Liang *et al.*
EURASIP Journal on Audio, Speech, and Music Processing (2024) 2024:13
https://doi.org/10.1186/s13636-024-00333-x

EURASIP Journal on Audio, Speech, and Music Processing

EMPIRICAL RESEARCH

Open Access

Sound field reconstruction using neural processes with dynamic kernels

Zining Liang[1], Wen Zhang[1*] and Thushara D. Abhayapala[2]

Abstract

Accurately representing the sound field with high spatial resolution is crucial for immersive and interactive sound field reproduction technology. In recent studies, there has been a notable emphasis on efficiently estimating sound fields from a limited number of discrete observations. In particular, kernel-based methods using Gaussian processes (GPs) with a covariance function to model spatial correlations have been proposed. However, the current methods rely on pre-defined kernels for modeling, requiring the manual identification of optimal kernels and their parameters for different sound fields. In this work, we propose a novel approach that parameterizes GPs using a deep neural network based on neural processes (NPs) to reconstruct the magnitude of the sound field. This method has the advantage of dynamically learning kernels from data using an attention mechanism, allowing for greater flexibility and adaptability to the acoustic properties of the sound field. Numerical experiments demonstrate that our proposed approach outperforms current methods in reconstructing accuracy, providing a promising alternative for sound field reconstruction.

Keywords Sound field reconstruction, Gaussian processes, Kernels, Neural processes

1 Introduction

Accurately describing the characteristics of a sound field, including its spatial, temporal, and spectral properties, is crucial for spatial audio applications, which aims to create realistic auditory environments through loudspeakers or headphones [1, 2]. With recent advances in immersive and interactive sound field reproduction technologies, the ability to render dynamically variable sound fields that allow for listener and source movement within the audio scene has become increasingly important. While obtaining continuous spatial coverage measurements of a sound field over a large area is extremely challenging [3–6], sound field reconstruction offers a resourceful approach to estimate the sound field from a limited set of discrete observations. Such methods can help overcome the limitations of direct measurement techniques and enable realistic, immersive audio experiences in real-world applications.

General solutions for sound field reconstruction typically rely on conventional linear regression, where the sound field is measured at multiple points and represented as a linear combination of basis functions such as plane waves, cylindrical or spherical harmonics [7–10]. However, a large number of basis functions are needed to accurately represent sound fields over a large spatial region using conventional linear regression. Under specific acoustic assumptions, it is possible to represent the sound field using sparse representations, including plane-wave [11] or spherical wave [12] expansions, and modal decomposition [13, 14], as well as equivalent source methods [15–17]. Many of these techniques employ the principle of compressed sensing principles [18] to estimate undersampled data for sound field reconstruction.

*Correspondence:
Wen Zhang
wen.zhang@nwpu.edu.cn
[1] Center of Intelligent Acoustics and Immersive Communications, School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China
[2] Audio and Acoustic Signal Processing Group, College of Engineering and Computer Science, The Australian National University, Canberra, Australia

Another approach, known as kernel ridge regression, is based on infinite-dimensional analysis of sound fields to address the issue of basis function truncation [19–21]. In this field, the hierarchical kernel was proposed [22], which requires manual adjustment of the kernel parameters to align with the specific characteristics of the sound field. More recent works [20, 21] have focused on adaptive kernels, i.e. the usage of pre-defined kernels or sub-kernels with adaptively adapted parameters.

Recently, there have been several data-driven methods utilizing neural networks (NN) for specific tasks within the field of sound field reconstruction [23–28]. Many of these methods are inspired primarily by image restoration and segmentation techniques in computer vision. For example, convolutional neural network (CNN) architectures including U-Net [23] was proposed for reconstructing sound field magnitude [23], physics-informed CNNs was proposed for reconstructing sound fields generated by point sources [26], and MultiResUNet was used for microphone array based room impulse response interpolation [24].

Our research focuses on reconstructing various types of sound fields across an entire spatial region, using a limited number of discrete observations. The work is based on Gaussian processes (GPs), which are powerful probabilistic models that can be used to capture the spatial correlation in the field by employing a kernel function and also to handle the uncertainty associated with the field's variations. The work in [22] presents a pioneering approach to using GPs for sound field reconstruction, demonstrating the significant potential of this technique. However, one crucial aspect that strongly influences the performance of GP models is the choice of kernel function. At the moment, there are still several unresolved questions regarding kernel selection. Firstly, the current work employs the pre-defined kernels, with the kernel function parameters adapted solely from the observations, resulting in limited expressiveness. Secondly, the current work has primarily focused on sound field reconstruction of far-field sources or sparsely distributed sources in reverberant rooms. The kernel functions used in prior work do not adequately capture near-field acoustic properties. Hence, there is potential for further exploration into various types of sound fields, such as near-field sources and standing waves, etc.

In summary, identifying an appropriate kernel with optimal kernel function parameters for various types of sound fields can be challenging. To address this issue, this paper proposes a novel data-driven approach to reconstruct the magnitude of the sound pressure using neural processes (NPs) [29]. NPs enable us to parameterize GPs using a deep neural network. In addition, we introduce dynamic kernels that can effectively adapt to the properties of diverse sound fields by leveraging attention mechanisms. Note that here the motivations for modeling sound field magnitude as in [23] are as follows. (1) The human auditory system is more sensitive to changes in sound magnitude than to changes in phase. Therefore, capturing and reconstructing the magnitude can often be sufficient for achieving perceptually accurate results. (2) Reconstructing only the magnitude simplifies the training complexity.

In this paper, the primary objective is to achieve an accurate reconstruction of sound field magnitudes using minimal observations that are arbitrarily and irregularly distributed. The paper is structured as follows. Section 2 provides a review of the GPs model, including commonly used kernel functions, and highlights the limitations of this model. Building on this, Section 3 presents the conceptual framework and neural network architecture details of the proposed approach using NPs. Section 4 outlines the training procedure and presents results on the reconstruction accuracy of the proposed method, in comparison with the conventional linear regression models and data-driven models.

2 Overview of GPs

2.1 GPs methodology

The problem is defined as reconstructing a sound field within a specific area of interest, using only a limited and finite set of observations, which are denoted as $\tilde{\mathbf{u}} = [\tilde{u}(\mathbf{r}_1, \omega), \ldots \tilde{u}(\mathbf{r}_N, \omega)]$, where $\mathbf{r} \in \Omega$ is the spatial locations and ω is the angular frequency. Hereafter, ω is omitted for notation simplicity. The observed pressure $\tilde{u}(\mathbf{r})$ at a location $\mathbf{r}$ is represented as

$$\tilde{u}(\mathbf{r}) = f(\mathbf{r}) + e(\mathbf{r}), \tag{1}$$

where the true sound field $f(\mathbf{r})$ cannot be directly observed or measured and $e(\mathbf{r})$ denotes the measurement noise [22].

Assuming the sound field in the space is a zero mean complex GP, that is the distribution of sound pressure within that space follows a complex Gaussian distribution

$$\tilde{u}(\mathbf{r}) \sim \mathcal{CGP}\big(0, \kappa\left(\mathbf{r}, \mathbf{r}'\right)\big), \tag{2}$$

where the covariance function, or the kernel, $\kappa\left(\mathbf{r}, \mathbf{r}'\right)$ of the sound pressures between the spatial locations of $\mathbf{r}$ and $\mathbf{r}'$ is written as

$$\kappa\left(\mathbf{r}, \mathbf{r}'\right) = \mathbb{E}\left[u(\mathbf{r})u\left(\mathbf{r}'\right)\right]. \tag{3}$$

The measurement noise in (1) is also assumed complex Gaussian with zero mean

$$e(\mathbf{r}) \sim \mathcal{CGP}\big(0, \kappa_e\left(\mathbf{r}, \mathbf{r}'\right)\big). \tag{4}$$

To predict the sound pressure at a new location $\mathbf{r}_*$, we need to compute the posterior distribution of $u_*(\mathbf{r})$ given the observed data $\tilde{u}(\mathbf{r})$ and the kernel parameters. This can be done using the conditional distribution of a multivariate normal distribution [30],

$$u_*(\mathbf{r}) \mid \mathbf{r}_*, \mathbf{r}, \tilde{\mathbf{u}} \sim \mathcal{GP}\left(\mu_{u_*|\tilde{\mathbf{u}}}(\mathbf{r}), \kappa_{u_*|\tilde{\mathbf{u}}}(\mathbf{r}, \mathbf{r}_*)\right), \tag{5}$$

where $\mu_{u_*|\tilde{\mathbf{u}}}(\mathbf{r})$ is the predictive mean and $\kappa_{u_*|\tilde{\mathbf{u}}}(\mathbf{r}, \mathbf{r}_*)$ is the kernel between the observed position and predictive position.

The optimal sound field reconstruction is the posterior mean in (5), that is

$$\mu_{u_*|\tilde{\mathbf{u}}}(\mathbf{r}) = \boldsymbol{\kappa}^{\mathrm{H}}(\mathbf{K} + \boldsymbol{\Sigma})^{-1}\tilde{\mathbf{u}}, \tag{6}$$

where the kernel, $\boldsymbol{\kappa} = [\kappa(\mathbf{r}_1, \mathbf{r}_*) \cdots \kappa(\mathbf{r}_N, \mathbf{r}_*)]$, is the spatial correlation function between the N observed pressures and the predictive locations $\mathbf{r}_*$, and the covariance matrices $\mathbf{K}$ and $\boldsymbol{\Sigma}$ are defined as [31]

$$\begin{aligned} \boldsymbol{\Sigma} &= \mathbb{E}\left[\mathbf{e}\mathbf{e}^{\mathrm{H}}\right], \\ \mathbf{K} &= \mathbb{E}\left[\mathbf{f}\mathbf{f}^{\mathrm{H}}\right]. \end{aligned} \tag{7}$$

Obviously, the kernel function, which models the spatial correlation between the sound pressure measurements, is a crucial part of sound field reconstruction using GP. The choice of kernel function can have a significant impact on the accuracy and efficiency of the sound field reconstruction.

