Adaptive Metasurface-Based Acoustic Imaging using Joint Optimization

Yongjian Fu[1], Yongzhao Zhang[2], Yu Lu[2], Lili Qiu[3,4], Yi-Chao Chen[2], Yezhou Wang[2], Mei Wang[4], Yijie Li[2], Ju Ren*[5,6], Yaoxue Zhang[5,6]
[1]Central South University, China, [2]Shanghai Jiaotong University, China, [3]Microsoft Research Aisa, China, [4]UT Austin, USA, [5]Tsinghua University, China, [6]Zhongguancun Laboratory, China
Email:fuyongjian@csu.edu.cn
{zhangyongzhao,yulu01,yichao,yezhouwang,yijieli}@sjtu.edu.cn
liliqiu@microsoft.com,meiwang@utexas.edu,{renju,zhangyx}@tsinghua.edu.cn

ABSTRACT

Acoustic imaging is attractive due to its ability to work under occlusion, different lighting conditions, and privacy-sensitive environments. Existing acoustic imaging methods require large transceiver arrays or device movement, which makes it challenging to use in many scenarios. In this paper, we develop a novel acoustic imaging system for low-cost devices with few speakers and microphones without any device movement. To achieve this goal, we leverage a 3D-printed passive acoustic metasurface to significantly enhance the diversity of the measurement data, thereby improving the imaging quality. Specifically, we jointly design the transmission signal, transceivers' beamforming weights, metasurface, and imaging algorithm to minimize the imaging reconstruction error in an end-to-end manner. We further develop a scheme to dynamically adapt the imaging resolution based on the distance to the target. We implement a system prototype. Using extensive experiments, we show that our system yields high-quality images across a wide range of scenarios.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**; **Ubiquitous and mobile computing systems and tools**.

KEYWORDS

Acoustic imaging, compressive sensing, joint optimization.

ACM Reference Format:
Yongjian Fu[1], Yongzhao Zhang[2], Yu Lu[2], Lili Qiu[3,4], Yi-Chao Chen[2], Yezhou Wang[2], Mei Wang[4], Yijie Li[2], Ju Ren*[5,6], Yaoxue Zhang[5,6]. 2024. Adaptive Metasurface-Based Acoustic Imaging using Joint Optimization. In *The 22nd Annual International Conference on Mobile Systems, Applications and Services (MOBISYS '24), June 3–7, 2024, Minato-ku, Tokyo, Japan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3643832.3661863

*Corresponding author
Yongjian Fu, Yongzhao Zhang, Yu Lu, Yezhou Wang, Yijie Li did this work as interns at Microsoft Research Asia and Yi-Chao Chen did this work as a visiting researcher at Microsoft Research Asia.

MOBISYS '24, June 3–7, 2024, Minato-ku, Tokyo, Japan

ACM ISBN 979-8-4007-0581-6/24/06…$15.00
https://doi.org/10.1145/3643832.3661863

1 INTRODUCTION

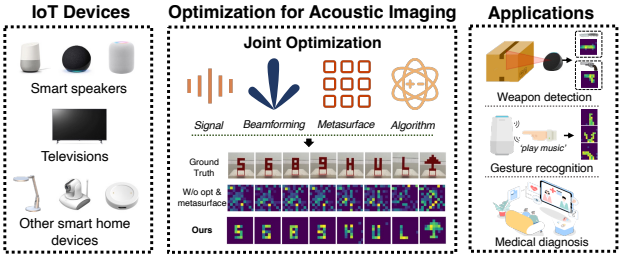


Figure 1: Illustration of MAJIC.

Motivation: Acoustic imaging uses sound waves reflected by a target object to reconstruct images of its shape. It complements widely used cameras since sound waves work under different lighting conditions and can penetrate through certain materials with high energy efficiency. Unlike RF-based imaging techniques, such as RFID [46], Wi-Fi [19, 20, 23, 35], mmWave [4, 7, 37], and Terahertz [14, 43], acoustic imaging can be easily deployed by using low-cost microphones and speakers on commercial devices. Such wide availability of speakers and microphones makes acoustic imaging attractive for many applications, including gesture recognition [42], activity detection [24, 29], and weapon detection [28].

To accurately reconstruct images, conventional acoustic imaging systems require large arrays of transceivers (e.g., 40 speakers and 40 microphones) to ensure adequate spatial sampling and signal-to-noise ratio (SNR), which results in dedicated, expensive, and complex hardware designs [15, 16, 21]. In contrast, existing commercial IoT devices typically have only a small number of transceivers to limit cost and energy consumption. For example, the latest Amazon Echo has seven microphones and three speakers [1], and the latest Apple HomePod has six microphones and eight speakers [17]. While they already have the largest number of transceivers in the market, it is still far from what is necessary to produce high-quality images.

In order to address the challenge of limited transceivers, AIM [28] made an early stride by introducing synthetic aperture radar (SAR). However, its reliance on mechanical movement spanning tens of centimeters makes it challenging to deploy in static devices (*e.g.*smart speakers). SPiDR [5] employed a 3D-printed metamaterial stencil to enhance spatial perception by dividing signals from the speakers into multiple replicas. Although this presented an intriguing initial solution, it uses a rather basic image reconstruction algorithm and

stencil design and requires multiple frames through robot motion to get clear images. Therefore, despite these advances, realizing acoustic imaging on low-cost IoT devices without device movement remains an open challenge.

Our approach: In this paper, as shown in Figure 1, we develop Metasurface Acoustic based Joint optimized Imaging sCheme (MAJIC), a low-cost and high-quality acoustic imaging system that employs a passive acoustic metasurface to achieve accurate acoustic imaging without device movement. We use a passive metasurface to turn a small transceiver array into a larger one, and jointly design the transmission signal, transceivers' beamforming, metasurface, and imaging algorithm to optimize the image reconstruction error.

Specifically, we cast the acoustic imaging problem as a linear inverse problem of inferring pixel values in an image, denoted as x, based on the received signal $y = Ax$, where A is the measurement matrix determined by the transmission signal and the channel. Our objective is to minimize the mean square error between the estimated image and the ground truth image. The imaging accuracy depends on (i) *the choice of measurement matrix A*, and (ii) *the imaging algorithm*.

For (i), we observe that the imaging quality can be improved by increasing the effective rank of A (*i.e.*, the number of singular values that account for 99% of the energy). In fact, the effective rank of A measures the similarity among spatial sampling points and is commonly improved by using more transceivers and frequencies for measurement. However, when the number of transceivers is fixed, the effective rank of A quickly saturates as the number of measurement frequencies increases. The rank at the saturation point is primarily governed by the number of transceivers. Therefore, unless a large number of speakers and microphones (*e.g.*, 40 of each) are used, the effective rank of A is much smaller than the image size, which significantly limits the imaging accuracy.

To tackle this challenge, we design a low-cost passive acoustic metasurface. The metasurface is composed of many sub-wavelength cells. Each cell can be considered as a small antenna independently controlling the outgoing acoustic signal. By carefully adjusting the amplitude and phase of each cell, the metasurface can be used to modulate the transceiver's wavefront, thereby increasing the diversity of the channel. With the assistance of such a metasurface, the rank of the measurement matrix can be effectively increased even under a small number of transceivers. We show that even a random metasurface can significantly improve the imaging quality.

To further enhance the effectiveness of the metasurface, we jointly design the transmission signal, transceivers' beamforming, metasurface, and imaging algorithm to optimize the imaging reconstruction error in an end-to-end manner. We develop an iterative process that first refines the signal, beamforming, and metasurface and then reconstructs the image based on the given configuration and iterates until the reconstructed image converges.

For (ii), many algorithms have been proposed to solve linear inverse problems $y = Ax$. However, two significant challenges render the existing algorithms inadequate for producing high-quality images: 1) even using an optimized metasurface and phased arrays, the effective rank of the measurement matrix is still insufficient to uniquely determine the image since the number of pixels is typically much larger the rank of the measurement matrix, and 2) real measurement may contain significant noise due to background noise, hardware artifacts, and angular deviations, which can significantly degrade the imaging quality.

One way to handle insufficient constraints is to use compressive sensing, which introduces a regularization term, such as sparsity of the unknowns, and use Alternating Direction Method of Multipliers (ADMM) to solve it. However, such explicit priors may not strictly hold in general. Instead, we propose a physics-informed image reconstruction algorithm to reconstruct an image. Inspired by [44], we unroll the ADMM into a neural network, which treats each iteration in ADMM as a neural layer. We go beyond [44] by introducing learnable neural priors and hyperparameters in each neural layer and refining the priors using image data. Turning ADMM into a neural network and introducing learnable priors allow us to exploit the unique characteristics in the underlying data, enhance imaging quality, and speed up convergence, while maintaining the white box design. To further enhance robustness against noise, we concatenate the neural network with a refined network to denoise the image and compensate for the mismatch caused by imperfect modeling.

Moreover, we find that increasing the distance between the target and transceiver array significantly degrades the imaging quality due to reduced angular resolution. To avoid sharp decay in imaging quality, we dynamically adapt the imaging resolution according to the distance to the target.

We implement our joint optimization algorithm, then develop a prototype by 3D printing the optimized metasurface and assembling it with an array of commodity speakers and microphones. Refer to [11] for our demo video.

