IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 December 2023, accepted 18 January 2024, date of publication 22 January 2024, date of current version 30 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3357117

RESEARCH ARTICLE

Multi-View Reconstruction Fusing Ultrasonic Phased Array and Camera for Mobile Robots in Simulation Environment

JIXIANG REN AND XIAOPING HONG, (Member, IEEE)
School of System Design and Intelligent Manufacturing, Southern University of Science and Technology, Shenzhen 518055, China

Corresponding author: Jixiang Ren (renjx2021@mail.sustech.edu.cn)

This work was supported in part by the Southern University of Science and Technology (SUSTech) Startup Fund under Grant Y01966105, in part by the SUSTech-DJI Joint Laboratory Fund under Grant K2096Z028, and in part by the Shenzhen Science and Technology Project under Grant JSGG20211029095803004 and Grant JSGG20201103100401004.

ABSTRACT Distance sensors are important for mobile robots to perceive surrounding environment. Typical sensors like LiDARs and depth cameras have been widely used, yet each has its limitations, such as LiDARs' relatively high cost, depth cameras' limitation to indoor use, and their poor performance in detecting transparent objects directly. On the other hand, ultrasonic phased array that integrates multiple ultrasonic sensors not only enables 3D ranging and imaging, but also provides advantages of strong environmental adaptability, being cost-effective and being able to detect transparent objects. To explore the application of in-air ultrasonic phased arrays for mobile robots, we simulate a 40 kHz 5 × 5 non-uniform sparse ultrasonic phased array. The simulator emulates the process of phased array transmission and reception, and utilizes algorithms such as beamforming and matched filtering to obtain depth information in three-dimensional space. Then, a multi-view indoor 3D reconstruction method fusing the ultrasonic phased array and a monocular camera is proposed, where two scanning strategies are developed to handle different scenarios. Finally, the method is validated in different Gazebo scenarios and compared with other baseline methods like LiDARs and depth cameras. The experimental results reveal the method's strong performance in terms of accuracy, consistency and completeness.

INDEX TERMS Phased arrays, reconstruction algorithms, sensor fusion, simulation, ultrasonic imaging.

I. INTRODUCTION

Three-dimensional imaging technology has seen unprecedented boom in recent years, especially those applied for mobile robots. Digitization of our 3D space can provide great insights in path planning, self driving, virtual reality and so on, assisting robots in perceiving surrounding world. This is achieved by a heterogeneous sensor fusion [1]. When it comes to 3D imaging, distance sensors are indispensable, whose typical characteristics are shown and compared in Table 1. LiDAR uses time-of-flight (ToF) technology to measure the distance of spatial points, describing surroundings with a set of accurate point clouds, whose limitations are its relatively high cost and poor performance to detect transparent objects such as glass directly, which is common in many office buildings [2], [3]. Depth camera (3D camera) projects structured light onto the object and capture it with a camera, obtaining the 3D structure by generating a 2D image along with the corresponding depth image. Although it's not as expensive as LiDAR, structured light is easily affected by other lighting conditions, which limits depth camera's applications to indoor environment without transparent objects [4].

Ultrasonic sensor detects the distance to objects by emitting high-frequency sound waves and calculating the time interval between emission and reception to determine the distance travelled, has been widely used in areas such as smart parking systems [5] and unmanned vehicles [6]. It's cheap and reliable in challenging environment [7] with

The associate editor coordinating the review of this manuscript and approving it for publication was Riccardo Carotenuto.

TABLE 1. Typical characteristics comparison of the most representative distance sensors.

Type	Cost	Resolution	Refresh Rate	Detection Range	Environmental Adaptability	3D Imaging Capability	Transparent Obj Detection
LiDAR	High	High	Medium	Long	Good	✓	-
Depth Camera	Medium	High	Fast	Medium	Fair	✓	-
Ultrasonic Sensor (single)	Low	Low	Slow	Short	Excellent	-	✓
Ultrasonic Sensor (array)	Low	Medium	Slow	Medium	Excellent	✓	✓

transparent objects detection capability. However, single ultrasonic sensor is unable to distinguish objects due to its low resolution. Inspired by sensor arrays, integrating multiple ultrasonic sensors namely ultrasonic phased array can generate beams with higher resolution and concentrated energy like LiDARs to acquire distance directly. Despite this, such ultrasonic point clouds are still too sparse to describe surroundings, which is attributed to ultrasonic beam's limited angular resolution. On the other hand, camera's dense pixels provide high resolution measurements but no distance directly. Hence, fusing camera and ultrasonic phased array together for 3D imaging can combine strengths of both. In this paper, we simulate an air-coupled ultrasonic phased array and propose a method fusing that and camera to reconstruct indoor rooms, whose simulation procedure is thoroughly demonstrated in Fig. 1. Besides, the method is applied and validated in different Gazebo scenarios. Specifically, the main contributions of this work are listed as follows:

1) We design a simulator for ultrasonic phased array based on acoustic models and principles, where multiple signal processing algorithms are applied. The frame rate of the array is significantly enhanced by adopting orthogonal frequency-division multiplexing (OFDM).
2) We propose a multi-view 3D reconstruction method for indoor environment using an ultrasonic phased array and a camera. Two scanning strategies are developed for different scenarios.
3) The proposed method is evaluated under different indoor scenarios in Gazebo simulation environment. Comparisons with other baseline methods verify the method's effectiveness.

II. RELATED WORK

A. IN-AIR ULTRASONIC 3D IMAGING

As demonstrated in Section I, ultrasonic sensor has promising applications due to its robustness in the presence of optically reflective or transparent objects, and more importantly, its low cost. However, a single ultrasonic sensor is utilized as a 1D range finder for limited directional information it provides. The idea of combining a number of sensors into a phased-array arrangement has been applied in medical sonography [8] and underwater sonar [9]. Unlike tissue or water, ultrasonic 3D imaging for mobile robots is challenging mainly for four reasons: sound speed limitation, acoustic energy attenuation, grating lobe suppression, and array volume selection.

Based on the fact that sound propagates about 340m/s in the air (15°C, 1 atm) which is four times slower than that in water or tissue, the frame rate is limited apparently when scanning. Some techniques aim at increasing the frame rate by reducing the number of scanning points, sacrificing resolution as a trade-off. e.g., fan-beam scanning [10] and diverging wave transmission [11]. Others applying advanced waveform encoding techniques like orthogonal frequency division multiplexing (OFDM) [12] allow scanning multiple points simultaneously. Nevertheless, transducers with a wide bandwidth are always needed under such cases. The next challenge is that acoustic energy dissipates easily in the air, limiting the max detection range. Phased array with multiple transducers is necessary, while a suitable signal frequency is equally important. Although a higher frequency provides a higher angular resolution, several in-air applications like levitation of small objects [13], power transfer [14] and acoustic vortex [15], prefer lower resonant frequency (typically lower than 100 kHz) to avoid strong acoustic energy attenuation.

