FAST ACOUSTIC SCATTERING USING CONVOLUTIONAL NEURAL NETWORKS

Ziqi Fan[1]*, *Vibhav Vineet*[2], *Hannes Gamper*[2], *Nikunj Raghuvanshi*[2]

[1]University of Florida, Gainesville
[2]Microsoft Research, Redmond

ABSTRACT

Diffracted scattering and occlusion are important acoustic effects in interactive auralization and noise control applications, typically requiring expensive numerical simulation. We propose training a convolutional neural network to map from a convex scatterer's cross-section to a 2D slice of the resulting spatial loudness distribution. We show that employing a full-resolution residual network for the resulting image-to-image regression problem yields spatially detailed loudness fields with a root-mean-squared error of less than 1 dB, at over 100x speedup compared to full wave simulation.

Index Terms— Diffraction, occlusion, scattering, convolutional neural network, wave simulation

1. INTRODUCTION

Fast evaluation of wave scattering and occlusion from general object shapes is important for diverse applications, such as optimizing baffle shape in outdoor noise control [1,2], and real-time auralization in games and mixed reality [3,4]. Modeling diffraction is critical since acoustical wavelengths span everyday object sizes. Wave solvers capture diffraction and can achieve real-time execution in restricted cases [5, 6] but in general they remain quite expensive, even with hardware acceleration [7, 8]. While pre-computed wave simulation is viable for real-time auralization [4], it disallows arbitrary shape changes at run-time. Geometric (ray-based) approaches can handle dynamic geometry but diffraction remains challenging due to the inherent zero-wavelength approximation [9].

We propose a machine learning approach for fast modeling of diffracted occlusion and scattering. Previously, machine learning has been successfully applied in acoustic signal processing problems including speech synthesis [10, 11], source localization [12, 13], blind estimation of room acoustic parameters from reverberated speech [14, 15], binaural spatialization [16], and structural vibration [17]. Pèrez et al. [18] used a fully-connected neural network to learn the effect of re-configuring the furniture layout of a single room on acoustical parameters, including reverberation time (T_{60}) and sound pressure level (SPL), at a few listener locations.

Pulkki and Svensson [19] trained a small fully-connected neural network to learn exterior scattering from rectangular plates as predicted by the Biot-Tolstoy-Medwin (BTM) diffraction model [20]. The input was a carefully designed low-dimensional representation of the geometric configuration of source, plate, and listener based on knowledge of diffraction physics. The output was a set of parameters of low-order digital filters meant to auralize the effect. The authors report plausible auralization of scattering effects despite some inaccuracies. However, due to relying on a hand-crafted low-dimensional parameterization, the method is not designed to generalize beyond rectangular plates.

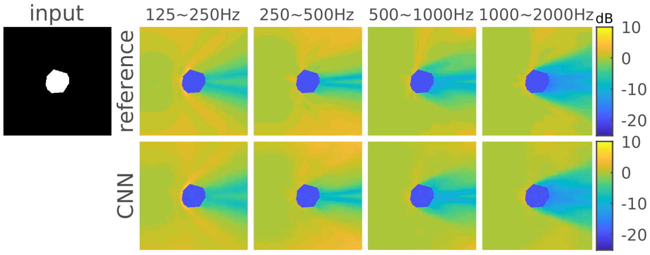

Fig. 1: Acoustic scattering formulated as 2D image-to-image regression. Input object shape is specified as a binary image (left). A point source, not shown, is placed to the left of the object. Numerical wave simulation is used to produce reference scattered loudness fields in frequency bands (top row). Our CNN produces a close approximation at over 100× speedup (bottom row).

In this paper, we report the first study on whether a neural network can effectively learn the mapping from a large class of shapes (convex prisms) to the resulting frequency-dependent loudness field, as illustrated in Fig. 1. We restrict the problem to convex shapes to rule out reverberation and resonance effects in this initial study. In contrast to [19], our goal is to design a neural network that generalizes well for a variety of input shapes by formulating the problem as high-dimensional image-to-image regression which allows application of state-of-the-art convolutional neural networks (CNNs) that have been successfully applied in computer vision [21–23].