Our contributions can be summarized as follows:

- We develop an end-to-end framework that jointly optimizes transmission signal, beamforming, metasurface, and imaging algorithm to minimize the error of reconstructed images. To the best of our knowledge, this is the first system that realizes high-quality acoustic imaging on low-cost IoT devices without device or target movement.
- We design a novel physics-informed image reconstruction algorithm to enhance the quality of images especially under high noise, and achieve good quality and fast convergence.
- We propose an effective scheme to adapt imaging resolution across varying distances.
- We implement a prototype and conduct extensive evaluation to demonstrate the effectiveness of the imaging capabilities. Our results demonstrate that MAJIC out-performs the baselines (w/o metasurface) by 22.6% to 83.1% in root mean square error (RMSE) reduction. We further show our imaging algorithm out-performs the traditional ADMM (using our optimized metasurface and beamforming) by 66.5% in RMSE reduction, and the adaptive resolution scheme increases imaging distance from $30cm$ to $135cm$. Our system out-performs SPiDR [5] by 88.42% and out-performs [52] by 80.75%. Moreover, our evaluations in various real-world scenarios, including occlusion, gesture recognition, and scene migration, demonstrate the practical applicability and effectiveness of MAJIC.

2 RELATED WORK

Acoustic sensing: Acoustic signals are increasingly used for wireless sensing as they are supported by many devices, such as smartphones, smart speakers, computers, and smart TVs etc. Various

Table 1: Summary of acoustic imaging systems using IoT devices.

	Microphones	Speakers	Distance	Movement	RMSE
AIM [28]	1	1	<50cm	Yes	\
SPiDR[5]	1	1	<20cm	Yes	≈0.2
MAJIC	4	6	<135cm	No	0.04

acoustic sensing systems have been developed based on correlation (*e.g.*, [34, 36]), FMCW (*e.g.*, [26, 33, 48]), Doppler shift (*e.g.*, [51]), phase (*e.g.*, [13, 49]), or Angle of Arrival (*e.g.*, [29, 41, 47]). Machine learning has also been applied to acoustic sensing (*e.g.*, [27, 29]). Existing acoustic sensing work mostly focuses on sensing location and movement trajectory by treating the target as a single point. Imaging requires more detailed geometry shape, which is more challenging than tracking.

Acoustic imaging: Acoustic imaging has been widely studied, because its greater integration, compactness and portability of IoT devices compared to RF signals, such as Wi-Fi [35] and mmWave [7]. However, achieving accurate acoustic imaging on IoT devices is not trivial. For example, [15] leverages 120 transceivers and $3.2kHz$ bandwidth to achieve high resolution. Since most commodity IoT devices have only a few speakers and microphones, we cannot directly apply these approaches. To enable acoustic imaging on IoT devices, some advanced works have been proposed, as shown in the Table 1. AIM [28] brings acoustic imaging to a mobile phone by letting a user hold a phone to swipe across an object to simulate multiple transceivers. However, it is not convenient and sometimes not feasible to move the mobile for imaging purpose (*e.g.*, in smart speakers). SPiDR [5] creates a stencil and passes transmission signals from the speaker through multiple tubes, enhancing spatial sensing using multipath encoding, and further uses robot movement to stack multiple frames for imaging. Different from [5], we combine the powerful wavefront shaping of the metasurface with dynamic adaptation of a small transceiver array to create rich beam patterns and generate diverse measurement data without movement of a device or target. We further design effective imaging reconstruction algorithm and adapt resolution to achieve high imaging quality under high noise and from a large distance.

Acoustic metasurface: Acoustic metasurface research primarily focuses on passive designs, often using coiling-up designs, Helmholtz resonators, or membrane types. Active acoustic metasurfaces, as indicated in prior studies [10, 18, 22], tend to be larger and more costly. For affordability and simple deployment, we use the coiling-up passive structure, chosen for its compact design and 3D printing compatibility. Unlike the bulkier Helmholtz resonators or the more complex membrane structures, it's easier to produce. The coiling-up structure [25, 31, 32] manipulates the phase of outgoing acoustic signals by configuring varied coiled paths within each cell. For instance, to achieve beamforming in a certain direction, we adjust the lengths of coiled paths across all cells to compensate for phase differences in incoming signals. Vari-Sound [31] designed 16 distinct unit cell structures corresponding to phase offsets from 0 to 15 times $2\pi/16$. The coiling-up structure is simple and easy to manufacture through 3D printing and is low cost and easy to implement. Therefore, considering the ubiquity and low cost of IoT devices, we adopt coiling-up structure as the unit cell, assembling various unit cells into a metasurface based on our optimization outcomes. If our optimization indicates that cell (i, j) should have a phase offset of p, we position the unit cell with the phase offset closest to our desired value at that location. Different from [31], which focuses on unit cell design (*i.e.*, microscopic design), we focus on optimizing the metasurface phase profile (*i.e.*, macroscopic design). Different from [52], which focus on increasing SNR in a certain direction, we focus on optimizing the imaging quality by creating rich beam patterns (*i.e.*, high rank of the measurement matrix).

3 BACKGROUND AND PRELIMINARY

In this section, we first introduce the basic idea of acoustic imaging using compressive sensing and the concept of acoustic metasurfaces. Then, we provide the intuition behind using acoustic metasurfaces for imaging.

3.1 Imaging using Compressive Sensing

Consider the signal $x_i(t)$ from N pixels of the target arrive at M microphones at a distance d_{nm}. The signal received by each microphone can be derived as follows:

$$y_m(t) = \sum_{n=0}^{N-1} \frac{e^{-j2\pi f \frac{d_{nm}}{c}}}{d_{nm}} x_n(t - d_{nm}/c) \tag{1}$$

where $y_m(t)$ is the sound pressure at the m^{th} microphone at time t, c is the acoustic signal propagation speed, and $x_n(t)$ is the reflected signal from the n^{th} target at time t.

The above relationship can also be captured using the following matrix form:

$$y = Ax + e \tag{2}$$

where A stands for a $M \times N$ measurement matrix, defined as: $A_{nm} = \frac{e^{-j2\pi f \frac{d_{nm}}{c}}}{d_{nm}}$, y is an $M \times 1$ vector representing the signal from all microphones, x is our target image and reshaped to an $N \times 1$ vector, and e represents Additive White Gaussian Noise (AWGN). x is a greyscale vector, which denotes the fraction of signal that is reflected by the target at each position. Due to the limited number of transceivers available on the low-cost IoT devices, acoustic imaging is typically an under-constrained inference problem, which may have an infinite number of solutions.

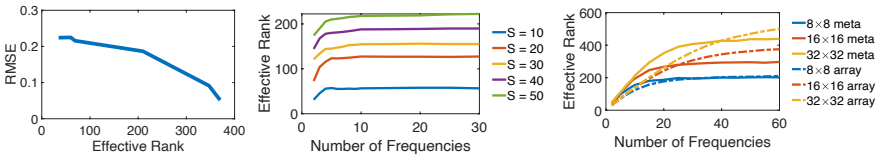
Compressive sensing can be used to reconstruct the image x if x or some transformation of x (*e.g.*, discrete cosine transformation (DCT)) is sparse. In this case, the image can be reconstructed by solving the following optimization problem:

$$\arg\min_x \|x\|_1 \quad s.t.\ \|Ax - y\|_2 < \epsilon \tag{3}$$

where $\|x\|_1$ leverages the sparsity prior and $\|Ax - y\|_2 < \epsilon$ enforces the accuracy of the reconstructed image. One can use iterative algorithms to solve this problem. In practice, the reconstruction error depends on the measurement matrix, noise, and the number of unknowns (i.e., pixels).

3.2 Measurement Matrix

The measurement matrix in the imaging task is determined by the channel, as shown in Eq. 1. The channel is dictated by two key parameters: frequency f and distance d_{nm} between each target position



(a) Impact of effective rank (b) Number of transceivers (c) Metasurface size

Figure 2: Properties of effective rank: (a) Higher effective rank lowers RMSE. (b) Changes with transceiver count (speakers, mics) at different frequencies. (c) Affected by frequency and metasurface/array size, using 6 speakers/mics for metasurfaces.

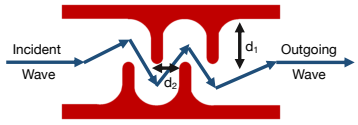
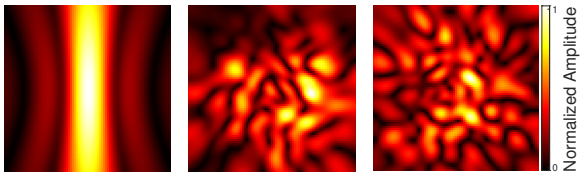


Figure 3: The metasurface unit design, where its response is controlled by d_1 and d_2.



(a) (b) (c)

Figure 4: Acoustic channel produced by (a) 6×1 speaker array (effective rank 54), (b) 6×1 speaker array with metasurface (effective rank 302), (c) 16×16 speaker array (effective rank 351).

and receiver. If the measurement matrix A satisfies the Restricted Isometry Property (RIP) [9, 39], then compressive sensing can be applied to accurately reconstruct the image.

In practice, it can be difficult to verify if a matrix satisfies the RIP property. Instead, we can consider using the rank of A as an approximation [8]. A larger rank suggests more linearly independent constraints for solving the optimization problem, and is preferred.

