3D imaging from multipath temporal echoes

Alex Turpin[1,†,*], Valentin Kapitany[2,†], Jack Radford[2], Davide Rovelli[2], Kevin Mitchell[2], Ashley Lyons[2], Ilya Starshynov[2], Daniele Faccio[2*]
[1] *School of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK*
[2] *School of Physics & Astronomy, University of Glasgow, Glasgow G12 8QQ, UK*
(Dated: March 15, 2021)

Echo-location is a broad approach to imaging and sensing that includes both man-made RADAR, LIDAR, SONAR and also animal navigation. However, full 3D information based on echo-location requires some form of scanning of the scene in order to provide the spatial location of the echo origin-points. Without this spatial information, imaging objects in 3D is a very challenging task as the inverse retrieval problem is strongly ill-posed. Here, we show that the temporal information encoded in the return echoes that are reflected multiple times within a scene is sufficient to faithfully render an image in 3D. Numerical modelling and an information theoretic perspective prove the concept and provide insight into the role of the multipath information. We experimentally demonstrate the concept by using both radio-frequency and acoustic waves for imaging individuals moving in a closed environment.

Introduction. In nature, detecting and locating objects from reflected echoes is generally possible only if two or more detectors are used. Animals such as bats or dolphins [1] and even humans [2] can emit pulses of sound to sense the environment they navigate through and identify objects. RADAR and LiDAR imaging systems operate in a similar way, albeit with electromagnetic (EM) radiation (radio waves and light, respectively): a series of EM pulses are used to scan and probe the scene and, by measuring the arrival time of the return echoes and correlating this with the direction from which they are detected, they can form a 3D estimate of the scene [3, 4]. This principle also holds for non-line-of-sight (NLOS) applications [5–9], where photon echoes of light, now scattered from multiple surfaces along indirect paths, are analysed with the goal of revealing the 3D shape and visual appearance of objects outside the direct line of sight. Although NLOS is typically deployed with optical sources, it has also been demonstrated with acoustic [10] and radio-frequency (RF) sources [11].

Locating objects in space and forming an image in 3D from their wave echoes using a single point detector without any form of scanning is, computationally-speaking, a strongly ill-posed problem and therefore considerably more challenging. However, recent work has shown that echoes contain a very rich structure in the time dimension that can be used to extract meaningful information about the scene [12–14]. In these cases, further assumptions of the scene are required in order eliminate ambiguities arising from the fact that the echo is single-path, i.e. the outgoing signal reflects only once from the scene objects. This leads to ambiguity in the form of an equal-distribution-probability for the echo origin point that is spread over a spherical dome centred on the detector and with a radius determined by the echo arrival time. The additional assumptions referred to above can be introduced, for example, in the form of additional information by means of a machine learning algorithm that exploits the knowledge of static objects in the scene background and a statistical knowledge of the objects that we want to image [12, 14].

The paradigm investigated here is the extension of echo detection to multipath trajectories of the return signal. The idea of using multipath reflections for sensing inside buildings, through walls or out of view, especially with RF waves, has been a topic of extensive study during the last decade [15–22]. However, these simple geometric approaches are typically limited to locating the position of objects (and not imaging), e.g. of humans inside known environments. Multipath sensing has also been combined with Bayesian inference [23] and convolutional neural networks [24] to localise sonic sources. In the optical domain, multipath interference, i.e. the contribution from light following multiple paths onto the same pixel, is generally considered problematic and has to be accounted for to acquire accurate depth maps [25–28]. However, recent works have explored multipath optical sensing both theoretically [29] and experimentally [30] by exploiting deterministic algorithms that provide mathematical proof for the ability to reconstruct the geometry of simple scenes from a single location.

In this work, we provide empirical evidence that 3D scenes can be reconstructed from temporal echoes alone. We make use of a data-driven approach that exploits multipath temporal echoes, i.e. echoes from waves that are reflected multiple times from surfaces and objects within a scene, to unambiguously reconstruct a meaningful 3D image in a fixed scenario. We first present numerical simulations that show how a simple artificial neural network can be trained to reconstruct a 3D scene. We then underline the importance of the multipath echoes, with a dominant role played by the first few reflections and a gradually decreasing importance of further bounces. These findings are supported by an information theoretic analysis applied to the raw multipath data that is independent of the image retrieval algorithm. We then

* Corresponding author: alex.turpin@glasgow.ac.uk / daniele.faccio@glasgow.uk

demonstrate our approach experimentally. Although our method could be in principle implemented with optical pulses, light suffers from severe diffused reflection, which would make it very hard to detect any optical signal after 2 reflections. We therefore concentrate on GHz EM radio-frequency (RF) and kHz acoustic waves, as these can be reflected multiple times by walls and objects. In both cases, we are able to precisely retrieve 3D images of a dynamic scene with a significant improvement beyond what is achievable using single-path echoes.

