Coherent 3D Acoustic Imaging on a Smartphone

Aryan Mahindra
Paul G. Allen School of Computer Science & Engineering
University of Washington
aryanm@uw.edu

March 2026
Preprint

Abstract

We present a smartphone-based acoustic imaging system that reconstructs a coherent three-dimensional point cloud of a nearby object using an unmodified iPhone's built-in acoustic transducers, with a motorized turntable providing controlled target rotation. The phone emits near-ultrasonic frequency-modulated continuous-wave (FMCW) chirps, records reflections, and performs reconstruction entirely on-device. Our central idea is to cast the problem as acoustic inverse synthetic aperture imaging: by keeping the phone fixed and rotating the object, we synthesize a circular aperture in the object frame without estimating a freehand sensor trajectory. Each acquired frame is converted to a complex range profile, and a GPU backprojection kernel coherently accumulates hundreds of views into a 3D voxel volume from which a point cloud is extracted. This design removes the dominant error mode of prior phone-based SAR imaging—unknown hand motion—while preserving the accessibility of commodity mobile hardware. The current prototype operates at approximately 0.30 m standoff, reconstructs a 0.30 m^3 volume sampled at 5 mm voxel spacing, and completes the on-device processing stage in under 30 s on an iPhone 15 Pro. Based on the published literature we could verify through March 2026, prior smartphone acoustic systems either produce 2D images, track only a small number of points, or reconstruct highly constrained structures such as hand skeletons. To the best of our knowledge, this is the first smartphone-based acoustic system aimed at coherent 3D volumetric reconstruction of arbitrary objects using only built-in phone acoustics and no external acoustic hardware.

1 Introduction

Active acoustic sensing on phones has advanced steadily over the last decade. Commodity speakers and microphones have been repurposed for around-device finger tracking, gesture sensing, indoor mapping, and coarse imaging in conditions where optical methods are impaired by darkness or occlusion [1, 2, 3, 4, 5]. The most relevant imaging result is AIM [7], which showed that a user can translate a phone along a line, synthesize a 1D aperture, and reconstruct a 2D acoustic image of a hidden object. That result established an important feasibility point: phone acoustics are not limited to single-point ranging and can support coherent image formation.

What has not been shown, however, is a phone-based system that produces a coherent 3D volumetric representation of an *arbitrary* object. Published smartphone acoustic systems fall into three neighboring categories. First, many systems estimate one or a few points in 1D–3D—for example fingertip, phone, or hand-joint positions—rather than reconstructing a dense scene representation [1, 2, 6, 10]. Second, phone acoustic imaging systems such as AIM [7], the Symmetry 2021 smartphone imaging system [8], and SONDAR [9] produce 2D acoustic images or silhouettes rather than 3D point clouds. Third, coherent 3D acoustic reconstruction has been demonstrated in laboratory synthetic-aperture sonar settings, but with specialized transducers, precision motion control, or off-device computation [12, 13, 14].

This paper closes that gap. We describe a prototype that uses a stationary iPhone and a motorized turntable to perform coherent 3D acoustic imaging in air. The turntable contributes only controlled object motion; all transmission, reception, signal processing, and reconstruction remain phone-resident. The key reparameterization is to treat object rotation as an object-centric synthetic aperture. In the object's frame, the stationary phone becomes a virtual speaker–microphone pair that traverses a circular path around the target. Because that trajectory is induced mechanically, the system avoids AIM's central difficulty: inferring a sufficiently accurate freehand sensor path from noisy measurements.

This first preprint focuses on system formulation, prototype characterization, and precise positioning relative to prior art. We report the prototype operating point, theoretical resolution limits, and on-device runtime, and we make the novelty boundary explicit rather than hiding it behind an unqualified "first" claim.

1.1 Contributions

1. A smartphone acoustic imaging design for coherent 3D volumetric reconstruction of arbitrary nearby objects using only the phone's built-in acoustic transducers, with a

motorized turntable supplying controlled object rotation.

2. An acoustic ISAR formulation that converts target rotation into a known circular synthetic aperture and removes the hand-trajectory estimation problem that dominates prior phone-based SAR approaches.
3. An end-to-end on-device pipeline—from FMCW acquisition to GPU backprojection and point-cloud extraction—that runs on commodity mobile hardware.
4. A clear prior-art boundary showing how the present system differs from 2D phone acoustic imaging, point-tracking systems, and laboratory 3D acoustic reconstruction.