We conduct an empirical evaluation to assess the impact of measurement matrix rank A on image reconstruction error, measured using Root Mean Square Error (RMSE). As illustrated in Figure 2 (a), as the effective rank of matrix A increases, the RMSE consistently decreases. Here, the effective rank is defined as the count of top singular values that collectively represent 99% of the total energy. This measure is preferred over strict rank determination because small singular values, although non-zero, may contribute minimal new information to the image reconstruction process. The observed monotonic relationship between RMSE and effective rank justifies the use of effective rank as a metric to quantify the effectiveness of the measurement matrix. Such well-known ill-posed problem leads to the need for carefully designed solutions, and the problem similarly often arises in channel estimation tasks [38].

There are two methods to increase the rank of the measurement matrix: increasing the number of transccivers or using more frequencies. As shown in Figure 2 (b), the effective rank increases with the number of frequencies, as well as the number of transccivers. But the benefit of increasing the number of frequencies tapers off after reaching a threshold because using more frequencies than the threshold yields linearly dependent constraints. Therefore, the performance of traditional acoustic imaging is limited by the number of transceivers, which is rather small on a typical IoT device.

However, to reconstruct a reasonably sized image with depth information, such as $10 \times 10 \times 10$, we require both 40 speakers and 40 microphones over 20 frequencies, which is unaffordable. In comparison, a typical IoT device with up to 6 speakers and microphones has a rank of only 38 using 20 frequencies. Further increasing the number of frequencies does not reduce the RMSE. These results indicate that it is necessary to develop additional mechanisms to generate an appropriate measurement matrix for acoustic imaging.

3.3 Impact of Acoustic Metasurface

Acoustic metasurface shapes acoustic fields using a carefully designed yet low-cost passive physical structure. Metasurface design involves two parts: 1) microscopic design, which determines a unit cell structure, and 2) macroscopic design, which determines the phase map across the entire metasurface (*i.e.*, which type of unit cell is placed at each location of the metasurface). We use the unit cell structure proposed in [31, 32] for the microscopic design. The microscopic design details are shown in Figure 3, in which there are two distance values d_1 and d_2 that determine its amplitude and phase response. It discretizes the unit cells into 16 types (*e.g.*, 0, $1/162\pi$, $2/162\pi$ ..., $15/162\pi$). Our joint optimization framework introduced in Section 4 will determine the macroscopic structure, which places the cells at appropriate positions to generate the desired beams.

The acoustic metasurface enhances the degree of freedom in both spatial and frequency domain for us to control the measurement matrix. Figure 2 (c) compares the effective rank with and without metasurfaces. We observe that using a metasurface significantly increases the rank under the same number of transceivers. For example, after the convergence, the effective rank increases from 60 without metasurface to 204, 297, and 442 using 8x8, 16x16, 32x32 metasurfaces, respectively. Their ranks come close to using the corresponding sizes of phased array. For example, the ranks of 8x8, 16x16, 32x32 speakers are 210, 379, and 509, respectively.

Figure 4 further compares some examples of acoustic channels with and without metasurfaces in the x-y plane. A small phased array alone has a very coarse beam pattern (Figure 4 (a)). Adding a 16x16 metasurface enriches the beam pattern (Figure 4 (b)). The richness of its beam pattern is just slightly lower than that of a 16x16 speaker array (Figure 4 (c)). These observations indicate that using passive metasurfaces effectively turns a small-scale phased array into a larger array while maintaining very low costs, thereby increasing the effective rank.

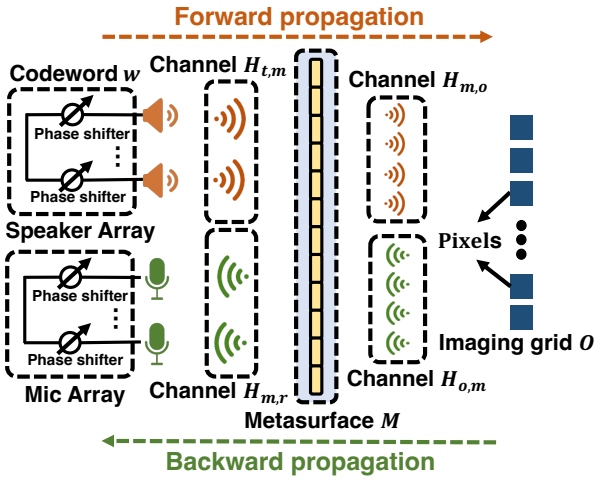


Figure 5: Channel modeling for imaging.

4 MAJIC DESIGN

In this section, we introduce our problem formulation and describe our system design.

4.1 Problem Formulation

As shown in Figure 5, let $H_{t,m}$ denote the acoustic channel from the speaker array to the metasurface, where $H_{t,m}(i,j)$ is the channel from the i-th speaker to the j-th metasurface cell. Similarly, we define $H_{m,r}$ as the channel from the metasurface to the microphone array. Suppose the target of interest is within a given 3D imaging area O. This is not a strong assumption, as the 3D imaging area can be redefined or sufficiently large enough to ensure it includes the target. Then we define another channel matrix $H_{m,o}$, where $H_{m,o}(j,k)$ denotes the channel from the j-th metasurface cell to the k-th grid in the 3D area. Note that if the k-th grid does not contain the target, there is no reflection from this grid. Similarly, we define $H_{o,m}$, where $H_{o,m}(k,j)$ denotes the channel from the k-th grid in the 3D area to the j-th metasurface cell. Finally, we let w denote the speakers' beamforming for a specific frequency.

Based on the above definitions, the received signals after going through speakers' beamforming, metasurface manipulation, and microphones' combining become as follows:

$$R_m = H_{m,r}M \cdot H_{o,m}O \cdot H_{m,o}M \cdot H_{t,m}w + e \tag{4}$$

M denotes the manipulation of each metasurface cells on acoustic signals (*i.e.*, signal attenuation and phase delay) and $\cdot$ denotes dot product. We use a dot product between the metasurface M and incoming signal because each metasurface cell manipulates multi-path signals coming through the cell in the same way regardless of which path the signal comes from. Similarly, we use a dot product to capture the interaction between the incoming signal and object. If the metasurface's configurations are fixed, one can infer our target object O based on the value of all the other terms. Note that all the channel matrices, including $H_{t,m}$, $H_{m,o}$, $H_{o,m}$, $H_{m,r}$, are known based on the relative position among the transceivers, metasurface M, and the 3D imaging area. Specifically, $H_{i,j} = a(d_{i,j})e^{-j2\pi f\frac{d_{i,j}}{c}}$, where $d_{i,j}$ is the distance from the source i to the destination j, c is the propagation speed of acoustic signals, and $a(d_{i,j})$ is the amount of signal attenuation at the distance $d_{i,j}$. Moreover, Eq. 4 can be further simplified as follows:

$$\begin{aligned} R_m &= H_{m,r}diag(M)H_{o,m}diag(H_{m,o}M \cdot H_{t,m}w)O + e \\ &= A_m(M,w)O + e \end{aligned} \tag{5}$$

where $A_m(M,w)$ is the measurement matrix for a specific frequency. To improve imaging performance, we try to suppress the channel noise and accumulate more constraints in Eq. 5.

Use beamforming at receiver side: Channel noise has significant impact on the image reconstruction error. An effective approach is to leverage multiple microphones at the receiver side, where beamforming can be used to harness spatial diversity and suppress channel noise. Let D denote the microphones' beamforming codebooks that consists of multiple weights, and R_m denote the received signal at all microphones. We have $R = DR_m$ to combine the received signals across microphones. For each frequency, the codebook size is equal to the number of microphones, since further increasing it does not yield new information due to linear dependence.

Use metasurface at multiple frequencies: A simple way to obtain more constraints is to use frequency diversity. However, as shown above, the benefit of frequency diversity is limited by the number of speakers and microphones. We jointly design a metasurface and the speakers and microphones' beamforming across multiple frequencies. To achieve this, we first need to derive the impact of the metasurface at different frequencies, including the phase shift and amplitude attenuation. The phase offset introduced by the metasurface can be directly calculated from the propagation distance of each metasurface cell's internal structure. To derive the impact on the amplitude, we observe the metasurface is designed to achieve close to 100% penetration at $20KHz$ (*i.e.*, $|M_{f=20kHz}| = 1$), and its penetration decays at other frequencies. This physical loss is complex to analyze. Therefore, we use COMSOL [3] (a finite-element-based multi-physical simulator) to simulate the impact of each of 16 different metasurface cells from $18kHz$ to $20kHz$ and generate a lookup table to record the resulting amplitude. We use this frequency range since it is inaudible and also supported by commodity devices. Since we employ both transmit and receive beamforming, our final measurement matrix can be denoted as $A(M,W,D)$.

Convert complex-valued matrix to real-valued matrix: Our formulation is derived in frequency domain, so that we can capture the impact of frequencies on the metasurface and the channel propagation model. As a result, $A(M,W,D)$ is complex-valued, which affects the image reconstruction process in two aspects: (i) most compressive sensing algorithms are developed for real-valued models [30] and (ii) complex-valued vectors are more likely to be linearly dependent since a linear multiplier can introduce a rotation and align two otherwise independent complex-valued vectors. We observe that all entries in our image O have values between 0 and 1, which are then multiplied by the real and imaginary parts of $A(M,W,D)$ separately. Therefore, we construct new M and R_m by stacking the real and imaginary parts from the original M and R_m, resulting in a real-valued model and effectively doubling the rank of the initial complex-valued measurement matrix.

