Deep Learning-Enabled High-Resolution and Fast Sound Source Localization in Spherical Microphone Array System

Soo Young Lee, Jiho Chang, and Seungchul Lee

Abstract—**While sound source localization (SSL) using a spherical microphone array system can be applied to obtain visual beam patterns of source distribution maps in a range of omnidirectional acoustic applications, the present challenges of the spherical measurement system on the valid frequency ranges and the spatial distortion as well as the grid-related limitations of data-driven SSL approaches raise the need to develop an appropriate method. Imbued by these challenges, this study proposes a deep learning (DL) approach to achieve the high-resolution performance of localizing multiple sound sources tailored for omnidirectional acoustic applications. First, we present a spherical target map representation that can panoramically pinpoint the position and strength information of multiple sound sources without any grid-related constraints. Then, a dual-branched spherical convolutional autoencoder is proposed to obtain high-resolution localization results from the conventional spherical beamforming maps while incorporating frequency-variant and distortion-invariant strategies to address the inherent challenges. We quantitatively and qualitatively assess our proposed method's localization capability for multiple sound sources and validate that the proposed method can achieve far more precise and computationally efficient results than the existing approaches. By extension, we newly present the experimental setup that can create omnidirectional acoustic scenarios for the multiple SSL. By evaluating our proposed method in this experimental setup, we demonstrate the effectiveness and applicability of the proposed method with the experimental data. Our study delivers the proposed approach's potential of being utilized in various SSL applications.**

Index Terms—**Acoustic beamforming, deep learning (DL), multiple sound source localization (SSL), real-time acoustic measurement, spherical microphone array (SMA) system.**

I. Introduction

Sound source localization (SSL) is the process of detecting the locations of sound sources as well as identifying their signal strengths. Numerous efforts to localize where and how much sound comes from have progressed for a variety of possible acoustic applications, such as fault diagnosis via noise source localization [1]–[3], the SSL applications in robots and unmanned aerial vehicles [4]–[6], and speaker source localization [7], [8]. Besides, from industrial perspectives, its potential applications include identifying anomalous acoustic sources and their causes in the manufacturing process, monitoring gas or liquid leakage, and so on.

Among several SSL approaches, beamforming-based methods have been developed to localize the sound sources via source distribution maps [9], [10]. While beamforming-based methods, e.g., delay-and-sum (DAS), minimum variance (MV), and multiple single classification (MUSIC), are designed by spatially filtering signals from microphone array measurements, it is possible to visualize beam patterns in certain directions, providing interpretability and intuition for the SSL. However, it is widely known that the spatial resolution of conventional beamforming (CB) methods is considerably limited by the sound sources' frequency ranges: poor main-lobe resolution at low frequencies and spurious sidelobes at high frequencies [11], [12].

Several techniques have been suggested to improve the CB methods, where those techniques are generally categorized into two approaches: model-based and data-driven approaches. First, as the model-based approaches, deconvolution methods [13]–[21] have been mainly developed to enhance the CB map's spatial resolution and reduce unwanted sidelobes, based on iterative calculations using point spread functions (PSFs). Although the deconvolution methods have increased the CB map's resolution to a certain degree, these methods inherently

Manuscript received December 2, 2021; revised February 20, 2022; accepted March 10, 2022. Date of publication March 23, 2022; date of current version April 11, 2022. The work of Soo Young Lee and Seungchul Lee was supported in part by the National Research Foundation of Korea (NRF) Grant funded through the Korea Government (MSIT) under Grant 2020R1A2C1009744, in part by the Institute of Civil Military Technology Cooperation funded through the Defense Acquisition Program Administration and Ministry of Trade, Industry and Energy of the Korean government under Grant 19-CM-GU-01, and in part by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) Grant funded through the Korean Government [Ministry of Trade, Industry, and Energy (MOTIE)] under Grant 20206610100290. The work of Jiho Chang was supported by the National Research Foundation of Korea (NRF) Grant funded through the Korea Government Ministry of Science and ICT (MSIT) under Grant 21011045. The Associate Editor coordinating the review process was Jing Lei. *(Corresponding authors: Jiho Chang; Seungchul Lee.)*

Soo Young Lee is with the Industrial Artificial Intelligence Laboratory, Department of Mechanical Engineering, Pohang University of Science and Technology (POSTECH), Gyeongbuk 37673, Republic of Korea (e-mail: leesays92@postech.ac.kr).

Jiho Chang is with the Acoustics, Ultrasound, and Vibration Research Group, Division of Physical Metrology, Korea Research Institute of Standards and Science (KRISS), Daejeon 34113, Republic of Korea (e-mail: jiho.chang@kriss.re.kr).

Seungchul Lee is with the Industrial Artificial Intelligence Laboratory, Department of Mechanical Engineering, Pohang University of Science and Technology (POSTECH), Gyeongbuk 37673, Republic of Korea, also with the Graduate School of Artificial Intelligence, Pohang University of Science and Technology (POSTECH), Gyeongbuk 37673, Republic of Korea, and also with the Institute for Convergence Research and Education in Advanced Technology, Yonsei University, Seoul 03722, Republic of Korea (e-mail: seunglee@postech.ac.kr).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIM.2022.3161693, provided by the authors.

Digital Object Identifier 10.1109/TIM.2022.3161693

involve high computational costs and somewhat restricted spatial resolution. On the other hand, deep learning (DL) as a data-driven approach has recently garnered increasing attention for the beamforming-based SSL [22], and here, a few works have been conducted in the current literature [23]–[27]. Those works can be described under two groups, i.e., grid-based and grid-free methods. While the grid-based methods refer to estimating ground-truth positions and strengths of sound sources located on particular discretized grid points, the grid-free methods refer to predicting continuous ground-truth positions and strengths of acoustic sources. However, the main limitations of these methods are that: 1) sound sources should be discretized on certain grid points, where this grid type of results causes discretization errors for estimating actual locations of sound sources, and 2) the number of sound sources should be predefined in advance when constructing the neural networks and predicting results from them. A schematic that describes the limitations of the grid-based and the grid-free methods is shown in Fig. 1.

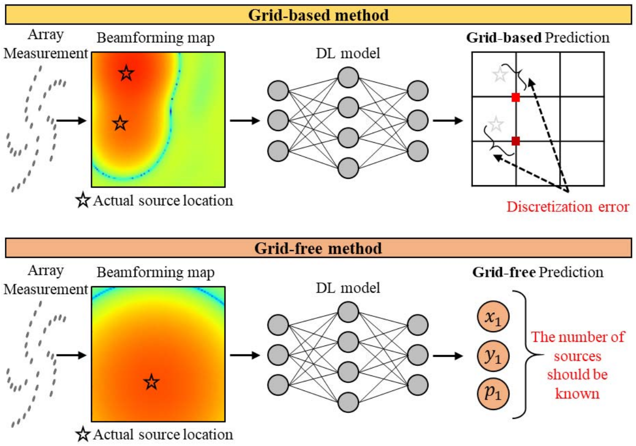

Fig. 1. Limitations of grid-based and grid-free methods for data-driven SSL.

In addition to the grid-related limitations mentioned above, developing the beamforming-based data-driven SSL method under a *spherical* microphone array (SMA) system introduces additional challenges as follows. While most of the relevant studies have been conducted using *planar* microphone array system, under this system, multiple microphones are placed on a two-dimensional measurement plane and sound sources are assumed to be within a half-infinite space. On the other hand, the SMA system has the advantage that the omnidirectional sound sources in a three-dimensional space can be detected by multiple microphones mounted on the rigid sphere surface (see Fig. 2). While the direction-of-arrival (DOA) distributions of sound sources in the SMA system can be represented using spherical beamforming (SB) maps, challenges arise in that the SB maps exhibit inconsistent resolution characteristics depending on the sources' frequency ranges. More specifically, the main issue in the low-frequency range is that the SB map's extremely large main-lobe area makes it difficult to resolve multiple adjacent sources well, whereas the key problem in the high-frequency range is the SB map's complicated and severe sidelobe patterns caused by the spatial aliasing effect [28] (Further details are introduced in Section II). In addition, the panoramic representation of this SB map, which contains the DOA information for multiple sound sources in azimuth and polar directions, inherently involves spatial distortions. As a result, these limitations of the SMA system on valid frequency range and spatial distortion raise the need to develop suitable methods for successfully localizing the sound sources in diverse omnidirectional acoustic scenarios.

Imbued by the challenges under grid-related constraints and the spherical measurement system, this article proposes a DL-based method for achieving high-resolution omnidirectional sound source localization (SSL), tailored for the SMA system. The main contributions of this article can be summarized as follows.

1) A novel spatial representation for the SMA system is proposed to overcome the existing limitations of grid-based and grid-free data-driven SSL approaches. The proposed representation, denoted as a spherical target map, is defined based on the spherical target function in order to spatially visualize positions and strengths for either on- or off-grid omnidirectional sound sources. While we transform the multiple SSL problem into the image-to-image prediction task from spherical beamforming (SB) maps to the proposed spherical target maps, this representation method is combined with the proposed neural networks to produce exact SSL results that are not limited to the grid after all.
2) A dual-branched spherical convolutional autoencoder network is proposed to address the challenges related to valid frequency ranges and the spatial distortion in the SMA system. More specifically, the proposed network is designed for the following purposes: 1) to effectively learn frequency-variant resolution characteristics between nonaliased SB maps at low frequencies and spatially aliased SB maps at high frequencies and 2) to learn distortion-invariant features from panoramic representations of the SB maps via spherical feature extraction operations.

Both qualitative and quantitative localization results are analyzed to evaluate the performance among our proposed method and existing comparative methods. We also demonstrate that the computational strength of the proposed method compared to the existing approaches, suggesting that it is more suitable for real-time SSL applications. Finally, the validation of our proposed method is expanded to the experimental setup that is specifically designed for various omnidirectional acoustic scenarios. Our findings indicate considerable SSL enhancement capability of our proposed method and its potential applicability for various omnidirectional SSL scenarios and applications.

