LEARNED ACOUSTIC RECONSTRUCTION USING SYNTHETIC APERTURE FOCUSING

Tim Straubinger, Robert Xiao, Helge Rhodin

University of British Columbia, Department of Computer Science

ABSTRACT

Many algorithmic approaches to 3D acoustic imaging have been devised which rely on a large abundance of receiving elements to produce images with delay-and-sum techniques, but these have found little use in air due to hardware complexity and low accuracy. Recent learning-based approaches to one-shot in-air acoustic reconstruction attempt to overcome these limitations using simple hardware and large datasets of geometry and echo pairs to train neural networks. However, existing learned models use spatially-dense representations and attempt to predict entire scenes at once, requiring an abundance of data to truly generalize.

We train an implicit neural network with no spatial awareness to predict the distance to the nearest obstacle at a single location from only time-delayed echoes. Using acoustic wave simulation, we show that our method yields better generalization and behaves more intuitively than competing methods while requiring only a fraction of the training data. Our code and data will be available at `https://timstr.github.io/learned-acoustic-reconstruction/`.

Index Terms— Acoustic imaging, deep learning

1. INTRODUCTION

The physical world around us can convey enormous amounts of information by the echoes it produces in response to emitted sounds. This is made clear by the many animals, such as bats, cetaceans, and even some humans [1] that are able to navigate by producing sound and listening to returning echoes. Thus, there is potential to be able to infer spatial information from acoustic echoes using technology for localization and geometric reconstruction.

Compared to vision, sound provides a comparable but distinct modality for sensing the environment. Notably, sound reacts primarily with changes in density, and travels slowly enough for its travel time to be easily measured, enabling accurate distance estimates with simple techniques. Furthermore, the wavelengths of sound are much larger than those of light, which directly limits the resolution that can be achieved. While this is a shortcoming, it can also be a benefit in human environments where privacy is a concern - for example, an acoustic sensing device may reveal the presence and pose but not the identity of a person nearby.

In medicine, acoustic imaging using ultrasound is commonly performed to study developing fetuses and internal organs. Using simple delay-and-sum techniques in software rather than the ray-based scanning method of typical B-mode systems, synthetic aperture ultrasound achieves better quality acoustic images at faster frame rates with the same hardware, typically using 64 to 256 acoustic elements [2, 3, 4, 5].

In underwater 3D acoustic imaging, the faster frame rates of synthetic aperture methods enables interactive exploration of large spaces with their longer echo times. Typical systems rely on a single acoustic emission and software processing with delay-and-sum techniques to disentangle the echoes from individual reflectors [6, 7, 8]. Hundreds of receivers are commonly used, which improves clarity but introduces additional challenges, such as finding efficient algorithms.

In both medical and underwater acoustic imaging, these classical approaches benefit from large amounts of data which improve imaging quality. Given the large number of receivers already in use, one would expect these methods to perform poorly when using a minimal number of receivers. Instead, one would need algorithms which are better able to reason about reflectors with less information. To the authors' knowledge, 3D reconstruction in air using similar techniques has only been performed on static scenes with large numbers of receivers and mechanical scanning [9, 10].

2. BACKGROUND

Machine learning has been investigated for in-air acoustic reconstruction, as it side-steps explicitly modeling wave dynamics and allows one to specialize in common scenarios using minimal hardware and single acoustic emissions [11, 12, 13, 14]. If designed correctly, a learning-based approach can efficiently extract information from echoes and predict the most likely geometry according to a trained model. But the use of learning brings additional challenges as well, such as the need for training data and the possibility of over-fitting.

In Bat-G Net, a large dataset of echoes is gathered by placing simple geometric solids along a horizontal plane and rotating them at fixed intervals [11] while recording the echoes from an ultrasonic FM sweep. This allows volumetric training labels to be generated procedurally in software but limits their spatial coverage. In BatVision, a different approach is taken using a depth camera to capture the 3D shape of diverse

indoor environments while recording their echoes in response to an FM sweep [13]. However, depth cameras capture only a projection of their surroundings, and any errors made during dataset curation are persisted and can affect learning.