2.2 Kernel functions

Kernels for sound field representation are typically categorized based on their properties of stationarity and isotropy. It is vital to choose or develop a kernel function that aligns with the characteristics of the sound field in GP methodology. For instance, a diffuse field demonstrates stationary and isotropic spatial correlation, while a plane wave field presents stationary but anisotropic spatial correlation. Below are some frequently applied kernel functions in audio and acoustics research [32].

2.2.1 RBF kernels

The definition of the isotropic radial basis function (RBF) kernel is

$$\kappa_{RBF_i}(\mathbf{r}, \mathbf{r}') = \alpha^2 \exp\left(-\frac{1}{2\rho^2}\|\boldsymbol{\delta}\|^2\right), \tag{8}$$

where α is the scaling factor that adjusts the kernel functions to match the size of the data, ρ is the length scale defining the decay rate of the kernel, and $\boldsymbol{\delta} \triangleq \mathbf{r} - \mathbf{r}'$ is the Euclidean distance between two points.

The definition of the anisotropic RBF kernel is

$$\kappa_{RBF_a}(\mathbf{r}, \mathbf{r}') = \alpha^2 \exp\left(-\frac{1}{2}\sum_{l=1}^{L}\frac{\left\|\mathbf{u}_l^{\mathrm{T}}\boldsymbol{\delta}\right\|^2}{\rho_l^2}\right), \tag{9}$$

where the unitary vector $\mathbf{u}_l \in \mathbb{R}^D$ defines the lth direction and ρ_l is the length scale of the corresponding direction.

The definition of the periodic RBF kernel, which is derived from Eq. (9), gives

$$\kappa_{RBF_p}(\mathbf{r}, \mathbf{r}') = \alpha^2 \exp\left(-\sum_{l=1}^{L}\frac{1}{2\rho_l^2}\sin^2\left(\frac{k\left\|\mathbf{u}_l^{\mathrm{T}}\boldsymbol{\delta}\right\|}{2}\right)\right), \tag{10}$$

where the kernel repeats every wavelength $\lambda = 2\pi/k$.

2.2.2 The plane waves kernels

Plane-wave expansions serve as a widely used method in sound field reconstruction. By decomposing the sound field into a sum of plane waves with varying amplitudes, directions, and frequencies, it becomes possible to reconstruct the field by determining their respective amplitudes and phases [14, 17, 33, 34]. That is, at the wavenumber k, the field at any point in space r can be expressed as

$$f(\mathbf{x}) = \sum_{l=1}^{L} w_l \mathrm{e}^{-\mathrm{j}\mathbf{k}_l^{\mathrm{T}}\mathbf{r}}, \tag{11}$$

where w_l are unknown weights, $\mathrm{e}^{-\mathrm{j}\mathbf{k}_l^{\mathrm{T}}\mathbf{r}}$is the elementary wave function, and $\mathbf{k}_l = k\mathbf{u}_l$ is the wavenumber vector.

If the weights w_l are also modeled as a complex Gaussian process such that

$$w_l \sim \mathcal{GP}\left(0, \sigma_l^2\right), \tag{12}$$

the kernel for the sound field that is generated by multiple sound sources [35, 36] is defined as

$$\kappa_m(\mathbf{r}, \mathbf{r}') = \sigma_{\mathbf{w}}^2 \sum_{l=1}^{L} \mathrm{e}^{-\mathrm{j}\mathbf{k}_l^{\mathrm{T}}\boldsymbol{\delta}}, \tag{13}$$

where the weights w_l share a same variance $\sigma_{\mathbf{w}}^2$. For a special case that the sound field is generated by only a few sources, which is normally characterized as sparse [37, 38], and the kernel is defined as

$$\kappa_s(\mathbf{r}, \mathbf{r}') = \sum_{l=1}^{L} \sigma_l^2 \mathrm{e}^{-\mathrm{j}\mathbf{k}_l^{\mathrm{T}}\boldsymbol{\delta}}, \tag{14}$$

where the variances of the weights w_l are independent and σ_l are considered as inverse gamma distributed [39]

$$\sigma_l \sim \Gamma^{-1}(a, b) = \frac{b^a}{\Gamma(a)}(1/\sigma_l)^{a+1}\exp\left(-b/\sigma_l\right), \tag{15}$$

where $a > 0$ is the shape parameter and $b > 0$ is the scale parameter of the density function. With a fixed prior a, smaller values of b promote sparser solutions.

The concept of the hierarchical kernel κ_h is introduced in [22]. In order to adapt to both normal and sparse sound field, the parameters σ_l in (15) is defined as

$$\sigma_h \sim \Gamma^{-1}(1, b), \quad b \sim \mathcal{N}(\mu_b, \sigma_b). \tag{16}$$

2.2.3 The diffuse field kernel

For a diffuse field driven by a pure tone, the spatial correlation and coherence can be modeled by the superposition of an infinite number of random phase plane waves [34]. That is, the diffuse field kernel function corresponding to (11) in the limit $L \to \infty$ is written as follows,

$$\kappa_f(\mathbf{r}, \mathbf{r}') = \sigma_{\mathbf{w}}^2 \lim_{L\to\infty} \sum_{l=1}^{L} \mathrm{e}^{-\mathrm{j}\mathbf{k}_l^{\mathrm{T}}\boldsymbol{\delta}}. \tag{17}$$

For the two-dimensional case, the kernel in (17) is the zeroth-order Bessel function

$$\kappa_b(\mathbf{r}, \mathbf{r}') = \frac{\sigma_{\mathbf{w}}^2}{2\pi} \int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{j}k\|\boldsymbol{\delta}\|\cos\varphi} \mathrm{d}\varphi = \sigma_{\mathbf{w}}^2 \, \mathrm{J}_0(k\|\boldsymbol{\delta}\|). \tag{18}$$

In summary, when attempting sound field reconstruction using GPs, it is necessary to understand the characteristics of the sound field and select the appropriate kernel function. Once the optimal kernel function is determined, Eq. (6) can be utilized to obtain the predictive sound field pressure. However, if there is no suitable kernel function available, a custom kernel may need to be derived. Nevertheless, developing a kernel that can effectively adapt to diverse acoustic environments can be challenging, particularly when dealing with complex sound fields. Additionally, estimating the optimal hyperparameters of the kernel through numerous experiments can be a time-consuming process.

3 Proposed method

In this work, we propose a novel approach to automatically obtain the optimal kernel from the magnitude of the sound field data for reconstruction, using a data-driven model based on NPs with attention mechanisms. Our proposed model generates dynamic kernels that can adapt to the unique properties of various sound fields and defines distributions over sound field functions similar to GPs. This combination provides a probabilistic, data-efficient, flexible, and computationally efficient solution for optimal kernel selection.

In this section, we first detail the overall architecture of our method in Section 3.1, and then introduce the proposed two-stream encoder and the efficient and lightweight decoder in Sections 3.2 and 3.3, respectively.

3.1 Architecture

As shown in Fig. 1, the proposed model is composed of an encoder and a decoder. Specifically, the encoder contains two paths: a GPs parameterized path, which models the global structure of the stochastic process realization, and a dynamic kernel path, which captures the spatial correlation between observations and predictions.

The encoder takes a limited set of observed sound field magnitude measurements along with their corresponding

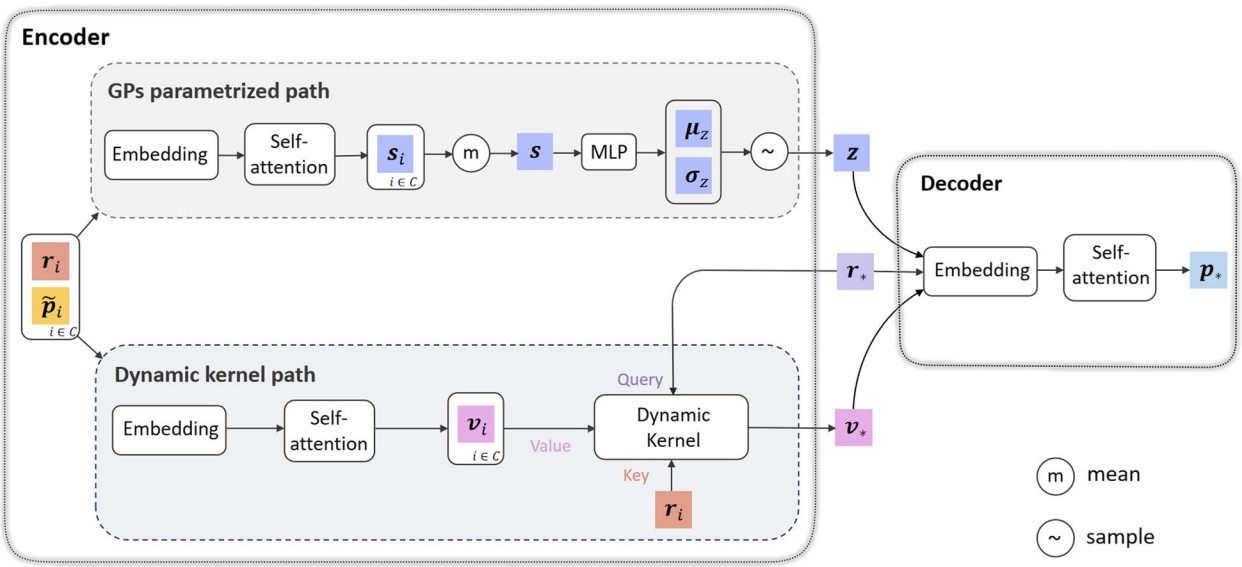

Fig. 1 Schematic diagram of the neural network architecture proposed for sound field reconstruction

locations $(\mathbf{r}, \tilde{\mathbf{p}})_{i \in C(0,N)}$ as input, where $\mathbf{p} = |\mathbf{u}|$ and C denotes the set of integers from 0 to N. Within the GPs parameterized path, the encoder outputs a latent variable $\mathbf{z}$, which encodes the global structure and uncertainty of the sound field distributions in the function space. In the dynamic kernel path, given the target location $\mathbf{r}_*$, the dynamic kernel mechanism outputs a correlation-specific representation $\mathbf{v}_*$. Since the dynamic kernel models the spatial correlation between observations and predictions using differentiable attention, which cannot be analytically obtained and acts as an implicit kernel, we visualize it in Section 4.6.