3D imaging with multipath temporal echoes. Our approach is conceptually sketched in Fig. 1. A source emits waves in the form of pulses that diverge with a wide angle so as to flash-illuminate the whole scene. The emitted pulses are then reflected by the room walls and the objects inside it and, finally, are detected by a single-pixel sensor with time-resolving capabilities. The timing of successive pulses is arranged so as to not temporally overlap with any returning echoes, i.e. each outgoing pulse and detection of return echoes are completely separate events from the emission of a successive pulse. The sensor collects and records the received energy over a wide angle and provides this information in the form of a temporal histogram. The process of pulsed waves bouncing multiple times inside the room is fully deterministic: with a complete knowledge of the distribution of objects within the room, the room dimensions, and their reflectivity, it is straightforward to predict the recorded temporal histogram. However, solving the inverse process, namely the reconstruction of the scene (including room and objects) in 3D dimensions from just the temporal histogram, is ill-posed: echoes arriving to the detector at time t_d are compatible with objects placed not just at a single point (as would be desired), but rather with the whole surface of a spherical dome represented by the equation $(ct_d)^2/2 = x^2 + y^2 + z^2$ (where c is the speed of the pulse). This ambiguity has been previously solved, although only in part, by utilising the fact that a moving object will obscure static background objects, therefore removing them from return echo patterns [14].

In contrast, in this work we highlight the strength of including multipath reflections in the data-driven solution to solve the ambiguity issue: using not only the first reflection but 2, 3, and more reflections breaks the degeneracy and helps the algorithm to reconstruct the position and shape of the object in 3D with high accuracy, thus making background objects not essential.

Numerical simulations. We first show numerical simulations based on Monte Carlo ray-tracing (see [31] for full details). Our scene consists of a closed room with walls, floor and ceiling that all have the same 100% reflectivity [Fig. 1(a)]. Inside this room, a rectangular cuboid is placed in different positions and the scene is imaged in 3D with a ToF camera providing a 2D depth map, see Fig. 1(c). We consider that an emitter emits probe pulses in all directions within azimuth and elevation angles θ and ϕ, both within $[-67.5°, 67.5°]$. The return echo amplitudes, i.e. the number of returning rays per time