We design a CNN that ingests convex prism geometries represented by their 2D cross-sections discretized onto binary occupancy grids. The predicted outputs are the corresponding loudness fields in octave bands along a horizontal slice passing through the source, represented as floating point images in decibels (dB). Our input–output mapping of acoustic scattering in terms of a spatial grid reveals spatial coherence, such as the smooth change in loudness across the geometric shadow edge. CNNs are particularly well-adapted to such tasks. Further, using CNNs allows us to train a single network unlike [19], where occluded and unoccluded cases had to be treated separately with distinct networks.

Experimental results and generalization tests indicate that the proposed neural network model is surprisingly effective at capturing detailed spatial variations in diffracted scattering and occlusion (e.g., compare top vs. bottom row in Fig. 1). Relative to wave simulated reference, the RMS error is below 1dB while providing over 100x speedup, with evaluation time of about 50ms on a high-end GPU. To foster further research, we have shared our complete dataset at: https://github.com/microsoft/AcousticScatteringData.

[*] Work done as research intern at Microsoft Research, Redmond

2. PROBLEM FORMULATION

2.1. Acoustic loudness fields

Consider the exterior acoustics problem of an object insonified with a point source at location $\mathbf{x}_0 = (x_0, y_0, z_0)$ emitting a Dirac impulse. Object shape can be abstractly described with an indicator function, $\mathcal{O}(\mathbf{x}) = \{0, 1\}$, where 1 indicates the object is present at a 3D spatial location $\mathbf{x}$, and 0 indicates otherwise. Scattering from the object results in a time-varying pressure field denoted by $G(\mathbf{x}, t; \mathbf{x}_0)$ termed the Green's function, which evaluates the pressure at any point $\mathbf{x} = (x, y, z)$ at time t. Semi-colon denotes parameters to be held fixed; in this case the source location, $\mathbf{x}_0$. The Green's function must satisfy the scalar wave equation,

$$\left[\partial_t^2 - c^2\nabla^2\right] G(\mathbf{x}, t; \mathbf{x}_0) = \delta(\mathbf{x} - \mathbf{x}_0, t), \quad (1)$$

where $c = 343$ m/s is the speed of sound and $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$ is the Laplacian operator, subject to the impedance boundary condition on the object's surface based on its material, and the Sommerfeld radiation condition at infinity [24].

Analytical solutions to (1) are unavailable beyond simple geometries such as a sphere or an infinite wedge [24]. Therefore, numerical solvers must be employed that perform computationally expensive sub-wavelength sampling of space. For applications such as modeling dynamic occlusion in virtual reality or optimizing baffle shape in noise control, the energetic properties of G are of particular interest, obtained by measuring its loudness in frequency bands. Therefore, the focus of our study is the sensitive, non-linear effect of object shape on the scattered loudness field.

Formally, denoting the temporal Fourier transform as $\mathcal{F}$, we define the Green's function in the frequency domain for angular frequency, ω: $\hat{G}(\mathbf{x}, \omega; \mathbf{x}_0) \equiv \mathcal{F}\left[G\left(\mathbf{x}, t; \mathbf{x}_0\right)\right]$ and define octave-band loudness fields as

$$L_i(\mathbf{x}; \mathbf{x}_0) \equiv 10 \log_{10} \frac{\|\mathbf{x} - \mathbf{x}_0\|^2}{\omega_{i+1} - \omega_i} \int_{\omega_i}^{\omega_{i+1}} |\hat{G}(\mathbf{x}, \omega; \mathbf{x}_0)|^2 d\omega, \quad (2)$$