We structure the remainder of this article as follows. In Section II, we convey the theoretical background on the spherical beamforming maps for the SSL and their conventional frequency limits. Section III describes the proposed spherical target map representation and the DL-based methods for high-resolution and fast source localization, followed by the evaluation strategies. Sections IV and V provide simulation and experimental results, respectively. Finally, Section VI concludes this article.

II. Background and Preliminaries

This section briefly introduces the fundamentals of spherical beamforming maps in the SMA system and conventional frequency limits to give readers background information. The relevant background provided in this section will serve as the basis for developing the proposed methods in the following sections.

A. Fundamentals of Spherical Beamforming Maps

This section concisely describes how the spherical beamforming (SB) maps are obtained from signal measurements in the SMA system. Suppose that the SMA is located at the origin as shown in Fig. 2, where Q microphones are mounted on the surface of the SMA. The position of the qth microphone $\vec{r}_{\text{mic}}^{(q)}$ is expressed with array radius a, polar angle $\theta_{\text{mic}}^{(q)}$, and azimuth angle $\phi_{\text{mic}}^{(q)}$. The position and the strength of the lth monopole sound source can be expressed as $\vec{r}_s^{(l)}$ and $q_s^{(l)}(\omega)$, respectively. An incident sound field generated by L monopole sources with the strength $q_s^{(l)}(\omega)$ can be expressed with the spherical harmonics as

$$P_{\text{inc}}(\vec{r}) = \sum_{l=1}^{L} q_s^{(l)}(\omega) \sum_{n=0}^{\infty} \sum_{m=-n}^{n} a_{nm}^{(l)} 4\pi i^n j_n(kr) Y_n^m(\theta,\phi) \tag{1}$$

where ω is the angular frequency (omitted in what follows) and k is the wavenumber $(=(\omega/c)$, c is speed of sound). While $Y_n^m(\theta,\phi)$ is the spherical harmonics [29], $a_{nm}^{(l)}$ denotes the weights as $a_{nm}^{(l)} = i^{1-n} k h_n^{(1)}(kr_s^{(l)}) Y_n^m(\theta_s^{(l)},\phi_s^{(l)})^*$. While $(\cdot)^*$ represents the conjugation, $h_n^{(1)}(\cdot)$ and $j_n(\cdot)$ are the nth order of the first kind spherical Hankel function and the spherical Bessel function, respectively.

Based on the scattering theory, the total field that consists of incident and scattered fields on the surface of SMA can be derived as

$$P_{\text{tot}}(a,\theta,\phi) = \sum_{l=1}^{L} q_s^{(l)} \sum_{n=0}^{\infty} \sum_{m=-n}^{n} a_{nm}^{(l)} b_n(ka) j_n(ka) Y_n^m(\theta,\phi) \tag{2}$$

where $b_n(ka) = 4\pi i^{n+1}/h_n'^{(1)}(ka)/(ka)^2$. Its weights in the spherical harmonics domain, $p_{mn}(ka) = \sum_{l=1}^{L} q_s^{(l)} a_{nm}^{(l)} b_n(ka)$, can be estimated with the microphone signals up to the Nth order

$$\mathbf{p}_{nm} = \mathbf{S} \cdot \mathbf{P}_{\text{mic}} \tag{3}$$

where

$$\begin{aligned} \mathbf{P}_{\text{mic}} &= \left[P_{\text{tot}}\left(a, \theta_{\text{mic}}^{(1)}, \phi_{\text{mic}}^{(1)}\right) \quad \cdots \quad P_{\text{tot}}\left(a, \theta_{\text{mic}}^{(Q)}, \phi_{\text{mic}}^{(Q)}\right) \right]^T \\ \mathbf{p}_{nm} &= \left[p_{00}^{(l)} \quad \cdots \quad p_{NN}^{(l)} \right]^T \end{aligned} \tag{4}$$

and $\mathbf{S}$ is a matrix for the discrete spherical Fourier transform [28]. In this study, $\mathbf{S}$ is defined as the pseudoinverse of a matrix $\mathbf{Y}$, where its each component is $Y_n^m(\theta_{\text{mic}}^{(q)}, \phi_{\text{mic}}^{(q)})$. The spherical beamformer output can be expressed as

$$b\left(\vec{r}_{\text{grid}}\right) = \mathbf{w}_{nm}^H\left(\vec{r}_{\text{grid}}\right) \mathbf{p}_{nm} \tag{5}$$

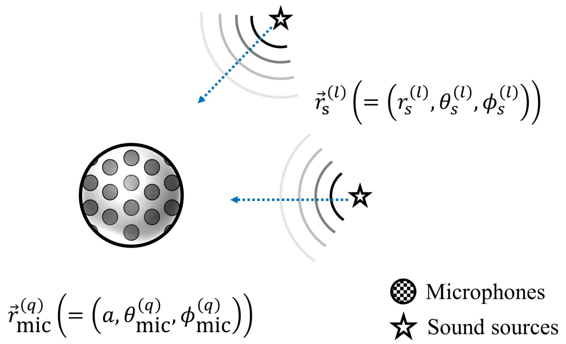
$\vec{r}_s^{(l)}\left(=\left(r_s^{(l)},\theta_s^{(l)},\phi_s^{(l)}\right)\right)$
$\vec{r}_{\text{mic}}^{(q)}\left(=\left(a,\theta_{\text{mic}}^{(q)},\phi_{\text{mic}}^{(q)}\right)\right)$
Microphones
Sound sources

Fig. 2. Multiple sound sources and SMA system.

where $\mathbf{w}_{nm}^H(\vec{r}_{\text{grid}})$ is a vector of the beamforming weights for an assumed position of a source $\vec{r}_{\text{grid}}$

$$\mathbf{w}_{nm}\left(\vec{r}_{\text{grid}}\right) = \left[w_{00}\left(\vec{r}_{\text{grid}}\right) \quad \cdots \quad w_{NN}\left(\vec{r}_{\text{grid}}\right) \right]^T. \tag{6}$$

Each component can be written as

$$w_{nm}^*\left(\vec{r}_{\text{grid}}\right) = \frac{d_n}{b_n(ka)} Y_n^m\left(\theta_{\text{grid}}, \phi_{\text{grid}}\right) \tag{7}$$

and the maximum directivity beamformer is known to have a constant value of $d_n = 4\pi/(N+1)^2$ [28]. The beamforming power $B(\vec{r}_{\text{grid}})$ can be defined as

$$B\left(\vec{r}_{\text{grid}}\right) = b\left(\vec{r}_{\text{grid}}\right) b\left(\vec{r}_{\text{grid}}\right)^H = \mathbf{w}_{nm}^H\left(\vec{r}_{\text{grid}}\right) \mathbf{p}_{nm} \mathbf{p}_{nm}^H \mathbf{w}_{nm}\left(\vec{r}_{\text{grid}}\right). \tag{8}$$

Substituting (3) into (8) leads to

$$B\left(\vec{r}_{\text{grid}}\right) = \mathbf{w}_{nm}^H\left(\vec{r}_{\text{grid}}\right) \mathbf{S} \cdot \mathbf{P}_{\text{mic}} \mathbf{P}_{\text{mic}}^H \cdot \mathbf{S}^H \mathbf{w}_{nm}\left(\vec{r}_{\text{grid}}\right). \tag{9}$$

If the beamforming power is averaged, using the cross-spectrum matrix $\mathbf{C}$ that is composed of the mean values of $\mathbf{P}_{\text{mic}} \mathbf{P}_{\text{mic}}^H$, the beamforming power is

$$B\left(\vec{r}_{\text{grid}}\right) = \mathbf{w}_{nm}^H\left(\vec{r}_{\text{grid}}\right) \mathbf{S} \cdot \mathbf{C} \cdot \mathbf{S}^H \mathbf{w}_{nm}\left(\vec{r}_{\text{grid}}\right). \tag{10}$$

Sources are assumed to be incoherent in this study, and the cross correlation terms of the matrix $\mathbf{C}$ vanish. In this wise, the beamforming power at all grid points $\vec{\mathbf{r}}_{\text{grid}}$ is composed of a spherical beamforming map $\mathbf{B}(\vec{\mathbf{r}}_{\text{grid}})$. An example of the spherical beamforming map is visualized in Fig. 3 (left).

B. Existing Frequency-Related Limitations

It is known that the valid frequency range is limited due to the radius a (for low frequency) and the microphone spacing (for high frequency). First, since the surface pressure has the spherical harmonics components up to $[ka]$ (nearest integer of ka), (5) leads to a large error of $\mathbf{p}_{nm}$ at low frequencies where $[ka]$ is less than N. To avoid this error, the high-order components of p_{nm} can be truncated by limiting N. In this study, N is determined as $[ka]+1$ if $[ka]+1 < N$. Despite this truncation, the multiple sources' main lobes within the SB maps are so wide that we often find poor-resolution SB maps with an overlapped main-lobe patterns. On the other hand, at high frequencies, the surface pressure has considerably higher order components than N that the given microphone spacing cannot acquire. In this case, when the number of spherical harmonics components is larger than that of the microphones, there occurs the high-frequency limit, i.e., spatial Nyquist frequency, due to the spatial aliasing effect [28]. Here, the SB maps not only have complex beam power distribution, but they can contain ghost sources in the spurious areas at which no sources are actually located. As a result, these limitations at low and high frequencies make it challenging to localize the sound sources precisely, which is one of the motivations of this study.

III. Proposed Method

This section mainly describes the proposed spherical target map that can spatially represent positional and strength information of both on- or off-grid omnidirectional sound sources, the proposed deep neural networks for target map prediction and its retrieval process to extract grid-free localization results, as well as model evaluation strategies.