Grating lobes occur when the transducer's size is too large to meet the $\lambda/2$ criterion for the maximum inter element spacing, resulting sound emissions in unintended directions, which not only brings ambiguities to acoustic imaging, but also causes harm to nearby users. To suppress grating lobes, non-uniform sparse array is proposed by breaking the periodicity of the element arrangement, which has been applied to beamforming in sound source localization [16], [17]. A more conservative approach is to meet the $\lambda/2$ criterion, which has been accomplished by developing tiny transducers based on PMUTS [18], CMUTS [19] or PVDF [20]. However, the tinier the transducer is, the higher cost it will take. This is also applicable to those attaching specific physical structures such as the 3D printed wave guide [21], [22] and the shrinking tubes by Konetzke et al. [23]. Considering the array volume selection, usually a large-aperture array with more elements has greater power and smaller angular resolution. On the other hand, an array with a large aperture is difficult to be integrated onto mobile robots for its large volume.

Our objective here is to design a relatively high frame rate ultrasonic phased array simulator for mobile robots, which can provide depth information by scanning points within its field of view (FOV) like LiDARs. Considering

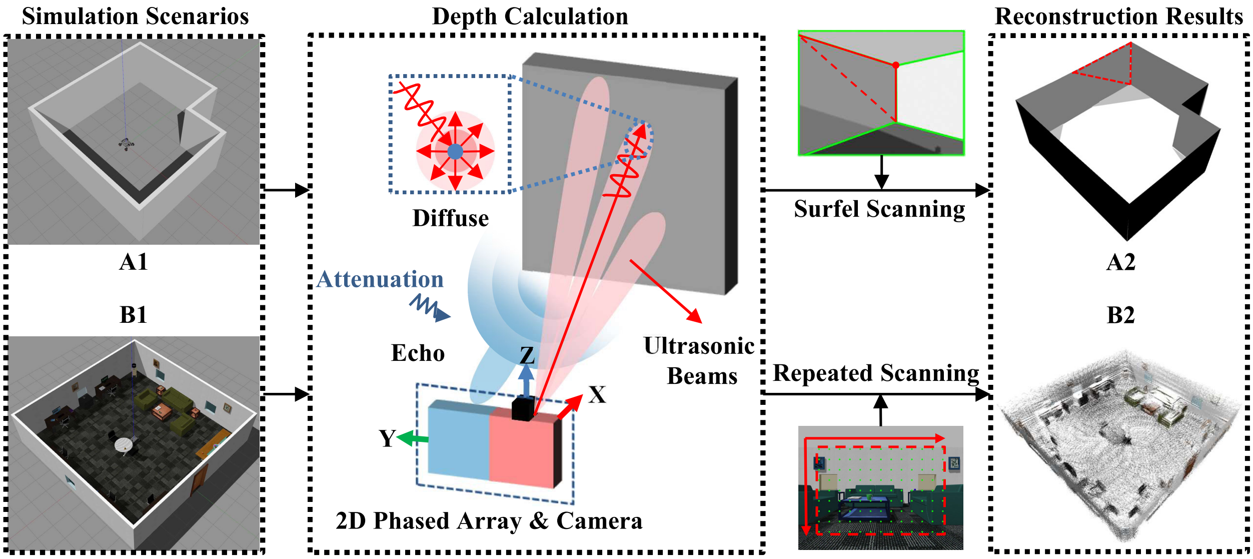


FIGURE 1. The schematic diagram of our reconstruction method. A1 is a scenario that is flat and lacks textures, a monocular camera identifies triangle surfels by ray-point-ray features, while a phased array scans these surfels to determine their positions with only a few points effectively, outputting a surfel map A2. B1 is a scenario with multiple objects and different textures, the phased array scans repeatedly within its field of view to generate a voxel map B2, whose colors and textures are provided by the camera.

the challenges mentioned above with these previous efforts, we simulate a non-uniform sparse 5 × 5 phased array with a 10 cm × 10 cm rectangular aperture, whose elements have a resonant frequency of 40 kHz. As a result, such an air-coupled ultrasonic phased array is a trade-off among the above challenges. Compared with other simulators like Field II, ours is not limited to sound intensity distribution or beam pattern calculation, but includes both transmitter and receiver to simulate the complete process of beam scanning and depth calculation. Apart from being validated in MATLAB, the simulator has been packaged as a C++ shared library, which enables its applications in the ROS Gazebo simulation environment.

B. RECONSTRUCTION METHODS

Reconstruction is an important application for mobile robots to perceive surrounding environment. As a range finder, a single ultrasonic sensor is typically capable of converting surroundings into a 2D occupied grid map for navigation and localization [24]. With the ultrasonic 3D imaging techniques mentioned above, an additional information of dimension is available. However, the number of works on in-air ultrasonic 3D reconstruction is still limited, which is mostly attributed to ultrasonic speckle's relatively large size. Compared with the laser speckle by LiDARs, the ultrasonic speckle is so large that significant errors occur when the beams hitting the edge of an object or the incident angle is too large, limiting its usage to detecting flat surfaces. e.g., the L-shaped obstacle reconstruction [25] whose scanning points are fitted by RANSAC method [26] to further reduce errors. Despite learning methods like supervised variational autoencoder (VAE) have been applied to reconstruct objects regardless of their surfaces [27], a training data is necessary here to approximate their shapes.

Compared with the methods above, a more natural idea is to identify those flat regions in the environment, whose information can be easily provided by images as detailed in Section I. Furuhashi et al. [28] has put this idea into practice. They use a camera to measure the shape of an object and an ultrasonic array sensor with 16 receivers to obtain a 3D image, whose results are combined into a depth image to reveal the shape of the object. Despite obtaining depth images, this method is only applicable to objects that are directly facing the sensors and relatively flat, like the aluminum plate in the study. Moreover, the 3D images acquisition cannot be done in real-time, limiting its application in the field of mobile robots.

To handle different scenarios, our proposed reconstruction method consists of two modes. The first mode, surfel scanning mode, is valid for flat, texture-less and structural scenarios, adopting the similar idea in [28]. The difference lies in that those flat regions are identified by a camera based on ray-point-ray [29] features, after which only a sparse set of scanning points is required to determine these regions in space. Compared with the method in [28], ours not only operates in real-time but also can measure inclined objects such as tilted walls. More importantly, because our transmitter is also a phased array rather than a single ultrasonic sensor, better beam directionality and detection range are available. The second mode, repeated scanning mode, is applicable to a wider range of indoor environments. We increase ultrasonic phased array's frame rate by employing OFDM technology to generate an ultrasonic point cloud map, and output a 3D voxel map finally.Although this mode can also be applicable to environment where surfel scanning mode operates, surfels can reconstruct with fewer points and higher efficiency. Therefore, they complement rather than replace each other.

III. METHODOLOGY

A. OVERVIEW

Fig. 2 shows the overall pipeline of the framework. Initially, a monocular camera and an ultrasonic phased array are calibrated extrinsically to match their FOV. Then for the surfel scanning mode, the image segmentation module extracts line features from the grayscale images by camera to generate two-dimensional triangle surfels as representations of flat regions. These flat regions are scanned by the transmitting array using ultrasonic beams, whose echoes are fed to the echo processing module and processed by our receiver. In echo processing, a delay and sum beam former is firstly performed, whose output is fed into a matched filter for increasing the signal to noise ratio. Next, the pulse compressed signal is enveloped whose maximum value is taken as echo's time-of-arrival, after which the depth information is available.

