

# A Residual Neural Network Framework for Rotating Sound Source Localization

**Yifeng Yao**

Shanghai Jiao Tong University

**Ce Zhang**

AECC Commercial Aircraft Engine Co., Ltd.

**Cheng Wei Lee**

Shanghai Jiao Tong University

**Wei Ma**

[mawei@sjtu.edu.cn](mailto:mawei@sjtu.edu.cn)

Shanghai Jiao Tong University

---

## Research Article

**Keywords:** rotating source localization, cross-spectral matrix, MCB, ROSI, deep learning, ResNet

**Posted Date:** October 20th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11116388/v1>

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

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# A Residual Neural Network Framework for Rotating Sound Source Localization

Yifeng Yao<sup>1</sup>, Ce Zhang<sup>2</sup>, Cheng Wei Lee<sup>1</sup>, Wei Ma<sup>1\*</sup>

<sup>1\*</sup>Shanghai Jiao Tong University, Shanghai, 200240, China.

<sup>2</sup>AECC Commercial Aircraft Engine Co., Ltd., Shanghai, 200240, China.

\*Corresponding author(s). E-mail(s): [mawei@sjtu.edu.cn](mailto:mawei@sjtu.edu.cn);

Contributing authors: [yaoyifeng@sjtu.edu.cn](mailto:yaoyifeng@sjtu.edu.cn);

[zhangee\\_phd@foxmail.com](mailto:zhangee_phd@foxmail.com); [chengwei-97@sjtu.edu.cn](mailto:chengwei-97@sjtu.edu.cn);

## Abstract

The identification of rotating sound sources with phased microphone arrays is of great interest in many industrial applications. Rotating beamforming algorithms, such as the time-domain rotating source identifier (ROSI) method and the frequency-domain mode composition beamforming (MCB) method, suffer from significant ghost contributions in rotating sound source localization. Deep learning has emerged as a novel approach for stationary sound source localization. However, since conventional cross-spectral matrix (CSM) are unsuitable for rotating sound source localization, this paper proposes a ResNet-based network model that employs MCB-derived CSM to address rotational Doppler effects in rotating sound source localization. Numerical simulations demonstrate that the proposed method achieves a localization accuracy of up to 98.83% under the adopted evaluation metric, with inference time within 50 milliseconds. Using a 16-microphone array, all sources are successfully identified.

**Keywords:** rotating source localization, cross-spectral matrix, MCB, ROSI, deep learning, ResNet

## 1 Introduction

Acoustic source identification via phased microphone arrays has gained widespread adoption since Johnson’s foundational work [1]. While conventional delay-and-sum beamforming maps source distributions from microphone pressure signals [1], its

initial deployment was restricted to stationary sources. The evolution to moving platforms began with Barsikow’s de-Dopplerization approach for train noise analysis [2], later extended to aircraft noise identification [3, 4], enabling source localization on dynamically changing platforms.

Rotating sound sources present critical challenges in engineering contexts including wind turbine noise analysis and axial fan diagnostics [5]. Sijtsma et al. [5] developed the time-domain rotating source identifier (ROSI) algorithm, which compensates for source motion via transfer functions from rotating sources to stationary microphones. This approach enabled Minck et al. [6] to localize blade-mounted rotating sources in fan systems. However, Ma and Zhang [7] demonstrated that ROSI fails to fully eliminate Doppler effects, inducing ghost contributions at all sideband frequencies (relative to fundamental frequency) with modulation equal to rotation frequency—particularly for non-zero sources. The primary limitation of time-domain methods remains excessive post-processing time, especially with large sample sets.

To improve computational efficiency, frequency-domain rotating beamforming techniques were developed. Lewis and Joseph [8, 9] introduced the virtual rotating array (VRA), designed with rotational speed matching the sound source. This framework was extended for broadband fan noise imaging by Dougherty and Walker [10], generalized to free-field conditions by Pannert and Maier [11], and formalized as linear interpolation for resampling by Herold and Sarradj [12]. Comparative studies [13] confirmed the accuracy of these frequency-domain methods. While VRA beamforming offers significant computational efficiency and enables deconvolution algorithms, its applicability requires a coaxial ring array configuration to satisfy constraints.

To overcome the array constraint inherent in VRA beamforming, Ma and Zhang [14] proposed mode composition beamforming (MCB), which relaxes the rigid array geometry requirements of VRA. However, both time-domain ROSI [7] and frequency-domain MCB [14] fail to fully eliminate ghost contributions under unconstrained array configurations, where arbitrary array geometries and microphone placements are allowed.

Deep learning enables automatic feature extraction directly from raw measurements, representing a paradigm shift from traditional signal processing. Ma and Liu [15] established high-resolution stationary source localization using deep learning, a foundation subsequently advanced by multiple studies. Wallace et al. [16] developed a two-stage microphone array fusion framework to enhance accuracy, while Sun et al. [17] integrated physical constraints via physics-informed neural networks. Chen et al. [18] further improved monopole identification by combining vision transformers with beamforming. Current deep learning methods for rotating sound source localization continue to rely on conventional CSM, which deliver high accuracy for stationary sources but suffer from degraded performance under rotation.

To address this limitation, this paper innovatively employs MCB-derived CSM as the input to a ResNet-based network model [19], enabling robust suppression of ghost contributions without explicit array geometry constraints. Systematic numerical experiments investigate the impact of source frequency, intensity, rotational speed, dataset scale, and network architecture on localization performance.

The remainder of this paper is organized as follows. Section 2 introduces the proposed algorithm and the associated data preprocessing procedures. Section 3 presents the application of ResNet method. Section 4 provides comparative analyses between the ResNet method and conventional rotating sound source localization algorithms. Finally, Section 5 presents the conclusion.

## 2 Algorithm and Data Preprocessing

### 2.1 Network Architecture

Rotating sound sources induce Doppler-related spectral distortion and non-stationary statistical characteristics in MCB-derived CSM, posing significant challenges for learning-based localization frameworks. In such scenarios, increasing the depth of convolutional neural network (CNN) [20] often leads to performance degradation during training convergence, a phenomenon that cannot be simply attributed to overfitting. Instead, deeper architectures may suffer from optimization difficulties, resulting in higher training errors despite increased representational capacity.

To address this issue, this study adopts ResNet architecture [19], which introduces identity mappings between stacked convolutional layers. Let  $H(x)$  denote the desired underlying mapping to be approximated by the network, where  $x$  represents the input CSM. In a residual block, the mapping is reformulated as

$$H(x) = F(x) + x, \quad (1)$$

where  $F(x)$  represents the residual function learned by the stacked layers. This formulation allows the network to focus on learning residual variations rather than directly fitting the full mapping, significantly reducing optimization difficulty in deep networks.

For rotating sound source localization, this residual learning mechanism is particularly advantageous. The identity mappings preserve the fundamental correlation structure of the CSM, while the residual branches are encouraged to learn Doppler-induced spectral distortions and rotation-related acoustic features. As a result, the network can maintain stable convergence while effectively suppressing ghost contributions caused by frequency shifts.

As illustrated in Fig. 1, when the input and output dimensions of a residual block are identical, the shortcut connection directly implements identity mapping. When the dimensions differ, two strategies can be adopted [19]: (1) identity mapping with zero-padding for dimension alignment, or (2) projection mapping using  $1 \times 1$  convolutions. In this study, the latter is employed to ensure consistent feature dimensionality across network stages.