A. Spherical Target Map Representation

We propose a spatial representation (denoted as a spherical target map $\mathbf{B}_T$) that can panoramically visualize ground-truth positions and strengths of multiple acoustic sources with no grid constraints. The spherical target map $\mathbf{B}_T$ is mainly represented with extremely high-resolution and clean representation, i.e., sharp main lobes and no sidelobes, where it serves as a ground-truth map that contains both position and strength information of either on- or off-grid omnidirectional sound sources. The proposed spherical target map representation can be utilized in various omnidirectional measurement scenarios, where it is first defined based on the spherical target function ξ as

$$\xi\left(\vec{r}_{\text{grid}}, \vec{r}_s^{(l)}\right) = \frac{\epsilon}{\alpha \cdot R_{\text{geod}}^{\nu} + \epsilon} \tag{11}$$

where R_{geod} represents the geodesic distance of the shortest arc length from the actual source location $\vec{r}_s^{(l)}$ to each grid point $\vec{r}_{\text{grid}}$ of the spherical source plane. R_{geod} can be expressed as

$$R_{\text{geod}} = \left|\vec{r}_{\text{sph}}\right| \cos^{-1}\left(\frac{\vec{r}_{\text{grid}} \cdot \vec{r}_s^{(l)}}{\left|\vec{r}_{\text{grid}}\right|\left|\vec{r}_s^{(l)}\right|}\right) \tag{12}$$

where the radius of the spherical source plane $\left|\vec{r}_{\text{sph}}\right|$ is set to be 2 m in this study. We determine values of ϵ, α, and ν such that ξ decreases 20, 40, and 60 dB (decline value of 20 dB) for a half, one, and two grid distances, respectively. More detailed discussion on the spherical target function and the decline value can be found in Supplementary Information and Supplementary Fig. 1. Note that each grid distance is set to be a 5° difference in this study, which leads the $\mathbf{B}_T$'s dimension to 37 grid points (0°–180°) for polar angle θ and 72 grid points (0°–355°) for azimuth angle ϕ. Based on the ξ described above, the spherical target map $\mathbf{B}_T$ can be calculated as

$$\mathbf{B}_T\left(\vec{\mathbf{r}}_{\text{grid}}\right) = \sum_{l=1}^{L} \xi\left(\vec{\mathbf{r}}_{\text{grid}}, \vec{r}_s^{(l)}\right) \cdot q_{s,r}{}^{(l)\,2} \tag{13}$$

where $q_{s,r}^{(l)}$ represents the distance-normalized strength of the lth sound source ($|q_s^{(l)}|/r_s^{(l)}$). As a result, using this proposed representation, the locations and strengths of both on-the-grid and out-of-grid acoustic sources can be mapped within the spherical target map in a pixel-level manner.

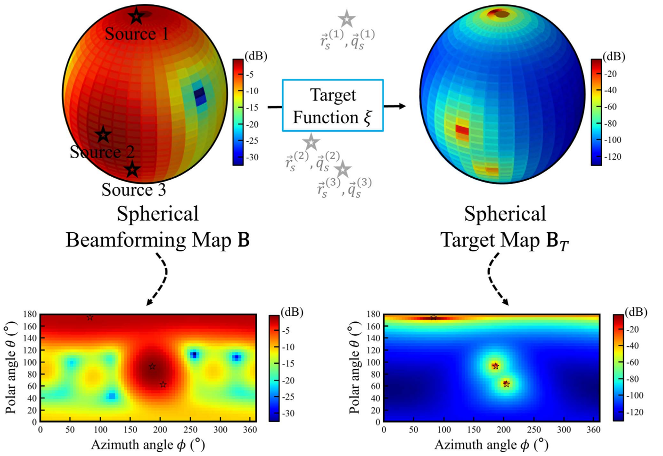

Fig. 3. Example ($f = 1$ kHz, $L = 3$) of the spherical beamforming map and the proposed spherical target map representation.

Fig. 3 shows an example of the spherical beamforming map $\mathbf{B}$ and the spherical target map $\mathbf{B}_T$ for $f = 1$ kHz, $L = 3$ case. While the actual locations of three sound sources ($L = 3$) that are not located on the grid points are marked as black stars ($\star$), it is observed that $\mathbf{B}$ shows unclear source distribution maps with broad main-lobe patterns at $f = 1$ kHz. In contrast, the $\mathbf{B}_T$ can distinctively represent multiple sound sources' locations and corresponding strength information in the form of a high-resolution and clean source distribution map with sharp main-lobe and no sidelobe patterns. It is worth mentioning that the pixel values of $\mathbf{B}_T$ represent the sound sources' strength information based on the distances from the sources' positions to the pixel grids, and hence, both position and strength information of multiple sound sources can be spatially visualized within $\mathbf{B}_T$. Note that the equirectangular projection is considered for $\mathbf{B}$ and $\mathbf{B}_T$ in this study for evaluating the proposed method with the existing approaches. More diverse examples of $\mathbf{B}$ and $\mathbf{B}_T$ depending on the frequency f and the number of sound sources L will be further introduced in Section IV-A.

B. Dual-Branched Spherical Convolutional Autoencoder

While the goal of the SSL can be formulated as the image-level prediction of the spherical target maps $\mathbf{B}_T$s from the spherical beamforming maps $\mathbf{B}$s, we propose a dual-branched fully convolutional autoencoder network to predict them precisely. The architecture of the proposed model is shown in Fig. 4. First, from the conventional frequency-related limitations in Section II-B, we hypothesize that the different feature learning approaches are required for two groups of the $\mathbf{B}$s distinctively divided into low- and high-frequency ranges according to the spatial aliasing limit. This frequency-variant characteristics of the $\mathbf{B}$s generate two encoding paths, i.e., encoder network for the nonaliased E_N and encoder network for the spatially aliased E_S. Given two-dimensional input beamforming map $\mathbf{B}$, hidden layers of the proposed model

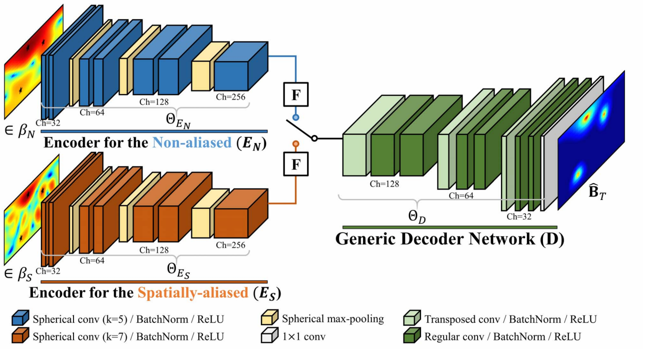
B ∈ β_N
Ch=32
Ch=64
Ch=128
Ch=256
Θ_{E_N}
Encoder for the Non-aliased (E_N)
F
F
B ∈ β_S
Ch=32
Ch=64
Ch=128
Ch=256
Θ_{E_S}
Encoder for the Spatially-aliased (E_S)
Ch=128
Ch=64
Ch=32
Θ_D
$\hat{\mathbf{B}}_T$
Generic Decoder Network (D)
Spherical conv (k=5) / BatchNorm / ReLU
Spherical conv (k=7) / BatchNorm / ReLU
Spherical max-pooling
1×1 conv
Transposed conv / BatchNorm / ReLU
Regular conv / BatchNorm / ReLU

Fig. 4. Architectural description of the proposed dual-branched spherical convolutional autoencoder network.

mainly consist of successive 2-D fully convolutional layers as

$$F_j^{(k)} = \varphi\left(\sum_i F_i^{(k-1)} \otimes \theta_w^{(k)} + \theta_b^{(k)}\right) \tag{14}$$

where $F_j^{(k)}$ is the jth feature map in the kth convolutional layer, $F_i^{(k-1)}$ is the ith feature map in the former layer, $\theta_w^{(k)}$ and $\theta_b^{(k)}$ represent trainable parameters of weights and biases in the kth convolutional layer, respectively, and $\otimes$ denotes the 2-D convolution operation via sliding window of the trainable filter $\Theta_w^{(k)}$. φ is a nonlinear activation function, where a rectified linear unit (ReLU) $\varphi(\cdot) = \max(0, \cdot)$ is utilized in this study. Batch normalization [30] is appended to each convolutional layer for increasing the stability of the learning process [31], while max pooling is used to gradually reduce the dimension of the convolved feature maps. The latent feature maps $\mathbf{F}$ extracted from the encoder networks can be expressed as

$$\mathbf{F} = \begin{cases} E_N\left(\mathbf{B}; \Theta_{E_N}\right), & \text{if } \mathbf{B} \in \beta_N \\ E_S\left(\mathbf{B}; \Theta_{E_S}\right), & \text{if } \mathbf{B} \in \beta_S \end{cases} \tag{15}$$

where β_N and β_S are two sets of $\mathbf{B}$s divided based on the spatial aliasing limit, which is set to be 5.16 kHz in this study (further details are introduced in Section IV-A). Θ_{E_N} and Θ_{E_S} represent collections of weights and biases in E_N and E_S, respectively. Subsequently, in the latter part of the model, high-level hierarchical feature maps $\mathbf{F}$ extracted from the encoder networks propagate to a generic decoder network D for obtaining predicted spherical target map $\hat{\mathbf{B}}_T$

$$\hat{\mathbf{B}}_T = D(\mathbf{F}; \Theta_D). \tag{16}$$

By alternately training the parameters of the two combinations of the autoencoder networks (E_N-D and E_S-D), the proposed model learns to predict high-resolution and accurate $\hat{\mathbf{B}}_T$s from low to high frequencies with or without spatial aliasing effect. It should be noted that $\hat{\mathbf{B}}_T$ is predicted by selectively utilizing one of the encoder paths corresponding to the input map for a given frequency, rather than by feeding several input maps from both groups at once.

In particular, we take advantage of spherical convolution and pooling operations to extract distortion-invariant spherical features from the panoramic form of $\mathbf{B}$s (see Fig. 5). The spherical operations are designed to apply the adaptive kernel sampling technique for the standard convolutional filters via the inverse gnomonic projection [32]. Since the equirectangular projection for the $\mathbf{B}$s inevitably introduces severe spatial distortions, particularly in the polar regions, we exploit the spherical convolutions and poolings as the main operations in the encoder networks to capture $\mathbf{B}$s' spatial characteristics without distortions. Besides, we empirically employ different kernel sizes of the spherical convolutions for two different encoder networks, where 7 × 7 kernels are assigned to E_S for recognizing diffused sidelobe patterns through larger receptive fields.