The decoder takes the latent variable $\mathbf{z}$, the correlation-specific representation $\mathbf{v}_*$, and the target location $\mathbf{r}_*$ as input and produces the predictive sound field magnitude $\mathbf{p}_*$ of the target location. This process can be understood as analogous to reconstructing the sound field using an appropriate kernel, utilizing the neural network to carry out the calculation described by (5).

3.2 Encoder

In this section, we introduce the structure and mechanism of the encoder with the two distinct paths.

3.2.1 GPs parameterized using NPs

The GPs parameterized path is designed to learn distributions over sound field functions from observations. To represent a GP using a neural network, we assume that $F(x) \sim \mathcal{GP}(\mu, \sigma)$ can be parameterized by a high-dimensional random vector $\mathbf{z}$, i.e., the latent variable [29]. We can then write $F(x) = g(x, \mathbf{z})$ for some fixed and learnable function g, where $\mathbf{z}$ models different realizations of the data-generating GPs [40]. The motivation for introducing $\mathbf{z}$ is to enable our model to capture different types of sound fields.

In the GPs parameterized path, the observed sound field magnitudes in the frequency-spatial domain $(\mathbf{r}, \tilde{\mathbf{p}})_i$ are embedded from the input space to the representation space using fully connected layers with Gaussian Error Linear Unit (GELU) [41] activation functions. In our approach, we incorporate a self-attention (SA) mechanism [42], denoted as $\mathbf{s}_i = SA(\mathbf{r}_i, \tilde{\mathbf{p}}_i)$, to model higher-order interactions within the sound field. The SA mechanism allows us to capture the interactions among the observations, enabling the learning of global structural features of the sound field, and obtaining richer representations of the observations. The mean aggregator is used to combine the features as $\mathbf{s} = m(\mathbf{s}_i)$ and generate a single global representation by Multi-Layer Perceptron (MLP), which parameterizes the latent distribution $\mathbf{z} \sim \mathcal{GP}(\boldsymbol{\mu}_z, \boldsymbol{\sigma}_z)$. Finally, each sample of $\mathbf{z}$ corresponds to one realization of the GPs, capturing the global uncertainty.

In summary, the GPs parameterized path learns the mapping from the observed data to the latent distribution of the GPs, representing Eq. (5) by the neural network. Following this framework, the kernel function is not explicitly defined but is learned through the neural network's parameters, which is described in detail below.

3.2.2 Dynamic kernel-based attention mechanism

In GPs, the kernel function captures the relationship between pairs of inputs by computing the dot product between their corresponding feature maps. Here, the kernel is defined as $\kappa(x, , x') = \langle \Phi(x), \Phi(x') \rangle = \Phi(x)^\top \Phi(x')$, where Φ represents the feature map that maps the inputs into a higher-dimensional feature space. The advantage of using such a kernel is that it allows us to design algorithms based on dot-product spaces [43]. In our approach, we introduce a dynamic kernel mechanism inspired by the Scaled Dot-Product Attention (SDPA) [42]. This dynamic kernel mechanism enables us to model the spatial correlation presented in diverse sound fields. More specifically, the target location $\mathbf{r}_*$ is treated as a query, while the observations $(\mathbf{r}, \mathbf{v})_i$ are treated as key-value pairs. Here, $\mathbf{v}_i$ represents the transformation of $\tilde{\mathbf{p}}_i$ into a higher-dimensional space through embedding. Similarly, both $\mathbf{r}_i$ and $\mathbf{r}_*$ undergo embedding within the dynamic kernel mechanism. The SDPA mechanism allows us to calculate weights that determine the correlation of each observation with respect to the target location, enabling accurate prediction of the sound field magnitude $\mathbf{p}_*$ at the target location.

Suppose we have n key-value pairs arranged as matrices $\mathbf{R} \in \mathbb{R}^{n \times d_r}$, $\mathbf{V} \in \mathbb{R}^{n \times d_v}$, and m queries $\mathbf{R}_* \in \mathbb{R}^{m \times d_r}$. The dynamic kernel mechanism calculates correlation weights $\boldsymbol{\kappa}_d$ by taking the dot-product of the queries and keys scaled by d_r, i.e., the kernel form [44], and assigns $\boldsymbol{\kappa}_d$ to $\mathbf{V}$ to obtain the output $\mathbf{V}_*$, which gives

$$\begin{aligned} \boldsymbol{\kappa}_d &= \text{softmax}\left(\mathbf{R}_* \mathbf{R}^\top / \sqrt{d_r}\right), \\ \mathbf{V}_* &= \boldsymbol{\kappa}_d \mathbf{V} \in \mathbb{R}^{d_v}. \end{aligned} \tag{19}$$

In addition to using a single dynamic kernel, we further propose using a multi-dynamic kernel to achieve linear smoother query values [42, 44]. As shown in Eq. (20), the multi-dynamic kernel is obtained by the sum of h kernels mapping with different weights $\mathbf{W}$, defined by

$$\begin{aligned} \boldsymbol{\kappa}_i &= \text{softmax}\left(\mathbf{R}_* \mathbf{W}_{*i} (\mathbf{R}\mathbf{W}_\mathbf{i})^\top / \sqrt{d_k}\right), \\ \mathbf{V}_* &= (\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_h)\mathbf{V} \in \mathbb{R}^{d_v}, i \in [1, h]. \end{aligned} \tag{20}$$

For each target location $\mathbf{r}_*$, the dynamic kernel generates an attention map between $\mathbf{r}_*$ and observations $(\mathbf{r}, \tilde{p})_i$,

which are totally learned from the data. This allows our proposed model to make more accurate predictions in environments with different acoustic properties. The visualization of this part is shown in Section 4.6.

3.3 Decoder

The decoder takes the latent variable $\mathbf{z}$, the correlation-specific representation $\mathbf{v}_*$, and the target location $\mathbf{r}_*$ as input. We define a Gaussian likelihood to describe the decoder, that is

$$\pi(\mathbf{p}_* \mid \mathbf{z}, \mathbf{v}_*, \mathbf{r}_*) = \mathcal{N}\Big(\mathbf{p}_* \mid g_\theta(\mathbf{r}_*, \mathbf{z}), \mathbf{v}_*, \tau^{-1}\mathbf{I}\Big), \quad (21)$$

where $\mathbf{z}$ is a global latent variable, $g_\theta(\mathbf{r}_*, \mathbf{z})$ is a decoder function to generate a prediction for target sound field magnitude $\mathbf{p}_*$ at a location $\mathbf{r}_*$, which is implemented as a deep neural network with parameters θ, and τ^{-1} is the variance of observation noise [29, 45]. Specifically, the likelihood $\pi(\mathbf{p}_* \mid \mathbf{z}, \mathbf{v}_*, \mathbf{r}_*)$ is defined as a factorized Gaussian distribution across the predictions $(\mathbf{r}_*, \mathbf{p}_*)$ with mean and variance determined by $\mathbf{z}$ and correlation-specific representation $\mathbf{v}_*$.

To generate the predictive sound field magnitude $\mathbf{p}_*$, the proposed model is defined by

$$\pi(\mathbf{p}_*, \mathbf{z} \mid \mathbf{v}_*, \mathbf{r}_*) = \pi(\mathbf{z} \mid \mathbf{s})\mathcal{N}\Big(\mathbf{p}_* \mid g_\theta(\mathbf{r}_*, \mathbf{z}), \mathbf{v}_*, \tau^{-1}\mathbf{I}\Big). \quad (22)$$

Since the conditional prior $\pi(\mathbf{z} \mid \mathbf{s})$ in Eq. (22) is intractable, it is approximated using the variational posterior [29]

$$q(\mathbf{z} \mid \mathbf{s}) = \mathcal{N}\big(\mathbf{z} \mid \boldsymbol{\mu}_z(m(\mathbf{s}_i)), \boldsymbol{\sigma}_z(m(\mathbf{s}_i))\big). \quad (23)$$

where $m(\cdot)$ is a mean aggregator function, and $\boldsymbol{\mu}_\omega(\cdot)$ and $\boldsymbol{\sigma}_\omega(\cdot)$ parameterize a normal distribution from which $\mathbf{z}$ is sampled.