The ResNet18 architecture, shown in Fig. 2, is adopted in this work due to its favorable balance between computational efficiency and representational capability. The network consists of 18 convolutional layers organized into four residual stages, followed by a fully connected layer that outputs a  $1 \times S$  vector representing the predicted acoustic source distribution over the predefined spatial grid.

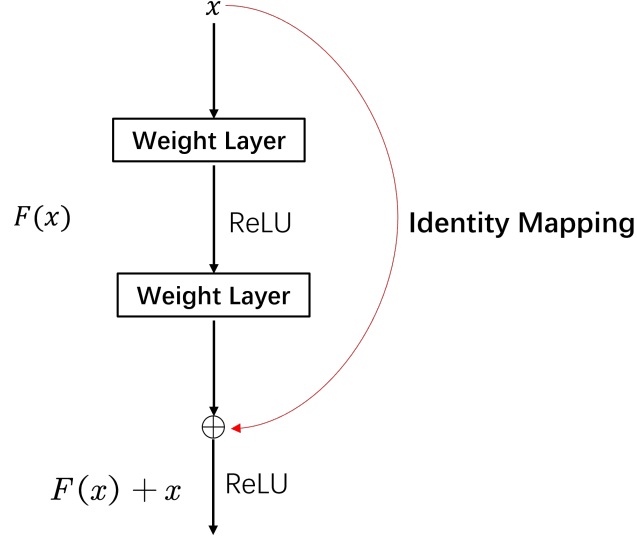


**Fig. 1** Schematic illustration of a residual block

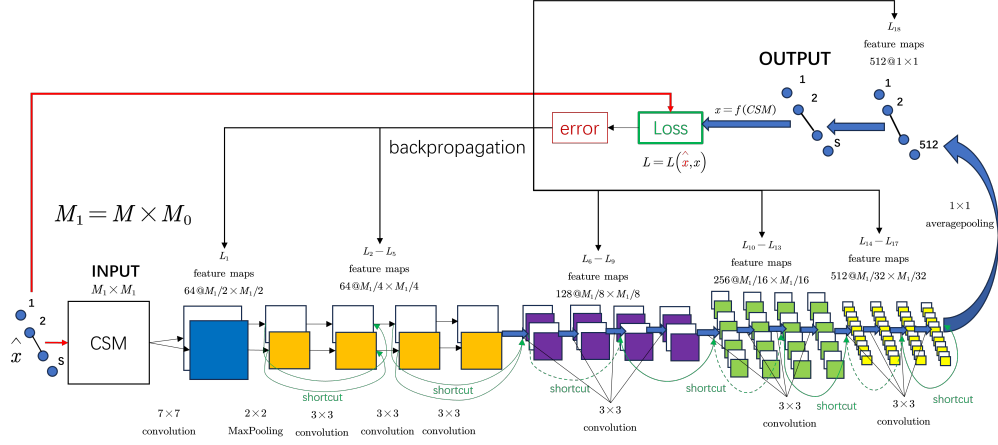


**Fig. 2** Architecture of the ResNet18 network used in this study

The overall training pipeline is as follows. Given a known source distribution and microphone array configuration, the corresponding MCB-derived CSM is first constructed and preprocessed. The processed MCB-derived CSM is then fed into the ResNet18 network to produce a predicted source distribution vector. The discrepancy between the predicted and reference distributions is quantified using a loss function, and network parameters are updated iteratively via backpropagation.

## 2.2 Data Preprocessing

In this study, the MCB-derived CSM is constructed following the MCB formulation [21]. Specifically, the MCB-derived CSM is defined as

$$\text{MCB-derived CSM} = P(\omega)^* P(\omega), \quad (2)$$

where  $(\cdot)^*$  denotes the complex conjugate transpose. The resulting matrix  $C \in \mathbb{C}^{(M \times M_0) \times (M \times M_0)}$  represents the cross-spectral matrix formed by stacking modal frequency components across all microphones. Here,  $M$  denotes the number of microphones, and  $M_0$  is the total number of acoustic modes considered in the MCB framework.

The matrix  $P(\omega)$  is constructed as

$$P(\omega) = \begin{bmatrix} P_1(\omega + (\frac{-M_0}{2} + 1)\Omega) & \cdots & P_1(\omega + \frac{M_0}{2}\Omega) \\ \vdots & \ddots & \vdots \\ P_m(\omega + (\frac{-M_0}{2} + 1)\Omega) & \cdots & P_m(\omega + \frac{M_0}{2}\Omega) \\ \vdots & \ddots & \vdots \\ P_M(\omega + (\frac{-M_0}{2} + 1)\Omega) & \cdots & P_M(\omega + \frac{M_0}{2}\Omega) \end{bmatrix}, \quad (3)$$

where  $P_m(\cdot)$  denotes the frequency-domain spectrum of the  $m$ -th microphone signal,  $\Omega$  represents the rotational angular frequency, and the frequency-shifted components account for the Doppler-induced modulation associated with rotating sound sources. This construction embeds both spatial and rotational information into the CSM, which serves as the network input after preprocessing.

In microphone array-based acoustic measurements, the magnitude of the MCB-derived CSM is strongly influenced by multiple physical and operational factors, including acoustic source power, rotational speed, source frequency, propagation distance, and measurement noise. In rotating sound source scenarios, Doppler-induced spectral spreading and time-varying phase modulation further amplify the dynamic range variation of MCB-derived CSM entries. As a result, raw MCB-derived CSM obtained under different operating conditions may differ by several orders of magnitude, even when the underlying spatial source distribution remains similar.

From an engineering perspective, such scale-dominated variations obscure the spatial correlation patterns embedded in the MCB-derived CSM, which are the primary cues for sound source localization. When directly used as network inputs, large-magnitude components tend to dominate the gradient updates during training, causing the learning process to bias toward source intensity variations rather than spatial distribution characteristics. This effect becomes more pronounced in rotating conditions, where Doppler effects introduce additional non-stationarity into the MCB-derived CSM statistics.

To mitigate these issues and to ensure numerical stability during network training, a standardization-based preprocessing strategy is adopted in this study. By normalizing the real and imaginary parts of the MCB-derived CSM independently, the proposed approach suppresses amplitude-driven variations while preserving relative

phase relationships and spatial correlation structures across the microphone array. This preprocessing enables the network to focus on learning geometry-related and rotation-related acoustic features, rather than being dominated by absolute signal magnitude differences across samples.

To this end, the real and imaginary components of the MCB-derived CSM are processed separately, as detailed below. Let the complex-valued MCB-derived CSM be expressed as

$$\text{MCB-derived CSM} = A + jB, \quad (4)$$

where  $A$  and  $B$  denote the real and imaginary parts, respectively. Each component is standardized independently as

$$A_{\text{std}} = \frac{A - \text{mean}(A)}{\text{std}(A)}, \quad B_{\text{std}} = \frac{B - \text{mean}(B)}{\text{std}(B)}, \quad (5)$$

where  $\text{mean}(\cdot)$  and  $\text{std}(\cdot)$  denote the mean value and standard deviation of all matrix elements, respectively. This normalization suppresses scale-dominated variations and emphasizes correlation structures that are invariant to source intensity, thereby facilitating stable training under Doppler-induced non-stationarity.

After preprocessing, the dataset is randomly shuffled at the beginning of each training epoch to enhance generalization by preventing the network from learning spurious data ordering patterns.

