

Learning an Interpretable End-to-End Network for Real-Time Acoustic Beamforming

Hao Liang, Guanxing Zhou, Xiaotong Tu*, Andreas Jakobsson, Xinghao Ding, Yue Huang

Abstract—Recently, many forms of audio industrial applications, such as sound monitoring and source localization, have begun exploiting smart multi-modal devices equipped with a microphone array. Regrettably, model-based methods are often difficult to employ for such devices due to their high computational complexity, as well as the difficulty of appropriately selecting the user-determined parameters. As an alternative, one may use deep network-based methods, but these are often difficult to generalize, nor can they generate the desired beamforming map directly. In this paper, a computationally efficient acoustic beamforming algorithm is proposed, which may be unrolled to form a model-based deep learning network for real-time imaging, here termed the DAMAS-FISTA-Net. By exploiting the natural structure of an acoustic beamformer, the proposed network inherits the physical knowledge of the acoustic system, and thus learns the underlying physical properties of the propagation. As a result, all the network parameters may be learned end-to-end, guided by a model-based prior using back-propagation. Notably, the proposed network enables an excellent interpretability and the ability of being able to process the raw data directly. Extensive numerical experiments using both simulated and real-world data illustrate the preferable performance of the DAMAS-FISTA-Net as compared to alternative approaches.

Index Terms—Acoustic beamforming, acoustic imaging, source localization, array signal processing, model-based deep learning, interpretable network.

I. INTRODUCTION

The acoustic camera, a multi-modal imaging device, which can visualize sound as a heatmap, has been widely applied in various fields, such as transportation, noise monitoring, and industrial anomaly detection [1]–[6]. As illustrated in Fig. 1, where the acoustic camera is able to make sound “visible”. It processes the multi-dimension signals acquired by the microphone array, employing acoustic beamforming algorithms to generate the sound pressure level distribution on the scanning plane, and displays the results as images or videos. Utilizing the phase-shifting properties of the signal propagation, the delay-and-sum (DAS) algorithm provides a robust, fast, and intuitive imaging result via delay compensation and weighted summation [7], [8]. However, it

The work was supported in part by National Natural Science Foundation of China under Grants 52105126, 82172033, 82272071, and the China Fundamental Research Funds for the Central Universities (Grant No. 20720220082). (*Corresponding author: Xiaotong Tu; Electronic mail: xttx@xmu.edu.cn)

Hao Liang, Guanxing Zhou, and Xiaotong Tu are with the School of Informatics, Xiamen University, Xiamen 361005, China (E-mail: haoliang@stu.xmu.edu.cn; joaquinchou@stu.xmu.edu.cn; xttx@xmu.edu.cn).

Andreas Jakobsson is with the Centre for Mathematical Sciences, Lund University, Sweden (E-mail: aj@maths.lth.se).

Xinghao Ding and Yue Huang are with the Institute of Artificial Intelligent, Xiamen University, Xiamen 361005, China (E-mail: dxh@xmu.edu.cn; yhuang2010@xmu.edu.cn).

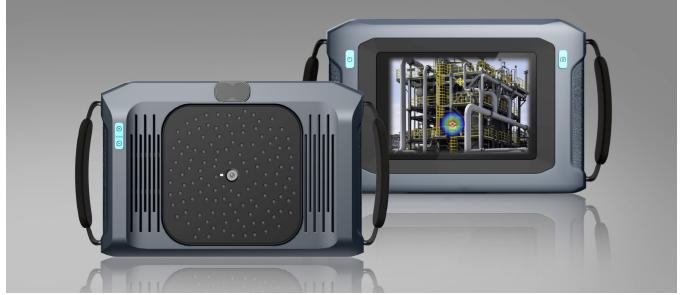


Fig. 1. Schematic of an acoustic camera. The light spot indicates the location and power of an industrial anomalous sound source.

is subject to the Sparrow resolution limit [9], and thus yields low-resolution results with high sidelobes [10].

In order to eliminate these side-lobe interferences, various deconvolution approaches have been proposed. In [11], the CLEAN-PSF method was introduced to iteratively remove the influence of the point spread function (PSF). Then, based on the spatial source coherence, Sijtsma further extended this method to allow for coherent sources, resulting in the so-called CLEAN-SC algorithm [12]. Later, a high-resolution extension of CLEAN-SC was developed in [13]. However, the CLEAN-based methods may erroneously “clean” the true source point, and the deconvolution may as a result become unfocused in the acoustic field. Assuming that the beamforming map is a linear combination of the sources and the PSF, the deconvolution approach for the mapping of incoherent acoustic sources (DAMAS), which removes the PSF’s influence by solving an inverse problem, was introduced in [14]. Building on this, a series of DAMAS-based methods exploiting some structured sparsity and intrinsic acoustic properties have been derived [15]–[20], all aiming to form higher quality and more robust estimates. Although these methods enable a significant improvement of the results, they usually require a high computational cost, preventing their usage in real-time imaging, as well as necessitating an often difficult selection of various user-defined parameters.

Recently, several deep network-based acoustic beamforming approaches have also been proposed to reconstruct acoustic sources [21]–[24]. Compared with the model-based procedures, these methods can dramatically reduce the necessary time complexity, while providing reliable estimates as a result of the strong feature extraction capability of the neural network. However, most existing deep network-based algorithms tend to train the neural network as a black box, inevitably leading to the absence of domain knowledge guidance, which

often causes the resulting network to be difficult to generalize in a robust manner, limiting the practical application of these methods.

In this work, we propose a computationally efficient implementation of the DAMAS, termed the DAMAS-FISTA, which employs the Fast Iterative Shrinkage Thresholding Algorithm (FISTA) [25]. Using this, we then design an interpretable end-to-end network, dubbed the DAMAS-FISTA-Net, by mapping the DAMAS-FISTA to render a high-quality beamforming map for real-time imaging. In contrast to most deep network architectures, the proposed network can directly handle the raw data without pre-beamforming. Notably, the proposed DAMAS-FISTA-Net can also learn the underlying physical properties of the acoustic environment by exploiting the model-based method, allowing the resulting network an excellent ability to be generalized. All the required network parameters may be learned end-to-end, guided by the model-based prior, by using back-propagation. Numerical simulations and experimental data analysis show the effectiveness and advantages of the proposed method.

In summary, the main contributions of this paper are:

- We introduce a computationally efficient implementation of the DAMAS estimator by employing a FISTA framework, forming the so-called DAMAS-FISTA method, which may then be used to guide the design of the proposed subsequent network.
- We propose the novel and computationally efficient DAMAS-FISTA-Net, which not only combines the benefit of the model-based method with those of deep learning, but which inherits related domain knowledge.
- The proposed network exhibits an excellent interpretability and ability to be generalized, while still being capable of processing the raw data directly.
- Extensive experiments illustrate that the DAMAS-FISTA-Net achieves promising performance for both simulated and real-world data, illustrating the practical potential of the method.

The remainder of this paper is organized as follows: In the next section, related work, including several acoustic beamforming methods and model-based deep learning methods, is summarized. Then, Section III presents the acoustic signal model and some related notations, and the DAMAS beamformer is briefly reviewed. In Section IV, the DAMAS-FISTA algorithm is proposed, followed by the proposed DAMAS-FISTA-Net method. Section V provides a comparison and validation of the proposed methods, formed by analyzing both simulated and real-world signals. Finally, our conclusions are summarized in Section VI.

II. RELATED WORK

There are two main categories of related acoustic beamforming methods, including model-based methods, such as the deconvolution-based algorithms, and deep network-based methods. A brief review of these two approaches is found below. Additionally, we provide a concise review of the recent model-based deep learning approaches which are relevant to our work.

Model-based acoustic beamforming. In the field of acoustic beamforming, the DAMAS algorithm introduced the use of deconvolution to increase the resolution and remove disturbing side-lobe interference [14]. However, this algorithm is unsuitable for coherence sources and suffers from being computationally cumbersome limitation. Notable efforts have been made to address these issues. In [16], the cross-power between coherent sources have been considered, while the conjugate relationship is used to reduce the computational complexity. However, computationally intensive operations are still unavoidable. Assuming that the PSF is shift-invariant, Dougherty proposed the DAMAS2, significantly improving the computational efficiency [15]. Later, the computational overhead was further reduced by using efficient solvers [17], [18]. More recently, Ding *et al.* proposed two high-resolution source localization methods, which exploit the sparsity of the beamforming map [20]. Although these methods perform well, a degree of shift-variance is introduced due to the shift-invariant assumption, causing the source power to be underestimated. Generally, all these model-based methods require hundreds or thousands of iterations even when using efficient solvers, which will inevitably increase the computational burden and restrict the application for real-time imaging. Additionally, many of these algorithms depend on a series of user-defined parameters, such as sparsity or penalty parameters, which typically are non-trivial to select in an optimal manner, but which will greatly influence the accuracy of the imaging results.