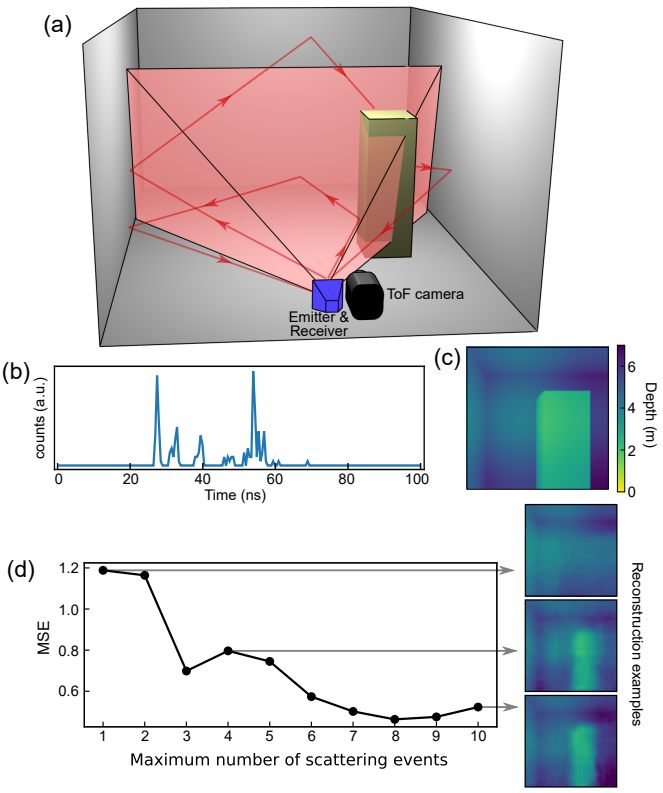

FIG. 1. (a) 3D visualisation of our physical system: a rectangular cuboid (yellow) moves within a room. Rays are emitted within a pyramid-like volume and illuminate the scene. Red arrows indicate examples of multipath reflections, which eventually reach the detector (blue) that records their arrival time. (b) An example of a recorded time histogram. (c) Color-depth encoded 3D view of the scene. (d) Mean mean-squared-error (MSE) with increasing multipath contributions, calculated between the ground truth 3D scene and the neural network reconstruction, averaged over 100 3D images. Insets show depth image reconstruction examples obtained for 1-path, 4-path, and 10-path events.

[Fig. 1(b)], are recorded in time at the detector that is co-located with the emitter. Each scene is sampled with 10000 rays per object position, for 2000 objects positions. This provides a data set of temporal trace-3D image pairs that we use to train a convolutional deep neural network, shaped such as to force information through a bottleneck (see [31] for details) to extract features from data. We then test the neural network with histograms that were never used during training and render an estimate of the scene in 3D. We repeat this analysis for an increasing number of path events, starting from single-path until 10-path events, and we analyse the quality of the reconstructions in terms of the mean-square error (MSE) between the ground truth and the retrieved images (see [31–34] for further details). To avoid specificity of the training by the deep neural network architecture, we re-train the network 10 times for each path event, such that for every training round we leave the starting weights of the neural

network random. This procedure guarantees a slightly different image reconstruction every time the algorithm is trained. Then, we average our reconstruction-quality metrics over these 10 networks. Our results, summarised in Fig. 1(d), show that the MSE decreases as the number of multipath events is increased. In particular, we see that the first 2-4 multipath echoes are the most important and significantly improve scene reconstruction. This can be seen clearly not only in the MSE but also in the insets to Fig. 1(d) that show examples of a reconstruction for 1, 4 and 10 path events. We clearly see that whilst for single-path it is hard to distinguish the object position due to blurring arising from the above mentioned ambiguities, multipath information cures this problem and allows to clearly resolve the 3D scene (see [31] for further examples). We quantify the gain in information when including an increasing number of paths using the concepts of Shannon entropy, mutual information and joint entropy as derived in Information Theory [35–37]. The Shannon entropy gives the expectation value of uncertainty reduction when observing a variable X at values x_i, which occur with probability $p(x_i)$:

$$H(X) = -\sum_{i=1}^{N} p(x_i) \log_2 p(x_i). \quad (1)$$

More specifically, we take a set of 2000 examples of individual temporal histograms from the numerical model described above, within which we identify histogram shapes x_i that occur with probability $p(x_i)$. We can then calculate the joint entropy $H(X,Y)$ for single-path histograms X and 2-path histograms Y:

$$H(X,Y) = -\sum_{j=1}^{M}\sum_{i=1}^{N} p(x_i,y_j) \log_2 p(x_i,y_j). \quad (2)$$

This can be extended to calculate the joint entropy for data containing $< n$ bounces and $< (n+1)$ bounces. The mutual information, $MI(X;Y)$, then describes the information shared by the two random variables due to correlations within the data:

$$MI(X;Y) = H(X) + H(Y) - H(X,Y). \quad (3)$$

We rearrange Eq. (3) to find the additional *uncorrelated* information, UI, in the multipath data Y, i.e. the mutual information $MI(X;Y)$ subtracted from the total information, $H(Y)$. In other words, the additional information that is gained by including photons from a second or multiple reflections/paths is given by $UI(X;Y) = H(X,Y) - H(X)$.

Figure 2(a) shows $UI(X-1;X)$ in log scale for increasing number of reflections/paths. As can be seen, significant additional (uncorrelated) information is gained from the 2nd and 3rd reflections but becomes negligible after 4 reflections. Remarkably, in this configuration UI for a 2-path signal is larger than the information contained in the direct 1-path (standard LIDAR, single reflection) signal. An intuitive insight into understanding this gain in information from multipath data is shown in Fig. 2(b): the 3D dimensional rendering of a scene, as would be observed by a camera placed at the detection point, appears very similar to what would be observed in a room of mirrors. The first reflection (in black) provides only direct line-of-sight information of the object; the first 4 reflections (in red) show different effective viewpoints (side-view and back-view) that would otherwise be inaccessible and therefore increase the information; all successive reflections (in light blue) are replicas of the first 4 reflections and do not contain additional useful information. That said, we underline that in real life scenarios, the noisy-channel coding theorem [35] indicates that adding redundant replicas of information in the form of higher order paths, could still lead to preservation of information that is lost due e.g. to measurement noise.