In the post processing, once the position of a reflector is determined, its ultrasonic point cloud is registered into a global map, followed by 3D triangle surfels described by 4 points (3 for generation and 1 for verification). Plane optimizations are performed on these surfels through cosine similarity judgment and global least square optimization, during which their positions are adjusted and optimized to achieve a better accuracy. These 3D surfels are fused with the image information for vertex calculation, and their corresponding point clouds are updated accordingly during this process. The post processing will run continuously until the array stops scanning and no more point clouds are generated in the map.

For the repeated scanning mode, the calibration process is the same as the surfel scanning mode. However, multiple ultrasonic beams are generated at once, thus the echo is the summation of them. Then an OFDM module is performed to decode the echo and obtain depth information for multiple points, whose point clouds are calculated and projected onto RGB images to generate a point cloud map like a Lidar does. The final step involves a filtering module to eliminate outliers, downsample point clouds, and output a voxel map, serving as the representation of surrounding environment.

B. ULTRASONIC PHASED ARRAY SIMULATION

Considering the challenges detailed in Section II, we simulate a 5×5 non-uniform sparse array with a 10 cm×10 cm rectangle aperture (Fig. 3). To suppress grating lobes, we start with a 5×5 uniform array and set the peak side lobe level (PSLL) within the FOV as our objective function, followed by adjusting the positions of the array elements with a genetic algorithm [30] to minimize the objective function. The 10 cm aperture meets the miniaturization requirements for integration with mobile robots while maintaining an angular resolution of about 5° based on the 3dB definition (the right beam pattern in Fig. 3 when the steering angle is 0°). The transducer we simulate is TCT40-10T/R with a 40 kHz resonant frequency, whose simulated beam pattern at 0° steering angle is shown on the left of Fig. 3 with an obvious directionality. Compared with other commonly used frequency in the air like 75 kHz, this frequency achieves a balance between energy dissipation and angular resolution.

Fig. 4 shows the schematic of the simulated ultrasonic system. As depicted in Fig. 4a, the transmitter array is located on the XY plane, whose normal is oriented along the positive Z axis. The reflector P_i is located at the point $(\rho_i, \theta_i, \varphi_i)$ which needs to be determined. The distance between the origin of the array and the reflector is given by ρ_i, while θ_i and φ_i are elevation (0°-90°) and azimuth (0°-360°) angles respectively. Our goal here is to obtain the response of the reflectors. We consider each array element as a vibrating sound source emitting a sinusoidal pulse consisting of 20 cycles for one time, which are corresponded with the excitation and decay processes. Then we can model a reflector's response:

$$s^R(t) = \sum_{i=1}^{25} \frac{d_i^T E_0}{r_i^T} \sin\left[\omega(t - t_i^D) - kr_i^T + \phi_i\right] \tag{1}$$

where s^R is the response obtained by superimposing the vibrations of 25 array elements. Unlike ideal spherical waves, our transducer has directionality, whose beam pattern has been shown in Fig. 3, hence d_i^T describes the distribution of emitted energy in all directions. E_0 represents a single transducer's power, and r_i^T denotes the distance between the ith element and the reflector [31], which is available in the Gazebo simulation environment:

$$r_i^T = \sqrt{(x_i^T - x)^2 + (y_i^T - y)^2 + z^2} \tag{2}$$

The part within brackets in (1) describes phase changes, where ωt_i^D denotes the phase change by phased array's time delayed pulses and kr_i^T denotes that by radial distance r_i^T. Besides, considering that the excitation time of each transducer can be different, the φ_i here is to describe the initial phase of each transducer, which can be obtained by array calibration.

Once an ultrasonic beam hits the surface of an object, the corresponding speckle is determined consisting of multiple reflectors due to array's 5° angular resolution (as shown in Fig. 4b). Here we divide the speckle into 25 reflectors (the orange dots) with a spacing of 1.25 degree, whose responses are superimposed together and then received by each microphone:

$$s^M(t) = \sum_{i=1}^{25} \frac{d_i^R \alpha\beta R(Z_1, Z_2)}{r_i^R} s^R(t - r_i^R/c) \tag{3}$$

where c is the speed of sound (≈340 m/s in the air) and r_i^R is the distance between the ith reflector and the microphone. Similar to the transmitting transducer, the directional characteristics of the microphone result in different energy reception intensities from different angles, which is described by d_i^R.

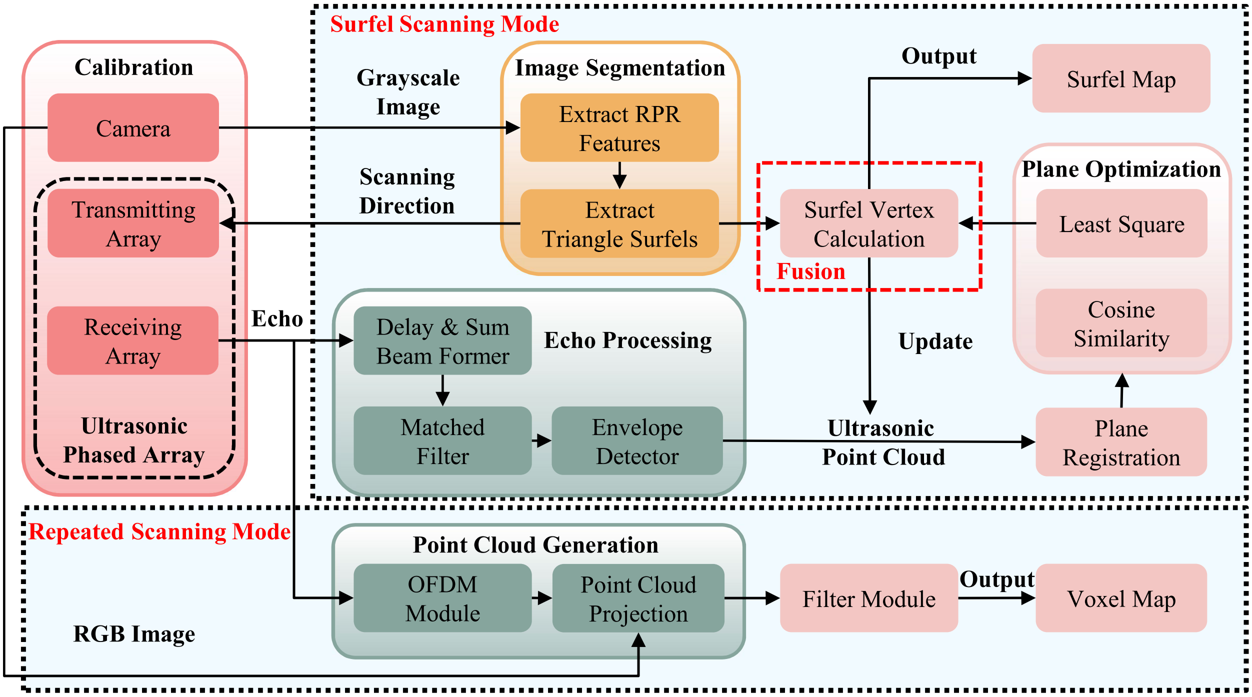


FIGURE 2. The pipeline of the framework combines ultrasonic point clouds and monocular camera images to generate a surfel map or a voxel map for different scanning modes respectively. For the surface scanning mode, images from the camera are segmented into triangular surfels firstly. Then an ultrasonic phased array scan these surfels to determine their positions, followed by an optimization of surfels belonging to the same plane. Finally, a surfel map is generated. For the repeated scanning mode, point clouds are generated by an OFDM module and subsequently projected onto an image to acquire color information. After passing a filter module, a voxel map is generated.