where $i \in \{1, 2, 3, 4\}$ denotes the index of four octave bands $[\omega_i, \omega_{i+1})$, $\omega_i \equiv 2\pi \times 125 \times 2^{i-1}$ rad/s, which together span the frequency range of [125, 2000] Hz. The factor $\|\mathbf{x} - \mathbf{x}_0\|^2$ normalizes L_i for free-space distance attenuation, so that in the absence of any geometry, $L_i(\mathbf{x}; \mathbf{x}_0) = 0$. That is, all loudness fields are 0 dB everywhere in the absence of a scatterer and they capture the *perturbation* on free-space energy distribution induced by the presence of object geometry, which is often the primary quantity of interest. Distance attenuation can be easily included later via a compensating factor of $1/\|\mathbf{x} - \mathbf{x}_0\|^2$.

2.2. Scattering functional

From (1), the loudness fields L_i depend both on the object geometry $\mathcal{O}$ and source location $\mathbf{x}_0$. We observe that the latter can be restricted to the negative x-axis, simplifying the formulation, as the D'Alembert operator, $\left[\partial_t^2 - c^2\nabla^2\right]$ is invariant to the choice of frame of reference [24]. Thus, given any $\mathbf{x}_0$ in one frame of reference with origin at object center, one can find a unique coordinate system rotation $\mathcal{R}$ such that $\mathcal{R}(\mathbf{x}_0)$ lies on the negative x-axis in the new coordinate system. The object must also be rotated so that $\mathcal{R}(\mathcal{O})$ and evaluations of the loudness fields similarly transformed to the rotated system. Therefore, the source can be restricted to the negative x-axis without any loss of generality because we are approximating scattering from *arbitrary* convex shapes and rotation preserves convexity.

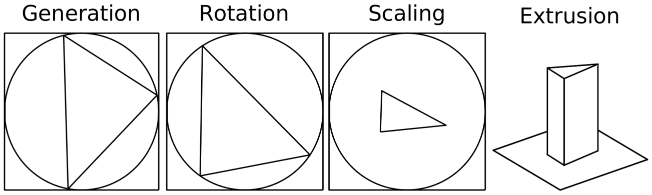

Fig. 2: Random object generating process. A polygon with 3-20 vertices is randomly generated by sampling angles on a circle, then rotated, scaled, and extruded in height to yield a convex prism object.

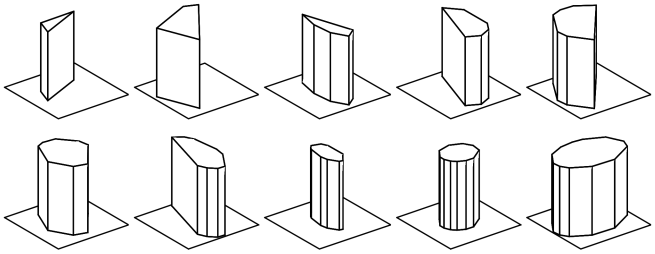

Fig. 3: Examples objects in our training dataset.

The remaining free parameter for the source is its radial distance to object center. In this initial study, distance is assumed to be fixed. This simplification allows dropping the dependence on $\mathbf{x}_0$ entirely. The problem can then be formalized as computing the *scattering functional*, $\mathcal{S} : \mathcal{O} \mapsto \{L_i\}$ which takes object shape as input and outputs a set of loudness fields in frequency bands. The functional is typically evaluated using a numerical solver for (1) coupled with an encoder that implements (2), such as in the "Triton" system [4] that we employ as baseline. The underlying solver has been validated in outdoor scenes [25]. Here we investigate whether neural networks may be used to provide a substantially faster approximation of $\mathcal{S}$.