3.4 Loss function

The parameters of the encoder and decoder are learned by maximizing the evidence lower-bound (ELBO),

$$\begin{aligned} L_{\text{ELBO}} = & -\mathbb{E}_{q(\mathbf{z}|\mathbf{s}_*)}[\log \pi(\mathbf{p}_* \mid \mathbf{z}, \mathbf{v}_*, \mathbf{r}_*)] \\ & + \mathrm{KL}(q(\mathbf{z} \mid \mathbf{s}_*) \| q(\mathbf{z} \mid \mathbf{s})). \end{aligned} \quad (24)$$

The objective function consists of two terms. The first term is the reconstruction error (RE), which is equivalent to the mean squared error (MSE) [46]. We denote this term as L_D, and it measures the discrepancy between the predicted output $\mathbf{p}_*$ and the corresponding ground truth $\mathbf{p}_\bullet$. The MSE is computed over all the elements, denoted as $\mathcal{N}$. The second term is called the Kullback-Leibler(KL) divergence [47], which is a measure of dissimilarity between two probability distributions. It quantifies the difference between the distribution of observed data $q(\mathbf{z} \mid \mathbf{s})$ and the distribution of predicted data $q(\mathbf{z} \mid \mathbf{s}_*)$ during the training process.

$$L_{\text{D}} = \frac{1}{\mathcal{N}} \sum_{i \in \mathcal{N}} |\mathbf{p}_*(\boldsymbol{r}_i) - \mathbf{p}_\bullet(\boldsymbol{r}_i)|^2. \quad (25)$$

To achieve a balance between data reconstruction and meaningful representation learning, we assign equal weights to both terms during training.

4 Simulation experiments

We evaluated the performance of our proposed sound field reconstruction model in comparison to the GPs and data-driven models. The sound fields we reconstructed included both spatially stationary and non-stationary fields, such as a diffuse field and point sources in the near-field. Additionally, we reconstructed simulated room transfer functions (RTFs) using the image source method [48] and modal theory [34]. Our reconstruction was carried out on a two-dimensional grid composed of 32 by 32 uniformly spaced points along the relevant dimensions. The absolute distance between input points is determined by the room size. Specifically, the distance between points along the x-axis is $l_x/32$, and the distance between points along the y-axis is $l_y/32$. To ensure scale independence in the learning process, it is common to standardize the input for each frequency. This standardization involves transforming the input values such that they have a mean of 0 and a standard deviation of 1.

4.1 Evaluation metrics

We use two metrics to evaluate the performance of our models. The first metric is the normalized mean square error (NMSE) between the ground truth $\mathbf{p}_\bullet$ and the predictions $\mathbf{p}_*$ for each frequency point k, which is calculated as follows

$$\mathrm{NMSE_k} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{\|\mathrm{p}_\bullet(\mathbf{r}_\mathrm{i}, \omega_\mathrm{k}) - \mathrm{p}_*(\mathbf{r}_\mathrm{i}, \omega_\mathrm{k})\|_2^2}{\|\mathrm{p}_\bullet(\mathbf{r}_\mathrm{i}, \omega_\mathrm{k})\|_2^2}. \quad (26)$$

The second metric is the Modal Assurance Criterion (MAC) [49] for each frequency point k, which is defined as follows,

$$\mathrm{MAC_k} = \frac{\left\|\mathbf{p}_{\bullet\mathrm{k}}^{\mathrm{T}} \mathbf{p}_{*\mathrm{k}}\right\|_2^2}{\left(\mathbf{p}_{\bullet\mathrm{k}}^{\mathrm{T}} \mathbf{p}_{\bullet\mathrm{k}}\right)\left(\mathbf{p}_{*\mathrm{k}}^{\mathrm{T}} \mathbf{p}_{*\mathrm{k}}\right)}. \quad (27)$$

The MAC measure evaluates the level of spatial similarity by determining how well the model predicts the overall shape of the pressure distribution in the sound field for each frequency point. The MAC values range from 0 (indicating maximum dissimilarity) to 1

(representing identical shapes), providing a quantitative measure of the quality of the model's predictions.

4.2 Training procedure

Our proposed model can be trained end-to-end on simulated data. To optimize the model, we use the Adam optimizer [50] and train it for 300 epochs. The base learning rate is initially set to 1e−4 and decays to 1e−5 after 200 epochs. Moreover, to achieve better performance and stability during the training process, we implement an exponential warm-up strategy throughout the first 20 epochs.

4.3 Spatially stationary field

In this section, we explore the reconstruction of the diffuse field, which is modeled by the superposition of an infinite number of random phase plane waves, as shown in Eq. (17). This type of sound field is particularly relevant to the sound field present in reverberation rooms [51].

To evaluate the performance of our proposed model, we conduct experiments on simulated data. Specifically, we estimate the sound field magnitudes in the frequency band [30, 500] Hz on a 32 by 32 grid, given 10 observations arbitrarily placed. The simulated data is generated by using m plane waves with unit magnitude and random phase, i.e., $\angle \mathbf{u}_l \sim \mathcal{U}[0, 2\pi)$ and random direction of propagation, i.e., $\mathbf{k}_l \sim \mathcal{U}[-k, k]$. Here, m is randomly sampled from the range of $m \in (1000, 3000)$. To train our proposed model, we use a diverse set of 8000 diffuse fields according to the above parameter settings.

In order to evaluate the effectiveness of our proposed model, we compare it against GPs with different kernels, including the Bessel kernel, hierarchical kernel, and RBF kernels. The prior densities of parameters in Eq. (8)–(10) are defined as $\alpha \sim \mathcal{N}(0, 1)$, ρ and $\rho_l \sim \Gamma^{-1}(a_\rho, b_\rho)$, where $a_\rho = 5$ and $b_\rho = 5$. For the hierarchical kernel in Eq. (16), the parameters are set as $b = 10^{-b_{\log}}$ and $b_{\log} \sim \mathcal{N}(2, 1)$ [22]. In order for the mean magnitude of the fields to be 1 Pa, the fields are normalized. The parameters settings and scaling align with the original work [52].

In Table 1, we present the mean performance of our proposed model compared to GPs on a diverse set of 1000 diffuse fields. The results clearly demonstrate that our model exhibits significantly improved reconstruction performance. To provide a detailed visualization of the reconstruction process, we selected a sound field from the test set. Figure 2 depicts the sound field magnitudes of the reconstructed data at various frequencies. The Bessel kernel performs relatively well due to its aptitude to coincide with the diffuse field. The hierarchical kernel exhibits a certain level of adaptability to the property of the sound field, enabling it to capture the structure of the diffuse field. However, in regions where there are no observations, such as the upper left corner, all kernels poorly extrapolate the sound field, particularly at 500 Hz. This phenomenon highlights the limitations of the GPs method in accurately capturing the complex behavior of the sound field in sparsely sampled regions.

In comparison, the proposed model achieves the best performance due to the proposed attention-based dynamic kernel mechanism, which enables the model to effectively capture the global sound field and obtain richer representations. This enhances the model's

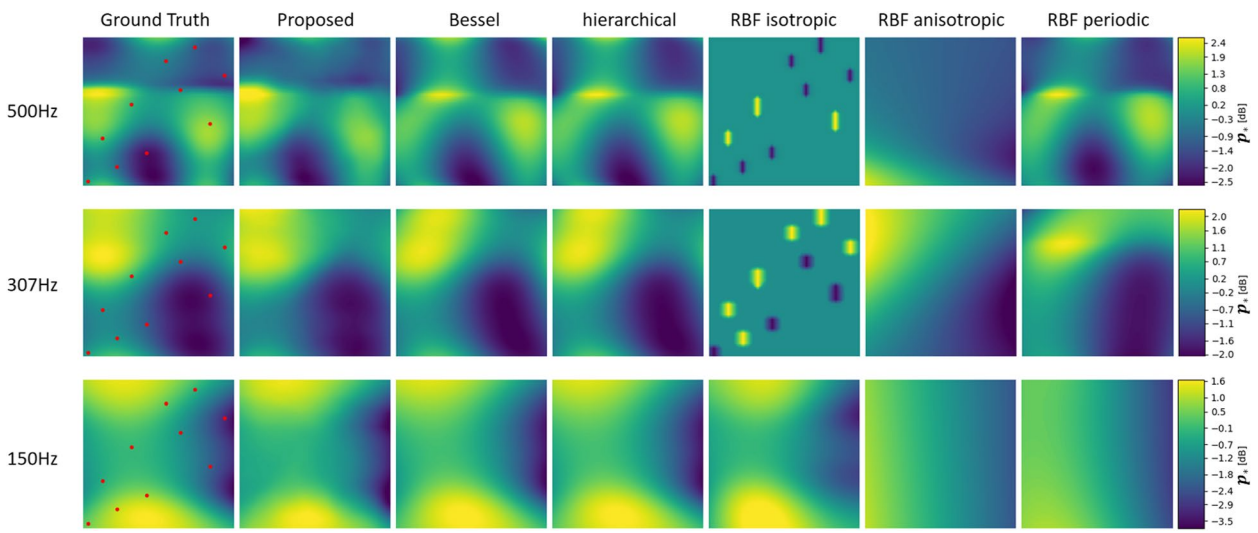

Fig. 2 Reconstructed diffuse-field magnitudes of different frequencies given 10 observations arbitrarily placed. The red dots indicate the locations used for reconstructing predicted sound field magnitudes

Table 1 The mean of NMSE and MAC of diffuse-field test dataset of different frequencies given 10 observations arbitrarily placed

	NMSE (dB)			MAC		
Method	**150 Hz**	**307 Hz**	**500 Hz**	**150 Hz**	**307 Hz**	**500 Hz**
Proposed	**−22.8639**	**−19.6237**	**−11.3852**	**0.9948**	**0.9893**	**0.9278**
Bessel	−9.5050	−7.6729	−1.2261	0.9227	0.9642	0.7874
Hierarchical	−10.2242	−7.9026	−1.4814	0.9305	0.9648	0.7877
RBF isotropic	−7.1763	−0.0536	−0.0431	0.9224	0.0080	0.0112
RBF anisotropic	2.9634	1.3958	2.4845	0.8075	0.1106	0.0063
RBF periodic	−2.0407	1.4845	1.8934	0.8606	0.7885	0.6728

overall performance, enabling it to outperform other approaches.