2 Related Work and Prior-Art Boundary

2.1 Phone acoustics: tracking, mapping, and 2D imaging

A large literature uses smartphone or wearable acoustics to estimate a small number of geometric degrees of freedom. FingerIO [1] tracks a fingertip in 2D using active sonar. LLAP [2] achieves millimeter-scale displacement tracking using phase changes in acoustic signals. MilliSonic [3] pushes tracking precision further, while ReflecTrack [6] estimates 3D position using a dual-microphone smartphone and environmental reflections. These systems are impressive sensing results, but their output is a trajectory or sparse set of points rather than a spatial image.

Two lines of work moved phone acoustics closer to imaging. AIM [7] uses linear sensor motion to synthesize an aperture and reconstructs 2D acoustic images. Li *et al.* [8] also reconstruct a 2D contour-like heat map by sweeping a phone while recording reflections and accelerometer data. SONDAR [9] adapts inverse synthetic aperture ideas to commodity devices for size and shape measurement, but its output remains a 2D acoustic image. BatMapper [4] and SAMS [5] reconstruct room or floor-plan structure from phone acoustics, yet they target large planar reflectors and produce maps rather than arbitrary-object point clouds.

2.2 3D output that does not meet the target setting

A smaller set of acoustic systems does produce a 3D output of some kind, but not under the joint constraints we target. SonicHand [10] reconstructs a 3D hand skeleton from smartphone audio, but the output is a fixed-topology set of hand joints learned under a strong hand prior. Shih and Rowe's "Can a Phone Hear the Shape of a Room?" [11] reconstructs room geometry, but the setup uses a separate fixed speaker module and targets room surfaces rather than arbitrary tabletop objects. At laboratory scale, recent in-air synthetic-aperture sonar datasets and neural volumetric reconstruction methods show high-fidelity coherent 3D reconstruction with specialized hardware and workstation-class processing [12, 13]. Classic acoustic ISAR has also been demonstrated with research sonar hardware [14].

2.3 Novelty boundary

Table 1 summarizes the boundary. The closest published works satisfy at most two of the three defining properties we care about: phone-native acoustics, coherent 3D volumetric output, and arbitrary-object generality. AIM and related phone imaging systems satisfy the hardware constraint and object generality but stop at 2D [7, 8, 9]. SonicHand satisfies the phone-native hardware constraint and produces a 3D output, but only for a learned hand skeleton [10]. Laboratory SAS and room-geometry systems produce 3D results, but not in the target phone-only sensing setting [11, 12, 13].

We therefore phrase the novelty claim carefully: *to the best of our knowledge, and based on the literature we could verify through March 2026, we are not aware of a prior published system that combines built-in smartphone acoustics, coherent 3D volumetric reconstruction, and arbitrary-object generality in a single phone-resident pipeline.*

3 System Overview

Figure 1 shows the operating principle. An iPhone is placed at a fixed standoff of roughly 0.30 m from a motorized turntable. The phone emits near-ultrasonic FMCW chirps, records the echoes on a microphone channel, and stores the raw samples in a real-time buffer. As the object rotates, each frame captures a slightly different aspect. Instead of reasoning in the phone frame, we transform to the object frame: a stationary phone observing a rotating target is equivalent to a virtual speaker–microphone pair moving around a stationary object. That object-centric view makes the aperture geometry explicit and lets reconstruction proceed using standard synthetic-aperture logic.

3.1 Acquisition

The prototype operates in the near-ultrasonic band where phone hardware still provides usable output but the emissions are less intrusive than audible chirps. The phone is configured for raw capture: echo cancellation, automatic gain control, and other system-level speech DSP are disabled so the captured waveform preserves phase and amplitude as much as possible. Frames are acquired at approximately 13 Hz. At a turntable period of roughly 15 s per revolution, this yields about 200 aspect samples per turn; multi-turn scans aggregate roughly 200–600 frames depending on the selected quality mode.