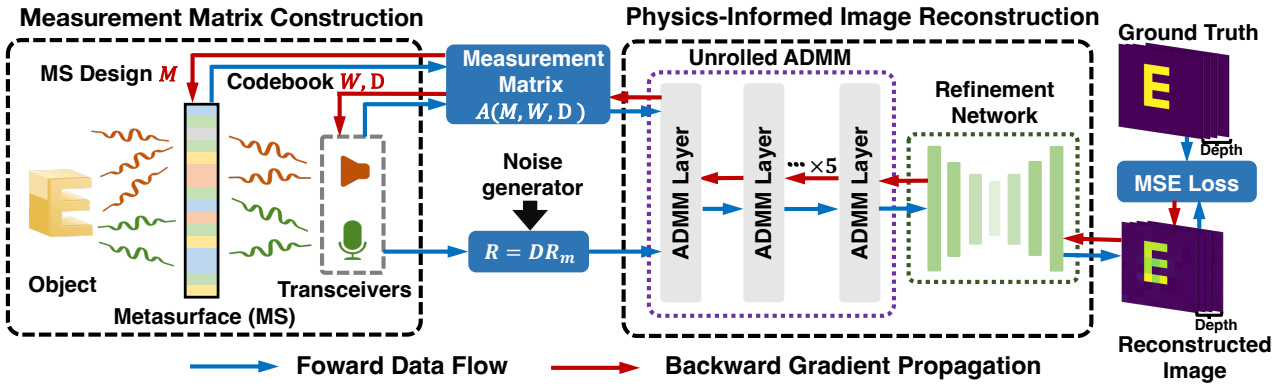


Figure 6: Joint optimization framework.

4.2 System Overview

As shown in Figure 6, to achieve high-quality imaging, we propose a joint optimization framework that determines the codebooks W and D across all frequencies and the metasurface M to minimize imaging error. This involves two aspects: (i) optimizing the measurement matrix, and (ii) reconstructing images using a physics-informed neural network. We integrate these two together to derive imaging error, which is then back propagated to update learnable parameters. Moreover, we introduce a scheme to adapt the imaging resolution according to the distance to avoid sharp decay in image quality.

4.3 Image Reconstruction

In this section, we describe how to reconstruct images.

4.3.1 Image inference model. When the constraints are insufficient to yield a unique solution, we need to leverage additional information about the object. One commonly used regularization term is sparsity, which indicates the target occupies a small portion of the 3D imaging area. This could be enforced by selecting an appropriate imaging region. This leads to the following optimization problem:

$$\hat{O} = \arg\min_{O} \|A(W, M)O - R\|_2^2 + \alpha\|O\|_1 \tag{6}$$

where $\|O\|_1$ is a commonly used L_1 norm to promote the sparsity in O and α captures the relative importance of the sparsity vs. fitting error. [12] shows that the solution with the minimal L_1 norm usually coincides with the sparsest solution for under-determined linear systems. Alternatively, we can also apply the sparsity regularization to the linear transformation of O (*e.g.*, $DCT(O)$) in case the target occupies a large portion of the imaging area. In the interest of brevity, the following description uses $\|O\|_1$ as the regularization term. We evaluate the sparsity regularization applied to both O and $DCT(O)$.

4.3.2 Classic ADMM imaging algorithm. The image reconsruction problem can be solved in a number of ways. A commonly used approach is ADMM. It is an iterative method that optimizes one variable at a time in each iteration while fixing the other variables [6]. We can rewrite Eq.6 as follows:

$$\min_{O,z} \|A(M, W, D)O - R_m\|_2^2 + \alpha\|z\|_1 \quad s.t. O - z = 0 \tag{7}$$

The ADMM algorithm is based on the *augmented Lagrangian*:

$$L_{\alpha,\rho} = \|A(M, W, D)O - R_m\|_2^2 + \alpha\|z\|_1 + \mu^T(O - z) + \frac{\rho}{2}\|O - z\|_2^2 \tag{8}$$

and performs sequential minimization of the O and z variables followed by the following dual variable updates:

$$\begin{aligned} O^{k+1} &= \underset{O}{argmin}\{\|A(M, W, D)O - R_m\|_2^2 + \frac{\rho}{2}\|O - z^k + u^k\|_2^2\} \\ z^{k+1} &= \mathcal{S}(O^{k+1}, u^k) = \underset{z}{argmin}\{\alpha\|z\|_1 + \frac{\rho}{2}\|O^{k+1} - z + u^k\|_2^2\} \\ u^{k+1} &= u^k + O^{k+1} - z^{k+1} \end{aligned} \tag{9}$$

for some arbitrary $O^0 \in \mathcal{R}^n, z^0 \in \mathcal{R}^n$, and $\mu^0 \in \mathcal{R}^n$ and $u = \frac{\mu}{\rho}$. $\mathcal{S}(\cdot)$ is a sparsity term based on domain knowledge (*e.g.*, $\mathcal{S}(\cdot) = \|z\|_1$). Compressive sensing solves the under-determined inverse problem by leveraging the sparsity assumption, and incorporates this assumption during the update of z^{k+1}. ρ and α are hyperparameters used to adjust the importance of the fidelity term and sparsity term during optimization.

Generally, the classic ADMM algorithm can produce satisfactory reconstructed images, but it still has several issues in practice. First, the hand-picked priors (*i.e.*, the sparsifying transform $\mathcal{S}$) and hyperparameters (*i.e.*, α and ρ) may not work well in our scenarios. Additionally, it takes hundreds of iterations to converge, which results in long running time.

4.3.3 Physics-Informed Image Reconstruction. We replace the classic ADMM algorithm with its neural-enhanced unrolled version to solve this inverse problem, which we call physics-informed learning model. It has two parts: unrolled ADMM with learnable layers and followed by a refinement network. The unrolled ADMM performs the bulk of image reconstruction and includes knowledge of the forward physical model, while the refinement nework denoises the image and corrects model mismatch errors.

Unrolled ADMM. The structure of our unrolled ADMM is inspired by ADMM-Net [44], which unfolds each iteration of traditional ADMM into a layer with learnable hyperparameters. As shown in Figure 7, our design goes a step further by replacing the sparse prior with a CNN, and learning imaging priors from imaging process in a data-driven manner. The intuition behind is that the sparsity may not strictly hold in real images and it is best to directly

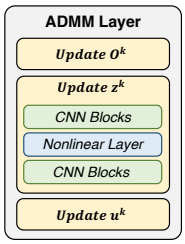


Figure 7: Structure of ADMM Layer.

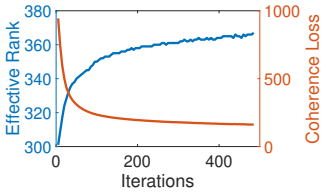


Figure 8: The change of effective rank and coherence loss over iterations.

learn a prior from real data; using a prior from real data can enhance the image quality and speed up inference. Therefore, we have a ADMM layer for each iteration, with the hyperparameter ρ and the learnable CNN $\mathcal{N}$. More specifically, the update equation in Eq. 9 becomes:

$$z^{k+1} = \mathcal{N}(O^{k+1}, u^k) \tag{10}$$

Eq. 10 does not impose sparsity constraints when updating z^{k+1}. Since we no longer use the sparsity term $\|z\|_1$, the hyperparameter α is also removed.

Our unrolled ADMM algorithm has several advantages over ADMM: 1) it uses a prior from real data, which is better than handcrafted sparsity regularization, thereby enhancing the imaging quality; 2) each layer in our neural network is associated with its own hyperparameters (*e.g.*, step size and penalty parameters) and these hyperparameters are learned from real data, whereas the corresponding parameters in ADMM are set in an ad hoc manner and fixed during the iteration process, which slows down the convergence. For example, it takes 8 iterations for our network to converge, while ADMM takes over 200 iterations to converge.

Refinement Network. To enhance the algorithm's robustness against noise and further correct mismatch errors caused by imperfect modeling, we concatenate our unrolled ADMM network with a refinement network $\mathcal{U}$ using a structure similar to UNet [40]. To accommodate our data format and enhance efficiency, our refinement network consists of a 3-layer encoder concatenated with a 3-layer decoder and outputs a 1D vector, which is then reshaped to the desired imaging dimensions (*e.g.*, $10 \times 10 \times 10$ 3D images). Such a design effectively improves system robustness since it is well-suited for noise reduction in images, thanks to its capability to capture detailed features through skip connections and hierarchical processing.

Training. To improve the actual system while minimizing the need for extensive dataset collection, we use real images from public datasets like FashionMNIST [50] to train our physics-informed model. The MSE is computed between the reconstructed images and the ground-truth images, as demonstrated below:

$$\min_{\{\rho,\mathcal{N},\mathcal{U}\}} L = \frac{1}{N}\|O - \hat{O}\|_2^2 \tag{11}$$

where N is the number of grids in a 3D scene. Then we back-propagate the error to update all learnable parameters. Since the training objective is to directly minimize the distance to the ground truth image instead of intermediate indicators, the quality of reconstructed images can be significantly optimized. Note that we only use the existing image dataset for training, and use our testbed measurement for evaluation.

4.4 End-to-End Optimization

In this section, we build an end-to-end optimization framework to jointly optimize the imaging algorithm and the measurement matrix (*i.e.*, the metasurface design $M_{f=20kHz}$, codebook for speakers W and microphones D across all frequencies). We first initialize the measurement matrix by minimizing the coherence of the matrix. This allows us to quickly obtain a reasonable measurement matrix as a starting point. Then, we concatenate the measurement matrix optimization with the image reconstruction and iteratively optimize them in an end-to-end manner.