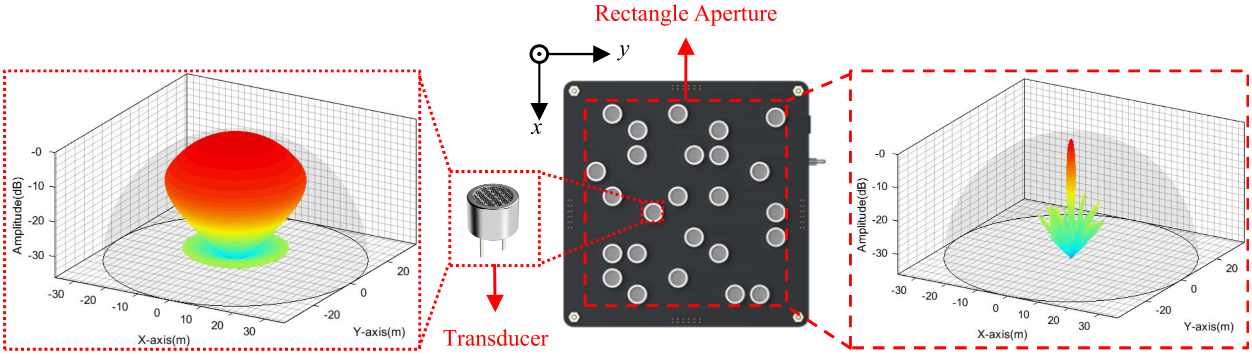


FIGURE 3. The transducer (TCT40-10T/R) and rectangle sparse array we simulate (middle), along with their beam patterns respectively (left and right).

For coefficients α and β, β represents the ratio of scattered energy to the total incident energy when the ultrasonic beam hits the interface. While α represents the proportion of the echo received by the array in relation to the total scattered energy, which can be obtained by calculating its solid angle. The reflectivity R determined by the acoustic impedance Z_1 and Z_2 on both sides is given by:

$$R(Z_1, Z_2) = \left(\frac{Z_2 - Z_1}{Z_2 + Z_1} \right)^2 \tag{4}$$

After these reflections are received by microphones in the 25-elements array individually, we step into the echo processing module including a beamformer, a matched filter and an envelope detector as Fig. 2 shows. In the first step, the beamformer acts as an acoustic lens which is steered into a particular direction ψ in the frontal hemisphere. Similar to the transmitting array, here $\psi = [\theta \ \varphi]^T$ where θ is the elevation angle and φ is the azimuth angle. In our case a delay and sum (or time-domain Bartlett) beamformer is used for its simplicity and robustness, moreover, it is capable of achieving better peak side lobe level to improve detection accuracy. We add time-delays $t_i^B(\psi)$ to each channel to compensate for the angle-dependent difference caused by the array's geometry:

$$s_\psi^B(t) = \sum_{i=1}^{25} s_i^M \left[t + t_i^B(\psi) \right] \tag{5}$$

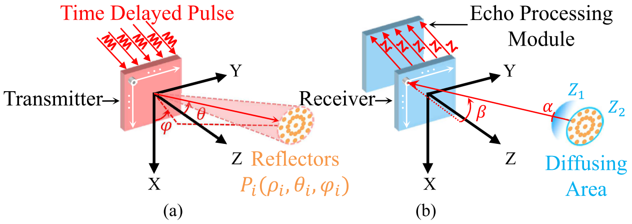


FIGURE 4. Schematic of the simulated ultrasonic system (a) Transmitter's working principles. The transmitter is driven by a series of time delayed pulses, and reflectors (orange dots) are calculated when the ultrasonic beam (θ,φ) hits an object. Here φ and θ are the azimuth and elevation angles respectively. (b) Receiver's working principles. Each transducer's echo is the sum of reflector responses within the diffusing area. Echo wave's amplitude decreases due to impedance change (Z_1,Z_2), solid angle energy loss (α) and transducer's gain loss (β). These echoes are processed in the Echo Processing Module to obtain depth information.

The next process is a matched filter which not only increases the signal to noise ratio (SNR) but also compresses the signal into its auto-correlation function. This filtering process [16] leads to the pulse compressed signal $s^F(t)$ and can be described by the following equation:

$$s^F(t) = \mathcal{F}^{-1}\left\{S^B_\psi[j\omega] \cdot S^*_e[j\omega]\right\} \tag{6}$$

Here $\mathcal{F}^{-1}$ is the inverse Discrete Fourier Transform (DFT) applied on the discrete fourier transforms of the signal after beamforming ($S^B_\psi[j\omega]$) and its complex conjugate of the fourier transform of the emitted signal ($S^*_e[j\omega]$). Followed by an envelope fitting method [32] to detect the envelope of the filtered signal, the time-of-arrival is available, and the object's depth d is computed as follows:

$$d = \frac{l^2 - b^2}{2(l - b\cos\theta)} \tag{7}$$

In this equation, l represents the distance calculated from the time-of-arrival, b refers to the baseline between the transmitting and receiving arrays, and θ is the elevation angle of the emitted ultrasonic beam. The depth d here is relative to the origin of the transmitting array.

C. SCANNING FRAME RATE ENHANCEMENT

Considering indoor scenarios and the speed of sound limitations, the range of the simulated ultrasonic phased array is set to 5 meters, which means each transmission and reception cycle must be completed within approximately 30 ms. If we detect one point at once, the scanning frame rate is too low to perform repeated scanning like a LiDAR. Orthogonal frequency division multiplexing (OFDM) allows to select a group of mutually orthogonal frequencies or their combinations when transmission, and then decoding occurs using correlated signals when reception, enabling the detection of multiple points during a single transmission and reception cycle. If we have two signals: $S_1(t)$ and $S_2(t)$, then their orthogonality can be mathematically described as:

$$\int_0^T sin(\omega_1 t + \phi_1)sin(\omega_2 t + \phi_2)dt \tag{8}$$

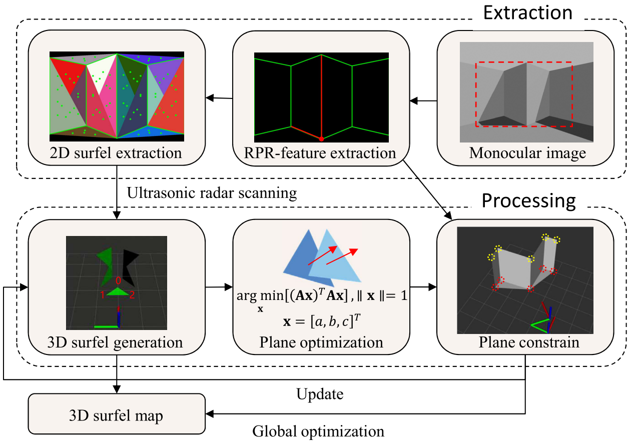


FIGURE 5. Surfel extraction and processing workflow.