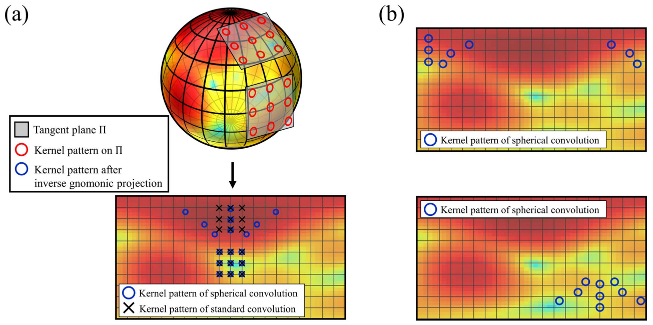
(a)
(b)
Tangent plane Π
Kernel pattern on Π
Kernel pattern after inverse gnomonic projection
Kernel pattern of spherical convolution
Kernel pattern of standard convolution
Kernel pattern of spherical convolution
Kernel pattern of spherical convolution

Fig. 5. Adaptive kernel sampling patterns of spherical operation: (a) comparison in terms of the selected kernel pattern between spherical convolution and standard convolution and (b) examples showing the adaptive kernel patterns of spherical convolutional for varying spatial regions.

Finally, several training strategies are employed to train our proposed network. In terms of the loss function, we consider a multipart loss function combined with the pixel-wise mean absolute error (MAE) loss (J_{MAE}) and the structural similarity (SSIM) index loss (J_{SSIM}) [33], [34]. While J_{MAE} plays a role in describing each paired pixels' distance between actual spherical target map $\mathbf{B}_T$ and predicted spherical target map $\hat{\mathbf{B}}_T$, J_{SSIM} can be used to measure spatial and structural closeness between those two maps. First, J_{MAE} can be expressed as

$$J_{\text{MAE}}\left(\mathbf{B}_T, \hat{\mathbf{B}}_T\right) = \frac{1}{U}\sum_u \left|\mathbf{B}_T\left(r_{\text{grid}}^{(u)}\right) - \hat{\mathbf{B}}_T\left(r_{\text{grid}}^{(u)}\right)\right| \tag{17}$$

where $\mathbf{B}_T(r_{\text{grid}}^{(u)})$ and $\hat{\mathbf{B}}_T(r_{\text{grid}}^{(u)})$ are the uth pixel values in the actual and predicted spherical target map, respectively, and U denotes the total number of the pixels within the target map. Besides, J_{SSIM} can be expressed as [33], [34]

$$J_{\text{SSIM}}\left(\mathbf{B}_T, \hat{\mathbf{B}}_T\right) = \frac{1}{X}\sum_x \frac{\left(2\mu_{\mathbf{B}_T^{(x)}}\mu_{\hat{\mathbf{B}}_T^{(x)}}\right)\left(2\sigma_{\mathbf{B}_T^{(x)},\hat{\mathbf{B}}_T^{(x)}}\right)}{\left(\mu^2_{\mathbf{B}_T^{(x)}} + \mu^2_{\hat{\mathbf{B}}_T^{(x)}}\right)\left(\sigma^2_{\mathbf{B}_T^{(x)}} + \sigma^2_{\hat{\mathbf{B}}_T^{(x)}}\right)} \tag{18}$$

where $\mathbf{B}_T^{(x)}$ and $\hat{\mathbf{B}}_T^{(x)}$ are the xth 11 × 11 sliding window patches from $\mathbf{B}_T$ and $\hat{\mathbf{B}}_T$, respectively, μ and σ^2 represent local averages and local variances within $\mathbf{B}_T^{(x)}$ and $\hat{\mathbf{B}}_T^{(x)}$, respectively, and $\sigma_{\mathbf{B}_T^{(x)},\hat{\mathbf{B}}_T^{(x)}}$ denotes the local covariance value of them. In (14), the SSIM loss value J_{SSIM} can be obtained by averaging the local SSIM values computed from the entire windows X. Finally, the combined loss function J_c can be defined as

$$J_c\left(\mathbf{B}_T, \hat{\mathbf{B}}_T\right) = \lambda \cdot J_{\text{MAE}} + (1-\lambda)\cdot(1 - J_{\text{SSIM}}) \tag{19}$$

where λ denotes a parameter that controls significance between J_{MAE} and J_{SSIM}, while λ of 0.85 is adopted in this study. The best value of λ is manually searched by evaluating the model's validation results while adjusting the parameter value from 0 to 1. The influence of λ on the model's validation results is visualized in Supplementary Fig. 2. While it is found that the proposed model's validation loss decreases to some extent as λ increases, its performance is observed to rather deteriorate when λ becomes too large. In terms of training process of the proposed model, trainable parameters of the encoder networks and the decoder network are alternately updated based on a mini-batch gradient descent algorithm in a way that J_c is minimized. Model training and regularization settings are obtained from the random search. While we utilize AdamOptimizer with a batch size of 60, we assign the learning rate of $1\mathrm{e}^{-4}$, which can be reduced via plateaus of the validation loss. Besides, L2 regularization with its coefficient of $2\mathrm{e}^{-6}$ and early stopping strategy (patience value of 10) during entire 200 training epochs is considered for the regularization settings. The model is implemented using Python 3.7.8 and PyTorch 1.9.0 Framework and trained using a NVIDIA GeForce 2080 Ti. It should be noted that the training and validation processes of the proposed dual-branched framework converge stably, whereas the single-branched model induces a severe overfitting problem, as visualized in Supplementary Fig. 3.

C. Deep Neural Network-Based Retrieval Process

We also suggest a fast retrieval process from the predicted target map $\hat{\mathbf{B}}_T$ based on a convolutional neural network (CNN)-based regression model. Once the proposed model in Section III-B is trained to predict the spherical target map in a spatial manner, the retrieval process is required to extract multiple sources' exact locations and strengths from the two-dimensional representation of $\hat{\mathbf{B}}_T$ so that the final localization results of the proposed model are not limited to the grid points. Fig. 6 shows the detailed descriptions for the proposed retrieval method. The first step of the retrieval process begins with finding several local maxima in $\hat{\mathbf{B}}_T$, where three different local peaks are detected in the example shown in Fig. 6. Second, we select a 7 × 7 subregion around each local maximum via the adaptive pattern sampling mentioned earlier. Here, a set of three patches is extracted from $\hat{\mathbf{B}}_T$ to describe each source's distributional pattern in terms of position $\hat{\theta}$, $\hat{\phi}$, and strength $\hat{B}_T$. Finally, each set of patches is fed into a pretrained CNN-based retrieval model. The retrieval network structure consists of two types of hidden layers, i.e., convolutional layers and fully connected (FC) layers. To begin, the convolutional layers are used to extract high-dimensional feature maps by sequential 2-D convolution processes, as previously described in (13). The following step is to flatten convolved feature maps and propagate them into a series of FC layers, where linear operations in FC layers are described as follows:

$$n_j^{(l)} = \varphi\left(\sum_i n_i^{(l-1)}\theta_w^{(l)} + \theta_b^{(l)}\right) \tag{20}$$

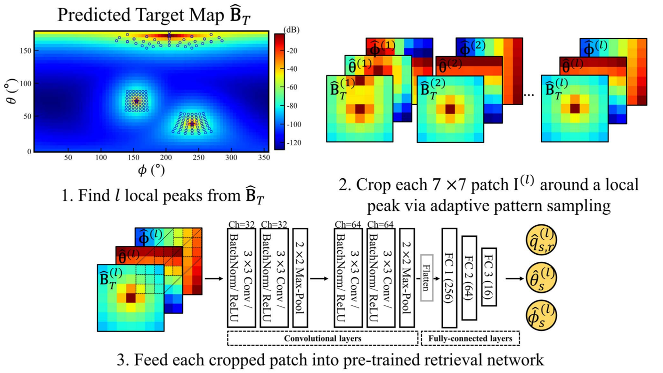
Predicted Target Map $\hat{\mathbf{B}}_T$
(dB)
-20
-40
-60
-80
-100
-120
150
100
50
0
θ (°)
50
100
150
200
250
300
350
ϕ (°)
1. Find l local peaks from $\hat{\mathbf{B}}_T$
$\hat{\phi}^{(1)}$
$\hat{\theta}^{(1)}$
$\hat{\mathbf{B}}_T^{(1)}$
$\hat{\phi}^{(2)}$
$\hat{\theta}^{(2)}$
$\hat{\mathbf{B}}_T^{(2)}$
...
$\hat{\phi}^{(l)}$
$\hat{\theta}^{(l)}$
$\hat{\mathbf{B}}_T^{(l)}$
2. Crop each 7 ×7 patch $\mathrm{I}^{(l)}$ around a local peak via adaptive pattern sampling
Ch=32
3 ×3 Conv / BatchNorm/ ReLU
Ch=32
3 ×3 Conv / BatchNorm/ ReLU
2 ×2 Max-Pool
Ch=64
3 ×3 Conv / BatchNorm/ ReLU
Ch=64
3 ×3 Conv / BatchNorm/ ReLU
2 ×2 Max-Pool
Flatten
FC 1 (256)
FC 2 (64)
FC 3 (16)
$\hat{q}_{s,r}^{(l)}$
$\hat{\theta}_s^{(l)}$
$\hat{\phi}_s^{(l)}$
Convolutional layers
Fully-connected layers
3. Feed each cropped patch into pre-trained retrieval network

Fig. 6. Step-by-step details of the deep neural network-based retrieval process.

where $n_j^{(l)}$ is the jth output neuron in the lth FC layer, $n_i^{(l-1)}$ is the ith input neuron in the previous FC layer, and $\theta_w^{(l)}$ and $\theta_b^{(l)}$ denote weights and biases in the lth FC layer, respectively.