3.2 Calibration

Each scan begins with a short empty-turntable calibration. This provides two pieces of information. First, it supplies a background return that can be subtracted to suppress static clutter and part of the phone's self-interference signature. Second, it allows the system to estimate the effective turntable

Table 1: Prior-art boundary for smartphone acoustic reconstruction. "Phone-native acoustics" means sensing uses only built-in phone transducers; our system still uses a motorized turntable for controlled object motion, but no external acoustic sensor, array, or compute server.

System	Phone-native acoustics	3D volumetric output	Arbitrary objects	Notes
AIM [7]	Yes	No	Yes	2D acoustic image from linear phone motion.
Li *et al.* [8]	Yes	No	Yes	2D contour-like heat map from swept-phone measurements.
SONDAR [9]	Yes / commodity	No	Yes	ISAR-style 2D acoustic image used for shape/size measurement.
SonicHand [10]	Yes	No (3D joints only)	No	Reconstructs a fixed hand skeleton, not a general surface or volume.
Shih and Rowe [11]	No	Yes	No	Reconstructs room surfaces and uses a separate speaker module.
AirSAS + neural SAS [12, 13]	No	Yes	Yes	Laboratory synthetic-aperture sonar with specialized hardware.
Ours	Yes*	Yes	Yes	Coherent 3D reflectivity volume and point cloud; turntable only supplies controlled motion.

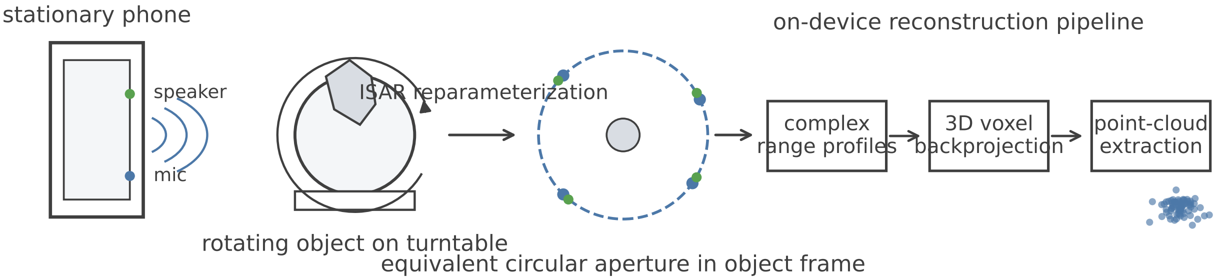

Figure 1: System overview. A stationary iPhone emits near-ultrasonic FMCW chirps while a motorized turntable rotates the target. In the target frame, the fixed phone becomes an equivalent circular synthetic aperture, producing angle-indexed complex range profiles that are coherently backprojected into a 3D voxel volume on-device.

center and standoff distance in the acoustic model. This step is essential because the reconstruction is highly sensitive to the assumed geometry.

3.3 Signal processing pipeline

The processing pipeline has three stages.

(1) Frame formation. Each chirp is matched to the transmitted waveform and converted to a complex baseband representation. The output is a complex range profile that preserves both magnitude and phase.

(2) Coherence maintenance. The system performs background subtraction, phase normalization, and consistency checks to keep the frame sequence coherent across hundreds of measurements. Without this step, the later backprojection would blur rather than focus.

(3) 3D reconstruction. Given the known turntable angle for each frame, the system synthesizes a sequence of object-centric acquisition poses and coherently backprojects the complex profiles into a 3D voxel grid. After accumulation, thresholding and local-maxima suppression convert the reflectivity volume into a renderable point cloud.

4 Acoustic ISAR Formulation

4.1 Transmit and receive model

Let the transmitted FMCW chirp be

$$s_{\mathrm{tx}}(t) = \exp\left(j2\pi \left(f_0 t + \tfrac{1}{2}\beta t^2\right)\right), \qquad 0 \le t \le T, \tag{1}$$

where f_0 is the start frequency, $\beta = B/T$ is the chirp rate, and B is the swept bandwidth. For a candidate voxel $\mathbf{v} \in \mathbb{R}^3$ and frame k, the bistatic path length is

$$d_k(\mathbf{v}) = \|\mathbf{v} - \mathbf{t}_k\| + \|\mathbf{v} - \mathbf{r}_k\|, \tag{2}$$

where $\mathbf{t}_k$ and $\mathbf{r}_k$ are the speaker and microphone positions expressed in the object frame. The corresponding delay is $\tau_k(\mathbf{v}) = d_k(\mathbf{v})/c$, with c the speed of sound in air.