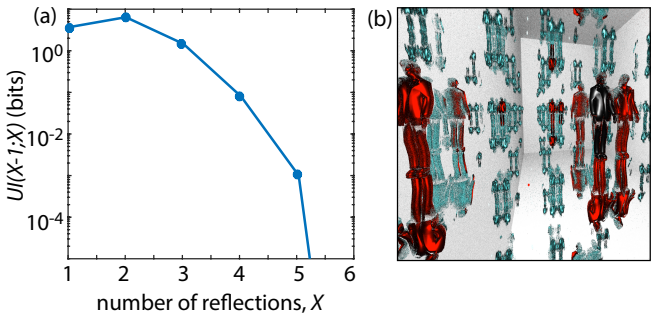

FIG. 2. (a) The gain in information when including photons in the temporal data which have experienced an increasing number of reflections within the scene. (b) A simulation of a multipath scene as would be viewed by a camera. The various reflections show different viewpoints of the mannequin therefore intuitively explaining why multipath echoes contain additional information but also why beyond the 4^{th} bounce, there is little or no gain of information (see text for details).

Experiments. We show the validity of our approach with experiments using two different sources of waves, namely GHz radio-frequency (RF) and kHz-frequency acoustic pulses. The experimental set-up in both cases is identical to Fig. 1(a), where the emitter/detector is an RF-antenna or a speaker and microphone, for the RF and acoustic experiments, respectively.

For the experiments with RF waves, we use a transceiver module (TI-AWR1642), which operates in the frequency modulated continuous wave regime [38], with a range resolution and maximal unambiguous range of 4.4 cm and 9 m respectively. The transceiver probes the scene with an angular aperture of 20° in the vertical plane and 180° in the horizontal plane (-3 dB FWHM). An analog-to-digital converter samples the signal with 120 ns temporal resolution and 133 Hz rate.

The experiments were conducted with a human individual walking around in a room with approximate dimensions of $3 \times 4 \times 2.5\,\mathrm{m}^3$. The echo recordings from the RF antenna are acquired in parallel to 3D (ground truth) im-