Both Bat-G Net and BatVision use a convolutional neural network to learn to predict 3D entire scenes in a dense representation from a single multi-channel acoustic recording, such as the volumetric occupancy map of Bat-G Net [11] or the two-dimensional depthmaps in BatVision [13]. While neural networks are able to generalize when shown enough training examples, the use of dense representations means that models must learn to disentangle the myriad interactions between all points in the input audio and the output geometry. The challenges of gathering large datasets means that models should either be capable of meaningfully interpolating and extrapolating from the data they have been shown or else risk over-fitting, producing meaningless results on previously-unseen examples. Implicit function neural networks [15, 16] are an alternative to dense representations, but are able to over-fit to the coordinates they receive as inputs.

Due to the tension in learned acoustic reconstruction - the practical challenges of gathering a large dataset combined with the needs of existing neural networks for an enormous variety of training examples - we suggest that a different approach is needed. Although many input representations have been have been explored such as waveform audio [13] and spectrograms [13, 11], synthetic aperture focusing as used in existing acoustic imaging methods provides an input signal that is clearly rich in information but which has not yet been applied in learned acoustic reconstruction.

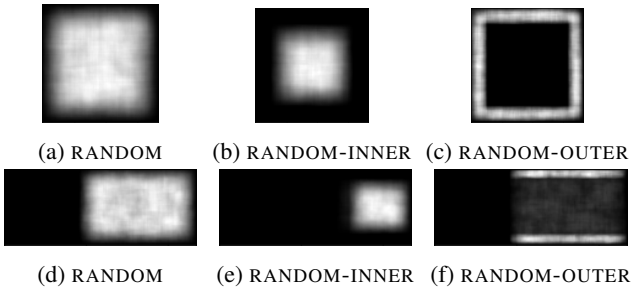
3. METHOD

We propose to use deep learning for acoustic reconstruction, with synthetic aperture focusing as an input pre-processing step and signed distance fields as a geometric representation. Rather than learning to predict entire scenes at once from whole audio recordings as with existing methods such as BatVisison [13] and Bat-G Net [11], our neural network functions implicitly on individual points in space, receiving cropped audio signals according to the expected round-trip time of a wave deflected at that point. This leaves our network with no spatial awareness other than what can be inferred from the time-aligned signals.

Unlike BatVision and Bat-G Net, we use an acoustic wave simulation to gather datasets and perform our experimentation. This enables greater experimental control and provides us access to both accurate and diverse geometric information.

3.1. Simulated datasets

We use the K-Wave acoustic simulation toolkit to generate large datasets of geometric obstacles and the echoes they produce [17]. Each example in our dataset is implemented as a



(a) RANDOM (b) RANDOM-INNER (c) RANDOM-OUTER

(d) RANDOM (e) RANDOM-INNER (f) RANDOM-OUTER

Fig. 1: Spatially-varying density of obstacles in our three datasets, as seen from the front in the top row and the side in the bottom row. Images were created by summing occupancy maps across each the entire dataset.

$177 \times 69 \times 69$ cm volume at a resolution of 7.5 mm per grid cell. We place a virtual emitter surrounded by 4 receivers in a planar arrangement spanning 34.5 cm in both directions centered near the low end of the volume. A $108 \times 69 \times 69$ cm region at the opposing end is reserved for placing obstacles.

We randomly place between 1 and 4 rectangular prisms and spheres in each example, with diameters varying from 2 to 20 cm. Grid locations inside obstacles are given the acoustic properties of wood, and all other locations are modeled as air at standard conditions. We simulate a linear FM chirp from 18 to 22 kHz at the emitter location and record the amplitude at each receiver at a 96 kHz sampling rate for 2048 samples. Due to the computational cost of each simulation, we do not consider higher frequencies. In this manner, we generate a total of 3 datasets, each with 5000 examples for training, 500 for validation, and 500 for testing. The first dataset, termed RANDOM, contains obstacles placed all throughout the experimental volume. The remaining two datasets, RANDOM-INNER and RANDOM-OUTER, are used for cross-validation purposes, and, restrict obstacles to lie in the inner and outer halves of the volume respectively, as seen from the receiver. These distributions are shown in Figure 1.