Deep network-based acoustic beamforming. Recently, many studies have shown that a deep network may be a useful tool for acoustic beamforming by taking advantage of its powerful learning capability. For instance, Reiter and Bell employed the widely used AlexNet [26] to identify point source locations with its input being the pre-beamformed result [21]. Later, Allman *et al.* treated the source localization problem as a classification task [22], designing a two-module deep network based on the VGGNet [27] and Fast R-CNN [28]. However, these two methods only estimate the localization information, without providing any intensity prediction. To handle this issue, Kujawski *et al.* proposed to use the ResNet architecture [29] to process the pre-beamformed result to estimate the position as well as the intensity of a single source [23]. To avoid information loss during pre-beamforming, Zhou *et al.* further proposed the Acoustic-Net [24], which employs the time-frequency distribution of multi-dimension signals as input. Combined with the RepVGG architecture [30], the inference time can then be significantly reduced. However, the above networks are only suitable for single-source situations. Indeed, some recent works (see, e.g., [31], [32]) may be suitable for multi-source situations. Unfortunately, to be efficient, these methods generally need notable prior knowledge, such as the number of sources. The main feature of deep network-based methods is that they are completely data-driven, without employing any physical mechanism, resulting in a network that is data-sensitive and lack structural diversity (repetitively employing fully-connected or convolutional layers). Additionally, neither of these methods can directly generate beamforming maps.

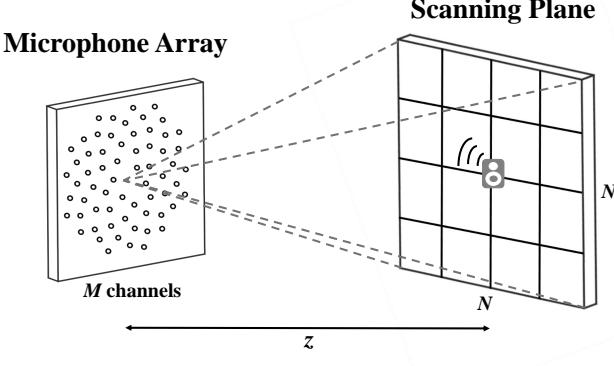


Fig. 2. The assumed model of the acoustic signal propagation.

Model-based deep learning. In order to build a bridge between model-based methods and deep network-based methods, the model-based deep learning framework has also been investigated and has attracted notable attention. A representative method is the learned iterative shrinkage and thresholding algorithm (LISTA) [33], which unrolls the iterative shrinkage and thresholding (ISTA) algorithm [34] as a special network to learn fast approximations of sparse coding. Later, Wang *et al.* developed a deep l_0 encoder to solve l_0 sparse regularization problems [35]. Furthermore, Yang *et al.* considered a more generalized compressive sensing problem, resulting in the ADMM-Net [36], [37]. Also, a recurrent neural network, dubbed DeepWave, has been designed for three-dimension imaging [38]. In [39], an optimization-inspired ISTA-Net was proposed, which may be seen as an extension of LISTA. More recently, Xiang *et al.* further proposed the FISTA-Net by introducing a momentum module and a parameter tweaking strategy [40]. Essentially, the proposed DAMAS-FISTA may be viewed as a special extension of the FISTA-Net, which also provides insights into the acoustic field. Notably, in comparison with the utilization of deep convolutional networks in the FISTA-Net, the proposed network utilizes fewer although significant parameters detailing the natural structure, which thereby effectively avoid the overfitting problem.

III. PRELIMINARIES

A. Traditional Beamforming

The traditional DAS beamforming algorithm exploits the phase difference generated by the relative time-delay on each microphone channel to synchronize the multi-dimensional acoustic signals [7], [8]. Specifically, as shown in Fig.2, the signals resulting from unknown sources are measured by a microphone array with M channels, with its scanning plane at a distance of z m from the scanning plane being divided into $N \times N$ grid points. As a delay in the time domain equals a phase shift in the frequency domain, one may organize these delays into a vector, termed the steering vector, which, for the n -th grid point, may be expressed as [10]

$$\mathbf{g}_n = \left[\frac{r_0}{r_{1,n}} e^{-jk(r_{1,n}-r_0)} \dots \frac{r_0}{r_{M,n}} e^{-jk(r_{M,n}-r_0)} \right]^T, \quad (1)$$

where r_0 denotes the distance from the array center to the n -th grid point, where $r_{m,n}$ is the distance from the m -th microphone to the n -th grid point, with k representing the wave number, i.e.,

$$k = \frac{2\pi f}{c}, \quad (2)$$

where f denotes the scanning frequency and c the speed of sound. A further discussion about the formulations of the steering vector is available in [41]. According to the definition of the steering vector in (1), the weighted steering vector, for the n -th grid point, may be expressed as

$$\begin{aligned} \mathbf{w}_n &= \frac{\mathbf{g}_n}{\|\mathbf{g}_n\|_2^2} \\ &= \left[\frac{r_{1,n}}{r_0} e^{-jk(r_{1,n}-r_0)} \dots \frac{r_{M,n}}{r_0} e^{-jk(r_{M,n}-r_0)} \right]^T. \end{aligned} \quad (3)$$

This allows the measured multi-dimensional signals to be divided into J frames to form a cross-spectral matrix (CSM), which is the correlation matrix of the signals received by the multiple channels in the frequency domain, i.e.,

$$\mathbf{C} = \frac{1}{J} \sum_{j=1}^J \mathbf{p}_j(f) \mathbf{p}_j(f)^H, \quad (4)$$

where $(\cdot)^H$ denotes the conjugate transpose, and

$$\mathbf{p}_j(f) = [p_{j,1}(f) \ p_{j,2}(f) \ \dots \ p_{j,M}(f)]^T, \quad (5)$$

with $p_{j,m}(f)$ denoting the j -th frame frequency-domain signal of i -th microphone. Forming the frequency representation for all grid points, the acoustic source power, i.e., the beamforming map, may be expressed as

$$\mathbf{b} = [b_1 \ b_2 \ \dots \ b_{N^2}]^T, \quad (6)$$

where b_n denotes the power at the n -th grid point, i.e.,

$$b_n = \frac{1}{M^2} \mathbf{w}_n^H \mathbf{C} \mathbf{w}_n. \quad (7)$$

The DAS is an efficient method for acoustic beamforming that has been widely applied in various fields, due to its robustness, and as it is computationally efficient and intuitive. However, it is subject to Sparrow's resolution limit, generally implying a low spatial resolution and the presence of spurious estimates.

B. Deconvolution Approach for Acoustic Beamforming

Employing the characteristics of acoustic source propagation, the DAMAS [14] constructs the inverse problem

$$\mathbf{b} = \mathbf{Ax}, \quad (8)$$

where \mathbf{x} denotes a column vector containing the unknown source power and \mathbf{A} denotes the propagation matrix, i.e.,

$$\mathbf{A} = \frac{1}{M^2} \text{abs}(\mathbf{W}^* \mathbf{G}^T) \circ \text{abs}(\mathbf{W}^* \mathbf{G}^T), \quad (9)$$

where $\mathbf{G} \in R^{N^2 \times M}$ and $\mathbf{W} \in R^{N^2 \times M}$ denote the steering matrix and the weighted steering matrix, respectively, such that

$$\mathbf{G} = \begin{bmatrix} \mathbf{g}_1^T \\ \vdots \\ \mathbf{g}_{N^2}^T \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} \mathbf{w}_1^T \\ \vdots \\ \mathbf{w}_{N^2}^T \end{bmatrix}, \quad (10)$$