4.4.1 Initialization of Measurement Matrix. Before the end-to-end optimization, we initialize $A(M, W, D)$ by minimizing the coherence of the matrix. The coherence measures the maximum inner product between any two columns of a matrix. Matrices with low coherence tend to be easier to invert and can lead to better reconstruction error in compressive sensing. Therefore, we use this property to increase the diversity of $A(M, W, D)$ so that each measurement provides new information. The coherence-based optimization can be modeled as $\min_{i\neq k} \sum G_{ik}$, where $G_{ik} = A(M, W_i)^T A(M, W_k)$ computes the correlation between the i-th and k-th rows in the measurement matrix. We can solve this optimization problem using a simple gradient descent approach and ignore the physical constraints on the tunable parameters for simplicity, as shown in Eq. 11. Note that we use the coherence as the initialization criteria because it is easier to optimize than the RIP property or effective rank. As shown in Figure 2 (a), the RMSE monotonically decreases with the effective rank. Figure 8 further shows that reducing coherence increases the effective rank, which justifies that we can minimize the coherence of the measurement matrix to provide a good initialization.

4.4.2 End-to-End Training. The objective function of end-to-end training is the same as Eq. 11 and the training process iteratively performs the following two steps: (1) designing the measurement matrix and (2) updating the physics-informed imaging reconstruction network. Step (1) treats the imaging reconstruction network as known and optimizes M and W, D. Step (2) treats M, W, and D as known and optimize the network's parameters. Then we iterate until convergence.

During the update of the measurement matrix, the learnable parameters should satisfy the following constraints: (i) $|W_{ij}| \leq 1$ imposes restriction on the maximum transmission power. (ii) $|M_{f=20kHz,i}| = 1$ according to [31, 52]. For the channel coefficient of other frequencies, we use the lookup table derived from COMSOL as described in Sec. 4.1. (iii) $\sum_{i=\#Mics} |D_i| = 1$ imposes constraints on the sum of each set of beamforming weights.

Similar to traditional neural network training, we solve this optimization problem using Adam optimizer in Pytorch, which is an extended version of the stochastic gradient descent algorithm. We adapt learning rates for different optimization parameters. To enforce the above constraints, we project the outputs from the Adam optimizer to the nearest feasible set.

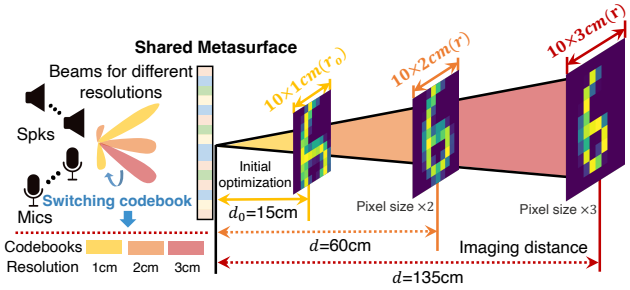


Figure 9: Adapting imaging resolution for different distance. End-to-End optimizing r_0 resolution at distance d_0, then select resolution r based on distance d and adjust codebooks to improve SNR.

The end-to-end optimization framework incorporates the interdependency between optimizing measurement matrix and image reconstruction by directly minimizing the imaging error, thereby out-performing coherence-based initialization.

4.5 Adapting Image Resolution for Different Distances

The imaging quality inevitably degrades with the distance due to two reasons: (i) reduced SNR and (ii) reduced angular resolution. To better support imaging across a varying distance, we propose a novel adaptive resolution image reconstruction scheme.

We scale the resolution between two pixels of the target as a function of the distance d between the target and transceiver. Specifically, let Δd denote the resolution of the reflected pulse (*i.e.*, two closest peaks in the reflected pulse that can be separated). Let r denote the separation between two closest pixels on an image plane that can be separated. We have the following relationship: $\Delta d = \sqrt{d^2 + r^2} - d$ according to Pythagorean Theorem, where d denotes the distance between the closest pixel on the target plane and the transceiver. Therefore, we have $r = \sqrt{2\Delta d d + \Delta d^2} \approx \sqrt{2\Delta d d}$. The intuition behind this equation is that the path length change of two adjacent pixels should be equal at different ranges to ensure the same resolvability (or equivalently, the time delay) of the corresponding reflected pulses. As shown in Figure 9, since we know the location of the target imaging area, we can derive the appropriate imaging resolution based on their relative distance to the transceivers. Note that one resolution can support a range of distances (*e.g.*, $1cm$ resolution for imaging within $15cm$, and $3cm$ resolution for imaging between $15cm - 135cm$).

In our implementation, we use the end-to-end optimization to determine the metasurface design and codebook using the finest resolution for fabrication and deployment. To support a longer distance, we use the current distance to determine a new resolution. We then run the end-to-end optimization to derive a new codebook while fixing the variables associated with the metasurface according to our deployed setup. This is because the codebook can change dynamically while the metasurface is fixed after it is manufactured.

5 PERFORMANCE EVALUATION

We describe our implementation and evaluations as follows.

5.1 Implementation

As depicted in Figure 10, our experimental setup includes a metasurface, a Bela board [2] as the controller, a 3×2 speaker array with a PAM8406 power amplifier as the transmitter, and a 4×1 microphone array as the receiver. The speakers are spaced $25mm$ apart horizontally and $44.6mm$ vertically, while the microphones are $14mm$ apart. We optimized the transmission signal, beamforming weights, metasurface, and imaging algorithm using our end-to-end framework described in Section 4.4. The optimization process is completed on a server equipped with NVIDIA 3090, and the total optimization time is 4.25 hours. The main cost of optimization time is due to the joint optimization convergence of the imaging algorithm and the measurement matrix. The metasurface is derived from the optimization results and physically assembled. Figure 10 (b) shows the front view of our cost-effective (< 5 dollars) optimized metasurface, positioned $3cm$ from the speaker to maximize signal transmission. The metasurface has a side length of $15.6cm$ and an area of $240cm^2$, and can be easily integrated into the shell of a smart speaker.

For imaging, we place the target within a predefined 3D area and transmit FMCW signals (18 to $20kHz$) at a $48kHz$ sampling rate using our optimized beamforming weights (codebooks W) to obtain measurement constraints. It takes the transmitter $120ms$ for a single scan, allocating $20ms$ per beamforming weight. When there is no object in the imaging area, we pre-record the signal to eliminate the primary influence of the direct signal. Specifically we record the signals after putting an object in the imaging area and subtract the pre-recorded signals. After that, we apply temporal window filtering to remove most of the environmental interference cancellation outside the imaging area. Next, receiver beamforming (codebooks D) and FFT are applied to the signals, yielding complex constraints from frequency bins, each spanning $25Hz$. With 6 transmitter and 4 receiver beamforming weights, and 81 frequencies, we extract 1944 complex constraints per transmission.

We produce 150 3D-printed objects as the test set for imaging evaluation, including numbers, letters, and other common patterns. We evaluate our scheme to image 2D objects (30×30) and 3D objects ($10 \times 10 \times 10$). Note that the pattern of our test set images is completely different from the open source dataset Fashion-MNIST, where the patterns of Fashion-MNIST are almost clothes, shoes, etc. Further, in order to ensure that the trained model will not be overfitted, we let the training be performed on the simulator, and the testing is performed on the real prototype. Unless stated otherwise, we place the objects 15cm away from the metasurface and use $1cm$ resolution. All results are collected from our testbed.

Before testing MAJIC, we calibrate it to account for the real interaction between the transceiver and metasurface and to adjust for any phase and attenuation discrepancies caused by imperfections in the 3D-printed metasurface cells. The calibration involves two main steps: first, we create a calibration matrix that captures the difference between the ideal reflection from a point reflector on an imaging grid and the actual signal received from the simulation; second, we use 10 random calibration images from our testbed to fine-tune the neural network in our physics-informed imaging algorithm, addressing any model inaccuracies. This one-time calibration, typically done during

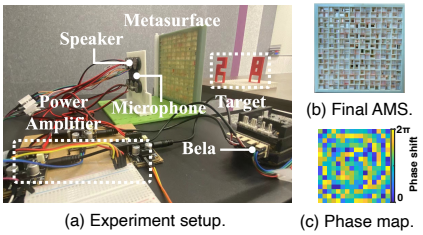


Figure 10: System prototype. (a) shows the testbed. (b) and (c) show final acoustic metasurface (AMS) and its phase map, respectively.

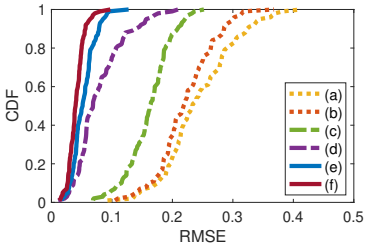


Figure 11: CDF of RMSE using scheme from (a) to (f).

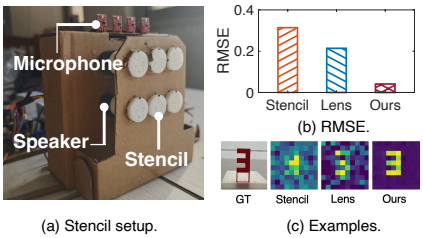


Figure 12: Comparision with other metasurfaces. (a) shows the stencil setup. (b) and (c) show RMSE and examples, respectively.

manufacturing, allows for straightforward deployment in various settings without extra work.