2.3. Acoustic scattering as image-to-image regression

In order to learn $\mathcal{S}$ successfully using a neural network, the choice of discrete representation for input $\mathcal{O}$, output L_i, and neural network architecture are critical inter-dependent considerations. We observe that shapes and loudness fields exhibit joint spatial coherence, containing smoothly varying regions, occasionally interrupted by abrupt changes such as near the object's edges, or near the geometric shadow boundary. Convolutional neural networks (CNNs) have been used extensively in the computer vision community for signals with such piece-wise smooth characteristics, motivating our current investigation. However, CNNs typically work on images represented as 2D grids of values. Therefore, we cast our input–output representation to 2D by restricting our shapes to convex prisms that have a uniform cross-section in height, i.e., along the z-axis, and training the neural network to map from this 2D convex cross-section to a 2D slice of the resulting 3D loudness fields. The simulation setup is shown in Fig. 4 and detailed in Section 3.2. Thus, the task is simplified to that of image-to-image regression, as illustrated in Fig. 1. The input is a binary image specifying presence of object, $\mathcal{O}$, at each pixel, and output is a multi-channel image with four channels corresponding to the four octave-band loudness fields L_i.

3. DATA GENERATION

The data generation consists of generating random convex-prism input shapes and computing the corresponding output loudness fields.

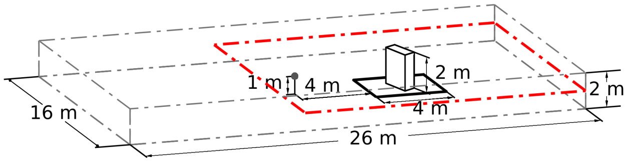

Fig. 4: A convex-prism object is insonified with a point source marked with gray dot. Simulation is performed inside a containing cuboidal region. Object and loudness field data is extracted on a 2D slice shown with dashed red square. Dimensions not to scale.

3.1. Input shape generation

The generation of random convex prisms is illustrated in Fig. 2. Given a target number of vertices, N, of the convex cross-section, the angles $\theta_i, i = [1, \cdots, \mathrm{N}]$ are drawn randomly from $[0, 2\pi]$ and then sorted. The ordered set of points $(x, y) = (2\cos\theta_i, 2\sin\theta_i)$ describes a convex polygon with all its vertices on the inscribed circle of a 4×4 m^2 *object region*. A random rotation in $[0, 2\pi)$ is performed about the origin, followed by scaling in x and y independently with scaling factors drawn randomly from $[0.25, 1]$. Finally, the rotated and scaled convex polygon is extruded along the z-axis to obtain a convex prism. All random numbers are drawn from the uniform distribution. The procedure results in objects with significant cross-section diversity, see Fig. 3. For each cross-section vertex count, $\mathrm{N} \in [3, 20]$, K random convex prisms are generated, where $\mathrm{K}_{\mathrm{tr}} = 6000$ for the training set, $\mathrm{K}_{\mathrm{cv}} = 60$ for the validation set, $\mathrm{K}_{\mathrm{te}} = 20$ for the test set, resulting in a total of 108 000, 1080 and 360 samples for training, validation and test, respectively.

3.2. Output loudness field generation

For each convex prism object, we compute the corresponding output loudness fields using the Triton system [4] that employs the fast ARD pseudo-spectral wave solver [26] to solve (1) combined with a streaming encoder to evaluate (2). The scattering object resides in a $4 \times 4 \times 2$ m^3 object region. The center of this object region is the origin of our coordinate system. We assume a nearly-rigid and frequency-independent acoustic impedance corresponding to Concrete material, with pressure reflectivity of 0.95. High reflectivity is chosen to ensure there is substantial reflection from the object.

The simulation is performed on a larger $26 \times 16 \times 2$ m^3 cuboidal region of space, as illustrated in Fig. 4, with perfectly matched layers absorbing any wavefronts exiting this region. A point sound source is placed on the negative x-axis at $(-6, 0, 0)$ m. The solver is configured for a usable bandwidth up to 2000 Hz, resulting in an update rate of 11 765 Hz. The solver executes on a discrete spatial grid with uniform spacing of 6.375 cm in all dimensions.