4.4 Spatially non-stationary field

In this section, we discuss the process of reconstructing the sound field in the near-field created by multiple point sources. This type of sound field is particularly relevant to the direct component of the Room Impulse Response (RIR) [21, 53]. The direct component of the RIR provides critical information about the room geometry [54].

To train our model, we created a dataset consisting of 8000 simulated sound fields. Each field is composed of a random number of point sources, denoted as $j \sim \mathcal{U}[1, 6]$, which are randomly distributed. Each point source is positioned at a radial distance, represented by $d \sim \mathcal{U}(\lambda, 3\lambda)$, from the central point of the reconstruction area. The parameters of GPs method are set as Section 4.3.

Table 2 shows the mean performance of our proposed model and GPs on a diverse set of 1000 near-fields. To provide a detailed reconstruction demonstration, we selected a sound field from the test set for visualization. Figure 3 shows the reconstruction of the near-field produced by five point sources evenly distributed at 2λ m from the center of the reconstructed area. From the figure, we see that the GPs method with existing kernels fails to accurately follow the distance inverse law in terms of the pressure amplitude reconstruction. This discrepancy arises from the mismatch between the kernel functions and the properties of the sound field. Specifically, the magnitude of the reconstructed sound field is relatively small near the sources (i.e., the edge of the reconstruction area), while the magnitude is excessively large at locations further away from the sources (i.e., the center of the reconstruction area). In addition, the kernels are poor for source localization, making it difficult to distinguish the location or even the number of sound sources from Fig. 3.

As predicted, the proposed model demonstrates superior performance in accurately reconstructing

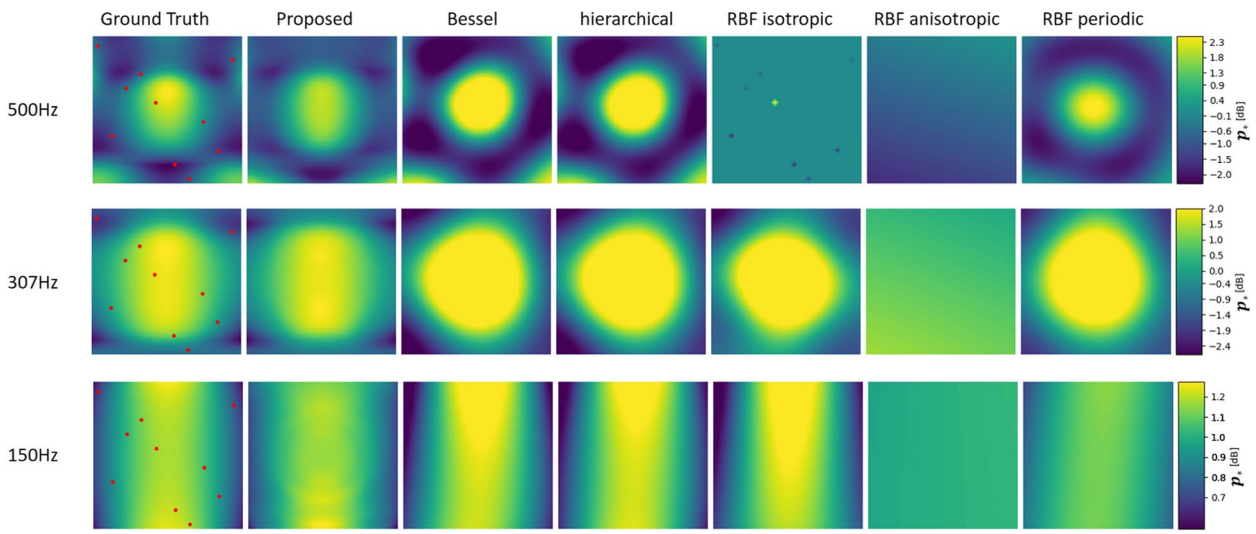

Fig. 3 Reconstructed near-field magnitudes of different frequencies given 10 observations arbitrarily placed. The red dots indicate the location used for reconstructing predicted sound field magnitude

Table 2 The mean of NMSE and MAC of near-field test dataset of different frequencies given 10 observations arbitrarily placed

	NMSE (dB)			MAC		
Method	**150 Hz**	**307 Hz**	**500 Hz**	**150 Hz**	**307 Hz**	**500 Hz**
Proposed	**−17.8494**	**−10.5829**	**−7.0664**	**0.9872**	**0.9134**	**0.8077**
Bessel	−13.7646	−1.5615	2.4769	0.9793	0.9272	0.7050
Hierarchical	−15.1681	−1.7293	2.2114	0.9849	0.9294	0.7083
RBF isotropic	−15.7987	−1.1729	3.1793	0.9753	0.0088	0.0112
RBF anisotropic	−8.7447	3.9365	3.6229	0.9387	0.0946	0.0019
RBF periodic	−15.1838	−1.1221	3.2672	0.9866	0.9136	0.5357

sources with varying numbers, orientations, and distances, particularly at 500 Hz. This outcome highlights the remarkable ability of the proposed model to generalize effectively and reconstruct diverse sound fields.

4.5 RTF magnitude reconstruction

RTFs are a crucial component for achieving immersive and interactive sound field reproduction in virtual reality applications [13]. They represent the frequency-domain representation of RIRs, which typically comprise direct and reverberant components that can be modeled by spherical waves and diffuse fields, respectively [31]. In Sections 4.3 and 4.4, we demonstrated the remarkable superiority of our proposed model over the GPs method in both near-field and diffuse-field sound field reconstruction. To provide a fair comparison, we further compare our proposed model with a data-driven sound field reconstruction method based on a U-net-like neural network [23]. The training process and settings are in line with the original work [55]. We employed two simulation methods, the Image-Source Method (ISM) and Modal Theory (MT), to generate RTF datasets. We tested the ability of our model to reconstruct sound fields in simple small-sized rooms, as well as complex rooms with standing waves. Note that the trained networks are not specific to any particular room geometries or wall reflective properties but only leverage the limited set of observations within the reconstruction area of interest, demonstrating the versatility and practicality of our proposed approach.

4.5.1 ISM-RTFs dataset

The ISM for generating RIRs is widely used in sound field reconstruction [56, 57], with the RIR generator [48] being a popular tool due to its simplicity and computational efficiency. The ISM-based approach is well-suited for small room sizes and simple geometries. In the frequency domain, the generated RTFs are represented as

$$p(\omega, \mathbf{r} \mid \mathbf{r_0}) = \sum_{\beta}^{B} \sum_{\gamma=-\infty}^{\infty} A(\omega) \frac{\mathrm{e}^{\mathrm{j}(\omega \mathrm{t} - \mathrm{k}\|\mathbf{r}_\beta + \mathbf{r}_\gamma\|)}}{4\pi \left\|\mathbf{r}_\beta + \mathbf{r}_\gamma\right\|}, \quad (28)$$

where $\mathbf{r}_\beta$ are the vectors corresponding to the permutations of $(x_0 \pm x, y_0 \pm y, z_0 \pm z)$, γ is the integer vector triplet (n_x, n_y, n_z), and $\mathbf{r}_\gamma = 2(n_x L_x, n_y, n_z L_z)$ [31].

In our simulations, we investigated point source radiation in 2D rooms within the frequency range of [30, 500] Hz, where $B = 4$ and $z = 0$ as specified in Eq. (28). We conducted the simulations in 11,000 rectangular rooms with floor areas randomly sampled from 12 to 20 m^2. In each room, an omnidirectional source was placed in a uniformly sampled random location. We set reverberation time $T_{60} = 0.4s$, the sampling frequency to $fs = 48$ kHz, and simulate reflections up to the 3rd order. To assess the performance of our proposed model with a limited number of observations, we placed 10, 30, and 50 microphones in a 32 by 32 grid in an arbitrary manner. We used 10,000 and 1000 rooms for training and testing the model, respectively, from the dataset. We then analyzed the mean performance of the model across these test rooms.

As shown in Fig. 4, our proposed algorithm consistently outperforms the U-net model. Specifically, the proposed model achieves similar results with only 30 observations, while U-net requires 50 observations to achieve comparable performance. This improvement can be attributed to the dynamic kernel that incorporates global information more comprehensively than the partial convolution employed in U-net [58]. This demonstrates the potential of our proposed model to reduce the number of required samples while maintaining its effectiveness.

Furthermore, we observe that the performance of our proposed model improves as the number of available observations increases. Although the performance slightly degrades with increasing frequency, the model still exhibits good performance in reconstructing RTFs in small rooms across most frequencies. These outcomes suggest that our algorithm is effective for reconstructing RTFs in small rooms.