3.2. Synthetic aperture focusing

We perform synthetic focusing at a given sampling location as follows. Given the known emitter location and receiver locations, respectively, we first compute the linear distance from the emitter to the sampling location d^e, and the distance from the sampling location to the i-th receiver d_i^r. The sum of d^e and d_i^r gives us the round-trip distance, and from this we can compute the total expected time of flight Δt_i of a deflected wave for each receiver as $\Delta t_i = \frac{d^e + d_i^r}{c}$ where $c = 343$ m/s is the speed of sound in air. The per-receiver delay Δt_i is then used to align and window each received signal S_i, resulting in synthetically focused audio signals $\hat{S}_i$, defined as

$$\hat{S}_i(t) = k \left(d^e\right)^2 \left(d_i^r\right)^2 S_i \left(t + f_s \Delta t_i - \frac{W}{2}\right) \qquad (1)$$

where $t \in \{0, ..., W-1\}$, $k = 30$ dB is an experimentally-tuned gain constant, $(d^e)^2 (d_i^r)^2$ applies an amplitude compensation based on the distance traveled assuming spherical wave propagation to and from a small deflector, and $W = 256$ is the window length, in samples. The expected moment of arrival occurs in the middle of the focused audio.

3.3. Geometric representation

We use the signed distance to the nearest obstacle surface as our choice of geometric representation for training neural networks on synthetically focused audio. Signed distance fields provide a smoother representation for an implicit model compared to occupancy grids and depth maps, which jump discontinuously at object boundaries. Secondly, they are trivially convertible to binary occupancy fields using a threshold.

3.4. Neural network model

In our experiments, we use a simple convolutional neural network to map from synthetically-focused audio $\hat{S}$ to the scalar signed distance d at a point. The four-channel audio is passed through 3 convolutional layers, each with a kernel size of 31, a stride of 2, and 128 hidden features. The final layer has 32 output hidden features and is passed to a fully-connected layer with a single output neuron. We apply batch normalization to all layer inputs [18] and use the Leaky ReLU activation function with a negative slope of 0.1. In total, our model has only a fraction of the number of learnable parameters compared to others as shown in Table 1.

Model	Learnable Parameters
Bat-G Net	25585129
BatVision	71533697
Ours	652585

Table 1: Total number of learnable scalar parameters for each model used in our experiments.

To train our network, we sample training locations according to a custom distribution which emphasizes the surfaces and interiors of obstacles, unlike uniform random sampling which would mostly fall into empty space. Given the signed distance field at all grid locations $\mathrm{sdf}(i, j, k)$, we assign the relative weight w to each grid location as

$$w(i, j, k) = e^{(-r \times \max(\mathrm{sdf}(i,j,k) - d_{\min}, 0))}, \qquad (2)$$

where $r \approx 69.318\ \mathrm{m}^{-1}$ is a spatial decay rate causing a decrease of 50% every centimeter, and $d_{\min} = 2$ cm is the distance below which the weight is limited and remains constant. We then normalize all weights to sum to 1 across all grid locations and draw samples accordingly. The randomly selected grid locations are used to create focused audio and training labels. We train our network by minimizing the loss $\mathcal{L} = \mathbb{E}\left[\left|d - \hat{d}\right|\right]$ where d is the ground truth distance and $\hat{d}$ is the network output.

4. RESULTS

To evaluate how our implicit model compares to Bat-G Net and BatVision, we train all models for 72 hours on each of our datasets separately and measure their performance. Bat-G Net is trained on pairs of spectrograms, half favouring frequency resolution and half offering better temporal resolution, to produce unoccluded obstacle maps, and BatVision is trivially modified to accept 4 spectrograms rather than 2 to produce normalized depth maps in the range of 0 to 1.

To test against Bat-G Net which produces an occupancy map, we densely evaluate our network at every point in space, yielding a signed distance field, which we threshold to produce a compatible binary occupancy map. We then compute the F1 score and Intersection over Union (IOU) of each prediction against the ground truth. Similarly, to compare against BatVision which produces a depthmap and thus cannot represent occlusions, we fill in all occluded regions of our models's predicted occupancy grid, and perform an inverse projection on the depthmaps produced by BatVision in which all occluded locations are similarly occupied. Though we report the same metrics when comparing against Bat-G Net and BatVision, the use back-filled occupancy for BatVision means that the test scores have different interpretations and should not be used to compare Bat-G Net directly to BatVision, and so we state whether back-filling was used whenever giving results.

As a basic comparison, we train all models on the entire RANDOM training set containing 5000 examples and evaluate on the RANDOM test set. We then re-train each model on progressively smaller subsets of the training set before testing again on the full test set. We give the quantitative results of each model in Table 2 and show 3D renderings of sample predictions in Figure 2.