For extracting the input-output data for training purposes our region of interest is the 16×16 m^2 2D slice that symmetrically contains the object region, with corners $(-8, -8, 0)$ to $(+8, +8, 0)$ m, shown with red square in Fig. 4. The solver already discretizes the object and fields onto a 3D spatial grid for simulation purposes, so we merely extract the relevant samples lying on our 2D slice of interest from the 3D arrays, without requiring any interpolation. The extracted 2D arrays are then padded to 256×256 pixel images. This results in a pair of an input binary image for an object and an output set of four loudness fields, constituting one entry in our dataset.

We ensure the training and test sets are disjoint by exhaustively checking that none of the object binary images in the test set have an exact match in the training set. Dataset generation was run in parallel for all shapes on a high-performance cluster, taking 3 days. Each entry took 4 minutes for simulation and encoding, excluding task preparation time. An example is shown in Fig. 1, top row.

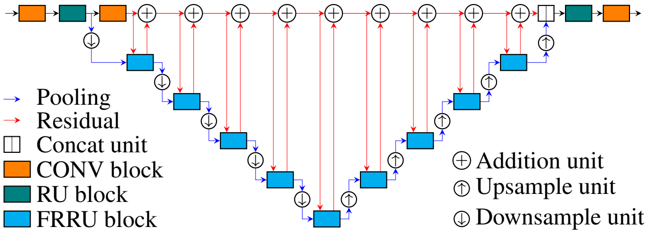

Fig. 5: Full-resolution residual network (FRRN), adapted from [27].

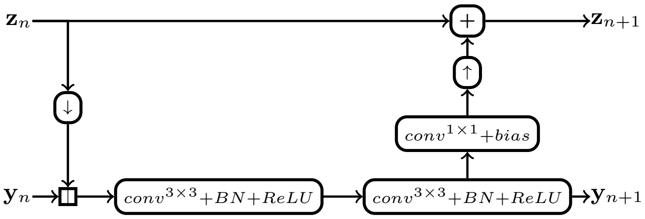

Fig. 6: Full-resolution residual unit (FRRU), adapted from [27].

4. FULL-RESOLUTION RESIDUAL NETWORK (FRRN)

We adopt the full-resolution residual network (FRRN) [27] to model the scattering functional defined in Section 2.2 using the training data generated in the previous section. As shown in Fig. 5, an FRRN is composed of two basic streams: a pooling stream and a residual stream. In general, data abstraction with multiple resolutions in the pooling stream enables the FRRN to integrate both fine local details and general transitions of loudness fields. The residual stream of full resolution ensures that loudness fields are output at the input spatial resolution and that backpropagation converges faster [21, 28].

The core component of an FRRN is the full-resolution residual unit (FRRU), shown in Fig. 6. There are 27 FRRUs in our FRRN (3 in each FRRU block in Fig. 5), which is the depth of our neural network. In each FRRU, the full-resolution input residual stream $\mathbf{z}_n$ is down-sampled to the same resolution as the input pooling stream $\mathbf{y}_n$ and is then concatenated to $\mathbf{y}_n$. The concatenation is fed into two consecutive convolutional units to generate the output pooling stream $\mathbf{y}_{n+1}$, which serves as the input pooling stream of the next FRRU. Further, the stream $\mathbf{y}_{n+1}$ propagates into another convolutional unit and is upsampled to the same resolution as $\mathbf{z}_n$. The upsampled stream is added back to $\mathbf{z}_n$ to form the output residual stream $\mathbf{z}_{n+1}$, which is subsequently added back to the main stream of full resolution. Such bidirectional downsampling and upsampling of features between the residual and pooling streams allows to learn features at successive layers of FRRN at different spatial resolutions.