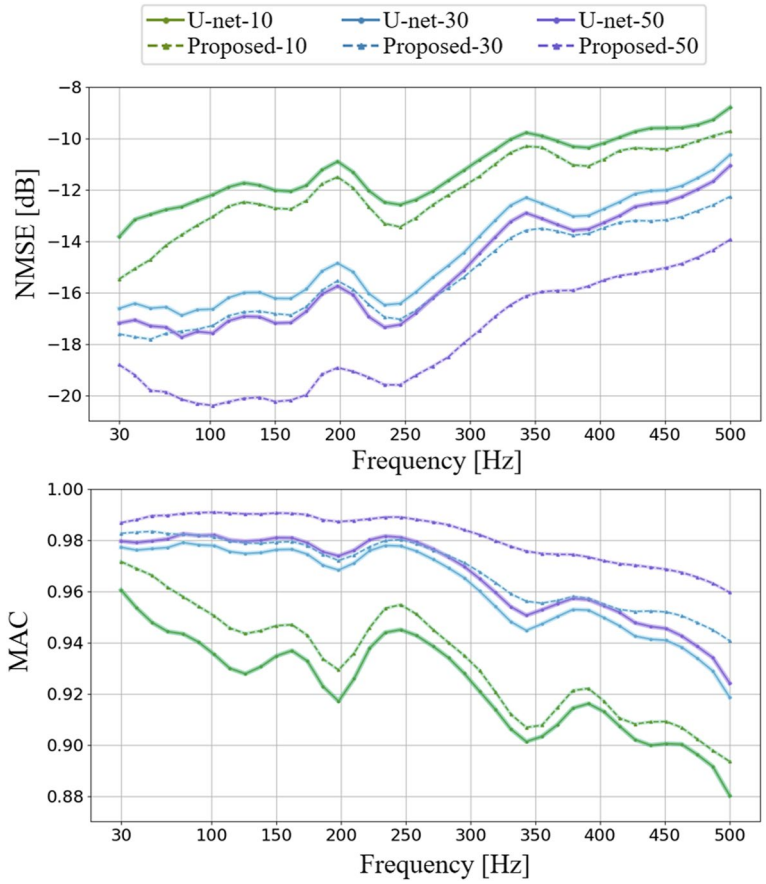

Fig. 4 Normalized mean square error (NMSE) in dB and Modal Assurance Criterion (MAC) estimated from ISM's RTF dataset given 10, 30, and 50 observations arbitrarily placed

4.5.2 MT-RTFs dataset

In order to investigate the potential of our proposed model for reconstructing complex sound field with standing waves, we generated a dataset using MT [55], i.e., the following equation

$$G(\mathbf{r}, \mathbf{r}_0, w) \approx -\frac{1}{V} \sum_{N} \frac{\psi_N(\mathbf{r})\psi_N(\mathbf{r}_0)}{(\omega/c)^2 - (\omega_N/c)^2 - j\omega/\tau_N}, \tag{29}$$

where $\sum_N$ is a triple summation across the modal order in each dimension (n_x, n_y, n_z) of the room, V is the room volume, $\psi_N(\cdot)$ is the eigenfunctions (representing the mode shape), and ω_N denotes eigenfrequencies (representing the resonance frequency). The time constant τ_N represents the characteristic time for a specific mode in a room to decay. It is a constant obtained by dividing the total sound energy in the room by the sound power absorbed by the walls related to that particular mode. Specifically, for each mode, τ_N is calculated from the absorption coefficient determined using Sabine's equation [23]. Here, we focus on 2D rectangular rooms within the frequency band [30, 500] Hz. We incorporate all room modes with eigenfrequencies f_m below 600 Hz, and specifically set n_z to 0 in Eq. (29). Consequently, the total number of modes can be calculated using the formula $N = {f_m}^2/(c^2/4n_xn_y)$ [34]. A reverberation time of $T_{60} = 0.4s$ is assumed. The training and test sets are split, and the room size and sound source location settings are the same as in Section 4.5.1.

Figure 5 depicts the mean performance of the proposed model in reconstructing MT's RTF dataset. The performance of the proposed model given 10, 30, and 50 observations consistently outperforms the U-net, indicating its potential for effectively reconstructing standing waves. Particularly in the low-frequency range, the proposed model exhibits a significant advantage. As the reconstruction frequency approaches the highest eigenfrequency, the complexity of the modes increases, which leads to a decrease in the reconstruction performance. This phenomenon aligns with theoretical expectations, suggesting that a higher number of observations is required to improve

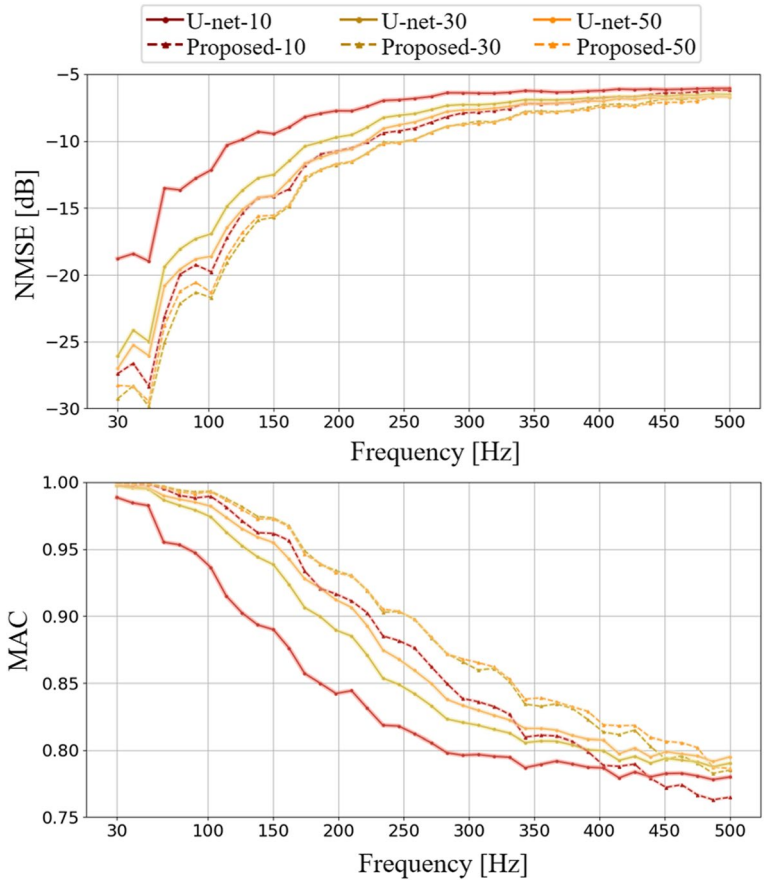

Fig. 5 Normalized mean square error (NMSE) in dB and Modal Assurance Criterion (MAC) estimated from MT's RTF dataset given 10, 30, and 50 observations arbitrarily placed

robustness and overcome the challenges posed by undersampling [23, 59].

In addition, comparing Figs. 4 and 5, the method's performance deteriorates with increasing frequency, which is more noticeable in Fig. 5. The reason for this phenomenon is the ISM-RTFs dataset is more homogeneous than the MT-RTFs dataset. Specifically, the sound fields generated by IMS are produced in shoebox rooms with image reflections up to the 3rd order. This indicates a relatively sparse sound field with wavefronts in the space-time domain. Due to the transient nature of the wavefronts, this type of sound field is dense in the frequency domain. In contrast, the sound fields generated by MT are relatively sparse in the modal region of the sound field (up to Schroeder's frequency). As frequency approaches Schroeder's frequency, the sound fields have increasingly more modes and eventually become diffuse.

4.6 Dynamic kernel visualization

In this section, we demonstrate the spatial correlation between observations and target locations using the proposed dynamic kernel Eq. (19). We select multiple rooms from both IMS-RTFs and MT-RTFs datasets to visualize the sound field and their spatial correlation at specific frequencies.

Figure 6a and b demonstrate that for the IMS-RTFs dataset, the correlation is stronger between observations in close proximity to the target location. Additionally, the dynamic kernel assigns relatively more attention to locations where the sound source is situated, i.e., the bottom left of Fig. 6a and the middle left of Fig. 6b, and less to areas where the sound field characteristics are less prominent, such as the top side of Fig. 6a and right side of Fig. 6b. This reflects the validity of the dynamic kernel in apportioning attention to the global sound field. Additionally, it provides an explanation for the experimental results in Section 4.3, as the sound field reconstructed by the proposed method reflects the locations of sources.