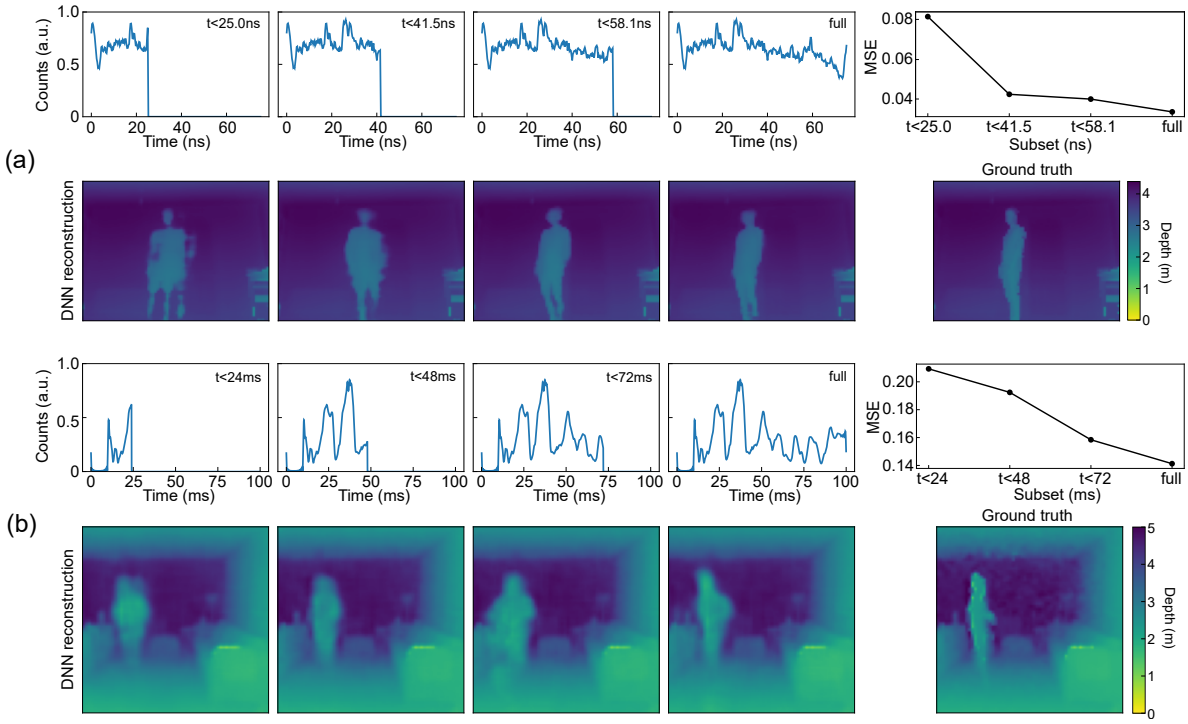

FIG. 3. (a) RF and (b) acoustic results. The top rows of (a) and (b) show the time histograms that are truncated at increasing times, therefore including an increasing number of multipath echoes. The last plot of first rows show the quality of the image reconstructions in terms of mean-squared error (MSE) compared to the ground truth for a set of 100 scenes, for increasing multipath events. The second row in (a) and (b) shows the corresponding images retrieved with the deep neural network, and the ToF camera ground truth image.

ages via a ToF camera (Basler), which provides 80×60 pixel color-encoded depth images.

For the acoustic measurements, we replace the RF antenna with a PC speaker (Logitech Z333 system, consisting of 2 speakers + 1 subwoofer) and a PC microphone (integrated in a Logitech C270 webcam). The speakers emit a pulsed wave with centre frequency of 5 kHz ($\lambda \approx 6.7$ cm) and a bandwidth of 1 kHz, with duration of 50 ms and repetition rate ≈ 10 Hz. The microphone, co-located with the speakers, records the returning echoes for 100 ms at a sampling rate of 192 kHz. The data is Fourier filtered so as to select only signals at (5 ± 0.5) kHz. The ToF 3D camera used to train the deep learning algorithm was an Intel Realsense D435 capturing 64×64 color-encoded depth images. The room used for this experiment had dimensions $7 \times 6 \times 2.5\,\mathrm{m}^3$). Note that the recording time window in both cases, respectively of 80 ns and 100 ms for the RF and acoustic experiments, is long enough to ensure that the waves can reach the furthest corner of the rooms and return to the detector.

For both the RF and acoustic measurements, we use the pairs of ground truth ToF images and RF (or acoustic) echo temporal traces to train a deep learning algorithm based on convolutional layers followed by a Rectified Linear Unit activation function (see [31] for details). We use 9000 and 5000 pairs of data for training the neural networks for RF and acoustic data respectively, after which, full 3D images can be retrieved from a single (previously unseen) RF (or acoustic) temporal trace.

Figures 3(a) and (b) show the results for the RF and acoustic cases, respectively (see also [31, 39] for videos). To explore the role of multipath events, we trained and tested our neural network with successively increased temporal extension of the time histograms: truncation of the data at short times corresponds to single path data, calculated as the ToF to the farthest wall in the room. We increase the truncation time (indicated in the figures) by evaluating the longest ToF value for 2-path and 3-path events in the room so as to include 2 and 3 bounces, thus gradually increasing the information from higher order path contributions. The retrieved 3D scenes [second row in Figs. 3(a) and (b)] show that networks trained solely on 1-path events [first column of Figs. 3(a) and (b)] struggle to provide a sharp 3D image as there are many possible scenes that correspond to the same single-path time histogram. Increasing the number of multipath events provides an increasingly improved reconstruction. This improvement can be quantified by calculating the MSE between the retrieved image and the ground truth, averaged over 865 and 500 different measurements, for RF and acoustics respectively. The MSE in Figs. 3(a) and (b) (far-right graph), decreases monotonically with increasing multipath contributions, in good agreement with our modelling and experimentally shows the significant

3D imaging capability achieved with multipath temporal echoes. Note that our technique can exploit training on a single individual to operate successfully on different individuals, recovering general shape and position, see [31]. In this work we focused only on imaging human individuals. Evidence from other work suggests that training with additional objects and geometrical shapes should also be possible [14] and generic imaging functionality has been shown in a different but related multipath setting [40].