The network is trained using the mean squared error (MSE) loss function, defined as

$$\text{Loss} = \frac{1}{S} \sum_{i=1}^S (\hat{x}_i - x_i)^2 = \frac{1}{S} \sum_{i=1}^S (|\hat{q}_i|^2 - |q_i|^2)^2, \quad (6)$$

where  $\hat{x}_i$  and  $x_i$  denote the predicted and reference acoustic source power intensities at the  $i$ -th spatial grid, and  $\hat{q}_i$  and  $q_i$  represent the corresponding sound pressure values.

Model parameters are optimized using the Adam optimizer. The initial learning rate is set to  $10^{-3}$ , with  $10^{-4}$  adopted in certain cases to improve stability. A batch size of 50 samples is used, and training is performed for a maximum of 45 epochs. An exponential learning rate decay strategy is employed:

$$\text{LearningRate} = 0.9^{\text{epoch}}, \quad (7)$$

where epoch denotes the number of completed training epochs. This strategy enables rapid initial convergence while preventing oscillations near the optimum during later training stages.

### 2.3 Evaluation Metrics

To comprehensively assess localization performance under different acoustic conditions, two complementary evaluation metrics are adopted. The first metric follows the criterion proposed by Ma and Liu [15] and is applicable to scenarios with homogeneous

source power distributions. The localization accuracy is defined as

$$\text{Accuracy}_1 = \frac{N_{\text{correct}}}{N_{\text{group}}}, \quad (8)$$

where  $N_{\text{group}}$  is the total number of test samples and  $N_{\text{correct}}$  is the number of correctly localized samples. For a sample containing  $N_s$  sources, the predicted power intensities are sorted in descending order to form a grid index set  $\{y_1, \dots, y_{N_s}\}$ . A prediction is considered correct if

$$\{y_1, \dots, y_{N_s}\} = \{\hat{y}_1, \dots, \hat{y}_{N_s}\}, \quad (9)$$

where  $\{\hat{y}_1, \dots, \hat{y}_{N_s}\}$  denotes the reference source locations.

Eq. 8 defines the criterion for correctly identifying an entire sound source group. For instance, in a sound source group with 4 actual sources, the group is considered correctly predicted if and only if the top 4 regions ranked by predicted intensity perfectly overlap with the 4 regions containing actual sources. Even a single mispredicted region results in the entire group being classified as incorrect, and the number of correctly predicted sources within the group is not reflected in the accuracy rate of Eq. 8. To comprehensively evaluate the model's performance, this paper proposes Eq. 10, which uses the number of correctly predicted sources as the numerator (instead of the number of correctly predicted groups) and the total number of sound sources as the denominator (rather than the total number of source groups). Eq. 8 and Eq. 10 are independent of each other as accuracy evaluation criteria.

$$\text{Accuracy}_2 = \frac{N_{\text{source\_correct}}}{N_{\text{source\_num}}}. \quad (10)$$

Where  $N_{\text{source\_num}}$  is the total number of sources and  $N_{\text{source\_correct}}$  is the number of correctly localized sources. A source is considered correctly localized if its position is correctly identified and the predicted power intensity satisfies

$$|\hat{x}_i - x_i| < 0.1, \quad (11)$$

where  $\hat{x}_i$  and  $x_i$  denote the predicted and reference power intensities at the corresponding grid. This metric allows for small tolerances in intensity prediction while evaluating each source individually, providing a more informative assessment of model performance in complex acoustic environments.

## 3 Application

### 3.1 Uniform-Intensity Rotating Source Localization

In this subsection, the microphone array is positioned parallel to the rotating sound source plane. This configuration is adopted to systematically investigate the coupled effects of source frequency and rotational speed on rotating sound source localization performance. With frequencies randomly distributed within the range of 4500–5000 Hz, the sound sources have a rotational speed varying from 500 to 2500 rpm and a root

mean square (rms) of 1. In addition, the influence of network architecture depth and training dataset scale on localization accuracy is quantitatively analyzed. The key experimental parameters are summarized as Table 1.

**Table 1** Information of simulations

| Category       | Parameter     | Value                     |
|----------------|---------------|---------------------------|
| Source         | frequency     | 4500–5000 Hz              |
|                | modes         | 14                        |
|                | number        | 2–6                       |
| Array          | microphones   | 16                        |
|                | aperture      | 0.465 m                   |
|                | center        | (0, 0, 0) m               |
| Scanning plane | grid points   | $10 \times 10$            |
|                | resolution    | 0.16 m                    |
|                | position      | $z = 0.5$ m               |
| Rotating       | speed         | 500–2500 rpm              |
|                | axis          | z-axis                    |
|                | center        | (0, 0, 1.00) m            |
| Sampling       | rate          | 25600 Hz                  |
|                | samples       | $1920 \times 5 \times 10$ |
| Training       | learning rate | see Eq. 7                 |
|                | accuracy      | see Eq. 8                 |

A summary of the adopted network architectures and dataset scales is provided in Table 2. The corresponding microphone array configuration is illustrated in Fig. 3.

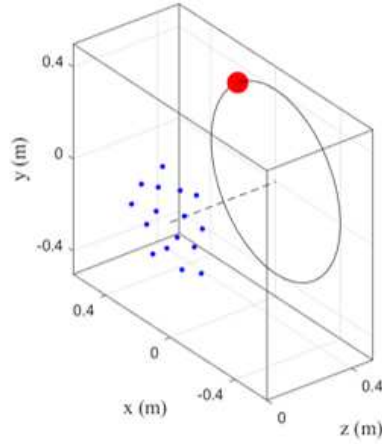


**Fig. 3** Microphone array schematic parallel to the rotating sound source plane

**Table 2** Localization accuracy for sound sources with identical intensity  
( $f = [4500, 5000]$  Hz,  $R = [500, 2500]$  rpm)

| Training Set ( $\times 10^4$ ) | Validation Set ( $\times 10^4$ ) | Test Set ( $\times 10^4$ ) | Network  | Accuracy |
|--------------------------------|----------------------------------|----------------------------|----------|----------|
| 8.32                           | 2.08                             | 0.42                       | ResNet18 | 52.43%   |
| 12.40                          | 3.10                             | 0.68                       | ResNet18 | 89.32%   |
| 12.40                          | 3.10                             | 0.68                       | ResNet34 | 94.31%   |

As shown in Table 2, when trained with only 83,200 samples, ResNet18 achieves a localization accuracy of 52.43%. This result indicates insufficient coverage of the joint frequency–rotational speed parameter space. Since both parameters are randomized, certain parameter combinations in the test set are not encountered during training, leading to poor generalization capability and large performance variance across test cases.

By increasing the training dataset to 123,200 samples while maintaining the same training strategy, the localization accuracy of ResNet18 improves significantly to 89.32%. This confirms that the initial performance degradation primarily originates from inadequate training data rather than intrinsic limitations of the network architecture.

With the same dataset scale, the deeper ResNet34 further improves localization accuracy to 94.31%, demonstrating the benefit of increased network depth when sufficient data are available. The residual connections effectively mitigate the degradation problem commonly observed in deep networks. However, ResNet34 requires substantially longer training time compared with ResNet18. From an engineering perspective, ResNet18 provides a more favorable trade-off between accuracy, computational cost, and data efficiency. The prediction results of ResNet34 are shown in Fig. 4. As shown in Fig 4, according to Eq. 9, all four predicted results are correct. In these four figures, the actual sound source intensities are equal, but the predicted results show unequal values, indicating that under the 124,000 samples, although the accuracy rate exceeds 90%, the prediction quality remains poor in practice.