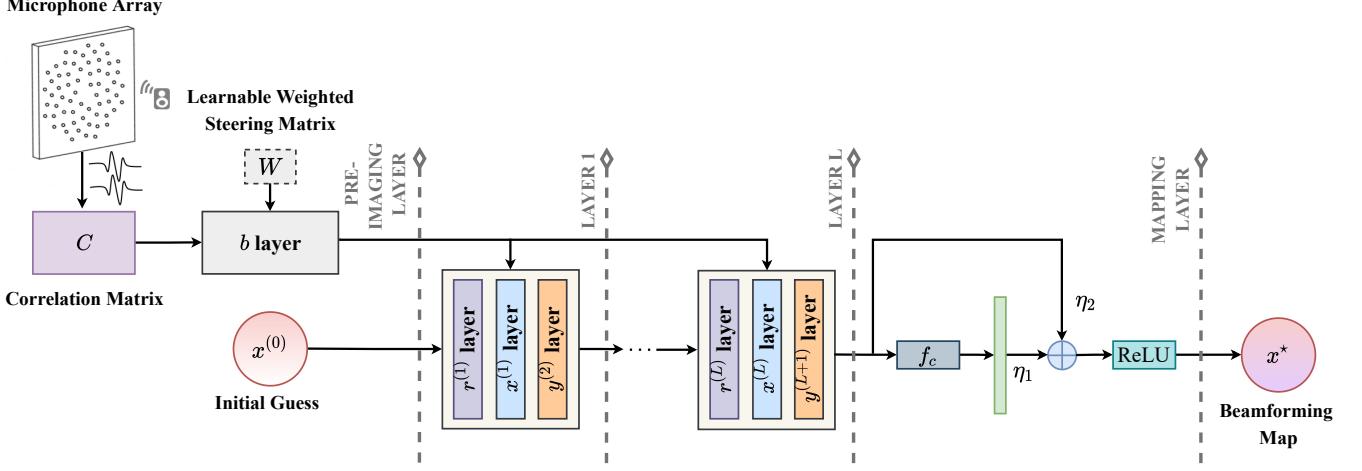


Fig. 3. The overall architecture of the proposed DAMAS-FISTA-Net with L iteration layers. This framework mainly consists of five types of modules: Pre-imaging (b), reconstruction ($r^{(k)}$), nonlinear transform ($x^{(k)}$), momentum ($y^{(k+1)}$), and mapping (x^*). A correlation matrix, C , directly obtained from the raw data, may be successively processed, and finally generates a beamforming map.

with \circ denoting the Hadamard product, $(\cdot)^*$ the conjugate, and $\text{abs}(\cdot)$ the element-wise absolute value. The resulting inverse problem in (8) may be solved efficiently using, for instance, a Gauss–Seidel iterative method. We refer the reader to [14] where the details of the resulting optimization process are discussed. The overall computational complexity of the DAMAS algorithm is $O(N^6)$, making it problematic to employ in real-time acoustic beamforming.

IV. THE PROPOSED METHOD

In this section, we first propose an iterative solver, dubbed the DAMAS-FISTA, to improve the efficiency of the DAMAS estimator. We then proceed to construct the DAMAS-FISTA-Net, an interpretable end-to-end network, by studying the DAMAS-FISTA update equations.

A. The DAMAS-FISTA Algorithm

Inspired by [17], assuming that the DAS result, b , can be modelled as experiencing a white Gaussian noise, and exploiting the prior information that the source power, x , is nonnegative, we propose to recover the sound power map by solving the optimization problem

$$\begin{aligned} \min_{\mathbf{x}} \quad & \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \\ \text{s.t.} \quad & x_i \geq 0. \end{aligned} \quad (11)$$

Since (11) is a non-negative convex optimization problem, the FISTA, which preserves the computational simplicity of ISTA but with a global rate of convergence which is significantly improved, both theoretically and practically [25], [34], may be introduced to solve it by iterating the following update steps:

$$\mathbf{r}^{(k)} = \mathbf{y}^{(k)} - \frac{1}{\mathcal{L}} \mathbf{A}^T (\mathbf{A}\mathbf{y}^{(k)} - \mathbf{b}), \quad (12)$$

$$\mathbf{x}^{(k)} = \max(\mathbf{r}^{(k)}, 0), \quad (13)$$

Algorithm 1 The DAMAS-FISTA algorithm

Initialize: The Lipschitz constant, \mathcal{L} , the initial guess, $\mathbf{x}^{(0)}$, the auxiliary vector, $\mathbf{y}^{(1)}$, the step parameter, $t^{(1)} = 1$, the stopping threshold, ε , and the maximum number of iterations.

repeat

step1. Update $\mathbf{x}^{(k)}$

$$\mathbf{x}^{(k)} = \max \left(\mathbf{y}^{(k)} - \frac{1}{\mathcal{L}} \mathbf{A}^T (\mathbf{A}\mathbf{y}^{(k)} - \mathbf{b}), 0 \right),$$

step2. Update $t^{(k+1)}$

$$t^{(k+1)} = \frac{1 + \sqrt{1 + 4(t^{(k)})^2}}{2},$$

step3. Update $\mathbf{y}^{(k+1)}$

$$\mathbf{y}^{(k+1)} = \mathbf{x}^{(k)} + \frac{t^{(k)} + 1}{t^{(k+1)}} (\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}),$$

until $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_2 < \varepsilon \|\mathbf{x}^{(k-1)}\|_2$ or k reaches the maximum number of iterations.

Output: The acoustic beamforming result, \mathbf{x}^* .

$$t^{(k+1)} = \frac{1 + \sqrt{1 + 4(t^{(k)})^2}}{2}, \quad (14)$$

$$\mathbf{y}^{(k+1)} = \mathbf{x}^{(k)} + \frac{t^{(k)} + 1}{t^{(k+1)}} (\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}), \quad (15)$$

where \mathcal{L} denotes a Lipschitz constant, i.e., the maximum eigenvalue of the Hessian matrix $\mathbf{A}^T \mathbf{A}$, whereas $\mathbf{r}^{(k)}$, $t^{(k)}$, and $\mathbf{y}^{(k)}$ are auxiliary variables. The proposed DAMAS-FISTA algorithm is summarized in Algorithm 1. It is suggested that the stopping threshold ε is selected in the range 10^{-2} to 10^{-4} , with the vectors $\mathbf{x}^{(0)}$ and $\mathbf{y}^{(1)}$ being initialized as a zero vector.

B. Network Architecture

It is worth noting that the above DAMAS-FISTA estimator is an iterative solver, for which one may, reminiscent to [33], [35]–[40], construct a corresponding interpretable end-to-end network architecture, here termed the DAMAS-FISTA-Net. The basic idea in order to do so is to design a non-linear, parameterized, feedforward architecture with a fixed depth, L , which can be trained to approximate the optimal beamforming map. Notably, in contrast to the existing networks for acoustic beamforming, we here design a steering matrix learning module, which can directly construct a beamforming map from raw data, forming a more flexible end-to-end network. DAMAS-FISTA-Net is tailored to the acoustic beamforming problem, mainly generalizing the five types of operations to have learnable parameters as network layers, i.e., a pre-imaging layer (\mathbf{b}), reconstruction layers ($\mathbf{r}^{(k)}$), nonlinear transform layers ($\mathbf{x}^{(k)}$), momentum layers ($\mathbf{y}^{(k+1)}$), and a mapping layer (\mathbf{x}^*). Among them, the $\mathbf{r}^{(k)}$ layers, the $\mathbf{x}^{(k)}$ layers, and the $\mathbf{y}^{(k+1)}$ layers are collectively termed the iteration layers forming a fixed depth, L . Fig.3 illustrates the architecture of the proposed DAMAS-FISTA-Net. We proceed to examine these layers in further detail.

Pre-imaging layer \mathbf{b} : This layer directly processes the raw data (the CSM of the input signal), and then generates a pre-imaging power map, \mathbf{b} , corresponding to the DAS procedure. As mentioned above, there are several options for the weighted steering matrix. Here, we allow this matrix to be learned directly from the sound field information in a data-driven manner. Specifically, given the correlation matrix, \mathbf{C} , the output of this layer may be written as

$$\mathbf{b} = \frac{1}{M^2} \text{Real} \left(\text{RowSum} (\mathbf{W}^* \circ (\mathbf{C} \cdot \mathbf{W}))^T \right), \quad (16)$$

where $\text{Real}(\cdot)$ denotes the operation that returns the real part, and $\text{RowSum}(\cdot)$ the operation that returns the sum of each row of the input matrix, with \mathbf{W} representing the learnable weighted steering matrix.