Here T is the duration of the signal, ω_1 and ω_2 are angular frequencies of $S_1(t)$ and $S_2(t)$, while ϕ_1 and ϕ_2 are their corresponding initial phases. After decomposition and integration in equation (8), we obtain orthogonal conditions between signal $S_1(t)$ and $S_2(t)$, where f_1 denotes the frequency of $S_1(t)$, and $\Delta f = f_2 - f_1$ represents the frequency difference between two signals above:

$$(2f_1 + \Delta f)T = k, \Delta fT = k, k \in Z \tag{9}$$

The conditions in (9) imply that only when both $(2f_1 + \Delta f)T$ and ΔfT are integers, the signal $S_1(t)$ and $S_2(t)$ are orthogonal. In our case, each sinusoidal pulse consists of 20 cycles, thus T is 0.5 ms. If we select $\Delta f = 4$ kHz, then equation (9) is satisfied. As a rule of thumb, the operational frequencies should be centered around the transducers' resonant frequency, thus we select 5 frequencies ranging from 32 kHz to 48 kHz, spaced at intervals of 4 kHz, to create 25 combinations in pairs for our transmit pulses within 30 ms. By this way, the detection time for each point is reduced from 30 ms to 1.7 ms, making it possible for repeated scanning.

D. SURFEL SCANNING RECONSTRUCTION

For the surfel scanning mode, although the depth information is provided by the ultrasonic phased array as mentioned above, either the frame rate or the accuracy is limited. Hence inspired by the existing reconstruction methods, we propose a method fusing ultrasonic point clouds and monocular camera images for indoor 3D reconstruction. Corresponded with the image segmentation and post processing in Fig. 2, we use a monocular camera to identify those flat regions by performing an image segmentation, followed by the point clouds generated by the array scanning, and the points that belong to the same region are organized in the form of triangle surfels, serving as the representation of the environment. Once the 3D surfels are generated, they are registered into a global map and continuously optimized and updated, achieving a globally consistent reconstructed map. The process of fusion can be divided into two steps (Fig. 5):

1) Surfel Extraction: The camera we simulate is MER2-202-60GM/C with a frame rate of 25 frames per second and a resolution of 1600 (H)×1200 (V). The process starts with extrinsic calibration performed between the array and camera, where the point clouds are projected onto a monocular image to match their coordinate systems. Here, the camera's intrinsic parameters are specified by the URDF file, whose coordinate system has its origin positioned 7.5cm above the array's origin (as shown in Fig. 1 and Fig. 3). Hence, the extrinsic parameters are also derived accordingly. For a new monocular image, since the array's FOV is smaller, only the part within the FOV is kept, after which a Fast Line Detector (FLD) is performed on the cropped image to extract line features. During this step, the features belonging to the same line are merged while those with a length smaller than 50 pixels are discarded. Based on the fact that a plane can be determined by any two line segments in space with a common intersection point, triangle surfels (the red dashed triangle) are extracted by such line segments as representations of the flat regions in the environment. Generally, each surfel can be uniquely determined by 3 non-collinear points in space. Here we select three points that are one-third of the way along the line connecting the centroid and the endpoints of the surfel. With the previously calibrated intrinsic parameters, each point's position can be measured by an ultrasonic beam, whose elevation and azimuth angle (θ, φ) can be derived as:

$$\theta = \arctan(\sqrt{(u-u_c)^2 + (v-v_c)^2}, f) \tag{10}$$

$$\varphi = \arctan(v - v_c, u - u_c) \tag{11}$$

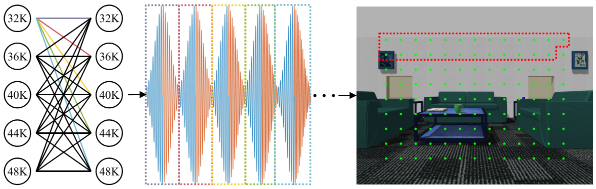
where (u_c, v_c) is the center coordinate of the image, and f is the focal length of the camera. Note that the range of φ is 0° to 360°, here a 2π should be added if $\varphi < 0$. Our array will scan the three points counterclockwise, along with the surfel's centroid (4 points in total). Only when the distance between the measured centroid and the plane formed by the measured three points is below a threshold (0.15 m), will the corresponding surfel represented by 4 points be registered into a global map. The registration process is divided into three steps: Firstly, a ground truth pose of the sensor relative to the world coordinate system $\mathbf{q}_i$ can be obtained by a wheel odometer in the simulation environment. Next, the pose is decomposed into a rotation matrix $\mathbf{R}_i$ and a translation vector $\mathbf{t}_i$. Since the coordinate of the point relative to the sensor $\hat{\mathbf{p}}_i$ has been measured, the final step is to transform $\hat{\mathbf{p}}_i$ into the world coordinate system, given by:

$$\mathbf{p}_i = \mathbf{R}_i \hat{\mathbf{p}}_i + \mathbf{t}_i \tag{12}$$

2) Surfel Processing: The surfel processing is performed after the surfels are determined and registered. In 3D surfel generation, surfel $\mathbf{S}_k$ is visualized by a triangle that contains the vertices:

$$\mathbf{S}_k = [\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3];\ \mathbf{p}_i = [x_i, y_i, z_i]^T, i = 1, 2, 3 \tag{13}$$

the next step is the plane optimization where a cosine similarity is performed to determine whether two surfels belong to the same plane, followed by a least square method to optimize the vertices of these surfels, which can be described by the following minimization problem:



(a) Frequency Combinations (b) Transmit in Sequence (c) Multiple Points Detection

FIGURE 6. Transmit different frequency combinations sequentially in one cycle and detect multiple points using OFDM: (a) 25 frequency combinations ranging from 32 kHz to 48 kHz, with an interval of 4 kHz. (b) Transmit different combinations in sequence within 30 ms. (c) Detect 25 points simultaneously (red dashed box).

$$\arg\min_{\mathbf{x}} \left[(\mathbf{A}\mathbf{x})^T \mathbf{A}\mathbf{x} \right] \tag{14}$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{p}_1 - \bar{\mathbf{p}} \\ \vdots \\ \mathbf{p}_n - \bar{\mathbf{p}} \end{bmatrix}, \mathbf{x} = [a, b, c]^T \tag{15}$$

Here $\mathbf{A}$ contains a set of surfel vertices to be optimized, organized as an $n \times 3$ matrix. $\bar{\mathbf{p}} = [\bar{x}, \bar{y}, \bar{z}]$ is the centroid of all points. $\mathbf{x}$ represents the plane to be fitted, whose L_2 norm is 1. After the vertices of the origin surfels are projected onto the fitted plane, these planes will automatically expand until they intersect or exceed the field of view of the sensor (the red dashed circles in Fig. 5). Besides, they are constrained by the line segments extracted from the mono image (the yellow dashed circles). After going through these two steps, the vertices are further optimized to make the corresponding surfels more precise. Up to now a submap has been generated, and this part is performed among submaps from different poses for a global consistency, until no more surfels are generated.