### 3.2 Non-Uniform Intensity Rotating Source Localization

The following simulation experiments place the microphone array Parallel to the rotating sound source plane. After achieving stable localization accuracy for sound sources with frequencies and rotational speeds within predefined ranges, the constraint on sound source intensity is further relaxed. The rms values are discretely sampled from the interval  $[0.5, 1.0]$  with a step size of 0.1. This configuration better reflects realistic engineering scenarios where sound sources typically exhibit non-uniform intensity distributions. The effects of dataset scale and network architecture on localization accuracy are systematically investigated. The experimental parameters are summarized as Table 3. The corresponding results are summarized in Table 4.

When intensity variations are introduced, insufficient training data lead to an underfitting regime, particularly for shallower networks. After increasing the training

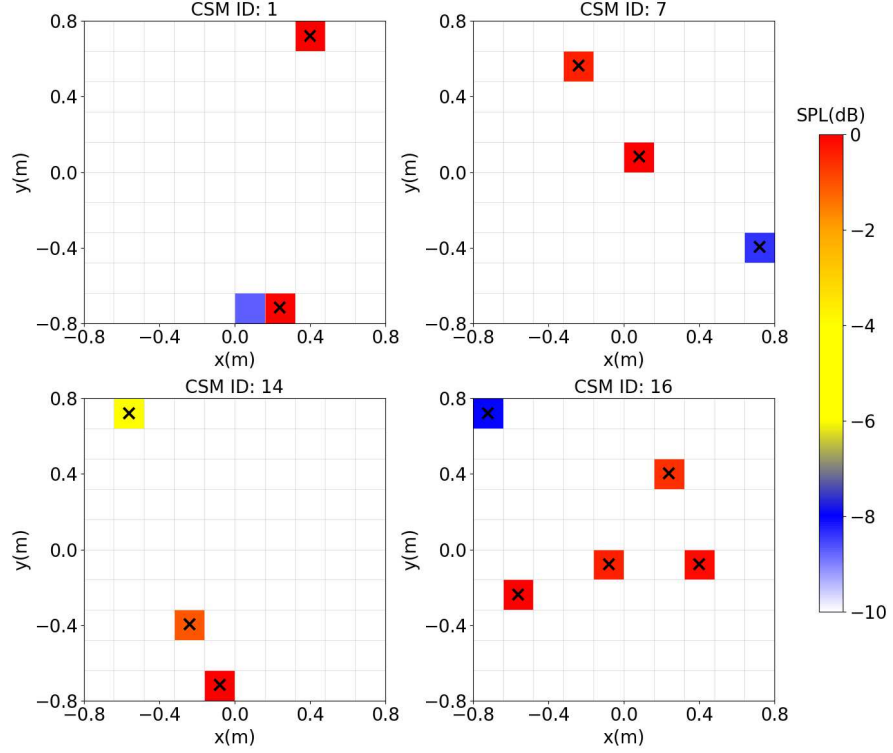


**Fig. 4** ResNet34-based prediction of rotating sound sources with uniform intensity under parallel array configuration ( $f = [4500, 5000]$  Hz,  $R = [500, 2500]$  rpm)

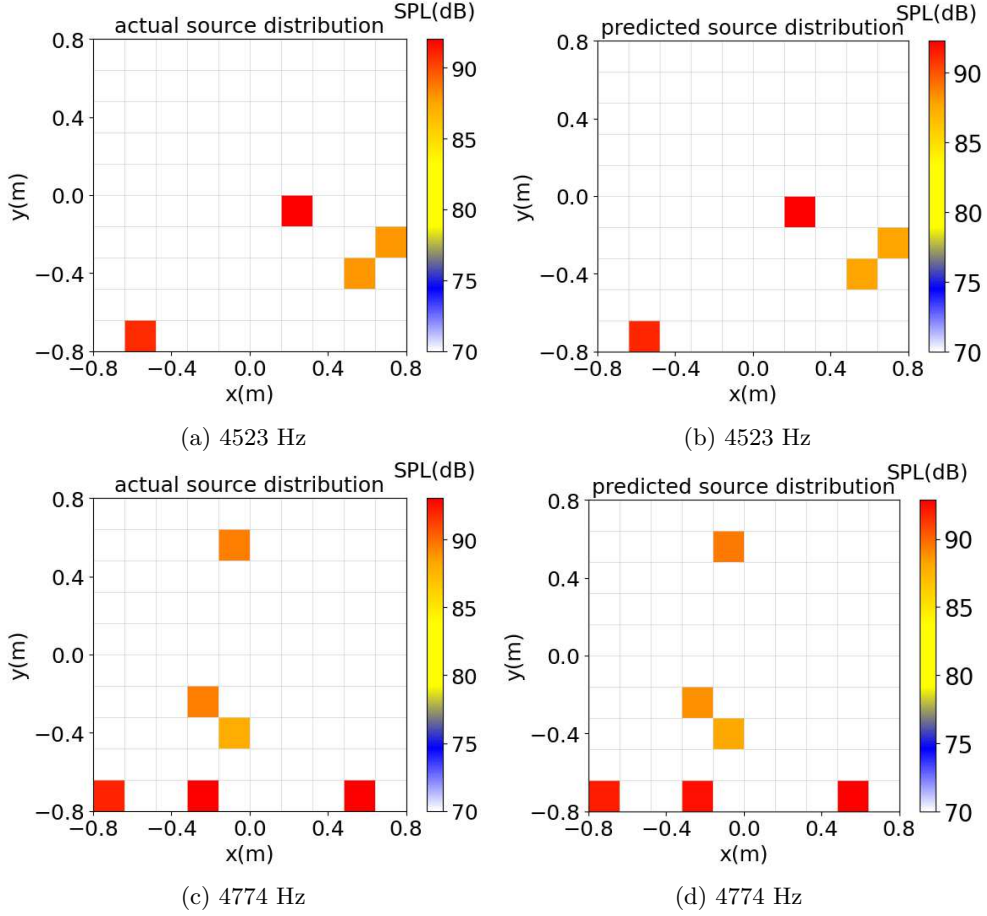
**Table 3** Simulation parameters

| Category       | Parameter     | Value                         |
|----------------|---------------|-------------------------------|
| Source         | frequency     | 4500–5000 Hz                  |
|                | modes         | 14                            |
|                | number        | 2–6                           |
| Array          | microphones   | 16                            |
|                | aperture      | 0.465 m                       |
|                | position      | $z = 0$ m, center (0, 0, 0) m |
| Scanning plane | grid points   | $10 \times 10$                |
|                | resolution    | 0.16 m                        |
|                | position      | $z = 0.5$ m                   |
| Rotation       | speed         | 500–2500 rpm                  |
|                | axis          | z-axis                        |
|                | center        | (0, 0, 1.00) m                |
| Sampling       | rate          | 25600 Hz                      |
|                | samples       | $1920 \times 5 \times 10$     |
| Training       | learning rate | see Eq. 7                     |
|                | accuracy      | see Eq. 10                    |

**Table 4** Localization accuracy for sound sources with non-uniform intensity

| Training Set ( $\times 10^4$ ) | Validation Set ( $\times 10^4$ ) | Test Set ( $\times 10^4$ ) | Network  | Accuracy |
|--------------------------------|----------------------------------|----------------------------|----------|----------|
| 18.90                          | 2.10                             | 3.94                       | ResNet18 | 61.00%   |
| 17.01                          | 1.89                             | 5.90                       | ResNet34 | 77.40%   |
| 24.30                          | 2.70                             | 15.75                      | ResNet34 | 92.48%   |

dataset by approximately 70,000 samples, the deeper ResNet34 achieves a localization accuracy of 92.48% under the per-source evaluation criterion (see Eq. 10). Representative prediction results are illustrated in Fig. 5.