Although the phone is physically stationary, the object rotates. If the turntable angle at frame k is θ_k about the vertical axis, then the equivalent object-frame acquisition poses are

$$\mathbf{t}_k = \mathbf{R}_z(-\theta_k)\,\mathbf{t}_0, \qquad \mathbf{r}_k = \mathbf{R}_z(-\theta_k)\,\mathbf{r}_0, \tag{3}$$

where $\mathbf{t}_0$ and $\mathbf{r}_0$ are the fixed phone transducer locations in a turntable-centered coordinate system. This is the key ISAR reparameterization: target rotation becomes a known synthetic aperture.

4.2 Coherent backprojection

After dechirping and matched filtering, frame k produces a complex range profile $b_k(\rho)$ indexed by path length ρ. The voxel intensity is formed by coherent accumulation:

$$V(\mathbf{v}) = \sum_{k=1}^{K} w_k\, b_k(d_k(\mathbf{v}))\exp\!\left(-j\,\frac{2\pi}{\lambda_c} d_k(\mathbf{v})\right), \tag{4}$$

where $\lambda_c = c/f_c$ is the center-wavelength and w_k is an optional confidence weight. The final volumetric reflectivity is $I(\mathbf{v}) = |V(\mathbf{v})|$. A point cloud is extracted by thresholding $I(\mathbf{v})$ and retaining spatially isolated maxima.

Equation (4) highlights why phase matters. If the assumed geometry is correct and the received frames remain coherent, contributions from a true scatterer add constructively at its voxel. If the geometry or phase is wrong, the same contributions cancel and the reconstruction defocuses.

4.3 Why ISAR is preferable to freehand phone SAR

In phone SAR, as in AIM [7], the user must move the phone along a prescribed path and the system must recover that path accurately enough for coherent focusing. That estimation problem is difficult because the same acoustic signal is being used both to infer the sensor trajectory and to image the target. By contrast, our turntable-based ISAR setup makes the acquisition path deterministic. The phone remains fixed, the object rotates at known angular increments, and the resulting synthetic aperture is induced mechanically rather than estimated from ambiguous echoes. This does not make the problem easy, but it moves the burden from freehand motion estimation to calibration and phase stability, which is a much better trade in the phone setting.

5 Implementation and Prototype Parameters

The current implementation targets an unmodified iPhone 15 Pro. Reconstruction is performed on-device using the phone GPU. The reconstruction lattice is a $60 \times 60 \times 60$ voxel grid with 5 mm spacing, covering a 0.30 m cube centered on the turntable. We emphasize that voxel spacing is a *sampling choice*, not a statement of physical resolution; the grid intentionally oversamples the expected acoustic point-spread function so that the focused peaks can be localized smoothly.

Table 2: Prototype operating point.

Parameter	Value
Device	iPhone 15 Pro
Acoustic hardware	built-in speaker + microphone
Operating band	near-ultrasonic
Frame rate	$\sim$13 Hz
Turntable period	$\sim$15 s/rev
Standoff distance	$\sim$0.30 m
Frames per scan	$\sim$200–600
Reconstruction volume	$0.30\,\mathrm{m}^3$
Voxel spacing	5 mm
External hardware	motorized turntable only
Processing location	entirely on-device

The dominant computation is the volumetric backprojection, whose complexity scales as $O(N_{\mathrm{vox}}N_{\mathrm{frames}})$. A phone CPU implementation is too slow for interactive use, so the system uses a GPU kernel that evaluates Eq. (4) in parallel across the voxel grid. In the current prototype, backprojection is the largest contributor to runtime.

6 Prototype Characterization

Because this first version focuses on feasibility, we characterize the prototype through its operating parameters, theoretical resolution limits, and runtime rather than a large benchmark suite.

6.1 Resolution

The bandwidth-limited range resolution is the familiar FMCW quantity

$$\Delta r \approx \frac{c}{2B}. \tag{5}$$

With $B \approx 4\,\mathrm{kHz}$ and $c \approx 343\,\mathrm{m/s}$, the theoretical range resolution is approximately 4.3 cm. This is coarse compared with optical depth sensing and is the main reason the reconstruction should be interpreted as a reflectivity point cloud rather than a watertight metric mesh.