E. REPEATED SCANNING RECONSTRUCTION

For the repeated scanning mode, our beam scanning directions are no longer determined by the camera. Instead, they are determined by the ultrasonic phased array's FOV and beam angle. From the simulation results in Section B, we obtain a 60° ×40° FOV and a 5° beam angle, which means each frame contains 13 × 9 points. Our strategy here is to transmit the 25 frequency combinations sequentially within 42.5 milliseconds (Fig. 6). By this means, the time cost for each frame is 208.5 milliseconds, reaching a 5 Hz frame rate.

As Fig. 2 depicts, after passing the OFDM module, the generated points clouds are projected onto the RGB images according to the intrinsic and extrinsic parameters obtained during the calibration. The projection process adds RGB color information to each point cloud, followed by transforming the point's coordinate into the world coordinate system in equation (12). The next step is map generation, where a sparse point cloud map is generated first. Then to achieve a better reconstruction outcome, the final step involves filtering

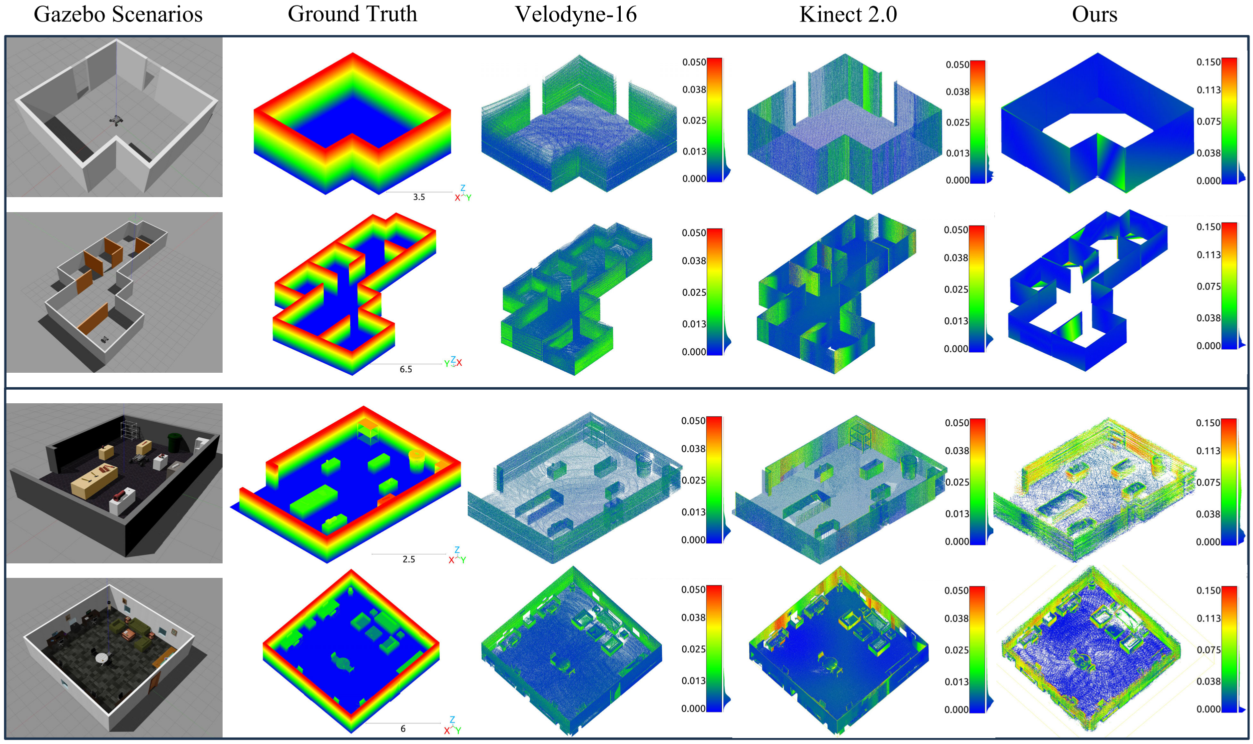


FIGURE 7. Reconstruction comparisons with a Velodyne-16 and a Kinect 2.0 in different Gazebo scenarios: a single room with glass windows, an indoor structure with multiple rooms, a small workshop and a large office.

and voxelizing the point cloud map, utilizing voxels as a representation of the surrounding environment. A voxel's color is the average of the points' colors contained within the voxel.

IV. EXPERIMENTS AND RESULTS

In this section, we select four scenarios to evaluate the effectiveness of our method in Gazebo simulation environment, two of which are ideal indoor structures lacking textures and objects, aiming at validating the surfel scanning mode (Section III-D). Another two scenarios are a small workshop and a large office, which are selected to validate the repeated scanning mode (Section III-E). Additionally, we compare our proposed method with a Velodyne-16 LiDAR and a Kinect 2.0 in each experiment. For ease of control, each sensor is mounted on a Scout Mini with an offset of 0.1 meters and 0.5 meters in the x-direction and z-direction from the center of the chassis respectively. All experiments are carried out on the same PC platform with an Intel Core i7-11700K @ 3.6GHz with 16GB memory.

A. SURFEL SCANNING RESULTS

As depicted in Fig. 7, the first scene is a single room which covers an area of 45 square meters with two glass windows, while the second scene is a structure consisting of multiple rooms and interior walls (orange), both scenes are lack of textures with light effect. A ground truth point cloud map is generated by sampling on the scene models with 1 cm resolution for accuracy evaluation. To match this resolution, once surfels are generated and fused into one global map, they are sampled with the same resolution. To maintain the consistency of the reconstruction results, the poses of the sensors are obtained uniformly by the robot's wheel odometer. We use CloudCompare to compute the point-to-point distance error between the ground truth and the reconstructed point cloud.