In addition, we also compare with stencils in [5] and acoustic lens in [52]. [5] develops a metasurface as a stencil for imaging. Following [5], we make a few stencils with a diameter of 20mm and a height of 30mm. The path lengths of 10 tubes are randomized. We choose the one that has the best signal diversity distribution according to [5]. [52] maximizes the SNR in a given direction. For both schemes, we use the same codewords, speakers, and microphones placement as those in our scheme for comparison. For fair comparison, all schemes use static setups.

Performance metric: The imaging quality is quantified using root mean square error (RMSE): $\sqrt{(e-t)^2}$, where e and t are the estimated and ground truth values. Although RMSE cannot fully represent the imaging effect in some special cases, it reflects the system performance in most cases with reference to the field of computer vision.

Baseline schemes: Table 2 summarizes the baseline schemes used in our evaluation for comparison. (a) Without metasurface and beamforming. (b) Optimizing speakers' beamforming. (c) Using a random metasurface and optimizing speakers' beamforming. (d) Jointly optimizing both metasurface and speakers' beamforming. (e) Jointly optimizing speakers' beamforming and the shared metasurface for both speakers and microphones. (f) Jointly optimizing beamforming and the shared metasurface for both speakers and microphones. Note that the microphone is placed at the top of the metasurface in schemes (c) and (d) to ensure that the reflected signals from the objects do not pass through the metasurface.

5.2 Overall Performance

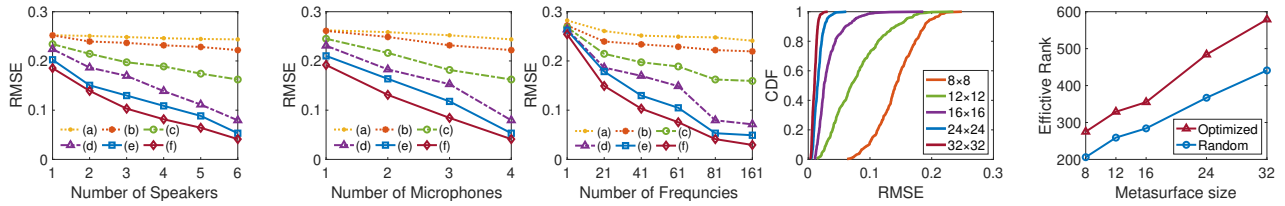
We evaluate the performance of schemes (a)-(f) using our testbed. Figure 11 shows the cumulative distribution function (CDF) of the RMSE. As expected, scheme (f) performs the best. Some examples of imaging results with various scheme and depth information are shown in Figure 14 and Figure 15, respectively. Compared with scheme (a) to (e), scheme (f) reduces RMSE by 83.1%, 81.4%, 74.8%, 48.11%, and 22.6%, respectively. The benefits of adding beamforming to the speaker array without metasurface are limited. Its restricted channel customization capability contributes marginally to the rank of the measurement matrix. Interestingly, we find that adding a

Table 2: Summary of schemes for comparison.

	speaker		microphone	
	beamforming	**metasurface**	**beamforming**	**metasurface**
(a)	no	no	no	no
(b)	opt.	no	no	no
(c)	opt.	random	no	no
(d)	joint opt.	joint opt.	no	no
(e)	joint opt.	joint opt.	no	joint opt.
(f)	joint opt.	joint opt.	joint opt.	joint opt.

randomly-built metasurface for the speaker array is already beneficial. Specifically, compared with (b), (c) reduce RMSE by 27.0% from 0.222 to 0.162. With the optimized metasurface (scheme (d)), we can further reduce RMSE by 0.079, which is a 51.2% reduction. The significant improvement from scheme (d) to (e) highlights the importance of considering the microphone array in metasurface optimization. One possible explanation is that the signal reflected from the target is weak, and the optimized metasurface can act as a reflection signal collector and redirect the reflected signals going through the metasurface toward the microphone array. The performance can be further improved by 22.6% using microphone-side beamforming in the scheme (f), as microphone-side beamforming technology can reduce the noise of the received signal.

Furthermore, we compare the imaging performance using two alternative acoustic metasurfaces – the stencil in SPiDR [5] and acoustic lens from [52] while the other components remain the same as the scheme (f) (*i.e.*, beamforming and imaging reconstruction). This allows us to assess the performance of different metasurface designs. As shown in Figure 12, our approach out-performs the stencil and acoustic lens by 88.42% and 80.75%, respectively. The results show the design of a metasurface has significant impact on the imaging performance and our joint optimization is effective. In comparison, the metasurface used in SPiDR [5] is ad hoc and limits its benefit. SPiDR [5] has to resort to robot movement to extract more constraints for imaging, while our design can achieve high imaging accuracy using a static setup. The metasurface in [52] aims to maximize SNR in a given direction and is less effective for imaging since imaging requires diverse measurements (*i.e.*, a measurement matrix with a high rank) in addition to high SNR.



(a) Speakers. (b) Microphones. (c) Frequencies. (d) RMSE of AMS size. (e) Effective rank of AMS size.

Figure 13: Effect of each component on the measurement matrix.

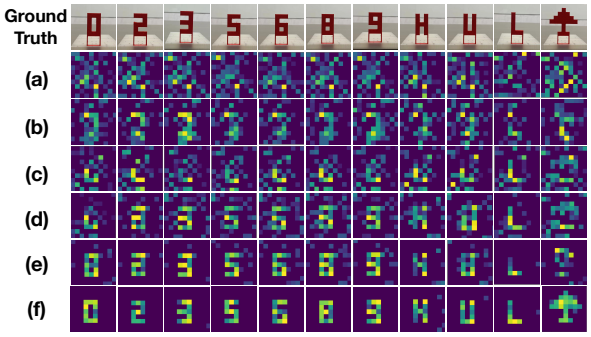


Figure 14: Examples of imaging results using scheme (a)-(f) at the same depth.

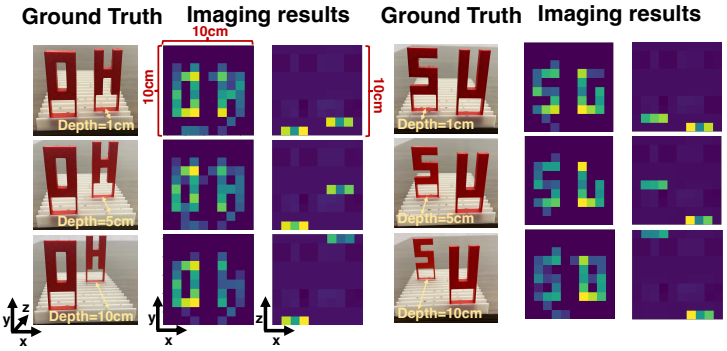


Figure 15: Examples of imaging at various depths.

5.3 Microbenchmarks

In this section, we evaluate the impacts of various design parameters.

5.3.1 Vary number of transceivers. Figure 13 plots the impact of varying the number of transceiver antennas. When there is an optimized metasurface (*i.e.*, scheme (d)-(f)), increasing the number of speakers and microphones provides additional signal diversity, resulting in a significant reduction in median RMSE. Specifically, using scheme (f) with 2, 3, 4, 5, and 6 speakers reduces median RMSE from 0.18 under a single speaker to 0.139, 0.103, 0.082, 0.064, and 0.041, respectively. Using scheme (f) with 2, 3, and 4 microphones reduces the RMSE from 0.192 under a single microphone to 0.131, 0.084, and 0.041, respectively. Interestingly, for schemes (a) and (b) where there is no metasurface, increasing the number of transceivers yields little performance benefit because without a metasurface the diversity is too limited to provide sufficient constraints for image reconstruction.

5.3.2 Vary number of frequencies. Next, we evaluate the impact of varying the number of frequencies on the measurement matrix. As shown in Figure 13(c), when using a single frequency, the metasurface provides virtually no gain for imaging since the constraints are too few to make a difference. As more frequencies are used, more constraints are introduced to the measurement matrix, which can improve the reconstruction performance, where the median RMSE in scheme (f) reduces from 0.254 to 0.041 using 81 frequencies. As we use more than 81 frequencies, the benefits become marginal while the required length of the sampling window for FFT increases. This indicates that 81 frequencies are sufficient for our scenario.

5.3.3 Vary metasurface size. As shown in Figure 13(d), increasing the metasurface size from 8×8 to 12×12, 16×16, 24×24, and 32×32 results in a median RMSE of 0.147, 0.089, 0.039, 0.018, and 0.011, respectively. However, we observe that the marginal benefit of using a larger metasurface diminishes beyond a certain point. For instance, the reduction in RMSE from increasing the size from 8×8 to 16×16 is more significant than that from increasing the size from 16×16 to 32×32. From the rank perspective, enlarging the metasurface and optimizing its configuration can effectively enhance the effective rank of the measurement matrix. Our optimization reduces the size of the metasurface, thereby cutting costs and space. For example, an optimized 16×16 metasurface has a rank of 355, which is close to the effective rank of a 24×24 random metasurface – 367, saving 50% cost and space.

5.3.4 2D images. We evaluate the performance of 3D imaging in our testbed. Next we consider imaging 2D objects in a simulator (the only simulation results in our evaluation). Without the depth information, we can support a larger 2D images. We achieve RMSE of 0.052 when imaging 30×30 2D objects, and RMSE of 0.075 when imaging 50×50 2D objects. Figure 21 shows example 50×50 images. The clear shapes of the reconstructed images show the effectiveness of our method for high-resolution 2D imaging.