For the MT-RTFs dataset shown in Fig. 6c and d, similar conclusions can be drawn, with closer observations displaying a stronger correlation with the target location. Interestingly, the observations that correlate most

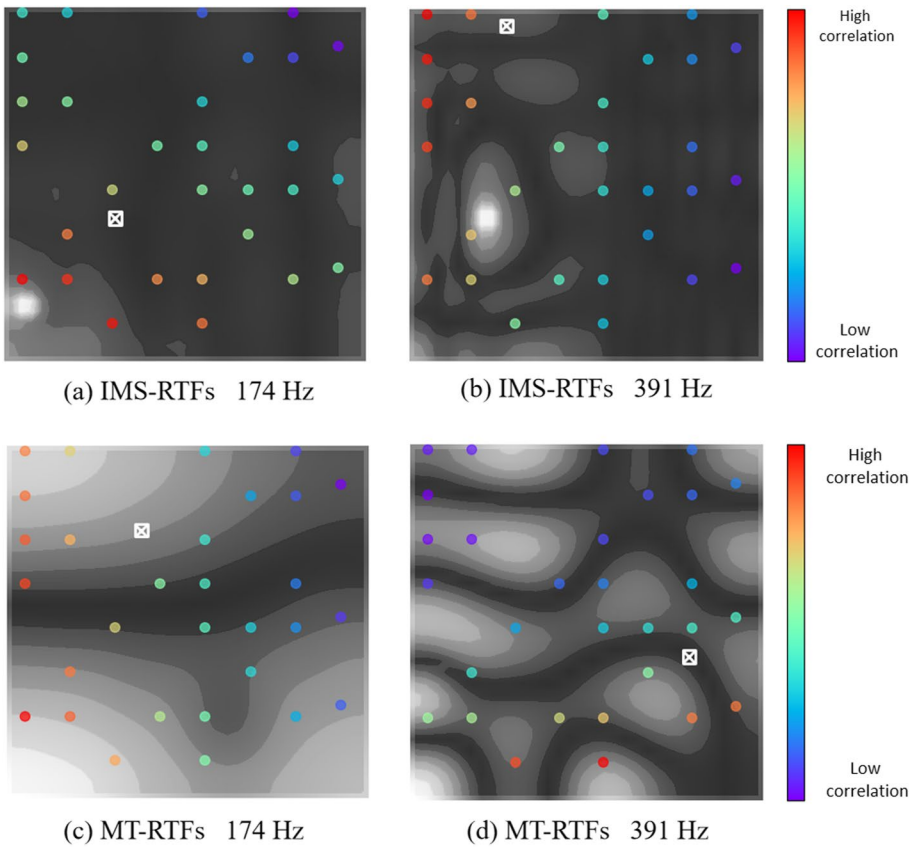

Fig. 6 Visualization of spatial correlation of RTFs at a specific frequency. The dots indicate the location of the observations that were used to reconstruct the output of our model, and the white square denotes the target location that needs to predict its magnitude. The color of the dots reflects the strength of the correlation between the observations and the target

strongly with the target point are not in proximity to it but rather at the left bottom of the Fig. 6c and the bottom of the Fig. 6d, where the structural features of the sound field are noticeable. This highlights the dynamic kernel's ability to learn from data. Furthermore, it is apparent that the sound field environment in the MT-RTFs dataset is more intricate than that of the IMS-RTFs dataset at the same frequency. This difference explains the proposed model's performance degradation in reconstructing the MT-RTFs dataset at higher frequencies.

4.7 Model generalization

To assess the generalization ability of our model, we combined the four datasets mentioned in Sections 4.3, 4.4, and 4.5 into a diverse dataset for both training and testing. We conducted experiments on four types of sound fields, where 10 observations were arbitrarily placed. In our comparisons between U-net and GPs, we employed the best-performing hierarchical kernel for GPs.

As illustrated in Fig. 7, we observed a decline in performance for the model trained on the diverse dataset when compared to training on each individual dataset separately. This decline can be attributed to the varying data distributions present in each dataset. However, it is important to note that even with this decline, our proposed model still exhibited strong performance, particularly in terms of robustness at high frequencies.

This outcome serves as a testament to our model's ability to learn from diverse data and highlights its applicability across various sound field scenarios. While the varying data distributions affected the model's performance to some extent, our model showcased resilience and delivered notable results, particularly in capturing sound characteristics at higher frequencies.

4.8 Computational complexity analysis

Apart from enhancing the accuracy of reconstruction, the proposed model also offers a significant advantage in terms of computational complexity during the inference process. With a model size of 4.3 million parameters, the deterministic inference time is around 0.016 s on a Nvidia Tesla K80 GPU. This estimation is based on the observation of 1000 different room predictions. In our experiments, we conducted model training for

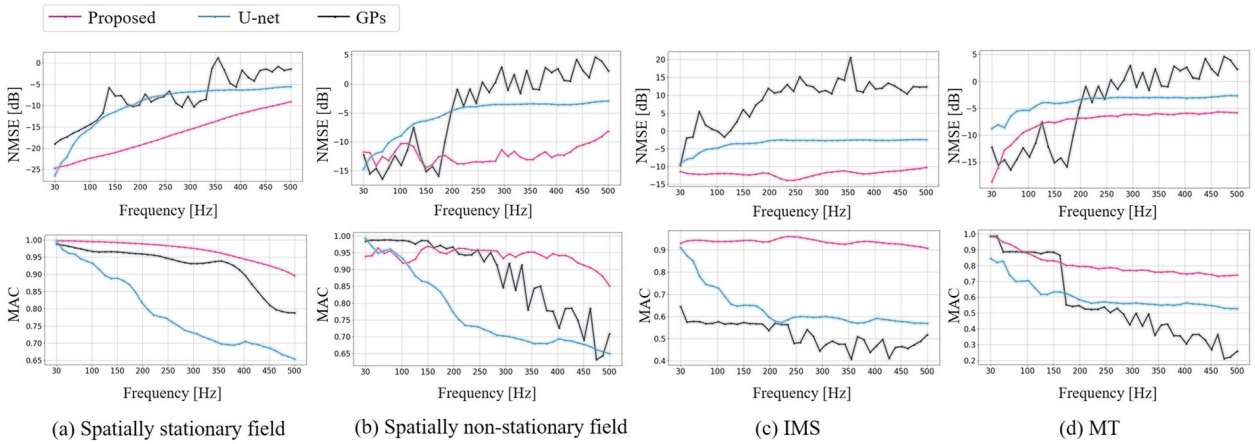

Fig. 7 Normalized mean square error (NMSE) in dB and Modal Assurance Criterion (MAC) estimated from four datasets given 10 observations arbitrarily placed

300 epochs on the training set. Each type of sound field required approximately 12 h of training time. The U-net model size is 3.9 million parameters resulting in a deterministic inference time of approximately 0.083 s on a Nvidia Tesla K80 GPU. Each type of sound field required approximately 24 h of training time for 300 epochs.

5 Conclusion

In this work, we proposed a novel method that parameterizes GPs using a deep neural network based on Neural Processes. Our method allows for the learning of dynamic kernels from simulated data with the introduction of attention, enabling the method to obtain a kernel that adapts to the acoustic properties of the sound field without many functional design restrictions. Numerical experiment results demonstrate that our proposed method outperforms current methods in terms of reconstructing accuracy for a diverse range of sound fields. Future work involves validating our approach using real-world data and further developing the methodology for complex sound field reconstruction.

Abbreviations

CNN	Convolutional neural networks
PICNN	Physics-informed convolutional neural networks
ULA	Uniform linear array
GPs	Gaussian processes
NPs	Neural processes
GELU	Gaussian Error Linear Unit
MLP	Multi-layer perceptron
CA	Cross attention
ELBO	Evidence lower-bound
RTF	Room transfer functions
NMSE	Normalized mean square error
MAC	Model assurance criteria
ISM	Image source method
MT	Modal theory

Acknowledgements

Not applicable.

Authors' contributions

WZ and ZL formalized and conceptualized the problem. ZL performed the experiments. WZ and TDA supervised the research. All authors read and proved the published version of the manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 61831019 and 62271401.

Availability of data and materials

The dataset used and analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 2 June 2023 Accepted: 7 February 2024
Published online: 20 February 2024

References

1. A. Plinge, S.J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, E.A. Habets, in *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*, Six-degrees-of-freedom binaural audio reproduction of first-order ambisonics with distance information (Audio Engineering Society, 2018)
2. M. Cobos, J. Ahrens, K. Kowalczyk, A. Politis, An overview of machine learning and other data-based methods for spatial audio capture, processing, and reproduction. EURASIP J. Audio Speech Music Process. **2022**(1), 1–21 (2022)
3. I.B. Witew, M. Vorländer, N. Xiang, Sampling the sound field in auditoria using large natural-scale array measurements. J. Acoust. Soc. Am. **141**(3), EL300–EL306 (2017)
4. S. Koyama, T. Nishida, K. Kimura, T. Abe, N. Ueno, J. Brunnström, in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Meshrir: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods (IEEE, 2021), pp. 1–5