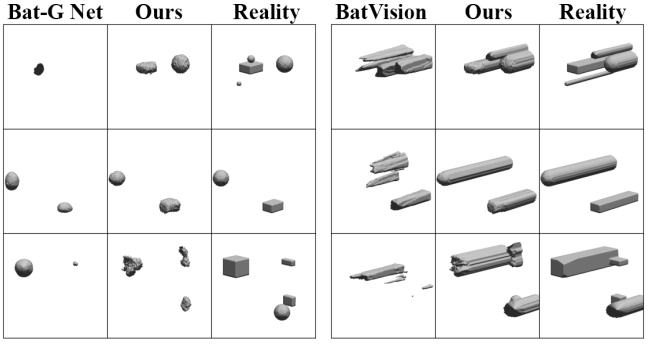


Fig. 2: Visualized model predictions on the RANDOM dataset.

Next, to measure how each model is able to extrapolate to locations outside those seen during training time, we re-train each model on the RANDOM-INNER training set and cross-evaluate on RANDOM-OUTER. Similary, to measure how each

Train on Subset of RANDOM, Test on RANDOM				
Examples	Model	Backfill	F1 Score ↑	IOU ↑
5000	Bat-G Net	No	0.4094	0.2913
	Ours	No	**0.6817**	**0.5506**
500	Bat-G Net	No	0.0317	0.0185
	Ours	No	**0.5603**	**0.4190**
50	Bat-G Net	No	0.0116	0.0063
	Ours	No	**0.3566**	**0.2405**
5	Bat-G Net	No	0.0067	0.0038
	Ours	No	**0.1109**	**0.0666**
5000	BatVision	Yes	0.2770	0.1883
	Ours	Yes	**0.7520**	**0.6289**
500	BatVision	Yes	0.0873	0.0512
	Ours	Yes	**0.6657**	**0.5253**
50	BatVision	Yes	0.0441	0.0245
	Ours	Yes	**0.4669**	**0.3353**
5	BatVision	Yes	0.0334	0.0179
	Ours	Yes	**0.1628**	**0.1018**

Table 2: Results on the full RANDOM test set after training each model on RANDOM training set and subsets thereof.

model is able to interpolate to obstacles in the same total spatial extent as the training data but in a region of space not seen during training, we train all models on RANDOM-OUTER before testing on RANDOM-INNER. The test results of each model between and across distrubtions are in Table 3.

5. DISCUSSION

Across all our experiments, we find that our method significantly outperforms both Bat-G Net and BatVision, both in terms of in-distribution performance, as well as when extrapolating and interpolating across nearby but disjoint spatial distributions. While all models suffer in performance when evaluating on datasets other than what they were trained on, Bat-G Net and BatVision both performed extremely poorly in our cross-validation experiments, failing to predict anything when cross-evaluated on our partitioned datasets. We attribute this to their dense representations, which by design give separate treatment to different spatial locations.

From our experiment on small datasets, we find that while performance in all cases decreases as may be expected when the training data is reduced, our model consistently outperforms Bat-G Net with $10\times$ fewer examples, and BatVision with $100\times$ fewer examples. We believe this to be due to our implicit representation, in which each echo and geometry training example effectively contains a continuum of examples, which in turn allows for improved sample efficiency.

Our synthetic aperture focusing clearly provides enough information to our implicit neural network for it to generalize well from relatively few training examples over larger networks using more data. The usefulness of this input signal is further demonstrated by the performance of our network despite its rather small size.

Model	Backfill	F1 Score ↑	IOU ↑
Train on RANDOM-INNER, test on RANDOM-INNER			
Bat-G Net	No	0.6129	0.4968
Ours	No	**0.7958**	**0.6880**
BatVision	Yes	0.6248	0.5071
Ours	Yes	**0.8421**	**0.7490**
Train on RANDOM-INNER, test on RANDOM-OUTER			
Bat-G Net	No	0.0000	0.0000
Ours	No	**0.3134**	**0.2363**
BatVision	Yes	0.0001	0.0000
Ours	Yes	**0.3644**	**0.2902**
Train on RANDOM-OUTER, test on RANDOM-OUTER			
Bat-G Net	No	0.2620	0.1810
Ours	No	**0.5863**	**0.4561**
BatVision, SG.	Yes	0.0000	0.0000
Ours	Yes	**0.6594**	**0.5288**
Train on RANDOM-OUTER, test on RANDOM-INNER			
Bat-G Net	No	0.0000	0.0000
Ours	No	**0.4635**	**0.3245**
BatVision, SG.	Yes	0.0000	0.0000
Ours	Yes	**0.6312**	**0.4841**

Table 3: Network test performance on datasets with both equal and opposite spatial distributions.