Reconstruction layer $\mathbf{r}^{(k)}$: This layer reconstructs the power map based on the gradient descent operation of (12). In order to reduce parameter redundancy, the reconstructed dictionary, \mathbf{A} , is fixed. Simultaneously, to preserve the DAMAS-FISTA structure while increasing the network flexibility, the step size may be allowed to vary across iterations, and a weighted parameter may be used. Then, given $\mathbf{y}^{(k)}$, the output of this layer is defined as

$$\mathbf{r}^{(k)} = \iota^{(k)} \mathbf{y}^{(k)} - \rho^{(k)} \mathbf{A}^T (\mathbf{A} \mathbf{y}^{(k)} - \mathbf{b}). \quad (17)$$

Nonlinear transform layer $\mathbf{x}^{(k)}$: This layer aims to ensure the non-negativity of the power estimate by limiting the reconstructed result to lie in a non-negative domain. Given $\mathbf{r}^{(k)}$, i.e., the reconstructed result in stage k , the output is

$$\mathbf{x}^{(k)} = \text{ReLU} (\mathbf{r}^{(k)}), \quad (18)$$

where $\text{ReLU}(\cdot)$ denotes the linear rectification function, i.e.,

$$\text{ReLU}(\mathbf{x}) = \max(\mathbf{x}, 0). \quad (19)$$

Momentum layer $\mathbf{y}^{(k+1)}$: This layer performs a momen-

tum update operation by replacing the constant update step, $t^{(k)}$, in (15) with two learnable parameters, $\tau^{(k)}$ and $\mu^{(k)}$, making the network learn flexible features autonomously from dataset. Given $\mathbf{x}^{(k-1)}$ and $\mathbf{x}^{(k)}$, the output may be expressed as

$$\mathbf{y}^{(k+1)} = \tau^{(k)} \mathbf{x}^{(k)} + \mu^{(k)} (\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}). \quad (20)$$

Mapping layer \mathbf{x}^* : In this layer, inspired by [29], a weighted residual learning module may be designed to map the preliminary output of the L -th nonlinear transform layer to the target domain. Specifically, we utilize a fully-connected layer, f_c , to learn a residual function with reference to the layer inputs. Then, a weighted attention mechanism may be used to learn the mapping feature. Notably, the output power should be non-negative. Thus, the output of this layer with input $\mathbf{x}^{(L)}$ is finally defined as

$$\mathbf{x}^* = \text{ReLU} (\eta_1 \mathbf{x}^{(L)} + \eta_2 f_c(\mathbf{x}^{(L)})), \quad (21)$$

where η_1 and η_2 are both learnable weighted parameters, whereas \mathbf{x}^* denote the acoustic beamforming result learned by the DAMAS-FISTA-Net.

C. Network Parameters

The learnable parameters of the DAMAS-FISTA-Net are

$$\Theta = \{\mathbf{W}\} \cup \{\iota^{(k)}, \rho^{(k)}, \tau^{(k)}, \mu^{(k)}\}_{k=1}^L \cup \{\eta_1, \eta_2, f_c(\theta)\}, \quad (22)$$

where L denotes the depth of the iteration layers and $f_c(\theta)$ the parameters in the fully-connected layer. All these parameters are considered as the network parameters to be learned.

D. Initialization

Instead of random initialization, we propose to employ a warm-starting manner to initialize the network. Specifically, the weighted steering vector, $\mathbf{W}^{(0)}$, may be initialized according to (3), whereas the weighted parameter, $\rho^{(k)}$, can be initialized as the maximum eigenvalue of $\mathbf{A}^T \mathbf{A}$. For the initial guess, $\mathbf{x}^{(0)}$ may be initialized as a zero vector, as well as $\mathbf{y}^{(1)} = \mathbf{x}^{(0)}$. The fully-connected layer, f_c , is initialized with Xavier's algorithm [42]. The parameters $\{\iota^{(k)}, \tau^{(k)}, \mu^{(k)}\}_{k=1}^L$ and $\{\eta_1, \eta_2\}$ are all initialized as 1.

E. Loss Function Design

Given a training data pairs $(\mathbf{p}_i, \mathbf{b}_i^{(gt)})$, the DAMAS-FISTA-Net first transforms the raw data from microphone array, \mathbf{p}_i , to a cross-spectral matrix without information loss. Then, the cross-spectral matrix may be taken as input and generates a high-quality beamforming map, $\hat{\mathbf{b}}_i(\mathbf{p}_i, \Theta)$, as output. A training set Γ is constructed containing pairs of raw data and ground-truth. The network parameters are optimized by minimizing the mean squared error (MSE) loss function, i.e.,

$$\mathcal{L}_{mse} = \frac{1}{|\Gamma|} \sum_{(\mathbf{p}_i, \mathbf{b}_i^{(gt)}) \in \Gamma} \sqrt{\|\hat{\mathbf{b}}_i(\mathbf{p}_i, \Theta) - \mathbf{b}_i^{(gt)}\|_2^2}, \quad (23)$$

which seeks to reduce the discrepancy between the network estimate and the ground truth.

F. Implementation Details

In our work, we use the PyTorch [43] framework to carry out our experiments with a single 3090Ti GPU. We employ the Adam [44] optimizer, with the initial learning rate being 1×10^{-3} , the weight decay being 1×10^{-2} , and the momentum being 0.9. The mini-batch size during training is set to 64, and the depth of the iteration layer, L , is set to 5.

V. NUMERICAL RESULTS AND DISCUSSIONS

To evaluate the capability of the DAMAS-FISTA and the DAMAS-FISTA-Net¹, we examine both simulated and real-world data, comparing the bias, Rényi entropy, and runtime performance of the proposed network to alternative state-of-the-art methods, namely the model-based DAS [8], DAMAS [14], FFT-NNLS [17], and FFT-DFISTA [20] estimators, as well as the network-based methods Acoustic-AlexNet [21], Acoustic-ResNet [23], and Acoustic-Net [24]. In comparing the methods, we examine:

- i) The concentration of the beamforming map energy, evaluated using the Rényi entropy [20], [45], i.e.,

$$R(\alpha) = \frac{1}{1-\alpha} \log_2 \left(\frac{\iint_{\mathbb{R}^2} |\mathbf{B}|^\alpha dx dy}{\iint_{\mathbb{R}^2} |\mathbf{B}| dx dy} \right), \quad (24)$$

where \mathbf{B} denotes the estimated beamforming map with α usually being chosen as 3 for evaluation.

- ii) The location bias, which forms a measure of the performance of the precision of the estimated location, being defined as

$$\Delta L = \|L_{\text{estimated}} - L_{gt}\|_2, \quad (25)$$

where L_{gt} and $L_{\text{estimated}}$ denote the true and estimated locations, respectively.

A. Simulated Results

To validate these algorithms, we initially generate two types of signal sources, including both one-point and two-point situations².

1) *One-point simulation source setting*: In the one-point simulation, the signal source is formed as a sinusoidal with frequency 2000 Hz, being generated by the acoustic software Acoular [46], an open source Python library, by using the 56 spiral array microphone array shown in Fig.4. The signal is measured with a sampling frequency of 51200 Hz, with a sampling period of 0.02 s. Each measurement contains a single source placed in the x-y plane at a distance of $z = 2.5$ m from the microphone array, with the individual source locations being sampled from a bivariate uniform distribution to form a suitable dataset of 2000 data points. The DAMAS-FISTA-Net is trained by splitting the data points into a training and validation set containing 70% and 30% of the data, respectively.

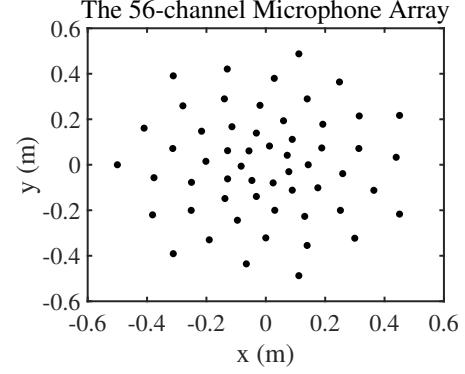


Fig. 4. Spatial position of the 56-channel microphone array.