Regarding the pretraining process, we assume that the network is optimized when the R^2 score for the validation set reaches above 0.99. From our manual search processes of the retrieval network design, it is observed that choosing the proper depth of FC layers affects achieving the target validation R^2 score of higher than 0.99. Average validation R^2 scores according to several different FC layer depths are summarized in Supplementary Table I. As for utilized dataset, randomly generated 50 000 samples and 10 000 samples are utilized as the training set and the validation set, respectively. The retrieval network is trained using a mini-batch gradient descent algorithm with a mean squared error (MSE) loss, utilizing AdamOptimizer with a batch size of 80. We consider a learning rate of $1\mathrm{e}^{-5}$ and L2 regularization coefficient of $1\mathrm{e}^{-6}$ during 200 training epochs. The model is implemented using Python 3.7.8 and PyTorch 1.9.0 Framework and trained using a NVIDIA GeForce 2080 Ti. As a result, the proposed retrieval method achieves a fast inference process that can be more suitable for real-time applications while rapidly extracting exact positional and strength information from $\hat{\mathbf{B}}_T$ so that the estimated results are not restricted to the grid after all. To summarize, our proposed approach for achieving high-resolution omnidirectional SSL can be visualized in Supplementary Fig. 4.

D. Evaluation Strategies

We first define *recall* and *precision* metrics to quantitatively compare the localization results of the proposed method with those of the existing methods [13]–[18]. A schematic and several examples are shown in Fig. 7. First, we define the *recall* that indicates the ratio of the number of the detected true sources (denoted as well-predicted sources **W**) to the total number of true sources (the sum of the number of undetected sources **U** and that of **W**). We also define the *precision* metric that is the ratio of the number of the well-predicted sources **W** to that of whole detections (the sum of the number of spurious ghost sources **G** and that of **W**). From the definitions above, the recall can be used to assess the model's resolvability among the multiple sources mainly at the low-frequency range, while the precision can measure the model's ghost source robustness primarily at the high-frequency range. Detailed procedures are given as follows.

1) *Whole Detections* (**W**+**G**)*:* After detecting all local maxima from the predicted map of a certain model, we select the local maxima as the model's whole detections if their absolute differences to the global maximum value in the predicted map are within 15 dB.
2) *Well-Predicted Sources* (**W**)*:* We compare the positions of the whole detections and those of the ground-truth sources and define the detections whose angular distances compared to the actual positions show no more than 10° as the well-predicted sources of the model.
3) The number of undetected sources (**U**) can be obtained by subtracting the number of the model's well-predicted sources from the total number of true sources.

In addition, we evaluate the absolute differences between the ground-truth and the predicted values in terms of angular distance error and strength error, for well-predicted sources **W** of the comparative methods. It is inevitable to assess angular distance error and strength error values only for **W** since the number of **W** of each model can be different from that of true sources ($\mathbf{U} + \mathbf{W}$). For this reason, we comprehensively evaluate the performance of the comparative methods with two groups of metrics (precision–recall and angular distance error-strength error); the former group measures the ratios between omitted, detected, and spurious sources, and then, the latter one evaluates how precise the methods estimate the positions and strengths as to the well-detected sources **W**.

IV. Synthetic Experimental Results

A. Simulation Setup

In this study, we first evaluate the localization performance of our proposed method via simulation experiments. Regarding the SMA system, we consider a 4th-order higher order ambisonics (HOA) microphone array with 32 elements mounted on a rigid sphere of 4.2 cm radius, which is in line with the em32 Eigenmike. As the spatial Nyquist frequency of the system is determined to be 5.16 kHz by its maximum order and the radius, we consider each of the three frequencies for two groups categorized by the spatial aliasing limit, i.e., 1, 2, 4 kHz for the nonaliased group and 6, 8, and 10 kHz for the spatially aliased group. Several examples of **B**s and $\mathbf{B}_T$s according to different frequencies are visualized in Fig. 8. At 1 kHz, lower directivity is expected because ka is less than 1, while directivity increases as the frequency increases. On the other hand, the spatial aliasing effect occurs in 6 kHz and becomes severe at higher frequencies.

The data generation process is considered as follows. While up to three monopole sound sources are considered, the distance from the origin to each source is also randomly chosen between 1.9 and 2.1 m. Each source's location, i.e., θ and ϕ, is also arbitrarily selected. Besides, a random sampling range of each source's strength is determined based on a maximum difference of 6 dB for the distance-normalized strength $q_{s,r}$, which leads to a range between 1 and 1.8095. It is worth mentioning that we consider noise-perturbed data to reflect possible errors of practical situations in our simulation study. While ±3-m/s speed of sound error is employed, we consider the radius and the angular errors of 1 mm and 1° concerning the microphone positioning errors, respectively. The signal-to-noise ratio (SNR) is set to be 30 dB in our experiments. As a result, we generate 28 500 training samples, 1500 validation samples, and 1200 test samples for each frequency with up to three sound sources, which entirely makes 171 000 samples for the training set, 9000 samples for the validation set, and 7200 samples for the test set, respectively. Note that the same amount of data is used for each of the number of sound sources and frequency. Detailed information for the train set, validation set, and test set is summarized in Supplementary Table II.

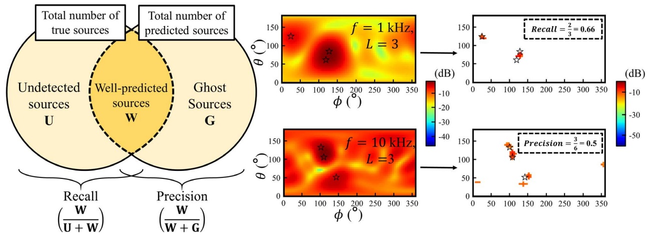
Total number of true sources
Total number of predicted sources
Undetected sources U
Well-predicted sources W
Ghost Sources G
Recall
$\left(\frac{\mathbf{W}}{\mathbf{U}+\mathbf{W}}\right)$
Precision
$\left(\frac{\mathbf{W}}{\mathbf{W}+\mathbf{G}}\right)$
f = 1 kHz, L = 3
f = 10 kHz, L =3
θ (°)
ϕ (°)
(dB)
-10
-20
-30
-40
-50
$Recall = \frac{2}{3} = 0.66$
$Precision = \frac{3}{6} = 0.5$
0
50
100
150
200
250
300
350

Fig. 7. Schematic diagram for recall and precision metrics and several illustrating examples.

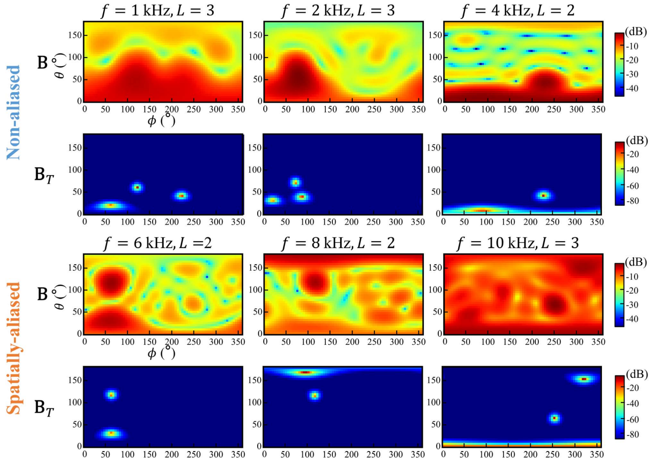

Fig. 8. Examples of the simulated data in our experiments. While odd rows represent the input spherical beamforming maps **B**s according to different frequency f, even rows indicate their corresponding spherical target maps $\mathbf{B}_T$s.

B. Qualitative Analysis

First, we evaluate the SSL performance of our proposed method with qualitative results, where the predicted spherical target maps $\widehat{\mathbf{B}}_T$s visually represent how well the method resolves the multiple sound sources' distribution maps from spherical beamforming maps **B**s. The proposed method is compared with several existing methods of the widely recognized deconvolution approaches, i.e., CLEAN [13], [14], deconvolution approach for the mapping of sound sources (DAMAS) [15], [16], and nonnegative least