**Fig. 5** ResNet18-based prediction of rotating sound sources with non-uniform intensity under parallel array configuration

The predicted heatmaps exhibit excellent agreement with the ground truth in both spatial location and intensity level. This capability is essential for practical rotating

noise diagnostics, where identifying dominant and secondary noise sources is critical for targeted noise mitigation.

The following simulation experiments place the microphone array perpendicular to the rotating sound source plane. All results discussed above are obtained with the microphone array plane parallel to the rotating source plane. The performance of the proposed deep learning framework under this challenging configuration is therefore investigated. The key experimental parameters are summarized as Table 5.

**Table 5** Simulation parameters

| Category       | Parameter     | Value                               |
|----------------|---------------|-------------------------------------|
| Source         | frequency     | 4500–5000 Hz                        |
|                | modes         | 14                                  |
|                | number        | 2–6                                 |
| Array          | microphones   | 16                                  |
|                | diameter      | 0.8 m                               |
|                | position      | $x = 0.8$ m, center $(0.8, 0, 0)$ m |
| Scanning plane | grid points   | $10 \times 10$                      |
|                | resolution    | 0.16 m                              |
|                | position      | $z = 0.5$ m                         |
| Rotation       | speed         | 500–2500 rpm                        |
|                | axis          | z-axis                              |
|                | center        | $(0, 0, 1.00)$ m                    |
| Sampling       | rate          | 25600 Hz                            |
|                | samples       | $1920 \times 5 \times 10$           |
| Training       | learning rate | see Eq. 7                           |
|                | accuracy      | see Eq. 10                          |

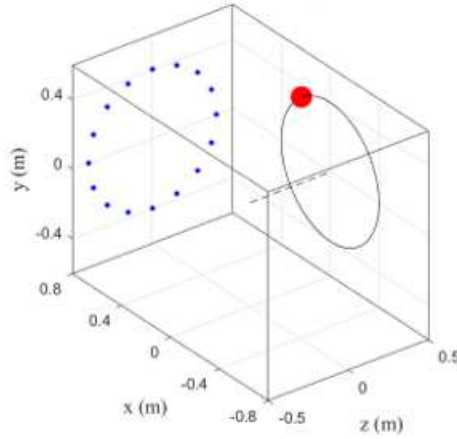


**Fig. 6** Microphone array perpendicular to the plane of rotating sound sources

The microphone array employed in this study is shown in Fig. 6 and the results for different network architectures and dataset scales are summarized in Table 6.

**Table 6** Localization accuracy for sound sources with non-uniform intensity under perpendicular array configuration

| Training Set ( $\times 10^4$ ) | Validation Set ( $\times 10^4$ ) | Test Set ( $\times 10^4$ ) | Network     | Accuracy |
|--------------------------------|----------------------------------|----------------------------|-------------|----------|
| 27.00                          | 3.00                             | 8.28                       | ResNet18    | 87.87%   |
| 27.00                          | 3.00                             | 8.28                       | ResNet34    | 88.23%   |
| 27.00                          | 3.00                             | 8.28                       | ResNet50    | 59.78%   |
| 27.00                          | 3.00                             | 8.28                       | MobileNetV3 | 50.39%   |
| 44.64                          | 4.96                             | 9.02                       | ResNet18    | 95.20%   |
| 44.64                          | 4.96                             | 9.02                       | ResNet34    | 95.86%   |
| 44.64                          | 4.96                             | 9.02                       | ResNet50    | 95.26%   |
| 44.64                          | 4.96                             | 9.02                       | MobileNetV3 | 64.59%   |
| 58.50                          | 6.50                             | 9.43                       | ResNet18    | 96.80%   |
| 58.50                          | 6.50                             | 9.43                       | ResNet34    | 96.01%   |
| 58.50                          | 6.50                             | 9.43                       | ResNet50    | 96.30%   |
| 58.50                          | 6.50                             | 9.43                       | ResNet101   | 96.49%   |
| 71.10                          | 7.90                             | 7.90                       | ResNet18    | 98.66%   |
| 71.10                          | 7.90                             | 7.90                       | ResNet34    | 98.83%   |
| 71.10                          | 7.90                             | 7.90                       | ResNet50    | 98.82%   |
| 71.10                          | 7.90                             | 7.90                       | ResNet101   | 97.78%   |

Fig. 7 compares the ground truth and predicted sound source distributions. Even under lateral array deployment, the proposed model achieves accurate localization of both source positions and intensities, demonstrating robustness in scenarios where conventional beamforming methods typically fail.

As the randomization constraints on source parameters are progressively relaxed, the required training dataset size increases accordingly. Localization accuracy approaches saturation near 100%, while heatmap visualization enables clear discrimination of relative source intensities. When the microphone array configuration matches that used in training and all parameters fall within the randomized ranges, the model generates predictions within 50 milliseconds.

The most time-consuming stage remains simulation data generation. While generating large datasets on a personal computer may require several weeks, high-performance computing resources can reduce this time to hours. Model training with ResNet18 completes within approximately one hour on a RTX4060 GPU. Training curves indicate stable convergence without evident overfitting or underfitting.

### 3.3 Multi-source Ghost Contributions Testing

The sound source count is further increased to six by selecting two frequencies (4593 Hz and 4927 Hz), with three spatially separated sources assigned to each frequency. The key experimental parameters are summarized as Table 7.

The prediction results obtained using the array configuration shown in Fig. 3 are presented in Fig. 8, while the prediction results obtained using the array configuration shown in Fig. 6 are shown in Fig. 9.

