based on hierarchical sparse bayesian inference," *Applied Acoustics*, vol. 231, p. 110514, 2025.

[30] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[31] M. Hussain, J. J. Bird, and D. R. Faria, "A study on cnn transfer learning for image classification," in *Advances in Computational Intelligence Systems: Contributions Presented at the 18th UK Workshop on Computational Intelligence, September 5-7, 2018, Nottingham, UK*. Springer, 2019, pp. 191–202.

[32] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11 531–11 539.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[34] E. Ozanich, P. Gerstoft, and H. Niu, "A feedforward neural network for direction-of-arrival estimation," *J. Acoust. Soc. Am.*, vol. 147, no. 3, pp. 2035–2048, 2020.

[35] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5987–5995.

[36] B. M. Dillon, T. Plehn, C. Sauer, and P. Sorrenson, "Better latent spaces for better autoencoders," *SciPost Phy.*, vol. 11, no. 3, p. 061, 2021.

[37] R. J. Weber and Y. Huang, "Analysis for capon and music doa estimation algorithms," in *2009 IEEE Antennas and Propagation Society International Symposium*, 2009, pp. 1–4.

[38] H. Cao, W. Wang, L. Su, H. Ni, P. Gerstoft, Q. Ren, and L. Ma, "Deep transfer learning for underwater direction of arrival using one vector sensor," *J. Acoust. Soc. Am.*, vol. 149, no. 3, pp. 1699–1711, 2021.