TABLE I
THE EXPERIMENTAL INDICATORS BY DIFFERENT ALGORITHMS FOR THE SIMULATED ONE-POINT CASE

Methods	$R(\alpha) \downarrow$	$\Delta L \downarrow$	Time \downarrow
DAS [8]	-0.2984	0	0.0655 s
DAMAS [14]	-6.6220	0	11.9595 s
FFT-NNLS [17]	-2.2609	0.0987	0.3547 s
FFT-DFISTA [20]	-4.1067	0.1151	5.2308 s
Acoustic-AlexNet [21]	-	0	0.0721 s
Acoustic-ResNet [23]	-	0.0288	0.0711 s
Acoustic-Net [24]	-	0.0231	0.0588 s
DAMAS-FISTA (Ours)	-6.6431	0	1.0689 s
DAMAS-FISTA-Net (Ours)	-6.5235	0.0028	0.0125 s

TABLE II
THE EXPERIMENTAL INDICATORS BY DIFFERENT ALGORITHMS FOR THE SIMULATED TWO-POINT CASE

Methods	$R(\alpha) \downarrow$	$\Delta L \downarrow$	Time \downarrow
DAS [8]	0.8298	0.0174	0.0654 s
DAMAS [14]	-4.4065	0.0371	8.4904 s
FFT-NNLS [17]	-1.1762	0.1703	0.5524 s
FFT-DFISTA [20]	-2.8402	0.1639	6.3166 s
DAMAS-FISTA (Ours)	-4.6407	0.0314	1.4553 s
DAMAS-FISTA-Net (Ours)	-5.5567	0.0584	0.0124 s

2) *Two-point simulation source setting*: Each measurement results from two signal sources placed in the x-y plane at a distance of $z = 2.5$ m from the microphone array, using the same signal models as in the one-point case. As before, the individual source locations are sampled from a bivariate uniform distribution to form a suitable dataset of 2000 data points. Also in this case, the DAMAS-FISTA-Net is trained by splitting the data points into a training and validation set containing 70% and 30% of the data, respectively.

Table I and Table II summarize the performance of the different methods for the one-point and two-point source cases, respectively. As may be observed, as compared to the DAMAS, the proposed DAMAS-FISTA always achieves better performance and obtains a higher efficiency. The results

¹Our implementation will be available online upon publication.

²The simulated data will be available online upon publication.

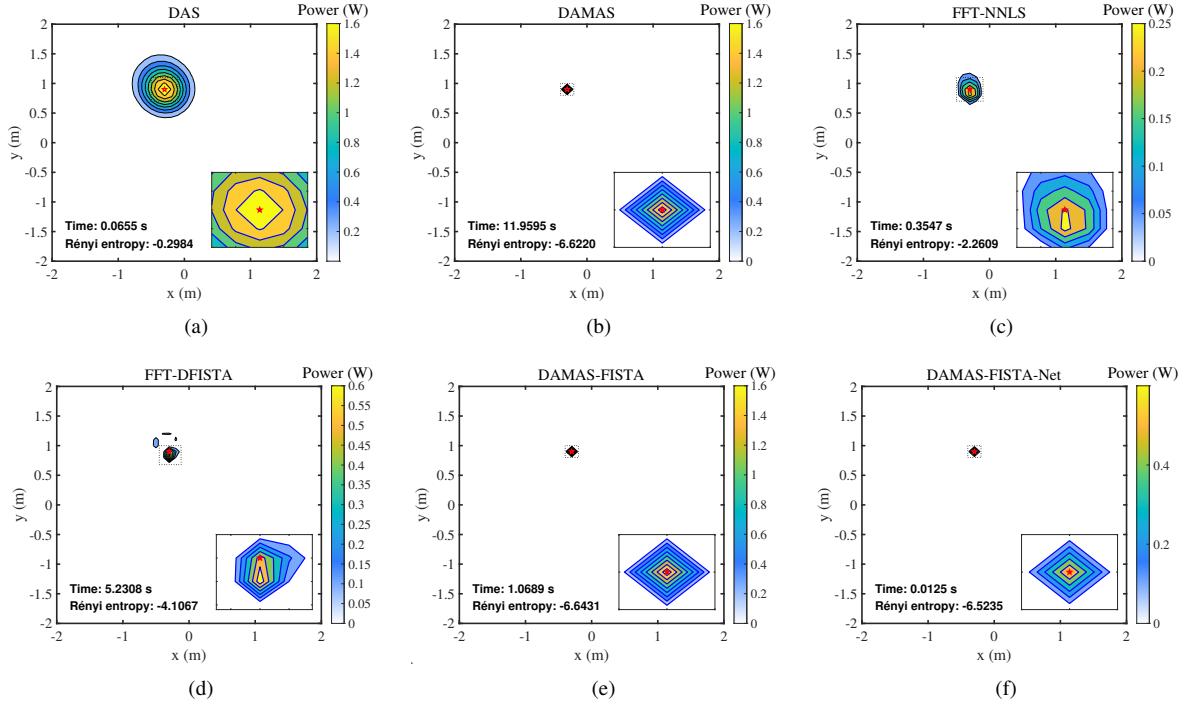


Fig. 5. The beamforming power map of a simulated single-point source by different algorithms: (a) DAS, (b) DAMAS, (c) FFT-NNLS, (d) FFT-DFISTA, (e) DAMAS-FISTA, and (f) DAMAS-FISTA-Net.

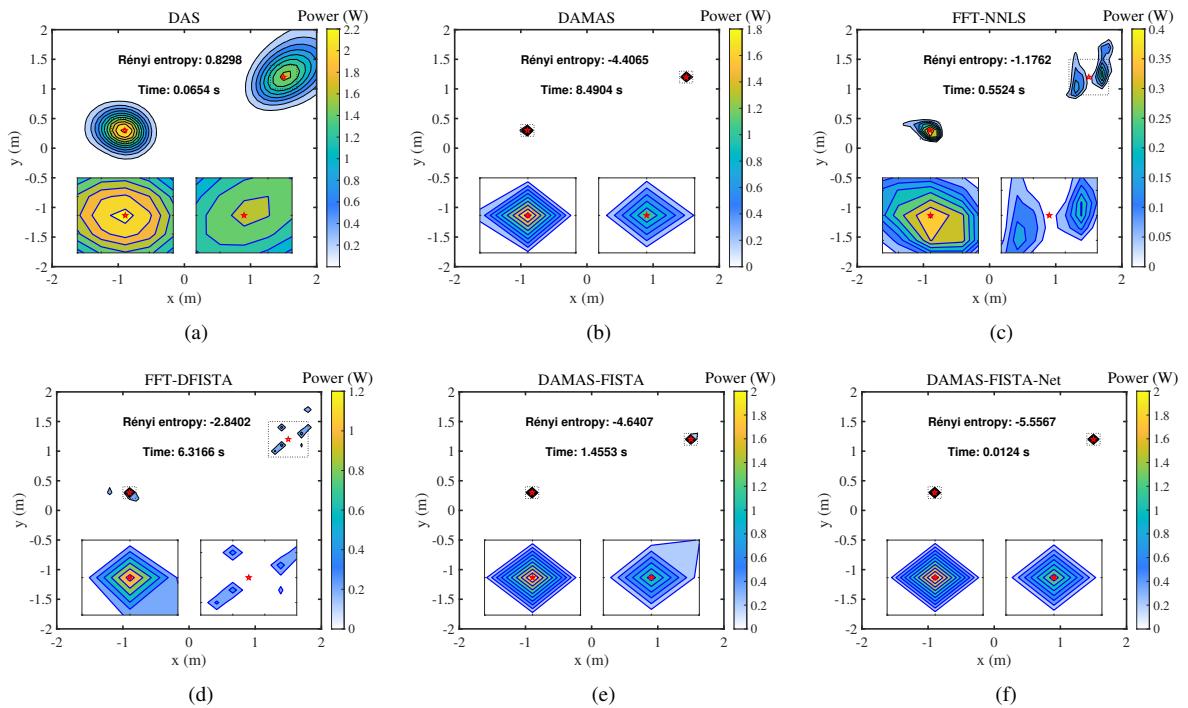


Fig. 6. The beamforming power map of a simulated two-point source by different algorithms: (a) DAS, (b) DAMAS, (c) FFT-NNLS, (d) FFT-DFISTA, (e) DAMAS-FISTA, and (f) DAMAS-FISTA-Net.