Conclusions. In summary, we have shown that multipath temporal echoes and deep learning can be used to provide full 3D images of a scene. Applications of these ideas might be found in imaging in closed environments so as to enable efficient generation of multipath echoes, for example with healthcare applications for homes and hospitals of the future. Interesting developments might include the generalisation to dynamic background scenarios, to open-air scenes, and to scenes incorporating information from different viewpoints, thus opening applications in NLOS imaging and 3D mapping of complex object geometries. More in general, multipath echo imaging offers interesting opportunities, considering that RF antennas can also be extremely compact (and are currently present in cell phones) and that the acoustic results were obtained with standard computer speakers and microphones, thus effectively transforming everyday household items into full 3D imaging systems.

Acknowledgements. We thank Hanoz Bhamgara, Mark Jarvis, and Marton Szafian for technical support with the RF system. D.F. acknowledges financial support from the Royal Academy of Engineering and from EPSRC (UK, grant no. EP/T00097X/1). V.K. acknowledges funding from Horiba. A.T. acknowledges support as an LKAS Fellow. Data relevant to this work are available for download at Ref. [41].

† A.T. and V.K. contributed equally to this work.

[1] J. A. Thomas, C. F. Moss, and M. Vater, *Echolocation in bats and dolphins* (University of Chicago Press, 2004).
[2] L. Thaler, S. R. Arnott, and M. A. Goodale, PLoS one **6**, e20162 (2011).
[3] P. Dong and Q. Chen, *LiDAR Remote Sensing and Applications* (CRC Press, 2017).
[4] M. I. Skolnik *et al.*, *Introduction to radar systems*, Vol. 3 (McGraw-hill New York, 1980).
[5] D. Faccio, A. Velten, and G. Wetzstein, Nat. Rev. Phys. **2**, 318–327 (2020).
[6] M. O'Toole, D. B. Lindell, and G. Wetzstein, Nature **555**, 338 (2018).
[7] X. Liu, I. Guillén, M. La Manna, J. H. Nam, S. A. Reza, T. H. Le, A. Jarabo, D. Gutierrez, and A. Velten, Nature **572**, 620 (2019).
[8] C. A. Metzler, F. Heide, P. Rangarajan, M. M. Balaji, A. Viswanath, A. Veeraraghavan, and R. G. Baraniuk, Optica **7**, 63 (2020).
[9] S. Chan, R. E. Warburton, G. Gariepy, J. Leach, and D. Faccio, Opt. Express **25**, 10109 (2017).
[10] D. B. Lindell, G. Wetzstein, and V. Koltun (2019).
[11] N. Scheiner, F. Kraus, F. Wei, B. Phan, F. Mannan, N. Appenrodt, W. Ritter, J. Dickmann, K. Dietmayer, B. Sick, *et al.*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020) pp. 2068–2077.
[12] P. Caramazza, A. Boccolini, D. Buschek, M. Hullin, C. F. Higham, R. Henderson, R. Murray-Smith, and D. Faccio, Scientific Reports **8**, 11945 (2018).
[13] C. A. Metzler, D. B. Lindell, and G. Wetzstein, arXiv preprint **arXiv:1912.06727** (2019).
[14] A. Turpin, G. Musarra, V. Kapitany, F. Tonolini, A. Lyons, I. Starshynov, F. Villa, E. Conca, F. Fioranelli, R. Murray-Smith, and D. Faccio, Optica **7**, 900 (2020).
[15] J. L. Krolik, J. Farrell, and A. Steinhardt, in *2006 IEEE Conference on Radar* (IEEE, 2006) pp. 4–pp.
[16] P. Setlur, M. Amin, and F. Ahmad, IEEE Transactions on Geoscience and Remote Sensing **49**, 4021 (2011).
[17] S. Sen and A. Nehorai, IEEE Transactions on Signal Processing **59**, 78 (2010).
[18] M. Leigsnering, F. Ahmad, M. Amin, and A. Zoubir, IEEE Transactions on Aerospace and Electronic Systems **50**, 920 (2014).
[19] A. H. Muqaibel, M. G. Amin, and F. Ahmad, International Journal of Antennas and Propagation **2015**, 510720 (2015).
[20] F. Fuschini, S. Häfner, M. Zoli, R. Müller, E. Vitucci, D. Dupleich, M. Barbiroli, J. Luo, E. Schulz, V. Degli-Esposti, *et al.*, Journal of Infrared, Millimeter, and Terahertz Waves **38**, 727 (2017).
[21] A. A. Goulianos, A. L. Freire, T. Barratt, E. Mellios, P. Cain, M. Rumney, A. Nix, and M. Beach, in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)* (IEEE, 2017) pp. 1–5.
[22] W. G. Neubauer, *Acoustic reflection from surfaces and shapes* (Naval Research Laboratory, 1986).
[23] J. H. Lim and D. W. Chof, Robotica **14**, 527 (1996).
[24] E. L. Ferguson, S. B. Williams, and C. T. Jin, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2018) pp. 2386–2390.
[25] A. Bhandari, A. Kadambi, R. Whyte, C. Barsi, M. Feigin, A. Dorrington, and R. Raskar, Optics letters **39**, 1705 (2014).
[26] D. Freedman, Y. Smolin, E. Krupka, I. Leichter, and M. Schmidt, in *European Conference on Computer Vision* (Springer, 2014) pp. 234–249.
[27] D. Shin, F. Xu, F. N. Wong, J. H. Shapiro, and V. K. Goyal, Optics express **24**, 1873 (2016).
[28] J. Marco, Q. Hernandez, A. Munoz, Y. Dong, A. Jarabo, M. H. Kim, X. Tong, and D. Gutierrez, ACM Transactions on Graphics (ToG) **36**, 1 (2017).
[29] I. Gkioulekas, S. J. Gortler, L. Theran, and T. Zickler, arXiv preprint arXiv:1709.03936 (2017).
[30] J. H. Nam and A. Velten, Applied Sciences **10**, 6458 (2020).
[31] The Supplementary material provides: (i) details of a simple analytical model showing that multipath temporal echoes can be used to determine the location of an object in 2D; (ii) a more in depth overview of the phys-