5. EXPERIMENTAL EVALUATION AND DISCUSSION

We employ the source code of FRRN provided by [29] in our study. Since the FRRN from [29] was originally designed for classifying pixels of images into multiple categories and modeling the scattering functional is a regression problem, we modified the source code and selected the mean squared error (MSE) as our loss function. We also modified the implementation so that the input and output of the neural network are respectively one-channel and four-channel 256×256 images, indicated by Fig. 1. We set the batch size as 8 and adopted a stochastic-gradient-descent (SGD) optimizer, a learning

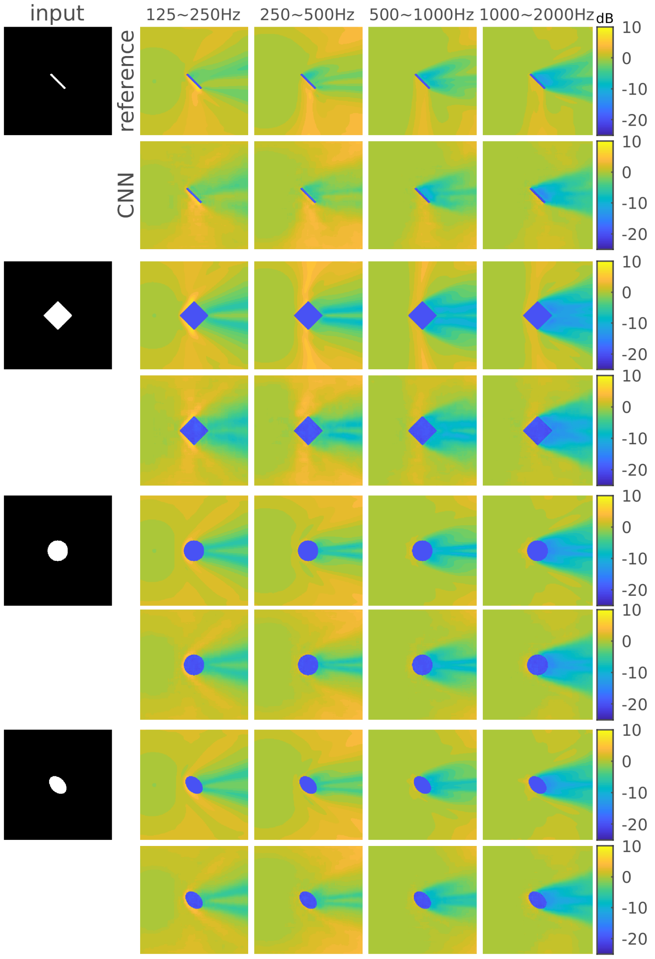

Fig. 7: Generalization tests comparing reference vs CNN prediction. CNN is able to model detailed scattering and occlusion variations.

rate of 1.0e−4, a momentum of 0.99 and a weight decay of 5.0e−4. The FRRN was trained on 108 000 examples for 50 000 iterations on a Tesla P100 GPU. Evaluating the CNN after training takes about 50 ms. The wave simulation takes 4 minutes on a multi-core CPU and can be accelerated by 10× if also performed on the GPU [30]. Adjusting for hardware differences, then our method is 100-1000× faster.

To test the generalization capability of our model, we created four prisms extruded from a bar, square, circle and ellipse, all of which cause the network to extrapolate beyond the randomly-generated training set. The CNN provides a surprisingly good reproduction of the spatial loudness variation, as shown in Fig. 7. Notice the reflected lobe from the bar prism (top row), which shows scattered energy propagating downwards. In comparison, reflection from the square prism (second row) is symmetric about the x-axis. The CNN successfully predicts these different acoustic features, even though it has only learned from random polygons. The CNN does introduce a degree of spatial smoothing on the scattered lobes, a trend we observe consistently. Also note the brightening at low frequencies at edges facing the source. This is due to constructive interference between incident and reflected signals. The CNN is also able to capture diffracted shadowing behind the object in all cases, along with smooth variation around the geometric shadow that gets more abrupt as frequency increases. Our results indicate that learning spatial *fields* of perceptual acoustic quantities is quite advantageous compared to learning acoustic responses at a few points, since fields provide the network with extensive information about the constraints on spatial variation imposed by wave physics.