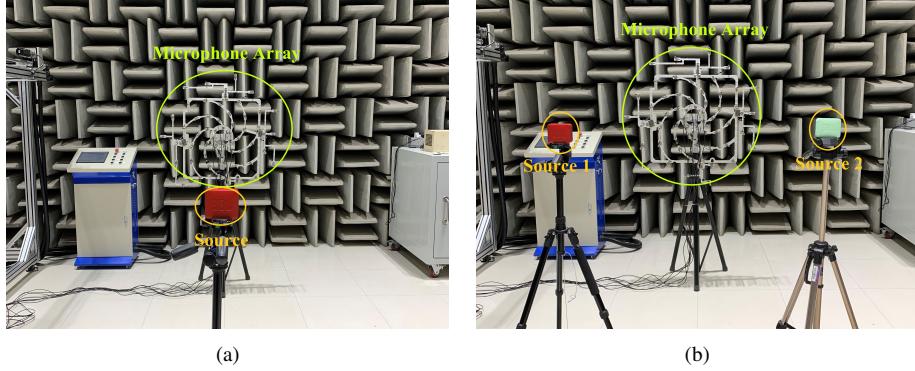


Fig. 7. The layout experimental setup: (a) a one-point source, (b) a two-point source.

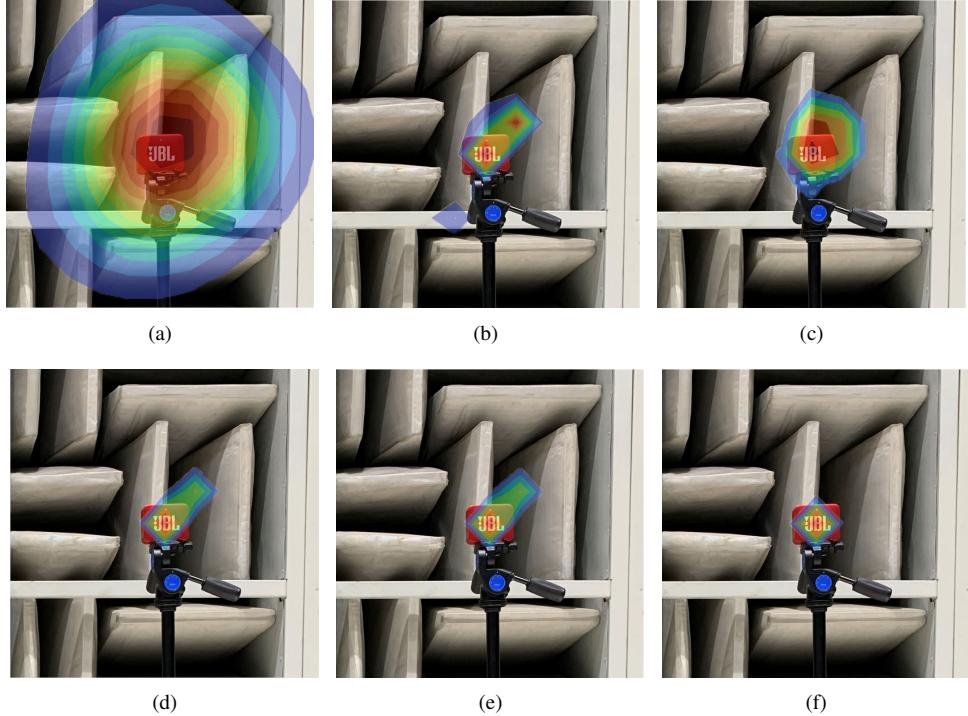


Fig. 8. The multi-modal beamforming map of a real-world one-point source by different algorithms: (a) DAS, (b) DAMAS, (c) FFT-NNLS, (d) FFT-DFISTA, (e) DAMAS-FISTA, and (f) DAMAS-FISTA-Net.

clearly illustrate the lower computational complexity of the network methods, although it is clear that these also have a higher bias. It should also be noted that these methods cannot generate the beamforming map directly, and, as noted, may only be used in the one-point source case. In contrast, the proposed DAMAS-FISTA-Net consistently achieves the fastest runtime performance, while having an acceptable bias, clearly indicating the computational advantage of the proposed network.

Fig.5 and Fig.6 show the corresponding beamforming maps. As may be seen, although the DAS has a low bias and an acceptable runtime, the energy of its beamforming map is not concentrated, resulting in a low-resolution imaging result. As for the FFT-based methods, as mentioned above, the shift-invariant PSF assumption will result in a degree of shift-variance, yielding a larger bias in the estimated location. Although the resolution of DAMAS is high, it may be seen

to require a considerable amount of computations to form the imaging. Conversely, the DAMAS-FISTA-Net is able to construct the high-resolution beamforming map efficiently, making it a good candidate for real-time imaging.

B. Real-World Data

We proceed to validate our method using real-world experiments. Generally, in real-world scenarios, there is not adequate data with precise labels to allow for reliable network training. Fortunately, our network may learn from the simulated dataset, allowing it to be directly applied to the real-world data. In this section, we evaluate the generalization ability of the DAMAS-FISTA-Net by applying the learned net from the simulated one-point and two-point dataset to real-world one-point and two-point measurement data³, respectively. This experiment

³The real-world data will be available online upon publication.

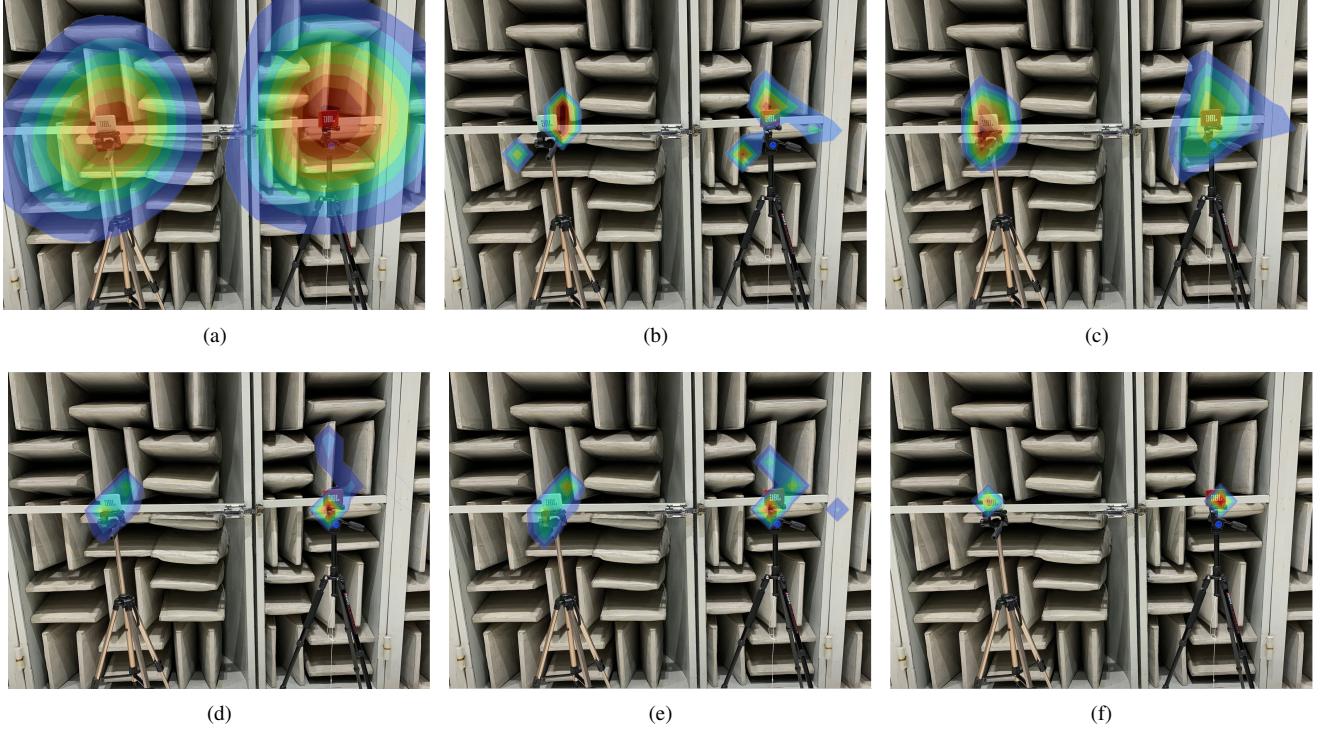


Fig. 9. The multi-modal beamforming map of a real-world two-point source by different algorithms: (a) DAS, (b) DAMAS, (c) FFT-NNLS, (d) FFT-DFISTA, (e) DAMAS-FISTA, and (f) DAMAS-FISTA-Net.

is performed in an anechoic chamber to reduce the effect of noise.