5. M.S. Kristoffersen, M.B. Møller, P. Martínez-Nuevo, J. Østergaard, Deep sound field reconstruction in real rooms: introducing the isobel sound field dataset. (2021). arXiv preprint arXiv:2102.06455
6. P.N. Samarasinghe, T.D. Abhayapala, M.A. Poletti, in *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*, 3d spatial soundfield recording over large regions (VDE, 2012), pp. 1–4
7. D.B. Ward, T.D. Abhayapala, Reproduction of a plane-wave sound field using an array of loudspeakers. IEEE Trans. Speech Audio Process. **9**(6), 697–707 (2001)
8. N. Ueno, S. Koyama, H. Saruwatari, Sound field recording using distributed microphones based on harmonic analysis of infinite order. IEEE Signal Process. Lett. **25**(1), 135–139 (2017)
9. T. Betlehem, T.D. Abhayapala, Theory and design of sound field reproduction in reverberant rooms. J. Acoust. Soc. Am. **117**(4), 2100–2111 (2005)
10. P. Samarasinghe, T. Abhayapala, M. Poletti, T. Betlehem, An efficient parameterization of the room transfer function. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(12), 2217–2227 (2015). https://doi.org/10.1109/TASLP.2015.2475173
11. S.A. Verburg, E. Fernandez-Grande, Reconstruction of the sound field in a room using compressive sensing. J. Acoust. Soc. Am. **143**(6), 3770–3779 (2018). https://doi.org/10.1121/1.5042247
12. M. Pezzoli, M. Cobos, F. Antonacci, A. Sarti, in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Sparsity-based sound field separation in the spherical harmonics domain (2022), pp. 1051–1055. https://doi.org/10.1109/ICASSP43922.2022.9746391
13. O. Das, P. Calamia, S.V.A. Gari, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Room impulse response interpolation from a sparse set of measurements using a modal architecture (IEEE, 2021), pp. 960–964
14. R. Mignot, G. Chardon, L. Daudet, Low frequency interpolation of room impulse responses using compressed sensing. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(1), 205–216 (2013)
15. S. Lee, Review: The use of equivalent source method in computational acoustics. J. Comput. Acoust. **25**(1), 1630001 (2017). https://doi.org/10.1142/S0218396X16300012
16. I. Tsunokuni, K. Kurokawa, H. Matsuhashi, Y. Ikeda, N. Osaka, Spatial extrapolation of early room impulse responses in local area using sparse equivalent sources and image source method. Appl. Acoust. **179**, 108027 (2021). https://doi.org/10.1016/j.apacoust.2021.108027
17. N. Antonello, E. De Sena, M. Moonen, P.A. Naylor, T. Van Waterschoot, Room impulse response interpolation using a sparse spatio-temporal representation of the sound field. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(10), 1929–1941 (2017)
18. D.L. Donoho, Compressed sensing. IEEE Trans. Inf. Theory. **52**(4), 1289–1306 (2006). https://doi.org/10.1109/TIT.2006.871582
19. N. Ueno, S. Koyama, H. Saruwatari, Sound field recording using distributed microphones based on harmonic analysis of infinite order. IEEE Sig. Process. Lett. **25**(1), 135–139 (2018).https://doi.org/10.1109/LSP.2017.2775242
20. R. Horiuchi, S. Koyama, J.G.C. Ribeiro, N. Ueno, H. Saruwatari, in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Kernel learning for sound field estimation with l1 and l2 regularizations (2021), pp. 261–265. https://doi.org/10.1109/WASPAA52581.2021.9632731
21. J.G. Ribeiro, S. Koyama, H. Saruwatari, in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kernel interpolation of acoustic transfer functions with adaptive kernel for directed and residual reverberations (IEEE, 2023), pp. 1–5
22. D. Caviedes-Nozal, N.A. Riis, F.M. Heuchel, J. Brunskog, P. Gerstoft, E. Fernandez-Grande, Gaussian processes for sound field reconstruction. J. Acoust. Soc. Am. **149**(2), 1107–1119 (2021)
23. F. Lluis, P. Martinez-Nuevo, M. Bo Møller, S. Ewan Shepstone, Sound field reconstruction in rooms: inpainting meets super-resolution. J. Acoust. Soc. Am. **148**(2), 649–659 (2020)
24. M. Pezzoli, D. Perini, A. Bernardini, F. Borra, F. Antonacci, A. Sarti, Deep prior approach for room impulse response reconstruction. Sensors **22**(7), 2710 (2022)
25. E. Fernandez-Grande, X. Karakonstantis, D. Caviedes-Nozal, P. Gerstoft, Generative models for sound field reconstruction. J. Acoust. Soc. Am. **153**(2), 1179–1190 (2023)
26. K. Shigemi, S. Koyama, T. Nakamura, H. Saruwatari, in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, Physics-informed convolutional neural network with bicubic spline interpolation for sound field estimation (IEEE, 2022)
27. A.A. Figueroa Durán, E. Fernandez Grande, in *Proceedings of the 24th International Congress on Acoustics*, Reconstruction of room impulse responses over an extended spatial domain using block-sparse and kernel regression methods (ICA, Korea, 2022)
28. M. Hahmann, S.A. Verburg, E. Fernandez-Grande, Spatial reconstruction of sound fields using local and data-driven functions. J. Acoust. Soc. Am. **150**(6), 4417–4428 (2021)
29. M. Garnelo, J. Schwarz, D. Rosenbaum, F. Viola, D.J. Rezende, S. Eslami, Y.W. Teh, Neural processes. (2018). arXiv preprint arXiv:1807.01622
30. C.E. Rasmussen, C. Williams, *Gaussian Processes for Machine Learning* (The MIT Press, 2005)
31. E. Fernandez-Grande, D. Caviedes-Nozal, M. Hahmann, X. Karakonstantis, S.A. Verburg, in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, Reconstruction of room impulse responses over extended domains for navigable sound field reproduction (IEEE, 2021), pp. 1–8
32. A. Liutkus, R. Badeau, G. Richard, Gaussian processes for underdetermined source separation. IEEE Trans. Signal Proc. **59**, 3155–3167 (2011)
33. J.M. Schmid, E. Fernandez-Grande, M. Hahmann, C. Gurbuz, M. Eser, S. Marburg, Spatial reconstruction of the sound field in a room in the modal frequency range using Bayesian inference. J. Acoust. Soc. Am. **150**(6), 4385–4394 (2021)
34. F. Jacobsen, P.M. Juhl, *Fundamentals of general linear acoustics* (Elsevier Inc., 2013)
35. M. Nolan, E. Fernandez-Grande, J. Brunskog, C.H. Jeong, A wavenumber approach to quantifying the isotropy of the sound field in reverberant spaces. J. Acoust. Soc. Am. **143**, 2514–2526 (2018)
36. E. Fernandez-Grande, Sound field reconstruction using a spherical microphone array. J. Acoust. Soc. Am. **139**, 1168–1178 (2016)
37. K.L. Gemba, S. Nannuru, P. Gerstoft, W.S. Hodgkiss, Multi-frequency sparse Bayesian learning for robust matched field processing. J. Acoust. Soc. Am. **141**, 3411–3420 (2017)
38. K.L. Gemba, S. Nannuru, P. Gerstoft, Robust ocean acoustic localization with sparse Bayesian learning. IEEE J. Sel. Top. Signal Process. **13**, 49–60 (2019)
39. K.P. Murphy, *Machine learning: a probabilistic perspective* (The MIT Press, 2012)
40. H. Kim, A. Mnih, J. Schwarz, M. Garnelo, A. Eslami, D. Rosenbaum, O. Vinyals, Y.W. Teh, Attentive neural processes. (2019). arXiv preprint arXiv:1901.05761
41. D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus). (2016). arXiv preprint arXiv:1606.08415
42. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need. Adv. Neural Inf. Process. Syst. **30**, 5998-6008 (2017)
43. T. Hofmann, B. Schölkopf, A.J. Smola, Kernel methods in machine learning. The Annals of Statistics, **36**(3), 1171–1220 (2008)
44. Y.H.H. Tsai, S. Bai, M. Yamada, L.P. Morency, R. Salakhutdinov, Transformer dissection: a unified understanding of transformer's attention via the lens of kernel. (2019). arXiv preprint arXiv:1908.11775
45. T.G. Rudner, V. Fortuin, Y.W. Teh, Y. Gal, in *Workshop on Bayesian Deep Learning, NeurIPS*, On the connection between neural processes and gaussian processes with deep kernels (NeurIPS, 2018), p. 14
46. L.A.P. Rey, V. Menkovski, J.W. Portegies, Diffusion variational autoencoders. (2019). arXiv preprint arXiv:1901.08991
47. S. Kullback, R.A. Leibler, On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)
48. E.A. Habets, Room impulse response generator. (2014). https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator. Accessed 10 July 2022
49. M. Pastor, M. Binda, T. Harčarik, Modal assurance criterion. Procedia Eng. **48**, 543–548 (2012)
50. D.P. Kingma, J. Ba, Adam: a method for stochastic optimization. (2014). arXiv preprint arXiv:1412.6980
51. M. Nolan, S.A. Verburg, J. Brunskog, E. Fernandez-Grande, Experimental characterization of the sound field in a reverberation room. J. Acoust. Soc. Am. **145**(4), 2237–2246 (2019)

52. D. Caviedes-Nozal. Acoustic gaussian processes (2021). https://github.com/d-caviedes/acoustic_gps. Accessed 2 May 2021
53. J.B. Allen, D.A. Berkley, Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
54. I. Dokmanić, R. Parhizkar, A. Walther, Y.M. Lu, M. Vetterli, Acoustic echoes reveal room shape. Proc. Natl. Acad. Sci. **110**(30), 12186–12191 (2013)
55. F.Lluis. Sound-field-neural-network. (2020). https://github.com/francescelluis/sound-field-neural-network. Accessed 9 Mar 2023
56. M. Fu, J.R. Jensen, Y. Li, M.G. Christensen, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Sparse modeling of the early part of noisy room impulse responses with sparse Bayesian learning (IEEE, 2022), pp. 586–590
57. S. Damiano, F. Borra, A. Bernardini, F. Antonacci, A. Sarti, in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Soundfield reconstruction in reverberant rooms based on compressive sensing and image-source models of early reflections (IEEE, 2021), pp. 366–370
58. G. Liu, F.A. Reda, K.J. Shih, T.C. Wang, A. Tao, B. Catanzaro, in *Proceedings of the European conference on computer vision (ECCV)*, Image inpainting for irregular holes using partial convolutions (ECCV, 2018), pp. 85–100
59. R. Mignot, L. Daudet, F. Ollivier, Room reverberation reconstruction: interpolation of the early part using compressed sensing. IEEE Trans. Audio Speech Lang. Process. **21**(11), 2301–2312 (2013)

Publisher's Note