The lateral focusing behavior is better because it benefits from coherent integration across the synthetic aperture. At a center frequency near 20 kHz, the acoustic wavelength is

Table 3: Prototype characterization.

Quantity	Value	Dominant factor
Range resolution	∼4.3 cm	4 kHz acoustic bandwidth
Best-case lateral scale	∼8.6 mm	$\lambda/2$ at 20 kHz
Voxel spacing	5 mm	grid sampling only
On-device processing time	<30 s	signal processing + GPU backprojection

roughly 17 mm. For well-conditioned scatterers observed over nearly the full rotation, the best-case tangential localization is on the order of $\lambda/2$, i.e., roughly 8.6 mm. In practice, however, the resolution is anisotropic: a single-height circular aperture has a missing-cone effect, so vertical or axis-aligned structure is less well conditioned than tangential structure. We therefore report a best-case lateral scale rather than claiming uniform isotropic 9 mm resolution in all directions.

6.2 Runtime

For the highest-quality mode with several hundred frames, the signal-processing and reconstruction stages complete in under 30 s on an iPhone 15 Pro. This figure excludes the optional additional scan time incurred when collecting multiple turntable revolutions. A single-revolution preview mode is correspondingly faster but provides fewer aspect samples and therefore lower robustness.

6.3 Interpretation of the output

The output of the system is a 3D *reflectivity* volume and a derived point cloud, not a watertight surface mesh. Strong reflectors, edges, and material discontinuities dominate the reconstruction. Weakly reflecting regions or surfaces oriented unfavorably with respect to the phone may be underrepresented. This is an expected property of coherent acoustic imaging and not a software artifact.

7 Challenges and Limitations

Several challenges distinguish coherent 3D acoustic imaging on a phone from both prior phone sensing and laboratory SAS.

Self-interference. The phone's speaker and microphone are separated by only about 16 mm, so the direct path is much stronger than echoes from an object at 30 cm. This leakage can exceed target returns by tens of decibels and must be suppressed without erasing weak reflections.

Phase coherence. Coherent backprojection is unforgiving. At 20 kHz, the wavelength is only about 17 mm, so even small phase errors accumulate into visible blur. Clock drift, thermal variation, speaker nonlinearity, and changing system latency all threaten coherence across a long scan.

Geometry sensitivity. The ISAR pose model depends on the assumed turntable center and standoff. Errors of only a few centimeters can shift predicted path lengths enough to defocus the reconstruction globally.

Bandwidth limits. Commodity phone speakers typically provide usable energy only over a narrow high-frequency band. That limits raw range resolution to a few centimeters even with perfect reconstruction.

Aperture anisotropy. A single-height circular aperture does not sample 3D spatial frequencies uniformly. The resulting missing-cone effect means some directions are inherently less well constrained than others.

Controlled-setting requirement. The present prototype requires the target to rotate on a turntable. That makes acquisition repeatable and coherent, but it also restricts the system to controlled tabletop scenarios.

8 Discussion

The comparison with prior work suggests a useful way to think about the contribution. This system is not a replacement for optical 3D scanning, nor is it a claim that phone acoustics can suddenly deliver dense metric geometry comparable to LiDAR. Instead, it establishes that coherent 3D *volumetric* acoustic reconstruction is reachable on a commodity phone when the acquisition geometry is chosen carefully and the entire pipeline preserves phase.

That distinction matters. Earlier phone acoustic systems showed that commodity hardware can track points, classify gestures, and form 2D images [1, 2, 7, 5]. Laboratory acoustic systems showed that coherent 3D reconstruction is possible with synthetic apertures [12, 13, 14]. The present design sits between those two literatures. Its main conceptual contribution is not merely "using a turntable," but recognizing that object-centric ISAR is a natural fit for the phone setting because it removes the freehand trajectory-estimation bottleneck.

Several extensions are worth pursuing. First, PowerPhone [15] suggests that some phones can be software-reconfigured for much higher sampling rates, which could materially improve acoustic resolution if integrated into a coherent imaging pipeline. Second, a richer treatment of the phone's leakage path and phase drift could improve robustness and permit longer acquisitions. Third, one could combine the physics-based backprojection described here with learned volumetric priors, similar in spirit to recent neural SAS methods [13], to regularize the severe bandwidth limitations of phone audio.