1) *One-point real source setting*: As shown in Fig.7 (a), we consider a one-point source at a distance of $z = 2.5$ m from the microphone array, having a sampling frequency of 51200 Hz. The signals of the source are generated via a Bluetooth speaker transmitting a 2000 Hz tone, collected by a spiral microphone array with 56 sensors (their coordinates are the same as the array in the simulation).

2) *Two-point real source setting*: As shown in Fig.7 (b), we consider a two-point source at a distance of $z = 2.5$ m from the microphone array, using the same experiment setup as in the one-point case.

As can be seen in Table III and Table IV, the reference deep network-based methods yield a large estimated deviation due to their poor generalization ability, whereas the model-based methods basically generate a low-resolution beamforming map and run slower. Conversely, the proposed DAMAS-FISTA-Net can efficiently form a high-resolution beamforming map. Notably, our network is able to break down the barriers between the simulated and the real-world data, i.e., it only learns from the simulated dataset, and is then able to perform well also on the real-world data, demonstrating the network's superior generalization ability. Furthermore, the multi-modal beamforming map fused with optical images are shown in Fig.8 and Fig.9, showing the practical potential of our proposed network.

C. Discussions

In the above experiments, the proposed DAMAS-FISTA-Net is able to produce highly accurate results for both simulated

TABLE III
THE EXPERIMENTAL INDICATORS BY DIFFERENT ALGORITHMS FOR THE REAL-WORLD ONE-POINT CASE

Methods	$R(\alpha) \downarrow$	$\Delta L \downarrow$	Time \downarrow
DAS [8]	-0.0133	0	0.0748 s
DAMAS [14]	-4.4425	0	2.3775 s
FFT-NNLS [17]	-3.3147	0	0.5355 s
FFT-DFISTA [20]	-4.5839	0	2.8353 s
Acoustic-AlexNet [21]	-	0.4243	0.0784 s
Acoustic-ResNet [23]	-	1.6733	0.0797 s
Acoustic-Net [24]	-	0.5439	0.0565 s
DAMAS-FISTA (Ours)	-4.9180	0	1.5821 s
DAMAS-FISTA-Net (Ours)	-5.7627	0	0.0072 s

TABLE IV
THE EXPERIMENTAL INDICATORS BY DIFFERENT ALGORITHMS FOR THE REAL-WORLD TWO-POINT CASE

Methods	$R(\alpha) \downarrow$	$\Delta L \downarrow$	Time \downarrow
DAS [8]	1.0571	0.1000	0.0734 s
DAMAS [14]	-3.0036	0.1207	13.7383 s
FFT-NNLS [17]	-1.5530	0	0.6463 s
FFT-DFISTA [20]	-3.1235	0	4.0994 s
DAMAS-FISTA (Ours)	-3.4238	0.0707	2.0231 s
DAMAS-FISTA-Net (Ours)	-4.6702	0.1000	0.0070 s

and real-world data. Proceeding, we examine the generaliza-

TABLE V
GENERALIZATION STUDIES

Methods	Training	Validation	$R(\alpha) \downarrow$	$\Delta L \downarrow$	Time \downarrow
DAMAS-FISTA-Net _{s1 → s1}	Simulated one-point data	Simulated one-point data	-6.5235	0.0028	0.0125 s
DAMAS-FISTA-Net _{s2 → s1}	Simulated two-point data	Simulated one-point data	-6.0999	0.0049	0.0123 s
DAMAS-FISTA-Net _{s1 → s2}	Simulated one-point data	Simulated two-point data	-5.2107	0.0371	0.0126 s
DAMAS-FISTA-Net _{s2 → s2}	Simulated two-point data	Simulated two-point data	-5.5567	0.0584	0.0124 s
DAMAS-FISTA-Net _{s1 → r1}	Simulated one-point data	Real-world one-point data	-5.7627	0	0.0072 s
DAMAS-FISTA-Net _{s2 → r1}	Simulated two-point data	Real-world one-point data	-5.7849	0	0.0074 s
DAMAS-FISTA-Net _{s1 → r2}	Simulated one-point data	Real-world two-point data	-4.0919	0.100	0.0078 s
DAMAS-FISTA-Net _{s2 → r2}	Simulated two-point data	Real-world two-point data	-4.6702	0.100	0.0070 s

TABLE VI
THE EXPERIMENTAL INDICATORS BY DIFFERENT ALGORITHMS FOR THE SIMULATED ONE-POINT CASE UNDER DIFFERENT NOISE LEVELS

Noise levels	Indicators	DAS [8]	DAMAS [14]	FFT-NNLS [17]	FFT-DFISTA [20]	Acoustic-AlexNet [21]	Acoustic-ResNet [23]	Acoustic-Net [24]	DAMAS-FISTA (Ours)	DAMAS-FISTA-Net (Ours)
-10 dB	$R(\alpha) \downarrow$	-0.2847	-6.4321	-2.2467	-4.0855	-	-	-	-6.5271	-6.5161
	$\Delta L \downarrow$	0	0	0.0983	0.1157	0	0.0290	1.4609	0	0.0028
	Time \downarrow	0.0658 s	10.7178 s	0.3559 s	5.4640 s	0.0722 s	0.0705 s	0.0579 s	1.2101 s	0.0123 s
0 dB	$R(\alpha) \downarrow$	-0.2966	-6.5740	-2.2705	-4.1005	-	-	-	-6.6252	-6.5228
	$\Delta L \downarrow$	0	0	0.0993	0.1162	0	0.0288	1.3063	0	0.0028
	Time \downarrow	0.0661 s	11.1752 s	0.3646 s	5.2969 s	0.0712 s	0.0709 s	0.0584 s	1.2234 s	0.0124 s
10 dB	$R(\alpha) \downarrow$	-0.2983	-6.6102	-2.2673	-4.1046	-	-	-	-6.6410	-6.5234
	$\Delta L \downarrow$	0	0	0.1000	0.1159	0	0.0288	0.7741	0	0.0028
	Time \downarrow	0.0665 s	11.5698 s	0.3653 s	5.1992 s	0.0716 s	0.0715 s	0.0588 s	1.1177 s	0.0124 s
Noise-free	$R(\alpha) \downarrow$	-0.2984	-6.6220	-2.2609	-4.1067	-	-	-	-6.6431	-6.5235
	$\Delta L \downarrow$	0	0	0.0987	0.1151	0	0.0288	0.0231	0	0.0028
	Time \downarrow	0.0655 s	11.9595 s	0.3547 s	5.2308 s	0.0721 s	0.0711 s	0.0588 s	1.0689 s	0.0125 s

TABLE VII
COMPARISON OF DIFFERENT NETWORK DEPTHS IN THE DAMAS-FISTA-NET

Network Depths	$L = 3$	$L = 4$	$L = 5$	$L = 6$	$L = 7$
$R(\alpha) \downarrow$	-6.4632	-6.5075	-6.5235	-6.5068	-6.4744
$\Delta L \downarrow$	0.0097	0.0027	0.0028	0.0057	0.0057
Time \downarrow	0.0103 s	0.0114 s	0.0125 s	0.0133 s	0.0145 s

tion ability of DAMAS-FISTA-Net and its sensitivity to noise. The effects of different network depths are also evaluated.

1) *Generalization Studies*: As evaluated in Section V-B, the proposed network is able to generalize well from simulated data to real-world data. Here, we further test its generalization ability in the multi-source situation, applying the learned network from the simulated one-point source dataset to simulated two-point source data. Table V summarizes the indicators of the different algorithms in constructing the beamforming map of the different situations. Notably, our network can be seen to achieve competitive performance even when the training and validation data are biased (e.g., the network learns from the simulated one-point source dataset but is validated on the two-point source scene), again revealing its robust generalization ability. The reason for this robustness may be due to our hybrid approach, which combines both the model-based scheme and the deep learning framework, allowing the network to inherit relevant domain knowledge. In contrast, the deep network-based methods can only form an estimate for the one-point source situation due to their network output settings, which may severely limit their applicability.

2) *Sensitivity to Noise*: We next extend the proposed method to process array data with Gaussian white noise in order to demonstrate the robustness of the DAMAS-FISTA-Net. Here, we add -10 dB, 0 dB, and 10 dB Gaussian white noise to the signals in the validation set, respectively. Table VI shows the comparison results under different noise levels. As can be observed, the acoustic imaging performance of the network is only mildly influenced by the added noise, e.g., the Rényi entropy is only increased by 0.0074 by the addition of noise with noise levels -10 dB. Notably, we do not need to retrain the DAMAS-FISTA-Net for different noise levels, implying that the learned network has a certain robustness to the influence of noise.