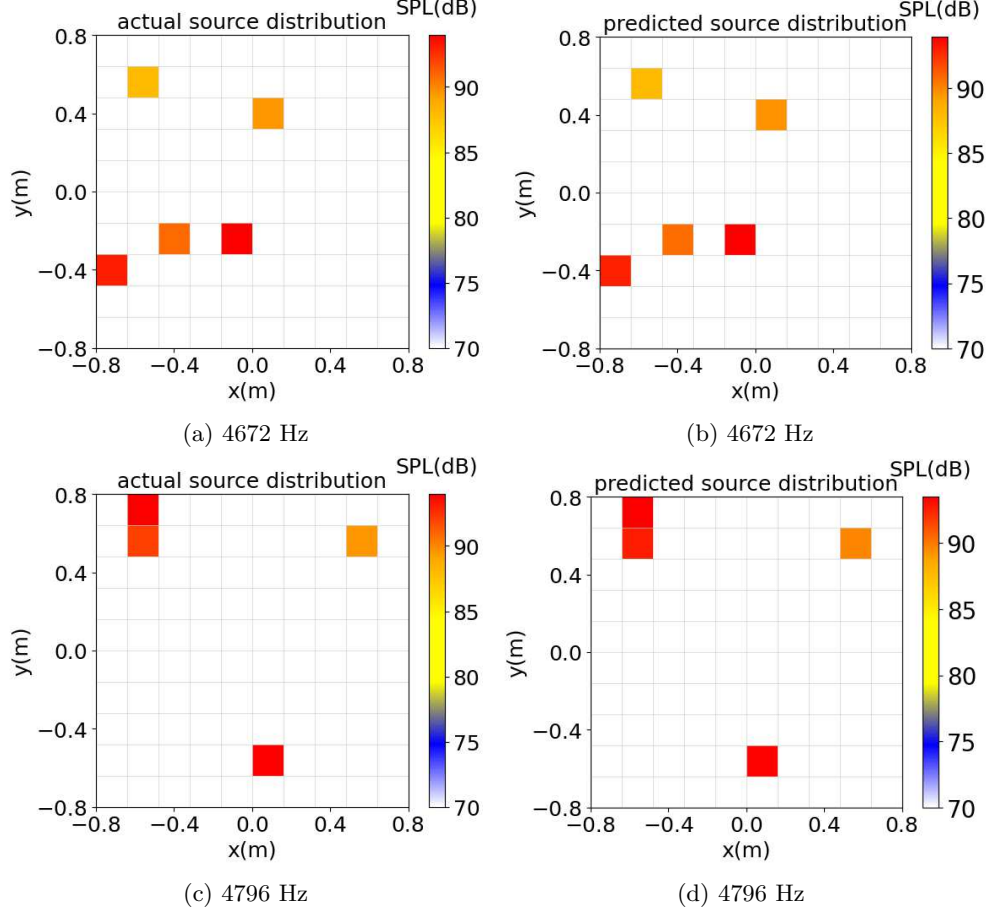


**Fig. 7** ResNet18-based prediction of rotating sound sources with non-uniform intensity under perpendicular array configuration

Under the condition where the microphone array is parallel to the rotating sound source plane, Fig. 8 shows the heatmap of sound source intensity distributions for different frequencies. Only when the frequency matches 4593 Hz or 4927 Hz (the frequencies of the sound sources) do the prediction results reveal three distinct sound sources. At other frequencies, no sound sources are predicted. Notably, the ResNet method does not exhibit ghost contributions. Furthermore, under the perpendicular array configuration (Fig. 6), no ghost contributions are observed either, demonstrating that array orientation does not affect ResNet method’s ability to compensate for Doppler-induced ghost contributions.

If ResNet method fails to compensate for Doppler-induced ghost contributions, it may generate false sound source predictions under specific conditions. Specifically, when the scanning frequency ( $f_{\text{scan}}$ ) deviates from the actual source frequency ( $f_{\text{source}}$ ) by an integer multiple of the rotational speed  $R$  (i.e.,  $\Delta f = |f_{\text{scan}} - f_{\text{source}}| = n \cdot R$ , where  $\Delta f$  is the frequency deviation in Hz,  $f_{\text{scan}}$  and  $f_{\text{source}}$  denote scanning and

**Table 7** Simulation parameters

| Category       | Parameter     | Value                     |
|----------------|---------------|---------------------------|
| Source         | frequency     | 4593, 4927 Hz             |
|                | modes         | 14                        |
|                | number        | 6                         |
| Scanning plane | grid points   | $10 \times 10$            |
|                | resolution    | 0.16 m                    |
|                | position      | $z = 0.5$ m               |
| Rotation       | speed         | 1500 rpm                  |
|                | axis          | z-axis                    |
|                | center        | (0, 0, 1.00) m            |
| Sampling       | rate          | 25600 Hz                  |
|                | samples       | $1920 \times 5 \times 10$ |
| Training       | learning rate | see Eq. 7                 |

true source frequencies respectively,  $R$  is the rotational speed in Hz, and  $n \in \mathbb{Z}^+$  is a positive integer representing the harmonic order), the heatmap will erroneously detect non-existent sources, leading to invalid localization results.

## 4 Comparative Analysis with Conventional Rotating Beamforming Methods

Current methods such as MCB, and ROSI based on microphone arrays have been applied to rotating sound source localization. ROSI method operates in the time domain for rotating source beamforming with no array type restrictions but suffers from slow computation speed and inability to perform deconvolution. In contrast, MCB offers array-type flexibility and supports deconvolution. To highlight differences in source localization performance, this study compares MCB, ROSI, and ResNet using simulation data. The comparative results are presented in Figs. 10 and 11.

The simulation employs  $M = 16$  microphones with  $M_0 = 14$  selected modes, using the microphone array configuration illustrated in Fig. 3. The sound source group consists of sources at 5000 Hz rotating at a constant speed of 1500 rpm, with identical source intensities. Actual source positions are indicated by black crosses in ResNet prediction maps for reference. Three test cases were randomly selected for illustration, in which all sound sources have uniform intensity, a rotational speed of 1500 rpm, and a frequency of 5000 Hz. As shown in Fig. 10, Case 1 contains five sound sources. The MCB method correctly identifies only three sources, while ROSI yields an extra predicted source, reflecting a localization inaccuracy. In contrast, the ResNet model accurately identifies all sources without false positives. The ROSI and MCB results are nearly identical, failing to resolve two adjacent sources. In Case 2 (Fig. 11), MCB and ROSI are prone to generating spurious source detections. Although the falsely predicted sources are spatially close to the true sources, they hinder accurate

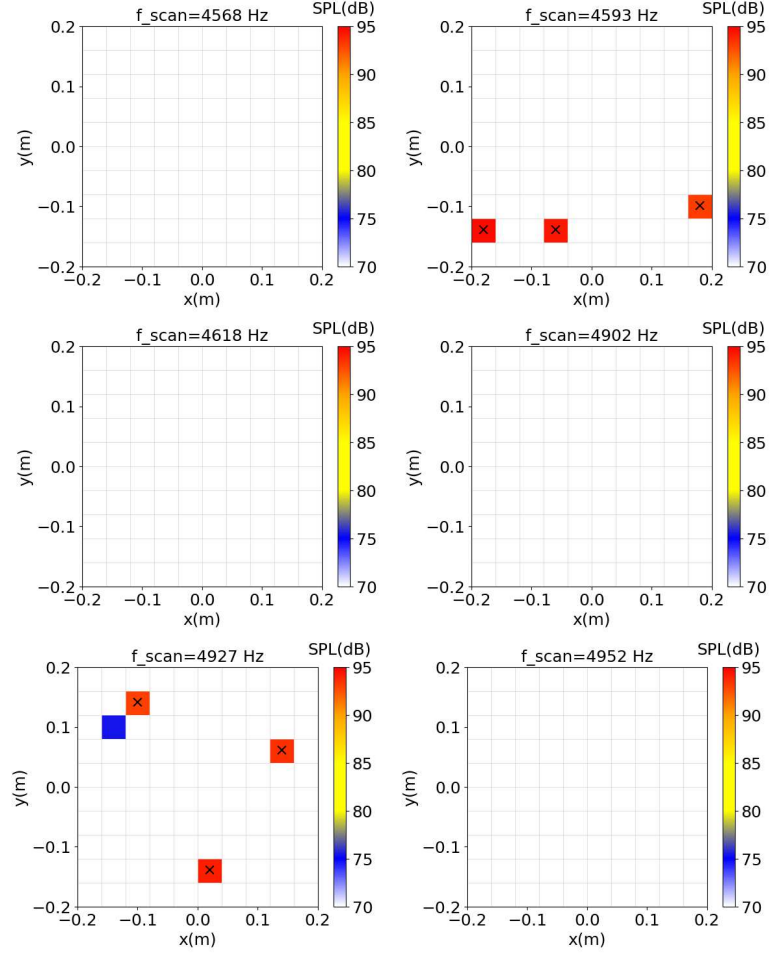


**Fig. 8** Ghost contribution suppression results for six sound sources under parallel array configuration

identification of the actual number of sound sources. ResNet maintains perfect localization accuracy. In contrast, ResNet accurately predicts the source count, position, and intensity.