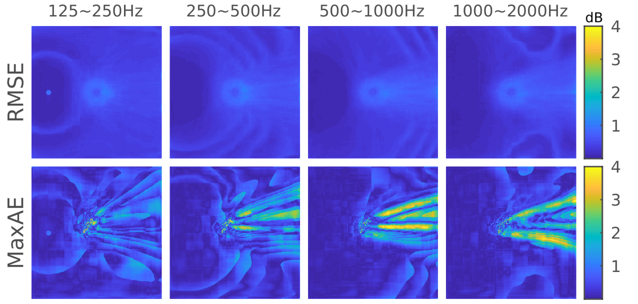

Fig. 8: Root-mean-squared error (RMSE) and maximum absolute error (MaxAE) computed over 360 test cases for each frequency band.

As a statistical test of accuracy, we fed the 360 objects in the test set into the trained network and evaluated the root-mean-squared errors (RMSE) and the maximum absolute errors (MaxAE) on all pixels in all frequency bands against the reference simulated results. These are shown in Fig. 8. The RMS errors are below 1 dB for all frequency bands. MaxAE provides a more detailed look at errors within particular test cases. At each pixel it shows the largest absolute error over all 360 test cases. As illustrated in Fig. 8, the errors are concentrated in the occluded region behind the object. This phenomenon can be explained as follows. Observe in Fig. 7 that our CNN is able to successfully predict the spatially-detailed light and dark streaks in the occluded region. These streaks are interference fringes due to diffracted wave-fronts wrapping around the object and meeting behind. Fringes oscillate faster in space for smaller wavelengths so that slight displacements in the fringes can cause large per-pixel errors due to subtracting two oscillatory functions with a small relative translation. This explanation fits the observation that MaxAE has a worsening trend with increasing frequency. Even so, our pessimistic MaxAE estimate is of the order of 4 dB, which, while larger than the best-case just-noticeable-difference of 1 dB, is sufficient for plausible auralization with spatially smooth effects.

6. CONCLUSION AND OUTLOOK

We investigated the application of convolutional neural networks (CNNs) to the problem of acoustic scattering from arbitrary convex prism shapes. By formulating the problem as 2D image-to-image regression and employing full-resolution residual networks we show that surprisingly detailed predictions can be obtained. Generalization tests indicate that the network hasn't just memorized. Network evaluation is over 100× faster than direct simulation. Our results suggest that CNNs are a promising avenue for the tough problem of fast acoustic occlusion and scattering, meriting further study.

This initial study had several restrictions: convex prism shapes only, fixed object material, fixed source distance, and training on 2D slices. Our formulation is designed so it generalizes beyond these restrictions. A natural extension of the current approach could be to employ 3D CNNs [31] for handling arbitrary shapes and corresponding 3D loudness fields. The limitation of fixed source distance could be addressed by providing an additional floating point input to the neural network that parameterizes the input–output mapping. We intend to pursue such extensions in future work, and hope our results and dataset foster parallel investigations in this exciting direction.

7. REFERENCES

[1] D. A. Bies, C. Hansen, and C. Howard, *Engineering noise control*. CRC press, 2017.

[2] L. L. Beranek and I. L. Ver, "Noise and vibration control engineering-principles and applications," *Noise and vibration control engineering-Principles and applications John Wiley & Sons, Inc., 814 p.*, 1992.

[3] M. Vorländer, *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality (RWTHedition)*, 1st ed. Springer, Nov. 2007.

[4] N. Raghuvanshi and J. Snyder, "Parametric directional coding for precomputed sound propagation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 108:1–108:14, July 2018.

[5] L. Savioja, "Real-Time 3D Finite-Difference Time-Domain Simulation of Mid-Frequency Room Acoustics," in *13th International Conference on Digital Audio Effects*, Sept. 2010.

[6] A. Allen and N. Raghuvanshi, "Aerophones in Flatland: Interactive Wave Simulation of Wind Instruments," *ACM Trans. Graph.*, vol. 34, no. 4, July 2015.