Since the reconstruction results are characterized by point clouds, we consider their accuracy, consistency and completeness as metrics for evaluation. For our results (Fig. 7), accuracy is represented by color variations, consistency is depicted by the continuity of color changes, while completeness is indicated by those missing parts when comparing the reconstruction results with ground truth. In terms of accuracy, our surfel scanning method ensures a level of precision comparable to that of LiDARs and depth cameras, namely, 90% of the points' distance error is less than 2 cm. Regarding the consistency of point clouds, the results from the depth camera exhibit multiple abrupt color variations, while ours is on par with LiDARs and surpassed that of depth cameras. This is because simply transforming each frame into the world coordinate system using odometry leads to abrupt accuracy variations. In spite of this, our global optimization of surfels ensures effective global consistency. As for the completeness of point clouds, our method outperforms both LiDARs and depth cameras. Although it is challenging to reconstruct extensive ground areas, our camera's FOV ensures completeness of walls and corners, while ultrasonic beams are able to detect the glass

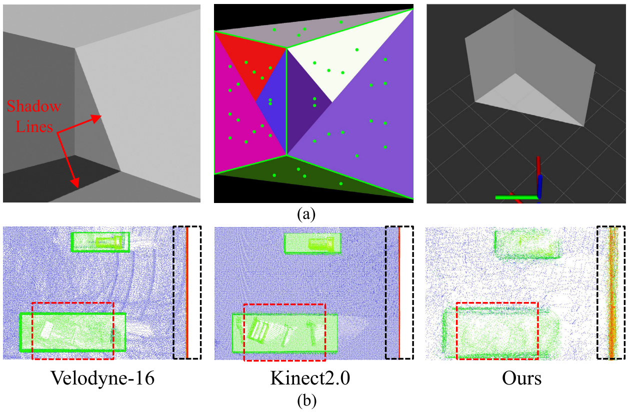


FIGURE 8. Challenging and failure cases: (a) A challenging case with shadow lines due to light effect. (b) A failure case with small objects and high movement speed.

windows directly. Additionally, owing the representation method of surfels, only minimal points are needed to generate the densest point clouds in our method.

B. REPEATED SCANNING RESULTS

In order to maintain the consistency of results, we employ an experimental setup similar to Section IV. A both in the workshop and office scenarios. The only difference is the utilization of repeated scanning for our sensor, not surfel scanning. As shown in Fig. 7, in terms of accuracy, our repeated scanning method ensures about 70% of the points' distance error within 2 cm both in the small workshop and large office, while that of the remaining points range from 2 cm to 15 cm. Regarding completeness and consistency, unlike LiDARs and depth cameras, although it is challenging for our ultrasonic beams to reconstruct small objects such as tools in the workshop and cups in the office with sparse point clouds, they still effectively capture larger objects like tables, chairs and sofas, as well as the small glass windows in the large office. That is to say, a colorful point cloud map similar to that of LiDARs and depth cameras, which is generated by a low-cost ultrasonic phased array and a monocular camera, is available in this scanning mode, along with the good performance in detecting transparent objects.

C. CHALLENGING AND FAILURE CASES

Besides the reconstruction results presented above, to further demonstrate our method's characteristics, we show one challenging case and two failure cases of our method (Fig. 8). As for the surfel scanning method, Fig. 8a is a challenging case with shadow lines caused by light effect, which always leads to errors in 2D surfel extraction. Despite this, our method will recognize and remove these excessive line features to make surfel extraction correct, obtaining the well-reconstructed corner (two vertical walls and the floor), showing our strategy's good performance.

As for the repeated scanning method, it is developed to address the limitation that surfel scanning is not suitable for more complex indoor environment, and Fig. 8b shows two cases when it fails. The first case depicts its inability to reconstruct very small objects and detailed surfaces due to the relatively large angular resolution of our ultrasonic beams, compared with LiDARs and depth cameras. Hence our method is more suitable for detecting indoor structures and larger objects, as well as close-range supplementary navigation for mobile robots. The second case depicts its inapplicability in high-speed scenarios, which is determined by the speed of sound and sensor frame rate. As can be seen in Fig. 8b, under the same pose transformation, our method yields thicker wall surfaces, compared with LiDARs and depth cameras. These errors stem from factors like OFDM calculation and odometry drift, but they are influenced by our robot's movement speed mostly.

V. CONCLUSION

In this paper we propose a method of fusing ultrasonic point clouds and camera images to generate a map for mobile robots in simulation environment. Two scanning strategies are developed for different scenarios. The surfel scanning mode is suitable for flat and texture-less scenarios. With the feature and depth information provided by a monocular camera and a simulated ultrasonic phased array respectively, 3D triangle surfels are generated to represent surrounding environment, which are transformed into a surfel map by post processing algorithms. For more complex scenarios, our repeated scanning mode selects multiple frequency combinations to simultaneously detect multiple points using OFDM, achieving a scanning frame rate of 5 Hz. The experimental results demonstrate promising reconstruction capabilities for these scenarios. Compared with LiDARs and depth cameras, ours utilizes camera and low-cost ultrasonic sensor to realize multi-view 3D reconstruction with good accuracy, consistency and completeness. Besides, challenging and failure cases of two scanning strategies are discussed. We hope this work would be helpful to mobile robots research or areas.

ACKNOWLEDGMENT

The author Jixiang Ren thanks DJI for DJI-Scholarship. The author Xiaoping Hong thanks the SUSTech-DJI Joint Laboratory.

REFERENCES

[1] Z. Wang, Y. Wu, and Q. Niu, "Multi-sensor fusion in automated driving: A survey," *IEEE Access*, vol. 8, pp. 2847–2868, 2020.

[2] H. Tibebu, J. Roche, V. De Silva, and A. Kondoz, "LiDAR-based glass detection for improved occupancy grid mapping," *Sensors*, vol. 21, no. 7, p. 2263, Mar. 2021.

[3] A. M. Wallace, A. Halimi, and G. S. Buller, "Full waveform LiDAR for adverse weather conditions," *IEEE Trans. Veh. Technol.*, vol. 69, no. 7, pp. 7064–7077, Jul. 2020.

[4] Y. Zhang, M. Ye, D. Manocha, and R. Yang, "3D reconstruction in the presence of glass and mirrors by acoustic and visual fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1785–1798, Aug. 2018.

[5] A. Fahim, M. Hasan, and M. A. Chowdhury, "Smart parking systems: Comprehensive review based on various aspects," *Heliyon*, vol. 7, no. 5, May 2021, Art. no. e07050.

[6] J. N. Yasin, S. A. S. Mohamed, M.-H. Haghbayan, J. Heikkonen, H. Tenhunen, and J. Plosila, ''Low-cost ultrasonic based object detection and collision avoidance method for autonomous robots,'' *Int. J. Inf. Technol.*, vol. 13, no. 1, pp. 97–107, Feb. 2021.

[7] N. Balemans, P. Hellinckx, and J. Steckel, ''Predicting LiDAR data from sonar images,'' *IEEE Access*, vol. 9, pp. 57897–57906, 2021.

[8] T. L. Szabo, *Diagnostic Ultrasound Imaging: Inside Out*. New York, NY, USA: Academic, 2004.

[9] H. Tian, S. Guo, P. Zhao, M. Gong, and C. Shen, ''Design and implementation of a real-time multi-beam sonar system based on FPGA and DSP,'' *Sensors*, vol. 21, no. 4, p. 1425, Feb. 2021.