ical model we used for our numerical simulations; (iii) details of the metrics used to estimate the performance of the approach; (iv) a description of the neural network-based image retrieval algorithm (including its operation when trained and tested on different individuals); and (v) full details of the information theory analysis we conduct. Supplementary Video 1 shows how our Monte Carlo-based physical model works. Supplementary Video 2 shows the quality of the reconstruction of our simulations for different path events. Finally, Supplementary Videos 3 and 4 show, respectively, the implementation of our approach with RF and acoustic echoes.

[32] D. P. Kingma and J. Ba, arXiv preprint arXiv:1412.6980 (2014).

[33] M. Abadi *et al*, "TensorFlow: Large-scale machine learning on heterogeneous systems," (2015).

[34] F. Chollet, "keras," `https://github.com/fchollet/keras` (2015).

[35] C. E. Shannon, The Bell system technical journal **27**, 379 (1948).

[36] D. J. C. MacKay, *Information theory, inference, and learning algorithms* (Cambridge University Press, Cambridge, 2003).

[37] T. M. Cover and J. A. Thomas, *Elements of information theory* (Wiley-Interscience, 2006).

[38] T. Instruments, *AWR1642 Single-Chip 77- and 79-GHz FMCW Radar sensor datasheet (Rev. B)* (2020 (accessed 12 October 2020)).

[39] "Supplementary Videos 1, 2, 3, and 4 can be found, respectively, in the following links:," `https://www.youtube.com/watch?v=cjPkSo9kZuI&ab_channel=ExtremeLightGroupUniversityofGlasgow` `https://www.youtube.com/watch?v=e4Ywbveb6sw&ab_channel=ExtremeLightGroupUniversityofGlasgow` `https://www.youtube.com/watch?v=mD-Qqu6aCPw&ab_channel=ExtremeLightGroupUniversityofGlasgow` `https://www.youtube.com/watch?v=Ws_4WIkWuAw&ab_channel=ExtremeLightGroupUniversityofGlasgow`.

[40] W. Chen, F. Wei, K. N. Kutulakos, S. Rusinkiewicz, and F. Heide, ACM Transactions on Graphics (TOG) **39**, 1 (2020).

[41] `http://researchdata.gla.ac.uk/1102/`.