5.3.5 Vary AoA. To evaluate the impact of different angles of arrival (AoAs), we place targets at positions ranging from 0° to 60° relative to the center of the transceivers. As shown in Figure 19, as the AoA increases from 0° to 20°, 40°, and 60°, the median RMSE increases from 0.041 to 0.077, 0.154, and 0.235, respectively, while

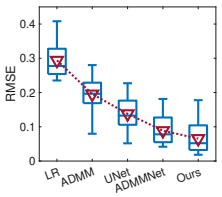


Figure 16: Comparison with other algorithms.

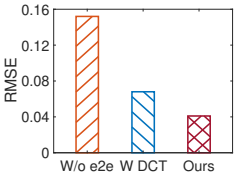


Figure 17: Impact of using end-to-end optimization and DCT.

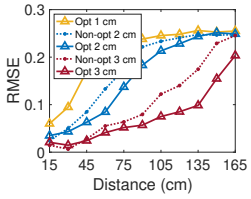


Figure 18: Adaptive resolution for distance.

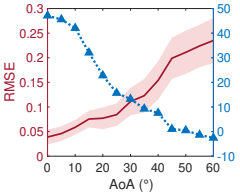


Figure 19: Robustness to AoAs.

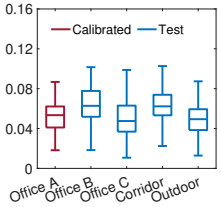
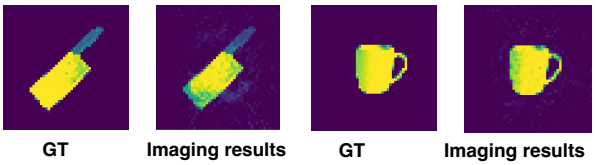


Figure 20: Robustness to various environments.



(a) Knife. (b) Cup.

Figure 21: Imaging results of 50×50 (2500 pixels) with $0.5cm$ resolution.

Table 3: Time to reconstruct $10\times10\times10$ images using a GeForce RTX 3090 on the server.

	LR	ADMM	UNet	ADMMNet	Ours
Time (ms)	4.4	175.1	1.7	9.3	9.7
Model size (MB)	\	\	116.2	1.8	3.6

the SNR decreases from 47.095 to 22.792, 7.550, and $-2.567dB$, respectively. The results show that even if the object is not strictly aligned with the predefined imaging grid, our system can accurately image an area within 30°. However, as the angle increases, the imaging performance degrades rapidly due to less reflection from the target.

5.3.6 Performance of Image Reconstruction Algorithm. We further compare with four commonly used algorithms in our testbed using our optimized metasurface and beamforming: (i) linear regression (LR), (ii) traditional ADMM, (iii) UNet (integrating preceding fully connected layers), and (iv) ADMMNet. Figure 16 shows that all algorithms incur increasing RMSE as the SNR decreases. Our reconstruction algorithm consistently achieves the lowest RMSE by leveraging learned imaging prior and denoising neural networks. In comparison, ADMMNet only has the learnable hyperparameter but lacks a denoising step, while UNet has remarkable denoising capability but lacks prior learned from data. Therefore, their performance is considerably worse. Figure 22 show examples of reconstructed images, further illustrating our imaging algorithm's benefit.

Next we evaluate the impact of end-to-end optimization. As shown in Figure 17, compared to the methods without end-to-end optimization (*i.e.*, using only coherence-based optimization), our algorithm reduces the median RMSE from 0.152 to 0.041, which is 73.1% improvement. The results indicate that jointly optimizing the measurement matrix and reconstruction algorithm is crucial. We also observe that due to the sufficient sparsity of our targets within the 3D imaging area, employing the DCT does not provide additional gains. For generality, we can still consider using DCT.

Moreover, we evaluate the overhead of our reconstruction algorithm, including both time and storage. As shown in Table. 3, our algorithm requires 9.7 ms to reconstruct an image, which is much faster than the 175.1 ms required by the traditional ADMM. Although the frame rate of our imaging is still limited by the scanning time at the transmitting end, we can effectively replace speakers with more microphones for high frame rate (i.e., $\geq$ 30 FPS) to support real-time applications with higher frame rates.

5.3.7 Effectiveness of Adaptive Resolution Scheme. We conduct experiments with targets at varying distances from $15cm$ to $165cm$ away from the transceivers. Note that our $4.5W$ device range can be further extended using higher power devices, like Apple's HomePod with $33W+$ power. We use grid patterns of $1cm$, $2cm$, and $3cm$ resolutions for imaging. The codebooks for transceivers are re-optimized at $60cm$ and $135cm$ for $2cm$ and $3cm$ resolutions. Figure 18 shows RMSE increases with distance across all resolutions. Our system achieves up to $30cm$ distance for high-quality imaging (RMSE < 0.1) at 1cm resolution. Without codebook re-optimization, the maximum distances for $2cm$ and $3cm$ resolutions are $45cm$ and $90cm$, respectively. Optimization extends the ranges to $60cm$ and $135cm$ for $2cm$ and $3cm$ resolutions, respectively. Re-optimizing for larger distances may slightly reduce performance at closer ranges ($< 30cm$), but this is minor compared to the more significant performance gain at a longer distance. If our imaging area occupies close and far away areas, we may use different grid sizes at different distances. In conclusion, our adaptive resolution strategy maintains a low RMSE across varying distances.

5.3.8 Impact of environments. The performance of the time window filtering algorithm may be affected by scene changes outside the imaging area. Therefore, we evaluate our system's robustness in five different environments, including the office's A to C, corridor, and outdoors. Office A is an unoccupied conference room; Office B is the same size as Office A, with people walking around, and Office C has a smaller area. The results demonstrate that our system has excellent environmental adaptability with only one pre-calibration at one location. On the other hand, changes outside the imaging area have little impact on system performance. As shown in Figure 20, our system can maintain robust imaging capabilities in different

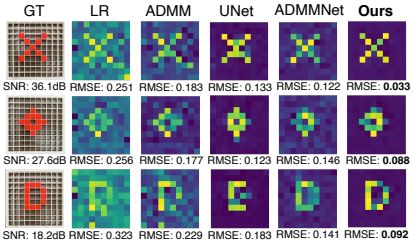


Figure 22: Imaging using various algorithms on the same optimized testbed.

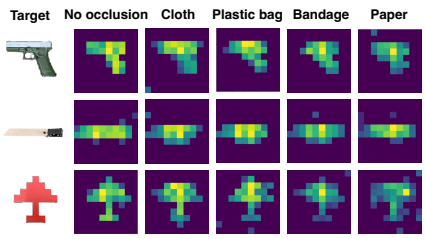
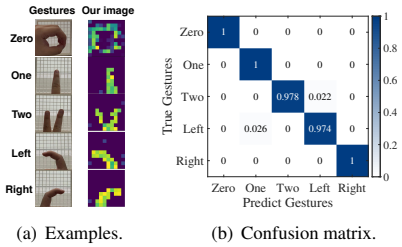


Figure 23: Imaging under the covering of various materials.



(a) Examples. (b) Confusion matrix.

Figure 24: The performance of gesture recognition.

scenarios, with mean RMSE of 0.052, 0.063, 0.052, 0.064, and 0.049, respectively. Moreover, the low RMSE in Offices B and C indicates that our background interference elimination method can effectively reduce interference outside the imaging area.

5.3.9 Impact on low-frequency sound. We place microphones and speakers at opposite ends of a metasurface to assess its impact on the attenuation of low-frequency sounds. Specifically, we transmit chirp signals ranging from $100Hz$ to $4kHz$ at 30-degree intervals within an incidence angle range of 0 to 180 degrees. Compared to scenarios without the metasurface, the average energy attenuation at various angles is 0.53dB. These results indicate that the metasurface minimally impacts applications involving low-frequency sounds, such as music playback.

5.4 Applications

5.4.1 Imaging under occlusion. Acoustic signals can penetrate objects to some extent, which makes it possible to image under occlusion (*e.g.*, covered by plastic or cloth bags). Figure 23 shows examples of the reconstructed images with various materials covering the target. The reconstructed images show that we can still clearly distinguish the shape of the targets. This potentially helps identify dangerous items (*e.g.*, weapons) under a covering.

5.4.2 Gesture recognition. Furthermore, we demonstrate the potential of our system for privacy-preserved gesture recognition, where the hand is placed within a pre-defined imaging area. Figure 24 shows the imaging results of five gestures. We further employ SVM to classify the five gestures and observe the high classification accuracy. This enables a touchless user interface and is more privacy-preserving than camera-based solutions.

6 DISCUSSION

Frequency range selection. Our current use of the 18kHz to 20kHz frequency range is not only because it is inaudible to most people but also because the metasurface maintains high transmission for signals within this range. However, this frequency band might be audible to pets or children, and Chirp modulation could potentially cause discomfort to them. For a more practical system, we could mitigate this effect through special modulation technique [45], or opt for a higher frequency band, such as 24kHz, which can be achieved by adjusting the size of the units. Additionally, a broader bandwidth could offer more diversity, but it also raises the design requirements for the metasurface's broadband transmission rate. One of our future directions is to explore metasurfaces that support a wider frequency range to enhance frequency diversity.

Dynamic imaging. The continuous acoustic imaging of dynamic objects plays a crucial role and has the potential to catalyze numerous innovative applications. At present, the relatively short processing times offered by MAJIC present a promising possibility for acoustic imaging to capture videos with the same ease and flexibility as optical cameras. However, realizing this potential fully involves overcoming a myriad of challenges. These include not only expanding the imaging space and enhancing the resolution to much finer scales but also effectively managing the phase variations introduced by the movement of objects within the imaging field. Such aspirations necessitate a move towards more intricate imaging models that take into account the dynamics of objects, the devices capturing the images, and even the surrounding environment's influence on the imaging process. This complex and layered approach opens up a new frontier in acoustic imaging research, promising exciting developments and applications in the future.