The ResNet model demonstrates no ghost contributions when the microphone array is perpendicular to the rotating source plane—a critical limitation observed in MCB [14]. This indicates that deep learning effectively learns CSM features to overcome challenges inherent in conventional algorithms. Additionally, ResNet achieves up to 98.83% accuracy for sound sources varying within specific ranges of rotational speed, frequency, and intensity. The model processes input CSM to output predictions within 50 milliseconds, whereas MCB requires more than 800 milliseconds and ROSI requires more than 4 seconds per case. This computational efficiency makes ResNet particularly suitable for large-scale data processing applications. Compared to MCB and ROSI, ResNet eliminates the need for propagation functions, requiring only 16 microphones

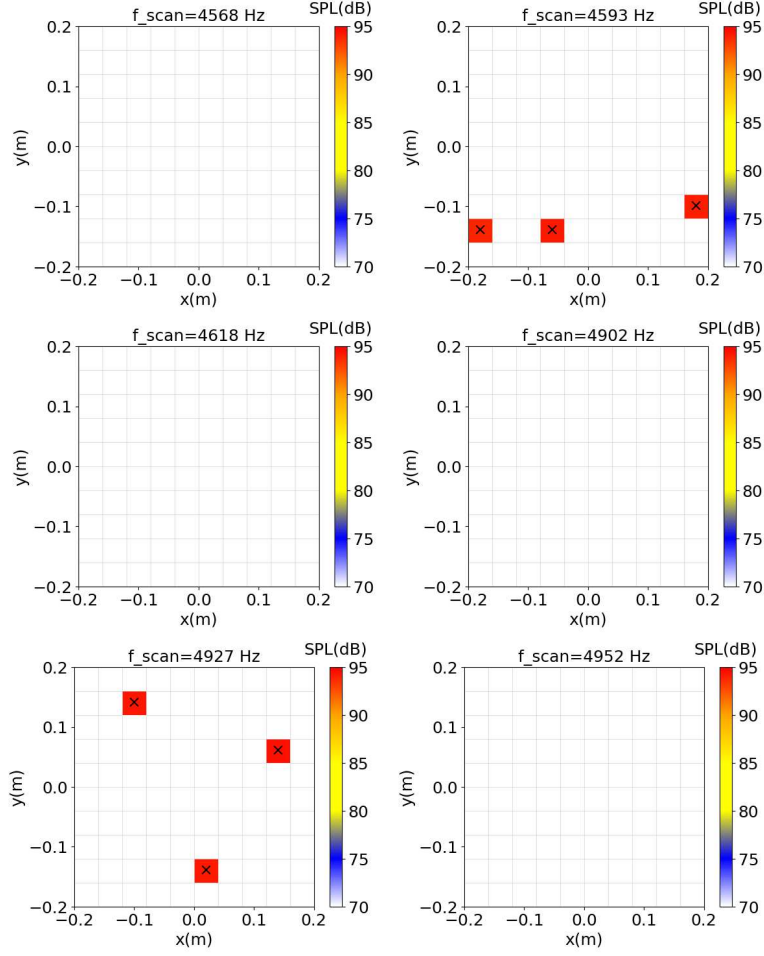


**Fig. 9** Ghost contribution suppression results for six sound sources under perpendicular array configuration

to localize 6 rotating sources—a significant reduction in input requirements relative to conventional beamforming techniques.

However, three primary challenges remain. First, training data generation involves substantial time and storage costs. Three categories of data are considered: simulated data with known propagation functions (current approach), computational aeroacoustics (CAA)-generated data for unknown propagation scenarios, and experimental data requiring resource-intensive accumulation. While replacing double-precision with single-precision floating-point storage reduces data volume by 50%, storage requirements remain significant. Future studies will explore hybrid training with combined simulation and experimental data to save time costs. Second, trained models exhibit strong specialization. Changes in microphone array configurations or array-source plane relationships necessitate new simulation data generation, limiting data reuse

capabilities. Third, the physical mechanisms underlying ResNet’s predictions remain opaque. Future research will incorporate additional acoustic parameters to enhance localization interpretability and provide clearer mechanistic explanations.

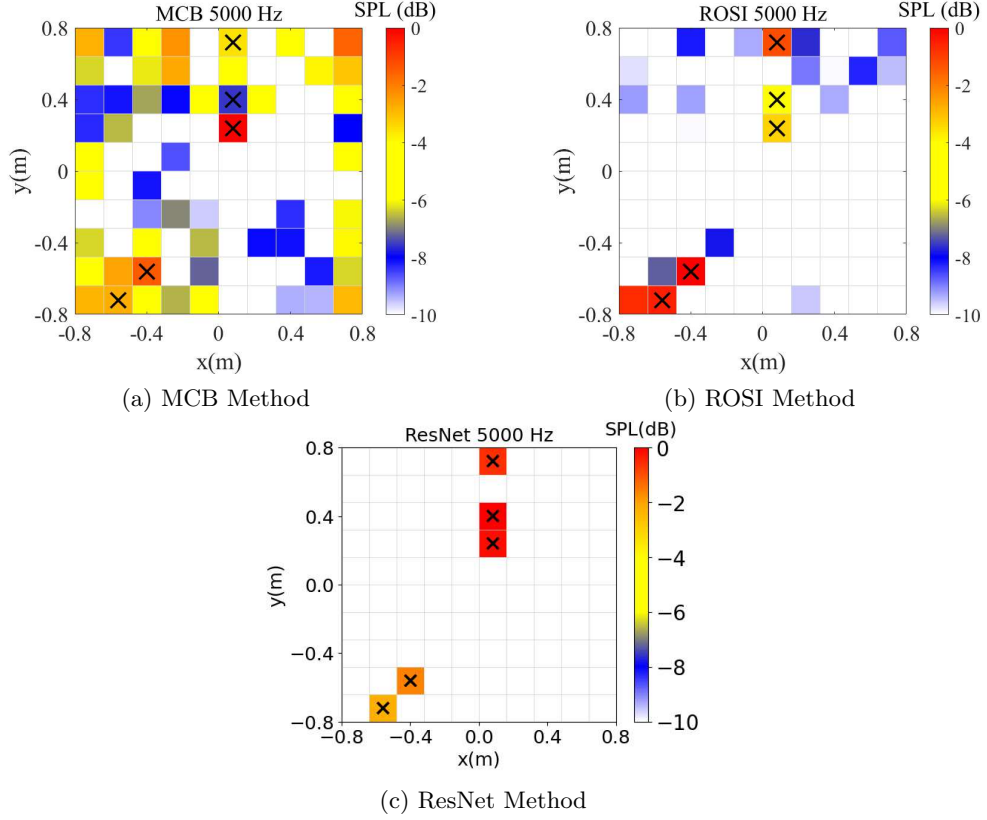


**Fig. 10** Comparative Results of MCB, ROSI, and ResNet (Case 1)

## 5 Conclusion

This study demonstrates the first application of ResNet for rotating sound source localization. The ResNet model takes as input a standardized MCB-derived CSM and outputs a one-dimensional vector representing the sound source intensity distribution, achieving a 98.83% localization accuracy with only 16 microphones. For identical test cases, MCB requires more than 800 milliseconds and ROSI requires more than 4 seconds to get the result, while ResNet delivers predictions within 50 milliseconds, representing orders of magnitude faster computation. Furthermore, the ResNet-based approach completely avoids ghost contribution. This breakthrough establishes ResNet as a novel algorithm for rotating source localization with microphone arrays. Critically, it eliminates the necessity for prior knowledge of propagation functions and