[7] Z. Fan, T. Arce, C. Lu, K. Zhang, T. W. Wu, and K. McMullen, "Computation of head-related transfer functions using graphics processing units and a pereptual validation of the computed HRTFs against measured HRTFs," in *Proc. Conf. Audio Eng. Soc.*, Aug 2019.

[8] N. Raghuvanshi, B. Lloyd, N. Govindaraju, and M. C. Lin, "Efficient numerical acoustic simulation on graphics processors using adaptive rectangular decomposition," in *Proceedings of the EAA Symposium on Auralization*. European Acoustics Association, June 2009.

[9] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 708–730, Aug. 2015.

[10] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7962–7966.

[11] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4460–4464.

[12] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 74–79.

[13] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2386–2390.

[14] J. Eaton, N. D. Gaubitch, H. Moore, Alastair, and P. A. Naylor, "Estimation of room acoustic parameters: The ace challenge," *IEEE Trans. Audio, Speech, Language Processing*, vol. 24, no. 10, pp. 1681–1693, 2016.

[15] A. F. Genovese, H. Gamper, V. Pulkki, N. Raghuvanshi, and I. J. Tashev, "Blind room volume estimation from single-channel noisy speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, May 2019, pp. 231–235.

[16] R. A. Tenenbaum, F. O. Taminato, and V. S. Melo, "Room acoustics modeling using a hybrid method with fast auralization with artificial neural network techniques," in *Proc. International Congress on Acoustics (ICA)*, 2019, pp. 6420–6427.

[17] D. E. Tsokaktsidis, T. V. Wysocki, F. Gauterin, and S. Marburg, "Artificial neural network predicts noise transfer as a function of excitation and geometry," in *Proc. International Congress on Acoustics (ICA)*, 2019, pp. 4392–4396.

[18] R. F. Pérez, "Machine-learning-based estimation of room acoustic parameters," Master's thesis, Aalto University, School of Electrical Engineering, 2018.

[19] V. Pulkki and U. P. Svensson, "Machine-learning-based estimation and rendering of scattering in virtual reality," *J. Acoust. Soc. Am.*, vol. 145, no. 4, pp. 2664–2676, 2019.

[20] U. P. Svensson, R. I. Fred, and J. Vanderkooy, "An analytic secondary source model of edge diffraction impulse responses," *J. Acoust. Soc. Am.*, vol. 106, no. 5, pp. 2331–2344, Nov. 1999.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *ICCV*, 2017, pp. 2980–2988.

[24] A. D. Pierce, *Acoustics: An Introduction to Its Physical Principles and Applications*. Acoustical Society of America, 1989.

[25] R. Mehra, N. Raghuvanshi, A. Chandak, D. G. Albert, D. Keith Wilson, and D. Manocha, "Acoustic pulse propagation in an urban environment using a three-dimensional numerical simulation," *The Journal of the Acoustical Society of America*, vol. 135, no. 6, pp. 3231–3242, 2014.

[26] N. Raghuvanshi, R. Narain, and M. C. Lin, "Efficient and Accurate Sound Propagation Using Adaptive Rectangular Decomposition," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 5, pp. 789–801, 2009.

[27] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 4151–4160.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. European Conf. Computer Vision (ECCV)*. Springer, 2016, pp. 630–645.

[29] "Semantic segmentation architectures implemented in pytorch," https://github.com/meetshah1995/pytorch-semseg/, 2017, accessed: June, 2019.

[30] R. Mehra, N. Raghuvanshi, L. Savioja, M. C. Lin, and D. Manocha, "An efficient GPU-based time domain solver for the acoustic wave equation," *Applied Acoustics*, vol. 73, no. 2, pp. 83–94, Feb. 2012.

[31] A. Dai, A. X. Chang, M. Savva, M. Halber, T. A. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2432–2443.