3) *Evaluation for Different Network Depths*: Finally, we investigate the effect of different network depths on the network's performance. The testing beamforming results in Table VII clearly show that within a certain number of layers, using more deeper layers could effectively improve the imaging quality, although at the cost of a larger computational overhead. However, it can also be seen that adding more layers to the network, e.g., when the network depths, L , is larger than

5, may also degrade the network's performance, which may be due to the overfitting of the dataset.

VI. CONCLUSION

In this paper, we propose a computationally efficient model-based acoustic beamforming method, termed the DAMAS-FISTA. Inspired by the model-based deep learning theory, we further propose a hybrid structured model-based deep learning network for real-time imaging, dubbed the DAMAS-FISTA-Net, which is able to process the raw measurement data directly. In order to achieve real-time execution and generalizable performance, the network exploits the structure of the model-based estimator, enabling the network to inherit physical knowledge and learn the underlying physical properties of the acoustic scene. The performance of the proposed network is via extensive simulated and real-world experiments shown to offer preferable performance as compared to alternative approaches, while also indicating the potential practical applicability of the estimator.

REFERENCES

- [1] J. A. Wall, L. J. McDaid, L. P. Maguire, and T. M. McGinnity, "Spiking neural network model of sound localization using the interaural intensity difference," *IEEE transactions on neural networks and learning systems*, vol. 23, no. 4, pp. 574–586, 2012.
- [2] R. Savitha, S. Suresh, and N. Sundararajan, "Projection-based fast learning fully complex-valued relaxation neural network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 4, pp. 529–541, 2013.
- [3] J. Dávila-Chacón, J. Liu, and S. Wermter, "Enhanced robot speech recognition using biomimetic binaural sound source localization," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 1, pp. 138–150, 2018.
- [4] D. Salvati, C. Drioli, and G. L. Foresti, "Diagonal unloading beamforming in the spherical harmonic domain for acoustic source localization in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2001–2012, 2020.
- [5] K. SongGong, W. Wang, and H. Chen, "Acoustic source localization in the circular harmonic domain using deep learning architecture," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2475–2491, 2022.
- [6] H. Sun and L. Demanet, "Beyond correlations: Deep learning for seismic interferometry," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [7] D. H. Johnson and D. E. Dudgeon, *Array signal processing: concepts and techniques*. Simon & Schuster, Inc., 1992.
- [8] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE assp magazine*, vol. 5, no. 2, pp. 4–24, 1988.
- [9] C. M. Sparrow, "On spectroscopic resolving power," *The Astrophysical Journal*, vol. 44, p. 76, 1916.
- [10] R. Merino-Martínez, P. Sijtsma, M. Snellen, T. Ahlefeldt, J. Antoni, C. J. Bahr, D. Blacodon, D. Ernst, A. Finez, S. Funke *et al.*, "A review of acoustic imaging methods using phased microphone arrays," *CEAS Aeronautical Journal*, vol. 10, no. 1, pp. 197–230, 2019.
- [11] J. Högbom, "Aperture synthesis with a non-regular distribution of interferometer baselines," *Astronomy and Astrophysics Supplement Series*, vol. 15, p. 417, 1974.
- [12] P. Sijtsma, "Clean based on spatial source coherence," *International journal of aeroacoustics*, vol. 6, no. 4, pp. 357–374, 2007.
- [13] P. Sijtsma, R. Merino-Martínez, A. M. Malgoezar, and M. Snellen, "High-resolution clean-sc: Theory and experimental validation," *International Journal of Aeroacoustics*, vol. 16, no. 4-5, pp. 274–298, 2017.
- [14] T. F. Brooks and W. M. Humphreys, "A deconvolution approach for the mapping of acoustic sources (damas) determined from phased microphone arrays," *Journal of sound and vibration*, vol. 294, no. 4-5, pp. 856–879, 2006.
- [15] R. Dougherty, "Extensions of damas and benefits and limitations of deconvolution in beamforming," in *11th AIAA/CEAS aeroacoustics conference*, 2005, p. 2961.
- [16] T. Brooks and W. Humphreys, "Extension of damas phased array processing for spatial coherence determination (damas-c)," in *12th AIAA/CEAS Aeroacoustics Conference (27th AIAA Aeroacoustics Conference)*, 2006, p. 2654.
- [17] K. Ehrenfried and L. Koop, "Comparison of iterative deconvolution algorithms for the mapping of acoustic sources," *AIAA journal*, vol. 45, no. 7, pp. 1584–1595, 2007.
- [18] O. Lylloff, E. Fernández-Grande, F. Agerkvist, J. Hald, E. Tiana Roig, and M. S. Andersen, "Improving the efficiency of deconvolution algorithms for sound source localization," *The journal of the acoustical society of America*, vol. 138, no. 1, pp. 172–180, 2015.
- [19] G. Herold, T. F. Geyer, and E. Sarradj, "Comparison of inverse deconvolution algorithms for high-resolution aeroacoustic source characterization," in *23rd AIAA/CEAS aeroacoustics conference*, 2017, p. 4177.
- [20] X. Ding, H. Liang, A. Jakobsson, X. Tu, and Y. Huang, "High-resolution source localization exploiting the sparsity of the beamforming map," *Signal Processing*, vol. 192, p. 108377, 2022.
- [21] A. Reiter and M. A. L. Bell, "A machine learning approach to identifying point source locations in photoacoustic data," in *Photons Plus UltraSound: Imaging and Sensing 2017*, vol. 10064. International Society for Optics and Photonics, 2017, p. 100643.
- [22] D. Allman, A. Reiter, and M. A. L. Bell, "Photoacoustic source detection and reflection artifact removal enabled by deep learning," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1464–1477, 2018.
- [23] A. Kujawski, G. Herold, and E. Sarradj, "A deep learning method for grid-free localization and quantification of sound sources," *The Journal of the Acoustical Society of America*, vol. 146, no. 3, pp. EL225–EL231, 2019.
- [24] G. Zhou, H. Liang, X. Ding, Y. Huang, X. Tu, and S. Abbas, "Acousticnet: A novel neural network for sound localization and quantification," 2022.
- [25] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.
- [26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012.
- [27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [28] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
- [29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [30] X. Ding, Y. Guo, G. Ding, and J. Han, "Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1911–1920.
- [31] S. Y. Lee, J. Chang, and S. Lee, "Deep learning-based method for multiple sound source localization with high resolution and accuracy," *Mechanical Systems and Signal Processing*, vol. 161, p. 107959, 2021.
- [32] L. Feng, M. Zan, L. Huang, and Z. Xu, "A double-step grid-free method for sound source identification using deep learning," *Applied Acoustics*, vol. 201, p. 109099, 2022.
- [33] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proceedings of the 27th international conference on international conference on machine learning*, 2010, pp. 399–406.
- [34] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 57, no. 11, pp. 1413–1457, 2004.
- [35] Z. Wang, Q. Ling, and T. Huang, "Learning deep 10 encoders," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.
- [36] Y. Yang, J. Sun, H. Li, and Z. Xu, "Deep admm-net for compressive sensing mri," in *NIPS*, 2016, pp. 10–18.
- [37] ———, "Admm-csnet: A deep learning approach for image compressive sensing," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 3, pp. 521–538, 2018.
- [38] M. SIMEONI, S. Kashani, P. Hurley, and M. Vetterli, "Deepwave: A recurrent neural-network for real-time acoustic imaging," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

- [39] J. Zhang and B. Ghanem, "Ista-net: Interpretable optimization-inspired deep network for image compressive sensing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1828–1837.
- [40] J. Xiang, Y. Dong, and Y. Yang, "Fista-net: Learning a fast iterative shrinkage thresholding network for inverse problems in imaging," *IEEE Transactions on Medical Imaging*, vol. 40, no. 5, pp. 1329–1339, 2021.
- [41] E. Sarradj, "Three-dimensional acoustic source mapping with different beamforming steering vector formulations," *Advances in Acoustics and Vibration*, vol. 2012, 2012.
- [42] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 249–256.
- [43] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.
- [44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [45] L. Stanković, "A measure of some time-frequency distributions concentration," *Signal Processing*, vol. 81, no. 3, pp. 621–631, 2001.
- [46] E. Sarradj, "Acoulear acoustic testing and source mapping software, 2019," *Online [Accessed 22 December 2021]* <http://www.acoulear.org>.