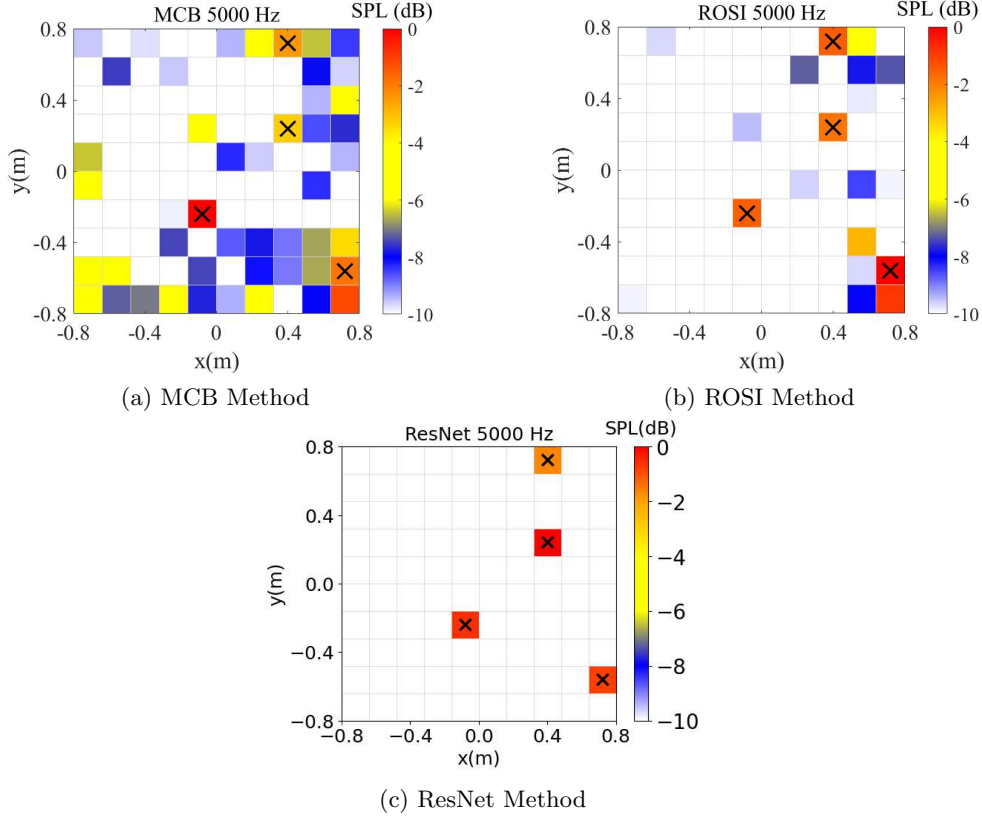


**Fig. 11** Comparative Results of MCB, ROSI, and ResNet (Case 2)

microphone positions, directly determining source distributions from MCB-derived CSM with near-perfect fidelity. These advantages make ResNet a promising framework meriting further investigation in rotational acoustics applications.

**Acknowledgments.** This work is supported by the National Natural Science Foundation of China (W2442003). The authors gratefully acknowledge the NSFC's support in advancing rotating sound source localization techniques.

## References

- [1] Johnson, D.H., Dudgeon, D.: Array Signal Processing: Concepts and Techniques. Prentice Hall, ??? (1993)
- [2] Barsikow, B.: Experiences with various configurations of microphone arrays used to locate sound sources on railway trains operated by the DBAG. *Journal of Sound and Vibration* **193**(1), 283–293 (1996)
- [3] Michel, U., Barsikow, B., Helbig, J., Hellming, M., Schuttpelz, M.: Flyover noise measurements on landing aircraft with a microphone array. In: *Proceedings of*

the AIAA Aviation Technology, Integration, and Operations (ATIO) Conference (AIAA-1998-2336) (1998)

- [4] Siller, H., Hage, W., Schumacher, T.: Source localisation on aircraft in flight: new measurements with the DLR research aircraft airbus 320 ATRA. In: Proceedings of the European Conference on Noise Control (BeBeC) (BeBeC-2018-D01) (2018)
- [5] Sijtsma, P., Oerlemans, S., Holthusen, H.H.: Location of rotating sources by phased array measurements. In: Proceedings of the 7th AIAA/CEAS Aeroacoustics Conference (2001)
- [6] Minck, O., Binder, N., Cherrier, O., Lamotte, L., Budinger, V.: Fan noise analysis using a microphone array. In: Proceedings of the 18th AIAA/CEAS Aeroacoustics Conference (2012)
- [7] Ma, W., Zhang, C.: Doppler effect in the time-domain beamforming for rotating acoustic source identification. *The Journal of the Acoustical Society of America* **148**, 430–443 (2020)
- [8] Lowis, C.R., Joseph, P.: Inversion technique for determining the strength of rotating broadband sources in ducts. In: Proceedings of the 11th AIAA/CEAS Aeroacoustics Conference (2005)
- [9] Lowis, C.R., Joseph, P.: Determining the strength of rotating broadband sources in ducts by inverse methods. *Journal of Sound and Vibration* **335**, 614–632 (2006)
- [10] Dougherty, R.P., Walker, B.E.: Virtual rotating microphone imaging of broadband fan noise. In: Proceedings of the 15th AIAA/CEAS Aeroacoustics Conference (2009)
- [11] Pannert, W., Maier, C.: Rotating beamforming: Motion compensation in the frequency-domain and application of high-resolution beamforming algorithms. *Journal of Sound and Vibration* **333**, 1899–1912 (2014)
- [12] Herold, G., Sarradj, E.: Microphone array method for the characterization of rotating sound sources in axial fans. *Noise Control Engineering Journal* **63**(6), 546–551 (2015)
- [13] Herold, G., Ocker, C., Sarradj, E., Pannert, W.: A comparison of microphone array methods for the characterization of rotating sound sources. In: Proceedings of the Berlin Beamforming Conference (2018)
- [14] Ma, W., Zhang, C.: A frequency-domain beamforming for rotating sound source identification. *The Journal of the Acoustical Society of America* **148**(3), 1602–1613 (2020)
- [15] Ma, W., Liu, X.: Phased microphone array for sound source localization with

- deep learning. *Aerospace Systems* **2**, 71–81 (2019)
- [16] Shahmohammadimehrjardi, M., Wallace, B., Chan, A.D.C., Goubran, R., Xi, P.C.: Two-step microphone array fusion algorithm for enhanced indoor sound source localization. In: *Proceedings of the IEEE Sensors Applications Symposium (SAS)* (2025)
  - [17] Sun, B., Xue, M., Su, M.: Physics-informed neural network enhanced acoustic source identification via cross-spectral correlation of microphone array signals. *Mechanical Systems and Signal Processing* **236**, 113027 (2025)
  - [18] Chen, J., Ma, B., Lee, C.W., Liu, X., Ma, W.: Enhancing the accuracy of monopole and dipole source identification with vision transformer. *Acoustics* **7**(4) (2025)
  - [19] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (2016)
  - [20] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105 (2012)
  - [21] Zhang, C., Ma, W.: Deconvolution of mode composition beamforming for rotating source localization. *Aerospace Systems* **7**, 727–734 (2024)