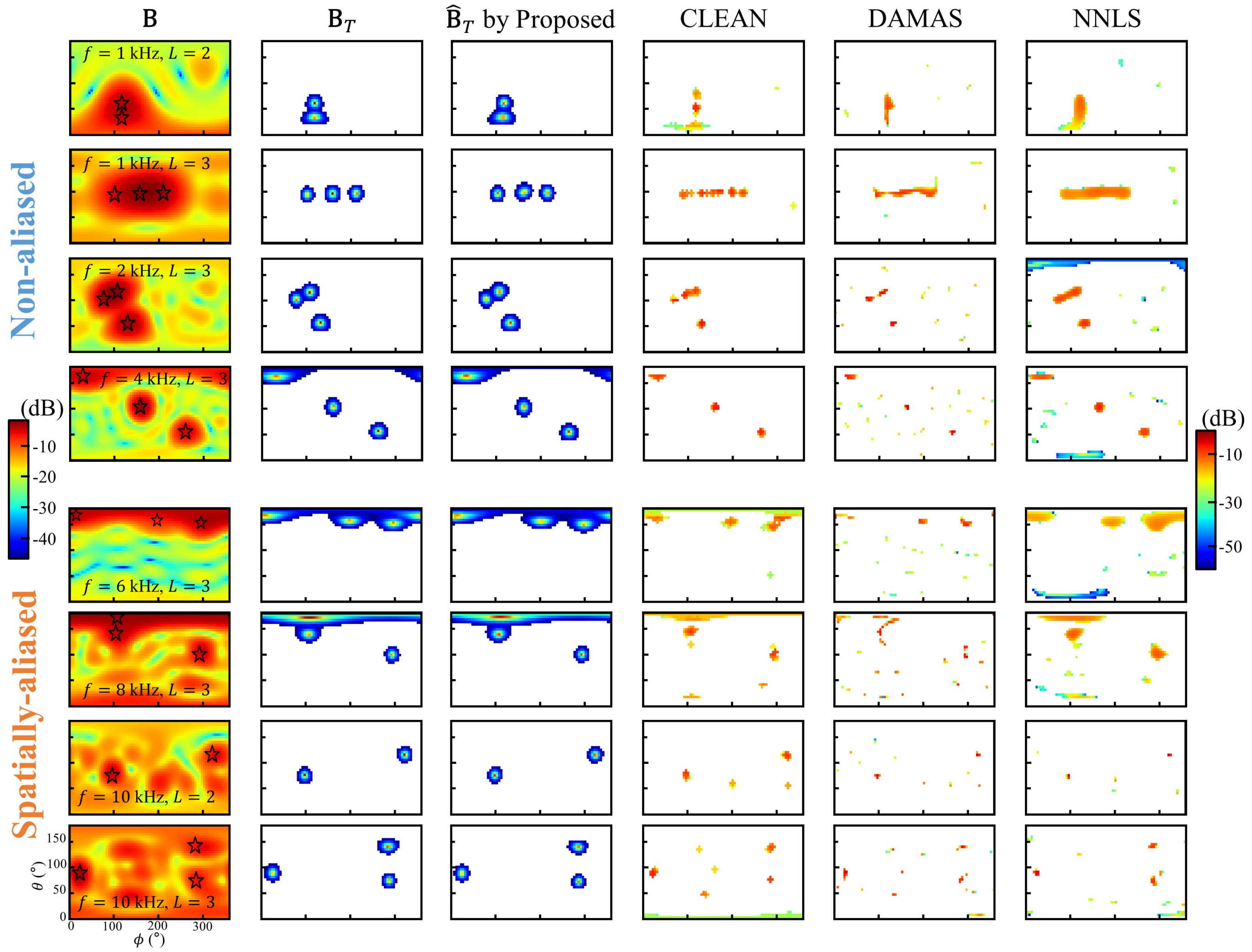

Fig. 9. Qualitative examples of the sound source distribution maps predicted by the proposed method and the existing approaches. Multiple sound sources' ground-truth locations are indicated by black stars (⋆) within spherical beamforming maps **B**s.

squares (NNLS) [17], [18]. These approaches have been developed to improve the spatial resolution of the CB map, where their methodologies are based on deconvolution using the PSF [9]. In this study, the PSF was calculated as the sound pressure on the rigid sphere surface by a sound source with unit strength. Note that shift-invariant property does not hold, and the PSF for all grid points was obtained. Although more recent deconvolution methods have been studied, we chose these widely recognized methods for comparison with the proposed method since these methods have been mainly used for comparison with other methods, as well as the recent methods have additional assumptions on sparsity or correlation. In terms of parameter settings of the comparison methods, the beamwidth of CLEAN is set to be 5°, given that 5° difference for each grid distance of $\mathbf{B}$ and $\widehat{\mathbf{B}}_T$ is considered in this study. Besides, we empirically selected the safety factor of 0.1 for CLEAN that generally showed the lowest deconvolution errors of CLEAN. It should be noted that the considered parameter values of CLEAN are in line with the previous studies [21], [35]. As for the number of iterations for CLEAN, DAMAS, and NNLS, we evaluate the source distribution maps estimated by those deconvolution methods after 300 iterations, where the deconvolution errors of the three comparison models converged. Examples of deconvolution errors with respect to the parameter settings can be found in Supplementary Fig. 5.

Fig. 9 shows the qualitative examples among the comparative methods for estimating the multiple sound sources' positions and strengths at several low and high frequencies. As previously stated, at the low frequencies (the nonaliased group), much attention is drawn to examine how well the certain model decomposes the multiple sources from their wide and overlapped main lobes, while it is important to assess how well one can achieve spurious source robustness when the spatial aliasing effect occurs at the high frequencies (the spatially aliased group). Fig. 9 (column 1) shows the spherical beamforming maps **B**s according to different frequency f and the number of sound sources L, where the ground-truth locations of multiple sound sources are indicated by black

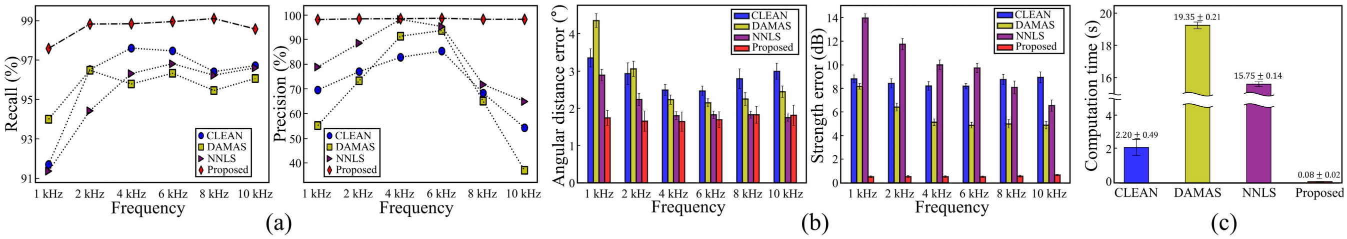

Fig. 10. Quantitative comparison of the SSL performance: (a) recall and precision plots according to the frequencies, (b) bar charts showing angular distance and strength errors according to the frequencies, and (c) computation time comparison.

TABLE I
DETAILED QUANTITATIVE RESULTS FOR FIG. 10

Methods	Metrics	Frequency (f) 1 kHz	2 kHz	4 kHz	6 kHz	8 kHz	10 kHz	Methods	Metrics	Frequency (f) 1 kHz	2 kHz	4 kHz	6 kHz	8 kHz	10 kHz
CLEAN	Recall (%)	91.7±1.7	96.5±1.7	97.6±1.3	97.4±1.1	96.4±1.8	96.7±1.4	DAMAS	Recall (%)	94.0±1.9	96.4±1.8	95.7±2.2	96.3±1.7	95.4±2.6	96.0±2.1
	Precision (%)	69.4±3.7	76.9±4.4	82.7±3.1	85.3±2.4	68.2±4.3	54.1±2.4		Precision (%)	55.0±3.3	73.2±4.2	91.3±2.4	93.6±2.2	64.9±4.4	36.8±3.2
	Angular distance error (°)	3.36±0.03	2.94±0.30	2.49±0.15	2.47±0.13	2.80±0.27	2.99±0.21		Angular distance error (°)	4.37±0.19	3.06±0.21	2.23±0.13	2.15±0.11	2.25±0.17	2.45±0.15
	Strength error (dB)	8.81±0.32	8.42±0.40	8.20±0.37	8.18±0.23	8.75±0.43	8.93±0.46		Strength error (dB)	8.17±0.24	6.41±0.33	5.12±0.28	4.88±0.24	4.99±0.35	4.88±0.31
	Computation time (s)	2.20±0.49							Computation time (s)	19.35±0.21					
NNLS	Recall (%)	91.3±1.9	94.4±2.5	96.3±2.1	96.8±1.6	96.2±2.1	96.6±2.0	**Proposed**	Recall (%)	97.6±0.7	98.8±0.9	98.8±0.6	98.9±0.3	99.1±0.4	98.6±0.8
	Precision (%)	78.8+2.9	88.5±3.0	98.2±1.4	95.3±1.7	71.6±3.4	64.8±3.6		Precision (%)	98.1±1.5	94.8±0.9	98.5±1.0	98.6±0.9	98.2±1.0	98.2±1.1
	Angular distance error (°)	2.89±0.15	2.24±0.16	1.81±0.11	1.83±0.10	1.83±0.10	1.75±0.10		Angular distance error (°)	1.74±0.19	1.65±0.27	1.64±0.25	1.69±0.21	1.82±0.22	1.81±0.26
	Strength error (dB)	13.97±0.35	11.74±0.48	9.99±0.41	9.72±0.39	8.07±0.54	6.53±0.48		Strength error (dB)	0.51±0.07	0.52±0.08	0.51±0.07	0.51±0.06	0.55±0.08	0.65±0.06
	Computation time (s)	15.75±0.14							Computation time (s)	0.08±0.02					

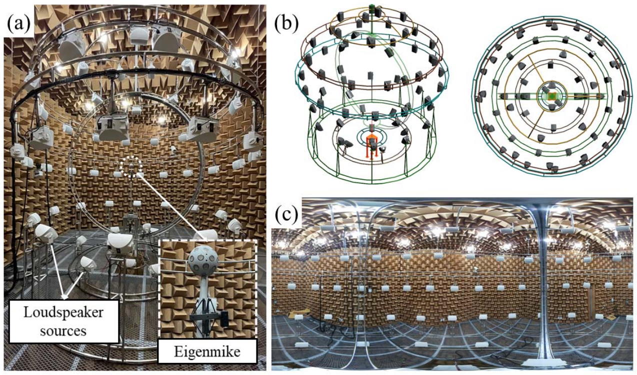

Fig. 11. Experimental configuration for various acoustic scenarios of localizing multiple omnidirectional sound sources: (a) loudspeaker array with 56 loudspeakers placed on a virtual sphere and the em32 Eigenmike, (b) isometric-view/top-view layout of the setup, and (c) panoramic image obtained by 360° camera in the center of the loudspeaker array.

stars ($\star$). The followings summarize the qualitative results of the comparative methods (Fig. 9, columns 3–6).

1) Even though the deconvolution methods can improve the spatial resolution of the **B**s and suppress the sidelobes to some degree, it is observed that there exist uncertain and ambiguous source distribution maps, which fail to yield either highly resolved or extremely clear SSL results. Since CLEAN extracts the source distribution maps based on peak regions of **B**s [36], it is observed that the method performs relatively well in a certain case, e.g., $f = 4$ kHz and $L = 3$, when multiple sources' peak regions are prominent and separate. However, it is also found that several spurious sources are detected at high frequencies, e.g., $f = 8$ kHz and $f = 10$ kHz. In addition, notable is that NNLS mainly produces smeared results at the low frequencies, whereas DAMAS exhibits the largest amount of ghost sources at high frequencies.
2) In contrast, from Fig. 9 (column 3, $\widehat{\mathbf{B}}_T$ by proposed), we observe that the proposed method can successfully predict the spherical target maps for both nonaliased and spatially aliased cases from low to high frequencies. At the low frequencies, e.g., $f = 1$ kHz and $f = 2$ kHz, we find that even adjacent sound sources' distributions can be resolved, demonstrating the high resolvability of the proposed model compared to the other methods. Furthermore, the proposed method is able to produce multiple sources' distributions without any sidelobe patterns under severe ghost imaging conditions at high frequencies, e.g., $f = 8$ kHz and $f = 10$ kHz, which qualitatively show the ghost source robustness of the proposed model.