9 Conclusion

We presented a coherent 3D acoustic imaging system for a commodity smartphone. The system uses near-ultrasonic FMCW chirps, a motorized turntable, and an object-centric ISAR formulation to synthesize a circular aperture while keeping the phone stationary. Reconstruction runs entirely on-device and produces a 3D reflectivity volume and point cloud. Just as importantly, we clarified the prior-art boundary: the closest published smartphone acoustic systems either stop at 2D imaging, return only sparse tracked points, or reconstruct constrained structures such as hand joints. This prototype therefore marks a concrete step toward phone-native volumetric acoustic imaging of arbitrary objects.

References

[1] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota. FingerIO: Using active sonar for fine-grained finger tracking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016. DOI: 10.1145/2858036.2858580.

[2] 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 (MobiCom)*, 2016. DOI: 10.1145/2973750.2973764.

[3] A. Wang and S. Gollakota. MilliSonic: Pushing the limits of acoustic motion tracking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019. DOI: 10.1145/3290605.3300248.

[4] B. Zhou, M. Elbadry, R. Gao, and F. Ye. BatMapper: Acoustic sensing based indoor floor plan construction using smartphones. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2017, pp. 42–55. DOI: 10.1145/3081333.3081363.

[5] S. Pradhan, G. Baig, W. Mao, L. Qiu, G. Chen, and B. Yang. Smartphone-based acoustic indoor space mapping. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):75, 2018. DOI: 10.1145/3214278.

[6] Y. Zhuang, Y. Wang, Y. Yan, X. Xu, and Y. Shi. ReflecTrack: Enabling 3D acoustic position tracking using commodity dual-microphone smartphones. In *Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2021. DOI: 10.1145/3472749.3474805.

[7] W. Mao, M. Wang, and L. Qiu. AIM: Acoustic imaging on a mobile. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2018. DOI: 10.1145/3210240.3210325.

[8] C. Li, J. Wang, X. Ding, and N. Zhang. Acoustic imaging using the built-in sensors of a smartphone. *Symmetry*, 13(6):1065, 2021. DOI: 10.3390/sym13061065.

[9] X. Liang, Z. Wei, D. Li, J. Xiong, and J. Gummeson. SONDAR: Size and shape measurements using acoustic imaging. In *Proceedings of the Twenty-Fifth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc)*, 2024, pp. 361–370. DOI: 10.1145/3641512.3686359.

[10] S. Wang, X. Wang, W. Jiang, C. Miao, Q. Cao, H. Wang, K. Sun, H. Xue, and L. Su. Towards smartphone-based 3D hand pose reconstruction using acoustic signals. *ACM Transactions on Sensor Networks*, 20(5):106, 2024. DOI: 10.1145/3677122.

[11] O. Shih and A. Rowe. Can a phone hear the shape of a room? In *Proceedings of the 18th International Conference on Information Processing in Sensor Networks (IPSN)*, 2019. DOI: 10.1145/3302506.3310407.

[12] T. E. Blanford, D. P. Williams, J. D. Park, B. T. Reinhardt, K. S. Dalton, S. F. Johnson, and D. C. Brown. An in-air synthetic aperture sonar dataset of target scattering in environments of varying complexity. *Scientific Data*, 11:1196, 2024. DOI: 10.1038/s41597-024-04050-0.

[13] A. W. Reed, J. Kim, T. Blanford, A. Pediredla, D. C. Brown, and S. Jayasuriya. Neural volumetric reconstruction for coherent synthetic aperture sonar. *ACM Transactions on Graphics*, 42(4):113, 2023. DOI: 10.1145/3592141.

[14] P. Serafin, M. Okoń-Fąfara, M. Szugajew, C. Leśnik, and A. Kawalec. 3-D inverse synthetic aperture sonar imaging. In *2017 18th International Radar Symposium (IRS)*, 2017. DOI: 10.23919/IRS.2017.8008209.

[15] S. Cao, D. Li, S. I. Lee, and J. Xiong. PowerPhone: Unleashing the acoustic sensing capability of smartphones. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking (MobiCom)*, 2023. DOI: 10.1145/3570361.3613270.