7 CONCLUSION

In this paper, we develop a novel end-to-end framework for acoustic imaging, which jointly optimizes the transmission signal, transceivers beamforming, metasurface, and imaging reconstruction algorithm. Our design realizes high-quality acoustic imaging using a small number of speakers and microphones that are available on typical low-cost IoT devices. This opens the door to many exciting applications that we hope to explore in the future, including flexible user interfaces, activity recognition, weapon detection and medical diagnosis.

ACKNOWLEDGMENTS

We are grateful for Yuanchao Shu's insightful feedback and anonymous reviewers' helpful comments. This research was supported in part by the National Natural Science Foundation of China under Grant No. 62122095, 62341201 and 62072472, and by a grant from the Guoqiang Institute, Tsinghua University.

REFERENCES

[1] All-new Echo (4th Gen). https://www.amazon.com/Echo-4th-Gen/dp/B07XKF5RM3.
[2] Bela: create beautiful interaction with sensors and sound. https://bela.io/, 2022.
[3] COMSOL: simulate real-world designs, devices, and processes with multiphysics software from comsol. https://www.ti.com/product/LM386, 2023.2.

[4] A. Adhikari, H. Regmi, S. Sur, and S. Nelakuditi. Mishape: Accurate human silhouettes and body joints from commodity millimeter-wave devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3):1–31, 2022.

[5] Y. Bai, N. Garg, and N. Roy. Spidr: Ultra-low-power acoustic spatial sensing for micro-robot navigation. In *ACM MobiSys*, 2022.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[7] P. Cai and S. Sur. Millipcd: Beyond traditional vision indoor point cloud generation via handheld millimeter-wave devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–24, 2023.

[8] E. J. Candes and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[9] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, pages 589–592, May 2008.

[10] X. Chen, P. Liu, Z. Hou, and Y. Pei. Magnetic-control multifunctional acoustic metasurface for reflected wave manipulation at deep subwavelength scale. *Scientific reports*, 7(1):1–9, 2017.

[11] Our acoustic imaging demo. https://youtu.be/ASEbUA11IbY.

[12] D. Donoho. For most large underdetermined systems of linear equations, the minimal l1-norm nearsolution approximates the sparsest near-solution.

[13] Y. Fu, S. Wang, L. Zhong, L. Chen, J. Ren, and Y. Zhang. Svoice: Enabling voice communication in silence via acoustic sensing on commodity devices. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 622–636, 2022.

[14] Y. Ghasempour, C.-Y. Yeh, R. Shrestha, Y. Amarasinghe, D. Mittleman, and E. W. Knightly. LeakyTrack: Non-coherent single-antenna nodal and environmental mobility tracking with a leaky-wave antenna. In *ACM SenSys*, 2020.

[15] J. Hald. Time domain acoustical holography and its applications. *Sound and Vibration*, 35(2):16–25, 2001.

[16] B. P. Hildebrand. *An introduction to acoustical holography*. Springer Science & Business Media, 2013.

[17] Product environmental report. https://www.apple.com/environment/pdf/products/homepod/HomePod_PER_Jan2023.pdf.

[18] F.-L. Hsiao, T.-K. Li, P.-C. Chen, S.-C. Wang, K.-W. Lin, W.-L. Lin, Y.-P. Tsai, W.-K. Lin, and B.-S. Lin. Phase resonance and sensing application of an acoustic metamaterial based on a composite both-sides-open disk resonator arrays. *Sensors and Actuators A: Physical*, 339:113524, 2022.

[19] D. Huang, R. Nandakumar, and S. Gollakota. Feasibility and limits of wi-fi imaging. In *ACM SenSys*, 2014.

[20] C. R. Karanam and Y. Mostofi. 3d through-wall imaging with unmanned aerial vehicles using WiFi. In *ACM/IEEE IPSN*, 2017.

[21] H. Lee. *Acoustical Sensing and Imaging*. CRC Press, 2016.

[22] K. H. Lee, K. Yu, A. Xin, Z. Feng, Q. Wang, et al. Sharkskin-inspired magnetoactive reconfigurable acoustic metamaterials. *Research*, 2020, 2020.

[23] C. Li, Z. Liu, Y. Yao, Z. Cao, M. Zhang, and Y. Liu. Wi-Fi see it all: generative adversarial network-augmented versatile Wi-Fi imaging. In *ACM SenSys*, 2020.

[24] J. Lian, X. Yuan, M. Li, and N.-F. Tzeng. Fall detection via inaudible acoustic sensing. In *UbiComp*, 2021.

[25] Z. Liang and J. Li. Extreme acoustic metamaterial by coiling up space. *Physical review letters*, 108(11):114301, 2012.

[26] W. Mao, J. He, and L. Qiu. CAT: high-precision acoustic motion tracking. In *Proc. of ACM MobiCom*, 2016.

[27] W. Mao, W. Sun, M. Wang, and L. Qiu. Deeprange: Ranging via deep learning. In *Proc. of UbiComp*, 2021.

[28] W. Mao, M. Wang, and L. Qiu. Aim: Acoustic imaging on a mobile. In *ACM MobiSys*, 2018.

[29] W. Mao, M. Wang, W. Sun, L. Qiu, S. Pradhan, and Y.-C. Chen. Rnn-based room scale hand motion tracking. In *The 25th Annual International Conference on Mobile Computing and Networking*, pages 1–16, 2019.

[30] M. Mardani, E. Gong, J. Y. Cheng, S. Vasanawala, G. Zaharchuk, M. Alley, N. Thakur, S. Han, W. Dally, J. M. Pauly, et al. Deep generative adversarial networks for compressed sensing automates mri. *arXiv preprint arXiv:1706.00051*, 2017.

[31] G. Memoli, M. Caleap, M. Asakawa, D. R. Sahoo, B. W. Drinkwater, and S. Subramanian. Metamaterial bricks and quantization of meta-surfaces. *Nature Communication*, 2017.

[32] G. Memoli, L. Chisari, J. P. Eccles, M. Caleap, B. W. Drinkwater, and S. Subramanian. Vari-sound: A varifocal lens for sound. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA, 2019. Association for Computing Machinery.

[33] R. Nandakumar, S. Gollakota, and N. Watson. Contactless sleep apnea detection on smartphones. In *Proc. of ACM MobiSys*, 2015.

[34] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota. FingerIO: Using active sonar for fine-grained finger tracking. In *Proc. of ACM CHI*, pages 1515–1525, 2016.

[35] A. Pallaprolu, B. Korany, and Y. Mostofi. Wiffract: a new foundation for RF imaging via edge tracing. In *ACM MobiCom*, 2022.

[36] C. Peng, G. Shen, Y. Zhang, Y. Li, and K. Tan. BeepBeep: a high accuracy acoustic ranging system using COTS mobile devices. In *Proc. of ACM SenSys*, 2007.

[37] K. Qian, Z. He, and X. Zhang. 3D point cloud generation with millimeter-wave radar. *ACM IMWUT*, 2020.

[38] M. Rezvani and R. Adve. Channel estimation for dynamic metasurface antennas. *IEEE Transactions on Wireless Communications*, 2023.

[39] Lecture note on rip.

[40] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[41] S. Shen, D. Chen, Y.-L. Wei, Z. Yang, and R. R. Choudhury. Voice localization using nearby wall reflections. In *Proc. of ACM MobiCom*, 2020.

[42] Y. Shibata, Y. Kawashima, M. Isogawa, G. Irie, A. Kimura, and Y. Aoki. Listening human behavior: 3d human pose estimation with acoustic signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13323–13332, 2023.

[43] R. I. Stantchev, X. Yu, T. Blu, and E. Pickwell-MacPherson. Real-time terahertz imaging with a single-pixel detector. *Nature communications*, 2020.

[44] J. Sun, H. Li, Z. Xu, et al. Deep admm-net for compressive sensing mri. *Advances in neural information processing systems*, 29, 2016.

[45] A. Wang, J. E. Sunshine, and S. Gollakota. Contactless infant monitoring using white noise. In *The 25th Annual International Conference on Mobile Computing and Networking*, pages 1–16, 2019.

[46] J. Wang, J. Xiong, X. Chen, H. Jiang, R. K. Balan, and D. Fang. TagScan: Simultaneous target imaging and material identification with commodity RFID devices. In *ACM MobiCom*, 2017.

[47] M. Wang, W. Sun, and L. Qiu. Mavl: Multiresolution analysis of voice localization. In *Proc. of NSDI*, 2021.

[48] S. Wang, L. Zhong, Y. Fu, L. Chen, J. Ren, and Y. Zhang. Uface: Your smartphone can" hear" your facial expression! *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–27, 2024.

[49] W. Wang, A. X. Liu, and K. Sun. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, pages 82–94. ACM, 2016.

[50] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[51] S. Yun, Y. chao Chen, and L. Qiu. Turning a mobile device into a mouse in the air. In *Proc. of ACM MobiSys*, May 2015.

[52] Y. Zhang, Y. Wang, L. Yang, M. Wang, Y.-C. Chen, L. Qiu, Y. Liu, G. Xue, and J. Yu. Acoustic sensing and communication using metasurface. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 1359–1374, 2023.