C. Quantitative Analysis

We now focus on quantitatively analyzing the SSL performances among the proposed method and the comparative methods. Overall quantitative results for the entire test set are described in Fig. 10 (detailed results are summarized in Table I). First, Fig. 10(a) shows the SSL performance comparison in recall and precision metrics. As described in Section III-D, the recall is designed to assess the model's high-resolution capability among the multiple sources, while the precision can be used to capture one's spurious source robustness. As shown in Fig. 10(a), it is observed that the proposed method achieves better localization performance than CLEAN, DAMAS, and NNLS for both recall and precision metrics regardless of the frequency range. While the deconvolution methods generally show a substantial decrease in the recall values at low frequencies, e.g., 1 and 2 kHz, we observe that the proposed model exhibits relatively robust recall performance at low frequencies, showing a mean recall of 97.6% at 1 kHz. Furthermore, while the precision performances of the existing methods drastically deteriorate as the frequency

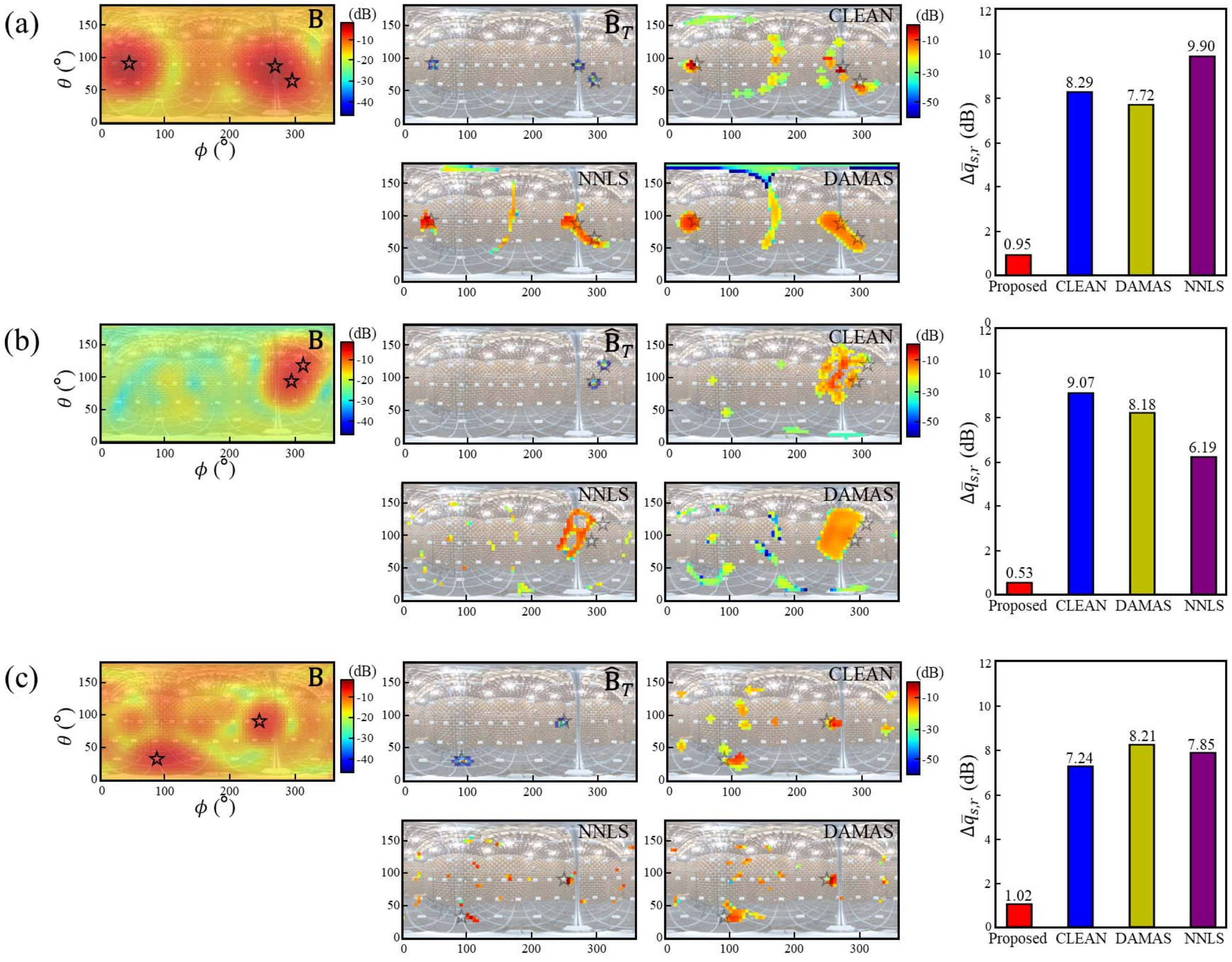

Fig. 12. Several experimental results of the comparative models for multiple omnidirectional SSL in the setup of the VASE Lab: (a) $f = 1$ kHz and $L = 3$ case, (b) $f = 4$ kHz and $L = 2$ case, and (c) $f = 8$ kHz and $L = 2$ case. Note that the sound sources' ground-truth locations are indicated by black stars ($\star$) within spherical beamforming maps $\mathbf{B}$s and the models' results.

goes beyond the spatial Nyquist frequency, it is found that the proposed model outperforms them by attaining high and frequency-independent precision performance.

We also examine the angular distance errors and strength errors of the proposed method and the comparative methods, as shown in Fig. 10(b). As previously mentioned, it should be stressed that only the well-predicted sources $\mathbf{W}$ are considered in error calculations in which it is possible to obtain the numerical error values only in these cases. Note that these errors can be mainly generous to the deconvolution methods, considering that these methods show relatively low recall and precision performance. Nevertheless, the proposed method outperforms other existing methods in terms of the angular distance error and the strength error across all frequencies considered in this study. While the NNLS and the proposed method are likely to show similar mean angular distance errors at the high-frequency range, great gaps can be observed at the lower frequencies, achieving a mean angular distance error of less than 2° on the whole. In particular, it is shown that the strength error of the proposed method is considerably lower than the existing approaches, yielding a mean strength error of approximately 0.5 dB, less than 1 dB. To summarize, we demonstrate that the proposed method can quantitatively fulfill more precise and robust SSL performance than the existing methods.

Finally, statistical computation time results among the proposed and comparative methods are compared in Fig. 10(c). For comparison purposes, we identically utilize the Intel Core i7-6700 CPU for implementing every considered method. Concerning the deconvolution methods, we find that high computational costs are required in order of DAMAS, NNLS, and CLEAN, which show 19.35, 15.75, and 2.20 s, respectively, for processing each beamforming map. On the other hand, the proposed method can exhibit significantly lower computational time (0.08 s) than the other methods, accelerating the prediction 241×, 196×, and 27× faster than DAMAS, NNLS, and CLEAN, respectively. It is worth mentioning that one can speed up the computation process of the proposed method even more using the GPU-based mini-batch computation. As a result, this confirms the potential suitability of the proposed method for real-time acoustic applications and scenarios compared to the existing approaches.

V. Experimental Verification

We further demonstrate the effectiveness of our proposed method through experimental results. In particular, noteworthy is that we design and manufacture the experimental setup, which is called a Various Acoustic Scene Experiment (VASE) Lab, to create omnidirectional acoustic scenarios for the SSL. The configuration of the VASE Lab is visually described in Fig. 11. The VASE Lab consists of two main elements: 1) a loudspeaker array for emitting various acoustic signals in the surrounding directions and 2) an SMA system located in the center of the loudspeaker array for sensing them. While a total of 56 loudspeakers placed on a 2-m-radius virtual sphere are installed in an anechoic chamber, we utilize em32 Eigenmike for the SMA. Besides, to increase the visual explanation of the SSL, we obtain a panoramic image with an omnidirectional camera (Samsung Gear 360) after calibrating its position same as the microphone array [see Fig. 11(c)]. We randomly select multiple combinations of several sound sources and their amplitudes within the 6-dB range to generate the experimental data.

Fig. 12 conveys the experimental SSL results attained by the proposed method and the existing methods, i.e., CLEAN, DAMAS, and NNLS, in the setup of the VASE Lab. We visualize each spherical beamforming map $\mathbf{B}$, the predicted spherical target map $\widehat{\mathbf{B}}_T$, and localization results of the existing methods for several cases of frequency and the number of sound sources. Implementation settings for the comparative models are identical to those used in the simulation study. For visual understanding, the panoramic scene images are overlapped in the background of $\mathbf{B}$ and the models' results, while actual locations of the sound sources are also indicated by black stars ($\star$) within $\mathbf{B}$ and the models' results. Besides, we specify each comparative model's averaged strength error $\Delta\bar{q}_{s,r}$, which quantifies the difference between the ground-truth normalized strength $q_{s,r}$ and the estimated one. As visualized in Fig. 12, overall localization results of the comparative models are observed to be in line with our findings in the simulation study. It can be seen that our proposed method can not only separate adjacent sources at the aliasing-free condition but also eliminate spurious sources at the spatial-aliasing condition, achieving high-resolvability and ghost source robustness at the same time. On the other hand, it is found that the results of other methods involve poor resolution as well as severe ghost source phenomena, as observed in the previous study. In addition, the experimental results show that the proposed approach can achieve significantly low strength errors in comparison to the other methods. As for the localization speed, the averaged computation time of the proposed, CLEAN, DAMAS, and NNLS is observed to be 0.09, 2.42, 20.19, and 17.05 s, respectively, which suggests that the proposed method can expedite the SSL process. As a result, this demonstrates that the feasibility of the proposed approach for a variety of engineering applications that require high-resolution, robust, and rapid SSL performance.