6. CONCLUSION

We have proposed a novel audio representation for learned acoustic reconstruction inspired by synthetic aperture techniques, and shown in simulation that this representation leads to far better performance and generalization than that of competing models. Our implicit formulation means that trained models can be made much smaller and require smaller training datasets relative to other existing networks which use dense representations and consider entire scenes at once. Because of its sample efficiency, our model can be trained using fewer examples which makes it applicable in the real world where dataset curation remains a significant challenge.

7. ACKNOWLEDGEMENTS

We are grateful for the generous hardware resources provided by the UBC ARC Sockeye high-performance computational cluster [19], without which this research would not have been possible.

8. REFERENCES

[1] Daniel Kish, "Human echolocation: How to "see" like a bat," *New Scientist*, vol. 202, no. 2703, pp. 31–33, 2009.

[2] Jørgen A. Jensen, Svetoslav I. Nikolov, Kim L. Gammelmark, and Morten H. Pedersen, "Synthetic aperture ultrasound imaging," *Ultrasonics*, vol. 44, pp. e5–e15, 2006.

[3] Svetoslav Nikolov and Joergen A. Jensen, "Comparison between different encoding schemes for synthetic aperture imaging," in *Medical Imaging 2002: Ultrasonic Imaging and Signal Processing*. 2002, vol. 4687, pp. 1–12, SPIE.

[4] Y. Tasinkevych, I. Trots, A. Nowicki, and P. A. Lewin, "Modified synthetic transmit aperture algorithm for ultrasound imaging," *Ultrasonics*, vol. 52, no. 2, pp. 333–342, 2012.

[5] R. Y. Chiao and Xiaohui Hao, "Coded excitation for diagnostic ultrasound: a system developer's perspective," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 52, no. 2, pp. 160–170, 2005.

[6] Cheng Chi, SpringerLINK ebooks Engineering, and SpringerLink (Online service), *Underwater Real-Time 3D Acoustical Imaging: Theory, Algorithm and System Design*, Springer Singapore, Singapore, 1st 2019. edition, 2019.

[7] A. Trucco, M. Palmese, and S. Repetto, "Devising an affordable sonar system for underwater 3-d vision," *IEEE transactions on instrumentation and measurement*, vol. 57, no. 10, pp. 2348–2354, 2008.

[8] Cheng Chi and Zhaohui Li, "High-resolution real-time underwater 3-d acoustical imaging through designing ultralarge ultrasparse ultra-wideband 2-d arrays," *IEEE transactions on instrumentation and measurement*, vol. 66, no. 10, pp. 2647–2657, 2017.

[9] David B Lindell, Gordon Wetzstein, and Vladlen Koltun, "Acoustic non-line-of-sight imaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6780–6789.

[10] Nathan Blaunstein, Vladimir Yakubov, and Taylor & Francis eBooks A-Z, *Electromagnetic and acoustic wave tomography: direct and inverse problems in practical applications*, CRC Press, Taylor & Francis Group, Boca Raton, 2019.

[11] Gunpil Hwang, Seohyeon Kim, and Hyeon-Min Bae, "Bat-g net: Bat-inspired high-resolution 3d image reconstruction using ultrasonic echoes," in *Advances in Neural Information Processing Systems*, 2019, pp. 3720–3731.

[12] Seohyeon Kim, Gunpil Hwang, and Hyeon-Min Bae, "Bat-g2 net: Bat-inspired graphical visualization network guided by radiated ultrasonic call," *IEEE access*, vol. 8, pp. 189673–189683, 2020.

[13] Jesper Haahr Christensen, Sascha Hornauer, and Stella X. Yu, "Batvision: Learning to see 3d spatial layout with two ears," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 1581–1587.

[14] Jesper H. Christensen, Sascha Hornauer, and Stella Yu, "Batvision with gcc-phat features for better sound to vision predictions," 2020.

[15] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European conference on computer vision*. Springer, 2020, pp. 405–421.

[16] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein, "Implicit neural representations with periodic activation functions," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 7462–7473, Curran Associates, Inc.

[17] Bradley E Treeby and Benjamin T Cox, "k-wave: Matlab toolbox for the simulation and reconstruction of photoacoustic wave fields," *Journal of biomedical optics*, vol. 15, no. 2, pp. 021314, 2010.

[18] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[19] UBC Advanced Research Computing, "Ubc arc sockeye," 2019.