[10] M. Karaman, I. O. Wygant, O. Oralkan, and B. T. Khuri-Yakub, ''Minimally redundant 2-D array designs for 3-D medical ultrasound imaging,'' *IEEE Trans. Med. Imag.*, vol. 28, no. 7, pp. 1051–1061, Jul. 2009.

[11] B. Lokesh and A. K. Thittai, ''Diverging beam transmit through limited aperture: A method to reduce ultrasound system complexity and yet obtain better image quality at higher frame rates,'' *Ultrasonics*, vol. 91, pp. 150–160, 2019.

[12] Q. Yan, Q. Xia, Y. Wang, P. Zhou, and H. Zeng, ''URadio: Wideband ultrasound communication for smart home applications,'' *IEEE Internet Things J.*, vol. 9, no. 15, pp. 13113–13125, Aug. 2022.

[13] Z. Long, H. Zhao, H. Peng, M. Yao, Y. Pan, and Z. Li, ''Acoustic levitation for large particle based on concave spherical transducer arrays,'' *IEEE Sensors J.*, vol. 22, no. 18, pp. 18104–18113, Sep. 2022.

[14] S. Surappa and F. L. Degertekin, ''Characterization of a parametric resonance based capacitive ultrasonic transducer in air for acoustic power transfer and sensing,'' *Sens. Actuators A, Phys.*, vol. 303, Mar. 2020, Art. no. 111863.

[15] S. Guo, Z. Ya, P. Wu, and M. Wan, ''A review on acoustic vortices: Generation, characterization, applications and perspectives,'' *J. Appl. Phys.*, vol. 132, no. 21, Dec. 2022, Art. no. 210701.

[16] R. Kerstens, D. Laurijssen, and J. Steckel, ''ERTIS: A fully embedded real time 3D imaging sonar sensor for robotic applications,'' in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 1438–1443.

[17] T. Verellen, R. Kerstens, and J. Steckel, ''High-resolution ultrasound sensing for robotics using dense microphone arrays,'' *IEEE Access*, vol. 8, pp. 190083–190093, 2020.

[18] Y. Birjis, S. Swaminathan, H. Nazemi, G. C. A. Raj, P. Munirathinam, A. Abu-Libdeh, and A. Emadi, ''Piezoelectric micromachined ultrasonic transducers (PMUTs): Performance metrics, advancements, and applications,'' *Sensors*, vol. 22, no. 23, p. 9151, Nov. 2022.

[19] T. M. Khan, A. S. Tasdelen, M. Yilmaz, A. Atalar, and H. Köymen, ''High-intensity airborne CMUT transmitter array with beam steering,'' *J. Microelectromech. Syst.*, vol. 29, no. 6, pp. 1537–1546, Dec. 2020.

[20] S. A. Pullano, C. D. Critello, M. G. Bianco, M. Menniti, and A. S. Fiorillo, ''PVDF ultrasonic sensors for in-air applications: A review,'' *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 68, no. 7, pp. 2324–2335, Jul. 2021.

[21] G. Allevato, J. Hinrichs, M. Rutsch, J. P. Adler, A. Jäger, M. Pesavento, and M. Kupnik, ''Real-time 3-D imaging using an air-coupled ultrasonic phased-array,'' *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 68, no. 3, pp. 796–806, Mar. 2021.

[22] T. Maier, G. Allevato, M. Rutsch, and M. Kupnik, ''Single microcontroller air-coupled waveguided ultrasonic sonar system,'' in *Proc. IEEE Sensors*, Oct. 2021, pp. 1–4.

[23] E. Konetzke, M. Rutsch, M. Hoffmann, A. Unger, R. Golinske, D. Killat, S. N. Ramadas, S. Dixon, and M. Kupnik, ''Phased array transducer for emitting 40-kHz air-coupled ultrasound without grating lobes,'' in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Oct. 2015, pp. 1–4.

[24] S. K. A. Nair, S. Joladarashi, and N. Ganesh, ''Evaluation of ultrasonic sensor in robot mapping,'' in *Proc. 3rd Int. Conf. Trends Electron. Informat. (ICOEI)*, Apr. 2019, pp. 638–641.

[25] N. I. Giannoccaro, L. Spedicato, and C. di Castri, ''A new strategy for spatial reconstruction of orthogonal planes using a rotating array of ultrasonic sensors,'' *IEEE Sensors J.*, vol. 12, no. 5, pp. 1307–1316, May 2012.

[26] J. M. Martínez-Otzeta, I. Rodríguez-Moreno, I. Mendialdua, and B. Sierra, ''RANSAC for robotic applications: A survey,'' *Sensors*, vol. 23, no. 1, p. 327, Dec. 2022.

[27] R. Ohara, Y. Yasuda, R. Hamabe, I. Toru, S. Izumi, and H. Kawaguchi, ''3D reconstruction from outdoor ultrasonic image using variation autoencoder,'' in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Oct. 2022, pp. 1–5.

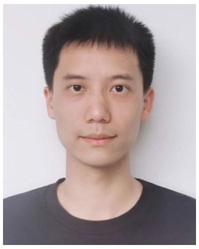
[28] H. Furuhashi, Y. Kuzuya, Y. Uchida, and M. Shimizu, ''Three-dimensional imaging sensor system using an ultrasonic array sensor and a camera,'' in *Proc. IEEE Sensors*, Nov. 2010, pp. 713–718.

[29] Y. He, X. Liu, X. Liu, and J. Zhao, ''Structure reconstruction using ray-point-ray features: Representation and camera pose estimation,'' in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 5388–5394.

[30] J. Li, S. Ren, and C. Guo, ''Synthesis of sparse arrays based on CIGA (convex improved genetic algorithm),'' *J. Microw., Optoelectronics Electromagn. Appl.*, vol. 19, pp. 444–456, Nov. 2020.

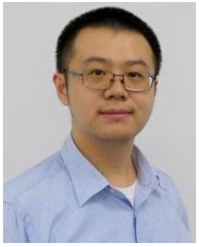
[31] D. B. Lindell, G. Wetzstein, and V. Koltun, ''Acoustic non-line-of-sight imaging,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6773–6782.

[32] Z. Qiu, Y. Lu, and Z. Qiu, ''Review of ultrasonic ranging methods and their current challenges,'' *Micromachines*, vol. 13, no. 4, p. 520, Mar. 2022.



JIXIANG REN received the B.E. degree in automation from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2021. He is currently pursuing the M.E. degree in intelligent manufacturing and robotics with the Southern University of Science and Technology (SUSTech), Shenzhen, China, under the supervision of Prof. Xiaoping Hong.

His research interests include novel sensor simulation and sensor fusion.



XIAOPING HONG (Member, IEEE) received the B.S. degree in physics from The Hong Kong University of Science and Technology (HKUST), Hong Kong, China, in 2009, and the Ph.D. degree in physics from the University of California at Berkeley, Berkeley, CA, USA, in 2014.

He joined the Southern University of Science and Technology (SUSTech), Shenzhen, China, as an Assistant Professor, in 2019. His current research interests include robotic sensors, and acoustic and optical imaging, with a focus on novel low-level sensor design and application.