VI. Conclusion

This study proposed a DL approach for localizing multiple omnidirectional sound sources using the SMA system. We first proposed the method of representing high-resolution and clean spherical target maps that can spatially pinpoint omnidirectional sound sources' positions and strengths without grid-related constraints. While we transform the multiple SSL task into image-to-image prediction problem using the proposed representation, the dual-branched spherical convolutional autoencoder network was then proposed to achieve the DL-enabled SSL performance with high resolution. Here, two strategies were integrated into the proposed model to address challenges related to the conventional frequency limits and distortion of the spherical beamforming map. We also proposed the deep neural network-based fast retrieval process from the predicted target map, enabling the estimated results not to be grid-constrained eventually and rapidly. Both quantitative and qualitative results showed that our proposed method outperforms the precedent approaches, where the proposed method showed high resolvability as well as spurious source robustness for nonaliased and spatially aliased conditions across the frequency ranges. In addition, we demonstrated that the proposed method can take advantage of the accelerated computation for the SSL, making it suitable for real-time acoustic localization applications. Finally, we presented the experimental setup specifically designed for creating omnidirectional SSL scenarios and demonstrated our proposed method's potential effectiveness from experimental results. To summarize, the main advantages of the proposed study are as follows: 1) the proposed method can make multiple sources' sound visible with high resolution and 2) the proposed method can significantly accelerate the SSL process without sacrificing its high resolution and precise performance. Nevertheless, it is worth mentioning that application conditions of this study are limited by up to three sound sources under six different frequencies. We believe that it is necessary to expand the data space that can describe more diverse and practical conditions, e.g., greater amount of sound sources, wider frequency ranges, and different SMA configurations, which can motivate a variety of future research directions.

References

[1] K. Park, Y. Motai, and J. R. Yoon, "Acoustic fault detection technique for high-power insulators," *IEEE Trans. Ind. Electron.*, vol. 64, no. 12, pp. 9699–9708, Dec. 2017.

[2] X. Lang, P. Li, Y. Guo, J. Cao, and S. Lu, "A multiple leaks' localization method in a pipeline based on change in the sound velocity," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 7, pp. 5010–5017, Jul. 2020.

[3] A. Glowacz, W. Glowacz, Z. Glowacz, and J. Kozik, "Early fault diagnosis of bearing and stator faults of the single-phase induction motor using acoustic signals," *Measurement*, vol. 113, pp. 1–9, Jan. 2018.

[4] F. Keyrouz, "Advanced binaural sound localization in 3-D for humanoid robots," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 9, pp. 2098–2107, Sep. 2014.

[5] J. S. Hu, C. Y. Chan, C. K. Wang, M. T. Lee, and C. Y. Kuo, "Simultaneous localization of a mobile robot and multiple sound sources using a microphone array," *Adv. Robot.*, vol. 25, nos. 1–2, pp. 135–152, 2011.

[6] D. Salvati, C. Drioli, G. Ferrin, and G. L. Foresti, "Acoustic source localization from multirotor UAVs," *IEEE Trans. Ind. Electron.*, vol. 67, no. 10, pp. 8618–8628, Oct. 2020.

[7] M. Zohourian, G. Enzner, and R. Martin, "Binaural speaker localization integrated into an adaptive beamformer for hearing aids," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 3, pp. 515–528, Mar. 2018.

[8] K. SongGong and H. Chen, "Robust indoor speaker localization in the circular harmonic domain," *IEEE Trans. Ind. Electron.*, vol. 68, no. 4, pp. 3413–3422, Apr. 2021.

[9] P. Chiariotti, M. Martarelli, and P. Castellini, "Acoustic beamforming for noise source localization—Reviews, methodology and applications," *Mech. Syst. Signal Process.*, vol. 120, pp. 422–448, Apr. 2019.

[10] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robot. Auton. Syst.*, vol. 96, pp. 184–210, Oct. 2017.

[11] R. Merino-Martinez *et al.*, "A review of acoustic imaging methods using phased microphone arrays," *CEAS Aeronaut. J.*, vol. 10, no. 1, pp. 197–230, Mar. 2019.

[12] M. R. Bai, J.-G. Ih, and J. Benesty, *Acoustic Array Systems: Theory, Implementation, and Application*. Hoboken, NJ, USA: Wiley, 2013.

[13] U. Schwarz, "Mathematical-statistical description of the iterative beam removing technique (method CLEAN)," *Astron. Astrophys.*, vol. 65, p. 345, Apr. 1978.

[14] Y. Wang, J. Li, P. Stoica, M. Sheplak, and T. Nishida, "Wideband RELAX and wideband CLEAN for aeroacoustic imaging," *J. Acoust. Soc. Amer.*, vol. 115, no. 2, pp. 757–767, Feb. 2004.

[15] T. F. Brooks and W. M. Humphreys, "A deconvolution approach for the mapping of acoustic sources (DAMAS) determined from phased microphone arrays," *J. Sound Vib.*, vol. 294, nos. 4–5, pp. 856–879, Jul. 2006.

[16] T. Brooks, W. Humphreys, and G. Plassman, "DAMAS processing for a phased array study in the NASA langley jet noise laboratory," in *Proc. 16th AIAA/CEAS Aeroacoustics Conf.*, Jun. 2010, p. 3780.

[17] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*. Philadelphia, PA, USA: SIAM, 1995.

[18] K. Ehrenfried and L. Koop, "Comparison of iterative deconvolution algorithms for the mapping of acoustic sources," *AIAA J.*, vol. 45, no. 7, pp. 1584–1595, Jul. 2007.

[19] T. Yardibi, J. Li, P. Stoica, and L. N. Cattafesta, III, "Sparsity constrained deconvolution approaches for acoustic source mapping," *J. Acoust. Soc. Amer.*, vol. 123, no. 5, pp. 2631–2642, 2008.

[20] P. Sijtsma, "CLEAN based on spatial source coherence," *Int. J. Aeroacoust.*, vol. 6, no. 4, pp. 357–374, 1972.

[21] Z. Chu, Y. Yang, and Y. He, "Deconvolution for three-dimensional acoustic source identification based on spherical harmonics beamforming," *J. Sound Vib.*, vol. 344, pp. 484–502, May 2015.

[22] M. J. Bianco *et al.*, "Machine learning in acoustics: Theory and applications," *J. Acoust. Soc. Amer.*, vol. 146, no. 5, pp. 3590–3628, 2019.

[23] W. Ma and X. Liu, "Phased microphone array for sound source localization with deep learning," *Aerosp. Syst.*, vol. 2, no. 2, pp. 71–81, Dec. 2019.

[24] P. Xu, E. J. G. Arcondoulis, and Y. Liu, "Acoustic source imaging using densely connected convolutional networks," *Mech. Syst. Signal Process.*, vol. 151, Apr. 2021, Art. no. 107370.

[25] A. Kujawski, G. Herold, and E. Sarradj, "A deep learning method for grid-free localization and quantification of sound sources," *J. Acoust. Soc. Amer.*, vol. 146, no. 3, pp. EL225–EL231, Sep. 2019.

[26] V. Varanasi, H. Gupta, and R. M. Hegde, "A deep learning framework for robust DOA estimation using spherical harmonic decomposition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1248–1259, 2020.

[27] P. Castellini, N. Giulietti, N. Falcionelli, A. F. Dragoni, and P. Chiariotti, "A neural network based microphone array approach to grid-less noise source localization," *Appl. Acoust.*, vol. 177, Jun. 2021, Art. no. 107947.

[28] B. Rafaely, *Fundamentals of Spherical Array Processing*, vol. 8. Berlin, Germany: Springer-Verlag, 2015.

[29] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. New York, NY, USA: Academic, 1999.

[30] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[31] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 2488–2498.

[32] B. Coors, A. P. Condurache, and A. Geiger, "Spherenet: Learning spherical representations for detection and classification in omnidirectional images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 518–533.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[34] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[35] Z. Chu, S. Zhao, Y. Yang, and Y. Yang, "Deconvolution using CLEAN-SC for acoustic source identification with spherical microphone arrays," *J. Sound Vib.*, vol. 440, pp. 161–173, Feb. 2019.

[36] Z. Chu and Y. Yang, "Comparison of deconvolution methods for the visualization of acoustic sources based on cross-spectral imaging function beamforming," *Mech. Syst. Signal Process.*, vol. 48, nos. 1–2, pp. 404–422, Oct. 2014.

Soo Young Lee received the B.S. degree from Chung-Ang University, Seoul, South Korea, in 2019. He is currently pursuing the Ph.D. degree with the Industrial Artificial Intelligence Laboratory, Pohang University of Science and Technology (POSTECH), Pohang, South Korea.

He was a Visiting Research Fellow for Sustainable Smart Manufacturing using IIoT and Artificial Intelligence with the University of Wisconsin–Madison, Madison, WI, USA, supported by the High-Potential Individuals Global Training Program of International Joint Research. He is currently a Guest Researcher with the Acoustics, Ultrasound, and Vibrations Research Group, Division of Physical Metrology, Korea Research Institute of Standards and Science (KRISS). His research interests include the development of artificial intelligence (AI) for intelligent systems and AI-aided engineering.

Jiho Chang received the B.S. degree from Seoul National University, Seoul, South Korea, in 2003, and the Ph.D. degree from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2011, both in mechanical engineering.

From 2011 to 2014, he was a Post-Doctoral Researcher with the Department of Electrical Engineering, Technical University of Denmark, Denmark. From 2014 to 2016, he was a Senior Engineer with Samsung Electronics. Since 2016, he has been a Senior Researcher with the Korea Research Institute of Standards and Science, Daejeon, South Korea. His research interests include acoustical array signal processing for loudspeaker arrays and microphone arrays, audio signal processing, spatial audio, and spatial hearing.

Seungchul Lee received the B.S. degree from Seoul National University, Seoul, South Korea, in 2001, and the M.S. and Ph.D. degrees from the University of Michigan, Ann Arbor, MI, USA, in 2008 and 2010, respectively.

He was an Assistant Professor with the Ulsan National Institute of Science and Technology, South Korea. He is currently an Associate Processor and a Principal Investigator of the Industrial Artificial Intelligence Laboratory, Pohang University of Science and Technology (POSTECH), Pohang, South Korea. His research interests include industrial artificial intelligence with mechanical systems, deep learning for machine healthcare, and the IoT-based smart manufacturing.